18

Using Regression Models to Estimate Program Effects

Charles S. Reichardt, Carol A. Bormann

Earlier chapters have described four research designs that are widely used for estimating program effects: the randomized experiment, the regression-discontinuity design, the nonequivalent comparison group design, and the interrupted time series design. This chapter explains how to analyze data from each of these four designs, using simple but effective statistical techniques that fall under the rubric of regression analysis.

The first section describes the purpose of statistical analysis. The four sections that follow explain how to analyze data from each of the four research designs. The purpose of the presentation is not to make you a statistical expert but rather to give you a sense of the logic behind the statistical analyses.

The Tasks of Statistical Analysis

The purpose of the four designs is to estimate the effects of a program or treatment. For example, an evaluation of the first year of "Sesame Street" estimated the effects of the television series on preschool children's learning and readiness for school (Ball and Bogatz, 1970). To estimate the effects of "Sesame Street," or of any other program, three tasks must be accomplished.

The Statistical Significance of a Treatment Effect

The first task is to show that the treatment effect is statistically significant. An introduction to statistical significance testing is given earlier in this volume in Chapter Seventeen by Newcomer. However, it is useful to review briefly the meaning of statistical significance tests.

Preparation of this chapter was supported, in part, by Grant U01-AA08778 from the National Institute on Alcohol Abuse and Alcoholism. The authors thank the editors of the volume for their helpful comments.

Perhaps the purpose of a statistical significance test can best be understood in the context of a randomized experiment. In the simplest of such experiments, individuals are randomly assigned to two treatment groups: an experimental group that receives the program being evaluated and a control or comparison group that does not receive the program being evaluated. If the program has an effect, the two groups will perform differently on an appropriate outcome measure. The problem is that even if the program being evaluated has no effect, the performance of the two groups on the relevant outcome measure will not be identical. One group would perform better on average than the other group simply because of chance differences introduced by the random assignment.

Therefore, the question facing the researcher is not whether there is *any* difference in performance between the two groups on the outcome measure, but whether the difference in performance is larger than would be expected by chance. This is the question that is answered by using a statistical significance test. If the results of the test are *statistically significant,* it means the observed difference is too large to be reasonably attributed to chance differences and therefore is indicative of a treatment effect.

Imagine a randomized experiment with five individuals assigned to each of two groups. The hypothetical data from this experiment are presented in Table 18.1. The first column in the table gives each individual's outcome score. The second column indicates whether the individual is in the experimental or comparison group. With these data, a statistical significance test can be performed using a t test. The results of such a test are a t value, degrees of freedom (df), and an obtained p value. By convention, the 5 percent *level of statistical significance* is used. This means that if the obtained p value is less than or equal to .05, the results are judged to be statistically significant; otherwise, they are not. For the data in Table 18.1, the results are $t = 3.60$, $df = 8$, $p = .007$. Because $p < .05$, one would conclude that the mean difference between the scores in the experimental and comparison conditions was statistically significant, and therefore, larger than could reasonably be expected by chance. In other words, the mean difference between the groups provides evidence in favor of a treatment effect.

Table 18.1. Data from a Hypothetical Randomized Experiment.

Outcome Score	*Group*
20	Experimental
24	Experimental
27	Experimental
18	Experimental
23	Experimental
16	Comparison
19	Comparison
10	Comparison
15	Comparison
11	Comparison

It is possible for a mean difference between treatment groups to be statistically *in*significant, even though a treatment effect is present, simply because the treatment effect is small relative to the background noise of chance differences. The *power* of a statistical significance test is a measure of the test's ability to detect small treatment effects when they are present. One way to increase the power of a statistical significance test is to increase the size of the sample (that is, the number of individuals included in the randomized experiment). Power also can be increased by adding covariates to the analysis, as described below. Wise researchers verify that the power of their statistical test is adequate given the size of the treatment effect that is likely to arise. Kraemer and Theimann (1987) provide computational procedures for calculating by hand the power of simple statistical significance tests and Borenstein and Cohen (1988) provide a program for calculating power using a computer (also see Cohen, 1977; Lipsey, 1990).

