# Statistical Analysis with The General Linear Model[1]

**Jeff Miller and Patricia Haden[2]**


**Version: February 14, 2013**



[2]We thank Profs. Wolf Schwarz and Rolf Ulrich for helpful comments and suggestions. Author contact address is Prof. Jeff Miller, Department of Psychology, University of Otago, Dunedin, New Zealand, miller@psy.otago.ac.nz. If you use this textbook, we would be delighted to hear about it.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Overview

## 1.1 The General Linear Model

This course is about a large and complex set of statistical methods tied together by a unifying conceptual framework known as "The General Linear Model" (GLM). This model can be used to answer an amazing variety of research questions within an infinite number of different experimental designs. Basically, the GLM can be used to test almost any hypothesis about a dependent variable (DV) that is measured *numerically* (e.g., height, income, IQ, age, time needed to run a 100-yard dash, grade point average, etc.; but not categorical DVs like eye color, sex, etc.).

In a first statistics course, students will have seen some special cases of the GLM without knowing it. For example, the various kinds of "t-tests" (one-sample, between-subjects, within-subjects, etc.) are special cases of the GLM. So are correlation, regression, and the Analysis of Variance (ANOVA). We will not assume any prior knowledge of these techniques, but students should realize, if the new methods seem familiar, that they are now seeing the big picture.

There are literally an infinite number of experimental designs and survey designs that can be analysed using the GLM. Naturally, it is not possible to teach students the possibilities on a case by case basis. Students must learn to use the GLM as an adaptable tool: how to apply it to a research design unlike any they have ever seen before. This requires an understanding of the technique well beyond the kind of pattern-recognition by which students in introductory statistics often learn to apply t-tests and the like.

Study of the GLM also teaches a lot about how to design experiments and surveys. The model and techniques of analysis highlight the issues that are most critical for drawing conclusions from data. Knowing these issues in advance will help us arrange to collect the data so that it will be maximally informative.

The techniques covered in this course are extremely common in psychology, sociology, education, and business. Students will find an understanding of the GLM to be useful whether they are conducting and analyzing research projects of their own or critically reading reports of research done by others. Computer programs to do these sorts of analyses are also available almost everywhere, so the necessary calculations can be performed easily and conveniently even on very large data sets.

For teaching purposes, we will break the GLM into three parts corresponding to its three main techniques:

1. Analysis of Variance (ANOVA).
2. Simple and Multiple Regression.
3. Analysis of Covariance (ANCOVA or ANOCOVA or ANOCVA).

Though the three techniques are closely related, they are designed to achieve different goals, as illustrated below. We will compare these goals in the context of an example about a medical researcher who wants to understand what influences blood cholesterol level.

### 1.1.1 GLM: ANOVA

Suppose a medical researcher was studying blood cholesterol level (a numerical DV) to see what influenced it. ANOVA would be used if the researcher wanted to compare average cholesterol level between different genders or different ethnic groups. In fact, ANOVA could be used to analyze the

effects of both gender and ethnic group at once, as in analyzing results similar to those shown in Figure 1.1.

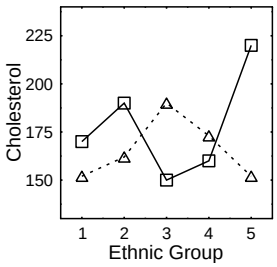


Figure 1.1: Average blood cholesterol levels (shown on the vertical axis) of males and females within each of five ethnic groups representing North America (NA), South America (SA), Europe (Eur), Asia, and the Pacific islands (Pac).

In ANOVA terminology, gender and ethnic group are called *independent variables* or *factors*. These terms will be discussed further in a subsequent section on terminology, but basically they refer to any categorical variable that defines the averages to be compared.

This is a good point to emphasize one aspect of statistical background that you are already supposed to have. Why do we need any statistical technique at all to analyze the results in this graph? Why can't we just interpret the averages that we see plotted before our eyes? The answer is that we need to take into account the possibility that the results are due to random influences (chance, random error, sampling error, etc.). In other words, we need statistics to rule out the possibility that the real population averages look totally different from what we have observed in our sample (e.g., they are really all equal). This concept should already be familiar: The main purpose of inferential statistics is to decide whether a certain pattern of results might have been obtained by chance (due to random error). All of the techniques available under the GLM have this as their basic goal, so it is well to keep this idea in mind.

Returning to the example, ANOVA will help us evaluate whether there are real (or just accidental) differences in average blood cholesterol between the different ethnic groups and between the two sexes. Furthermore, we can evaluate whether the difference between sexes is the same for all ethnic groups, or vice versa.

### 1.1.2 GLM: Regression

The medical researcher might also be interested in how blood cholesterol level is related to some numerical predictors like age, amount of exercise per week, weight, amount of fat in diet, amount of alcohol consumed, amount smoked, etc. The major difference from the previous case is that now we do not have distinct groups of people (e.g., European females), but instead each person may have a different value on the numerical predictor (e.g., weight). Logically, predictor variables are similar to factors or IVs, except that they take on numerical rather than categorical values.

Typically, data of this form are represented in a *scattergram* as shown in Figure 1.2.

This type of data is analyzed with *Regression*, which allows a researcher to see how the DV is related to each predictor variable (in this case age). *Simple Regression* relates the DV to one predictor variable, and it is a formal version of what we can do by eye when looking at the scattergram. *Multiple Regression* relates the DV to many predictor variables at once, with no theoretical limit on the number of predictor variables. Maybe you can visualize the case of two predictor variables by imagining a third axis coming out from the page (e.g., labelled "amount smoked"). But try to visualize the case of 100 predictor variables!

Again, the fundamental issue is to what extent the results may be attributable to chance. In looking at the above scattergram, we are tempted to say that blood cholesterol increases with age. Is this a real effect in the whole population, or could it just be the case that we found this pattern in

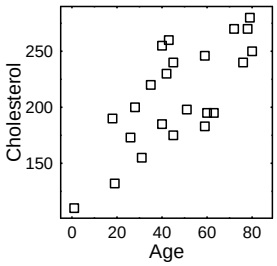


Figure 1.2: Scattergram showing the relationship between age and blood cholesterol level for a sample of 23 individuals, each represented by one square on the graph.

the sample by accident? This is the most basic question that the GLM can answer for us.[1]

### 1.1.3 GLM: ANCOVA

This very advanced topic is a combination of ANOVA and regression. The easiest way to think of it is that we want to look for predictive relationships (regression) that differ between groups (ANOVA). For example, suppose that the scattergram relating cholesterol to age used different symbols for males vs. females. Then, we could check whether the age/cholesterol relationship was or was not the same for the two sexes. The data might look like those in Figure 1.3.


Figure 1.3: Scattergram showing the relationship between age and blood cholesterol level separately for males and females.

After examining the above data, we might want to say that cholesterol level increases faster with age for males than for females. Whether or not this is really true or just a chance finding is something that ANCOVA will answer for us. But the important point is the comparison we are examining with the technique: Is the relationship between two variables quantitatively the same for two different groups? ANCOVA can also handle more than two groups and/or more than two variables in the predictive relationship.

[1] Some students notice that another approach to the analysis in the regression situation is to form groups of people by dividing the numerical predictor into separate ranges, and putting each person into one range. For example, we might classify people according to their ages being below 30, 31-60, and above 60. This approach lets us treat the data as having three distinct groups, so that we could use ANOVA just as we did with the categorical variables of sex and ethnic group. There are a few advantages, but mostly disadvantages, to using this approach instead of regression. After we have seen how both ANOVA and regression work, we will consider these advantages and disadvantages.

In summary, the three techniques of the GLM are used to achieve different goals, as follows:

**ANOVA:** Find out how a numerical DV is related to one or more categorical IVs.

**REGRESSION:** Find out how a numerical DV is related to one or more numerical predictor variables.

**ANCOVA:** Find out how a numerical DV is related to categorical IVs and numerical predictor variables at the same time.

## 1.2 Learning About the GLM

It should be clear from the above introduction that the GLM is a very large and complex topic. Most students' previous exposure to statistics probably consisted of a series of relatively straight-forward hypothesis testing techniques, such as the sign test, $t$-tests, tests involving the correlation between two variables, etc. In studying these hypothesis testing techniques, the typical strategy is to learn to recognize a single research question and data set for which the procedure is appropriate, to learn the appropriate formula, and to learn which numbers get plugged in where. To some extent, then, what students learn in beginning statistics courses may be described as a list of recipes for making certain specific statistical dishes.

Study of the GLM is quite different, simply because the technique is so flexible. Once a student understands the technique, he or she can analyze almost any data set involving numerical dependent variables. This book presents a unified picture, fitting together all three components of the GLM. In the end, we hope you will see a large but coherent picture that allows you to test all kinds of different hypotheses in the same general way. You might even be well-advised to forget what you already know about ANOVA or regression; if you try to hold on to a previous, narrow approach, it may well interfere with your seeing the big picture.

There are quite a few equations in this book, but the book is nevertheless intended for a practical, researcher's course—not a math course. In looking at the equations, we will emphasize the practical concepts behind the symbols rather than the pure math of the situation. This approach will minimize the required math background and maximize the contact with experimental design, interpretation of results, and other real-world considerations. It turns out that the equations and symbols are a very convenient conceptual notation as well as mathematical notation, so there is considerable practical value in looking at them. They help not only in learning what the computed numbers mean, but also in developing your understanding of the GLM to the point that you can apply it to new experimental designs never covered in class.

Of course, the equations also show how to do the necessary computations. Some of these computations are not difficult—the ones needed for ANOVA require only addition, subtraction, multiplication, and division. Be warned though: the equations presented here are not the fastest ones to use when doing computations (not even close), and for that reason they may look quite different from those found in other statistics courses or texts. The equations presented here were chosen to illustrate the concepts rather than to facilitate the computations, for two reasons. First, with computers getting faster, better, cheaper, easier to use, and more available, no one with a data set of any size will have to analyze it by hand. Second, because it is so easy to get fancy statistics by computer, it is all the more important to understand what the statistics mean. Therefore, we will leave all the tricky computational formulas to be used in computer programs. As humans, we will use formulas that require a few more arithmetic operations, but that are also more directly meaningful conceptually.

There are both advantages and disadvantages to studying a big, interrelated conceptual system. On the plus side is the general applicability of these methods, already mentioned. Also, because the material is so coherent, once it fits together for you the parts all tend to reinforce one another, making it easier to remember. The disadvantages arise from the massive amount of material. In your first statistics course you probably studied hypothesis testing procedures that could be covered in a single lecture, or a week at the most. This is not at all true of the GLM and its parts. One could easily spend a year studying ANOVA in a graduate course, and another year studying regression and ANCOVA. So, we are trying to pack the essential concepts from two years of graduate material into a book designed for a single 10-13 week course. Naturally, we will have to make a number of simplifications and leave out many details (e.g., all the computational formulas), but there are still many new concepts to be integrated. A related problem is that the topic is very tough to get started on. You have to master quite a few of the new concepts before they begin to fit together and you see some of the overall pattern. There is also a lot of terminology to be learned, not to mention a number

of computational strategies to be mastered. Often, students will think that they are hopelessly lost up through the middle of the section on ANOVA. But stick with it. Students very often report that they get a big "AHA" reaction after about 3-4 weeks of study.

## 1.3 Scientific Research and Alternative Explanations

The statistical techniques described in this course are tools for scientific research, so it is appropriate to include a few comments on the function and form of science, as well as the role of statistics in it. The function of science is to increase our understanding of objects, beings, and processes in the world, and the form is to collect and analyze various kinds of facts about what happens under various kinds of conditions. There are major limitations, however, because we can neither examine all possible conditions nor gather all possible facts. The best that researchers can do is focus their efforts on subsets of facts and conditions that seem likely (based on previous work and on scientific intuition) to show a meaningful pattern.

Beyond gathering facts, however, scientists are supposed to *explain* them. An explanation is usually a description that includes a causal chain between the circumstances and the observations. Given the explanation, one can usually make a prediction about what the observations ought to look like under some particular new, unexamined circumstances. If such a prediction can be made, and if subsequent research shows that the prediction is correct, then the explanation gains plausibility. On the other hand, theories have to be revised when predictions are disconfirmed.

For example, Isaac Newton studied a set of physical phenomena and came up with the hypothesis known as the theory of gravity. (Both hypotheses and theories are types of explanations. A hypothesis is a tentative explanation, and it gets promoted to the status of a theory after it makes a lot of correct predictions with only a few minor errors). Based on his hypothesis, he predicted that all objects should fall from a given height in the same time (neglecting friction of the air). The prediction was tested with the best methods available at the time (i.e. the tallest buildings and the most accurate watches), and the hypothesis gained acceptance because the results of experimental testing verified the prediction.

There is a significant logical problem with this strategy of confirming predictions, unfortunately. When a hypothesis makes a prediction that is confirmed, we tend to believe more strongly in the hypothesis. But no amount of confirmed predictions can ever *prove* the hypothesis. This is because there may be another hypothesis that makes all the same predictions– perhaps one that nobody has thought of yet. This other hypothesis would be called an *alternative explanation*, and it might be the correct one.

Such was exactly the case with Newton's hypothesis, although nobody thought of another hypothesis that made the same predictions until Einstein came along many years later. Einstein's relativity theory made essentially the same predictions as Newton's theory with respect to experiments that had already been performed. However, it also suggested some new experiments in which the two theories made different predictions, and relativity turned out to be right when these experiments were conducted. And the same thing may very well happen to relativity, if someone figures out a better hypothesis someday. Even if no one ever does, we will never really be certain that relativity is correct, because there is no way to be sure that someone won't come up with a better hypothesis tomorrow.

The upshot of all this, contrary to popular belief about science, is that scientists can never prove their theories absolutely. Though we can be certain of the results of particular experiments, we can never be certain of the explanations of those results. The truth may always be some alternative explanation that no one has yet thought of.

## 1.4 The Role of Inferential Statistics in Science

Statistics (particularly *inferential* statistics) are used to rule out the following alternative explanation of the facts: "The facts are wrong." More specifically, the alternative explanation is that the results were inaccurate because of random error.

There are almost always sources of random error in an experiment. For example, Newton probably could only measure times to about the nearest half-second, so any particular measurement could be off a random amount up to a half-second. Perhaps his average measurements, which agreed so well with his theory, were all wrong by a crucial half-second that was enough to disprove his theory. Or maybe all the different half-second errors cancelled out so the averages were in fact correct.

Inferential statistics are used to decide questions like this, by evaluating the possibility that a certain pattern of results was obtained by chance, due to random error. Using inferential statistics, researchers try to show that it is very unlikely that their results were obtained by chance, so that the alternative explanation of chance results is too remote to take seriously.

It is important to be able to rule out the alternative explanation that the results were obtained by chance (i.e., that the random influences in the experiment didn't happen to cancel out properly), because this explanation always has to be considered. In fact, this alternative explanation is a possibility in every experiment with random influences. Thus, anyone who doesn't want to accept a theory can always offer "chance" as an alternative explanation of the results supporting it. Inferential statistics are used to counter this particular alternative explanation.

# Part I

# Analysis of Variance

# Chapter 2

# Introduction to ANOVA

It may be difficult to remember that ANOVA is "Part 1" of something, because it is such a substantial topic in itself. But it is.

ANOVA is a tremendously general type of statistical analysis, useful in both very simple and very complicated experimental designs. Basically, ANOVA is used to find out how the average value of a numerical variable—called the *dependent variable* or *DV*—varies across a set of conditions that have all been tested within the same experiment. The various conditions being compared in the experiment are defined in terms of one or more categorical variables—called the *independent variables*, the *IVs*, or the *factors*. In short, ANOVA is used to find out how the average value of a numerical DV depends on one or more IVs.

The complexity and usefulness of ANOVA both come from the flexibility with which an experimenter can incorporate many IVs or factors into the statistical analysis. ANOVA is especially informative in research where many factors influence the dependent variable, because there is no limit to the number of distinct effects that ANOVA can reveal. Of course, before one can incorporate many factors into the statistical analysis, one must first incorporate them into the experimental design. It turns out that there is a lot of terminology regarding such designs, and we must identify these terms before we can begin to study ANOVA, so that we can discuss the designs properly.

## 2.1 Terminology for ANOVA Designs

It is useful to work with an example. Suppose we want to study the question of whether drinking a cup of coffee increases perceptual alertness. The dependent variable is perceptual alertness, and we want to see how it is influenced by the factor of coffee consumption. Assuming we can agree on a valid measure of perceptual alertness, it seems possible to answer this question by taking a random sample of people, giving coffee to half the sample and no coffee to the other half, and then testing everybody for perceptual alertness. The results from the two groups (coffee vs. no coffee) could be compared using a simple statistical method like a two-group t-test, and we would have data relevant to our empirical question. There would be no particular need to use ANOVA (ANOVA can be used in two-group designs, but the t-test is computationally simpler and equally informative).

While simplicity is surely a virtue, oversimplicity is not. For someone truly interested in the effects of coffee on alertness, it would be reasonable to consider a variety of other factors that might influence the results. For example, one might expect very different results for people who are regular coffee drinkers as opposed to people who never drink coffee. A cup of coffee might well give a lift to a regular coffee drinker, but it might just as easily give a stomach ache to someone who never drinks coffee except in our experiment. Thus, perceptual alertness might depend jointly on two factors: coffee consumption in the experiment, and personal history of coffee consumption prior to the experiment. If both factors have effects, the above two-group design is too simple and may lead us to some bad overgeneralizations (e.g., "if you want to make someone more alert, give them a cup of coffee"). It would be better to use a design with four groups: regular coffee drinkers given a cup of coffee, non-coffee drinkers given a cup of coffee, regular coffee drinkers given no coffee, and non-coffee drinkers given no coffee. This four-group design would allow us to see how drinking a cup of coffee affected alertness of the regular coffee drinkers separately from the non-coffee drinkers, and it would prevent us from making the erroneous overgeneralization noted above.

Of course, one could argue that the four-group design is still oversimplified, because there are still other factors that might have an important influence on perceptual alertness. For example, the effect

of a cup of coffee probably depends on the time of day at which people are tested for alertness (at least for regular coffee drinkers). If we tested regular coffee drinkers at 8:00 a.m. without providing them with coffee, we would probably find an average alertness around the level of a geranium. Regular coffee drinkers would be much more alert for the 8:00 a.m. test if they were first given a cup of coffee. If we tested in the afternoon, however, alertness of the regular coffee drinkers might be less dependent on the cup of coffee.

One could carry on with such arguments for quite a while, since there are surely many factors that determine how coffee influences alertness. The point of the example is simply that it is often reasonable to design experiments in which effects of several factors can be examined at the same time. ANOVA is a set of general purpose techniques that can analyse any experiment of this type, regardless of the number of factors. Simpler techniques, such as the t-test, would be inadequate once we got beyond two groups.

To keep the example workable we will stop with the design incorporating the three factors discussed above: 1) coffee or no coffee given before testing, 2) regular coffee drinker or non-coffee drinker, and 3) testing at 8:00 a.m. or 3:00 p.m.

In ANOVA terminology, the above coffee and alertness experiment is said to be a design with three independent variables or factors. Technically, any categorical variable that defines one of the comparisons we want to make is a factor.[1]

Each factor is said to have *levels*. The levels of the factor are the specific conditions of that factor compared in a particular experiment. Often, we state a factor just by describing its levels; for example, we might say that one factor is "whether or not a person is a regular coffee drinker". In other instances we use a name for the factor, apart from its levels; for example, the "time of testing" factor has two levels: 8:00 a.m. and 3:00 p.m.

ANOVA techniques require that we organize the analysis around the number of factors in the design and the number of levels of each factor. It is convenient and conventional, then, to describe a design by listing the number of levels of each factor. The above design would be described as a "2 by 2 by 2" or "2x2x2" design. Each number in the list corresponds to one factor, and it indicates the number of levels of that factor. The numbers corresponding to the different factors can be listed in any order. Thus, a "4 by 3" or "4x3" design would be a design in which there were two factors, one of which had four levels (e.g., college class = freshman, sophomore, junior, or senior) and the other of which had three levels (e.g., male, female, or undecided). Sometimes the description is abbreviated by giving only the number of factors, as in "one-factor design", "two-factor design", etc. In ANOVA, the term "way" is often substituted for "factor" in this context, so you will also see "one-way design" and "one-factor design" used interchangeably, as are "two-way design" and "two-factor design", etc.

The experimental design described above is called a *factorial design*, as are most of the designs described in this course. In a factorial design, the experimenter tests all of the possible combinations of levels of all the factors in the experiment. For example, one factor is whether or not a person is given coffee before being tested, and another factor is whether a person is a regular coffee drinker or does not normally drink coffee. To say that this is a factorial design simply means that the experimenter tested all four possible combinations of the two levels on each factor. If the experimenter for some reason decided not to test the regular coffee drinkers in the no-coffee condition, then it would not be a factorial design. Another phrase used to convey the same idea is that "factors are crossed". Two factors are said to be "crossed" with one another when the experimenter tests all of the possible combinations of levels on the two factors. This design is *fully crossed*, since all three factors are crossed with each other. Thus, a fully crossed design and a factorial design mean the same thing.

Designs are also said to have *cells*, each of which is a particular set of conditions tested (i.e., combination of levels across all factors). One cell in our example design, for instance, is the condition in which regular coffee drinkers are tested at 8:00 a.m. without being given a cup of coffee. In all there are eight cells in the design, as diagrammed in Table 2.1. The number of cells in a design is always equal to the product of the number of levels on all the factors (e.g., 8 = 2 times 2 times 2).

Observations or measurements taken in particular cells are generally referred to as *replications* in that cell. If we tested 10 people in each cell, we would say that we had "10 replications per cell". The number of replications per cell is also called the *cell size*, so we would also say that we had a "cell size of 10". Some students may wonder what happens if we have different cell sizes (different numbers of

[1] Usually, the term "factor" is used to refer to independent variables of interest to the experimenter. Levels on these factors have been selected specifically to reflect certain comparisons, and often the whole point of the experiment was to find out how the DV varied across these levels. Later, however, we will see examples of factors involving comparisons that we are not particularly interested in. Thus, our good general definition of a factor is in terms of a variable which enables us to make a comparison, even if we are not interested in it.

| | | Regular Coffee Drinkers | | | Coffee Nondrinkers | |
|---|---|---|---|---|---|---|
| | | Coffee | No Coffee | | Coffee | No Coffee |
| Time of | 8:00 | | | 8:00 | | |
| Testing | 3:00 | | | 3:00 | | |

Table 2.1: Cells in the Three-Factor Coffee/Alertness Design

replications per cell). This complicates matters seriously with respect to the computations done by ANOVA, and we will defer that topic until late in the book. For now, then, we restrict our attention to *equal cell-size* ANOVA.

The term *subject* refers to the randomly sampled experimental unit. In this example, the subjects are the individuals we randomly choose to test for perceptual alertness. The subjects in an experiment need not always be people. For example, if we wanted to find out whether high schools in California or New York had higher student/teacher ratios, we might randomly sample some high schools from each state. The high schools would then be the subjects. We would measure the DV (student/teacher ratio) for each subject, and we would compare averages across the two levels of the IV (state: CA vs. NY).

It is important to distinguish between two types of factors: *between-subjects* and *within-subjects*, because the ANOVA is calculated slightly differently for these two types of factors. A between-subjects factor is one for which different subjects are tested at the different levels of the factor. A within-subjects factor is one for which the same subjects are tested at all the different levels of the factor (such factors are also called *repeated-measures* factors). For example, if we tested each individual subject at either 8:00 a.m. or 3:00 p.m. but not both, then time of day would be a between-subjects factor. On the other hand, if we tested each individual subject at both 8:00 a.m. and 3:00 p.m., then time of day would be a within-subjects factor. Of course, many factors can only be tested between-subjects, not within subjects. For example, we must test gender as a between-subjects factor, with different individuals tested as males than as females, unless our subjects are willing to undergo sex change operations halfway through the experiment.

The phrases "between-subjects design" and "within-subjects design" (also called "repeated-measures design") refer to designs in which the factors are *all* either between-subjects or within-subjects, respectively. A *mixed design* is one in which some factors are between-subjects and some are within-subjects.

## 2.2 Summary of ANOVA Terminology

**Empirical Question:** Specific question to be answered in an experiment.

**Population:** The set of all individuals about whom/which the question is asked.

**Sample:** The set of specific individuals tested in the experiment.

**Subject (S):** One individual in the sample.

**Dependent Variable (DV):** The variable that is measured on each subject.

**Independent Variable (IV):** A variable defining different conditions or groups for comparison.

**Factor:** Same as Independent Variable.

**Levels:** The specific conditions or groups being compared under an IV.

**Conditions:** Treatments given to subjects.

**Groups:** Sets of subjects.

**Experimental Design:** The set of conditions or groups included in an experiment.

**Factorial Design:** An experimental design with two or more factors in which all possible combinations of factor levels are tested.

**Cell:** One particular combination of levels.

**Number of Replications:** Number of data values obtained in each cell.

**Cell Size:** Same as Number of Replications.

**Between-subjects Factor:** An experimental factor for which each subject is tested at only one level.

**Within-subjects Factor:** An experimental factor for which each subject is tested at all levels.

**Repeated-measures Factor:** Same as Within-subjects Factor.

**Between-subjects design:** A design in which all factors are between-Ss.

**Within-subjects design:** A design in which all factors are within-Ss.

**Mixed design:** A design with at least one between-Ss factor and at least one within-Ss factor.

## 2.3 Conclusions from ANOVA

In most cases, ANOVA is used to find out whether the average value of the DV differs across the categories being compared. The standard practice in research using ANOVA is to say that the IV "affects" or "has an effect on" the DV whenever the average value of the DV differs depending on the IV. This would seem quite reasonable in terms of the coffee experiment, for example. If we found higher perceptual alertness after drinking coffee than after drinking no coffee, we would probably feel comfortable concluding that coffee has an effect on perceptual alertness.

The terminology is questionable, however, because a causal relationship is not necessarily implied by such findings. For example, suppose we find that students who flunk out of college have lower SAT scores than those who do not. Obviously, we cannot conclude that college success (the IV) *causes* changes in SAT scores (the DV), because SAT score is measured before students even go to college. Nonetheless, using standard ANOVA terminology one might refer to the difference in average SAT scores by saying that there was an "effect" of college success on SAT. Be very careful to remember that this is not a causal statement, but rather a short-hand terminology for indicating that there is a difference between averages in different categories.

## 2.4 Overview of the GLM as used in ANOVA

Having dealt with the basic terminology, let us return to the problem of statistical analysis. As discussed above, the GLM provides a framework for analysis of factorial experiments. This framework is valuable because it accomplishes five functions:

1. Represents structure of experiment.
2. Measures how much effect each factor has by itself.
3. Measures interactive effects of several factors working together.
4. Measures amount of sampling error.
5. Tests hypotheses based on preceding measurements.

First we will see how the GLM represents the structure of an experiment (Section 2.5). Then, we will see how measurement and hypothesis testing work. For that task, we will start with the simplest design, in which there is a single factor (Chapter 3).

## 2.5 How the GLM Represents the Structure of an Experiment

There are various aspects of an experiment that must be taken into account by the statistical analysis, and the General Linear Model represents them all. Table 2.2 shows what these aspects are, and shows what component of the GLM represents each one. The aspects are discussed one by one below.

The GLM integrates together the effects shown in Table 2.2 with a very simple scheme: addition. In essence, the model is based on the idea that each value of the DV is determined by a sum of the various effects, as shown in the equations in Table 2.3.

| Aspect of Experiment | Component of GLM | Notation |
|---|---|---|
| Numerical Dependent Variable | Data Values ("Scores") | $Y$ |
| Common Influences in Experiment | Baseline Value or "Overall mean" | $\mu$ |
| Effects of Experimental Factors | "Main Effects" | $A$, $B$, $C$ ... |
| Interactive Effects Between Factors | "Interactions" | $AB$, $BC$, $AC$ ... |
| Random Error | Random Error | $S(Group)$ |

Table 2.2: How the GLM represents various aspects of an experiment

| Score | = | Overall mean | + | Factor Effects | + | Interactions | + | Error |
|---|---|---|---|---|---|---|---|---|
| $Y$ | = | $\mu$ | + | $A + B + \ldots$ | + | $AB + AC + \ldots$ | + | $S(Group)$ |

Table 2.3: Equations Representing Data Values in terms of the Structure of the Experiment

This is not quite a complete notation for the model, because we will need subscripts on most of the symbols. For example, the subscripts on the Y values are used to differentiate among the different data values (e.g., which subject it came from, which group that subject was in). The subscripts will be discussed later.

The first aspect to be represented is the numerical DV, which will be present in any experiment on which we use ANOVA. The model has a component for the DV, called the *data values* or *scores*, and the symbol $Y$ is used to stand for this component.

The second aspect of the experiment is the set of common conditions that influence all the scores. For example, if we measure perceptual alertness, then one sort of thing that is common across all the data values is our method of measurement. We measure alertness in certain units with a certain test or measuring device, in a room with a certain constant amount of light and noise, and we probably give certain standard instructions to all our subjects. These common aspects influence the numbers we obtain (e.g., they may determine whether the DV ranges from 0 to 100 vs. 100 to 200), and they should have the same influence on all the scores. The GLM represents these common aspects with a component that has a common influence on all of the scores. This component is a baseline or overall mean level of scores, symbolized by "$\mu$" and pronounced "mu". It is useful to think of all the scores as starting from the baseline, but deviating from it due to effects of experimental factors and random error.

We also have various conditions in our experiment, as determined by the factors we choose to include in the study. The model makes allowance for each factor (i.e., represents it) by allowing it to have an effect on the obtained data scores. For example, the $A$ term could be used to represent the overall effect of coffee on perceptual alertness. The $B$ term could represent the effect of time of day, and so on.

In an experiment with more than one factor, another possible aspect of the experiment is interaction effects between factors. For instance, in the coffee/alertness example discussed above, we considered the possibility that there might be synergistic effects for particular combinations of coffee and type of person (e.g., giving coffee to someone who doesn't ordinarily drink it could produce a bad reaction—a negative interaction effect). To allow for such effects, the model includes components called *interactions*. These interactions are caused by a particular combination of factor levels, and these are represented in the model with terms like $AB$, $AC$, $BC$ (two-way interactions), $ABC$, $ABD$, $BCD$ (three-way interactions), and so on. Don't worry if the concept of an interaction is not entirely clear at this point. We will study it in more detail when we get to two-factor designs.

Finally, random error is an aspect of the experiment and a component of the model. We might imagine many sorts of random error in an experiment—some might be contributed by random fluctuations in the device we use to measure the DV on each subject, some are contributed by random sampling (i.e., which particular subjects we select for testing from the population), and so on. It turns out that we will eventually learn to distinguish between several different types of random error, but for right now we will lump them all together. The notation for this composite error emphasizes the random sampling component: "S(group)" is intended to symbolize the random influence associated with sampling each individual subject within a group.

As shown in Equation 2.3, the GLM represents the different aspects of the experiment in an additive fashion. That is, each data value is represented as a sum of various components, corresponding to

a baseline, plus the effects of individual factors, plus the interaction effects between combinations of factors, plus error. The structure of the experiment is thus represented by which components are summed to make up each score.

We will use this model by 1) estimating the values of each component from the data, and 2) testing hypotheses about which components have effects that are too large to be due to random error. We will see how these two steps are carried out for a large variety of designs.

ANOVA will be covered with the following organization. First, we will consider between-subject designs. We will start with one-factor, between-subject designs, and then proceed to two- and three-factor between-subject designs, eventually formulating general rules for between-subject designs with any number of factors. Second, we will consider within-subject designs. We will start with one-factor within-subject designs, and proceed to two- and three-factor designs, eventually generalizing to any number of factors. Third, we will consider mixed designs. We will start with designs containing one between-subjects factor and one within-subjects factor, and then add more factors of each type.

# Chapter 3

# One-Factor, Between-Subject Designs

This chapter introduces the computational concepts underlying ANOVA. The eventual goal, as described earlier, is to come up with a set of hypothesis testing procedures that can be used in factorial designs having any number of experimental factors. We will start, naturally, with one of the simplest cases, in which there is only one experimental factor. Even this simple case, however, illustrates the majority of the concepts of ANOVA.

In this section, we will use as an example the following experiment in education. Suppose we want to compare three methods for teaching children how to spell. We randomly select three groups of children, teach one group with each of the three methods, and give them all a standard spelling test at the end. We have the experimental hypothesis that the three methods are not all equally good, so the null hypothesis is that the methods are equally good. If we call $\mu_1$, $\mu_2$, and $\mu_3$ the true average test scores for people learning spelling by Method 1, Method 2, and Method 3, respectively, then we could state our null hypothesis as:

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

A one-factor ("one-way") ANOVA can be used to test this hypothesis.

In the following sections we will present many ideas intended to justify the calculations necessary for a one-way ANOVA. Because of all this theoretical discussion, the actual computations will be spread out over many pages. Therefore, the final subsection of this chapter presents a step-by-step summary of the computational procedure for a one-way ANOVA.

## 3.1 The "Variance" in Analysis of Variance

An introduction to ANOVA must begin with a discussion of the concept of variance. Most students will recognize from previous exposure to descriptive statistics that the variance of a sample (the square of the standard deviation) is computed as:

$$s^2 = \frac{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}{N}$$

(Depending on the use, sometimes the divisor is $N - 1$. Don't worry about this.) This quantity is important as a measurement of how different the various scores in a sample are from each other, or how much dispersion there is among the scores.

It is worth focusing attention briefly on one aspect of the formula for variance that will play a particularly important role in ANOVA: The formula is mainly concerned with the quantities:

$$Y_i - \bar{Y} = \text{Deviation of the } i^{\text{th}} \text{ score from the mean}$$

If these deviations are large, the variance will be large. If these deviations are small, the variance will be small. The deviations, then, are really the operative quantities in the formula for variance. It is useful to think of the deviation as a distance from a score to a baseline, the baseline being another name for the average value in the set of scores.

How do we use variance, and in what sense do we analyze it? We use variance in two ways. One is to measure variability *within groups*. The variance of the scores in a group tells how much spread

or dispersion there is around the group mean. This is a measurement of random error. All of the individuals in a given group are considered equivalent by the experimenter because they receive the same treatment, so they should all receive the same score, except for random error. For instance, in the experiment on spelling methods, the experimenter has no way to account for different scores within a single group except to attribute them to random error.[1]

The second use of variability is to measure the size of differences *between groups*. To do this, we will pretend that the observed group means are the actual data scores, and we will compute a variance based on these group means. Suppose, for example, that we observed group means of 16, 15, and 17. The variance of these three numbers is 0.67. Though it is perhaps a new idea to compute a variance based on group means, there is no reason why we cannot plug the three numbers into the variance formula and compute the result. Suppose instead that the observed group means had been 24, 15, and 9. The variance of these three numbers is 38. Thus, the larger the differences among the group means, the larger the variance of the group means.

The between-group variance is important because it measures the extent to which the averages contradict the null hypothesis. If the between-group variance is large, then the group means are quite different, and that suggests that $H_0$ is false. If the between group variance is small, then the group means are similar, and that suggests that $H_0$ is not false. Thus, the bigger the between-group variance, the more evidence we have against $H_0$.

Many students will already have studied a closely related statistical test called the $t$-test. This test is used to compare the means for two groups, in order to test the null hypothesis that the two group means are equal. The $t$-test is based on subtracting one mean from the other, thereby getting a difference between the two observed group means to be used as a measure of how inconsistent the data are with the null hypothesis (larger difference implies more inconsistent). This measure works fine for an experiment with two groups, but it is impossible to extend to experiments with more than two groups. With three groups, there is no single subtraction that tells how much all three observed means differ from one another. As shown above, though, the variance of group means can be used to measure the extent to which the group means differ from one another. Thus, the variance is a more general way to measure differences than is a simple subtraction, because it works well regardless of the number of groups.

To summarize, there are two kinds of variance that we look at in experimental designs with one experimental factor: within-group and between-group.[2] To illustrate the differences between these two types of variance, Table 3.1 shows two illustrative data sets that might have been obtained in the study of different methods for teaching spelling. Both data sets have three groups with four subjects per group. Considering all 12 numbers as a whole, there is exactly the same variance in data set A as in Data Set B. In fact, the 12 numbers are the same in the two data sets, except they have been rearranged across groups.

With respect to the null hypothesis of no difference between groups, the important difference between the two data sets is in how much of this variance is within-group variance and how much is between-group variance. Data Set A has relatively small differences within groups, and relatively large differences between the group averages. Data Set B is just the reverse, having relatively large differences within groups and relatively small differences between the group averages.

We can represent this situation graphically as shown in Figure 3.1, which shows the data values for the two sets plotted on a number line. The Group 1 data points are shown as *'s, the Group 2 data points are shown as o's, and the Group 3 data points are shown as +'s. The plot for Data Set A illustrates that the variance between groups is large relative to the variance within groups, because the scores from the three groups form three more-or-less distinct clusters with almost no overlap. In Data Set B, however, the within-group differences are much larger and there are no obvious differences between groups. The total spread of all scores (i.e., total variance) is the same for the two data sets, but in Data Set A most of the variance results from differences between groups while in Data Set B most of the variance results from differences within groups.

Again, the crucial question is: Do we have strong evidence against the null hypothesis that the three groups are equal on the average? A moment's reflection should convince you that Data Set A provides stronger evidence against $H_0$ than does Data Set B. All we need now is a formal model that can quantify these intuitions.

[1] One might argue that the scores are really different because there are important factors that the experimenter has not taken into account, but are not really random error. Philosophically, that is probably right, but from a practical point of view, we simply use the term "random error" to refer to effects of all factors we have not taken into account. Thus, it is definitional to say that the scores within a group are only different because of random error.

[2] There is an unfortunate similarity of this terminology to that of within-subjects factors and between-subjects factors. The ideas are quite unrelated and should not be confused.

Data Set A:

| Method 1 | | Method 2 | | Method 3 | |
|---|---|---|---|---|---|
| Subject | $Y_{ij}$ | Subject | $Y_{ij}$ | Subject | $Y_{ij}$ |
| 1 | 124 | 1 | 101 | 1 | 76 |
| 2 | 129 | 2 | 88 | 2 | 91 |
| 3 | 115 | 3 | 107 | 3 | 84 |
| 4 | 112 | 4 | 92 | 4 | 81 |
| Average | 120 | Average | 97 | Average | 83 |

Overall mean = 100

Data Set B:

| Method 1 | | Method 2 | | Method 3 | |
|---|---|---|---|---|---|
| Subject | $Y_{ij}$ | Subject | $Y_{ij}$ | Subject | $Y_{ij}$ |
| 1 | 124 | 1 | 101 | 1 | 76 |
| 2 | 88 | 2 | 129 | 2 | 91 |
| 3 | 84 | 3 | 107 | 3 | 115 |
| 4 | 112 | 4 | 81 | 4 | 92 |
| Average | 102.0 | Average | 104.5 | Average | 93.5 |

Overall mean = 100

Important: There are different subjects in the three groups. For example, subject 2 is a different person in Method 1 than Method 2.

Table 3.1: Sample Data Sets for Spelling Experiment

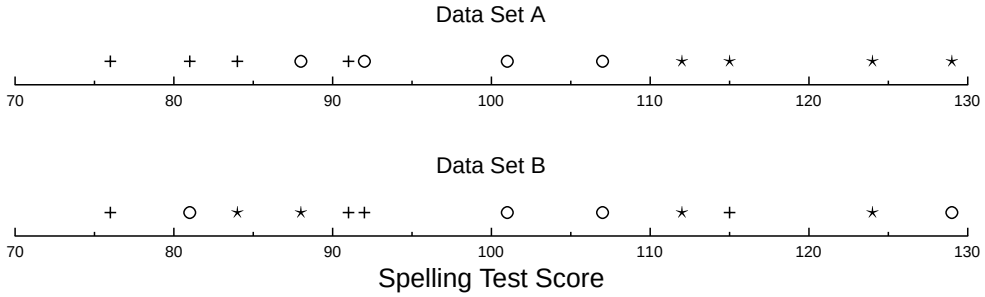


Figure 3.1: Plot of Individual Student's Scores on the Spelling Test (each point shows the score for one student).

## 3.2 Measuring Between- and Within-Group Variance

The GLM provides a method of measuring between- and within-group variance that generalizes easily to experiments with more than one experimental factor. In the next few subsections, we will illustrate this method using the example data in Data Set A. In this section, we will see how to measure the effects of the experimental factor and random error on each individual data value. In a later section, we will see how to compute a composite measure of within- and between-group variability across the whole data set. These will be the measures of within- and between-group variance that we can use to test $H_0$: No effect of methods.

Consider the form of the GLM for the spelling-method experiment. For this one factor design, the exact GLM is:

$$Y_{ij} = \mu + A_i + S(A)_{ij} \quad \text{for } i = 1 \ldots 3 \text{ and } j = 1 \ldots 4$$

This model provides one equation for each data value, and the subscripts are used to separate the different equations, as discussed below.

As discussed before, the $Y$ term stands for the scores or data values, but now subscripts $i$ and $j$ have been added. The subscript $i$ is used to indicate the group from which each score came (Spelling Method 1, 2, or 3), and the subscript $j$ is used to indicate which subject within each group contributed

that score. For example, $Y_{23}$ stands for the score for subject 3 in spelling method 2 (which happens to be 107 in both of the data sets in Table 3.1).

$\mu$ is the baseline. One way to think of this value is as the true average score, averaging across all of the populations and/or conditions tested in the experiment (i.e., what the true average score really should have been, except for random error, across this whole experiment). There is only one true average value for any whole experiment, so this term doesn't need a subscript.

$A$ is the effect of the experimental factor (method for teaching spelling). There are three different methods, and each one might have a different effect. Thus, we put the subscript $i$ on $A$, so that the model allows for a different quantitative effect for each method.

$S(A)$ is the error term. We use $S$ to stand for the random contribution of each individual subject, and the $(A)$ is read as "within A." This notation is designed to emphasize the fact that there are different subjects at each level of Factor A, so that for example Subject 1 in $A_1$ is a different person from Subject 1 in $A_2$. Now, each score is influenced by error (e.g., due to random sampling of the subject who gave that score), so we must allow for a different error contribution to each score. To do this, we use the subscripts $i$ and $j$ on the error term, just as we did on the original scores ($Y$'s). Thus, $S(A)_{23}$ is the contribution of random error to that score of 107. This contribution is conceptually distinct from the contribution of the baseline and the effect of Method 2.

Because of all the different subscripts, the GLM really provides 12 different equations for Data Set A—one equation for each data point. The twelve different equations are shown in Table 3.2.

$$
\begin{array}{l}
Y_{11} = \mu + A_1 + S(A)_{11} \\
Y_{12} = \mu + A_1 + S(A)_{12} \\
Y_{13} = \mu + A_1 + S(A)_{13} \\
Y_{14} = \mu + A_1 + S(A)_{14} \\
Y_{21} = \mu + A_2 + S(A)_{21} \\
Y_{22} = \mu + A_2 + S(A)_{22} \\
Y_{23} = \mu + A_2 + S(A)_{23} \\
Y_{24} = \mu + A_2 + S(A)_{24} \\
Y_{31} = \mu + A_3 + S(A)_{31} \\
Y_{32} = \mu + A_3 + S(A)_{32} \\
Y_{33} = \mu + A_3 + S(A)_{33} \\
Y_{34} = \mu + A_3 + S(A)_{34}
\end{array}
$$

Table 3.2: GLM Breakdown of Spelling Data Set A

Consider one of these equations—say the seventh one, $Y_{23} = \mu + A_2 + S(A)_{23}$. This equation states that the score of Subject 3 in Group 2 equals the baseline, plus the effect of having been taught with Method 2, plus a random error contribution specific to that individual subject. Thus, the model describes every data value as being a sum of specific components.

To use this model for partitioning variance (and subsequently testing $H_0$), we must come up with best guesses or "estimates" for the numerical values of the components. The process of coming up with these values is called *estimating the terms in the model*. Every term in the model must be estimated before we can use it. Thus, for this experiment, we must estimate 16 values: $\mu$, $A_1$, $A_2$, $A_3$, $S(A)_{11}$, $S(A)_{12}$, $S(A)_{13}$, $S(A)_{14}$, $S(A)_{21}$, $S(A)_{22}$, $S(A)_{23}$, $S(A)_{24}$, $S(A)_{31}$, $S(A)_{32}$, $S(A)_{33}$, and $S(A)_{34}$.

The equations used to estimate these 16 values (called the *estimation equations*) are shown in Table 3.3. We will not show any mathematical proofs that these are the right equations to use for getting the best guess for each term, though such proofs do exist. Instead, we will offer some conceptually-oriented explanations of each equation.

$$
\begin{array}{ll}
\hat{\mu} & = Y_{..} \\
\hat{A}_i & = Y_{i.} - \hat{\mu} \\
\widehat{S(A)}_{ij} & = Y_{ij} - \hat{\mu} - \hat{A}_i
\end{array}
$$

Table 3.3: Estimation Equations for a One-Factor ANOVA

There are two new types of notation in this table. First, some subscripts on the $Y$'s have been replaced with dots, as in $Y_{i.}$. A dot is used to refer to *the average across different values of the dotted*

*subscript*. For example, $Y_{1\cdot}$ refers to the average of the scores in Group 1—that is, the average of $Y_{11}$, $Y_{12}$, $Y_{13}$, and $Y_{14}$. In Data Set A, that average has the numerical value of 120. Similarly, $Y_{\cdot\cdot}$ refers to the average of all the scores in the data set, which is 100.

The second type of notation introduced in this table is the use of "hats" over some terms, as in $\hat{\mu}$. Hats are always used to remind us that the indicated numerical values are estimated from random samples rather than being "true" values measured for the population as a whole. For example, after we compute 100 for $\mu$, we must remember that it is our best estimate based on this sample, but the true population value is very likely at least somewhat different.

Using these estimation equations on Data Set A, we get the estimated values in Table 3.4. Note that the estimation equations for $A_i$ and $S(A)_{ij}$ are used more than once, because it is necessary to estimate $A$ for each group and $S(A)$ for each subject.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $\hat{\mu}$ | $= Y_{\cdot\cdot}$ | = | | | | | | | 100 |
| $\hat{A}_1$ | $= Y_{1\cdot}$-$\hat{\mu}$ | = | 120 | - | 100 | | | = | +20 |
| $\hat{A}_2$ | $= Y_{2\cdot}$-$\hat{\mu}$ | = | 97 | - | 100 | | | = | -3 |
| $\hat{A}_3$ | $= Y_{3\cdot}$-$\hat{\mu}$ | = | 83 | - | 100 | | | = | -17 |
| $\widehat{S(A)}_{11}$ | $= Y_{11}$-$\hat{\mu}$-$\hat{A}_1$ | = | 124 | - | 100 | - | (+20) | = | +4 |
| $\widehat{S(A)}_{12}$ | $= Y_{12}$-$\hat{\mu}$-$\hat{A}_1$ | = | 129 | - | 100 | - | (+20) | = | +9 |
| $\widehat{S(A)}_{13}$ | $= Y_{13}$-$\hat{\mu}$-$\hat{A}_1$ | = | 115 | - | 100 | - | (+20) | = | -5 |
| $\widehat{S(A)}_{14}$ | $= Y_{14}$-$\hat{\mu}$-$\hat{A}_1$ | = | 112 | - | 100 | - | (+20) | = | -8 |
| $\widehat{S(A)}_{21}$ | $= Y_{21}$-$\hat{\mu}$-$\hat{A}_2$ | = | 101 | - | 100 | - | (-3) | = | +4 |
| $\widehat{S(A)}_{22}$ | $= Y_{22}$-$\hat{\mu}$-$\hat{A}_2$ | = | 88 | - | 100 | - | (-3) | = | -9 |
| $\widehat{S(A)}_{23}$ | $= Y_{23}$-$\hat{\mu}$-$\hat{A}_2$ | = | 107 | - | 100 | - | (-3) | = | +10 |
| $\widehat{S(A)}_{24}$ | $= Y_{24}$-$\hat{\mu}$-$\hat{A}_2$ | = | 92 | - | 100 | - | (-3) | = | -5 |
| $\widehat{S(A)}_{31}$ | $= Y_{31}$-$\hat{\mu}$-$\hat{A}_3$ | = | 76 | - | 100 | - | (-17) | = | -7 |
| $\widehat{S(A)}_{32}$ | $= Y_{32}$-$\hat{\mu}$-$\hat{A}_3$ | = | 91 | - | 100 | - | (-17) | = | +8 |
| $\widehat{S(A)}_{33}$ | $= Y_{33}$-$\hat{\mu}$-$\hat{A}_3$ | = | 84 | - | 100 | - | (-17) | = | +1 |
| $\widehat{S(A)}_{34}$ | $= Y_{34}$-$\hat{\mu}$-$\hat{A}_3$ | = | 81 | - | 100 | - | (-17) | = | -2 |

Table 3.4: Model Estimates for Data Set A

The best guesses obtained by using the estimation equations are usually summarized in a "decomposition matrix", shown in Table 3.5. The decomposition matrix has columns corresponding to the GLM, which is written at the top as a pattern. Each column of the matrix corresponds to one term of the model. For example, under the $Y_{ij}$ column are listed all of its values (the data). Under the $\mu$ column are listed all of its values, and so on. Thus, the decomposition matrix shows how each individual score was obtained, according to the model: as the sum of a baseline (100), a group effect (20, -3, or -17), and an error score (many values).

| $Y_{ij}$ | = | $\hat{\mu}$ | + | $\hat{A}_i$ | + | $\widehat{S(A)}_{ij}$ |
|---|---|---|---|---|---|---|
| 124 | = | 100 | + | 20 | + | 4 |
| 129 | = | 100 | + | 20 | + | 9 |
| 115 | = | 100 | + | 20 | - | 5 |
| 112 | = | 100 | + | 20 | - | 8 |
| 101 | = | 100 | - | 3 | + | 4 |
| 88 | = | 100 | - | 3 | - | 9 |
| 107 | = | 100 | - | 3 | + | 10 |
| 92 | = | 100 | - | 3 | - | 5 |
| 76 | = | 100 | - | 17 | - | 7 |
| 91 | = | 100 | - | 17 | + | 8 |
| 84 | = | 100 | - | 17 | + | 1 |
| 81 | = | 100 | - | 17 | - | 2 |

Table 3.5: Decomposition Matrix for Data Set A

## 3.3 Conceptual Explanations of Estimation Equations

The estimation equations make some intuitive sense. For example, it makes sense to estimate the overall baseline as the average of all the scores in the experiment, because the overall baseline is assumed to have a common effect on all of these scores. As already noted, statisticians like to distinguish between theoretical parameters and numerical estimates of these parameters by writing "hats" over the parameter names, so we would say that for Data Set A of Table 3.1:

$$\hat{\mu} = 100$$

This is only an estimate because there is sampling error in the experiment, but it is the best estimate of $\mu$ available from that data set.

In order to measure the effect of Factor A, we need to consider the effect of being in each group. What does being in Group 1 tend to do to a spelling test score, for example? Well, on average, it increases the scores 20 points above the baseline. Thus, $\hat{A}_1$ is +20. Note that an $\hat{A}$ is estimated as the difference between the group mean and the baseline or overall average. That is,

$$\hat{A}_i = (\text{mean for group i}) - \hat{\mu}$$

It follows that:

$$\hat{\mu} + \hat{A}_i = \text{mean for group i}$$

Conceptually, the effect of a particular method is to move scores up or down relative to the baseline.

Because the baseline is defined as the average of the three group means, the $\hat{A}$'s (deviations of group means from baseline) must add up to 0. That is,

$$\sum_{i=1}^{\text{Number of Groups}} \hat{A}_i = 0$$

In this example, the +20, -3, and -17 sum to 0, and in fact it must always be true that some groups are above the baseline and some below by corresponding amounts. This is a useful fact to know when doing computations for these problems, because it provides an check to help make sure that you have done the computations correctly. We will see two additional computational checks later.

The values of $S(A)_{ij}$ are computed as the difference between each score and the (baseline + group effect). In other words, the error associated with a score is *whatever is left over* after we have taken into account the baseline and the group effect. This makes sense, because the design of our experiment takes into account *only* the effects of baseline and group (Factor A). Any other effect (e.g., what the subject had for breakfast on the day of testing) is not explainable in this design, so it is attributed to random error.

Note that the estimates of error must also add up to 0. This fact is true not only for the data set as a whole, but also for each group separately. Thus,

$$\sum_j \widehat{S(A)}_{ij} = 0 \quad \text{for every i}$$

providing a second check on the computations. The same sort of argument used to justify the $\hat{A}$'s summing to 0 can be applied here. The error scores are deviations of individual scores from the group mean. Overall, some people in the group will be above average, and some below, by corresponding amounts. Across all the scores in the group, the positive and negative deviations from the mean must exactly cancel when the error scores are summed.

How do these estimation equations relate to the idea of variance? Each individual score can be regarded as having its own individual variation:

$$\text{Individual score - Mean} = \text{Individual variation}$$

The general linear model says that this individual variation is made up of two different components:

$$\text{Individual variation} = (\text{Variation due to group}) + (\text{Random variation})$$

One component is variation due to the difference between the group mean and the baseline. For example, consider the first score under Method 1 in Data Set A: 124. Part of the reason that this score is above the mean is because it is in Method 1, which tends to give high scores. $\hat{A}_1$ measures

this component of the individual variation. This component of variation is called the "effect of group" or the "group effect", and we say that being in Group 1 tends, on average, to increase scores by 20 units.

The other component is the variation of the individual score within its group, generally referred to as "random error". The data score, in addition to being well above the baseline, is 4 units above the mean for its own group. The within-group variation or random error associated with this score, then, would be +4.

## 3.4 Summary Measures of Variation

In the last two sections we saw how to measure the effects of the baseline, experimental factor, and random error on each score, by partitioning each score into components contributed by each aspect of the experiment. In this subsection we will see how to get summary measures of these effects for the data set as a whole. We need summary measures so that we can make a single, overall test of the

$$H_0\colon\ u_1 = u_2 = u_3$$

The original discussion of testing this $H_0$ indicated that we would like to compare variation between groups against variation within groups. To the extent that variation between is large relative to variation within, we have evidence against the null hypothesis.

We must obtain composite or summary measures of the variation between and within groups from the parameters we have estimated. This is done in two steps. First, we compute sums of squares ($SS$'s). We compute one to measure between-group variation and one to measure within-group variation. These are found simply by squaring and summing numbers in the appropriate column of the decomposition matrix.

To get the between-group $SS$, we use the column corresponding to the $\hat{A}$ term in the GLM. Square each number in that column, and sum these squared values. This is the *between-group sum of squares* or $SS_{between}$. It is better to call it the "sum of squares due to A" or $SS_A$, though, because this terminology will be more useful in multifactor designs.

In our Data Set A, for example, the $SS_A$ is 2,792. It should be apparent that the $SS_A$ will be large to the extent that there are large differences among the group means. Since the $\hat{A}$'s were estimated as differences between group means and overall means, the $\hat{A}$'s will all be close to 0 if the group means are all about the same. To the extent that the $\hat{A}$'s are close to 0, of course, the $SS_A$ will be small. When group means are quite different from one another, at least some of them must differ considerably from the baseline, so some of the $\hat{A}$'s will have to be fairly large (i.e., far from 0 in absolute value). This will tend to produce a large $SS_A$. Thus $SS_A$ is a measurement of variation between group means, with larger $SS_A$ indicating more differences among the groups.

The $SS_{within}$ is also found by squaring and adding all the numbers in the corresponding column of the decomposition matrix. The corresponding column for $SS_{within}$ is $S(A)_{ij}$. To the extent that variation within groups is large, $SS_{within}$ will also be large. Therefore, this is a composite measure of the within-group variation in our data set. Again, it is convenient for more complex designs to give this $SS$ a more specific name than $SS_{within}$. The most specific name identifies the model term from which the $SS$ was computed, so we had best call this $SS_{S(A)}$, even though that is a little hard to read.

In the same manner, we can also compute $SS$'s for the data and the baseline by squaring and adding up the corresponding columns of the decomposition matrix. It is worthwhile to do this, even though we won't need these values to test our null hypothesis, because they provide a useful check on the calculations. As discussed further below, the $SS_\mu$, $SS_A$, and $SS_{S(A)}$ should add up to $SS_Y$, if all the computations have been done correctly. Table 3.6 summarizes the $SS$ computations.

It is also worth noting that there is another computational check involving the sums of squares. Specifically, it must always be the case that:

$$SS_Y = SS_\mu + SS_A + SS_{S(A)}$$

Actually, this equation has considerable conceptual significance in addition to serving as a useful computational check, but we shall defer discussion of that topic until Section 3.8.

## 3.5 Degrees of Freedom

If you have been following the arguments closely, you may think that we are done. We have measured $SS_A$ and $SS_{S(A)}$, which are directly related to the between- and within-group variances. Shouldn't we

Decomposition Matrix:

| $Y_{ij}$ | = | $\hat{\mu}$ | + | $\hat{A}_i$ | + | $\widehat{S(A)}_{ij}$ |
|---|---|---|---|---|---|---|
| 124 | = | 100 | + | 20 | + | 4 |
| 129 | = | 100 | + | 20 | + | 9 |
| 115 | = | 100 | + | 20 | - | 5 |
| 112 | = | 100 | + | 20 | - | 8 |
| 101 | = | 100 | - | 3 | + | 4 |
| 88 | = | 100 | - | 3 | - | 9 |
| 107 | = | 100 | - | 3 | + | 10 |
| 92 | = | 100 | - | 3 | - | 5 |
| 76 | = | 100 | - | 17 | - | 7 |
| 91 | = | 100 | - | 17 | + | 8 |
| 84 | = | 100 | - | 17 | + | 1 |
| 81 | = | 100 | - | 17 | - | 2 |

Sum of squares of all values in each column:

| | | | | | | |
|---|---|---|---|---|---|---|
| 123,318 | = | 120,000 | + | 2,792 | + | 526 |

Table 3.6: Sums of Squares for Data Set A

now be able to compare these two values somehow to decide whether the differences between groups are too large to be due to random error? Unfortunately, it's not quite that simple.

There is one very subtle problem that prevents us from comparing $SS_A$ and $SS_{S(A)}$ directly. The problem is that these values are based on different numbers of independently estimated parameters. Specifically, there are 12 different $\widehat{S(A)}_{ij}$'s, but only 3 $\hat{A}_i$'s.

The size of an $SS$ is related to the number of different numerical values assigned to its term in the model (the more values, the larger its $SS$ tends to be). Think of it this way: Since there are more $\widehat{S(A)}$'s than $\hat{A}$'s, there are more chances for some really big $\widehat{S(A)}$'s to make the sum of squares for $S(A)$ large relative to the sum of squares for $A$. This will be illustrated below with an example involving height prediction.

To get comparable measures of variation, we have to form an average measure of variance called a *mean square* ($MS$). This average reflects the $SS$ per parameter value that the model term can explain, and the $MS$'s can be compared directly.

To obtain the mean square, we divide each $SS$ by the number of independent values that the associated term in the model can take on. The number of independent values is called the *degrees of freedom* or *df* associated with that term in the model. The mean square and degrees of freedom for each term in the model are symbolized with $MS_{\text{model term}}$ and $df_{\text{model term}}$.

For example, the $A$ term can take on two independent values: $\hat{A}_1$ and $\hat{A}_2$. $\hat{A}_3$ is not counted, because it is not independent of $\hat{A}_1$ and $\hat{A}_2$. In fact, since the $\hat{A}$'s must sum to 0, $\hat{A}_3$ is fully determined by $\hat{A}_1$ and $\hat{A}_2$. Thus the degrees of freedom for $A$ ($df_A$) is 2, and the mean square for $A$ is

$$MS_A = \frac{SS_A}{df_A}$$

To compute the $MS_{S(A)}$, we must divide the $SS_{S(A)}$ by the $df_{S(A)}$. But how many independent $\widehat{S(A)}$'s did we estimate? Recall that there were 3 groups with 4 subjects per group for a total of 12 data points, and we estimated one $\widehat{S(A)}$ per data point. However, these 12 $\widehat{S(A)}$'s were not all estimated independently. Within each group of 4 subjects, the $\widehat{S(A)}$'s had to add up to 0. That means that we really only estimated 3 independent $\widehat{S(A)}$'s per group; the fourth was determined by the first 3, since the four had to sum to 0. Altogether, then, we estimated 3 independent $\widehat{S(A)}$'s in each of the 3 groups, for a total of 9 independent $\widehat{S(A)}$'s. Thus the $df_{S(A)}$ is 9. In general, the number of degrees of freedom for subject is always equal to the number of subjects minus the number of groups (here, 12 - 3 = 9).

Another way to look at the $df_A$ is as a count of the number of differences we are looking at. The $SS_A$ is supposed to measure the variation between groups, and it makes sense that the more groups we look at the larger the differences we should find, if only by chance. The $df_A$ counts the number of

independent differences that can be obtained from our groups. Note that with 2 groups, there is one difference ($df = 2 - 1 = 1$). With three groups we can look at the difference between Group 1 and Group 2 as one difference, and the difference between Group 1 and Group 3 as a second difference. The difference between Groups 2 and 3 is not counted, because it is entirely determined by the first two comparisons (i.e., it is redundant once we have looked at the first two differences). Thus with 3 groups there are 2 independent differences that we can look at ($df = 3 - 1 = 2$). In general, whenever there are K groups, there are K - 1 independent differences that we can look at. The same argument applies to $df_{S(A)}$, except that now we are looking at differences among people within a group rather than differences among groups.

Though the concept of degrees of freedom is almost always mysterious when first encountered, we will try to explain with an example why the sum of squares associated with a term in the model tends to increase with the number of degrees of freedom associated with the model. (That, after all, is why we need the correction factor.)

For this example we will suppose that we are given appropriate data for a particular sample of 96 people, and asked to see whether a person's height (DV) is related to the time of day at which he or she was born (Factor A). We will consider two different ways to analyze the data, one with only two groups (2 values for $\hat{A}$) and one with 24 groups (24 values for $\hat{A}$). We will see that the computed $SS_A$ is likely to be much larger in the analysis with more groups, even though the true $SS_A$ is zero in both cases. The discussion of this example assumes that there are in fact (i.e., across the whole population) no differences in height depending on the time of day of birth. That is, we are assuming that the true $SS_A$ is 0 for both analyses.

### 3.5.1 Analysis 1

If we had to explain variation in height in terms of whether a person was born in the a.m. or the p.m. (a factor with 2 levels), we would probably not be able to do very well. We would have to try to find a difference between those born in the a.m. hours and those born in the p.m. hours. Since there are no differences across the whole population between those born in the a.m. and those born in the p.m., and since we would probably have about 48 people in each group, we would probably find a rather small difference in average height between the a.m. and the p.m. halves of our sample. This would lead to a fairly small $SS_A$ for the factor of a.m. versus p.m. We would probably be forced to conclude that a.m. versus p.m. birth could not account for much of the variation in height.

### 3.5.2 Analysis 2

Again we are asked to explain variation in height for a sample of 96 people, only this time we are allowed to explain it in terms of the *hour* of birth rather than simply a.m. versus p.m. Using hour of birth, we could divide up the sample of 96 into 24 groups of approximately 4 each. Then, we could look for differences among the average heights in the 24 groups. With 24 groups, of course, each group would be fairly small. Even though these groups do not differ across the population as a whole (by assumption), we might well find sizable differences between the different groups *in our sample*. There is considerable random variation in height, and with small numbers of individuals in each group these chance variations would not be likely to cancel out as completely as they would in the previous example with about 48 people per group. Therefore, whereas we were confident that our two group means would be nearly equal in Analysis 1, with about 48 people per group, we cannot be confident of that in Analysis 2, with, on average, around 4 people per group. Thus, we expect larger chance differences between groups in Analysis 2 than in Analysis 1. Because of the way $SS_A$ is computed, that is the same as saying that we expect a larger $SS_A$ in Analysis 2.

### 3.5.3 Comparison of Analyses

What this example shows is that, other things being equal (i.e., true $SS_A$ of 0), we expect to obtain a larger $SS_A$ if we divide the sample into more groups. But we don't want our statistical measure to claim there are larger differences between groups just because there are more groups; we want the test to take the number of groups into account and correct for it. This illustrates the need for a correction factor for the $SS_A$, to take into account the number of groups into which we divided the total sample. That is exactly what $df_A$ is used for.

In terms of the time of birth and height example, why do we also need a correction factor for the $SS_{S(A)}$? One argument is that $SS_{S(A)}$ is the complement of $SS_A$, since we divided up total variation

into variation due to two components. If using 24 groups rather than 2 groups increases the $SS_A$, then it must reduce the $SS_{S(A)}$. It would not make sense to apply a correction factor to one of the components but not the other. Furthermore, our correction factors for the two components should trade off: if we increase the correction factor for $SS_A$ then we should decrease the correction factor for $SS_{S(A)}$, and vice versa. Another way to write the formula for $df_{S(A)}$ is useful here:

$$df_{S(A)} = \text{Number of subjects} - \text{Number of groups}$$

In this formula, it is pretty clear that the correction factor decreases as the number of groups increases. Interestingly, if we divided up the sample into as many groups as we had data points, then $df_{S(A)}$ would be 0. This makes sense from the point of view that if each data point were in its own group, we would have no way to estimate within-group variance, so the analysis would be impossible.

### 3.5.4 Summary of degrees of freedom

Two points summarize the concept of degrees of freedom. First, the *df* associated with a term in the GLM equals the number of independent values of that term. For example, how many independent values are associated with the term $\mu$? One, clearly, the overall mean. Thus, the *df* for $\mu$ is always 1. Similarly, how many independent values are associated with the term $Y$ (the data values)? Each individual data point is measured independently of all the other ones, and knowing the value of some data points will not allow computation of any of the other ones. Thus, the total number of degrees of freedom associated with the $Y$'s is equal to the total number of $Y$'s observed.

Second, the *df* is used as a correction factor for the $SS$ associated with that term. The $SS$ divided by the *df* always gives the $MS$, and the $MS$ is corrected in the sense that it is an average amount of variation attributable to each of the independent parameters that we estimated.

## 3.6 Summarizing the Computations in an ANOVA Table

We have done quite a few different calculations in analysing the single factor experiment of this chapter, and it is standard procedure to summarize these computations with an ANOVA table of the form shown in Table 3.7.

| Source | *df* | $SS$ | $MS$ | $F$ | Error Term |
|---|---|---|---|---|---|
| $\mu$ | 1 | 120000.0 | 120000.0 | 2053.232 | $S(A)$ |
| $A$ | 2 | 2792.0 | 1396.0 | 23.886 | $S(A)$ |
| $S(A)$ | 9 | 526.0 | 58.4 | | |
| Total | 12 | 123318.0 | | | |

Table 3.7: Summary ANOVA Table for Data Set A

There are five main columns in the ANOVA table, shown as the first five columns of Table 3.7. The source column lists the sources of variation that went into the experiment, and of course these sources will vary depending on the factors included in the experiment. There is always one source for each term in the linear model. The baseline and total lines are included in the table more for completeness than for utility, though we will see a few situations later where they will be useful. Some ANOVA books do not bother to include the baseline at all, perhaps because it is odd to think of the mean as a source of variation. Perhaps the mean is best thought of as a source of variation from 0, where 0 is some great cosmic baseline.

Each source has associated with it a *df* value and an $SS$ value. We have already seen where these values come from for the $A$ and $S(A)$ sources, and we can get analogous values for the mean. The $SS_\mu$ and $SS_Y$ can be found using the same procedure we used to get $SS_A$ and $SS_{S(A)}$. To get $SS_\mu$, go to the mean column of the matrix in Table 3.6, and add up the squares of all the numbers in the corresponding columns. To get $SS_Y$, find the sum of squares for the numbers in the $Y$ column (i.e., the data values).

Having found the *df*'s and $SS$'s for all the sources, the difficult computations are all done. The $MS$'s are found simply by dividing the $SS$ on each line by the associated *df* on the same line. That takes care of the whole $MS$ column.

Finally, we compute the observed value of the hypothesis testing statistic $F_{observed}$ (also known as the *F ratio*):

$$F_{observed} \text{ for Factor A} = \frac{MS_A}{MS_{S(A)}}$$

Note that this is the ratio of the between-group variance to the within-group variance, as measured by the GLM. If this value is large, we can reject the null hypothesis and conclude that there are more than just chance differences between methods for teaching spelling.

How large does $F_{observed}$ have to be to reject $H_0$? For right now, we will consider this question from a strictly practical point of view; section 3.7 explains some of the theoretical background behind these practical procedures.

In practice, $F_{observed}$ is compared against an $F_{critical}$, which we look up in a table of $F_{critical}$ values. Most tables of $F_{critical}$ are organized like Table 3.8. There is a separate page in the table for each confidence or alpha level at which you might want to test the null hypothesis; in this table, the values are shown for 95% confidence (alpha = .05). The table has rows and columns corresponding to degrees of freedom of the terms in the numerator and denominator of $F_{observed}$, and $F_{critical}$ is the number at the intersection of the appropriate row and column.

In the analysis of Data Set A, for example, $F_{observed}$ was computed with $MS_A$ in the numerator, and the corresponding degrees of freedom for the numerator is $df_A = 2$. $F_{observed}$ had $MS_{S(A)}$ in the denominator, and the corresponding degrees of freedom for the denominator is $df_{S(A)} = 9$. Thus, $F_{critical}$ is the number in the column labelled "2" and the row labelled "9": specifically, 4.26.

| df in denominator | df in numerator 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| 1 | 161.40 | 199.50 | 215.70 | 224.60 | 230.20 | 234.00 | 236.80 | 238.90 |
| 2 | 18.51 | 19.00 | 19.16 | 19.25 | 19.30 | 19.33 | 19.35 | 19.37 |
| 3 | 10.13 | 9.55 | 9.28 | 9.12 | 9.01 | 8.94 | 8.89 | 8.85 |
| 4 | 7.71 | 6.94 | 6.59 | 6.39 | 6.26 | 6.16 | 6.09 | 6.04 |
| 5 | 6.61 | 5.79 | 5.41 | 5.19 | 5.05 | 4.95 | 4.88 | 4.82 |
| 6 | 5.99 | 5.14 | 4.76 | 4.53 | 4.39 | 4.28 | 4.21 | 4.15 |
| 7 | 5.59 | 4.74 | 4.35 | 4.12 | 3.97 | 3.87 | 3.79 | 3.73 |
| 8 | 5.32 | 4.46 | 4.07 | 3.84 | 3.69 | 3.58 | 3.50 | 3.44 |
| 9 | 5.12 | 4.26 | 3.86 | 3.63 | 3.48 | 3.37 | 3.29 | 3.23 |

Table 3.8: Values of $F_{critical}$ for 95% confidence (alpha = .05)

Once we get $F_{critical}$, we apply the following *rejection rule* to decide whether or not to reject $H_0$:

Reject $H_0$ if $F_{observed} > F_{critical}$.
Fail-to-reject $H_0$ if $F_{observed} < F_{critical}$.

The data from Data Set A gave an $F_{observed}$ of 23.886, which is much greater than 4.26, so these data are sufficient to reject the null hypothesis that the three group means are equal. The conclusion to be drawn from this statistical procedure is that the three groups do not all do equally well on the spelling test. That is, performance on the spelling test does depend on which method was used to teach spelling, and the differences we observed among the three groups were not simply due to chance. It seems clear from looking at the means (and knowing that the differences are not all simply due to chance) that the best method for teaching spelling is the one used with Group 1, so this a tempting final conclusion based on our experimental design and statistical analysis. Strictly speaking, however, this would be going too far in our conclusions. All that is justified by the ANOVA is to conclude that the groups do not all have the same mean. Further analyses called *multiple comparisons* are needed before we can draw specific conclusions about which groups are best and which are worst. For example, one strategy would be to redo the analysis including only Groups 1 and 2, in order to see if we had statistical justification for the specific statement that Group 1 is better than Group 2. Many other and better methods for making multiple comparisons are available, but are outside the scope of the present discussion.

Note that there are two $F_{observed}$ values in Table 3.7. Besides the one for A, which we have just discussed, there is another one for $\mu$. This $F$ is computed as

$$F_{observed} \text{ for } \mu = \frac{MS_\mu}{MS_{S(A)}}$$

It tests the null hypothesis that the true value of $\mu$ (i.e., true baseline or overall average) is 0. This is not a particularly interesting null hypothesis to test here, because there is no doubt that people will learn some words regardless of how they are taught to spell. We include it in the computations, though, because in some cases it is interesting to test this hypothesis. For example, suppose we were comparing three different methods of psychotherapy and measuring the improvement after six months of each treatment. In this case, we can find out whether there is an overall trend of improvement (or no change) by testing the null hypothesis that the baseline is 0. Because of the fact that we are measuring improvement, a baseline of 0 is an interesting value.

**Notation:** From now on, we will generally use the abbreviation $F_{\mu}$ in place of the more cumbersome "$F_{observed}$ for $\mu$," and use $F_A$ in place of "$F_{observed}$ for $A$." Thus, whenever you see an $F$ with a subscript that is a term in the GLM you should interpret it to mean the $F_{observed}$ that is used to test the null hypothesis about that term in the model.

Incidentally, we should repeat that there are considerably easier ways to compute $SS_{\mu}$, $SS_A$, and $SS_{S(A)}$. As mentioned in the introduction, the goals of this book involve presenting computational tools that are 1) as intuitive as possible, and 2) easily generalized to more complex situations. Doing it the hard way will cost a little time on the easy problems, but it is good preparation for the more complicated problems that lie ahead. While you may feel that we are using a rifle to kill a mosquito here, we need the practice because there are grizzly bears waiting just around the bend.

## 3.7 Epilogue 1: Why Compare $F_{observed}$ to $F_{critical}$?

The comparison of $F_{observed}$ to $F_{critical}$ that we use to decide whether to reject $H_0$ may appear somewhat arbitrary, but of course it is not. This section describes some of the theoretical background for this comparison, in case you aren't happy just taking it on faith.

The first thing to realize is that $F_{observed}$ is a random quantity ("random variable" is the more technical term for this). This simply means that the value you compute for $F_{observed}$ depends on the exact random sample of data that you happened to obtain. If you repeated the experiment (i.e. took a new random sample), it is virtually certain that you would get (at least slightly) different data values and that you would therefore compute a different value of $F_{observed}$. In short, the value of $F_{observed}$ is random because it varies from one sample to the next.

There are sometimes predictable patterns to randomness, however. As a simple example, flipping a fair coin should theoretically result in equal numbers of heads and tails in the long run. Thus, even though flipping a coin is a random process and we won't get a 50/50 split in every set of flips, we can describe the pattern of resulting randomness by listing all of the observations that might be obtained (heads/tails) and the theoretically expected relative frequency of each observation (50% each). As a more complex example, suppose we select a random person out of a population and measure their weight. This too is random, because each person's weight is different. Again, though, we can describe the pattern of random results. As you probably know, many biological measurements have a normal distribution across the population. Thus, we would expect the random weights to follow a normal distribution pattern in this case.

Analogously to the simpler coin and weight examples, it is possible to describe the patterns of randomness in $F_{observed}$. It may seem amazingly complicated to do this,[3] but it can be proved that any $F_{observed}$ value comes from something called the $F$ distribution, named after its inventor Sir R. A. Fisher. The formula for this distribution can be written down and used to compute the exact relative frequencies of the different $F_{observed}$ values.

Figure 3.2 shows examples of different $F$ distributions illustrating the patterns in the random values of $F_{observed}$. As you can see, the $F$ distribution depends on the numbers of dfs in the numerator and denominator. The $F$ distribution also depends on whether $H_0$ is true and, if it is not true, on the size of the true experimental effect. Typically, though, we mainly consider the $F$ distribution that is predicted when $H_0$ is true, because the $F_{critical}$ value is obtained from this distribution, as is described below. Thus, all of the example distributions shown in Figure 3.2 were computed under the assumption that $H_0$ was true.[4] Note that the value of $F_{observed}$ can vary randomly quite a bit from

[3] After all, each $F_{observed}$ is affected by all of the observations in the experiment, each of which is determined randomly. How could anyone take all those potential effects into consideration to characterize the pattern of randomness that you would get for $F_{observed}$? Those maths folks are pretty smart.

[4] If you were sufficiently industrious, you could obtain one of these distributions empirically. Specifically, you could repeat an experiment thousands of times, taking care that $H_0$ was true each time, and tabulate the distribution of the different $F_{observed}$ values that you obtained. Of course it is not practical to do this, nor is it necessary because of Fisher's work, but it may help understand the meaning of the $F$ test to imagine constructing its distribution under $H_0$

one experiment to the next—even if the null hypothesis is true—although there is less variation when there are more degrees of freedom (in either the numerator or denominator).

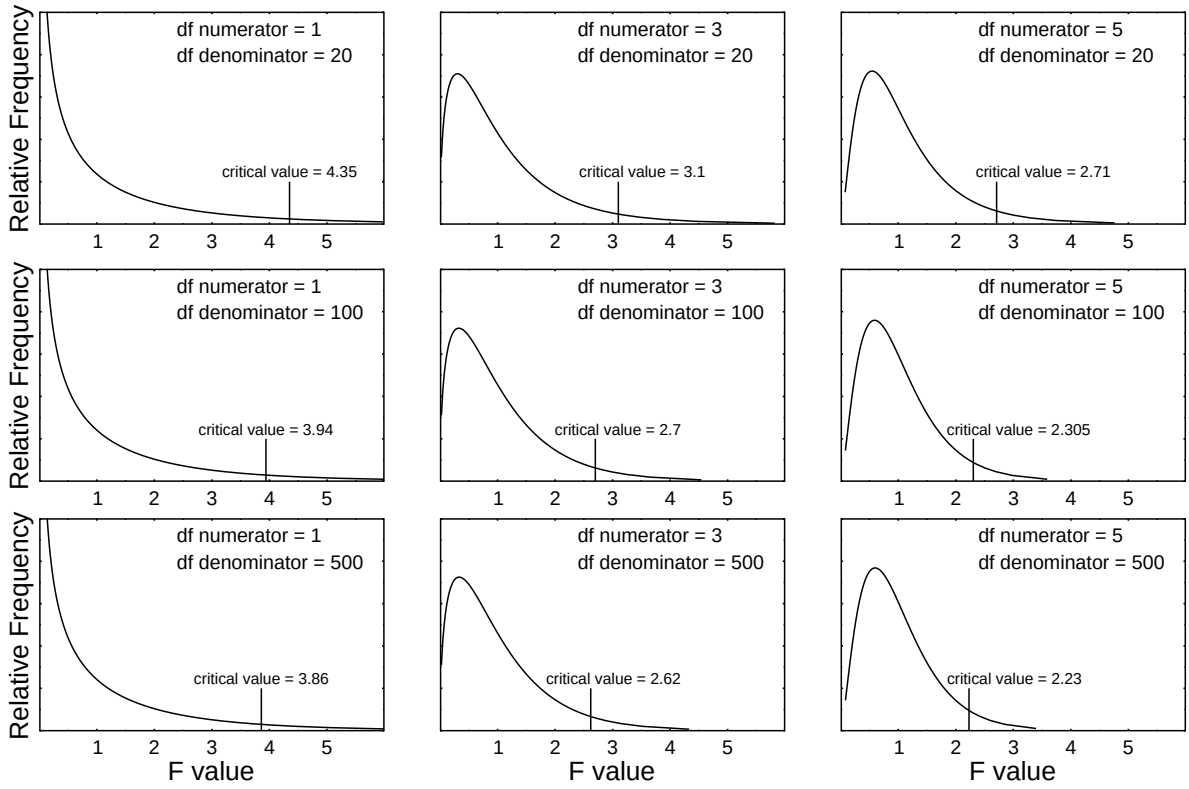


Figure 3.2: Example $F$ Distributions. Each panel shows the theoretical distribution of $F_{observed}$ for the indicated number of degrees of freedom in the numerator and denominator. All distributions were computed under the assumption that the null hypothesis is true. For each distribution, the critical value is indicated by the vertical line. This is the value that cuts off the upper 5% of the distribution—i.e., that $F_{observed}$ will exceed only 5% of the time.

Now the standard approach within hypothesis-testing statistics is to summarize each $F$ distribution with a single "critical" value that can be computed from the formulas describing that distribution. This critical value is a number that will be exceeded only 5% of the time (5% is the conventional choice, but others, such as 1%, are also possible). Tables of $F_{critical}$ simply give this value for various different combinations of degrees of freedom values.

With all of this background, it is finally possible to see the rationale for comparing $F_{observed}$ with $F_{critical}$. First, suppose $F_{observed}$ is less than $F_{critical}$. If the null hypothesis is true, this is not a surprising result, because $F_{observed}$ should usually be less than $F_{critical}$ (i.e., 95% of the time, given that we chose $F_{critical}$ so that a larger value would be obtained only 5% of the time). Therefore, if we find from our data that $F_{observed}$ is less than $F_{critical}$, then we have obtained a result that is quite consistent with $H_0$. In other words, finding $F_{observed}$ less than $F_{critical}$ gives us no reason to reject $H_0$.

Conversely, suppose $F_{observed}$ is greater than $F_{critical}$. If the null hypothesis were true, this *would be* a surprising result, because $F_{critical}$ was chosen to be one of the largest values that we would ever get if $H_0$ were true. Faced with this surprising result, we have two choices. We can either conclude that something unlikely has happened (i.e., we have accidentally observed an *unusually* extreme value of $F_{observed}$), or else we can conclude that $H_0$ must not be true after all. The standard approach is to conclude the latter and reject $H_0$. Of course, sometimes we will be fooled because actually we just got a large $F_{observed}$ by accident. But at least we have arranged it so that won't happen too often across lots of experiments.

To gain further insight into the $F$ test, it is also worth noting that $F_{observed}$ values tend to be larger in situations where $H_0$ is false. Compared to the distributions shown in Figure 3.2, for example,

in this fashion.

the distribution of $F_{observed}$ would be pulled farther to the right if $H_0$ happened to be false (i.e., if the true group means were different). Therefore, when $H_0$ is false, the chance that $F_{observed}$ value will exceed $F_{critical}$ is greater than 5%. This is good, because it means you will reject $H_0$ more often when $H_0$ is false. In fact, if $H_0$ is *very false*—by which nonsensical terminology I mean that the group means being compared differ by a lot—you will tend to reject $H_0$ much more than 5% of the time. The exact probability of rejecting $H_0$ with a certain experiment is known as the *power* of that experiment. Unfortunately, we never know the actual power of an experiment in any practical situation, because the power depends on the difference between group means, and we don't know this (if we knew it, we wouldn't bother to do the experiment). Also unfortunately, power is never 100%, because there is always the potential for error when using a random sample.

## 3.8 Epilogue 2: The Concept of Partitioning

Conceptually, the main idea behind ANOVA is that of *partitioning*, or breaking into components. We have seen how the GLM partitions each data score into a sum of components, with each component reflecting a different aspect of the experimental design (baseline, error, etc.).

The fact that the $\hat{A}_i$'s and $\widehat{S(A)}_{ij}$'s sum to 0 illustrates how the different terms in the model contribute independent pieces of information about each score. By independent, we mean that each new component of the linear model brings in information that could not be conveyed by any simpler terms in the model. For example, suppose for a minute that the $\hat{A}_i$'s summed to a positive number instead of 0. That would indicate that there was a general tendency, across all groups, for scores to be slightly above baseline. This would violate the concept of the baseline, however, since the baseline is supposed to represent the middle point of the data set as a whole. Thus, the $\hat{A}$'s have to add up to zero because they convey a different type of information from that provided by the baseline.

It is important to realize that the GLM partitions not only the individual scores ($Y$'s) but also the total sum of squares and the total degrees of freedom, as shown in Table 3.9. In all three cases a total something (score, sum of squares, degrees of freedom) is broken down into separate and logically independent subcomponents.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Scores: | $Y_{ij}$ | = | $\mu$ | + | $A_i$ | + | $S(A)_{ij}$ |
| Sums of Squares: | $SS_Y$ | = | $SS_\mu$ | + | $SS_A$ | + | $SS_{S(A)}$ |
| Degrees of freedom: | $df_Y$ | = | $df_\mu$ | + | $df_A$ | + | $df_{S(A)}$ |

Table 3.9: Partitioning Equations for One-Way ANOVA

To interpret the partitioning of scores, we said that each data value was made up of three effects: the baseline, the effect of Factor A, and random error. The partitioning of $SS$'s and $df$'s make analogous statements about the data set as a whole rather than about each individual data value.

To interpret the partitioning of $SS$'s, we can say that the total of the squared scores is partly due to the squared baselines, partly due to the squared effects of Factor A, and partly due to the squared effects of error. Conceptually, then, the $SS_{total}$ provides another illustration of the concept of breaking up scores, and variance between scores, into components. Not only do the scores themselves reflect the sum of their components, but the total variance of the scores also reflects the sum of the variances due to different components.

Why do we care about squared scores in the first place, though? The main reason is that sums of squares are convenient composite measures of variation. Remember, if we just added up the $\hat{A}_i$ values instead of the squares of them, we would always get 0 (not such a useful composite measure). When we instead add up the *squares* of the $\hat{A}_i$ values, we obtain a useful measure of the total variation between groups. Of course we could add up absolute values or fourth powers of $\hat{A}_i$'s, instead of squares. The reason that squares are best is essentially a mathematical point—not critical in practice—deriving ultimately from the formula for the normal distribution.

To interpret the partitioning of $df$'s, we can say that the total number of independent data values (the number of $Y$'s) is equal to the number of independent values of $\mu$ plus the number of independent values of $A$ plus the number of independent values of $S(A)$. Why do we care about the number of independent values? Because we need these as corrections for the $SS$'s.

In summary, ANOVA works by a parallel partitioning of scores, sums of squares (variance measures), and degrees of freedom. Each partition takes into account the structure of the experimental

design, including random error.[5]

## 3.9 Summary of Computations

In the preceding sections we have given the rationale for each computational step as we described it, and so the description of the computational procedure is spread out over many pages. Here is a step-by-step summary, for convenience in working problems.

1. Write down the model: $Y_{ij} = \mu + A_i + S(A)_{ij}$
2. Write down the estimation equations:
   - $\hat{\mu} = Y_{..}$ (Average of all $Y$'s).
   - $\hat{A}_i = Y_{i.} - \hat{\mu}$ (Difference between Group i average and overall average).
   - $\widehat{S(A)}_{ij} = Y_{ij} - \hat{\mu} - \hat{A}_i$ (Left-over in score).
3. Compute estimates for all the terms in model.
4. Summarize the estimation process in a decomposition matrix. There is one line in the matrix for each data point, and one column for each term in the model.
5. Compute the Sum of Squares, degrees of freedom, and Mean Square for each term in the model:
   - SS Rule: To find the Sum of Squares corresponding to any term in the model, square all the values in the corresponding column in the decomposition matrix, and total the squared values down the column. (This rule works for decomposition matrices in ALL ANOVA designs, including multifactor, within-subjects and mixed designs.)
   - *df* Rules:
     - The *df* for the mean is always 1.
     - The *df* for the factor is its number of levels - 1.
     - The *df* for error is the total number of subjects - number of groups.
   - For each term, Mean square = Sum of Squares / df.
6. Construct a summary ANOVA Table.
   - Headings are "Source *df* SS MS F Error Term"
   - There is one source line for each term in the model. Start with $\mu$ and end with $Y$ (called "Total").
   - Copy into the table the *df*'s, $SS$'s, and $MS$'s computed in the previous step.
   - Compute the $F_{observed}$'s for both $\mu$ and $A$.
     - $F_{\mu} = MS_{\mu}/MS_{S(A)}$
     - $F_A = MS_A/MS_{S(A)}$
7. Compare the $F_{observed}$ values with the $F_{critical}$ values, rejecting $H_0$ if $F_{observed} > F_{critical}$. Draw conclusions. The $F_{observed}$ for $\mu$ tests $H_0$: $\mu = 0$, and the $F_{observed}$ for Factor A tests $H_0$: All $A_i$'s $= 0$.

[5] There is a sense in which the linear model is a post-hoc, tautological description of the data. It is post-hoc because we use the data to compute the terms of the model. It is tautological because the procedures we use to estimate the terms in the model are guaranteed to give us numerically correct equations of the form

$$Y_{ij} = \hat{\mu} + \hat{A}_i + \widehat{S(A)}_{ij}$$

These are not serious faults, however, since the data are rewritten in terms of conceptually meaningful components. Furthermore, these components are completely independent of one another. Those aspects of the data described by $\hat{\mu}$ are quite separate from those aspects described by the $\hat{A}$'s, which are in turn separate from those described by the $\widehat{S(A)}$'s.

## 3.10 One-Factor Computational Example

**Design:** An instructor wishes to compare quiz performances of freshmen, sophomores, juniors, and seniors. He randomly selects three students from each class level, and compares quiz scores. This design has a single between-subjects factor—class level—which will be called Factor A. There are 3 subjects per group.

**Data:**

| | Levels of Factor A | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $A_1$ | | | $A_2$ | | | $A_3$ | | | $A_4$ | | |
| Scores: | 39 | 41 | 40 | 35 | 35 | 32 | 24 | 25 | 23 | 22 | 23 | 21 |

**Model:** $Y_{ij} = \mu + A_i + S(A)_{ij}$

**Decomposition Matrix:**

| $Y_{ij}$ | | $\hat{\mu}$ | | $\hat{A}_i$ | | $\widehat{S(A)}_{ij}$ |
|---|---|---|---|---|---|---|
| 39 | = | 30 | + | 10 | − | 1 |
| 41 | = | 30 | + | 10 | + | 1 |
| 40 | = | 30 | + | 10 | + | 0 |
| 35 | = | 30 | + | 4 | + | 1 |
| 35 | = | 30 | + | 4 | + | 1 |
| 32 | = | 30 | + | 4 | − | 2 |
| 24 | = | 30 | − | 6 | + | 0 |
| 25 | = | 30 | − | 6 | + | 1 |
| 23 | = | 30 | − | 6 | − | 1 |
| 22 | = | 30 | − | 8 | + | 0 |
| 23 | = | 30 | − | 8 | + | 1 |
| 21 | = | 30 | − | 8 | − | 1 |

**ANOVA Table:**

| Source | *df* | *SS* | *MS* | *F* | Error Term |
|---|---|---|---|---|---|
| $\mu$ | 1 | 10800.0 | 10800.0 | 7200.00 | $S(A)$ |
| $A$ | 3 | 648.0 | 216.0 | 144.00 | $S(A)$ |
| $S(A)$ | 8 | 12.0 | 1.5 | | |
| Total | 12 | 11460.0 | | | |

**Decision:** The $F_{critical}$ with 3 and 8 *df* is 4.07. Since $144 > 4.07$, reject $H_0$ for Factor A.

**Conclusions:** The researcher has enough evidence to reject $H_0$ and conclude that the different class levels do not all have the same average on his quiz. From the graph, it appears that the average is highest for Freshman (around 40), next highest for Sophomores (around 34), and lowest for Juniors (around 24) and Seniors (around 22), and multiple comparison methods could be used to make more specific comparisons if they were desired.

## 3.11 Review Questions about One-Factor ANOVA

1. When do you use a one-way ANOVA? What null hypothesis is it designed to test?
2. Explain what $MS_A$ and $MS_{S(A)}$ are intended to measure about the data. Why are these measurements relevant to testing the null hypothesis?
3. Give the formula for $F_A$.
4. How do you know whether or not to reject $H_0$ once you have $F_{observed}$?
5. What do you conclude if you reject $H_0$? What do you conclude if you fail-to-reject $H_0$?

## 3.12 Computational Exercises

For all exercises, use 95% confidence in testing hypotheses.

**1.** A medical researcher wants to know whether the standard dose of Anesthetic A or B causes surgical patients to stay unconscious longer. He randomly assigns three of his next six surgical patients to be given Drug A, and the other three to be given Drug B. In all cases, he measures the time from when the patient passes out from the drug, to when the patient's eyes first open after surgery. Here are the results:

| | | | |
|---|---|---|---|
| Drug A: | 27 | 18 | 33 |
| Drug B: | 6 | 11 | 25 |

Graph the averages. What should the researcher conclude?

**2.** A computer science instructor wanted to know whether freshman, sophomores, juniors, or seniors spent different amounts of time using the computer for a class assignment. He randomly selected 5 students from each class level, and secretly monitored how much time each spent using the computer during the week before the assignment was due. Here are the results:

| Freshmen | | | | | Sophomores | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 32 | 30 | 28 | 30 | 25 | 27 | 27 | 30 | 30 | 21 |

| Juniors | | | | | Seniors | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 27 | 22 | 23 | 23 | 25 | 22 | 20 | 22 | 23 | 13 |

Graph the averages. What can the instructor conclude?

**3.** Two corn plants of each of six varieties were planted in a field at the beginning of May. All were the same height. At the end of July, the heights were as follows:

| | | | | | |
|---|---|---|---|---|---|
| Variety 1: | 52 | 56 | Variety 4: | 33 | 43 |
| Variety 2: | 47 | 57 | Variety 5: | 57 | 51 |
| Variety 3: | 58 | 54 | Variety 6: | 44 | 48 |

Graph the averages. Is this enough evidence to conclude that the different varieties produce plants that grow at different rates?

## 3.13 Answers to Exercises

**1.** This is a one-factor, between-Ss design, with 2 groups and 3 subjects per group.

**Model:** $Y_{ij} = \mu + A_i + S(A)_{ij}$

**Estimation Equations:**

$$\begin{aligned} \hat{\mu} &= Y_{..} \\ \hat{A}_i &= Y_{i.} - \hat{\mu} \\ \widehat{S(A)}_{ij} &= Y_{ij} - \hat{\mu} - \hat{A}_i \end{aligned}$$

**Decomposition Matrix:**

| $Y_{ij}$ | = | $\hat{\mu}$ | + | $\hat{A}_i$ | + | $\widehat{S(A)}_{ij}$ |
|---|---|---|---|---|---|---|
| 27 | = | 20 | + | 6 | + | 1 |
| 18 | = | 20 | + | 6 | − | 8 |
| 33 | = | 20 | + | 6 | + | 7 |
| 6 | = | 20 | − | 6 | − | 8 |
| 11 | = | 20 | − | 6 | − | 3 |
| 25 | = | 20 | − | 6 | + | 11 |

**ANOVA Table:**

| Source | *df* | *SS* | *MS* | *F* | Error Term |
|---|---|---|---|---|---|
| $\mu$ | 1 | 2400.0 | 2400.0 | 31.169 | $S(A)$ |
| $A$ | 1 | 216.0 | 216.0 | 2.805 | $S(A)$ |
| $S(A)$ | 4 | 308.0 | 77.0 | | |
| Total | 6 | 2924.0 | | | |

**Decision:** $F_{critical}$ with $df$ = 1,4 is 7.71, and $2.805 < 7.71$. Therefore, fail-to-reject $H_0$.

**Conclusions:** Though the patients receiving Drug A stayed unconscious longer on average, there is not enough evidence to reject the explanation that this finding was obtained by chance. Of course there may be a difference between drugs, but this experiment hasn't demonstrated it.

**2.** This is a one-factor, between-Ss design, with 4 groups and 5 subjects per group. Note that the model and estimation equations are the same as in the previous problem.

**Model:** $Y_{ij} = \mu + A_i + S(A)_{ij}$

**Estimation Equations:**

$$\begin{aligned} \hat{\mu} &= Y_{..} \\ \hat{A}_i &= Y_{i.} - \hat{\mu} \\ \widehat{S(A)}_{ij} &= Y_{ij} - \hat{\mu} - \hat{A}_i \end{aligned}$$

**Decomposition Matrix:**

| $Y_{ij}$ | = | $\hat{\mu}$ | + | $\hat{A}_i$ | + | $\widehat{S(A)}_{ij}$ |
|---|---|---|---|---|---|---|
| 32 | = | 25 | + | 4 | + | 3 |
| 30 | = | 25 | + | 4 | + | 1 |
| 28 | = | 25 | + | 4 | − | 1 |
| 30 | = | 25 | + | 4 | + | 1 |
| 25 | = | 25 | + | 4 | − | 4 |
| 27 | = | 25 | + | 2 | + | 0 |
| 27 | = | 25 | + | 2 | + | 0 |
| 30 | = | 25 | + | 2 | + | 3 |
| 30 | = | 25 | + | 2 | + | 3 |
| 21 | = | 25 | + | 2 | − | 6 |
| 27 | = | 25 | − | 1 | + | 3 |
| 22 | = | 25 | − | 1 | − | 2 |
| 23 | = | 25 | − | 1 | − | 1 |
| 23 | = | 25 | − | 1 | − | 1 |
| 25 | = | 25 | − | 1 | + | 1 |
| 22 | = | 25 | − | 5 | + | 2 |
| 20 | = | 25 | − | 5 | + | 0 |
| 22 | = | 25 | − | 5 | + | 2 |
| 23 | = | 25 | − | 5 | + | 3 |
| 13 | = | 25 | − | 5 | − | 7 |

**ANOVA Table:**

| Source | *df* | *SS* | *MS* | *F* | Error Term |
|---|---|---|---|---|---|
| $\mu$ | 1 | 12,500 | 12,500.0 | 1,219.51 | $S(A)$ |
| $A$ | 3 | 230 | 76.7 | 7.48 | $S(A)$ |
| $S(A)$ | 16 | 164 | 10.3 | | |
| Total | 20 | 12,894 | | | |

**Decision:** $F_{critical}$ with $df$ = 3,16 is 3.24. Since $7.48 > 3.24$, reject $H_0$.

**Conclusions:** The instructor can conclude that students in different classes spent different times working on the computer in the week before the assignment was due. From the averages, it appears that more advanced students spent less time.

**3.** This is a one-factor, between-Ss design, with 6 groups and 2 subjects per group.

**Model:** $Y_{ij} = \mu + A_i + S(A)_{ij}$

**Estimation Equations:**

$$\begin{array}{lcllll} \hat{\mu} & = & Y_{..} & & & \\ \hat{A}_i & = & Y_{i.} & - & \hat{\mu} & \\ \widehat{S(A)}_{ij} & = & Y_{ij} & - & \hat{\mu} & - \quad \hat{A}_i \end{array}$$

**Decomposition Matrix:**

| $Y_{ij}$ | = | $\hat{\mu}$ | + | $\hat{A}_i$ | + | $\widehat{S(A)}_{ij}$ |
|---|---|---|---|---|---|---|
| 52 | = | 50 | + | 4 | − | 2 |
| 56 | = | 50 | + | 4 | + | 2 |
| 47 | = | 50 | + | 2 | − | 5 |
| 57 | = | 50 | + | 2 | + | 5 |
| 58 | = | 50 | + | 6 | + | 2 |
| 54 | = | 50 | + | 6 | − | 2 |
| 33 | = | 50 | − | 12 | − | 5 |
| 43 | = | 50 | − | 12 | + | 5 |
| 57 | = | 50 | + | 4 | + | 3 |
| 51 | = | 50 | + | 4 | − | 3 |
| 44 | = | 50 | − | 4 | − | 2 |
| 48 | = | 50 | − | 4 | + | 2 |

**ANOVA Table:**

| Source | *df* | *SS* | *MS* | *F* | Error Term |
|---|---|---|---|---|---|
| $\mu$ | 1 | 30,000 | 30,000.0 | 1,267.60 | $S(A)$ |
| $A$ | 5 | 464 | 92.8 | 3.92 | $S(A)$ |
| $S(A)$ | 6 | 142 | 23.7 | | |
| Total | 12 | 30,606 | | | |

**Decision:** $F_{critical}$ with $df = 5$ and 6 is 4.39. Since $3.92 < 4.39$, fail-to-reject $H_0$.

**Conclusions:** The researcher does not have enough evidence to conclude that the varieties grow at different rates. The differences in the averages could be due to chance.

# Chapter 4

# Two-Factor, Between-Subject Designs

## 4.1 The Information in Two-Factor Experiments

ANOVA gets a lot more interesting, and only a little more complex conceptually and computationally, when we consider designs with two factors (called *two-factor* or *two-way* designs). Before we consider the statistical analysis of two factor experiments, though, we should consider the possible results. After all, a statistical analysis only indicates whether or not the results are due to chance. You must know what sorts of results you are looking for before you ask whether they are due to chance.

As an example experiment, suppose we were interested in finding out what determines how well college students learn foreign languages, and we believed that two important factors were the gender of the student and the gender of the teacher. (For example, we might want to disprove a sexist tale about how females are better than males at learning languages.) We would thus have two experimental factors, gender of student and gender of teacher, each with the two levels "male" and "female" (a 2 by 2 design).

To examine the effects of these two factors, it is most natural to set up a completely crossed factorial design with four groups, as shown in Table 4.1. Two-factor designs are almost always displayed in such a table, with rows corresponding to the levels of one factor and columns corresponding to the levels of the other factor.[1] We could randomly assign male and female college students to male and female teachers at the beginning of a year of language study, and give all students a standardized language achievement test at the end of the year.[2] The DV would be the score on the year-end test.

| | | Teacher: Male | Teacher: Female |
|---|---|---|---|
| Student: | Male | | |
| | Female | | |

Table 4.1: Experimental Design to Study Effects of Student and Teacher Gender on Learning

What sorts of results might be obtained in this experiment? Obviously, there are quite a few different patterns of possible results, and it is instructive to consider a number of them. We will first consider idealized results (without random error), just to illustrate some logically different patterns. Later, we will see how ANOVA can be used to see if more realistic results (i.e., with random error) are due to chance or to real effects of the experimental factors.

Table 4.2 shows a number of different patterns of results that might be obtained in different samples. For each sample data set, the number within each cell is the average language test score (DV) for that cell. The averages for each row and column are written at the right side and bottom of each table, including the overall average at the bottom right. These averages are called the *marginals*, and they

[1] It would be equally correct to make the table with Gender of Student along the top and Gender of Teacher along the side. It is just a question of which way makes more sense to the researcher.

[2] This experiment is trickier than it looks, because of the problem of selecting teachers as well as students. To make this example statistically correct, we will make two assumptions: 1) Teachers are randomly selected also. 2) No two students in the experiment had the same teacher– that is, every student is from a different class. The latter assumption is needed to satisfy the assumption that all the scores in the analysis are independent.

are almost always included with a table of results. Figure 4.1 shows the factorial plot corresponding to each set of sample results, using a conventional format explained below.

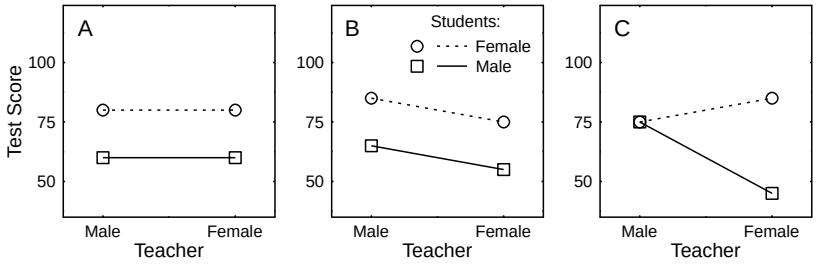


Figure 4.1: Factorial plots showing three different sets of possible results for an experiment testing amount of learning as a function of gender of student and gender of teacher.

Consider Sample A. Female students score higher than male students, but gender of the teacher has no effect. In ANOVA terminology, Sample A is said to have a *main effect* of Gender of Student but no main effect of Gender of Teacher. A factor has a main effect whenever the average scores from the different levels of that factor are different, averaging across the levels of the other factor. In simple language, a factor has a main effect when its levels are different overall. Thus, the main effect of Gender of Student is revealed in the marginal averages at the right side of the table (60 vs. 80), and the lack of main effect of Gender of Teacher is revealed in the marginal averages at the bottom (70 vs. 70).

The graph corresponding to Sample A in Figure 4.1 also shows that Gender of Student has a main effect but Gender of Teacher does not. This type of graph is called a *factorial plot.* It is conventional to draw these graphs so that:

- The average value of the DV is on the vertical axis.
- The levels of one factor appear along the horizontal axis.
- The levels of the other factor appear as different lines on the graph, with one line for each level.

In factorial plots, it is equally correct to use either factor on the horizontal axis. For example, in this case it would be equally correct to make the graph with Gender of Student along the horizontal axis, using different lines for male vs. female teachers. The choice of how to orient the graph is up to the researcher, and should be determined by which way makes more intuitive sense to look at. It is not unreasonable to draw the graphs both ways, just to see how they look.

To see main effects in factorial plots, you have to do some "averaging by eye", which is usually fairly easy. For example, to see whether there is a main effect of Gender of Teacher in the graph for Sample A, visualize the point corresponding to the average height for male teachers (average on the left side of the graph), visualize the average height for female teachers (average on the right side), and compare the two visualized averages. It is easy to see that Gender of Teacher has no main effect, because the points on the left side of the graph are no higher or lower, on average, than the points on the right side of the graph. Conversely, to see whether there is a main effect of Gender of Student, you have to visualize the average height of each line (averaging across left and right sides—Gender of Teacher). In this graph, the main effect of Gender of Student is evident from the fact that the line for female students has a higher average than the line for male students.

Sample B shows a data set with both main effects present. Looking at the marginal means, we can see that the main effect of Gender of Student is the same size as in Sample A, and the main effect of Gender of Teacher is half that size. (Note that scores are higher for male teachers than female ones. Obviously, the data must be fictitious.) We can see the same thing by looking at the factorial plot for these data. The line for female students is clearly higher, on average, than the line for male students, reflecting the main effect of Gender of Student. Furthermore, the points on the left side of the graph are higher, on average, than the points on the right side, reflecting the main effect of Gender of Teacher. It is also interesting to note that each factor has a main effect that is the same

Sample A:

| | | Teacher Gender: Male | Female | Average |
|---|---|---|---|---|
| Student Gender: | Male | 60 | 60 | 60 |
| | Female | 80 | 80 | 80 |
| | Average | 70 | 70 | 70 |

Sample B:

| | | Teacher Gender: Male | Female | Average |
|---|---|---|---|---|
| Student Gender: | Male | 65 | 55 | 60 |
| | Female | 85 | 75 | 80 |
| | Average | 75 | 65 | 70 |

Sample C:

| | | Teacher Gender: Male | Female | Average |
|---|---|---|---|---|
| Student Gender: | Male | 75 | 45 | 60 |
| | Female | 75 | 85 | 80 |
| | Average | 75 | 65 | 70 |

Table 4.2: Student Gender by Teacher Gender: Sample Results

for both levels of the other factor. For example, scores for male teachers are ten points higher than those for female teachers. This is true not only in the marginal averages, but it is also true when you look only at the male students or only at the female students. Thus, the effect of Gender of Teacher is the same for male and female students. Similarly, scores for female students are 20 points higher than those for male students, and this is true for both male teachers and for female teachers. When, as in this case, the effects of each factor are the same across all levels of the other factor, we say that there is no *interaction* between the factors. This point will become clearer after we consider the next data set, in which an interaction is present.

Sample C shows a more complicated pattern of results. First, note that the marginal averages are the same in Sample C as they were in Sample B. In other words, the main effects of both factors are the same for both samples. What is very different between B and C, though, is what goes on "inside" the table. In these data, Gender of Teacher has a very different effect for male students than for female students. For male students, scores are higher with male teachers than with female teachers. For female students it is just the reverse: scores are higher with female teachers than with male teachers. In ANOVA terms, this data set is said to have an *interaction* between the two experimental factors. An interaction is said to be present whenever the effect of one factor (e.g., having a male or female teacher) is different across the levels of the other factor (e.g., being a male or female student).

