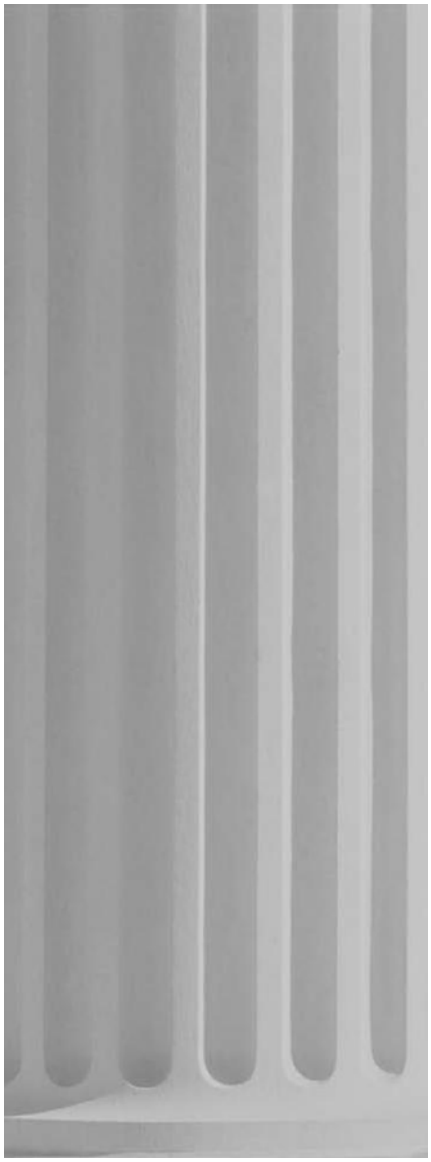
CHAPTER 14

Analysis for Research Design

KEY TERMS

alpha level
analysis of variance (ANOVA)
bell curve
causal
confidence intervals
control group
covariates
Cronbach's alpha
degrees of freedom (*df*)
descriptive statistics
dummy variable
effect size
error term
general linear model (GLM)
hypothesis
inferential statistics
interaction effect
least squares
linear model
main effects
mean
measurement error
model specification
nonequivalent-groups design (NEGD)
null case
null hypothesis
p value
propensity score analysis
quasi-experimental research design
random assignment
randomized block designs (RD)
regression analysis
regression line
regression point displacement design (RPD)
relationship
reliability
residuals
sampling distribution
selection threat
slope
standard deviation
standard error
standard error of the difference
true score theory
t-test
t-value
variables
variance



OUTLINE

causal
Pertaining to a cause-effect relationship.

relationship
Refers to the correspondence between two variables.

descriptive statistics
Statistics used to describe the basic features of the data in a study.

hypothesis
A specific statement of prediction.

random assignment
Process of assigning your sample into two or more subgroups by chance. Procedures for random assignment can vary from flipping a coin to using a table of random numbers to using the random number capability built into a computer.

inferential statistics
Statistical analyses used to reach conclusions that extend beyond the immediate data alone.

quasi-experimental research design
Research designs that have several of the key features of randomized experimental designs, such as pre-post measurement and treatment-control group comparisons, but lack random assignment to a treatment group.

The heart of the quantitative data analysis—the part where you answer the major research questions in a quantitative study—is inextricably linked to the research design. Especially in **causal** research, the research design frames the entire endeavor, specifying how the measures and participants are brought together. So, it shouldn't surprise you that the research design also frames the data analysis, determining the type of analysis that you can and cannot do.

This chapter describes the **relationship** between design and analysis. I begin with inferential statistics, which differ from **descriptive statistics** in that they are explicitly constructed to address a research question or **hypothesis**. I then present the general linear model (GLM). Even though each specific design has its own design quirks and idiosyncrasies, things aren't as confusing or distinct as they may at first seem. The GLM underlies all of the analyses presented here, so if you get a good understanding of what that's all about, the rest should be a little easier to handle. (Note that I said a little easier. I didn't say it was going to be easy.) I then move on to consider the basic randomized experimental designs, starting with the simplest—the two-group, posttest-only experiment—and moving to more complex designs. Finally, I take you into the world of quasi-experimental analysis where the quasi nature of the design leads to all types of analytic problems (some of which may even make you queasy). You'll learn that you pay a price, analytically speaking, when you move away from **random assignment**. By the time you're through with all of this, you'll have a pretty firm grasp on how analysis is crafted to your research design and about the perils of applying the analysis that seems most obvious to the wrong design structure.

14-1 Inferential Statistics

Inferential statistics is the process of trying to reach conclusions that extend beyond the immediate data. You are trying to use the data as the basis for drawing broader inferences (thus, the name) rather than just describing the data. For instance, you use inferential statistics to try to infer from the sample data what the population might think. Or, you use inferential statistics to make judgments about the probability that an observed difference between groups is a dependable one or one that might have happened by chance in your study. Thus, you use inferential statistics to make inferences from your data to general conditions; you use descriptive statistics simply to describe what's going on in the data.

In this chapter, I concentrate on inferential statistics, which are useful in experimental and **quasi-experimental research design** or in program-outcome evaluation. To understand inferential statistics there are two issues I need you to consider, one somewhat general and theoretical and the other more concrete and

methodological. You should also know that this presentation differs from what you typically see in a statistics text. Here we are taking a top-down approach, that is, going from the big picture to the specific test situation. In most statistics textbooks, you start with the relatively simple analyses such as the ***t*-test** and work your way up to the more complex statistical modeling approaches. Imagine that instead of teaching you how to analyze data, I was trying to teach you how to make beer. If that was the case (pun intended), I would start with the big picture about what beer is made from and how you go from water, malt, hops, and yeast to a particular kind of beer. So here we start with the most general model and then consider some specific kinds of analysis that are derived from it. For the following discussion, you may actually find that a beer or two will help you along.

***t*-test**
A statistical test of the difference between the means of two groups, often a program and comparison group. The *t*-test is the simplest variation of the one-way analysis of variance (ANOVA).

First, virtually all the major inferential statistics come from a general family of statistical models known as the general linear model (GLM). You can't get much more general than the GLM. This includes the *t*-test, analysis of variance (ANOVA), analysis of covariance (ANCOVA), regression analysis, (all of which are described later in this chapter), and many of the multivariate methods like factor analysis, multidimensional scaling, cluster analysis, discriminant function analysis, and so on. Given the importance of the GLM, it's a good idea for any serious social researcher to become familiar with it. The discussion of the GLM here is elementary and considers only the simplest straight-line model, but it will familiarize you with the idea of a **linear model** and help prepare you for the more complex analyses described in the rest of this chapter.

linear model
Any statistical model that uses equations to estimate lines.

Second, on a more concrete and methodological note, you can't truly understand how the GLM is used to analyze data from research designs unless you learn what a dummy variable is and how it is used. The name doesn't suggest that you are using **variables** that aren't smart or, even worse, that the analyst who uses them is a dummy! Perhaps these variables would be better described as proxy variables. Essentially a dummy variable is one that uses discrete numbers, usually 0 and 1, to represent different groups in your study in the equations of the GLM. The concept of dummy variables is a simple idea that enables some complicated things to happen. For instance, by including a simple dummy variable in a model, you can model two separate lines (one for each treatment group) with a single equation. All this will be clarified here.

variables
Any entity that can take on different values. For instance, age can be considered a variable because age can take on different values for different people at different times.

confidence intervals
Technically, 1-alpha. The confidence interval is the probability of correctly concluding that there is no treatment effect.

effect size
An estimate of the effect of a treatment or program. The effect size is a signal to noise ration where the numerator (top) represents the effect you are trying to assess (e.g., a difference in averages between two groups) and the denominator (bottom) represents the variability or noise in the data.

14-1a Significance Testing

Remember that in Chapter 2 in this text you learned that in most research situations, we do not study the entire population of interest by measuring everyone; instead, we attempt to study a sample and then see what we can infer from the data about the whole population. But how do we decide whether to make those inferences? Surely not every result is as important as every other result. The guidelines for significance testing help us make those decisions, and the more recent (more recently taken seriously, that is) guidelines about **confidence intervals** and **effect size** help us interpret the precision and magnitude of our results. There has been a very important reformation in statistical methods in the past 10 years, including standards for reporting results. This is related to common misunderstanding of what was meant by a "***p* value" and the procedures for **null hypothesis** significance testing generally. Here I'll introduce these ideas, and then I'll expand on the discussion of confidence intervals and effect sizes in Chapter 16.

***p* value**
The estimate of the probability for a test of an hypothesis. Usually the p value is compared to the significance level when testing a hypothesis. If the p-value exceeds the designated significance level the alternative hypothesis is accepted; if it does not, the null hypothesis is accepted.

null hypothesis
The hypothesis that describes the possible outcomes other than the alternative hypothesis. Usually, the null hypothesis predicts there will be no effect of a program or treatment you are studying.

Sir Ronald Fisher remains one of the most important statisticians in history for many reasons, one of which is that his work established the basic guidelines used to determine whether the results of a statistical test were worth paying attention to, or "significant." Sir Ronald suggested the notion that a statistical result could be considered significant if it could be shown that the probability of the result being due to chance was 5 percent or less. For those of you who just connected this to the discussion of statistical power and conclusion validity in Chapter 12, congratulations,

you've experienced a key insight! Yes, this 5% rule, or the "*p* is less than or equal to .05" cutoff is where our actual hypothesis test is typically performed. We compare our observed *p*-value, the estimated probability for our hypothesis, with the alpha-level cutoff criterion. If *p* is less than alpha (which is usually set at .05, a 5% chance of being wrong or having a Type I error), we conclude that oour alternative hypothesis is correct and reject the no-difference null hypothesis. If the *p*-value is greater than the significance level or alpha value, we reject our alternative hypothesis and accept the null hypothesis. The significance level is the level of risk we are willing to accept as the price of our inference from sample to population. But even though that idea might seem straightforward, it has been widely misunderstood. It's no wonder—look how difficult it is to describe! In fact, surveys of even doctoral-level practitioners show that Fisher's guidelines have become so widely misunderstood that some leaders in the field suggested a ban on all significance testing. This might be the equivalent of solving our energy crisis by banning machines. The trade-offs in that solution would obviously be impractical and would have some very serious negative consequences. The American Psychological Association convened a task force to confront this issue and make recommendations (Wilkinson & the Task Force on Statistical Inference, 1999). Thus, instead of getting rid of *p* values, we are now encouraged to supplement them in very sensible ways, with estimates of the precision and importance or our results, formally known as *confidence intervals* and *effect sizes*.

14-1b Confidence Intervals and the Effect Size

sampling distribution
The theoretical distribution of an infinite number of samples of the population of interest in your study.

standard error
The spread of the averages around the average of averages in a sampling distribution.

I now ask you to remember another idea from earlier in the book: the **sampling distribution** (in fact, it might be a good idea to review Chapter 2 and think about external validity as you study the ideas in this chapter). In Chapter 2, you learned that in every sample of data there is some error, reflecting the inevitable inaccuracy that occurs when we observe only some of the population. You also learned that if the amount of error is small, quantified in the **standard error**, then the estimate provided by your statistics is relatively precise. If the standard error is large, then there is a considerable range in which the next estimate from the next sample might fall. Maybe you've just had another insight—that we could actually use the standard error to report on just how big that range is, that is, how precise our estimate is. This is exactly what we do when we define a confidence interval (CI). A 95% CI is defined as the range that encompasses plus or minus two standard errors (i.e., 95% of the area under a normal curve distribution) around a statistic. The term *confidence interval* really means what it sounds like here: We are confident that 95 percent of the time our estimate will be in the interval defined by two times the standard error. So with both the *p* value and CI, we have a pretty good idea of the probability of our result and how it fits our level of risk (significance level) cutoff criterion for testing our hypothesis.

effect size
An estimate of the effect of a treatment or program. The effect size is a signal to noise ration where the numerator (top) represents the effect you are trying to assess (e.g., a difference in averages between two groups) and the denominator (bottom) represents the variability or noise in the data.

But there's one other thing we'd really like to know: How big is the result? This issue is handled by the **effect size** (ES). The ES is essentially another kind of signal-to-noise ratio. It specifically tells us how far the signal-to-noise ratio deviates from zero. If the null hypothesis of no relationship (what Thompson [2006] called the *nil null*) is true, then the signal-to-noise ratio in our data is zero and so is the effect size. Analogously, in sports we want to know who won, but we also want to know by how much and whether the game was preseason, regular season or playoffs.

Although careful reading of the history of statistics shows us that statisticians have always been concerned about the arbitrariness of significance testing as well as the context of research findings (Wainer & Robinson, 2003), practice has tended toward a very narrow use of the concept of what is to be considered significant. In the extreme, picture an "analyst" scanning scan through printouts of large numbers of tests until finding something with a *p* value of .05 or less and shouting, "Bingo!" In fact, this practice has more to do with the random results of a Bingo game than it does with science. If you think about the signal-to-noise concept again, you can see why.

The signal-to-noise ratio has two components: a numerator and a denominator. The numerator is the signal which we can refer to as the effect. But the relative size of the effect is dependent on the denominator, or the noise level in the estimate. Since you can always diminish the noise level by increasing the sample size, your estimate of the effect and the *p* value associated with it will reflect the sample size. This is why we do power analysis, but it is also why we have come to realize the importance of reporting effect sizes—so we can put the effect in perspective. However, there is another aspect to putting the effect size into context, which has to do with the practical or clinical significance of an effect. This very important topic will be covered in Chapter 16.

14-2 General Linear Model

The **general linear model (GLM)** underlies most of the statistical analyses that are used in applied and social research. It is the foundation for the *t*-test, ANOVA, ANCOVA, regression analysis, and many of the multivariate methods, including factor analysis, cluster analysis, multidimensional scaling, discriminant function analysis, canonical correlation, and others. Because of its generality, the model is important for students of social research. Although a deep understanding of the GLM requires some advanced statistical training, I will attempt here to introduce the concept and provide a nonstatistical description.

general linear model (GLM)
A system of equations that is used as the mathematical framework for most of the statistical analyses used in applied social research.

14-2a The Two-Variable Linear Model

The easiest point of entry into understanding the GLM is with the two-variable case. Figure 14–1a shows a bivariate plot of two variables. These may be any two continuous variables, but in the discussion that follows, think of them as a pretest (on the *x*-axis) and a posttest (on the *y*-axis). Each dot on the plot represents the pretest and posttest score for an individual. The pattern clearly shows a positive relationship because, in general, people with higher pretest scores also have higher posttests and vice versa.

The goal in data analysis is to summarize or describe accurately what is happening in the data. The bivariate plot shows the data. How might you best summarize this data? Figure 14–1b shows that a straight line through the cloud of data points would effectively describe the pattern in the bivariate plot. Although the line does not perfectly describe any specific point (because no point falls precisely on the line), it does accurately describe the pattern in the data. When you fit a line to data, you are using a linear model. The term *linear* refers to the fact that you are fitting a line. The term *model* refers to the equation that summarizes the line that you fit. A line like the one shown in Figure 14–1c is often referred to as a **regression line** (a description of the relationship between two variables) and the analysis that produces it is often called *regression analysis*.

regression line
A line that describes the relationship between two or more variables.