The Size of a Treatment Effect

The results of a statistical significance test reveal whether an estimated treatment effect is larger than reasonably could be expected by chance. But the results of a statistical significance test do not reveal how *large* the treatment effect is. The second task of statistical analysis is to estimate the size of the treatment effect. Without this information, one cannot judge whether the effect is of practical importance and whether the treatment is worth the extra cost and effort required to implement it. To make informed decisions, policymakers need to know the size of treatment effects.

Unfortunately, the size of a treatment effect can never be known exactly. In a randomized experiment, for example, the mean difference between the outcome scores in the experimental and comparison groups is an estimate of the average effect of the treatment. However, this mean difference in outcome scores will not be exactly equal to the average effect of the treatment. The mean difference also will reflect the effects of differences between the groups that were inevitably introduced by the vagaries of random assignment. Because the size of these chance differences cannot be known exactly, neither can the size of the treatment effect. The best that can be done is to estimate the size of the treatment effect within a margin of error for a given level of confidence. The margin of error takes account of the effects of chance differences.

The margin of error depends on the level of confidence that one desires to have in the results. Conventional practice is to use the 95 percent level of confidence when calculating the margin of error. Once the level of confidence is chosen, the treatment effect estimate and the margin of error are packaged together in what is called a *confidence interval.*

For example, an estimate of the average treatment effect for the data in Table 18.1 is equal to the mean difference between the groups on the outcome variable, which is 8.20. The 95 percent margin of error for this estimate can be easily calculated using a computer program and is equal

to 5.26. So a 95 percent confidence interval is equal to 8.20 ± 5.26. This means that we can be 95 percent confident that the average treatment effect in the population from which the sample was drawn is between 2.94 and 13.46, assuming there are no other threats to validity in the study.

It is important to include the confidence interval when reporting the size of a treatment effect to keep readers from being misled. For example, there is a substantial difference between estimating the average effect of a program as 10 plus or minus 5 with 95 percent confidence, and estimating the average effect of a program as 10 plus or minus 50 with 95 percent confidence. In the first case, the effect of the program is almost certainly positive, while in the second there is a good possibility that the program's effect is negative. Presenting the estimate of the program's effect as 10 without reporting a margin of error would fail to convey the appropriate degree of uncertainty about the estimate (Reichardt and Gollob, 1987).

The size of the margin of error for a given level of confidence is called the *precision* of the estimate of the treatment effect. For a given level of confidence, one would like the margin of error to be as small as possible, just as one would like the power of a statistical test to be as high as possible. Like power, precision can be increased by increasing the sample size or by adding covariates, as described later.

Discovering and Removing Biases

The estimate of an effect can be biased by a variety of threats to validity. It is important to recognize these potential sources of error and to try to remove their biasing effects. A recurring theme in the present chapter will be the value of drawing pictures of the data both to get a feel for the information they contain and to spot potential sources of bias.

Although most of what follows concerns statistical analysis, the reader should keep in mind that fancy statistical procedures may not be the most efficient means of reducing or removing biases. For this purpose, thoughtfulness in data collection often is superior to sophistication in statistics. Glass (1988) provides an illustration of this principle based on the data in Figure 18.1. This figure shows the enrollment in Denver public schools from 1928 to 1975. As marked by the arrow, court-ordered desegregation began in 1969. The research question of interest is how much of the decline in enrollment following 1969 was due to the court mandate and subsequent "white flight." No degree of statistical machination could answer this question satisfactorily using the data in the figure. However, the question could be answered well with a few thoughtfully chosen refinements in data collection. As Glass (1988, p. 460) explains, this question

> could be resolved fairly conclusively by breaking down and plotting in several alternative ways the total enrollment series in Figure [18.1]. Breaking the enrollment data down by grade might cast a little light on things. If it's really white flight that