The graph for Sample C illustrates the interaction very nicely. The lines for male and female students are tilted in opposite directions, so it is easy to see that male teachers are better than female teachers for male students and that the reverse is true for female students.

We can also see the two main effects in the graph for Sample C, though it is more difficult than in the graph for Sample B. The main effect of Gender of Student is evident from the fact that the line for female students is higher, *on average*, than the line for male students. Similarly, the main effect of Gender of Teacher can be seen by noting that the average of the points on the left side of the graph is higher than the average of the points on the right.

The above patterns of results considered in Samples A, B, and C illustrate three different types of information that we can obtain from this two-factor design:

1. Information about the main effect of Gender of Student,
2. Information about the main effect of Gender of Teacher, and
3. Information about the effect of an interaction between Gender of Teacher and Gender of Student.

Analogous types of information are present in all two-factor experiments: the main effect of one factor, the main effect of the other factor, and the effect of an interaction between the two factors.

The purpose of the statistical analysis will be to help us decide which of these effects are too large to be due to chance.

## 4.2 The Concept of a Two-Factor Interaction

The concept of an interaction is very important in the analysis of all multifactor designs, and it probably unfamiliar to most students. At an intuitive level, an interaction is a special effect that arises because two factors work together in some particular way. That is, a *particular combination* of levels on two factors gives rise to some effect that couldn't have been predicted from knowing just the main effects of the two factors. The term "synergy" is sometimes used outside of statistics to refer to similar results dependent on factors having special effects in certain combinations.

Within the context of ANOVA, the concept of an interaction can be specified much more precisely than this. This section will first define the concept as it is used in ANOVA, and then illustrate it with several additional examples.

**Definition:** There is a two-way *interaction* between two experimental factors when the effect of one factor *differs across* the levels of the other factor. Or, equivalently: There is a two-way interaction between two experimental factors when the effect of one factor *depends on* the level of the other factor. We say that there is a large interaction when the effect of one factor depends a great deal on the level of the other factor. We say that there is a small interaction (or no interaction) when the effect of one factor depends only a little (or not at all) on the level of the other factor.

We will illustrate this definition with three examples. The first two are clear (if perhaps slightly silly) cases in which we would expect very large interactions between two factors (i.e., the effect of one factor should vary quite a bit depending on the level of the second factor).

First, suppose we ask people how much they like certain sandwiches, letting them respond on a 1-to-5 scale with 1 being not at all and 5 being very, very much. One experimental factor is whether or not the sandwich has bologna on it. A second experimental factor is whether or not the sandwich has peanut butter on it. These two factors define four types of sandwiches that we would include in the experiment: 1) a sandwich with no bologna and no peanut butter (i.e., two slices of bread), 2) a sandwich with bologna and no peanut butter, 3) a sandwich with peanut butter but no bologna, and 4) a sandwich with both peanut butter and bologna. It seems pretty clear that how much people like having peanut butter on a sandwich will depend on whether or not the sandwich has bologna on it, as in the hypothetical results shown in Table 4.3 and Figure 4.2.

| | | Peanut Butter | | |
|---|---|---|---|---|
| | | No | Yes | Average |
| Bologna | No | 2.1 | 3.7 | 2.9 |
| | Yes | 3.5 | 1.1 | 2.3 |
| | Average | 2.8 | 2.4 | 2.6 |

Table 4.3: Effects of Peanut Butter and Bologna on Sandwiches

Looking at the top row of the table, it is clear that people like a sandwich with bread plus peanut butter more than they like a sandwich with just bread. In this comparison, the effect of putting peanut butter on a sandwich is to make the sandwich more likeable. Looking at the bottom row, however, it is clear that people like a sandwich with bread plus bologna sandwich better than they like one with bread plus bologna plus peanut butter sandwich. In this comparison, the effect of putting peanut butter on a sandwich is to make the sandwich less likeable. Thus putting peanut butter on the sandwich (Factor A) can either increase or decrease the likability of the sandwich, depending on whether or not bologna is on the sandwich (Factor B).[3]

A second example of an extreme interaction effect would be found in an experiment on the effects of vision on markspersonship. Suppose that we had people fire a rifle at a target, and we measured distance from the bull's eye to the point where the bullet hit as the DV. The two experimental factors are A) whether the right eye is open or closed, and B) whether the left eye is open or closed. A large interaction effect would almost certainly be obtained in this experiment: markspersonship would be

[3]We could turn this around, too, because it is also true that the effect of putting bologna on the sandwich also depends on whether or not peanut butter is on the sandwich. In fact, interaction effects are always symmetric in this way, as will be discussed shortly.

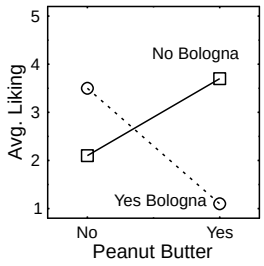


Figure 4.2: Factorial plots of hypothetical results showing liking for a sandwich as a function of presence/absence of peanut butter and presence/absence of bologna.

only slightly affected by whether the right eye was open or closed when the left eye was open. However, when the left eye was closed, there would surely be a huge effect of whether the right eye was open or closed. In the latter case, both eyes are closed and so the person is shooting blind! Sample results are shown in Table 4.4 and Figure 4.3.

| | | Right Eye: Open | Closed | Average |
|---|---|---|---|---|
| Left Eye: | Open | 0.6 | 0.4 | 0.5 |
| | Closed | 0.2 | 100.8 | 50.5 |
| | Average | 0.4 | 50.6 | 25.25 |

Table 4.4: Effects of Left and Right Eye on Target Shooting

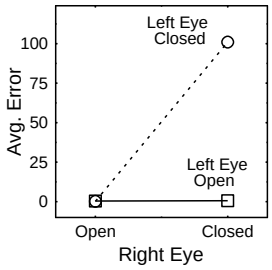


Figure 4.3: Factorial plots of hypothetical results showing sharp-shooting accuracy as a function of right eye open vs. closed and left eye open vs. closed.

A third and more realistic example of an interaction can be derived from the discussion in Section 2.1 of a hypothetical experiment looking for an effect of coffee on perceptual alertness. As discussed previously, it seems reasonable to expect that the effect of a drinking or not drinking cup of coffee (Factor A) would depend on the time of day at which alertness was tested (Factor B). Presumably, coffee would have a bigger effect at 8:00 a.m. than at 3:00 p.m., because everyone is wide awake and doesn't need coffee that much in the afternoon. According to this argument, in an experiment with these two factors we would expect an interaction like that shown in Table 4.5.

In these results, coffee increases perceptual alertness at both times of day. There is still an interaction effect, though, because the increase is larger in the morning (30 points) than in the afternoon (10 points). Note that this is an easy way to check for an interaction: Compute the effect of one factor

| | Time of Day: 8:00 a.m. | 3:00 p.m. | Average |
|---|---|---|---|
| Coffee | 40 | 50 | 45 |
| No Coffee | 10 | 40 | 25 |
| Average | 25 | 45 | 35 |

Note: Higher scores indicate greater alertness.

Table 4.5: Effects of Coffee and Time of Day on Alertness

(e.g., coffee) at each level of the other factor (e.g., time of day). If the effect is different, there is an interaction.[4]

All three of the above examples illustrate the theoretical point that two-factor interaction effects are always symmetric. That is, if the effect of Factor A differs across the levels of Factor B, then the effect of Factor B differs across the levels of Factor A. In the sandwich example, we noted that the effect of putting peanut butter on the sandwich depended on whether there was bologna or not. But we could equally have said it the other way around: the effect of putting bologna on the sandwich depended on whether there was peanut butter or not. Similarly, in the marksmanship example, the effect of having the right eye open or closed depends on whether the left eye is open or closed, *and vice versa*. Similarly, we could say that the difference between morning and afternoon alertness (effect of time of day) is much less with coffee than without it.

This symmetry means that we can choose to look at a two-factor interaction effect in either of two ways, depending on our intuition rather than some statistical consideration. The researcher can always view this interaction as a change in the effect of either factor depending on the level of the other factor, choosing whichever point of view is intuitively more attractive. In fact, some times it is informative to think about the interaction first from one point of view and then from the other—we may realize additional implications of the data by doing this.

Having defined and illustrated the concept of an interaction effect and its symmetry, we will now discuss some more practical issues concerning these effects.

First, standard ANOVA terminology refers to an interaction effect as simply an *interaction*, dropping the word "effect". We will therefore adopt this practice too. Keep in mind, though, that an interaction is something that influences the results just like a main effect.

Second, the easiest way to spot an interaction is always by looking at a factorial plot. In a factorial plot, parallel lines indicate no interaction, and nonparallel lines indicate an interaction.

Consider Figure 4.1A—the graph of a data set without an interaction. As advertised, the lines in the plot are parallel. To see why "no interaction" is equivalent to "parallel lines", remember an interaction indicates that the effect of a factor is different across levels of the other factor. The effect of gender of teacher is shown (separately for male and female students) by the slope of a line. If there is no interaction, then the effect must be the same for male and female students, which must mean that the slopes are the same (hence that the lines are parallel). If we want to look at it in terms of the effect of gender of student, we should note that the difference between the two lines indicates the effect of gender of student. That this difference is the same for male and female teachers indicates that the lines are equally far apart for male and female teachers, which again must mean that they are parallel.

To the extent that there is an interaction, the lines on the factorial plot must have different slopes. The larger the difference in slopes, the larger the interaction. Of course, we will need a statistical test to tell us whether a difference in slopes is too large to be due to chance. Figure 4.1C shows one of many possible patterns in which there is an interaction of two factors. All such patterns have in common the property that the lines in the factorial plot are not parallel, but of course there are many ways for the lines to be nonparallel. The point is that whenever the effect of one factor depends on the level of the other factor, the lines cannot be parallel.

Third, interactions can have an important influence on the conclusions that we draw from an experiment. To see how and why, let's return to the example datasets concerning effects of Gender of Teacher and Student on language learning (Table 4.1).

In Sample B there was no interaction at all: female students were better than male students by exactly the same amount for male teachers as for female teachers. The effect of Gender of Student

[4]When checking the effect, be sure to subtract the levels in the same order. A common error is to just take the largest minus the smallest at each level, rather than $A_1$ minus $A_2$ at each level of $B$.

does not depend on Gender of Teacher, and vice versa, so there is no interaction. In the absence of an interaction, it is possible to draw conclusions simply in terms of the main effects: The results indicate that female students learn languages better than male students, and that male teachers teach languages better than female ones.

In Sample C, however, there was a large interaction. Male teachers were much better than female teachers for male students, but female teachers were better than male teachers for female students. Thus the effect of Gender of Teacher depended very much on whether the student was male or female, and this is what we would conclude from the results. To give an even more complete description of the results, we could summarize the form of the interaction by saying that students learned much better from teachers of the same gender than from teachers of the opposite gender.

The interaction observed in Sample C could easily have important consequences for theories being tested. The results would support, for example, a theory asserting that an important determinant of language learning was the acoustic match between the voices of the teacher and learner (perhaps due to some very specific type of vocal tract modeling). The results would contradict any theory predicting that everyone should learn better from female teachers (e.g., because our linguistic apparatus is more sensitive to high frequency voices), and any theory predicting that everyone should learn better from a teacher of the opposite gender (e.g., because students will work much harder to impress a teacher of the opposite gender).

Interactions are important because they lead to interesting conclusions that are often considerably more sophisticated than the original research questions that the researcher had in mind (e.g., "are male and female students equally good at learning", and "are male and female teachers equally good at teaching"). The procedures we use to analyze two-factor designs often remind us of the possibility that the world is a more complicated place than we assumed in our initial research questions.

Sample C especially illustrates the practical importance of interactions for drawing conclusions from experimental results. With these results, we might draw some seriously erroneous conclusions if we looked only at the main effects. First, looking at the main effect of Gender of Student, we might conclude that female students are better than male students. If we were to apply such a general conclusion to a situation with only male teachers, however, we would get a big surprise. While the conclusion was true in the overall averages, it was not true for male teachers. Similarly, looking only at the main effect of Gender of Teacher, we might also go wrong by concluding oversimplistically that male teachers are better than female. Such a conclusion is too general, obviously, because the reverse is true for female students.

This example shows that we must be careful about drawing general conclusions from main effects when interactions are present in the results. In the presence of an interaction, we should be very careful to qualify the conclusions we draw about main effects. For example, we could conclude that female students are better than male *on average*, but we must also add the qualification that this does not hold for male teachers.

Finally, we should emphasize the point that the opportunity to evaluate an interaction is often the main reason why a researcher does a two-factor experiment in the first place. As we have seen, the presence or absence of an interaction says a lot about the generality of a factor's effect, and this can have both theoretical and practical implications. Sometimes an empirical question can only be answered by looking at an interaction, as in these examples: 1) Is it true that rich people donate more money to the Republican party and poor people donate more to the Democratic party (DV = amount donated, Factor A = Rich vs. poor, Factor B = Republican vs. Democrat)? 2) Is it true that males have more mathematical ability and females have more verbal ability (DV = ability, Factor A = Gender, Factor B = math vs. verbal)?

Of course, two-way designs also allow a researcher to look at the effects of two independent variables at the same time rather than doing two separate experiments, and this is also part of the reason for doing two-factor experiments. To the extent that it is easier to do one two-factor experiment than two one-factor experiments, it might just seem like two-factor experiments lead to faster, more efficient research. But the interaction information is even more important, because this information cannot easily be obtained from single-factor experiments.

## 4.3 The GLM for Two-Factor Between-Subjects Designs

The statistical analysis of data from a two-factor experiment relies heavily on the GLM. Here is the model for a design with two between-subjects factors:

$$Y_{ijk} = \mu + A_i + B_j + AB_{ij} + S(AB)_{ijk}$$

Note that the data values have three subscripts in the two-factor design. By convention, the one in the first position going from left to right, here $i$, always indicates the level of Factor A; the one in the second position, here $j$, indicates the level of Factor B; and the one in the third position, here $k$, indicates the subject within each group. Thus, $Y_{123}$ is the data value from the third subject in the group with level 1 of Factor A and level 2 of Factor B. When we get to designs with even more factors, even more subscripts will be added to the $Y$ values, and the position of the subscript will continue to tell you which factor it corresponds to. The first subscript always indicates the level of Factor A, the second indicates the level of Factor B, the third indicates the level of Factor C, and so on as far as needed; the last subscript indicates the subject.

Unfortunately, there is no consensus across statistics books about what symbol to use for the interaction term in a two-factor linear model. In this class, we will use a concatenation of factor letters to indicate an interaction of the factors; the advantage of this notation is that it is easily extended to designs with many factors. It is important to read $AB_{ij}$ as a separate term all of its own, where the term measures the effect due to interaction of Factors A and B. Do not read this term as anything involving multiplication of A with B.

All the terms in the two-factor model measure theoretically important and intuitively meaningful quantities, just as they did in the one-factor design. This model has a baseline, a main effect of each factor (i.e., one term for the effect of Factor A and one term for the effect of Factor B), and an effect of subject within group (i.e., sampling error). Each of these terms has the same intuitive meaning as in the one-factor model.

The model for the two-way design also has an additional term to measure the $AB$ interaction. Specifically, this term measures an effect due to the particular combination of levels of the two factors. That is why this term has subscripts specifying levels of both factors: To indicate the combination of levels.

In summary, the two-factor model takes into account five distinctly different effects on the data values:

1. The effect of the baseline ($\mu$).
2. The effect of Factor A ($A_i$).
3. The effect of Factor B ($B_j$).
4. The effect of the AB interaction ($AB_{ij}$).
5. The effect of within-group variability or error ($S(AB)_{ijk}$).

We will use the model to measure each effect and test null hypotheses saying that each effect is just due to chance.

## 4.4 Computations for Two-Factor Between-Subjects Designs

The computations in the two-factor analysis are essentially the same as in the one-way design: Estimate the terms of the model, get the sum of squares and degrees of freedom associated with each term, form $MS$'s, and compute $F_{observed}$'s to test null hypotheses. The major difference from the one-way analysis is that the two-way model has more terms.

It is easiest to estimate the terms of the linear model from left to right. The estimate of $\mu$, naturally, is the overall mean in the data set $Y_{...}$. (Remember, the dot notation indicates that you average over the dotted subscript, as described in Section 3.2).

Estimates of the main effects of Factors A and B are obtained as in the one-way design, except now we must average over another factor:

$$\begin{aligned} \hat{A}_i &= Y_{i..} - \hat{\mu} \\ \hat{B}_j &= Y_{.j.} - \hat{\mu} \end{aligned}$$

In words, the estimated effect for a certain level of a factor is the average for all the observations at that level of the factor minus the overall average of all observations.

Skipping over the interaction terms briefly, since they require extra discussion, we estimate the error terms with either of these two equivalent formulas:

$$\begin{aligned} \widehat{S(AB)}_{ijk} &= Y_{ijk} - Y_{ij.} \\ \widehat{S(AB)}_{ijk} &= Y_{ijk} - \hat{\mu} - \hat{A}_i - \hat{B}_j - \widehat{AB}_{ij} \end{aligned}$$

According to the first equation, the estimate of error associated with an observation ($Y_{ijk}$) is the difference between that observation and the average of all observations in the exact same experimental condition ($Y_{ij.}$). The only difference from the one-way design is that now the experimental conditions are defined by two factors rather than only one. According to the second equation, the estimate of error is whatever is left over of an observation after we subtract the parts we can account for with the other terms in the model. In a sense, the estimate of error is whatever is needed to make the model be correct for that observation.

The interaction terms are estimated using this equation:

$$\widehat{AB}_{ij} = Y_{ij.} - \hat{\mu} - \hat{A}_i - \hat{B}_j$$

In words, the interaction term associated with a particular cell is that portion of the cell mean $Y_{ij.}$ that is left over after we subtract out the overall baseline, the deviation from baseline due to the effect of Factor A, and the deviation from baseline due to the effect of Factor B. As already mentioned, the interaction term measures whatever is special about the particular combination of levels of the two factors. In other words, the interaction term indicates how much of the cell mean cannot be explained simply by the overall baseline, the overall effect of Factor A, and the overall effect of Factor B.

To fully understand interaction terms, it is important to see why

$$\begin{aligned} \sum_j \widehat{AB}_{ij} &= 0 \quad \text{for every } i \\ \sum_i \widehat{AB}_{ij} &= 0 \quad \text{for every } j \end{aligned}$$

These equations simply say that if we total the $\widehat{AB}$'s across any row or column of the design, we must get a result of 0. Another way to say this is that the average interaction effect must be 0 for every row and column, because if the total is 0 the average must also be 0.

Intuitively, the justification for this is that interaction terms measure deviations from what is expected given the baseline and main effects. The overall deviation of a given row or column from the baseline is what determines the estimate of the main effect for that row or column. If the main effect terms do the jobs as intended, no overall deviation from baseline can be left over for the interaction terms.

There is also a graphic justification for the idea that the $\widehat{AB}_{ij}$'s must add to 0 across both $i$'s and $j$'s, given that the interaction terms measure the extent to which lines are parallel in the factorial plot. If the $\widehat{AB}$'s did not average 0 for one of the lines, for example, the non-zero average would tend to move the whole line either up or down without changing its slope. That would change the overall separation between lines, which is what the main effect terms are supposed to measure. A similar argument can be made about the average $\widehat{AB}$'s for a set of points above a particular mark on the horizontal axis. In that case, though, the non-zero average $\widehat{AB}$ would move all the points above that mark up or down, and would change the slopes of all lines equally. Since the interaction terms are supposed to measure differences in slopes, something that changes all slopes equally is irrelevant.

Finally, a brief algebraic argument will easily establish that the $\widehat{AB}_{ij}$'s must average to 0 for every i, averaging across j. We know from the equation used to estimate the $\widehat{AB}$'s that

$$Y_{ij.} = \hat{\mu} + \hat{A}_i + \hat{B}_j + \widehat{AB}_{ij}$$

Now we take the average across $j$ of both sides of this equation, and get:

$$Y_{i..} = \hat{\mu} + \hat{A}_i + \widehat{AB}_{i.}$$

But, from the equation to estimate $\hat{A}_i$, we also know that

$$Y_{i..} = \hat{\mu} + \hat{A}_i$$

We now have two expressions both equal to each $Y_{i..}$, so they must equal each other:

$$\hat{\mu} + \hat{A}_i = \hat{\mu} + \hat{A}_i + \widehat{AB}_{i.}$$

It is clear from this equation that $\widehat{AB}_{i.} = 0$. The same proof can easily be rewritten to show that the $\widehat{AB}_{ij}$'s must average to 0 for every j, averaging across i.

To relate the estimates of interaction terms back to the earlier discussion of what an interaction means, it is important to see that interaction estimates actually measure the extent to which the observed effect of one factor depends on the level of the other factor. More specifically, it is important to see that $\widehat{AB}$'s will be far from 0 when very different effects of one factor are observed at different levels of another factor, and $\widehat{AB}$'s will be close to 0 when the effects of each factor are the same at all the levels of the other factor. The best way to understand this is to see what happens when you plug different values for $\hat{\mu}$, the $\hat{A}$'s, the $\hat{B}$'s, and the $\widehat{AB}$'s into the following equation:

$$Y_{ij.} = \hat{\mu} + \hat{A}_i + \hat{B}_j + \widehat{AB}_{ij}$$

Try making up different values. If all the $\widehat{AB}$'s are set to 0 then you will get parallel lines regardless of what numbers you choose for the other terms. (Don't forget that $\hat{A}_i$ and $\hat{B}_j$ have to sum to 0).

Another way to get an intuitive grasp of the $\widehat{AB}$'s is to think of them as measuring the effect of one factor (say A) separately for each level of the other factor (say B), relative to the overall effect of factor A. That is, the $\widehat{AB}$'s indicate whether the effect of A observed for $B_j$ is more or less than the average A effect. If the A effect is the same for all the levels of B, then the $\widehat{AB}$'s are all 0, because no level of B shows an A effect different from the average A effect. To the extent that the A effects differ across levels of B, though, the $\widehat{AB}$'s will differ from 0. Viewing the problem in this way, we could think of the equation

$$Y_{ij.} = \hat{\mu} + \hat{A}_i + \hat{B}_j + \widehat{AB}_{ij}$$

as describing the cell mean as the sum of the baseline, an overall effect of A, an overall effect of B, and a correction for the effect of A specific to the particular level of B. If we get the same effect of A at all levels of B, then the correction is 0. To the extent that we get more and more different effects of A at the different levels of B, though, larger and larger corrections will be needed.

The discussion in the previous paragraph requires us to focus on the effect of one particular factor and think of the variation in its effect across levels of the other factor. This is a more specific way to think about an interaction, since the emphasis is on the effect of one factor. Thinking in these specific terms will not get us into trouble, however, as long as we remember that we could just as well have switched the roles of the two factors in our thought processes.

Now that we have seen how to estimate terms of the linear model for the two-way design, we can do a complete analysis and test all the null hypotheses that are involved in the two-way design. Table 4.6 shows a sample data set (3 subjects per cell) for the Gender of Teacher by Gender of Student design discussed earlier. The cell and marginal means are at the bottom of the table.

Raw Data:

| | | Teacher Gender: Male | Teacher Gender: Female |
|---|---|---|---|
| Student Gender: | Male | 73, 73, 70 | 67, 68, 69 |
| | Female | 73, 76, 73 | 83, 81, 82 |

Cell and Marginal Means:

| | | Teacher Gender: Male | Teacher Gender: Female | Average |
|---|---|---|---|---|
| Student Gender: | Male | 72 | 68 | 70 |
| | Female | 74 | 82 | 78 |
| | Average | 73 | 75 | 74 |

Table 4.6: Data: Effects of Student and Teacher Gender on Learning

Table 4.7 summarizes the linear model and the estimation equations from the discussion above. As in the one-way design, we simply use the estimation equations and the data to get numerical values for each of the terms in the linear model. Table 4.8 shows exactly how this is done using the data in Table 4.6.

As in the one-way ANOVA, the estimated values are summarized in a decomposition matrix, shown in Table 4.9. Sums of squares are computed by squaring and adding the terms in each column of the decomposition matrix.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $\hat{\mu}$ | = | $Y_{...}$ | | | | | | | |
| $\hat{A}_i$ | = | $Y_{i..}$ | - | $\hat{\mu}$ | | | | | |
| $\hat{B}_j$ | = | $Y_{.j.}$ | - | $\hat{\mu}$ | | | | | |
| $\widehat{AB}_{ij}$ | = | $Y_{ij.}$ | - | $\hat{\mu}$ | - | $\hat{A}_i$ | - | $\hat{B}_j$ | |
| $\widehat{S(AB)}_{ijk}$ | = | $Y_{ijk}$ | - | $\hat{\mu}$ | - | $\hat{A}_i$ | - | $\hat{B}_j$ | - $\widehat{AB}_{ij}$ |

Table 4.7: Estimation Equations for the model $Y_{ijk} = \mu + A_i + B_j + AB_{ij} + S(AB)_{ijk}$

$$
\begin{aligned}
\hat{\mu} &= Y_{...} = 74 \\
\hat{A}_1 &= Y_{1..} - \hat{\mu} = 70 - 74 = -4 \\
\hat{A}_2 &= Y_{2..} - \hat{\mu} = 78 - 74 = +4 \\
\hat{B}_1 &= Y_{.1.} - \hat{\mu} = 73 - 74 = -1 \\
\hat{B}_2 &= Y_{.2.} - \hat{\mu} = 75 - 74 = +1 \\
\widehat{AB}_{11} &= Y_{11.} - \hat{\mu} - \hat{A}_1 - \hat{B}_1 \\
&= 72 - 74 - (-4) - (-1) = +3 \\
\widehat{AB}_{12} &= Y_{12.} - \hat{\mu} - \hat{A}_1 - \hat{B}_2 \\
&= 68 - 74 - (-4) - (+1) = -3 \\
\widehat{AB}_{21} &= Y_{21.} - \hat{\mu} - \hat{A}_2 - \hat{B}_1 \\
&= 74 - 74 - (+4) - (-1) = -3 \\
\widehat{AB}_{22} &= Y_{22.} - \hat{\mu} - \hat{A}_2 - \hat{B}_2 \\
&= 82 - 74 - (+4) - (+1) = +3 \\
S\widehat{(AB)}_{111} &= Y_{111} - \hat{\mu} - \hat{A}_1 - \hat{B}_1 - \widehat{AB}_{11} \\
&= 73 - 74 - (-4) - (-1) - (+3) = +1 \\
S\widehat{(AB)}_{112} &= Y_{112} - \hat{\mu} - \hat{A}_1 - \hat{B}_1 - \widehat{AB}_{11} \\
&= 73 - 74 - (-4) - (-1) - (+3) = +1 \\
S\widehat{(AB)}_{113} &= Y_{113} - \hat{\mu} - \hat{A}_1 - \hat{B}_1 - \widehat{AB}_{11} \\
&= 70 - 74 - (-4) - (-1) - (+3) = -2 \\
S\widehat{(AB)}_{121} &= Y_{121} - \hat{\mu} - \hat{A}_1 - \hat{B}_2 - \widehat{AB}_{12} \\
&= 67 - 74 - (-4) - (+1) - (-3) = -1 \\
S\widehat{(AB)}_{122} &= Y_{122} - \hat{\mu} - \hat{A}_1 - \hat{B}_2 - \widehat{AB}_{12} \\
&= 68 - 74 - (-4) - (+1) - (-3) = 0 \\
S\widehat{(AB)}_{123} &= Y_{123} - \hat{\mu} - \hat{A}_1 - \hat{B}_2 - \widehat{AB}_{12} \\
&= 69 - 74 - (-4) - (+1) - (-3) = +1 \\
S\widehat{(AB)}_{211} &= Y_{211} - \hat{\mu} - \hat{A}_2 - \hat{B}_1 - \widehat{AB}_{21} \\
&= 73 - 74 - (+4) - (-1) - (-3) = -1 \\
S\widehat{(AB)}_{212} &= Y_{212} - \hat{\mu} - \hat{A}_2 - \hat{B}_1 - \widehat{AB}_{21} \\
&= 76 - 74 - (+4) - (-1) - (-3) = +2 \\
S\widehat{(AB)}_{213} &= Y_{213} - \hat{\mu} - \hat{A}_2 - \hat{B}_1 - \widehat{AB}_{21} \\
&= 73 - 74 - (+4) - (-1) - (-3) = -1 \\
S\widehat{(AB)}_{221} &= Y_{221} - \hat{\mu} - \hat{A}_2 - \hat{B}_2 - \widehat{AB}_{22} \\
&= 83 - 74 - (+4) - (+1) - (+3) = +1 \\
S\widehat{(AB)}_{222} &= Y_{222} - \hat{\mu} - \hat{A}_2 - \hat{B}_2 - \widehat{AB}_{22} \\
&= 81 - 74 - (+4) - (+1) - (+3) = -1 \\
S\widehat{(AB)}_{223} &= Y_{223} - \hat{\mu} - \hat{A}_2 - \hat{B}_2 - \widehat{AB}_{22} \\
&= 82 - 74 - (+4) - (+1) - (+3) = 0
\end{aligned}
$$

Table 4.8: Estimates for Data in Table 4.6

| | $Y_{ijk}$ | = | $\hat{\mu}$ | + | $\hat{A}_i$ | + | $\hat{B}_j$ | + | $\widehat{AB}_{ij}$ | + | $S\widehat{(AB)}_{ijk}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 73 | = | 74 | + | $(-4)$ | + | $(-1)$ | + | $(+3)$ | + | $(+1)$ |
| | 73 | = | 74 | + | $(-4)$ | + | $(-1)$ | + | $(+3)$ | + | $(+1)$ |
| | 70 | = | 74 | + | $(-4)$ | + | $(-1)$ | + | $(+3)$ | + | $(-2)$ |
| | 67 | = | 74 | + | $(-4)$ | + | $(+1)$ | + | $(-3)$ | + | $(-1)$ |
| | 68 | = | 74 | + | $(-4)$ | + | $(+1)$ | + | $(-3)$ | + | $(0)$ |
| | 69 | = | 74 | + | $(-4)$ | + | $(+1)$ | + | $(-3)$ | + | $(+1)$ |
| | 73 | = | 74 | + | $(+4)$ | + | $(-1)$ | + | $(-3)$ | + | $(-1)$ |
| | 76 | = | 74 | + | $(+4)$ | + | $(-1)$ | + | $(-3)$ | + | $(+2)$ |
| | 73 | = | 74 | + | $(+4)$ | + | $(-1)$ | + | $(-3)$ | + | $(-1)$ |
| | 83 | = | 74 | + | $(+4)$ | + | $(+1)$ | + | $(+3)$ | + | $(+1)$ |
| | 81 | = | 74 | + | $(+4)$ | + | $(+1)$ | + | $(+3)$ | + | $(-1)$ |
| | 82 | = | 74 | + | $(+4)$ | + | $(+1)$ | + | $(+3)$ | + | $(0)$ |
| $SS$: | $SS_{total}$ | = | $SS_{\mu}$ | + | $SS_A$ | + | $SS_B$ | + | $SS_{AB}$ | + | $SS_{S(AB)}$ |
| | $66,040$ | = | $65,712$ | + | 192 | + | 12 | + | 108 | + | 16 |

Table 4.9: Decomposition Matrix

After computing the sums of squares, the next step is to calculate the degrees of freedom. The concept of degrees of freedom is the same as it was in the one-way design: the *df* for a term is the number of independent values of that term. Thus, the *df* for the mean is still 1, because there is one value of $\mu$. The *df* for the main effect of any factor is the number of levels of that factor minus 1. As in the one-way design, we estimate one value of the main effect term for each level of the factor, but one degree of freedom is lost because the values have to add up to zero across levels.

Similarly, the *df* for the $AB$ interaction corresponds to the number of independent values of $AB_{ij}$. Now we computed four different values for this term: $AB_{11}$, $AB_{12}$, $AB_{21}$, and $AB_{22}$. But there is really only one independent value, because the values have to add up to 0 across rows and columns. Note that once we compute the value of $AB_{ij}$ for any one cell, we can fill in the values for the other cells by knowing that the interaction terms must sum to zero for every row and column.

There is a convenient fact about interaction *df*'s that will save much thought about how many interaction terms are independent. The *df* for the interaction of Factors A and B is always the product of $df_A$ and $df_B$. Thus, once we figure out the *df*'s for the main effects, we can easily get the interaction *df*'s by multiplication. Knowing this fact saves an enormous amount of time, especially when we get to more complicated designs.

The *df* for $S(AB)$–symbolized $df_{S(AB)}$–can also be found by extending the logic of the one-way design. In the one-way design, we noted that the degrees of freedom for the subjects term was equal to the number of subjects minus the number of groups. This was because we estimated one value of the subjects term for each subject, but these terms were constrained to add up to 0 within each group. This is also true in the two-way design. Thus, within each cell all of the $S(AB)$'s but one are independent. In short, the $df_{S(AB)}$ is again equal to the total number of subjects minus the number of groups (i.e., $12 - 4 = 8$). Note the number of groups in a 2-way between-Ss design is the number of cells in the design, because a different group of subjects is tested in each cell.

Once we get $SS$'s and *df*'s, of course, the rest of the ANOVA summary table is easy to fill in, as shown in Table 4.10. A $MS$ is always the corresponding $SS$ divided by the corresponding *df*. Observed $F$ ratios are obtained for each source by taking that source's $MS$ divided by the $MS_{S(AB)}$. These $F_{observed}$'s are compared against the appropriate $F_{critical}$'s from the table of $F_{critical}$'s, and the null hypothesis associated with each source is rejected if $F_{observed} > F_{critical}$.

In this case, the $F_{critical}$ associated with 95% confidence is 5.32, so we can reject the null hypothesis associated with each $F_{observed}$:

- $32,856 > 5.32$ so reject the $H_0$: True mean test score = 0.
- $96 > 5.32$ so reject the $H_0$: True means equal for male and female students.
- $6 > 5.32$ so reject the $H_0$: True means equal for male and female teachers.
- $54 > 5.32$ so reject the $H_0$: True difference between male/female teachers is the same for male as for female students (or vice versa).

| Source | *df* | *SS* | *MS* | *F* | Error Term |
|---|---|---|---|---|---|
| $\mu$ | 1 | 65,712 | 65,712 | 32,856 | $S(AB)$ |
| $A$ (Student) | 1 | 192 | 192 | 96 | $S(AB)$ |
| $B$ (Teacher) | 1 | 12 | 12 | 6 | $S(AB)$ |
| $AB$ | 1 | 108 | 108 | 54 | $S(AB)$ |
| $S(AB)$ | 8 | 16 | 2 | | |
| Total | 12 | 66,040 | | | |

Table 4.10: ANOVA Summary Table for Data in Table 4.6

## 4.5 Drawing Conclusions About Two-Factor Designs

Practical interpretations of experimental results can begin after the above ANOVA is done. It is always useful to refer to a factorial plot when drawing conclusions from the ANOVA, because the ANOVA only tells whether an effect is real; it does not indicate the direction or size of the effect.[5] Figure 4.1 shows the factorial plot for the data set analyzed in the preceding section, which we now proceed to interpret.

Because the main effect of Gender of Teacher was significant (i.e., $F_{observed} > F_{critical}$), we can say that, on average, students studying a foreign language under a female teacher score higher on standardized language achievement tests than students studying under male teachers. The observed overall difference in our data was fairly small, but the $F$ test allows us to be 95% sure that it was not simply due to chance.

Because the main effect of Gender of Student was also significant, we can conclude that female students really do score higher on average than male students after a year of language study. Our observed difference was roughly 10%, and the statistical test allows us to be 95% sure that the difference is more than would be expected by chance.

The earlier discussion of interaction emphasized that the presence of an interaction complicates the conclusions that can be drawn about main effects, and that, in the presence of an interaction, one must be cautious. Thus, the finding of a significant interaction complicates the interpretation of both of the above main effects. Technically, the significant interaction forces us to conclude that the difference between male and female students is not the same for male teachers as for female teachers, and vice versa. Looking at the data, we see that in fact this difference is smaller for male teachers than for female ones, and this should be stated explicitly as one of the conclusions of the study.[6] The statistical test is necessary to tell us that this pattern, apparent in the factorial plot, was not simply obtained by chance.

It is worth emphasizing that the interaction only allows us to draw conclusions about *a difference in effects*, not conclusions about particular cells in the design. For example, from looking at the cell means it seems reasonable to conclude that female students learn better from female teachers than from male teachers, but the analysis does *not* justify this conclusion. Remember, the ANOVA made three comparisons: male vs. female teachers overall, male vs. female students overall, and difference between male and female students with male teachers vs. difference between male and female students with female teachers. So, none of the comparisons we have made directly addressed the specific question of whether female students learn better from female teachers than from male teachers. Further analysis would be required to answer this question, using multiple comparison techniques, as mentioned in connection with the one-way designs. The simplest paired-comparison technique, for example, is to compute a one-way ANOVA to see whether there is a significant effect of Gender of Teacher using only data from female students.

It is important to be explicit about the various kinds of tempting conclusions that are **not** justified by the ANOVA. Another conclusion that we have not established is that there is no real difference between male and female students under the condition with male teachers. Aside from the consideration that we can never fully establish null hypotheses, the problem here is that we have not even conducted an explicit test of this hypothesis. Again, this hypothesis requires a comparison between means of individual cells, and we have only performed overall hypothesis tests involving all cells. Paired-comparison methods are needed to answer questions like this.

[5] In fact, R. A. Fisher himself—the inventor of the $F$ test—once said that ANOVA is nothing more than a tool to be used in deciding whether the patterns we see in graphs are real or due to chance.

[6] In this study the interaction may also be summarized neatly by saying that there is a real tendency for students to learn better from a teacher of the same gender than from a teacher of the opposite gender.

Apart from the comparison of specific cells, another sort of tempting but unjustified conclusion concerns the relative sizes of the effects of Gender of Teacher and Gender of Student. It certainly appears from that results that scores are more influenced by the gender of the student than by the gender of the teacher. However, we have not made any explicit comparisons between the two effects, so we cannot say for sure that this was not just a chance finding. Again, other methods are needed to answer questions of this kind.

## 4.6 Summary of Computations

**Model:** $Y_{ijk} = \mu + A_i + B_j + AB_{ij} + S(AB)_{ijk}$

**Estimation Equations:**

$$\begin{array}{lcl}
\hat{\mu} & = & Y_{...} \\
\hat{A}_i & = & Y_{i..} - \hat{\mu} \\
\hat{B}_i & = & Y_{.j.} - \hat{\mu} \\
\widehat{AB}_{ij} & = & Y_{ij.} - \hat{\mu} - \hat{A}_i - \hat{B}_j \\
\widehat{S(AB)}_{ijk} & = & Y_{ijk} - \hat{\mu} - \hat{A}_i - \hat{B}_j - \widehat{AB}_{ij}
\end{array}$$

**Decomposition Matrix:**

$$\begin{array}{lcccccccccc}
Y_{ijk} & = & \hat{\mu} & + & \hat{A}_i & + & \hat{B}_j & + & \widehat{AB}_{ij} & + & \widehat{S(AB)}_{ijk} \\
Y_{111} & = & \hat{\mu} & + & \hat{A}_1 & + & \hat{B}_1 & + & \widehat{AB}_{11} & + & \widehat{S(AB)}_{111} \\
Y_{112} & = & \hat{\mu} & + & \hat{A}_1 & + & \hat{B}_1 & + & \widehat{AB}_{11} & + & \widehat{S(AB)}_{112} \\
\vdots & & & & & & & & & &
\end{array}$$

**Degrees of Freedom:**

- *df* for $\mu$= 1.
- *df* for main effect = Number of levels - 1.
- *df* for interaction = product of *df*'s for main effects.
- *df* for $S(AB)$ = Number of subjects - Number of groups.

**ANOVA Table:** Obtain the $SS$ for a term by computing the total of the squared values in the corresponding column of decomposition matrix.

| Source | *df* | $SS$ | $MS$ | $F$ | Error Term |
|---|---|---|---|---|---|
| $\mu$ | | | | | $S(AB)$ |
| $A$ | | | | | $S(AB)$ |
| $B$ | | | | | $S(AB)$ |
| $AB$ | | | | | $S(AB)$ |
| $S(AB)$ | | | | | |
| Total | | | | | |

**Interpretation:** The only general guideline for interpretation is provided by a list of the null hypotheses tested.

$F_\mu$**:** $H_0$: The overall mean is 0.

$F_A$**:** $H_0$: Overall, the levels of Factor A have the same averages.

$F_B$**:** $H_0$: Overall, the levels of Factor B have the same averages.

$F_{AB}$**:** $H_0$: The effect of Factor A is the same at every level of Factor B (and vice versa).

## 4.7 Relationships Between One- and Two-Factor ANOVA

To broaden our understanding of how two-way ANOVA decomposes the overall variability into variability due to various terms, it is instructive to analyze the previous experiment as if it were a one-way design with four groups, as shown in Table 4.11. This is not at all legitimate from an experimentalist's viewpoint, since we know that the groups really constitute a 2 by 2 factorial. Nonetheless, if we pretend briefly that we regard the four groups as four levels of a single experimental factor, we can compute a one-way ANOVA with these numbers. It is instructive to compare the results of such a one-way analysis with the results of the appropriate two-way analysis that we have already carried out.

| | Group 1 | Group 2 | Group 3 | Group 4 |
|---|---|---|---|---|
| Data: | 73, 73, 70 | 67, 68, 69 | 73, 76, 73 | 83, 81, 82 |
| Mean: | 72 | 68 | 74 | 82 |

Decomposition Matrix:

| $Y_{ij}$ | = | $\hat{\mu}$ | + | $\hat{A}_i$ | + | $\widehat{S(A)}_{ij}$ |
|---|---|---|---|---|---|---|
| 73 | = | 74 | + | (-2) | + | (+1) |
| 73 | = | 74 | + | (-2) | + | (+1) |
| 70 | = | 74 | + | (-2) | + | (-2) |
| 67 | = | 74 | + | (-6) | + | (-1) |
| 68 | = | 74 | + | (-6) | + | ( 0) |
| 69 | = | 74 | + | (-6) | + | (+1) |
| 73 | = | 74 | + | ( 0) | + | (-1) |
| 76 | = | 74 | + | ( 0) | + | (+2) |
| 73 | = | 74 | + | ( 0) | + | (-1) |
| 83 | = | 74 | + | (+8) | + | (+1) |
| 81 | = | 74 | + | (+8) | + | (-1) |
| 82 | = | 74 | + | (+8) | + | ( 0) |

| Source | *df* | *SS* | *MS* | *F* | Error Term |
|---|---|---|---|---|---|
| $\mu$ | 1 | 65,712 | 65,712 | 32,856 | $S(A)$ |
| $A$ | 3 | 312 | 104 | 52 | $S(A)$ |
| $S(A)$ | 8 | 16 | 2 | | |
| Total | 12 | 66,040 | | | |

Table 4.11: Reanalysis of Table 4.6 Data

Using the estimation equations associated with the one-way design, we obtain the decomposition matrix and ANOVA table shown in Table 4.11. We will compare these with the decomposition matrix and ANOVA table computed in the previous section, looking particularly at parameter estimates, *SS*'s, and *df*'s.

First, note some similarities. Since the same data scores are being analyzed, $SS_{total}$'s and $df_{total}$'s are the same for the one- and two-way designs. This also leads to the same $\hat{\mu}$ in the two designs, so the $SS_{\mu}$'s are the same. Finally, identical values for the within-cell error terms ($S(A)$ and $S(AB)$ are obtained in the two designs, so the error *SS*'s ($SS_{S(A)}$ and $SS_{S(AB)}$) are the same. This makes sense, because within cell error is measured as the deviation of a score from its group mean in both designs. With respect to this measurement, it does not matter whether the groups constitute multiple levels of a single factor or a factorial crossing of two (or more) factors.

The most interesting relationship between these two designs has to do with the way the differences between groups are broken down, as shown in Table 4.12.

| | One-Way | | Two-Way | | | | |
|---|---|---|---|---|---|---|---|
| Effects: | $A_i$ | = | $A_i$ | + | $B_j$ | + | $AB_{ij}$ |
| Sums of Squares: | $SS_A$ | = | $SS_A$ | + | $SS_B$ | + | $SS_{AB}$ |
| Degrees of Freedom: | $df_A$ | = | $df_A$ | + | $df_B$ | + | $df_{AB}$ |

Table 4.12: Group Breakdowns for One- and Two-Way Designs

Contrasting the linear models, it is apparent that the term $A_i$ in the one-way model must account for all the differences among the four groups. In the two-way design, differences among groups can be accounted for with three terms: $A_i$, $B_j$, and $AB_{ij}$. The two-way design thus divides the overall difference among groups into more components than does the one-way design. That is, the one-way design simply has differences among groups, but the two-way design specifies to what extent the groups are different because of 1) differences between male and female students, 2) differences between male and female teachers, and 3) differences due to an interaction between Gender of Student and Gender of Teacher.

One point of this comparison between the two designs is that the two-way design is more informative because it divides an overall between-groups difference into several meaningful terms. A second point is that there is a close mathematical relationship between the two analyses. This relationship further illustrates the idea that ANOVA breaks up overall variation into variation due to different factors within a particular experimental design. If, for some reason, we were to ignore some of the factors in an experimental design, the variation they produced would be lumped together and called by a different name. In a sense, the total variance is determined by the data, and the goal of the data analyst is to divide it up into terms that are as meaningful as possible.

As long as we are considering different ways to analyze the Gender of Teacher by Gender of Student design, it is worth mentioning that there are two other legitimate ways to conceptualize and analyze the experiment. One is shown in Table 4.13, and the other is analogous except the roles of teacher and student are reversed. In Table 4.13, it is clear that there are two levels of the Gender of Teacher factor, but these two levels have been labeled relative to the Gender of Student factor. It is arguable whether this is a peculiar way to think of the Gender of Teacher factor, but it is certainly not wrong. If we had been motivated to do the experiment by some theory about teachers serving as role models for students, for instance, we might have said that the experimental factor of primary interest was whether the student was the same or opposite gender as the teacher. In any case, this way of conceptualizing the experimental design leads to the exact same $MS$'s and $F$'s as our original conceptualization, and is therefore no less valid statistically. The major difference is semantic. What we were previously calling the Gender of Teacher (male vs. female) by Gender of Student (male vs. female) interaction is now called the main effect of whether the teacher and student are the same or different genders. What we were previously calling the effect of Gender of Teacher (male vs. female) is now called the interaction of Gender of Student (male vs. female) with Gender of Teacher (same vs. different).

| | | Gender of Teacher: Same as Student | Gender of Teacher: Different From Student |
|---|---|---|---|
| Gender of Student: | Male | 73, 73, 70 | 67, 68, 69 |
| | Female | 83, 81, 82 | 73, 76, 73 |

| Source | *df* | *SS* | *MS* | *F* | Error Term |
|---|---|---|---|---|---|
| $\mu$ | 1 | 65,712 | 65,712 | 32,856 | $S(AB)$ |
| $A$ (Student, M vs. F) | 1 | 192 | 192 | 96 | $S(AB)$ |
| $B$ (Teacher, S vs. D) | 1 | 108 | 108 | 54 | $S(AB)$ |
| $AB$ | 1 | 12 | 12 | 6 | $S(AB)$ |
| $S(AB)$ | 8 | 16 | 2 | | |
| Total | 12 | 66,040 | | | |

Table 4.13: Reconceptualization of Teacher Gender X Student Gender Design

This aspect of two-way ANOVA is worth some consideration because it helps to illustrate a critical point that students of ANOVA often miss: The point of view of the experimenter often determines how a given data set is analyzed. People working in a particular scientific area generally have points of view that are sufficiently similar that they would analyze the same data in the same way. This is neither always nor necessarily the case, though, as we can see if we imagine someone regarding the Gender of Teacher by Gender of Student design as in Table 4.13. It is worth remembering that statistical techniques are computational (and sometimes conceptual) tools. Like all tools, these techniques require the user to have particular plans and goals in mind. Students tend to come away from the study of ANOVA with the idea that ANOVA dictates a single correct method of data analysis for any experiment. From an epistemological point of view, this is slightly naive. It is analogous to

thinking that the structure of our language dictates a single correct form for any statement we want to make, or even that screwdrivers and wrenches determine a single optimal way to put together a car.

## 4.8 Review Questions about Two-Factor ANOVA

1. What three null hypotheses is a two-factor ANOVA designed to test?
2. What is an interaction?
3. What is the equation to estimate interaction terms in a two-way design?
4. What is the significance of an interaction for interpreting main effects?

## 4.9 Exercises on Two-Way Between-Ss ANOVA

For each of the data sets described below, construct a table of cell and marginal means, a factorial plot, and do the appropriate ANOVA. Interpret the results, using 95% confidence for all hypothesis tests.

**1.** An experimenter wanted to ascertain the effects of Personality Type (Introvert vs. Extrovert) and Type of Motivation (Praise/Blame) on performance of a complicated motor task. Personality Type was determined with a questionnaire given at the beginning of the experiment, and half of the people in each personality type were randomly assigned to each motivation condition. In the praise condition, people were told they were doing very well on the motor task, and in the blame condition they were told they were doing very poorly. These were the results (higher scores mean better performance):

| | Type of Motivation (A) | |
|---|---|---|
| | Praise | Blame |
| Introvert | 48, 39, 48 | 31, 37, 25 |
| Extrovert | 29, 45, 37 | 47, 51, 43 |

**2.** A navy researcher was interested in the effects of Alcohol and Sleep Loss on performance in a submarine control room. Experienced operators were tested with or without each of these stressors, and the number of errors over the test session was recorded. These are the results:

| | | Sleep Loss (A) | | |
|---|---|---|---|---|
| | | No | Yes | Average |
| Alcohol | No | 9, 11 | 15, 17 | 13 |
| (B) | Yes | 21, 23 | 39, 41 | 31 |
| | Average | 16 | 28 | 22 |

**3.** An educational psychologist was interested in the effects of three types of study strategies on learning by students at various class levels. Twelve freshman, twelve sophomores, twelve juniors, and twelve seniors taking the same course were randomly assigned to the three types of study strategies, and their final exam scores were obtained. Here are the results:

| | Study Strategy (A) | | |
|---|---|---|---|
| Class Level: | Music On | TV On | Book Under Pillow |
| Freshman | 32, 41, 36, 39 | 20, 27, 28, 17 | 24, 12, 20, 16 |
| Sophomore | 27, 24, 28, 37 | 28, 30, 27, 19 | 20, 20, 22, 42 |
| Junior | 31, 22, 23, 28 | 27, 24, 20, 17 | 37, 32, 34, 17 |
| Senior | 26, 18, 25, 11 | 10, 16, 7, 3 | 36, 36, 41, 23 |

## 4.10 Answers to Exercises

**1.** Personality and Motivation.

**Design:**

- Factor A: Method of Motivation. 2 levels (Praise/Blame), Between-Ss.
- Factor B: Type of Personality. 2 levels (Intro/Extro), Between-Ss.
- 3 Subjects per group.

**Model:** $Y_{ijk} = \mu + A_i + B_j + AB_{ij} + S(AB)_{ijk}$

**Estimation Equations:**

$$\begin{array}{lcllllllll} \hat{\mu} & = & Y_{...} & & & & & & & \\ \hat{A}_i & = & Y_{i..} & - & \hat{\mu} & & & & & \\ \hat{B}_i & = & Y_{.j.} & - & \hat{\mu} & & & & & \\ \widehat{AB}_{ij} & = & Y_{ij.} & - & \hat{\mu} & - & \hat{A}_i & - & \hat{B}_j & \\ S\widehat{(AB)}_{ijk} & = & Y_{ijk} & - & \hat{\mu} & - & \hat{A}_i & - & \hat{B}_j & - & \widehat{AB}_{ij} \end{array}$$

**Decomposition Matrix:**

| $Y_{ijk}$ | = | $\hat{\mu}$ | + | $\hat{A}_i$ | + | $\hat{B}_j$ | + | $\widehat{AB}_{ij}$ | + | $S\widehat{(AB)}_{ijk}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 48 | = | 40 | + | 1 | - | 2 | + | 6 | + | 3 |
| 39 | = | 40 | + | 1 | - | 2 | + | 6 | - | 6 |
| 48 | = | 40 | + | 1 | - | 2 | + | 6 | + | 3 |
| 31 | = | 40 | - | 1 | - | 2 | - | 6 | + | 0 |
| 37 | = | 40 | - | 1 | - | 2 | - | 6 | + | 6 |
| 25 | = | 40 | - | 1 | - | 2 | - | 6 | - | 6 |
| 29 | = | 40 | + | 1 | + | 2 | - | 6 | - | 8 |
| 45 | = | 40 | + | 1 | + | 2 | - | 6 | + | 8 |
| 37 | = | 40 | + | 1 | + | 2 | - | 6 | + | 0 |
| 47 | = | 40 | - | 1 | + | 2 | + | 6 | + | 0 |
| 51 | = | 40 | - | 1 | + | 2 | + | 6 | + | 4 |
| 43 | = | 40 | - | 1 | + | 2 | + | 6 | - | 4 |

**ANOVA Table:**

| Source | *df* | *SS* | *MS* | *F* | Error Term |
|---|---|---|---|---|---|
| $\mu$ | 1 | 19,200 | 19,200.0 | 537.06 | $S(AB)$ |
| $A$ | 1 | 12 | 12.0 | 0.34 | $S(AB)$ |
| $B$ | 1 | 48 | 48.0 | 1.34 | $S(AB)$ |
| $AB$ | 1 | 432 | 432.0 | 12.08 | $S(AB)$ |
| $S(AB)$ | 8 | 286 | 35.8 | | |
| Total | 12 | 19,978 | | | |

**Interpretation:** Based on the $F_{critical}$ with 1 and 8 *df*, we see that only the baseline and $AB$ interaction have significant effects. Regarding the former, we can conclude that the mean performance is greater than zero, which is not too interesting. (This will be the last example in which we mention an *uninteresting* baseline.) Regarding the latter, we can conclude that the effect of Motivation Type is different for Introverts than Extroverts. Looking back at the cell means or factorial plot, we can conclude even more specifically that the difference (Praise minus Blame) is greater for Introverts than Extroverts.

**2.** Sleep Loss and Alcohol.

**Design:**

- Factor A: Sleep loss. 2 levels (No/Yes), Between-Ss.
- Factor B: Alcohol. 2 levels (No/Yes), Between-Ss.
- 2 Subjects per group.

**Model:** Same as previous exercise.

**Estimation Equations:** Same as previous exercise.

**Decomposition Matrix:**

| $Y_{ijk}$ | = | $\hat{\mu}$ | + | $\hat{A}_i$ | + | $\hat{B}_j$ | + | $\widehat{AB}_{ij}$ | + | $S\widehat{(AB)}_{ijk}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 9 | = | 22 | - | 6 | - | 9 | + | 3 | - | 1 |
| 11 | = | 22 | - | 6 | - | 9 | + | 3 | + | 1 |
| 15 | = | 22 | + | 6 | - | 9 | - | 3 | - | 1 |
| 17 | = | 22 | + | 6 | - | 9 | - | 3 | + | 1 |
| 21 | = | 22 | - | 6 | + | 9 | - | 3 | - | 1 |
| 23 | = | 22 | - | 6 | + | 9 | - | 3 | + | 1 |
| 39 | = | 22 | + | 6 | + | 9 | + | 3 | - | 1 |
| 41 | = | 22 | + | 6 | + | 9 | + | 3 | + | 1 |

**ANOVA Table:**

| Source | *df* | *SS* | *MS* | *F* | Error Term |
|---|---|---|---|---|---|
| $\mu$ | 1 | 3,872 | 3,872 | 1936 | $S(AB)$ |
| $A$ | 1 | 288 | 288 | 144 | $S(AB)$ |
| $B$ | 1 | 648 | 648 | 324 | $S(AB)$ |
| $AB$ | 1 | 72 | 72 | 36 | $S(AB)$ |
| $S(AB)$ | 4 | 8 | 2 | | |
| Total | 8 | 4888 | | | |

**Interpretation:** Both alcohol and sleep loss tend to increase errors. Furthermore, the increase caused by both together is more than the sum of the two increases produced by either stressors.

**3.** Class Level and Study Strategy.

**Design:**

- Factor A: Study Strategy. 3 levels (Music/TV/Pillow), Between-Ss.
- Factor B: Class Level. 4 levels (Fresh/Soph/Junior/Senior), Between-Ss.
- 4 Subjects per group.

**Cell Means:**

| | Study Strategy (Factor A) | | | |
|---|---|---|---|---|
| Class Level (B): | Music On | TV On | Book Under Pillow | Average |
| Freshman | 37 | 23 | 18 | 26 |
| Sophomore | 29 | 26 | 26 | 27 |
| Junior | 26 | 22 | 30 | 26 |
| Senior | 20 | 9 | 34 | 21 |
| Average | 28 | 20 | 27 | 25 |

**Model:** Same as previous exercises.

**Estimation Equations:** Same as previous exercises.

**Decomposition Matrix:**

| $Y_{ijk}$ | = | $\hat{\mu}$ | + | $\hat{A}_i$ | + | $\hat{B}_j$ | + | $\widehat{AB}_{ij}$ | + | $S\widehat{(AB)}_{ijk}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 32 | = | 25 | + | 3 | + | 1 | + | 8 | - | 5 |
| 41 | = | 25 | + | 3 | + | 1 | + | 8 | + | 4 |
| 36 | = | 25 | + | 3 | + | 1 | + | 8 | - | 1 |
| 39 | = | 25 | + | 3 | + | 1 | + | 8 | + | 2 |
| 20 | = | 25 | - | 5 | + | 1 | + | 2 | - | 3 |
| 27 | = | 25 | - | 5 | + | 1 | + | 2 | + | 4 |
| 28 | = | 25 | - | 5 | + | 1 | + | 2 | + | 5 |
| 17 | = | 25 | - | 5 | + | 1 | + | 2 | - | 6 |
| 24 | = | 25 | + | 2 | + | 1 | - | 10 | + | 6 |
| 12 | = | 25 | + | 2 | + | 1 | - | 10 | - | 6 |
| 20 | = | 25 | + | 2 | + | 1 | - | 10 | + | 2 |
| 16 | = | 25 | + | 2 | + | 1 | - | 10 | - | 2 |
| 27 | = | 25 | + | 3 | + | 2 | - | 1 | - | 2 |
| 24 | = | 25 | + | 3 | + | 2 | - | 1 | - | 5 |
| 28 | = | 25 | + | 3 | + | 2 | - | 1 | - | 1 |
| 37 | = | 25 | + | 3 | + | 2 | - | 1 | + | 8 |
| 28 | = | 25 | - | 5 | + | 2 | + | 4 | + | 2 |
| 30 | = | 25 | - | 5 | + | 2 | + | 4 | + | 4 |
| 27 | = | 25 | - | 5 | + | 2 | + | 4 | + | 1 |
| 19 | = | 25 | - | 5 | + | 2 | + | 4 | - | 7 |
| 20 | = | 25 | + | 2 | + | 2 | - | 3 | - | 6 |
| 20 | = | 25 | + | 2 | + | 2 | - | 3 | - | 6 |
| 22 | = | 25 | + | 2 | + | 2 | - | 3 | - | 4 |
| 42 | = | 25 | + | 2 | + | 2 | - | 3 | + | 16 |
| 31 | = | 25 | + | 3 | + | 1 | - | 3 | + | 5 |
| 22 | = | 25 | + | 3 | + | 1 | - | 3 | - | 4 |
| 23 | = | 25 | + | 3 | + | 1 | - | 3 | - | 3 |
| 28 | = | 25 | + | 3 | + | 1 | - | 3 | + | 2 |
| 27 | = | 25 | - | 5 | + | 1 | + | 1 | + | 5 |
| 24 | = | 25 | - | 5 | + | 1 | + | 1 | + | 2 |
| 20 | = | 25 | - | 5 | + | 1 | + | 1 | - | 2 |
| 17 | = | 25 | - | 5 | + | 1 | + | 1 | - | 5 |
| 37 | = | 25 | + | 2 | + | 1 | + | 2 | + | 7 |
| 32 | = | 25 | + | 2 | + | 1 | + | 2 | + | 2 |
| 34 | = | 25 | + | 2 | + | 1 | + | 2 | + | 4 |
| 17 | = | 25 | + | 2 | + | 1 | + | 2 | - | 13 |
| 26 | = | 25 | + | 3 | - | 4 | - | 4 | + | 6 |
| 18 | = | 25 | + | 3 | - | 4 | - | 4 | - | 2 |
| 25 | = | 25 | + | 3 | - | 4 | - | 4 | + | 5 |
| 11 | = | 25 | + | 3 | - | 4 | - | 4 | - | 9 |
| 10 | = | 25 | - | 5 | - | 4 | - | 7 | + | 1 |
| 16 | = | 25 | - | 5 | - | 4 | - | 7 | + | 7 |
| 7 | = | 25 | - | 5 | - | 4 | - | 7 | - | 2 |
| 3 | = | 25 | - | 5 | - | 4 | - | 7 | - | 6 |
| 36 | = | 25 | + | 2 | - | 4 | + | 11 | + | 2 |
| 36 | = | 25 | + | 2 | - | 4 | + | 11 | + | 2 |
| 41 | = | 25 | + | 2 | - | 4 | + | 11 | + | 7 |
| 23 | = | 25 | + | 2 | - | 4 | + | 11 | - | 11 |

**ANOVA Table:**

| Source | *df* | *SS* | *MS* | *F* | Error Term |
|---|---|---|---|---|---|
| $\mu$ | 1 | 30,000 | 30,000.0 | 727.76 | $S(AB)$ |
| $A$ | 2 | 608 | 304.0 | 7.37 | $S(AB)$ |
| $B$ | 3 | 264 | 88.0 | 2.13 | $S(AB)$ |
| $AB$ | 6 | 1,576 | 262.7 | 6.37 | $S(AB)$ |
| $S(AB)$ | 36 | 1,484 | 41.2 | | |
| Total | 48 | 33,932 | | | |

**Interpretation:** The significant $F_{observed}$ for A indicates that the three study strategies are not all equivalent, and the significant $F_{observed}$ for AB indicates that the relative pattern across the three study strategies is not the same for all class levels. These conclusions are a bit vague, but they are exactly what is implied by the $F$'s. Looking back at the marginal means for A, we see that the marginal means for Music-On and Book-Under-Pillow are about the same, both much higher than TV-On. Because we know there is a real effect here, this pattern makes it clear what the effect is: It is worst to study with the TV on. Because the interaction is significant, we should also look back at the cell means or factorial plot. Unfortunately, there is no convenient summary of all the cells which seems to describe the results. It appears that Freshman do especially well studying with Music On and especially poorly with the Book Under Pillow strategy, whereas Seniors do especially well with the Book Under Pillow and especially badly with the TV on (maybe they watch different shows). But we cannot conclude any of these things for sure without special paired comparison techniques, so we must stop with the overall conclusion that the pattern of strategy effects changes across class levels. Of course, we can present the factorial plot together with this conclusion and let the readers make their own inferences about what caused the interaction.

# Chapter 5

# Three-Factor, Between-Subject Designs

## 5.1 A More General Conceptual Overview of ANOVA

Before moving on to more complicated experimental designs, it is useful to step back and consolidate the concepts developed so far. In both one- and two-way designs, the linear model provides a way to express data values as the sum of a number of components:

- One-Way Model: $Y_{ij} = \mu + A_i + S(A)_{ij}$
- Two-Way Model: $Y_{ijk} = \mu + A_i + B_j + AB_{ij} + S(AB)_{ijk}$

There is an important sense in which the terms of the model are *independent descriptors* of the data. They are *independent* in that successive terms in the model (going from left to right) measure new effects beyond what has already been measured. Main effects, for example, are measured as deviations from the baseline. Interactions are deviations from cell means expected on the basis of the baseline and the main effects. Because each term is measured relative to previous terms in the model, we say that the successive terms are independent. In other words, each term is used to describe a property of the data not described by any previous term in the model; the terms are non-redundant.

Terms in the model can be regarded as *descriptors* of the data in the sense that they provide a way of rewriting the data. We have used the linear model to form a decomposition matrix, in which each data score is written as the sum of a number of terms. It is possible to write out the whole set of data simply by listing the actual scores (the numbers to the left of the equals signs in the decomposition matrix), and it is also possible to write out the data set by listing the values of the various terms in the linear model (the numbers to the right of the equals signs in the decomposition matrix). Since the terms in the linear model are chosen so that each data value is exactly the sum of the terms associated with it, the linear model is always "true" for every data point. This the sense in which the linear model provides us with a way of describing or rewriting the data.

What good is the linear model, if it only rewrites the data for us? The main advantage of the linear model is that it rewrites the data in terms of independent and meaningful components: baseline, main effects, interactions, and error. For example, in the one-way design, there is a term in the model explicitly designed to capture the differences between groups (i.e., the effect of the factor). In the two-way design, there is one term for each of the two experimental factors. The main effect terms in the two-way design are especially useful, because they measure the effect of one factor averaged over the levels of the other factor (i.e., independent of the second factor).

As we saw in the two-way design, it is not always possible to describe the data just in terms of $\mu$, $A_i$, $B_j$, and $S(AB)_{ijk}$. An additional term, namely $AB_{ij}$, is needed to the extent that an interaction is present (i.e., the effect of one factor depends on the level of the other factor). In other words, an $AB$ interaction term might be needed to describe the data, because the cell means might not be exactly predicted by the baseline, the main effect of $A_{\cdot}$ and the main effect of $B_{\cdot}$.

For example, in a three by three design the cell means can always be written as shown in Table 5.1. The parameter estimates will always add up to the exact cell mean. However, if the AB term were dropped from the model, it would be easy to come up with data sets for which no possible assignment of numbers to $\hat{\mu}$, $\hat{A}_i$, and $\hat{B}_j$ could reproduce the cell means. Thus, we need the interaction term to describe the data.

| | $B_1$ | $B_2$ | $B_3$ |
|---|---|---|---|
| $A_1$ | $\hat{\mu}+\hat{A}_1+\hat{B}_1+\widehat{AB}_{11}$ | $\hat{\mu}+\hat{A}_1+\hat{B}_2+\widehat{AB}_{12}$ | $\hat{\mu}+\hat{A}_1+\hat{B}_3+\widehat{AB}_{13}$ |
| $A_2$ | $\hat{\mu}+\hat{A}_2+\hat{B}_1+\widehat{AB}_{21}$ | $\hat{\mu}+\hat{A}_2+\hat{B}_2+\widehat{AB}_{32}$ | $\hat{\mu}+\hat{A}_2+\hat{B}_3+\widehat{AB}_{23}$ |
| $A_3$ | $\hat{\mu}+\hat{A}_3+\hat{B}_1+\widehat{AB}_{31}$ | $\hat{\mu}+\hat{A}_3+\hat{B}_2+\widehat{AB}_{32}$ | $\hat{\mu}+\hat{A}_3+\hat{B}_3+\widehat{AB}_{33}$ |

Table 5.1: Model for Cell Means in 3 x 3 Design

The point of the above discussion is that various logically different types of effects can be present in a two-factor design, and our model must allow us to measure all of them by including corresponding terms. As we consider experimental designs with more factors, we will be forced to add more and more complicated terms to the model to measure all of the different sorts of interactions that can be present. These terms will be correspondingly more difficult to interpret, but that is not a flaw in the model. It simply reflects the fact that the real world can produce very complicated patterns of data.

## 5.2 ANOVA Computations for Three-Factor Between-Ss Designs

Having studied between-subjects designs with one and two experimental factors, we need only slightly generalize our concepts to deal with between-subjects designs involving three or more factors. These designs are not much more complex than the two-way design computationally, and yet the addition of a third factor can lead to real difficulties in interpreting the results.

As an example design, we will work with a hypothetical experiment concerning the influence of gender on sentences recommended by jurors. Imagine that the subjects are jurors in mock trials, and exactly the same evidence is presented concerning the defendant's guilt in all conditions of the experiment. At the end of the trial each juror is asked to recommend a prison sentence for the defendant (0-15 years), and this number is the dependent variable. The experimental factors are the Gender of the Juror (A), the Gender of the Defendant (B), and the Gender of the Prosecuting Attorney (C). Hypothetical data from this three-factor experiment are shown in Table 5.2, with four observations per cell.

| | | Male Prosecutor: Defendant: Male | Male Prosecutor: Defendant: Female |
|---|---|---|---|
| Juror: | Male | 10, 11, 10, 11 | 5, 4, 5, 4 |
| | Female | 4, 4, 5, 4 | 5, 5, 5, 6 |

| | | Female Prosecutor: Defendant: Male | Female Prosecutor: Defendant: Female |
|---|---|---|---|
| Juror: | Male | 3, 4, 4, 3 | 5, 5, 6, 6 |
| | Female | 5, 5, 4, 5 | 6, 5, 6, 6 |

Table 5.2: Sample Data For Three-Way Between-Ss Design

### 5.2.1 Model

The first step in the analysis is to construct the linear model, which looks like this for a three-way, between-subjects design:

$$Y_{ijkl} = \mu + A_i + B_j + AB_{ij} + C_k + AC_{ik} + BC_{jk} + ABC_{ijk} + S(ABC)_{ijkl}$$

where,

$i$ indicates the level of A (1 denotes male juror, 2 denotes female).

$j$ indicates the level of B (1 denotes male defendant, 2 denotes female).

$k$ indicates the level of C (1 denotes male prosecutor, 2 denotes female).

$l$ indicates the subject within cell (1, 2, 3, or 4).

As in the one- and two-way designs, the model has terms for the overall baseline and for the main effect of each experimental factor ($A_i$, $B_j$, and $C_k$).[1] We also see three different two-way interaction terms ($AB_{ij}$, $AC_{ik}$, and $BC_{jk}$), because there are three pairs of factors that might produce special effects in combination. Specifically, the effect of Defendant Gender may depend on Prosecutor Gender, the effect of Defendant Gender may depend on Juror Gender, or the effect of Prosecutor Gender may depend on Juror Gender (and vice versa for each of these). The other familiar looking term in the model is the subject error term at the end. As usual, the model allows for some error associated with each score.

The new term in the linear model for the three-way design is the three-way interaction term, $ABC_{ijk}$. A three-factor interaction can be understood as an extension of the concept of a two-factor interaction in any of several ways. First, recall that one way to think about a two-factor interaction term was as a measurement of the extent to which the results depended on the particular pair of levels being tested. Under this conception of a two-factor interaction, it measured the extent to which a cell mean differed from what would be predicted given the baseline and the two main effects. Similarly, the three-factor interaction can be thought of as measuring the extent to which the results depend on a particular combination of all three factors. In other words, the three-factor interaction measures how much the cell means differ from what would be expected given the baseline, the main effects, and the two-factor interactions.

Another conception of the two-factor interaction was the extent to which one factor has an effect that depends on the level of a second factor. This conception provides a more concrete way to view the three-factor interaction: the extent to which a particular two-factor interaction depends on the level of a third factor. For example, we could think about the two-factor interaction between Gender of Juror and Gender of Prosecutor. Based on our observed results, we could plot this interaction twice: once for male defendants and once for female defendants. The question is whether these two plots show the same two-way interaction. If they do (i.e., if the two-factor interaction is the same at both levels of the third factor), then there is no three-way interaction. But if the plots show a different pattern (i.e., if the two-factor interaction is different for male defendants than for female defendants), then there is a three-way interaction.

### 5.2.2 Estimation

Estimation in the three-way design is a simple extension of estimation in the two-way design, and we can now begin to formulate a general rule for estimation in any design. The exact estimation equations are shown in Table 5.3.

| | | |
|---|---|---|
| $\hat{\mu}$ | = | $Y_{....}$ |
| $\hat{A}_i$ | = | $Y_{i...}$-$\hat{\mu}$ |
| $\hat{B}_j$ | = | $Y_{.j..}$-$\hat{\mu}$ |
| $\widehat{AB}_{ij}$ | = | $Y_{ij..}$-$\hat{\mu}$-$\hat{A}_i$-$\hat{B}_j$ |
| $\hat{C}_k$ | = | $Y_{..k.}$-$\hat{\mu}$ |
| $\widehat{AC}_{ik}$ | = | $Y_{i.k.}$-$\hat{\mu}$-$\hat{A}_i$-$\hat{C}_k$ |
| $\widehat{BC}_{jk}$ | = | $Y_{.jk.}$-$\hat{\mu}$-$\hat{B}_j$-$\hat{C}_k$ |
| $\widehat{ABC}_{ijk}$ | = | $Y_{ijk.}$-$\hat{\mu}$-$\hat{A}_i$-$\hat{B}_j$-$\widehat{AB}_{ij}$- $\hat{C}_k$-$\widehat{AC}_{ik}$-$\widehat{BC}_{jk}$ |
| $S(\widehat{ABC})_{ijkl}$ | = | $Y_{ijkl}$-$\hat{\mu}$-$\hat{A}_i$-$\hat{B}_j$-$\widehat{AB}_{ij}$- $\hat{C}_k$-$\widehat{AC}_{ik}$-$\widehat{BC}_{jk}$-$\widehat{ABC}_{ijk}$ |
| | = | $Y_{ijkl}$ - $Y_{ijk.}$ |

Table 5.3: Estimation Equations for Three-Way Between-Ss ANOVA

The general pattern of the estimation process works from left to right across the linear model, just as it did in the one- and two-way designs. To estimate $\mu$, we use the grand mean across all the data. The only change from the simpler designs is that the grand mean is obtained by averaging across more experimental factors. These factors have their own terms in the model, however, so the existence of these factors does not change the estimate of $\mu$.

[1] The terms in a linear model can be written in any order. In this course, we will use an ordering principle designed to make it easy to construct the model, described in section 6.1.1. This principle does not keep all the main effects together, all the two-way interactions together, and so on. That is a price we will pay.

As in the one- and two-way designs, the main effect term associated with a particular level of a factor is estimated as the difference between the average of all scores at that level of the factor minus the baseline. The only change from the simpler designs is that there are more factors to average across to get the mean for the specified level.

The estimates of the two-factor interaction terms are also computed just as they were in the two-way design, except that there is again an extra subscript to average over. Again, the two-factor interaction effect associated with a particular pair of levels is estimated by taking the mean at that combination of levels, and then subtracting the baseline and the main effect terms corresponding to those two levels. As discussed above, a change from the two-way design is that now we have three two-way interactions instead of just one, because there are three possible pairs of factors.

In a three-way design, there is an estimated three-way interaction term for each cell in the design. The estimated value for a cell is the cell mean minus that which is explained by the baseline, the main effects, and the two-way interactions. Whatever is not explained by those terms is a special effect due to the three-way combination of levels at that cell. As is discussed extensively below, it can be interpreted as a change in one of the two-factor interactions depending on the level of the third factor.

Finally, an estimate of $S(ABC)_{ijkl}$ is obtained for each data value. As in previous designs, this error term is estimated as the score minus all those parts of the score explained by other terms in the model. Again, one can think of the error as whatever the model needs to come out right. A computationally simpler way to get the estimates of $S(ABC)_{ijkl}$ is to take the difference between each score and its cell mean. We will not emphasize this method, however, because it does not generalize well to within-subjects designs. The notion of error as something unexplained by other terms in the model generalizes better.

Before attempting to use the estimation equations to break down the individual scores, it is useful to calculate all the cell and marginal means, as shown in Table 5.4. It is easiest to begin by computing the cell means. From these, compute all the possible two-way tables of marginal means by averaging out one factor at a time. For example, the marginal table for the juror by defendant interaction is obtained by averaging across male and female prosecutors at each possible combination of gender of juror and gender of defendant. After we have all the two-way tables, we compute the marginal means for the main effects and the overall mean.

Cell Means: $(Y_{ijk.})$

| | | Male Prosecutors: Defendant: Male | Male Prosecutors: Defendant: Female | | | Female Prosecutors: Defendant: Male | Female Prosecutors: Defendant: Female |
|---|---|---|---|---|---|---|---|
| Juror: | Male | 10.50 | 4.50 | Juror: | Male | 3.50 | 5.50 |
| | Female | 4.25 | 5.25 | | Female | 4.75 | 5.75 |

Marginal Means:

$Y_{ij..}$

| | | Defendant: Male | Defendant: Female |
|---|---|---|---|
| Jur: | Male | 7.0 | 5.0 |
| | Fem. | 4.5 | 5.5 |

$Y_{i.k.}$

| | | Prosecutor: Male | Prosecutor: Female |
|---|---|---|---|
| Jur: | Male | 7.50 | 4.50 |
| | Fem. | 4.75 | 5.25 |

$Y_{.jk.}$

| | | Prosecutor: Male | Prosecutor: Female |
|---|---|---|---|
| Def: | Male | 7.375 | 4.125 |
| | Fem. | 4.875 | 5.625 |

| | | |
|---|---|---|
| Jurors: | $Y_{i...}$ | Male = 6.000, Female = 5.000 |
| Defendants: | $Y_{.j..}$ | Male = 5.750, Female = 5.250 |
| Prosecutors: | $Y_{..k.}$ | Male = 6.125, Female = 4.875 |
| Grand Mean: | $Y_{....}$ | 5.500 |

Table 5.4: Cell and Marginal Means for Mock Jury Experiment

### 5.2.3 Decomposition Matrix and $SS$'s

With the estimation equations and the cell and marginal means, it is easy (though tedious) to do all of the computation necessary to arrive at the decomposition matrix shown in Table 5.5.

Once we have the decomposition matrix, we can start to write down the ANOVA table, shown in Table 5.6. As always, there is one source line for each term in the linear model. The sum of squares for each term can be computed by squaring and totaling the corresponding column of the decomposition matrix, as always.

| $Y_{ijkl}$ | = | $\hat{\mu}$ | + | $\hat{A}_i$ | + | $\hat{B}_j$ | + | $\hat{C}_{ij}$ | + | $\widehat{AB}_k$ | + | $\widehat{AC}_{ik}$ | + | $\widehat{BC}_{jk}$ | + | $\widehat{ABC}_{ijk}$ | + | $S(\widehat{ABC})_{ijkl}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 10 | = | 5.5 | + | 0.5 | + | 0.25 | + | 0.625 | + | 0.75 | + | 0.875 | + | 1 | + | 1 | − | 0.5 |
| 11 | = | 5.5 | + | 0.5 | + | 0.25 | + | 0.625 | + | 0.75 | + | 0.875 | + | 1 | + | 1 | + | 0.5 |
| 10 | = | 5.5 | + | 0.5 | + | 0.25 | + | 0.625 | + | 0.75 | + | 0.875 | + | 1 | + | 1 | − | 0.5 |
| 11 | = | 5.5 | + | 0.5 | + | 0.25 | + | 0.625 | + | 0.75 | + | 0.875 | + | 1 | + | 1 | + | 0.5 |
| 3 | = | 5.5 | + | 0.5 | + | 0.25 | − | 0.625 | + | 0.75 | − | 0.875 | − | 1 | − | 1 | − | 0.5 |
| 4 | = | 5.5 | + | 0.5 | + | 0.25 | − | 0.625 | + | 0.75 | − | 0.875 | − | 1 | − | 1 | + | 0.5 |
| 4 | = | 5.5 | + | 0.5 | + | 0.25 | − | 0.625 | + | 0.75 | − | 0.875 | − | 1 | − | 1 | + | 0.5 |
| 3 | = | 5.5 | + | 0.5 | + | 0.25 | − | 0.625 | + | 0.75 | − | 0.875 | − | 1 | − | 1 | − | 0.5 |
| 5 | = | 5.5 | + | 0.5 | − | 0.25 | + | 0.625 | − | 0.75 | + | 0.875 | − | 1 | − | 1 | + | 0.5 |
| 4 | = | 5.5 | + | 0.5 | − | 0.25 | + | 0.625 | − | 0.75 | + | 0.875 | − | 1 | − | 1 | − | 0.5 |
| 5 | = | 5.5 | + | 0.5 | − | 0.25 | + | 0.625 | − | 0.75 | + | 0.875 | − | 1 | − | 1 | + | 0.5 |
| 4 | = | 5.5 | + | 0.5 | − | 0.25 | + | 0.625 | − | 0.75 | + | 0.875 | − | 1 | − | 1 | − | 0.5 |
| 5 | = | 5.5 | + | 0.5 | − | 0.25 | − | 0.625 | − | 0.75 | − | 0.875 | + | 1 | + | 1 | − | 0.5 |
| 5 | = | 5.5 | + | 0.5 | − | 0.25 | − | 0.625 | − | 0.75 | − | 0.875 | + | 1 | + | 1 | − | 0.5 |
| 6 | = | 5.5 | + | 0.5 | − | 0.25 | − | 0.625 | − | 0.75 | − | 0.875 | + | 1 | + | 1 | + | 0.5 |
| 6 | = | 5.5 | + | 0.5 | − | 0.25 | − | 0.625 | − | 0.75 | − | 0.875 | + | 1 | + | 1 | + | 0.5 |
| 4 | = | 5.5 | − | 0.5 | + | 0.25 | + | 0.625 | − | 0.75 | − | 0.875 | + | 1 | − | 1 | − | 0.25 |
| 4 | = | 5.5 | − | 0.5 | + | 0.25 | + | 0.625 | − | 0.75 | − | 0.875 | + | 1 | − | 1 | − | 0.25 |
| 5 | = | 5.5 | − | 0.5 | + | 0.25 | + | 0.625 | − | 0.75 | − | 0.875 | + | 1 | − | 1 | + | 0.75 |
| 4 | = | 5.5 | − | 0.5 | + | 0.25 | + | 0.625 | − | 0.75 | − | 0.875 | + | 1 | − | 1 | − | 0.25 |
| 5 | = | 5.5 | − | 0.5 | + | 0.25 | − | 0.625 | − | 0.75 | + | 0.875 | − | 1 | + | 1 | + | 0.25 |
| 5 | = | 5.5 | − | 0.5 | + | 0.25 | − | 0.625 | − | 0.75 | + | 0.875 | − | 1 | + | 1 | + | 0.25 |
| 4 | = | 5.5 | − | 0.5 | + | 0.25 | − | 0.625 | − | 0.75 | + | 0.875 | − | 1 | + | 1 | − | 0.75 |
| 5 | = | 5.5 | − | 0.5 | + | 0.25 | − | 0.625 | − | 0.75 | + | 0.875 | − | 1 | + | 1 | + | 0.25 |
| 5 | = | 5.5 | − | 0.5 | − | 0.25 | + | 0.625 | + | 0.75 | − | 0.875 | − | 1 | + | 1 | − | 0.25 |
| 5 | = | 5.5 | − | 0.5 | − | 0.25 | + | 0.625 | + | 0.75 | − | 0.875 | − | 1 | + | 1 | − | 0.25 |
| 5 | = | 5.5 | − | 0.5 | − | 0.25 | + | 0.625 | + | 0.75 | − | 0.875 | − | 1 | + | 1 | − | 0.25 |
| 6 | = | 5.5 | − | 0.5 | − | 0.25 | + | 0.625 | + | 0.75 | − | 0.875 | − | 1 | + | 1 | + | 0.75 |
| 6 | = | 5.5 | − | 0.5 | − | 0.25 | − | 0.625 | + | 0.75 | + | 0.875 | + | 1 | − | 1 | + | 0.25 |
| 5 | = | 5.5 | − | 0.5 | − | 0.25 | − | 0.625 | + | 0.75 | + | 0.875 | + | 1 | − | 1 | − | 0.75 |
| 6 | = | 5.5 | − | 0.5 | − | 0.25 | − | 0.625 | + | 0.75 | + | 0.875 | + | 1 | − | 1 | + | 0.25 |
| 6 | = | 5.5 | − | 0.5 | − | 0.25 | − | 0.625 | + | 0.75 | + | 0.875 | + | 1 | − | 1 | + | 0.25 |

Table 5.5: Decomposition Matrix for Mock Jury Experiment

### 5.2.4 Degrees of Freedom

To fill in the column for degrees of freedom, we need only reapply the rules already learned.

- The *df* for the mean is 1.
- The *df* for any main effect is the number of levels of the factor minus 1.
- The *df* for any interaction (including three-factor) is found by multiplying together the *df*'s for the factors involved in the interaction. This multiplication rule holds true even for the higher-way interactions: multiply together the *df*'s for *all* the factors involved in the interaction.
- The *df* for $S(ABC)$ can be found with either of two methods. Since the total *df* is known to equal the number of observations, the *df* for $S(ABC)$ can be found by subtraction once the rest of the *df*'s are known. Alternatively, this *df* can be found using the formula:

$$df_{S(ABC)} = \text{Number of subjects - Number of groups}$$

As in the two-factor between-Ss design, the number of groups is the number of cells in the design.

### 5.2.5 ANOVA Table

Computation of $MS$'s and $F$'s is easy once we get this far. The $MS$ for any line of the table is the $SS$ for that line divided by the *df* for that line. The $F$ for any line is the $MS$ for that line divided by $MS_{S(ABC)}$. All that remains is to look up the $F_{critical}$ for the appropriate *df*'s—in this case 1 and 24, yielding 4.26. Now the computations are done, and the remaining problem is interpretation of the results.

| Source | *df* | *SS* | *MS* | *F* | Error Term |
|---|---|---|---|---|---|
| $\mu$ | 1 | 968.0 | 968.0 | 3318.8 | $S(ABC)$ |
| $A$ (Juror) | 1 | 8.0 | 8.0 | 27.4 | $S(ABC)$ |
| $B$ (Defendant) | 1 | 2.0 | 2.0 | 6.857 | $S(ABC)$ |
| $AB$ | 1 | 18.0 | 18.0 | 61.7 | $S(ABC)$ |
| $C$ (Prosecutor) | 1 | 12.5 | 12.5 | 42.8 | $S(ABC)$ |
| $AC$ | 1 | 24.5 | 24.5 | 84.0 | $S(ABC)$ |
| $BC$ | 1 | 32.0 | 32.0 | 109.7 | $S(ABC)$ |
| $ABC$ | 1 | 32.0 | 32.0 | 109.7 | $S(ABC)$ |
| $S(ABC)$ | 24 | 7.0 | 0.2917 | | |
| Total | 32 | 1104.0 | | | |

Table 5.6: ANOVA Table for Mock Jury Experiment

## 5.3 Interpreting Three-Factor ANOVAs

The general approach to interpreting the results is to proceed down the ANOVA table, saying what each significant effect or interaction means. The initial interpretation can be general, like "there is an effect of Factor A" or "the effect of Factor A depends on the level of Factor B." A more specific interpretation should then be developed by examining the data and stating, in the general terms of the ANOVA tests, exactly what the effect is and/or how it depends on another factor. We illustrate this approach below by discussing the data analyzed in Section 5.2. Of course, it is often desirable to perform further analyses in order to make more specific statements about the results.

The significant main effects of Gender of Juror, Gender of Defendant, and Gender of Prosecutor indicate that the observed differences in marginal means for these factors were not simply due to chance. Looking at the means, then, we can interpret these $F$'s as indicating that: 1) On average, male jurors give longer sentences than female jurors, 2) On average, male defendants get longer sentences than female defendants, and 3) On average, longer sentences are given when the prosecutor is male than female. The "on average" qualification is critically important, however. Significant higher-order interactions indicate that the pattern of results is more complicated than we can describe using simple statements about main effects. Given these interactions, it is certain that the effects of some factors depended on the levels of some other factors, so any overall statements about *the* effects of these factors will be oversimplified.

To interpret two-way interactions properly, it is advisable to look at the marginal means or factorial plots associated with the two-way interactions. Each significant interaction says that, on average, the effect of one factor depends on the level of the other factor, but that statement is simply too general to be an interesting conclusion. Note that the qualification "on average" crops up again. There were three factors in the experiment, and each two-way interaction is an average across the levels of the third factor. Since the three-way interaction was significant, though, we know that each two-way interaction changes depending on the level of the third factor. Any conclusion about the interaction must take this averaging into account. Just as we were forced to be cautious in our interpretations of main effects if there were significant two-way interactions, so too are we forced to qualify our interpretations of the two-way interactions when there is a significant three-way interaction.

In practice, it is much easier to understand an interaction when it is graphed than when it is presented in a table of means. The three panels in Figure 5.1 show graphs of the means associated with each of the two-way interactions observed in this experiment, and our interpretations are really little more than descriptions of these graphs.

The Gender of Juror by Gender of Defendant interaction in Panel A arises because sentences are longer overall when the defendant and juror are the same gender than when they are opposite genders. It appears that, overall, male jurors give much longer sentences to male defendants than to female defendants, while female jurors are slightly more severe with female defendants than with male ones. Alternatively, one could view the interaction as male defendants getting longer sentences, overall, from male jurors than from female jurors, while female defendants get longer sentences, overall, from female jurors than male jurors. Neither of these two conclusions is strictly justified without more specific analyses, because both make statements about specific means that are really only part of the total interaction (e.g., "male jurors give longer sentences to male defendants than to female defendants"). To come up with a proper interpretation, the conclusion must somehow refer to the whole interaction, as did the original conclusion relying on same vs. different gender.

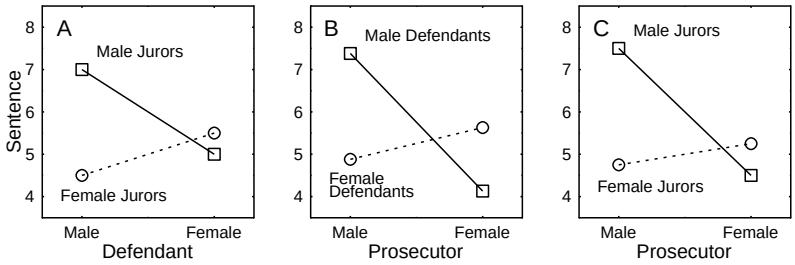


Figure 5.1: Sentence as a function of genders of A) defendant and juror; B) defendant and prosecutor; and C) prosecutor and juror.

The other two two-way interactions also reflect an overall tendency toward longer sentences in the same-gender combinations than in the opposite-gender combinations, as can be seen by studying Panels B and C. Conclusions about these interactions would be analogous to those drawn from the Gender of Juror by Gender of Defendant interaction.

The interaction that is really difficult to interpret is the three-way. Two different strategies are available for trying to understand the nature of the interaction, but neither one is guaranteed to give a really clear picture. Sometimes, you end up just concluding that there is something complicated going on that you can't understand.[2]

If you do use one strategy and find that it leads you to a satisfactory interpretation, it is still worthwhile to try the other strategy. After all, interactions can be looked at in many different ways. The more different points of view we take in considering an interaction, the better chance we have of fully understanding all its implications.

### 5.3.1 Strategy 1: How does a two-way interaction change?

The first strategy is to focus your point of view on a particular two-way interaction (any one you like– you can try it from all points of view to see which makes the most sense). Graph this two-way interaction separately for each level of the third factor. For example, Figure 5.2 shows the Gender of Juror by Gender of Defendant interaction graphed separately for male and female prosecutors.

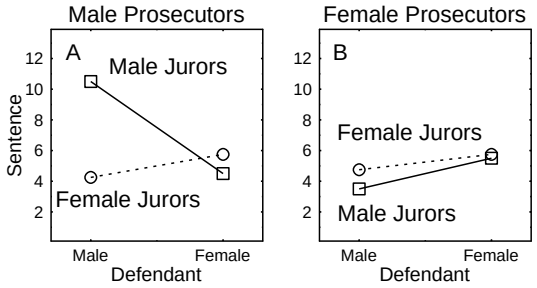


Figure 5.2: Sentence as a function of defendant and juror genders, separately for male and female prosecutors.

The significant three-way interaction indicates that the interaction is not the same in these two graphs. Can you see the difference? In Figure 5.2 it appears that there is a bigger difference in slopes

[2]This conclusion is not all that uncommon in published scientific articles. Usually the article will have contributed some new piece of information with a clear theoretical significance, but there will be other aspects of the data that remain unexplained.

for the male prosecutors than for the female prosecutors. Notice also that the Gender of Juror by Gender of Defendant interactions go in opposite directions with male and female prosecutors. With male prosecutors, the observed average sentences were longer when the juror and defendant were the same gender, while with female prosecutors the observed average sentences were longer when the juror and the defendant were opposite genders. While we cannot conclude that the true means exactly follow the pattern of the observed averages (note that we have made statements comparing particular cells, rather than sticking to a summary of the total interaction pattern), we can conclude that the interaction of Gender of Juror with Gender of Defendant is different when the prosecutor is male than when the prosecutor is female.

That last conclusion might make sense to someone schooled in ANOVA, but what if we wanted to interpret the three-way interaction for a layperson? We might be able to communicate the idea by saying that, with male prosecutors more than with female prosecutors, sentences are longer when the juror and defendant are the same gender as opposed to opposite genders. You can illustrate this statement with a table like 5.7. The value in the table is the size of the Gender of Juror by Gender of Defendant interaction: average sentences with (same gender juror and defendant) - average sentences with (opposite gender juror and defendant). Clearly, this difference is larger with male than female prosecutors.

| Prosecutor | Size of Interaction |
|---|---|
| Male: | $\frac{10.5+5.75}{2} - \frac{4.5+4.25}{2} = 3.75$ |
| Female: | $\frac{3.5+5.75}{2} - \frac{5.5+4.75}{2} = -0.5$ |

Table 5.7: Effect of Prosecutor on Juror By Defendant Interaction

We may be able to make other interesting statements about these results by repeating this process focusing on one of the other pairwise interactions.

### 5.3.2 Strategy 2: How does a main effect change?

Another way to look at a three-way interaction is to focus on a main effect, and see how that main effect depends on the interaction of the other two factors. This new strategy looks at one particular main effect as a function of the factorial design defined by the other two factors.

For example, look at the difference between sentences given to male and female defendants (say, male minus female). From the cell means in the upper part of Table 5.4, we can compute this difference for each possible combination of male and female jurors and prosecutors. (For example, the difference for male jurors with a male prosecutor is $10.50 - 4.50 = 6.00$.) A graph of the results is shown in Figure 5.3. The vertical axis measures the size of the Gender of Defendant effect, which we might be tempted to call a "bias" against male defendants, since the experiment involved identical evidence in all conditions. The Gender of Defendant effect ("anti-male bias," if you will) is shown for each gender of juror and prosecutor.

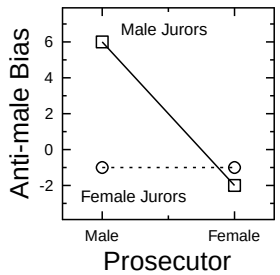


Figure 5.3: Bias against male defendants as a function of juror and prosecutor genders.

Inspecting Figure 5.3 gives a slightly new view on the three-way interaction, and shows how we

should qualify our interpretation of the overall Gender of Defendant effect (males get longer sentences than females). First of all, note that the significant overall main effect of Gender of Defendant corresponds to the fact that the average of the four points plotted on this figure is above 0. Interestingly, though the average is above 0, only one of the four points is actually above 0. Thus the overall anti-male defendant bias is the result of a big bias in one condition rather than a consistent bias across all conditions. Furthermore, we can see that Gender of Juror and Gender of Prosecutor produce both main effects and an interactive effect on this bias. The fact that the bias is bigger overall for male jurors than for female jurors reflects the Gender of Juror by Gender of Defendant interaction. The fact that the bias is bigger overall with male prosecutors than with female prosecutors reflects the Gender of Prosecutor by Gender of Defendant interaction.

### 5.3.3 Summary

It is clear by now that three-way interactions can be very difficult to understand and describe intuitively. It may be especially difficult to figure out how to *start* looking at a three-way interaction. Is it better to start with strategy 1 or 2? Whichever strategy you use, how do you know which interaction or main effect to focus on?

There are no general statistical answers to such questions. In some situations the experimenter will have a strong theoretical interest in some particular main effect or interaction, and that will determine which point of view to adopt. In other situations there will be no particular reason to adopt one point of view rather than another. In both cases the careful researcher will look at an interaction from several, maybe even all possible points of view. The only cost of trying a new point of view is researcher time, and the potential payoff is an important insight into what is going on in the experiment. If that potential payoff is not exciting enough to justify the extra effort, then the researcher ought to look for an easier way to make a living.

## 5.4 Exercises: Three-Factor, Between-Ss ANOVA

**1.** A nutritionist was interested in the determining the benefit produced by having different combinations of three different amino acids (A, B, and C) in the diet. He randomly divided 24 rats into the 8 groups shown below. Each rat was fed a diet including the indicated combination of amino acids, and the life span of the rat was measured, with the results shown in the table. Perform the ANOVA and interpret the results.

| | B Absent: | | B Present: | |
|---|---|---|---|---|
| | A Absent | A Present | A Absent | A Present |
| C Absent: | 104 104 152 | 60 102 54 | 91 81 38 | 114 90 126 |
| C Present: | 51 87 72 | 114 130 74 | 134 115 99 | 122 118 168 |

**2.** The educational psychologist interested in three study strategies (see exercises for Two-Way Between-Ss Designs) decided to add Student Gender as a third factor (Factor A: 1 = Female, 2 = Male) in a repetition of the study with two subjects per cell. The new results are shown below; what do they suggest?

| | A1 | | | | | | | | A2 | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | C1 | | C2 | | C3 | | C4 | | C1 | | C2 | | C3 | | C4 | |
| B1: | 5 | 5 | 5 | 7 | 8 | 6 | 8 | 12 | 6 | 0 | 6 | 2 | 5 | 5 | 11 | 5 |
| B2: | 13 | 11 | 13 | 17 | 18 | 16 | 25 | 23 | 4 | 4 | 2 | 4 | 7 | 7 | 9 | 3 |
| B3: | 13 | 13 | 17 | 13 | 7 | 11 | 24 | 22 | 12 | 10 | 5 | 5 | 8 | 10 | 22 | 16 |

## 5.5 Answers to Exercises

### 1.

**Design:**

- 2 x 2 x 2, Between-Ss, with 3 subjects per group.
- Factor A = absence/presence of amino acid a in diet.
- Factor B = absence/presence of amino acid b in diet.
- Factor C = absence/presence of amino acid c in diet.
- DV= Length of life of rat, in months.

**Cell Means:**

| | B Absent: | | B Present: | |
|---|---|---|---|---|
| | A Absent | A Present | A Absent | A Present |
| C Absent: | 120 | 72 | 70 | 110 |
| C Present: | 70 | 106 | 116 | 136 |

**Marginal Means:**

Averaged across A:

| | B Absent | B Present | Average |
|---|---|---|---|
| C Absent: | 96 | 90 | 93 |
| C Present: | 88 | 126 | 107 |
| Average: | 92 | 108 | 100 |

Averaged across B:

| | A Absent | A Present | Average |
|---|---|---|---|
| C Absent: | 95 | 91 | 93 |
| C Present: | 93 | 121 | 107 |
| Average: | 94 | 106 | 100 |

Averaged across C:

| | A Absent | A Present | Average |
|---|---|---|---|
| B Absent: | 95 | 89 | 92 |
| B Present: | 93 | 123 | 108 |
| Average: | 94 | 106 | 100 |

**Estimation equations:** See Table 5.3.

**Decomposition Matrix:**

| $Y_{ijkl}$ = | $\hat{\mu}$ | + | $\hat{A}_i$ | + | $\hat{B}_j$ | + | $\widehat{AB}_{ij}$ | + | $\hat{C}_k$ | + | $\widehat{AC}_{ik}$ | + | $\widehat{BC}_{jk}$ | + | $\widehat{ABC}_{ijk}$ | + | $S(\widehat{ABC})_{ijkl}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 104 = | 100 | - | 6 | - | 8 | + | 9 | - | 7 | + | 8 | + | 11 | + | 13 | - | 16 |
| 104 = | 100 | - | 6 | - | 8 | + | 9 | - | 7 | + | 8 | + | 11 | + | 13 | - | 16 |
| 152 = | 100 | - | 6 | - | 8 | + | 9 | - | 7 | + | 8 | + | 11 | + | 13 | + | 32 |
| 60 = | 100 | + | 6 | - | 8 | - | 9 | - | 7 | - | 8 | + | 11 | - | 13 | - | 12 |
| 102 = | 100 | + | 6 | - | 8 | - | 9 | - | 7 | - | 8 | + | 11 | - | 13 | + | 30 |
| 54 = | 100 | + | 6 | - | 8 | - | 9 | - | 7 | - | 8 | + | 11 | - | 13 | - | 18 |
| 91 = | 100 | - | 6 | + | 8 | - | 9 | - | 7 | + | 8 | - | 11 | - | 13 | + | 21 |
| 81 = | 100 | - | 6 | + | 8 | - | 9 | - | 7 | + | 8 | - | 11 | - | 13 | + | 11 |
| 38 = | 100 | - | 6 | + | 8 | - | 9 | - | 7 | + | 8 | - | 11 | - | 13 | - | 32 |
| 114 = | 100 | + | 6 | + | 8 | + | 9 | - | 7 | - | 8 | - | 11 | + | 13 | + | 4 |
| 90 = | 100 | + | 6 | + | 8 | + | 9 | - | 7 | - | 8 | - | 11 | + | 13 | - | 20 |
| 126 = | 100 | + | 6 | + | 8 | + | 9 | - | 7 | - | 8 | - | 11 | + | 13 | + | 16 |
| 51 = | 100 | - | 6 | - | 8 | + | 9 | + | 7 | - | 8 | - | 11 | - | 13 | - | 19 |
| 87 = | 100 | - | 6 | - | 8 | + | 9 | + | 7 | - | 8 | - | 11 | - | 13 | + | 17 |
| 72 = | 100 | - | 6 | - | 8 | + | 9 | + | 7 | - | 8 | - | 11 | - | 13 | + | 2 |
| 114 = | 100 | + | 6 | - | 8 | - | 9 | + | 7 | + | 8 | - | 11 | + | 13 | + | 8 |
| 130 = | 100 | + | 6 | - | 8 | - | 9 | + | 7 | + | 8 | - | 11 | + | 13 | + | 24 |
| 74 = | 100 | + | 6 | - | 8 | - | 9 | + | 7 | + | 8 | - | 11 | + | 13 | - | 32 |
| 134 = | 100 | - | 6 | + | 8 | - | 9 | + | 7 | - | 8 | + | 11 | + | 13 | + | 18 |
| 115 = | 100 | - | 6 | + | 8 | - | 9 | + | 7 | - | 8 | + | 11 | + | 13 | - | 1 |
| 99 = | 100 | - | 6 | + | 8 | - | 9 | + | 7 | - | 8 | + | 11 | + | 13 | - | 17 |
| 122 = | 100 | + | 6 | + | 8 | + | 9 | + | 7 | + | 8 | + | 11 | - | 13 | - | 14 |
| 118 = | 100 | + | 6 | + | 8 | + | 9 | + | 7 | + | 8 | + | 11 | - | 13 | - | 18 |
| 168 = | 100 | + | 6 | + | 8 | + | 9 | + | 7 | + | 8 | + | 11 | - | 13 | + | 32 |

**ANOVA:**

| Source | *df* | *SS* | *MS* | *F* | Error Term |
|---|---|---|---|---|---|
| $\mu$ | 1 | 240000.0 | 240000.0 | 398.423 | $S(ABC)$ |
| $A$ | 1 | 864.0 | 864.0 | 1.434 | $S(ABC)$ |
| $B$ | 1 | 1536.0 | 1536.0 | 2.550 | $S(ABC)$ |
| $AB$ | 1 | 1944.0 | 1944.0 | 3.227 | $S(ABC)$ |
| $C$ | 1 | 1176.0 | 1176.0 | 1.952 | $S(ABC)$ |
| $AC$ | 1 | 1536.0 | 1536.0 | 2.550 | $S(ABC)$ |
| $BC$ | 1 | 2904.0 | 2904.0 | 4.821 | $S(ABC)$ |
| $ABC$ | 1 | 4056.0 | 4056.0 | 6.733 | $S(ABC)$ |
| $S(ABC)$ | 16 | 9638.0 | 602.4 | | |
| Total | 24 | 263654.0 | | | |

**Interpretation:** First look at the significant BC interaction. It is clear that there is a mutual facilitation when both these amino acids are present, because there is one value in the BC marginal means that is much higher than all the others (126). (Technically, we do need to do paired comparisons to establish this, but the data are so clear-cut that in practice few would quarrel with the conclusion.)

Given that the BC interaction was significant, it seems reasonable to start by looking at the ABC interaction in terms of how the BC interaction changes across levels of A. Here, we find a pattern that is startling in terms of how we believe amino acids work. With A present, there is a big mutual facilitation of B and C, just as we concluded from looking at the simple BC interaction. But with A absent, the largest effect is that you should have either both or neither of B and C, but not one or the other. This suggests that there is something positively harmful about having just one of the three amino acids. We don't know if that is a realistic possibility or not, but that is what these (fake) data suggest.