Figure 14–1c shows the equation for a straight line. You may remember this equation from your high school algebra classes, where it is often stated in the form $y = mx + b$. This equation has the following components:

y = the *y*-axis variable, the outcome or posttest
x = the *x*-axis variable, the pretest
b_0 = the intercept (value of y when $x = 0$)
b_1 = the **slope** of the line

slope
The change in *y* for a change in *x* of one unit.

The slope of the line is the change in the posttest given in pretest units. As mentioned previously, this equation does not perfectly fit the cloud of points in Figure 14–1a. If it did, every point would fall on the line. You need one more component to describe the way this line fits to the bivariate plot.

Figure 14–1d shows the equation for the two-variable or bivariate linear model, in this example for a pretest-posttest situation. The component added to the

FIGURE 14–1a A bivariate plot

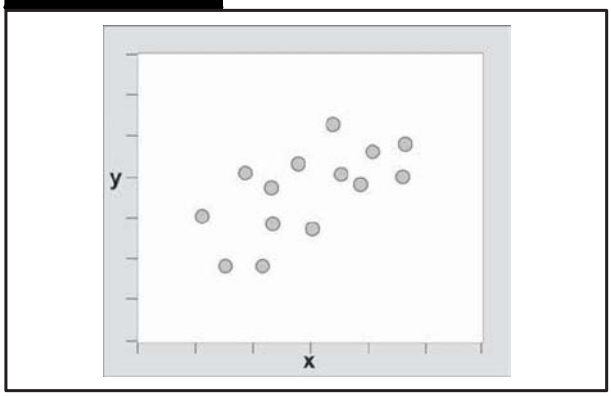


FIGURE 14–1b A straight-line summary of the data

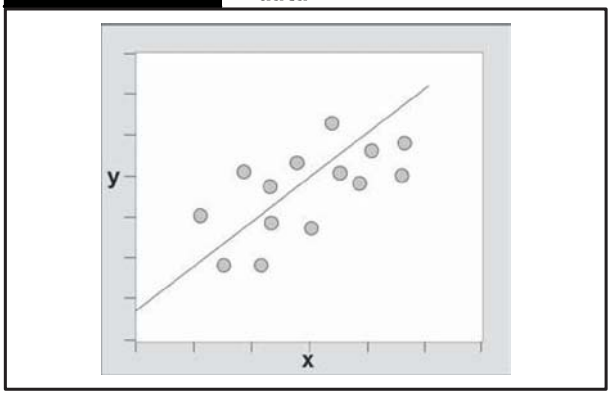


FIGURE 14–1c The straight-line model

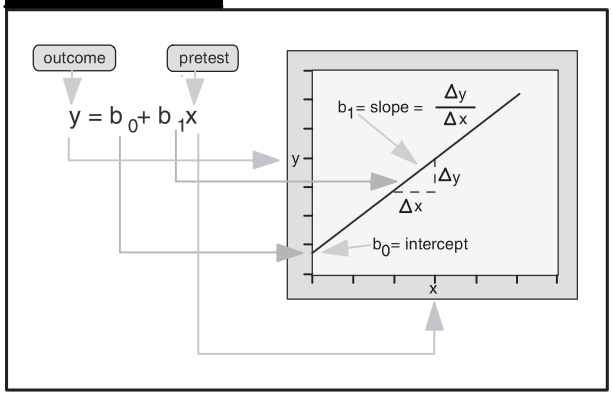


FIGURE 14–1d The two-variable linear model

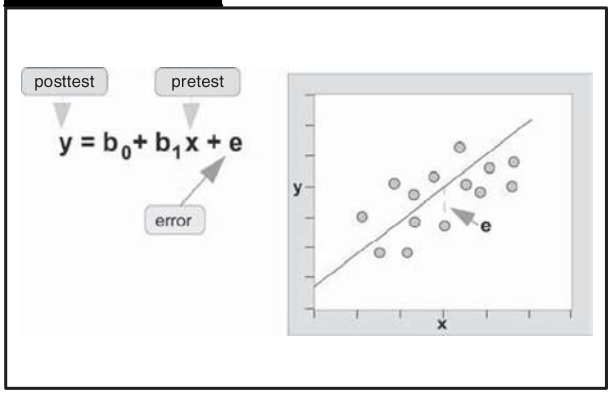
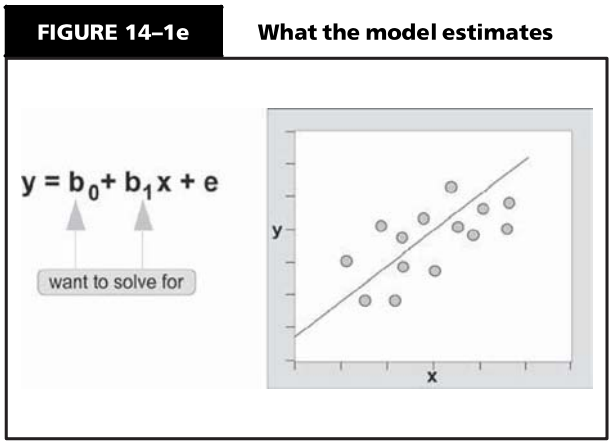


FIGURE 14–1e What the model estimates



error term
A term in a regression equation that captures the degree to which the line is in error (that is, the residual) in describing each point.

mean
A description of the central tendency in which you add all the values and divide by the number of values.

equation in Figure 14–1d is an **error term** that describes the vertical distance from the straight line to each point. This component is called *error* because it is the degree to which the line is in error in describing each point. When you fit the two-variable linear model to your data, you have an x and y score for each person in your study. You input these value pairs into a computer program. The program estimates the b_0 and b_1 values as indicated in Figure 14–1e. You will actually get two numbers back that are estimates of those two values.

You can think of the two-variable regression line like any other descriptive statistic; it simply describes the relationship between two variables much as a **mean**

describes the central tendency of a single variable. Just as the mean does not accurately represent every value in a distribution, the regression line does not accurately represent every value in the bivariate distribution. You use these summaries because they show the general patterns in your data and allow you to describe these patterns in more concise ways than showing the entire distribution would allow.

14-2b Extending the General Linear Model to the General Case

With this brief introduction to the two-variable case in mind, let's extend the model to its most general case—the GLM. Essentially the GLM looks the same as the two-variable model shown in Figure 14–1e; it is an equation. The big difference is that each of the four terms in the GLM can represent a *set* of variables instead of representing only a single variable. So, the general linear model can be written as follows:

$$y = b_0 + bx + e$$

where

y = a set of outcome variables
x = a set of pre-program variables or **covariates**
b_0 = the set of intercepts (value of each y when each $x = 0$)
b = a set of coefficients, one each for each x

covariates
Variables you adjust for in your study.

This model allows you to include an enormous amount of information. In an experimental or quasi-experimental study, you would represent the program or treatment with one or more dummy-coded variables, each represented in the equation as an additional x-value. (Although the convention is to use the symbol Z to indicate that the variable is a dummy-coded x.) If your study has multiple outcome variables, you can include them as a set of y-values. If you have multiple pretests, you can include them as a set of x-values. For each x-value (and each Z-value), you estimate a b-value that represents an x,y relationship. The estimates of these b-values and the statistical testing of these estimates is what enables you to test specific research hypotheses about relationships between variables or differences between groups.

The GLM allows you to summarize a variety of research outcomes. The major problem for the researcher who uses the GLM is **model specification**, which the user must enact to specify the exact equation that best summarizes the data for a study. If the model is misspecified, the estimates of the coefficients (the b-values) are likely to be biased (wrong) and the resulting equation will not describe the data accurately. In complex situations, this model specification problem can be a serious and difficult one (see, for example, the discussion of model specification in the statistical analysis of the regression-discontinuity design later in this chapter).

model specification
The process of stating the equation that you believe best summarizes the data for a study.

The GLM is one of the most important tools in the statistical analysis of data. It represents a major achievement in the advancement of social research in the 20th century. In the latter part of the 20th century, a convergence of computing power and statistical methods produced new techniques for testing complex models. These newer techniques include structural equation modeling (SEM), which allows a researcher to test the goodness of fit of a theoretical model to a multivariate system of equations that includes both observed and unobserved or latent variables and allows for correlations or relationships among error terms. As Kline (2005) noted, SEM is actually a "family" of methods. His excellent text provides a very readable overview with many examples of the kinds of problems that can be addressed with SEM. Another technique that has extended the tool kit of researchers is hierarchical linear modeling (HLM). This method is similar to SEM in the sense of testing an overall model, but it is particularly well suited to the common situation in which people (and their data) are organized in hierarchical ways, such as

in schools (by grade, school, school district etc.). HLM allows researchers to examine the direct influence of such variables as well as their interactions.

14-2c Dummy Variables

dummy variable
A variable that uses discrete numbers, usually 0 and 1, to represent different groups in your study in the equations of the GLM.

control group
A group, comparable to the program group, that did not receive the program.

A **dummy variable** is a numerical variable used in regression analysis to represent subgroups of the sample in your study. It is not a variable used by dummies. In fact, you have to be pretty smart to figure out how to use dummy variables. In research design, a dummy variable is typically used to distinguish different treatment groups. In the simplest case, you would use a 0,1 dummy variable, where a person is given a value of 0 if placed in the **control group** or a 1 if in the treated group.

Dummy variables are useful because they enable you to use a single regression equation to represent multiple groups. This means that you don't need to write out separate equation models for each subgroup. The dummy variables act like *switches* that turn various parameters on and off in an equation. Another advantage of a 0,1 dummy-coded variable is that even though it is a nominal-level variable, you can treat it statistically like an interval-level variable. (If this made no sense to you, you probably should refresh your memory on levels of measurement covered in Section 3-3, Levels of Measurement.) For instance, if you take an average of a 0,1 variable, the result is meaningful—the proportion or percentage of 1s in the distribution.

To illustrate dummy variables, consider the simple regression model for a posttest-only two-group randomized experiment shown in Figure 14–2. This model is mathematically identical to conducting a *t*-test on the posttest means for two groups or conducting a one-way ANOVA (as described later in this chapter). The key term in the model is β_1, the estimate of the difference between the groups. To see how dummy variables work, I'll use this simple model to show you how dummy variables can be used to pull out the separate subequations for each subgroup. Then I'll show how to estimate the difference between the subgroups by subtracting their respective equations. You'll see that you can pack an enormous amount of information into a single equation using dummy variables. All I want to show you here is that (β_1) is the difference between the treatment and control groups.

To see this, the first step is to compute what the equation would be for each of the two groups separately (Figure 14–3). For the control group, $Z = 0$. When you substitute that into the equation, and recognize that by assumption the error term averages to 0, you find that the predicted value for the control group is β_0, the intercept. Now, to figure out the treatment-group line, you substitute the value of 1 for Z, again recognizing that by assumption, the error term averages to 0. The equation for the treatment group indicates that the treatment group value is the sum of the two beta values.

FIGURE 14–2 **Use of a dummy variable in a regression equation**

$$y_i = \beta_0 + \beta_1 Z_i + e_i$$

where:
y_i = outcome score for the i^{th} unit
β_0 = coefficient for the *intercept*
β_1 = coefficient for the *slope*
Z_i = 1 if i^{th} unit is in the treatment group
0 if i^{th} unit is in the control group
e_i = residual for the i^{th} unit

FIGURE 14–3 **Using a dummy variable to create separate equations for each dummy variable value**

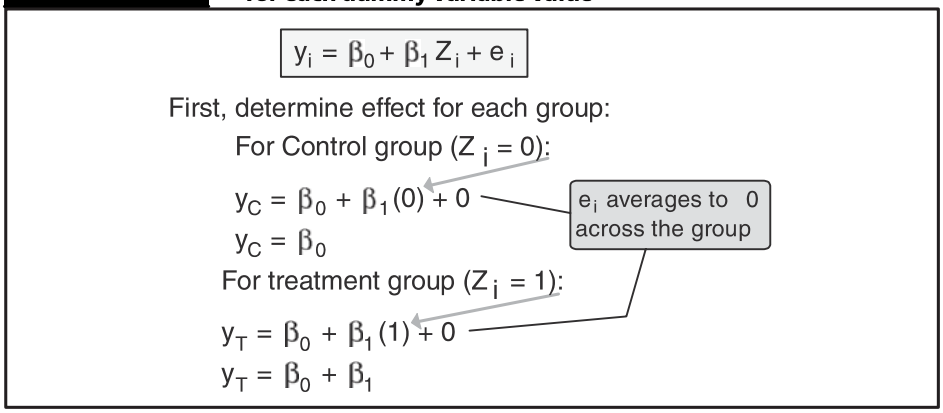
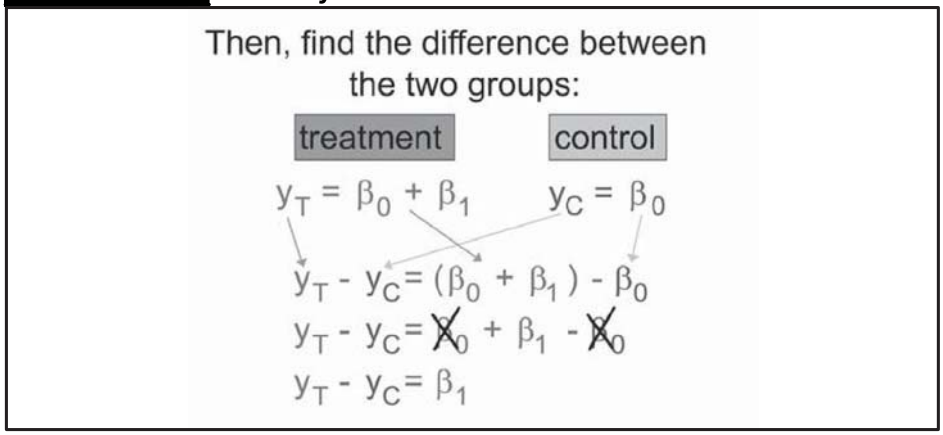


FIGURE 14–4 **Determine the difference between two groups by subtracting the equations generated through their dummy variables**



Now you're ready to move on to the second step—computing the difference between the groups. How do you determine that? Well, the difference must be the difference between the equations for the two groups that you worked out previously. In other words, to find the difference between the groups, you find the difference between the equations for the two groups! It should be obvious from Figure 14–4 that the difference is β_1. Think about what this means. The difference between the groups is β_1. Okay, one more time just for the sheer heck of it: The difference between the groups in this model is β_1 in the equation at the top of Figure 14–3!

Whenever you have a regression model with dummy variables, you can always see how the variables are being used to represent multiple subgroup equations by following the two steps described in Figures 14–3 and 14–4 as follows:

- Create separate equations for each subgroup by substituting the dummy values (as in Figure 14–3).
- Find the difference between groups by finding the difference between their equations (as in Figure 14–4).

14-3 Experimental Analysis

I turn now to the discussion of the experimental designs and how they are analyzed. Perhaps one of the simplest inferential tests is used when you want to compare the average performance of two groups on a single measure to see whether there is a

difference. This simple two-group, posttest-only randomized experiment is usually analyzed with the simple *t*-test, which is actually just the simplest variation of the one-way ANOVA. You might want to know whether eighth-grade boys and girls differ in math test scores or whether a program group differs on the outcome measure from a control group. The factorial experimental designs are usually analyzed with the ANOVA model. **Randomized block designs (RD)** use a special form of the ANOVA-blocking model that uses dummy-coded variables to represent the blocks. The analysis of covariance experimental design uses, not surprisingly, the analysis of covariance (ANCOVA) statistical model.

randomized block designs (RD)
Experimental designs in which the sample is grouped into relatively homogeneous subgroups or blocks within which your experiment is replicated. This procedure reduces noise or variance in the data.