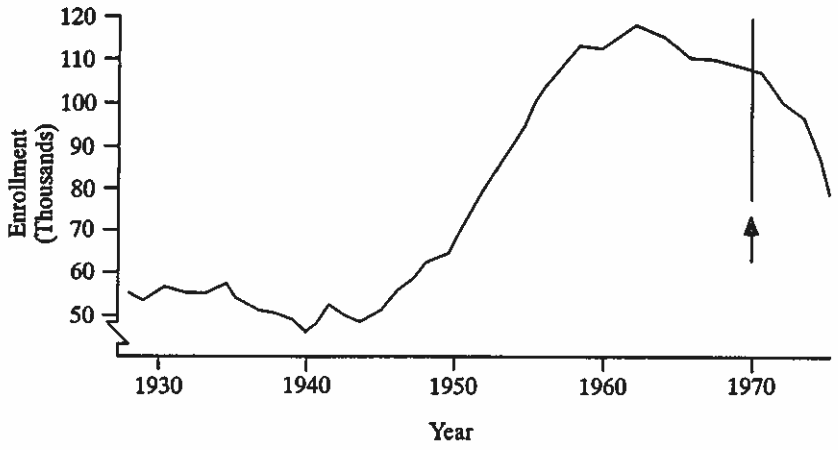
is causing the decline, one might expect a larger decline at the elementary grades than at the secondary grades, particularly grades 11 and 12 where parents would likely decide to stick it out for the short run. If enrollment data existed separately for different ethnic groups, these time series would provide a revealing test. If they showed roughly equal declines across all ethnic groups, the "white flight" hypothesis would suffer a major setback. Data on enrollment that could be separated by individual school, neighborhood, or census tract would be exceptionally valuable. These various units could be ranked prior to looking at the data on their susceptibility to white flight. Such a ranking could be based on variables like "pre-1969 ethnic mixture," or "mobility of families based on percentage of housing values mortgaged or amount of disposable income." If the large enrollment declines fell in the highly susceptible regions, the pattern would constitute some degree of support for the white flight hypothesis.

As you conduct statistical analyses, ask yourself the following questions. Is the effect statistically significant? How large is the effect? How might the estimate of the effect be biased and how can these biases be removed?

Randomized Experiment

In a randomized experiment, individuals are assigned to treatment conditions at random. After the different treatments are administered, the indi-

Figure 18.1. Enrollment in Denver Public Schools.



Note: The arrow marks the beginning of court-ordered desegregation.
Source: Glass, 1988. Copyright 1988 by the American Educational Research Association. Reprinted by permission of the publisher.

viduals are assessed on an outcome variable. The difference between the mean of the outcome scores in the different treatment groups is an estimate of the average effect of the different treatments. A test of the statistical significance of the estimate and a confidence interval for the size of the effect can be calculated as described above. Detailed discussion of randomized experiments is provided earlier in this volume in Chapter Eight by Dennis.

In interpreting the results of a randomized experiment, it is useful to draw a picture of the distribution of the outcome scores for each treatment group separately and to examine these pictures to get a sense of how the scores differ. One thing to look for is a difference between the treatment groups in the variability of the scores. This is evidence that the effect of the treatment varies across different individuals, a topic discussed further below. In addition, standard statistical significance tests and confidence intervals assume that the variability is roughly equal in the two groups. If the variability of the outcome scores appears dramatically unequal across the groups *and* if the sample sizes in the groups are dramatically different, alternative statistical procedures (for example, an unpooled-variance *t* test) might be appropriate. A statistical consultant can help with these determinations.

It is also useful to examine the pictures of the data to learn whether the distributions are symmetric or skewed. A positive skew means that many scores are piled up at the low end of the distribution, with scores trailing off at the high end so there are some high scores that are quite far removed from the rest of the pack. Income and net worth, for example, are usually positively skewed since most incomes pile up at the low end but a few people have quite high incomes. Negative skew is the opposite; most of the scores pile up at the high end with scores trailing off at the low end.

When a distribution of scores is skewed, the mean can be a poor way to characterize the center of the scores and so perhaps should not be used for estimating a treatment effect. The mean also can be a poor way to characterize the center of a distribution if there are a few very extreme scores (outliers), especially if the sample size is small. If the mean is suspect for either reason, calculate the median (that is, the score at the 50th percentile) because it usually is more representative of the center when a distribution is skewed or has outliers. Then see whether the difference between the medians tells the same story as the difference between the means. Also try repeating the analysis using the means but with the outliers removed. If the difference between the medians is dramatically different from the difference between the means or if the results change with the outliers removed, you may want to use alternative methods for calculating statistical significance tests and confidence intervals. Possible alternatives might be nonparametric procedures or procedures using trimmed distributions. See a statistical consultant for assistance.