**2. Design:** 2 x 3 x 4, Between-Ss, with 2 Ss per group.

**Cell Means:**

| | A1 | | | | A2 | | | |
|---|---|---|---|---|---|---|---|---|
| | C1 | C2 | C3 | C4 | C1 | C2 | C3 | C4 |
| B1 | 5 | 6 | 7 | 10 | 3 | 4 | 5 | 8 |
| B2 | 12 | 15 | 17 | 24 | 4 | 3 | 7 | 6 |
| B3 | 13 | 15 | 9 | 23 | 11 | 5 | 9 | 19 |

**Marginal Means:**

| | A1 | A2 |
|---|---|---|
| B1 | 7 | 5 |
| B2 | 17 | 5 |
| B3 | 15 | 11 |

| | C1 | C2 | C3 | C4 |
|---|---|---|---|---|
| A1 | 10 | 12 | 11 | 19 |
| A2 | 6 | 4 | 7 | 11 |

| | C1 | C2 | C3 | C4 |
|---|---|---|---|---|
| B1 | 4 | 5 | 6 | 9 |
| B2 | 8 | 9 | 12 | 15 |
| B3 | 12 | 10 | 9 | 21 |

**Decomposition Matrix:**

| $Y_{ijkl}$ | = | $\hat{\mu}$ | + | $\hat{A}_i$ | + | $\hat{B}_j$ | + | $\widehat{AB}_{ij}$ | + | $\hat{C}_k$ | + | $\widehat{AC}_{ik}$ | + | $\widehat{BC}_{jk}$ | + | $\widehat{ABC}_{ijk}$ | + | $S(\widehat{ABC})_{ijkl}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 5 | = | 10 | + | 3 | - | 4 | - | 2 | - | 2 | - | 1 | + | 0 | + | 1 | + | 0 |
| 5 | = | 10 | + | 3 | - | 4 | - | 2 | - | 2 | - | 1 | + | 0 | + | 1 | + | 0 |
| 6 | = | 10 | - | 3 | - | 4 | + | 2 | - | 2 | + | 1 | + | 0 | - | 1 | + | 3 |
| 0 | = | 10 | - | 3 | - | 4 | + | 2 | - | 2 | + | 1 | + | 0 | - | 1 | - | 3 |
| 13 | = | 10 | + | 3 | + | 1 | + | 3 | - | 2 | - | 1 | - | 1 | - | 1 | + | 1 |
| 11 | = | 10 | + | 3 | + | 1 | + | 3 | - | 2 | - | 1 | - | 1 | - | 1 | - | 1 |
| 4 | = | 10 | - | 3 | + | 1 | - | 3 | - | 2 | + | 1 | - | 1 | + | 1 | + | 0 |
| 4 | = | 10 | - | 3 | + | 1 | - | 3 | - | 2 | + | 1 | - | 1 | + | 1 | + | 0 |
| 13 | = | 10 | + | 3 | + | 3 | - | 1 | - | 2 | - | 1 | + | 1 | + | 0 | + | 0 |
| 13 | = | 10 | + | 3 | + | 3 | - | 1 | - | 2 | - | 1 | + | 1 | + | 0 | + | 0 |
| 12 | = | 10 | - | 3 | + | 3 | + | 1 | - | 2 | + | 1 | + | 1 | + | 0 | + | 1 |
| 10 | = | 10 | - | 3 | + | 3 | + | 1 | - | 2 | + | 1 | + | 1 | + | 0 | - | 1 |
| 5 | = | 10 | + | 3 | - | 4 | - | 2 | - | 2 | + | 1 | + | 1 | - | 1 | - | 1 |
| 7 | = | 10 | + | 3 | - | 4 | - | 2 | - | 2 | + | 1 | + | 1 | - | 1 | + | 1 |
| 6 | = | 10 | - | 3 | - | 4 | + | 2 | - | 2 | - | 1 | + | 1 | + | 1 | + | 2 |
| 2 | = | 10 | - | 3 | - | 4 | + | 2 | - | 2 | - | 1 | + | 1 | + | 1 | - | 2 |
| 13 | = | 10 | + | 3 | + | 1 | + | 3 | - | 2 | + | 1 | + | 0 | - | 1 | - | 2 |
| 17 | = | 10 | + | 3 | + | 1 | + | 3 | - | 2 | + | 1 | + | 0 | - | 1 | + | 2 |
| 2 | = | 10 | - | 3 | + | 1 | - | 3 | - | 2 | - | 1 | + | 0 | + | 1 | - | 1 |
| 4 | = | 10 | - | 3 | + | 1 | - | 3 | - | 2 | - | 1 | + | 0 | + | 1 | + | 1 |
| 17 | = | 10 | + | 3 | + | 3 | - | 1 | - | 2 | + | 1 | - | 1 | + | 2 | + | 2 |
| 13 | = | 10 | + | 3 | + | 3 | - | 1 | - | 2 | + | 1 | - | 1 | + | 2 | - | 2 |
| 5 | = | 10 | - | 3 | + | 3 | + | 1 | - | 2 | - | 1 | - | 1 | - | 2 | + | 0 |
| 5 | = | 10 | - | 3 | + | 3 | + | 1 | - | 2 | - | 1 | - | 1 | - | 2 | + | 0 |
| 8 | = | 10 | + | 3 | - | 4 | - | 2 | - | 1 | - | 1 | + | 1 | + | 1 | + | 1 |
| 6 | = | 10 | + | 3 | - | 4 | - | 2 | - | 1 | - | 1 | + | 1 | + | 1 | - | 1 |
| 5 | = | 10 | - | 3 | - | 4 | + | 2 | - | 1 | + | 1 | + | 1 | - | 1 | + | 0 |
| 5 | = | 10 | - | 3 | - | 4 | + | 2 | - | 1 | + | 1 | + | 1 | - | 1 | + | 0 |
| 18 | = | 10 | + | 3 | + | 1 | + | 3 | - | 1 | - | 1 | + | 2 | + | 0 | + | 1 |
| 16 | = | 10 | + | 3 | + | 1 | + | 3 | - | 1 | - | 1 | + | 2 | + | 0 | - | 1 |
| 7 | = | 10 | - | 3 | + | 1 | - | 3 | - | 1 | + | 1 | + | 2 | + | 0 | + | 0 |
| 7 | = | 10 | - | 3 | + | 1 | - | 3 | - | 1 | + | 1 | + | 2 | + | 0 | + | 0 |
| 7 | = | 10 | + | 3 | + | 3 | - | 1 | - | 1 | - | 1 | - | 3 | - | 1 | - | 2 |
| 11 | = | 10 | + | 3 | + | 3 | - | 1 | - | 1 | - | 1 | - | 3 | - | 1 | + | 2 |
| 8 | = | 10 | - | 3 | + | 3 | + | 1 | - | 1 | + | 1 | - | 3 | + | 1 | - | 1 |
| 10 | = | 10 | - | 3 | + | 3 | + | 1 | - | 1 | + | 1 | - | 3 | + | 1 | + | 1 |
| 8 | = | 10 | + | 3 | - | 4 | - | 2 | + | 5 | + | 1 | - | 2 | - | 1 | - | 2 |
| 12 | = | 10 | + | 3 | - | 4 | - | 2 | + | 5 | + | 1 | - | 2 | - | 1 | + | 2 |
| 11 | = | 10 | - | 3 | - | 4 | + | 2 | + | 5 | - | 1 | - | 2 | + | 1 | + | 3 |
| 5 | = | 10 | - | 3 | - | 4 | + | 2 | + | 5 | - | 1 | - | 2 | + | 1 | - | 3 |
| 25 | = | 10 | + | 3 | + | 1 | + | 3 | + | 5 | + | 1 | - | 1 | + | 2 | + | 1 |
| 23 | = | 10 | + | 3 | + | 1 | + | 3 | + | 5 | + | 1 | - | 1 | + | 2 | - | 1 |
| 9 | = | 10 | - | 3 | + | 1 | - | 3 | + | 5 | - | 1 | - | 1 | - | 2 | + | 3 |
| 3 | = | 10 | - | 3 | + | 1 | - | 3 | + | 5 | - | 1 | - | 1 | - | 2 | - | 3 |
| 24 | = | 10 | + | 3 | + | 3 | - | 1 | + | 5 | + | 1 | + | 3 | - | 1 | + | 1 |
| 22 | = | 10 | + | 3 | + | 3 | - | 1 | + | 5 | + | 1 | + | 3 | - | 1 | - | 1 |
| 22 | = | 10 | - | 3 | + | 3 | + | 1 | + | 5 | - | 1 | + | 3 | + | 1 | + | 3 |
| 16 | = | 10 | - | 3 | + | 3 | + | 1 | + | 5 | - | 1 | + | 3 | + | 1 | - | 3 |

**ANOVA:**

| Source | *df* | $SS$ | $MS$ | $F$ | Error Term |
|---|---|---|---|---|---|
| $\mu$ | 1 | 4800.0 | 4800.0 | 886.154 | $S(ABC)$ |
| $A$ | 1 | 432.0 | 432.0 | 79.754 | $S(ABC)$ |
| $B$ | 2 | 416.0 | 208.0 | 38.400 | $S(ABC)$ |
| $AB$ | 2 | 224.0 | 112.0 | 20.677 | $S(ABC)$ |
| $C$ | 3 | 408.0 | 136.0 | 25.108 | $S(ABC)$ |
| $AC$ | 3 | 48.0 | 16.0 | 2.954 | $S(ABC)$ |
| $BC$ | 6 | 128.0 | 21.3 | 3.938 | $S(ABC)$ |
| $ABC$ | 6 | 64.0 | 10.7 | 1.969 | $S(ABC)$ |
| $S(ABC)$ | 24 | 130.0 | 5.4 | | |
| Total | 48 | 6650.0 | | | |

**Interpretation:** The significant sources are $A$, $B$, $C$, $AB$, and $BC$. Overall, we see that females score better than males, the Music-On strategy is worse than the other two, and seniors score higher than the other classes. The advantage for females over males is especially large with the TV-On strategy, intermediate with the Book Under Pillow strategy, and smallest (possibly nonexistent) with the Music On strategy (AB interaction). The three different strategies show different patterns of results across class level ($BC$ interaction). The Music-On strategy is least affected by class level, and Book Under Pillow is most affected. Also, whereas test scores increase from freshmen to juniors with the Music On and TV On strategies, they decrease with the Book Under Pillow strategy (further analysis is needed to see if this differential class effect is statistically significant). The theoretical implications of this pattern are not immediately apparent.

# Chapter 6

# Generalization of Between-Subject Designs

In the preceding sections, we have considered in detail ANOVA for between-subjects designs with equal sizes having one, two, or three experimental factors. ANOVA can also be used with designs having four, five, ten, or (in principle) any number of factors. However, it would be both infinitely tedious and unnecessary to work step-by-step through an example of each size design. Instead, we can formulate a set of general rules for computing ANOVA given an equal-cell-size, between-subjects design with any number of factors. Many of these general rules will also apply to within-subjects and mixed designs.

While studying ANOVA for one, two, and three factor designs, you may have noticed a pattern emerging; a similarity in the steps required to perform each ANOVA:

1. Identify the experimental factors.
2. Construct the appropriate linear model.
3. Estimate the components of the model.
4. Compute SS's and df's associated with each term in the model.
5. Compute MS's and F's.
6. Interpret the significant F's.

These six steps constitute a general algorithm for computing ANOVA for an experimental design having any number of factors. This chapter presents a general explanation of the correct way to perform steps 2 through 6 (step 1 is self-explanatory) for between-subjects designs. Careful study of this material should not only provide you with the tools necessary to perform ANOVA on data from any arbitrary between-subjects experimental design, but should also help you to see the material presented earlier as part of an orderly and coherent system.

## 6.1 Constructing the model

The linear model describes a data point as the sum of logically different sources of variance that contribute to that data value, so the appropriate model must contain a term for each source of variance. Very simply, the model must contain a term for the baseline, a term for each main effect, a term for each interaction, and a term for error. Formally, the linear model for an N-factor between-subjects design is always of the form:

$$Y = \mu + \text{main effects} + \text{2-way interactions} + \text{3-way interactions} + \ldots + \text{N-way interaction} + \text{S(Between-Ss factors)}$$

leaving off the subscripts for simplicity.

There is always one main effect term for each experimental factor, one 2-way interaction term for each pair of factors, one 3-way interaction term for each combination of three factors, etc. Counting

the $\mu$ and error terms, there are a total of $2^N+1$ terms in the linear model for a design with N factors. For example, a four-factor design will have the following linear model:

$$\begin{aligned} Y_{ijklm} &= \mu + A_i + B_j + C_k + D_l + AB_{ij} + AC_{ik} + AD_{il} + BC_{jk} + BD_{jl} + CD_{kl} \\ &\quad + ABC_{ijk} + ABD_{ijl} + ACD_{ikl} + BCD_{jkl} + ABCD_{ijkl} + S(ABCD)_{ijklm} \end{aligned}$$

As stated above, a correct linear model should contain $2^N+1$ terms. This fact can be used to insure that no terms have been omitted from the model. In the four factor example above, we should have $2^4+1=17$ terms on the right-hand side of the equals sign, as in fact we do. This is an encouraging, but not conclusive sign. We must also be sure that the terms in the model are all correct. Great care must be exercised in construction of the linear model, as it is the model that determines all subsequent calculations.[1]

### 6.1.1 The Order of the Terms in the Model

Although it seems most natural to write out the model with the terms arranged as shown above (all main effects, then all two-way interactions, then all three-way interactions, etc.), it is easier to *generate* the correct model using a strategy that puts the terms down in another order, and so we will actually use this alternate ordering of terms.

This strategy starts with the terms for baseline, the first two main effects, and their interaction, like this:

$$Y = \mu + A_i + B_j + AB_{ij}$$

To this starting model, you add the next main effect ($C_k$). Then *go back to the first main effect* and form an interaction of this new main effect with *each previous term in the model*. At the beginning of the model we find the $A$ term, and so we form its interaction with $C$ to get $AC$. After $A$ in the model comes $B$, so we next form the $BC$ interaction. After $B$ in the model comes $AB$, so we next form the $ABC$ interaction. That brings us back to the factor we are adding ($C$), so we are done with this step. Here is the result obtained so far, with the terms generated by adding the $C$ factor written directly below their "matching" terms from the earlier part of the model:

$$\begin{array}{rcccccccc} Y &=& \mu & + & A_i & + & B_j & + & AB_{ij} & + \\ & & C_k & + & AC_{ik} & + & BC_{jk} & + & ABC_{ijk} & \end{array}$$

Of course, stringing all the terms out on a single line, the model constructed so far looks like this:

$$Y = \mu + A_i + B_j + AB_{ij} + C_k + AC_{ik} + BC_{jk} + ABC_{ijk}$$

To add another factor $D$, we repeat the process. First, add $D$ on the end of the model. Then, go back to the $A$ term at the beginning and proceed rightward forming interactions of $D$ with each term already in the model, resulting in $AD$, $BD$, $ABD$, $CD$, etc, as shown below.

$$\begin{array}{rcccccccc} Y &=& \mu & +A_i & +B_j & +AB_{ij} & +C_k & +AC_{ik} & +BC_{jk} & +ABC_{ijk} \\ & & D_l & +AD_{il} & +BD_{jl} & +ABD_{ijl} & +CD_{kl} & +ACD_{ikl} & +BCD_{jkl} & +ABCD_{ijkl} \end{array}$$

To add another factor $E$, we repeat the process. Put the $E$ on the end. Go back to the $A$ term and form interactions of $E$ with each term already in the model.

The process can be repeated for each new factor to add indefinitely many factors. Note that the model doubles in size each time we add a factor, because we get not only a new main effect but an interaction with all the previous terms in the model. This means that the model starts to have

[1] For those students with some background in probability or combinatorics, note that the interaction terms in the model comprise the set of all possible combinations of size K from a set of size N for K equal 2 to N. That is, the two-way interaction terms are the set of all pairs of size two from N items; the three-way interaction terms are all combinations of size three, etc. The number of such combinations for a given K is "N-choose-K", computed as

$$\left( \begin{array}{c} N \\ K \end{array} \right) = \frac{N!}{K! \times (N-K)!}$$

Computing N-choose-K for each K can provide a further check on the accuracy of your linear model, as it will indicate the correct number of terms for each size interaction. In the above example, 4-choose-2 equals 6 (4 factorial divided by 2 factorial times 2 factorial). If, when constructing our model we had fewer than 6 two-way interaction terms, we would know that we had omitted some terms from the model.

unreasonably many terms when we get up to experiments with six, seven or more factors (depending on your cutoff for "reasonable"). But that is not a problem with the model. It simply reflects the fact that an experiment with so many factors addresses an enormous number of separate questions. In fact, it is quite useful to have the ANOVA model around to remind us of all the questions we can ask, lest we forget some.

The last step in generating the model is to add the subject error term, which is $S()$, with a list of the between-subjects factors inside the parentheses.

## 6.2 Subscripts

It is very important to get the correct subscripts on all the terms in the model, because the subscripts are used to determine the correct estimation equation. Here are the rules for subscripting:

1. The number of subscripts on the $Y$ term will be one more than the number of experimental factors. By convention, the initial subscripts are used to indicate the levels of the experimental factors at which a particular $Y$ value was observed. The last subscript is used to indicate which observation is being referenced within a particular cell.

2. Each main effect term has a single subscript. Traditionally, the leftmost main effect term is associated with the first subscript, the second main effect term is associated with the second subscript, etc. Of course the results of the ANOVA are completely independent of the order in which terms occur in the model, and of the letters assigned to them as subscripts. These notational conventions merely provide for standardization and computational convenience.

3. Each interaction term inherits the subscripts of the main effect terms it involves. Thus the term signifying the three-way interaction of $A_i$, $C_k$, and $E_m$ is $ACE_{ikm}$.

4. The error term $S(ABC.....)$ has the same subscripts as $Y$, the individual data value. Note that this is logically correct, as each data value has is own unique influence of random error.

It is well to remember that the subscript letters ($i$, $j$, $k$, etc.) are *symbols* meant to be replaced by integers. For example, $A_i$ is only a symbol—it has no actual value until it becomes a specific $A$ like $A_1$ or $A_2$. Similarly, there is no value of $ABC_{ijk}$, although there are values of $ABC_{213}$, etc.

Conceptually, the model-construction step can be summarized by saying that the model allows for effects of baseline, subject error, each experimental factor, and each combination of factors (pairs of factors, triples of factors, etc.).

## 6.3 Estimation

This is the most complex step in the ANOVA algorithm, both to describe and to perform. To introduce it, we will first use it to describe the estimation process used in the three-way design, and then we will give the general statement of the algorithm.

In estimating interactions for the three-way design (see Table 5.3), we estimated each interaction term starting with the $Y$ having the same subscripts, averaging over any subscripts not on the interaction term. Then we subtracted the baseline and any previously estimated model terms having the same subscripts. The result was the estimate of the interaction. This definition in terms of subscripts seems very abstract, but it turns out to be a very general rule in ANOVA.

Estimation of the values of the three-way interaction term $ABC_{ijk}$ can be viewed as another application of this same subscript-based process. To estimate each value of $ABC_{ijk}$, we start with the data value obtained by averaging over the subscripts not present on the term we are trying to estimate ($Y_{ijk.}$). From this data mean, subtract the baseline and all previously estimated model terms with subscripts appearing on the term we are currently estimating.

Finally, the estimate of $S(ABC)_{ijkl}$ can also be viewed in the terms stated above:

1. Start with the data mean obtained by averaging over any subscripts not on the term to be estimated. (All subscripts appear on this term, so we just start with the single data value indicated.)

2. Subtract the values of the baseline and any previously estimated terms of the model which have subscripts that are a subset of the subscripts on the term currently being estimated. (In this

step we subtract off all of the estimated terms, since they all have subscripts contained in those of $S(ABC)_{ijkl}$).

This leads us, then, to a general statement of the estimation algorithm for the linear model used with equal cell size designs:

1. Always begin by setting $\hat{\mu}$ equal to the grand mean of all the data values. Subsequent terms are estimated from left to right across the model, following steps 2-4 *for each term*:

2. Find the average data value for the condition described by the term. For example, in computing $\hat{A}_1$ this step is to average all the values obtained at level one of factor $A$. In computing $\widehat{AC}_{13}$ this step is to average all the data values obtained at level one of factor $A$ and level 3 of factor $C$. Another way to describe this average is to say that it is the average across all the subscripts *not* attached to the model term.

3. Subtract the baseline from the data average obtained in the previous step, and then subtract all model terms that can be constructed from subsets of the factors in the term currently being estimated. When computing $\hat{A}_1$, for example, you just subtract off the baseline, because there are no models terms that are subsets of $A$. But in computing $\widehat{ABC}_{143}$, you subtract off the baseline and $\hat{A}_1$, $\hat{B}_4$, $\widehat{AB}_{14}$, $\hat{C}_3$, $\widehat{AC}_{13}$, and $\widehat{BC}_{43}$, because all these model terms are made from subsets of the letters in the term being estimated (i.e., $ABC$).

4. The result of the subtraction(s) in Step 3 is the desired estimate.

For example, in a four-factor design ($Y_{ijklm}$) the estimation equation for $\widehat{ABD}_{ijl}$ is:

$$\widehat{ABD}_{ijl} = Y_{ij.l.} - \hat{\mu} - \hat{A}_i - \hat{B}_j - \widehat{AB}_{ij} - \hat{D}_l - \widehat{AD}_{il} - \widehat{BD}_{jl}$$

Conceptually, the estimation process can be summarized as follows. Each term in the model is estimated by taking the average of all data values to which that term contributes, and then subtracting off the effects of all the simpler terms. Thus, the estimate of a term is whatever is unexplained in the corresponding data average after we take into account all the other effects on that data average.

## 6.4 Computing $SS$'s and $df$'s

To compute the $SS$'s, write out the decomposition matrix in which each score is expressed as the sum of its estimated linear model components. The $SS$ associated with any term is then the sum of the squared values in the appropriate column of the decomposition matrix. Conceptually, the sum of squares summarizes the total effect of a term on the data set as a whole.

To compute *df*'s, use the following rules:

- The *df* for $\mu = 1$.
- The *df* for each main effect equals the number of levels in the associated factor minus 1.
- The *df* for any interaction equals the product of the *df*'s of all the factors involved in interaction.
- The *df* for $S(ABC...)$ equals the number of subjects minus the number of groups. The number of subjects is the total number of randomly sampled individuals in the experiment. The number of groups equals the number of cells in the design, which is also equal to the product of the numbers of levels on all the between-Ss factors.
- The total *df* equals the total number of observations.

Conceptually, *df*'s are counters of the number of independent values taken on by each term in the model. For example, $\mu$ can take on one value, a main effect term takes on one independent value for each level except the last, and so on. The rules generating *df*'s for $\mu$, main effects, and subjects have been explained in connection with simpler designs, and the same explanations hold for the general case. The *df*'s for interaction can be deduced from the constraints that they sum to zero for each of their subscripts, but it is probably sufficient just to know that they are counters and that the product rule always works.

## 6.5 Computing $MS$'s and $F$'s

This is trivially easy.

- $MS$ for any term = $SS$ for that term divided by *df* for that term.
- $F_{observed}$ for any term = $MS$ for that term divided by $MS_{S(ABC...)}$.

## 6.6 Interpreting significant $F$'s

This step can not be described in any general way. There is no simple set of rules which will lead you to the most correct and insightful interpretation of the results of an ANOVA. Significant $F$'s can only be fully understood in the context of a complete experiment, and often only with reference to hundreds of experiments that have gone before. The researcher hopes that the results of his or her study will provide insight into his or her field, be it physics, biology, psychology, etc. Finding this insight, relating it to preexisting knowledge in the field, using it to increase scientific understanding—these are the art of science.

The researcher's success in interpreting the ANOVA depends in part on careful design of the experiment itself, in part on comprehensive knowledge of the system under study, and sometimes in part on luck—those sudden ideas that come to you while doing the dishes or driving down the freeway. These things can't be rigorously formulated.

However there are some rules of thumb that can save you from the embarrassment of making claims that ANOVA cannot support.

First, conclusions should be stated in terms of the full design, not just a few cells. Each $F$ is computed using all the data points, and the interpretation should be as general as the statistic is. Second, be sure to qualify statements about any significant source of variance if that source is contained in a significant higher-way interaction. The conclusion must be stated as an "on average" or "overall" result, and the importance of this qualification must be kept in mind when discussing more general implications of the research. Third, use lots of graphs to try to understand and explain the significant interactions.

While on the topic of interpreting $F$'s, we would like to suggest that it would be instructive at this point to look at some actual published experiments, where ANOVA is used, and where $F$'s are interpreted in the context of a complete field of study. Your local campus library will have many professional journals containing articles reporting current experiments. For psychologists, one good example is the *Journal of Experimental Psychology: General*, which reports experiments from many different areas within the field. Skim through some articles in this journal or a comparable one for studies where ANOVA is used. Identify the experimental factors and their levels, and observe the interpretation process. This is ANOVA in the "real world."

## 6.7 Exercise: Four-Factor Design

To give you an opportunity to practice applying the general, between-subjects ANOVA algorithm, below is presented a complete ANOVA for a four factor, fully between-subjects design. This exercise is intended to illustrate computations only, so there is no associated experiment and the results are not interpreted.

**Design:**

- Factor A: 2 levels, Between-Ss.
- Factor B: 2 levels, Between-Ss.
- Factor C: 2 levels, Between-Ss.
- Factor D: 2 levels, Between-Ss.
- 2 Subjects per group.

**Data:**

| | A1, B1 | | | | A1, B2 | | | | A2, B1 | | | | A2, B2 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | C1 | | C2 | | C1 | | C2 | | C1 | | C2 | | C1 | | C2 | |
| D1: | 39 | 33 | 16 | 16 | 56 | 56 | 47 | 41 | 19 | 21 | 12 | 12 | 20 | 20 | 3 | 5 |
| D2: | 18 | 14 | 11 | 5 | 43 | 37 | 26 | 22 | 13 | 11 | 16 | 16 | 0 | 0 | -3 | -5 |

**Model:**

$$\begin{aligned} Y_{ijklm} &= \mu + A_i + B_j + AB_{ij} + C_k + AC_{ik} + BC_{jk} + ABC_{ijk} + D_l + AD_{il} + BD_{jl} \\ &\quad + ABD_{ijl} + CD_{kl} + ACD_{ikl} + BCD_{jkl} + ABCD_{ijkl} + S(ABCD)_{ijklm} \end{aligned}$$

**Estimation Equations:**

$$\begin{aligned} \hat{\mu} &= Y_{.....} \\ \hat{A}_i &= Y_{i....} - \hat{\mu} \\ \hat{B}_j &= Y_{.j...} - \hat{\mu} \\ \widehat{AB}_{ij} &= Y_{ij...} - \hat{\mu} - \hat{A}_i - \hat{B}_j \\ \hat{C}_k &= Y_{..k..} - \hat{\mu} \\ \widehat{AC}_{ik} &= Y_{i.k..} - \hat{\mu} - \hat{A}_i - \hat{C}_k \\ \widehat{BC}_{jk} &= Y_{.jk..} - \hat{\mu} - \hat{B}_j - \hat{C}_k \\ \widehat{ABC}_{ijk} &= Y_{ijk..} - \hat{\mu} - \hat{A}_i - \hat{B}_j - \widehat{AB}_{ij} - \hat{C}_k - \widehat{AC}_{ik} - \widehat{BC}_{jk} \\ \hat{D}_l &= Y_{...l.} - \hat{\mu} \\ \widehat{AD}_{il} &= Y_{i..l.} - \hat{\mu} - \hat{A}_i - \hat{D}_l \\ \widehat{BD}_{jl} &= Y_{.j.l.} - \hat{\mu} - \hat{B}_j - \hat{D}_l \\ \widehat{ABD}_{ijl} &= Y_{ij.l.} - \hat{\mu} - \hat{A}_i - \hat{B}_j - \widehat{AB}_{ij} - \hat{D}_l - \widehat{AD}_{il} - \widehat{BD}_{jl} \\ \widehat{CD}_{kl} &= Y_{..kl.} - \hat{\mu} - \hat{C}_k - \hat{D}_l \\ \widehat{ACD}_{ikl} &= Y_{i.kl.} - \hat{\mu} - \hat{A}_i - \hat{C}_k - \widehat{AC}_{ik} - \hat{D}_l - \widehat{AD}_{il} - \widehat{BD}_{jl} \\ \widehat{BCD}_{jkl} &= Y_{.jkl.} - \hat{\mu} - \hat{B}_j - \hat{C}_k - \widehat{BC}_{jk} - \hat{D}_l - \widehat{BD}_{jl} - \widehat{CD}_{kl} \\ \widehat{ABCD}_{ijkl} &= Y_{ijkl.} - \hat{\mu} - \hat{A}_i - \hat{B}_j - \widehat{AB}_{ij} - \hat{C}_k - \widehat{AC}_{ik} - \widehat{BC}_{jk} - \widehat{ABC}_{ijk} \\ &\quad - \hat{D}_l - \widehat{AD}_{il} - \widehat{BD}_{jl} - \widehat{ABD}_{ijl} - \widehat{CD}_{kl} - \widehat{ACD}_{ikl} - \widehat{BCD}_{jkl} \\ S(\widehat{ABCD})_{ijklm} &= Y_{ijklm} - \hat{\mu} - \hat{A}_i - \hat{B}_j - \widehat{AB}_{ij} - \hat{C}_k - \widehat{AC}_{ik} - \widehat{BC}_{jk} - \widehat{ABC}_{ijk} - \hat{D}_l \\ &\quad - \widehat{AD}_{il} - \widehat{BD}_{jl} - \widehat{ABD}_{ijl} - \widehat{CD}_{kl} - \widehat{ACD}_{ikl} - \widehat{BCD}_{jkl} - \widehat{ABCD}_{ijkl} \\ &= Y_{ijklm} - Y_{ijkl.} \end{aligned}$$

**Decomposition Matrix:**

| $Y_{ijklm}$ | = | $\mu$ | + | $A_i$ | + | $B_j$ | + | $AB_{ij}$ | + | $C_k$ | + | $AC_{ik}$ | + | $BC_{jk}$ | + | $ABC_{ijk}$ | + | $D_l$ | + | ... |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 39 | = | 20 | + | 10 | - | 3 | - | 8 | + | 5 | + | 2 | - | 1 | + | 1 | + | 6 | + | ... |
| 33 | = | 20 | + | 10 | - | 3 | - | 8 | + | 5 | + | 2 | - | 1 | + | 1 | + | 6 | + | ... |
| 19 | = | 20 | - | 10 | - | 3 | + | 8 | + | 5 | - | 2 | - | 1 | - | 1 | + | 6 | + | ... |
| 21 | = | 20 | - | 10 | - | 3 | + | 8 | + | 5 | - | 2 | - | 1 | - | 1 | + | 6 | + | ... |
| 56 | = | 20 | + | 10 | + | 3 | + | 8 | + | 5 | + | 2 | + | 1 | - | 1 | + | 6 | + | ... |
| 56 | = | 20 | + | 10 | + | 3 | + | 8 | + | 5 | + | 2 | + | 1 | - | 1 | + | 6 | + | ... |
| 20 | = | 20 | - | 10 | + | 3 | - | 8 | + | 5 | - | 2 | + | 1 | + | 1 | + | 6 | + | ... |
| 20 | = | 20 | - | 10 | + | 3 | - | 8 | + | 5 | - | 2 | + | 1 | + | 1 | + | 6 | + | ... |
| 16 | = | 20 | + | 10 | - | 3 | - | 8 | - | 5 | - | 2 | + | 1 | - | 1 | + | 6 | + | ... |
| 16 | = | 20 | + | 10 | - | 3 | - | 8 | - | 5 | - | 2 | + | 1 | - | 1 | + | 6 | + | ... |
| 12 | = | 20 | - | 10 | - | 3 | + | 8 | - | 5 | + | 2 | + | 1 | + | 1 | + | 6 | + | ... |
| 12 | = | 20 | - | 10 | - | 3 | + | 8 | - | 5 | + | 2 | + | 1 | + | 1 | + | 6 | + | ... |
| 47 | = | 20 | + | 10 | + | 3 | + | 8 | - | 5 | - | 2 | - | 1 | + | 1 | + | 6 | + | ... |
| 41 | = | 20 | + | 10 | + | 3 | + | 8 | - | 5 | - | 2 | - | 1 | + | 1 | + | 6 | + | ... |
| 3 | = | 20 | - | 10 | + | 3 | - | 8 | - | 5 | + | 2 | - | 1 | - | 1 | + | 6 | + | ... |
| 5 | = | 20 | - | 10 | + | 3 | - | 8 | - | 5 | + | 2 | - | 1 | - | 1 | + | 6 | + | ... |
| 18 | = | 20 | + | 10 | - | 3 | - | 8 | + | 5 | + | 2 | - | 1 | + | 1 | - | 6 | + | ... |
| 14 | = | 20 | + | 10 | - | 3 | - | 8 | + | 5 | + | 2 | - | 1 | + | 1 | - | 6 | + | ... |
| 13 | = | 20 | - | 10 | - | 3 | + | 8 | + | 5 | - | 2 | - | 1 | - | 1 | - | 6 | + | ... |
| 11 | = | 20 | - | 10 | - | 3 | + | 8 | + | 5 | - | 2 | - | 1 | - | 1 | - | 6 | + | ... |
| 43 | = | 20 | + | 10 | + | 3 | + | 8 | + | 5 | + | 2 | + | 1 | - | 1 | - | 6 | + | ... |
| 37 | = | 20 | + | 10 | + | 3 | + | 8 | + | 5 | + | 2 | + | 1 | - | 1 | - | 6 | + | ... |
| 0 | = | 20 | - | 10 | + | 3 | - | 8 | + | 5 | - | 2 | + | 1 | + | 1 | - | 6 | + | ... |
| 0 | = | 20 | - | 10 | + | 3 | - | 8 | + | 5 | - | 2 | + | 1 | + | 1 | - | 6 | + | ... |
| 11 | = | 20 | + | 10 | - | 3 | - | 8 | - | 5 | - | 2 | + | 1 | - | 1 | - | 6 | + | ... |
| 5 | = | 20 | + | 10 | - | 3 | - | 8 | - | 5 | - | 2 | + | 1 | - | 1 | - | 6 | + | ... |
| 16 | = | 20 | - | 10 | - | 3 | + | 8 | - | 5 | + | 2 | + | 1 | + | 1 | - | 6 | + | ... |
| 16 | = | 20 | - | 10 | - | 3 | + | 8 | - | 5 | + | 2 | + | 1 | + | 1 | - | 6 | + | ... |
| 26 | = | 20 | + | 10 | + | 3 | + | 8 | - | 5 | - | 2 | - | 1 | + | 1 | - | 6 | + | ... |
| 22 | = | 20 | + | 10 | + | 3 | + | 8 | - | 5 | - | 2 | - | 1 | + | 1 | - | 6 | + | ... |
| -3 | = | 20 | - | 10 | + | 3 | - | 8 | - | 5 | + | 2 | - | 1 | - | 1 | - | 6 | + | ... |
| -5 | = | 20 | - | 10 | + | 3 | - | 8 | - | 5 | + | 2 | - | 1 | - | 1 | - | 6 | + | ... |

| + | $AD_{il}$ | + | $BD_{jl}$ | + | $ABD_{ijl}$ | + | $CD_{kl}$ | + | $ACD_{ikl}$ | + | $BCD_{jkl}$ | + | $ABCD_{ijkl}$ | + | $S(ABCD)_{ijklm}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| + | 2 | - | 2 | + | 1 | + | 2 | - | 1 | + | 1 | + | 1 | + | 3 |
| + | 2 | - | 2 | + | 1 | + | 2 | - | 1 | + | 1 | + | 1 | - | 3 |
| - | 2 | - | 2 | - | 1 | + | 2 | + | 1 | + | 1 | - | 1 | - | 1 |
| - | 2 | - | 2 | - | 1 | + | 2 | + | 1 | + | 1 | - | 1 | + | 1 |
| + | 2 | + | 2 | - | 1 | + | 2 | - | 1 | - | 1 | - | 1 | + | 0 |
| + | 2 | + | 2 | - | 1 | + | 2 | - | 1 | - | 1 | - | 1 | + | 0 |
| - | 2 | + | 2 | + | 1 | + | 2 | + | 1 | - | 1 | + | 1 | + | 0 |
| - | 2 | + | 2 | + | 1 | + | 2 | + | 1 | - | 1 | + | 1 | + | 0 |
| + | 2 | - | 2 | + | 1 | - | 2 | + | 1 | - | 1 | - | 1 | + | 0 |
| + | 2 | - | 2 | + | 1 | - | 2 | + | 1 | - | 1 | - | 1 | + | 0 |
| - | 2 | - | 2 | - | 1 | - | 2 | - | 1 | - | 1 | + | 1 | + | 0 |
| - | 2 | - | 2 | - | 1 | - | 2 | - | 1 | - | 1 | + | 1 | + | 0 |
| + | 2 | + | 2 | - | 1 | - | 2 | + | 1 | + | 1 | + | 1 | + | 3 |
| + | 2 | + | 2 | - | 1 | - | 2 | + | 1 | + | 1 | + | 1 | - | 3 |
| - | 2 | + | 2 | + | 1 | - | 2 | - | 1 | + | 1 | - | 1 | - | 1 |
| - | 2 | + | 2 | + | 1 | - | 2 | - | 1 | + | 1 | - | 1 | + | 1 |
| - | 2 | + | 2 | - | 1 | - | 2 | + | 1 | - | 1 | - | 1 | + | 2 |
| - | 2 | + | 2 | - | 1 | - | 2 | + | 1 | - | 1 | - | 1 | - | 2 |
| + | 2 | + | 2 | + | 1 | - | 2 | - | 1 | - | 1 | + | 1 | + | 1 |
| + | 2 | + | 2 | + | 1 | - | 2 | - | 1 | - | 1 | + | 1 | - | 1 |
| - | 2 | - | 2 | + | 1 | - | 2 | + | 1 | + | 1 | + | 1 | + | 3 |
| - | 2 | - | 2 | + | 1 | - | 2 | + | 1 | + | 1 | + | 1 | - | 3 |
| + | 2 | - | 2 | - | 1 | - | 2 | - | 1 | + | 1 | - | 1 | + | 0 |
| + | 2 | - | 2 | - | 1 | - | 2 | - | 1 | + | 1 | - | 1 | + | 0 |
| - | 2 | + | 2 | - | 1 | + | 2 | - | 1 | + | 1 | + | 1 | + | 3 |
| - | 2 | + | 2 | - | 1 | + | 2 | - | 1 | + | 1 | + | 1 | - | 3 |
| + | 2 | + | 2 | + | 1 | + | 2 | + | 1 | + | 1 | - | 1 | + | 0 |
| + | 2 | + | 2 | + | 1 | + | 2 | + | 1 | + | 1 | - | 1 | + | 0 |
| - | 2 | - | 2 | + | 1 | + | 2 | - | 1 | - | 1 | - | 1 | + | 2 |
| - | 2 | - | 2 | + | 1 | + | 2 | - | 1 | - | 1 | - | 1 | - | 2 |
| + | 2 | - | 2 | - | 1 | + | 2 | + | 1 | - | 1 | + | 1 | + | 1 |
| + | 2 | - | 2 | - | 1 | + | 2 | + | 1 | - | 1 | + | 1 | - | 1 |

**ANOVA:**

| Source | *df* | $SS$ | $MS$ | $F$ | Error Term |
|---|---|---|---|---|---|
| $\mu$ | 1 | 12800.0 | 12800.0 | 2133.333 | $S(ABCD)$ |
| $A$ | 1 | 3200.0 | 3200.0 | 533.333 | $S(ABCD)$ |
| $B$ | 1 | 288.0 | 288.0 | 48.000 | $S(ABCD)$ |
| $AB$ | 1 | 2048.0 | 2048.0 | 341.333 | $S(ABCD)$ |
| $C$ | 1 | 800.0 | 800.0 | 133.333 | $S(ABCD)$ |
| $AC$ | 1 | 128.0 | 128.0 | 21.333 | $S(ABCD)$ |
| $BC$ | 1 | 32.0 | 32.0 | 5.333 | $S(ABCD)$ |
| $ABC$ | 1 | 32.0 | 32.0 | 5.333 | $S(ABCD)$ |
| $D$ | 1 | 1152.0 | 1152.0 | 192.000 | $S(ABCD)$ |
| $AD$ | 1 | 128.0 | 128.0 | 21.333 | $S(ABCD)$ |
| $BD$ | 1 | 128.0 | 128.0 | 21.333 | $S(ABCD)$ |
| $ABD$ | 1 | 32.0 | 32.0 | 5.333 | $S(ABCD)$ |
| $CD$ | 1 | 128.0 | 128.0 | 21.333 | $S(ABCD)$ |
| $ACD$ | 1 | 32.0 | 32.0 | 5.333 | $S(ABCD)$ |
| $BCD$ | 1 | 32.0 | 32.0 | 5.333 | $S(ABCD)$ |
| $ABCD$ | 1 | 32.0 | 32.0 | 5.333 | $S(ABCD)$ |
| $S(ABCD)$ | 16 | 96.0 | 6.0 | | |
| Total | 32 | 21088.0 | | | |

# Chapter 7

# Within-Subjects Designs

## 7.1 Within- vs. Between-Subject Designs

All of the experimental designs we have considered so far have been fully between-subjects designs. In such designs each subject is tested in exactly one condition (i.e., at one level of the experimental factor or one combination of levels of the experimental factors). In essence, this means that each subject is tested in just one cell of the design. For example, in a one-way between-subjects design, if a subject is tested at level 1 of the factor, he is never tested at level 2 or 3 or any other level (e.g., in an experiment where college class was a factor, each subject would be tested as either a freshman, sophomore, junior, or senior, but not as more than one of these). In a two-way between-subjects design, if a subject is tested in the condition defined by level 1 of Factor A and level 4 of Factor B ($AB_{14}$, e.g., male senior), he is never tested in cell $AB_{24}$ (female senior) or $AB_{12}$ (male sophomore) and so on.

We will now begin considering a new class of experimental design, in which each subject is tested in *every* cell of the experiment. This type of design is called a *within-subjects design.*[1] For example, in a one-factor within-subjects design exploring the effect of being in a particular college class, each subject would have to be measured four times: once as a freshman, again as a sophomore, again as a junior, and finally as a senior. If drunk vs. sober were a second within-Ss factor in the design, each subject would have to be tested twice each year (once drunk and once sober), for a total of eight tests per subject. Obviously, not all factors can be tested within-subjects. For example, it is impossible to test each subject at both levels of the gender factor. Similarly, you can't compare two different methods for teaching French in a within-subjects design. Once you have taught subjects with one method, you can't start over from the beginning and teach them with the other method.

Between-subjects designs can be very useful in many research situations, especially those in which it is easy to test lots of subjects in each condition. Furthermore, they are the only option with factors like gender. Nonetheless, many researchers prefer to use within-subjects designs whenever possible, because within-subjects designs are less influenced by random error than between-subjects designs. Within-subjects designs thus have greater power[2] than between-subjects designs. In the remainder of this section we will explain why within-subjects designs are less affected by random variation. In subsequent sections we will explain how to perform ANOVA on within-subjects designs. Finally, we will present some example data sets and analyses.

Random error causes two types of problems in between-subjects designs. The first is that the average results can be severely influenced by accidental differences between groups of subjects tested in different conditions. Imagine for example that you are testing motor task performance under conditions of sleep loss vs. normal sleep. You test subjects' motor performance by having them try to throw a basketball through a basketball hoop. One group performs the task after a good night's sleep; the other after staying awake for 24 hours. You randomly select 10 subjects and randomly assign five to each group. Unfortunately, your random selection and assignment happens to place all five starting members of the college basketball team in one group and five dedicated couch potatoes

[1] More formally, we have the following definitions: An experimental factor for which each subject is tested at every level is called a *within-subjects* factor. An experimental design composed entirely of within-subjects factors is called a *fully within-subjects design.*

[2] In statistical terms, *power* is the probability of rejecting $H_0$ when it is false—i.e., of detecting the error in the $H_0$. Other things being equal (e.g., sample size), an experiment will have more power to the extent that the scores are less influenced by random variability.

in the other. (Granted, this is an unlikely occurrence, but with random selection and assignment, any given pattern is logically possible.)

If you happen to place the basketball players in the normal sleep condition and the couch potatoes in the sleep loss condition, your data will probably suggest that sleep loss adversely affects motor performance much more than it really does. Worse yet, if you happen to place the basketball players in the sleep loss condition and the couch potatoes in the rested condition, your data may well show that the sleep loss group performs significantly better than the rested group.[3] Will you conclude that sleep loss *enhances* motor performance? Will this be the end of your career as a research scientist? Either way the groups are assigned, though, it is clear that an accidentally occurring difference between your two groups of subjects greatly distorts the true effect of your experimental manipulation.

The second problem created by error in between-subjects designs can be seen when we consider the character of the $F_{observed}$. An $F$ is essentially the ratio of effect size to error, where error is measured as the differences between subjects in the same condition. A large $F_{observed}$ convinces us that the effect could not have occurred due to chance, since our obtained within-cell error estimates the disparity of scores expected in this population due to chance. However, in a between-subjects design, we may obtain a small $F_{observed}$ even when there exists a true effect of our experimental factor, simply because the effect is small relative to the random variation in the population.

To illustrate this second disadvantage, let us consider another farfetched example (we will see some realistic examples later). You want to find out whether people weigh more with their clothes on than with their clothes off. You randomly select 10 people and weigh them fully dressed; you randomly select another 10 people and weigh them without clothes. The average person probably wears about five pounds of clothes, so the mean of the dressed group should be about five pounds greater than the mean of the not dressed group. However, the weight variation among a randomly selected group of 10 people will be very large compared to that five pound difference. If you select from a population of college undergraduates, your subjects' weight may range from under 100 to over 250 pounds. Thus, the within-cell error term in your ANOVA will be very large relative to the main effect term, and you may not obtain a significant main effect of the clothes on/off factor. Although increasing the sample size could overcome this random variation (because the mean of a sample is known more precisely as the sample size increases), the general point is still valid: With any given sample size, you have less chance of rejecting a false $H_0$ (i.e., less power) when random variation between subjects contributes to the error term.

Within-subjects designs solve both of the problems caused by random variation between subjects. In the basketball example, you could have each subject shoot baskets one day when fully rested and another day when sleep deprived. There would then be no way that the people tested in the rested group could be inherently more or less skilled that the people tested in the sleep-deprived group, because they would be the very *same people*.

In the clothes on/off example, you could weigh a subject with his clothes on and then weigh the same subject with his clothes off. Although there will still be variation in weight from one subject to the next, this variation will not conceal the main effect of clothes because within-subjects designs allow us to compute an error term that is unaffected by this kind of random variation between subjects. We will explain how this is done below. Put very simply, the main effect of clothes will be clear because each subject will show it (i.e., every subject will weigh more when dressed than when not).

In summary, the main advantages of the within-subjects design over the between-subjects design are that it eliminates effects of a priori group differences and provides a more sensitive measure of error than the simple within-cell variation used as the error term in between-subjects designs.[4] Thus, within-subjects designs are preferred in many types of experiments, including tests of various medications, of teaching techniques, of manipulations expected to effect motor coordination, visual processing, reading ability, reflexes, etc. Essentially, a within-subjects design should be seriously considered any time the subjects differ greatly from one another in their overall scores on the DV.

There are certain practical problems associated with within-subjects factors, however. As we have already seen, some manipulations simply do not lend themselves to this type of design. Gender and ethnic group for example, must always be between-subjects factors; age can only be a within-subjects factor if subjects can be tested repeatedly over a period of some years. Second, use of within-subjects factors introduces *order effects*. That is, the results of an experiment may vary depending on the order

[3] In this example, the true impact of sleep loss on motor performance can clearly be masked by an *a priori* difference between the two groups of subjects. Probably, sleep loss does adversely affect motor performance, but, if the tired basketball team can outshoot the rested couch potatoes, you will not observe the true effect, and may observe its opposite.

[4] Be careful of the somewhat confusing terminology: *Between*-subjects designs have *within*-cell error terms.

in which subjects are tested in the within-subjects conditions. For example, it can become impossible to measure the effects of a second medication if the first medication given cures the condition; a comparison of techniques for teaching reading must consider the improvement in skill resulting from Condition 1 when testing Condition 2; even our basketball example may be subject to order effects if subjects' ability to shoot baskets was improved by the practice they received in the first condition. It is possible to control for order effects when testing within-subjects factors, although this can be very tricky; we will discuss one approach after studying mixed designs (i.e., those having both between- and within-subjects factors).

## 7.2 Models for Within-Ss Designs

ANOVA is computed very similarly for within- and between-subjects designs, but there are two important differences: 1) use of a within-subjects design introduces some new terms to the model, and 2) it complicates the selection of error terms for computing $F_{observed}$ values. In this section, we consider how models are different in within-subjects designs. To illustrate these new terms, we will consider a specific example of a within-subjects design.

Suppose that we are interested in studying problem solving by congressmen. We might randomly select six congressmen, give each one a series of logic problems, and measure the time they required to solve each one. Furthermore, suppose that we want to compare problem solving under different conditions likely to be encountered in office. For example, we might want to compare problem solving speeds under four stress conditions: 1) normal working conditions, 2) after 2 glasses of wine, 3) after being awakened in the middle of the night, and 4) in the midst of a bad head cold. The data might look like those in Table 7.1.

| | Decision-Making Condition: | | | | |
|---|---|---|---|---|---|
| Congressman | Normal | Wine | Awaken | Cold | Average |
| Mr. Rockford | 121 | 134 | 133 | 128 | 129 |
| Mr. Columbo | 121 | 139 | 141 | 135 | 134 |
| Mrs. Fletcher | 113 | 125 | 129 | 113 | 120 |
| Mr. Mason | 123 | 139 | 137 | 129 | 132 |
| Mr. Chan | 126 | 140 | 143 | 143 | 138 |
| Miss Marple | 110 | 121 | 133 | 144 | 127 |
| Average | 119 | 133 | 136 | 132 | 130 |

Table 7.1: Decision Making by Congressmen

The appropriate model for these data is:

$$Y_{ij} = \mu + A_i + S_j + AS_{ij}$$

As always, we use $Y$ to stand for the scores themselves. We need two subscripts to uniquely identify each of the scores. The subscript $i$ indicates the level of the conditions factor and $j$ indicates the congressman (i.e., subject).

The within-subjects model still has the baseline ($\mu$) and the main effect of conditions ($A_i$). The rationale is exactly as before: We expect each score to be influenced by the overall baseline level of scoring and by the experimental manipulation.

The model also contains some new terms, peculiar to within-subjects designs. First, consider the new $S_j$ term. Recall that the function of the linear model is to divide up data values into components due to various sources. For example, the function of a main effect term is to describe an overall deviation from the baseline of a whole set of scores (e.g., all those obtained after two glasses of wine). These scores are expected to have a common overall deviation from the baseline because they are all obtained under conditions that are somehow similar. In earlier designs for example, we used main effect terms to describe a common deviation from baseline due to scores coming from people of the same gender. The idea was that people of the same gender have something in common, so their scores might have some common component as well.

What is new in a within-subjects design is that we have several scores from each subject. Multiple scores from the same person certainly have something in common, and the model represents this commonality by including the $S_j$ term to take into account a main effect of subject. Just as the model

has a term to reflect what is common about all the scores taken in a sleep-deprived condition, for example, it has a term to reflect what is common to all of Congressman Rockford's scores (or any other subject). This subjects term has special importance in within-subject designs, so we will use the letter $S$ as the subject main effect term in all models, to distinguish it from experimental factors $A$, $B$, $C$, etc.

In short, $S_j$ reflects the effect of Subject $j$ on a set of scores for that subject, just as $A_i$ reflects the effect of Condition $i$ on a set of scores for that condition. Naturally, we will use the data to get estimates, $\hat{S}_j$, of the true values of $S_j$. What can we infer from the values of $\hat{S}_j$ that we estimate? As we will see later, each $\hat{S}_j$ is calculated as the difference between the average score of Subject $j$ and the average of all scores in the data set. Thus, it will be distant from zero to the extent that Subject $j$'s scores are different from the average of all scores from all subjects. In plainer language, if subjects perform very differently from each other on the average across experimental conditions, the $\hat{S}_j$ values will be large (positively and negatively); if subjects' average performances are very similar, the $\hat{S}_j$'s will be small.

The model also has a term representing the interaction between the experimental factor, $A$, and the subjects factor. This $AS_{ij}$ interaction term is analogous to an ordinary interaction term: it measures the extent to which the effect of Factor A differs from subject to subject. For example, one congressman might solve problems more slowly in the middle of the night than after having two glasses of wine, while another congressman might show the reverse pattern.

In between-subjects designs, we saw that interaction terms are large to the extent that the effect of one factor depends on the level of the other. Analogously, the $\widehat{AS}_{ij}$'s are large to the extent that the effect of Factor A depends on which subject is being tested.

It is extremely important to understand that $S_j$ and $AS_{ij}$ provide information about two very distinct kinds of differences among subjects. $S_j$ indicates how Subject $j$ differs from the other subjects with respect to *overall* level of performance—i.e., performance on average across the conditions tested. For example, we might test one congressman who is methodical and thorough by nature and another congressman who is quicker and more impetuous. The methodical congressman would tend to take more time to solve problems in all conditions of the experiment, which would produce a positive $S_j$ for him, since the DV is decision time. The quicker congressman would tend to solve problems more rapidly in all conditions of the experiment, which would make his $S_j$ negative. Therefore, $S_j$ measures individual differences that affect a person's performance in the same direction in *all* conditions of the experiment.

$AS_{ij}$ measures a very different sort of variation among individuals: variation in how they respond to the different conditions of the experiment. To illustrate that this is indeed a different sort of variation, we will discuss two pairs of congressman.

First, consider again the thorough and impetuous congressmen discussed above. Whereas they are very different in overall level of performance, they might respond exactly the same way to the experimental manipulation. For example, both might take exactly 5 minutes more to solve a problem in the middle of the night than during the day. Thus, we see that it is possible for two subjects to have very different values of $S_j$ and yet have very similar values of $AS_{ij}$.

Second, consider two equally thorough congressmen, with identical large values of $S_j$. But suppose one is a very light sleeper and the other very sound sleeper. The light sleeper might solve problems almost as fast in the middle of the night as during the day, whereas the sound sleeper might work much more slowly in the middle of the night than during the day. If so, we might obtain a much bigger effect of condition of testing for the sound sleeper than for the light sleeper. Thus, we see that it is also possible for two subjects to have very similar values of $S_j$ and yet have very different values of $AS_{ij}$ for a given condition.

In summary, $S_j$ measures *overall* variation from subject to subject. $AS_{ij}$, on the other hand, measures *condition-specific* variation from subject to subject. Alternatively, one can also say that $AS_{ij}$ measures how much the experimental effect varies from subject to subject.

To better understand the difference between $S_j$ and $AS_{ij}$, it is useful to compare the four data sets in Table 7.2. The data from each set are plotted in Figure 7.1, using a format similar to that of a factorial plot.[5]

[5]Strictly speaking, these are not true factorial plots, because individual subject data rather than mean values are plotted. Note the similarity to a factorial plot, though: Levels of one factor ($A$) are arranged on the horizontal axis, and levels of another factor ($S$) are indicated by different lines on the plot.

**A:** Small values of $S$ and $AS$

| | Decision-Making Condition: | | | | |
|---|---|---|---|---|---|
| Congressman | Normal | Wine | Awaken | Cold | Average |
| Mr. Rockford | 123 | 136 | 141 | 132 | 133 |
| Mr. Columbo | 120 | 134 | 137 | 125 | 129 |
| Mrs. Fletcher | 121 | 135 | 137 | 131 | 131 |
| Mr. Mason | 123 | 137 | 137 | 131 | 132 |
| Mr. Chan | 121 | 137 | 140 | 130 | 132 |
| Miss Marple | 106 | 119 | 124 | 143 | 123 |
| Average | 119 | 133 | 136 | 132 | 130 |

**B:** Large $S$ and small $AS$

| | Decision-Making Condition: | | | | |
|---|---|---|---|---|---|
| Congressman | Normal | Wine | Awaken | Cold | Average |
| Mr. Rockford | 118 | 132 | 135 | 127 | 128 |
| Mr. Columbo | 123 | 138 | 140 | 135 | 134 |
| Mrs. Fletcher | 101 | 115 | 118 | 110 | 111 |
| Mr. Mason | 130 | 143 | 147 | 140 | 140 |
| Mr. Chan | 140 | 154 | 157 | 149 | 150 |
| Miss Marple | 102 | 116 | 119 | 131 | 117 |
| Average | 119 | 133 | 136 | 132 | 130 |

**C:** Small $S$ and large $AS$

| | Decision-Making Condition: | | | | |
|---|---|---|---|---|---|
| Congressman | Normal | Wine | Awaken | Cold | Average |
| Mr. Rockford | 133 | 126 | 134 | 127 | 130 |
| Mr. Columbo | 130 | 134 | 152 | 112 | 132 |
| Mrs. Fletcher | 119 | 134 | 145 | 126 | 131 |
| Mr. Mason | 107 | 123 | 143 | 147 | 130 |
| Mr. Chan | 109 | 137 | 151 | 127 | 131 |
| Miss Marple | 116 | 144 | 91 | 153 | 126 |
| Average | 119 | 133 | 136 | 132 | 130 |

**D:** Large values of $S$ and $AS$

| | Decision-Making Condition: | | | | |
|---|---|---|---|---|---|
| Congressman | Normal | Wine | Awaken | Cold | Average |
| Mr. Rockford | 136 | 129 | 137 | 130 | 133 |
| Mr. Columbo | 139 | 143 | 161 | 121 | 141 |
| Mrs. Fletcher | 137 | 152 | 163 | 144 | 149 |
| Mr. Mason | 99 | 115 | 135 | 139 | 122 |
| Mr. Chan | 98 | 126 | 140 | 116 | 120 |
| Miss Marple | 105 | 133 | 80 | 142 | 115 |
| Average | 119 | 133 | 136 | 132 | 130 |

Table 7.2: Sample Data Sets Illustrating $S$ and $AS$

In Data Set A, both $S$ and $AS$ have small effects. The overall level of performance does not vary greatly from one subject to the next, nor does the pattern of condition effects vary much from subject to subject. In the graph, this is apparent from the fact that the lines for the different subjects are all very similar.

In Data Set B, $S$ has a large effect but $AS$ has a small one. The former can be seen from the fact that the overall level of performance varies greatly from one subject to another. For example, Mrs. Fletcher's scores are in the low 100's, but Mr. Chan's scores are around 150. Thus there is a large difference between subjects, and we would expect the $S_j$ terms for this data set to be very different from zero.

Note also that the pattern of effect caused by the experimental factor is almost identical for all sub-

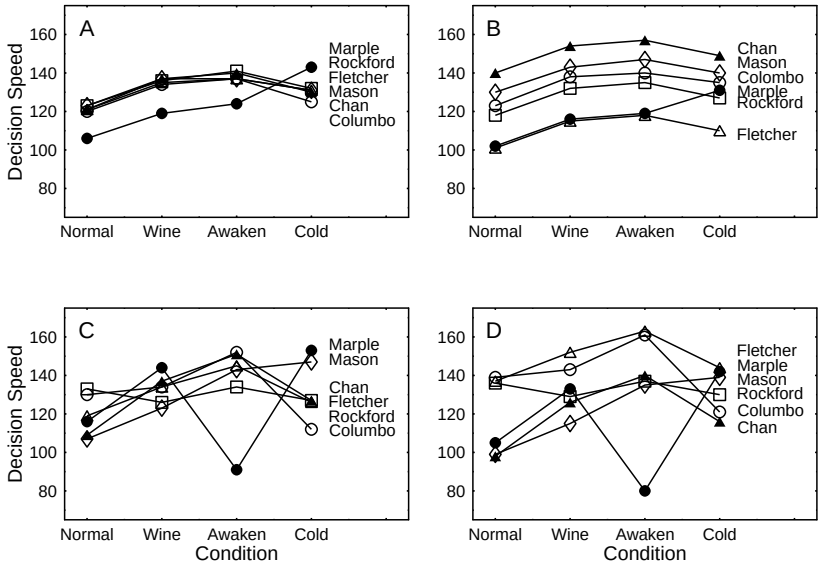


Figure 7.1: Decision time as a function of condition and congressman.

jects in Data Set B. Since every subject shows nearly the same pattern across the four conditions, the $AS$ interaction is small. On average, scores are 11 points lower than baseline in the Normal condition, and 3, 6, and 2 points above baseline for the other three conditions, respectively. More importantly, every subject shows about this same pattern (note the nearly parallel lines in Figure 7.1B). Thus the effect of the experimental factor does not seem to depend on which subject is being tested. Whenever an experimental effect is very consistent from subject to subject, the $AS_{ij}$ terms will be close to zero.

Data Set C shows what happens when $S$ is small but $AS$ is large. The subjects have about the same averages across the different conditions of the experiment—in other words, there is little difference between subjects in overall performance. But different subjects show very different patterns of effects across the experimental conditions, as reflected in the decidedly nonparallel lines of Figure 7.1C. For example, most subjects score highest in the Awaken condition, but Miss Marple gets her lowest score by far in this condition.

Data Set D shows what happens with relatively large values for both $S$ and $AS$. Here, the subjects vary quite a bit in overall level, and also show very different patterns of effects across the four conditions.

To summarize this set of four examples, $S_j$ and $AS_{ij}$ reflect two different types of variation from subject to subject: overall level vs. pattern of effect. These two types of variation are separate from one another, as demonstrated by the fact that it is possible to construct data sets in which they are independently large or small.

It is also instructive to consider why we have the $AS_{ij}$ interaction term in the model for a within-subjects design but we do not have it in the model for a between-subjects design. By definition, a within-subjects design gives us a score for each subject at each level of the experimental factor. Thus, we can measure the effect of Factor A separately for each individual subject. For example, look at any single line in Figure 7.1. The effect of A, for the subject represented by that line, is measured by how much the line deviates from a straight, horizontal line. Because we can measure the effect of A for each subject, we can see how much it varies from subject to subject, and this is what $AS_{ij}$ measures.

In contrast, between-subjects designs do not allow us to measure the effect of an experimental factor for each subject. Each subject is observed at only one level of the factor, so we can only

evaluate the effect of the factor by comparing groups of subjects, not by looking at the data from any individual subject.

We can easily extend this treatment of within-subjects models to a design having two within-subjects factors, A and B. In such a design, each subject would have to be tested at all combinations of levels on Factors A and B. Assuming that each experimental factor has two levels, the data would have the structure shown in Table 7.3.

| | Condition | | | |
|---|---|---|---|---|
| Subject | $AB_{11}$ | $AB_{12}$ | $AB_{21}$ | $AB_{22}$ |
| 1 | 443 | 467 | 453 | 437 |
| 2 | 363 | 387 | 373 | 357 |
| 3 | 353 | 377 | 363 | 347 |
| 4 | 402 | 426 | 416 | 396 |
| 5 | 404 | 428 | 410 | 398 |
| Average | 393 | 417 | 403 | 387 |

Table 7.3: Sample Data for a Two-Factor, Within-Ss Design

The linear model for this design is:

$$Y_{ijk} = \mu + A_i + B_j + AB_{ij} + S_k + AS_{ik} + BS_{jk} + ABS_{ijk}$$

As before, the model contains a baseline, plus main effect and interaction terms for the experimental factors ($A$, $B$, $AB$). We again have a main effect term for subjects, so the model can take into account what is common to all scores from a given subject.

We also have terms for the two-way interactions $AS$ and $BS$. The former measures the extent to which the effect of Factor A varies from subject to subject, and the latter measures the extent to which the effect of Factor B varies from subject to subject. Both of these terms can be understood in just the same way as the $AS$ term in the model for the one-factor within-subjects design above (i.e., Congressmen experiment). The only complication is that now we have to do a little averaging to see each main effect on a given subject. For example, to see the main effect of Factor A on Subject 1, we have to compare the average of the scores at $A_1$ (443 and 467) against the average of the scores at $A_2$ (453 and 437).

The entirely new term is $ABS_{ijk}$, which represents the three-way interaction between Factors A, B, and S. This term is analogous to a three-way interaction term in a fully between-subjects design: it measures the extent to which the $AB_{ij}$ interaction term depends on the level of subject (i.e., varies from subject to subject). Again, the major idea is that we have measured this interaction for each subject separately (note that we could plot the AB interaction for subject 1, for subject 2, etc.), so we can see how much the interaction varies from subject to subject.

In summary, constructing the model for a within-subjects design is very similar to constructing a model for a between-subjects design. Models for within-subjects designs contain all the terms for experimental factors and their interactions that are found in between-subjects models. Instead of a single error term, though, within-subjects models contain a main effect term for subjects (always called $S$), and terms for the interaction of subjects with each main effect and interaction of experimental factors.

The rationale behind the introduction of the subjects factor can be summarized as follows: The linear model is used to describe scores in terms of their common components. When we perform an experiment using a within-subjects design, we obtain more than one score from each subject. These scores have something in common by virtue of the fact that they all came from the same individual, so the linear model must have a term that represents this common influence.

The rationale for the terms representing the interaction of subjects with experimental factors can be summarized like this: Because we have measured each experimental effect on every subject, we can measure how much it varies from one subject to the next. Because we have measured each interaction of experimental factors on every subject, we can also measure how much the interaction varies from subject to subject.

Finally, we can state the general rules needed to construct the linear model for any within-subjects design.

- $Y$ has one subscript for each experimental factor, and one for subjects.
- Include $\mu$, main effects and interactions of all experimental factors.
- Include an $S$ main effect.
- Include an interaction of $S$ with all main effects and interactions of experimental factors. That is, include:
  - All two-way interactions of an experimental factor and $S$.
  - All three-way interactions of two experimental factors and $S$.
  - All four-way interactions of three experimental factors and $S$.
  - And so on, up to the interaction of all experimental factors and $S$.

## 7.3 Estimation and Decomposition

Once we have the correct model for a within-subjects design, no new ideas or procedures are needed to estimate the values of model terms or to compute $SS$'s and *df*'s. The value of a term is estimated by taking the relevant data average and subtracting off whatever is explainable by simpler terms in the model. This process can be repeated as often as necessary to construct a decomposition matrix. As always, the sum of squares for a term can be found by squaring and totalling the numbers in the corresponding column of the decomposition matrix.

Degrees of freedom are found in the usual way for all experimental factors and their interactions. The *df* for subjects is still the number of subjects minus the number of groups, but in this type of design there is only one group. When there are no between-Ss factors, all of the subjects are logically equivalent.

We will now illustrate these general rules with a computational example, to demonstrate precisely how to perform the calculations. In this example we will proceed only as far as the computation of mean squares. The correct computation of $F_{observed}$ is discussed in detail in Section 7.5.

Our first example data are given in Table 7.4. These data are a hypothetical comparison of people's speed of response to the onset of differently colored lights (blue, green and red). Each subject sat in front of a computer monitor and pressed a button as soon as any light appeared. The DV is the time from the onset of the light to the key press response, measured in milliseconds (one-thousandths of a second). Each subject was tested once with each light.

| Subject | Blue | Green | Red | Average |
|---|---|---|---|---|
| 1 | 405 | 399 | 423 | 409 |
| 2 | 405 | 402 | 426 | 411 |
| 3 | 407 | 401 | 428 | 412 |
| 4 | 403 | 398 | 423 | 408 |
| Average | 405 | 400 | 425 | 410 |

Table 7.4: Sample Data for Calculations In One-Factor Within-Ss Design

We begin as always by constructing the linear model for the experimental design. As discussed above, the linear model for a one-factor within-subjects design is:

$$Y_{ij} = \mu + A_i + S_j + AS_{ij}$$

Next we compute estimates for each term in the model. This is done exactly as though the $S$ term in our model was $B$. That is, the estimation process is identical to that for a two-way fully between-subjects design. Thus, $\mu$ is estimated as:

$$\hat{\mu} = Y_{..} = 410$$

$\hat{A}_i$ is estimated by the average of all scores at level $i$ of Factor A, minus $\hat{\mu}$:

$$\begin{aligned}\hat{A}_i &= Y_{i.} - \hat{\mu} \\ \hat{A}_1 &= 405 - 410 = -5 \\ \hat{A}_2 &= 400 - 410 = -10 \\ \hat{A}_3 &= 425 - 410 = 15\end{aligned}$$

To estimate the subject terms, recall that $S_j$ is computationally just another factor in the model. Thus, analogous to the computations just performed for $A$, $\hat{S}_j$ is the average of all scores from subject $j$, minus $\hat{\mu}$.

$$\begin{aligned}\hat{S}_i &= Y_{.j} - \hat{\mu} \\ \hat{S}_1 &= 409 - 410 = -1 \\ \hat{S}_2 &= 411 - 410 = 1 \\ \hat{S}_3 &= 412 - 410 = 2 \\ \hat{S}_4 &= 408 - 410 = -2\end{aligned}$$

The estimation equation for $\widehat{AS}_{ij}$ is exactly what it would be for an $\widehat{AB}_{ij}$ interaction term, with $S$ replacing $B$ throughout.

$$\begin{aligned}\widehat{AS}_{ij} &= Y_{ij} - \hat{\mu} - \hat{A}_i - \hat{S}_j \\ \widehat{AS}_{11} &= 405 - 410 - (-5) - (-1) = 1 \\ &\vdots\end{aligned}$$

Using this equation for each pair of $i$ and $j$ values, we obtain the following table of $\widehat{AS}_{ij}$ values.

| | A1 | A2 | A3 |
|---|---|---|---|
| S1: | 1 | 0 | -1 |
| S2: | -1 | 1 | 0 |
| S3: | 0 | -1 | 1 |
| S4: | 0 | 0 | 0 |

Having estimated all the terms in the model, we can easily construct the decomposition matrix for this analysis—just set each score equal to the sum of its estimated components, as always. The matrix for this data set is shown in Table 7.5.

Given the decomposition matrix, the $SS$ for each source is computed by squaring and totalling all scores in each column of the matrix. The $MS$ for a term is equal to its $SS$ divided by its $df$, as always. So, we have done enough so far to calculate the results shown in the following partial ANOVA table (note that no $F$'s are included—this requires discussion of error terms).

| Source | *df* | *SS* | *MS* | *F* | Error Term |
|---|---|---|---|---|---|
| $\mu$ | 1 | 2,017,200 | 2,017,200 | | |
| $A$ | 2 | 1,400 | 700 | | |
| $S$ | 3 | 30 | 10 | | |
| $AS$ | 6 | 6 | 1 | | |
| Total | 12 | 2,018,636 | | | |

For each source, *df* is calculated using the rules given above. The $df_S$ equals three: four subjects minus one group. The $df_{AS}$ equals $df_A \times df_S = 6$.

This computational example shows that estimation and decomposition proceed in just the same way using a within-subjects model as they do using between-subjects models. Additional computational examples, including some with more factors, are given in Section 7.7.

| $Y_{ij}$ | = | $\mu$ | + | $A_i$ | + | $S_j$ | + | $AS_{ij}$ |
|---|---|---|---|---|---|---|---|---|
| 405 | = | 410 | - | 5 | - | 1 | + | 1 |
| 399 | = | 410 | - | 10 | - | 1 | + | 0 |
| 423 | = | 410 | + | 15 | - | 1 | - | 1 |
| 405 | = | 410 | - | 5 | + | 1 | - | 1 |
| 402 | = | 410 | - | 10 | + | 1 | + | 1 |
| 426 | = | 410 | + | 15 | + | 1 | + | 0 |
| 407 | = | 410 | - | 5 | + | 2 | + | 0 |
| 401 | = | 410 | - | 10 | + | 2 | - | 1 |
| 428 | = | 410 | + | 15 | + | 2 | + | 1 |
| 403 | = | 410 | - | 5 | - | 2 | + | 0 |
| 398 | = | 410 | - | 10 | - | 2 | + | 0 |
| 423 | = | 410 | + | 15 | - | 2 | + | 0 |

Table 7.5: Decomposition Matrix for Colored Light Data

## 7.4 "Random-Effects" vs. "Fixed-Effects" Factors

Before we begin our discussion of how to compute $F_{observed}$ values in within-subjects designs, we must consider an important conceptual difference between the subjects factor and the other experimental factors. In most experiments we want to do more than simply draw conclusions specific to the particular subjects we tested. We generally think of our subjects as representative of some large population of subjects, and we want to draw conclusions that apply to the whole population. Because we want to generalize our conclusions to the whole population, we *randomly* select subjects from that population for testing in our experiment, and we hope that the randomly selected subjects will be representative of the whole population. Because we randomly select the levels of our subjects factor it is called a *random factor* or *random effects factor*.

In the formal language of ANOVA, we say that a factor is a random factor when it satisfies two criteria. A factor is a random factor if and only if *both* of the following conditions apply:

**Random Selection:** The levels of the factor actually included in the experiment were chosen randomly from a larger population of potential levels.

**Generalized Conclusions:** Conclusions are intended to apply to the whole population of potential levels of the random factor, not just the ones actually included in the experiment.

If a factor fails to meet either or both of these criteria, it is called a "fixed effects" or "fixed" factor in the design.

All of the factors we have considered other than the subjects factor were fixed effects factors because we intentionally chose exactly which levels to test. In fact, it is almost always the case that the interesting experimental factors are fixed effects factors, since it is rare that we are interested in differences among conditions that we have sampled randomly. However, the subjects factor is almost always a random effects factor, since we generally select subjects randomly from a population and then try to generalize the results to everyone in that population.

Suppose, for example, that we wanted to compare problem solving by congressmen under various conditions, as discussed above. If we had three particular congressmen that we were interested in comparing (say we were trying to decide which of three candidates for a particular office to vote for), then subjects would be a fixed effects factor. Alternatively, we might be interested in drawing general conclusions about problem solving speed under various conditions, and we would want those conclusions to apply to all congressmen. In this case we would not be interested in any individual congressmen except as representatives of the whole population of congressmen. We would randomly select our subjects, and the subjects factor would then be a random effects factor.

If we were actually interested in generalizing our conclusions to all congressmen, we would surely want to test a large sample of them (say 30) in all the conditions of the experiment, rather than testing only a few congressmen as described above. Testing more congressmen would give us a better chance of having a representative sample, and so would be more likely to lead to accurate conclusions. Note, however, that the number of congressmen tested makes little difference to the structure of the linear model. With 30 subjects instead of 6, we would have the identical model, except the subscript on the subjects term ($S_j$) would range from 1 to 30 instead of 1 to 6. Testing more congressmen does

not change the structure of the experimental design in any way, it merely adds levels to an already existing factor.

## 7.5 Choice of Error Term & Computation of $F_{observed}$

### 7.5.1 Error as a "Yardstick"

In between-subjects designs, the $F_{observed}$ for any term is computed as:

$$F_{Term} = \frac{MS_{Term}}{MS_{S(ABC...)}}$$

Intuitively, the $MS_{S(ABC...)}$ estimates the size of observed effects that one would expect to get entirely by chance—with these subjects, under these conditions. The $MS_{Term}$ measures the difference between groups of subjects in different conditions. If the $F_{Term}$ is large, then the difference between groups was larger than one would expect to get by chance, and we conclude that the observed difference reflects a real population effect. The important idea is that $MS_{S(ABC...)}$ functions as an "error yardstick," against which we measure the observed factor effects.

In a between-subjects design, the $MS$ for subjects is our only measure of random variability between subjects, and is thus the only candidate for the role of yardstick. In within-subjects designs however, we have several measures of subject variability. For example, we have seen that in a one-factor design, $S_j$ measures overall subject variability, while $AS_{ij}$ measures the extent to which the effect of Factor A varies depending on the subject being tested. Both of these sources in some sense reflect sampling error (i.e., how much variability we would expect to find in our data due to chance), yet they represent different sources of that variability. Which $MS$ should serve as the denominator for the $F_{observed}$?

The answer is that different $MS$'s are used to compute $F$'s for different sources. To choose the correct error $MS$ for each $F$, we must consider carefully which error measure provides the appropriate yardstick for each experimental effect or interaction.

### 7.5.2 Error Term for $\mu$

We begin by finding the appropriate yardstick for testing $H_0$: $\mu = 0$. $\hat{\mu}$ provides an estimate of the overall population mean; thus, our confidence that the population mean differs from zero depends both on the size of $\hat{\mu}$ and on the amount of variability in our data set. If $\hat{\mu}$ is very large (positively or negatively), we are more confident that the population mean is in fact non-zero than if $\hat{\mu}$ is very small. However, to the extent that the scores in our data set are wildly variable (some positive and some negative), we will have our doubts even about a very large $\hat{\mu}$. That is, a small $\hat{\mu}$ in a very homogenous data set may reflect a non-zero population mean while a larger $\hat{\mu}$ from a very heterogenous data set may not.

But which type of variability is relevant, that reflected in $S_j$ or that reflected in $AS_{ij}$? The answer is $S_j$. To see why, compare Data Sets B and C in Table 7.2. In both cases, the overall average is $\hat{\mu} = 130$. In which case, though, are you more sure that the true $\mu$ is close to 130, say between 128 and 132?

Consider Data Set B. If we were to randomly sample two new subjects, we could reasonably expect to get new subjects whose averages were anywhere in the range of about 110 to 150, because we have already gotten subject averages in that range (see "average" column at far right of table). It is pretty easy to imagine getting two new subjects with means in the 140 to 150 range, which would raise the group average well above 130. In other words, because of the high variability in the subject average column, the overall experiment average of 130 does not seem like a very stable number.

Now consider Data Set C. Here, if we were to sample two new subjects, we would expect their averages to fall in approximately the range of 125-135. There is little variability in the averages of the subjects we have already observed, and so we would be surprised to see a new subject who was really discrepant from the overall average of 130. Thus, because of the low variability in the subject average column, the overall experiment average of 130 seems more stable than the average in Data Set B.

What this comparison shows is that $S$ variability is a good yardstick for uncertainty about $\mu$. When $S$ variability is high (Data Set B but not C), we have relatively more uncertainty about $\mu$. When $AS$ variability is high (Data Set C but not B), though, it does not increase uncertainty about $\mu$. That is why we use $S$ variability (i.e., $MS_S$) rather than $AS$ variability (i.e., $MS_{AS}$) as the error yardstick to test a hypothesis about $\mu$.

In summary, when we are interested in whether the overall mean might be different from zero by chance, the relevant yardstick is the variability from subject to subject in overall score, and this variability is indexed by the $S_j$ term. If the subjects are comparatively similar to one another in overall level, then the average score observed in the experiment is probably pretty close to the true average. If subjects vary wildly in overall level, then the average score observed in the experiment may not be so close to the true average. Thus,

$$F_\mu = \frac{MS_\mu}{MS_S}$$

measures the ratio of the size of $\hat{\mu}$ to the amount of variability in the subject averages.

### 7.5.3 Error Term for a Main Effect

Let us now consider how to compute $F_{observed}$ for a main effect. If we observe a large difference between various levels of an experimental factor, we are more likely to believe that the effect actually exists in the population than if we observe a very small difference. However, just as in all previous calculations of $F_{observed}$, we need a comparison value reflecting how much of an observed difference might arise by chance.

The key insight is that any given observed difference is likely to be real (i.e., unlikely to have arisen by chance) *if each of the subjects showed approximately the same effect*. On the other hand, a given difference is less likely to be real (i.e., more likely to have arisen by chance) if each of the subjects showed a rather different effect. In other words, the effect is probably real if it is consistent across subjects, but it is probably just chance if it is very inconsistent across subjects.

To develop your understanding of this point, examine the data sets plotted in Figure 7.1. The effect of decision-making condition is the same, on average, in all four data sets, as is apparent from the identical marginal averages for Condition in Table 7.2. But the data sets are not equally convincing that this effect is representative of the whole population from which these subjects were sampled. In Data Set A, the effect certainly appears real. The results are very consistent across subjects, and so we can expect that we would get pretty much the same results if we sampled some other subjects.

Now consider Data Set B. These subjects are pretty different from one another in overall level (i.e., separation between lines in the figure, caused by large values of $S_j$), but they all show pretty *similar effects of decision-making condition* (i.e., nearly parallel lines in the figure, caused by small values of $AS_{ij}$). If we take some new subjects from this population, we would certainly expect substantial variation in the overall level of the subjects' performance, as discussed in the previous section. But we would not expect much variation in the effect of decision-making condition; this effect is very consistent across these subjects, and so we would expect it to be consistent across other subjects from this population as well. Obviously the effect is more likely to be real if it is there for all subjects. This means that either of the two data sets, A or B, would be pretty good evidence that the effect of decision-making seen in the averages is representative of the whole population, precisely because *the size of the effect is consistent across subjects*. Not coincidentally, the $AS_{ij}$ terms are small in both of these data sets.

In contrast, Data Sets C and D show relatively unconvincing evidence for an effect of decision-making condition, even though the effect on average is the same size as in Data Sets A and B. The reason is apparent in the plots: the effect of decision-making is very inconsistent across subjects in each case. Looking at either of these plots, it is easy to imagine taking some new subjects with effects quite different from those shown in the means, and so we can have relatively little confidence that the means of the data set are representative of the population as a whole.

To sum up: To see whether a main effect is real or due to chance, we need an error yardstick that measures how much the effect varies from subject to subject. Low variation suggests a real effect, whereas high variation suggests a chance effect.

In within-subjects designs, we have exactly the yardstick we require: the $MS_{AS}$ interaction term. By definition, this measures the extent to which the effect of Factor A varies from subject to subject. A large $MS_{AS}$ means that the effect of Factor A varies considerably from subject to subject, whereas a small $MS_{AS}$ means that Factor A has a very consistent effect. Thus, we compute the $F$ for a main effect as:

$$F_A = \frac{MS_A}{MS_{AS}}$$

### 7.5.4 Error Terms for Interactions

The above discussion can be easily extended to higher-order terms in within-subjects models. For example, suppose we want to determine whether an observed AB interaction is real or due to chance. We will tend to believe the interaction is real if it is consistent—that is, if every subject shows the same pattern of interaction, or nearly so. Thus, we want to compare the observed AB interaction against a term which measures how much the AB interaction varies from subject to subject. But that variation is exactly what the ABS interaction term measures. Thus the $F_{AB}$ is:

$$F_{AB} = \frac{MS_{AB}}{MS_{ABS}}$$

Again, the rationale is exactly like that for the main effect: If the interaction is consistent from subject to subject we can conclude that it is real, but if it is inconsistent (highly variable) we have more reason to think that the pattern in the AB cell averages was observed by chance.

You can make the same argument for an interaction of any complexity you care to consider. Is an observed ABCD interaction real or due to chance? Well, is it consistent across subjects (low $MS_{ABCDS}$) or not?

### 7.5.5 Error Term Summary

The concept of consistency underlies the rules for choosing a correct error yardstick—formally known as "error term"—in within-subjects designs.

- To test $H_0$: $\mu$= 0, compute $F_{\mu} = \frac{MS_{\mu}}{MS_S}$.
- To test $H_0$: No effect of a factor, compute $F_{factor} = \frac{MS_{factor}}{MS_{factor\ by\ S\ interaction}}$.
- To test $H_0$: No interaction of factors ABC..., compute $F_{ABC...} = \frac{MS_{ABC...}}{MS_{ABC...S}}$.

Note that we do not normally compute *F*'s for the effect of subjects or interactions involving subjects. One reason is that it is not regarded as meaningful to test the hypothesis that some randomly selected subjects are different from one another, or that randomly selected subjects show different effects of an experimental manipulation. Generally, this is assumed, or known from previous research. If subjects were not different from one another, why would we bother to sample more than one?

Occasionally, it is desirable to test the $H_0$'s that the particular subjects in an experiment are all the same in overall level, all have the same effect of Factor A, all show the same AB interaction, etc. The most appropriate way to do this is to treat Subjects as a fixed-effects factor and use an analysis for a between-subjects design. For this analysis, you are not regarding the individuals tested as randomly selected subjects, but rather as a fixed sample. In fact, for this analysis the random "subject" really corresponds to separate tests of a given individual. This analysis requires more than one observation per subject in each condition. Then, you can do a between-subjects analysis treating subjects as a regular experimental factor, and you will get the appropriate *F*'s for testing the hypotheses about subjects. Note that in this analysis the *F*'s for the other factors test whether or not there are effects *for these subjects*. It would be inappropriate to generalize significant effects on these subjects to the whole population, because this analysis uses the wrong error term.

### 7.5.6 Comparison with Between-Subjects Designs

Now that you understand where error terms come from in within-subjects designs, you are in a good position to understand the old experimentalist's rubric that within-subjects designs have more power than between-subjects designs. This is an important concept in experimental design, so it is worth a brief diversion from ANOVA computations.

The major flaw of between-subjects designs is that they do not provide an estimate of how the effect of an experimental factor varies from subject to subject. Each subject is tested in only one condition, so we cannot see how an individual subject's score changes across conditions. That is, we cannot measure the effect of conditions *for any single subject*, so obviously we cannot see how that effect varies from one subject to the next. More or less by default, these designs force us to use within-cell error as our yardstick, simply because there is no alternative.

Consider what this within-cell error really is, however, in terms of the concepts of within-subjects designs. In a between-subjects design, two subjects in the same cell may differ from one another either because they are different in overall level or because they respond differently to that level of the factor. That is,

$$S(A)_{ij} = S_j + AS_{ij}$$

where the left side of the equation shows the between-subjects error term and the right side shows the corresponding sum of within-subjects terms.

We now see why within-subjects designs tend to have smaller error terms (and thus be more powerful) than between-subjects designs. In a within-subjects design we can essentially reduce our error by separating out random variability in overall subject scores from random variability in the size of each subject's main effect. If the main effect is consistent across subjects, then the error term will be small even if subjects vary widely in overall level. Between-subjects designs are forced to lump both sources of error together, which may obscure a true population effect. For example, consider the experiment to see whether people weigh more with their clothes on or off, discussed in Section 7.1. In the within-subjects design, only variation in the weight of the clothes enters into the error term, not variation in overall weight of the subject. In the between-subjects version both types of variation enter into the error term, which is thereby much larger. We will discuss this concept in more detail below when we look at some data sets and see the actual computations.

## 7.6 Interpretation of $F$'s

The interpretation of results in within-subjects designs is no different from that in between-subjects designs. Significant main effects and interactions should be described and graphed. The fact that the factors were tested within-subjects rather than between-subjects generally has no influence on the interpretation.[6]

## 7.7 Computational Examples

**1.** Design with one within-subjects factor. Subjects have to make a number of decisions in each of three conditions varying in the type of pressure they are under. In a control condition, there is no pressure. In a high-speed condition, there is time pressure (i.e., subjects have to decide quickly). In a high-importance condition, a lot depends on whether the subject gets the decision right or not (e.g., big money bonuses for correct decisions, like on TV game shows). The DV is the percentage of correct decisions in each condition.

**Data:**

| | Type of Pressure | | |
|---|---|---|---|
| | None | Time | Importance |
| Subject 1: | 70 | 51 | 44 |
| Subject 2: | 82 | 66 | 59 |
| Subject 3: | 90 | 63 | 66 |
| Subject 4: | 70 | 59 | 57 |
| Subject 5: | 77 | 62 | 71 |
| Subject 6: | 61 | 53 | 69 |

**Model:** $Y_{ij} = \mu + A_i + S_j + AS_{ij}$

[6] There are some examples where effects are really larger in a within-subjects design than in a between-subjects design, or vice versa. Such findings are not a statistical matter, however, but reflect on the particular area of research in which they are obtained.

**Estimation Equations:**

$$\begin{aligned} \hat{\mu} &= Y_{\cdot\cdot} \\ \hat{A}_i &= Y_{i\cdot} - \hat{\mu} \\ \hat{S}_i &= Y_{\cdot j} - \hat{\mu} \\ \widehat{AS}_{ij} &= Y_{ij} - \hat{\mu} - \hat{A}_i - \hat{S}_j \end{aligned}$$

**Decomposition Matrix:**

| $Y_{ij}$ | = | $\mu$ | + | $A_i$ | + | $S_j$ | + | $AS_{ij}$ |
|---|---|---|---|---|---|---|---|---|
| 70 | = | 65 | + | 10 | - | 10 | + | 5 |
| 51 | = | 65 | - | 6 | - | 10 | + | 2 |
| 44 | = | 65 | - | 4 | - | 10 | - | 7 |
| 82 | = | 65 | + | 10 | + | 4 | + | 3 |
| 66 | = | 65 | - | 6 | + | 4 | + | 3 |
| 59 | = | 65 | - | 4 | + | 4 | - | 6 |
| 90 | = | 65 | + | 10 | + | 8 | + | 7 |
| 63 | = | 65 | - | 6 | + | 8 | - | 4 |
| 66 | = | 65 | - | 4 | + | 8 | - | 3 |
| 70 | = | 65 | + | 10 | - | 3 | - | 2 |
| 59 | = | 65 | - | 6 | - | 3 | + | 3 |
| 57 | = | 65 | - | 4 | - | 3 | - | 1 |
| 77 | = | 65 | + | 10 | + | 5 | - | 3 |
| 62 | = | 65 | - | 6 | + | 5 | - | 2 |
| 71 | = | 65 | - | 4 | + | 5 | + | 5 |
| 61 | = | 65 | + | 10 | - | 4 | - | 10 |
| 53 | = | 65 | - | 6 | - | 4 | - | 2 |
| 69 | = | 65 | - | 4 | - | 4 | + | 12 |

**ANOVA:**

| Source | *df* | $SS$ | $MS$ | $F$ | Error Term |
|---|---|---|---|---|---|
| $\mu$ | 1 | 76050.0 | 76050.0 | 551.09 | $S$ |
| $A$ | 2 | 912.0 | 456.0 | 9.01 | $AS$ |
| $S$ | 5 | 690.0 | 138.0 | | |
| $AS$ | 10 | 506.0 | 50.6 | | |
| Total | 18 | 78158.0 | | | |

**2.** Design with two within-subjects factors.

**Design:**

- Factor A: 2 levels, Within-subjects.
- Factor B: 2 levels, Within-subjects.
- 3 Subjects per group. (Factor $S$ has 3 levels.)

**Data:**

| | B1 | | B2 | | Subject |
|---|---|---|---|---|---|
| | A1 | A2 | A1 | A2 | Averages |
| Subject 1 | 38 | 28 | 44 | 34 | 36 |
| Subject 2 | 41 | 35 | 41 | 39 | 39 |
| Subject 3 | 32 | 42 | 62 | 44 | 45 |
| Averages | 37 | 35 | 49 | 39 | 40 |

**Model:** $Y_{ijk} = \mu + A_i + B_j + AB_{ij} + S_k + AS_{ik} + BS_{jk} + ABS_{ijk}$

**Estimation Equations:**

$$\begin{aligned}
\hat{\mu} &= Y_{...} \\
\hat{A}_i &= Y_{i..} - \hat{\mu} \\
\hat{B}_j &= Y_{.j.} - \hat{\mu} \\
\widehat{AB}_{ij} &= Y_{ij.} - \hat{\mu} - \hat{A}_i - \hat{B}_j \\
\hat{S}_k &= Y_{..k} - \hat{\mu} \\
\widehat{AS}_{ik} &= Y_{i.k} - \hat{\mu} - \hat{A}_i - \hat{S}_k \\
\widehat{BS}_{jk} &= Y_{.jk} - \hat{\mu} - \hat{B}_j - \hat{S}_k \\
\widehat{ABS}_{ijk} &= Y_{ijk} - \hat{\mu} - \hat{A}_i - \hat{B}_j - \widehat{AB}_{ij} - \hat{S}_k - \widehat{AS}_{ik} - \widehat{BS}_{jk}
\end{aligned}$$

**Decomposition Matrix:**

| $Y_{ijk}$ | = | $\hat{\mu}$ | + | $\hat{A}_i$ | + | $\hat{B}_j$ | + | $\widehat{AB}_{ij}$ | + | $\hat{S}_k$ | + | $\widehat{AS}_{ik}$ | + | $BS_{jk}$ | + | $\widehat{ABS}_{ijk}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 38 | = | 40 | + | 3 | - | 4 | - | 2 | - | 4 | + | 2 | + | 1 | + | 2 |
| 28 | = | 40 | - | 3 | - | 4 | + | 2 | - | 4 | - | 2 | + | 1 | - | 2 |
| 44 | = | 40 | + | 3 | + | 4 | + | 2 | - | 4 | + | 2 | - | 1 | - | 2 |
| 34 | = | 40 | - | 3 | + | 4 | - | 2 | - | 4 | - | 2 | - | 1 | + | 2 |
| 41 | = | 40 | + | 3 | - | 4 | - | 2 | - | 1 | - | 1 | + | 3 | + | 3 |
| 35 | = | 40 | - | 3 | - | 4 | + | 2 | - | 1 | + | 1 | + | 3 | - | 3 |
| 41 | = | 40 | + | 3 | + | 4 | + | 2 | - | 1 | - | 1 | - | 3 | - | 3 |
| 39 | = | 40 | - | 3 | + | 4 | - | 2 | - | 1 | + | 1 | - | 3 | + | 3 |
| 32 | = | 40 | + | 3 | - | 4 | - | 2 | + | 5 | - | 1 | - | 4 | - | 5 |
| 42 | = | 40 | - | 3 | - | 4 | + | 2 | + | 5 | + | 1 | - | 4 | + | 5 |
| 62 | = | 40 | + | 3 | + | 4 | + | 2 | + | 5 | - | 1 | + | 4 | + | 5 |
| 44 | = | 40 | - | 3 | + | 4 | - | 2 | + | 5 | + | 1 | + | 4 | - | 5 |

**ANOVA Table:**

| Source | *df* | *SS* | *MS* | *F* | Error Term |
|---|---|---|---|---|---|
| $\mu$ | 1 | 19200.0 | 19200.0 | 228.571 | *S* |
| *A* | 1 | 108.0 | 108.0 | 9.000 | *AS* |
| *B* | 1 | 192.0 | 192.0 | 3.692 | *BS* |
| *AB* | 1 | 48.0 | 48.0 | 0.632 | ABS |
| *S* | 2 | 168.0 | 84.0 | | |
| *AS* | 2 | 24.0 | 12.0 | | |
| *BS* | 2 | 104.0 | 52.0 | | |
| *ABS* | 2 | 152.0 | 76.0 | | |
| Total | 12 | 19996.0 | | | |

**3.** Another design with two within-subjects factors.

**Design:**

- Factor A: 2 levels, Within-subjects.
- Factor B: 3 levels, Within-subjects.
- 5 Subjects per group.

**Data:**

| | B1 | | B2 | | B3 | | Subject |
|---|---|---|---|---|---|---|---|
| | A1 | A2 | A1 | A2 | A1 | A2 | Average |
| Subject 1 | 63 | 31 | 71 | 53 | 58 | 36 | 52 |
| Subject 2 | 74 | 40 | 75 | 53 | 55 | 51 | 58 |
| Subject 3 | 81 | 35 | 83 | 51 | 73 | 49 | 62 |
| Subject 4 | 79 | 39 | 87 | 59 | 65 | 49 | 63 |
| Subject 5 | 33 | 15 | 29 | 29 | 49 | 25 | 30 |
| Average | 66 | 32 | 69 | 49 | 60 | 42 | 53 |

**Model and Estimation Equations:** Same as previous design.

**Decomposition Matrix:**

| $Y_{ijk}$ | = | $\hat{\mu}$ | + | $\hat{A}_i$ | + | $\hat{B}_j$ | + | $\widehat{AB}_{ij}$ | + | $\hat{S}_k$ | + | $\widehat{AS}_{ik}$ | + | $BS_{jk}$ | + | $\widehat{ABS}_{ijk}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 63 | = | 53 | + | 12 | - | 4 | + | 5 | - | 1 | + | 0 | - | 1 | - | 1 |
| 31 | = | 53 | - | 12 | - | 4 | - | 5 | - | 1 | + | 0 | - | 1 | + | 1 |
| 71 | = | 53 | + | 12 | + | 6 | - | 2 | - | 1 | + | 0 | + | 4 | - | 1 |
| 53 | = | 53 | - | 12 | + | 6 | + | 2 | - | 1 | + | 0 | + | 4 | + | 1 |
| 58 | = | 53 | + | 12 | - | 2 | - | 3 | - | 1 | + | 0 | - | 3 | + | 2 |
| 36 | = | 53 | - | 12 | - | 2 | + | 3 | - | 1 | + | 0 | - | 3 | - | 2 |
| 74 | = | 53 | + | 12 | - | 4 | + | 5 | + | 5 | - | 2 | + | 3 | + | 2 |
| 40 | = | 53 | - | 12 | - | 4 | - | 5 | + | 5 | + | 2 | + | 3 | - | 2 |
| 75 | = | 53 | + | 12 | + | 6 | - | 2 | + | 5 | - | 2 | + | 0 | + | 3 |
| 53 | = | 53 | - | 12 | + | 6 | + | 2 | + | 5 | + | 2 | + | 0 | - | 3 |
| 55 | = | 53 | + | 12 | - | 2 | - | 3 | + | 5 | - | 2 | - | 3 | - | 5 |
| 51 | = | 53 | - | 12 | - | 2 | + | 3 | + | 5 | + | 2 | - | 3 | + | 5 |
| 81 | = | 53 | + | 12 | - | 4 | + | 5 | + | 9 | + | 5 | + | 0 | + | 1 |
| 35 | = | 53 | - | 12 | - | 4 | - | 5 | + | 9 | - | 5 | + | 0 | - | 1 |
| 83 | = | 53 | + | 12 | + | 6 | - | 2 | + | 9 | + | 5 | - | 1 | + | 1 |
| 51 | = | 53 | - | 12 | + | 6 | + | 2 | + | 9 | - | 5 | - | 1 | - | 1 |
| 73 | = | 53 | + | 12 | - | 2 | - | 3 | + | 9 | + | 5 | + | 1 | - | 2 |
| 49 | = | 53 | - | 12 | - | 2 | + | 3 | + | 9 | - | 5 | + | 1 | + | 2 |
| 79 | = | 53 | + | 12 | - | 4 | + | 5 | + | 10 | + | 2 | + | 0 | + | 1 |
| 39 | = | 53 | - | 12 | - | 4 | - | 5 | + | 10 | - | 2 | + | 0 | - | 1 |
| 87 | = | 53 | + | 12 | + | 6 | - | 2 | + | 10 | + | 2 | + | 4 | + | 2 |
| 59 | = | 53 | - | 12 | + | 6 | + | 2 | + | 10 | - | 2 | + | 4 | - | 2 |
| 65 | = | 53 | + | 12 | - | 2 | - | 3 | + | 10 | + | 2 | - | 4 | - | 3 |
| 49 | = | 53 | - | 12 | - | 2 | + | 3 | + | 10 | - | 2 | - | 4 | + | 3 |
| 33 | = | 53 | + | 12 | - | 4 | + | 5 | - | 23 | - | 5 | - | 2 | - | 3 |
| 15 | = | 53 | - | 12 | - | 4 | - | 5 | - | 23 | + | 5 | - | 2 | + | 3 |
| 29 | = | 53 | + | 12 | + | 6 | - | 2 | - | 23 | - | 5 | - | 7 | - | 5 |
| 29 | = | 53 | - | 12 | + | 6 | + | 2 | - | 23 | + | 5 | - | 7 | + | 5 |
| 49 | = | 53 | + | 12 | - | 2 | - | 3 | - | 23 | - | 5 | + | 9 | + | 8 |
| 25 | = | 53 | - | 12 | - | 2 | + | 3 | - | 23 | + | 5 | + | 9 | - | 8 |

**ANOVA Table:**

| Source | *df* | *SS* | *MS* | *F* | Error Term |
|---|---|---|---|---|---|
| $\mu$ | 1 | 84270.0 | 84270.0 | 76.332 | $S$ |
| $A$ | 1 | 4320.0 | 4320.0 | 49.655 | $AS$ |
| $B$ | 2 | 560.0 | 280.0 | 5.283 | $BS$ |
| $AB$ | 2 | 380.0 | 190.0 | 4.691 | $ABS$ |
| $S$ | 4 | 4416.0 | 1104.0 | | |
| $AS$ | 4 | 348.0 | 87.0 | | |
| $BS$ | 8 | 424.0 | 53.0 | | |
| $ABS$ | 8 | 324.0 | 40.5 | | |
| Total | 30 | 95042.0 | | | |

**4.** Design with 3 Within-subjects Factors.

**Design:**

- Factor A: 2 levels, Within-subjects.
- Factor B: 3 levels, Within-subjects.
- Factor C: 3 levels, Within-subjects.
- 3 Subjects per group.