14-3a The *t*-Test

To analyze the two-group, posttest-only randomized experimental design you need an analysis that meets the following requirements:

- Has two groups
- Uses a post-only measure
- Has two distributions (measures), each with an average and variation
- Assesses treatment effect = statistical (non-chance) difference between the groups

The *t*-test fits the bill perfectly. The *t*-test assesses whether the means of two groups are *statistically* different from each other. Why is it called the *t*-test? Because when the statistician who invented this analysis first wrote out the formula, he used the letter "*t*" to symbolize the value that describes the difference between the groups. Why? Beats me. You remember the formula for the straight line from your high school algebra? You know, the one that goes $y = mx + b$? Well, using the name *t*-test is like calling that formula the *y*-formula. Maybe the statisticians decided they would come up with more interesting names later. Maybe they were in the same fix as the astronomers who had so many stars to name they just assigned temporary numbers until someone noteworthy enough came along. Whatever the reason, don't lose any sleep over it. The *t*-test is just a name and, as the bard says, what's in a name?

Before you can proceed to the analysis itself, it is useful to understand what we mean by the term "difference" in the question, "Is there a difference between the groups?" Each group can be represented by a bell-shaped curve that describes the group's distribution on a single variable. You can think of the **bell curve** as a smoothed histogram or bar graph describing the frequency of each possible measurement response.

bell curve
Smoothed histogram or bar graph describing the expected frequency for each value of a variable. The name comes from the fact that such a distribution often has the shape of a bell.

Figure 14–5 shows the distributions for the treated (dotted line) and control (solid line) groups in a study. Actually, the figure shows the idealized or smoothed

FIGURE 14–5 **Idealized distributions for treated and control group posttest values**

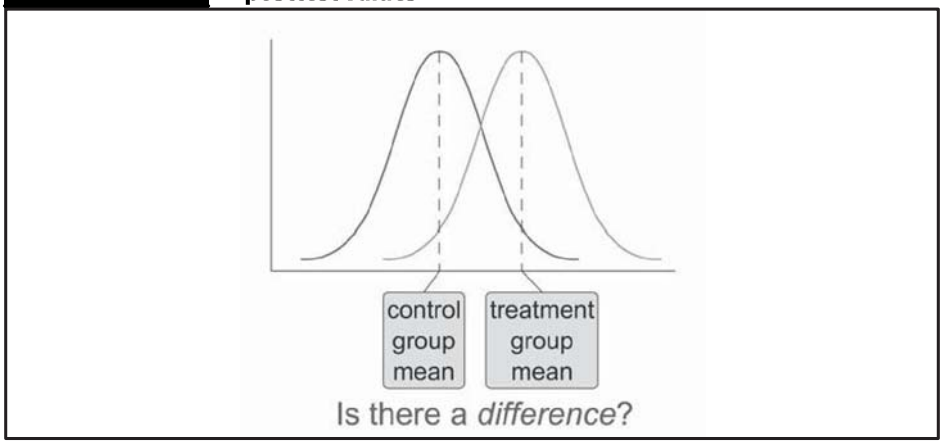
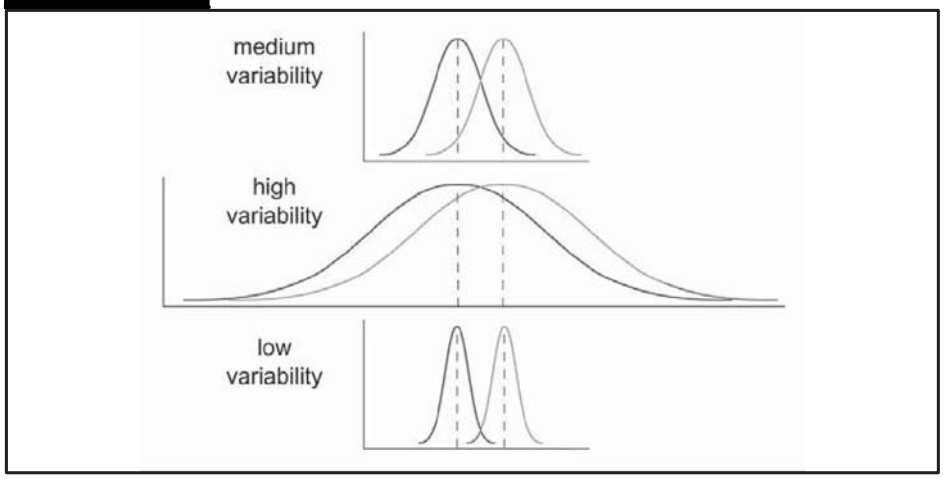


FIGURE 14–6 **Three scenarios for differences between means**



distribution—the actual distribution would usually be depicted with a histogram or bar graph. The figure indicates where the control and treatment group means are located. The question the *t*-test addresses is whether the means are statistically different.

What does it mean to say that the averages for two groups are statistically different? Consider the three situations shown in Figure 14–6. The first thing to notice about the three situations is that *the difference between the means is the same in all three.* But, you should also notice that the three situations don't look the same; they tell different stories. The top example shows a case with moderate variability of scores within each group. The second situation shows the high-variability case. The third shows the case with low variability. Clearly, you would conclude that the two groups appear most different or distinct in the bottom or low-variability case. Why? Because there is relatively little overlap between the two bell-shaped curves. In the high-variability case, the group difference appears least striking (even though the difference between groups is identical) because the two bell-shaped distributions overlap so much.

This leads to an important conclusion: When you are looking at the differences between scores for two groups, you have to judge the difference between their means relative to the spread or variability of their scores. The *t*-test does just this.

Statistical Analysis of the *t*-Test So how does the *t*-test work? The formula for the *t*-test is a ratio. The top part of the ratio is the difference between the two means or averages. The bottom part is a measure of the variability or dispersion of the scores. This formula is essentially another example of the signal-to-noise metaphor in research; the difference between the means is the signal that, in this case, you think your program or treatment introduced into the data; the bottom part of the formula is a measure of variability that is essentially noise that might make it harder to see the group difference. The ratio that you compute is called a ***t*-value** and describes the difference between the groups relative to the variability of the scores in the groups. Figure 14–7a shows the formula for the *t*-test and how the numerator and denominator are related to the distributions.

The top part of the formula is easy to compute—just find the difference between the means. The bottom part is called the **standard error of the difference**. To compute it, take the **variance** (see Chapter 12) for each group and divide it by the number of people in that group. You add these two values and then take their square root. The specific formula is given in Figure 14–7b. Remember, that the variance is simply the square of the **standard deviation**. The final formula for the *t*-test is shown in Figure 14–7c.

***t*-value**
The estimate of the difference between the groups relative to the variability of the scores in the groups.

standard error of the difference
A statistical estimate of the standard deviation one would obtain from the distribution of an infinite number of estimates of the difference between the means of two groups.

variance
A statistic that describes the variability in the data for a variable. The variance is the spread of the scores around the mean of a distribution. Specifically, the variance is the sum of the squared deviations from the mean divided by the number of observations minus 1.

standard deviation
The spread or variability of the scores around their average in a *single sample.* The standard deviation, often abbreviated SD, is mathematically the square root of the variance. The standard deviation and variance both measure dispersion, but because the standard deviation is measured in the same units as the original measure and the variance is measured in squared units, the standard deviation is usually more directly interpretable and meaningful.

FIGURE 14–7a Formula for the *t*-test

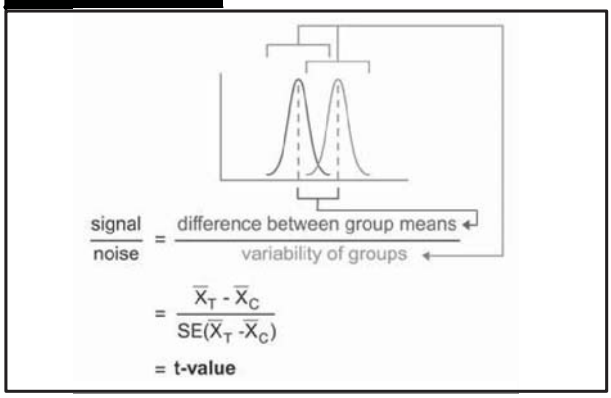


FIGURE 14–7b Formula for the standard error of the difference between the means

$$SE(\bar{X}_T - \bar{X}_C) = \sqrt{\frac{var_T}{n_T} + \frac{var_C}{n_C}}$$

FIGURE 14–7c Formula for the *t*-test

$$t = \frac{\bar{X}_T - \bar{X}_C}{\sqrt{\frac{var_T}{n_T} + \frac{var_C}{n_C}}}$$

alpha level
The significance level. Specifically, alpha is the Type I error, or the probability of concluding that there is a treatment effect when, in reality, there is not.

degrees of freedom (*df*)
A statistical term that is a function of the sample size. In the *t*-test formula, for instance, the df is the number of persons in both groups minus 2.

The *t*-value will be positive if the first mean is larger than the second value and negative if it is smaller. After you compute the *t*-value, you have to look up the probability or *p*-value associated with your *t*-value (many statistical programs automatically provide the *p*-value) in a table of significance to test whether the *t*-ratio is large enough to say that the difference between the groups is not likely to have been a chance finding. To test the significance, you need to set a risk level (called the **alpha level**, as described in Chapter 12). In most social research, the rule of thumb is to set the alpha level at .05. This means that 5 times out of 100, you would find a statistically significant difference between the means even if there were none (meaning by chance). You also need to determine the **degrees of freedom (*df*)** for the test. In the *t*-test, the *df* is the sum of the persons in both groups minus 2. Given the alpha level, the *df*, and the *t*-value, you can look the *t*-value up in a standard table of significance to determine the *p*-value. By comparing the *p*-value with the significance level or alpha you can determine whether the *t*-value is large enough to be significant. If it is, you can conclude that the difference between the means for the two groups is different (even given the variability). Fortunately, statistical computer programs routinely print the significance test results and save you the trouble of looking them up in a table.

You can estimate the treatment effect for the posttest-only randomized experiment in three ways. All three yield mathematically equivalent results, a fancy way of saying that they give you the exact same answer. So why are there three different ones? In large part, these three approaches evolved independently and only after that was it clear that they are essentially three ways to do the same thing. So, what are the three ways? First, you can compute an independent *t*-test as described here. Second, you could compute a one-way ANOVA between two independent groups. Finally, you can use regression analysis to regress the posttest values onto a dummy-coded treatment variable. Of these three, the regression analysis approach is the most general. In fact, I describe the statistical models for all the experimental and quasi-experimental designs in regression-model terms. You just need to be aware that the results from all three methods are identical. Okay, so here's the statistical model for the *t*-test in regression form (Figure 14–8).

Look familiar? It is identical to the formula I showed in Figure 14–2 to introduce dummy variables. Also, you may not realize it (although I hope against hope that you do), but essentially this formula is the equation for a straight line with a random error term (e_i) thrown in. Remember high school algebra? Remember high school? Okay, for those of you with faulty memories, you may recall that the equation for a straight line is often given as follows:

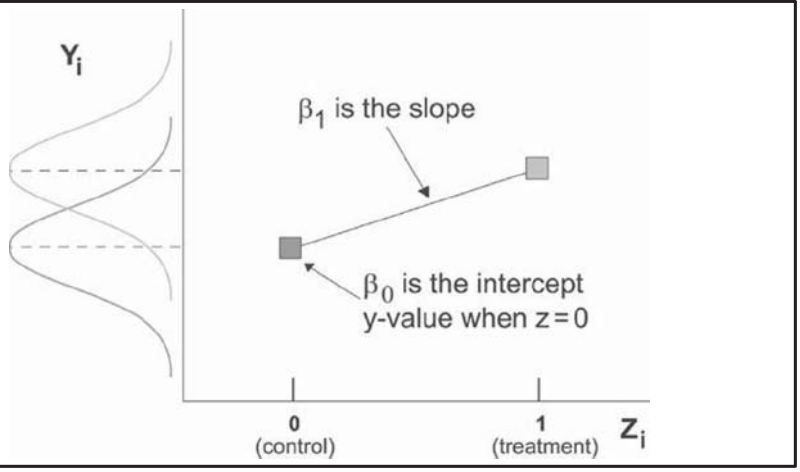
$$y = mx + b$$

FIGURE 14–8 **The regression formula for the *t*-test or the two-group one-way analysis of variance (ANOVA)**

$$y_i = \beta_0 + \beta_1 Z_i + e_i$$

where:
y_i = outcome score for the i^{th} unit
β_0 = coefficient for the *intercept*
β_1 = coefficient for the *slope*
Z_i = 1 if i^{th} unit is in the treatment group
0 if i^{th} unit is in the control group
e_i = residual for the i^{th} unit

FIGURE 14–9 **The Elements of the Equation in Figure 14–8 in Graphic Form**



which, when rearranged can be written as follows:

$$y = b + mx$$

(The complexities of the commutative property make you nervous? If this gets too tricky, you may need to stop for a break. Have something to eat, make some coffee, or take the poor dog out for a walk.) Now you should see that in the statistical model y_i is the same as y in the straight line formula, β_0 is the same as b, β_1 is the same as m, and Z_i is the same as x. In other words, in the statistical formula, β_0 is the intercept and β_1 is the slope (Figure 14–9).

It is critical that you understand that the slope, β_1, is the same thing as the posttest difference between the means for the two groups. How can a slope be a difference between means? To see this, you have to look at a graph of what's going on, which I provide for you in Figure 14–9. The graph shows the posttest on the vertical axis. This is exactly the same as the two bell-shaped curves shown in Figures 14–5 and 14–6 except that here they're turned on their sides and are graphed on the vertical dimension. On the horizontal axis, the Z variable is plotted. This variable has only two possible values: a 0 if the person is in the control group or a 1 if the person is in the program group. This kind of variable is a dummy variable because it is a stand-in variable that represents the program or treatment conditions with its two values (see the discussion of dummy variables earlier in this chapter). The two points in the graph indicate the average posttest value for the control ($Z = 0$) and treated ($Z = 1$) cases. The line that connects the two dots is included only for visual enhancement purposes; because there are no Z values between 0 and 1, there can

be no values plotted where the line is. Nevertheless, you can meaningfully speak about the slope of this line—the line that would connect the posttest means for the two values of *Z*. Do you remember the definition of slope? (Here we go again, back to high school!) The *slope* is the change in *y* over the change in *x* (or, in this case, *Z*). Remember, the change in *Z* between the groups is always equal to 1 (that is, $1 - 0 = 1$). Therefore, the slope of the line must be equal to the difference between the average *y*-values for the two groups. That's what I set out to show (reread the first sentence of this paragraph). β_1 is the same value that you would get if you subtracted the two means from each other. (In this case, because the treatment group equals 1, you are subtracting the control group out of the treatment group value. A positive value implies that the treatment-group mean is higher than the control-group mean; a negative value means it's lower.)