Including Pretests in the Study

In contrast to the outcome (or posttest) measure that is collected after the different treatments are administered, a pretest measure is collected before

the treatments are administered. Pretest measures need not be included in a randomized experiment, but there can be substantial advantages to including them.

Checking Random Assignment. In field studies, the random assignment of individuals to treatment conditions is often corrupted (Boruch and Wothke, 1985; Braucht and Reichardt, 1993; Conner, 1977). If random assignment was successfully implemented, the distributions of pretest scores should be similar across the treatment groups. However, if random assignment was corrupted, the distributions might be quite different. Therefore, pretest scores can be used to check the integrity of the random assignment procedure. If random assignment appears to have been compromised, it might be necessary to use the analysis strategies described in the section on the nonequivalent comparison group design.

Coping with Differential Attrition. Some of the participants in the study might drop out before the outcome measure is collected. Such attrition is a potential source of bias in the study, especially if the rate of attrition differs across the treatment groups. In particular, bias is introduced if individuals who would score high (or low) on the outcome measure tend to drop out from one treatment group more (or less) than from the other treatment group. Pretest measures are necessary to try to correct for this bias.

Understanding the nature of any differential attrition requires comparing, across the treatment conditions, the pretest scores of individuals who dropped out of the study with the pretest scores of the individuals who did not drop out. Taking account of differential attrition requires making adjustments in the outcome scores based on differences between the groups on the pretest scores. The methods for making these adjustments are the same as the methods for analyzing data from the nonequivalent comparison group design, which is described below. The best course of action is to try to avoid attrition as much as possible.

Increasing Power and Precision. Including one or more pretest measures in a statistical analysis as covariates can increase the power or the precision of the results. One such analysis is called an analysis of covariance. In regression terminology, you regress the outcome measure onto both the pretest measure and a variable representing treatment-group membership. Alternatively, pretests could be used as blocking variables rather than covariates, but this is a bit more complicated and will not be considered here (see Reichardt, 1979). In either case, you might want to ask a statistical consultant for assistance with the analysis.

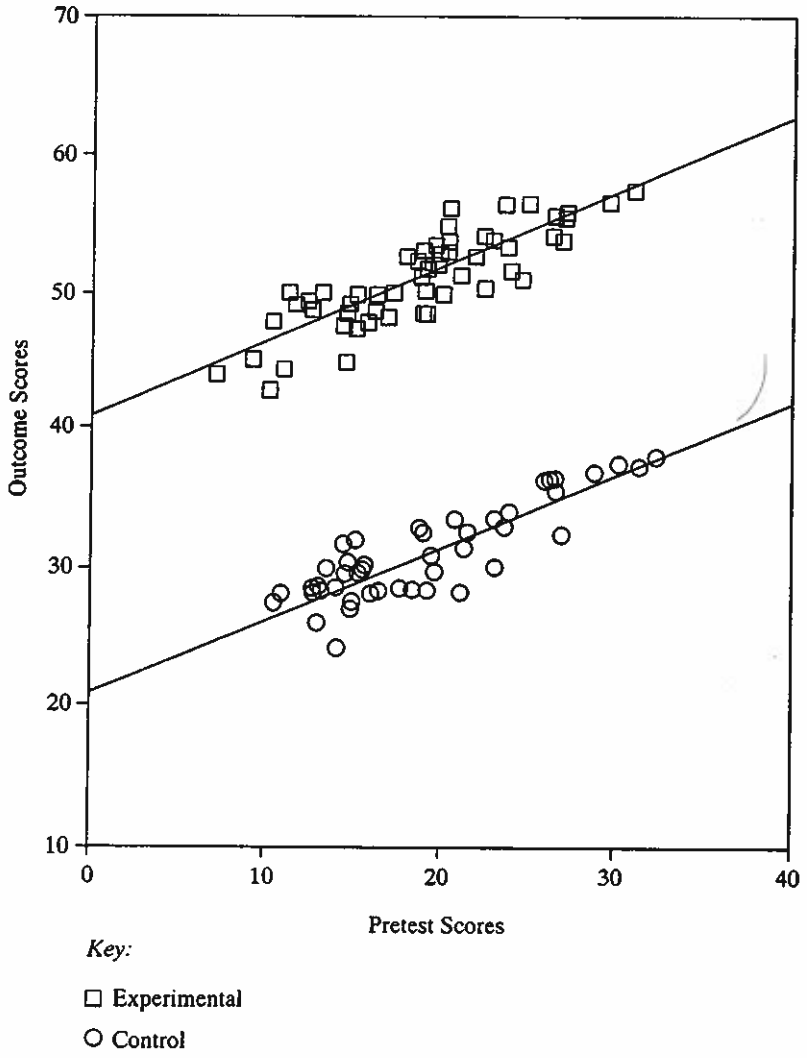
The increase in power and precision that can be obtained by including a pretest in the analysis as a covariate depends on the correlation between the pretest and the outcome measure. The higher the correlation, the greater the power and precision. For example, adding a pretest that correlates .5 with the outcome measure increases power and precision as much as increasing the sample size by 33 percent. Adding a pretest correlated .75 with the outcome measure increases power and precision as much as increasing the sample size by 128 percent. Since collecting data on a pretest often