**Data:**

| | A1 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | C1 | | | C2 | | | C3 | | |
| Subject | B1 | B2 | B3 | B1 | B2 | B3 | B1 | B2 | B3 |
| 1 | 54 | 51 | 33 | 50 | 45 | 49 | 22 | 27 | 38 |
| 2 | 48 | 47 | 34 | 48 | 49 | 47 | 27 | 30 | 39 |
| 3 | 75 | 70 | 56 | 70 | 65 | 63 | 74 | 66 | 46 |

| | A2 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | C1 | | | C2 | | | C3 | | |
| Subject | B1 | B2 | B3 | B1 | B2 | B3 | B1 | B2 | B3 |
| 1 | 74 | 65 | 65 | 54 | 45 | 57 | 52 | 49 | 52 |
| 2 | 86 | 69 | 58 | 66 | 57 | 57 | 67 | 60 | 47 |
| 3 | 71 | 58 | 66 | 84 | 63 | 75 | 76 | 56 | 72 |

**Model:**

$$\begin{aligned} Y_{ijkl} &= \mu + A_i + B_j + AB_{ij} + C_k + AC_{ik} + BC_{jk} + ABC_{ijk} \\ &\quad + S_l + AS_{il} + BS_{jl} + ABS_{ijl} + CS_{kl} + ACS_{ikl} + BCS_{jkl} + ABCS_{ijkl} \end{aligned}$$

**Estimation Equations:**

$$\begin{aligned}
\hat{\mu} &= Y_{....} \\
\hat{A}_i &= Y_{i...} - \hat{\mu} \\
\hat{B}_j &= Y_{.j..} - \hat{\mu} \\
\widehat{AB}_{ij} &= Y_{ij..} - \hat{\mu} - \hat{A}_i - \hat{B}_j \\
\hat{C}_k &= Y_{..k.} - \hat{\mu} \\
\widehat{AC}_{ik} &= Y_{i.k.} - \hat{\mu} - \hat{A}_i - \hat{C}_k \\
\widehat{BC}_{jk} &= Y_{.jk.} - \hat{\mu} - \hat{B}_j - \hat{C}_k \\
\widehat{ABC}_{ijk} &= Y_{ijk.} - \hat{\mu} - \hat{A}_i - \hat{B}_j - \widehat{AB}_{ij} - \hat{C}_k - \widehat{AC}_{ik} - \widehat{BC}_{jk} \\
\hat{S}_l &= Y_{...l} - \hat{\mu} \\
\widehat{AS}_{il} &= Y_{i..l} - \hat{\mu} - \hat{A}_i - \hat{S}_l \\
\widehat{BS}_{jl} &= Y_{.j.l} - \hat{\mu} - \hat{B}_j - \hat{S}_l \\
\widehat{ABS}_{ijl} &= Y_{ij.l} - \hat{\mu} - \hat{A}_i - \hat{B}_j - \widehat{AB}_{ij} - \hat{S}_l - \widehat{AS}_{il} - \widehat{BS}_{jl} \\
\widehat{CS}_{kl} &= Y_{..kl} - \hat{\mu} - \hat{C}_k - \hat{S}_l \\
\widehat{ACS}_{ikl} &= Y_{i.kl} - \hat{\mu} - \hat{A}_i - \hat{C}_k - \widehat{AC}_{ik} - \hat{S}_l - \widehat{AS}_{il} - \widehat{CS}_{kl} \\
\widehat{BCS}_{ikl} &= Y_{.jkl} - \hat{\mu} - \hat{B}_j - \hat{C}_k - \widehat{BC}_{jk} - \hat{S}_l - \widehat{BS}_{jl} - \widehat{CS}_{kl} \\
\widehat{ABCS}_{ijkl} &= Y_{ijkl} - \hat{\mu} - \hat{A}_i - \hat{B}_j - \widehat{AB}_{ij} - \hat{C}_k - \widehat{AC}_{ik} - \widehat{BC}_{jk} - \widehat{ABC}_{ijk} \\
&\quad - \hat{S}_l - \widehat{AS}_{il} - \widehat{BS}_{jl} - \widehat{ABS}_{ijl} - \widehat{CS}_{kl} - \widehat{ACS}_{ikl} - \widehat{BCS}_{ikl}
\end{aligned}$$

**Decomposition Matrix:**

| $Y_{ijk}$ | = | $\hat{\mu}$ | + | $\hat{A}_i$ | + | $\hat{B}_j$ | + | $\widehat{AB}_{ij}$ | + | $\hat{C}_k$ | + | $\widehat{AC}_{ik}$ | + | $\widehat{BC}_{jk}$ | + | $\widehat{ABC}_{ijk}$ | + | ... |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 54 | = | 56 | - | 7 | + | 5 | - | 2 | + | 4 | - | 1 | + | 3 | + | 1 | + | ... |
| 74 | = | 56 | + | 7 | + | 5 | + | 2 | + | 4 | + | 1 | + | 3 | - | 1 | + | ... |
| 51 | = | 56 | - | 7 | - | 2 | + | 3 | + | 4 | - | 1 | + | 2 | + | 1 | + | ... |
| 65 | = | 56 | + | 7 | - | 2 | - | 3 | + | 4 | + | 1 | + | 2 | - | 1 | + | ... |
| 33 | = | 56 | - | 7 | - | 3 | - | 1 | + | 4 | - | 1 | - | 5 | - | 2 | + | ... |
| 65 | = | 56 | + | 7 | - | 3 | + | 1 | + | 4 | + | 1 | - | 5 | + | 2 | + | ... |
| 50 | = | 56 | - | 7 | + | 5 | - | 2 | + | 2 | + | 3 | - | 1 | + | 0 | + | ... |
| 54 | = | 56 | + | 7 | + | 5 | + | 2 | + | 2 | - | 3 | - | 1 | + | 0 | + | ... |
| 45 | = | 56 | - | 7 | - | 2 | + | 3 | + | 2 | + | 3 | - | 2 | + | 0 | + | ... |
| 45 | = | 56 | + | 7 | - | 2 | - | 3 | + | 2 | - | 3 | - | 2 | + | 0 | + | ... |
| 49 | = | 56 | - | 7 | - | 3 | - | 1 | + | 2 | + | 3 | + | 3 | + | 0 | + | ... |
| 57 | = | 56 | + | 7 | - | 3 | + | 1 | + | 2 | - | 3 | + | 3 | + | 0 | + | ... |
| 22 | = | 56 | - | 7 | + | 5 | - | 2 | - | 6 | - | 2 | - | 2 | - | 1 | + | ... |
| 52 | = | 56 | + | 7 | + | 5 | + | 2 | - | 6 | + | 2 | - | 2 | + | 1 | + | ... |
| 27 | = | 56 | - | 7 | - | 2 | + | 3 | - | 6 | - | 2 | + | 0 | - | 1 | + | ... |
| 49 | = | 56 | + | 7 | - | 2 | - | 3 | - | 6 | + | 2 | + | 0 | + | 1 | + | ... |
| 38 | = | 56 | - | 7 | - | 3 | - | 1 | - | 6 | - | 2 | + | 2 | + | 2 | + | ... |
| 52 | = | 56 | + | 7 | - | 3 | + | 1 | - | 6 | + | 2 | + | 2 | - | 2 | + | ... |
| 48 | = | 56 | - | 7 | + | 5 | - | 2 | + | 4 | - | 1 | + | 3 | + | 1 | + | ... |
| 86 | = | 56 | + | 7 | + | 5 | + | 2 | + | 4 | + | 1 | + | 3 | - | 1 | + | ... |
| 47 | = | 56 | - | 7 | - | 2 | + | 3 | + | 4 | - | 1 | + | 2 | + | 1 | + | ... |
| 69 | = | 56 | + | 7 | - | 2 | - | 3 | + | 4 | + | 1 | + | 2 | - | 1 | + | ... |
| 34 | = | 56 | - | 7 | - | 3 | - | 1 | + | 4 | - | 1 | - | 5 | - | 2 | + | ... |
| 58 | = | 56 | + | 7 | - | 3 | + | 1 | + | 4 | + | 1 | - | 5 | + | 2 | + | ... |
| 48 | = | 56 | - | 7 | + | 5 | - | 2 | + | 2 | + | 3 | - | 1 | + | 0 | + | ... |
| 66 | = | 56 | + | 7 | + | 5 | + | 2 | + | 2 | - | 3 | - | 1 | + | 0 | + | ... |
| 49 | = | 56 | - | 7 | - | 2 | + | 3 | + | 2 | + | 3 | - | 2 | + | 0 | + | ... |
| 57 | = | 56 | + | 7 | - | 2 | - | 3 | + | 2 | - | 3 | - | 2 | + | 0 | + | ... |
| 47 | = | 56 | - | 7 | - | 3 | - | 1 | + | 2 | + | 3 | + | 3 | + | 0 | + | ... |
| 57 | = | 56 | + | 7 | - | 3 | + | 1 | + | 2 | - | 3 | + | 3 | + | 0 | + | ... |
| 27 | = | 56 | - | 7 | + | 5 | - | 2 | - | 6 | - | 2 | - | 2 | - | 1 | + | ... |
| 67 | = | 56 | + | 7 | + | 5 | + | 2 | - | 6 | + | 2 | - | 2 | + | 1 | + | ... |
| 30 | = | 56 | - | 7 | - | 2 | + | 3 | - | 6 | - | 2 | + | 0 | - | 1 | + | ... |
| 60 | = | 56 | + | 7 | - | 2 | - | 3 | - | 6 | + | 2 | + | 0 | + | 1 | + | ... |
| 39 | = | 56 | - | 7 | - | 3 | - | 1 | - | 6 | - | 2 | + | 2 | + | 2 | + | ... |
| 47 | = | 56 | + | 7 | - | 3 | + | 1 | - | 6 | + | 2 | + | 2 | - | 2 | + | ... |
| 75 | = | 56 | - | 7 | + | 5 | - | 2 | + | 4 | - | 1 | + | 3 | + | 1 | + | ... |
| 71 | = | 56 | + | 7 | + | 5 | + | 2 | + | 4 | + | 1 | + | 3 | - | 1 | + | ... |
| 70 | = | 56 | - | 7 | - | 2 | + | 3 | + | 4 | - | 1 | + | 2 | + | 1 | + | ... |
| 58 | = | 56 | + | 7 | - | 2 | - | 3 | + | 4 | + | 1 | + | 2 | - | 1 | + | ... |
| 56 | = | 56 | - | 7 | - | 3 | - | 1 | + | 4 | - | 1 | - | 5 | - | 2 | + | ... |
| 66 | = | 56 | + | 7 | - | 3 | + | 1 | + | 4 | + | 1 | - | 5 | + | 2 | + | ... |
| 70 | = | 56 | - | 7 | + | 5 | - | 2 | + | 2 | + | 3 | - | 1 | + | 0 | + | ... |
| 84 | = | 56 | + | 7 | + | 5 | + | 2 | + | 2 | - | 3 | - | 1 | + | 0 | + | ... |
| 65 | = | 56 | - | 7 | - | 2 | + | 3 | + | 2 | + | 3 | - | 2 | + | 0 | + | ... |
| 63 | = | 56 | + | 7 | - | 2 | - | 3 | + | 2 | - | 3 | - | 2 | + | 0 | + | ... |
| 63 | = | 56 | - | 7 | - | 3 | - | 1 | + | 2 | + | 3 | + | 3 | + | 0 | + | ... |
| 75 | = | 56 | + | 7 | - | 3 | + | 1 | + | 2 | - | 3 | + | 3 | + | 0 | + | ... |
| 74 | = | 56 | - | 7 | + | 5 | - | 2 | - | 6 | - | 2 | - | 2 | - | 1 | + | ... |
| 76 | = | 56 | + | 7 | + | 5 | + | 2 | - | 6 | + | 2 | - | 2 | + | 1 | + | ... |
| 66 | = | 56 | - | 7 | - | 2 | + | 3 | - | 6 | - | 2 | + | 0 | - | 1 | + | ... |
| 56 | = | 56 | + | 7 | - | 2 | - | 3 | - | 6 | + | 2 | + | 0 | + | 1 | + | ... |
| 46 | = | 56 | - | 7 | - | 3 | - | 1 | - | 6 | - | 2 | + | 2 | + | 2 | + | ... |
| 72 | = | 56 | + | 7 | - | 3 | + | 1 | - | 6 | + | 2 | + | 2 | - | 2 | + | ... |

| ... | + | $\hat{S}_k$ | + | $\widehat{AS}_{ik}$ | + | $BS_{jk}$ | + | $\widehat{ABS}_{ijk}$ | + | $\widehat{CS}_k$ | + | $\widehat{ACS}_{ik}$ | + | $BCS_{jk}$ | + | $\widehat{ABC}S_{ijk}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ... | - | 7 | - | 1 | - | 3 | + | 1 | + | 4 | - | 2 | + | 2 | + | 1 |
| ... | - | 7 | + | 1 | - | 3 | - | 1 | + | 4 | + | 2 | + | 2 | - | 1 |
| ... | - | 7 | - | 1 | + | 0 | - | 1 | + | 4 | - | 2 | + | 1 | + | 1 |
| ... | - | 7 | + | 1 | + | 0 | + | 1 | + | 4 | + | 2 | + | 1 | - | 1 |
| ... | - | 7 | - | 1 | + | 3 | + | 0 | + | 4 | - | 2 | - | 3 | - | 2 |
| ... | - | 7 | + | 1 | + | 3 | + | 0 | + | 4 | + | 2 | - | 3 | + | 2 |
| ... | - | 7 | - | 1 | - | 3 | + | 1 | - | 1 | + | 3 | + | 1 | + | 1 |
| ... | - | 7 | + | 1 | - | 3 | - | 1 | - | 1 | - | 3 | + | 1 | - | 1 |
| ... | - | 7 | - | 1 | + | 0 | - | 1 | - | 1 | + | 3 | - | 1 | + | 0 |
| ... | - | 7 | + | 1 | + | 0 | + | 1 | - | 1 | - | 3 | - | 1 | + | 0 |
| ... | - | 7 | - | 1 | + | 3 | + | 0 | - | 1 | + | 3 | + | 0 | - | 1 |
| ... | - | 7 | + | 1 | + | 3 | + | 0 | - | 1 | - | 3 | + | 0 | + | 1 |
| ... | - | 7 | - | 1 | - | 3 | + | 1 | - | 3 | - | 1 | - | 3 | - | 2 |
| ... | - | 7 | + | 1 | - | 3 | - | 1 | - | 3 | + | 1 | - | 3 | + | 2 |
| ... | - | 7 | - | 1 | + | 0 | - | 1 | - | 3 | - | 1 | + | 0 | - | 1 |
| ... | - | 7 | + | 1 | + | 0 | + | 1 | - | 3 | + | 1 | + | 0 | + | 1 |
| ... | - | 7 | - | 1 | + | 3 | + | 0 | - | 3 | - | 1 | + | 3 | + | 3 |
| ... | - | 7 | + | 1 | + | 3 | + | 0 | - | 3 | + | 1 | + | 3 | - | 3 |
| ... | - | 4 | - | 4 | + | 0 | - | 3 | + | 1 | - | 2 | + | 2 | - | 1 |
| ... | - | 4 | + | 4 | + | 0 | + | 3 | + | 1 | + | 2 | + | 2 | + | 1 |
| ... | - | 4 | - | 4 | + | 2 | - | 2 | + | 1 | - | 2 | - | 1 | + | 1 |
| ... | - | 4 | + | 4 | + | 2 | + | 2 | + | 1 | + | 2 | - | 1 | - | 1 |
| ... | - | 4 | - | 4 | - | 2 | + | 5 | + | 1 | - | 2 | - | 1 | + | 0 |
| ... | - | 4 | + | 4 | - | 2 | - | 5 | + | 1 | + | 2 | - | 1 | + | 0 |
| ... | - | 4 | - | 4 | + | 0 | - | 3 | + | 0 | + | 2 | - | 1 | + | 2 |
| ... | - | 4 | + | 4 | + | 0 | + | 3 | + | 0 | - | 2 | - | 1 | - | 2 |
| ... | - | 4 | - | 4 | + | 2 | - | 2 | + | 0 | + | 2 | + | 1 | + | 1 |
| ... | - | 4 | + | 4 | + | 2 | + | 2 | + | 0 | - | 2 | + | 1 | - | 1 |
| ... | - | 4 | - | 4 | - | 2 | + | 5 | + | 0 | + | 2 | + | 0 | - | 3 |
| ... | - | 4 | + | 4 | - | 2 | - | 5 | + | 0 | - | 2 | + | 0 | + | 3 |
| ... | - | 4 | - | 4 | + | 0 | - | 3 | - | 1 | + | 0 | - | 1 | - | 1 |
| ... | - | 4 | + | 4 | + | 0 | + | 3 | - | 1 | + | 0 | - | 1 | + | 1 |
| ... | - | 4 | - | 4 | + | 2 | - | 2 | - | 1 | + | 0 | + | 0 | - | 2 |
| ... | - | 4 | + | 4 | + | 2 | + | 2 | - | 1 | + | 0 | + | 0 | + | 2 |
| ... | - | 4 | - | 4 | - | 2 | + | 5 | - | 1 | + | 0 | + | 1 | + | 3 |
| ... | - | 4 | + | 4 | - | 2 | - | 5 | - | 1 | + | 0 | + | 1 | - | 3 |
| ... | + | 11 | + | 5 | + | 3 | + | 2 | - | 5 | + | 4 | - | 4 | + | 0 |
| ... | + | 11 | - | 5 | + | 3 | - | 2 | - | 5 | - | 4 | - | 4 | + | 0 |
| ... | + | 11 | + | 5 | - | 2 | + | 3 | - | 5 | + | 4 | + | 0 | - | 2 |
| ... | + | 11 | - | 5 | - | 2 | - | 3 | - | 5 | - | 4 | + | 0 | + | 2 |
| ... | + | 11 | + | 5 | - | 1 | - | 5 | - | 5 | + | 4 | + | 4 | + | 2 |
| ... | + | 11 | - | 5 | - | 1 | + | 5 | - | 5 | - | 4 | + | 4 | - | 2 |
| ... | + | 11 | + | 5 | + | 3 | + | 2 | + | 1 | - | 5 | + | 0 | - | 3 |
| ... | + | 11 | - | 5 | + | 3 | - | 2 | + | 1 | + | 5 | + | 0 | + | 3 |
| ... | + | 11 | + | 5 | - | 2 | + | 3 | + | 1 | - | 5 | + | 0 | - | 1 |
| ... | + | 11 | - | 5 | - | 2 | - | 3 | + | 1 | + | 5 | + | 0 | + | 1 |
| ... | + | 11 | + | 5 | - | 1 | - | 5 | + | 1 | - | 5 | + | 0 | + | 4 |
| ... | + | 11 | - | 5 | - | 1 | + | 5 | + | 1 | + | 5 | + | 0 | - | 4 |
| ... | + | 11 | + | 5 | + | 3 | + | 2 | + | 4 | + | 1 | + | 4 | + | 3 |
| ... | + | 11 | - | 5 | + | 3 | - | 2 | + | 4 | - | 1 | + | 4 | - | 3 |
| ... | + | 11 | + | 5 | - | 2 | + | 3 | + | 4 | + | 1 | + | 0 | + | 3 |
| ... | + | 11 | - | 5 | - | 2 | - | 3 | + | 4 | - | 1 | + | 0 | - | 3 |
| ... | + | 11 | + | 5 | - | 1 | - | 5 | + | 4 | + | 1 | - | 4 | - | 6 |
| ... | + | 11 | - | 5 | - | 1 | + | 5 | + | 4 | - | 1 | - | 4 | + | 6 |

**ANOVA Table:**

| Source | *df* | *SS* | *MS* | *F* | Error Term |
|---|---|---|---|---|---|
| $\mu$ | 1 | 169344.0 | 169344.0 | 101.161 | $S$ |
| $A$ | 1 | 2646.0 | 2646.0 | 7.000 | $AS$ |
| $B$ | 2 | 684.0 | 342.0 | 5.700 | $BS$ |
| $AB$ | 2 | 252.0 | 126.0 | 1.077 | $ABS$ |
| $C$ | 2 | 1008.0 | 504.0 | 4.800 | $CS$ |
| $AC$ | 2 | 252.0 | 126.0 | 1.312 | $ACS$ |
| $BC$ | 4 | 360.0 | 90.0 | 3.333 | $BCS$ |
| $ABC$ | 4 | 72.0 | 18.0 | 0.514 | $ABCS$ |
| $S$ | 2 | 3348.0 | 1674.0 | | |
| $AS$ | 2 | 756.0 | 378.0 | | |
| $BS$ | 4 | 240.0 | 60.0 | | |
| $ABS$ | 4 | 468.0 | 117.0 | | |
| $CS$ | 4 | 420.0 | 105.0 | | |
| $ACS$ | 4 | 384.0 | 96.0 | | |
| $BCS$ | 8 | 216.0 | 27.0 | | |
| $ABCS$ | 8 | 280.0 | 35.0 | | |
| Total | 54 | 180730.0 | | | |

## 7.8 Exercises

**1.** Write out the "source" and "error term" columns of an ANOVA table for a within-subjects design with factors A, B, and C. Indicate the appropriate error term that would be used to compute the $F_{observed}$ for each line of the table. If no $F_{observed}$ is computed for a given line, leave the error term blank for that line.

**2.** NASA has 6 astronauts in their SKY LAB program, and they are trying to decide which one to put in charge of the next mission. They want to choose the one who reacts most quickly to certain types of emergency, so they decide to run a comparison. Each astronaut is tested twice under zero gravity, normal gravity, and double gravity. On one of the tests the astronaut has been awake for less than 12 hours, and on the other one s/he has been awake for more than 24 hours.

(a) Identify the factors in this experiment. For each one, indicate whether it is a fixed- or random-effects factor.

(b) In words, what does the $MS$ for Gravity Condition measure?

(c) In words, what does the $MS$ for Astronaut measure?

(d) In words, what does the $MS$ for the Gravity Condition by Time Since Sleep interaction measure?

(e) In words, what does the $MS$ for the Gravity Condition by Astronaut interaction measure?

(f) In words, what does the $MS$ for the Gravity Condition by Time Since Sleep by Astronaut interaction measure?

(g) True or False:

- The Gravity Condition by Sleep interaction cannot be significant unless both Gravity Condition and Sleep are significant alone.
- The Astronaut by Sleep interaction cannot be significant unless the Astronaut effect is significant.
- The Gravity Condition by Sleep by Astronaut interaction cannot be significant unless one of the main effects or two-way interactions is significant.

**3.** Here are the data and some of the estimates from a two-factor within-subjects design, with 2 levels of Factor A and 3 levels of Factor B.

| Subject | A1B1 | A2B1 | A1B2 | A2B2 | A1B3 | A2B3 |
|---|---|---|---|---|---|---|
| 1 | 61 | 75 | 26 | 56 | 30 | 40 |
| 2 | 75 | 75 | 34 | 50 | 47 | 43 |
| 3 | 74 | 78 | 32 | 52 | 50 | 50 |
| 4 | 58 | 68 | 31 | 57 | 46 | 52 |
| 5 | 62 | 64 | 32 | 50 | 47 | 45 |

$$
\begin{aligned}
\hat{\mu} &= 52 \\
\hat{A}_1 &= -5 \quad \hat{A}_2 = 5 \\
\hat{B}_1 &= 17 \quad \hat{B}_2 = -10 \quad \hat{B}_3 = -7 \\
\widehat{AB}_{11} &= +2 \quad \widehat{AB}_{12} = -6 \quad \widehat{AB}_{13} = +4 \\
\widehat{AB}_{21} &= -2 \quad \widehat{AB}_{22} = +6 \quad \widehat{AB}_{23} = -4
\end{aligned}
$$

Compute estimates of the following terms, showing the values used in the computations:

- $\hat{S}_2$
- $\widehat{AS}_{12}$
- $\widehat{BS}_{12}$
- $\widehat{ABS}_{112}$

## 7.9 Answers to Exercises

**1.**

| Source: | Error Term |
|---|---|
| $\mu$ | $S$ |
| $A$ | $AS$ |
| $B$ | $BS$ |
| $AB$ | $ABS$ |
| $C$ | $CS$ |
| $AC$ | $ACS$ |
| $BC$ | $BCS$ |
| $ABC$ | ABCS |
| $S$ | |
| $AS$ | |
| $BS$ | |
| $ABS$ | |
| $CS$ | |
| $ACS$ | |
| $BCS$ | |
| $ABCS$ | |

**2.**

(a) The three factors are Gravity Condition, Time Since Sleep, and Astronaut, and all are fixed-effects factors. In particular, Astronaut cannot be a random-effects factor, because the levels (i.e., astronauts) were not sampled randomly, nor is there any intent to generalize the results to a larger population of astronauts.

(b) The extent to which overall performance is different in the three gravity conditions.

(c) The extent to which overall performance is different for the six different astronauts.

(d) The extent to which the effect of gravity condition on performance is different depending on the time since sleep (or vice versa).

(e) The extent to which the effect of gravity condition on performance is different for the different astronauts.

(f) The extent to which the interaction of gravity condition and time since sleep is different for the different astronauts.

(g) The True/False questions are all false.

**3.**

- $\hat{S}_2 = 2$
- $\widehat{AS}_{12} = 3$
- $\widehat{BS}_{12} = 4$
- $\widehat{ABS}_{112} = 0$

# Chapter 8

# Mixed Designs—The General Case

Now that you know how to analyze designs with between-subjects factors and designs with within-subjects factors, it is time to consider designs with both types of factors in the same experiment. Such designs are the topic of this section. Mixed designs are very common, because most dependent variables depend both on factors that vary between subjects and on factors that vary within subjects across time. The simplest mixed design has two factors—one within and one between. For example, a researcher might compare GPAs for males vs. females in math vs. language courses. Gender would be a between-subjects factor, of course. If both math and language GPAs were obtained for each subject, the Course factor would be within-subjects. Obviously, a more complicated mixed design could be developed from this example by adding factors. A second between-subjects factor might be Class level (freshman, sophomore, junior, senior), and a third might be university (Harvard, Berkeley, Slippery Rock). An additional within-subjects factor could be Time of Day (assuming that each student had at least one course of each type in the morning and another in the afternoon). And so on.

The same steps are needed to analyze mixed designs as to analyze between- and within-subjects designs:

- Write down the model.
- Write down the estimation equations.
- Obtain $SS$'s, either by estimating individual values and forming the decomposition matrix or by using the shortcut method.
- Determine the values of degrees of freedom.
- Select the appropriate error term(s) and compute $F_{observed}$'s.

Thus, in this chapter we will see how to merge the methods of the previous analyses for designs with all factors between- or within-subjects. In fact, the rules to be developed in this chapter are generalizations of both sets of rules developed previously. This means that you can use the rules stated here to analyze between-subjects designs or within-subjects designs, in addition to the mixed designs.

To illustrate the rules we develop, we will discuss two examples throughout the chapter. In both these examples, a physician is interested in the dependent variable of heart rate. The simpler example has only two factors: Age (between-subjects) and presence vs. absence of a new drug being tested (within-subjects). The data for this example are shown in Table 8.1. The more complex example augments this design with the between-subjects factor of Gender and the within-subjects factor of Stress Level (i.e., sleeping vs. walking on a treadmill). The data for this example appear in Table 8.2, and the corresponding cell and marginal means are shown in Table 8.3.

## 8.1 Linear Model

Table 8.4 shows the general rules for constructing a linear model. These rules are general enough to apply to the mixed designs considered in this chapter; they also work for entirely between- or within-subjects designs considered in earlier chapters.

Following the rules in this table, the model for the Age x Drug design is:

$$Y_{ijk} = \mu + A_i + B_j + AB_{ij} + S(A)_{ik} + BS(A)_{ijk}$$

| | Old Subjects | | | Young Subjects | |
|---|---|---|---|---|---|
| Subject | No Drug | Drug | Subject | No Drug | Drug |
| 1 | 73 | 59 | 1 | 61 | 63 |
| 2 | 81 | 67 | 2 | 60 | 50 |
| 3 | 86 | 60 | 3 | 59 | 61 |

Cell Means:

| Age: | No Drug | Drug |
|---|---|---|
| Old | 80 | 62 |
| Young | 60 | 58 |

Table 8.1: Example Data Illustrating Dependence of Heart Rate on Age (Factor A, between-Ss) and Drug (Factor B, within-Ss)

Old Males

| | No Drug | | Drug | |
|---|---|---|---|---|
| Subject | Treadmill | Sleep | Treadmill | Sleep |
| 1 | 131 | 73 | 41 | 47 |
| 2 | 121 | 83 | 47 | 41 |
| Average | 126 | 78 | 44 | 44 |

Old Females

| | No Drug | | Drug | |
|---|---|---|---|---|
| Subject | Treadmill | Sleep | Treadmill | Sleep |
| 1 | 101 | 71 | 45 | 59 |
| 2 | 75 | 65 | 23 | 49 |
| Average | 88 | 68 | 34 | 54 |

Young Males

| | No Drug | | Drug | |
|---|---|---|---|---|
| Subject | Treadmill | Sleep | Treadmill | Sleep |
| 1 | 100 | 64 | 52 | 48 |
| 2 | 112 | 84 | 36 | 56 |
| Average | 106 | 74 | 44 | 52 |

Young Females

| | No Drug | | Drug | |
|---|---|---|---|---|
| Subject | Treadmill | Sleep | Treadmill | Sleep |
| 1 | 71 | 61 | 31 | 41 |
| 2 | 89 | 59 | 45 | 59 |
| Average | 80 | 60 | 38 | 50 |

Table 8.2: Example Data Illustrating Dependence of Heart Rate on Age (Factor A, between-Ss), Drug (Factor B, within-Ss), Gender (Factor C, between-Ss), and Stress (Factor D, within-Ss)

and the model for the Age x Drug x Gender x Stress design is:

$$\begin{aligned} Y_{ijklm} &= \mu + A_i + B_j + AB_{ij} + C_k + AC_{ik} + BC_{jk} + ABC_{ijk} \\ &\quad + D_l + AD_{il} + BD_{jl} + ABD_{ijl} + CD_{kl} + ACD_{ikl} + BCD_{jkl} + ABCD_{ijkl} \\ &\quad + S(AC)_{ikm} + BS(AC)_{ijkm} + DS(AC)_{iklm} + BDS(AC)_{ijklm} \end{aligned}$$

These models deserve a little discussion. There is nothing surprising in the presence of $\mu$ or the main effects and interactions involving the regular experimental factors (i.e., A, B, C, and D). Such terms are at least somewhat familiar because they had the same form in both between- and within-subjects designs, and they retain the same meaning in mixed designs.

The novel feature of these models is the pattern of subjects terms. In between-subjects designs there was a single subjects term, with all the factors listed after it in parentheses. In within-subjects designs there was an S term with no parentheses, and there were also interactions of S with all of the other factors. Here we follow both rules at once: The between-subjects factors are listed in parentheses after S, and the S term appears in interactions with all of the within-subjects factors.

There is nothing mysterious about these rules involving the subject term—on the contrary, their conceptual basis is quite straightforward. In all cases, *the terms in the model simply reflect the types of information that are available from the experimental design.*

| | No Drug | | Drug | |
|---|---|---|---|---|
| | Tread | Sleep | Tread | Sleep |
| Old | 107 | 73 | 39 | 49 |
| Young | 93 | 67 | 41 | 51 |
| Average | 100 | 70 | 40 | 50 |

| | Male | | Female | |
|---|---|---|---|---|
| | Tread | Sleep | Tread | Sleep |
| No Drug | 116 | 76 | 84 | 64 |
| Drug | 44 | 48 | 36 | 52 |
| Average | 80 | 62 | 60 | 58 |

| | Old | | Young | |
|---|---|---|---|---|
| | Male | Female | Male | Female |
| Tread | 85 | 61 | 75 | 59 |
| Sleep | 61 | 61 | 63 | 55 |
| Average | 73 | 61 | 69 | 57 |

| | Old | | Young | |
|---|---|---|---|---|
| | No Drug | Drug | Mo Drug | Drug |
| Male | 102 | 44 | 90 | 48 |
| Female | 78 | 44 | 70 | 44 |
| Average | 90 | 44 | 80 | 46 |

| | Old | Young | Average |
|---|---|---|---|
| Tread | 73 | 67 | 70 |
| Sleep | 61 | 59 | 60 |
| Average | 67 | 63 | 65 |

| | Male | Female | Average |
|---|---|---|---|
| No Drug | 96 | 74 | 85 |
| Drug | 46 | 44 | 45 |
| Average | 71 | 59 | 65 |

Table 8.3: Means and Marginal Means for Data of Table 8.2.

- $Y$ has one subscript for each experimental factor, between or within, and one for subjects.
- There is a $\mu$ term.
- There is a main effect term $A$, $B$, $C$, ... for each experimental factor, between or within.
- There are all possible two-way, three-way, four-way, ... interaction terms that can be constructed from subsets of the experimental factors.
- There is a subjects term, $S$, which is written with a list of the between-subjects factors inside parentheses.
- There are terms for the interactions of subjects with every within-subjects factor and every interaction involving only within-subjects factors.

Table 8.4: General Rules for Constructing ANOVA Models

Consider the simpler design first.

- There is a $S(A)_{ik}$ term, because the design allows us to estimate the specific effect on each score due to the particular randomly sampled subject from whom that score was obtained. We denote this effect by $S(A)_{ik}$ so that the notation will reflect the fact that Young Subject 1 is a different individual than Old Subject 1 (and may therefore have a different effect).
- There is a $BS(A)_{ijk}$ term, because the design allows us to estimate the relative responsiveness of each individual to the two drug conditions. The design allows this because each subject is tested both with and without the drug. If Joe shows little effect of the drug whereas Sam shows a big effect, for example, then the values of $BS(A)_{ijk}$ will be near-zero for Joe and far-from-zero for Sam. In short, this is the model term that shows us how much the drug effect varies from person to person—a type of information that is always present for any within-subjects factor.
- The model *does not contain* an $AS(A)$ term, which could potentially measure the responsiveness of each individual subject to the Age factor (i.e., the interaction of Age and Subject). The reason is quite clear: Since each subject was *only measured in one age group*, it is impossible to compare the responses of different subjects to the change in age.

In a nutshell: Models for mixed designs include terms for the interaction of subjects with within-subjects factors, because it is possible to see how within-subjects effects differ from subject to subject. Models for mixed designs do not include terms for the interaction of subjects with between-subjects

factors, because it is impossible to see between-subjects effects on an individual subject (and therefore it is impossible to see how they differ from subject to subject).

The same principles apply to the more complex design. All of the standard experimental factors (A, B, C, and D) and their interactions are included in the model, because we can estimate the effects of these factors, as in any factorial experiment. Likewise, there is a term to indicate the effect of the particular subject in each group; this time, the four groups of subjects are differentiated by two between-subjects factors. In addition, there are two two-way interaction terms involving subjects: $BS(AC)_{ijkm}$ and $DS(AC)_{iklm}$. The first indicates any special effect of Drug on a subject, and it is included because having Drug as a within-subjects factor enables us to measure the effect of Drug separately for each subject. The second indicates any special effect of Stress on a subject, and it is included because having Stress as a within-subjects factor enables us to measure the effect of Stress separately for each subject.

Finally, there is a three-way interaction term involving the two within-subjects factors, together with subjects: $BDS(AC)_{ijklm}$. The justification for this term was already given in the chapter on within-subjects designs: The design allows us to estimate the BD interaction separately for each subject, so the term will indicate the size of the BD interaction for each subject. As before, we can use this term to see how the BD interaction varies from subject to subject.

It is also worth discussing the terms that are *missing from* the model for the complex example. For example, the Age by subjects interaction term is missing, for the same reason as in the simpler example: we cannot estimate the effect of age on an individual subject in this design (to do so would require testing the same subject in both age groups). The same comment applies to the Gender by subjects interaction and the Age by Gender interaction.

More interesting is the absence of the Age by Drug by subjects interaction. Age is between-subjects, but Drug is within—so why is the three-way interaction missing? As always, the rule is that the interaction is missing because the design does not allow us to estimate it. In this case, to measure the Age by Drug by Subjects interaction we would have to be able to measure the Age by Drug interaction separately for each subject. We can so measure the Drug effect—Drug being a within factor—but not the change in Drug effect with age (i.e., the Drug x Age interaction). Because each subject is observed in only one age group, it is impossible to see how the effect of drug changes with age *for that subject*. Thus, the design does not allow measurement of the Drug x Age x Subjects interaction, and the term is not included in the model.

A little reflection shows that the points made about the factors of these examples apply in general to *all* within- and between-subject factors. Thus, these ideas really justify the general rules for the linear model given in Table 8.4.

## 8.2 Estimation Equations and $SS$'s

For mixed designs, estimation proceeds just as it did for between- and within-subject designs.

- For any term, the estimation equation has the form:

  Estimate of Term = Data Average - Estimates of Subterms

- The data average has the same subscripts as those on the term to be estimated, and it is averaged over all the subscripts absent from that term.

- The *subterms* of a term are those simpler terms that are logically contained within the term. More formally, term "Sub" is a subterm of term "Super" if every subscript on term "Sub" is also a subscript on term "Super." $\mu$ is a subterm of every other term.

Tables 8.5 and 8.6 show the estimation equations for the simple and complex examples involving the dependent variable of heart rate. It is easy to see that there are no new principles at work here, although the complicated nature of the model terms may at first cause confusion. It is always possible to get the right answer, though, by working from left to right through the model and checking subscripts carefully.

Once the estimation equations have been written down, it is a simple though often very tedious matter to compute the $SS$ for each term in the model. This can be done either by estimating each individual term and writing down the decomposition matrix (see examples in Tables 8.7, 8.8, and 8.9), or by using the shortcut method.

$$
\begin{aligned}
\hat{\mu} &= Y_{...} \\
\hat{A}_i &= Y_{i..} - \hat{\mu} \\
\hat{B}_j &= Y_{.j.} - \hat{\mu} \\
\widehat{AB}_{ij} &= Y_{ij.} - \hat{\mu} - \hat{A}_i - \hat{B}_j \\
\widehat{S(A)}_{ik} &= Y_{i.k} - \hat{\mu} - \hat{A}_i \\
\widehat{BS(A)}_{ijk} &= Y_{ijk} - \hat{\mu} - \hat{A}_i - \hat{B}_j - \widehat{AB}_{ij} - \widehat{S(A)}_{ik}
\end{aligned}
$$

Table 8.5: Estimation Equations for the Simple Design of Table 8.1

## 8.3 Degrees of Freedom

The rules for computing degrees of freedom are also quite familiar.

- $df_\mu = 1$.
- The main effect of a factor with $K$ levels has $K - 1$ *df*.
- The number of *df*'s for the main effect of subjects is again the number of subjects minus the number of groups. The number of groups is the product of the numbers of levels of all the between-Ss factors.
- Interaction *df*'s are found by multiplying together the *df*'s of the associated main effects.
- $df_{total}$ = the number of data values.

Note that these rules also give the correct *df* values for between- and within-subject designs.

## 8.4 Error Terms

A surprisingly simple rule describes the proper selection of error terms for computing $F_{observed}$'s in mixed designs as well as between- and within-subjects designs:

> *To compute the $F_{observed}$ with any given source in the numerator, use in the denominator the source which is the interaction of subjects by all of the within-subjects factors contained in the source in the numerator.*

Before discussing mixed designs, we hasten to point out that this rule also applies to between- and within-subjects designs; that is, it gives the same error terms for these designs as we described in previous chapters. In between-subjects designs, for example, "the interaction of subjects by all of the within-subjects factors ..." is simply the subjects term used to compute all $F_{observed}$'s. This is because there are no within-subjects factors in any source, and the interaction of "subjects by nothing" is just "subjects." Conversely, in within-subjects designs, "the interaction of subjects by all of the within-subjects factors ..." is simply the numerator by subjects interaction term, because all of the factors in the numerator are necessarily within-subjects. As we have already seen in Chapter 7, this is the appropriate error term.

Now to mixed designs. Tables 8.10 and 8.11 show the ANOVA tables for the two examples being discussed in this chapter, and it worth examining the pattern of the error terms to see how they are generated by the rule. You can see that, for any given source, the error term is simply the interaction of subjects with the within-subjects factors contained in the source.

Of course we must also discuss the rationales behind the error terms, in order to justify the rule. For main effects, the rationales are pretty familiar. According to the rule, we use subjects (i.e., the subjects by nothing interaction) as the error term for a between-subjects main effect. As noted in the chapters on between-subjects designs, this is because the subjects term is a measure of how much difference we can expect between subjects in the same condition(s), and this difference tells us how large the between-group differences might be due to chance. Conversely, we use the subject by factor interaction as the error term for a within-subjects main effect. As noted in the chapter on within-subjects designs, this is because the subject by factor interaction is a measure of how inconsistent the

$$
\begin{aligned}
\hat{\mu} &= Y_{.....} \\
\hat{A}_i &= Y_{i....} - \hat{\mu} \\
\hat{B}_j &= Y_{.j...} - \hat{\mu} \\
\widehat{AB}_{ij} &= Y_{ij...} - \hat{\mu} - \hat{A}_i - \hat{B}_j \\
\hat{C}_k &= Y_{..k..} - \hat{\mu} \\
\widehat{AC}_{ik} &= Y_{i.k..} - \hat{\mu} - \hat{A}_i - \hat{C}_k \\
\widehat{BC}_{jk} &= Y_{.jk..} - \hat{\mu} - \hat{B}_j - \hat{C}_k \\
\widehat{ABC}_{ijk} &= Y_{ijk..} - \hat{\mu} - \hat{A}_i - \hat{B}_j - \widehat{AB}_{ij} - \hat{C}_k - \widehat{AC}_{ik} - \widehat{BC}_{jk} \\
\hat{D}_l &= Y_{...l.} - \hat{\mu} \\
\widehat{AD}_{il} &= Y_{i..l.} - \hat{\mu} - \hat{A}_i - \hat{D}_l \\
\widehat{BD}_{jl} &= Y_{.j.l.} - \hat{\mu} - \hat{B}_j - \hat{D}_l \\
\widehat{ABD}_{ijl} &= Y_{ij.l.} - \hat{\mu} - \hat{A}_i - \hat{B}_j - \widehat{AB}_{ij} - \hat{D}_l - \widehat{AD}_{il} - \widehat{BD}_{jl} \\
\widehat{CD}_{kl} &= Y_{..kl.} - \hat{\mu} - \hat{C}_k - \hat{D}_l \\
\widehat{ACD}_{ikl} &= Y_{i.kl.} - \hat{\mu} - \hat{A}_i - \hat{C}_k - \widehat{AC}_{ik} - \hat{D}_l - \widehat{AD}_{il} - \widehat{CD}_{kl} \\
\widehat{BCD}_{jkl} &= Y_{.jkl.} - \hat{\mu} - \hat{B}_j - \hat{C}_k - \widehat{BC}_{jk} - \hat{D}_l - \widehat{BD}_{jl} - \widehat{CD}_{kl} \\
\widehat{ABCD}_{ijkl} &= Y_{ijkl.} - \hat{\mu} - \hat{A}_i - \hat{B}_j - \widehat{AB}_{ij} - \hat{C}_k - \widehat{AC}_{ik} - \widehat{BC}_{jk} - \widehat{ABC}_{ijk} \\
&\quad -\hat{D}_l - \widehat{AD}_{il} - \widehat{BD}_{jl} - \widehat{ABD}_{ijl} - \widehat{CD}_{kl} - \widehat{ACD}_{ikl} - \widehat{BCD}_{jkl} \\
\widehat{S(AC)}_{ikm} &= Y_{i.k.m} - \hat{\mu} - \hat{A}_i - \hat{C}_k - \widehat{AC}_{ik} \\
\widehat{BS(AC)}_{ijkm} &= Y_{ijk.m} - \hat{\mu} - \hat{A}_i - \hat{B}_j - \widehat{AB}_{ij} - \hat{C}_k - \widehat{AC}_{ik} - \widehat{BC}_{jk} - \widehat{ABC}_{ijk} - \widehat{S(AC)}_{ikm} \\
\widehat{DS(AC)}_{iklm} &= Y_{i.klm} - \hat{\mu} - \hat{A}_i - \hat{C}_k - \widehat{AC}_{ik} - \hat{D}_l - \widehat{AD}_{il} - \widehat{CD}_{kl} - \widehat{ACD}_{ikl} - \widehat{S(AC)}_{ikm} \\
\widehat{BDS(AC)}_{ijklm} &= Y_{ijklm} - \hat{\mu} - \hat{A}_i - \hat{B}_j - \widehat{AB}_{ij} - \hat{C}_k - \widehat{AC}_{ik} - \widehat{BC}_{jk} - \widehat{ABC}_{ijk} - \hat{D}_l \\
&\quad -\widehat{AD}_{il} - \widehat{BD}_{jl} - \widehat{ABD}_{ijl} - \widehat{CD}_{kl} - \widehat{ACD}_{ikl} - \widehat{BCD}_{jkl} - \widehat{ABCD}_{ijkl} \\
&\quad -\widehat{S(AC)}_{ikm} - \widehat{BS(AC)}_{ijkm} - \widehat{DS(AC)}_{iklm}
\end{aligned}
$$

Table 8.6: Estimation Equations for the Complex Design of Table 8.2

effect is across subjects, and because our confidence in the reality of an effect is inversely related to its inconsistency across subjects (i.e., the more consistent it is across subjects, the more confident we are that it's real). The rationales are also familiar for interactions that are entirely between-subjects (i.e., involve only between-subjects factors) or entirely within-subjects (i.e., involve only within-subjects factors). In the former case we are interested in the size of between-group differences that might arise by chance, and "subjects" is the relevant error term because it measures the inconsistency between subjects within a given condition. In the latter case we are interested in the (in)consistency of the interaction in question across subjects, and the "subjects by interaction" error term measures this for us.

To summarize: For entirely between main effects or interactions, we use the same error term that we would use if none of the within-subjects factors had been included in the design. For entirely within main effects or interactions, we use the same error term that we would use if none of the between-subjects factors had been included in the design. This makes some intuitive sense, since including extra factors in the design and analysis doesn't really change the nature of the information we have about any given factor.

The new and somewhat tricky case involves interactions including both between- and within-subjects factors. In the simple example, for instance, we need an error term for the Age (between) by Drug (within) interaction. In the complex example, we need error terms for this same interaction plus a variety of other "mixed" ones (e.g., Age x Drug x Gender, Age x Drug x Gender x Stress). The rule for these cases is quite clear of course—use the "subjects by within" interactions for the error terms. But what is the rationale?

The rationale is subtle: The design allows us to measure how much the within-subjects effect varies randomly from person to person. Therefore, in order to see whether the within-subjects effect really

| $Y_{ijk}$ | = | $\mu$ | + | $A_i$ | + | $B_j$ | + | $AB_{ij}$ | + | $S(A)_{ik}$ | + | $BS(A)_{ijk}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 73 | = | 65 | + | 6 | + | 5 | + | 4 | - | 5 | - | 2 |
| 59 | = | 65 | + | 6 | - | 5 | - | 4 | - | 5 | + | 2 |
| 81 | = | 65 | + | 6 | + | 5 | + | 4 | + | 3 | - | 2 |
| 67 | = | 65 | + | 6 | - | 5 | - | 4 | + | 3 | + | 2 |
| 86 | = | 65 | + | 6 | + | 5 | + | 4 | + | 2 | + | 4 |
| 60 | = | 65 | + | 6 | - | 5 | - | 4 | + | 2 | - | 4 |
| 61 | = | 65 | - | 6 | + | 5 | - | 4 | + | 3 | - | 2 |
| 63 | = | 65 | - | 6 | - | 5 | + | 4 | + | 3 | + | 2 |
| 60 | = | 65 | - | 6 | + | 5 | - | 4 | - | 4 | + | 4 |
| 50 | = | 65 | - | 6 | - | 5 | + | 4 | - | 4 | - | 4 |
| 59 | = | 65 | - | 6 | + | 5 | - | 4 | + | 1 | - | 2 |
| 61 | = | 65 | - | 6 | - | 5 | + | 4 | + | 1 | + | 2 |

Table 8.7: Decomposition Matrix for the Data of Table 8.1

varies between groups, we need to see whether the group-to-group variation in the effect is larger than we can explain in terms of random person-to-person variation in the effect.

In the simple design, for example, we note that the drug effect is somewhat different for the Old people (18 beats per minute) than for the Young people (2 beats per minute). This seems like a large effect, but to be sure we must compare it to the person-to-person random variation in the Drug effect—that is, to the Drug by Subjects(Age) interaction term. If there were enough random variation in the size of the drug effect among both the young and the old subjects, perhaps we would decide that the difference between 18 and 2 could just be random error.

A more elaborate version of the same basic rationale explains the selection of error terms in the complex data set as well. Consider the most difficult examples: the three and four way interactions including Drug by Stress. Now the Drug by Stress interaction is measured within-subjects, so the model allows us to estimate the extent to which this interaction varies from subject to subject. Furthermore, this estimate provides an index of the amount of random variation in the Drug by Stress interaction. Naturally, then, when we want to see whether the Drug by Stress interaction is different for Males vs. Females (i.e., test Gender x Drug x Stress), we compare the size of the three-way interaction against this index of random variation in the Drug x Stress interaction. The same idea justifies testing both Age x Drug x Stress and Age x Gender x Drug x against the Drug x Stress x Subjects interaction.[1]

## 8.5 Order Effects

We close this chapter with a discussion of an example involving the use of a mixed design to allow for *order effects* on a within-subjects factor.[2] As noted in Section 7.1, an apparent disadvantage of within-subjects testing is that the process of testing a subject in one condition may produce serious changes that affect the results in conditions tested later. Consider, for example, an experiment in which we seek to determine the effect of alcohol on motor coordination. A test of motor coordination is given to each subject twice—once under the influence of alcohol and once not under the influence. Clearly, we would have to worry about the possibility of order effects, because subjects might show a general improvement at the second test of coordination, just due to learning.

One obvious experimental precaution would be to test different groups of subjects in different orders. We would want to test half of the subjects first in the No-Alcohol condition and then in the Alcohol condition, with the other half of the subjects tested in the reverse order. This experimental precaution is known as *counterbalancing* order, and it is needed to prevent confounding order with Alcohol vs. No-Alcohol, as would happen if all subjects were tested in the same order.

[1] Our discussion here focuses on the Drug x Stress interaction and how it changes across levels of other factors (e.g., Gender). Some readers, remembering the symmetric nature of interactions, may wonder whether the discussion could just as well have focused on (say) the Gender by Drug interaction and how it changed across levels of Stress. The answer is "no", and this is one sense in which interactions fail to be completely symmetric. Whereas interactions are completely symmetric with respect to their *interpretations*, they are asymmetric with respect to their error terms, because the error term always involves the random variability in the within-subjects component.

[2] Such effects are sometimes called *carry-over* effects.

| $Y_{ijklm}$ | = | $\mu$ | + | $A_i$ | + | $B_j$ | + | $AB_{ij}$ | + | $C_k$ | + | $AC_{ik}$ | + | $BC_{jk}$ | + | $ABC_{ijk}$ | + | $D_l$ | + | $AD_{il}$ | + | $BD_{jl}$ | + | $ABD_{ijl}$ | + ... |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 131 | = | 65 | + | 5 | + | 20 | + | 10 | + | 6 | + | 4 | + | 5 | + | 1 | + | 2 | + | 1 | + | 3 | + | 1 | +... |
| 73 | = | 65 | - | 5 | + | 20 | - | 10 | + | 6 | - | 4 | + | 5 | - | 1 | + | 2 | - | 1 | + | 3 | - | 1 | +... |
| 41 | = | 65 | + | 5 | - | 20 | - | 10 | + | 6 | + | 4 | - | 5 | - | 1 | + | 2 | + | 1 | - | 3 | - | 1 | +... |
| 47 | = | 65 | - | 5 | - | 20 | + | 10 | + | 6 | - | 4 | - | 5 | + | 1 | + | 2 | - | 1 | - | 3 | + | 1 | +... |
| 121 | = | 65 | + | 5 | + | 20 | + | 10 | + | 6 | + | 4 | + | 5 | + | 1 | + | 2 | + | 1 | + | 3 | + | 1 | +... |
| 83 | = | 65 | - | 5 | + | 20 | - | 10 | + | 6 | - | 4 | + | 5 | - | 1 | + | 2 | - | 1 | + | 3 | - | 1 | +... |
| 47 | = | 65 | + | 5 | - | 20 | - | 10 | + | 6 | + | 4 | - | 5 | - | 1 | + | 2 | + | 1 | - | 3 | - | 1 | +... |
| 41 | = | 65 | - | 5 | - | 20 | + | 10 | + | 6 | - | 4 | - | 5 | + | 1 | + | 2 | - | 1 | - | 3 | + | 1 | +... |
| 101 | = | 65 | + | 5 | + | 20 | + | 10 | - | 6 | - | 4 | - | 5 | - | 1 | + | 2 | + | 1 | + | 3 | + | 1 | +... |
| 71 | = | 65 | - | 5 | + | 20 | - | 10 | - | 6 | + | 4 | - | 5 | + | 1 | + | 2 | - | 1 | + | 3 | - | 1 | +... |
| 45 | = | 65 | + | 5 | - | 20 | - | 10 | - | 6 | - | 4 | + | 5 | + | 1 | + | 2 | + | 1 | - | 3 | - | 1 | +... |
| 59 | = | 65 | - | 5 | - | 20 | + | 10 | - | 6 | + | 4 | + | 5 | - | 1 | + | 2 | - | 1 | - | 3 | + | 1 | +... |
| 75 | = | 65 | + | 5 | + | 20 | + | 10 | - | 6 | - | 4 | - | 5 | - | 1 | + | 2 | + | 1 | + | 3 | + | 1 | +... |
| 65 | = | 65 | - | 5 | + | 20 | - | 10 | - | 6 | + | 4 | - | 5 | + | 1 | + | 2 | - | 1 | + | 3 | - | 1 | +... |
| 23 | = | 65 | + | 5 | - | 20 | - | 10 | - | 6 | - | 4 | + | 5 | + | 1 | + | 2 | + | 1 | - | 3 | - | 1 | +... |
| 49 | = | 65 | - | 5 | - | 20 | + | 10 | - | 6 | + | 4 | + | 5 | - | 1 | + | 2 | - | 1 | - | 3 | + | 1 | +... |
| 100 | = | 65 | + | 5 | + | 20 | + | 10 | + | 6 | + | 4 | + | 5 | + | 1 | - | 2 | - | 1 | - | 3 | - | 1 | +... |
| 64 | = | 65 | - | 5 | + | 20 | - | 10 | + | 6 | - | 4 | + | 5 | - | 1 | - | 2 | + | 1 | - | 3 | + | 1 | +... |
| 52 | = | 65 | + | 5 | - | 20 | - | 10 | + | 6 | + | 4 | - | 5 | - | 1 | - | 2 | - | 1 | + | 3 | + | 1 | +... |
| 48 | = | 65 | - | 5 | - | 20 | + | 10 | + | 6 | - | 4 | - | 5 | + | 1 | - | 2 | + | 1 | + | 3 | - | 1 | +... |
| 112 | = | 65 | + | 5 | + | 20 | + | 10 | + | 6 | + | 4 | + | 5 | + | 1 | - | 2 | - | 1 | - | 3 | - | 1 | +... |
| 84 | = | 65 | - | 5 | + | 20 | - | 10 | + | 6 | - | 4 | + | 5 | - | 1 | - | 2 | + | 1 | - | 3 | + | 1 | +... |
| 36 | = | 65 | + | 5 | - | 20 | - | 10 | + | 6 | + | 4 | - | 5 | - | 1 | - | 2 | - | 1 | + | 3 | + | 1 | +... |
| 56 | = | 65 | - | 5 | - | 20 | + | 10 | + | 6 | - | 4 | - | 5 | + | 1 | - | 2 | + | 1 | + | 3 | - | 1 | +... |
| 71 | = | 65 | + | 5 | + | 20 | + | 10 | - | 6 | - | 4 | - | 5 | - | 1 | - | 2 | - | 1 | - | 3 | - | 1 | +... |
| 61 | = | 65 | - | 5 | + | 20 | - | 10 | - | 6 | + | 4 | - | 5 | + | 1 | - | 2 | + | 1 | - | 3 | + | 1 | +... |
| 31 | = | 65 | + | 5 | - | 20 | - | 10 | - | 6 | - | 4 | + | 5 | + | 1 | - | 2 | - | 1 | + | 3 | + | 1 | +... |
| 41 | = | 65 | - | 5 | - | 20 | + | 10 | - | 6 | + | 4 | + | 5 | - | 1 | - | 2 | + | 1 | + | 3 | - | 1 | +... |
| 89 | = | 65 | + | 5 | + | 20 | + | 10 | - | 6 | - | 4 | - | 5 | - | 1 | - | 2 | - | 1 | - | 3 | - | 1 | +... |
| 59 | = | 65 | - | 5 | + | 20 | - | 10 | - | 6 | + | 4 | - | 5 | + | 1 | - | 2 | + | 1 | - | 3 | + | 1 | +... |
| 45 | = | 65 | + | 5 | - | 20 | - | 10 | - | 6 | - | 4 | + | 5 | + | 1 | - | 2 | - | 1 | + | 3 | + | 1 | +... |
| 59 | = | 65 | - | 5 | - | 20 | + | 10 | - | 6 | + | 4 | + | 5 | - | 1 | - | 2 | + | 1 | + | 3 | - | 1 | +... |

Table 8.8: Part 1 of Decomposition Matrix for the Data of Table 8.2

It is not enough just to counterbalance order in the experimental design—we must also take order into account in the statistical analysis. Mixed designs are very useful in examining order effects. Include in the design a between-subjects factor defined by the different orders in which the different groups of subjects are tested across the levels of the within-subjects factor(s). The model terms for this factor will allow for an effect of order on overall performance level and an effect of order on the size of the experimental effects (e.g., Alcohol). In most cases, these terms will help to reduce the error terms, because any large effects of or interactions with Order will go into these error terms if Order is not included as a factor in the design.

It may at first seem undesirable to include Order in the analysis, because in most cases order is probably not a factor that we as experimenters would have any interest in. Nonetheless, it would be a mistake to omit Order from the analysis, because doing so could well inflate the error term(s) for the effects we were interested in. Besides, order effects and their interactions might turn out to be more interesting than we realized ahead of time. If the experiment provides information on order effects, why not look to see what the results are?

Data for this example are presented in Table 8.12, with 2 groups containing 4 subjects per group. The model is a the standard model for a mixed design with one between- and one within-subjects factor:

$$Y_{ijk} = \mu + A_i + B_j + AB_{ij} + S(B)_{jk} + AS(B)_{ijk}$$

where $A$ is the term for the Alcohol factor and $B$ is the term for the Order factor.

The ANOVA table for this example is shown in Table 8.13. The effect of Alcohol is clearly significant, indicating that there is an effect of alcohol on motor coordination. Interestingly, there is also a big interaction of Alcohol and Order. Looking at the means, we see that the Alcohol effect is much larger for those tested with Alcohol first and No-Alcohol second than for those tested in the reverse order. This probably just means that there is a substantial practice effect, upon which the

| $\ldots + CD_{kl}$ | | $+ ACD_{ikl}$ | | $+ BCD_{jkl}$ | | $+ ABCD_{ijkl}$ | | $+ S(AC)_{ikm}$ | | $+ BS(AC)_{ijkm}$ | | $+ DS(AC)_{iklm}$ | | $+ BDS(AC)_{ijklm}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\ldots +$ 0 | + | 2 | + | 1 | + | 0 | + | 0 | + | 1 | + | 0 | + | 4 |
| $\ldots +$ 0 | - | 2 | + | 1 | + | 0 | + | 0 | - | 1 | + | 0 | - | 4 |
| $\ldots +$ 0 | + | 2 | - | 1 | + | 0 | + | 0 | + | 1 | + | 0 | - | 4 |
| $\ldots +$ 0 | - | 2 | - | 1 | + | 0 | + | 0 | - | 1 | + | 0 | + | 4 |
| $\ldots +$ 0 | + | 2 | + | 1 | + | 0 | + | 0 | - | 1 | + | 0 | - | 4 |
| $\ldots +$ 0 | - | 2 | + | 1 | + | 0 | + | 0 | + | 1 | + | 0 | + | 4 |
| $\ldots +$ 0 | + | 2 | - | 1 | + | 0 | + | 0 | - | 1 | + | 0 | + | 4 |
| $\ldots +$ 0 | - | 2 | - | 1 | + | 0 | + | 0 | + | 1 | + | 0 | - | 4 |
| $\ldots +$ 0 | - | 2 | - | 1 | + | 0 | + | 8 | + | 4 | + | 0 | + | 1 |
| $\ldots +$ 0 | + | 2 | - | 1 | + | 0 | + | 8 | - | 4 | + | 0 | - | 1 |
| $\ldots +$ 0 | - | 2 | + | 1 | + | 0 | + | 8 | + | 4 | + | 0 | - | 1 |
| $\ldots +$ 0 | + | 2 | + | 1 | + | 0 | + | 8 | - | 4 | + | 0 | + | 1 |
| $\ldots +$ 0 | - | 2 | - | 1 | + | 0 | - | 8 | - | 4 | + | 0 | - | 1 |
| $\ldots +$ 0 | + | 2 | - | 1 | + | 0 | - | 8 | + | 4 | + | 0 | + | 1 |
| $\ldots +$ 0 | - | 2 | + | 1 | + | 0 | - | 8 | - | 4 | + | 0 | + | 1 |
| $\ldots +$ 0 | + | 2 | + | 1 | + | 0 | - | 8 | + | 4 | + | 0 | - | 1 |
| $\ldots +$ 0 | - | 2 | - | 1 | + | 0 | - | 3 | + | 4 | - | 5 | - | 2 |
| $\ldots +$ 0 | + | 2 | - | 1 | + | 0 | - | 3 | - | 4 | - | 5 | + | 2 |
| $\ldots +$ 0 | - | 2 | + | 1 | + | 0 | - | 3 | + | 4 | + | 5 | + | 2 |
| $\ldots +$ 0 | + | 2 | + | 1 | + | 0 | - | 3 | - | 4 | + | 5 | - | 2 |
| $\ldots +$ 0 | - | 2 | - | 1 | + | 0 | + | 3 | - | 4 | + | 5 | + | 2 |
| $\ldots +$ 0 | + | 2 | - | 1 | + | 0 | + | 3 | + | 4 | + | 5 | - | 2 |
| $\ldots +$ 0 | - | 2 | + | 1 | + | 0 | + | 3 | - | 4 | - | 5 | - | 2 |
| $\ldots +$ 0 | + | 2 | + | 1 | + | 0 | + | 3 | + | 4 | - | 5 | + | 2 |
| $\ldots +$ 0 | + | 2 | + | 1 | + | 0 | - | 6 | - | 2 | + | 2 | - | 3 |
| $\ldots +$ 0 | - | 2 | + | 1 | + | 0 | - | 6 | + | 2 | + | 2 | + | 3 |
| $\ldots +$ 0 | + | 2 | - | 1 | + | 0 | - | 6 | - | 2 | - | 2 | + | 3 |
| $\ldots +$ 0 | - | 2 | - | 1 | + | 0 | - | 6 | + | 2 | - | 2 | - | 3 |
| $\ldots +$ 0 | + | 2 | + | 1 | + | 0 | + | 6 | + | 2 | - | 2 | + | 3 |
| $\ldots +$ 0 | - | 2 | + | 1 | + | 0 | + | 6 | - | 2 | - | 2 | - | 3 |
| $\ldots +$ 0 | + | 2 | - | 1 | + | 0 | + | 6 | + | 2 | + | 2 | - | 3 |
| $\ldots +$ 0 | - | 2 | - | 1 | + | 0 | + | 6 | - | 2 | + | 2 | + | 3 |

Table 8.9: Part 2 of Decomposition Matrix for the Data of Table 8.2

alcohol effect is superimposed. Alternatively, it may mean that the effect of alcohol truly does depend on previous experience with the task. These data do not decide that question. But they do show that Alcohol has at least some effect, and they demonstrate this with order taken into account.

| Source | *df* | *SS* | *MS* | *F* | Error Term |
|---|---|---|---|---|---|
| $\mu$ | 1 | 50700.0 | 50700.0 | 1584.375 | $S(A)$ |
| $A$ (Age) | 1 | 432.0 | 432.0 | 13.500 | $S(A)$ |
| $B$ (Drug) | 1 | 300.0 | 300.0 | 12.500 | $BS(A)$ |
| $AB$ | 1 | 192.0 | 192.0 | 8.000 | $BS(A)$ |
| $S(A)$ | 4 | 128.0 | 32.0 | | |
| $BS(A)$ | 4 | 96.0 | 24.0 | | |
| Total | 12 | 51848.0 | | | |

Table 8.10: ANOVA for the Data of Table 8.1

| Source | *df* | *SS* | *MS* | *F* | Error Term |
|---|---|---|---|---|---|
| $\mu$ | 1 | 135200 | 135200 | 620.183 | $S(AC)$ |
| $A$ (Age) | 1 | 128 | 128 | 0.587 | $S(AC)$ |
| $B$ (Drug) | 1 | 12800 | 12800 | 220.690 | $BS(AC)$ |
| $AB$ | 1 | 288 | 288 | 4.966 | $BS(AC)$ |
| $C$ (Gender) | 1 | 1152 | 1152 | 5.284 | $S(AC)$ |
| $AC$ | 1 | 0 | 0 | 0.000 | $S(AC)$ |
| $BC$ | 1 | 800 | 800 | 13.793 | $BS(AC)$ |
| $ABC$ | 1 | 32 | 32 | 0.552 | $BS(AC)$ |
| $D$ (Stress) | 1 | 800 | 800 | 10.811 | $DS(AC)$ |
| $AD$ | 1 | 32 | 32 | 0.432 | $DS(AC)$ |
| $BD$ | 1 | 3200 | 3200 | 53.333 | $BDS(AC)$ |
| $ABD$ | 1 | 32 | 32 | 0.533 | $BDS(AC)$ |
| $CD$ | 1 | 512 | 512 | 6.919 | $DS(AC)$ |
| $ACD$ | 1 | 128 | 128 | 1.730 | $DS(AC)$ |
| $BCD$ | 1 | 32 | 32 | 0.533 | $BDS(AC)$ |
| $ABCD$ | 1 | 0 | 0 | 0.000 | $BDS(AC)$ |
| $S(AC)$ | 4 | 872 | 218 | | |
| $DS(AC)$ | 4 | 296 | 74 | | |
| $BS(AC)$ | 4 | 232 | 58 | | |
| $BDS(AC)$ | 4 | 240 | 60 | | |
| Total | 32 | 156776 | | | |

Table 8.11: ANOVA for the Data of Table 8.2

| Order 1: Alcohol 1st | | | | Order 2: No-Alcohol 1st | | | |
|---|---|---|---|---|---|---|---|
| Subject | Alcohol | No-Alcohol | Average | Subject | Alcohol | No-Alcohol | Average |
| 1 | 34 | 12 | 23 | 1 | 13 | 19 | 16 |
| 2 | 27 | 13 | 20 | 2 | 24 | 24 | 24 |
| 3 | 23 | 7 | 15 | 3 | 22 | 28 | 25 |
| 4 | 20 | 8 | 14 | 4 | 21 | 25 | 23 |
| Average | 26 | 10 | 18 | | 20 | 24 | 22 |

Table 8.12: Sample Data for Order Effect Example

| Source | *df* | *SS* | *MS* | *F* | Error Term |
|---|---|---|---|---|---|
| $\mu$ | 1 | 6400 | 6400 | 184 | $S(B)$ |
| $A$ | 1 | 144 | 144 | 21.5 | $AS(B)$ |
| $B$ | 1 | 64 | 64 | 1.8 | $S(B)$ |
| $AB$ | 1 | 400 | 400 | 59.7 | $AS(B)$ |
| $S(B)$ | 6 | 208 | 34.7 | | |
| $AS(B)$ | 6 | 40 | 6.7 | | |
| Total | 16 | 7256 | | | |

Table 8.13: ANOVA Table for Mixed Design With Order Factor

# Chapter 9

# Shortcut for Computing Sums of Squares

This chapter explains how to use a computational shortcut for getting sums of squares. Basically, it is possible to get the sums of squares *without* obtaining the numerical estimates for all of the terms in the model, as is required to obtain $SS$'s from the decomposition matrix. Since you don't actually have to estimate each term in the model, you save a lot of subtractions.

This computational shortcut is not given in the earlier chapters because it does not promote intuitive understanding. We present it here in spite of this major defect, just in case you are ever caught away from a computer and want to do a sizeable ANOVA by hand.

The computational shortcut works for all equal cell size designs: between-subjects, within-subjects, and mixed. Before you can use it, though, you must first be able to write out the model. Then, to compute the $SS$ for any term in the model:

1. Write out its estimation equation, which has the form:

   $$\text{Term} = \text{Data average - estimates of other model terms}$$

2. Write out the corresponding shortcut equation, which has the form:

   $$SS_{Term} = n \times \sum(\text{Data average}^2) - SS's \text{ for other model terms}$$

   To construct and use this equation you need to know two things:

   - $n$ is the number of scores that are averaged together to compute each of the data averages to which the equation refers.
   - The "other model terms" for which you subtract off the $SS$'s are the very same terms for which you subtract off estimates in the corresponding estimation equation.

   That is all there is to it.

## 9.1 One-Factor Between-Ss Example

To illustrate the shortcut in the simplest case, we will use it to get $SS$'s for the data of Section 3.10. The model is:

$$Y_{ij} = \mu + A_i + S(A)_{ij}$$

and so we only need to compute $SS_\mu$, $SS_A$, and $SS_{S(A)}$.

**$\mu$:** The estimation equation is

$$\hat{\mu} = Y_{\cdot\cdot}$$

Thus, the corresponding shortcut $SS$ equation is

$$SS_\mu = n \times Y_{\cdot\cdot}{}^2$$

- $n$ is 12, because $Y_{\cdot\cdot}$ is an average over 12 data values.

- No "other model terms" are subtracted from the data value in the estimation equation, so no $SS$'s are subtracted in the shortcut equation.

Substituting in the numbers gives

$$SS_\mu = 12 \times 30^2 = 10,800$$

**$A$:** The estimation equation is

$$\hat{A}_i = Y_{i.} - \hat{\mu}$$

Thus, the corresponding shortcut $SS$ equation is

$$SS_A = n \times \sum({Y_{i.}}^2) - SS_\mu$$

- $n$ is 3, because each $Y_{i.}$ is an average over 3 data values.
- $\hat{\mu}$ is subtracted from the data value in the estimation equation, so $SS_\mu$ is subtracted in the shortcut equation.

Thus,

$$SS_A = 3 \times (40^2 + 34^2 + 24^2 + 22^2) - 10,800 = 648$$

**$S(A)$:** The estimation equation is

$$\widehat{S(A)}_i = Y_{ij} - \hat{\mu} - \hat{A}_i$$

Thus, the corresponding shortcut $SS$ equation is

$$SS_{S(A)} = n \times \sum({Y_{ij}}^2) - SS_\mu - SS_A$$

- $n$ is 1, because each $Y_{ij}$ is an average over 1 data value (i.e., it is just a data value).
- $\hat{\mu}$ and $\hat{A}_i$ are subtracted from the data value in the estimation equation, so $SS_\mu$ and $SS_A$ are subtracted in the shortcut equation.

Thus,

$$\begin{aligned} SS_{S(A)} &= (39^2 + 41^2 + 40^2 + 35^2 + 35^2 + 32^2 + 24^2 + 25^2 + 23^2 + 22^2 + 23^2 + 21^2) \\ &\quad -10,800 - 648 \\ &= 12 \end{aligned}$$

## 9.2 Three-Factor Between-Ss Example

Table 9.1 shows sample data from an experiment examining the amount learned by students as a function of their own gender, the gender of their teachers, and the type of courses being taken (science vs. language). The cell and marginal means are also shown.

To compute the $SS$'s with the shortcut method, proceed as follows:

**$\mu$:** The estimation equation is:

$$\hat{\mu} = Y_{....}$$

The corresponding shortcut equation is therefore:

$$SS_\mu = n \times {Y_{....}}^2$$

with $n = 24$, because 24 data values were averaged together to get $Y_{....}$. Thus,

$$SS_\mu = 24 \times 100^2 = 240,000$$

Data:

Science Courses

| Students | Male Teachers | | | Female Teachers | | |
|---|---|---|---|---|---|---|
| Male | 101 | 99 | 103 | 101 | 105 | 97 |
| Female | 90 | 92 | 97 | 91 | 96 | 92 |

Language Courses

| Students | Male Teachers | | | Female Teachers | | |
|---|---|---|---|---|---|---|
| Male | 109 | 107 | 105 | 97 | 99 | 101 |
| Female | 99 | 94 | 92 | 114 | 114 | 105 |

Cell Means:

Science Courses

| Students: | Male Teachers | Female Teachers |
|---|---|---|
| Male | 101 | 101 |
| Female | 93 | 93 |

Language Courses

| Students: | Male Teachers | Female Teachers |
|---|---|---|
| Male | 107 | 99 |
| Female | 95 | 111 |

Marginal means:

| Students: | Male Teachers | Female Teachers | Average |
|---|---|---|---|
| Male | 104 | 100 | 102 |
| Female | 94 | 102 | 98 |
| Average | 99 | 101 | 100 |

| Courses: | Male Teachers | Female Teachers | Average |
|---|---|---|---|
| Science | 97 | 97 | 97 |
| Language | 101 | 105 | 103 |

| Courses: | Male Students | Female Students | Average |
|---|---|---|---|
| Science | 101 | 93 | 97 |
| Language | 103 | 103 | 103 |
| Average | 102 | 98 | 100 |

Table 9.1: Sample Data for Teacher x Student x Course Experiment

$A$**:** The estimation equation is:

$$\hat{A}_i = Y_{i...} - \hat{\mu}$$

and the corresponding shortcut equation is:

$$SS_A = n \times \sum({Y_{i...}}^2) - SS_\mu$$

$n = 12$, because 12 data values are averaged together to get each $Y_{i...}$. Thus,

$$\begin{aligned} SS_A &= 12 \times (99^2 + 101^2) - SS_\mu \\ &= 240,024 - 240,000 = 24 \end{aligned}$$

$B$**:** The estimation equation is:

$$\hat{B}_j = Y_{.j.} - \hat{\mu}$$

and the corresponding shortcut equation is:

$$SS_B = n \times \sum({Y_{.j.}}^2) - SS_\mu$$

$n = 12$, because 12 data values are averaged together to get each $Y_{.j.}$. Thus,

$$\begin{aligned} SS_B &= 12 \times (102^2 + 98^2) - SS_\mu \\ &= 240,096 - 240,000 = 96 \end{aligned}$$

*AB*: The estimation equation is:

$$\widehat{AB}_{ij} = Y_{ij..} - \hat{\mu} - \hat{A}_i - \hat{B}_j$$

and the corresponding shortcut equation is:

$$SS_{AB} = n \times \sum({Y_{ij..}}^2) - SS_\mu - SS_A - SS_B$$

$n = 6$, because 6 data values are averaged together to get each $Y_{ij..}$. Thus,

$$\begin{aligned} SS_{AB} &= 6 \times (104^2 + 94^2 + 100^2 + 102^2) - 240,000 - 24 - 96 \\ &= 240,336 - 240,120 = 216 \end{aligned}$$

*C*: The estimation equation is:

$$\hat{C}_k = Y_{..k.} - \hat{\mu}$$

and the corresponding shortcut equation is:

$$SS_C = n \times \sum({Y_{..k.}}^2) - SS_\mu$$

$n = 12$, because 12 data values are averaged together to get each $Y_{..k.}$. Thus,

$$\begin{aligned} SS_C &= 12 \times (97^2 + 103^2) - SS_\mu \\ &= 240,216 - 240,000 = 216 \end{aligned}$$

*AC*: The estimation equation is:

$$\widehat{AC}_{ik} = Y_{i.k.} - \hat{\mu} - \hat{A}_i - \hat{C}_k$$

and the corresponding shortcut equation is:

$$SS_{AC} = n \times \sum({Y_{i.k.}}^2) - SS_\mu - SS_A - SS_C$$

$n = 6$, because 6 data values are averaged together to get each $Y_{i.k.}$. Thus,

$$\begin{aligned} SS_{AC} &= 6 \times (97^2 + 101^2 + 97^2 + 105^2) - 240,000 - 24 - 216 \\ &= 240,264 - 240,240 = 24 \end{aligned}$$

*BC*: The estimation equation is:

$$\widehat{BC}_{jk} = Y_{.jk.} - \hat{\mu} - \hat{B}_j - \hat{C}_k$$

and the corresponding shortcut equation is:

$$SS_{BC} = n \times \sum({Y_{.jk.}}^2) - SS_\mu - SS_B - SS_C$$

$n = 6$, because 6 data values are averaged together to get each $Y_{.jk.}$. Thus,

$$\begin{aligned} SS_{BC} &= 6 \times (101^2 + 103^2 + 93^2 + 103^2) - 240,000 - 96 - 216 \\ &= 240,408 - 240,312 = 96 \end{aligned}$$

*ABC*: The estimation equation is:

$$\widehat{ABC}_{ijk} = Y_{ijk.} - \hat{\mu} - \hat{A}_i - \hat{B}_j - \widehat{AB}_{ij} - \hat{C}_k - \widehat{AC}_{ik} - \widehat{BC}_{jk}$$

and the corresponding shortcut equation is:

$$SS_{ABC} = n \times \sum({Y_{ijk.}}^2) - SS_\mu - SS_A - SS_B - SS_{AB} - SS_C - SS_{AC} - SS_{BC}$$

$n = 3$, because 3 data values are averaged together to get each $Y_{ijk.}$. Thus,

$$\begin{aligned} SS_{ABC} &= 3 \times (101^2 + 93^2 + 101^2 + 93^2 + 107^2 + 95^2 + 99^2 + 111^2) \\ &\quad -240,000 - 24 - 96 - 216 - 216 - 24 - 96 \\ &= 216 \end{aligned}$$

$S(ABC)$: The estimation equation is:

$$S(\widehat{ABC})_{ijkl} = Y_{ijkl} - \hat{\mu} - \hat{A}_i - \hat{B}_j - \widehat{AB}_{ij} - \hat{C}_k - \widehat{AC}_{ik} - \widehat{BC}_{jk} - \widehat{ABC}_{ijk}$$

and the corresponding shortcut equation is:

$$SS_{S(ABC)} = n \times \sum({Y_{ijkl}}^2) - SS_\mu - SS_A - SS_B - SS_{AB} - SS_C - SS_{AC} - SS_{BC} - SS_{ABC}$$

$n = 1$, because each $Y_{ijkl}$ is simply one data value. Thus,

$$\begin{aligned} SS_{S(ABC)} &= 1 \times (101^2 + 99^2 + 103^2 + 109^2 + 107^2 + 105^2 + 90^2 + 92^2 \\ &\quad +97^2 + 99^2 + 94^2 + 92^2 + 101^2 + 105^2 + 97^2 + 97^2 + 99^2 + 101^2 \\ &\quad +91^2 + 96^2 + 92^2 + 114^2 + 114^2 + 105^2) \\ &\quad -240,000 - 24 - 96 - 216 - 216 - 24 - 96 - 216 \\ &= 176 \end{aligned}$$

## 9.3 Two-Factor Within-Ss Example

Here is a within-subjects example, using the data of Problem 3 of Section 7.7.

$\mu$: The estimation equation is:

$$\hat{\mu} = Y_{...}$$

and the corresponding shortcut equation is:

$$SS_\mu = n \times {Y_{...}}^2$$

$n$ is 30 because $Y_{...}$ is an average over 30 data values, so

$$SS_\mu = 30 \times 53^2 = 84,270$$

$A$: The estimation equation is:

$$\hat{A}_i = Y_{i..} - \hat{\mu}$$

and the corresponding shortcut equation is:

$$SS_A = n \times \sum({Y_{i..}}^2) - SS_\mu$$

$n = 15$, because 15 data values are averaged together to get each $Y_{i..}$. Thus,

$$SS_A = 15 \times (65^2 + 41^2) - 84,270 = 4,320$$

$B$: The estimation equation is:

$$\hat{B}_j = Y_{.j.} - \hat{\mu}$$

and the corresponding shortcut equation is:

$$SS_B = n \times \sum({Y_{.j.}}^2) - SS_\mu$$

$n = 10$, because 10 data values are averaged together to get each $Y_{.j.}$. Thus,

$$SS_B = 10 \times (49^2 + 59^2 + 51^2) - 84,270 = 560$$

$AB$: The estimation equation is:

$$\widehat{AB}_{ij} = Y_{ij.} - \hat{\mu} - \hat{A}_i - \hat{B}_j$$

and the corresponding shortcut equation is:

$$SS_{AB} = n \times \sum({Y_{ij.}}^2) - SS_\mu - SS_A - SS_B$$

$n = 5$, because 5 data values are averaged together to get each $Y_{ij.}$. Thus,

$$\begin{aligned} SS_{AB} &= 5 \times (66^2 + 32^2 + 69^2 + 49^2 + 60^2 + 42^2) - 84,270 - 4,320 - 560 \\ &= 380 \end{aligned}$$

$S$: The estimation equation is:

$$\hat{S}_k = Y_{..k} - \hat{\mu}$$

and the corresponding shortcut equation is:

$$SS_S = n \times \sum({Y_{..k}}^2) - SS_\mu$$

$n = 6$, because 6 data values are averaged together to get each $Y_{..k}$. Thus,

$$\begin{aligned} SS_S &= 6 \times (52^2 + 58^2 + 62^2 + 63^2 + 30^2) - 84,270 \\ &= 4,416 \end{aligned}$$

$AS$: The estimation equation is:

$$\widehat{AS}_{ik} = Y_{i.k} - \hat{\mu} - \hat{A}_i - \hat{S}_k$$

and the corresponding shortcut equation is:

$$SS_{AS} = n \times \sum({Y_{i.k}}^2) - SS_\mu - SS_A - SS_S$$

$n = 3$, because 3 data values are averaged together to get each $Y_{i.k}$. Thus,

$$\begin{aligned} SS_{AS} &= 3 \times (64^2 + 40^2 + 68^2 + 48^2 + ... + 37^2 + 23^2) - 84,270 - 4,320 - 4,416 \\ &= 348 \end{aligned}$$

$BS$: The estimation equation is:

$$\widehat{BS}_{jk} = Y_{.jk} - \hat{\mu} - \hat{B}_j - \hat{S}_k$$

and the corresponding shortcut equation is:

$$SS_{BS} = n \times \sum({Y_{.jk}}^2) - SS_\mu - SS_B - SS_S$$

$n = 2$, because 2 data values are averaged together to get each $Y_{.jk}$. Thus,

$$\begin{aligned} SS_{BS} &= 2 \times (47^2 + 62^2 + 47^2 + ... + 24^2 + 29^2 + 37^2) - 84,270 - 560 - 4,416 \\ &= 424 \end{aligned}$$

$ABS$: The estimation equation is:

$$\widehat{ABS}_{ijk} = Y_{ijk} - \hat{\mu} - \hat{A}_i - \hat{B}_j - \widehat{AB}_{ij} - \hat{S}_k - \widehat{AS}_{ik} - \widehat{BS}_{jk}$$

and the corresponding shortcut equation is:

$$SS_{ABS} = n \times \sum({Y_{ijk}}^2) - SS_\mu - SS_A - SS_B - SS_{AB} - SS_S - SS_{AS} - SS_{BS}$$

$n = 1$, because each $Y_{ijk}$ is simply a single data value. Thus,

$$\begin{aligned} SS_{ABS} &= 1 \times (63^2 + 31^2 + 71^2 + \ldots + 29^2 + 49^2 + 25^2) - 84,270 - 4,320 - 560 \\ &\quad -380 - 4,416 - 348 - 424 \\ &= 324 \end{aligned}$$

# Part II

# Regression

# Chapter 10

# Introduction to Correlation and Regression

Correlation and regression analyses use the GLM to examine the relationship between a numerical DV (as in ANOVA) and one or more independent variables that are also numerical (unlike ANOVA). For example, such analyses would be used to answer questions like,

- "What is the relationship between intelligence and height?" (Note these are both numerical variables.)
- "How does attitude toward the death penalty vary with income?" (Income is obviously numerical; for a regression analysis, attitude toward the death penalty would have to be numerical as well, perhaps measured on a 5-point scale.)
- "Is the amount learned by a statistics student related to that student's math aptitude?" (Again, both amount learned and math aptitude are numerical variables.)

There are two major uses for correlation and regression analyses: hypothesis testing and prediction. As an example of hypothesis testing, you might wonder whether there is *any* relationship between statistics learning and math aptitude. The null hypothesis would say that there was no relationship, and the goal of the study would be to find out whether this null hypothesis could be rejected. As an example of prediction, you might want to predict a student's statistics score (the "dependent" or "to-be-predicted" variable) from his or her score on a math aptitude test (the "predictor" variable). Such a prediction might be used, for example, to identify students who needed special tutoring in order to pass the class. This book will emphasize hypothesis testing, because that is the main use of these techniques, at least within the social sciences.

We will also pay considerable attention to the issue of drawing causal conclusions from correlation and regression analyses. As will be seen, random assignment is usually not possible in the types of studies analyzed with these statistical techniques, and it is not possible to draw firm causal conclusions without random assignment. Nonetheless, as illustrated by the main conceptual example of this chapter, it is often possible to find evidence that strongly *suggests* causal relationships. Thus, the first studies in a new research area often use correlation and regression to identify the most likely causal relationships, and later studies in the area can study these relationships more definitively in controlled experiments using random assignment.

Before starting, we should also note that "correlation" and "regression" refer to two slightly different versions of essentially the same statistical technique. The two versions address the same sorts of questions and produce the same answers, so it hardly matters which one you use. They use different terminology, however, and we will generally emphasize regression because we prefer its terminology. We will also present the basic correlation terminology, though, so that you will be able to understand it when you encounter it in others' research reports.

## 10.1 A Conceptual Example

The remainder of this chapter presents an example illustrating the use of regression analysis from a conceptual viewpoint. Several important concepts and refinements will be added to this picture in

the next several chapters, but this example captures many of the key ideas of regression analysis, especially with respect to hypothesis testing and inferences about causality.

Table 10.1 provides an example "case by variable" data set of the sort needed for correlation and regression analysis. The term "case" refers to the subjects (i.e., randomly selected entities) in the study. In this example, the cases are students in a statistics class. The "variables" refer to the different measurements taken on each case. In this example, each student was measured on three variables: the percentage of 10 statistics homework assignments that were completed (HWRK%), the average percentage scored on all the exams in the statistics class (EXAM%), and the average percentage scored by the student in all of his or her *other classes* in the university (UNIV%).

| Case | HWRK% | EXAM% | UNIV% |
|---|---|---|---|
| 1 | 50 | 91 | 88 |
| 2 | 70 | 77 | 50 |
| 3 | 50 | 55 | 62 |
| 4 | 70 | 85 | 85 |
| 5 | 70 | 58 | 65 |
| 6 | 80 | 53 | 73 |
| 7 | 50 | 73 | 83 |
| 8 | 50 | 77 | 70 |
| 9 | 60 | 61 | 92 |
| 10 | 50 | 50 | 58 |
| 11 | 60 | 75 | 82 |
| 12 | 80 | 60 | 84 |

Table 10.1: Example of a "case by variable" data set. Each case is one student in a statistics class. Each student was measured on three variables: HWRK%, EXAM%, and UNIV%.

Although the data in Table 10.1 are fictitious, we actually did collect real data like these on a sample of students in one of our classes. The purpose of the study was to help motivate our students to do assigned homework problems, even though these did not count much toward their final grades in the class. We thought we could motivate the students by presenting data to suggest that doing the homework significantly helped students improve their scores on exams, which did count greatly toward the final grades.

It will take several chapters to present the mechanics of regression in detail, but the basic conceptual issues can be illustrated within the context of this data set. The first question is whether there is any association between students' homework diligence and their exam scores. In order to motivate our students, we obviously needed to demonstrate that there is, and that in particular exam scores are greater for students who did more homeworks.

In fact, our analysis did indeed show that students who did more homeworks tended to score higher on the exams. More specifically, the analysis suggested a relationship like that shown in Table 10.2. This table shows various possible homework percentages; beside each homework percentage, it shows the exam percentage that would be predicted based on the statistical analysis. For example, if a student did 60% of the homeworks, the analysis suggests that the exam score is most likely to be approximately 72%. In contrast, if another student did 80% of the homeworks, the exam score is likely around 78%. The crucial feature of Table 10.2 is that the predicted exam score is *greater* for students who did more of the homework assignments. Given these results, we could try to motivate our students by saying, "Look, students who did more homeworks also tended to do better on their exams. So, you too should work hard on the homeworks—doing so will improve your exam scores."

Are you skeptical? You should be. Essentially, we're claiming that doing homework causes higher exam scores—that is, we're arguing for a causal connection. Our data are certainly *consistent with* that causal connection: If doing homeworks increases exam scores, then students who do more homeworks would be expected to get higher exam scores. But our data don't *demonstrate* that causal connection, because there are other possible explanations of how higher homework percentages could predict higher exam scores. For example, a skeptical student might suggest this reasonable alternative explanation of the results:

> Better students not only do more homeworks, they also do all kinds of other things that might help them get better grades. Specifically, they study more in lots of ways: for example, they read the book more carefully, go over their lecture notes more often, attend

| Actual HWRK% | Predicted EXAM% |
|---|---|
| 60 | 72 |
| 70 | 75 |
| 80 | 78 |
| 90 | 81 |

Table 10.2: A possible predictive relationship of the sort that might be established by regression analysis of data like those shown in Table 10.1. Each line corresponds to one case. Based on the actual HWRK% score, the exam score would be predicted to be the value shown. Note that the predicted exam score increases with the homework percentage.

more tutorials, and so on. They also prepare themselves better physically and mentally for their exams (e.g., by getting more sleep and avoiding alcohol the night before). Perhaps their good exam scores are caused by some or all of these other things they do, and the homeworks per se had nothing to do with their high exam scores. In that case, the relationship between homework diligence and exam scores is merely incidental—not causal.

Now we personally believe that students would be foolish to neglect their homework on the basis of this skeptic's argument, because this is clearly a "Pascal's bet" sort of situation[1]. Grudgingly, however, we'd have to admit that the skeptic is correct in claiming that our data do not establish a causal connection and that this "all-around good students" alternative explanation is equally consistent with the predictive relationship established by our data.

If that were the end of the story, we'd have to conclude that it is basically impossible to argue strongly for a causal connection on the basis of this type of correlational data. In that case, you'd have to do an experiment to have a strong argument for a causal connection. Specifically, you would have to randomly assign some students to do many homeworks and others to do few or none. Only with random assignment could you be certain (within statistical limits, of course) that the number of homeworks done was not hopelessly confounded with the all-around goodness of the student.

Fortunately, that is not the end of the story, because regression analysis allows us to make predictions using more than one variable. Glossing over a lot more computational details, further statistical analyses of the data showed that exam scores could be predicted from *both* the percentage of homeworks and the overall university average grades (UNIV%) approximately as shown in Table 10.3. In this table, you can see that students who do more homeworks are again predicted to score higher on the exams, and that this is true *even for students who are equally good overall, as measured by UNIV%*. As discussed next, this pattern rules out the all-around good student hypothesis and thus strengthens our case considerably.

| Actual HWRK% | Actual UNIV% | Predicted EXAM% | Actual HWRK% | Actual UNIV% | Predicted EXAM% |
|---|---|---|---|---|---|
| 70 | 70 | 74 | 70 | 80 | 83 |
| 80 | 70 | 77 | 80 | 80 | 86 |
| 90 | 70 | 80 | 90 | 80 | 89 |

Table 10.3: A possible predictive relationship of the sort that might be established by regression analysis of data like those shown in Table 10.1. On both the left and right sides of the table, each line corresponds to one case, for a total of six cases in all. Based on the actual HWRK% and UNIV% scores for the case, the exam score would be predicted to be the value shown. Note that the predicted exam score increases with the homework percentage, even for students with a fixed UNIV% (e.g., the three cases on the left).

The argument is a bit tricky. First, look at the left-most three columns. Here you see predictions for three students with *identical* UNIV% values. It seems reasonable to argue that these are equally good students *overall*, because their UNIV%s are all the same. But even considering only these three students, according to this analysis we should still predict that exam scores will increase with

[1]The 17th century mathematician and philosopher Blaise Pascal was reported to have said he was betting in favor of the existence of God because there would be little cost of being wrong in making this bet, as compared with a much larger cost of being wrong in making the opposite bet.

the homework percentage! This increase cannot be explained by the "all-around good students" hypothesis, because the students are equally good overall. Therefore, the increase is stronger support for our claim of a causal connection between homework and exam scores—stronger than the original predictive relationship shown in Table 10.2. Of course, we would have to show that the predicted increase was not simply due to chance and that the increase was not just predicted for students with UNIV%=70 but also for students with UNIV%=80 (right-most three columns) and all others as well. Assuming we could show both of these things, though, we would be able to argue that the data demonstrated a predictive relationship between homework diligence and exam scores *beyond that explainable by the all-around good student hypothesis*. Once we concluded that the data were consistent with our proposed causal connection and were inconsistent with the all-around good student hypothesis, the ball would go back into the skeptics court. They would have to either generate another plausible alternative explanation or grant that we had pretty good evidence of a causal connection after all. If they did generate another plausible explanation, of course, we would try to conduct a further analysis with a new variable to rule out this explanation, just as UNIV% allowed us to rule out the all-around good student hypothesis.

Of course the data might not come out as shown in Table 10.3. We might carry out further analyses to predict exam scores from HWRK% and UNIV%, and we might instead obtain a pattern like that shown in Table 10.4. Note that in this case the predicted exam score depends only on the student's overall university score (UNIV%), not at all on the homework score. This pattern strongly supports the all-around good student hypothesis and it contradicts our proposed causal connection.

| Actual HWRK% | Actual UNIV% | Predicted EXAM% | Actual HWRK% | Actual UNIV% | Predicted EXAM% |
|---|---|---|---|---|---|
| 70 | 70 | 74 | 70 | 80 | 83 |
| 80 | 70 | 74 | 80 | 80 | 83 |
| 90 | 70 | 74 | 90 | 80 | 83 |

Table 10.4: A possible predictive relationship of the sort that might be established by regression analysis of data like those shown in Table 10.1. On both the left and right sides of the table, each line corresponds to one case, for a total of six cases in all. Based on the actual HWRK% and UNIV% scores for the case, the exam score would be predicted to be the value shown. Note that the predicted exam score does not increase with the homework percentage if you look only at students with identical UNIV%s.

As a student of statistics, you may be interested to know how the data actually did come out for this example. We analyzed scores of 209 students in 1996 and 1997 statistics classes, and the results showed a highly reliable trend of the sort shown in Table 10.3. Specifically, even equating for a student's overall UNIV% in university classes, there was a significant tendency for students who did more homework to get higher exam scores. Thus, the real data support our claim that doing homework causes higher exam scores.

Of course these results still only *suggest* a causal relationship—they do not prove one—because it is still possible that some other variable is responsible for the relationship between homework scores and exam scores. For example, perhaps students vary in their enjoyment of statistics, and the ones who enjoy it more tend both to do more homeworks and to study more for exams. Then, it could be the extra studying rather than the extra homeworks that caused higher exam performance. To rule out this explanation, we would have to ask students how much they liked statistics and control for that just as we controlled for overall university grades.

Nonetheless, even though the analysis with overall university grades doesn't prove a causal effect of doing homeworks on exam scores, it should be clear that the argument for such a connection is strengthened by including overall university marks in the analysis. Specifically, the analysis provides evidence against one plausible alternative explanation of the correlation (the "all-around good students" hypothesis). This is often the situation with regression data sets: They cannot demonstrate causal connections conclusively, but they can build a potentially very strong case that such a connection is present.

## 10.2 Overview

Regression analysis allows you to make predictions about one numerical dependent variable based on one or more numerical predictor variables. You can first of all see whether the dependent variable (e.g., EXAM%) is related to a certain predictor variable (e.g., HWRK%). There are two possible outcomes:

1. If such a relation is found, it is consistent with the hypothesis of a causal influence of the predictor variable on the dependent variable. The causal influence is not proved, however. As discussed in the example of this chapter, one possibility is that some other variable actually has the causal influence (i.e., UNIV%).

2. If instead no relation is found, then the results suggest that there is no causal influence of the predictor variable on the dependent variable. The absence of a causal influence is not proved, however, for reasons discussed in later chapters. For example, one possibility is that there is too much error to demonstrate a causal influence that really is present.

In either case, further information may be gained by including additional variables in the analysis, as discussed starting in chapter 14.

1. If a relation was found, it is possible to see whether the relation still holds when additional predictors (e.g., UNIV%) are taken into account. As discussed in this chapter, such analyses can help determine whether the other variables actually have the causal influence.

2. If no relation was found, it is possible to see whether a relation may emerge when other variables are taken into account. This can happen for a variety of reasons, as will be discussed in chapter 16.

In the end, regression analysis can never be as definitive about causality as a randomized experiment can be. Regardless of how many variables are taken into account and how many alternative hypotheses are ruled out, there are always more potential variables to look at and more hypotheses to consider. But as illustrated by the example of this chapter, regression analysis can often be used to gain strongly suggestive evidence about what variables do and do not have a causal influence on the dependent variable, even in situations where it is impossible or impractical to conduct an experiment using random assignment.

# Chapter 11

# Simple Correlation

Correlation analysis is used to detect and quantify relationships among numerical variables. It uses data of the "case by variable" format introduced in Chapter 10. As an illustrative example, Table 11.1 presents hypothetical data from a study of reading abilities in school children. Note that each child in the sample was measured with respect to several numerical variables. The study was conducted to find out more about what makes children better readers, by identifying which variables are most closely associated with reading ability.

| Case | Abil | IQ | Home | TV |
|---|---|---|---|---|
| 1 | 61 | 107 | 144 | 487 |
| 2 | 56 | 109 | 123 | 608 |
| 3 | 45 | 81 | 108 | 640 |
| 4 | 66 | 100 | 155 | 493 |
| 5 | 49 | 92 | 103 | 636 |
| 6 | 62 | 105 | 161 | 407 |
| 7 | 61 | 92 | 138 | 463 |
| 8 | 55 | 101 | 119 | 717 |
| 9 | 62 | 118 | 155 | 643 |
| 10 | 61 | 99 | 121 | 674 |
| 11 | 51 | 104 | 93 | 675 |
| 12 | 48 | 100 | 127 | 595 |
| 13 | 50 | 95 | 97 | 673 |
| 14 | 50 | 82 | 140 | 523 |
| 15 | 67 | 114 | 151 | 665 |
| 16 | 51 | 95 | 112 | 663 |
| 17 | 55 | 94 | 102 | 684 |
| 18 | 54 | 103 | 142 | 505 |
| 19 | 57 | 96 | 127 | 541 |
| 20 | 54 | 104 | 102 | 678 |
| 21 | 52 | 98 | 124 | 564 |
| 22 | 48 | 117 | 87 | 787 |
| 23 | 61 | 100 | 141 | 582 |
| 24 | 54 | 101 | 117 | 647 |
| 25 | 48 | 94 | 111 | 448 |

Table 11.1: Example data for simple correlation analyses. A sample of 25 8-year-old children was obtained from a local school, and each child was measured on several variables: a standardized test of reading ability (Abil), intelligence (IQ), the number of minutes per week spent reading in the home (Home), and the number of minutes per week spent watching TV (TV).

"Simple" correlation refers to a correlation analysis that uses only two variables at a time (also sometimes called "bivariate" correlation). In contrast, "multiple" correlation uses three or more variables at a time, as discussed starting in Chapter 14. In this chapter, we will discuss only simple correlation analyses of the example data in Table 11.1. Even though there are more than two variables in the data set, the analyses presented in this chapter are "simple" because they never involve more

than two variables within the same calculations.

## 11.1 The Scattergram

The best starting point for simple correlation analysis is the scattergram, which is a type of graph designed to reveal the relationship between two variables. For example, Figure 11.1 is a scattergram showing the relationship between reading ability and home reading time for the data in Table 11.1. To construct a scattergram, you first choose one variable to put on the horizontal axis and another variable to put on the vertical axis. The choice of which variable goes on which axis is somewhat arbitrary, but the usual convention is to put the dependent (to-be-predicted) variable on the vertical axis and the independent (predictor) variable on the horizontal axis. You then put numerical labels on each axis, being sure to extend each axis slightly past the largest and smallest values on its variable within the sample. Once the axes are labelled, you plot each case as one point on the scattergram. The location of the case is determined by its values on the variables being plotted in the scattergram. For example, case 3 is marked in Figure 11.1: Its values of Home=108 and Abil=45 dictate where it is placed on the scattergram.

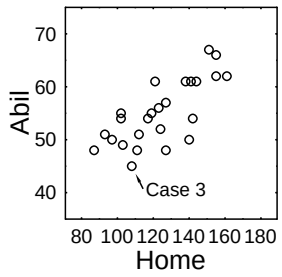


Figure 11.1: Scattergram displaying the relationship between reading ability (Abil) and home reading time (Home) for the data in Table 11.1.

The scattergram is a visual summary of a relationship between two variables, not really a formal statistical procedure. In Figure 11.1, for example, you can probably see a tendency for the points to go from the lower left to the upper right of the scattergram. This suggests that there is some relationship between the two variables: Cases with smaller values of Abil tend to have smaller values of Home, and cases with larger values on one variable tend also to have larger values on the other. This relationship is by no means perfect, but the trend is clear enough that you can see it by eye.

How would you interpret the association between reading ability and home reading, assuming you could rule out the possibility that it was simply found by chance (you will see how to do this later in the chapter)? Really, all you can say is that children with more reading ability tend to read more at home than do children with less reading ability. There are various causal explanations of the association, but you can't tell which is right just by establishing the trend. For example, it could be that more home reading causes reading ability to improve. Conversely, having better reading ability might cause children to read more at home (e.g., because reading is more enjoyable if it is more fluent). A third possibility is that both reading ability and amount of home reading are determined by some third variable, such as the degree of encouragement provided by the parents. In principle, you could eliminate the parental encouragement explanation by measuring this variable and taking it into account, just as overall university grades were taken into account in Chapter 10. But there is no way to distinguish between home reading causing better reading ability and the reverse, unless you can randomly assign the children to spend different amounts of time reading at home.

## 11.2 Types of Bivariate Relationships

There is no limit to the number of different types of relationships that might be observed between two variables. A few of the possibilities are depicted in Figure 11.2.

Panel A shows a scattergram in which there is no relationship between the two variables. As you can see, knowing the value of one variable doesn't tell you anything at all about the value of the

other variable, because the values of the two variables are unrelated to one another (statistically, they are often said to be "independent" of one another). This null relationship represents the situation *according to the null hypothesis* in correlation research. The question is always whether you have observed a relationship that is too systematic to have been sampled by chance out of a population that really looks like this.

Panels B and C show what are called *linear* relationships between variables. Panel B shows an increasing or *positive* linear relationship: Larger values on one variable are associated with larger values on the other variable; moreover, each 1-point increase on one variable tends to associated with the same size increase on the other variable, so that the curve increases at a constant rate going across the graph. This is the same sort of relationship as the one shown in Figure 11.1, but it is stronger in that the points lie more nearly in a straight line. Panel C shows a decreasing or *negative* relationship: Larger values on one variable are associated with *smaller* values on the other variable. You might expect a relationship like this, for example, between time spent reading and time spent watching TV in the data of Table 11.1: Kids who spend more time watching TV would likely spend less time reading, and vice versa.

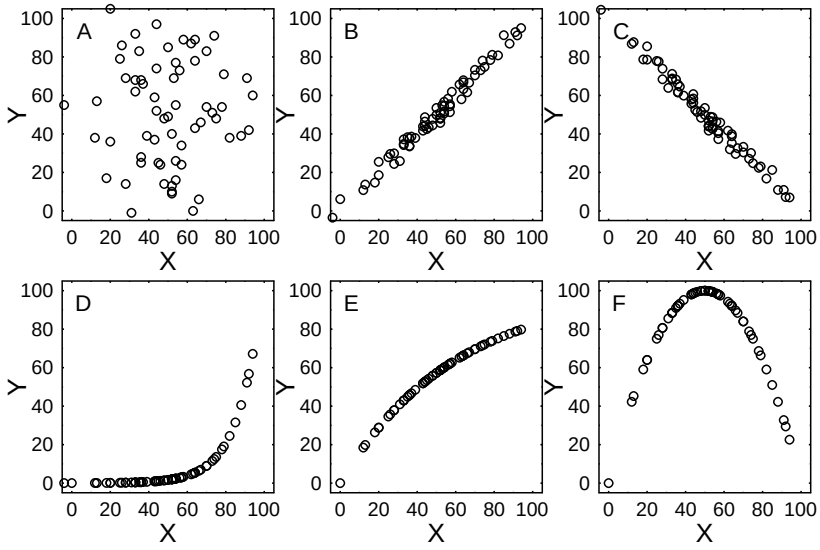


Figure 11.2: Scattergrams displaying some possible relationships between two variables.

Panels D, E, and F show three of the many possible *curvilinear* relationships. In Panel D, $Y$ increases with $X$, but the increase accelerates as $X$ gets larger, causing the curve to rise more and more rapidly. As an example, this curve depicts the relationship between time ($X$) and human technological capability ($Y$). Technological progress was quite slow over the first many centuries of our species, but the rise has become more and more rapid over the past several hundred years. Panel E depicts an increasing relationship that decelerates, reflecting a situation that economists refer to as "diminishing returns." For example, you might expect a relationship like this between amount of time spent exercising ($X$) and physical fitness ($Y$). For people who don't exercise much, increasing (say) 2 hours per week might provide a substantial increase in fitness. For people who already exercise a lot and are near peak fitness, however, an increase of 2 hours per week would provide little further gain. Finally, Panel F shows one possible *nonmonotonic* relationship, usually called an "inverted U." Note that as $X$ increases, $Y$ at first increases too (left side of scattergram) but then levels off and starts to decrease (right side). This might be the relationship, for example, between a person's aggressiveness ($X$) and his or her success as a salesperson ($Y$). The most successful sales people would have to be medium-aggressive, and sales people who were not aggressive enough or too aggressive would be less successful.

An important caveat about correlation and regression analysis is that they are designed primarily to be used when variables have a *linear* relationship. It is possible to adapt the general linear model to handle situations with nonlinear relationships, and this is more or less difficult depending on the exact nature of the nonlinearity. For right now, though, we will restrict our consideration to the analysis of linear relationships.

## 11.3 The Correlation Coefficient

The scattergram is an effective visual depiction of the relationship between two variables, but it is not very quantitative. A quantitative measure is also needed, both to index the relative strengths of relationships and to test whether an observed relationship is too strong to have occurred by chance. The correlation coefficient is one such measure.

The formula for the correlation coefficient—more technically known as Pearson's $r$ after its inventor Karl Pearson—is shown here:

$$r = \frac{\sum_{i=1}^{N} X_i \times Y_i - N \times \bar{X} \times \bar{Y}}{\sqrt{\left(\sum_{i=1}^{N} {X_i}^2 - N \times \bar{X}^2\right) \times \left(\sum_{i=1}^{N} {Y_i}^2 - N \times \bar{Y}^2\right)}} \tag{11.1}$$

Table 11.2 illustrates the correlation computation using the IQ and Abil data of Table 11.1. For this computation we are considering IQ to be $X$ and Abil to be $Y$, but this is arbitrary (i.e., the same value of $r$ is obtained if the two variables are reversed). Note that Table 11.2 includes not only columns for IQ and Abil themselves, but also two more columns for the squares of these variables and one further column for their product. The three extra columns are needed for the computation of $r$.

| Case | IQ | Abil | IQ$^2$ | Abil$^2$ | IQ×Abil |
|---|---|---|---|---|---|
| 1 | 107 | 61 | 11449 | 3721 | 6527 |
| 2 | 109 | 56 | 11881 | 3136 | 6104 |
| 3 | 81 | 45 | 6561 | 2025 | 3645 |
| 4 | 100 | 66 | 10000 | 4356 | 6600 |
| 5 | 92 | 49 | 8464 | 2401 | 4508 |
| 6 | 105 | 62 | 11025 | 3844 | 6510 |
| 7 | 92 | 61 | 8464 | 3721 | 5612 |
| 8 | 101 | 55 | 10201 | 3025 | 5555 |
| 9 | 118 | 62 | 13924 | 3844 | 7316 |
| 10 | 99 | 61 | 9801 | 3721 | 6039 |
| 11 | 104 | 51 | 10816 | 2601 | 5304 |
| 12 | 100 | 48 | 10000 | 2304 | 4800 |
| 13 | 95 | 50 | 9025 | 2500 | 4750 |
| 14 | 82 | 50 | 6724 | 2500 | 4100 |
| 15 | 114 | 67 | 12996 | 4489 | 7638 |
| 16 | 95 | 51 | 9025 | 2601 | 4845 |
| 17 | 94 | 55 | 8836 | 3025 | 5170 |
| 18 | 103 | 54 | 10609 | 2916 | 5562 |
| 19 | 96 | 57 | 9216 | 3249 | 5472 |
| 20 | 104 | 54 | 10816 | 2916 | 5616 |
| 21 | 98 | 52 | 9604 | 2704 | 5096 |
| 22 | 117 | 48 | 13689 | 2304 | 5616 |
| 23 | 100 | 61 | 10000 | 3721 | 6100 |
| 24 | 101 | 54 | 10201 | 2916 | 5454 |
| 25 | 94 | 48 | 8836 | 2304 | 4512 |
| Sum: | 2501.00 | 1378.00 | 252163 | 76844 | 138451 |
| Mean: | 100.04 | 55.12 | | | |

Table 11.2: Illustration of computations for correlation between IQ and reading ability.

The formula for $r$ uses six quantities:

**$N$:** The number of cases in the data set; in this example, 25.

**$\bar{X}$:** The sample mean of the $X$ values; in this example, 100.04.

**$\bar{Y}$:** The sample mean of the $Y$ values; in this example, 55.12.

**$\sum_{i=1}^{N} {X_i}^2$:** The sum of the squared $X$ values; in this example, 252,163.

**$\sum_{i=1}^{N} {Y_i}^2$:** The sum of the squared $Y$ values; in this example, 76,844.

$\sum_{i=1}^{N} X_i \times Y_i$**:** The sum of the products obtained by multiplying the $X$ and $Y$ values for each case; in this example, 138,451.

These values are plugged into the formula as follows:

$$\begin{aligned} r &= \frac{138451.00 - 25 \times 100.04 \times 55.12}{\sqrt{(252163.00 - 25 \times 100.04^2) \times (76844.00 - 25 \times 55.12^2)}} \\ &= 0.451 \end{aligned}$$

The value of $r$ computed from a sample must always lie between -1 and +1, as a mathematical consequence of the formula for $r$. To help give you a feel for the meaning of different values of $r$, Figure 11.3 displays scattergrams of samples with various different correlations. As you can see, a sample correlation of $r = 0$ (upper left corner) indicates that the two variables are unrelated. For larger values of $r$, however, there is a positive linear relationship between the two variables. With $r = 0.2$, the relationship is difficult to see, but with $r = 0.4$ or greater you can see the trend increasingly well. Note that $r = +1$ (lower right corner) indicates a perfect linear relationship (i.e., the points lie exactly on a straight line).

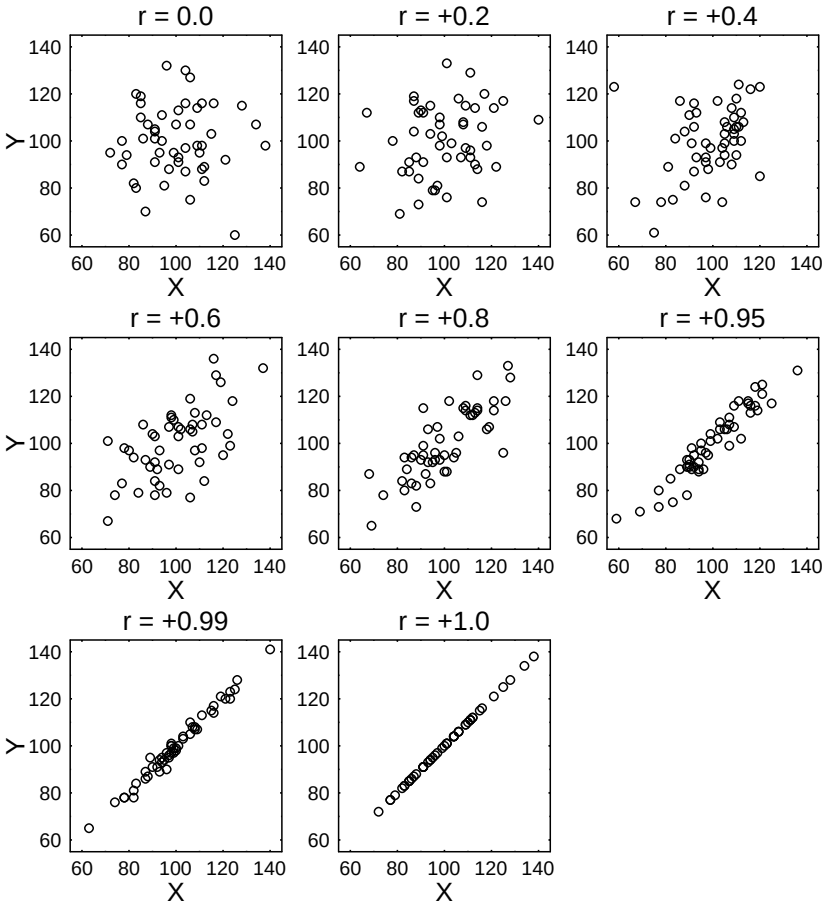


Figure 11.3: Scattergrams displaying samples with different positive correlation ($r$) values.

Figure 11.4 displays scattergrams of samples with various negative correlations. The relationships are exactly analogous to those shown in Figure 11.3 except that the linear trends are negative, going from upper left to lower right.

Some intuition about the formula for $r$ can be gained by looking at the following alternative version of the formula, which is mathematically equivalent but computationally less convenient:

$$r = \frac{\sum_{i=1}^{N} \left[ \left( X_i - \bar{X} \right) \times \left( Y_i - \bar{Y} \right) \right]}{N \times s_x \times s_y}$$

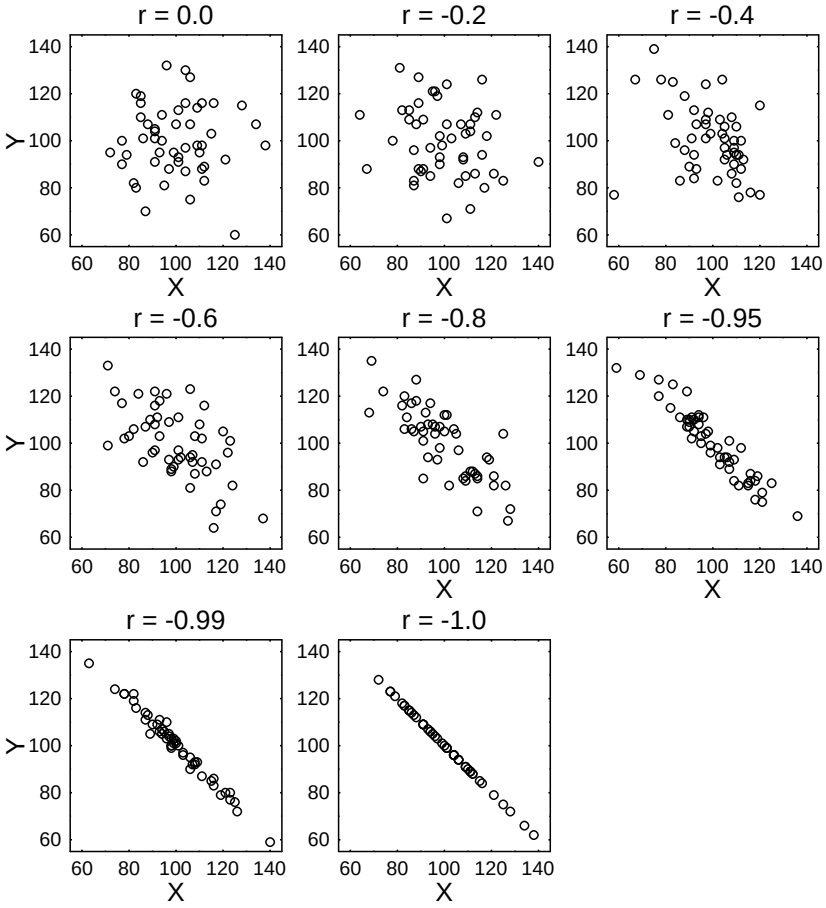


Figure 11.4: Scattergrams displaying samples with different negative correlation ($r$) values.