But remember, at the beginning of this discussion, I pointed out that just knowing the difference between the means was not good enough for estimating the treatment effect because it doesn't take into account the variability or spread of the scores. So how do you do that here? Every regression-analysis program will give, in addition to the beta values, a report on whether each beta value is statistically significant. They report a *t*-value that tests whether the beta value differs from zero. It turns out that the *t*-value for the β_1 coefficient is the exact same number that you would get if you did a *t*-test for independent groups. And, it's the same as the square root of the *F*-value in the two-group one-way ANOVA (because $t^2 = F$).

Here's a few conclusions from all this:

- The *t*-test, one-way ANOVA, and regression analysis all yield the *same* results in this case.
- The regression-analysis method utilizes a dummy variable (*Z*) for treatment.
- Regression analysis is the most *general* model of the three.

14-3b Factorial Design Analysis

Now that you have some understanding of the GLM and dummy variables, I can present the models for other experimental designs rather easily. Figure 14–10 shows the regression model for a simple 2 × 2 factorial design.

In this design, you have one factor for time in instruction (1 hour/week versus 4 hours/week) and one factor for setting (in-class or pull-out). The model uses a dummy variable (represented by a *Z*) for each factor. In two-way factorial designs like this, you have two **main effects** and one **interaction effect**. In this model, the

main effects
An outcome that shows consistent differences between all levels of a factor.

interaction effect
An effect that occurs when differences on one factor depend on which level you are on another factor.

FIGURE 14–10 **Regression model for a 2 × 2 factorial design**

$$y_i = \beta_0 + \beta_1 Z_{1i} + \beta_2 Z_{2i} + \beta_3 Z_{1i} Z_{2i} + e_i$$

where:
y_i = outcome score for the i^{th} unit
β_0 = coefficient for the intercept
β_1 = mean difference on factor 1
β_2 = mean difference on factor 2
β_3 = interaction of factor 1 and factor 2
Z_{1i} = dummy variable for factor 1
(0 = 1 hr/wk, 1 = 4 hrs/wk)
Z_{2i} = dummy variable for factor 2
(0 = in class, 1 = pull-out)
e_i = residual for the i^{th} unit

FIGURE 14–11 **Regression model for a randomized block design**

$$y_i = \beta_0 + \beta_1 Z_{1i} + \beta_2 Z_{2i} + \beta_3 Z_{3i} + \beta_4 Z_{4i} + e_i$$

where:
y_i = outcome score for the i^{th} unit
β_0 = coefficient for the intercept
β_1 = mean difference for treatment
β_2 = blocking coefficient for block 2
β_3 = blocking coefficient for block 3
β_4 = blocking coefficient for block 4
Z_{1i} = dummy variable for treatment (0 = control, 1 = treatment)
Z_{2i} = 1 if block 2, 0 otherwise
Z_{3i} = 1 if block 3, 0 otherwise
Z_{4i} = 1 if block 4, 0 otherwise
e_i = residual for the i^{th} unit

main effects are the statistics associated with the beta values that are adjacent to the *Z* variables. The interaction effect is the statistic associated with β_3 (that is, the *t*-value for this coefficient) because it is adjacent in the formula to the multiplication of (interaction of) the dummy-coded *Z* variables for the two factors. Because there are two dummy-coded variables, and each has two values, you can write out $2 \times 2 = 4$ separate equations from this one general model. (Go ahead, I dare you. If you need to refresh your memory, check back to the discussion of dummy variables presented earlier in this chapter.) You might want to see if you can write out the equations for the four cells. Then, look at some of the differences between the groups. You can also write two equations, one for each *Z* variable. These equations represent the main effect equations. To see the difference between levels of a factor, subtract the equations from each other.

14-3c Randomized Block Analysis

The statistical model for the randomized block design (RD) can also be presented in regression analysis notation. Figure 14–11 shows the model for a case where there are four blocks or homogeneous subgroups.

Notice that a number of dummy variables are used to specify this model. The dummy variable Z_1 represents the treatment group. The dummy variables Z_2, Z_3, and Z_4 indicate blocks 2, 3, and 4, respectively. Analogously, the beta values (β_s) reflect the treatment and blocks 2, 3, and 4. What happened to Block 1 in this model? To see what the equation for the Block 1 comparison group is, fill in your dummy variables and multiply through. In this case, all four *Z*s are equal to 0, and you should see that the intercept (β_0) is the estimate for the Block 1 control group. For the Block 1 treatment group, $Z_1 = 1$ and the estimate is equal to $\beta_0 + \beta_1$. By substituting the appropriate dummy variable switches, you should be able to figure out the equation for any block or treatment group.

The data matrix that is entered into this analysis would consist of five columns and as many rows as you have participants: the posttest data and one column of 0s or 1s for each of the four dummy variables.

14-3d Analysis of Covariance

The statistical model for the ANCOVA, which estimates the difference between the groups on the posttest after adjusting for differences on the pretest, can also be

FIGURE 14–12 **Regression model for the ANCOVA**

$$y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + e_i$$

where:
y_i = outcome score for the i^{th} unit
β_0 = coefficient for the intercept
β_1 = pretest coefficient
β_2 = mean difference for treatment
Z_i = dummy variable for treatment (0 = control, 1 = treatment)
e_i = residual for the i^{th} unit

given in regression analysis notation. The model shown in Figure 14–12 is for a case where there is a single covariate, a treated group, and a control group.

The dummy variable Z_i represents the treatment group. The beta values(s) are the parameters being estimated. The value β_0 represents the intercept. In this model, it is the predicted posttest value for the control group for a given X value (and, when $X = 0$, it is the intercept for the control-group regression line). Why? Because a control group case has a $Z = 0$, and since the Z variable is multiplied with β_2, that whole term would drop out.

The data matrix that is entered into this analysis would consist of three columns—the posttest data, one column of 0s or 1s to indicate which treatment group the participant is in, and the covariate score—and as many rows as you have participants.

This model assumes that the data in the two groups are well described by straight lines that have the same slope. If this does not appear to be the case, you have to modify the model appropriately. How do you do that? Well, I'll tell you the short answer, but for the complete one you need to take an advanced statistics course. The short answer is that you add the term $b_i X_i Z_i$ to the model in Figure 14–12. If you've been following along, you should be able to create the separate equations for the different values of the dummy variable Z_i and convince yourself that the addition of this term will allow for the two groups to have different slopes.

14-4 Quasi-Experimental Analysis

nonequivalent-groups design (NEGD)
A pre-post two-group quasi-experimental design structured like a pretest-posttest randomized experiment, but lacking random assignment to group.

measurement error
Any influence on an observed score not related to what you are attempting to measure.

regression point displacement design (RPD)
A pre-post quasi experimental research design where the treatment is given to only one unit in the sample, with all remaining units acting as controls. This design is particularly useful to study the effects of community-level interventions, where outcome data is routinely collected at the community level.

The quasi-experimental designs differ from the experimental ones in that they don't use random assignment to assign units (people) to program groups. The lack of random assignment in these designs tends to complicate their analysis considerably. For example, to analyze the **nonequivalent-groups design (NEGD)**, you have to adjust the pretest scores for **measurement error**. One easily understood method of doing this is called a *reliability-corrected analysis of covariance model* (a more sophisticated approach, not covered here, is called *propensity score analysis*). In the RD design, you need to be especially concerned about curvilinearity and model misspecification. Consequently, I recommend a conservative analysis approach based on polynomial regression that starts by overfitting the likely true function and then reduces the model based on the results. The **regression point displacement design (RPD)** has only a single treated unit. Nevertheless, the analysis of the RPD design is based directly on the traditional ANCOVA model.

In all fairness, I have to warn you that this section is not for the faint-of-heart. The experimental designs discussed previously have fairly straightforward analysis models. You'll see here that you pay a price for not using random assignment like they do; the analyses are considerably more complex.

FIGURE 14–13 **Design notation for the NEGD**

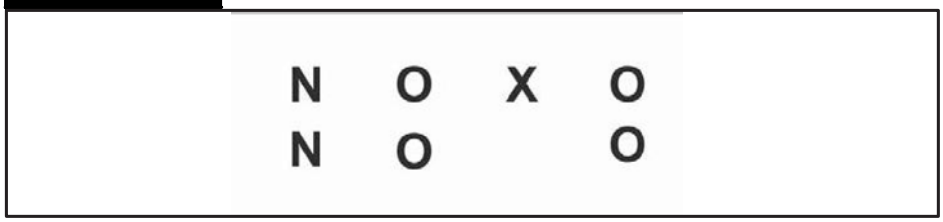
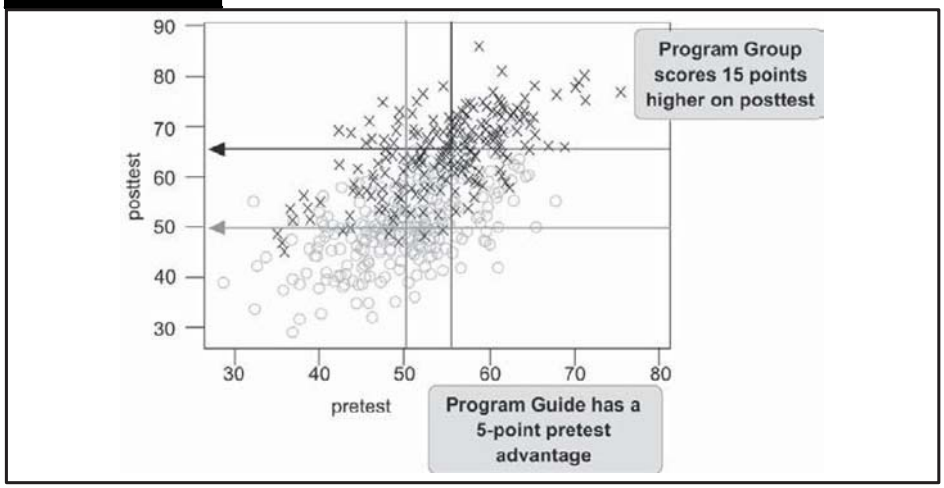


FIGURE 14–14 **Simulated data for the NEGD**



14-4a Nonequivalent-Groups Analysis

The design notation for the NEGD shows two groups—a program and comparison group—and each is measured pre and post (Figure 14–13). The statistical model that you might intuitively expect to be used in this situation would have a pretest variable, posttest variable, and a dummy variable that describes which group the person is in. These three variables would be the input for the statistical analysis.

In this example, assume you are interested in estimating the difference between the groups on the posttest after adjusting for differences on the pretest. This is essentially the ANCOVA model as described in connection with randomized experiments. (Review the Section 9-5, Covariance Designs.) There's only one major problem with this model when used with the NEGD—it doesn't work! Here, I'll tell you the story of why the ANCOVA model fails and one relatively simple way you can adjust it so it works correctly.

A Simulation Example To see what happens when you use the ANCOVA analysis on data from a NEGD, I created a computer simulation to generate hypothetical data. (You can learn how to do these computer simulations yourself at *http://www.socialresearchmethods.net/simul/simul.htm*) used a random number generator to create scores for 500 hypothetical persons, with 250 in the program and 250 in the comparison condition. Because this is a nonequivalent design, I made the groups nonequivalent on the pretest by adding 5 points to each program group person's pretest score. Then, I added 15 points to each program person's posttest score. When I take the initial 5-point advantage into account, I should find a 10-point program effect. The bivariate plot in Figure 14–14 shows the data from this simulation.

I then analyzed the data with the ANCOVA model. Remember that the way I set this up, I should observe approximately a 10-point program effect if the ANCOVA analysis works correctly. The results are presented in Figure 14–15.

FIGURE 14–15 **Results of the original ANCOVA analysis of the simulated data in Figure 14–14**

$$y_i = 18.7 + .626X_i + 11.3Z_i$$

Predictor	Coef	StErr	t	p
Constant	18.714	1.969	9.50	0.000
pretest	0.62600	0.03864	16.20	0.000
Group	11.2818	0.5682	19.85	0.000

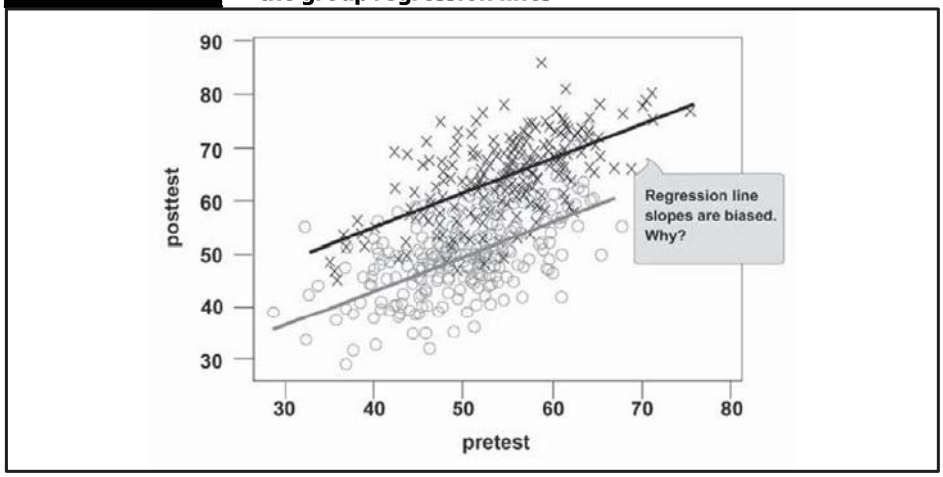
- $CI_{.95(\beta_2 = 10)} = \beta_2 \pm 2SE(\beta_2)$
  $= 11.2818 \pm 2(.5682)$
  $= 11.2818 \pm 1.1364$
- CI = 10.1454 to 12.4182

In this analysis, I created three scores for each person: a pretest score (X), a posttest score (y), and either a 0 or 1 to indicate whether the person was in the program ($Z = 1$) or comparison ($Z = 0$) group. The table in Figure 14–15 shows the equation that the ANCOVA model estimates.

The equation has the three values I put in (X, Y, and Z) and the three coefficients that the program estimates. The key coefficient is the one next to the program variable Z (labeled Group in the table). This coefficient estimates the average difference between the program and comparison groups (because it's the coefficient paired with the dummy variable indicating what group the person is in). The value should be 10 because I put in a 10-point difference. In this analysis, the actual value I got was 11.3 (or 11.2818, to be more precise). Well, that's not too bad, you might say. It's fairly close to the 10-point effect I created. But I need to determine whether the obtained value of 11.2818 is statistically different from the true value of 10. To see whether it is, I have to construct a confidence interval around the estimate and examine the difference between 11.2818 and 10 relative to the variability in the data. Fortunately, the program does this automatically. If you look in the table in Figure 14–15, you'll see that the third line shows the coefficient associated with the difference between the groups, the standard error for that coefficient (an indicator of variability), the t-value, and the probability value. All the t-value shows is that the coefficient of 11.2818 is statistically different from zero. But I want to know whether it is different from the true treatment effect that I put in when constructing the simulation, a value of 10. To determine this, I construct a confidence interval around the t-value, using the standard error. The 95 percent confidence interval is the coefficient plus or minus two times the standard error value. The calculation shows that the 95 percent confidence interval for the 11.2818 coefficient is 10.1454 to 12.4182. Any value falling within this range can't be considered different beyond a 95 percent level from the obtained value of 11.2818. But the true value of 10 points falls *outside* the range. In other words, the estimate of 11.2818 is significantly *different* from the true value. In still other words, the results of this analysis are biased. I got the wrong answer. In this example, the estimate of the program effect is significantly larger than the true program effect (even though the difference between 10 and 11.2818 doesn't seem that much larger, it exceeds chance levels). So, you have a problem when you apply the analysis model that intuitively makes the most sense for the NEGD. To understand why this bias occurs, look a little more deeply at how the statistical analysis works in relation to the NEGD.