is less expensive than increasing the sample size, it is worthwhile to spend some time contemplating the types of pretest measures that are likely to be highly correlated with the outcome. Often a pretest that is operationally identical to the posttest is the best choice.

Assessing Treatment-Effect Interactions. In addition to estimating the average effect of the treatment, it is also valuable to study treatment-effect interactions, which means studying how the size of the treatment effect varies across different types of individuals. The meaning of a treatment-effect interaction can perhaps best be understood graphically.

Figure 18.2 presents a scatterplot of the results of a hypothetical randomized experiment. Outcome scores vary along the vertical axis while

Figure 18.2. Data from a Hypothetical Randomized Experiment with a Positive Treatment Effect and No Treatment-Effect Interaction.



pretest scores vary along the horizontal axis. The scores for individuals in the experimental group are denoted by squares. The scores for individuals in the control group are denoted by circles. The regression line for the regression of the outcome scores on the pretest scores is drawn in the figure for each group separately. The upward slope of the regression lines means that individuals who were high on the pretest also tend to be high on the posttest or outcome, and vice versa.

Notice that the mean of the squares and circles along the horizontal (pretest) dimension are close to equal (that is, the groups are not displaced horizontally). This shows that individuals were randomly assigned to the treatment groups. Also notice that the squares are higher than the circles on the vertical (outcome) dimension (that is, the regression lines are displaced vertically). This reflects the effect of the treatment. The squares are about twenty points higher on the outcome variable than the circles, revealing an average treatment effect of about twenty points. The 95 percent confidence interval for the average treatment effect in the plotted data runs from 19.7 to 21.1. Notice that the regression line in the experimental group is also about twenty points higher than the regression line in the control group. When the pretest is added to the analysis as a covariate, the treatment effect is literally estimated as the vertical displacement between the regression lines, rather than as the difference between the outcome means in the two groups, as would be the case without the pretest.