Look at the numerator of this fraction. It is a sum across cases of the products $\left(X_i - \bar{X}\right) \times \left(Y_i - \bar{Y}\right)$. Note that a case will produce a positive product if both $X_i$ and $Y_i$ are above their respective means, and also if both $X_i$ and $Y_i$ are below their respective means (in the latter case, the product is a negative times a negative, which equals a positive). With positive relationships (e.g., Figure 11.3), $X_i$ and $Y_i$ tend to be both small or both large in the same cases, so the products all tend to be positive, leading to a large positive numerator.

On the other hand, a case will produce a negative product if $X_i$ is above its mean and $Y_i$ is below its mean, or vice versa. In both of these cases, the product is a positive times a negative, which yields a negative result. This is common with negative relationships (e.g., Figure 11.4), because here a large value of $X_i$ tends to occur together with a small value of $Y_i$, and vice versa. In negative relationships, then, the sum in the numerator is a negative number.

Overall, then, the sign of the numerator is determined by whether there is an overall positive or negative relationship between $X$ and $Y$. In the denominator of the formula, $s_x$ and $s_y$ are the standard deviations of $X$ and $Y$, respectively, so the denominator of the fraction is always positive. Thus, the sign of the numerator also determines the overall sign of $r$ itself.

## 11.4 Testing the Null Hypothesis of No Correlation

The value of $r$ is the sample correlation, so it may be influenced by sampling error. That is, just as the value of $F_{observed}$ varies randomly from one experiment to the next (review section 3.7), so too does the value of $r$ vary randomly depending on which particular individuals happened to be included in the sample. The next problem, then, is to make an inference about the true value of the correlation within the population as a whole, known as "rho" ($\rho$).

We will only consider the problem of testing the null hypothesis $\rho = 0$, which is the most common

situation. According to this null hypothesis, there is no linear relationship between the two variables; that is, as one variable increases, the other remains constant, on average. If the true correlation is zero, then the observed value of $r$ should presumably be close to zero too. So, if $r$ is sufficiently far from zero, we should reject the null hypothesis that $\rho = 0$.

How far from zero is sufficient? Leaving aside all the mathematical details, the answer is simple. Look up the critical value of $r$ in appendix B.2. Normally, you look in the column corresponding to the $p$ level of .05, although other columns can be consulted if needed. Look down to the row corresponding to the number of cases in the sample, and the entry for that row and column is the "$r_{\text{critical}}$." The null hypothesis should be rejected if the $r$ value computed from the sample is further from zero than $r_{\text{critical}}$ (i.e., reject if $r > r_{\text{critical}}$ or $r < -r_{\text{critical}}$).

With the data of Table 11.1, for example, there are 25 cases, so the critical value from appendix B.2 is $r_{\text{critical}} = .396$ with a significance level of $p = .05$. Given that the observed $r = 0.451$ computed from the sample is greater than $r_{\text{critical}} = 0.396$, you can reject the null hypothesis in this example. (The appropriate conclusion is discussed in the next section.)

If you scan down the .05 column of the table of $r_{\text{critical}}$ values in appendix B.2, you will see that even a pretty small correlation can be statistically significant if the number of cases is large enough. Figure 11.5 makes this same point graphically. Each of the different panels in the figure shows a data sets in which the correlation is just strong enough to be significant. For a small number of cases ($N = 10$), the correlation has to be quite strong to be significant ($r = .63$), and you can easily see the increasing trend from the lower left to the upper right. With $N = 20$ or 40, a smaller $r$ is needed to give a significant result (i.e., $r = .44$ or .31), but you can still probably see the increasing trend. When the sample gets large enough, though, a statistically significant result may be found even when the trend is too weak to see. You can look at this fact in either of two ways:

- The correlation technique is so sensitive that it can detect really subtle relationships.
- With a large sample, a statistically significant result may be so weak as to be nearly useless for predictive purposes.

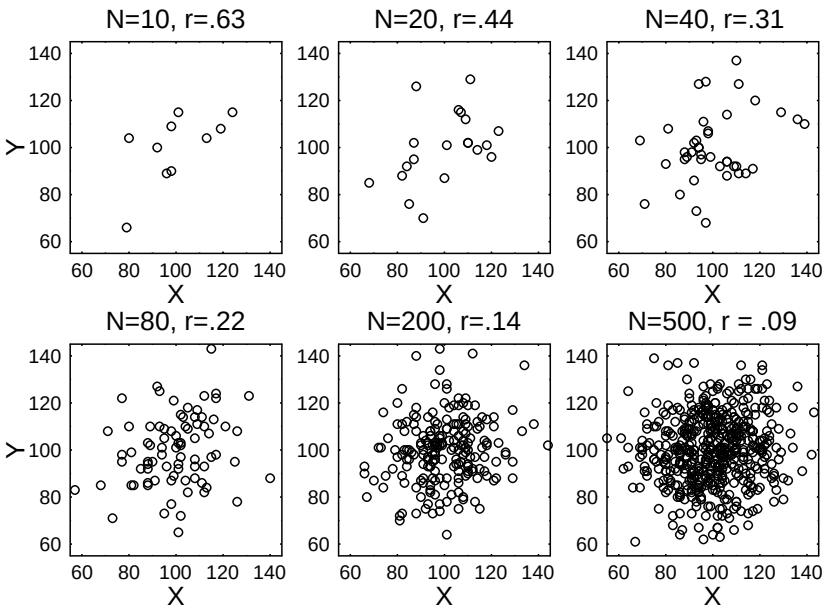


Figure 11.5: Scattergrams displaying samples with correlation ($r$) values significant at $p < .05$.

## 11.5 Conclusions from Significant and Nonsignificant Correlations

### 11.5.1 Significant Correlations

When a sample correlation is significant, the null hypothesis of no linear association can be rejected. You can thus conclude that there is an increasing trend (if $r$ is positive) or a decreasing trend (if $r$ is negative) beyond that explainable by chance.

The exact phrasing of this conclusion depends on whether the $X$ values were randomly assigned to cases or not. Normally, they are not, and the conclusion should be stated this way for a positive $r$.

> The data demonstrate that larger values of $X$ are associated with larger values of $Y$. This association is consistent with the idea that larger values of $X$ cause increases in $Y$, but it does not demonstrate the causal relationship because the causality could go in the opposite direction or some third variable could be responsible for the association by simultaneously causing values on both $X$ and $Y$.

For a negative $r$, the conclusion is the same except you say "are associated with *smaller* values of $Y$" and "cause *decreases* in $Y$."

If $X$ values were randomly assigned to cases, then a stronger conclusion is justified, as stated this way for a positive $r$.

> The data demonstrate that larger values of $X$ cause increases in $Y$.

For a negative $r$, the conclusion is the same except you say "cause *decreases* in $Y$."

Using the data of Table 11.1, for example, we found a significant correlation of $r = 0.451$ between Abil and IQ. Since IQ values were not randomly assigned to cases, the conclusion is:

> The data demonstrate that larger IQs are associated with greater reading ability. This association is consistent with the idea that larger IQs cause better reading ability, but it does not demonstrate the causal relationship because the causality could go in the opposite direction or some third variable could be responsible for the association by causing values on both reading ability and IQ. For example, it could be that an intellectually enriched environment causes both high IQ and good reading ability.

It may also be illuminating to consider a situation where $X$ values *were* assigned randomly to cases, because this helps to clarify the nature of the "causal" interpretations supported by this type of analysis. Suppose we conduct a study in which a sample of pets are randomly assigned to owners, with random assignment carried out by us as experimenters. We look for a relation between the owner's IQ ($X$) and the pet's longevity ($Y$), and suppose we do find that pets with higher-IQ owners do tend to live longer (i.e., a significant correlation). In that case we could legitimately conclude

> The data demonstrate that a pet owner having a higher IQ tends to cause the pet to live longer.

Now you can probably think of a number of different "causal mechanisms" that might be responsible for this pattern. It could be that owners with higher IQs are more knowledgeable about their pets and therefore take better care of them. That would be a pretty direct causal connection between high owner IQs and increased pet longevity. Alternatively, perhaps owners with higher IQs simply tend to be more affluent and therefore to live in less crowded suburb environments where their pets are less likely to get hit by cars. That would be a pretty indirect causal connection, and you might be skeptical about someone claiming that the owner's IQ had much to do with it. Either way (direct or indirect), though, it would be true that randomly assigning a pet to a higher-IQ owner would tend to raise its life expectancy—that is exactly the sense of "causality" under examination with this type of statistical analysis.

### 11.5.2 Nonsignificant Correlations

When a correlation is not significant, the null hypothesis of no linear association may be true. The interpretation in that case would go something like this:

> The data do not demonstrate any tendency for larger values of $X$ to be consistently associated with larger or smaller values of $Y$. This is consistent with the claim that $X$ and $Y$ are causally unrelated to one another. It does not prove that claim, however, for a variety of reasons. For example, the causal relation may be so weak that a larger sample size is needed to demonstrate it.

(Chapter 13 discusses some other possible reasons why you might not get a significant correlation even when there is a causal connection between two variables.)

In the data of Table 11.1, for example, the correlation between IQ and Home is only 0.20, which is not significant. The interpretation of this correlation would therefore be something like this:

> The data do not demonstrate any tendency for larger values of IQ to be associated with larger or smaller amounts of time spent reading in the home. This is consistent with the claim that IQ and home reading are causally unrelated to one another. It does not prove that claim, however, for a variety of reasons. For example, the causal relation may be so weak that a larger sample size is needed to demonstrate it.

## 11.6 The Correlation Matrix

Strictly speaking, the scattergram and correlation procedures discussed in this chapter are bivariate techniques, because they consider only two variables at a time. Nonetheless, these procedures can be used to summarize data sets with more than two variables—it's just that you need lots of summaries and each applies to only two variables at a time.

As a graphical illustration, Figure 11.6 shows a set of six scattergrams, each of which displays the relationship between two of the variables in Table 11.1. These scattergrams allow you to examine each of the possible relations you might want to look at, and thus convey much of the information about the relationships among variables.

Table 11.3 shows an analogous summary of the data known as a "correlation matrix," in two slightly different versions. In each version, the table shows the correlation between each pair of variables in the data set, just as Figure 11.6 shows the scattergram for each pair. To find a correlation between two variables, look in the row corresponding to one of the variables and the column corresponding to the other. The number at the intersection of that row and that column is the correlation between those two variables. The values on the diagonal have to be 1.00, because the correlation between a variable and itself must always be one.

Note that Version 1 of the table has more numbers than Version 2 but really conveys no extra information. Version 1 lists each correlation twice (e.g., the correlation of 0.45 between Abil and IQ is listed in both the second column of the Abil row and in the first column of the IQ row), whereas Version 2 lists each correlation once.

| | Version 1 | | | | | Version 2 | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Abil | IQ | Home | TV | | Abil | IQ | Home | TV |
| Abil | 1.00 | 0.45 | 0.74 | -0.29 | Abil | 1.00 | 0.45 | 0.74 | -0.29 |
| IQ | 0.45 | 1.00 | 0.20 | 0.25 | IQ | | 1.00 | 0.20 | 0.25 |
| Home | 0.74 | 0.20 | 1.00 | -0.65 | Home | | | 1.00 | -0.65 |
| TV | -0.29 | 0.25 | -0.65 | 1.00 | TV | | | | 1.00 |

Table 11.3: Two versions of a correlation matrix showing the correlations between all pairs of variables in Table 11.1.

Although it may at first appear that all the correlations in a correlation matrix are independent of one another, that is not actually true. As an example, suppose that $X_1$ is perfectly correlated with $X_2$ ($r = 1$) and that $X_2$ is perfectly correlated with $X_3$. In that case, it is a mathematical requirement that $X_1$ and $X_3$ are perfectly correlated as well. In fact, the exact mathematical constraints among the different correlations in a correlation matrix are too complex to specify exactly here, but it is worth knowing that not all combinations of numbers between -1 and +1 are possible within a correlation matrix.

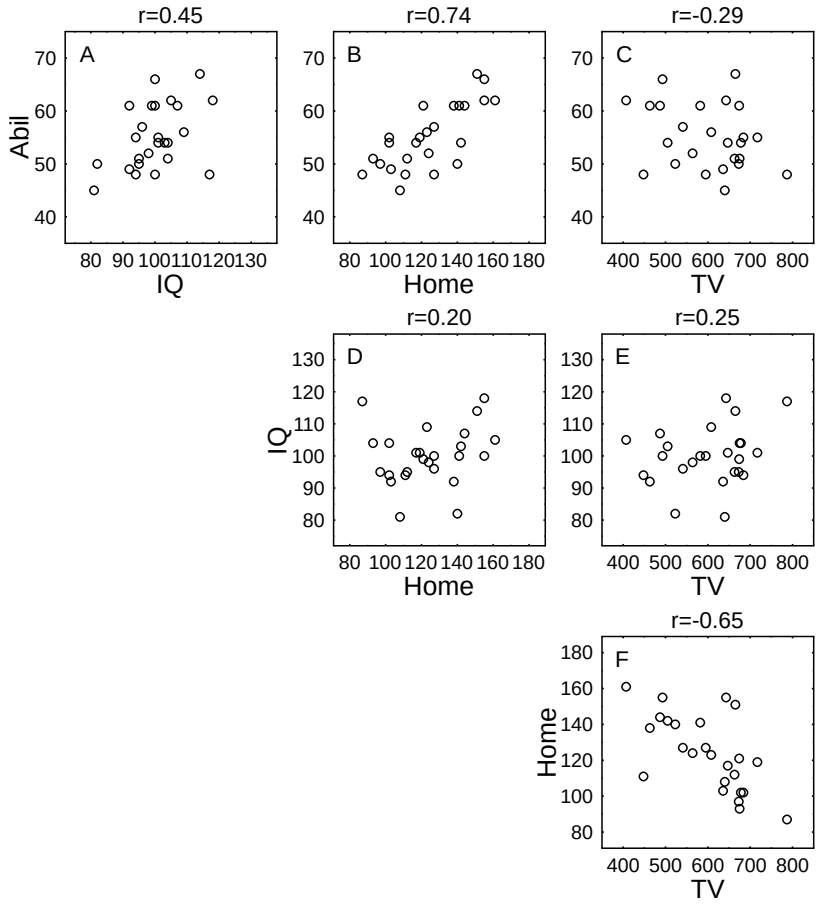


Figure 11.6: Scattergrams displaying the relationships between all pairs of variables in Table 11.1.

# Chapter 12

# Simple Regression

Like correlation, regression is used to detect and quantify relationships among numerical variables. It also uses data of the "case by variable" format introduced in Chapter 10, and for our illustrative example we will continue using the reading ability data of Table 11.1. Other similarities between regression and correlation—as well as some important differences—will become apparent as we proceed.

## 12.1 The Simple Regression Model

Equation 12.1 shows the version of the general linear model used in simple regression analysis.

$$Y_i = a + b \times X_i + e_i \tag{12.1}$$

$Y$ is the dependent variable. As in ANOVA, it is the variable that the researcher wants to learn about. In the reading ability data of Table 11.1, for example, reading ability (Abil) would most likely be the $Y$ variable, because the researcher is mainly interested in finding out what influences this variable. In regression analysis, the $Y$ variable is usually called the "to-be-predicted" variable, because the model can be used to make predictions about $Y$. The subscript $i$ is needed on $Y_i$ to distinguish among the different cases, because each case has its own value of $Y$, just as in correlation analysis.

$X$ is the independent variable—the variable to which $Y$ might be related. In the reading ability data of Table 11.1, for example, you might choose any of the other variables (IQ, Home, or TV) as the $X$ variable, and the analysis would then reveal how Abil was related to whichever independent variable you chose. The $X$ variable is called the "predictor" variable, because this variable is used to compute predicted values of $Y$. Again, the subscript $i$ is used to distinguish among the different cases. Note that in the present *simple regression* model there is only one $X$ variable; in *multiple regression* (see Chapter 14), there are two or more predictor variables (i.e., $X$'s).

As illustrated in Figure 12.1, the parameters $a$ and $b$ quantify the linear relationship—if any—between $X$ and $Y$[1]. The parameter $b$ is called the "slope" of the regression equation. As shown in Figure 12.1, $b$ determines how $Y$ changes with changes in $X$. In the upper left panel, for example, $Y$ increases relatively gradually as $X$ increases. Specifically, there is a 0.5 point increase in $Y$ for each 1-point increase in $X$, which is easiest to see by examining the ends of the lines. The middle line of the panel, for example, starts at the point with $X = 0$ and $Y = 0$, and it increases to $X = 80$ and $Y = 40$—i.e., $Y$ increases half as fast as $X$. In the upper right panel, on the other hand, $Y$ increases 2 points for each 1-point increase in $X$, which is a relatively steep increase. When $b$ is negative, as shown in the lower panels, $Y$ tends to decrease as $X$ increases, so the line goes down from left to right. Again, the change in $Y$ is more gradual when $b$ is closer to zero.

The parameter $a$ is called the "Y-intercept" or simply "intercept" of the regression equation. More specifically, $a$ is the value of $Y$ that you predict when $X$ equals zero, because $a + b \times 0 = a$. As you can see from Figure 12.1, changing $a$ shifts the whole line up or down on the graph. The parameter $a$ is also sometimes called the "regression constant" or just "constant", because it is a constant value added in to all the $Y$ values.

In real data, of course, the points do not usually lie exactly on a straight line as they do in the examples shown in Figure 12.1. Instead, there is almost always some variation around the line, as

[1] Recall that the equation for a straight line is

$$Y = a + b \times X$$

which is almost the same as the general linear model for simple regression.

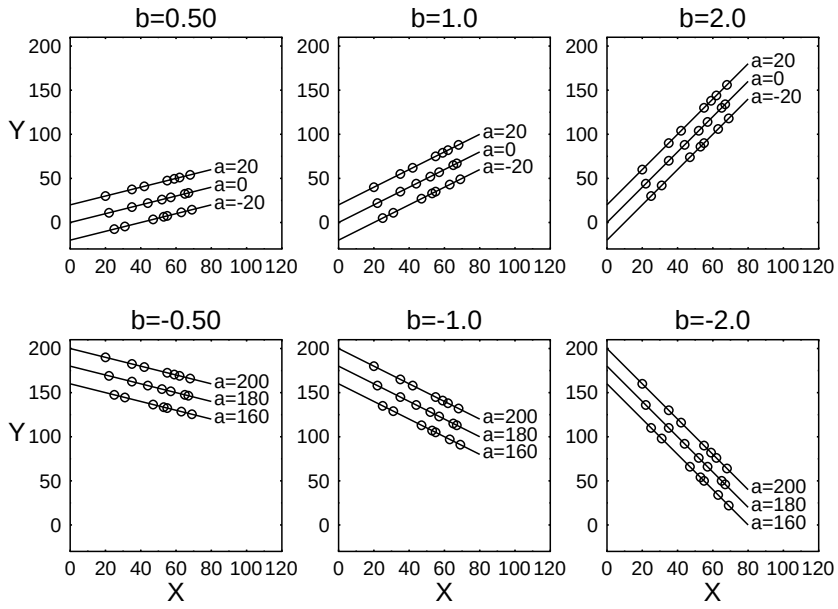


Figure 12.1: Illustration of how the intercept ($a$) and slope ($b$) values determine a straight line. Each line shows the points consistent with the equation $Y = a + b \times X$ for the indicated values of $a$ and $b$. The line's value of $a$ is shown next to it, and the value of $b$ is shown at the top of each panel ($b$ is the same for all the lines within one panel). For example, the equation of the top line in the upper panel on the left is $Y = 20 + 0.5 \times X$.

shown in Figure 12.2. Within the model, this variation is attributed to the random error component, $e_i$. In other words, this term of the model reflects the effects of random sampling and measurement error on the $Y_i$ values. $e_i$ is also subscripted by $i$, because the model allows each $Y_i$ to have its own unique error component. Graphically, the value of $e_i$ is the vertical distance from the point on the scattergram to the best-fitting straight line, as shown by the dotted lines in Figure 12.2.

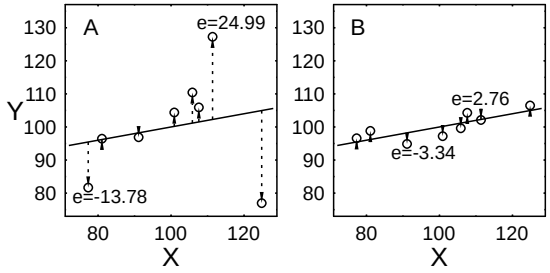


Figure 12.2: Illustration of the error component in a regression equation. The points represent data from eight cases, and the solid line is the regression line through those data. For each point, the error, $e_i$, is the vertical distance from the point to the regression line, as indicated by the dotted line. Panel A shows a data set for which the errors are large, and panel B shows a data set for which they are small.

Finally, it should be emphasized that $b$ is the most important parameter of the model for purposes of hypothesis testing. When $Y$ is unrelated to $X$ (i.e., their correlation is zero), then $b$ equals zero, as shown in Figure 12.3. In this case the regression model reduces to:

$$\begin{aligned} Y_i &= a + 0 \times X_i + e_i \\ &= a + e_i \end{aligned}$$

so you can see that $X$ effectively drops out of the equation. In essence, when $Y$ is not related to $X$,

the estimated value of the Y-intercept, $\hat{a}$, turns out to be simply the mean value of $Y$, so you always predict $Y$ to be at its mean value, regardless of what $X$ is. This makes intuitive sense: If $X$ tells you nothing about $Y$, you might as well predict $Y$ to be at its mean and ignore $X$. Thus, the question "Is $Y$ correlated with $X$?" is the same as the question "Is $b$ different from zero?" The null hypothesis is that $b = 0$ (i.e., there is no relation), and the general linear model provides a statistical test to see whether the data allow this hypothesis to be rejected.

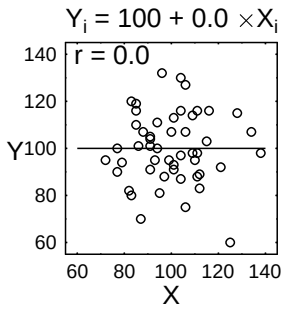


Figure 12.3: Scattergram and best-fitting regression line for a sample with a zero correlation between $X$ and $Y$. The slope $b$ is zero and so the term $b \times X_i$ effectively drops out of the model. The estimate of $a$ is equal to the mean $Y$.

## 12.2 Fitting the Model to the Data

To use the regression model, you must estimate values for the terms in the model and then divide up sums of squares and degrees of freedom, just as you did in ANOVA. This section describes the necessary computations in a step-by-step fashion, and Table 12.1 provides an example illustrating the computations. For this example, we will try to predict Abil from IQ using the data of Table 11.1.

**Step 1:** Estimate the value of $b$. The estimate is called $\hat{b}$, and the formula for it is

$$\hat{b} = \frac{\sum_{i=1}^{N} X_i \times Y_i - N \times \bar{X} \times \bar{Y}}{\sum_{i=1}^{N} {X_i}^2 - N \times \bar{X}^2}$$

For the data of Table 12.1, the computation is

$$\begin{aligned} \hat{b} &= \frac{138451 - 25 \times 100.04 \times 55.12}{252163 - 25 \times 100.04^2} \\ &= 0.30 \end{aligned}$$

**Step 2:** Estimate the value of $a$. The estimate is called $\hat{a}$, and the formula for it is

$$\hat{a} = \bar{Y} - \hat{b} \times \bar{X}$$

For the data of Table 12.1, the computation is

$$\begin{aligned} \hat{a} &= 55.12 - 0.30 \times 100.04 \\ &= 24.75 \end{aligned}$$

**Step 3:** For each case, estimate the "predicted $Y$ value" denoted $Y_i'$. The predicted value $Y_i'$ is what the $Y_i$ value should have been, according to the model, if there had been zero error for that case. The formula for the predicted value, $Y_i'$, is

$$Y_i' = \hat{a} + \hat{b} \times X_i$$

Table 12.1 shows how $Y'$ is computed for each case: The actual value of $X_i$ for that case is multiplied by $\hat{b}$, and $\hat{a}$ is added to it. A useful computational check is that the sum of the predicted values, $\sum_{i=1}^{N} Y_i'$, should equal the sum of the actual $Y_i$ values, $\sum_{i=1}^{N} Y_i$, except for a small difference due to rounding error.

**Step 4:** For each case, estimate the error. The error estimate is denoted $\hat{e}_i$, and the formula for it is

$$\begin{aligned} \hat{e}_i &= Y_i - Y_i' \\ &= Y_i - \hat{a} - \hat{b} \times X_i \end{aligned}$$

Thus, the error score is simply the difference between the actual observed $Y_i$ value and the predicted value, $Y_i'$. Table 12.1 also shows the computation of error for each case. A useful computational check is that the error scores must sum to zero across all the cases, except for some small rounding error, just as the estimated error scores did in ANOVA.

| Case | IQ | Abil | IQ×Abil | Predicted $= \hat{a} + \hat{b} \times X_i$ | | Error=Abil-Predicted | |
|---|---|---|---|---|---|---|---|
| 1 | 107 | 61 | 6527 | 56.85 | $= 24.75 + 0.30 \times 107$ | 4.15 | $= 61 - 56.85$ |
| 2 | 109 | 56 | 6104 | 57.45 | $= 24.75 + 0.30 \times 109$ | -1.45 | $= 56 - 57.45$ |
| 3 | 81 | 45 | 3645 | 49.05 | $= 24.75 + 0.30 \times 81$ | -4.05 | $= 45 - 49.05$ |
| 4 | 100 | 66 | 6600 | 54.75 | $= 24.75 + 0.30 \times 100$ | 11.25 | $= 66 - 54.75$ |
| 5 | 92 | 49 | 4508 | 52.35 | $= 24.75 + 0.30 \times 92$ | -3.35 | $= 49 - 52.35$ |
| 6 | 105 | 62 | 6510 | 56.25 | $= 24.75 + 0.30 \times 105$ | 5.75 | $= 62 - 56.25$ |
| 7 | 92 | 61 | 5612 | 52.35 | $= 24.75 + 0.30 \times 92$ | 8.65 | $= 61 - 52.35$ |
| 8 | 101 | 55 | 5555 | 55.05 | $= 24.75 + 0.30 \times 101$ | -0.05 | $= 55 - 55.05$ |
| 9 | 118 | 62 | 7316 | 60.15 | $= 24.75 + 0.30 \times 118$ | 1.85 | $= 62 - 60.15$ |
| 10 | 99 | 61 | 6039 | 54.45 | $= 24.75 + 0.30 \times 99$ | 6.55 | $= 61 - 54.45$ |
| 11 | 104 | 51 | 5304 | 55.95 | $= 24.75 + 0.30 \times 104$ | -4.95 | $= 51 - 55.95$ |
| 12 | 100 | 48 | 4800 | 54.75 | $= 24.75 + 0.30 \times 100$ | -6.75 | $= 48 - 54.75$ |
| 13 | 95 | 50 | 4750 | 53.25 | $= 24.75 + 0.30 \times 95$ | -3.25 | $= 50 - 53.25$ |
| 14 | 82 | 50 | 4100 | 49.35 | $= 24.75 + 0.30 \times 82$ | 0.65 | $= 50 - 49.35$ |
| 15 | 114 | 67 | 7638 | 58.95 | $= 24.75 + 0.30 \times 114$ | 8.05 | $= 67 - 58.95$ |
| 16 | 95 | 51 | 4845 | 53.25 | $= 24.75 + 0.30 \times 95$ | -2.25 | $= 51 - 53.25$ |
| 17 | 94 | 55 | 5170 | 52.95 | $= 24.75 + 0.30 \times 94$ | 2.05 | $= 55 - 52.95$ |
| 18 | 103 | 54 | 5562 | 55.65 | $= 24.75 + 0.30 \times 103$ | -1.65 | $= 54 - 55.65$ |
| 19 | 96 | 57 | 5472 | 53.55 | $= 24.75 + 0.30 \times 96$ | 3.45 | $= 57 - 53.55$ |
| 20 | 104 | 54 | 5616 | 55.95 | $= 24.75 + 0.30 \times 104$ | -1.95 | $= 54 - 55.95$ |
| 21 | 98 | 52 | 5096 | 54.15 | $= 24.75 + 0.30 \times 98$ | -2.15 | $= 52 - 54.15$ |
| 22 | 117 | 48 | 5616 | 59.85 | $= 24.75 + 0.30 \times 117$ | -11.85 | $= 48 - 59.85$ |
| 23 | 100 | 61 | 6100 | 54.75 | $= 24.75 + 0.30 \times 100$ | 6.25 | $= 61 - 54.75$ |
| 24 | 101 | 54 | 5454 | 55.05 | $= 24.75 + 0.30 \times 101$ | -1.05 | $= 54 - 55.05$ |
| 25 | 94 | 48 | 4512 | 52.95 | $= 24.75 + 0.30 \times 94$ | -4.95 | $= 48 - 52.95$ |
| Sum: | 2501 | 1378 | 138451 | 1378.00 | | 0 | |
| Mean: | 100.04 | 55.12 | | 55.12 | | 0 | |
| SS: | 252163 | 76844 | | 76133.02 | | 710.98 | |

Table 12.1: Illustration of computations for simple regression model predicting Abil from IQ using the data of Table 11.1. The "SS" in the bottom line of the table stands for "sum of squares".

## 12.3 ANOVA Table for Simple Regression

After fitting the simple regression model, we summarize the computations in an ANOVA table with sums of squares and degrees of freedom. This table is then used to test the null hypothesis that $b = 0$. Perhaps surprisingly, we do *not* compute the sums of squares from a decomposition matrix in regression analysis as we did in ANOVA. The reasons for this are more mathematical and technical than conceptual, so we will not emphasize them[2].

[2]But for the interested reader, here is a brief sketch of the problem with the decomposition matrix for regression: Although we never explicitly said so, the decomposition matrix for ANOVA only worked because the sum of the estimated values was zero for every term except the baseline, $\mu$ (e.g., the $\hat{A}_i$'s had to sum to zero, so did the $\hat{B}_j$'s, the $\widehat{AB}_{ij}$'s, and so on). In the standard regression model, there are two terms for which the estimated values do not usually sum to zero: $\hat{a}$ and $\hat{b} \times X_i$. Because there are two terms like this instead of only one, the decomposition breakdown does not correspond to sums of squares as it did in ANOVA. It would be possible to rewrite the general linear model for simple regression in a slightly more complicated form for which the decomposition matrix would work; namely, $Y_i = \mu + b \times (X_i - \bar{X})$. This would depart from standard practice, however.

Table 12.2 shows the most general version of the "long-form" ANOVA table for simple regression. It is not as common as a shortened version discussed below, but it is presented here because it is most closely analogous to the ANOVA tables that were used in previous chapters.

| Source | *df* | *SS* | *MS* | *F* |
|---|---|---|---|---|
| $a$ | 1 | $N \times \bar{Y}^2$ | $SS_a/df_a$ | $MS_a/MS_e$ |
| $b$ | 1 | $\sum_{i=1}^{N} Y_i'^2 - SS_a$ | $SS_b/df_b$ | $MS_b/MS_e$ |
| Error | $N-2$ | $\sum_{i=1}^{N} e_i^2$ | $SS_{error}/df_e$ | |
| Total | $N$ | $\sum_{i=1}^{N} Y_i^2$ | | |

Table 12.2: General version of the long-form ANOVA table for simple regression.

In this ANOVA table, the total degrees of freedom is the number of $Y$ values being analyzed, $N$, just as it was in ANOVA. $a$ and $b$ each have one degree of freedom, because each is a single free parameter that is estimated from the data (analogous to the $\mu$ term in ANOVA). That leaves $N-2$ degrees of freedom for the error term. Now we saw that there is an error estimate, $\hat{e}_i$, for each case, which means that we estimated $N$ values for this term. We also saw that these had to add up to zero, which provides a constraint that takes away one degree of freedom. There is also one more subtle constraint, which is why the $\hat{e}$'s have $N-2$ degrees of freedom rather than $N-1$. This constraint is that *there must be a zero correlation between the $X_i$ values and the $\hat{e}_i$ values.* We will not prove this assertion mathematically, but Figure 12.4 illustrates it graphically. For this figure, the constraint says that the $\hat{e}$ values must not consistently increase or decrease with $X$, and indeed they do not. Intuitively, the $\hat{e}_i$ values have to add up to zero for both relatively small $X$'s (left side of the figure) and for relatively large $X$'s (right side of the figure). This is guaranteed to happen because $\hat{b}$ is the optimal slope.

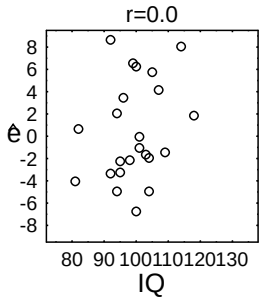


Figure 12.4: A scattergram illustrating the fact that the estimated error scores, $\hat{e}_i$, are uncorrelated with the $X_i$ values used for prediction. The 25 data points correspond to the 25 cases of Table 12.1. Each case is plotted according to its IQ and the value of $\hat{e}_i$ computed for that case (rightmost column of Table 12.1). Note that the scattergram displays a zero correlation between these two variables.

Table 12.2 also shows the general versions of the formulas for the sums of squares in simple regression. Three of these SS formulas should be familiar from the analysis of variance:

$$\begin{aligned} SS_{total} &= \sum_{i=1}^{N} {Y_i}^2 \\ SS_{error} &= \sum_{i=1}^{N} {\hat{e}_i}^2 \\ SS_a &= N \times \bar{Y}^2 \end{aligned}$$

The total sum of squares is the sum of the squared data values, and the sum of squares for error is the sum of the squared error estimates. The $SS_a$ is the same as the $SS_\mu$ in analysis of variance, namely $N$ times the square of the mean $Y$.

The new sum of squares is the one for $b$, $SS_b$, and it can be computed in either of two ways:

$$\begin{aligned} SS_b &= \sum_{i=1}^{N} {Y'_i}^2 - SS_a = \sum_{i=1}^{N} \left(Y'_i - \bar{Y}\right)^2 \\ &= SS_{total} - SS_a - SS_{error} \end{aligned}$$

The upper formula emphasizes the connection between this sum of squares and the predictive model: $SS_b$ is the sum of the squared predictions minus what is due to the baseline ($SS_a$). The lower formula is probably easier to use in practice. Once you have computed $SS_{total}$, $SS_a$, and $SS_{error}$, it is easiest to get $SS_b$ by subtraction.

Using the values in Table 12.1, the required quantities are computed as follows:

$$\begin{aligned} df_a &= 1 \\ df_b &= 1 \\ df_{error} &= 25 - 2 = 23 \\ df_{total} &= 25 \\ SS_a &= 25 \times 55.12^2 = 75955.36 \\ SS_b &= 76133.02 - 25 \times 55.12^2 = 177.66 \\ SS_{error} &= 710.98 \\ SS_{total} &= 76844 \\ MS_a &= \frac{75955.36}{1} = 75955.36 \\ MS_b &= \frac{177.66}{1} = 177.66 \\ MS_{error} &= \frac{710.98}{23} = 30.91 \\ F_a &= \frac{75955.36}{30.91} = 2457.31 \\ F_b &= \frac{177.66}{30.91} = 5.75 \end{aligned}$$

These computations are summarized more compactly in the summary regression ANOVA table in Table 12.3.

| Source | *df* | *SS* | *MS* | *F* |
|---|---|---|---|---|
| *a* | 1 | 75955.36 | 75955.36 | 2457.31 |
| *b* | 1 | 177.66 | 177.66 | 5.75 |
| Error | 23 | 710.98 | 30.91 | |
| Total | 25 | 76844.00 | | |

Table 12.3: Long-form regression ANOVA table for predicting Abil from IQ using the data and computations of Table 12.1.

Table 12.4 shows a more common and somewhat abbreviated "short-form" version of the ANOVA table for simple regression. This is the version found in most textbooks and provided by most computer packages. This form is changed in two ways compared to the long form:

1. The line for $a$ is omitted. As discussed below, this line is rarely of any interest, so there is little loss in omitting it. In case it is needed for prediction, though, the estimated value of $\hat{a}$ is always printed out somewhere. It is often called the "regression constant", and generally appears either in an explicit equation or in a table of parameters ($a$, $b$) and their estimates.

2. The "Total" line now represents what should properly be called a "Corrected Total." Note that it has only $N - 1$ degrees of freedom, whereas in fact there are truly $N$ total degrees of freedom in the data. The lost degree of freedom represents a correction for the baseline, which has effectively been "taken away" from the data. Conceptually, it is as if someone subtracted $\bar{Y}$ from all the $Y$ scores so that these scores were forced to add up to zero. The corrected total sum

of squares has also been reduced by subtracting out $SS_a$. This too represents what the sum of squares would have been if $\bar{Y}$ had been subtracted from all of the $Y$s.

| Source | *df* | *SS* | *MS* | *F* |
|---|---|---|---|---|
| $b$ | 1 | $\sum_{i=1}^{N} Y_i'^2 - SS_a$ | $SS_b/df_b$ | $MS_b/MS_e$ |
| Error | $N-2$ | $\sum_{i=1}^{N} e_i^2$ | $SS_{error}/df_e$ | |
| Corrected Total | $N-1$ | $\sum_{i=1}^{N} Y_i^2 - SS_a$ | | |

Table 12.4: General version of the short-form ANOVA table for simple regression.

Computations for the short-form regression ANOVA table generally use the same equations as those for the long-form table, as illustrated already for the data of Table 12.1. The two new values are the $df_{corrected\ total}$ and $SS_{corrected\ total}$, computed as follows for these data:

$$\begin{aligned} df_{corrected\ total} &= 25 - 1 = 24 \\ SS_{corrected\ total} &= 76844 - 25 \times 55.12^2 = 888.64 \end{aligned}$$

The short-form computations for the data of Table 12.1 are summarized in Table 12.5. Note that the $F$ for $b$ is the identical in both the short and long versions of the table, so it does not matter which one you use for testing the null hypothesis "$H_0$: $b = 0$".

| Source | *df* | *SS* | *MS* | *F* |
|---|---|---|---|---|
| $b$ | 1 | 177.66 | 177.66 | 5.75 |
| Error | 23 | 710.98 | 30.91 | |
| Corrected Total | 24 | 888.64 | | |

Table 12.5: Short-form regression ANOVA table for predicting Abil from IQ using the data and computations of Table 12.1.

## 12.4 Critical $F$'s and Conclusions

As already mentioned, the most important term in the regression model is $b$, because this term captures the relationship of $Y$ to $X$. The null hypothesis is:

$$H_0 : b = 0$$

and it is tested by the observed $F$ on the $b$ line of the regression ANOVA table. This $F_{observed}$ is compared against a critical $F$ with 1 degree of freedom in the numerator and $N-2$ degrees of freedom for error, corresponding to the MS's in the numerator and denominator of the $F$ computation. As usual, the null hypothesis is rejected if $F_{observed} > F_{critical}$.

### 12.4.1 Significant Slope

When a slope is significant, the conclusion is essentially identical to the conclusion following from a significant correlation, as described in section 11.5.1. Specifically, the null hypothesis of no linear association can be rejected. You can conclude that larger $X$s tend to be associated with larger $Y$s (if $b$ is positive) or tend to be associated with smaller $Y$s (if $b$ is negative). Moreover, conclusions about causality again depend on whether the values of $X$ were randomly assigned to cases. If they were, you may conclude that larger values of $X$ *cause* values of $Y$ to be larger (if $b$ is positive) or cause values of $Y$ to be smaller (if $b$ is negative). If they were not, you can only draw the weaker conclusion that the association is consistent with—but does not prove—the hypothesis of a causal relationship.

### 12.4.2 Nonsignificant Slope

When the $F_b$ is not significant, the null hypothesis of $b = 0$ (i.e., no linear association) may be true. The interpretation in this case is exactly the same as the interpretation of a nonsignificant correlation, as described in section 11.5.2.

### 12.4.3 Intercept

The long form of the regression ANOVA table also provides an $F$ for $a$. This $F$ tests the null hypothesis:

$$H_0\text{: the true mean } Y \text{ is zero}$$

just like the $F_\mu$ in ANOVA.[3] The observed $F_a$ is compared against the critical $F$ with one degree of freedom in the numerator and $N-2$ degrees of freedom for error, just as the observed $F_b$ was. Again, the null hypothesis is rejected if the observed $F_a$ is greater than $F_{critical}$.

The interpretation of a significant $F_a$ is simply that the true mean $Y$ value is greater than zero (if $\bar{Y}$ is positive) or is less than zero (if $\bar{Y}$ is negative). Conversely, the interpretation of a nonsignificant $F_a$ is that the true mean $Y$ value may be zero. These conclusions are rarely of any interest in regression situations, which is why the short-form of the regression ANOVA table is so common.

## 12.5 Relation of Simple Regression to Simple Correlation

By now it should be clear that regression and correlation are closely related statistical techniques. One way to look at the relation between regression and correlation is to look at how the regression slope and intercept change with the correlation in a sample. For example, Figure 12.5 shows examples of regression lines for samples with a variety of correlations (cf. Figure 11.3). The data were constructed with means of 100 and standard deviations of 15 on both $X$ and $Y$ in all panels, so the only things changing across panels are the correlation between the two variables and the slope. In fact, since $X$ and $Y$ have the same standard deviation, the slope is the same as the correlation in every panel.

First, note that when the correlation is 0.0 (upper left panel), the slope is zero and the intercept is 100—the mean $Y$. The zero slope was already discussed in connection with point (1) of Table 12.6. Since none of the $Y$ score is predicted to come from $X$, all of it must come from the intercept, hence the intercept is set to the best guess for $Y$, namely its mean.

Second, as the correlation increases, the slope increases proportionally and the intercept decreases proportionally. To some extent these exact numbers depend on the equality of means and standard deviations of $X$ and $Y$, but the panels do illustrate the general principle that *increasing the strength of a positive correlation increases the slope and decreases the intercept*. The maximum possible slope is determined by the ratio of the standard deviations of $X$ and $Y$; specifically, the slope is at most $\frac{s_y}{s_x}$. Naturally, the converse is true for negative correlations: increasing their strength decreases the slope (i.e., makes it more negative) and increases the intercept.

It is also possible to look at the relationship between correlation and regression more mathematically. Table 12.6 summarizes the relationships already identified and lists some additional ones.

| | Regression | Relation | Correlation |
|---|---|---|---|
| (1) | $\hat{b} = 0$ | is the same as | $r = 0$ |
| | $\hat{b} > 0$ | is the same as | $r > 0$ |
| | $\hat{b} < 0$ | is the same as | $r < 0$ |
| (2) | $\hat{b}$ | equals | $r \times \frac{s_y}{s_x}$ |
| (3) | $\hat{b}$ is significant | is the same as | $r$ is significant |
| (4) | $SS_b = 0$ | is the same as | $r = 0$ |
| | $SS_{error} = 0$ | is the same as | $r = 1$ or $r = -1$ |
| (5) | $\frac{SS_b}{SS_{total} - SS_a}$ | equals | $r^2$ |

Table 12.6: Summary of relationships between regression and correlation.

Point (1) of Table 12.6 indicates that the regression slope has the same sign as the correlation: positive, negative, or zero. Mathematically, the reason for this is given in point (2), which says that the slope relating $Y$ to $X$, $\hat{b}$, equals the correlation of $X$ and $Y$ multiplied by the ratio of the standard deviations, $s_y$ divided by $s_x$. Standard deviations must always be positive numbers, so this ratio must also be positive. Thus, the sign of $r$ determines the sign of $\hat{b}$.

[3] It is perhaps surprising that this $F$ tests a hypothesis about the true mean of $Y$ rather than a hypothesis about the intercept (e.g., the true intercept equals 0), but this is in fact the case. Other methods are available for testing hypotheses about the intercept, but these will not be covered here because they are rarely used.

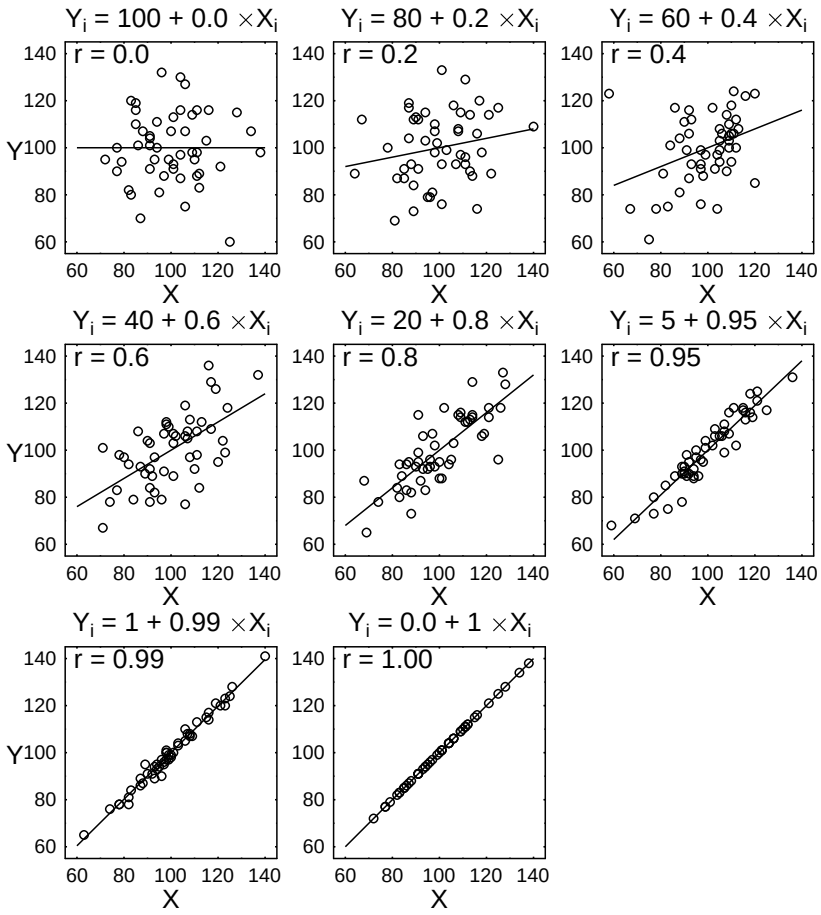


Figure 12.5: Scattergrams displaying samples with different positive correlations ($r$), with the best-fitting regression line indicated on each scattergram. For these data sets, the correlation is the same as the slope ($b$) on each graph.

Point (3) of Table 12.6 indicates that regression and correlation always give the same answer to the question of whether there is a statistically reliable relationship between $X$ and $Y$. The two hypothesis tests look different (i.e. "$H_0$: $b = 0$" versus "$H_0$: $\rho = 0$"), but they both assert that $Y$ is not linearly related to $X$. For any data set, both tests will be significant or neither one will; it is impossible to get a significant result with one test but not the other. In practice, of course, this means that you can test the null hypothesis of no relationship using whichever test you prefer. The correlation coefficient is a little easier to compute, but the regression equation gives you a little more information (i.e., the equation for the linear relationship).

Point (4) shows how the most extreme linear relationships appear in each of the two techniques. If there is no correlation at all between two variables, then there must be zero sum of squares for the prediction of one variable from the other—namely, zero $SS_b$. Conversely, if all the points of the scattergram lie exactly on a straight line, then the correlation is perfect (+1 or -1, depending on whether the line goes up or down from left to right). In this case, there will also be no error for the regression model. That is, all the $\hat{e}_i$ values will be zero, and so will the $SS_{error}$.

Finally, point (5) shows that the square of the correlation coefficient is related to the regression sums of squares for less extreme values as well. The square of the correlation coefficient, $r^2$, is an important quantity in correlation terminology. This value is usually referred to as the "percent of variance predicted," the "percent of variance explained" or the "percent of variance accounted for," and it is the relationship of this quantity to regression sums of squares that gives rise to these names. First, let's look at

$$SS_{total} - SS_a \quad = \quad \sum_{i=1}^{N} \left(Y_i - \bar{Y}\right)^2$$

which is sometimes called the "total variance" (and which we have seen before as the "corrected total" of the short-form regression ANOVA table). Unlike the total sum of squares, the total variance is only influenced by the *variation* in the $Y_i$ scores, not by their overall mean. If we could predict exactly how much each $Y$ would be above or below the mean, then we would be predicting all of $Y$'s variance (i.e., we would be predicting all of its fluctuations around the mean). Second, look at $SS_b$. This can be thought of as the sum of squares that is predicted by the model (or "explained" or "accounted for"). Moreover, the maximum possible $SS_b$ is $SS_{total} - SS_a$, because

$$\begin{aligned} SS_{total} &= SS_a + SS_b + SS_{error} \\ SS_{total} - SS_a &= SS_b + SS_{error} \\ SS_{total} - SS_a &\geq SS_b \end{aligned}$$

because $SS_{error} \geq 0$. Thus, the ratio $\frac{SS_b}{SS_{total} - SS_a}$ can be regarded as the proportion of $Y$'s total variance that the model succeeds in predicting. Mathematically, this fraction is always the same as $r^2$.

# Chapter 13

# Traps and Pitfalls in Regression Analysis

In practice, regression analysis is trickier than it looks, and it is easy to be fooled by regression data, ultimately reaching incorrect conclusions. This chapter describes a number of potential pitfalls to watch out for. They will be described for the situation with one predictor variable, because that is where they are easiest to see, but each can also arise in regression with multiple predictors.

## 13.1 Effects of Pooling Distinct Groups

Suppose you have a sample including two or more distinct groups, such as males and females. Intuitively, it seems logical to include both groups in your correlation and regression analyses to maximize the sample size and get the most powerful tests. This is quite dangerous, however, because *a pooled analysis can yield patterns quite different from those present in either group individually.* For simplicity we will discuss these "pooling effects" mainly in terms of correlations, but the same considerations apply just as strongly in regression analysis.

Figure 13.1 illustrates some of the possible effects of pooling. In panel A, there is no correlation between $X$ and $Y$ for either the males or the females, and the pooled data show no correlation either, just as you would expect. In panel B, there is still no correlation between $X$ and $Y$ within either group, but now the pooled data do show a positive correlation. The positive correlation arises in the pooled sample because the average score is higher for females than for males on both $X$ and $Y$. In panel C, the pooled correlation is even stronger than in panel B, because the difference in averages is even larger. Note that if the females had larger scores on $X$ but smaller scores on $Y$, then pooling would produce a negative correlation. Panels D–F illustrate the same general idea in an even more extreme version. For each group, the correlation between $X$ and $Y$ is -0.8. Nonetheless, a positive correlation can be present in the overall sample if the groups differ in average $X$ and $Y$.

In the examples shown in Figure 13.1, group differences in mean $X$ and $Y$ are responsible for the pooling effects—that is, they *create* correlations in the pooled groups that are not present in either group in isolation. Alternatively, pooling can *conceal* correlations. As shown in Figure 13.2, this can happen even when groups have similar means on $X$, on $Y$, or on both variables. In panels A and B there is a high correlation in both groups, but the correlation is drastically reduced when the groups are pooled, due to the separation between groups on one of the two variables. In panel C the groups have opposite correlations, and naturally there is no correlation when the groups are pooled.

### 13.1.1 Implications for Interpretation

Pooling effects provide an alternative explanation for both nonsignificant and significant correlations, other than the usual explanations that we would like to draw.

For example, suppose you find no significant correlation between exercise ($X$) and longevity ($Y$) in a sample. It could be that longevity is simply unrelated to exercise, as you would be tempted to conclude. Alternatively, it could be that unbeknownst to anyone there are two genetically different subsets of the population, call them A's and B's. The A's have a certain gene such that more exercise leads to greater longevity, but the B's have a different gene such that more exercise leads to lesser longevity. Pooling across these two genetic types that you are not aware of, you get counteracting correlations

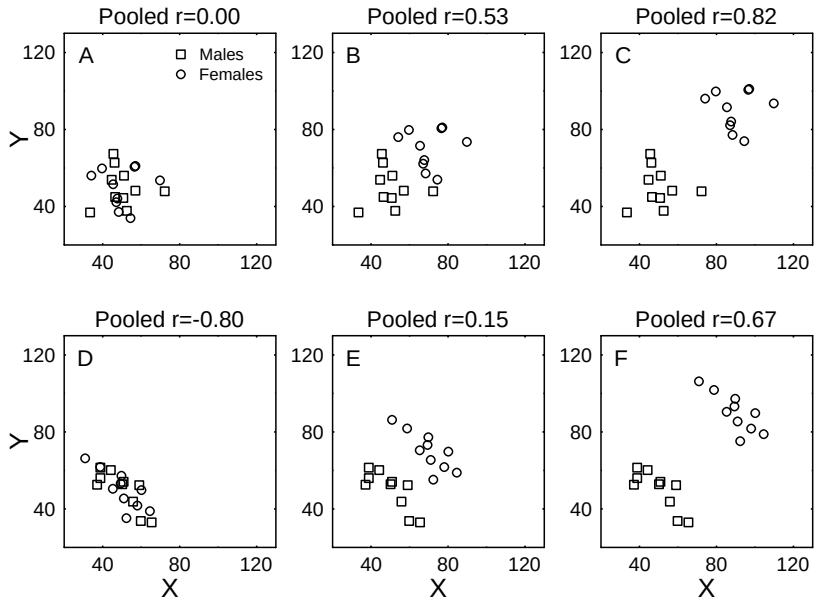


Figure 13.1: Scattergrams showing how pooling can generate correlations between two variables. In panels A, B, and C, the correlation between $X$ and $Y$ is 0.00 for both the males and females. When the two genders are pooled into a single sample, however, the pooled correlation can be positive if the gender with the larger mean on $X$ also has a larger mean on $Y$. In panels D, E, and F, the correlation between $X$ and $Y$ is -0.80 for both the males and females. When the two genders are pooled into a single sample, however, the pooled correlation can be positive if the gender with the larger mean on $X$ also has a larger mean on $Y$.

like those shown in Figure 13.2c, and these counteracting correlations produce the nonsignificant result in your sample. Surely if you really knew about the A and B genes you would not want to conclude that longevity is unrelated to exercise; instead, it would be better to conclude that there is a relation that differs depending on genetic type.

Now let's modify the example so that you do have a significant correlation between exercise and longevity, in which case you would be tempted to conclude that people could extend their lives by exercising. Pooling, however, provides a possible alternative explanation of the association that means the causal conclusion is not airtight. For example, suppose again that there are two genetic types A and B, but now suppose that the A's have a certain "robustness' gene that causes them *both* to live a long time and to like to exercise a lot. The B's do not have this gene, so they neither live so long nor exercise so much. Suppose further that actual longevity is unrelated to the amount of exercise for people in both genetic groups, so nobody can actually live longer by exercising more. Now, the pooled picture could look just like Figure 13.1c, which gives us a significant correlation even though there is no causal connection within either group. In essence, pooling problems are merely one possible manifestation of the alternative, non-causal explanations that arise whenever there is self-selection—i.e., whenever predictor values are *not* randomly assigned to cases.

### 13.1.2 Precautions

To avoid pooling effects—or at least be aware of them—it is a good idea to examine any subgroups that are identifiable in your sample. First, it is worth comparing their means on all variables; if there are differences, then you must watch out for the possibility that these differences induce correlations (cf., Figure 13.1) or conceal them (cf., Figure 13.2). Second, you should compute the desired correlations (and regression models) separately for each of the identifiable subgroups. If the results within a subgroup differ substantially from the pooled results, then pooling has probably affected the results. If you do find pooling effects, then you must abandon the regression model in favor of a model that can take into account both grouping effects and $X/Y$ correlations simultaneously. These are the *analysis of covariance* models of Part 3.

Unfortunately, even with the above precautions you can never really be sure you have avoided

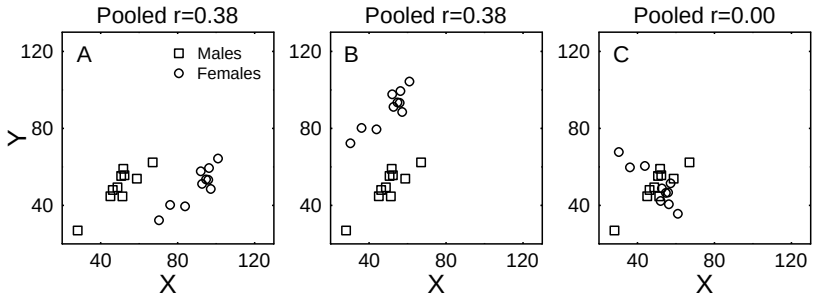


Figure 13.2: Scattergrams showing how pooling can conceal correlations between two variables. In panels A and B, the correlation between $X$ and $Y$ is 0.90 for both the males and females, but the correlation is greatly reduced in the pooled sample. In each panel, the males and females differ on only one variable: They differ in mean $X$ in panel A, and they differ in mean $Y$ in panel B. In panel C the correlation between $X$ and $Y$ is 0.80 for the males and -0.80 for the females, and it is zero for the pooled samples.

pooling effects. The danger is that your data set contains some groups *that you are not aware of.* You cannot look at them separately if you are not aware of them, so your overall results may be contaminated by pooling across them.

## 13.2 Range Restriction Reduces Correlations

Another practical issue in regression concerns the ranges of $X$ and $Y$ within the sample. If the range of either variable is somehow restricted by the sampling process, the correlation between the variables will tend to be reduced.

The standard example of this effect concerns the correlation between IQ and grades in school, which has often been observed to decrease for more advanced students (e.g., high-school students versus university students versus graduate students). The reason for this effect is illustrated in Figure 13.3. A sample of high-school students will include just about the full range of abilities, as shown in panel A, because almost everyone goes to high school. A sample of university students will include a narrower range of abilities, as shown in panel B, because the least able students tend not to go on to college. Comparing panels A and B, it is visually clear that the sample is more elongated in panel A, and this elongation translates into a stronger correlation ($r = 0.70$ versus $r = 0.50$). The range restriction is even more severe when the sample is restricted to graduate students, as shown in panel C, because now the sample will tend to include an even more restricted range of abilities. Correspondingly, the sample is less elongated and the correlation is smaller ($r = 0.19$).

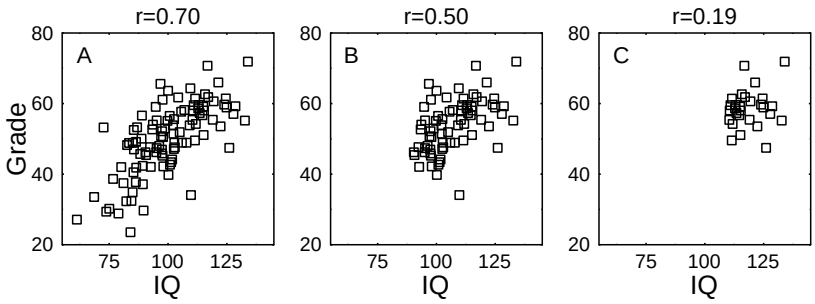


Figure 13.3: Scattergrams showing the effect of range restriction on correlation. Panel A shows the relation between IQ and grades for a sample reflecting the full range of abilities across the population. Panel B shows a subset of the data—just the subsample with IQ in the top 70%, and panel C shows just the subsample with IQ in the top 30%. Note that the correlation decreases as the IQ range decreases.

### 13.2.1 Implications for Interpretation

As shown in Figure 13.3, range restriction tends to reduce correlations. Thus, it implies that the correlation in a sample might be much weaker than the true correlation in the whole population. Obviously, this is most important when you have found a nonsignificant correlation in your sample. Of course the correlation might be nonsignificant because the two variables are truly not very strongly related, as you would tend to conclude. However, range restriction might have been responsible for the absence of correlation. When examining a nonsignificant correlation, then, it is always a good idea to consider whether the sample contained the full population range on the apparently uncorrelated variables.

### 13.2.2 Precautions

The best time to take precautions against range restriction is when collecting the sample. Range restriction can be a very real problem with many sorts of "convenient" samples that are commonly taken instead of random ones. For example, a sample of college students is likely to be very homogeneous with respect to many variables (e.g., age, IQ, socioeconomic status), and it would therefore be a poor choice for examining correlations among such variables.

It is best to avoid range restriction by selecting a truly random sample; with a truly random sample, the likelihood of range restriction is small enough to ignore. Even if a true random sample is impractical, you should still try to obtain a sample containing the full range of variation on the variables you are measuring. And if there is normative data on the ranges of your variables across the whole population, you should compare the range in your sample with the range in the population to make sure that you have succeeded in sampling across the full range. The range of IQs in the normal adult population, for example, might be 70–130, depending on the IQ test. If you measured IQs in your sample, you would want to find that your sample covered approximately this same range, or else be prepared to face the consequences of range restriction. Finally, if you cannot avoid range restriction in your sample, you may still be able to correct for it after the fact using certain formulas beyond the scope of this book. These formulas must be used with extreme caution, however, as they are only valid if certain strong assumptions are satisfied.

## 13.3 Effects of Measurement Error

Many variables can be measured exactly, such as height, weight, age, income, number of children, and so on. Other variables cannot. For example, an IQ test measures IQ with some random error, partly due to the test itself (e.g., exactly which version of the test was used) and partly due to the state of the person taking the test (e.g., how much sleep they had).

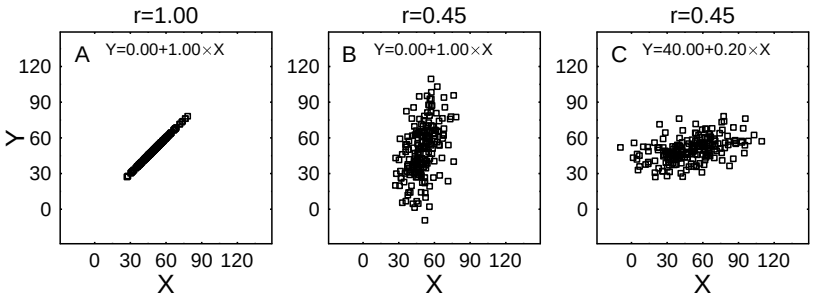


Figure 13.4: Scattergrams showing the effect of measurement errors on correlation and regression. Panel A shows a sample with a perfect $X$-$Y$ correlation in a situation where both $X$ and $Y$ can be measured without error. Panel B shows the same sample, except that a random number has been added to each value of $Y$ to simulate errors in measuring $Y$. Panel C shows the same sample, except that a random number has been added to each value of $X$ to simulate errors in measuring $X$. Note that (1) errors in measuring either $X$ or $Y$ tend to reduce the correlation, and (2) errors in measuring $X$ alter the slope and intercept of the regression equation, but errors in measuring $Y$ do not.

As illustrated in Figure 13.4 and summarized in Table 13.1, measurement error can have dramatic

effects on correlation and regression analyses. Panel A of the figure shows a hypothetical sample in which the true values of $X$ and $Y$ are perfectly correlated. (This is an extreme example, but it makes the effects of measurement error more obvious.) The next two panels show how the data change when random numbers are added to the data to simulate random error in measuring one of the variables. In panel B the random numbers have been added to $Y$, so each point in panel A has been moved randomly up or down by some amount. In panel C the random numbers have been added to $X$, so each point in panel A has been moved randomly left or right by some amount. Note that the correlation is much lower in panels B and C than in panel A. In both cases, the correlation drops because the random measurement error clearly breaks up the perfect relation between $X$ and $Y$. Thus, one lesson to be learned from this example is that measurement error tends to reduce correlations[1].

A second and more subtle effect of measurement error can be seen by comparing the regression equations computed for the three samples. Panel A shows the "true" regression equation that could be measured in the absence of any measurement error, namely:

$$Y = 0.00 + 1.00 \times X$$

Interestingly, the very same regression equation is obtained using the data of Panel B, for which measurement error was added to the $Y$ values. Thus, measurement error in $Y$ does not necessarily change either the slope or the intercept when predicting $Y$ from other variables. That is good news; even though the correlation is harder to see, at least the regression equation tends to give the true slope and intercept. It is not too hard to see why this is true. Any given $Y$ is moved up or down randomly by the measurement error. But for any given $X$, the average of its $Y$'s will be just about the same as the average without measurement error. That is, the measurement errors of the different $Y$'s associated with a given $X$ will tend to cancel each other out, so the predicted $Y$ for that $X$ (i.e., the average $Y$ for that $X$) will be the same with or without measurement error in $Y$.

The news is not as good when there is measurement error in $X$, as illustrated in panel C. In this case, the slope is greatly reduced and the intercept increased, relative to the true values that would have been observed without measurement error in $X$. The slope is reduced (i.e., flatter regression line) for two reasons:

1. The data have been stretched along the $X$ axis. Note that the range of $X$ is approximately 30 to 70 without measurement error (panel A) but -10 to 110 with error (panel C). Because $Y$ changes the same amount from lowest to highest in both cases, the change in $Y$ per unit $X$ would have to be much smaller with the larger range even if the correlation were perfect in both cases.

2. The change in $Y$ is not as strongly associated with the change in $X$, because the measurement error also jumbles the points in the left–right dimension. This jumbling disturbs the perfect correlation of $X$ and $Y$, just as the error in $Y$ did in panel B.

The combination of these two effects reduces the slope and increases the intercept when there is error in $X$. Note also that analogous slope changes would also occur if the correlation were negative; in this case, adding measurement error to $X$ would again flatten the line, this time by increasing the slope (i.e., making it less negative). Note also that the estimated regression intercept also changes in the presence of measurement error in $X$. The intercept change is a direct consequence of the slope change, because as the line flattens it will necessarily cross the $Y$ axis at a different point.

### 13.3.1 Implications for Interpretation

You must consider measurement error any time you are interpreting a nonsignificant correlation or regression. Since measurement error in either $X$ or $Y$ tends to reduce the correlation (Figure 13.4), it is possible that a given correlation was nonsignificant because there was too much measurement error.

You must also consider the possibility of measurement error in $X$ before interpreting the value of a regression slope or intercept, since these values may be biased. Some studies of subliminal perception provide an interesting example (Greenwald, Draine, & Abrams, 1996; Greenwald, Klinger, & Schuh, 1995). In brief, the researchers separately measured the amount of information perceived consciously

[1] You might be suspicious about generalizing from this artificial example with a perfect $X$-$Y$ correlation in the absence of measurement error. Here, measurement error had to break up the correlation, because there was nowhere to go but down. But suppose the correlation had been smaller, say 0.5, without any measurement error. Then, wouldn't it be possible for measurement error to *increase* the correlation within a sample instead of reducing it? The answer is that it would be possible for a particularly fortuitous set of measurement errors to increase the correlation within a particular sample, but it is very unlikely that this would happen. In most random samples, measurement error tends to drive even small correlations closer to zero.

| | Measurement Error in | |
|---|---|---|
| Effect on: | $X$ | $Y$ |
| Correlation | Reduces r. | Reduces r. |
| Slope | Slope flattened. | No systematic effect. |
| Intercept | Intercept biased because of change in slope. | No systematic effect. |

Table 13.1: Effects of Measurement Error on Correlation and Regression Analysis

($X$) and unconsciously ($Y$), and then computed a regression to predict $Y$ from $X$. The crucial finding was that the intercept was greater than zero, which they interpreted as evidence some information was perceived unconsciously even when none was perceived consciously (i.e., $Y$ greater than zero when $X$=0). Others have questioned their methodology, however, because their measure of $X$ was subject to some measurement error, and this error would have biased the intercept to be greater than zero (Dosher, 1998; Miller, 2000).

### 13.3.2 Precautions

The obvious precaution is to eliminate measurement error, especially error in $X$ if a regression equation is to be computed. While this is easy enough to recommend within the pristine confines of a statistics text, it is often impossible to achieve in the messy real world of data collection. In situations where measurement error is unavoidable, the only hope is to measure it and to estimate its probable effects on the correlation and regression analyses. Techniques for accomplishing this are beyond the scope of this book, however.

A final comment about measurement error is that in regression analysis, unlike ANOVA, it cannot be completely overcome by increasing the sample size. In ANOVA, random error tends to average out as sample size increases, so the observed pattern of cell means approaches the true pattern as sample size increases. Measurement error does not distort the pattern of means, but merely increases the sample size needed to approach the true pattern. In regression, though, measurement error does not magically go away as the sample size is increased. Consider, for example, the difference between panels A and B of Figure 13.4. The data patterns differ because of the measurement error in $Y$, and the difference will persist no matter how large a sample is included in panel B.

## 13.4 Applying the Model to New Cases

In some situations it is desirable to use the results of a regression analysis to make predictions for new cases. This is a type of extrapolation, because the model was developed on one set of cases and then is being used to make predictions about a new set. It is natural to wonder how accurate your predictions will be when you extrapolate like this. As an example, suppose you developed a model to predict the outcomes of horse races, intending to use the model to guide your bets and win some money. To develop the model, you would collect a data set with a to-be-predicted variable (e.g., each horse's finishing time) and one or more predictor variables. You would then fit a regression model to predict $Y$ from the $X$'s, and if the model fit well you would use it to predict the outcomes of new races on which you intended to bet. How accurate would you expect these predictions to be?

It seems natural to expect that the errors in your new predictions would be about the same as the estimated errors ($\hat{e}_i = Y_i - Y_i'$) that you computed for your sample. If the $MS_{error}$ in your regression analysis were 100, for example, you might expect the average prediction error to be approximately the square root of that—around 10. Unfortunately, this expectation is too optimistic, and your real average error would very likely be greater than 10.

Why does error tend to increase when making predictions for new cases? The main reason is the sampling error in the regression slope and intercept values. The key point is that $\hat{a}$ and $\hat{b}$ are the optimal values leading to the minimum $MS_{error}$ *for the original sample,* not for the population as a whole. When you use the sample values to make predictions about the new cases, then, you are not really using the best values—the true population values would be better. The prediction error for new cases will be inflated as a result.

To illustrate this increase in prediction error when extrapolating to new cases, consider a simple hypothetical situation in which samples are taken from a known population. Imagine that the scat-

tergram shown in panel A of Figure 13.5 represents the $X$ and $Y$ values for *an entire population of 250 individuals,* perhaps animals of some endangered species. As indicated in the panel, the optimal regression line for this population is $Y_i = 85.0 + 0.300 \times X_i$.

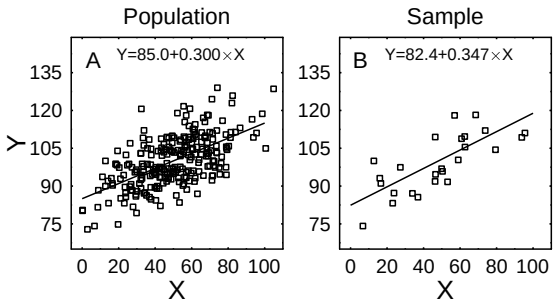


Figure 13.5: Scattergrams to illustrate why prediction error tends to increase when predictions are made for new cases. Panel A shows the relation between $X$ and $Y$ for a full population of 250 cases; panel B shows the relation for a random sample of 25 cases from this population.

Now imagine that you have only a random sample of 25 of these cases, shown in panel B, as you might get in a real research situation. From these cases you compute the regression line $Y_i = 82.4 + 0.347 \times X_i$. Note that the sample's slope and intercept differ from their true population values; this is, of course, because of sampling error. Furthermore, this regression analysis yields $MS_{error} = 61.196$, which might lead you to expect a typical prediction error to be approximately $\sqrt{61.196} = 7.82$ when making predictions for new cases. But if you apply this predictive model to the remaining 225 "new" cases in the population, you'll get a value of $\sqrt{65.612} = 8.10$ as the resulting average error.[2] Thus, a typical prediction error when making predictions for new cases is about 3.5% larger than you would have guessed from looking at the analysis of the random sample.

You might think that you could avoid this problem by including each new case in the sample and refitting the model every time a new data value is obtained. Note, however, that you need both $X$ and $Y$ for the new case in order to refit the model including this case. But if you wait to have both $X$ and $Y$, then it is too late to act on your predicted $Y$ value. In the case of betting on horse races, for example, you must place your bet before the finishing time ($Y$) is determined—that is, before it is possible to refit the model including that case.

### 13.4.1 Implications for Interpretation

The only implication is the obvious one: that prediction error should be expected to be somewhat larger for new cases than for the sample itself.

### 13.4.2 Precautions

There is no way to avoid the increase in prediction error, so the only possible precaution is to estimate it and allow for it. The technical issue of how much increase in prediction error to expect is beyond the scope of this book. Two qualitative points will suffice for our purposes:

- Prediction error increases more if the model has more predictors.
- Prediction error increases less if the model is based on a larger sample.

## 13.5 Regression Towards the Mean

Regression towards the mean (also known as "regression to the mean" or "statistical regression") is a reasonably simple concept, but its applications are often difficult to spot. This section will explain the basic concept, and the next section on "Implications for Interpretation" will explain why you need to know about it.

In brief, "regression towards the mean" is simply a name for this fact:

[2] To be more specific, for each of the other 225 cases I computed $\hat{e}_i = Y_i - 82.4 + 0.347 \times X_i$. 65.612 is the average of the 225 squared $\hat{e}_i$ values.

*When you make a prediction with a regression model, the predicted value will be closer to its mean than the predictor is to its mean.*

That is, $Y_i'$ is closer to $\bar{Y}$ than $X_i$ is to $\bar{X}$, (ignoring a few mathematical loose ends that will be dealt with shortly).

Panel A of Figure 13.6 provides an illustrative data set. The to-be-predicted variable is the height of a son, and the predictor variable is the father's height. These are fictitious data, but they are pretty realistic except that for convenience the fathers and sons have identical mean heights (180 cm) and standard deviations of height (10 cm).

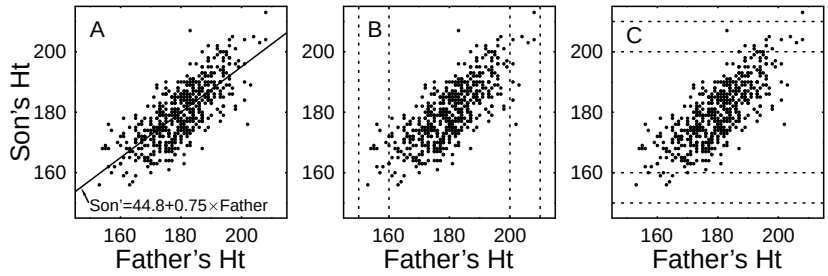


Figure 13.6: Panel A shows a scattergram of 500 cases showing the relationship between heights of fathers and heights of their sons. The best-fitting regression line and its equation are shown on the graph. Panels B and C show the same scattergram with certain groups of cases indicated by dashed lines, as discussed in the text.

The best-fitting regression line to predict son's height from father's height, shown on the graph, is $Son' = 44.8 + 0.75 \times Father$. Table 13.2 shows some example predictions computed using this equation. Note that for each case in this table the value of $Y_i'$ is closer to the mean 180 than is $X_i$. That is, for small values of $X_i$, $Y_i'$ is larger than $X_i$, whereas for large values of $X_i$, $Y_i'$ is smaller than $X_i$. This is simply the phenomenon of regression towards the mean as it applies to these cases.

| Father's Height ($X_i$) | Predicted Son's Height | $Y_i'$ |
|---|---|---|
| 150 | $44.8 + 0.75 \times 150 =$ | 157.3 |
| 160 | $44.8 + 0.75 \times 160 =$ | 164.8 |
| 170 | $44.8 + 0.75 \times 170 =$ | 172.3 |
| 190 | $44.8 + 0.75 \times 190 =$ | 187.3 |
| 200 | $44.8 + 0.75 \times 200 =$ | 194.8 |
| 210 | $44.8 + 0.75 \times 210 =$ | 202.3 |

Table 13.2: Example predicted values using regression equation to predict son's height from father's height.

The regression toward the mean illustrated in Table 13.2 is not just an accident of this particular data set, but it reflects a phenomenon that holds across all possible data sets. One relevant equation is this:

$$\left|\frac{Y_i' - \bar{Y}}{s_y}\right| = |r_{xy}| \times \left|\frac{X_i - \bar{X}}{s_x}\right| \tag{13.1}$$

In words, it says that the absolute distance from the predicted $Y_i'$ to $\bar{Y}$ (measured relative to the standard deviation of the $Y$ scores) is equal to the absolute value of the sample correlation coefficient times the absolute distance from the predictor value $X_i$ to $\bar{X}$ (measured relative to the standard deviation of the $X$ scores). Since the absolute value of the correlation is virtually always less than 1.0, it follows that $Y_i'$ is virtually always closer to $\bar{Y}$ than $X_i$ is to $\bar{X}$, with distances on both variables measured relative to their respective standard deviations.

Intuitively, one way to explain this is as follows: If you have no information on which to base a prediction, your best strategy is to guess the mean of the variable you are trying to predict. For example, if you had to guess a son's height without knowing anything at all, your best guess would be the mean height, 180 cm. Now if you get some useful predictive information, such as the father's height, this causes you to move your prediction away from the mean. For example, if the father's

height is one standard deviation above average, then we would expect the son's height to be above average too. But how much above average? Not a full standard deviation, as the father's height was, because father's and son's heights are only imperfectly correlated. In fact, if the correlation is only $r_{xy} = 0.5$, then you should expect the son's height to be one-half a standard deviation above average.

It is important to realize that regression toward the mean is actually a common phenomenon in the real world, not just an abstract mathematical property of this particular statistical model. That is, the regression model is "right" to keep its predictions relatively close to the mean $Y$, because these predictions will be most accurate in the long run. For example, tall fathers really do tend to have somewhat shorter sons, on average, and very short fathers really do tend to have somewhat taller sons. This is part of what makes regression models appropriate and useful in real-world situations.

This property is inherent in all normally distributed populations, and it was first observed by Francis Galton in an analysis of the genetic transmission of height in the 1880s, around the time of Darwin's then-new genetic theory. The basic data set was a collection of father's and son's heights, as we have been considering (e.g., Figure 13.6). Regression analysis had not yet been invented yet, so the investigators used their data to construct a table like that shown in Table 13.3. First, cases were grouped according to the heights of the fathers. For example, one group contained all the cases with fathers heights in the range from 150–160 cm. Second, each group of cases was summarized by computing the average height of the fathers in that group and the average height of the sons in that group.

| Group | Father's Height Range for Group | Number of Cases | Average Height of Fathers | Average Height of Sons |
|---|---|---|---|---|
| 1 | 150–160 | 11 | 156.0 | 166.6 |
| 2 | 160–170 | 61 | 165.3 | 169.2 |
| 3 | 170–180 | 163 | 174.9 | 175.4 |
| 4 | 180–190 | 182 | 184.2 | 183.1 |
| 5 | 190–200 | 73 | 193.5 | 191.2 |
| 6 | 200–210 | 10 | 203.5 | 196.4 |
| Total | 150–210 | 500 | 180.0 | 180.0 |

Table 13.3: A summary tabulation of the cases shown in Figure 13.6. Each case was assigned to one group depending on the height of the father. After the groups had been determined, the average heights of fathers and sons in each group were determined.

The results shown in Table 13.3 are representative of what was actually obtained, and they illustrate clear regression toward the mean. For example, the fathers with heights in the 150–160 cm range tended to have somewhat taller sons, with an average height of 166.6 cm. Similarly, the fathers with heights in the 200–210 cm range tended to have somewhat shorter sons, with an average height of 196.4 cm. For each group, the average of the son's heights is closer to the overall mean height (180 cm) than the average height of the fathers in that group.

As an historical aside, inspection of data like that shown in Table 13.3 initially led scientists to conclude that something about the mechanisms of genetic transmission worked to even out heights across many generations. They apparently believed that individual differences in height had been greater in the past, were diminishing with successive generations, and would eventually all but disappear. In fact, the term "regression" originated within statistics in precisely this context, to describe this pattern of apparent change over time.

It was not long before these researchers realized that the pattern of regression shown in Table 13.3 resulted from some sort of statistical effect rather than genetic mechanisms. This was discovered when someone decided to have another look at the same data, this time grouping the cases according to the sons' heights instead of the fathers'. The new tabulation yielded a pattern like the one shown in Table 13.4.

Contrary to genetic explanations of the regression (and to most people's intuitions at this point!), the exact opposite pattern of regression was obtained when the cases were grouped according to the sons instead of the fathers. As you can see, for every group in Table 13.4, the average height of the fathers is closer to the mean than the average height of the sons. That is, the fathers' heights regress toward the mean, not the sons heights. This suggests that heights are getting more extreme from one generation to the next, not less extreme as suggested by the tabulation in Table 13.3. How can the same data lead to two opposite conclusions?

| Group | Son's Height Range for Group | Number of Cases | Average Height of Fathers | Average Height of Sons |
|---|---|---|---|---|
| 1 | 150–160 | 5 | 161.2 | 157.4 |
| 2 | 160–170 | 71 | 169.7 | 165.9 |
| 3 | 170–180 | 175 | 175.9 | 174.7 |
| 4 | 180–190 | 163 | 183.6 | 184.4 |
| 5 | 190–200 | 68 | 189.7 | 193.2 |
| 6 | 200–210 | 17 | 195.5 | 202.4 |
| 7 | 210–220 | 1 | 208.0 | 213.0 |
| Total | 150–210 | 500 | 180.0 | 180.0 |

Table 13.4: A summary tabulation of the cases shown in Figure 13.6. Each case was assigned to one group depending on the height of the son. After the groups had been determined, the average heights of fathers and sons in each group were determined.

The answer is that in both tables the regression towards the mean results from grouping the bivariate distribution:

*When cases are grouped according to one variable, the group averages on a second variable tend to be less extreme than the group averages on the first (grouping) variable.*

First, look back at Tables 13.3 and 13.4 to see that this is an accurate summary. Next, look back at panels B and C of Figure 13.6 to see *why* this is true. Panel B illustrates grouping according to fathers' heights. The two columns set off by vertical dashes correspond to the cases in which the fathers' heights were in the ranges of 150–160 cm and 200–210 cm. If you examine the points captured in the left-most column (i.e., fathers heights of 150–160 cm), you can see that the average height of the sons in this group is clearly more than 160 cm. Because of the scattergram's cloud shape (i.e., because the two variables have normal distributions), many of the cases with very low values on father's height have not-so-low values on son's height, and these points bring up the average sons' height for that group. Similarly, if you examine the points captured in the right-most column (i.e., fathers heights of 200–210 cm), you can see that the average height of the sons in this group is clearly less than 200 cm. Again because of the scattergram's cloud shape, many of the cases with very high values on father's height have not-so-high values on son's height, and these points bring down the average sons' height for that group. To summarize, when you select out the extreme fathers, their sons tend to be more toward the middle of the cloud because that is where most of the points are.

Panel C illustrates the same point, now grouping according to sons' heights instead of fathers heights. The two rows set off by horizontal dashes correspond to the cases in which the sons' heights were in the ranges of 150–160 cm and 200–210 cm. If you examine the points captured in the lower row (i.e., sons heights of 150–160 cm), you can see that the average height of the fathers in this group is clearly more than 160 cm. Again, this is because of the scattergram's cloud shape. Similarly, if you examine the points captured in the upper row (i.e., sons heights of 200–210 cm), you can see that the average height of the fathers in this group is clearly less than 200 cm. Likewise, then, when you select out extreme sons, their fathers tend to be more toward the middle of the cloud because that is where most of the points are.

One final but important point about regression toward the mean is that it only applies *on the average*, not necessarily for every case. In the data set shown in Figure 13.6, for example, there was a father in the tallest group (208 cm) who had an even taller son (213 cm). Similarly, there was a son in the tallest group (204 cm) who had an even taller father (208 cm). Obviously, regression toward the mean did not operate in these two individual cases. With normal distributions, though, it must operate on the average across large numbers of cases.

### 13.5.1 Implications for Interpretation

The phenomenon of regression toward the mean must be kept in mind whenever cases are grouped or selected using one (or more) variables and then tested or measured on a different variable. You have to keep in mind that the pattern observed on the measured variable might be due to a purely statistical phenomenon: regression toward the mean.

For example, assume that the following was demonstrated: "Women who go to University tend to marry men who are less intelligent than they are." You can imagine a variety of causal explanations

of this that you might read in a newspaper or magazine: For example, someone might argue from this observation that University somehow makes women less desirable marriage partners and so they have less intelligent suitors than they would have had otherwise. Before entertaining any such causal explanation, however, you must first realize that this phenomenon could simply be another example of regression toward the mean. Suppose we consider married couples to be cases. We now select out a group in which the IQs of the wives are higher than 110 (the wives must have relatively high IQs, because they went to University), and measure the group average IQ of the husbands. Naturally it will be lower than the wive's average, just due to the phenomenon of regression toward the mean. The situation is depicted in Figure 13.7, which is analogous to Figure 13.6. Critically, there is no basis to conclude that going to University has any causal influence on its own. There may be such influences, but this effect doesn't demonstrate them because the effect can be explained entirely by regression toward the mean. (And of course the converse should also be true—University men should tend to marry women who are less intelligent, on average, than they are—for exactly the same reason.)

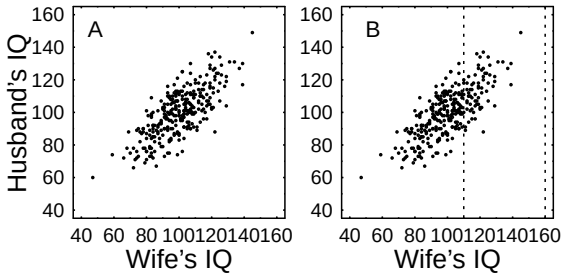


Figure 13.7: Panel A shows a scattergram of 300 cases showing the relationship between IQs of wives and husbands. Panel B shows the same scattergram with the wives who attended University indicated by dashed lines, as discussed in the text. The mean IQ of these wives is 119, whereas the mean IQ of their husbands is 113. This diagram is an oversimplification of the true situation, because in reality there is no absolute cutoff IQ separating women who do and do not attend University. Despite this oversimplification, the diagram illustrates why the mean IQ of the wives would likely be greater than the mean IQ of their husbands.

Another classic example is that people are often disappointed when they go back for a second meal at a restaurant where they had particularly wonderful food on their first visit. Whether they order the same dish or a new one on that second trip, they will often find that the food just isn't as pleasing as it was before. This is to be expected on the basis of regression toward the mean. In this example, the restaurant is the case, and your subjective impressions of the first and second meals are the two variables. When you select out the subset of restaurants with really good scores on the first meal, the average of the second meal impressions is bound to be lower. You would notice the analogous phenomenon if you selected restaurants where the second meal was especially outstanding (i.e., the first meals would not have been as good, on average), but this is not as salient a comparison psychologically.

As a third and slightly more complicated example, suppose you are evaluating two new University programs designed to meet the needs of special students. One program is aimed at high achievers. Students are invited to join after they achieve (say) an average of at least 85% across all their courses for one semester. Once they are in the program, they are invited to special lectures and University functions designed to help them develop their potential to the fullest. The other program is aimed at underachievers. Students are invited to join this one if they score less than 60% in a semester, and they are given special tutoring and remedial help. At the end of one semester, the results come back. Disappointingly, in the "high achievers" program, the students' average has dropped from 87% before the program to 83% in it. The other program is apparently going much better, however, because the underachievers raised their averages from 56% to 63%. Naturally, you are tempted to conclude that the high achievers program is at best useless—possibly harmful (e.g., too distracting)—and should be scrapped. In contrast, the underachievers program is producing big gains. Obviously, you should put all your efforts into the underachievers program, right?

Wrong. Students were selected into groups ("high achievers" versus "underachievers") based on scores in one semester, and then their performance was assessed based on new scores from another semester. The phenomenon of regression toward the mean dictates that, on average, the high scores

must decrease and the low scores must increase. That is exactly what happened! Based on these results, you can't say for sure that either program did anything at all, because regression toward the mean provides a perfectly good explanation of the results all by itself. It could be that one or both programs are useful, of course, and in fact the high achievers program could even be the better one. You just can't tell from the information you have at hand.

### 13.5.2 Precautions

Regression toward the mean is a real statistical phenomenon, present whenever variables are normally or near-normally distributed. There is no way to eliminate it, so you must be aware of it. If you consider it when designing a study, you may be able to control for its effects, as discussed next. Even if you don't consider it until the data are in, you can at least avoid making conclusions that are invalid because it might be operating.

In many situations, the effects of regression toward the mean can be determined by including a control group. Then, you can see if the experimental effects are larger than those in the control group; if so, you can dispense with regression toward the mean as an explanation.

As an example, consider again evaluating the programs for high achievers and underachievers. Instead of inviting all qualified students into the two programs, you would be better off to invite only a randomly selected half of each group into each program. The other, uninvited halves could be held out of the program for a semester to serve as control groups, and changes in their scores would include the effects of regression toward the mean (as well as seasonal effects, etc). At the end of one semester, you would have data for these four groups:

| High Achievers Program | |
|---|---|
| In Program | Control |
| 1 | 2 |

| Underachievers Program | |
|---|---|
| In Program | Control |
| 3 | 4 |

You could legitimately compare groups 1 and 2 to evaluate the high achievers program. Because of regression toward the mean, you would expect both these groups to have somewhat lower average scores in this semester than in the previous one. If the program is helpful, though, the dip should be smaller for the students who were in it. Similarly, you could compare groups 3 and 4 to evaluate the underachievers program. Both groups should increase relative to the previous semester due to regression towards the mean. If the program is helpful, though, group 3 should show a larger increase than group 4.

# Chapter 14

# Multiple Regression

## 14.1 Introduction

A multiple regression model is used to quantify the relationship of a $Y$ variable and two or more $X$ variables. Just as in simple regression, the purposes of quantifying the relationship are (a) prediction, and (b) looking for signs of causal relationships.

A multiple regression model is used with data in the "Case by Variable" format, an example of which is shown in Table 14.1. As in simple regression, there is a single sample of cases, and each case is measured with respect to a numerical $Y$ variable. In multiple regression, each case is also measured with respect to two or more $X$ variables.

| Case | PERF | TRA | EXP | IQ | Predicted | Error |
|---|---|---|---|---|---|---|
| 1 | 62 | 13 | 15 | 85 | 60.49 = -22.61 + 2.37 × 13 + 1.60 × 15 + 0.33 × 85 | 1.51 = 62 - 60.49 |
| 2 | 67 | 11 | 14 | 123 | 66.76 = -22.61 + 2.37 × 11 + 1.60 × 14 + 0.33 × 123 | 0.24 = 67 - 66.76 |
| 3 | 73 | 12 | 13 | 127 | 68.86 = -22.61 + 2.37 × 12 + 1.60 × 13 + 0.33 × 127 | 4.14 = 73 - 68.86 |
| 4 | 75 | 11 | 20 | 120 | 75.39 = -22.61 + 2.37 × 11 + 1.60 × 20 + 0.33 × 120 | -0.39 = 75 - 75.39 |
| 5 | 54 | 7 | 19 | 99 | 57.33 = -22.61 + 2.37 × 7 + 1.60 × 19 + 0.33 × 99 | -3.33 = 54 - 57.33 |
| 6 | 69 | 10 | 13 | 114 | 59.80 = -22.61 + 2.37 × 10 + 1.60 × 13 + 0.33 × 114 | 9.20 = 69 - 59.80 |
| 7 | 64 | 12 | 15 | 124 | 71.07 = -22.61 + 2.37 × 12 + 1.60 × 15 + 0.33 × 124 | -7.07 = 64 - 71.07 |
| 8 | 59 | 10 | 10 | 124 | 58.31 = -22.61 + 2.37 × 10 + 1.60 × 10 + 0.33 × 124 | 0.69 = 59 - 58.31 |
| 9 | 57 | 12 | 13 | 99 | 59.56 = -22.61 + 2.37 × 12 + 1.60 × 13 + 0.33 × 99 | -2.56 = 57 - 59.56 |
| 10 | 49 | 10 | 12 | 80 | 46.91 = -22.61 + 2.37 × 10 + 1.60 × 12 + 0.33 × 80 | 2.09 = 49 - 46.91 |
| 11 | 55 | 8 | 11 | 113 | 51.52 = -22.61 + 2.37 × 8 + 1.60 × 11 + 0.33 × 113 | 3.48 = 55 - 51.52 |
| 12 | 42 | 8 | 11 | 112 | 51.19 = -22.61 + 2.37 × 8 + 1.60 × 11 + 0.33 × 112 | -9.19 = 42 - 51.19 |
| 13 | 63 | 8 | 18 | 107 | 60.76 = -22.61 + 2.37 × 8 + 1.60 × 18 + 0.33 × 107 | 2.24 = 63 - 60.76 |
| 14 | 60 | 7 | 15 | 116 | 56.56 = -22.61 + 2.37 × 7 + 1.60 × 15 + 0.33 × 116 | 3.44 = 60 - 56.56 |
| 15 | 50 | 10 | 11 | 107 | 54.27 = -22.61 + 2.37 × 10 + 1.60 × 11 + 0.33 × 107 | -4.27 = 50 - 54.27 |
| Sum: | 899 | | | | 899.00 | 0.00 |
| Mean: | 59.93 | | | | 59.93 | 0.00 |
| SumSq: | 55049 | | | | 54740.62 | 308.38 |

Table 14.1: Hypothetical example data set for multiple regression. The researcher is interested in finding out how on-the-job performance (PERF) by computer operators is related to weeks of training (TRA), months of experience (EXP), and IQ. The right-most two columns ("Predicted" and "Error") are not part of the data set but emerge in doing the computations, as described below.

## 14.2 The Model for Multiple Regression

All multiple regression models have this form:

$$Y_i = a + b_1 \times X_{1i} + b_2 \times X_{2i} + e_i$$

The $Y_i$ values in the model are the observed values of the to-be-predicted variable, analogous to the dependent variable in ANOVA. For the example in Table 14.1, the $Y_i$ values are the PERF scores. The subscript $i$ is used to distinguish the different cases; for example, $Y_{13}$ is 63.

The $a$ term in the model is a constant, analogous to a baseline in ANOVA. As in simple regression, it is rarely of much interest in multiple regression but it must nonetheless be included in the model for computational reasons.

The model terms of primary interest are

- $b_1 \times X_{1i}$
- $b_2 \times X_{2i}$
- $b_3 \times X_{3i}$
- and so on.

These terms capture the relationships of the $X$ variables to $Y$.

Each of these $b \times X$ terms involves a different one of the predictor variables $X_{1i}$, $X_{2i}$, etc. The first subscript (1,2,3,...) is used to differentiate among the different predictor variables, and the $i$ subscript is used to differentiate among the cases just as it does with $Y_i$. In Table 14.1, for example, the values of $X_{1i}$ are the values of the first predictor variable, namely TRA, and $X_{1,6}$ is the TRA value for the 6th case—i.e., 10. Continuing on from left to right as is standard, $X_{2i}$ denotes the values of the EXP variable and $X_{3i}$ denotes the values of the IQ variable.

The terms $b_1$, $b_2$, $b_3$, and so on are the "regression slopes." Each slope captures the linear relationship between $Y$ and its associated predictor or $X$ variable, just as a slope does in simple regression. Unlike simple regression, however, each multiple regression slope should be thought of as indicating how $Y$ would change with $X$ *if all of the other X's were held constant.* If you examine the multiple regression equation, you can see that if $b_1$ is zero, then the term $b_1 \times X_{1i}$ is also zero, and the first predictor variable totally drops out of the model. In other words, if $X_{1i}$ does not help predict $Y_i$ when all of the other $X$'s are taken into account, then $b_1$ will be adjusted to (nearly) zero and the model will (essentially) ignore the value of $X_{1i}$ in predicting $Y_i$. On the other hand, if the $X_{1i}$ values do help predict $Y_i$ (for a constant set of other $X$'s), then $b_1$ will be adjusted to an appropriate positive or negative value. For example, if $Y_i$ tends to increase as $X_{1i}$ increases (with all of the other $X$'s held constant), then $b_1$ will be set to a positive value. That will mean that larger values of $X_{1i}$ will produce larger predicted values of $Y_i$, with the other $X$'s remaining constant.[1]

Thinking in ANOVA terms, each $b \times X$ term is analogous to a main effect of that $X$ variable. In a sense, then, the regression model includes main effects. But it includes only these main effects (and error)—it does not include anything corresponding to an interaction of two or more factors. In principle one could include interaction terms in regression models, but this is an advanced topic that we will not cover.

Finally, the $e_i$ term is the random error term in the model. As usual, the model attributes to random error anything that it cannot explain. Also as usual, the size of the estimated random error is used as a yardstick to decide whether other effects are "real" or "small enough that they might be due to chance."

## 14.3 A Limitation: One $F$ per Model

There is an important limitation of the regression model that many students find puzzling at first: With a multiple regression model, you only get one $F$ that jointly tests whether $Y$ is related to *any* of the predictors. This is very different from ANOVA, where we got one $F$ to test each factor's main effect separately (and more $F$'s for interactions). In contrast, with a multiple regression model you only get one $F$ for the whole model, regardless of the number of predictors.

The reasons for this limitation can be explained more easily after we have seen how the multiple regression model itself is used and how it can be extended with a technique called the "Extra Sum of Squares Comparison", described in Chapter 15. In essence, though, the idea is that the different predictors work together in ways that are difficult to measure separately.

Perhaps an analogy to team sports will be helpful: In many team sports where the players all work together to try to score (e.g., soccer, rugby, net ball), it is very difficult to measure the value of each player to the team's scoring. Obviously, just measuring the points scored by that player would not be a very good measure, because some players might be extremely valuable to the offense even if they don't score many points. For example, a player might be exceptionally good at making "assists"—that is, at getting the ball into a position where *other* players on their team can score. Clearly, such a player would make a large contribution to the team's scoring, even if he or she never

[1] It is tedious to keep specifying the condition "with the other $X$'s remaining the same" or equivalent, so henceforth we will omit this condition, but it should always be understood to be implicitly present in the multiple regression situation.

personally scored any points at all, as could be demonstrated by removing the player from the team and seeing its scores fall. It is easy to see that the possibility of such assists makes it more difficult to quantify each player's contribution to the scoring in this type of sport. In contrast, there are other sports (e.g., cricket, baseball) where scoring is done more by individual effort, and in these sports it is correspondingly easier to measure each player's individual contribution to the offense.

For now, suffice it to say that multiple regression is more like soccer than like cricket, so it is relatively difficult to measure out each predictor's separate contribution to the model; ANOVA is more like cricket. For now, it will be enough to be aware that a multiple regression model (unlike an ANOVA) does not give you a separate $F$ to isolate the predictive contribution of each predictor. To get information on a predictor's separate contribution to the model requires a type of extra analysis, typically done after the multiple regression, that will be covered in Chapter 15.

## 14.4 Computations

Although the computations are more complex, the principles underlying the multiple regression model are remarkably similar to those used in simple regression. As usual, you:

- Write down the model.
- Estimate the parameters of the model—namely, compute $\hat{a}$, $\hat{b}_1$, $\hat{b}_2$, ....
- Compute the predicted $Y$ value and error for each case.
- Compute sums of squares.
- Compute a summary table and test null hypotheses.

### 14.4.1 Estimation of $a$ and the $b$'s

The main difference between multiple regression and simple regression is that you need a computer to estimate the values of $a$ and the $b$'s. I will not even present the estimation equations, because they cannot be written down very compactly without using matrix algebra. For our purposes, then, it will remain mysterious how these values are computed. Note, however, that the computation leads to the *best possible* values of $a$ and the $b$'s. That is, the computation gives us the unique set of values that lead to the most accurate predictions of $Y$ for our sample. That is, if the $X$'s are good predictors, the $\hat{a}$ and $\hat{b}$ values will produce predicted $Y$'s quite close to the actual $Y$s. Thus, the $a$ and $b$ values are exactly those numbers that best capture the relationships among the variables in the current data set.

For the example data set in Table 14.1, the best values turn out to be:

$$\begin{array}{rcr} \hat{a} & = & \text{-22.609} \\ \hat{b}_1 & = & 2.371 \\ \hat{b}_2 & = & 1.604 \\ \hat{b}_3 & = & 0.332 \end{array}$$

This means that the best predictions of performance are obtained by using the following predictive equation for PERF, which is a specific instantiation of the more general regression prediction equation written below it:

$$\begin{array}{ccccccccccccc} \text{PERF'} & = & \text{-22.609} & + & 2.371 & \times & \text{TRA} & + & 1.604 & \times & \text{EXP} & + & 0.332 & \times & \text{IQ} \\ Y_i' & = & \hat{a} & + & \hat{b}_1 & \times & X_{1i} & + & \hat{b}_2 & \times & X_{2i} & + & \hat{b}_3 & \times & X_{3i} \end{array}$$

### 14.4.2 Predicted $Y$ Values

Once you have the values of $a$ and the $b$'s, the prediction equation is used to compute a predicted $Y$ value for each case. In general, the formula is

$$Y_i' = \hat{a} + \hat{b}_1 \times X_{1i} + \hat{b}_2 \times X_{2i} + \ldots$$

Note the "prime" after the $Y$, which again denotes that it is a predicted value rather than an observed one. For each case, you get an overall predicted performance value like this:

1. plug in the numerical values of TRA, EXP, and IQ for that case,

2. multiply each by its associated value of the $\hat{b}$,
3. add these products, and
4. add in the estimated $\hat{a}$.

For example, the "Predicted" column in Table 14.1 shows the computation of predicted values for that data set, using the best predictive model, already given as:

PERF' = -22.609 + 2.371 × TRA + 1.604 × EXP + 0.332 × IQ

Note that the TRA, EXP, and IQ values for each case are substituted into the right side of the formula.

### 14.4.3 Error Estimates

Once you have the predicted values for all cases, the next step is to compute the error scores, also case by case. As usual, each error value is simply the number needed to set the two sides of the equation to equality. The simplest way to compute this value is with the equation:

$$\hat{e}_i = Y_i - Y'_i \tag{14.1}$$

The "Error" column in Table 14.1 shows the computations of the $\hat{e}_i$ values for that data set.

### 14.4.4 Sums of Squares

As in simple regression, four sums of squares are normally computed in connection with the regression model. The computational formulas for these are:

$$\begin{aligned} SS_{\mu} &= N \times \bar{Y} \times \bar{Y} \\ SS_{model} &= \sum_{i=1}^{N} {Y'_i}^2 - SS_{\mu} \\ SS_{error} &= \sum_{i=1}^{N} {\hat{e}_i}^2 \\ SS_{corrected\ total} &= \sum_{i=1}^{N} {Y_i}^2 - SS_{\mu} \end{aligned}$$

Here are the specific versions of these equations for the data shown in Table 14.1:

$$\begin{aligned} SS_{\mu} &= 15 \times 59.93 \times 59.93 = 53880.07 \\ SS_{model} &= 54740.62 - 53880.07 = 860.55 \\ SS_{error} &= 308.38 \\ SS_{corrected\ total} &= 55049.00 - 53880.07 = 1168.93 \end{aligned}$$

The $SS_{\mu}$ is computed in the same way as in ANOVA. In regression analysis, it is only used to correct other sums of squares when the mean $Y$ is not zero. (*Note: The $SS_{\mu}$ is sometimes referred to as $SS_a$.)*

The interpretation of the $SS_{error}$ is the simplest: It is a measure of the total error in prediction, across all the cases.

The $SS_{corrected\ total}$ can be interpreted as the total variation (in $Y$) that the model is attempting to predict. It is perhaps illuminating to look at a more conceptually informative alternative version of the formula for this quantity:

$$SS_{corrected\ total} = \sum_{i=1}^{N} (Y_i - \bar{Y})^2 \tag{14.2}$$

Note that this sum of squares will be zero if the $Y$ values are all the same—i.e., if they do not vary. To the extent that they do vary, $SS_{corrected\ total}$ will be greater than zero, and the goal of the model is to predict this variation in the $Y$'s.

Finally, the $SS_{model}$ is a measure of the extent to which the model actually does predict the observed variation in $Y$. Again, this is easiest to see when looking at a more conceptually informative alternative version of the formula:

$$SS_{model} = \sum_{i=1}^{N} (Y_i' - \bar{Y})^2 \tag{14.3}$$

Note that this sum of squares will be the same as $SS_{corrected\ total}$ if the predictions are perfect. In that case, the predicted $Y$ will be equal to the observed $Y$ for every case (that's just what is meant by perfect predictions), and so the right sides of Equations 14.2 and 14.3 would be the same.

It is also worth noting that the easiest way to compute $SS_{model}$ is to use this equation, after computing $SS_{corrected\ total}$ and $SS_{error}$:

$$SS_{model} = SS_{corrected\ total} - SS_{error} \tag{14.4}$$

### 14.4.5 Degrees of Freedom

The rules are simple:

$df_{model}$ is the number of predictor variables. In Table 14.1, for example, it is three.

$df_{corrected\ total}$ is the number of cases minus one. One is subtracted because the $SS_{\mu}$ has been subtracted out.

$df_{error}$ is $df_{corrected\ total}$ minus $df_{model}$.

These rules are really the same as in simple regression, but they are applied in situations with two or more predictor variables.

### 14.4.6 Summary ANOVA Table

Table 14.2 shows is the summary ANOVA table for the regression data set in Table 14.1. The general formulas for the df's and SS's were already given above, and as usual each MS is the SS divided by the df on the same line. There is also an $F$ for the model, computed as

$$F_{model} = \frac{MS_{model}}{MS_{error}} \tag{14.5}$$

To look up the critical $F$, the number of degrees of freedom in the numerator is the number of predictors (3, in this example), and the number of dfs in the denominator is the df for error (11, in this example). If the computed $F$ for the model is greater than the critical $F$, you reject the null hypothesis. The interpretations of significant and nonsignificant results for the $F_{model}$ are explained in the next section.

| Source | *df* | *SS* | *MS* | *F* |
|---|---|---|---|---|
| Model | 3 | 860.55 | 286.85 | 10.23 |
| Error | 11 | 308.38 | 28.03 | |
| Total | 14 | 1168.93 | | |

Table 14.2: The summary ANOVA table for the 3-predictor multiple regression model using the data of Table 14.1.

## 14.5 Interpretation & Conclusions

### 14.5.1 The Null Hypothesis

In general, the $F$ for a multiple regression model tests the null hypothesis:

*all the predictors together do not really predict Y.*

For the example in Table 14.1, the null hypothesis is that knowing TRA, EXP, and IQ does not actually allow you to predict PERF any better than chance.

In terms of the parameters of the model, the null hypothesis can be stated as:

*the true values of all the b's are zero.*

Finally, in terms of a comparison between models, the null hypothesis says that:

*the model with all the predictors in it is not really any better than the model* $Y_i = a + e_i$.

### 14.5.2 What to Conclude When $H_0$ is Rejected

When $H_0$ is rejected, the conclusion is that $Y$ values are predicted at better than chance levels from the combination of all predictor variables. This finding is consistent with the idea that at least one of the predictor variables has some causal effect on the $Y$ variable, although it does not prove a causal relationship because of the possibility that some confounding variable—not included in the analysis—provides the causal link between $Y$ and one or more of the $X$'s.

In the example of Table 14.1, we would conclude that on-the-job performance is predicted by the combination of weeks of training, months of experience, and IQ. This is consistent with the idea that one or more of these variables is what *causes* good on-the-job performance, although it does not prove such a causal relationship. For example, the relationship of performance to training and experience might actually be due to the confounding variable of enthusiasm for the job. Perhaps what really determines performance is enthusiasm, and more enthusiastic people also put more time into training and also stick with the job longer (giving them more experience). It is more difficult to come up with a confounding variable to explain the relationship of performance and IQ, but such a confounding is still possible. One possibility is that both performance on the job and score on the IQ test used in this study were related to how well people function under stress. In that case PERF and IQ scores might be positively correlated, even if ability to function under stress is not supposed to be part of the definition of IQ.