The Problem Take a look at Figure 14–16, which shows the regression lines for simulated data presented originally in Figure 14–14. These lines may look like they fit the data well. But in the previous section, I showed that they give a biased estimate—in this case an overestimate—of the treatment effect.

FIGURE 14–16 **Bivariate plot for a nonequivalent-groups design showing the group regression lines**



Why is the ANCOVA analysis biased when used with the NEGD? And, why isn't it biased when used with a pretest-posttest randomized experiment? Actually, several things happen to produce the bias, which is why it's somewhat difficult to understand (and counterintuitive). Here are the two reasons for the bias:

- Pretest measurement error leads to the attenuation or flattening of the slopes in the regression lines.
- Groups are nonequivalent.

The first problem actually also occurs in randomized studies, but it doesn't lead to biased treatment effects because the groups are equivalent (at least probabilistically). The combination of both these conditions causes the problem. And, understanding the problem is what leads us to a simple solution in this case.

Regression and Measurement Error I begin my attempt to explain the source of the bias by asking you to consider how error in measurement affects regression analysis. I'll provide three different measurement-error scenarios to demonstrate what the error does.[1] In all three scenarios, assume that there is no true treatment effect, that the null hypothesis is true.

The first scenario is the case of no measurement error at all. In this hypothetical case, all of the points fall right on the regression lines themselves. The second scenario introduces measurement error on the posttest, but not on the pretest. Figure 14–17 shows that when you have posttest error, you are disbursing the points vertically—up and down—from the regression lines. Imagine a specific case, one person in a study. Without measurement error, the person would be expected to score on the regression line itself. With posttest measurement error, that person would do better or worse on the posttest than he or she would otherwise. This would lead the score to be displaced vertically. In the third scenario, measurement error occurs only on the pretest. It stands to reason in this scenario that the cases would be displaced horizontally—left and right—off of the regression lines. For these three hypothetical cases, none of which would occur in reality, you can see how data points would be disbursed.

[1]This discussion of measurement error in nonequivalent-groups designs draws heavily from Charles Reichardt's Chapter 4 in Cook, T. D., and Campbell, D. T. (1979). *Quasi-experimentation: Design and Analysis Issues for Field Settings.* Houghton Mifflin, Boston. Chip (Charles' nickname) and I were in graduate school together at Northwestern, with him preceding me by a few years. He spent many an afternoon patiently trying to explain these complex statistical issues to me. He gets all the credit for anything I might have right and none of the blame for any of my errors.

FIGURE 14–17 **Measurement error in nonequivalent-groups designs**

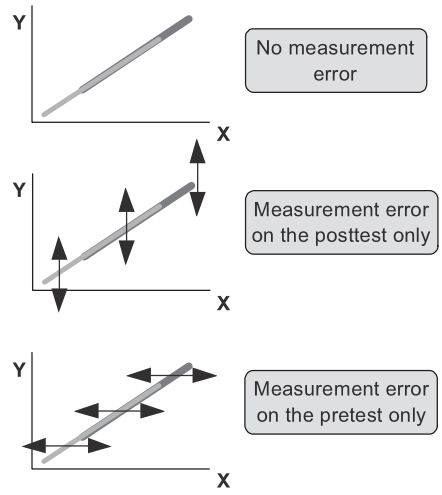
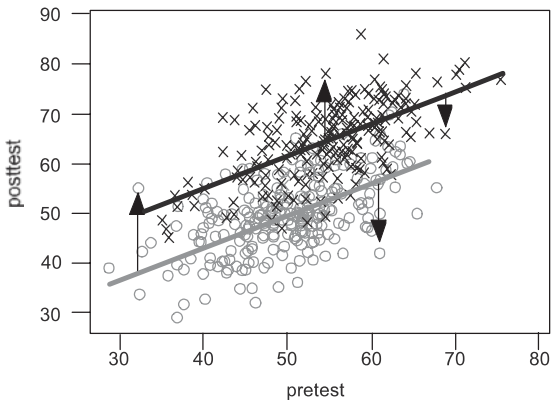


FIGURE 14–18 **Regression analysis as a least squares procedure**



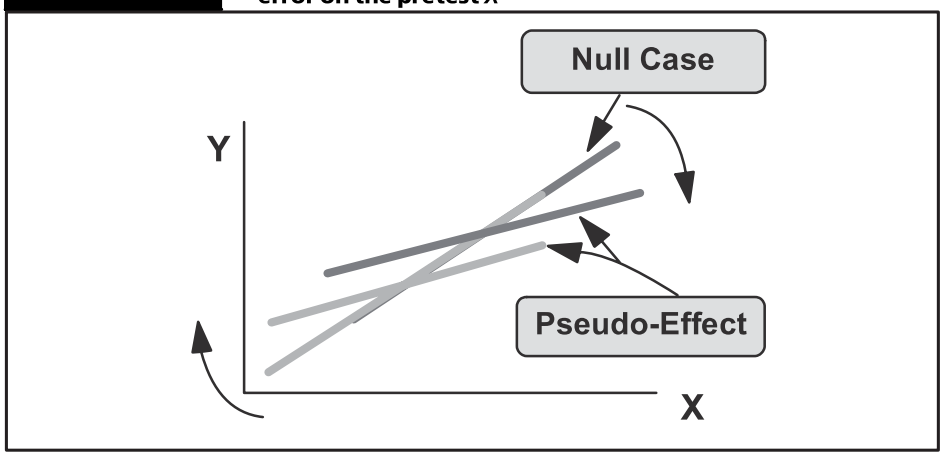
least squares
The criterion for fitting a regression line so that you minimize the sum of the squares of the residuals from the regression line.

residuals
The vertical distance from the regression line to each point. The residual in regression analysis refers to the portion of the outcome or dependent variable that you cannot predict with your regression equation.

How Regression Fits Lines Regression analysis is a **least squares** analytic procedure. The actual criterion for fitting the line is to fit it so that you minimize the sum of the squares of the **residuals** from the regression line. Let's deconstruct this sentence a bit. The key term is *residual*. The *residual* is the vertical distance from the regression line to each point.

Figure 14–18 shows four residuals, two for each group. Two of the residuals fall above their regression line and two fall below. What is the criterion for fitting a line through the cloud of data points? Take all of the residuals within a group. (Fit separate lines for the program and comparison group.) If the data points are above their respective line, the residuals will be positive; if they're below, they'll be negative values. Square all the residuals in each group. Compute the sum of the squares of the residuals—just add them. That's it. Regression analysis fits a line through the data for each group that yields the smallest sum of the squared residuals. How it does this is another matter, but you should now understand what it's doing. The key thing to notice is that the regression line *is fit in terms of the residuals, and the residuals are always and only vertical displacements from the regression line.*

FIGURE 14–19 The pseudo-effect or bias that results from measurement error on the pretest *X*



How Measurement Error Affects Slope Now let's put the ideas of the previous two sections together. Consider the three measurement-error scenarios described previously. When there is no measurement error, the slopes of the regression lines are unaffected. Figure 14–17 shows the regression lines in this no-error condition. Notice that there is no treatment effect in any of the three graphs shown in the figure. (There would be a treatment effect only if there was a vertical displacement between the two lines and a corresponding difference in intercepts.) Now, consider the case where there is measurement error on the posttest. Will the slopes be affected? The answer is no. Why? Because in regression analysis, you fit the line relative to the vertical displacements of the points. Posttest measurement error affects the vertical dimension, and, if the errors are random, you would get as many residuals pushing up as down and the slope of the line would, on average, remain the same as in the **null case**. There would, in this posttest measurement-error case, be more variability of data around the regression line, but the line would be located in the same place as in the no-error case.

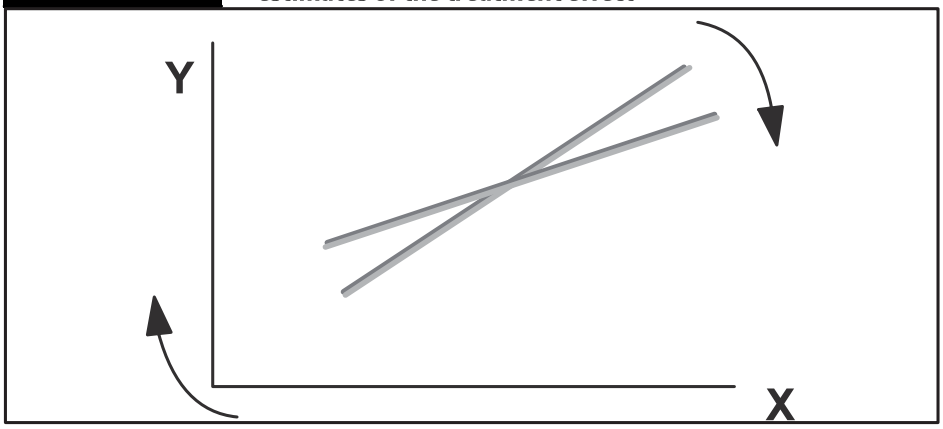
null case
A situation in which the treatment has no effect.

Now let's consider the case of measurement error on the pretest (the bottom panel in Figure 14–17). In this scenario, errors are added along the horizontal dimension, but regression analysis fits the lines relative to vertical displacements. So how will this affect the slope?

Figure 14–19 illustrates what happens. If there is no error, the lines would overlap as indicated for the null case in the figure. When you add in pretest measurement error, you are in effect elongating the horizontal dimension without changing the vertical. Since regression analysis fits to the vertical, this would force the regression line to stretch to fit the horizontally elongated distribution. The only way it can do this is by rotating around its center point. The result is that the line has been flattened or attenuated; the slope of the line will be lower when there is pretest measurement error than it should actually be. You should be able to see that flattening the line in each group by rotating it around its own center introduces a displacement between the two lines that was not there in the null case. Although there was no treatment effect in the null case, a false or pseudo-effect was introduced when pretest measurement error is introduced. The biased estimate of the slope that results from pretest measurement error introduces a phony treatment effect. In this example, it introduced an effect where there was none. In the simulated example shown in Figure 14–15, it exaggerated the actual effect that was constructed for the simulation.

Why Doesn't the Problem Occur in Randomized Designs? So, why doesn't this pseudo-effect occur in the randomized ANCOVA design? Figure 14–20

FIGURE 14–20 Pretest measurement error in a randomized experiment does attenuate slopes but does not lead to biased estimates of the treatment effect



shows that even in the randomized design, pretest measurement error does cause the slopes of the lines to be flattened.

But, a pseudo-effect in the randomized case doesn't occur even though the attenuation does. Why? Because in the randomized case, the two groups are equivalent on the pretest; there is no horizontal difference between the lines. The lines for the two groups overlap perfectly in the null case. So, when the attenuation occurs, it occurs the same way in both lines (and around the same rotation point) and there is no vertical displacement introduced between the lines. Compare Figures 14–19 and 14–20. You should now see that the difference is the NEGD case shows the attenuation of slopes *and* the initial nonequivalence between the groups. Under these circumstances, the flattening of the lines introduces a displacement. In the randomized case, flattening also occurs, but there is no displacement because there is no nonequivalence between the groups initially.

Summary of the Problem So where does this leave us? The ANCOVA statistical model seemed at first glance to have all of the right components to correctly model data from the NEGD, but it didn't work correctly. The estimate of the treatment effect was biased. Upon examination, you saw that the bias was due to two major factors: the attenuation of slope that results from pretest measurement error *coupled with* the initial nonequivalence between the groups. The problem is not caused by posttest measurement error because of the criterion that is used in regression analysis to fit the line. It does not occur in randomized experiments because there is no pretest nonequivalence. In other words, there is no selection bias. It is the combination of measurement error and selection bias that is causing the problem, and because randomized designs aren't usually susceptible to selection bias, this problem is not an issue for them. You might also guess from these arguments that the bias will be greater with greater nonequivalence between groups; the less similar the groups, the bigger the problem. In real-life research, as opposed to simulations, you can count on measurement error in all measurements. Measurement is never perfect. Therefore, in nonequivalent-groups designs, the ANCOVA analysis that seemed intuitively sensible can be expected to yield incorrect results!

The Solution Now that you understand the problem in the analysis of the NEGD, you can go about trying to fix it. Since the problem is caused in part by measurement error on the pretest, one way to deal with it would be to address the measurement-error issue. If you could remove the pretest measurement error and approximate the no pretest error case, there would be no attenuation or flattening of the regression lines and no pseudo-effect introduced. To see how you might

FIGURE 14–21 **Reliability defined in terms of true score theory**

$$\frac{\text{var}(T)}{\text{var}(T) + \text{var}(e)}$$

adjust for pretest measurement error, you need to recall what you know about measurement error and its relation to **reliability** of measurement.

Recall from reliability theory (see Chapter 3) and the idea of **true score theory** that reliability can be defined as the ratio shown in Figure 14–21.

In this ratio, T is the true ability or level on the measure and e is measurement error. It follows that the reliability of the pretest is directly related to the amount of measurement error. If there is no measurement error on the pretest, the var(e) term in the denominator is zero and reliability = 1. If the pretest is nothing but measurement error, the Var(T) term is 0 and the reliability is 0. That is, if the measure is nothing but measurement error, it is totally unreliable. If half of the measure is true score and half is measurement error, the reliability is .5. This shows that there is a direct relationship between measurement error and reliability; reliability reflects the proportion of measurement error in your measure. Since measurement error on the pretest is a necessary condition for bias in the NEGD (if there is no pretest measurement error there is no bias even in the NEGD), if you correct for the measurement error, you correct for the bias. But, you can't see measurement error directly in your data. (Remember, only God can see how much of a score is true score and how much is error, and she isn't telling.) However, you can estimate the reliability. Since reliability is directly related to measurement error, you can use the reliability estimate as a proxy for how much measurement error is present, and you can adjust pretest scores using the reliability estimate to correct for the attenuation of slopes and remove the bias in the NEGD.

reliability
The degree to which a measure is consistent or dependable; the degree to which it would give you the same result over and over again, assuming the underlying phenomenon is not changing.

true score theory
A theory that maintains that every measurement is an additive composite of two components: the true ability of the respondent and random error.

The Reliability-Corrected ANCOVA To solve the bias in ANCOVA treatment effect estimates for the NEGD, you use a reliability correction that adjusts the pretest for measurement error. Figure 14–22 shows what a reliability correction looks like.

The top graph shows the pretest distribution as you observe it, with measurement error included in it. Remember that I said previously that adding measurement error widens or elongates the horizontal dimension in the bivariate distribution. In the frequency distribution shown in the top graph, the distribution is wider than it would be if there were no error in measurement. The second graph shows that what you really want to do to adjust the pretest scores is squeeze the pretest distribution inwards by an amount proportionate to the amount that measurement error elongated or widened it. You make this adjustment separately for the program and comparison groups. The third graph shows what effect squeezing the pretest would have on the regression lines; it would increase their slopes rotating them back to where they truly belong and remove the bias that was introduced by the measurement error. In effect, you are doing the opposite of what measurement error did so that you can correct for the measurement error.