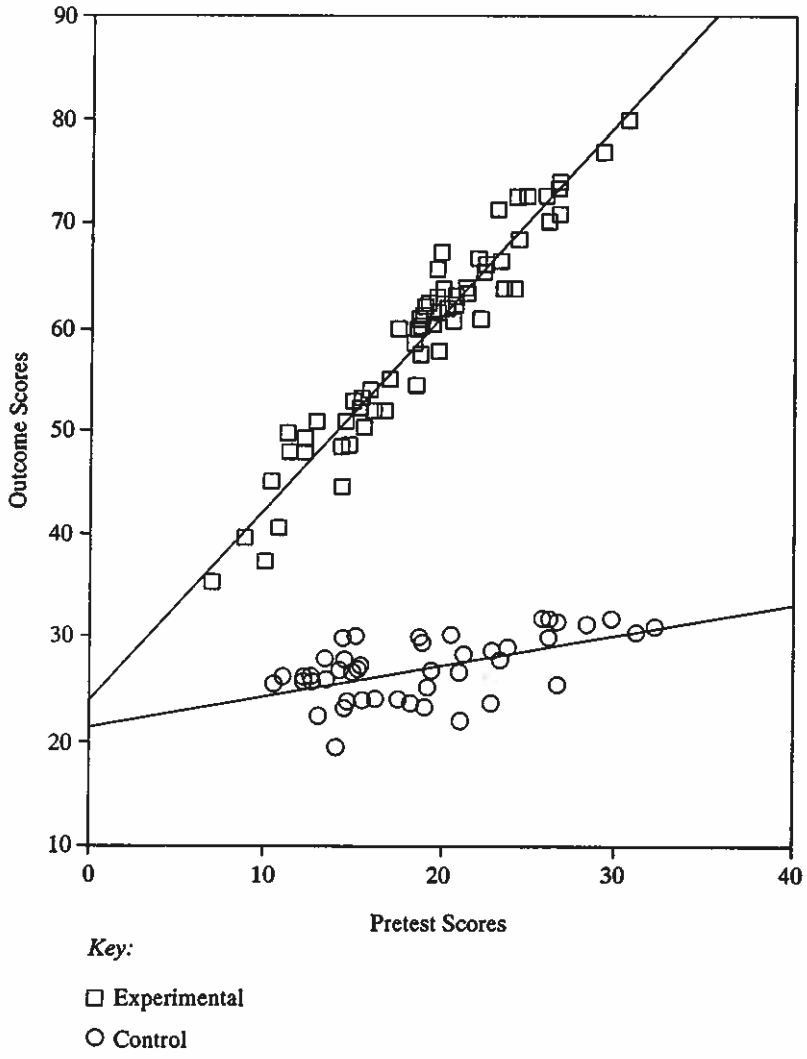
Also notice that the treatment effect is the same regardless of the individual's pretest score. For example, the treatment effect is about twenty points both for individuals with relatively high pretest scores (say 30) and for individuals with relatively low pretest scores (say 12). This effect is readily apparent from the observation that the regression lines are parallel, indicating that the effect of the treatment does *not* interact with the pretest scores.

Now consider Figure 18.3. The squares are higher than the circles and the regression line for the experimental group is displaced above the regression line for the control group, both of which reflect the average effect of the treatment. But the size of the treatment effect varies with the individual's pretest score. Individuals with high pretest scores (say 30) have a treatment effect of about fifty points (the 95 percent confidence interval runs from 46.6 to 51.5), while individuals with low pretest scores (say 12) have a treatment effect of about twenty points (the 95 percent confidence interval runs from 19.6 to 23.1). This result shows that the effect of the treatment interacts with the pretest and is readily apparent from the observation that the regression lines are not parallel.

In Figure 18.4, the interaction between pretest and treatment is even more extreme. The effect of the treatment is positive in the population on average and for individuals with high pretest scores, but the treatment effect is negative for individuals with low pretest scores.

The implication is that the data analyst needs to pay attention to both the average and the interactive effects of the treatment if appropriate policy implications are to be drawn. For example, although a novel teaching method

Figure 18.3. Data from a Hypothetical Randomized Experiment with a Treatment-Effect Interaction.