### 14.5.3 What to Conclude When $H_0$ is Not Rejected

The conclusion is that $Y$ values are not predicted at better than chance levels from the combination of all predictor variables. This finding is consistent with the idea that none of the predictor variables has any causal effect on the $Y$ variable, although it does not prove that there is no causal relationship because

- the causal relationship may have been too weak to detect with the actual sample size.
- the causal relationship may have been concealed by pooling effects, large measurement error, range restriction, etc.

In the example of Table 14.1, if the computed $F$ had been much smaller (say 1.0), we would have concluded that on-the-job performance is not predicted by the combination of weeks of training, months of experience, and IQ. This would be consistent with the idea that none of these variables has a causal influence on on-the-job performance, although it does not prove that there is no causal relationship for the reasons mentioned above.

## 14.6 Discussion

One of the advantages of multiple regression models over simple regression models is that using several predictors sometimes leads to much more accurate predictions. Consider the example in Table 14.1. From the ANOVA in Table 14.2, we see that the $MS_{error}$ associated with the multiple regression model is 28.03. In contrast, if you fit a one-predictor model to predict PERF from just TRA, you'll get an $MS_{error}$ of 71.05; if you fit a one-predictor model to predict PERF from just EXP, you'll get an $MS_{error}$ of 69.89; if you fit a one-predictor model to predict PERF from just IQ, you'll get an $MS_{error}$ of 68.21. In other words, the prediction error using all three predictors is far less than the prediction error using any one of the predictors.

This same point is illustrated with scattergrams in Figure 14.1. Panels A, B, and C show the scattergrams and one-predictor regressions relating on-the-job performance to each of the three predictors in isolation using the data of Table 14.1. The regression lines shown in these three panels are simple

regressions using just the $Y$ variable and the indicated single $X$ variable and temporarily ignoring the other $X$'s. Note that there is considerable scatter of the observed points around the regression lines—that is, considerable prediction error.

In contrast, panel D shows the relation between predicted values and observed values for the multiple regression model. Clearly, the points tend to be closer to the prediction line than they were in the other three panels, indicating that the predictions are substantially more accurate (less error-prone).

A novel feature of the scattergram in panel D is the horizontal axis: It is the predicted $Y$ value, not one of the $X$'s as in panels A–C. You can see that it is a bit of a challenge to make a scattergram for a multiple regression model, because there is only one horizontal axis and at first it seems that we can only use it to show one $X$ predictor at a time. A useful trick is to put the predicted $Y$ on the horizontal axis, instead of one of the $X$'s. Remember, the predicted $Y$ value takes into account the information from *all* the $X$ predictors, so in a sense this type of scattergram lets you see the relationship of $Y$ to all the predictors at once.

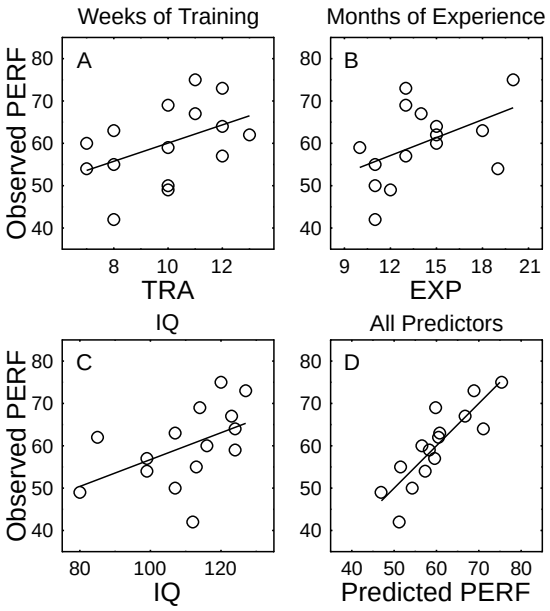


Figure 14.1: Scattergrams showing on-the-job performance as predicted from weeks of training (panel A), from months of experience (panel B), from IQ (panel C), and from all three predictors together in a multiple regression model. The solid line in each panel shows the predictions of the model using the indicated predictors. These figures display the results of analyses using the data of Table 14.1.

# Chapter 15

# Extra Sum of Squares Comparisons

## 15.1 Overview and Example

As already discussed, a limitation of multiple regression models is that they only provide an overall $F$ to test whether the $Y$ variable is predicted by the model as a whole, including all its predictor variables. Usually, it much more important to separate out the individual contributions of specific predictor variables. The technique of making extra sum of squares comparisons (ESSCs, for short) provides a way to do this, at least to some extent.

You might wonder why we need a new technique to look at the contributions of individual predictors, given that we have already seen how a simple one-variable regression can be used to study the relation between $Y$ and an individual predictor. The reason goes back to the discussion of soccer versus cricket in Section 14.3. Different predictors work together to predict $Y$ in multiple regression, just as different players work together to score in soccer. By looking at an individual predictor in a one-variable regression, all we can tell is how well that predictor does on its own. The new technique will enable us to see how well a certain predictor does as part of a certain team, and that is different. This idea will be illustrated with a further example soon.

Computationally, an ESSC is always a comparison between two regression models. There is a "smaller model" with a certain set of predictors, and it is compared against a "larger model" with that same set of predictors plus at least one extra predictor. The test allows you to determine whether the larger model is significantly better than the smaller model; if so, you can conclude that the extra predictor(s) do indeed make significant separate and unique contributions to the overall prediction.

For example, suppose you are trying to determine whether students' success in a statistics class depends on their knowledge of maths. One simple study would be:

**Study 1:** At the beginning of a statistics class, you give each student a maths achievement test (MAT). At the end of the statistics class, you obtain each student's percentage mark for the class (STAT). You then carry out a simple linear regression to try to predict STAT from MAT.

If it turns out that knowledge of maths does *not* predict statistics marks, you will conclude that statistics marks do no strongly depend on knowledge of maths, subject to the usual caveats.

If it turns out that knowledge of maths *does* predict statistics marks, however (as seems likely!), your conclusions will be weaker. This result is consistent with the idea that success in statistics does depend on maths knowledge, but it is very weak support for that idea because you have ignored an obvious mediating factor: namely, overall ability (as opposed to math-specific ability, which you claim to be focusing on). It could be that high-ability students tend to do better in all subjects, and that the relationship between maths knowledge and statistics marks results from the common influence of ability on both of them, not from an influence of maths knowledge on statistics scores. To put it more concretely, we might have found just as strong a relationship to STAT if we had measured achievement on a verbal test or a history test, instead of a maths test. Thus, we cannot be sure that success in statistics really depends on knowledge of maths per se.

A better study could be conducted to try to control for overall student ability as follows:

**Study 2:** At the beginning of a statistics class, you give each student a maths achievement test (MAT). You also obtain their overall average score in other university classes (AVG) as a measure of overall ability. At the end of the statistics class, you obtain each student's percentage mark for the class (STAT). Now, you use multiple regression to see how the statistics mark is related

to both of the other two variables. Specifically, the question is whether STAT is specifically related to knowledge of maths *beyond* what is explainable in terms of overall ability (AVG). This question can be answered with an ESSC, even though it cannot be answered by the simple regression of Study 1.

A hypothetical data set for Study 2 is shown in Table 15.1. As discussed in Chapter 14, you can obtain the best-fitting two-predictor multiple regression model from a computer program, which for these data yields the model:

$$\text{STAT} = -6.724 + 0.422 \times \text{MAT} + 0.644 \times \text{AVG} \tag{15.1}$$

Then, these $\hat{a}$ and $\hat{b}$ values can be used to compute the predicted values and errors shown in Table 15.1. The overall summary ANOVA for the two-predictor regression is shown in Table 15.2.

| Case | STAT | MAT | AVG | Predicted | | Error | |
|---|---|---|---|---|---|---|---|
| 1 | 67 | 70 | 75 | 71.12 | = -6.72 + 0.42 × 70 + 0.64 × 75 | -4.12 | = 67 - 71.12 |
| 2 | 56 | 62 | 66 | 61.94 | = -6.72 + 0.42 × 62 + 0.64 × 66 | -5.94 | = 56 - 61.94 |
| 3 | 62 | 61 | 67 | 62.17 | = -6.72 + 0.42 × 61 + 0.64 × 67 | -0.17 | = 62 - 62.17 |
| 4 | 65 | 62 | 74 | 67.10 | = -6.72 + 0.42 × 62 + 0.64 × 74 | -2.10 | = 65 - 67.10 |
| 5 | 85 | 79 | 80 | 78.13 | = -6.72 + 0.42 × 79 + 0.64 × 80 | 6.87 | = 85 - 78.13 |
| 6 | 80 | 74 | 88 | 81.18 | = -6.72 + 0.42 × 74 + 0.64 × 88 | -1.18 | = 80 - 81.18 |
| 7 | 77 | 68 | 70 | 67.05 | = -6.72 + 0.42 × 68 + 0.64 × 70 | 9.95 | = 77 - 67.05 |
| 8 | 75 | 66 | 76 | 70.07 | = -6.72 + 0.42 × 66 + 0.64 × 76 | 4.93 | = 75 - 70.07 |
| 9 | 59 | 53 | 58 | 52.99 | = -6.72 + 0.42 × 53 + 0.64 × 58 | 6.01 | = 59 - 52.99 |
| 10 | 64 | 65 | 79 | 71.58 | = -6.72 + 0.42 × 65 + 0.64 × 79 | -7.58 | = 64 - 71.58 |
| 11 | 67 | 57 | 71 | 63.05 | = -6.72 + 0.42 × 57 + 0.64 × 71 | 3.95 | = 67 - 63.05 |
| 12 | 73 | 74 | 78 | 74.74 | = -6.72 + 0.42 × 74 + 0.64 × 78 | -1.74 | = 73 - 74.74 |
| 13 | 67 | 66 | 71 | 66.85 | = -6.72 + 0.42 × 66 + 0.64 × 71 | 0.15 | = 67 - 66.85 |
| 14 | 62 | 71 | 75 | 71.54 | = -6.72 + 0.42 × 71 + 0.64 × 75 | -9.54 | = 62 - 71.54 |
| 15 | 55 | 61 | 67 | 62.17 | = -6.72 + 0.42 × 61 + 0.64 × 67 | -7.17 | = 55 - 62.17 |
| 16 | 78 | 57 | 88 | 74.00 | = -6.72 + 0.42 × 57 + 0.64 × 88 | 4.00 | = 78 - 74.00 |
| 17 | 77 | 74 | 70 | 69.58 | = -6.72 + 0.42 × 74 + 0.64 × 70 | 7.42 | = 77 - 69.58 |
| 18 | 58 | 68 | 62 | 61.90 | = -6.72 + 0.42 × 68 + 0.64 × 62 | -3.90 | = 58 - 61.90 |
| Sum: | 1227.00 | | | 1227.00 | | 0.00 | |
| Mean: | 68.17 | | | 68.17 | | 0.00 | |
| SumSq: | 85003.00 | | | 84433.85 | | 569.15 | |

Table 15.1: Hypothetical data for a sample of 18 students. The goal was to find out whether the student's mark in a statistics class (STAT) was related to their knowledge of maths (MAT), also taking into account their overall average university mark (AVG). A computer program reports that the best-fitting two-predictor model is STAT $= -6.724 + 0.422 \times$ MAT $+ 0.644 \times$ AVG, and the predicted values and error were computed using this model.

$$\begin{aligned} SS_{\mu} &= 18 \times 68.17 \times 68.17 = 83640.50 \\ SS_{model} &= 84433.85 - 83640.50 = 793.35 \\ SS_{error} &= 569.15 \\ SS_{corrected\ total} &= 85003.00 - 83640.50 = 1362.50 \end{aligned}$$

| Source | *df* | *SS* | *MS* | *F* |
|---|---|---|---|---|
| Model | 2 | 793.35 | 396.68 | 10.45 |
| Error | 15 | 569.15 | 37.94 | |
| Total | 17 | 1362.50 | | |

Table 15.2: Further multiple regression computations and summary ANOVA table for the two-predictor model shown in Table 15.1.

For this example, the significant $F_{model}$ in Table 15.2 indicates that the two-predictor model does indeed predict statistics marks at a better-than-chance level. Of course this is not at all surprising. Even if statistics mark were totally unrelated to maths knowledge, we would still expect STAT to be predictable from AVG. Students with higher overall marks averaged across all classes should score relatively well in statistics classes too; the motivation, intelligence, study habits, and so on that helped them do well in other classes should also tend to help in statistics. Thus, the overall $F_{model}$ tells us nothing about whether statistics marks depend specifically on maths knowledge.

Within the context of this example, the crucial question is whether maths knowledge makes a significant unique contribution to the predictions of the two-predictor model. What we really want to know is whether we can predict better using *both* AVG and MAT than just using AVG by itself. If so, it must be because statistics mark is specifically related to maths knowledge, not just related to average overall mark of all types. In short, the question is whether we really need MAT in the two-predictor model.

This is exactly the sort of question that the ESSC was designed to answer. In terms of our earlier definition, the smaller model is the model with just AVG, and the larger model uses both AVG and MAT. By comparing these two models, we can see if MAT improves predictions beyond those available from just AVG. If it does, then we can conclude that the data demonstrate a separate predictive contribution of MAT, above and beyond the predictive ability provided by AVG.

## 15.2 Computations

Table 15.3 provides a schematic diagram of the extra sum of squares comparison. The middle two columns of the table represent the smaller and the larger of the two models being compared, respectively. Listed in each column are the model's predictors, its sum of squares, and its dfs. The column at the far right represents the comparison between the two models, listing on each line the additional or "extra" portion that is present in the larger model but not in the smaller one. The most crucial value in this column is the "extra sum of squares" or $SS_{extra}$. Intuitively, this represents the improvement in predictability brought about by adding the extra predictor(s) into the model, over and above the predictability obtainable with the smaller model. The $F$ in the lower panel of the table will be explained in a moment.

Table 15.5 shows how these columns are filled in for the extra sum of squares comparison to see whether MAT improves the prediction of STAT over and above using only AVG.

- The smaller model is the one-predictor model predicting STAT from AVG. Its sum of squares and degrees of freedom are computed using the standard formulas associated with a simple (i.e., one-predictor) regression, computed as if STAT and AVG were the only two variables in the data set. This model and its ANOVA table are shown in Table 15.4[1]. The SS and df for this model are simply copied from that table into the "smaller model" column of Table 15.5.

- The larger model is the two-predictor model summarized in Table 15.2. Its sum of squares and degrees of freedom are copied into the "larger model" column of Table 15.5.

- The right-most column is then formed by taking the difference between the second and third columns. Intuitively, the "extra sum of squares" in this column represents the improvement in predictability due to adding MAT, over and above the predictability obtainable from just AVG.

| *Extra Sum of Squares Comparison:* | | | |
|---|---|---|---|
| | **Smaller model** | **Larger model** | **Extra** |
| Predictors: | smaller set | smaller set additional | additional |
| SS model: | $SS_{smaller}$ | $SS_{larger}$ | $SS_{extra} = SS_{larger} - SS_{smaller}$ |
| df model: | $df_{smaller}$ | $df_{larger}$ | $df_{extra} = df_{larger} - df_{smaller}$ |
| Extra $F = \frac{SS_{extra}/df_{extra}}{MS_{error,\ larger\ model}} = F_{observed}$ for additional predictors | | | |

Table 15.3: The general pattern used in making an extra sum of squares comparison.

As in this example, the $SS_{larger}$ is always numerically larger than the $SS_{smaller}$, so the $SS_{extra}$ is always positive. Conceptually speaking, the larger model *always* does at least a little bit better. But this may be just due to chance, because it has an extra degree of freedom to use in fitting the data. The question is: *Is the larger model enough better that we can exclude chance as the explanation of the improvement?*

To answer this question, we compute an Extra $F$. The general version of the Extra $F$ computation is shown at the bottom of Table 15.3. The numerator of the Extra $F$ is the $SS_{extra}$ divided by the

[1] We will not describe the computation of this simple regression because this topic has already been covered.

STAT = 10.40 + 0.79 × AVG

| Source | *df* | *SS* | *MS* | *F* |
|---|---|---|---|---|
| Model | 1 | 669.62 | 669.62 | 15.46 |
| Error | 16 | 692.88 | 43.30 | |
| Total | 17 | 1362.50 | | |

Table 15.4: Summary ANOVA table for a simple regression model predicting STAT from AVG with the data shown in Table 15.1.

| *Extra Sum of Squares Comparison:* | | | |
|---|---|---|---|
| | **Smaller model** | **Larger model** | **Extra** |
| Predictors: | AVG | AVG MAT | MAT |
| SS model: | 669.62 | 793.35 | 123.73 |
| df model: | 1 | 2 | 1 |
| Extra $F = \frac{123.73/1}{37.94} = 3.26$ for adding MAT | | | |

Table 15.5: Computation of extra sum of squares comparison to see whether success in statistics is related to knowledge of maths.

$df_{extra}$, often called the "extra mean square" or $MS_{extra}$[2]. The denominator is always the $MS_{error}$ from the larger model. It makes sense to use the $MS_{error}$ from the larger model rather than from the smaller model, because the smaller model's $MS_{error}$ could be inflated (thereby reducing the $F$ and the chances of a significant result) if it ignores important predictors that are included in the larger model. (This comment about inflating error will be clearer after we have covered the topic of "error reduction" in section 16.3, but for now just make the analogy to ANOVA: If we omitted factors from an ANOVA, their effects went into the error term and reduced $F$'s and power.)

An example of the Extra $F$ computation is shown at the bottom of Table 15.5. The $SS_{extra}$ and $df_{extra}$ in the numerator of the $F$ are taken directly from the right-most column of the table. The $MS_{error}$ in the denominator is taken from the ANOVA table summarizing the fit of the larger model—in this case, Table 15.2.

As usual, the Extra $F$ is compared against a critical $F$ from the $F$ table. The critical $F$ has $df_{extra}$ degrees of freedom in the numerator, because the $MS_{extra}$ is associated with $df_{extra}$ degrees of freedom. In the denominator, it has the number of degrees of freedom associated with the $MS_{error}$ of the larger model, which is the number of cases minus one minus the number of predictors in the larger model. In the current example, then, the critical $F$ has 1 and 15 dfs, so $F_{critical} = 4.54$. The observed $F$ of 3.26 is less than $F_{critical}$, so the null hypothesis is not rejected.

## 15.3 Conclusions & Interpretation

### 15.3.1 The Null Hypothesis

The Extra $F$ tests the null hypothesis that:

> *the extra predictors add nothing to the predictability of Y given that the predictors in the smaller model are already being used.*

In terms of the parameters of the model, this can also be stated as:

> *the true values of the b's for the added predictors in the larger model are really zero.*

In terms of a model comparison, this can also be stated as:

> *the larger model is really no better than the smaller model.*

[2]In most situations the larger model has just one more predictor than the smaller one, so the $df_{extra}$ is one and the $MS_{extra}$ equals the $SS_{extra}$. But if the larger model has two or more extra predictors relative to the smaller one, the extra sum of squares has to be divided by the number of extra predictors, just as all sums of squares are divided by their degrees of freedoms before computation of an $F$. See Section 15.4.3 for an example.

### 15.3.2 Interpretation of a Significant Extra $F$

When the Extra $F$ is significant, the interpretation is:

> There is a separate predictive relationship between $Y$ and at least one of the extra predictors beyond that explainable in terms of the predictors in the smaller model. This is consistent with the hypothesis that one or more of the extra predictors has an additional causal influence on $Y$ beyond any influences related to the $X$'s in the smaller model. The causal influence is not proved, however, because there may be other mediating variables not included in the smaller model.

As a matter of terminology, common alternative phrasings of this conclusion sound like this:

- There is a separate predictive relationship between $Y$ and the extra predictors *that is not mediated by* the predictors in the smaller model.
- There is a separate predictive relationship between $Y$ and the extra predictors *controlling for* the predictors in the smaller model.
- There is a separate predictive relationship between $Y$ and the extra predictors *taking into account* the predictors in the smaller model.
- There is a separate predictive relationship between $Y$ and the extra predictors *partialling out* the predictors in the smaller model.

All of these phrasings refer to the same extra sum of squares comparison.

In the example involving the statistics mark, if the extra $F$ had been significant the interpretation would have been:

> There is a significant predictive relationship between maths knowledge and statistics mark "beyond that explainable in terms of" overall average university mark *(or "not mediated by" or "controlling for").* That is, we know that the relationship between statistics marks and maths knowledge is not completely explainable as "better overall students have more maths knowledge and also get better statistics marks." (This rules out the mediating variable explanation that would have contaminated Study 1, as discussed earlier.) The unique relationship of maths knowledge and statistics mark, controlling for overall average, is consistent with the idea that statistics mark is causally determined by maths knowledge. That causal relationship is not proved, however, because perhaps some other variable mediates the relationship between statistics marks and maths knowledge.

What other variable might mediate this relationship? Here is one possible scenario: Suppose that what really makes people good at statistics is a tendency to think in a very orderly fashion, and this "mental orderliness" also tends to make them score higher on the maths test. In that case MAT and STAT could be related because both were causally determined by orderliness, not because math knowledge had a causal influence on statistics mark. So, orderliness could be the mediating variable in this case. In the end, if you really want to be certain of a causal relationship, then you just have to do an experiment:

**Experiment:** Assign students randomly to two groups. Pre-teach the students in one group ("high maths knowledge group") some extra maths, and teach the other group ("control group") something else (Latin?) to control for placebo effects, generalized learning experience, and so on. After both groups had been pre-taught, teach them statistics and see if their statistics performance levels were different at the end.

Obviously, this experiment is much more work than either of the correlational studies described earlier, because of the pre-teaching that is required.

But before we go on, here is some reassurance for the mathematically challenged: Take heart, this $F$ is generally not significant in real studies looking at this relationship (e.g., Ware & Chastain, 1989). That is, real data do not suggest any specific relationship between maths skill and statistics mark.

### 15.3.3 Interpretation of a Nonsignificant Extra $F$

When the Extra $F$ is not significant, the interpretation is:

> There is no clear evidence of a separate predictive relationship between $Y$ and the extra predictors beyond that explainable in terms of the predictors in the smaller model. This is consistent with the hypothesis that none of the extra predictors has an additional causal influence on $Y$ beyond the influences related to the $X$'s in the smaller model. The absence of a causal influence is not proved, however, because the causal relationship may have been too weak to detect with the actual sample size or may have been concealed by pooling effects, large measurement error, range restriction, etc.

In the example involving the statistics mark, the interpretation of the nonsignificant is:

> There is no clear evidence of a significant predictive relationship between maths knowledge and statistics mark beyond that explainable in terms of overall average university mark. This is consistent with the idea that statistics mark is not causally determined by maths knowledge. The absence of a causal relationship is not proved, however, because the causal relationship may have been too weak to detect with the actual sample size or may have been concealed by pooling effects, large measurement error, range restriction, etc.

## 15.4 Discussion

### 15.4.1 Measurement by subtraction.

The idea of an extra sum of squares comparison is a new version of an old, familiar procedure: Measurement by subtraction. For example, how would you determine the weight of a liter of oil? You would find a large enough container, weigh it empty, pour in the oil, weigh it again, and see how much the weight had gone up. That's quite analogous to an extra sum of squares comparison, except you are measuring weight rather than "sum of squares explained by the model." The container is analogous to the smaller model: First, you measure it by itself. Then you add the portion you are really interested in (oil or predictors), and measure the new total. Finally, the two measurements are subtracted, and the difference is taken as the measurement of the added part.

Measurement by subtraction also underlies most before/after comparisons. Suppose you test your strength in January, undergo a weight-training program for six months, and then test your strength again in July. The difference in strength from before to after would be a measure of the effectiveness of the program. Your original self is the smaller part, and yourself augmented by the training program is the larger one, with an increment due to the program.

Another example, discussed in Chapter 14, was the example of measuring a player's offensive contribution in a game where players must work together to score, like soccer. Although impractical at best, the idea of measurement by subtraction could be applied in this case as well. Imagine that we wanted to evaluate Joe Bloggs's offensive contribution. We could have Joe's team play (say) eight games against another team, with Joe playing in four of the games and sitting out the other four. We could then find the difference between the points scored by his team when he was playing minus the points scored by those same players when he was sitting out. Obviously, if the team scored lots more when he was playing, then he must have been making a big contribution to the scoring—even if he never actually scored a goal personally[3].

### 15.4.2 Controlling for potential confounding variables.

Extra sum of squares comparisons allow you to examine the relationship between $Y$ and a certain $X$ variable (or group of $X$ variables) *controlling for* the effects of certain other $X$ variables. In essence, this allows you to take into account some of the confounding or mediating variables that weaken interpretations in correlation and regression analysis. When they are taken into account, they are eliminated as alternative explanations.

In the statistics mark study, for example, you could see whether there was a specific relation between statistics mark and math knowledge that was not just due to the correlation of both of these

[3] I am assuming that eight games would be enough for the averages to be pretty stable (i.e., for random error to mostly average out). If not, we'd need more games, but the basic idea would be the same.

variables with overall university mark, which presumably takes account of various general study habits, etc, that are not specifically associated with math knowledge. If you got a significant STAT/MAT relationship after you eliminated this mediating correlation, you would have better evidence that maths knowledge really does causally influence statistics mark (better than if you had not eliminated this mediating variable).

As another example, suppose you were investigating the relationship between life expectancy and coffee drinking. For each person in your sample, you record how long they lived and also somehow determine how much coffee they drank. Computing a simple regression, you find that life expectancy is significantly negatively associated with amount of coffee drunk. In other words, people who drink more coffee do not live as long. Of course you cannot conclude that there is a causal relationship—that drinking lots of coffee shortens your life. This is because the coffee/life-expectancy relationship could be mediated by some other variable—very possibly amount smoked. We know that people who tend to smoke more also tend to drink more coffee, and it could be the extra smoking rather than the extra coffee that is reducing life expectancy. This is one possible way to explain the coffee/life-expectancy relationship without a direct causal link from coffee to health.

To rule out this explanation of the coffee/life-expectancy relationship, you could measure smoking and control for it statistically. In practice, this means making an extra sum of squares comparison of the form shown in Table 15.6. The smaller model includes just the amount smoked, which is the variable you want to control for. The larger model includes amount of coffee drunk in addition. Thus, the difference is a measure of the specific predictive association between coffee drinking and life expectancy *beyond that mediated by the amount smoked.*

| *Extra Sum of Squares Comparison:* | | | |
|---|---|---|---|
| | **Smaller model** | **Larger model** | **Extra** |
| Predictors: | SMOK | SMOK COFF | COFF |
| SS model: | $SS_{SMOK}$ | $SS_{SMOK+COFF}$ | $SS_{extra} = SS_{COFF}$ |
| df model: | 1 | 2 | 1 |
| Extra $F = \frac{SS_{extra}/1}{MS_{error,\ larger\ model}} = F_{observed}$ for COFF | | | |

Table 15.6: Format of a possible extra sum of squares comparison to test for a specific effect of amount of coffee drunk (COFF) on life expectancy, controlling for amount smoked (SMOK).

### 15.4.3 When would you use more than one "extra" predictor?

In all of the comparisons discussed so far, the larger model has exactly one more predictor than the smaller model, and indeed this is by far the most common situation. However, the basic idea of the ESSC can also be applied to evaluate the specific contribution of a group of variables, simply by comparing a smaller model against a larger model with more than one additional variable. This is usually done only when a group of variables all measure the same sort of thing.

For example, suppose we were trying to evaluate environmental and genetic contributions to IQ. We obtain a sample of adopted children, measure their IQs, the IQs of their adoptive parents, and the IQs of their biological parents. In principle we could measure both mothers' and fathers' IQs, yielding a data set containing five variables:

- Child's IQ ($Y$)
- AdoM, Adoptive mother's IQ ($X_1$)
- AdoF, Adoptive father's IQ ($X_2$)
- BioM, Biological mother's IQ ($X_3$)
- BioF, Biological father's IQ ($X_4$)

With respect to our interest in environmental versus genetic determinants of IQ, it seems clear that we have two predictors of each type: The adoptive mother's and father's IQs should be related to the child's IQ to the extent that there is an environmental influence. Conversely, the biological mother's and father's IQs should be related to the extent that there is a genetic influence. Because the variables

are paired up like this, it would be sensible to make extra sum of squares comparisons evaluating the specific predictive contributions of *both* environmental predictors at once, as shown in Table 15.7, or of both genetic predictors at once, as shown in Table 15.8.

| *Extra Sum of Squares Comparison:* | | | |
|---|---|---|---|
| | **Smaller model** | **Larger model** | **Extra** |
| Predictors: | BioM+BioF | BioM+BioF AdoM+AdoF | AdoM+AdoF |
| SS model: | $SS_{BioM+BioF}$ | $SS_{BioM+BioF+AdoM+AdoF}$ | $SS_{extra} = SS_{AdoM+AdoF}$ |
| df model: | 2 | 4 | 2 |
| Extra $F = \frac{SS_{extra}/2}{MS_{error,\ larger\ model}} = F_{observed}$ for AdoM+AdoF | | | |

Table 15.7: Format of a possible extra sum of squares comparison to test for evidence of a specific environmental effect on IQs of adopted children.

For example, Table 15.7 shows the ESSC to test for an environmental influence on IQ. Both of the "genetic" predictors are in the smaller model, so they are controlled for in the comparisons. Then, both of the environmental variables are added at once, so we can see if knowing a lot about the environment improves the prediction.

Suppose this ESSC yields a significant extra $F$. In that case, we could conclude that there is a unique relationship between the child's IQ and the IQ of at least one of the adoptive parents, above and beyond any relationship of the child's IQ with the IQs of its biological parents. This would support the idea that IQ is at least partly determined by the environment.

On the other hand, suppose this ESSC does not yield a significant extra $F$. In that case, we could conclude that neither of adoptive parents' IQs has a separate, unique relationship to the child's IQ, above and beyond any relationship of the child's IQ with the IQs of its biological parents. This would be consistent with the idea that IQ is not very strongly determined by the environment.

| *Extra Sum of Squares Comparison:* | | | |
|---|---|---|---|
| | **Smaller model** | **Larger model** | **Extra** |
| Predictors: | AdoM+AdoF | AdoM+AdoF BioM+BioF | BioM+BioF |
| SS model: | $SS_{AdoM+AdoF}$ | $SS_{AdoM+AdoF+BioM+BioF}$ | $SS_{extra} = SS_{BioM+BioF}$ |
| df model: | 2 | 4 | 2 |
| Extra $F = \frac{SS_{extra}/2}{MS_{error,\ larger\ model}} = F_{observed}$ for BioM+BioF | | | |

Table 15.8: Format of a possible extra sum of squares comparison to test for evidence of a specific genetic effect on IQs of adopted children.

Conversely, suppose the extra sum of squares comparison in Table 15.8 yields a significant $F$. In that case, we could conclude that the IQs of the biological parents do have a separate, unique relationship to the child's IQ, above and beyond any relationship of the child's IQ with the IQs of its adoptive parents. This would support the idea that IQ is at least partly determined by genetics. Of course it is possible that IQ is determined by both the environment and genetics, in which case both of the preceding extra sum of squares comparisons might be significant.

In summary, it makes sense to add two or more predictor variables together when you think they all reflect the type of effects you are trying to assess—e.g., environmental or genetic.

## 15.5 $F_{add}$ and $F_{drop}$

The extra sum of squares comparison plays such an important role in regression analysis that its $F$ value is referred to with different names in different situations (just as the Eskimos have several different words for *snow*). This section explains the most common alternative names for the "extra $F$". The alternatives will be used extensively in Chapter 17.

As we have seen, the extra sum of squares comparison always involves a larger model and a smaller model. The comparison can be stated as either:

- "Is the larger model better than the smaller one?" or
- "Is the smaller model worse than the larger one?"

Even though these two statements are equivalent mathematically, in practice many people use different terminology depending on whether the smaller model was fit first, and then the larger, or the reverse. As summarized in Table 15.9, when the smaller model is fit first and the larger is fit second, the $F_{extra}$ is often called an "$F$ to add" or an "$F$ to enter." When the larger model is fit first and the smaller is fit second, the $F_{extra}$ is called an "$F$ to drop" or an "$F$ to remove."

| Order of model fits | Names for $F_{extra}$ |
|---|---|
| Smaller then larger | "$F$ to add" or "$F$ to enter" |
| Larger then smaller | "$F$ to drop" or "$F$ to remove" |

Table 15.9: Terminology for the $F_{extra}$ depending on the order of model fits.

Sometimes, the distinction between adding and dropping variables is inherent in the type of analysis being carried out. You will see some examples of this in Chapter 17, which presents some procedures for building good regression models by adding or dropping one predictor at a time.

At other times, the distinction between adding and dropping variables is really only a matter of perspective, like the cup that is half empty or half full. As an example of how perspective plays a role, the following two cases illustrate two slightly different situations where you would be more likely to talk about the "$F$ to add" (case 1) or the "$F$ to drop" (case 2). Clearly, the situations are not very different, and you would be getting the same $F$ value from the same ESS comparison in both cases. Because of your slightly different perspective in the two cases, however, you would probably use different terms to describe this $F$.

**Case 1:** A researcher named Joe Bloggs proposes that IQ is determined only by genetics, not by the environment. You are skeptical and want to show that the environment also plays a role. So, you first fit a model predicting IQ from genetic variables alone, and then show that adding in environmental variables makes the model predict significantly better.

In this example, the "starting model" is the smaller, genetics-only model proposed by Joe Bloggs. Your focus is on showing that the model is improved by the addition of the extra environmental variables. Thus, to emphasize your point, you might well describe the significant $F_{extra}$ as a significant "$F$-to-add" for the environmental variables. The phrase "$F$-to-enter" is used equivalently; the idea is that the model is improved by entering additional predictors into it.

**Case 2:** A researcher named Joe Bloggs proposes that IQ is determined by both genetics and the environment. You are skeptical and want to show that the environment plays no role. So, you first fit a model predicting IQ from both types of variables, and then show that removing the environmental variables does not make the model predict significantly worse.

In this example, the "starting model" is the larger model with both types of predictors, as proposed by Joe Bloggs. Your focus is on showing that the model is not worsened by removing the environmental variables. Thus, to emphasize your point, you might well describe the non-significant $F_{extra}$ as a nonsignificant "$F$-to-drop" for the environmental variables. The phrase "$F$-to-remove" is used equivalently; the idea is that the model not worsened by removing certain predictors from it.

# Chapter 16

# Relationships Among Predictors

Thus far, we have concentrated on the relationship between $Y$ and each of the predictors used in the multiple regression model. In this chapter, we consider something of equal importance in multiple regression analysis: the relationships of the different predictor variables to one another. To simplify the discussion, we will for the time being consider only situations where $Y$ has positive or zero correlations with all the predictors variables. After analyzing the effects of various types of relationships in this simpler situation first, we will consider in Section 16.6 what happens when $Y$ is negatively correlated with some of the predictors.

## 16.1 Context Effects

This chapter discusses several specific patterns of results that can arise when extra sum of squares comparisons are made in multiple regression analysis. These are not the only possible patterns of results, but they are worthy of discussion because they can be quite puzzling. Some of the patterns are reasonably common, but at least one is quite rare. Nonetheless, they are all worth understanding, because they illustrate the different ways in which the relationships among predictors can influence the results.

Each of the patterns is an illustration of what I call a "context effect." A context effect occurs when the apparent usefulness of a certain predictor depends on which other predictor variables are included in the model. In terms of an ESS comparison, I think of the predictors in the smaller model as the "context" in which the extra predictor is being evaluated, and say that the context has an effect if $F$ for that extra predictor changes depending on which variables are in the smaller model.

Context effects are puzzling because they violate our preconceptions about what sorts of results we are likely to get when we perform "measurement by subtraction." Weighing a liter of oil, for example, you would expect to get just the same weight whichever container you used, because the weight of the container is supposed to cancel out in the subtraction. You would be very puzzled if you weighed the same liter of oil in two different containers and found, for example, that it weighed twice as much when measured in one container as when measured in another container. But exactly that sort of effect can happen in extra sum of squares comparisons—we will even see an example analogous to having the oil completely weightless in one measurement but quite heavy in another one!

As you can probably guess from the previous paragraph, weighing oil is not a great analogy for an extra sum of squares comparison in regression, even though it does illustrate measurement by subtraction. The disanalogy is that oil has a "real" weight independent of its container. Predictors don't, because the "unique" predictive relationship associated with a predictor depends on what other predictors are being used already. Evaluating Joe Bloggs's contribution to a soccer team's offense is a better example, because you can imagine that his contribution would depend on which other players were used on his team. For example, he might fit very well with a team of highly aggressive and unpredictable players—he might contribute a lot to their scoring—and that team might score a lot more with him than without him. Alternatively, he might fit very poorly with a team of more patient and methodical players, and they might score about as well without him as with him. The point is that Joe's "offensive contribution" is not just determined by his own characteristics, but also by how these characteristics mesh with the other players on his team. The same is true for predictors, as illustrated by the various kinds of context effects discussed below.

## 16.2 Redundancy

One puzzling pattern of results that can be obtained in a multiple regression situation is a combination of these two basic outcomes:

1. A certain predictor, say $X_2$, is a significant predictor of $Y$ when considered as the only predictor.
2. The ESS for this predictor is not significant when the predictor is added to a smaller model with another predictor, say $X_1$.

Together, these two outcomes seem to create a paradox. The first outcome tells us that $X_2$ is a good predictor of $Y$, and yet the second outcome tells us that $X_2$ is *not*. How can these both be true?

In general terms, the explanation for this pattern is that $X_1$ and $X_2$ are redundant. By "redundant," I mean that they convey approximately the same predictive information about $Y$. Since they convey the same information, the model with both variables is not really any better than the model with just $X_1$, producing the nonsignificant ESS comparison mentioned in outcome 2 above, even if $X_2$ is a good predictor by itself.

As a silly example, suppose you wanted to predict people's weights from two predictors: $X_1$ is the person's height in centimeters, and $X_2$ is his/her height in inches. Because taller people tend to weigh more, both $X_1$ and $X_2$ would be good predictors in isolation. Your predictions would be no better with both predictors than with just one, however, because the two predictors are completely redundant—they both just measure height.

As a more realistic example, consider this thought problem that I've given to hundreds of students over the years. About three-quarters get it right—give it a try:

> Imagine that you are trying to predict the heights of daughters. You have already measured one predictor variable—their mother's heights. And now you are going to measure a second predictor variable which must be either (a) their maternal grandfather's height, or (b) their paternal grandfather's height. Your problem is to choose whichever one of these second predictors would help most with the predictions. Given that you are already using the mother's height, then, which second variable do you think would be better—the maternal or paternal grandfather?

*My answer:* It seems reasonable to assume that a daughter's height is approximately equally determined by the genes from both sides of the family: her mother's and her father's. Once we use the mother's height as a predictor, we have excellent information about the genetic contribution from that side. Adding the maternal grandfather's height would give us somewhat better information about the mother's side, but I suspect it would not help very much because we already know a lot about the mother's side (i.e., we know the mother's side). In contrast, at this point we know nothing at all about the father's side. That's why I'd prefer the paternal grandfather to the maternal one. In essence, I am suggesting that a first person from the father's side is likely to be more useful than a second person from the mother's. That is, I think you would do better to use the *paternal* grandfather's height as the second predictor.

| DauHt | MomHt | DadHt | MGFHt | PGFHt | DauSES | DauInc | MomSES |
|---|---|---|---|---|---|---|---|
| 165 | 162 | 185 | 161 | 183 | 56 | 29093 | 58 |
| 164 | 164 | 171 | 167 | 169 | 47 | 22676 | 50 |
| 170 | 175 | 179 | 174 | 166 | 62 | 35386 | 75 |
| 172 | 163 | 174 | 183 | 169 | 56 | 33891 | 40 |
| 157 | 158 | 179 | 156 | 170 | 67 | 35230 | 76 |
| 157 | 162 | 180 | 168 | 183 | 60 | 28244 | 58 |
| 164 | 169 | 185 | 171 | 168 | 61 | 30801 | 66 |
| 171 | 178 | 182 | 173 | 189 | 50 | 26872 | 67 |
| 177 | 175 | 180 | 162 | 185 | 48 | 23796 | 69 |
| 172 | 183 | 160 | 178 | 165 | 66 | 36843 | 70 |
| 173 | 168 | 182 | 170 | 184 | 68 | 34047 | 64 |
| 160 | 159 | 181 | 159 | 183 | 69 | 31046 | 61 |
| 176 | 169 | 168 | 165 | 176 | 55 | 28572 | 53 |
| 171 | 178 | 172 | 165 | 166 | 68 | 33679 | 78 |
| 170 | 179 | 159 | 179 | 165 | 61 | 30909 | 62 |
| 187 | 170 | 182 | 171 | 173 | 59 | 30298 | 43 |
| 175 | 165 | 169 | 159 | 173 | 66 | 34305 | 61 |
| 181 | 178 | 191 | 181 | 180 | 60 | 29115 | 58 |
| 155 | 165 | 162 | 167 | 171 | 50 | 24419 | 66 |
| 179 | 176 | 180 | 174 | 179 | 70 | 34762 | 68 |
| 184 | 171 | 176 | 154 | 176 | 74 | 38623 | 59 |
| 166 | 179 | 164 | 170 | 167 | 35 | 16403 | 54 |
| 159 | 167 | 177 | 175 | 173 | 50 | 29159 | 61 |
| 166 | 176 | 171 | 166 | 167 | 60 | 35382 | 70 |
| 188 | 188 | 180 | 182 | 171 | 65 | 31747 | 64 |
| 164 | 164 | 156 | 172 | 165 | 76 | 34734 | 58 |
| 179 | 190 | 167 | 179 | 167 | 53 | 27273 | 66 |
| 164 | 155 | 188 | 158 | 194 | 56 | 26526 | 51 |
| 173 | 159 | 187 | 161 | 183 | 60 | 33873 | 55 |
| 163 | 155 | 167 | 161 | 161 | 57 | 29807 | 45 |
| 181 | 174 | 171 | 169 | 177 | 58 | 30493 | 59 |
| 171 | 160 | 172 | 168 | 171 | 64 | 33715 | 40 |
| 164 | 166 | 170 | 162 | 178 | 57 | 29530 | 71 |
| 160 | 178 | 169 | 173 | 168 | 55 | 29350 | 70 |
| 156 | 149 | 164 | 147 | 168 | 52 | 29353 | 52 |
| 158 | 162 | 154 | 178 | 147 | 60 | 27056 | 52 |
| 173 | 176 | 185 | 172 | 175 | 64 | 33461 | 62 |
| 167 | 171 | 181 | 180 | 174 | 61 | 29573 | 60 |
| 177 | 176 | 174 | 177 | 178 | 53 | 25412 | 59 |
| 169 | 174 | 160 | 179 | 176 | 71 | 35281 | 63 |
| 163 | 172 | 175 | 172 | 159 | 44 | 21690 | 56 |
| 152 | 162 | 183 | 175 | 172 | 59 | 31022 | 57 |
| 183 | 178 | 180 | 170 | 171 | 74 | 35610 | 65 |
| 181 | 180 | 178 | 188 | 194 | 58 | 29936 | 59 |
| 185 | 172 | 176 | 175 | 177 | 86 | 46392 | 57 |
| 171 | 162 | 177 | 164 | 183 | 60 | 30679 | 46 |
| 173 | 170 | 175 | 163 | 177 | 49 | 26730 | 55 |
| 181 | 187 | 178 | 186 | 188 | 58 | 31707 | 70 |
| 167 | 170 | 157 | 156 | 164 | 70 | 39356 | 73 |
| 167 | 163 | 168 | 183 | 181 | 62 | 36144 | 51 |

Table 16.1: Example data set illustrating most of the effects discussed in this chapter. In each analysis, we will try to predict the height of daughters (DauHt) from some subset of the seven available predictor variables. Four of these predictor variables are the heights of the parents (MomHt and DadHt) and the heights of the maternal and paternal grandfathers (MGFHt and PGFHt). Two additional predictors are measures of the childhood environment of the daughter, intended to provide information about the diet and health care provided to her during her growing years (say birth to age 15). These are the socioeconomic status of the family during these years (DauSES) and the average family income during these years (DauInc). The final predictor is a measure of the socioeconomic status of the mother during her (i.e., the mother's) own growing years (MomSES).

To be more concrete still, consider an example data set for this problem shown in Table 16.1. (For now, just consider the to-be-predicted variable of daughter's height and the predictor variables of mother's height, maternal grandfather's height, and paternal grandfather's height; the other variables will come into play later.) For these sample data, fitting simple regression equations shows that daughters' heights are predicted about equally well from the heights of the maternal and paternal grandfather when either predictor is used by itself, as shown by these summary ANOVAs:

$$\text{DauHt} = 118.823 + 0.301 \times \text{MGFHt}$$

| Source | *df* | *SS* | *MS* | *F* |
|---|---|---|---|---|
| Model | 1 | 358.777 | 358.777 | 4.773 |
| Error | 48 | 3608.203 | 75.171 | |
| Total | 49 | 3966.980 | | |

$$\text{DauHt} = 118.571 + 0.296 \times \text{PGFHt}$$

| Source | *df* | *SS* | *MS* | *F* |
|---|---|---|---|---|
| Model | 1 | 348.657 | 348.657 | 4.625 |
| Error | 48 | 3618.323 | 75.382 | |
| Total | 49 | 3966.980 | | |

Actually, in isolation the maternal grandfather's height seems to be a slightly better predictor.

Now let's see what happens when we start with MomHt in the model (call this $X_1$) and try to add each one of the grandfathers' heights. The single-predictor model with just the mother's height is

$$\text{DauHt} = 67.556 + 0.603 \times \text{MomHt}$$

| Source | *df* | *SS* | *MS* | *F* |
|---|---|---|---|---|
| Model | 1 | 1433.527 | 1433.527 | 27.160 |
| Error | 48 | 2533.453 | 52.780 | |
| Total | 49 | 3966.980 | | |

The two two-predictor models are

$$\text{DauHt} = 73.703 + 0.656 \times \text{MomHt} - 0.090 \times \text{MGFHt}$$

| Source | *df* | *SS* | *MS* | *F* |
|---|---|---|---|---|
| Model | 2 | 1454.056 | 727.028 | 13.598 |
| Error | 47 | 2512.924 | 53.466 | |
| Total | 49 | 3966.980 | | |

$$\text{DauHt} = 17.040 + 0.601 \times \text{MomHt} + 0.292 \times \text{PGFHt}$$

| Source | *df* | *SS* | *MS* | *F* |
|---|---|---|---|---|
| Model | 2 | 1773.682 | 886.841 | 19.004 |
| Error | 47 | 2193.298 | 46.666 | |
| Total | 49 | 3966.980 | | |

Computing the extra sum of squares comparisons yields

| *Extra Sum of Squares Comparison:* | | | |
|---|---|---|---|
| | **Smaller model** | **Larger model** | **Extra** |
| Predictors: | MomHt | MomHt MGFHt | MGFHt |
| SS model: | 1433.527 | 1454.056 | 20.529 |
| df model: | 1 | 2 | 1 |
| Extra $F = \frac{20.529/1}{53.466} = 0.538$ for MGFHt | | | |

| *Extra Sum of Squares Comparison:* | | | |
|---|---|---|---|
| | **Smaller model** | **Larger model** | **Extra** |
| Predictors: | MomHt | MomHt PGFHt | PGFHt |
| SS model: | 1433.527 | 1773.682 | 340.154 |
| df model: | 1 | 2 | 1 |
| Extra $F = \frac{340.154/1}{46.666} = 7.289$ for PGFHt | | | |

In summary, adding MGFHt does not improve the model using MomHt as a predictor, but adding PGFHt does improve that model. That is, the better predictor in isolation (MGFHt) is not the better predictor given MomHT.

As already discussed, the reason for this anomaly is that MGFHt is fairly redundant with MomHt, but PGFHt is not. Of course this makes biological sense, because the mother is genetically related to the maternal grandfather, not the paternal one. In addition, the redundancy is evident in the pattern

of correlations among these variables, shown in Table 16.2. In the second row, we can see that the MGFHt and MomHt are strongly correlated ($r = 0.60$), which indicates that they measure the same thing to a great extent. Really, there is little more to their redundancy than this high correlation. Conversely, PGFHt and MomHt are essentially uncorrelated ($r = 0.01$), which indicates that they measure quite different things. Based on the idea that we want to add predictors that bring in *new* information, then, it is obvious that we should prefer PGFHt over MGFHt when trying to improve the model using MomHt. And of course it would work just the other way around if our initial model had included DadHt; in that case, it would be more helpful to add MGFHt to get new information, not PGFHt.

| | MomHt | DadHt | MGFHt | PGFHt | DauSES | DauInc | MomSES |
|---|---|---|---|---|---|---|---|
| DauHt | 0.601 | 0.247 | 0.301 | 0.296 | 0.293 | 0.293 | -0.009 |
| MomHt | | -0.010 | 0.596 | 0.006 | 0.003 | 0.003 | 0.499 |
| DadHt | | | -0.001 | 0.611 | 0.000 | -0.012 | -0.013 |
| MGFHt | | | | 0.004 | 0.005 | 0.006 | -0.001 |
| PGFHt | | | | | 0.004 | -0.001 | -0.013 |
| DauSES | | | | | | 0.898 | 0.209 |
| DauInc | | | | | | | 0.204 |

Table 16.2: Correlation matrix for the data set shown in Table 16.1.

More generally, what this example is meant to illustrate is that you want to avoid *redundancy* when you are assembling multiple predictors. Even if a predictor is useful by itself, it is not necessarily useful in combination with other predictors if it is redundant with those other predictors. As an extreme example, suppose your first predictor of daughter's height was the height of the mother in centimeters, and someone suggested that as a second predictor you should also use the height of the mother in inches. Common sense should tell you that this suggestion is silly. Adding this second predictor would do you no good, because once you have the mother's height once, you don't need it again. In other words, the second predictor is useless because it is redundant with another predictor already being used.

### 16.2.1 SS Thinning

Redundancy can also be responsible for a second puzzling pattern of results that is sometimes obtained, represented by a combination of these two basic outcomes:

1. Each of two predictors, say $X_1$ and $X_2$, is a significant predictor of $Y$ when considered as the only predictor.

2. The multiple regression model containing both $X_1$ and $X_2$ is not significant.

Again we seem to have a paradox: From the single-predictor regressions, we can conclude that each predictor is useful in predicting $Y$ at a better than chance level. The result of the two-predictor regression, however, is consistent with the view that *neither* predictor is useful. How can the two-predictor model fail to be significant if each predictor is significant in isolation?

An example of this arises when we try to predict daughters' heights (DauHt) from the measures of their environments during their growing years (DauSES and DauInc) using the data in Table 16.1. Here are the two single-predictor models and the two-predictor model:

$$\text{DauHt} = 152.430 + 0.293 \times \text{DauSES}$$

| Source | *df* | *SS* | *MS* | *F* |
|---|---|---|---|---|
| Model | 1 | 339.485 | 339.485 | 4.492 |
| Error | 48 | 3627.495 | 75.573 | |
| Total | 49 | 3966.980 | | |

$$\text{DauHt} = 153.678 + 0.001 \times \text{DauInc}$$

| Source | *df* | *SS* | *MS* | *F* |
|---|---|---|---|---|
| Model | 1 | 340.505 | 340.505 | 4.507 |
| Error | 48 | 3626.475 | 75.552 | |
| Total | 49 | 3966.980 | | |

DauHt = 152.153 + 0.152 × DauSES + 0.0005 × DauInc

| Source | *df* | *SS* | *MS* | *F* |
|---|---|---|---|---|
| Model | 2 | 358.225 | 179.112 | 2.333 |
| Error | 47 | 3608.755 | 76.782 | |
| Total | 49 | 3966.980 | | |

Note that the two one-predictors models are both significant, but the two-predictor model is not.

The explanation for this puzzling pattern has two parts. First, the two predictors are correlated (i.e., redundant), so to some degree they provide the same information about $Y$. In the example data, for instance, there is a 0.9 correlation between DauSES and DauInc. Because of this strong correlation, the two-predictor model does only a little better than the one-predictor models ($SS_{model}$'s of approximately 339 and 340 versus 358). Second, the $df_{model}$ increases from one to two going from the one-predictor models to the two-predictor model. Since the $SS_{model}$ stays approximately the same, that means that the $MS_{model}$ drops approximately in half for the two-predictor model compared to the one-predictor models ($MS_{model}$'s of approximately 339 and 340 versus 179). This reduction in $MS_{model}$ is crucial, because the $F_{model}$ is the ratio of the $MS_{model}$ to the $MS_{error}$. Although the $MS_{error}$ is approximately the same for all three models, the $F$ for the two-predictor model is only about half of that for the one-predictor models, because of its extra $df_{model}$.

In summary, as you include more and more redundant predictors in a model, the $SS_{model}$ does not increase as fast as the $df_{model}$, so the $F_{model}$ drops. You might say that the same SS is getting thinned out by being spread over too many degrees of freedom. After there has been enough thinning (i.e., enough redundant predictors added), the $F_{model}$ can eventually drop below the $F_{critical}$, even if each of the predictors is significant individually.

Conceptually, thinning occurs because a regression model *should* be able to do better as more predictors are added, even if only by chance. If added predictors are redundant, however, there is no improvement in prediction—not even by chance. As more and more redundant predictors are added, eventually the model appears to be doing no better than you would expect by chance with that number of predictors, and so it's $F_{model}$ may drop below the level needed for significance.

## 16.3 Reduction of Error in $Y$

A third puzzling pattern of results is essentially the opposite of the first one:

1. A certain predictor, say $X_2$, is not a significant predictor of $Y$ when considered as the only predictor.
2. The ESS for this predictor is significant when the predictor is added to a smaller model with another predictor, say $X_1$.

Now, the first outcome tells us that $X_2$ is not a good predictor of $Y$, and yet the second outcome tells us that $X_2$ is. How can these both be true?

In the data of Table 16.1, the variables of DadHt and MomHt provide an example of this. By itself, DadHt is not a significant predictor of DauHt, as indicated by this one-predictor model[1]:

DauHt = 126.771 + 0.249 × DadHt

| Source | *df* | *SS* | *MS* | *F* |
|---|---|---|---|---|
| Model | 1 | 242.312 | 242.312 | 3.123 |
| Error | 48 | 3724.668 | 77.597 | |
| Total | 49 | 3966.980 | | |

Yet DadHt does add significantly to MomHt as a predictor, as indicated by the following two-predictor model and associated ESS comparison:

DauHt = 22.755 + 0.605 × MomHt + 0.255 × DadHt

| Source | *df* | *SS* | *MS* | *F* |
|---|---|---|---|---|
| Model | 2 | 1688.315 | 844.158 | 17.412 |
| Error | 47 | 2278.665 | 48.482 | |
| Total | 49 | 3966.980 | | |

[1] I agree this is not a very realistic example data set, because there is no obvious reason why the heights of daughters should be more strongly correlated with the heights of their mothers than with the heights of their fathers. The data were constructed mainly to illustrate the various outcomes discussed in this chapter, not to be realistic.

| *Extra Sum of Squares Comparison:* | | | |
|---|---|---|---|
| | **Smaller model** | **Larger model** | **Extra** |
| Predictors: | MomHt | MomHt DadHt | DadHt |
| SS model: | 1433.527 | 1688.315 | 254.788 |
| df model: | 1 | 2 | 1 |
| Extra $F = \frac{254.788/1}{48.482} = 5.255$ for DadHt | | | |

If DadHt is not significant by itself, how can it add significantly to MomHt?

This pattern arises because MomHt reduces the error term, thereby allowing us to see the statistical significance of DadHt, even though it is only weakly related to DauHt. Note that the one-predictor model with DadHt has a $MS_{error}$ of 77, whereas the two-predictor model with both parents' heights has a much smaller $MS_{error}$ of 48. In essence, including MomHt as a predictor removed much apparent error from the analysis. As always, having a smaller error term produces larger $F$'s, because the error term is the divisor of the $F$ ratio.

Note that the error reduction we are discussing here is the same idea as the error reduction we discussed in connection with ANOVA. When a factor is omitted from an ANOVA, its effects go into the error term, thereby reducing the $F$'s for the other factors. Similarly, if an important variable (e.g., MomHt) is omitted from a regression analysis, effects that it could explain are included in the error term and reduce $F$'s for the other predictors.

## 16.4 Suppression of Error in $X$

A fourth puzzling pattern of results is characterized by these outcomes:

1. A certain predictor, say $X_2$, has a correlation with $Y$ of zero, or very nearly so.
2. The ESS for this predictor is significant when the predictor is added to a smaller model with another predictor, say $X_1$.

At first this looks like just an extension of the previous puzzling result, "Reduction of Error in $Y$": A predictor $X_2$ looks no good by itself, but is good when added to another predictor. This case is more extreme, however, because the poor predictor actually has a zero correlation with $Y$, not a small correlation that turns out to be significant when error is reduced. How could a predictor that has a zero correlation with $Y$ *ever* be a useful predictor? After all, multiple regression simply uses linear relationships between $Y$ and the $X$'s to predict $Y$. If there is not even a hint of a linear relation between $Y$ and a certain $X$, how can that $X$ possibly be a useful predictor?

In the data of Table 16.1, the variables of MomHt and MomSES provide an example. As shown in Table 16.2, the correlation of DauHt with MomSES is essentially zero. And indeed it makes sense that the height of the daughter should be unrelated to the socioeconomic status that the mother enjoyed when she was growing up, especially if there is considerable social mobility from one generation to the next. The lack of relation between DauHt and MomSES can also be seen in this simple regression predicting DauHt from MomSES:

DauHt = 170.567 - 0.009 × MomSES

| Source | *df* | *SS* | *MS* | *F* |
|---|---|---|---|---|
| Model | 1 | 0.328 | 0.328 | 0.004 |
| Error | 48 | 3966.652 | 82.639 | |
| Total | 49 | 3966.980 | | |

Despite the zero correlation between DauHt and MomSES, MomSES is quite a good predictor to add to a model with MomHt in it. Here is the single-predictor model with just MomHt, repeated from above:

DauHt = 67.556 + 0.603 × MomHt

| Source | *df* | *SS* | *MS* | *F* |
|---|---|---|---|---|
| Model | 1 | 1433.527 | 1433.527 | 27.160 |
| Error | 48 | 2533.453 | 52.780 | |
| Total | 49 | 3966.980 | | |

And here is the two-predictor model with MomHt and MomSES, plus the ESS comparison to evaluate the effect of adding MomSES:

$$\text{DauHt} = 57.253 + 0.809 \times \text{MomHt} - 0.412 \times \text{MomSES}$$

| Source | *df* | *SS* | *MS* | *F* |
|---|---|---|---|---|
| Model | 2 | 1938.854 | 969.427 | 22.466 |
| Error | 47 | 2028.126 | 43.152 | |
| Total | 49 | 3966.980 | | |

| *Extra Sum of Squares Comparison:* | | | |
|---|---|---|---|
| | **Smaller model** | **Larger model** | **Extra** |
| Predictors: | MomHt | MomHt MomSES | MomSES |
| SS model: | 1433.527 | 1938.854 | 505.327 |
| df model: | 1 | 2 | 1 |

Extra $F = \frac{505.327/1}{43.152} = 11.711$ for MomSES

The ESS comparison yields a highly significant $F$ for MomSES, despite the absence of any correlation with $Y$.

This pattern of results is often called a "suppressor effect", and the variable that is useful despite being uncorrelated with $Y$ is called a "suppressor variable." For example, in this case MomSES is the suppressor variable.

The pattern arises because the suppressor variable suppresses error in the other predictor variable, thereby making the other variable even better correlated with $Y$. In the example of Table 16.1, the explanation would go something like this:

> MomHt is basically determined by two components: genetics plus the mother's environment during her growth years. Now the genetic component will be passed along to the daughter, so this component is quite relevant for predicting the daughter's height. For example, if the mother is short because she has the genes to be short, then the daughter is quite likely to be short as well. But the environment of the mother during her growing years is not directly relevant to the daughter's height, so this is less relevant to predicting the daughter's height. For example, if the mother is short because of malnutrition or disease during her growth years, the daughter is not particularly likely to be short as well (assuming that the growing environments of successive generations are not too strongly correlated).
>
> The implication of our "two-component" theory of height is that we could predict the daughter's height better if we could measure the mother's "genetic height" instead of her actual height. That is, we would like to measure the height to which she should have grown, barring any unusual environmental circumstances. Unfortunately, we can't measure that, because we don't yet know how to relate genes to expected height. But we can attempt to arrive at the genetic component in a two-step process: first, we measure actual height; second, we remove the effects of environment by correcting for the mother's SES. Based on our two-component model, we should be left with just the genetic component of the mother's height, and this should be a *better* predictor of the daughter's height than was the original MomHt.

To see how the regression model uses the mother's SES to correct for environmental effects, it is instructive to look closely at the two-predictor model:

$$\text{DauHt} = 57.253 + 0.809 \times \text{MomHt} - 0.412 \times \text{MomSES}$$

Note the $\hat{b}_1$ of 0.809 associated with MomHt. This tells us to predict that the daughter will be taller when the mother is taller, which makes sense. Note also the $\hat{b}_2$ of -0.412 associated with MomSES, which tells us to predict that the daughter will be shorter when the mother had a higher SES. This is where the correction happens. When the mother has a high SES, 0.809 × MomHt is *too much* to add in to the predicted daughter's height. In this case some of the mother's tallness was probably due to environment rather than genetics, and the environmental component will not be passed along to the daughter. So, when the mother's SES was high, we need to *take back* some of the predicted tallness assigned to the daughter. Note that we are not assuming any direct causal relationship between the mother's SES and the daughter's height.

## 16.5 Venn Diagrams—*Optional Section*

This section uses Venn diagrams as a form of pictorial representation to help portray the relationships among variables that lead to redundancy, error reduction, and suppressor effects. The material presented here is not used elsewhere in the book, but it may aid some readers in conceptualizing context effects.

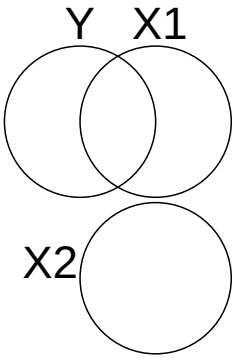


Figure 16.1: Use of Venn diagrams to represent correlations between variables. $Y$ and $X_1$ are correlated with each other, but neither is correlated with $X_2$.

In a Venn diagram, each variable is represented as a circle. Points inside the circle represent factors influencing the variable, and points outside the circle represent factors that do not influence it. The extent of correlation between two variables is represented by the overlap between their circles, with higher correlation represented by greater overlap. This makes sense: If two variables have mostly common influences then it makes sense that they should be well correlated. In Figure 16.1, for example, $X_1$ and $Y$ are correlated, but $X_2$ and $Y$ are not. With respect to the prediction of $Y$ from a given $X$ variable, the part explained by the model is the part that overlaps with that $X$; the remainder of $Y$, not overlapping with that $X$, is unexplained error.

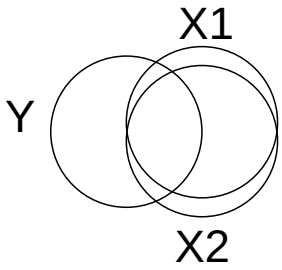


Figure 16.2: A pictorial representation of redundancy.

Figure 16.2 shows a pictorial representation of redundancy using Venn diagrams. In Figure 16.2, redundancy is apparent in the fact that both of the predictors, $X_1$ and $X_2$, overlap with $Y$ in pretty much the same area. That is, both predictors have the same things in common with $Y$, so either one of them—but not both—is needed to explain that aspect of $Y$.

Figure 16.3 depicts the error reduction situation with a Venn diagram. The key point is that $X_1$ and $X_2$ overlap with different components of $Y$. In the two-predictor model, the error in $Y$ is the part that is not overlapped by *either* $X$ variable. Note that there is much less unexplained $Y$ (i.e., error) when both variables are included in the model than when only one is, because the two predictors bite off different pieces of $Y$. Thus, each $X$ can reduce error for the other. Note also that $X_1$ and $X_2$ do not overlap with each other, so they are not redundant. This is the situation that maximizes error reduction, although some error reduction could still occur if $X_1$ and $X_2$ overlapped slightly, as long as the area of overlap was small relative to the separate overlaps with $Y$.

Figure 16.4 shows a pictorial representation of suppression in terms of Venn diagrams. $Y$ is correlated with $X_2$ but not with the suppressor variable $X_1$ (in the example, DauHt is $Y$, MomHt is

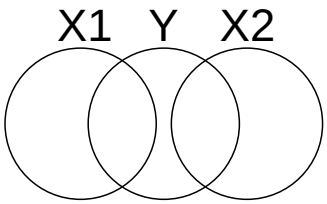


Figure 16.3: A pictorial representation of error reduction.

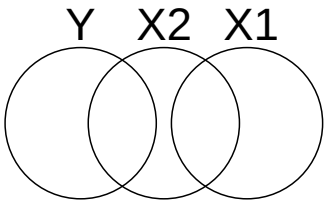


Figure 16.4: A pictorial representation of error suppression. $X_1$ improves the predictions of $Y$ from $X_2$, because $X_1$ eliminates some irrelevant information from $X_2$, essentially allowing $X_2$'s relevant part to stand out better.

$X_2$, and MomSES is $X_1$). $X_1$ is correlated with the part of $X_2$ which *it does not share* with $Y$. By including $X_1$ in the model, we can essentially factor out some of the irrelevant part of $X_2$ and more accurately predict $Y$ from the remaining, relevant part of $X_2$.

## 16.6 Mixing Positive and Negative Correlations

This section will discuss various possibilities where $Y$ has a positive correlation with some predictors but a negative correlation with others. All of the same phenomena can still arise: redundancy, error reduction, and error suppression. It is slightly more difficult to spot these phenomena, however, because the mixed signs of the correlations can be confusing.

### 16.6.1 Redundancy

Redundancy arises when two predictors are both related to the to-be-predicted variable in pretty much the same way. Interestingly, this can happen even when the two predictors are negatively correlated.

For example, consider trying to predict an athlete's long jump distance from his or her performance in a 10K running race. Race performance could be measured as the total time needed to run the 10K ($X_1$). Alternatively, it could be measured as the runner's average speed for the race ($X_2$; e.g., in kilometers per hour). Time and speed have an almost perfect negative correlation, because larger speeds produce smaller times, and vice versa[2]. Despite the negative correlation, of course, we know that both $X_1$ and $X_2$ really measure the same thing: how fast the runner went.

| | Time | Speed |
|---|---|---|
| Jump distance | -0.50 | 0.50 |
| Time | | -0.95 |

Table 16.3: Hypothetical correlation matrix for time and speed in a 10K race and long-jump distance.

Table 16.3 shows a plausible correlation matrix for $Y$, $X_1$, and $X_2$ in this example. There is a strong negative correlation between time and speed, as already discussed. In addition, the correlations of jump distance with the two predictors have different signs—one positive and one negative. We expect

[2]The correlation is not quite perfect because time and speed are not linearly related. For example, three runners finishing the race in 1, 2, or 3 hours would have average speeds of 10, 5, and 3.33 K per hour. Plotting these points on a scattergram reveals a nonlinear relationship.

longer jumps from those who run faster, so jump distance should be positively correlated with running speed and negatively correlated with running time.

Intuitively, it makes sense that long jump performance would not be predicted much better from both time and speed than from either one of them—that is, the two-predictor model should not do much better than either one-predictor model. This is exactly the pattern we saw in previous cases of redundancy as well, and it makes perfect sense here because the two predictors are redundant. The only difference from the examples seen earlier is that there are some negative correlations here. But note that the negative correlations simply arise because speed is measured using opposite scales (high is faster versus low is faster). Because of that, the to-be-predicted variable necessarily has a positive correlation with one of the redundant predictors and a negative correlation with the other one.

### 16.6.2 Reduction of Error in $Y$

Error reduction occurs when two predictors are related to different components of $Y$ and not to each other. In that case, including them both in the model reduces error so that each one's predictive contribution can be seen more clearly.

| Example 1 | | | Example 2 | | |
|---|---|---|---|---|---|
| | $X_1$ | $X_2$ | | $X_1$ | $X_2$ |
| $Y$ | 0.50 | 0.50 | $Y$ | -0.50 | 0.50 |
| $X_1$ | | 0.00 | $X_1$ | | 0.00 |

Table 16.4: Two hypothetical correlation matrices illustrating reduction of error in $Y$.

Table 16.4 shows two examples of correlation matrices that would produce identical amounts of error reduction. Example 1 has only positive correlations, as in the examples seen earlier. Example 2 differs in that one of the predictors has a negative correlation with $Y$. This difference is unimportant, however, because $Y$ can be predicted just as well from a negative correlation as from a positive one. The essential point is still that the two predictors provide different information about $Y$ (i.e., the correlation between predictors is 0), so each one will tend to reduce error for the other.

### 16.6.3 Suppression of Error in $X$

One predictor (say $X_2$) may be useful in predicting $Y$ even if it is completely uncorrelated with $Y$, as long as it provides information that can suppress the irrelevant part of another predictor (say $X_1$). The key features of this effect are that:

1. $X_1$ is correlated with $Y$,
2. $X_1$ is correlated with $X_2$, and
3. $X_2$ is not correlated with $Y$.

The critical point is that the correlations mentioned in the first two of these points can be either positive or negative, independently of one another.

| Example 1 | | | Example 2 | | |
|---|---|---|---|---|---|
| | $X_1$ | $X_2$ | | $X_1$ | $X_2$ |
| $Y$ | 0.50 | 0.00 | $Y$ | -0.50 | 0.00 |
| $X_1$ | | 0.50 | $X_1$ | | 0.50 |
| Example 3 | | | Example 4 | | |
| | $X_1$ | $X_2$ | | $X_1$ | $X_2$ |
| $Y$ | 0.50 | 0.00 | $Y$ | -0.50 | 0.00 |
| $X_1$ | | -0.50 | $X_1$ | | -0.50 |

Table 16.5: Two hypothetical correlation matrices illustrating reduction of error in $Y$.

For example, Table 16.5 illustrates four example data sets that would all yield the same suppressor effect for $X_2$. In each case, $X_2$ does not predict $Y$ but it does predict $X_1$. Thus, $X_2$ can be used to suppress the irrelevant part of $X_1$ and thereby improve predictions from it.

# Chapter 17

# Finding the Best Model

In some situations a researcher has a large pool of potential predictor variables and just wants to see how these can best be used to predict the variable of interest. This situation often arises in exploratory research, and there may be no specific hypotheses to test, contrary to our emphasis in the previous regression chapters. For example, an educational researcher might be interested in finding out what predicts difficulties in learning to read, so that kids likely to have problems can be given a special head-start program. The to-be-predicted variable would be some available measure of reading problems, and the researcher might use as predictors a whole host of different variables that can easily be assembled from school records and other sources. Alternatively, a medical researcher might be interested in finding out what predicts susceptibility to a certain disease for which preventive measures are available. Or, a personnel psychologist might be interested in what predicts on-the-job success so that the best candidates can be selected for a certain job. Finally, a gambler might be interested in predicting which of two teams will win a sporting event. In all of these cases, the goal is simply to come up with the best possible predictive model.

Mainly, this chapter presents some computational procedures that can be used when the problem is simply to find the best model. In using these procedures, you compute many $F$'s using ESS comparisons. You do not interpret these $F$'s, however (although you could interpret them as discussed in section 15.3, if you wanted to). Instead, in these procedures the $F$'s are simply used to decide what to do next.

In addition to the computational procedures, there are two conceptual points to consider: First, just what are the criteria for the "best" model? This point is considered in the next section. Second, how much faith should we have in the predictions from the best model after we find it? This point is considered in the final section.

## 17.1 What is the Best Model?

The first issue is exactly how we should define "the best model." One goal is obviously to minimize prediction error, but a second goal is also to keep the model as simple as possible. If they were equally accurate, we would certainly prefer a two-predictor model to a twenty-predictor model, because it would be easier to use (i.e., you would only have to measure 2 predictors, not 20). A third goal is to have a model that works well not only on the current cases but also on future cases to which we may want to apply it, as discussed in Section 13.4. There is some room for debate about exactly how to combine these goals in defining what the best model is, but in practice people pretty much seem to accept the following operational definition. The best model meets two criteria:

1. It includes only predictors that make a significant ($p < .05$) unique contribution to the model.
2. It has the smallest $MS_{error}$ of all the models for which the first criterion is satisfied.

This definition seems reasonable. The first criterion helps keep the number of predictors small (and it also gives us some assurance that the model will still be useful when we apply it to new cases for reasons discussed in Section 13.4), whereas the second minimizes prediction error.

Given these criteria for the best model, the next question is how to go about finding it, given an actual set of data. The next four sections describe various procedures that can be used—along with their limitations. These procedures rely heavily on the terminology of the $F_{add}$ and $F_{drop}$, as explained in 15.5, so it might be useful to review that terminology briefly before proceeding.

Table 17.1 summarizes the data set that will be used as an example in this chapter. It contains a to-be-predicted $Y$ variable and seven potential predictors $X_1$–$X_7$. The table shows the results of fitting every possible model to predict $Y$ from some or all of these seven predictors (127 possible models). For each model, the table shows the predictors in the model and the $SS_{model}$ and $MS_{error}$ obtained when $Y$ is predicted from the indicated $X$'s. The $SS_{model}$ and $MS_{error}$ are the only values needed from each model for the computational procedures demonstrated in this chapter.

*Note to students:* In the following sections, each of the procedures is illustrated by working through the example data set in excruciating detail. If you think you understand a procedure before reading to the end of an example, it is a good idea to stop reading and see whether you can carry on with the procedure by hand, using the text only as a check.

## 17.2 All Possible Models Procedure

The only approach *guaranteed* to find the best model is to try all the possible models. That is, you would have to compute a multiple regression for every possible combination of predictors, as shown in Table 17.1. With this table, you apply the step-by-step procedure shown in Table 17.2 to find the best model.

### 17.2.1 Illustration with data of Table 17.1.

Here is how to apply the procedure of Table 17.2 to the data shown in Table 17.1.

**Step 1:**

**Part 1:** The smallest $MS_{error}$ is 0.265, corresponding to the model with $X_1$, $X_2$, $X_3$, $X_4$, $X_5$, and $X_7$.

**Part 2:** Here are the $F_{drop}$ values for the predictors in this model:

| Variable: | $X_1$ | $X_2$ | $X_3$ | |
|---|---|---|---|---|
| $F_{drop}$: | $\frac{19594.670-19593.845}{0.265}$ | $\frac{19594.670-19539.776}{0.265}$ | $\frac{19594.670-19591.576}{0.265}$ | |
| | 3.113 | 207.147 | 11.675 | |

| Variable: | $X_4$ | $X_5$ | $X_6$ | $X_7$ |
|---|---|---|---|---|
| $F_{drop}$: | $\frac{19594.670-19567.293}{0.265}$ | $\frac{19594.670-19592.730}{0.265}$ | | $\frac{19594.670-19592.868}{0.265}$ |
| | 103.309 | 7.321 | | 6.800 |

**Part 3:** No, the $F_{drop}$ for $X_1$ is not significant. Therefore, cross out this model and go on to the next step.

**Step 2:**

**Part 1:** After eliminating the model considered in Step 1, the smallest remaining $MS_{error}$ is 0.268, corresponding to the model with $X_1$, $X_2$, $X_3$, $X_4$, $X_6$, and $X_7$.

**Part 2:** Here are the $F_{drop}$ values for the predictors in this model:

| Variable: | $X_1$ | $X_2$ | $X_3$ | |
|---|---|---|---|---|
| $F_{drop}$: | $\frac{19594.541-19593.936}{0.268}$ | $\frac{19594.541-19537.663}{0.268}$ | $\frac{19594.541-19591.497}{0.268}$ | |
| | 2.257 | 212.231 | 11.358 | |

| Variable: | $X_4$ | $X_5$ | $X_6$ | $X_7$ |
|---|---|---|---|---|
| $F_{drop}$: | $\frac{19594.541-19560.888}{0.268}$ | | $\frac{19594.541-19592.730}{0.268}$ | $\frac{19594.541-19592.748}{0.268}$ |
| | 125.571 | | 6.757 | 6.690 |

**Part 3:** No, the $F_{drop}$ for $X_1$ is not significant. Therefore, cross out this model and go on to the next step.

**Step 3:**

**Part 1:** The smallest remaining $MS_{error}$ is 0.270, corresponding to the model with all seven predictors, $X_1$–$X_7$.

**Part 2:** Here are the $F_{drop}$ values for the predictors in this model:

| Variable: | $X_1$ | $X_2$ | $X_3$ | |
|---|---|---|---|---|
| $F_{drop}$: | $\frac{19594.697-19593.957}{0.270}$ | $\frac{19594.697-19539.776}{0.270}$ | $\frac{19594.697-19591.608}{0.270}$ | |
| | 2.741 | 203.411 | 11.441 | |

| Variable: | $X_4$ | $X_5$ | $X_6$ | $X_7$ |
|---|---|---|---|---|
| $F_{drop}$: | $\frac{19594.697-19569.604}{0.270}$ | $\frac{19594.697-19594.541}{0.270}$ | $\frac{19594.697-19594.670}{0.270}$ | $\frac{19594.697-19593.151}{0.270}$ |
| | 92.937 | 0.578 | 0.100 | 5.726 |

**Part 3:** No, several of the $F_{drop}$'s are not significant. Therefore, cross out this model and go on to the next step.

**Step 4:**

**Part 1:** The smallest remaining $MS_{error}$ is now 0.275, corresponding to the model with $X_2$, $X_3$, $X_4$, $X_6$, and $X_7$.

**Part 2:** Here are the $F_{drop}$ values for the predictors in this model:

| Variable: | $X_1$ | $X_2$ | $X_3$ | |
|---|---|---|---|---|
| $F_{drop}$: | | $\frac{19593.936-19390.454}{0.275}$ | $\frac{19593.936-19591.448}{0.275}$ | |
| | | 739.935 | 9.047 | |

| Variable: | $X_4$ | $X_5$ | $X_6$ | $X_7$ |
|---|---|---|---|---|
| $F_{drop}$: | $\frac{19593.936-10408.521}{0.275}$ | | $\frac{19593.936-19592.487}{0.275}$ | $\frac{19593.936-19591.234}{0.275}$ |
| | 33401.509 | | 5.269 | 9.825 |

**Part 3:** Yes, all the $F_{drop}$'s are significant. Therefore, stop. The best model has these predictors. (If you look further, however, you will see that this is a very close decision, because the model with predictors $X_2$, $X_3$, $X_4$, $X_5$, and $X_7$ has very nearly the same $MS_{error}$ and also has all significant predictors.)

### 17.2.2 Discussion

The main drawback of the "All Possible Models" procedure is that it requires too much work when the number of predictors gets reasonably large. With $k$ predictors, there are $2^k - 1$ possible models. Table 17.3 shows some examples of the number of possible models for different numbers of predictors, and you can see that the number of possible models increases very rapidly as the number of predictors increases. On a computer, you might be able to work through the "All Possible Models" procedure for five to seven or eight predictors—maybe even ten predictors if you really cared enough to do a lot of work. But with enough predictors it would just not be practical to use this approach, because even computers are not that fast yet.

Because of the difficulty of trying all possible models, various shortcut procedures have been developed to search for the best model in situations where there are many predictors. The next three sections describe the three most common shortcut procedures: Forward Selection, Backward Elimination, and Stepwise. As you can see in Table 17.3, these shortcuts consider many fewer models and are therefore much less work. The downside, however, is that none of them is guaranteed to find the best model.

Each of the three shortcut procedures works in a step-by-step fashion, changing the model by adding or dropping one variable at each step. At each step, you make a series of extra sum of squares comparisons to decide how to change the model. To carry out these comparisons, you need to know the $SS_{model}$ and $MS_{error}$ for certain models that the procedure considers, as explained in detail below.

We will continue using the example data shown in Table 17.1 to illustrate these procedures. Note that this table actually contains more information than you need to carry out these procedures, because the procedures do not examine all of the possible regression models—only a subset, illustrated as we go along.

## 17.3 Forward Selection Procedure

The forward selection procedure tries to find the best model by adding the best available predictor at each step. The rules for the procedure are summarized in Table 17.4, and we will now illustrate the application of these rules using the data in Table 17.1.

### 17.3.1 Illustration with data of Table 17.1.

**Step 0:** The procedure begins with the model having no predictors at all: $Y = a$.

**Step 1:** At step 1, the procedure chooses the single best predictor from the available pool. To do this, it computes each of the possible one-predictor models and gets an $F_{add}$ for each one using Equation 17.1 in Table 17.4. For example, the $F_{add}$ for $X_1$ is

$$F_{add} \text{ for } X_1 = \frac{614.095 - 0.000}{395.666} = 1.552$$

The values of 614.095 and 395.666 are the $SS_{model}$ and $MS_{error}$ for the model with $X_1$ as the only predictor (see Table 17.1). The value of 0.000 is the $SS_{model}$ for the model from the previous step (i.e., $Y = a$), because a model with no predictors always has $SS_{model} = 0$. Similarly here are the $F_{add}$ values for the other predictors at this step, each computed using the one-predictor model with that predictor:

| Variable: | $X_1$ | $X_2$ | $X_3$ |
|---|---|---|---|
| $F_{add}$: | $\frac{614.095-0.000}{395.666}$ | $\frac{1166.328-0.000}{384.161}$ | $\frac{49.416-0.000}{407.430}$ |
| | 1.552 | 3.036 | 0.121 |

| Variable: | $X_4$ | $X_5$ | $X_6$ | $X_7$ |
|---|---|---|---|---|
| $F_{add}$: | $\frac{7698.263-0.000}{248.079}$ | $\frac{3533.240-0.000}{334.850}$ | $\frac{3562.206-0.000}{334.247}$ | $\frac{1388.865-0.000}{379.525}$ |
| | 31.031 | 10.552 | 10.657 | 3.659 |

At step 1, the highest $F_{add}$ is the one obtained for $X_4$, namely 31.031. This extra $F$ has 1 and $50 - 1 - 1 = 48$ degrees of freedom, because there are 50 cases in the data set and because one df is used by $a$ and another is used by the $b$ for the single predictor in the larger model. Thus, the critical $F$ is slightly less than 4.08, and 31.031 is significant. We therefore add $X_4$ to the model, and the new model at the end of step one is:

$$Y = a + b \times X_4$$

**Step 2:** At step 2, the procedure chooses the best second predictor to add to $X_4$. That is, it determines which is the best additional predictor *given that $X_4$ is already in the model.* To do this, it computes each of the two-predictor models including $X_4$, and gets an $F_{add}$ for each added predictor again using Equation 17.1. For example, the $F_{add}$ for $X_1$ at this step is

$$F_{add} \text{ for } X_1 = \frac{11519.160 - 7698.263}{172.062} = 22.207$$

The values of 11519.160 and 172.062 are the $SS_{model}$ and $MS_{error}$ for the model with $X_1$ and $X_4$ as the predictors (see Table 17.1). The value of 7698.263 is the $SS_{model}$ for the model from the previous step (i.e., $Y = a + b \times X_4$). Here is the full table of $F_{add}$ values for this step, each computed using the two-predictor model with one potential new predictor and $X_4$:

| Variable: | $X_1$ | $X_2$ | $X_3$ | |
|---|---|---|---|---|
| $F_{add}$: | $\frac{11519.160-7698.263}{172.062}$ | $\frac{19525.407-7698.263}{1.716}$ | $\frac{7731.686-7698.263}{252.646}$ | |
| | 22.207 | 6892.275 | 0.132 | |

| Variable: | $X_4$ | $X_5$ | $X_6$ | $X_7$ |
|---|---|---|---|---|
| $F_{add}$: | | $\frac{8589.080-7698.263}{234.404}$ | $\frac{8289.696-7698.263}{240.774}$ | $\frac{7698.512-7698.263}{253.352}$ |
| | | 3.800 | 2.456 | 0.001 |

Note that there is no $F_{add}$ for $X_4$; this predictor is already included in the model from the previous step, so it cannot be added again.

At this step, the highest $F_{add}$ is the one obtained for $X_2$, namely 6892.275. This $F$ has 1 and $50-1-2=47$ degrees of freedom, because there are 50 cases in the data set and because one df is used by $a$ and another two are used by the two $b$'s for the two predictors in the larger model. Thus, the critical $F$ is slightly less than 4.08, and this $F_{add}$ is significant. We therefore add $X_2$ to the model, and the new model at the end of this step is:

$$Y = a + b_1 \times X_4 + b_2 \times X_2$$

**Step 3:** The next step is to choose the best third predictor to add to $X_4$ and $X_2$—that is, to determine which third predictor is best *given that these two predictors are already in the model*. To do this, it is necessary to compute each of the three-predictor models including $X_2$ and $X_4$, and get an $F_{add}$ for each possible added predictor using Equation 17.1. For example, the $F_{add}$ for $X_1$ at this step is

$$F_{add} \text{ for } X_1 = \frac{19528.127 - 19525.407}{1.694} = 1.606$$

The values of 19528.127 and 1.694 are the $SS_{model}$ and $MS_{error}$ for the model with $X_1$, $X_2$, and $X_4$ as the predictors (see Table 17.1). The value of 19525.407 is the $SS_{model}$ for the model from the previous step (i.e., $Y = a + b_1 \times X_4 + b_2 \times X_2$). Similarly, here are the $F_{add}$ values for the other predictors at this step, each computed using the three-predictor model with that predictor and the two already selected in previous steps:

| Variable: | $X_1$ | $X_2$ | $X_3$ | |
|---|---|---|---|---|
| $F_{add}$: | $\frac{19528.127-19525.407}{1.694}$ | | $\frac{19530.455-19525.407}{1.643}$ | |
| | 1.606 | | 3.072 | |

| Variable: | $X_4$ | $X_5$ | $X_6$ | $X_7$ |
|---|---|---|---|---|
| $F_{add}$: | | $\frac{19575.842-19525.407}{0.657}$ | $\frac{19586.890-19525.407}{0.417}$ | $\frac{19532.023-19525.407}{1.609}$ |
| | | 76.766 | 147.441 | 4.112 |

Now there is no $F_{add}$ for $X_2$ or $X_4$, because these predictors are already included in the model from the previous step.

At this step, the highest $F_{add}$ is the one obtained for $X_6$, namely 147.441. This $F$ has 1 and $50-1-3=46$ degrees of freedom, because there are 50 cases in the data set and because one df is used by $a$ and another three are used by the $b$'s for the three predictors in the larger model. Thus, the critical $F$ is slightly less than 4.08, and this $F_{add}$ is significant. We therefore add $X_6$ to the model, and the new model at the end of this step is:

$$Y = a + b_1 \times X_4 + b_2 \times X_2 + b_3 \times X_6$$

**Step 4:** The next step is to choose the best fourth predictor to add to $X_2$, $X_4$, and $X_6$—that is, to determine which fourth predictor is best *given that these three predictors are already in the model*. To do this, it is necessary to compute each of the four-predictor models including $X_2$, $X_4$, and $X_6$, and get an $F_{add}$ for each possible added predictor using Equation 17.1. For example, the $F_{add}$ for $X_1$ at this step is

$$F_{add} \text{ for } X_1 = \frac{19591.485 - 19586.890}{0.324} = 14.182$$

The values of 19591.485 and 0.324 are the $SS_{model}$ and $MS_{error}$ for the model with $X_1$, $X_2$, $X_4$, and $X_6$ as the predictors (see Table 17.1). The value of 19586.890 is the $SS_{model}$ for the model from the previous step (i.e., $Y = a + b_1 \times X_2 + b_2 \times X_4 + b_3 \times X_6$). Similarly, here are the $F_{add}$ values for the other predictors at this step:

| Variable: | $X_1$ | $X_2$ | $X_3$ | |
|---|---|---|---|---|
| $F_{add}$: | $\frac{19591.485-19586.890}{0.324}$ | | $\frac{19591.234-19586.890}{0.329}$ | |
| | 14.182 | | 13.204 | |

| Variable: | $X_4$ | $X_5$ | $X_6$ | $X_7$ |
|---|---|---|---|---|
| $F_{add}$: | | $\frac{19591.303-19586.890}{0.328}$ | | $\frac{19591.448-19586.890}{0.325}$ |
| | | 13.454 | | 14.025 |

At this step, the highest $F_{add}$ is the one obtained for $X_1$, namely 14.182. This $F$ has 1 and $50 - 1 - 4 = 45$ degrees of freedom, because there are 50 cases in the data set and because one df is used by $a$ and another four are used by the $b$'s for the four predictors in the larger model. Thus, the critical $F$ is slightly less than 4.08, and this $F_{add}$ is significant. We therefore add $X_1$ to the model, and the new model at the end of this step is:

$$Y = a + b_1 \times X_1 + b_2 \times X_2 + b_3 \times X_4 + b_4 \times X_6$$

**Step 5:** The next step is to choose the best fifth predictor to add to $X_1$, $X_2$, $X_4$, and $X_6$—that is, to determine which fifth predictor is best *given that these four predictors are already in the model*. To do this, it is necessary to compute each of the five-predictor models including $X_1$, $X_2$, $X_4$, and $X_6$, and get an $F_{add}$ for each possible added predictor using Equation 17.1. For example, the $F_{add}$ for $X_3$ at this step is

$$F_{add} \text{ for } X_3 = \frac{19592.748 - 19591.485}{0.302} = 4.182$$

The values of 19592.748 and 0.302 are the $SS_{model}$ and $MS_{error}$ for the model with $X_1$, $X_2$, $X_3$, $X_4$, and $X_6$ as the predictors (see Table 17.1). The value of 19591.485 is the $SS_{model}$ for the model from the previous step. Similarly, here are the $F_{add}$ values for the other predictors at this step:

| Variable: | $X_1$ | $X_2$ | $X_3$ | |
|---|---|---|---|---|
| $F_{add}$: | | | $\frac{19592.748-19591.485}{0.302}$ | |
| | | | 4.182 | |

| Variable: | $X_4$ | $X_5$ | $X_6$ | $X_7$ |
|---|---|---|---|---|
| $F_{add}$: | | $\frac{19591.592-19591.485}{0.329}$ | | $\frac{19591.497-19591.485}{0.331}$ |
| | | 0.325 | | 0.036 |

At this step, the highest $F_{add}$ is the one obtained for $X_3$, namely 4.182. This $F$ has 1 and $50 - 1 - 5 = 44$ degrees of freedom, because there are 50 cases in the data set and because one df is used by $a$ and another five are used by the $b$'s for the five predictors in the larger model. Thus, the critical $F$ is slightly less than 4.08, and this $F_{add}$ is significant. We therefore add $X_3$ to the model, and the new model at the end of this step is:

$$Y = a + b_1 \times X_1 + b_2 \times X_2 + b_3 \times X_3 + b_4 \times X_4 + b_5 \times X_6$$

**Step 6:** The next step is to choose the best sixth predictor to add to the five already chosen—that is, to determine which sixth predictor is best *given that these five predictors are already in the model*. To do this, it is necessary to compute each of the six-predictor models including the five predictors already selected, and get an $F_{add}$ for each possible added predictor using Equation 17.1. Here are the $F_{add}$ values for $X_5$ and $X_6$ at this step:

| Variable: | $X_1$ | $X_2$ | $X_3$ | |
|---|---|---|---|---|
| $F_{add}$: | | | | |

| Variable: | $X_4$ | $X_5$ | $X_6$ | $X_7$ |
|---|---|---|---|---|
| $F_{add}$: | | $\frac{19593.151-19592.748}{0.300}$ | | $\frac{19594.541-19592.748}{0.268}$ |
| | | 1.343 | | 6.690 |

At this step, the highest $F_{add}$ is the one obtained for $X_7$, namely 6.690. This $F$ has 1 and $50-1-6=43$ degrees of freedom, because there are 50 cases in the data set and because one df is used by $a$ and another six are used by the $b$'s for the six predictors in the larger model. Thus, the critical $F$ is slightly less than 4.08, and this $F_{add}$ is significant. We therefore add $X_7$ to the model, and the new model at the end of this step has every predictor except $X_5$.

**Step 7:** The next step is to choose the best seventh predictor to add to the six already chosen—that is, to determine which seventh predictor is best *given that these six predictors are already in the model*. This is a simple task in the current data set, because there is only one more predictor to consider adding. Here is the $F_{add}$ value for $X_5$ at this step:

$$F_{add} \text{ for } X_5 = \frac{19594.697 - 19594.541}{0.270} = 0.578$$

and this step is summarized with the mostly empty table:

| Variable: | $X_1$ | $X_2$ | $X_3$ | |
|---|---|---|---|---|
| $F_{add}$: | | | | |

| Variable: | $X_4$ | $X_5$ | $X_6$ | $X_7$ |
|---|---|---|---|---|
| $F_{add}$: | | $\frac{19594.697-19594.541}{0.270}$ | | |
| | | 0.578 | | |

At this step, the highest (only) $F_{add}$ is 0.578. This $F$ has 1 and $50-1-7=42$ degrees of freedom, because there are 50 cases in the data set and because one df is used by $a$ and another seven are used by the $b$'s for the seven predictors in the larger model. Thus, the critical $F$ is slightly less than 4.08, and this $F_{add}$ is not significant.

Since there is no significant $F_{add}$, we stop at this step without adding any variable. The final model is the one selected at the end of the previous step—in this case, the six predictor model with all predictors except $X_5$. This six predictor model is the best model that the forward selection procedure can find. Here is the exact equation for this model: $Y = 0.000 + 0.214 \times X_1 + 0.985 \times X_2 - 0.487 \times X_3 + 1.100 \times X_4 + 0.237 \times X_6 - 0.351 \times X_7$.

### 17.3.2 Discussion

Table 17.5 summarizes the results obtained when the forward selection procedure is applied to the data of Table 17.1. (Most computer programs that do forward selection provide the results in a table like this.) At each step, an $F_{add}$ is computed for each variable that has not yet been included in the model, and the variable yielding the highest $F_{add}$ is selected as the one to add at that step. The procedure stops when there is no significant $F_{add}$ at a step.

Although it seems like the forward selection procedure was a lot of work, it really was a shortcut compared to testing all possible models, as illustrated by Table 17.6. At the first step the procedure considered all seven of the possible one-predictor models, so there was no savings in terms of these models. But at the second step the procedure considered only six of the 21 possible two-predictor models, so there was considerable savings at this step. Similarly, the procedure ignored many of the possible three, four, and five predictor models in the next several steps, saving lots more computation[1].

[1]Remember, when you carry out this procedure for yourself, you will not start with the information about all models shown in Table 17.1. When I talk about "saving computation", I am mostly referring to the computation that went into generating that table.

A drawback of the procedure, though, is that one of these ignored models could have been the best model, in which case the forward selection procedure would have missed it. That is the price of the shortcut. For example, the forward selection procedure did not find the best model for this data set, which was already identified in the previous section as the model with predictors $X_2$, $X_3$, $X_4$, $X_6$, and $X_7$.

## 17.4 Backward Elimination Procedure

The backward elimination procedure is a second shortcut procedure that operates much like the forward selection procedure, except in reverse. For this procedure, the starting model is the one with *all* the predictors in it, and at each step the procedure tries to drop out one unnecessary (i.e., nonsignificant) predictor, stopping when all predictors are significant. Table 17.7 shows the rules for this procedure in more detail, and the following calculations show how these rules are applied in the data of Table 17.1.

### 17.4.1 Illustration with data of Table 17.1.

**Step 0:** The procedure begins with the model having all seven predictors in it.

**Step 1:** The procedure tries to drop whichever predictor is least useful to the model (i.e., has the smallest $F_{drop}$). To do this, it computes each of the possible six-predictor models and gets an $F_{drop}$ for dropping each of the seven predictors currently in the model, using Equation 17.2 from Table 17.7. For example, the $F_{drop}$ for $X_1$ is

$$F_{drop} \text{ for } X_1 = \frac{19594.697 - 19593.957}{0.270} = 2.741$$

The values of 19594.697 and 0.270 are the $SS_{model}$ and $MS_{error}$ for the larger model (i.e., the one with seven predictors), and the value of 19593.957 is the $SS_{model}$ for the six-predictor model omitting $X_1$ (see Table 17.1). Similarly, here are the $F_{drop}$ values for the other predictors:

| Variable: | $X_1$ | $X_2$ | $X_3$ | |
|---|---|---|---|---|
| $F_{drop}$: | $\frac{19594.697-19593.957}{0.270}$ | $\frac{19594.697-19539.776}{0.270}$ | $\frac{19594.697-19591.608}{0.270}$ | |
| | 2.741 | 203.411 | 11.441 | |

| Variable: | $X_4$ | $X_5$ | $X_6$ | $X_7$ |
|---|---|---|---|---|
| $F_{drop}$: | $\frac{19594.697-19569.604}{0.270}$ | $\frac{19594.697-19594.541}{0.270}$ | $\frac{19594.697-19594.670}{0.270}$ | $\frac{19594.697-19593.151}{0.270}$ |
| | 92.937 | 0.578 | 0.100 | 5.726 |

At this step, the smallest $F_{drop}$ is the one obtained for $X_6$, namely 0.100. This $F$ has 1 and $50-1-7=42$ degrees of freedom, because there are 50 cases in the data set and because one df is used by $a$ and another seven are used by the $b$'s for the seven predictors in the larger model. Thus, the critical $F$ is slightly less than 4.08, and this $F_{drop}$ is not significant. We therefore drop $X_6$ from the model, and the new model at the end of this step has every predictor except $X_6$.

**Step 2:** The procedure next tries to eliminate whichever of the *remaining* predictors is least useful to the model (i.e., has the smallest $F_{drop}$). To do this, it computes each of the possible five-predictor models omitting $X_6$ plus one other predictor to form the new smaller models. It gets an $F_{drop}$ for dropping each of the six predictors currently in the model, using Equation 17.2. For example, the $F_{drop}$ for $X_1$ is

$$F_{drop} \text{ for } X_1 = \frac{19594.670 - 19593.845}{0.265} = 3.113$$

The values of 19594.670 and 0.265 are the $SS_{model}$ and $MS_{error}$ for the new larger model chosen in Step 1 (i.e., the one with six predictors omitting $X_6$), and the value of 19593.845 is the $SS_{model}$ for the five-predictor model omitting $X_1$ and $X_6$ (see Table 17.1). Similarly, here are the $F_{drop}$ values for the other predictors:

| Variable: | $X_1$ | $X_2$ | $X_3$ | |
|---|---|---|---|---|
| $F_{drop}$: | $\frac{19594.670-19593.845}{0.265}$ | $\frac{19594.670-19539.776}{0.265}$ | $\frac{19594.670-19591.576}{0.265}$ | |
| | 3.113 | 207.147 | 11.675 | |

| Variable: | $X_4$ | $X_5$ | $X_6$ | $X_7$ |
|---|---|---|---|---|
| $F_{drop}$: | $\frac{19594.670-19567.293}{0.265}$ | $\frac{19594.670-19592.730}{0.265}$ | | $\frac{19594.670-19592.868}{0.265}$ |
| | 103.309 | 7.321 | | 6.800 |

At this step, the smallest $F_{drop}$ is the one obtained for $X_1$, namely 3.113. This $F$ has 1 and $50-1-6=43$ degrees of freedom, because there are 50 cases in the data set and because one df is used by $a$ and another six are used by the $b$'s for the six predictors in the larger model. Thus, the critical $F$ is slightly less than 4.08, and this $F_{drop}$ is not significant. We therefore drop $X_1$ from the model, and the new model at the end of this step has every predictor except $X_1$ and $X_6$.

**Step 3:** The procedure next tries to eliminate whichever of the remaining predictors is least useful to the model (i.e., has the smallest $F_{drop}$). To do this, it computes each of the possible four-predictor models omitting $X_1$ and $X_6$ plus one other predictor to form the new smaller models. It gets an $F_{drop}$ for dropping each of the five predictors currently in the model, using Equation 17.2. For example, the $F_{drop}$ for $X_2$ is

$$F_{drop} \text{ for } X_2 = \frac{19593.845 - 19466.104}{0.277} = 461.159$$

The values of 19593.845 and 0.277 are the $SS_{model}$ and $MS_{error}$ for the new larger model chosen in Step 2 (i.e., the one with five predictors omitting $X_1$ and $X_6$), and the value of 19466.104 is the $SS_{model}$ for the four-predictor model omitting $X_1$, $X_2$, and $X_6$ (see Table 17.1). Similarly, here are the $F_{drop}$ values for the other predictors:

| Variable: | $X_1$ | $X_2$ | $X_3$ | |
|---|---|---|---|---|
| $F_{drop}$: | | $\frac{19593.845-19466.104}{0.277}$ | $\frac{19593.845-19591.534}{0.277}$ | |
| | | 461.159 | 8.343 | |

| Variable: | $X_4$ | $X_5$ | $X_6$ | $X_7$ |
|---|---|---|---|---|
| $F_{drop}$: | $\frac{19593.845-16749.630}{0.277}$ | $\frac{19593.845-19592.487}{0.277}$ | | $\frac{19593.845-19591.078}{0.277}$ |
| | 10267.924 | 4.903 | | 9.989 |

At this step, the smallest $F_{drop}$ is the one obtained for $X_5$, namely 4.903. This $F$ has 1 and $50-1-5=44$ degrees of freedom, because there are 50 cases in the data set and because one df is used by $a$ and another five are used by the $b$'s for the five predictors in the larger model. Thus, the critical $F$ is slightly less than 4.08, and this $F_{drop}$ is significant. We therefore stop without dropping any more terms from the model. The final model selected by the backward elimination method is thus the one we started with at the beginning of this step, with all predictors except $X_1$ and $X_6$. The full model is: $Y = 0.000 + 0.880 \times X_2 - 0.422 \times X_3 + 1.233 \times X_4 + 0.211 \times X_5 - 0.368 \times X_7$.

### 17.4.2 Discussion

Table 17.8 summarizes the results obtained when the backward elimination procedure is applied to the data of Table 17.1. At each step, an $F_{drop}$ is computed for each variable that is still in the model, and the variable yielding the lowest $F_{drop}$ is selected as the one to drop at that step. The procedure stops when there is no nonsignificant $F_{drop}$ at a step.

Again, the backward elimination procedure was a shortcut compared to testing all possible models. At the first step the procedure considered all seven of the possible six-predictor models, so there was no savings. But at the second step the procedure considered only six of the 21 possible five-predictor models, so there was considerable savings at this step. Similarly, the procedure would have ignored most of the possible four and three predictor models in the next two steps, if it had continued on. As

with forward selection, a drawback of the procedure is that one of these ignored models could have been the best model, in which case the backward elimination procedure would have missed it. For example, the backward elimination procedure did not find the best model for this data set, which was already identified in Section 17.2 as the model with predictors $X_2$, $X_3$, $X_4$, $X_6$, and $X_7$. Interestingly, the backward elimination procedure selected a different "best" model than that selected by the forward selection procedure, too. This conflict will be discussed further in the final section of this chapter.

## 17.5 Stepwise Procedure

The stepwise procedure is essentially a combination of the forward selection and backward elimination procedures. Like forward selection, it begins without any predictors and tries to add them, step by step, until it can no longer improve the model. In addition, though, it also tries to drop out predictors at each step. The rationale for this is that a predictor added in an early step may become unnecessary once some further predictors are added, if those further predictors explain everything accounted for by the early predictor. In that case, the early predictor should be dropped out of the model to simplify it.

Table 17.9 shows the rules for the stepwise procedure.

### 17.5.1 Illustration with data of Table 17.1.

**Step 0:** The procedure begins with the model having no predictors at all: $Y = a$.

**Step 1:**

**Drop part:** There is nothing to do in this part at step 1, because the model does not yet have any predictors.

**Add part:** *This part is always identical to Step 1 of the forward selection procedure.* In this part, the procedure chooses the best single predictor from the available pool. To do this, it computes each of the possible one-predictor models and gets an $F_{add}$ for each one using Equation 17.1 from Table 17.4. For example, the $F_{add}$ for $X_1$ is

$$F_{add} \text{ for } X_1 = \frac{614.095 - 0.000}{395.666} = 1.552$$

The values of 614.095 and 395.666 are the $SS_{model}$ and $MS_{error}$ for the model with $X_1$ as the only predictor (see Table 17.1). The value of 0.000 is the $SS_{model}$ for the model from the previous step (i.e., $Y = a$). Here is the full table of $F_{add}$ values for this step, each computed using the one-predictor model with the indicated predictor:

| Variable: | $X_1$ | $X_2$ | $X_3$ | |
|---|---|---|---|---|
| $F_{add}$: | $\frac{614.095-0.000}{395.666}$ | $\frac{1166.328-0.000}{384.161}$ | $\frac{49.416-0.000}{407.430}$ | |
| | 1.552 | 3.036 | 0.121 | |
| Variable: | $X_4$ | $X_5$ | $X_6$ | $X_7$ |
| $F_{add}$: | $\frac{7698.263-0.000}{248.079}$ | $\frac{3533.240-0.000}{334.850}$ | $\frac{3562.206-0.000}{334.247}$ | $\frac{1388.865-0.000}{379.525}$ |
| | 31.031 | 10.552 | 10.657 | 3.659 |

The highest $F_{add}$ is the one obtained for $X_4$, namely 31.031. This extra $F$ has 1 and $50 - 1 - 1 = 48$ degrees of freedom, because there are 50 cases in the data set and because one df is used by $a$ and another is used by the $b$ for the one predictor in the larger model. Thus, the critical $F$ is slightly less than 4.08, and 31.031 is significant. We therefore add $X_4$ to the model, and the new model at the end of step one is:

$$Y = a + b \times X_4$$

**Step 2:** The current model for this step is $Y = a + b_1 \times X_4$

**Drop part:** *This part is only a formality at this step. It is not possible for a predictor to be dropped until there are at least three predictors in the model.* The procedure now computes an $F_{drop}$ for $X_4$, the variable that is in the model. This is pretty silly because that variable was just added, but it happens anyway because the procedure is automatic. The drop part does make more sense in later steps. Anyway, at this point it uses Equation 17.2 from Table 17.7 to compute

| Variable: | $X_1$ | $X_2$ | $X_3$ | |
|---|---|---|---|---|
| $F_{drop}$: | | | | |

| Variable: | $X_4$ | $X_5$ | $X_6$ | $X_7$ |
|---|---|---|---|---|
| $F_{drop}$: | $\frac{7698.263-0.000}{248.079}$ | | | |
| | 31.031 | | | |

The values of 7698.263 and 248.079 are the $SS_{model}$ and $MS_{error}$ for the model with $X_4$, because this was the current model for this step. The value of 0.000 is the $SS_{model}$ for the smaller model obtained by dropping $X_4$ from the current model for this step. Given that the $F_{drop}$ of 31.031 is significant, $X_4$ is *not* dropped.