All you need to know is how much to squeeze the pretest distribution in to adjust for measurement error correctly. The answer is in the reliability coefficient. Since reliability is an estimate of the proportion of your measure that is true score relative to error, it should tell you how much you have to squeeze. In fact, the formula for the adjustment is simple (Figure 14–23).

The idea in this formula is that you are going to construct new pretest scores for each person. These new scores will be adjusted for pretest unreliability by an amount proportional to the reliability. Each person's score will be closer to the pretest mean for their group. The formula tells you how much closer. Let's look at a

FIGURE 14–22 How adjusting for pretest reliability narrows the distribution of the pretest and "sharpens" the slope of the regression lines back to their true level

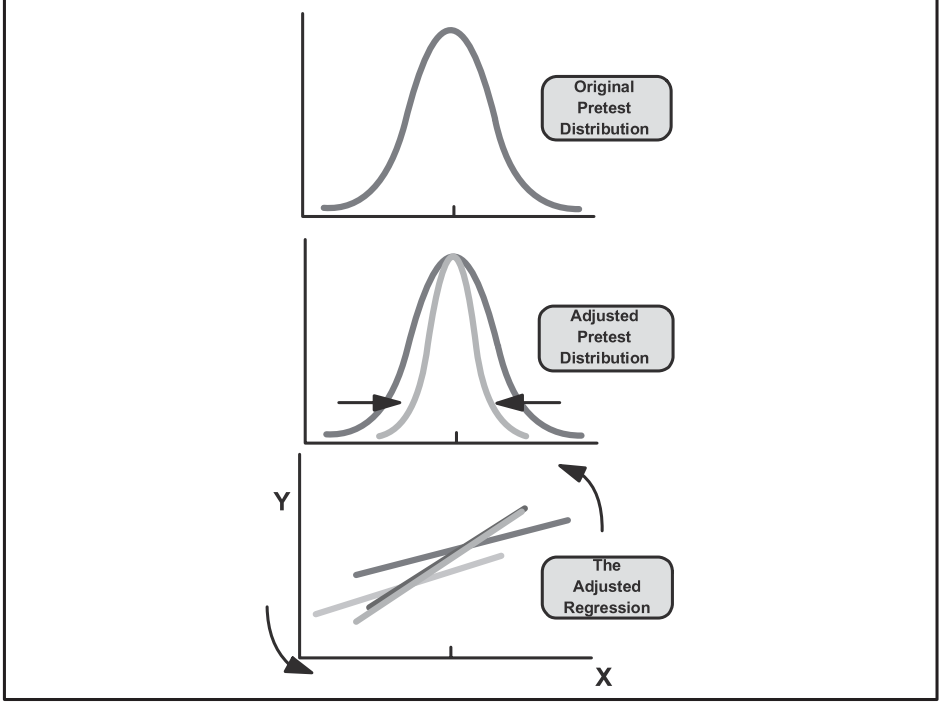


FIGURE 14–23 Formula for adjusting pretest values for unreliability in the reliability-corrected ANCOVA

$$X_{adj} = \bar{X} + r(X - \bar{X})$$

where:

X_{adj} = adjusted pretest value

$\bar{X}$ = original pretest value

r = reliability

few examples. First, let's look at the case where there is no pretest measurement error. Here, reliability would be 1. In this case, you actually don't want to adjust the data at all. Imagine that you have a person with a pretest score of 40, where the mean of the pretest for the group is 50. You would get the following adjusted score if there is no measurement error (i.e., reliability = 1):

$$\begin{aligned} X_{adj} &= 50 + 1(40 - 50) \\ X_{adj} &= 50 + 1(-10) \\ X_{adj} &= 50 - 10 \\ X_{adj} &= 40 \end{aligned}$$

Or, in other words, you wouldn't make any adjustment at all. That's what you want in the no-measurement-error case.

FIGURE 14–24 **The regression model for the reliability-corrected ANCOVA**

$$y_i = \beta_0 + \beta_1 X_{adj} + \beta_2 Z_i + e_i$$

where:

y_i = outcome score for the i^{th} unit
β_0 = coefficient for the intercept
β_1 = pretest coefficient
β_2 = mean difference for treatment
X_{adj} = transformed pretest
Z_i = dummy variable for treatment (0 = control, 1 = treatment)
e_i = residual for the i^{th} unit

Now, let's assume that reliability was moderately low, say .5. For a person with a pretest score of 40 where the group mean is 50, you would get the following:

$$\begin{aligned} X_{\text{adj}} &= 50 + .5(40 - 50) \\ X_{\text{adj}} &= 50 + .5(-10) \\ X_{\text{adj}} &= 50 - 5 \\ X_{\text{adj}} &= 45 \end{aligned}$$

In other words, when reliability is .5, you would move the pretest score halfway in towards the mean—halfway from its original value of 40 towards the mean of 50, or to 45.

Finally, let's assume that for the same case the reliability was stronger at .8. The reliability adjustment would be as follows:

$$\begin{aligned} X_{\text{adj}} &= 50 + .8(40 - 50) \\ X_{\text{adj}} &= 50 + .8(-10) \\ X_{\text{adj}} &= 50 - 8 \\ X_{\text{adj}} &= 42 \end{aligned}$$

That is, with reliability of .8 you would want to move the score in 20 percent toward its mean (because if reliability is .8, the amount of the score due to error is $1 - .8 = .2$).

You should be able to see that if you make this adjustment to all of the pretest scores in a group, you would be squeezing the pretest distribution in by an amount proportionate to the measurement error (1 − reliability). It's important to note that you need to make this correction separately for your program and comparison groups.

You're now ready to take this adjusted pretest score and substitute it for the original pretest score in the ANCOVA model (Figure 14–24).

Notice that the only difference is that the X in the original ANCOVA is changed to the term X_{adj}.

The Simulation Revisited So, let's go see how well these adjustments work. I'll use the same simulated data that I used earlier. The results are shown in Figure 14–25.

This time the estimate of the treatment effect is 9.3048 (instead of 11.2818). This estimate is closer to the true value of 10 points that I put into the simulated data. When I construct a 95 percent confidence interval for the adjusted estimate, the true value of 10 falls within the interval. That is, the analysis estimated a treatment effect that is not statistically different from the true effect; it is an unbiased estimate.

FIGURE 14–25 **Results of the reliability-corrected ANCOVA for the simulated data**

$$y_i = -3.14 + 1.06X_{adj} + 9.30Z_i$$

Predictor	Coef	StErr	t	p
Constant	-3.141	3.300	-0.95	0.342
adjpre	1.06316	0.06557	16.21	0.000
Group	9.3048	0.6166	15.09	0.000

- $CI_{.95(\beta_2 = 10)} = \beta_2 \pm 2SE(\beta_2)$
  $= 9.3048 \pm 2(.6166)$
  $= 9.3048 \pm 1.2332$
- CI = 8.0716 to 10.5380

You should also compare the slope of the lines in this adjusted model with the original slope. Now the slope is nearly 1 at 1.06316, whereas before it was .626—considerably lower or flatter. The slope in the adjusted model approximates the expected true slope of the line (which is 1) in the simulated data. The original slope showed the attenuation that the pretest measurement error caused.

So, the reliability-corrected ANCOVA model is used in the statistical analysis of the NEGD to correct for the bias that would occur as a result of measurement error on the pretest.

Which Reliability Should You Use? There's really only one more major issue to settle to finish this story. You know from reliability theory that you can't calculate the true reliability; you can only estimate it. A variety of reliability estimates exist, and they're likely to give you different values. **Cronbach's alpha** tends to be a high estimate of reliability. The test-retest reliability tends to be a lower-bound estimate of reliability. So which do you use in a correction formula? The answer is both! When analyzing data from the NEGD, it's safest to do two analyses: one with an upper-bound estimate of reliability and one with a lower-bound one. If you find a significant treatment effect estimate with both, you can be fairly confident that you have found a significant effect.

Cronbach's alpha
One specific method of estimating the reliability of a measure. Although not calculated in this manner, Cronbach's alpha can be thought of as analogous to the average of all possible split-half correlations.

This certainly doesn't feel like a satisfying conclusion to this rather convoluted story about the analysis of the NEGD, and it's not. In some ways, I look at this as the price you pay when you give up random assignment and use intact groups in a NEGD: Your analysis becomes more complicated as you deal with adjustments that are needed, in part, because of the nonequivalence between the groups. Nevertheless, there are also benefits in using nonequivalent groups instead of randomly assigning. You have to decide whether the trade-off is worth it.

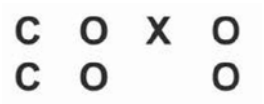
Statisticians have been well aware of this problems of selection bias and measurement error for several decades. They have developed a number of statistical approaches that are more sophisticated and versatile than the simple reliability-corrected analysis described here. My favorite is matching and propensity score modeling (Rubin and Thomas, 1996)[2], but there are a number of alternatives (Winship and Morgan, 1999[3]; Heckman and Robb, 1988)[4]. We introduce the idea of propensity modeling briefly in Section 14-4d, but the details of this and other approaches fall outside the scope of this modest introductory text. At the very least,

[2]Rubin, D. B. and Thomas, N. (1996). Matching using estimated propensity scores: relating theory to practice. *Biometrics*, 52: 249–64.

[3]Winship, C. and Morgan, S. L. (1999). The estimation of causal effects from observational data, *Annual Review of Sociology;* pps. 659–707.

[4]Heckman, J. J. and Robb, R. (1988). The value of longitudinal data for solving the problem of selection bias in evaluating the impact of treatment on outcomes. In Duncan, G. and Kalton, G. (Eds.) Panel Surveys. New York, Wiley, 512–38.

FIGURE 14–26 **Notation for the regression-discontinuity design**



this discussion should make you aware of the importance of the problem and the existence of a range of potential solutions.

14-4b Regression-Discontinuity Analysis

The basic regression-discontinuity analysis design (Trochim, 1984)[5] is a two-group, pretest-posttest model as indicated in the design notation (see Figure 14–26). As in other versions of this design structure (see Section 14-3d, Analysis of Covariance, and Section 14-4a, Nonequivalent-Groups Analysis), you will need a statistical model that includes a term for the pretest, one for the posttest, and a dummy-coded variable to represent the program.

Assumptions in the Analysis Before discussing the specific analytic model, it's important to understand the assumptions that must be met. This presentation assumes that you are dealing with the basic regression-discontinuity design as described in Chapter 10-2. There are five central assumptions that must be made for the analytic model to be appropriate, each of which is discussed in turn:

1. *The cutoff criterion.* The cutoff criterion must be followed without exception. When there is an incorrect assignment relative to the cutoff value (unless it is known to be random), a **selection threat** arises and estimates of the effect of the program are likely to be biased. An incorrect assignment relative to the cutoff, often termed a *fuzzy* regression-discontinuity design, introduces analytic complexities that are outside the scope of this discussion.
2. *The pre-post distribution.* It is assumed that the pre-post distribution is describable as a polynomial function. If the true pre-post relationship is logarithmic, exponential, or some other function, the following model is specified incorrectly and estimates of the effect of the program are likely to be biased. Of course, if the data can be transformed to create a polynomial distribution prior to analysis, the following model may be appropriate although it is likely to be more problematic to interpret. It is also sometimes the case that even if the true relationship is not polynomial, a sufficiently high-order polynomial will adequately account for whatever function exists. However, you are not likely to know whether this is the case.
3. *Comparison group pretest variance.* There must be a sufficient number of pretest values in the comparison group to enable adequate estimation of the true relationship (for example, the pre-post regression line) for that group. It is usually desirable to have variability in the program group as well although this is not strictly enforced because you can project the comparison group line to a single point for the program group.
4. *Continuous pretest distribution.* Both groups must come from a single, continuous pretest distribution with the division between groups determined by the cutoff. In some cases, you might be able to find intact groups (for example, two groups of patients from two different geographic locations) that serendipitously divide on some measure so as to imply some cutoff. Such naturally discontinuous groups must be used with caution because of the greater likelihood that if they differed naturally at the cutoff prior to the program such a difference could

selection threat
Any factor other than the program that leads to posttest differences between groups.

[5]Trochim, W. (1984). Research Design for Program Evaluation: The Regression-Discontinuity Approach. Beverly Hills, CA: Sage Publications.

FIGURE 14–27a **A curvilinear relationship**

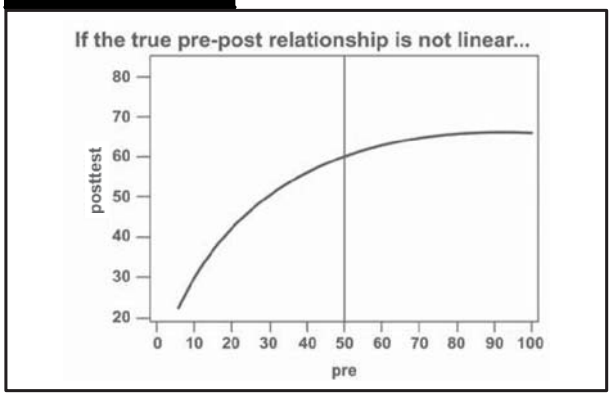


FIGURE 14–27b **Misspecification of the regression model in a regression-discontinuity design by fitting straight lines to a true curvilinear relationship**

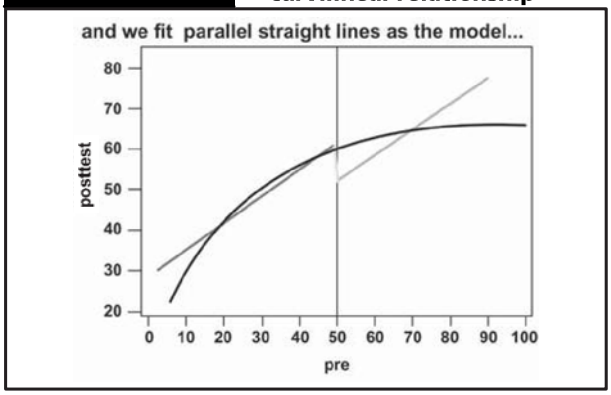
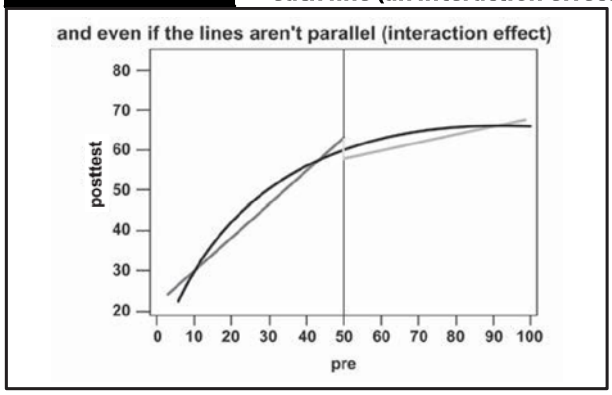


FIGURE 14–27c **A curvilinear relationship fit with a straight-line model with different slopes for each line (an interaction effect)**



reflect a selection bias that could introduce natural pre-post discontinuities at that point.