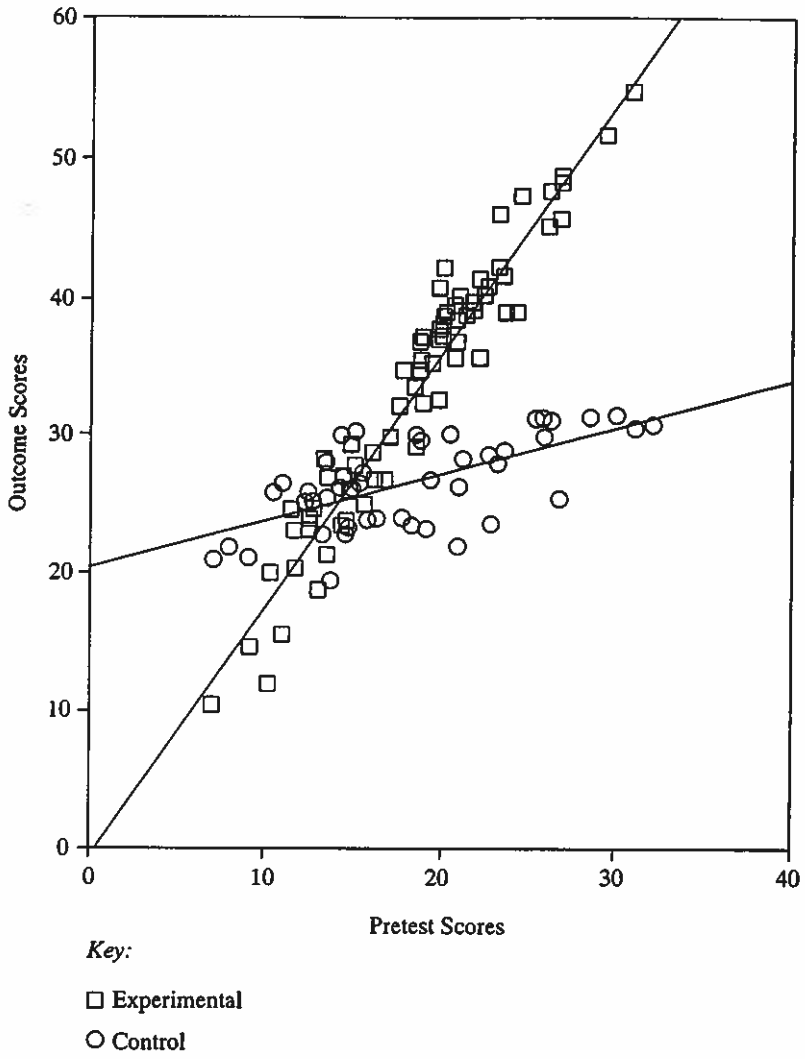


might be superior to the old teaching method on average, the old method might nonetheless be superior for low-ability students. In this case, it would be better to tailor the teaching method to the type of student rather than to apply the new teaching method blindly to all students.

Outliers and Curvilinearity

When pretest measures are included in the analysis, it is important to plot the data, as in Figures 18.2 through 18.4, and to examine both the plots and the fit of the regression lines. Look for interactions so they can be taken

Figure 18.4. Data from a Hypothetical Randomized Experiment with a Crossover Treatment-Effect Interaction.



into account in the analysis. Also look for outliers and evidence of curvilinearity. *Curvilinearity* means that the regression "lines" that best fit the data are curved rather than straight. An outlier can make it appear as if either an average effect or an interaction is present when it is not. If outliers are present, try removing them and repeating the analysis to see how much difference the outliers make. Curvilinearity that is not recognized can hide an interaction or lower the power and precision of the analysis. Curvilinearity can be taken into account either by transforming the data or by polynomial regression. More details are given in any standard regression text (for example, Hamilton, 1992; Draper and Smith, 1981), or see a statistical consultant.

An Example

In a simple randomized experiment described in Ryan, Joiner, and Ryan (1985), ninety-two students in an introductory statistics class recorded their pulse, height, weight, gender, how much they smoked, and how much they typically exercised. Each student then flipped a coin to determine his or her treatment assignment. Heads meant they were to run in place for a minute; tails meant they were to rest quietly. A minute later, the students took their pulse again. The resulting data are distributed with the Minitab computer software program and available in Ryan, Joiner, and Ryan (1985).