**Add part:** *This part is always the same as Step 2 of the forward selection procedure.* The procedure now chooses the best second predictor to add to $X_4$. That is, it determines which is the best additional predictor *given that $X_4$ is already in the model.* To do this, it computes each of the two-predictor models including $X_4$, and gets an $F_{add}$ for each added predictor using Equation 17.1. For example, the $F_{add}$ for $X_1$ at this step is

$$F_{add} \text{ for } X_1 = \frac{11519.160 - 7698.263}{172.062} = 22.207$$

The values of 11519.160 and 172.062 are the $SS_{model}$ and $MS_{error}$ for the model with $X_1$ and $X_4$ as the predictors (see Table 17.1). The value of 7698.263 is the $SS_{model}$ for the model from the previous step (i.e., $Y = a + b \times X_4$). Here is the full table of $F_{add}$ values for this step, each computed using the two-predictor model with the indicated predictor and $X_4$:

| Variable: | $X_1$ | $X_2$ | $X_3$ | |
|---|---|---|---|---|
| $F_{add}$: | $\frac{11519.160-7698.263}{172.062}$ | $\frac{19525.407-7698.263}{1.716}$ | $\frac{7731.686-7698.263}{252.646}$ | |
| | 22.207 | 6892.275 | 0.132 | |

| Variable: | $X_4$ | $X_5$ | $X_6$ | $X_7$ |
|---|---|---|---|---|
| $F_{add}$: | | $\frac{8589.080-7698.263}{234.404}$ | $\frac{8289.696-7698.263}{240.774}$ | $\frac{7698.512-7698.263}{253.352}$ |
| | | 3.800 | 2.456 | 0.001 |

Note that there is no $F_{add}$ for $X_4$; this predictor is already included in the model from the previous step, so it cannot be added again.

At this step, the highest $F_{add}$ is the one obtained for $X_2$, namely 6892.275. This $F$ has 1 and $50 - 1 - 2 = 47$ degrees of freedom, because there are 50 cases in the data set and because one df is used by $a$ and another two are used by the $b$'s for the two predictors in the larger model. Thus, the critical $F$ is slightly less than 4.08, and this $F_{add}$ is significant. We therefore add $X_2$ to the model, and the new model at the end of this step is:

$$Y = a + b_1 \times X_4 + b_2 \times X_2$$

**Step 3:** The current model for this step is:

$$Y = a + b_1 \times X_4 + b_2 \times X_2$$

**Drop part:** *This part is still only a formality.* The procedure tries to drop each of the predictors in this step's starting model by computing:

| Variable: | $X_1$ | $X_2$ | $X_3$ | |
|---|---|---|---|---|
| $F_{drop}$: | | $\frac{19525.407-7698.263}{1.716}$ | | |
| | | 6892.275 | | |

| Variable: | $X_4$ | $X_5$ | $X_6$ | $X_7$ |
|---|---|---|---|---|
| $F_{drop}$: | $\frac{19525.407-1166.328}{1.716}$ | | | |
| | 10698.764 | | | |

The values of 19525.407 and 1.716 are the $SS_{model}$ and $MS_{error}$ for the larger model with both $X_2$ and $X_4$ as predictors. The value of 7698.263 is the $SS_{model}$ for the model with just $X_4$ (i.e., dropping $X_2$), and the value of 1166.328 is the $SS_{model}$ for the model with just $X_2$ (i.e., dropping $X_4$). Both of these $F_{drop}$ values are highly significant, so neither term is dropped.

**Add part:** *This part is always the same as Step 3 of the forward selection procedure.* The next step is to choose the best third predictor to add to $X_4$ and $X_2$—that is, to determine which third predictor is best *given that these two predictors are already in the model.* To do this, it is necessary to compute each of the three-predictor models including $X_2$ and $X_4$, and get an $F_{add}$ for each possible added predictor using Equation 17.1.

| Variable: | $X_1$ | $X_2$ | $X_3$ | |
|---|---|---|---|---|
| $F_{add}$: | $\frac{19528.127-19525.407}{1.694}$ | | $\frac{19530.455-19525.407}{1.643}$ | |
| | 1.606 | | 3.072 | |

| Variable: | $X_4$ | $X_5$ | $X_6$ | $X_7$ |
|---|---|---|---|---|
| $F_{add}$: | | $\frac{19575.842-19525.407}{0.657}$ | $\frac{19586.890-19525.407}{0.417}$ | $\frac{19532.023-19525.407}{1.609}$ |
| | | 76.766 | 147.441 | 4.112 |

For example, the values of 19528.127 and 1.694 are the $SS_{model}$ and $MS_{error}$ for the model with $X_1$, $X_2$, and $X_4$ as the predictors (see Table 17.1). The value of 19525.407 is the $SS_{model}$ for the model from the previous step (i.e., $Y = a + b_1 \times X_4 + b_2 \times X_2$). Note that there is no $F_{add}$ for $X_2$ or $X_4$, because these predictors are already included in the model from the previous step.

At this step, the highest $F_{add}$ is the one obtained for $X_6$, namely 147.441. This $F$ has 1 and $50 - 1 - 3 = 46$ degrees of freedom, because there are 50 cases in the data set and because one df is used by $a$ and another three are used by the $b$'s for the three predictors in the larger model. Thus, the critical $F$ is slightly less than 4.08, and this $F_{add}$ is significant. We therefore add $X_6$ to the model, and the new model at the end of this step is:

$$Y = a + b_1 \times X_4 + b_2 \times X_2 + b_3 \times X_6$$

**Step 4:** The current model for this step is

$$Y = a + b_1 \times X_4 + b_2 \times X_2 + b_3 \times X_6$$

**Drop part:** *This is really the first step at which a predictor could possibly be dropped. That would happen if the second and third predictors added made the first unnecessary. From now on, the drop parts are no longer a formality. Furthermore, if a variable is dropped, then the subsequent add parts will no longer identical to the comparable steps in the forward selection procedure.* In this part the procedure tries to drop each of the variables in this step's current model. It computes:

| Variable: | $X_1$ | $X_2$ | $X_3$ | |
|---|---|---|---|---|
| $F_{drop}$: | | $\frac{19586.890-8289.696}{0.417}$ | | |
| | | 27091.592 | | |

| Variable: | $X_4$ | $X_5$ | $X_6$ | $X_7$ |
|---|---|---|---|---|
| $F_{drop}$: | $\frac{19586.890-4933.100}{0.417}$ | | $\frac{19586.890-19525.407}{0.417}$ | |
| | 35140.983 | | 147.441 | |

The values of 19586.890 and 0.417 are the $SS_{model}$ and $MS_{error}$ for the larger model with all three predictors (the current model for this step), and the values of 8289.696, 19525.407, and 4933.100 are the $SS_{model}$ values for the two-predictor models dropping the indicated variable. Since all of the $F_{drop}$ values are significant, no predictor is dropped.

**Add part:** *This part is the same as Step 4 of the forward selection procedure for this data set, but it could have been different if a predictor had been dropped in the Drop part of this step.* The next step is to choose the best fourth predictor to add to $X_2$, $X_4$, and $X_6$—that is, to determine which fourth predictor is best *given that these three predictors are already in the model*. To do this, it is necessary to compute each of the four-predictor models including $X_2$, $X_4$, and $X_6$, and get an $F_{add}$ for each possible added predictor using Equation 17.1. For example, the $F_{add}$ for $X_1$ at this step is

$$F_{add} \text{ for } X_1 = \frac{19591.485 - 19586.890}{0.324} = 14.182$$

The values of 19591.485 and 0.324 are the $SS_{model}$ and $MS_{error}$ for the model with $X_1$, $X_2$, $X_4$, and $X_6$ as the predictors (see Table 17.1). The value of 19586.890 is the $SS_{model}$ for the model from the previous step (i.e., $Y = a + b_1 \times X_2 + b_2 \times X_4 + b_3 \times X_6$). Similarly, here are the $F_{add}$ values for the other predictors at this step:

| Variable: | $X_1$ | $X_2$ | $X_3$ | |
|---|---|---|---|---|
| $F_{add}$: | $\frac{19591.485-19586.890}{0.324}$ | | $\frac{19591.234-19586.890}{0.329}$ | |
| | 14.182 | | 13.204 | |

| Variable: | $X_4$ | $X_5$ | $X_6$ | $X_7$ |
|---|---|---|---|---|
| $F_{add}$: | | $\frac{19591.303-19586.890}{0.328}$ | | $\frac{19591.448-19586.890}{0.325}$ |
| | | 13.454 | | 14.025 |

At this step, the highest $F_{add}$ is the one obtained for $X_1$, namely 14.182. This $F$ has 1 and $50 - 1 - 4 = 45$ degrees of freedom, because there are 50 cases in the data set and because one df is used by $a$ and another four are used by the $b$'s for the four predictors in the larger model. Thus, the critical $F$ is slightly less than 4.08, and this $F_{add}$ is significant. We therefore add $X_1$ to the model, and the new model at the end of this step is:

$$Y = a + b_1 \times X_1 + b_2 \times X_2 + b_3 \times X_4 + b_4 \times X_6$$

**Step 5:** The current model for this step is

$$Y = a + b_1 \times X_1 + b_2 \times X_2 + b_3 \times X_4 + b_4 \times X_6$$

**Drop part:** In this part the procedure tries to drop each of the variables in this step's current model. It computes:

| Variable: | $X_1$ | $X_2$ | $X_3$ | |
|---|---|---|---|---|
| $F_{drop}$: | $\frac{19591.485-19586.890}{0.324}$ | $\frac{19591.485-11806.414}{0.324}$ | | |
| | 14.182 | 24027.997 | | |

| Variable: | $X_4$ | $X_5$ | $X_6$ | $X_7$ |
|---|---|---|---|---|
| $F_{drop}$: | $\frac{19591.485-8640.421}{0.324}$ | | $\frac{19591.485-19528.127}{0.324}$ | |
| | 33799.580 | | 195.549 | |

The values of 19591.485 and 0.324 are the $SS_{model}$ and $MS_{error}$ for the larger model with all four predictors (the current model for this step), and the other values are the $SS_{model}$ values for the three-predictor models dropping the indicated variable. Since all of the $F_{drop}$ values are significant, no predictor is dropped.

**Add part:** *This part is the same as Step 5 of the forward selection procedure, but it could have been different.* The next step is to choose the best fifth predictor to add to $X_1$, $X_2$, $X_4$, and $X_6$—that is, to determine which fifth predictor is best *given that these four predictors*

*are already in the model.* To do this, it is necessary to compute each of the five-predictor models including $X_1$, $X_2$, $X_4$, and $X_6$, and get an $F_{add}$ for each possible added predictor using Equation 17.1. For example, the $F_{add}$ for $X_3$ at this step is

$$F_{add} \text{ for } X_3 = \frac{19592.748 - 19591.485}{0.302} = 4.182$$

The values of 19592.748 and 0.302 are the $SS_{model}$ and $MS_{error}$ for the model with $X_1$, $X_2$, $X_3$, $X_4$, and $X_6$ as the predictors (see Table 17.1). The value of 19591.485 is the $SS_{model}$ for the model from the previous step. Similarly, here are the $F_{add}$ values for the other predictors at this step:

| Variable: | $X_1$ | $X_2$ | $X_3$ | |
|---|---|---|---|---|
| $F_{add}$: | | | $\frac{19592.748-19591.485}{0.302}$ | |
| | | | 4.182 | |

| Variable: | $X_4$ | $X_5$ | $X_6$ | $X_7$ |
|---|---|---|---|---|
| $F_{add}$: | | $\frac{19591.592-19591.485}{0.329}$ | | $\frac{19591.497-19591.485}{0.331}$ |
| | | 0.325 | | 0.036 |

At this step, the highest $F_{add}$ is the one obtained for $X_3$, namely 4.182. This $F$ has 1 and $50 - 1 - 5 = 44$ degrees of freedom, because there are 50 cases in the data set and because one df is used by $a$ and another five are used by the $b$'s for the five predictors in the larger model. Thus, the critical $F$ is slightly less than 4.08, and this $F_{add}$ is significant. We therefore add $X_3$ to the model, and the new model at the end of this step is:

$$Y = a + b_1 \times X_1 + b_2 \times X_2 + b_3 \times X_3 + b_4 \times X_4 + b_5 \times X_6$$

**Step 6:** The current model for this step is

$$Y = a + b_1 \times X_1 + b_2 \times X_2 + b_3 \times X_3 + b_4 \times X_4 + b_5 \times X_6$$

**Drop part:** In this part the procedure tries to drop each of the variables in this step's current model. It computes:

| Variable: | $X_1$ | $X_2$ | $X_3$ | |
|---|---|---|---|---|
| $F_{drop}$: | $\frac{19592.748-19591.234}{0.302}$ | $\frac{19592.748-19445.446}{0.302}$ | $\frac{19592.748-19591.485}{0.302}$ | |
| | 5.013 | 487.755 | 4.182 | |

| Variable: | $X_4$ | $X_5$ | $X_6$ | $X_7$ |
|---|---|---|---|---|
| $F_{drop}$: | $\frac{19592.748-19560.857}{0.302}$ | | $\frac{19592.748-19592.608}{0.302}$ | |
| | 105.599 | | 0.464 | |

The values of 19592.748 and 0.302 are the $SS_{model}$ and $MS_{error}$ for the larger model with all five predictors (the current model for this step), and the other values are the $SS_{model}$ values for the four-predictor models dropping the indicated variable. The smallest $F_{drop}$ value is the one for $X_6$, and it is not significant. Thus, this predictor is dropped and we go on to the next step. The new model at the end of this step is:

$$Y = a + b_1 \times X_1 + b_2 \times X_2 + b_3 \times X_3 + b_4 \times X_4$$

**Add part:** This part is skipped because a predictor was dropped.

*Note that at last the stepwise procedure has diverged from the forward selection procedure!* This happened because $X_6$, added as a good predictor at an early step, was no longer needed after some other predictors had been added. This means that the other predictors convey the information provided by $X_6$, so it is redundant with them.

**Step 7:** The current model at the beginning of this step is:

$$Y = a + b_1 \times X_1 + b_2 \times X_2 + b_3 \times X_3 + b_4 \times X_4$$

**Drop part:** In this part the procedure tries to drop each of the variables in this step's current model. It computes:

| Variable: | $X_1$ | $X_2$ | $X_3$ |
|---|---|---|---|
| $F_{drop}$: | $\frac{19592.608-19530.455}{0.299}$ | $\frac{19592.608-15060.491}{0.299}$ | $\frac{19592.608-19528.127}{0.299}$ |
| | 207.870 | 15157.582 | 215.656 |

| Variable: | $X_4$ | $X_5$ | $X_6$ | $X_7$ |
|---|---|---|---|---|
| $F_{drop}$: | $\frac{19592.608-15056.763}{0.299}$ | | | |
| | 15170.050 | | | |

The values of 19592.608 and 0.299 are the $SS_{model}$ and $MS_{error}$ for the larger model with all four predictors (the current model for this step), and the other values are the $SS_{model}$ values for the three-predictor models dropping the indicated variable. Since all the $F_{drop}$ values are significant, no predictor is dropped. Here is the table summarizing the drop part:

**Add part:** The next step is to choose the best fifth predictor to add to the four already chosen—that is, to determine which fifth predictor is best *given that these four predictors are already in the model*. Here is the $F_{add}$ value for each of the unused predictors at this step:

| Variable: | $X_1$ | $X_2$ | $X_3$ |
|---|---|---|---|
| $F_{add}$: | | | |

| Variable: | $X_4$ | $X_5$ | $X_6$ | $X_7$ |
|---|---|---|---|---|
| $F_{add}$: | | $\frac{19592.868-19592.608}{0.300}$ | $\frac{19592.748-19592.608}{0.302}$ | $\frac{19592.730-19592.608}{0.303}$ |
| | | 0.867 | 0.464 | 0.403 |

The value 19592.608 is the $SS_{model}$ for the current four-predictor model. The other values are the $SS_{model}$ and $MS_{error}$ values for the five-predictor models with the indicated variables added.

Note that an $F_{add}$ is computed for $X_6$, even though this predictor was just dropped at the previous step. This is just a formality at this step, because we know from the previous $F_{drop}$ computation that $X_6$ cannot be significant here. Nonetheless, the rule is always to compute $F_{add}$'s for all predictors not in the model. Crucially, $X_6$ could be added back in to the model at a later step, even though it was dropped earlier. This might happen if some other variables added in the meantime greatly reduced error for $X_6$, allowing a significant $F$ to indicate that it was needed in the model after all.

At this step, the highest $F_{add}$ is 0.867. This $F$ has 1 and $50-1-5=44$ degrees of freedom, because there are 50 cases in the data set and because one df is used by $a$ and another five are used by the $b$'s for the five predictors in the larger model. Thus, the critical $F$ is slightly less than 4.08, and this $F_{add}$ is not significant, and we do not add a variable.

We stop at this step, because no variable was dropped or added. The final model is the starting one for this step—in this case, the four predictor model with $X_1$–$X_4$. This four predictor model is the best model that the stepwise procedure can find. Here is the exact equation for this model: $Y = 0.000 + 0.226 \times X_1 + 1.076 \times X_2 - 0.124 \times X_3 + 1.102 \times X_4$.

### 17.5.2 Discussion

Table 17.10 summarizes the results obtained when the stepwise procedure is applied to the data of Table 17.1. At each step, an $F_{drop}$ is computed for each variable that is still in the model, and the variable yielding the lowest $F_{drop}$ is selected as the one to drop at that step. If no variable is dropped, then an $F_{add}$ is computed for each variable that is not yet in the model, and the variable yielding the highest $F_{add}$ is selected as the one to add. The procedure stops when a step yields neither a nonsignificant $F_{drop}$ nor a significant $F_{add}$ (i.e., all are significant).

The stepwise procedure considers more models than the forward selection or the backward elimination procedure, because it attempts to both drop and add predictors at each step. It is still a

shortcut procedure, however, because it does not consider all the possible models. In fact, it only considers about twice as many models as the forward selection and backward elimination procedures, and this is still far short of the total number of models that might be considered. As was true for the other shortcuts, the stepwise procedure also failed to find the best model for this data set, which was already identified in a previous section as the model with predictors $X_2$, $X_3$, $X_4$, $X_6$, and $X_7$.

## 17.6 How Much Can You Trust the "Best Model"

Although the procedures covered in this chapter are the best available ways to address the problem of finding the best model, they are far from perfect. It is worthwhile to keep their defects in mind, even as you use them. There are two distinct problems: (1) You may not find the true best model for the sample, and (2) Even if you do, it may not be the best model for the population (i.e., not the truly best model).

### 17.6.1 Finding the Best Model for the Sample

The example data used in this chapter (Table 17.1) are very unusual—one might even say pathological—in that each of the shortcut procedures finds a different "suggested best" model, none of which is truly the best model (as revealed when we examined all possible models). Obviously, this example was selected to illustrate that this can happen and to emphasize the shortcomings of the shortcut procedures. One lesson to be learned from this is that all possible models must be examined in order to be *sure* of finding the true best model for the sample.

Even if all possible models cannot be examined, it is a good idea to try *all three* of the shortcut procedures. If all three procedures end up with the same suggested best model, then you can have a little more confidence that this is indeed the best model for the sample. If they end up with different best models, you should choose the one with the smallest $MS_{error}$, and you obviously have a better chance of having arrived at the actual best model for the sample than if you had only tried one of the shortcut procedures.

### 17.6.2 Finding the Best Model for the Population

Another important caveat for users of these procedures—including those who try all possible models—is that there is no guarantee that you will find the best model for the population. Here, the problem is that the results can be heavily influenced by chance, which means that the best model for the sample may not be the best model for the population.

As an illustration, take a close look at the four $F_{add}$ values obtained in Step 4 of the forward selection procedure (page 194). These four values are nearly identical, which means that it is just luck which one happened to be the highest. Put another way, the $F_{add}$ for $X_5$ might very well have been the highest with only a slightly different sample. In that case we would have added $X_5$ at this step and had $X_5$ in the final model. As it was, however, the forward selection procedure ultimately did not include $X_5$ in the final model. Thus, it seems clear that chance partly determined which variables were in the final model, in this case at least. The general lesson, then, is that selection of the best model is still subject to chance, and you are not at all guaranteed to reach the model that is truly best (i.e., for the population). Many researchers typically ignore this fact, because there are no simple ways to assess chance's influences on model selection. But it is a fact just the same.

| $SS_{model}$ | $MS_{error}$ | Predictors in model | $SS_{model}$ | $MS_{error}$ | Predictors in model |
|---|---|---|---|---|---|
| 614.095 | 395.666 | X1 | 19592.608 | 0.299 | X1 X2 X3 X4 |
| 1166.328 | 384.161 | X2 | 19567.137 | 0.865 | X1 X2 X3 X5 |
| 49.416 | 407.430 | X3 | 19560.857 | 1.004 | X1 X2 X3 X6 |
| 7698.263 | 248.079 | X4 | 19552.211 | 1.196 | X1 X2 X3 X7 |
| 3533.240 | 334.850 | X5 | 19591.576 | 0.322 | X1 X2 X4 X5 |
| 3562.206 | 334.247 | X6 | 19591.485 | 0.324 | X1 X2 X4 X6 |
| 1388.865 | 379.525 | X7 | 19580.974 | 0.557 | X1 X2 X4 X7 |
| 7194.491 | 264.076 | X1 X2 | 19268.992 | 7.490 | X1 X2 X5 X6 |
| 1079.135 | 394.190 | X1 X3 | 19541.859 | 1.426 | X1 X2 X5 X7 |
| 11519.160 | 172.062 | X1 X4 | 19543.893 | 1.381 | X1 X2 X6 X7 |
| 3593.878 | 340.685 | X1 X5 | 19403.431 | 4.503 | X1 X3 X4 X5 |
| 3666.388 | 339.142 | X1 X6 | 19445.446 | 3.569 | X1 X3 X4 X6 |
| 2378.688 | 366.540 | X1 X7 | 19525.283 | 1.795 | X1 X3 X4 X7 |
| 1207.748 | 391.453 | X2 X3 | 19175.507 | 9.568 | X1 X3 X5 X6 |
| 19525.407 | 1.716 | X2 X4 | 19286.151 | 7.109 | X1 X3 X5 X7 |
| 4567.078 | 319.978 | X2 X5 | 18983.568 | 13.833 | X1 X3 X6 X7 |
| 4933.100 | 312.190 | X2 X6 | 19456.758 | 3.318 | X1 X4 X5 X6 |
| 3380.478 | 345.225 | X2 X7 | 19528.607 | 1.721 | X1 X4 X5 X7 |
| 7731.686 | 252.646 | X3 X4 | 19528.571 | 1.722 | X1 X4 X6 X7 |
| 6440.870 | 280.110 | X3 X5 | 18305.457 | 28.902 | X1 X5 X6 X7 |
| 4868.371 | 313.568 | X3 X6 | 19591.078 | 0.333 | X2 X3 X4 X5 |
| 2407.721 | 365.922 | X3 X7 | 19591.234 | 0.329 | X2 X3 X4 X6 |
| 8589.080 | 234.404 | X4 X5 | 19592.487 | 0.301 | X2 X3 X4 X7 |
| 8289.696 | 240.774 | X4 X6 | 18454.647 | 25.587 | X2 X3 X5 X6 |
| 7698.512 | 253.352 | X4 X7 | 16749.630 | 63.476 | X2 X3 X5 X7 |
| 3595.973 | 340.640 | X5 X6 | 10408.521 | 204.390 | X2 X3 X6 X7 |
| 4866.327 | 313.611 | X5 X7 | 19591.303 | 0.328 | X2 X4 X5 X6 |
| 4138.791 | 329.091 | X6 X7 | 19591.534 | 0.323 | X2 X4 X5 X7 |
| 15056.763 | 98.898 | X1 X2 X3 | 19591.448 | 0.325 | X2 X4 X6 X7 |
| 19528.127 | 1.694 | X1 X2 X4 | 18071.980 | 34.090 | X2 X5 X6 X7 |
| 7747.165 | 257.802 | X1 X2 X5 | 19305.478 | 6.679 | X3 X4 X5 X6 |
| 8640.421 | 238.383 | X1 X2 X6 | 19466.104 | 3.110 | X3 X4 X5 X7 |
| 19040.863 | 12.287 | X1 X2 X7 | 19390.454 | 4.791 | X3 X4 X6 X7 |
| 15060.491 | 98.817 | X1 X3 X4 | 17972.720 | 36.296 | X3 X5 X6 X7 |
| 7459.122 | 264.064 | X1 X3 X5 | 18504.116 | 24.487 | X4 X5 X6 X7 |
| 5669.611 | 302.966 | X1 X3 X6 | 19592.868 | 0.300 | X1 X2 X3 X4 X5 |
| 2689.055 | 367.761 | X1 X3 X7 | 19592.748 | 0.302 | X1 X2 X3 X4 X6 |
| 12546.247 | 153.474 | X1 X4 X5 | 19592.730 | 0.303 | X1 X2 X3 X4 X7 |
| 11806.414 | 169.557 | X1 X4 X6 | 19569.496 | 0.831 | X1 X2 X3 X5 X6 |
| 19412.839 | 4.200 | X1 X4 X7 | 19567.293 | 0.881 | X1 X2 X3 X5 X7 |
| 3679.725 | 346.224 | X1 X5 X6 | 19560.888 | 1.026 | X1 X2 X3 X6 X7 |
| 5078.094 | 315.825 | X1 X5 X7 | 19591.592 | 0.329 | X1 X2 X4 X5 X6 |
| 4408.970 | 330.371 | X1 X6 X7 | 19591.576 | 0.329 | X1 X2 X4 X5 X7 |
| 19530.455 | 1.643 | X2 X3 X4 | 19591.497 | 0.331 | X1 X2 X4 X6 X7 |
| 7318.127 | 267.129 | X2 X3 X5 | 19547.474 | 1.331 | X1 X2 X5 X6 X7 |
| 6248.040 | 290.392 | X2 X3 X6 | 19470.127 | 3.089 | X1 X3 X4 X5 X6 |
| 5731.756 | 301.615 | X2 X3 X7 | 19539.776 | 1.506 | X1 X3 X4 X5 X7 |
| 19575.842 | 0.657 | X2 X4 X5 | 19537.663 | 1.554 | X1 X3 X4 X6 X7 |
| 19586.890 | 0.417 | X2 X4 X6 | 19341.817 | 6.005 | X1 X3 X5 X6 X7 |
| 19532.023 | 1.609 | X2 X4 X7 | 19528.607 | 1.760 | X1 X4 X5 X6 X7 |
| 4993.055 | 317.674 | X2 X5 X6 | 19591.325 | 0.335 | X2 X3 X4 X5 X6 |
| 6664.492 | 281.338 | X2 X5 X7 | 19593.845 | 0.277 | X2 X3 X4 X5 X7 |
| 6083.218 | 293.975 | X2 X6 X7 | 19593.936 | 0.275 | X2 X3 X4 X6 X7 |
| 9704.857 | 215.243 | X3 X4 X5 | 18480.148 | 25.589 | X2 X3 X5 X6 X7 |
| 8689.054 | 237.326 | X3 X4 X6 | 19591.581 | 0.329 | X2 X4 X5 X6 X7 |
| 7819.780 | 256.223 | X3 X4 X7 | 19474.513 | 2.990 | X3 X4 X5 X6 X7 |
| 13337.425 | 136.274 | X3 X5 X6 | 19593.151 | 0.300 | X1 X2 X3 X4 X5 X6 |
| 8954.562 | 231.554 | X3 X5 X7 | 19594.670 | 0.265 | X1 X2 X3 X4 X5 X7 |
| 6998.527 | 274.077 | X3 X6 X7 | 19594.541 | 0.268 | X1 X2 X3 X4 X6 X7 |
| 8993.468 | 230.708 | X4 X5 X6 | 19569.604 | 0.848 | X1 X2 X3 X5 X6 X7 |
| 8610.245 | 239.039 | X4 X5 X7 | 19591.608 | 0.336 | X1 X2 X4 X5 X6 X7 |
| 8293.475 | 245.926 | X4 X6 X7 | 19539.776 | 1.541 | X1 X3 X4 X5 X6 X7 |
| 17815.541 | 38.924 | X5 X6 X7 | 19593.957 | 0.281 | X2 X3 X4 X5 X6 X7 |
| | | | 19594.697 | 0.270 | X1 X2 X3 X4 X5 X6 X7 |

Table 17.1: Summary of fits of all possible models to predict $Y$ from $X_1$–$X_7$. Note that for some of the more complicated models the $MS_{error}$ terms are tiny compared with the $SS_{model}$ values. This is responsible for the extremely large $F$'s computed in certain comparisons in this chapter.

**Part 1:** Find the model with the smallest $MS_{error}$ of all remaining models.

**Part 2:** Compute the $F_{drop}$ associated with dropping each predictor in this model.

**Part 3:** Are all these $F_{drop}$ values significant?

**Yes:** Stop. This is the best model.

**No:** Cross out this model and return to Part 1.

Table 17.2: Rules of the "All Possible Models" procedure. Each step of the procedure has three parts.

| Number of Predictors | Number of Possible Models | Maximum number of models considered: | | |
|---|---|---|---|---|
| | | Forward Selection | Backward Elimination | Stepwise |
| 2 | 3 | 3 | 3 | 3 |
| 5 | 31 | 15 | 15 | 30 |
| 10 | 1,023 | 55 | 55 | 110 |
| 20 | 1,048,575 | 210 | 210 | 420 |
| 30 | 1,073,741,824 | 465 | 465 | 930 |

Table 17.3: Illustration of the number of possible models and the numbers of models considered by each of the three shortcut procedures considered later in this chapter: forward selection, backward elimination, and stepwise. The maximum number of models considered by the stepwise procedure is only a rough estimate, because this procedure takes an unpredictable number of steps.

- The starting model is $Y = a$.
- At each step, consider adding each variable not already in the model. For each one, compute:

$$F_{add} = \frac{SS_{larger} - SS_{smaller}}{MS_{error,\ larger\ model}} \tag{17.1}$$

- Add whichever variable gives the highest $F_{add}$, but only if that $F_{add}$ is significant.
- Stop when no variable can be added to the model with a significant $F_{add}$.

Table 17.4: Rules of the forward selection procedure.

| Step | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | $X_7$ |
|---|---|---|---|---|---|---|---|
| 1 | 1.552 | 3.036 | 0.121 | 31.031 | 10.552 | 10.657 | 3.659 |
| 2 | 22.207 | 6892.275 | 0.132 | | 3.800 | 2.456 | 0.001 |
| 3 | 1.606 | | 3.072 | | 76.766 | 147.441 | 4.112 |
| 4 | 14.182 | | 13.204 | | 13.454 | | 14.025 |
| 5 | | | 4.182 | | 0.325 | | 0.036 |
| 6 | | | | | 1.343 | | 6.690 |
| 7 | | | | | 0.578 | | |

Table 17.5: Summary of $F_{add}$ values for the forward selection procedure applied to the data of Table 17.1.

| Step | Number of predictors in larger model | Number of models examined | Number of models possible |
|---|---|---|---|
| 1 | 1 | 7 | 7 |
| 2 | 2 | 6 | 21 |
| 3 | 3 | 5 | 35 |
| 4 | 4 | 4 | 35 |
| 5 | 5 | 3 | 21 |
| 6 | 6 | 2 | 7 |
| 7 | 7 | 1 | 1 |

Table 17.6: The number of models examined by the forward selection procedure applied to the data in Table 17.1, as compared with the number of possible models that might have been examined. Comparing the right-most two columns, it is clear that the procedure considers all of the possible one-predictor models at step 1, but does not consider all of the possible models with two predictors, three predictors, and so on.

- The starting model is the one predicting $Y$ from *all* of the $X$'s.
- At each step, consider dropping each variable already in the model. For each one, compute:

$$F_{drop} = \frac{SS_{larger} - SS_{smaller}}{MS_{error,\ larger\ model}} \tag{17.2}$$

  Drop whichever variable gives the lowest $F_{drop}$, as long as this $F$ is not significant.
- Stop when every variable still in the model gives a significant $F_{drop}$.

Table 17.7: Rules of the backward elimination procedure.

| Step | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | $X_7$ |
|---|---|---|---|---|---|---|---|
| 1 | 2.741 | 203.411 | 11.441 | 92.937 | 0.578 | 0.100 | 5.726 |
| 2 | 3.113 | 207.147 | 11.675 | 103.309 | 7.321 | | 6.800 |
| 3 | | 461.159 | 8.343 | 10267.924 | 4.903 | | 9.989 |

Table 17.8: Summary of $F_{drop}$ values for the backward elimination procedure applied to the data of Table 17.1.

- The starting model is $Y = a$.
- Each step has two parts:

  **Drop part:** Consider dropping each variable already in the model. For each one, compute an $F_{drop}$ with Equation 17.2 from Table 17.7. Drop whichever variable gives the lowest $F_{drop}$, as long as this $F$ is not significant. If a variable is dropped, go on to the next step. If no variable is dropped, go on to the next part of this step.

  **Add part:** Consider adding each variable not already in the model. For each one, compute an $F_{add}$ with Equation 17.1. from Table 17.4. Add whichever variable gives the highest $F_{add}$, but only if that $F$ is significant.
- Stop when the model does not change at a step (i.e., no variable is dropped or added).

Table 17.9: Rules of the stepwise procedure.

| Step | Part | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | $X_7$ |
|---|---|---|---|---|---|---|---|---|
| 1 | Drop | | | | | | | |
| | Add | 1.552 | 3.036 | 0.121 | 31.031 | 10.552 | 10.657 | 3.659 |
| 2 | Drop | | | | 31.031 | | | |
| | Add | 22.207 | 6892.275 | 0.132 | | 3.800 | 2.456 | 0.001 |
| 3 | Drop | | 6892.275 | | 10698.764 | | | |
| | Add | 1.606 | | 3.072 | | 76.766 | 147.441 | 4.112 |
| 4 | Drop | | 27091.592 | | 35140.983 | | 147.441 | |
| | Add | 14.182 | | 13.204 | | 13.454 | | 14.025 |
| 5 | Drop | 14.182 | 24027.997 | | 33799.580 | | 195.549 | |
| | Add | | | 4.182 | | 0.325 | | 0.036 |
| 6 | Drop | 5.013 | 487.755 | 4.182 | 105.599 | | 0.464 | |
| | Add | | | | | | | |
| 7 | Drop | 207.870 | 15157.582 | 215.656 | 15170.050 | | | |
| | Add | | | | | 0.867 | 0.464 | 0.403 |

Table 17.10: Summary of $F_{add}$ and $F_{drop}$ values for the stepwise procedure applied to the data of Table 17.1.

# Part III

# Analysis of Covariance

# Chapter 18

# Dummy Variable Regression

## 18.1 Overview

Dummy variable regression (DVR) and analysis of covariance (ANCOVA) are methods of statistical analysis that combine analysis of variance and multiple regression. As in ANOVA, there are several groups of scores defined by the levels on one or more factors. As in regression, each value of the dependent variable $Y$ is associated with scores on one or more $X$ variables (called *covariates* rather than predictors, in DVR and ANCOVA terminology). ANCOVA is really a special case of DVR, with a certain specific goal, which we will describe later. For now, we will discuss only DVR, but you should keep in mind that these ideas all apply to ANCOVA as well.

DVR can be thought of as an analysis of the relationship of $Y$ to $X$ in several groups at the same time. Graphically, this corresponds to a scattergram with a number of different samples plotted on it, where each sample represents a different level of an ANOVA factor. Figure 18.1 shows the simplest kind of example: a single ANOVA factor with two levels (Gender) and a single covariate.

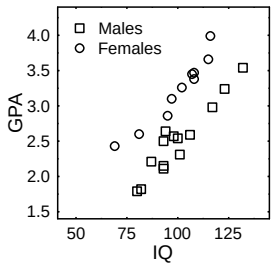


Figure 18.1: GPA as a function of IQ for males and females separately. Each symbol represents one individual.

## 18.2 The General Linear Model for DVR and ANCOVA

The general linear model for DVR includes terms for both the categorical variables of ANOVA (i.e., factor main effects and interactions) and the numerical variables of regression (i.e., $X$'s). For example, here is one way to write the GLM for the data set graphed above:

$$Y_{ij} = \mu + A_i + (b \times X_{ij}) + (b \times A_i \times X_{ij}) + S(A)_{ij}$$

Note that a DVR model is more or less a combination of all the terms from both the ANOVA and the regression model. We now have extra subscripts on the $X$'s because we have to indicate which $Y$ value each $X$ corresponds to, and there are several subscripts on the $Y$'s. The unfamiliar term $(b \times A_i \times X_{ij})$ reflects an interaction between the experimental factor A and the covariate $X$.

This term measures the extent to which the slope relating $X$ to $Y$ is different at the different levels of Factor A.

## 18.3 Computations For DVR

To use a DVR model, we follow the same general steps that we have followed in both ANOVA and regression.

1. Estimate the parameters in the model.
2. Compute sums of squares attributable to each term in the model (i.e., each source).
3. Count degrees of freedom for each source.
4. Compute mean squares and F's.

As we will see, the interpretations of significant $F$'s change only slightly when using a DVR model rather than simply an ANOVA or regression model.

Unfortunately, getting the estimates and breaking up the sum of squares is computationally very difficult. The estimation equations we learned for dealing with ANOVA do not work when there is a covariate in the model. Regression programs (e.g., Minitab) do not make distinctions among different groups of cases (corresponding to factor levels), because they work in a case by variable format.

The solution to the problem is to learn the new technique of *dummy variable regression. Dummy variables* are special $X$ variables that are constructed in order to represent ANOVA factors and interactions. After the dummy variables have been constructed, they are entered into a regression program just like any other $X$ variables. The regression program is then used to find the estimates and sums of squares for the dummy variables. Because of the way the dummy variables are set up, the estimates and $SS$'s obtained from the regression program are exactly the ones that correspond to the ANOVA factors and their interactions. In short, dummy variable regression is a technique for doing ANOVA with a regression program, and the dummy variables are the special $X$ variables needed to do that.

First, you must learn how to construct dummy variables to do ANOVA with a regression program. This may seem like a lot of work for nothing, since you can already do ANOVA. However, these same dummy variables can also be used to do DVR with a regression program, and that is a new technique.

## 18.4 Effects Coding

There are actually several methods of constructing or "coding" the variables that will work as described above; we will use one called "effects coding". Though it is a general method that can be used for both between-subjects and within-subjects factors, we will only study the method as applied to between-subjects designs.

The values of the dummy variables are uniquely determined by the structure of the experiment. In fact, once you get used to the different patterns of dummy variables you will find it easy to figure out the structure of the experimental design by looking at the dummy variables.

Use the following rules to construct the dummy variables: (As you read through the rules for the first time, refer to Example 1 which appears immediately after these rules.)

1. List the $Y$ values in one long column corresponding to the first variable in the case-by-variable format. Then perform Steps 2-5 once for each factor in the design:
2. For the factor you are currently coding, make column headings for $k-1$ dummy variables (i.e., $X$'s), where $k$ is the number of levels of the factor (i.e., there is one dummy variable for each *df* of the factor.) We will label these columns with headings of the form $XA_{ji}$. $XA$ means that the variable is a dummy variable for coding factor $A$. The subscript $j$ indexes the $k-1$ dummy variables needed to code this factor, and the subscript $i$ indexes the cases (i.e., the different values of $Y$ for which we need corresponding values of the dummy variable). Thus, $XA_{19}$ is the value of the first dummy variable coding Factor A for the ninth case. Similarly, $XB_{47}$ is the value of the fourth dummy variable used to code Factor B, for the seventh case.
3. Decide on an ordering of the levels of the factor. Call one level "level 1", another "level 2", and so on, up to "level k", which is the last level of the factor. The ordering of levels is arbitrary. For example, when coding a Gender factor, we might decide to call the males "level 1" and the females "level 2."

4. Assign a numerical value to each case on each of the columns you set up in Step 2. The numerical values are all either 0, +1, or -1, according to the following pattern:

   - Set $XA_{1i} = 1$ for $Y_i$ scores at level 1 of the factor, set $XA_{1i} = 0$ for $Y_i$ scores at level 2, 3, ..., k-1 of the factor, and set $XA_{1i} = -1$ for $Y_i$ scores at level k of the factor.
   - Set $XA_{2i} = 1$ for $Y_i$ scores at level 2 of the factor, set $XA_{2i} = 0$ for $Y_i$ scores at level 1, 3, 4, ..., k-1 of the factor, and set $XA_{2i} = -1$ for $Y_i$ scores at level k of the factor.
   - Set $XA_{3i} = 1$ for $Y_i$ scores at level 3 of the factor, set $XA_{3i} = 0$ for $Y_i$ scores at level 1, 2, 4, ..., k-1 of the factor, and set $XA_{3i} = -1$ for $Y_i$ scores at level k of the factor.
   - And so on for all $k - 1$ $XA$ dummy variables.

   It is possible to state this rule more succinctly, using notation that is a bit more complicated:

$$XA_{ji} = \begin{cases} 1, & \text{if } Y_i \text{ is a score at level } j \text{ of Factor A and } j \leq k-1 \\ 0, & \text{if } Y_i \text{ is a score at level } j' \neq j \text{ of Factor A and } j \leq k-1 \\ -1, & \text{if } Y_i \text{ is a score at level } k \text{ of Factor A.} \end{cases}$$

5. **Coding interactions.** (See Examples 2 and 3.) In multifactor designs, of course you have to set up dummy variables to code the interactions that would normally be examined with ANOVA. This is easier than it sounds.

   You don't need to do anything special when you code the first factor, because the procedure codes interactions of new factors with factors already coded; when you are coding the first factor, there is nothing already coded for it to interact with. When coding any factor after the first, you need to set up columns to code the interaction(s) of the factor you are currently coding with the factor(s) you have previously coded (and with the previous interactions, if you have previously coded two or more factors). Each new interaction column corresponds to a pair made by taking one of the columns set up for coding previous factors (or interactions) and one of the columns set up for the current factor.

   Proceed as follows:

   - Go back to the beginning of the previously set up $X$'s.
   - Take the first of the previous $X$'s, and make $k - 1$ new columns in which you pair it with each of the $k - 1$ dummy variables for the current factor.
   - Take the second of the previous $X$'s, and make $k - 1$ new columns in which you pair it with each of the $k - 1$ dummy variables for the current factor.
   - And so on, until you run out of previous $X$'s. (Stop when you get up to the $X$'s for the current factor.)
   - The numerical values in each column are found by multiplying (for each case) the numerical values in the two columns making up the pair.
   - For columns coding interactions, we will use labels like $XAB_{12i}$. The $XAB$ indicates that it is a dummy variable for coding the $AB$ interaction, and the subscripts 12 indicate that this column was formed as the product of the first dummy variable coding $A$ and the second dummy variable coding Factor B. As with $Y_i$, the subscript $i$ indexes the case. Similarly, $XABC_{231i}$ would be a dummy variable coding the $ABC$ interaction, formed by multiplying the first dummy variable coding $C$ times the $XAB_{23}$ interaction column.

## 18.5 Example 1

Here are the data from a one-factor, between-subjects design with four levels and two subjects per cell:

| | Freshman | Sophomore | Junior | Senior |
|---|---|---|---|---|
| Data: | 98, 87 | 43, 121 | 68, 90 | 35, 71 |

The dummy variable matrix looks like this:

| $Y_i$ | $XA_{1i}$ | $XA_{2i}$ | $XA_{3i}$ |
|---|---|---|---|
| 98 | 1 | 0 | 0 |
| 87 | 1 | 0 | 0 |
| 43 | 0 | 1 | 0 |
| 121 | 0 | 1 | 0 |
| 68 | 0 | 0 | 1 |
| 90 | 0 | 0 | 1 |
| 35 | -1 | -1 | -1 |
| 71 | -1 | -1 | -1 |

Following the step-by-step description of the procedure given in the previous section, the first step is simply to list the $Y_i$ values in the left-most column of the matrix. The second step is to generate the three columns to the right of the $Y$'s, and three columns are needed because Factor A has three *df*. The third step is to assign numbers to the levels of Factor A—we chose to use "Level 1" for the Freshmen, "Level 2" for the Sophomores, "Level 3" for the Juniors, and "Level 4" for the Seniors. The fourth step is to assign the values of 0, 1, and -1 to the $XA_{ji}$'s. The fifth step is to code the interactions, but that is not needed in this example because there is only one factor.

## 18.6 Example 2

This example shows how to code dummy variables for a 2 Gender (Factor A) x 3 Class Level (Factor B) design with two subjects per cell. The data are as follows:

| | Freshman | Sophomore | Junior |
|---|---|---|---|
| Male | 12, 14 | 17, 18 | 21, 35 |
| Female | 19, 16 | 24, 30 | 27, 25 |

The dummy variable matrix looks like this:

| $Y_i$ | $XA_{1i}$ | $XB_{1i}$ | $XB_{2i}$ | $XAB_{11i}$ | $XAB_{12i}$ |
|---|---|---|---|---|---|
| 12 | 1 | 1 | 0 | 1 | 0 |
| 14 | 1 | 1 | 0 | 1 | 0 |
| 17 | 1 | 0 | 1 | 0 | 1 |
| 18 | 1 | 0 | 1 | 0 | 1 |
| 21 | 1 | -1 | -1 | -1 | -1 |
| 35 | 1 | -1 | -1 | -1 | -1 |
| 19 | -1 | 1 | 0 | -1 | 0 |
| 16 | -1 | 1 | 0 | -1 | 0 |
| 24 | -1 | 0 | 1 | 0 | -1 |
| 30 | -1 | 0 | 1 | 0 | -1 |
| 27 | -1 | -1 | -1 | 1 | 1 |
| 25 | -1 | -1 | -1 | 1 | 1 |

## 18.7 Example 3

Here are the data from a three-factor design with three levels of Class Level (Factor A) by two Genders (Factor B) by three levels of Course Material (Factor C), and two subjects per cell.

Math Courses

| | Freshman | Sophomore | Junior |
|---|---|---|---|
| Male | 12, 14 | 17, 18 | 21, 35 |
| Female | 19, 16 | 24, 30 | 27, 25 |

Science Courses

| | Freshman | Sophomore | Junior |
|---|---|---|---|
| Male | 32, 34 | 37, 38 | 41, 35 |
| Female | 39, 36 | 44, 30 | 47, 45 |

Humanities Courses

| | Freshman | Sophomore | Junior |
|---|---|---|---|
| Male | 52, 54 | 57, 58 | 61, 55 |
| Female | 59, 56 | 64, 50 | 67, 65 |

Dummy Variable Matrix

| $Y_i$ | $XA_{1i}$ | $XA_{2i}$ | $XB_{1i}$ | $XAB_{11I}$ | $XAB_{21i}$ | $XC_{1i}$ | $XC_{2i}$ | $XAC_{11i}$ | $XAC_{12i}$ | ... |
|---|---|---|---|---|---|---|---|---|---|---|
| 12 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | ... |
| 14 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | ... |
| 17 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | ... |
| 18 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | ... |
| 21 | -1 | -1 | 1 | -1 | -1 | 1 | 0 | -1 | 0 | ... |
| 35 | -1 | -1 | 1 | -1 | -1 | 1 | 0 | -1 | 0 | ... |
| 19 | 1 | 0 | -1 | -1 | 0 | 1 | 0 | 1 | 0 | ... |
| 16 | 1 | 0 | -1 | -1 | 0 | 1 | 0 | 1 | 0 | ... |
| 24 | 0 | 1 | -1 | 0 | -1 | 1 | 0 | 0 | 0 | ... |
| 30 | 0 | 1 | -1 | 0 | -1 | 1 | 0 | 0 | 0 | ... |
| 27 | -1 | -1 | -1 | 1 | 1 | 1 | 0 | -1 | 0 | ... |
| 25 | -1 | -1 | -1 | 1 | 1 | 1 | 0 | -1 | 0 | ... |
| 32 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | ... |
| 34 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | ... |
| 37 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | ... |
| 38 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | ... |
| 41 | -1 | -1 | 1 | -1 | -1 | 0 | 1 | 0 | -1 | ... |
| 35 | -1 | -1 | 1 | -1 | -1 | 0 | 1 | 0 | -1 | ... |
| 39 | 1 | 0 | -1 | -1 | 0 | 0 | 1 | 0 | 1 | ... |
| 36 | 1 | 0 | -1 | -1 | 0 | 0 | 1 | 0 | 1 | ... |
| 44 | 0 | 1 | -1 | 0 | -1 | 0 | 1 | 0 | 0 | ... |
| 30 | 0 | 1 | -1 | 0 | -1 | 0 | 1 | 0 | 0 | ... |
| 47 | -1 | -1 | -1 | 1 | 1 | 0 | 1 | 0 | -1 | ... |
| 45 | -1 | -1 | -1 | 1 | 1 | 0 | 1 | 0 | -1 | ... |
| 52 | 1 | 0 | 1 | 1 | 0 | -1 | -1 | -1 | -1 | ... |
| 54 | 1 | 0 | 1 | 1 | 0 | -1 | -1 | -1 | -1 | ... |
| 57 | 0 | 1 | 1 | 0 | 1 | -1 | -1 | 0 | 0 | ... |
| 58 | 0 | 1 | 1 | 0 | 1 | -1 | -1 | 0 | 0 | ... |
| 61 | -1 | -1 | 1 | -1 | -1 | -1 | -1 | 1 | 1 | ... |
| 55 | -1 | -1 | 1 | -1 | -1 | -1 | -1 | 1 | 1 | ... |
| 59 | 1 | 0 | -1 | -1 | 0 | -1 | -1 | -1 | -1 | ... |
| 56 | 1 | 0 | -1 | -1 | 0 | -1 | -1 | -1 | -1 | ... |
| 64 | 0 | 1 | -1 | 0 | -1 | -1 | -1 | 0 | 0 | ... |
| 50 | 0 | 1 | -1 | 0 | -1 | -1 | -1 | 0 | 0 | ... |
| 67 | -1 | -1 | -1 | 1 | 1 | -1 | -1 | 1 | 1 | ... |
| 65 | -1 | -1 | -1 | 1 | 1 | -1 | -1 | 1 | 1 | ... |

continued on next page

| ... | $XAC_{21i}$ | $XAC_{22i}$ | $XBC_{11i}$ | $XBC_{12i}$ | $XABC_{111}$ | $XABC_{112}$ | $XABC_{211}$ | $XABC_{212}$ |
|---|---|---|---|---|---|---|---|---|
| ... | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |
| ... | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |
| ... | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 |
| ... | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 |
| ... | -1 | 0 | 1 | 0 | -1 | 0 | -1 | 0 |
| ... | -1 | 0 | 1 | 0 | -1 | 0 | -1 | 0 |
| ... | 0 | 0 | -1 | 0 | -1 | 0 | 0 | 0 |
| ... | 0 | 0 | -1 | 0 | -1 | 0 | 0 | 0 |
| ... | 1 | 0 | -1 | 0 | 0 | 0 | -1 | 0 |
| ... | 1 | 0 | -1 | 0 | 0 | 0 | -1 | 0 |
| ... | -1 | 0 | -1 | 0 | 1 | 0 | 1 | 0 |
| ... | -1 | 0 | -1 | 0 | 1 | 0 | 1 | 0 |
| ... | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 |
| ... | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 |
| ... | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 |
| ... | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 |
| ... | 0 | -1 | 0 | 1 | 0 | -1 | 0 | -1 |
| ... | 0 | -1 | 0 | 1 | 0 | -1 | 0 | -1 |
| ... | 0 | 0 | 0 | -1 | 0 | -1 | 0 | 0 |
| ... | 0 | 0 | 0 | -1 | 0 | -1 | 0 | 0 |
| ... | 0 | 1 | 0 | -1 | 0 | 0 | 0 | -1 |
| ... | 0 | 1 | 0 | -1 | 0 | 0 | 0 | -1 |
| ... | 0 | -1 | 0 | -1 | 0 | 1 | 0 | 1 |
| ... | 0 | -1 | 0 | -1 | 0 | 1 | 0 | 1 |
| ... | 0 | 0 | -1 | -1 | -1 | -1 | 0 | 0 |
| ... | 0 | 0 | -1 | -1 | -1 | -1 | 0 | 0 |
| ... | -1 | -1 | -1 | -1 | 0 | 0 | -1 | -1 |
| ... | -1 | -1 | -1 | -1 | 0 | 0 | -1 | -1 |
| ... | 1 | 1 | -1 | -1 | 1 | 1 | 1 | 1 |
| ... | 1 | 1 | -1 | -1 | 1 | 1 | 1 | 1 |
| ... | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 |
| ... | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 |
| ... | -1 | -1 | 1 | 1 | 0 | 0 | 1 | 1 |
| ... | -1 | -1 | 1 | 1 | 0 | 0 | 1 | 1 |
| ... | 1 | 1 | 1 | 1 | -1 | -1 | -1 | -1 |
| ... | 1 | 1 | 1 | 1 | -1 | -1 | -1 | -1 |

## 18.8 Using Dummy Variables To Perform ANOVA

Once the dummy variables have been coded and entered into the regression program, it is a snap to perform ANOVA. First, fit a regression model to predict the $Y$ variable from all of the $X$ dummy variables. The regression table obtained from this fit gives an $SS_{model}$ that is equal to the total $SS$ for *all* of the sources, because the model includes all the sources. This $SS$ is not used directly, but is needed for subtractions. The $SS_{error}$ from this regression table is directly useful, however. This $SS_{error}$ is the correct value of the $SS$ for S(Between factors), just as we would calculate it using the usual ANOVA methods. It is correct because the full set of dummy variables represents the entire ANOVA model, and anything left over must be error. Likewise, the $df_{error}$ is exactly right for the ANOVA.

Next, fit each of the submodels that we can obtain by dropping the dummy variables for one source from the full model. Each time you fit a submodel, drop *all* of the dummy variables corresponding to *one ANOVA source* (i.e., drop them all together). If a factor has more than two levels, it makes no sense to drop the $k-1$ dummy variables for that factor one at a time. All the dummies together code the factor, so if we want to measure the effect of the factor we have to treat these dummy variables as a unit.

Note that the previous paragraph refers to sources, not factors. This implies that we should drop the dummy variables for a factor (say $A$), but *leave in* the dummy variables for its interactions (e.g., $AB$). Some students object that it is conceptually odd to have a submodel with (say) an $AB$ interaction but no effect of Factor A (e.g., the first submodel fit in Examples 1-3). This objection makes intuitive sense, since if we don't acknowledge the main effect of a factor in the model, how can we acknowledge interactions involving that factor? Here is the answer: Interactions are logically separate sources of variance from main effects. When we drop the dummies for a factor, we are

evaluating the main effect of the factor, not pretending that the factor didn't exist.

After fitting each submodel, we compare the results against what we obtained when we fit the full model. The extra sum of squares principle (ESSP) is used to isolate the $SS$ and *df* due to the source that was dropped out when we fit this submodel. (Note: The submodels do *not* get smaller and smaller. That is, we drop each set of dummy variables from the full model, but then we *put them back in* when we go on to the next submodel.)

After we get the $SS$ and *df* for each source with these ESSP comparisons, these values can be used to construct the usual ANOVA table. For each source line, the $SS$ and *df* are the extra $SS$ and extra *df* identified with the comparison used to isolate that source, and the $MS$ is $SS/df$ as usual. The error source line in this ANOVA table has the $SS$, *df* and $MS$ of the error term in the full model's regression ANOVA table. Finally, $F$ are computed as the ratio of the source $MS$ to the $MS_{error}$, just as if we had gotten the $SS$'s and *df*'s in the usual manner rather than using dummy variables.

### 18.8.1 Computations for Example 1

All that is required is to fit the model predicting $Y_i$ from the three dummy variables $XA_{1i}$, $XA_{2i}$, and $XA_{3i}$. The model is:

$$Y_i = 76.625 + 15.875 \times XA_{1i} + 5.375 \times XA_{2i} + 2.375 \times XA_{3i}$$

and the associated ANOVA table is:

| Source | *df* | $SS$ | $MS$ | $F$ |
|---|---|---|---|---|
| Model | 3 | 1689.375 | 563.125 | 0.564 |
| Error | 4 | 3992.500 | 998.125 | |
| Total | 7 | 5681.880 | | |

The $F$ for the full model is the same as the $F$ for Class, since Class is the only source in the model (remember that the mean is not included in the model in most regression programs, and the "total" is actually the "corrected total"). Since the $F$ is not significant, we don't have enough evidence to say that the classes are different.

### 18.8.2 Computations for Example 2

**Full model:** Start by fitting the full model to predict $Y_i$ from the five $X$'s, obtaining the model:

$$Y_i = 21.5 - 2 \times XA_{1i} - 6.25 \times XB_{1i} + 0.75 \times XB_{2i} - 0.25 \times XAB_{11i} - 2.75 \times XAB_{12i}$$

with the associated ANOVA table:

| Source | *df* | $SS$ | $MS$ | $F$ |
|---|---|---|---|---|
| Model | 5 | 394 | 78.8 | 3.7824 |
| Error | 6 | 125 | 20.8 | |
| Total | 11 | 519 | | |

**Factor A:** To evaluate the effect of Factor A, we must use the ESSP. Fit the model without the dummy variables that code $A$, using just the four predictors $XB_1$, $XB_2$, $XAB_{11}$, and $XAB_{12}$. The model is:

$$Y_i = 21.5 - 6.25 \times XB_{1i} + 0.75 \times XB_{2i} - 0.25 \times XAB_{11i} - 2.75 \times XAB_{12i}$$

and the associated ANOVA table is:

| Source | *df* | $SS$ | $MS$ | $F$ |
|---|---|---|---|---|
| Model | 4 | 346 | 86.5 | 3.5 |
| Error | 7 | 173 | 24.7 | |

Then use the ESSP to determine the effect of Factor A:

$$\begin{aligned} ESS_A &= 394 - 346 = 48 \\ Edf_A &= 5 - 4 = 1 \\ F_A &= \frac{48/1}{20.8} = 2.3 \end{aligned}$$

This $F_A$ is less than $F_{critical}$(1,6), so it is not significant.

**Factor B:** The same procedure is used to evaluate the effect of Factor B. First, we fit the model omitting the dummy variables coding B, and just using the three predictors $XA_{1i}$, $XAB_{11i}$, and $XAB_{12i}$. This yields:

$$Y_i = 21.5 - 2 \times XA_1 - 0.25 \times XAB_{11} - 2.75 \times XAB_{12}$$

with the associated ANOVA table:

| Source | *df* | *SS* | *MS* | *F* |
|---|---|---|---|---|
| Model | 3 | 114.5 | 38.1667 | 0.754841 |
| Error | 8 | 404.5 | 50.5625 | |

The effect of Factor B is evaluated by comparing this reduced model against the full model:

$$\begin{aligned} ESS_B &= 394 - 114.5 = 279.5 \\ Edf_B &= 5 - 3 = 2 \\ F_B &= \frac{279.5/2}{20.833} = 6.708 \end{aligned}$$

This $F_B$ is greater than $F_{critical}(2,6)$, so it is significant.

**AB Interaction:** To examine the $AB$ interaction, fit the model without the dummy variable coding this interaction:

$$Y_i = 21.5 - 2 \times XA_1 - 6.25 \times XB_1 + 0.75 \times XB_2$$

| Source | *df* | *SS* | *MS* | *F* |
|---|---|---|---|---|
| Model | 3 | 327.5 | 109.1667 | 4.56049 |
| Error | 8 | 191.5 | 23.9375 | |

Then compare this reduced model against the full model with the ESSP:

$$\begin{aligned} ESS_{AB} &= 394 - 327.5 = 66.5 \\ Edf_{AB} &= 5 - 3 = 2 \\ F_{AB} &= \frac{66.5/2}{20.833} = 1.596 \end{aligned}$$

This $F_{AB}$ is less than $F_{critical}(2,6)$, and is not significant.

### 18.8.3 Computations for Example 3

**Full model:**

$$\begin{aligned} Y_i = \; & 39.28 - 4.03 \times XA_1 - 0.36 \times XA_2 - 2 \times XB_1 - 0.25 \times XAB_{11} \\ & +0.58 \times XAB_{21} - 17.78 \times XC_1 - 1.11 \times XC_2 - 2.22 \times XAC_{11} \\ & +1.11 \times XAC_{12} + 1.11 \times XAC_{21} - 0.56 \times XAC_{22} + 0 \times XBC_{11} \\ & +0 \times XBC_{12} + 0 \times XABC_{111} + 0 \times XABC_{112} - 3.33 \times XABC_{211} + 1.67 \times XABC_{222} \end{aligned}$$

| Source | *df* | *SS* | *MS* | *F* |
|---|---|---|---|---|
| Model | 17 | 8844.22 | 520.248 | 24.97192 |
| Error | 18 | 375 | 20.83333 | |
| Total | 35 | 9219.22 | | |

**Factor A:** Fit a reduced model without the dummy variables $XA_1$ and $XA_2$,

| Source | *df* | *SS* | *MS* | *F* |
|---|---|---|---|---|
| Model | 15 | 8416.83 | 561.122 | 13.986 |
| Error | 20 | 802.389 | 40.1194 | |

and then use the ESSP:

$$\begin{aligned} ESS_A &= 8844.22 - 8416.83 = 427.39 \\ Edf_A &= 17 - 15 = 2 \\ F_A &= \frac{427.39/2}{20.833} = 10.257 \end{aligned}$$

This $F_A$ is greater than $F_{critical}(2,18)$, so it is significant.

**Factor B:** Fit a reduced model without $XB_1$:

| Source | *df* | $SS$ | $MS$ | $F$ |
|---|---|---|---|---|
| Model | 16 | 8700.22 | 543.764 | 19.90658 |
| Error | 19 | 519 | 27.31579 | |

and then do the ESSP analysis:

$$\begin{aligned} ESS_B &= 8844.22 - 8700.22 = 144 \\ Edf_B &= 17 - 16 = 1 \\ F_B &= \frac{144/1}{20.833} = 6.912 \end{aligned}$$

This $F_B$ is greater than $F_{critical}$(1,18), so it is significant.

**Factor AB:** Fit a reduced model without $XAB_{11}$ and $XAB_{21}$:

| Source | *df* | $SS$ | $MS$ | $F$ |
|---|---|---|---|---|
| Model | 15 | 8838.06 | 589.204 | 30.9158 |
| Error | 20 | 381.167 | 19.05833 | |

and then do ESSP:

$$\begin{aligned} ESS_{AB} &= 8844.22 - 8838.06 = 6.17 \\ Edf_{AB} &= 17 - 15 = 2 \\ F_{AB} &= \frac{6.17/2}{20.833} = 0.148 \end{aligned}$$

This $F_{AB}$ is less than $F_{critical}$(2,18), and is not significant.

**Factor C:** Fit a reduced model without $XC_1$ and $XC_2$:

| Source | *df* | $SS$ | $MS$ | $F$ |
|---|---|---|---|---|
| Model | 15 | 755.333 | 50.3556 | 0.119 |
| Error | 20 | 8463.89 | 423.194 | |

and then do ESSP:

$$\begin{aligned} ESS_C &= 8844.22 - 755.36 = 8088.89 \\ Edf_C &= 17 - 15 = 2 \\ F_C &= \frac{8088.89/2}{20.833} = 194.13 \end{aligned}$$

This $F_C$ is greater than $F_{critical}$(2,18), and is significant.

**Factor AC:** Fit a reduced model without $XAC_{11}$, $XAC_{12}$, $XAC_{21}$, and $XAC_{22}$:

| Source | *df* | $SS$ | $MS$ | $F$ |
|---|---|---|---|---|
| Model | 13 | 8799.78 | 676.906 | 35.5039 |
| Error | 22 | 419.444 | 19.06566 | |

and then use the ESSP:

$$\begin{aligned} ESS_{AC} &= 8844.22 - 8799.78 = 44.44 \\ Edf_{AC} &= 17 - 13 = 4 \\ F_{AC} &= \frac{44.44/4}{20.833} = 0.533 \end{aligned}$$

This $F_{AC}$ is less than $F_{critical}$(4,18), and is not significant.

**Factor BC:** Fit a reduced model without $XBC_{11}$ and $XBC_{12}$:

| Source | *df* | *SS* | *MS* | *F* |
|---|---|---|---|---|
| Model | 15 | 8844.22 | 589.615 | 31.44612 |
| Error | 20 | 375 | 18.75 | |

and use ESSP:

$$\begin{aligned} ESS_{BC} &= 8844.22 - 8844.22 = 0 \\ Edf_{BC} &= 17 - 15 = 2 \\ F_{BC} &= \frac{0/2}{20.833} = 0 \end{aligned}$$

This $F_{BC}$ is less than $F_{critical}(2,18)$, and is not significant.

**Factor ABC:** Fit a reduced model without $XABC_{111}$, $XABC_{211}$, $XABC_{112}$, and $XABC_{212}$:

| Source | *df* | *SS* | *MS* | *F* |
|---|---|---|---|---|
| Model | 13 | 8710.89 | 670.068 | 28.99968 |
| Error | 22 | 508.333 | 23.10606 | |

and use ESSP:

$$\begin{aligned} ESS_{ABC} &= 8844.22 - 8710.89 = 133.33 \\ Edf_{ABC} &= 17 - 13 = 4 \\ F_{ABC} &= \frac{133.33/4}{20.833} = 1.6 \end{aligned}$$

This $F_{ABC}$ is less than $F_{critical}(4,18)$, and is not significant.

**Summary:** It is convenient to summarize all these ESS computations in a single ANOVA table:

| Source | *df* | *SS* | *MS* | *F* |
|---|---|---|---|---|
| Full Model | 17 | 8844.22 | 520.248 | 24.972 |
| A | 2 | 427.39 | 213.69 | 10.257 |
| B | 1 | 144.00 | 144.00 | 6.912 |
| AB | 2 | 6.17 | 3.08 | 0.148 |
| C | 2 | 8088.89 | 4044.44 | 194.13 |
| AC | 4 | 44.44 | 11.11 | 0.533 |
| BC | 2 | 0.00 | 0.00 | 0 |
| ABC | 4 | 133.33 | 33.33 | 1.600 |
| Error | 18 | 375.00 | 20.83 | |
| Total | 35 | 9219.22 | | |

## 18.9 The Relationship of ANOVA to DVR

### 18.9.1 Models and Parameters

In this section we will show how to recover the ANOVA parameter estimates from the regression analysis. This is useful in practice, but more importantly it helps to show why DVR is equivalent to ANOVA.

It is not difficult to recover the estimates of ANOVA parameters from the dummy variable regression analysis, because they are simply related to the estimated intercept $\hat{a}$ and slopes $\hat{b}_1$, $\hat{b}_2$, $\hat{b}_3$, ... obtained in this analysis. Looking at the relationship between ANOVA and regression estimates also helps to show *why* dummy variable regression is equivalent to ANOVA—a topic that has been completely ignored in the algorithmic approach taken thus far.

The key for both these goals is to discover that dummy variable regression provides *exactly the same decomposition matrix* as ANOVA, though the notation in the models appears somewhat different. Let's first look at the decomposition of $Y$ values produced by the dummy variable regression for scores in Example 1.

After the estimated intercept $\hat{a}$ and estimated $\hat{b}$ values are obtained, the first score will be broken down as follows:

$$98 = \hat{a} + \hat{b}_1 \times 1 + \hat{b}_2 \times 0 + \hat{b}_3 \times 0 + \hat{e}_1$$

The above equation simplifies quite a bit when you multiply through by the numerical values of the dummy variables:

$$98 = \hat{a} + \hat{b}_1 + \hat{e}_1$$

Now, compare the decompositions provided by ANOVA and regression, for one score from each level of Factor A (the other score at each level has the same decomposition, except for the subscript on error).

| Score | ANOVA Decomposition | Regression Decomposition |
|---|---|---|
| 98 | $= \hat{\mu} + \hat{A}_1 + \widehat{S(A)}_{11}$ | $= \hat{a} + \hat{b}_1 + \hat{e}_1$ |
| 43 | $= \hat{\mu} + \hat{A}_2 + \widehat{S(A)}_{21}$ | $= \hat{a} + \hat{b}_2 + \hat{e}_3$ |
| 68 | $= \hat{\mu} + \hat{A}_3 + \widehat{S(A)}_{31}$ | $= \hat{a} + \hat{b}_3 + \hat{e}_5$ |
| 35 | $= \hat{\mu} + \hat{A}_4 + \widehat{S(A)}_{41}$ | $= \hat{a} - (\hat{b}_1 + \hat{b}_2 + \hat{b}_3) + \hat{e}_7$ |

You can see that the differences between the ANOVA and regression decompositions are strictly cosmetic:

- The regression parameter $a$ is equivalent to the ANOVA parameter $\mu$, since both are simply added into all the $Y$ scores. Not only are they conceptually equivalent, but they are numerically equivalent as well (see below).

- Likewise the regression parameters $b_1$, $b_2$, and $b_3$ are equivalent to the ANOVA parameters $A_1$, $A_2$, and $A_3$, because each is just added in to the scores at one level of A.

  At the fourth level of A there appears to be a major difference in the models, but it is still only a difference in notation. Remember the ANOVA constraint that the $A_i$'s must sum to zero. This implies that:

  $$\hat{A}_4 = -(\hat{A}_1 + \hat{A}_2 + \hat{A}_3)$$

  If we write the last ANOVA decomposition equation using the right side of this equality rather than the left, then the equivalence of the regression $b$'s and ANOVA $A$'s is completely clear:

  $$35 = \hat{\mu} - (\hat{A}_1 + \hat{A}_2 + \hat{A}_3) + \widehat{S(A)}_{41} = \hat{a} - (\hat{b}_1 + \hat{b}_2 + \hat{b}_3) + \hat{e}_7$$

- Though the notation for the error scores is quite different, these are also conceptually and numerically the same quantities in ANOVA and regression. Note that there is one estimated error value for each $Y$ score, regardless of how many subscripts we use to keep track of them. The difference in symbols ($S(A)$ versus $e$) is unimportant, since in both cases the error is the left-over part of the score not attributed to any of the other components of the decomposition. Since the other components are identical, as described above, the left-overs must also be identical. Thus, if the regression program prints out the estimated error scores from the dummy variable regression, you can read the estimated ANOVA errors right off the printout.

We have now shown that the dummy variable regression model is completely equivalent to the ANOVA model, with equivalent terms measuring identical quantities. Only the names have been changed, to confuse the students. Thus it should come as no surprise that the numerical estimates given by the regression program are the very same numbers that we would compute by hand with the estimation formulas for the ANOVA model. For example, the estimated value of $a$ provided by the regression program will always be the overall average of the $Y$'s, just as $\hat{\mu}$ is in ANOVA. Likewise, the estimated values of the $b$'s are exactly the same numbers that we would compute by hand to estimate the corresponding $A$'s. As discussed above, the $\hat{A}$ for the highest level, which has no corresponding $\hat{b}$, is uniquely determined by the constraint that the $\hat{A}$'s sum to 0. To find the estimate of this effect, we compute the sum of the effects of all the other levels and then take the negative of that sum.

To summarize: The estimated value of $a$ is simply the overall mean of the $Y$ values, so it is also the estimate of $\mu$ in the ANOVA model. The estimated value of any $b$ is simply the estimate of the ANOVA term that it codes. Estimates of the "missing" terms (factor effects at the highest level of the factor, and interactions involving that level) can be found by subtraction.

Here are the estimates of ANOVA parameters for Example 1:

$$\begin{aligned}\hat{\mu} &= 76.625\\ \hat{A}_1 &= 15.875\\ \hat{A}_2 &= 5.375\\ \hat{A}_3 &= 2.375\\ \hat{A}_4 &= -(15.875 + 5.375 + 2.375) = -23.625\end{aligned}$$

Interactions work just like main effects. The $\hat{b}$'s corresponding to interaction columns estimate the numerical values of the interaction term to which they correspond (e.g., $AB_{11}$). The missing interaction terms can again be computed from the constraints—interaction terms must sum to zero across any subscript with the other subscript(s) held constant. (For example, two-factor interactions must sum to zero across each row and down each column.) The following table shows the breakdown of cell means for Example 2, with each cell mean written in the form $\hat{\mu} + \hat{A}_i + \hat{B}_j + \widehat{AB}_{ij}$:

| | Freshman | Sophomore | Junior |
|---|---|---|---|
| Male | 21.5-2-6.25-0.25 | 21.5-2+0.75-2.75 | 21.5-2+6.25-0.75+0.25+2.75 |
| Female | 21.5+2-6.25+0.25 | 21.5+2+0.75+2.75 | 21.5+2+6.25-0.75-0.25-2.75 |

If you are getting bored, see if you can figure out the cell means from the regression parameters for Example 3.

### 18.9.2 DVR Computations by Computer

Many students get the misconception that the regression program somehow "looks at" the dummy variable values to infer the factorial design, and then somehow does something special to accomplish ANOVA. This is not at all what happens. As just illustrated, the $X$ variables are treated as *numerical values*, like any other $X$ variables in a standard regression situation. The dummy variable method works because we choose the right numerical values for the $X$ variables: these values reflect the categorical factors of ANOVA, but they do so in a numerical way as required by the regression program.

### 18.9.3 ANOVA With Unequal Cell Sizes (Weighted Means Solution)

Now that you have learned to do ANOVA with a regression program, you can easily learn one of two methods for doing ANOVA when there are unequal numbers of subjects in the different cells of the experiment. This method, called the *weighted means solution*, is appropriate in relatively few cases, whereas the alternative and computationally more complicated "unweighted means solution" is appropriate in most cases (see G. Keppel, *Design and Analysis: A Researcher's Handbook*, Chapter 17).

To compute the weighted means solution, you set up dummy variables following the same rules illustrated above with equal cell sizes. The fact that there are different numbers of subjects per cell can be ignored. For example, we could repeat the analysis of any of the three examples given above, dropping a few subjects. This would just mean deleting a few lines from the case by variable format data table, but it would not change the lines that remained. (Note, though, that we must still have at least one subject per cell.) We could use the regression program to get $SS$'s and *df*'s as before.

### 18.9.4 Context effects in ANOVA

When dealing with multiple regression models, a major issue was that of context effects. We saw that the $SS$ for a given term in a model could decrease when another term was present (due to redundancy), or could increase (due to a suppressor effect). Don't we need to worry about context effects when doing ANOVA with dummy variable regression?

In the case of redundancy effects, the answer is "no", at least with equal cell size designs. The reason is that in these designs, the dummy variables for different sources are *always uncorrelated*. (Look back at the examples, and note the zero correlation of any two dummy variables coding different sources.) This means that the dummy variables are not at all redundant, so we don't have to worry about redundancy effects. In unequal cell size designs, however, the dummy variables for different

sources may be correlated, depending on the exact pattern of cell sizes. In this case, redundancy is a problem, and the concepts learned in multiple regression apply with dummy variables as well.

Suppressor effects are also not a problem when doing ANOVA with dummy variable regression, as long as we always compare each submodel against the *full* model. The full model takes into account all of the sources of variance in the experiment, so it explains all of the error that can be explained in terms of these source (just like the regular ANOVA).[1] In principle, suppressor effects could be a problem if we ignored some of the dummy variables when testing others of the dummy variables, because in that case any variance explained by the ignored dummy variables would go into the error term—making it harder to get a significant result for the dummy variables being tested. But, since we always compare against the full model, this "in principle" problem never actually arises, presuming that we do the dummy variable analysis correctly.

## 18.10 Interactions Of ANOVA Factors and Covariates

After the long excursion into coding ANOVA factors with dummy variables, we return to the topic of models with both ANOVA factors and numerical regression variables. (Since you have surely forgotten the initial discussion of these models, it would be a good idea to take a minute and review the initial description of DVR in the very first section of this chapter.)

We are still not quite ready to work with these models, because we have not yet discussed the interactions of ANOVA factors and covariates. What do these interactions measure, and how do we represent them within the models? Fortunately, both questions are easy to answer.

The interaction of any ANOVA source and a covariate measures the extent to which the *slope* relating the DV to the covariate *depends on* the ANOVA source. This interaction is represented by (of course) another set of variables in the case-by-variable format.

To represent the interaction of an ANOVA source and a covariate, form one new interaction variable corresponding to each of the dummy variables used to code that ANOVA source. Label each of the new variables with the concatenation of the name of the dummy variable for the source and the name of the covariate, separated by a multiplication sign ($\times$). The values for each new interaction variable are found by multiplying, case by case, the values on the covariate times the value on the dummy variable, exactly as the label suggests.

Look at Example 4 below (you can skip the story, for now). The first table shows the design: a single factor with 2 levels (Teaching Method), 6 subjects in each group, and one covariate (IQ). The second table shows the data and dummy variables laid out in the case-by-variable format, with 12 cases by 4 variables—the value of the DV ($Y$) is in the first column, then the dummy variable to code the factor, then the covariate. The last variable represents the new interaction being discussed here – the interaction of IQ and teaching method. There is only one interaction column because only one dummy variable was needed to code the factor (i.e., it had two levels). Note that each value of this interaction variable is the product of the values on the IQ variable and the dummy variable for that case.

Now let's see how to interpret the estimates of the $b$'s for the covariate (IQ) and the interaction of the covariate with Factor A. Fitting the full model yields:

$$Y_i = -15.78 - 22.61 \times XA_{1i} + 0.36 \times IQ_i + 0.2256 \times (XA_{1i} \times IQ_i)$$

The estimated *average* slope relating $Y$ to IQ is 0.36, indicating that, on the average across all subjects, the $Y$ score increased .36 points for each 1 point increase in IQ. This is the same value that you would get if you ran two separate regressions—one for each group—and took the average of the two slope values.

The estimated *effect on slope* of being at level 1 of Factor A is to increase the slope 0.2256 units. This indicates that:

$$\text{Slope for Group 1} = 0.36 + 0.2256 = 0.5856$$

Similarly, the estimated effect on slope of being at level 2 of A is to decrease the slope 0.2256 units:

$$\text{Slope for Group 2} = 0.36 - 0.2256 = 0.1344$$

In general, the slope for each group is found by taking the overall mean slope and adding in any factor effects on slope for that group (just as we found the average score for each group by taking the

[1] This is another reason why we leave interactions in when we drop out main effects. Interactions must be left in the model so that they can potentially suppress error in the comparisons looking for the main effects.

overall average score and adding in any factor effects on mean for that group). The factor effect for the highest level of a factor is the negative of the sum of the other level effects (for both mean and slope).

AN ASIDE: The estimates of both $a$ and the factor effect change when a covariate is added to the model (i.e., compare the estimated values above with the same values in a dummy variable regression without the covariate). This should not be surprising, because parameter values also change when more predictor variables are added to a multiple regression model. Basically, the estimated $a$ changes because the average value of the covariate is not zero. Whenever there is a covariate with a nonzero mean, this covariate adds its mean times its slope into the $Y$ scores on the average across the whole data set. This changes the amount we want to add in on the average with parameter $a$. The estimated factor effect changes to the extent that the dummy variable is redundant with the covariate—that is, the extent to which the groups differ on the covariate.

## 18.11 Principles of Data Analysis Using DVR

The general principles for using DVR are easy to state: Fit the full model, and then drop terms one at a time to see what is significant. The implications of this procedure, and what can be learned using it, are best illustrated by examples.

## 18.12 Example 4: One Factor and One Covariate

Suppose we are interested in studying foreign language learning. We want to compare two teaching methods to see which is better, and also look for any relationship between language learning and IQ. We randomly select 12 students, and give IQ tests to all 12 (IQ is the covariate). Then, we randomly assign 6 students to each teaching method. After the students have been taught, we give them all the same standardized test, and the DV ($Y_i$) is the score on this test. Here are the data:

| Teaching Method 1 | | | Teaching Method 2 | | |
|---|---|---|---|---|---|
| Subject | $Y_i$ | $IQ_i$ | Subject | $Y_i$ | $IQ_i$ |
| 1 | 20 | 101 | 1 | 18 | 82 |
| 2 | 22 | 104 | 2 | 19 | 91 |
| 3 | 12 | 84 | 3 | 21 | 109 |
| 4 | 29 | 114 | 4 | 19 | 90 |
| 5 | 15 | 94 | 5 | 23 | 120 |
| 6 | 23 | 103 | 6 | 24 | 126 |

The GLM includes the ANOVA factor (Teaching Method), the covariate (IQ), and their interaction:

$$Y_{ij} = \mu + A_i + b_1 \times X_{ij} + b_2 \times A_i \times X_{ij} + S(A)_{ij}$$

In regression-oriented notation, this model is:

$$Y_i = a + b_1 \times XA_{1i} + b_2 \times X_i + b_3 \times XA_{1i} \times X_i + e_i$$

where $XA_{1i}$ is the dummy variable used to code Factor A and $X_i$ is the covariate.

To fit this model with a regression program, start by putting the data into case-by-variable format with the appropriate dummy variable coding of the experimental factor. Form one column for the interaction of Teaching Method and IQ (since Teaching Method has only 1 dummy variable). The input matrix for the regression program looks like this:

| $Y_i$ | $XA_{1i}$ | $IQ_i$ | $XA_{1i} \times IQ_i$ |
|---|---|---|---|
| 20 | 1 | 101 | 101 |
| 22 | 1 | 104 | 104 |
| 12 | 1 | 84 | 84 |
| 29 | 1 | 114 | 114 |
| 15 | 1 | 94 | 94 |
| 23 | 1 | 103 | 103 |
| 18 | -1 | 82 | -82 |
| 19 | -1 | 91 | -91 |
| 21 | -1 | 109 | -109 |
| 19 | -1 | 90 | -90 |
| 23 | -1 | 120 | -120 |
| 24 | -1 | 126 | -126 |

As usual, the analysis begins by fitting the full model with the regression program. In this case we obtain:

$$Y_i = -15.78 - 22.61 \times XA_{1i} + 0.36 \times IQ_i + 0.2256 \times XA_{1i} \times IQ_i$$

and the resulting ANOVA table is:

| Source | *df* | *SS* | *MS* | *F* |
|---|---|---|---|---|
| Model | 3 | 206.03 | 68.68 | 79.80 |
| Error | 8 | 6.885 | 0.8607 | |

The overall $F$ for the model is significant, so we can conclude that the model predicts $Y$ at better than a chance level. This indicates that test scores are related to teaching method, IQ, and/or their interaction.

To find out which terms in the model are individually significant we have to fit the various submodels and use the ESSP. It is generally a good idea to test the terms from right to left within the model (i.e., starting with the terms measuring interactions of covariates and dummy variables). If you find that the interaction of a covariate and a factor is not significant, it is best to drop that interaction out of the model for testing other terms (i.e., the nonsignificant terms will be left out of the context or comparison model). The reason is that the interaction term is generally redundant with the dummy variable, as you can see by computing their correlation. It is also standard procedure to drop covariates out of the model if neither they nor any of their interactions contribute significantly.

Thus, the first reduced model drops the term for the factor by covariate interaction:

$$Y_i = -4.313 + 0.115 \times XA_{1i} + 0.244 \times IQ_i$$

with ANOVA table:

| Source | *df* | *SS* | *MS* | *F* |
|---|---|---|---|---|
| Model | 2 | 126.7109 | 63.3554 | 6.61439 |
| Error | 9 | 86.2058 | 9.5784 | |

Compared with the full model,

$$\begin{aligned} ESS_{IQ\times A} &= 206.03 - 126.71 = 79.32 \\ Edf_{IQ\times A} &= 3 - 2 = 1 \\ F_{IQ\times A} &= \frac{79.73}{0.86} = 92 \end{aligned}$$

This $F_{observed}$ is highly significant, so the interpretation is that the relation between amount learned and IQ has a different slope for the two teaching methods. As computed above, the slopes are .5856 and .1344 for Teaching Methods 1 and 2, respectively, so we can conclude that the dependence of amount learned on IQ is greater for Method 1 than 2. (Note the similarity to an interaction interpretation in regular ANOVA.)

Also as in ANOVA, the significant interaction is a warning that we will have to limit our conclusions about the overall effects of teaching method and IQ—the "main effects". This point is discussed further in connection with each of the effects tested below.

A practical consequence of the significance of this term is that it must be left in the model (i.e., in the context) when we test the significance of the other terms.

Next, we evaluate IQ, using a reduced model dropping the IQ term:

$$Y_i = 20.48 - 4.32 \times XA_{1i} + 0.04 \times XA_{1i} \times IQ_i$$

| Source | *df* | *SS* | *MS* | *F* |
|---|---|---|---|---|
| Model | 2 | 4.15481 | 2.077403 | 0.0895596 |
| Error | 9 | 208.7619 | 23.19576 | |

The ESS comparison gives

$$\begin{aligned} ESS_{IQ} &= 206.03 - 4.15 = 201.88 \\ Edf_{IQ} &= 3 - 2 = 1 \\ F_{IQ} &= \frac{201.88}{0.86} = 234.5 \end{aligned}$$

The interpretation of this highly significant $F_{observed}$ is that *the average slope*, across all teaching methods, *is nonzero*. Since the estimated value in the full model is positive we can conclude that $Y$ tends to increase with IQ, on the average across all groups. Because of the significant interaction of IQ with teaching method found previously, though, we must stress that this observed increase of $Y$ with IQ is only on the average, and may not be found in every group.

Next, evaluate the effect of the teaching method, with the reduced model:

$$Y_i = -4.76 + 0.248 \times IQ_i + 0.004 \times XA_{1i} \times IQ_i$$

| Source | *df* | *SS* | *MS* | *F* |
|---|---|---|---|---|
| Model | 2 | 128.5192 | 64.2596 | 6.85254 |
| Error | 9 | 84.3974 | 9.37749 | |

and the ESS comparison is:

$$\begin{aligned} ESS_A &= 206.03 - 128.5 = 77.5 \\ Edf_A &= 3 - 2 = 1 \\ F_A &= \frac{77.5}{0.86} = 90 \end{aligned}$$

The interpretation of this significant $F_{observed}$ is that the $Y$-intercepts of the functions relating $Y$ to $X$ are different beyond what can be explained by chance. Remember, $Y$-intercept refers to the value of $Y$ that would be expected at $X = 0$. Since the estimated $b_1$ in the full model is negative, we can conclude that *average* $Y$ scores are really higher at $X = 0$ with Teaching Method 2 than with Teaching Method 1.

If the interaction of IQ and teaching method—tested above—had *not* been significant, we could have reached the general conclusion that Teaching Method 2 was better than Teaching Method 1 across all IQs. Because the interaction was significant, however, we cannot draw any general conclusion about the effect of teaching method across all IQs. All we really know is that Teaching Method 2 is superior at the specific IQ of 0. Obviously, it is not much use to be able to conclude anything about what happens at $IQ = 0$, because that is not something that will ever happen. Fortunately, we can easily test for an effect of teaching method at any IQ we want by "changing the $Y$-intercept", as illustrated in the next section.

It is useful to summarize the ESS comparisons in this section with the following variation of an ANOVA table.

| Source | *df* | *SS* | *MS* | *F* |
|---|---|---|---|---|
| Model | 3 | 206.03 | 68.68 | 79.8 |
| Method | 1 | 77.50 | 77.50 | 90 |
| IQ | 1 | 201.88 | 201.88 | 234 |
| $IQ \times Method$ | 1 | 79.32 | 79.32 | 92 |
| Error | 8 | 6.88 | 0.86 | |

This table shows both the overall $SS$ for the model as a whole, and also the $SS$ associated with each term in the model. Note that the $SS$ for each term is the $SS$ for that term when added to the model *last*, because each was measured with the other terms in the context. Note also that the $SS$'s for the individual terms add up to more than the total $SS$ for the model. As usual, this reflects some redundancy between the various terms.

Before leaving this example, we should mention another analysis that would make a lot of sense with these data: a regular ANOVA with teaching method as the factor and IQ as the dependent variable. If you got a significant result, you would know that the groups were different in overall IQ

to begin with, and so you would know that you had been unlucky in your random assignment. This would make you very cautious about interpreting the rest of the study.

Actually, you would probably do this analysis to check on the random assignment before you actually ran the teaching part of the study. If you got a significant difference, you could do another random assignment, and in fact keep doing random assignments until you get a nonsignificant difference. Another strategy is to explicitly divide up the sample into two groups equating IQ in some fashion. That strategy is more efficient if certain assumptions can be met, but it is dangerous. See an advanced text on experimental design for more information on matching groups.

## 18.13 Changing the $Y$-Intercept

In the previous example, it would be interesting to ask whether the teaching methods were different for specific IQs other than 0. For example, we might ask whether there is any difference at the population average IQ, which is 100. Since the regression model tests the null hypothesis of no effect of teaching method on the $Y$-intercept, the only way to test this is to change the $Y$-intercept to the $X$ value we are interested in (100, in this case). This can be done simply by subtracting the value (100) from all of the $X$ scores, and rerunning the whole regression analysis using the new "adjusted" IQ scores instead of the original ones.

First, we make a new variable ${X_i}' = IQ_i - 100$. The $Y$-intercept of the function relating $Y$ to $X$' is the point where $X' = 0$. But $X' = 0$ is the same thing as $IQ = 100$. Thus, whatever conclusion we can draw about differences in the $Y$-intercepts using $X$' will apply to differences in $Y$ values at $IQ = 100$.

Next, we redo the whole model analysis. Fitting the full model with $X$' instead of IQ gives:

$$Y_i = 20.22 - 0.0485 \times XA_{1i} + 0.36 \times X'_i + 0.226 \times XA_{1i} \times X'_i$$

with associated ANOVA table:

| Source | *df* | *SS* | *MS* | *F* |
|---|---|---|---|---|
| Model | 3 | 206.0315 | 68.6772 | 79.7968 |
| Error | 8 | 6.8852 | 0.8607 | |
| Total | 11 | 212.9167 | | |

Be sure to note that this table didn't change when we changed from IQ to $X$'. Changing the measurement scale of IQ has no effect on its predictive power, as long as we do a *linear* change of scale.

To test for an effect of teaching method on the intercept of the function relating $Y$ to $X$', we fit the model dropping the dummy variable for Teaching Method:

$$Y_i = 20.21 + 0.36 \times X'_i + 0.225 \times XA_{1i} \times X'_i$$

yielding

| Source | *df* | *SS* | *MS* | *F* |
|---|---|---|---|---|
| Model | 2 | 206.0037 | 103.0018 | 134.0981 |
| Error | 9 | 6.91297 | 0.768108 | |

and the ESS comparison is:

$$\begin{aligned} ESS_A &= 206.0315 - 206.0037 = 0.0278 \\ Edf_A &= 3 - 2 = 1 \\ F_A &= \frac{0.0278}{0.86} = 0.03 \end{aligned}$$

Thus, there is no evidence of an effect of Teaching Method on amount learned for students with IQ's of 100.

## 18.14 Example: A Factorial Analysis of Slopes

The next example extends this type of analysis to a situation with two ANOVA factors. In particular, the question is how the two factors influence the slope relating $Y$ to $X$. The example is particularly

interesting, because it shows how the analyses we are studying might be used in a very "real-world" setting.

The basic scenario involves identifying employers guilty of violating civil rights laws. Suppose you are a lawyer for the Justice Department, working in the section responsible for enforcing civil rights legislation. You are assigned to investigate a company accused of unfair employment practices (i.e., discrimination). It is your job to decide whether or not the company should be prosecuted in court, based on all available evidence. In particular, your must find out whether the company treats the two sexes differently (Male vs. Female) and whether the company treats races differently (Whites vs. Non-whites). Thus, the two factors in your analysis are Gender (Factor A) and Race (Factor B).

One thing you might do is analyze the company's hiring and promotion records, to see whether Gender and/or Race influence salary ($Y$). You will need to take years of job experience into account too, because this is clearly going to influence salary, too. Different groups may have different job experience (on average) and that difference could be cited by the company as justification for any salary difference. (In particular it is often the case that women have less job experience than men on the average. The question is whether the difference in job experience is fully responsible for their lower pay, or whether there is an additional effect—presumably sex discrimination—at work). Thus, years of experience will be your covariate $X$. It would also be a good idea to select people originally hired to do the same or comparable jobs.

Suppose you randomly select 10 employees in each of the four combinations of the $2 \times 2$ design (Gender $\times$ Race), and record Salary and Years Experience for each employee, with these results:

Nonwhite Females

| Employee | Salary | Experience |
|---|---|---|
| 1 | 11.5 | 4 |
| 2 | 19.0 | 12 |
| 3 | 23.0 | 16 |
| 4 | 8.5 | 2 |
| 5 | 14.0 | 7 |
| 6 | 16.0 | 9 |
| 7 | 15.0 | 8 |
| 8 | 17.5 | 11 |
| 9 | 13.5 | 6 |
| 10 | 12.0 | 5 |

White Females

| Employee | Salary | Experience |
|---|---|---|
| 11 | 15.5 | 10 |
| 12 | 9.0 | 2 |
| 13 | 19.0 | 15 |
| 14 | 11.5 | 4 |
| 15 | 14.0 | 8 |
| 16 | 15.0 | 9 |
| 17 | 9.0 | 1 |
| 18 | 23.0 | 20 |
| 19 | 12.5 | 6 |
| 20 | 12.0 | 5 |

Nonwhite Males

| Employee | Salary | Experience |
|---|---|---|
| 21 | 40.5 | 20 |
| 22 | 23.0 | 9 |
| 23 | 31.0 | 14 |
| 24 | 43.5 | 22 |
| 25 | 27.0 | 11 |
| 26 | 13.0 | 2 |
| 27 | 37.0 | 18 |
| 28 | 41.5 | 21 |
| 29 | 34.5 | 16 |
| 30 | 35.0 | 17 |

White Males

| Employee | Salary | Experience |
|---|---|---|
| 31 | 48.5 | 20 |
| 32 | 16.0 | 2 |
| 33 | 53.0 | 23 |
| 34 | 34.0 | 12 |
| 35 | 46.0 | 19 |
| 36 | 37.0 | 14 |
| 37 | 26.0 | 8 |
| 38 | 50.0 | 21 |
| 39 | 57.0 | 25 |
| 40 | 23.0 | 6 |

The full model is designed to see how salary depends on Gender, Race, and Years of Experience ($YE$), allowing for possible interactions. It is:

$$\begin{aligned} Y_i &= a + b_1 \times XA_{1i} + b_2 \times XB_{1i} + b_3 \times XAB_{11i} \\ &\quad + b_4 \times YE_i + b_5 \times XA_{1i} \times YE_i + b_6 \times XB_{1i} \times YE_i + b_7 \times XAB_{11i} \times YE_i \end{aligned}$$

This looks quite intimidating, but conceptually it is not so bad. The first half of the model—up through $b_3 \times XAB_{11i}$—measures intercept effects (i.e., effects at $YE = 0$). In this example, these intercept effects have a natural interpretation in terms of differences in salary *at hiring*, because the point at which the employee has 0 experience is the point at which they are first hired. Obviously, if all the employees are doing the same job and if there are no discrimination effects, then there should be no significant effects at intercept.

The second half of the model measures slopes. In this example, these have a natural interpretation in terms of *raises*—how much pay increase an employee gets per year of experience. The $b_4 \times YE_i$ term

measures the average raise per year, averaged across all four groups of employees. If the company is doing well, we would certainly expect this term to be significant, with a positive value for $\hat{b}_4$. The last three terms measure the effects of Gender, Race, and the Gender by Race interaction on raises. In other words, if males and females are given differential raises, we would expect the term $b_5 \times XA_{1i} \times YE_i$ to be significant. Analogously, the next term indicates the raise differential for the races. The final term measures an interactive effect of Gender and Race on the rate of getting raises; we will discuss this in more detail later.

The full analysis uses data, dummy variables, covariate, etc. in the following form:

| Case | $Y_i$ | $XA_{1i}$ | $XB_{1i}$ | $XAB_{11i}$ | $YE_i$ | $XA_{1i} \times YE_i$ | $XB_{1i} \times YE_i$ | $XAB_{11i} \times YE_i$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 11.5 | 1 | 1 | 1 | 4 | 4 | 4 | 4 |
| 2 | 19.0 | 1 | 1 | 1 | 12 | 12 | 12 | 12 |
| 3 | 23.0 | 1 | 1 | 1 | 16 | 16 | 16 | 16 |
| 4 | 8.5 | 1 | 1 | 1 | 2 | 2 | 2 | 2 |
| 5 | 14.0 | 1 | 1 | 1 | 7 | 7 | 7 | 7 |
| 6 | 16.0 | 1 | 1 | 1 | 9 | 9 | 9 | 9 |
| 7 | 15.0 | 1 | 1 | 1 | 8 | 8 | 8 | 8 |
| 8 | 17.5 | 1 | 1 | 1 | 11 | 11 | 11 | 11 |
| 9 | 13.5 | 1 | 1 | 1 | 6 | 6 | 6 | 6 |
| 10 | 12.0 | 1 | 1 | 1 | 5 | 5 | 5 | 5 |
| 11 | 15.5 | 1 | -1 | -1 | 10 | 10 | -10 | -10 |
| 12 | 9.0 | 1 | -1 | -1 | 2 | 2 | -2 | -2 |
| 13 | 19.0 | 1 | -1 | -1 | 15 | 15 | -15 | -15 |
| 14 | 11.5 | 1 | -1 | -1 | 4 | 4 | -4 | -4 |
| 15 | 14.0 | 1 | -1 | -1 | 8 | 8 | -8 | -8 |
| 16 | 15.0 | 1 | -1 | -1 | 9 | 9 | -9 | -9 |
| 17 | 9.0 | 1 | -1 | -1 | 1 | 1 | -1 | -1 |
| 18 | 23.0 | 1 | -1 | -1 | 20 | 20 | -20 | -20 |
| 19 | 12.5 | 1 | -1 | -1 | 6 | 6 | -6 | -6 |
| 20 | 12.0 | 1 | -1 | -1 | 5 | 5 | -5 | -5 |
| 21 | 40.5 | -1 | 1 | -1 | 20 | -20 | 20 | -20 |
| 22 | 23.0 | -1 | 1 | -1 | 9 | -9 | 9 | -9 |
| 23 | 31.0 | -1 | 1 | -1 | 14 | -14 | 14 | -14 |
| 24 | 43.5 | -1 | 1 | -1 | 22 | -22 | 22 | -22 |
| 25 | 27.0 | -1 | 1 | -1 | 11 | -11 | 11 | -11 |
| 26 | 13.0 | -1 | 1 | -1 | 2 | -2 | 2 | -2 |
| 27 | 37.0 | -1 | 1 | -1 | 18 | -18 | 18 | -18 |
| 28 | 41.5 | -1 | 1 | -1 | 21 | -21 | 21 | -21 |
| 29 | 34.5 | -1 | 1 | -1 | 16 | -16 | 16 | -16 |
| 30 | 35.0 | -1 | 1 | -1 | 17 | -17 | 17 | -17 |
| 31 | 48.5 | -1 | -1 | 1 | 20 | -20 | -20 | 20 |
| 32 | 16.0 | -1 | -1 | 1 | 2 | -2 | -2 | 2 |
| 33 | 53.0 | -1 | -1 | 1 | 23 | -23 | -23 | 23 |
| 34 | 34.0 | -1 | -1 | 1 | 12 | -12 | -12 | 12 |
| 35 | 46.0 | -1 | -1 | 1 | 19 | -19 | -19 | 19 |
| 36 | 37.0 | -1 | -1 | 1 | 14 | -14 | -14 | 14 |
| 37 | 26.0 | -1 | -1 | 1 | 8 | -8 | -8 | 8 |
| 38 | 50.0 | -1 | -1 | 1 | 21 | -21 | -21 | 21 |
| 39 | 57.0 | -1 | -1 | 1 | 25 | -25 | -25 | 25 |
| 40 | 23.0 | -1 | -1 | 1 | 6 | -6 | -6 | 6 |

Note that this matrix has the following simple form, because each case within the same group has the same values on the dummy variables:

| Cases | $Y$ | $XA$ | $XB$ | $XAB$ | YE | $XA \times YE$ | $XB \times YE$ | $XAB \times YE$ |
|---|---|---|---|---|---|---|---|---|
| Nonwhite Female | $Y$ | 1 | 1 | 1 | YE | YE | YE | YE |
| White Female | $Y$ | 1 | -1 | -1 | YE | YE | -YE | -YE |
| Nonwhite Male | $Y$ | -1 | 1 | -1 | YE | -YE | YE | -YE |
| White Male | $Y$ | -1 | -1 | 1 | YE | -YE | -YE | YE |

As always, we begin the analysis by fitting the full model:

$$\begin{aligned} Y_i &= 9.31 - 1.70 \times XA_{1i} - 0.86 \times XB_{1i} + 0.33 \times XAB_{11i} + 1.26 \times YE \\ &\quad - 0.39 \times XA_{1i} \times YE_i - 0.0053 \times XB_{1i} \times YE + 0.13 \times XAB_{11i} \times YE_i \end{aligned}$$

with the resulting ANOVA table:

| Source | *df* | *SS* | *MS* | *F* |
|---|---|---|---|---|
| Model | 7 | 7633.85 | 1090.549 | 8880.65 |
| Error | 32 | 3.92962 | 0.1228007 | |
| Total | 39 | 7637.78 | | |

Table 18.1 summarizes the steps of the analysis. Proceeding as before, we try to drop terms from the right end of the model. Each term is evaluated using the ESSP; if it is significant, it is left in, if it is not significant, it is dropped out. Each line in the table represents one such attempt to drop a term. To save space, the reduced model and ANOVA table for each step is not shown, but only the key values ($ESS$, $Edf$, $MS_{error}$, and $F$).

| Term Dropped | $ESS_{Term}$ | $Edf_{Term}$ | $MS_{Error}$ | $F$ |
|---|---|---|---|---|
| $XAB_{11i} \times YE_i$ | 18.966 | 1 | 0.1228 | 154.4453 |
| $XB_{1i} \times YE_i$ | 0.04 | 1 | 0.1228 | 0.2558485 |
| $XA_{1i} \times YE_i$ | 177.08 | 1 | 0.1200 | 1475 |
| $YE_i$ | 1951.547 | 1 | 0.1200 | 16258.62 |
| $XAB_{11i}$ | 0.844072 | 1 | 0.1200 | 7.03209 |
| $XB_{1i}$ | 24.0142 | 1 | 0.1200 | 200.0654 |
| $XA_{1i}$ | 22.54357 | 1 | 0.1200 | 187.8138 |

Table 18.1: Analysis of Discrimination Data

Note that the $MS_{error}$ changed starting at $XA$. The reason is that the $XB \times YE_i$ term was not significant, so it was left out of the model in all further comparisons. The reason for leaving the term out is that it is somewhat redundant with the main effect of Factor B (note the high correlation of the column for this term with the column coding this main effect). Thus, the full model had only 6 predictors in all ESS comparisons after that point.

### 18.14.1 Interpretations of Intercept Effects

There are significant effects of Gender (A), Race (B), and their interaction on the *Y*-intercept of the function relating pay to experience. As noted above, this means that when we find a significant effect, we can conclude that there are differences in the salary at which workers are hired.

The significant intercept effect of Factor A indicates that males and females are not hired at the same pay, on average. Keeping in mind that females were level 1 and males level 2 of the Gender factor, we can use the parameter estimates in the full model to compute the average pay at hiring for females and males:

| | | | | | |
|---|---|---|---|---|---|
| Females: | 9.31 | - | 1.70 | = | 7.61 |
| Males: | 9.31 | + | 1.70 | = | 11.01 |

Thus we can conclude that males are hired at higher salaries than females.

By exactly the same process, we can conclude that whites are hired at higher salaries than nonwhites:

| | | | | | |
|---|---|---|---|---|---|
| Nonwhites: | 9.31 | - | 0.86 | = | 8.45 |
| Whites: | 9.31 | + | 0.86 | = | 10.17 |

The significant AB interaction indicates that the effect of Gender on hiring salary (intercept) is different for whites than nonwhites, and—equivalently—the effect of Race is different for males than females. Using the parameter estimates from the model, we recover the following estimates of average hiring pay for the four groups:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | Females | | | | |
| Nonwhites: | 9.31 | - | 1.70 | - | 0.86 | + | 0.33 | = | 7.08 |
| Whites: | 9.31 | - | 1.70 | + | 0.86 | - | 0.33 | = | 8.14 |
| | | | | | Males | | | | |
| Nonwhites: | 9.31 | + | 1.70 | - | 0.86 | - | 0.33 | = | 9.82 |
| Whites: | 9.31 | + | 1.70 | + | 0.86 | + | 0.33 | = | 12.20 |

The interaction can be summarized, then, by saying that the effect of Gender is larger for whites ($12.20 - 8.14 = 4.06$) than for nonwhites ($9.82 - 7.08 = 2.74$).

### 18.14.2 Interpretation of Slope Effects

The average slope term (YE) is significant, with a positive estimated $\hat{b}$. We conclude that salary tends to increase with experience, on the average across all four groups.

There is also a significant effect of Gender on slope (i.e., an interaction of Gender and Years Experience), indicating that the function relating salary to experience is different for males than females. The slopes are:

| | | | | | |
|---|---|---|---|---|---|
| Females: | 1.26 | - | 0.39 | = | 0.87 |
| Males: | 1.26 | + | 0.39 | = | 1.65 |

Thus, salary increases faster with experience for males than females.

The effect of Race on slope (i.e., interaction of Race and YE) is not significant, so we do not have enough evidence to conclude that salary increases at a different rate with experience for nonwhites and whites.

The slope also depends on the interaction of Gender and Race (i.e., there is a three-way interaction of Gender, Race, and YE). Alternatively, we can say that the effect of Gender on slope is different for the two races, or the effect of Race on slope is different for the two sexes. The slopes for the four groups are:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | Females | | | | |
| Nonwhites: | 1.26 | - | 0.39 | - | 0.01 | + | 0.13 | = | 0.99 |
| Whites: | 1.26 | - | 0.39 | + | 0.01 | - | 0.13 | = | 0.75 |
| | | | | | Males | | | | |
| Nonwhites: | 1.26 | + | 0.39 | - | 0.01 | - | 0.13 | = | 1.51 |
| Whites: | 1.26 | + | 0.39 | + | 0.01 | + | 0.13 | = | 11.79 |

The interaction is again discernable in the slopes. The difference between male and female slopes is much greater for whites ($1.79 - 0.75 = 1.04$) than for nonwhites ($1.51 - 0.99 = 0.52$). Looked at the other way around, we see that white males get salary increases faster than nonwhite males, but white females get raises slower than nonwhite females.

Summary: The data certainly suggest that the company is discriminating against both females and nonwhites. Inequalities based on gender and race are apparent in both the original hiring pay and the rate of salary increases. The company has a lot of Explaining To Do, in terms of other relevant variables, if it is to persuade us that it is not letting Gender and Race influence its salary decisions.

## 18.15 Multiple Covariates Example

The presence of multiple covariates in the data set does not much change either the analysis or the interpretation of the results. Here, we work through one example to illustrate.

Suppose we elaborate the example of Section 18.12 to study amount learned as a function of motivation as well as IQ and teaching method. Here are the motivation scores for the 12 subjects, as measured with a standardized test:

| Subject: | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Method 1: | 18 | 15 | 9 | 20 | 12 | 19 |
| Method 2: | 15 | 16 | 18 | 15 | 19 | 19 |

The analysis proceeds just as it did in the earlier case, even though there is an extra covariate. Starting with the full model, we find:

$$Y_i = -12.6 - 21.6 \times XA_1 + 0.34 \times IQ_i + 0.17 \times XA_{1i} \times IQ_i - 0.083 \times MOT_i + 0.263 \times XA_{1i} \times MOT_i$$

with an ANOVA table of:

| Source | *df* | *SS* | *MS* | *F* |
|---|---|---|---|---|
| Model | 5 | 206.6689 | 41.3338 | 39.6948 |
| Error | 6 | 6.24773 | 1.041288 | |
| Total | 11 | 212.9167 | | |

We now evaluate individual terms with ESSP, proceeding from right to left through the model and dropping any nonsignificant terms corresponding to interactions of factors and covariates. Note that there is a slightly arbitrary element to this analysis, depending on whether we put the IQ or MOT covariate at the far right of the model. If the two covariates are redundant, we might drop the interaction term of the first one we test, but not the second. Thus, we might get different models depending on whether we try to drop IQ before MOT or vice versa. This might lead to an incorrect conclusion that the slope changed for one but not the other. Before concluding that the covariates interacted differently with the factor, it would be best to reverse the order and make sure the same interaction was nonsignificant.