5. *Program implementation.* It is assumed that the program is uniformly delivered to all recipients, that is, that they all receive the same dosage, length of stay, amount of training, or whatever. If this is not the case, it is necessary to model explicitly the program as implemented, thus complicating the analysis somewhat.

The Curvilinearity Problem The major problem in analyzing data from the regression-discontinuity design is incorrect model specification. As I will show, when you specify the statistical model incorrectly, you are likely to get biased estimates of the treatment effect. To introduce this idea, let's begin by considering what happens if the data (the bivariate, pre-post relationship) is curvilinear and you fit a straight-line model to the data.

Figure 14–27a shows a simple curvilinear relationship. If the curved line in Figure 14–27a describes the pre-post relationship, you need to take this into account in your statistical model. Notice that, although there is a cutoff value at 50 in the figure, there is no jump or discontinuity in the line at the cutoff. This indicates that there is no effect of the treatment.

FIGURE 14–28a An exactly specified model

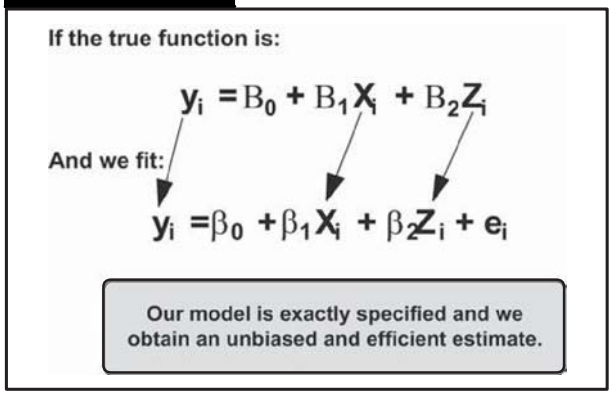


FIGURE 14–28b An overspecified model

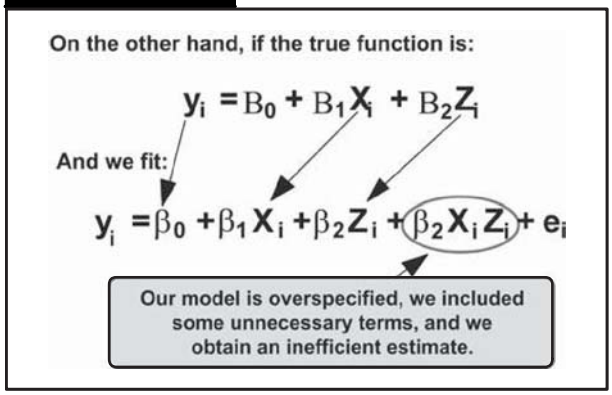
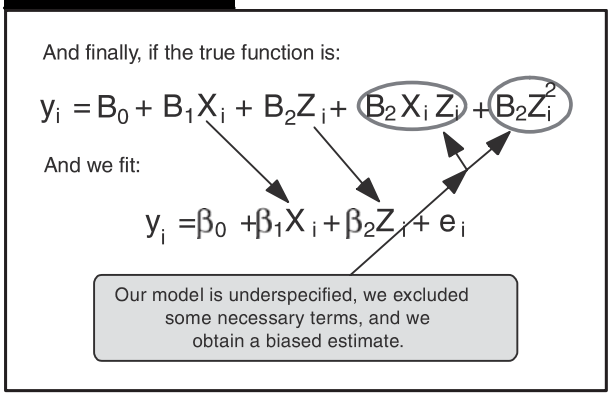


FIGURE 14–28c An underspecified model



Now look at Figure 14–27b. The figure shows what happens when you fit a straight-line model to the curvilinear relationship of Figure 14–27a. In the model, slopes of both straight lines are restricted and so must be the same. (For example, you did not allow for any interaction between the program and the pretest.) You can see that the straight-line model suggests that there is a jump at the cutoff, even though you can see that in the true function there is no discontinuity.

Even allowing the straight-line slopes to differ (i.e., allowing for a linear interaction effect) doesn't solve the problem (although it does help). Figure 14–27c shows what happens in this case. Although the pseudo-effect in this case is smaller than when the slopes are forced to be equal, you still obtain a pseudo-effect.

The conclusion is a simple one. If the true model is curved and you fit only straight lines, you are likely to conclude incorrectly that the treatment made a difference when it did not. This is a specific instance of the more general problem of model specification.

Model Specification To understand the model specification issue and how it relates to the regression-discontinuity design, you must try to distinguish three types of specifications. Figure 14–28a shows the case where the true model is *exactly specified*. What does "exactly specified" mean? The top equation describes the truth for the data. It describes a simple straight-line, pre-post relationship with a treatment effect. Notice that it includes terms for the posttest *Y*, the pretest *X*, and the dummy-coded treatment variable *Z*. The bottom equation shows the model that you specify in the analysis. It too includes a term for the posttest *Y*, the pretest *X*, and the dummy-coded treatment variable *Z*. That's all it includes; there are no

unnecessary terms in the model that you specify. When you exactly specify the true model, you get unbiased and efficient estimates of the treatment effect.

Now, let's look at the situation in Figure 14–28b. The true model is the same as in Figure 14–28a. However, this time an analytic model that includes an extra and unnecessary term is specified. In this case, because all of the necessary terms are included, the estimate of the treatment effect will be unbiased. However, you pay a price for including unneeded terms in your analysis; the treatment effect estimate will not be efficient. What does this mean? It means that the chance that you will conclude your treatment doesn't work when it in fact does increases. Including an unnecessary term in the analysis is like adding unnecessary noise to the data; it makes it harder to see the effect of the treatment even if it's there.

Finally, consider the example shown in Figure 14–28c. Here, the truth is more complicated than the model. In reality, only two of the necessary terms are included in the analysis. In this case, the treatment effect estimate is both biased and inefficient.

Analysis Strategy Given the discussion of model misspecification, you can develop a modeling strategy that is designed first, to guard against biased estimates and second, to ensure maximum efficiency of estimates. The best option would obviously be to specify the true model exactly. However, this is often difficult to achieve in practice because the true model is often obscured by the error in the data and because you're not God and don't know what the true model is. If you have to make a mistake—if you must specify the model incorrectly—the discussion of misspecification suggests you should overspecify the true model rather than underspecify. Overspecification ensures that you have included all necessary terms even at the expense of unnecessary ones. It will yield an unbiased estimate of the effect, even though it will be inefficient. Underspecification is the situation you most want to avoid because it yields both biased and inefficient estimates.

Given this preference sequence, you should begin your general analysis by specifying a model that you are fairly certain is overspecified. The treatment effect estimate for this model is likely to be unbiased although it will be inefficient. Then, in successive analyses, gradually remove higher-order terms until the treatment-effect estimate appears to differ from the initial one, or until the model diagnostics (for example, the residual plots) indicate that the model fits poorly.

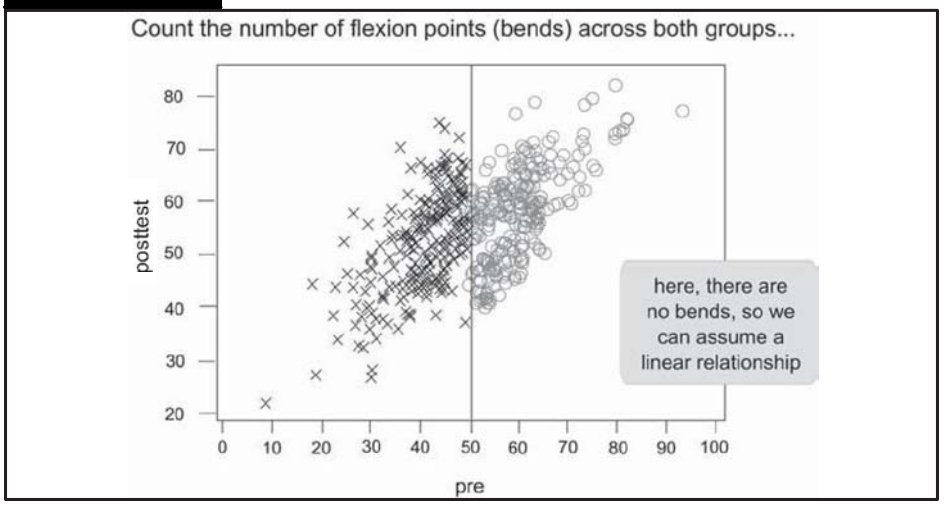
Steps in the Analysis The basic regression-discontinuity analysis involves five steps:

1. *Transform the pretest.* The analysis begins by subtracting the cutoff value from each pretest score, creating the modified pretest term shown in Figure 14–29. This is done to set the intercept equal to the cutoff value. How does this work? If you subtract the cutoff from every pretest value, the modified pretest will be equal to 0 where it was originally at the cutoff value. Since the intercept is by definition the *Y*-value when $X = 0$, what you have done is set X to 0 at the cutoff, making the cutoff the intercept point. You want to estimate the treatment effect at the cutoff and you achieve this by setting the cutoff equal to the intercept using this adjustment.
2. *Examine relationship visually.* You should look for two major things in a graph of the pre-post relationship. First it is important to determine whether there is any

FIGURE 14–29 **Transforming the pretest by subtracting the cutoff value**

$$\tilde{X}_i = X_i - X_c$$

FIGURE 14–30 **Bivariate distribution with no flexion points**



visually discernable discontinuity in the relationship at the cutoff. The discontinuity could be a change in level vertically (main effect), a change in slope (interaction effect), or both. If it is visually clear that there is a discontinuity at the cutoff, you should not be satisfied with analytic results that indicate no program effect. However, if no discontinuity is visually apparent, it may be that variability in the data is masking an effect and you must attend carefully to the analytic results.

The second thing to look for in the bivariate relationship is the degree of polynomial that may be required as indicated by the bivariate slope of the distribution, particularly in the comparison group. A good approach is to count the number of flexion points (that is, the number of times the distribution flexes or bends) that are apparent in the distribution. If the distribution appears linear, there are no flexion points. A single flexion point could be indicative of a second (quadratic) order polynomial. You use this information to determine the initial model that will be specified.

3. *Specify higher-order terms and interactions.* Depending on the number of flexion points detected in step 2, you next create transformations of the modified assignment variable, X. The rule of thumb here is that you go two orders of polynomial higher than was indicated by the number of flexion points. Thus, if the bivariate relationship appeared linear (there were no flexion points), you would want to create transformations up to a second-order $(0 + 2)$ polynomial. This is shown in Figure 14–30. There do not appear to be any inflexion points or bends in the bivariate distribution of Figure 14–30.

   The first-order polynomial already exists in the model (X) and so you would only have to create the second-order polynomial by squaring X to obtain X^2. For each transformation of X, you also create the interaction term by multiplying the polynomial by Z. In this example, there would be two interaction terms: $X_i Z_i$ and $X_i^2 Z_i$. Each transformation can be easily accomplished through straightforward multiplication on the computer. If there appeared to be two flexion points in the bivariate distribution, you would create transformations up to the fourth $(2 + 2)$ power and their interactions.

   Visual inspection need not be the only basis for the initial determination of the degree of polynomial that is needed. Certainly, prior experience modeling similar data should be taken into account. The rule of thumb given here implies that you should err on the side of overestimating the true polynomial function that is needed for reasons outlined previously in discussing model specification. For whatever power is initially estimated from visual inspection, you should

FIGURE 14–31 The initial model for the case of no flexion points (full quadratic model specification)

$$y_i = \beta_0 + \beta_1 \tilde{X}_i + \beta_2 Z_i + \beta_3 \tilde{X}_i Z_i + \beta_4 \tilde{X}_i^2 + \beta_5 \tilde{X}_i^2 Z_i + e_i$$

where:
y_i = outcome score for the i^{th} unit
β_0 = coefficient for the intercept
β_1 = pretest coefficient
β_2 = mean difference for treatment
β_3 = linear interaction
β_4 = quadratic pretest coefficient
β_5 = quadratic interaction
X_i = transformed pretest
Z_i = dummy variable for treatment (0 = control, 1 = treatment)
e_i = residual for the i^{th} unit

construct all transformations and their interactions up to that power. Thus if the fourth power is chosen, you should construct all four terms X to X^4 and their interactions.

4. *Estimate initial model.* At this point, you can begin the actual analysis. You can use any acceptable multiple-regression program on the computer to accomplish this. You simply regress the posttest scores, y, on the modified pretest, X, the treatment variable, Z, and all higher-order transformations and interactions created in step 3. The regression coefficient associated with the Z term (the group-membership variable) is the estimate of the main effect of the program. If there is a vertical discontinuity at the cutoff, it will be estimated by this coefficient. You can test the significance of the coefficient (or any other) by constructing a standard t-test using the standard error of the coefficient, which is invariably supplied in the computer program output. For the data in Figure 14–30, you would use the model given in Figure 14–31.

   If, during step 3, you correctly overestimated the polynomial function required to model the distribution, the estimate of the program effect will at least be unbiased. However, by including terms that may not be needed in the true model, the estimate is likely to be inefficient; that is, standard error terms will be inflated and hence the significance of the program effect may be underestimated. Nevertheless, if at this point in the analysis the coefficient is highly significant, it would be reasonable to conclude that there is a program effect. The direction of the effect is interpreted based on the sign of the coefficient and the direction of scale of the posttest. Interaction effects can also be examined. For instance, a linear interaction would be implied by a significant regression coefficient for the XZ term.

5. *Refining the model.* On the basis of the results of step 4, you might want to attempt to remove apparently unnecessary terms and reestimate the treatment effect with greater efficiency. This is a tricky procedure and should be approached cautiously to minimize the possibility of bias. To accomplish this, you should certainly examine the output of the regression analysis in step 4, noting the degree to which the overall model fits the data, the presence of any insignificant coefficients, and the pattern of residuals. A conservative way to decide how to refine the model would be to begin by examining the highest-order term in the current model and its interaction. If both coefficients are nonsignificant, and the goodness-of-fit measures and pattern of residuals indicate a good fit, you might drop these two terms and reestimate the resulting model. Thus, if you estimated up to a fourth-order polynomial, and found the coefficients for X^4 and X^4Z were

FIGURE 14–32 **Bivariate distribution for example regression-discontinuity analysis**

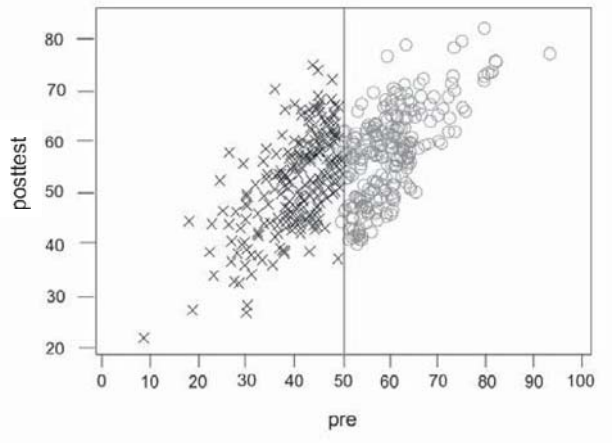


FIGURE 14–33 **Regression results for the full quadratic model**

The regression equation is

posttest = 49.1 + 0.972*precut + 10.2*group
-0.236*linint - 0.00539*quad + 0.00276 quadint

Predictor	Coef	Stdev	t-ratio	p
Constant	49.1411	0.8964	54.82	0.000
precut	0.9716	0.1492	6.51	0.000
Group	10.231	1.248	8.20	0.000
linint	-0.2363	0.2162	-1.09	0.275
quad	-0.005391	0.004994	-1.08	0.281
quadint	0.002757	0.007475	0.37	0.712

• s = 6.643 • R - sq = 47.4% • R - sq (adj) = 47.1%

nonsignificant, you could drop these terms and respecify the third-order model. You would repeat this procedure until either of the coefficients is significant, the goodness-of-fit measure drops appreciably, or the pattern of residuals indicates a poorly fitting model. The final model may still include unnecessary terms, but there are likely to be fewer of these and consequently, efficiency should be greater. Model-specification procedures that involve dropping any term at any stage of the analysis are more dangerous and more likely to yield biased estimates because of the considerable multicolinearity that will exist between the terms in the model.