First, we fit the model without the interaction of Teaching Method and Motivation. Comparing this reduced model against the full model, we find:

$$\begin{aligned} ESS_{XA\times MOT} &= 206.6689 - 206.442 = 0.2269 \\ Edf_{XA\times MOT} &= 5 - 4 = 1 \\ F_{XA\times MOT} &= \frac{0.2269/1}{1.041288} = 0.657 \end{aligned}$$

Since this interaction term is not significant, we will leave it out of the model.

Next, we fit a reduced model without the interaction of Teaching Method and IQ:

$$\begin{aligned} ESS_{XA\ \times IQ} &= 206.442 - 183.443 = 22.999 \\ Edf_{XA\ \times IQ} &= 4 - 3 = 1 \\ F_{XA\ \times IQ} &= \frac{22.999/1}{0.924951} = 24.865 \end{aligned}$$

Since this interaction term is significant, leave it in the model.

Next, we fit a reduced model without the main effect of Motivation:

$$\begin{aligned} ESS_{MOT} &= 206.442 - 206.032 = 0.410 \\ Edf_{MOT} &= 4 - 3 = 1 \\ F_{MOT} &= \frac{0.410/1}{0.924951} = 0.44386 \end{aligned}$$

Since this covariate is not significant, drop it from the model.

Since neither Motivation nor its interaction with Teaching Method was significant, we have dropped this covariate out of the model entirely. We are now back to the example of Section 18.12, and there is no point in repeating the analyses that were done earlier (the same ones are appropriate). If motivation had a significant ESS, we would have left it in the model, and proceeded onward to test the other terms. Of course, the reason for including multiple covariates in the study is that they may contribute significantly to the DV, even though this one does not.

# Chapter 19

# Analysis of Covariance

## 19.1 Goals of ANCOVA

So far, we have been looking at dummy variable regression in situations where the researcher is interested in 1) examining the relationship between the DV ($Y$) and one or more covariates, 2) seeing how this relationship differs between different groups, and 3) seeing how the DV differs between groups at specific values of the covariates (e.g., $X = 0$). ANCOVA is a special case of dummy variable regression distinguished mainly by its goal: the researcher *only* wants to look at *overall* differences between groups. In particular, the researcher wants to draw general conclusions about the effects of the ANOVA factor(s), not conclusions specific to certain values of the covariate(s). In these situations, it may still be worthwhile to include covariate(s) in the model in order to *reduce the error term* against which effects of factors are measured.

## 19.2 How ANCOVA Reduces Error

The advantages of ANCOVA are made very clear by looking at an example in which error reduction has a big effect on the analysis. Suppose a French teacher wants to know which of three textbooks to use in his French class. He takes a class of 15 students, randomly divides it into three equal groups, and has each group use one of the three books. At the end of the course, he gives all students the same test of amount learned, in order to see whether the amount learned depends on the book used. The results are shown in Table 19.1, and the corresponding ANOVA is shown in Table 19.2.

| Book 1 | | Book 2 | | Book 3 | |
|---|---|---|---|---|---|
| Student | Test | Student | Test | Student | Test |
| 1 | 34 | 1 | 46 | 1 | 59 |
| 2 | 56 | 2 | 61 | 2 | 61 |
| 3 | 64 | 3 | 66 | 3 | 72 |
| 4 | 69 | 4 | 75 | 4 | 76 |
| 5 | 77 | 5 | 77 | 5 | 84 |
| Average | 60 | Average | 65 | Average | 70.4 |

Table 19.1: Sample Data for French Book Experiment

| Source | *df* | *SS* | *MS* | *F* | Error Term |
|---|---|---|---|---|---|
| $\mu$ | 1 | 63635.30 | 63635.30 | 357.30 | $S(A)$ |
| $A$ | 2 | 270.54 | 135.27 | 0.76 | $S(A)$ |
| $S(A)$ | 12 | 2137.20 | 178.10 | | |
| Total | 15 | 66043 | | | |

Table 19.2: ANOVA for French Book Experiment

On the average, scores were highest for Book 3, next with Book 2, and lowest with Book 1. But are

the differences too large to be due to chance? This data set is nothing more than a one-factor, between-subjects ANOVAdesign, as was covered in Part 1. The appropriate ANOVA, shown in Table 19.2, indicates that the differences between groups are small enough that they might easily be due to chance (i.e., the $F_A$ is not significant).

Now suppose that the teacher had exactly the same students and test scores, but had measured the additional variable of grade point average (GPA) for each student, as shown in Table 19.3. (GPA is the student's average grade across many classes, based on a scale of A=4, B=3, C=2, ....)

| Book 1 | | | Book 2 | | | Book 3 | | |
|---|---|---|---|---|---|---|---|---|
| Student | Test | GPA | Student | Test | GPA | Student | Test | GPA |
| 1 | 34 | 3.24 | 1 | 46 | 3.28 | 1 | 59 | 3.35 |
| 2 | 56 | 3.44 | 2 | 61 | 3.43 | 2 | 61 | 3.39 |
| 3 | 64 | 3.54 | 3 | 66 | 3.48 | 3 | 72 | 3.47 |
| 4 | 69 | 3.59 | 4 | 75 | 3.58 | 4 | 76 | 3.52 |
| 5 | 77 | 3.65 | 5 | 77 | 3.63 | 5 | 84 | 3.62 |

Table 19.3: Augmented Data Set for French Book Experiment

We can incorporate the new GPA variable into the analysis by including it as a covariate in a dummy variable regression model. After we set up the appropriate dummy variables ($XA_{1i}$ and $XA_{2i}$) to code the main effect of book, the dummy variable regression model[1] is:

$$Y_i = a + b_1 \times XA_{1i} + b_2 \times XA_{2i} + b_3 \times GPA_i + e_i$$

Starting with this model, we use the ESSP to evaluate the main effect of Book. The full model is:

$$Y_i = -273.2 - 6.235 \times XA_{1i} - 0.06854 \times XA_{2i} + 97.2 \times GPA_i$$

with the associated ANOVA table:

| Source | *df* | *SS* | *MS* | *F* |
|---|---|---|---|---|
| Full Model | 3 | 2385.5 | 795.2 | 393.323 |
| Error | 11 | 22.2 | 2.0 | |

To evaluate Book, drop the two dummy variables used to code this factor. The ANOVA table for the reduced model is:

| Source | *df* | *SS* | *MS* | *F* |
|---|---|---|---|---|
| Reduced Model | 1 | 1994.5 | 1994.5 | 62.754 |
| Error | 13 | 413.2 | 31.8 | |

and the ESS comparison gives:

$$\begin{aligned} ESS_{book} &= 2385.5 - 1994.5 = 391 \\ Edf_{book} &= 3 - 1 = 2 \\ F_{book} &= \frac{391/2}{2.0} = 96.7 \end{aligned}$$

The ESS (i.e., the difference in the $SS$ for the full vs. reduced model) is attributable to the main effect of book, because the dummy variables for this factor were dropped to get the reduced model. Because the $F_{observed}$ corresponding to the ESS is highly significant, we can conclude that group effects are too large to be due to chance. Thus, we conclude that the differences among books are real.

The comparison between the ANOVA and the ANCOVA is at first paradoxical. The former analysis says the effect of books may be due to chance, but the latter analysis says that it is real. How can the two analyses disagree when the same $Y$ values were used in both?

The answer is that the *covariate reduced error*. Note that the error term used to compute $F_{book}$ was 178.10 in the ANOVA, but only 2.0 in the ANCOVA. Of course the $F$ is larger when the error term is reduced! The observed differences among book averages are not too large to be due to chance

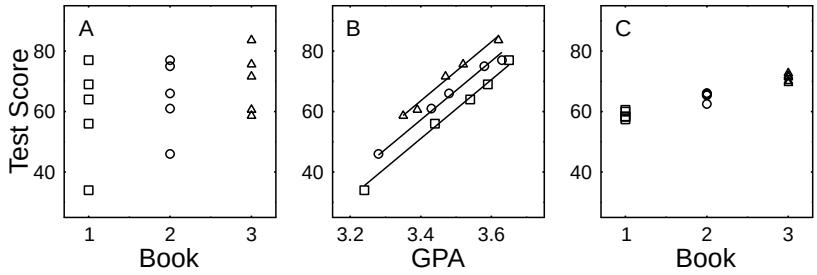


Figure 19.1: Data Representations in ANOVA vs. ANCOVA. The squares, circles, and triangles depict the results for students who used books 1, 2, and 3, respectively.

if error is about 178.1 units, but they certainly are too large to be due to chance if error is about 2.0 units.

Why is the error term so different in the ANOVA and ANCOVA? This is an important conceptual point illustrated by panels A and B of Figure 19.1. Panel A shows what the ANOVA "sees". There are three groups of test scores, represented by three different symbols. The individual data values within each group are shown, and it is clear that there is a lot of variation within each group. The averages for the groups appear somewhat different, but the difference in averages is small relative to the variation within groups. (Note that the variation causes the groups to overlap quite a bit in terms of test scores.) As far as the ANOVA can see, the variation within groups is simply due to error, because there is no other explanation for it. With this much error in the data and relatively small differences between group averages, ANOVA cannot be at all sure that the average differences are real.

Now, examine panel B, which shows what the ANCOVA "sees." When we add the extra variable of GPA into the picture, one thing we see immediately is that test score is very closely related to GPA. In other words, students with high GPA's learned more in the class, and that relationship is present for all three groups.

More importantly, though, we see that there is really very little "random error" in the data. Within each group, we see a nearly perfect linear relation between test score and GPA. This is why the error term in the ANCOVA is so much smaller than the error term in the ANOVA: Panel A includes a lot of *unexplained* variability in test scores, and this is all assigned to error. But in Panel B most of the variability in test scores is *explained by the covariate GPA*, and little is left over to be assigned to error.

Finally, looking across books in Panel B, we can see that the difference due to books is very consistent indeed. At any GPA you pick, the line relating test score to GPA is highest for Book 3, middle for Book 2, and lowest for Book 1. Thus, it really does make sense to conclude that the difference between books is real, just as the ANCOVA analysis indicated.

Another common way of understanding how ANCOVA reduces error involves the concept of *adjusted Y* scores. The adjusted $Y$ scores are *the values of Y that would be expected if all the cases had had exactly average scores on the covariate.*

Graphically, it is easy to see how the adjusted $Y$ score is computed for each case, and this process is illustrated in Figure 19.2. First, the best fitting line is determined for each group. Then, the point corresponding to each case is moved, *parallel to this line*, until it reaches the column where $X$ is at its overall average. The new value of $Y$ for this point is the adjusted $Y$ value.

This same adjustment can be made for every case, and Panel C of Figure 19.1 shows the results of this adjustment for the data of Panels A and B. Note that most of the within-group error that was present in Panel A has been removed from Panel C.

What ANCOVA does, conceptually, is to perform an ANOVA on the adjusted $Y$'s rather than on the original observed $Y$'s. Clearly, there will be a larger $F_{book}$ in the ANOVA on the adjusted scores than in the ANOVA on the original scores, precisely because within-group error has been drastically reduced. Conceptually, this ANOVA represents the results that would be expected if we had a very homogeneous sample of cases *in which there was no variation on X*. We know from the close relationship of $Y$ to $X$ that there would be much less within-group variation in $Y$ scores if there

[1]Note that we have omitted the terms reflecting slope differences among groups. The reason for this will become clear in Sections 19.3 and 19.4.

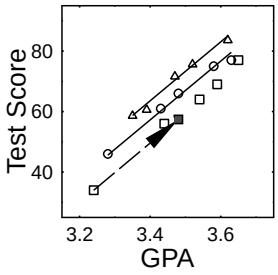


Figure 19.2: ANCOVA Adjustment of $Y$ Scores. The dashed line shows how the lower left point is adjusted to the mean GPA. The squares, circles, and triangles depict the results for students who used books 1, 2, and 3, respectively.

had been no variation in the $X$'s in the original sample, and the ANCOVA capitalizes on this.[2]

To summarize, we have seen that ANOVA attributes all variation not explained by the experimental factors to error. ANCOVA, on the other hand, can keep out of the error term any variation that is explainable in terms of a covariate. Thus, with a good covariate (one that is highly correlated with $Y$), one can potentially reduce error terms quite a lot by using ANCOVA.

## 19.3 ANCOVA Computational Procedure

From the example in Section 19.2, it would appear that ANCOVA is computationally simple. In that example, we fit a full model with the covariate (but no slope change terms) and a reduced model (without the factor dummies). The ESS was attributable to the factor, and the corresponding $F_{factor}$ was computed in the usual way.

In practice, though, the computational procedure is slightly more complicated than this, because there are two assumptions of ANCOVA that were ignored in Section 19.2. Thus, in practice, one must first perform computations to check these assumptions, and only then can one do the ESS comparison as shown in Section 19.2. In this section, we will describe the full computational procedure for ANCOVA, including the two computational checks. In the next section, we will explain why the assumptions and corresponding computational checks are necessary.

The computational procedure for ANCOVA has four steps:

1. Test for slope differences in the functions relating the DV to the covariate(s). That is, do a full dummy variable regression analysis of slopes as described in Section 18.14. If significant differences are found, you must stop—ANCOVA is not appropriate.

2. Test for effects of ANOVA factors on the covariate(s). This is done by performing a regular ANOVA *using the covariate as the DV*. If no significant effects or interactions are found, proceed to Step 3. However, if there are any significant results, determine whether these significant results are due to:

   (a) Nonrandom sampling. If this is the explanation, you can proceed to Step 3, but extra caution will be necessary in interpreting the results.

   (b) An effect of the factors on the covariate. If this is the explanation, you must stop—ANCOVA is not appropriate.

3. Do the ANCOVA ESS analysis. Do this by comparing the ANCOVA model including the ANOVA dummies and covariate(s) against reduced models formed by dropping the dummies coding any ANOVA factor or interaction. This is the part of the analysis that was illustrated in Section 19.2.

[2]Unfortunately, despite the appropriateness of this description at a conceptual level, carrying out the ANOVA on adjusted scores as described here will not necessarily give the same results as the true ANCOVA, and the differences can be fairly large (i.e., much more than just rounding error). A full explanation of the reasons for the discrepancy is beyond the scope of this book. Briefly, however, the problem is that the SS for the smaller ANCOVA model generally includes some of the book effect when the mean $X$ values differ across books. As a result, the ANCOVA tends to produce a smaller $ESS_{book}$ than the ANOVA on adjusted scores. Readers wanting further discussion of this point should consult Winer, Brown, and Michels (1991, pp. 785-6).

4. Compute adjusted means to interpret significant effects and interactions.

In the first step, we test to see whether the function relating the DV to the covariates has the same slope regardless of group. In the example of the French books, this means that we must find out whether the slope relating test score to GPA is the same for all French books.

This step is performed with a standard dummy variable regression analysis. The full model has all the dummy variables for ANOVA factors and their interactions, and all the covariates. It also has all the terms reflecting effects of ANOVA factors on slope, which we have so far been ignoring in ANCOVA. It is precisely these terms that we need to test for significance, so we form various reduced models dropping out these terms.[3]

In the current example, the computations are as follows: First, we fit the full model:

$$Y_i = -271 - 25.42 \times XA_{1i} + 20.46 \times XA_{2i} + 96.56 \times GPA_i + 5.5 \times XA_{1i} \times GPA_i - 5.89 \times XA_{2i} \times GPA_i + e_i$$

with associated ANOVA table:

| Source | *df* | *SS* | *MS* | *F* |
|---|---|---|---|---|
| Model | 5 | 2391.1 | 478.23 | 259.224 |
| Error | 9 | 16.6 | 1.84 | |
| Total | 14 | 2407.7 | | |

Then, we fit the reduced model (i.e., the ANCOVA model) dropping the terms for slope differences, and obtain:

$$Y_i = -273.21 - 6.24 \times XA_{1i} - 0.07 \times XA_{2i} + 97.21 \times GPA_i$$

with ANOVA table:

| Source | *df* | *SS* | *MS* | *F* |
|---|---|---|---|---|
| Model | 3 | 2385.5 | 795.17 | 393.321 |
| Error | 11 | 22.2 | 2.02 | |
| Total | 14 | 2407.7 | | |

The ESS comparison between these two models tests for group differences in the slope relating test score to GPA:

$$\begin{aligned} ESS_{\text{slope differences}} &= 2391.1 - 2385.5 = 5.6 \\ Edf_{\text{slope differences}} &= 5 - 3 = 2 \\ F_{\text{slope differences}} &= \frac{5.6/2}{1.8} = 1.527 \end{aligned}$$

Since 1.527 is not significant, there is not enough evidence to conclude that the slopes differ across groups.

At the end of Step 1, you look for any significant slope differences.[4] If any are found, you must abandon the plan to do ANCOVA, because it is just not meaningful in the presence of slope differences, for reasons explained in the next section. Of course you can still do all the sorts of model comparisons illustrated in the sections on dummy variable regression, but you have to keep in mind that your conclusions about experimental factors apply only to the expected $Y$ value when all $X$'s are 0.

In Step 2, it is necessary to see whether the ANOVA factors have any effect on the average value of the covariate. In the example comparing three French books, for instance, the first step is to see whether GPA is different between the groups assigned to the three books. The best way to do this is to perform an ANOVA with GPA as the dependent variable and book as factor A. Here is the ANOVA on GPA for the current example:

[3] As in the dummy variable regression analysis of slopes, we drop the slope effect terms in separate groups corresponding to the effect of Factor A on slope, the effect of Factor B on slope, the AB interaction effect on slope, and so on.

[4] Because we are testing an assumption which must be met for ANCOVA to be valid, the conservative thing to do is use a liberal significance level, typically $p = .10$. This makes it easier to reject the null hypothesis that the assumption is met. Thus, if we fail to reject the $H_0$, we can have more confidence that ANCOVA is appropriate. On the other hand, by using a liberal significance level, we more often conclude that ANCOVA is inappropriate when it actually is. This is a reasonable tradeoff to make because ANCOVA can be very misleading when its assumptions are not met (see Section 19.4). This footnote also applies to testing the second assumption (Step 2).

| Source | *df* | *SS* | *MS* | *F* | Error Term |
|---|---|---|---|---|---|
| $\mu$ | 1 | 181.73 | 181.73 | 9740.5 | S(Book) |
| Book | 2 | 0.0012208 | 0.0006104 | 0.0327154 | S(Book) |
| Error | 12 | 0.22388 | 0.0186564 | | |
| Total | 15 | 181.9507 | | | |

In more complicated examples, the ANOVA conducted at this step may have more than one factor. That does not change the nature of the step at all—we still check to see whether the covariate differs across groups. In such examples, though, it will require a multifactor ANOVA to do the check.

At the end of Step 2, we must examine the ANOVA for significant effects or, in the case of ANOVA's with more than one factor, significant interactions. If any are found, we must proceed very carefully. It may or may not be appropriate to carry out the ANCOVA analysis, depending on *what caused* the significant effects. If the significant effects are due to pre-existing differences between groups on the covariate (i.e., non-random sampling), then it is appropriate to do the ANCOVA analysis, although the results must be interpreted very carefully. If the significant effects are due to an influence of the experimental factor on the covariate, then it is not appropriate to do the ANCOVA analysis.[5] The reasons for this will become clear in Section 19.4, where the ANCOVA assumptions are discussed.

If neither Step 1 nor Step 2 requires us to stop, then the assumptions of ANCOVA have been met, and we can do the ANCOVA analysis. For this analysis (Step 3), we start with what we are calling the ANCOVA model—the model with dummy variables for the factor(s) and the average covariate effect, but no terms for slope differences. As we just saw, this is the reduced model in Step 2 (which was not significantly worse than the full model in that step).

In Step 3, we compare the ANCOVA model against a series of reduced models. Each reduced model is formed by dropping a set of dummy variables that code one factor or interaction. In the current example, there is only one reduced model to fit, because there is only one experimental factor:

$$Y_i = a + b_1 \times GPA_i + e_i$$

For this example, the computations for Step 3 are exactly the ESS comparison shown on page 236.

In situations with more than one experimental factor, the effect of each ANOVA factor and interaction is evaluated using the ESSP. The ESS is the difference between the $SS$ for the ANCOVA model and the $SS$ for the reduced model lacking dummy variables for that factor (or interaction). The *Edf* is the number of dummy variables dropped, and the $F_{observed}$ is computed using the error term for the ANCOVA model. The set of *F*'s generated in this way are the goal of the whole analysis. They test for significance of the experimental effects and interactions, using the reduced error term provided by the ANCOVA model.

## 19.4 ANCOVA Assumptions

Two assumptions are tested in Steps 1 and 2 of the computational procedure for ANCOVA. The first assumption is that the slope relating the DV to the covariate is the same for all levels of the independent variable. The second is that the average value of the covariate (e.g., GPA) is about the same at all levels of the independent variable (book). Why are these assumptions necessary?

The first assumption is necessary because we are trying to draw general conclusions about the effects of the ANOVA factors (remember, the covariate is included only to reduce error). If there are slope differences, however, then the effect of the ANOVA factor is different at different values of the covariate, as illustrated next.

Consider the data for the French book experiment, shown in Figure 19.1B. In that figure the lines for the three books have the same slope, and it is clear which book is best at every value of GPA. However, suppose that the data had looked like those in Figure 19.3. Clearly, the best-fitting straight lines relating test score to GPA would have quite different slopes for the three books. For Book 1 it increases, for Book 2 it is flat, and for Book 3 it decreases.

With the data of Figure 19.3, we cannot draw a conclusion about which book is best overall, because the effect of book depends on GPA. At the lowest GPAs, Book 3 is best; at the highest GPAs, Book 1 is best; in the middle, Book 2 is best.

[5]Some experts recommend that ANCOVA only be used when the covariable is measured before the experimental treatment is applied, to make sure that the factor cannot possibly influence the scores on the covariate. This prescription seems unnecessarily restrictive to us, though we agree that it is nonsensical to do ANCOVA when the factor affects covariate scores.

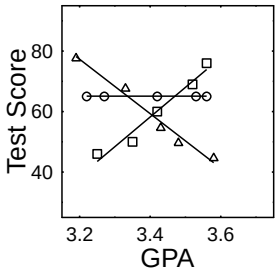


Figure 19.3: Unequal Slopes Relating Learning to GPA. The squares, circles, and triangles depict the results for students who used books 1, 2, and 3, respectively.

This example illustrates why, if we just want to draw simple conclusions about the effect of an ANOVA factor (e.g., Book), we cannot use ANCOVA when slopes are different. Whenever there are slope differences, the factor effect is not simple, but rather involves an interaction with the level of the covariate. We cannot draw any overall conclusion about the factor effect, but only conclusions specific to particular values of the covariate. This specificity of conclusion is inconsistent with the goal of ANCOVA, so the analysis cannot proceed when slopes are different.

The second assumption we tested was that there were no ANOVA effects or interactions on the average value of the covariate. If this assumption was violated, we had to determine what caused the effects. If nonrandom assignment caused the violation, then we can proceed with some caution. But if the violation was caused by an effect of the independent variable on the covariate, we must stop. Let us consider two examples to understand this step.

In the first example, we will consider an experiment evaluating French books in which we were not able to use random assignment. Suppose that there were three classes of five students available for the experiment, and that for practical reasons we wanted all the students within a class to use the same book. Furthermore, suppose that the classes had very different average GPAs (perhaps one is an honors class, one is a normal class, and one is a remedial class).

In this situation we might not be allowed to use random assignment to try to get equivalent groups for each book, because the school system might want to keep the three groups of students together. If so, whichever way we assigned books to classes, one book would get a group of brighter students, and another book would get a less bright group. We could never arrange it so that the books got equivalent groups, which is the goal of random assignment.

The data from this experiment might look something like that shown in Figure 19.4A or B. Note that in both figures the average test score is highest for book 1, second highest for book 2, and lowest for book 3. Also note that in both figures there are substantial differences between groups on GPA. Thus, Step 2 of the ANCOVA analysis would reveal significant differences between groups for either set of results. Of course, we would know that these differences were caused by nonrandom sampling, and we would have been expecting them right from the beginning of the experiment.

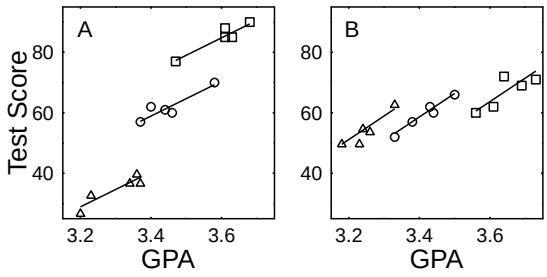


Figure 19.4: Examples with Group Differences on GPA. The squares, circles, and triangles depict the results for students who used books 1, 2, and 3, respectively.

Is it legitimate to conclude anything from the results? Consider first the data in Panel A. Book 1 not only had the highest average test score, but it also had the highest average GPA. Can we say

that the book was really the best, or could the higher test scores for this book have been due entirely to the GPA advantage?

According to ANCOVA, Book 1 really is the best given the results depicted in Panel A. If the test scores for all books are adjusted to the mean $X$ value, Book 1 still has the highest scores (see Figure 19.5A). This is easy to see, because the line for Book 1 is higher overall than the lines for Books 1 and 3. Thus, the higher scores for Book 1 are not just an artifact of the higher GPA for that group, but really reflect something good about the book.

Now consider the data in Figure 19.4B. Again Book 1 has the highest test scores and GPAs, but now ANCOVA tells us that Book 3 is the best. As before, the reason is that the line for Book 3 is the highest overall, and so are the adjusted $Y$ scores for this line (see Figure 19.5B). According to the ANCOVA, then, the only reason for the low test scores for this book is that it given to students with low GPAs. The ANCOVA indicates that these students would have had even lower test scores, given their GPAs, if the book had not been so good.

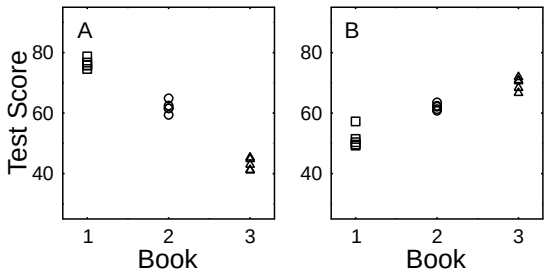


Figure 19.5: Learning Scores After Adjustment for Group Differences. The squares, circles, and triangles depict the results for students who used books 1, 2, and 3, respectively.

To summarize this example (both data sets), ANCOVA can potentially be used to adjust scores on the DV to remove the effects of nonrandom assignment. Thus, information about the covariate can potentially be used not only to remove error from the analysis, but also to correct for preexisting differences between groups.

Extra caution is needed when using ANCOVA to get rid of preexisting differences between groups. When you do this, you may have to depend heavily on the assumption that the relationship between $Y$ and $X$ holds, for each group, over a wider range of data than has actually been observed. Consider the results for Book 3 in Figure 19.4A, for example. There appears to be a linear relationship between GPA and test score as GPA varies from about 3.2 to 3.37 or 3.38. However, the ANCOVA represents a comparison of test scores adjusted to the average GPA: about 3.45. This adjustment could be pretty far off if the linear relationship changes (e.g., rises sharply) once we get above the top of the GPAs observed in that group. The point is that we did not observe any Book 3 cases with GPAs in the range we are adjusting to, so we have to assume that it is safe to *extrapolate* the best-fitting line. There is nothing in the statistics that can justify the extrapolation—the only justification can be common sense and prior knowledge of the area being studied.

Now, let us examine a second example violating the assumption tested in Step 2 of the computational procedure. In this example, the assumption will be violated because the ANOVA factor affects the covariate, not because of nonrandom sampling.

Again we will consider an experiment comparing three French books, and $Y$ will be the score on a standardized test. Now, however, $X$ is the score on a French vocabulary test rather than GPA. Both $X$ and $Y$ are measured at the end of the course.

As in the earlier example, it seems plausible that there would be a good correlation between $X$ and $Y$ people who do better on the test would probably know more vocabulary, and vice versa. Thus, we would expect a pattern of results like that shown in Figure 19.4A or B, with the book producing the highest test scores also producing highest vocabulary scores.

So far, we have discussed these data as if test score and vocabulary score are two separate dependent variables, as in fact they are. There is as yet nothing puzzling or peculiar about the results. But consider the following pseudo-statistical argument:

- It would be interesting to see which book produced the highest test scores *correcting for* differences in vocabulary. For example, one book might emphasize vocabulary while another book

might emphasize grammar instead. Naturally, scores on the standardized test will be partly influenced by vocabulary knowledge, but they will also be influenced by other factors like knowledge of grammar, knowledge of idioms, etc.

- To examine these issues, we can perform ANCOVA, using test score as $Y$ and vocabulary score as $X$. The ANCOVA will adjust the $Y$ scores to what they would have been if the $X$ scores had all been equal, and then we can compare the books with respect to the test scores they would have produced had they all been equally effective at teaching vocabulary.

- Applying this line of argument to the data in Figure 19.4A, we would conclude that Book 1 leads to the highest test scores, correcting for differences in vocabulary. But applying the argument to the data in Panel B, we conclude that Book 3 produces the highest test scores, correcting for differences in vocabulary.

What is wrong with this argument? The main problem is that the researcher is asking ANCOVA to simulate a situation which would never occur in the real world, and there is a logical problem in that. ANCOVA is asked to say what the results would be if all books produced the same average vocabulary score, *but they do not.* The vocabulary scores are not different by accident, but because of the very nature of the books. Because the books themselves *caused* the differences in vocabulary scores, it makes no sense to ask how the books would differ in test scores if there were no differences in vocabulary. If there were no differences in vocabulary, it couldn't be these books.

Perhaps an analogy would be useful. In the United States most people think that Democrats, who favor spending more money on social programs, give a higher priority to humanitarian concerns than Republicans, who favor spending less money on social programs. But we could ask: "Yes, but which party gives a higher priority to humanitarian concerns when you equate for how much money they want to spend on social programs?" This question just doesn't make much sense. Equating spending is removing some of the very difference between parties that is under study.

We should emphasize that the general goal of the researcher we have just been criticizing is not ridiculous. S/he wants to compare the different books separately with respect to how well they teach vocabulary, grammar, and other aspects of the language. For that purpose, though, the appropriate methodology is to devise a separate test for each aspect of the language of interest, and then measure each aspect as a separate dependent variable. ANCOVA cannot be used as a substitute for this methodology.

# Appendix A

# The AnoGen Program

The AnoGen program can be freely downloaded. For a current link, look on the first author's web page:

http://psy.otago.ac.nz/miller/index.htm#Software

## A.1 Introduction

This program was designed for use in teaching the statistical procedure known as *Analysis of Variance* (ANOVA). In brief, it generates appropriate data sets for use as examples or practice problems, and it computes the correct ANOVA for each data set. It handles between-subjects, within-subjects, and mixed designs, and can go up to six factors, with some restrictions on the numbers of levels, subjects, and model terms.

The program can be run in either of two modes: one designed for use by students; the other, by teachers.

The student mode is simpler: The student simply specifies the experimental design, and AnoGen generates an appropriate random data set. The student can then view the data set and answers, and save them to a file. Thus, students can fairly simply get as much computational practice as they want.

The teacher mode is more complicated: The teacher not only specifies the experimental design, but also controls the the cell means and error variance to obtain whatever $F$ values are desired for the example. Considerable familiarity with ANOVA is needed to use this mode.

## A.2 Step-by-Step Instructions: Student Mode

1. Log on to the Psychology network in the Psych computer lab, Goddard Laboratories, first floor.
2. You can start AnoGen with either of two methods, depending on whether you are running Windows (this is the default) or DOS:

   **Windows** Click on the AnoGen icon, which is located in the folder called "Course Related Applications."

   **DOS** Change to drive "L:" and then change into the "Jeff" directory. From there, type "AnoGen" to start the program.
3. Once AnoGen is running, type S to enter the student mode.
4. Specify the design:

   (a) To set the number of within-subjects factors, type W, and then type the number you want, followed by enter.

   (b) Similarly, type B to set the number of between-subjects factors,

   (c) Similarly, type S to set the number of subjects per group. A "group" is defined by one combination of levels of the between-subjects factors. For example, if you have between-subjects factors of Male/Female and Young/Old, then there are four groups corresonding to the four combinations.

| | |
|---|---|
| Group A1B1: | |
| Sub 1: | 95 |
| Sub 2: | 78 |
| Sub 3: | 97 |
| Group A2B1: | |
| Sub 1: | -19 |
| Sub 2: | -37 |
| Sub 3: | -10 |
| Group A1B2: | |
| Sub 1: | 55 |
| Sub 2: | 64 |
| Sub 3: | 73 |
| Group A2B2: | |
| Sub 1: | 58 |
| Sub 2: | 63 |
| Sub 3: | 71 |

Table A.1: An example of a problem display. This design has two between-subjects factors (A and B) with two levels each, and three subjects per group.

Note that you can set these numbers in any order, and you can change each one as often as you like. After you have the settings you want, type ctrl-Q to move on to the next step.

5. Now specify the number of levels of each factor. Type the letter corresponding to the factor you want to change (A, B, ...), and then enter the number of levels you want. Again, after you have the settings as you want them, type ctrl-Q to move on to the next step.

6. Type P to display the problem (i.e., the data set). Ideally, you would now do the computations by hand, for practice. (The information given in the problem display is intended to be self-explanatory, but some explanation is given in Section A.3.)

7. Type S to display the solution (i.e., cell means, ANOVA table, etc). This is where you check your solution and make sure you've done it correctly. The solution contains the various parts that I use in teaching ANOVA using the general linear model. (More explanation of the information given in the solution display is given in Section A.4.)

8. If you want, type F to save the problem and solution to a file. (The main reason to for doing this is to get a printed version of the problem and solution.) Enter the name of the file to which you want the information saved. *Otago Psych students: On the Psychology network, the command for printing an AnoGen output file is "APRT filename", which must be entered at a DOS prompt after you quit AnoGen.*

9. Type ctrl-Q to quit when you are done with this problem. AnoGen will then ask if you want to start over again: Type Y if you want to do another problem, or N to quit.

## A.3 Explanation of Problem Display

Table A.1 shows an example of a problem display. There is one line per subject, and the different groups correspond to the different levels of the between-subjects factor(s). For this example, the problem display fits on a single screen; with larger designs (i.e., more groups or more subjects per group), the problem display may be split across several screens.

Table A.2 shows an example of a problem display for a more complex experimental design. Note that the different conditions tested within-subjects are listed across the line, and the different subjects and groups organized as in the between-subjects design.

Group B1C1:

| | A1 | A2 |
|---|---|---|
| Sub 1: | 77 | 53 |
| Sub 2: | 84 | 56 |
| Sub 3: | 103 | 41 |

Group B2C1:

| | A1 | A2 |
|---|---|---|
| Sub 1: | 77 | 65 |
| Sub 2: | 54 | 64 |
| Sub 3: | 73 | 69 |

Group B1C2:

| | A1 | A2 |
|---|---|---|
| Sub 1: | 103 | 75 |
| Sub 2: | 100 | 78 |
| Sub 3: | 97 | 57 |

Group B2C2:

| | A1 | A2 |
|---|---|---|
| Sub 1: | 72 | 10 |
| Sub 2: | 74 | 18 |
| Sub 3: | 58 | 2 |

Table A.2: An example of a problem display. This design has a within-subjects factor (A) with two levels, two between-subjects factors (B and C) with two levels each, and three subjects per group.

## A.4 Explanation of Solution Display

The solution display has several components, as described below. Some of these components may be omitted if they do not fit well with the way your instructor teaches the material.

### A.4.1 Design

This shows a list of factors with the number of levels per factor. Also shown is the number of subjects per group.

### A.4.2 Cell Means

The cell means are given in a table of this form (these are the means for the problem in Table A.2):

| Cell: | Mean |
|---|---|
| u | 65 |
| A1 | 81 |
| A2 | 49 |
| B1 | 77 |
| B2 | 53 |
| A1 B1 | 94 |
| A1 B2 | 68 |
| A2 B1 | 60 |
| A2 B2 | 38 |
| C1 | 68 |
| C2 | 62 |
| A1 C1 | 78 |
| ... | |

The first line (u) shows the overall mean across all conditions. The next two lines (A1 and A2) show the means of all scores at levels 1 and 2 of factor A, respectively. The next two lines (B1 and B2) show the means of all scores at levels 1 and 2 of factor B, respectively. The next line (A1 B1) shows the mean of all scores at level 1 of factor A and level 1 of factor B, and then the next three lines show means for the other combinations of levels on these two factors. And so on.

### A.4.3 Model

The model section shows the form of the general linear model appropriate for this design. The main effect and interaction terms are denoted by capital letters (A, B, AB, etc), S is for subjects, and the subscripts are denoted by lower-case letters (i, j, k, etc).

### A.4.4 Estimation Equations

The estimation equations section shows the equation used to estimate each term in the linear model. The period subscript is used to denote averaging across levels of the factor corresponding to that subscript.

### A.4.5 Decomposition Matrix

The decomposition matrix shows the breakdown of all data values (numbers of the left sides of the equals signs) into the estimated values corresponding to each term in the linear model. The order of the numbers on the line is the same as the order of the terms in the model.

### A.4.6 ANOVA Table

The ANOVA table is in a relatively standard format. The $F$ value is marked with one asterisk if it is significant at the level of $p < .05$ and two asterisks if significant at $p < .01$. The error term used to compute each $F$ is shown at the far right side of the table.

# Appendix B

# Statistical Tables

## B.1 $F$-Critical Values

The following six pages give tables of $F$-critical values for $p = .10$, $p = .05$, $p = .025$, $p = .01$, $p = .005$, and $p = .001$, respectively. The most "standard" values are those with $p = .05$.

## F-Critical Values for $p = .10$

| DF Error (Denom) | Degrees of Freedom for Effect, Interaction, etc (Numerator) | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 12 | 15 | 20 | 24 | 30 | 40 | 60 | 120 | 250 |
| 1 | 39.863 | 49.500 | 53.593 | 55.833 | 57.240 | 58.204 | 58.906 | 59.439 | 59.858 | 60.195 | 60.705 | 61.220 | 61.740 | 62.002 | 62.265 | 62.529 | 62.794 | 63.061 | 63.203 |
| 2 | 8.526 | 9.000 | 9.162 | 9.243 | 9.293 | 9.326 | 9.349 | 9.367 | 9.381 | 9.392 | 9.408 | 9.425 | 9.441 | 9.450 | 9.458 | 9.466 | 9.475 | 9.483 | 9.487 |
| 3 | 5.538 | 5.462 | 5.391 | 5.343 | 5.309 | 5.285 | 5.266 | 5.252 | 5.240 | 5.230 | 5.216 | 5.200 | 5.184 | 5.176 | 5.168 | 5.160 | 5.151 | 5.143 | 5.138 |
| 4 | 4.545 | 4.325 | 4.191 | 4.107 | 4.051 | 4.010 | 3.979 | 3.955 | 3.936 | 3.920 | 3.896 | 3.870 | 3.844 | 3.831 | 3.817 | 3.804 | 3.790 | 3.775 | 3.768 |
| 5 | 4.060 | 3.780 | 3.619 | 3.520 | 3.453 | 3.405 | 3.368 | 3.339 | 3.316 | 3.297 | 3.268 | 3.238 | 3.207 | 3.191 | 3.174 | 3.157 | 3.140 | 3.123 | 3.114 |
| 6 | 3.776 | 3.463 | 3.289 | 3.181 | 3.108 | 3.055 | 3.014 | 2.983 | 2.958 | 2.937 | 2.905 | 2.871 | 2.836 | 2.818 | 2.800 | 2.781 | 2.762 | 2.742 | 2.732 |
| 7 | 3.589 | 3.257 | 3.074 | 2.961 | 2.883 | 2.827 | 2.785 | 2.752 | 2.725 | 2.703 | 2.668 | 2.632 | 2.595 | 2.575 | 2.555 | 2.535 | 2.514 | 2.493 | 2.481 |
| 8 | 3.458 | 3.113 | 2.924 | 2.806 | 2.726 | 2.668 | 2.624 | 2.589 | 2.561 | 2.538 | 2.502 | 2.464 | 2.425 | 2.404 | 2.383 | 2.361 | 2.339 | 2.316 | 2.304 |
| 9 | 3.360 | 3.006 | 2.813 | 2.693 | 2.611 | 2.551 | 2.505 | 2.469 | 2.440 | 2.416 | 2.379 | 2.340 | 2.298 | 2.277 | 2.255 | 2.232 | 2.208 | 2.184 | 2.171 |
| 10 | 3.285 | 2.924 | 2.728 | 2.605 | 2.522 | 2.461 | 2.414 | 2.377 | 2.347 | 2.323 | 2.284 | 2.244 | 2.201 | 2.178 | 2.155 | 2.132 | 2.107 | 2.082 | 2.068 |
| 11 | 3.225 | 2.860 | 2.660 | 2.536 | 2.451 | 2.389 | 2.342 | 2.304 | 2.274 | 2.248 | 2.209 | 2.167 | 2.123 | 2.100 | 2.076 | 2.052 | 2.026 | 2.000 | 1.985 |
| 12 | 3.177 | 2.807 | 2.606 | 2.480 | 2.394 | 2.331 | 2.283 | 2.245 | 2.214 | 2.188 | 2.147 | 2.105 | 2.060 | 2.036 | 2.011 | 1.986 | 1.960 | 1.932 | 1.918 |
| 13 | 3.136 | 2.763 | 2.560 | 2.434 | 2.347 | 2.283 | 2.234 | 2.195 | 2.164 | 2.138 | 2.097 | 2.053 | 2.007 | 1.983 | 1.958 | 1.931 | 1.904 | 1.876 | 1.861 |
| 14 | 3.102 | 2.726 | 2.522 | 2.395 | 2.307 | 2.243 | 2.193 | 2.154 | 2.122 | 2.095 | 2.054 | 2.010 | 1.962 | 1.938 | 1.912 | 1.885 | 1.857 | 1.828 | 1.812 |
| 15 | 3.073 | 2.695 | 2.490 | 2.361 | 2.273 | 2.208 | 2.158 | 2.119 | 2.086 | 2.059 | 2.017 | 1.972 | 1.924 | 1.899 | 1.873 | 1.845 | 1.817 | 1.787 | 1.770 |
| 16 | 3.048 | 2.668 | 2.462 | 2.333 | 2.244 | 2.178 | 2.128 | 2.088 | 2.055 | 2.028 | 1.985 | 1.940 | 1.891 | 1.866 | 1.839 | 1.811 | 1.782 | 1.751 | 1.734 |
| 17 | 3.026 | 2.645 | 2.437 | 2.308 | 2.218 | 2.152 | 2.102 | 2.061 | 2.028 | 2.001 | 1.958 | 1.912 | 1.862 | 1.836 | 1.809 | 1.781 | 1.751 | 1.719 | 1.702 |
| 18 | 3.007 | 2.624 | 2.416 | 2.286 | 2.196 | 2.130 | 2.079 | 2.038 | 2.005 | 1.977 | 1.933 | 1.887 | 1.837 | 1.810 | 1.783 | 1.754 | 1.723 | 1.691 | 1.673 |
| 19 | 2.990 | 2.606 | 2.397 | 2.266 | 2.176 | 2.109 | 2.058 | 2.017 | 1.984 | 1.956 | 1.912 | 1.865 | 1.814 | 1.787 | 1.759 | 1.730 | 1.699 | 1.666 | 1.648 |
| 20 | 2.975 | 2.589 | 2.380 | 2.249 | 2.158 | 2.091 | 2.040 | 1.999 | 1.965 | 1.937 | 1.892 | 1.845 | 1.794 | 1.767 | 1.738 | 1.708 | 1.677 | 1.643 | 1.625 |
| 21 | 2.961 | 2.575 | 2.365 | 2.233 | 2.142 | 2.075 | 2.023 | 1.982 | 1.948 | 1.920 | 1.875 | 1.827 | 1.776 | 1.748 | 1.719 | 1.689 | 1.657 | 1.623 | 1.604 |
| 22 | 2.949 | 2.561 | 2.351 | 2.219 | 2.128 | 2.060 | 2.008 | 1.967 | 1.933 | 1.904 | 1.859 | 1.811 | 1.759 | 1.731 | 1.702 | 1.671 | 1.639 | 1.604 | 1.585 |
| 23 | 2.937 | 2.549 | 2.339 | 2.207 | 2.115 | 2.047 | 1.995 | 1.953 | 1.919 | 1.890 | 1.845 | 1.796 | 1.744 | 1.716 | 1.686 | 1.655 | 1.622 | 1.587 | 1.568 |
| 24 | 2.927 | 2.538 | 2.327 | 2.195 | 2.103 | 2.035 | 1.983 | 1.941 | 1.906 | 1.877 | 1.832 | 1.783 | 1.730 | 1.702 | 1.672 | 1.641 | 1.607 | 1.571 | 1.552 |
| 25 | 2.918 | 2.528 | 2.317 | 2.184 | 2.092 | 2.024 | 1.971 | 1.929 | 1.895 | 1.866 | 1.820 | 1.771 | 1.718 | 1.689 | 1.659 | 1.627 | 1.593 | 1.557 | 1.537 |
| 26 | 2.909 | 2.519 | 2.307 | 2.174 | 2.082 | 2.014 | 1.961 | 1.919 | 1.884 | 1.855 | 1.809 | 1.760 | 1.706 | 1.677 | 1.647 | 1.615 | 1.581 | 1.544 | 1.523 |
| 27 | 2.901 | 2.511 | 2.299 | 2.165 | 2.073 | 2.005 | 1.952 | 1.909 | 1.874 | 1.845 | 1.799 | 1.749 | 1.695 | 1.666 | 1.636 | 1.603 | 1.569 | 1.531 | 1.511 |
| 28 | 2.894 | 2.503 | 2.291 | 2.157 | 2.064 | 1.996 | 1.943 | 1.900 | 1.865 | 1.836 | 1.790 | 1.740 | 1.685 | 1.656 | 1.625 | 1.592 | 1.558 | 1.520 | 1.499 |
| 29 | 2.887 | 2.495 | 2.283 | 2.149 | 2.057 | 1.988 | 1.935 | 1.892 | 1.857 | 1.827 | 1.781 | 1.731 | 1.676 | 1.647 | 1.616 | 1.583 | 1.547 | 1.509 | 1.488 |
| 30 | 2.881 | 2.489 | 2.276 | 2.142 | 2.049 | 1.980 | 1.927 | 1.884 | 1.849 | 1.819 | 1.773 | 1.722 | 1.667 | 1.638 | 1.606 | 1.573 | 1.538 | 1.499 | 1.477 |
| 40 | 2.835 | 2.440 | 2.226 | 2.091 | 1.997 | 1.927 | 1.873 | 1.829 | 1.793 | 1.763 | 1.715 | 1.662 | 1.605 | 1.574 | 1.541 | 1.506 | 1.467 | 1.425 | 1.401 |
| 60 | 2.791 | 2.393 | 2.177 | 2.041 | 1.946 | 1.875 | 1.819 | 1.775 | 1.738 | 1.707 | 1.657 | 1.603 | 1.543 | 1.511 | 1.476 | 1.437 | 1.395 | 1.348 | 1.320 |
| 120 | 2.748 | 2.347 | 2.130 | 1.992 | 1.896 | 1.824 | 1.767 | 1.722 | 1.684 | 1.652 | 1.601 | 1.545 | 1.482 | 1.447 | 1.409 | 1.368 | 1.320 | 1.265 | 1.230 |
| 250 | 2.726 | 2.324 | 2.106 | 1.967 | 1.870 | 1.798 | 1.741 | 1.695 | 1.657 | 1.624 | 1.572 | 1.515 | 1.450 | 1.414 | 1.375 | 1.330 | 1.279 | 1.217 | 1.176 |

Note: If this table does not include the value of DF Error that you want, you should either get a more complete table or else use the next *smaller* DF Error value in this table.

## F-Critical Values for $p = .05$

| DF Error (Denom) | Degrees of Freedom for Effect, Interaction, etc (Numerator) | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 12 | 15 | 20 | 24 | 30 | 40 | 60 | 120 | 250 |
| 1 | 161.45 | 199.50 | 215.71 | 224.58 | 230.16 | 233.99 | 236.77 | 238.88 | 240.54 | 241.88 | 243.91 | 245.95 | 248.01 | 249.05 | 250.10 | 251.14 | 252.20 | 253.25 | 253.80 |
| 2 | 18.51 | 19.00 | 19.16 | 19.25 | 19.30 | 19.33 | 19.35 | 19.37 | 19.38 | 19.40 | 19.41 | 19.43 | 19.45 | 19.45 | 19.46 | 19.47 | 19.48 | 19.49 | 19.49 |
| 3 | 10.13 | 9.55 | 9.28 | 9.12 | 9.01 | 8.94 | 8.89 | 8.85 | 8.81 | 8.79 | 8.74 | 8.70 | 8.66 | 8.64 | 8.62 | 8.59 | 8.57 | 8.55 | 8.54 |
| 4 | 7.71 | 6.94 | 6.59 | 6.39 | 6.26 | 6.16 | 6.09 | 6.04 | 6.00 | 5.96 | 5.91 | 5.86 | 5.80 | 5.77 | 5.75 | 5.72 | 5.69 | 5.66 | 5.64 |
| 5 | 6.61 | 5.79 | 5.41 | 5.19 | 5.05 | 4.95 | 4.88 | 4.82 | 4.77 | 4.74 | 4.68 | 4.62 | 4.56 | 4.53 | 4.50 | 4.46 | 4.43 | 4.40 | 4.38 |
| 6 | 5.99 | 5.14 | 4.76 | 4.53 | 4.39 | 4.28 | 4.21 | 4.15 | 4.10 | 4.06 | 4.00 | 3.94 | 3.87 | 3.84 | 3.81 | 3.77 | 3.74 | 3.70 | 3.69 |
| 7 | 5.59 | 4.74 | 4.35 | 4.12 | 3.97 | 3.87 | 3.79 | 3.73 | 3.68 | 3.64 | 3.57 | 3.51 | 3.44 | 3.41 | 3.38 | 3.34 | 3.30 | 3.27 | 3.25 |
| 8 | 5.32 | 4.46 | 4.07 | 3.84 | 3.69 | 3.58 | 3.50 | 3.44 | 3.39 | 3.35 | 3.28 | 3.22 | 3.15 | 3.12 | 3.08 | 3.04 | 3.01 | 2.97 | 2.95 |
| 9 | 5.12 | 4.26 | 3.86 | 3.63 | 3.48 | 3.37 | 3.29 | 3.23 | 3.18 | 3.14 | 3.07 | 3.01 | 2.94 | 2.90 | 2.86 | 2.83 | 2.79 | 2.75 | 2.73 |
| 10 | 4.96 | 4.10 | 3.71 | 3.48 | 3.33 | 3.22 | 3.14 | 3.07 | 3.02 | 2.98 | 2.91 | 2.85 | 2.77 | 2.74 | 2.70 | 2.66 | 2.62 | 2.58 | 2.56 |
| 11 | 4.84 | 3.98 | 3.59 | 3.36 | 3.20 | 3.09 | 3.01 | 2.95 | 2.90 | 2.85 | 2.79 | 2.72 | 2.65 | 2.61 | 2.57 | 2.53 | 2.49 | 2.45 | 2.43 |
| 12 | 4.75 | 3.89 | 3.49 | 3.26 | 3.11 | 3.00 | 2.91 | 2.85 | 2.80 | 2.75 | 2.69 | 2.62 | 2.54 | 2.51 | 2.47 | 2.43 | 2.38 | 2.34 | 2.32 |
| 13 | 4.67 | 3.81 | 3.41 | 3.18 | 3.03 | 2.92 | 2.83 | 2.77 | 2.71 | 2.67 | 2.60 | 2.53 | 2.46 | 2.42 | 2.38 | 2.34 | 2.30 | 2.25 | 2.23 |
| 14 | 4.60 | 3.74 | 3.34 | 3.11 | 2.96 | 2.85 | 2.76 | 2.70 | 2.65 | 2.60 | 2.53 | 2.46 | 2.39 | 2.35 | 2.31 | 2.27 | 2.22 | 2.18 | 2.15 |
| 15 | 4.54 | 3.68 | 3.29 | 3.06 | 2.90 | 2.79 | 2.71 | 2.64 | 2.59 | 2.54 | 2.48 | 2.40 | 2.33 | 2.29 | 2.25 | 2.20 | 2.16 | 2.11 | 2.09 |
| 16 | 4.49 | 3.63 | 3.24 | 3.01 | 2.85 | 2.74 | 2.66 | 2.59 | 2.54 | 2.49 | 2.42 | 2.35 | 2.28 | 2.24 | 2.19 | 2.15 | 2.11 | 2.06 | 2.03 |
| 17 | 4.45 | 3.59 | 3.20 | 2.96 | 2.81 | 2.70 | 2.61 | 2.55 | 2.49 | 2.45 | 2.38 | 2.31 | 2.23 | 2.19 | 2.15 | 2.10 | 2.06 | 2.01 | 1.98 |
| 18 | 4.41 | 3.55 | 3.16 | 2.93 | 2.77 | 2.66 | 2.58 | 2.51 | 2.46 | 2.41 | 2.34 | 2.27 | 2.19 | 2.15 | 2.11 | 2.06 | 2.02 | 1.97 | 1.94 |
| 19 | 4.38 | 3.52 | 3.13 | 2.90 | 2.74 | 2.63 | 2.54 | 2.48 | 2.42 | 2.38 | 2.31 | 2.23 | 2.16 | 2.11 | 2.07 | 2.03 | 1.98 | 1.93 | 1.90 |
| 20 | 4.35 | 3.49 | 3.10 | 2.87 | 2.71 | 2.60 | 2.51 | 2.45 | 2.39 | 2.35 | 2.28 | 2.20 | 2.12 | 2.08 | 2.04 | 1.99 | 1.95 | 1.90 | 1.87 |
| 21 | 4.32 | 3.47 | 3.07 | 2.84 | 2.68 | 2.57 | 2.49 | 2.42 | 2.37 | 2.32 | 2.25 | 2.18 | 2.10 | 2.05 | 2.01 | 1.96 | 1.92 | 1.87 | 1.84 |
| 22 | 4.30 | 3.44 | 3.05 | 2.82 | 2.66 | 2.55 | 2.46 | 2.40 | 2.34 | 2.30 | 2.23 | 2.15 | 2.07 | 2.03 | 1.98 | 1.94 | 1.89 | 1.84 | 1.81 |
| 23 | 4.28 | 3.42 | 3.03 | 2.80 | 2.64 | 2.53 | 2.44 | 2.37 | 2.32 | 2.27 | 2.20 | 2.13 | 2.05 | 2.01 | 1.96 | 1.91 | 1.86 | 1.81 | 1.78 |
| 24 | 4.26 | 3.40 | 3.01 | 2.78 | 2.62 | 2.51 | 2.42 | 2.36 | 2.30 | 2.25 | 2.18 | 2.11 | 2.03 | 1.98 | 1.94 | 1.89 | 1.84 | 1.79 | 1.76 |
| 25 | 4.24 | 3.39 | 2.99 | 2.76 | 2.60 | 2.49 | 2.40 | 2.34 | 2.28 | 2.24 | 2.16 | 2.09 | 2.01 | 1.96 | 1.92 | 1.87 | 1.82 | 1.77 | 1.74 |
| 26 | 4.23 | 3.37 | 2.98 | 2.74 | 2.59 | 2.47 | 2.39 | 2.32 | 2.27 | 2.22 | 2.15 | 2.07 | 1.99 | 1.95 | 1.90 | 1.85 | 1.80 | 1.75 | 1.72 |
| 27 | 4.21 | 3.35 | 2.96 | 2.73 | 2.57 | 2.46 | 2.37 | 2.31 | 2.25 | 2.20 | 2.13 | 2.06 | 1.97 | 1.93 | 1.88 | 1.84 | 1.79 | 1.73 | 1.70 |
| 28 | 4.20 | 3.34 | 2.95 | 2.71 | 2.56 | 2.45 | 2.36 | 2.29 | 2.24 | 2.19 | 2.12 | 2.04 | 1.96 | 1.91 | 1.87 | 1.82 | 1.77 | 1.71 | 1.68 |
| 29 | 4.18 | 3.33 | 2.93 | 2.70 | 2.55 | 2.43 | 2.35 | 2.28 | 2.22 | 2.18 | 2.10 | 2.03 | 1.94 | 1.90 | 1.85 | 1.81 | 1.75 | 1.70 | 1.67 |
| 30 | 4.17 | 3.32 | 2.92 | 2.69 | 2.53 | 2.42 | 2.33 | 2.27 | 2.21 | 2.16 | 2.09 | 2.01 | 1.93 | 1.89 | 1.84 | 1.79 | 1.74 | 1.68 | 1.65 |
| 40 | 4.08 | 3.23 | 2.84 | 2.61 | 2.45 | 2.34 | 2.25 | 2.18 | 2.12 | 2.08 | 2.00 | 1.92 | 1.84 | 1.79 | 1.74 | 1.69 | 1.64 | 1.58 | 1.54 |
| 60 | 4.00 | 3.15 | 2.76 | 2.53 | 2.37 | 2.25 | 2.17 | 2.10 | 2.04 | 1.99 | 1.92 | 1.84 | 1.75 | 1.70 | 1.65 | 1.59 | 1.53 | 1.47 | 1.43 |
| 120 | 3.92 | 3.07 | 2.68 | 2.45 | 2.29 | 2.18 | 2.09 | 2.02 | 1.96 | 1.91 | 1.83 | 1.75 | 1.66 | 1.61 | 1.55 | 1.50 | 1.43 | 1.35 | 1.30 |
| 250 | 3.88 | 3.03 | 2.64 | 2.41 | 2.25 | 2.13 | 2.05 | 1.98 | 1.92 | 1.87 | 1.79 | 1.71 | 1.61 | 1.56 | 1.50 | 1.44 | 1.37 | 1.29 | 1.23 |

Note: If this table does not include the value of DF Error that you want, you should either get a more complete table or else use the next *smaller* DF Error value in this table.

## F-Critical Values for $p = .025$

| DF Error (Denom) | Degrees of Freedom for Effect, Interaction, etc (Numerator) | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 12 | 15 | 20 | 24 | 30 | 40 | 60 | 120 | 250 |
| 1 | 647.79 | 799.50 | 864.16 | 899.58 | 921.85 | 937.11 | 948.22 | 956.66 | 963.28 | 968.63 | 976.71 | 984.87 | 993.10 | 997.25 | 1001.41 | 1005.60 | 1009.80 | 1014.02 | 1016.22 |
| 2 | 38.51 | 39.00 | 39.17 | 39.25 | 39.30 | 39.33 | 39.36 | 39.37 | 39.39 | 39.40 | 39.41 | 39.43 | 39.45 | 39.46 | 39.46 | 39.47 | 39.48 | 39.49 | 39.49 |
| 3 | 17.44 | 16.04 | 15.44 | 15.10 | 14.89 | 14.73 | 14.62 | 14.54 | 14.47 | 14.42 | 14.34 | 14.25 | 14.17 | 14.12 | 14.08 | 14.04 | 13.99 | 13.95 | 13.92 |
| 4 | 12.22 | 10.65 | 9.98 | 9.60 | 9.36 | 9.20 | 9.07 | 8.98 | 8.90 | 8.84 | 8.75 | 8.66 | 8.56 | 8.51 | 8.46 | 8.41 | 8.36 | 8.31 | 8.28 |
| 5 | 10.01 | 8.43 | 7.76 | 7.39 | 7.15 | 6.98 | 6.85 | 6.76 | 6.68 | 6.62 | 6.52 | 6.43 | 6.33 | 6.28 | 6.23 | 6.18 | 6.12 | 6.07 | 6.04 |
| 6 | 8.81 | 7.26 | 6.60 | 6.23 | 5.99 | 5.82 | 5.70 | 5.60 | 5.52 | 5.46 | 5.37 | 5.27 | 5.17 | 5.12 | 5.07 | 5.01 | 4.96 | 4.90 | 4.88 |
| 7 | 8.07 | 6.54 | 5.89 | 5.52 | 5.29 | 5.12 | 4.99 | 4.90 | 4.82 | 4.76 | 4.67 | 4.57 | 4.47 | 4.41 | 4.36 | 4.31 | 4.25 | 4.20 | 4.17 |
| 8 | 7.57 | 6.06 | 5.42 | 5.05 | 4.82 | 4.65 | 4.53 | 4.43 | 4.36 | 4.30 | 4.20 | 4.10 | 4.00 | 3.95 | 3.89 | 3.84 | 3.78 | 3.73 | 3.70 |
| 9 | 7.21 | 5.71 | 5.08 | 4.72 | 4.48 | 4.32 | 4.20 | 4.10 | 4.03 | 3.96 | 3.87 | 3.77 | 3.67 | 3.61 | 3.56 | 3.51 | 3.45 | 3.39 | 3.36 |
| 10 | 6.94 | 5.46 | 4.83 | 4.47 | 4.24 | 4.07 | 3.95 | 3.85 | 3.78 | 3.72 | 3.62 | 3.52 | 3.42 | 3.37 | 3.31 | 3.26 | 3.20 | 3.14 | 3.11 |
| 11 | 6.72 | 5.26 | 4.63 | 4.28 | 4.04 | 3.88 | 3.76 | 3.66 | 3.59 | 3.53 | 3.43 | 3.33 | 3.23 | 3.17 | 3.12 | 3.06 | 3.00 | 2.94 | 2.91 |
| 12 | 6.55 | 5.10 | 4.47 | 4.12 | 3.89 | 3.73 | 3.61 | 3.51 | 3.44 | 3.37 | 3.28 | 3.18 | 3.07 | 3.02 | 2.96 | 2.91 | 2.85 | 2.79 | 2.76 |
| 13 | 6.41 | 4.97 | 4.35 | 4.00 | 3.77 | 3.60 | 3.48 | 3.39 | 3.31 | 3.25 | 3.15 | 3.05 | 2.95 | 2.89 | 2.84 | 2.78 | 2.72 | 2.66 | 2.63 |
| 14 | 6.30 | 4.86 | 4.24 | 3.89 | 3.66 | 3.50 | 3.38 | 3.29 | 3.21 | 3.15 | 3.05 | 2.95 | 2.84 | 2.79 | 2.73 | 2.67 | 2.61 | 2.55 | 2.52 |
| 15 | 6.20 | 4.77 | 4.15 | 3.80 | 3.58 | 3.41 | 3.29 | 3.20 | 3.12 | 3.06 | 2.96 | 2.86 | 2.76 | 2.70 | 2.64 | 2.59 | 2.52 | 2.46 | 2.43 |
| 16 | 6.12 | 4.69 | 4.08 | 3.73 | 3.50 | 3.34 | 3.22 | 3.12 | 3.05 | 2.99 | 2.89 | 2.79 | 2.68 | 2.63 | 2.57 | 2.51 | 2.45 | 2.38 | 2.35 |
| 17 | 6.04 | 4.62 | 4.01 | 3.66 | 3.44 | 3.28 | 3.16 | 3.06 | 2.98 | 2.92 | 2.82 | 2.72 | 2.62 | 2.56 | 2.50 | 2.44 | 2.38 | 2.32 | 2.28 |
| 18 | 5.98 | 4.56 | 3.95 | 3.61 | 3.38 | 3.22 | 3.10 | 3.01 | 2.93 | 2.87 | 2.77 | 2.67 | 2.56 | 2.50 | 2.44 | 2.38 | 2.32 | 2.26 | 2.22 |
| 19 | 5.92 | 4.51 | 3.90 | 3.56 | 3.33 | 3.17 | 3.05 | 2.96 | 2.88 | 2.82 | 2.72 | 2.62 | 2.51 | 2.45 | 2.39 | 2.33 | 2.27 | 2.20 | 2.17 |
| 20 | 5.87 | 4.46 | 3.86 | 3.51 | 3.29 | 3.13 | 3.01 | 2.91 | 2.84 | 2.77 | 2.68 | 2.57 | 2.46 | 2.41 | 2.35 | 2.29 | 2.22 | 2.16 | 2.12 |
| 21 | 5.83 | 4.42 | 3.82 | 3.48 | 3.25 | 3.09 | 2.97 | 2.87 | 2.80 | 2.73 | 2.64 | 2.53 | 2.42 | 2.37 | 2.31 | 2.25 | 2.18 | 2.11 | 2.08 |
| 22 | 5.79 | 4.38 | 3.78 | 3.44 | 3.22 | 3.05 | 2.93 | 2.84 | 2.76 | 2.70 | 2.60 | 2.50 | 2.39 | 2.33 | 2.27 | 2.21 | 2.14 | 2.08 | 2.04 |
| 23 | 5.75 | 4.35 | 3.75 | 3.41 | 3.18 | 3.02 | 2.90 | 2.81 | 2.73 | 2.67 | 2.57 | 2.47 | 2.36 | 2.30 | 2.24 | 2.18 | 2.11 | 2.04 | 2.00 |
| 24 | 5.72 | 4.32 | 3.72 | 3.38 | 3.15 | 2.99 | 2.87 | 2.78 | 2.70 | 2.64 | 2.54 | 2.44 | 2.33 | 2.27 | 2.21 | 2.15 | 2.08 | 2.01 | 1.97 |
| 25 | 5.69 | 4.29 | 3.69 | 3.35 | 3.13 | 2.97 | 2.85 | 2.75 | 2.68 | 2.61 | 2.51 | 2.41 | 2.30 | 2.24 | 2.18 | 2.12 | 2.05 | 1.98 | 1.94 |
| 26 | 5.66 | 4.27 | 3.67 | 3.33 | 3.10 | 2.94 | 2.82 | 2.73 | 2.65 | 2.59 | 2.49 | 2.39 | 2.28 | 2.22 | 2.16 | 2.09 | 2.03 | 1.95 | 1.92 |
| 27 | 5.63 | 4.24 | 3.65 | 3.31 | 3.08 | 2.92 | 2.80 | 2.71 | 2.63 | 2.57 | 2.47 | 2.36 | 2.25 | 2.19 | 2.13 | 2.07 | 2.00 | 1.93 | 1.89 |
| 28 | 5.61 | 4.22 | 3.63 | 3.29 | 3.06 | 2.90 | 2.78 | 2.69 | 2.61 | 2.55 | 2.45 | 2.34 | 2.23 | 2.17 | 2.11 | 2.05 | 1.98 | 1.91 | 1.87 |
| 29 | 5.59 | 4.20 | 3.61 | 3.27 | 3.04 | 2.88 | 2.76 | 2.67 | 2.59 | 2.53 | 2.43 | 2.32 | 2.21 | 2.15 | 2.09 | 2.03 | 1.96 | 1.89 | 1.85 |
| 30 | 5.57 | 4.18 | 3.59 | 3.25 | 3.03 | 2.87 | 2.75 | 2.65 | 2.57 | 2.51 | 2.41 | 2.31 | 2.20 | 2.14 | 2.07 | 2.01 | 1.94 | 1.87 | 1.83 |
| 40 | 5.42 | 4.05 | 3.46 | 3.13 | 2.90 | 2.74 | 2.62 | 2.53 | 2.45 | 2.39 | 2.29 | 2.18 | 2.07 | 2.01 | 1.94 | 1.88 | 1.80 | 1.72 | 1.68 |
| 60 | 5.29 | 3.93 | 3.34 | 3.01 | 2.79 | 2.63 | 2.51 | 2.41 | 2.33 | 2.27 | 2.17 | 2.06 | 1.94 | 1.88 | 1.82 | 1.74 | 1.67 | 1.58 | 1.53 |
| 120 | 5.15 | 3.80 | 3.23 | 2.89 | 2.67 | 2.52 | 2.39 | 2.30 | 2.22 | 2.16 | 2.05 | 1.94 | 1.82 | 1.76 | 1.69 | 1.61 | 1.53 | 1.43 | 1.37 |
| 250 | 5.08 | 3.74 | 3.17 | 2.84 | 2.62 | 2.46 | 2.34 | 2.24 | 2.16 | 2.10 | 2.00 | 1.89 | 1.76 | 1.70 | 1.63 | 1.55 | 1.46 | 1.35 | 1.28 |
| 500 | 5.05 | 3.72 | 3.14 | 2.81 | 2.59 | 2.43 | 2.31 | 2.22 | 2.14 | 2.07 | 1.97 | 1.86 | 1.74 | 1.67 | 1.60 | 1.52 | 1.42 | 1.31 | 1.24 |

Note: If this table does not include the value of DF Error that you want, you should either get a more complete table or else use the next *smaller* DF Error value in this table.

## F-Critical Values for $p = .01$

| DF Error (Denom) | Degrees of Freedom for Effect, Interaction, etc (Numerator) | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 12 | 15 | 20 | 24 | 30 | 40 | 60 | 120 | 250 |
| 1 | 4052.2 | 4999.5 | 5403.3 | 5624.6 | 5763.6 | 5859.0 | 5928.4 | 5981.1 | 6022.5 | 6055.8 | 6106.3 | 6157.3 | 6208.7 | 6234.6 | 6260.6 | 6286.8 | 6313.0 | 6339.4 | 6353.1 |
| 2 | 98.50 | 99.00 | 99.17 | 99.25 | 99.30 | 99.33 | 99.36 | 99.37 | 99.39 | 99.40 | 99.42 | 99.43 | 99.45 | 99.46 | 99.47 | 99.47 | 99.48 | 99.49 | 99.50 |
| 3 | 34.12 | 30.82 | 29.46 | 28.71 | 28.24 | 27.91 | 27.67 | 27.49 | 27.35 | 27.23 | 27.05 | 26.87 | 26.69 | 26.60 | 26.50 | 26.41 | 26.32 | 26.22 | 26.17 |
| 4 | 21.20 | 18.00 | 16.69 | 15.98 | 15.52 | 15.21 | 14.98 | 14.80 | 14.66 | 14.55 | 14.37 | 14.20 | 14.02 | 13.93 | 13.84 | 13.75 | 13.65 | 13.56 | 13.51 |
| 5 | 16.26 | 13.27 | 12.06 | 11.39 | 10.97 | 10.67 | 10.46 | 10.29 | 10.16 | 10.05 | 9.89 | 9.72 | 9.55 | 9.47 | 9.38 | 9.29 | 9.20 | 9.11 | 9.06 |
| 6 | 13.75 | 10.92 | 9.78 | 9.15 | 8.75 | 8.47 | 8.26 | 8.10 | 7.98 | 7.87 | 7.72 | 7.56 | 7.40 | 7.31 | 7.23 | 7.14 | 7.06 | 6.97 | 6.92 |
| 7 | 12.25 | 9.55 | 8.45 | 7.85 | 7.46 | 7.19 | 6.99 | 6.84 | 6.72 | 6.62 | 6.47 | 6.31 | 6.16 | 6.07 | 5.99 | 5.91 | 5.82 | 5.74 | 5.69 |
| 8 | 11.26 | 8.65 | 7.59 | 7.01 | 6.63 | 6.37 | 6.18 | 6.03 | 5.91 | 5.81 | 5.67 | 5.52 | 5.36 | 5.28 | 5.20 | 5.12 | 5.03 | 4.95 | 4.90 |
| 9 | 10.56 | 8.02 | 6.99 | 6.42 | 6.06 | 5.80 | 5.61 | 5.47 | 5.35 | 5.26 | 5.11 | 4.96 | 4.81 | 4.73 | 4.65 | 4.57 | 4.48 | 4.40 | 4.35 |
| 10 | 10.04 | 7.56 | 6.55 | 5.99 | 5.64 | 5.39 | 5.20 | 5.06 | 4.94 | 4.85 | 4.71 | 4.56 | 4.41 | 4.33 | 4.25 | 4.17 | 4.08 | 4.00 | 3.95 |
| 11 | 9.65 | 7.21 | 6.22 | 5.67 | 5.32 | 5.07 | 4.89 | 4.74 | 4.63 | 4.54 | 4.40 | 4.25 | 4.10 | 4.02 | 3.94 | 3.86 | 3.78 | 3.69 | 3.64 |
| 12 | 9.33 | 6.93 | 5.95 | 5.41 | 5.06 | 4.82 | 4.64 | 4.50 | 4.39 | 4.30 | 4.16 | 4.01 | 3.86 | 3.78 | 3.70 | 3.62 | 3.54 | 3.45 | 3.40 |
| 13 | 9.07 | 6.70 | 5.74 | 5.21 | 4.86 | 4.62 | 4.44 | 4.30 | 4.19 | 4.10 | 3.96 | 3.82 | 3.66 | 3.59 | 3.51 | 3.43 | 3.34 | 3.25 | 3.21 |
| 14 | 8.86 | 6.51 | 5.56 | 5.04 | 4.69 | 4.46 | 4.28 | 4.14 | 4.03 | 3.94 | 3.80 | 3.66 | 3.51 | 3.43 | 3.35 | 3.27 | 3.18 | 3.09 | 3.05 |
| 15 | 8.68 | 6.36 | 5.42 | 4.89 | 4.56 | 4.32 | 4.14 | 4.00 | 3.89 | 3.80 | 3.67 | 3.52 | 3.37 | 3.29 | 3.21 | 3.13 | 3.05 | 2.96 | 2.91 |
| 16 | 8.53 | 6.23 | 5.29 | 4.77 | 4.44 | 4.20 | 4.03 | 3.89 | 3.78 | 3.69 | 3.55 | 3.41 | 3.26 | 3.18 | 3.10 | 3.02 | 2.93 | 2.84 | 2.80 |
| 17 | 8.40 | 6.11 | 5.18 | 4.67 | 4.34 | 4.10 | 3.93 | 3.79 | 3.68 | 3.59 | 3.46 | 3.31 | 3.16 | 3.08 | 3.00 | 2.92 | 2.83 | 2.75 | 2.70 |
| 18 | 8.29 | 6.01 | 5.09 | 4.58 | 4.25 | 4.01 | 3.84 | 3.71 | 3.60 | 3.51 | 3.37 | 3.23 | 3.08 | 3.00 | 2.92 | 2.84 | 2.75 | 2.66 | 2.61 |
| 19 | 8.18 | 5.93 | 5.01 | 4.50 | 4.17 | 3.94 | 3.77 | 3.63 | 3.52 | 3.43 | 3.30 | 3.15 | 3.00 | 2.92 | 2.84 | 2.76 | 2.67 | 2.58 | 2.54 |
| 20 | 8.10 | 5.85 | 4.94 | 4.43 | 4.10 | 3.87 | 3.70 | 3.56 | 3.46 | 3.37 | 3.23 | 3.09 | 2.94 | 2.86 | 2.78 | 2.69 | 2.61 | 2.52 | 2.47 |
| 21 | 8.02 | 5.78 | 4.87 | 4.37 | 4.04 | 3.81 | 3.64 | 3.51 | 3.40 | 3.31 | 3.17 | 3.03 | 2.88 | 2.80 | 2.72 | 2.64 | 2.55 | 2.46 | 2.41 |
| 22 | 7.95 | 5.72 | 4.82 | 4.31 | 3.99 | 3.76 | 3.59 | 3.45 | 3.35 | 3.26 | 3.12 | 2.98 | 2.83 | 2.75 | 2.67 | 2.58 | 2.50 | 2.40 | 2.35 |
| 23 | 7.88 | 5.66 | 4.76 | 4.26 | 3.94 | 3.71 | 3.54 | 3.41 | 3.30 | 3.21 | 3.07 | 2.93 | 2.78 | 2.70 | 2.62 | 2.54 | 2.45 | 2.35 | 2.30 |
| 24 | 7.82 | 5.61 | 4.72 | 4.22 | 3.90 | 3.67 | 3.50 | 3.36 | 3.26 | 3.17 | 3.03 | 2.89 | 2.74 | 2.66 | 2.58 | 2.49 | 2.40 | 2.31 | 2.26 |
| 25 | 7.77 | 5.57 | 4.68 | 4.18 | 3.85 | 3.63 | 3.46 | 3.32 | 3.22 | 3.13 | 2.99 | 2.85 | 2.70 | 2.62 | 2.54 | 2.45 | 2.36 | 2.27 | 2.22 |
| 26 | 7.72 | 5.53 | 4.64 | 4.14 | 3.82 | 3.59 | 3.42 | 3.29 | 3.18 | 3.09 | 2.96 | 2.81 | 2.66 | 2.58 | 2.50 | 2.42 | 2.33 | 2.23 | 2.18 |
| 27 | 7.68 | 5.49 | 4.60 | 4.11 | 3.78 | 3.56 | 3.39 | 3.26 | 3.15 | 3.06 | 2.93 | 2.78 | 2.63 | 2.55 | 2.47 | 2.38 | 2.29 | 2.20 | 2.15 |
| 28 | 7.64 | 5.45 | 4.57 | 4.07 | 3.75 | 3.53 | 3.36 | 3.23 | 3.12 | 3.03 | 2.90 | 2.75 | 2.60 | 2.52 | 2.44 | 2.35 | 2.26 | 2.17 | 2.11 |
| 29 | 7.60 | 5.42 | 4.54 | 4.04 | 3.73 | 3.50 | 3.33 | 3.20 | 3.09 | 3.00 | 2.87 | 2.73 | 2.57 | 2.49 | 2.41 | 2.33 | 2.23 | 2.14 | 2.08 |
| 30 | 7.56 | 5.39 | 4.51 | 4.02 | 3.70 | 3.47 | 3.30 | 3.17 | 3.07 | 2.98 | 2.84 | 2.70 | 2.55 | 2.47 | 2.39 | 2.30 | 2.21 | 2.11 | 2.06 |
| 40 | 7.31 | 5.18 | 4.31 | 3.83 | 3.51 | 3.29 | 3.12 | 2.99 | 2.89 | 2.80 | 2.66 | 2.52 | 2.37 | 2.29 | 2.20 | 2.11 | 2.02 | 1.92 | 1.86 |
| 60 | 7.08 | 4.98 | 4.13 | 3.65 | 3.34 | 3.12 | 2.95 | 2.82 | 2.72 | 2.63 | 2.50 | 2.35 | 2.20 | 2.12 | 2.03 | 1.94 | 1.84 | 1.73 | 1.66 |
| 120 | 6.85 | 4.79 | 3.95 | 3.48 | 3.17 | 2.96 | 2.79 | 2.66 | 2.56 | 2.47 | 2.34 | 2.19 | 2.03 | 1.95 | 1.86 | 1.76 | 1.66 | 1.53 | 1.46 |
| 250 | 6.74 | 4.69 | 3.86 | 3.40 | 3.09 | 2.87 | 2.71 | 2.58 | 2.48 | 2.39 | 2.26 | 2.11 | 1.95 | 1.87 | 1.77 | 1.67 | 1.56 | 1.43 | 1.34 |
| 500 | 6.69 | 4.65 | 3.82 | 3.36 | 3.05 | 2.84 | 2.68 | 2.55 | 2.44 | 2.36 | 2.22 | 2.07 | 1.92 | 1.83 | 1.74 | 1.63 | 1.52 | 1.38 | 1.28 |

Note: If this table does not include the value of DF Error that you want, you should either get a more complete table or else use the next *smaller* DF Error value in this table.

# F-Critical Values for $p = .005$

| DF Error (Denom) | Degrees of Freedom for Effect, Interaction, etc (Numerator) | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 12 | 15 | 20 | 24 | 30 | 40 | 60 | 120 | 250 |
| 1 | Too Big for my Computer! | | | | | | | | | | | | | | | | | | |
| 2 | 198.50 | 199.00 | 199.17 | 199.25 | 199.30 | 199.33 | 199.36 | 199.37 | 199.39 | 199.40 | 199.42 | 199.43 | 199.45 | 199.46 | 199.47 | 199.47 | 199.48 | 199.49 | 199.50 |
| 3 | 55.55 | 49.80 | 47.47 | 46.19 | 45.39 | 44.84 | 44.43 | 44.13 | 43.88 | 43.69 | 43.39 | 43.08 | 42.78 | 42.62 | 42.47 | 42.31 | 42.15 | 41.99 | 41.91 |
| 4 | 31.33 | 26.28 | 24.26 | 23.15 | 22.46 | 21.97 | 21.62 | 21.35 | 21.14 | 20.97 | 20.70 | 20.44 | 20.17 | 20.03 | 19.89 | 19.75 | 19.61 | 19.47 | 19.39 |
| 5 | 22.78 | 18.31 | 16.53 | 15.56 | 14.94 | 14.51 | 14.20 | 13.96 | 13.77 | 13.62 | 13.38 | 13.15 | 12.90 | 12.78 | 12.66 | 12.53 | 12.40 | 12.27 | 12.21 |
| 6 | 18.64 | 14.54 | 12.92 | 12.03 | 11.46 | 11.07 | 10.79 | 10.57 | 10.39 | 10.25 | 10.03 | 9.81 | 9.59 | 9.47 | 9.36 | 9.24 | 9.12 | 9.00 | 8.94 |
| 7 | 16.24 | 12.40 | 10.88 | 10.05 | 9.52 | 9.16 | 8.89 | 8.68 | 8.51 | 8.38 | 8.18 | 7.97 | 7.75 | 7.64 | 7.53 | 7.42 | 7.31 | 7.19 | 7.13 |
| 8 | 14.69 | 11.04 | 9.60 | 8.81 | 8.30 | 7.95 | 7.69 | 7.50 | 7.34 | 7.21 | 7.01 | 6.81 | 6.61 | 6.50 | 6.40 | 6.29 | 6.18 | 6.06 | 6.01 |
| 9 | 13.61 | 10.11 | 8.72 | 7.96 | 7.47 | 7.13 | 6.88 | 6.69 | 6.54 | 6.42 | 6.23 | 6.03 | 5.83 | 5.73 | 5.62 | 5.52 | 5.41 | 5.30 | 5.24 |
| 10 | 12.83 | 9.43 | 8.08 | 7.34 | 6.87 | 6.54 | 6.30 | 6.12 | 5.97 | 5.85 | 5.66 | 5.47 | 5.27 | 5.17 | 5.07 | 4.97 | 4.86 | 4.75 | 4.69 |
| 11 | 12.23 | 8.91 | 7.60 | 6.88 | 6.42 | 6.10 | 5.86 | 5.68 | 5.54 | 5.42 | 5.24 | 5.05 | 4.86 | 4.76 | 4.65 | 4.55 | 4.45 | 4.34 | 4.28 |
| 12 | 11.75 | 8.51 | 7.23 | 6.52 | 6.07 | 5.76 | 5.52 | 5.35 | 5.20 | 5.09 | 4.91 | 4.72 | 4.53 | 4.43 | 4.33 | 4.23 | 4.12 | 4.01 | 3.96 |
| 13 | 11.37 | 8.19 | 6.93 | 6.23 | 5.79 | 5.48 | 5.25 | 5.08 | 4.94 | 4.82 | 4.64 | 4.46 | 4.27 | 4.17 | 4.07 | 3.97 | 3.87 | 3.76 | 3.70 |
| 14 | 11.06 | 7.92 | 6.68 | 6.00 | 5.56 | 5.26 | 5.03 | 4.86 | 4.72 | 4.60 | 4.43 | 4.25 | 4.06 | 3.96 | 3.86 | 3.76 | 3.66 | 3.55 | 3.49 |
| 15 | 10.80 | 7.70 | 6.48 | 5.80 | 5.37 | 5.07 | 4.85 | 4.67 | 4.54 | 4.42 | 4.25 | 4.07 | 3.88 | 3.79 | 3.69 | 3.58 | 3.48 | 3.37 | 3.31 |
| 16 | 10.58 | 7.51 | 6.30 | 5.64 | 5.21 | 4.91 | 4.69 | 4.52 | 4.38 | 4.27 | 4.10 | 3.92 | 3.73 | 3.64 | 3.54 | 3.44 | 3.33 | 3.22 | 3.17 |
| 17 | 10.38 | 7.35 | 6.16 | 5.50 | 5.07 | 4.78 | 4.56 | 4.39 | 4.25 | 4.14 | 3.97 | 3.79 | 3.61 | 3.51 | 3.41 | 3.31 | 3.21 | 3.10 | 3.04 |
| 18 | 10.22 | 7.21 | 6.03 | 5.37 | 4.96 | 4.66 | 4.44 | 4.28 | 4.14 | 4.03 | 3.86 | 3.68 | 3.50 | 3.40 | 3.30 | 3.20 | 3.10 | 2.99 | 2.93 |
| 19 | 10.07 | 7.09 | 5.92 | 5.27 | 4.85 | 4.56 | 4.34 | 4.18 | 4.04 | 3.93 | 3.76 | 3.59 | 3.40 | 3.31 | 3.21 | 3.11 | 3.00 | 2.89 | 2.83 |
| 20 | 9.94 | 6.99 | 5.82 | 5.17 | 4.76 | 4.47 | 4.26 | 4.09 | 3.96 | 3.85 | 3.68 | 3.50 | 3.32 | 3.22 | 3.12 | 3.02 | 2.92 | 2.81 | 2.75 |
| 21 | 9.83 | 6.89 | 5.73 | 5.09 | 4.68 | 4.39 | 4.18 | 4.01 | 3.88 | 3.77 | 3.60 | 3.43 | 3.24 | 3.15 | 3.05 | 2.95 | 2.84 | 2.73 | 2.67 |
| 22 | 9.73 | 6.81 | 5.65 | 5.02 | 4.61 | 4.32 | 4.11 | 3.94 | 3.81 | 3.70 | 3.54 | 3.36 | 3.18 | 3.08 | 2.98 | 2.88 | 2.77 | 2.66 | 2.60 |
| 23 | 9.63 | 6.73 | 5.58 | 4.95 | 4.54 | 4.26 | 4.05 | 3.88 | 3.75 | 3.64 | 3.47 | 3.30 | 3.12 | 3.02 | 2.92 | 2.82 | 2.71 | 2.60 | 2.54 |
| 24 | 9.55 | 6.66 | 5.52 | 4.89 | 4.49 | 4.20 | 3.99 | 3.83 | 3.69 | 3.59 | 3.42 | 3.25 | 3.06 | 2.97 | 2.87 | 2.77 | 2.66 | 2.55 | 2.49 |
| 25 | 9.48 | 6.60 | 5.46 | 4.84 | 4.43 | 4.15 | 3.94 | 3.78 | 3.64 | 3.54 | 3.37 | 3.20 | 3.01 | 2.92 | 2.82 | 2.72 | 2.61 | 2.50 | 2.43 |
| 26 | 9.41 | 6.54 | 5.41 | 4.79 | 4.38 | 4.10 | 3.89 | 3.73 | 3.60 | 3.49 | 3.33 | 3.15 | 2.97 | 2.87 | 2.77 | 2.67 | 2.56 | 2.45 | 2.39 |
| 27 | 9.34 | 6.49 | 5.36 | 4.74 | 4.34 | 4.06 | 3.85 | 3.69 | 3.56 | 3.45 | 3.28 | 3.11 | 2.93 | 2.83 | 2.73 | 2.63 | 2.52 | 2.41 | 2.35 |
| 28 | 9.28 | 6.44 | 5.32 | 4.70 | 4.30 | 4.02 | 3.81 | 3.65 | 3.52 | 3.41 | 3.25 | 3.07 | 2.89 | 2.79 | 2.69 | 2.59 | 2.48 | 2.37 | 2.31 |
| 29 | 9.23 | 6.40 | 5.28 | 4.66 | 4.26 | 3.98 | 3.77 | 3.61 | 3.48 | 3.38 | 3.21 | 3.04 | 2.86 | 2.76 | 2.66 | 2.56 | 2.45 | 2.33 | 2.27 |
| 30 | 9.18 | 6.35 | 5.24 | 4.62 | 4.23 | 3.95 | 3.74 | 3.58 | 3.45 | 3.34 | 3.18 | 3.01 | 2.82 | 2.73 | 2.63 | 2.52 | 2.42 | 2.30 | 2.24 |
| 40 | 8.83 | 6.07 | 4.98 | 4.37 | 3.99 | 3.71 | 3.51 | 3.35 | 3.22 | 3.12 | 2.95 | 2.78 | 2.60 | 2.50 | 2.40 | 2.30 | 2.18 | 2.06 | 2.00 |
| 60 | 8.49 | 5.79 | 4.73 | 4.14 | 3.76 | 3.49 | 3.29 | 3.13 | 3.01 | 2.90 | 2.74 | 2.57 | 2.39 | 2.29 | 2.19 | 2.08 | 1.96 | 1.83 | 1.76 |
| 120 | 8.18 | 5.54 | 4.50 | 3.92 | 3.55 | 3.28 | 3.09 | 2.93 | 2.81 | 2.71 | 2.54 | 2.37 | 2.19 | 2.09 | 1.98 | 1.87 | 1.75 | 1.61 | 1.52 |
| 250 | 8.02 | 5.41 | 4.38 | 3.81 | 3.44 | 3.18 | 2.99 | 2.83 | 2.71 | 2.61 | 2.45 | 2.27 | 2.09 | 1.99 | 1.88 | 1.77 | 1.64 | 1.48 | 1.39 |
| 500 | 7.95 | 5.35 | 4.33 | 3.76 | 3.40 | 3.14 | 2.94 | 2.79 | 2.66 | 2.56 | 2.40 | 2.23 | 2.04 | 1.94 | 1.84 | 1.72 | 1.58 | 1.42 | 1.32 |

Note: If this table does not include the value of DF Error that you want, you should either get a more complete table or else use the next *smaller* DF Error value in this table.

## F-Critical Values for $p = .001$

| DF Error (Denom) | Degrees of Freedom for Effect, Interaction, etc (Numerator) | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 12 | 15 | 20 | 24 | 30 | 40 | 60 | 120 | 250 |
| 1 | Too Big for my Computer! | | | | | | | | | | | | | | | | | | |
| 2 | 998.50 | 999.00 | 999.17 | 999.25 | 999.30 | 999.33 | 999.36 | 999.38 | 999.39 | 999.40 | 999.42 | 999.43 | 999.45 | 999.46 | 999.47 | 999.47 | 999.48 | 999.49 | 999.50 |
| 3 | 167.03 | 148.50 | 141.11 | 137.10 | 134.58 | 132.85 | 131.58 | 130.62 | 129.86 | 129.25 | 128.32 | 127.37 | 126.42 | 125.94 | 125.45 | 124.96 | 124.47 | 123.97 | 123.71 |
| 4 | 74.14 | 61.25 | 56.18 | 53.44 | 51.71 | 50.53 | 49.66 | 49.00 | 48.47 | 48.05 | 47.41 | 46.76 | 46.10 | 45.77 | 45.43 | 45.09 | 44.75 | 44.40 | 44.22 |
| 5 | 47.18 | 37.12 | 33.20 | 31.09 | 29.75 | 28.83 | 28.16 | 27.65 | 27.24 | 26.92 | 26.42 | 25.91 | 25.39 | 25.13 | 24.87 | 24.60 | 24.33 | 24.06 | 23.92 |
| 6 | 35.51 | 27.00 | 23.70 | 21.92 | 20.80 | 20.03 | 19.46 | 19.03 | 18.69 | 18.41 | 17.99 | 17.56 | 17.12 | 16.90 | 16.67 | 16.44 | 16.21 | 15.98 | 15.86 |
| 7 | 29.25 | 21.69 | 18.77 | 17.20 | 16.21 | 15.52 | 15.02 | 14.63 | 14.33 | 14.08 | 13.71 | 13.32 | 12.93 | 12.73 | 12.53 | 12.33 | 12.12 | 11.91 | 11.80 |
| 8 | 25.41 | 18.49 | 15.83 | 14.39 | 13.48 | 12.86 | 12.40 | 12.05 | 11.77 | 11.54 | 11.19 | 10.84 | 10.48 | 10.30 | 10.11 | 9.92 | 9.73 | 9.53 | 9.43 |
| 9 | 22.86 | 16.39 | 13.90 | 12.56 | 11.71 | 11.13 | 10.70 | 10.37 | 10.11 | 9.89 | 9.57 | 9.24 | 8.90 | 8.72 | 8.55 | 8.37 | 8.19 | 8.00 | 7.90 |
| 10 | 21.04 | 14.91 | 12.55 | 11.28 | 10.48 | 9.93 | 9.52 | 9.20 | 8.96 | 8.75 | 8.45 | 8.13 | 7.80 | 7.64 | 7.47 | 7.30 | 7.12 | 6.94 | 6.85 |
| 11 | 19.69 | 13.81 | 11.56 | 10.35 | 9.58 | 9.05 | 8.66 | 8.35 | 8.12 | 7.92 | 7.63 | 7.32 | 7.01 | 6.85 | 6.68 | 6.52 | 6.35 | 6.18 | 6.08 |
| 12 | 18.64 | 12.97 | 10.80 | 9.63 | 8.89 | 8.38 | 8.00 | 7.71 | 7.48 | 7.29 | 7.00 | 6.71 | 6.40 | 6.25 | 6.09 | 5.93 | 5.76 | 5.59 | 5.50 |
| 13 | 17.82 | 12.31 | 10.21 | 9.07 | 8.35 | 7.86 | 7.49 | 7.21 | 6.98 | 6.80 | 6.52 | 6.23 | 5.93 | 5.78 | 5.63 | 5.47 | 5.30 | 5.14 | 5.05 |
| 14 | 17.14 | 11.78 | 9.73 | 8.62 | 7.92 | 7.44 | 7.08 | 6.80 | 6.58 | 6.40 | 6.13 | 5.85 | 5.56 | 5.41 | 5.25 | 5.10 | 4.94 | 4.77 | 4.69 |
| 15 | 16.59 | 11.34 | 9.34 | 8.25 | 7.57 | 7.09 | 6.74 | 6.47 | 6.26 | 6.08 | 5.81 | 5.54 | 5.25 | 5.10 | 4.95 | 4.80 | 4.64 | 4.47 | 4.39 |
| 16 | 16.12 | 10.97 | 9.01 | 7.94 | 7.27 | 6.80 | 6.46 | 6.19 | 5.98 | 5.81 | 5.55 | 5.27 | 4.99 | 4.85 | 4.70 | 4.54 | 4.39 | 4.23 | 4.14 |
| 17 | 15.72 | 10.66 | 8.73 | 7.68 | 7.02 | 6.56 | 6.22 | 5.96 | 5.75 | 5.58 | 5.32 | 5.05 | 4.78 | 4.63 | 4.48 | 4.33 | 4.18 | 4.02 | 3.93 |
| 18 | 15.38 | 10.39 | 8.49 | 7.46 | 6.81 | 6.35 | 6.02 | 5.76 | 5.56 | 5.39 | 5.13 | 4.87 | 4.59 | 4.45 | 4.30 | 4.15 | 4.00 | 3.84 | 3.75 |
| 19 | 15.08 | 10.16 | 8.28 | 7.27 | 6.62 | 6.18 | 5.85 | 5.59 | 5.39 | 5.22 | 4.97 | 4.70 | 4.43 | 4.29 | 4.14 | 3.99 | 3.84 | 3.68 | 3.59 |
| 20 | 14.82 | 9.95 | 8.10 | 7.10 | 6.46 | 6.02 | 5.69 | 5.44 | 5.24 | 5.08 | 4.82 | 4.56 | 4.29 | 4.15 | 4.00 | 3.86 | 3.70 | 3.54 | 3.46 |
| 21 | 14.59 | 9.77 | 7.94 | 6.95 | 6.32 | 5.88 | 5.56 | 5.31 | 5.11 | 4.95 | 4.70 | 4.44 | 4.17 | 4.03 | 3.88 | 3.74 | 3.58 | 3.42 | 3.34 |
| 22 | 14.38 | 9.61 | 7.80 | 6.81 | 6.19 | 5.76 | 5.44 | 5.19 | 4.99 | 4.83 | 4.58 | 4.33 | 4.06 | 3.92 | 3.78 | 3.63 | 3.48 | 3.32 | 3.23 |
| 23 | 14.20 | 9.47 | 7.67 | 6.70 | 6.08 | 5.65 | 5.33 | 5.09 | 4.89 | 4.73 | 4.48 | 4.23 | 3.96 | 3.82 | 3.68 | 3.53 | 3.38 | 3.22 | 3.14 |
| 24 | 14.03 | 9.34 | 7.55 | 6.59 | 5.98 | 5.55 | 5.23 | 4.99 | 4.80 | 4.64 | 4.39 | 4.14 | 3.87 | 3.74 | 3.59 | 3.45 | 3.29 | 3.14 | 3.05 |
| 25 | 13.88 | 9.22 | 7.45 | 6.49 | 5.89 | 5.46 | 5.15 | 4.91 | 4.71 | 4.56 | 4.31 | 4.06 | 3.79 | 3.66 | 3.52 | 3.37 | 3.22 | 3.06 | 2.97 |
| 26 | 13.74 | 9.12 | 7.36 | 6.41 | 5.80 | 5.38 | 5.07 | 4.83 | 4.64 | 4.48 | 4.24 | 3.99 | 3.72 | 3.59 | 3.44 | 3.30 | 3.15 | 2.99 | 2.90 |
| 27 | 13.61 | 9.02 | 7.27 | 6.33 | 5.73 | 5.31 | 5.00 | 4.76 | 4.57 | 4.41 | 4.17 | 3.92 | 3.66 | 3.52 | 3.38 | 3.23 | 3.08 | 2.92 | 2.84 |
| 28 | 13.50 | 8.93 | 7.19 | 6.25 | 5.66 | 5.24 | 4.93 | 4.69 | 4.50 | 4.35 | 4.11 | 3.86 | 3.60 | 3.46 | 3.32 | 3.18 | 3.02 | 2.86 | 2.78 |
| 29 | 13.39 | 8.85 | 7.12 | 6.19 | 5.59 | 5.18 | 4.87 | 4.64 | 4.45 | 4.29 | 4.05 | 3.80 | 3.54 | 3.41 | 3.27 | 3.12 | 2.97 | 2.81 | 2.72 |
| 30 | 13.29 | 8.77 | 7.05 | 6.12 | 5.53 | 5.12 | 4.82 | 4.58 | 4.39 | 4.24 | 4.00 | 3.75 | 3.49 | 3.36 | 3.22 | 3.07 | 2.92 | 2.76 | 2.67 |
| 40 | 12.61 | 8.25 | 6.59 | 5.70 | 5.13 | 4.73 | 4.44 | 4.21 | 4.02 | 3.87 | 3.64 | 3.40 | 3.14 | 3.01 | 2.87 | 2.73 | 2.57 | 2.41 | 2.32 |
| 60 | 11.97 | 7.77 | 6.17 | 5.31 | 4.76 | 4.37 | 4.09 | 3.86 | 3.69 | 3.54 | 3.32 | 3.08 | 2.83 | 2.69 | 2.55 | 2.41 | 2.25 | 2.08 | 1.99 |
| 120 | 11.38 | 7.32 | 5.78 | 4.95 | 4.42 | 4.04 | 3.77 | 3.55 | 3.38 | 3.24 | 3.02 | 2.78 | 2.53 | 2.40 | 2.26 | 2.11 | 1.95 | 1.77 | 1.66 |
| 250 | 11.09 | 7.10 | 5.59 | 4.77 | 4.25 | 3.88 | 3.61 | 3.40 | 3.23 | 3.09 | 2.87 | 2.64 | 2.39 | 2.26 | 2.12 | 1.97 | 1.80 | 1.60 | 1.48 |
| 500 | 10.96 | 7.00 | 5.51 | 4.69 | 4.18 | 3.81 | 3.54 | 3.33 | 3.16 | 3.02 | 2.81 | 2.58 | 2.33 | 2.20 | 2.05 | 1.90 | 1.73 | 1.53 | 1.39 |

Note: If this table does not include the value of DF Error that you want, you should either get a more complete table or else use the next *smaller* DF Error value in this table.

## B.2 Critical Values of Correlation Coefficient (Pearson $r$)

| Sample Size | Significance or "$p$" Level (Two-Tailed) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | .10 | .05 | .025 | .02 | .01 | .005 | .0025 | .001 | .0005 |
| 3 | 0.988 | 0.997 | 0.999 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 4 | 0.900 | 0.950 | 0.975 | 0.980 | 0.990 | 0.995 | 0.998 | 0.999 | 0.999 |
| 5 | 0.805 | 0.878 | 0.924 | 0.934 | 0.959 | 0.974 | 0.984 | 0.991 | 0.994 |
| 6 | 0.729 | 0.811 | 0.868 | 0.882 | 0.917 | 0.942 | 0.959 | 0.974 | 0.982 |
| 7 | 0.669 | 0.754 | 0.817 | 0.833 | 0.875 | 0.906 | 0.929 | 0.951 | 0.963 |
| 8 | 0.621 | 0.707 | 0.771 | 0.789 | 0.834 | 0.870 | 0.897 | 0.925 | 0.941 |
| 9 | 0.582 | 0.666 | 0.732 | 0.750 | 0.798 | 0.836 | 0.867 | 0.898 | 0.917 |
| 10 | 0.549 | 0.632 | 0.697 | 0.715 | 0.765 | 0.805 | 0.837 | 0.872 | 0.893 |
| 11 | 0.521 | 0.602 | 0.667 | 0.685 | 0.735 | 0.776 | 0.810 | 0.847 | 0.870 |
| 12 | 0.497 | 0.576 | 0.640 | 0.658 | 0.708 | 0.750 | 0.785 | 0.823 | 0.847 |
| 13 | 0.476 | 0.553 | 0.616 | 0.634 | 0.684 | 0.726 | 0.761 | 0.801 | 0.826 |
| 14 | 0.458 | 0.532 | 0.594 | 0.612 | 0.661 | 0.703 | 0.740 | 0.780 | 0.806 |
| 15 | 0.441 | 0.514 | 0.575 | 0.592 | 0.641 | 0.683 | 0.719 | 0.760 | 0.787 |
| 16 | 0.426 | 0.497 | 0.557 | 0.574 | 0.623 | 0.664 | 0.701 | 0.742 | 0.769 |
| 17 | 0.412 | 0.482 | 0.541 | 0.558 | 0.606 | 0.647 | 0.683 | 0.725 | 0.752 |
| 18 | 0.400 | 0.468 | 0.526 | 0.543 | 0.590 | 0.631 | 0.667 | 0.708 | 0.736 |
| 19 | 0.389 | 0.456 | 0.512 | 0.529 | 0.575 | 0.616 | 0.652 | 0.693 | 0.721 |
| 20 | 0.378 | 0.444 | 0.499 | 0.516 | 0.561 | 0.602 | 0.637 | 0.679 | 0.706 |
| 21 | 0.369 | 0.433 | 0.487 | 0.503 | 0.549 | 0.589 | 0.624 | 0.665 | 0.693 |
| 22 | 0.360 | 0.423 | 0.476 | 0.492 | 0.537 | 0.576 | 0.611 | 0.652 | 0.680 |
| 23 | 0.352 | 0.413 | 0.466 | 0.482 | 0.526 | 0.565 | 0.599 | 0.640 | 0.668 |
| 24 | 0.344 | 0.404 | 0.456 | 0.472 | 0.515 | 0.554 | 0.588 | 0.629 | 0.656 |
| 25 | 0.337 | 0.396 | 0.447 | 0.462 | 0.505 | 0.543 | 0.578 | 0.618 | 0.645 |
| 26 | 0.330 | 0.388 | 0.439 | 0.453 | 0.496 | 0.534 | 0.567 | 0.607 | 0.634 |
| 27 | 0.323 | 0.381 | 0.430 | 0.445 | 0.487 | 0.524 | 0.558 | 0.597 | 0.624 |
| 28 | 0.317 | 0.374 | 0.423 | 0.437 | 0.479 | 0.515 | 0.549 | 0.588 | 0.615 |
| 29 | 0.311 | 0.367 | 0.415 | 0.430 | 0.471 | 0.507 | 0.540 | 0.579 | 0.606 |
| 30 | 0.306 | 0.361 | 0.409 | 0.423 | 0.463 | 0.499 | 0.532 | 0.570 | 0.597 |
| 35 | 0.283 | 0.334 | 0.378 | 0.392 | 0.430 | 0.464 | 0.495 | 0.532 | 0.558 |
| 40 | 0.264 | 0.312 | 0.354 | 0.367 | 0.403 | 0.435 | 0.465 | 0.501 | 0.525 |
| 45 | 0.248 | 0.294 | 0.334 | 0.346 | 0.380 | 0.411 | 0.440 | 0.474 | 0.498 |
| 50 | 0.235 | 0.279 | 0.317 | 0.328 | 0.361 | 0.391 | 0.418 | 0.451 | 0.474 |
| 60 | 0.214 | 0.254 | 0.289 | 0.300 | 0.330 | 0.358 | 0.383 | 0.414 | 0.436 |
| 70 | 0.198 | 0.235 | 0.268 | 0.278 | 0.306 | 0.332 | 0.356 | 0.385 | 0.405 |
| 80 | 0.185 | 0.220 | 0.251 | 0.260 | 0.286 | 0.311 | 0.334 | 0.361 | 0.380 |
| 90 | 0.174 | 0.207 | 0.236 | 0.245 | 0.270 | 0.293 | 0.315 | 0.341 | 0.360 |
| 100 | 0.165 | 0.197 | 0.224 | 0.232 | 0.256 | 0.279 | 0.299 | 0.324 | 0.342 |
| 120 | 0.151 | 0.179 | 0.205 | 0.212 | 0.234 | 0.255 | 0.274 | 0.297 | 0.313 |
| 150 | 0.135 | 0.160 | 0.183 | 0.190 | 0.210 | 0.228 | 0.245 | 0.266 | 0.281 |
| 200 | 0.117 | 0.139 | 0.158 | 0.164 | 0.182 | 0.198 | 0.213 | 0.231 | 0.244 |
| 250 | 0.104 | 0.124 | 0.142 | 0.147 | 0.163 | 0.177 | 0.190 | 0.207 | 0.219 |
| 300 | 0.095 | 0.113 | 0.129 | 0.134 | 0.149 | 0.162 | 0.174 | 0.189 | 0.200 |
| 400 | 0.082 | 0.098 | 0.112 | 0.116 | 0.129 | 0.140 | 0.151 | 0.164 | 0.173 |
| 500 | 0.074 | 0.088 | 0.100 | 0.104 | 0.115 | 0.125 | 0.135 | 0.147 | 0.155 |

# Bibliography


[1] Dosher, B. A. (1998). The response-window regression method: Some problematic assumptions: Comment on Draine and Greenwald (1998). *Journal of Experimental Psychology: General, 127,* 311–317.

[2] Greenwald, A. G., Draine, S. C., & Abrams, R. L. (1996). Three cognitive markers of unconscious semantic activation. *Science, 273,* 1699–1702.

[3] Greenwald, A. G., Klinger, M. R., & Schuh, E. S. (1995). Activation by marginally perceptible ("Subliminal") stimuli: Dissociation of unconscious from conscious cognition. *Journal of Experimental Psychology: General, 124,* 22–42.

[4] Miller, J. O. (2000). Measurement error in subliminal perception experiments: Simulation analyses of two regression methods. *Journal of Experimental Psychology: Human Perception and Performance, 26,* 1461-1477.

[5] Ware, M. E., & Chastain, J. D. (1989). Computer-assisted statistical analysis: A teaching innovation? *Teaching of Psychology, 16,* 222–227.

[6] Winer, B. J., Brown, D. R., & Michels, K. M. (1991). *Statistical principles in experimental design.* 3rd edition. New York: McGraw-Hill.