Example Analysis Okay, so I've thrown a lot at you in this section. Here's where I think it will begin to make a little more sense. It's easier to understand how data from a regression-discontinuity design is analyzed by looking at an example. The data for this example is shown in Figure 14–32.

Several things are apparent visually. First, there is a whopping treatment effect in this simulated data. You'll never see an effect like this in real life! Figure 14–32 shows simulated data where the true treatment effect is 10 points. Second, both groups are well described by straight lines; there are no flexion points apparent. Thus, the initial model to specify is the full quadratic one shown in Figure 14–31.

The results of the initial specification are shown in Figure 14–33. The treatment effect estimate is the one next to the group variable. This initial estimate is

FIGURE 14–34 **Regression results for initial model without quadratic terms**

The regression equation is

posttest = 49.8 + 0.824*precut + 9.89*group -0.0196*linint

Predictor	Coef	Stdev	t-ratio	p
Constant	49.7508	0.6957	71.52	0.000
precut	0.82371	0.05889	13.99	0.000
Group	9.8939	0.9528	10.38	0.000
linint	-0.01963	0.08284	-0.24	0.813

• s = 6.639 • R - sq = 47.5% • R - sq (adj) = 47.2%

FIGURE 14–35 **Regression results for final model**

The regression equation is

posttest = 49.8 + 0.814*precut + 9.89*group

Predictor	Coef	Stdev	t-ratio	p
Constant	49.8421	0.5786	86.14	0.000
precut	0.81379	0.04138	19.67	0.000
Group	9.8875	0.9515	10.39	0.000

• s = 6.633 • R - sq = 47.5% • R - sq (adj) = 47.3%

10.231 (SE = 1.248)—close to the true value of 10 points—but notice that there is evidence that several of the higher-order terms are not statistically significant and may not be needed in the model. Specifically, the linear interaction term linint (XZ), and both the quadratic (X^2) and quadratic interaction (X^2Z) terms are not significant.

Although you might be tempted (and perhaps even justified) to drop all three terms from the model, if you follow the guidelines given in step 5, you begin by dropping only the two quadratic terms quad and quadint. The results for this model are shown in Figure 14–34.

You can see that in this model the treatment effect estimate is now 9.89 (SE = .95). Again, this estimate is close to the true 10-point treatment effect. Notice, however, that the SE is smaller than it was in the original model. This is the gain in efficiency you get when you eliminate the two unneeded quadratic terms. You can also see that the linear interaction term linint is still not significant. This term would be significant if the slopes of the lines for the two groups were different. Visual inspection shows that the slopes are the same and so it makes sense that this term is not significant.

Finally, let's drop out the nonsignificant linear interaction term and respecify the model. These results are shown in Figure 14–35.

You see in these results that the treatment effect and SE are almost identical to the previous model and that the treatment effect estimate is an unbiased estimate of the true effect of 10 points. You can also see that all the terms in the final model are statistically significant, suggesting that they are needed to model the data and should not be eliminated.

So, what does our model look like visually? Figure 14–36 shows the original bivariate distribution with the fitted regression model.

Clearly, the model fits well, both statistically and visually.

FIGURE 14–36 **Bivariate distribution with final regression model**

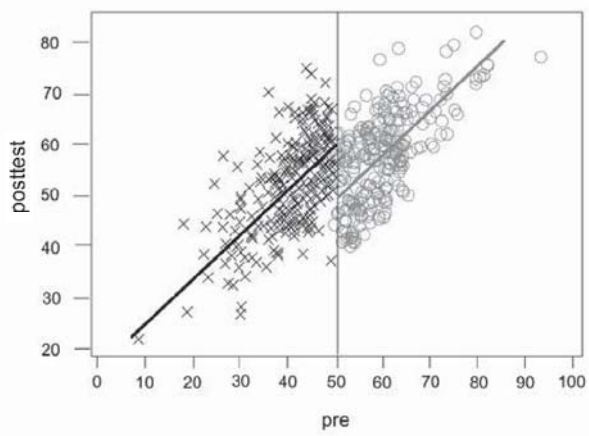


FIGURE 14–37 **Notation for the regression point displacement design**

$$\begin{array}{llll} N_{(n=1)} & O & X & O \\ N & O & & O \end{array}$$

14-4c Regression Point Displacement Analysis

I come now to the last of the quasi-experimental designs I want to discuss in reference to analysis—the regression point displacement (RDP) design (Trochim and Campbell, 1999)[6]. At this point in the chapter, you should be able to anticipate the kind of analysis I'm going to suggest. You'll see that the principles are the same here as for all of the other analyses, especially in that this analysis also relies on the GLM and regression analysis.

The notation for the RPD design (Figure 14–37) shows that the statistical analysis requires the following:

- A posttest score
- A pretest score
- A variable to represent the treatment group (where 0 = comparison and 1 = program)

These requirements are identical to the requirements for the ANCOVA model and should look very familiar by now. The only difference is that the RPD design has only a single treated group score.

Figure 14–38 shows a bivariate (pre-post) distribution for a hypothetical RPD design of a community-based AIDS education program. The new AIDS education program is piloted in one particular county in a state, with the remaining counties acting as controls. The state routinely publishes annual HIV-positive rates by county for the entire state. The *x*-values show the HIV-positive rates per 1000 people for the year preceding the program, while the *y*-values show the rates for the year following it. Our goal is to estimate the size of the vertical displacement of the treated unit from the regression line of all of the control units, indicated on the graph by

[6]Trochim, W. and Campbell, D. (1999). Design for Community-Based Demonstration Projects. In Campbell, D. T. and Russo, M. J. (Eds.). Social Experimentation, Sage Publications, 309–337.

FIGURE 14–38 **Bivariate plot for the RPD design**

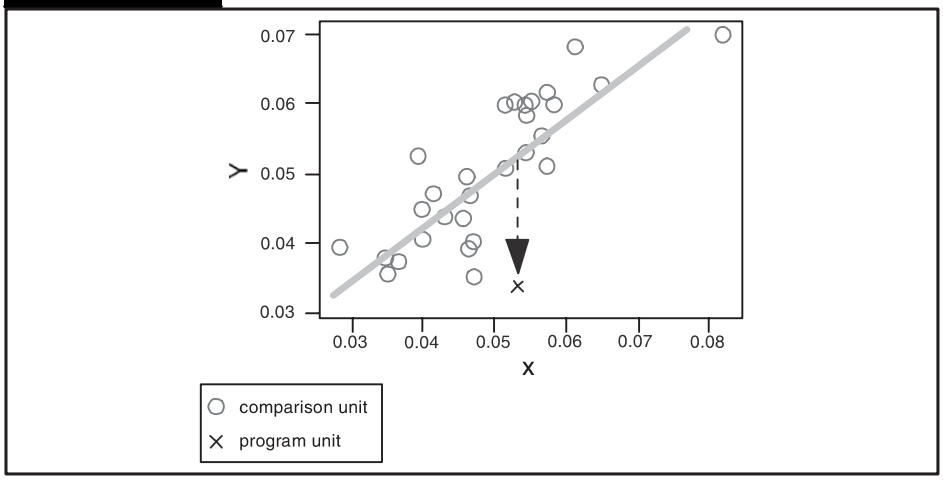


FIGURE 14–39 **The regression model for the RPD design assuming a linear pre-post relationship**

$$y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + e_i$$

where:

y_i = outcome score for the i^{th} unit

β_0 = coefficient for the intercept

β_1 = pretest coefficient

β_2 = mean difference for treatment

X_i = covariate

Z_i = dummy variable for treatment (0 = control, 1 = treatment [n = 1])

e_i = residual for the i^{th} unit

the dashed arrow. The model I'll use is the now-familiar ANCOVA model stated in regression model form (Figure 14–39).

When you fit the model to the simulated data, you obtain the regression table shown in Figure 14–40.

The coefficient associated with the dichotomous treatment variable is the estimate of the vertical displacement from the line. In this example, the results show that the program lowers HIV-positive rates by .019 and that this amount is statistically significant. This displacement is shown in the results graph in Figure 14–41.

14-4d Propensity Score Analysis

propensity score analysis
A statistical modeling approach for adjusting for selection bias by estimating the probability or "propensity" of assignment to treatment given a set of pre-program variables.

In addition to the RPD analysis, there is another method for analyzing data from quasi-experimental designs that can produce relatively strong causal inference in the absence of randomization: **propensity score analysis**. As you know, the great benefit of random assignment is to strengthen internal validity by eliminating bias in results due to preexisting differences in the members of treatment and comparison groups. But what if the preexisting differences could themselves be measured and then used in the analysis to refine our estimates of treatment effects accordingly? That is exactly what we do in propensity score analysis.

FIGURE 14–40 **Results of applying the regression model of Figure 14–39 to the data of Figure 14–38**

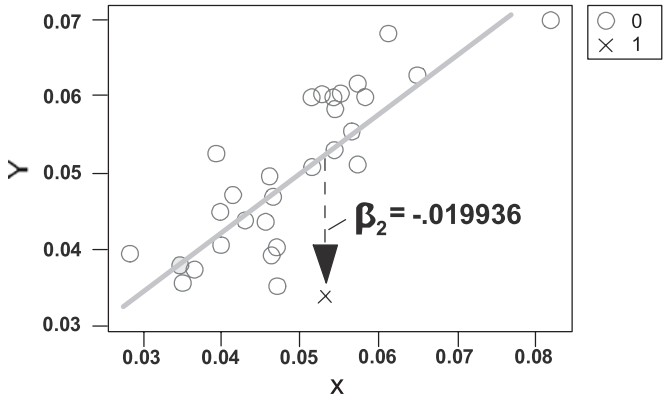
The regression equation is

Y = 0.0120 + 0.784 X - 0.0199 Z

Predictor	Coef	Stdev	t-ratio	p
Constant	0.011956	0.004965	2.41	0.023
X	0.78365	0.09864	7.94	0.000
Z	-0.019936	0.005800	-3.44	0.002

• s = 0.005689 • R - sq = 72.6% • R - sq (adj) = 70.6%

FIGURE 14–41 **The results of the RPD analysis showing the size of the treatment effect**



In a randomized design, we know the propensity, or probability, or any individual being in a treatment or control group is exactly .5 because of random assignment. In a quasi-experimental design, we can estimate the probability of a person being in a treatment or comparison group by building a statistical model of the variables that account for the *propensity* of a study participant to happen to be in a treatment or comparison group. Thus, a propensity score can be defined as the probability (ranging as always from 0 to 1) of being in a treatment or comparison group based on a set of covariates that predict membership in the treatment or comparison group (Luellen, Shadish, & Clark, 2005). In a study of participants in social or health services, some of the predictors of group membership might include age, gender, socioeconomic status, prior use of similar services, and perhaps measures of functioning in relevant domains. A propensity score can be modeled using a variety of advanced statistical methods, including logistic regression and classification trees, and then incorporated into the analysis of treatment effects. Successful use of this strategy is much more likely if the set of relevant predictors of group membership has been thoughtfully considered prior to the study so that the data is collected and ready for inclusion in the analysis.

TABLE 14–1 Summary of the Statistical Models for the Experimental and Quasi-Experimental Research Designs

Design	Analysis	Notes
Experimental Designs		
The two-group posttest-only randomized experiment	$y_i = \beta_0 + \beta_1 z_i + e_i$	No pretest term x_i needed. This is eqivalent to the *t*-test.
Factorial design (2←2)	$y_i = \beta_0 + \beta_1 z_{1i} + \beta_2 z_{2i} + \beta_3 z_{1i} z_{2i} + e_i$	No pretest term x_i needed; the z terms are dummy variables for each of the two factors.
Factorial blocks	$y_i = \beta_0 + \beta_1 z_{1i} + \beta_2 z_{2i} + \beta_3 z_{3i} + \beta_4 z_{4i} + e_i$	No pretest term x_i needed; z_1 is dummy for treatment; z_{2-4} are dummies for blocks 2–4 of 4.
Analysis of covariance (ANCOVA)	$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + e_i$	The analysis of covariance (ANCOVA) model
Quasi-Experimental Designs		
Nonequivalent groups design (NEGD)	$y_i = \beta_0 + \beta_1 x_{adj} + \beta_2 z_i + e_i$	$x_{adj} = \bar{x} + r(x - \bar{x})$ where $\bar{x}$ is the pretest average and r is the reliability estimate.
Regression discontinuity design (RD)	$y_i = \beta_0 + \beta_1 \tilde{x}_i + \beta_2 z_i + e_i$	$\tilde{x}_i$ = pretest (x_i) – cutoff value
Regression point displacement design (RPD)	$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + e_i$	Identical to ANCOVA except that for dummy treatment variable, there is only one pre-post score pair where $z_i = 1$ (treated case).

Summary

Phew! I'm hoping for your sake that you weren't assigned to read this chapter in a single night. And, I'm hoping you didn't put this off until the night before the examination. However, in case you did, let me summarize the salient points.

This chapter described data analysis as it relates to research design. When I talk about statistical analysis of research design, I'm in the area usually referred to as *inferential statistics*. With inferential statistics, we attempt to draw inferences (often causal) from the data based on our research design. To understand inferential statistics, you need to be familiar with the general linear model (GLM), which underlies virtually all of the statistics presented in this chapter. The GLM is the basic structure of the *t*-test, ANOVA, ANCOVA, regression analysis, and many of the multivariate methods, including factor analysis, multidimensional scaling, cluster analysis, discriminant function analysis, and so on. GLM is a deceptively simple formula that describes how a set of independent variables is related mathematically to a set of dependent variables. To understand how the GLM works in research design you have to grasp the idea of dummy variables that are used to represent the various treatment and control groups in the regression models for the different designs. The analyses of the different experimental designs are straightforward regression models that range from the simple *t*-test to the multiple dummy variable factorial and blocking models to the ANCOVA that includes both a dummy variable and a continuous pretest. The quasi-experimental designs make for more difficult and challenging modeling problems. While it might seem intuitively that the usual ANCOVA model would be appropriate for both the NEGD and the regression-discontinuity, it will often yield biased estimates if not appropriately modified. All of the models, grouped by design, are shown in Table 14–1.

Login to the Online Edition of your text at www.atomicdog.com to find additional resources located in the Study Guide at the end of each chapter.