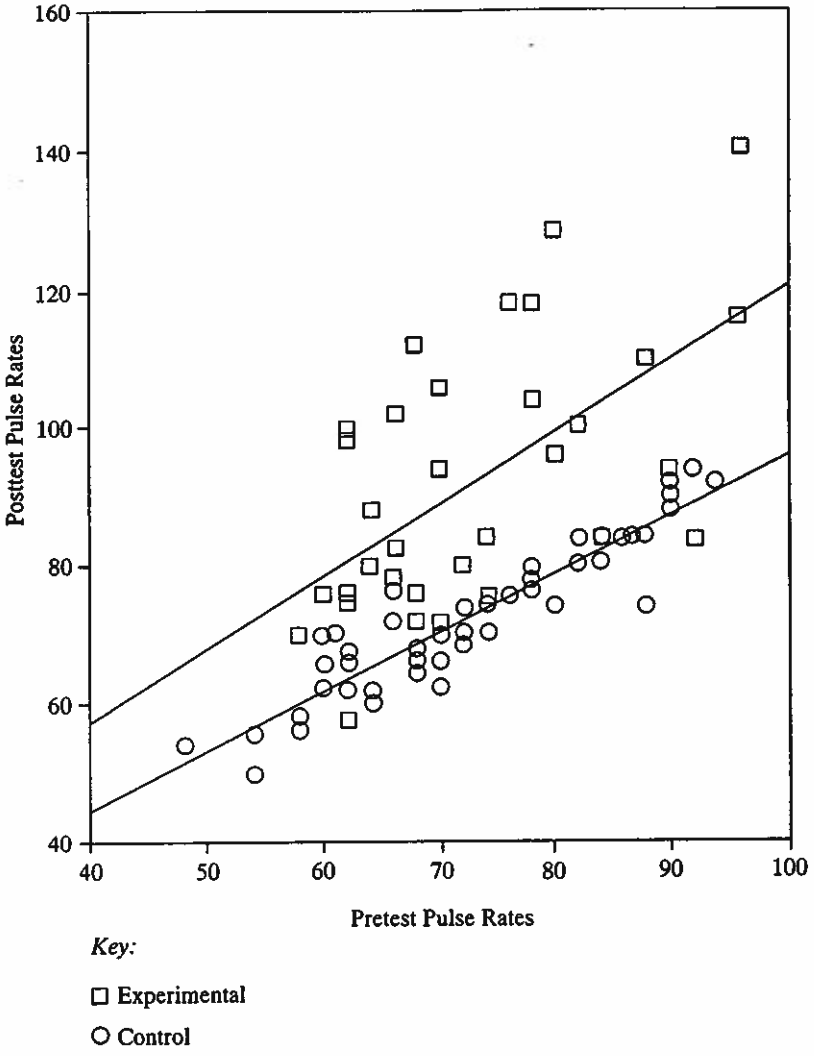
An examination of the data revealed an error requiring that one student's data (in row 54) be omitted from the analysis. For the remaining data, plots of the distributions of the pretest pulse data in the two treatment groups looked reasonably similar. However, there were statistically significant differences between the treatment groups on height ($t = 2.5$, $df = 89$, $p = .01$) and weight ($t = 2.29$, $df = 89$, $p = .03$), and there were significantly fewer women in the experimental group ($N = 10$) than in the control group ($N = 24$) according to a chi-square goodness-of-fit test ($\chi^2 = 5.76$, $df = 1$, $p = .02$). The corresponding sample sizes for the men were 24 and 32, respectively, and this difference was not statistically significant. It appears that the random assignment might have been somewhat compromised by women who chose not to run in place even though their coin showed heads.

The mean of the pulse rates at posttest was 72.3 in the control group and 91.9 in the experimental group. This mean difference of 19.6 beats per minute was statistically significant ($t = 6.48$, $df = 89$, $p < .001$, 95% confidence interval = 13.5 to 25.5). Adding the pretest pulse as a covariate in an analysis of covariance reduces the width of the confidence interval for the size of the effect by 29 percent (treatment effect estimate = 19.15; 95% confidence interval = 14.89 to 23.41). Based on either analysis, it is clear that running in place significantly raised the pulse in this population of individuals. A plot of the posttest pulse versus the pretest pulse for each group is given in Figure 18.5. This plot suggests that there is no interaction between the treatment and the pretest pulse ($t = .92$, $df = 87$, $p = .36$).

However, there is a statistically significant interaction between the treatment and weight ($t = 3.74$, $df = 87$, $p = .0003$). This interaction is revealed in the plot of the posttest pulse versus weight for each group in Figure 18.6. The interaction means that the treatment has a smaller effect for heavier individuals than for lighter individuals. In particular, for each pound increase in weight, the effect of the treatment on posttest pulse is reduced on average by .44 beats per minute (95% confidence interval = .21 to .67).

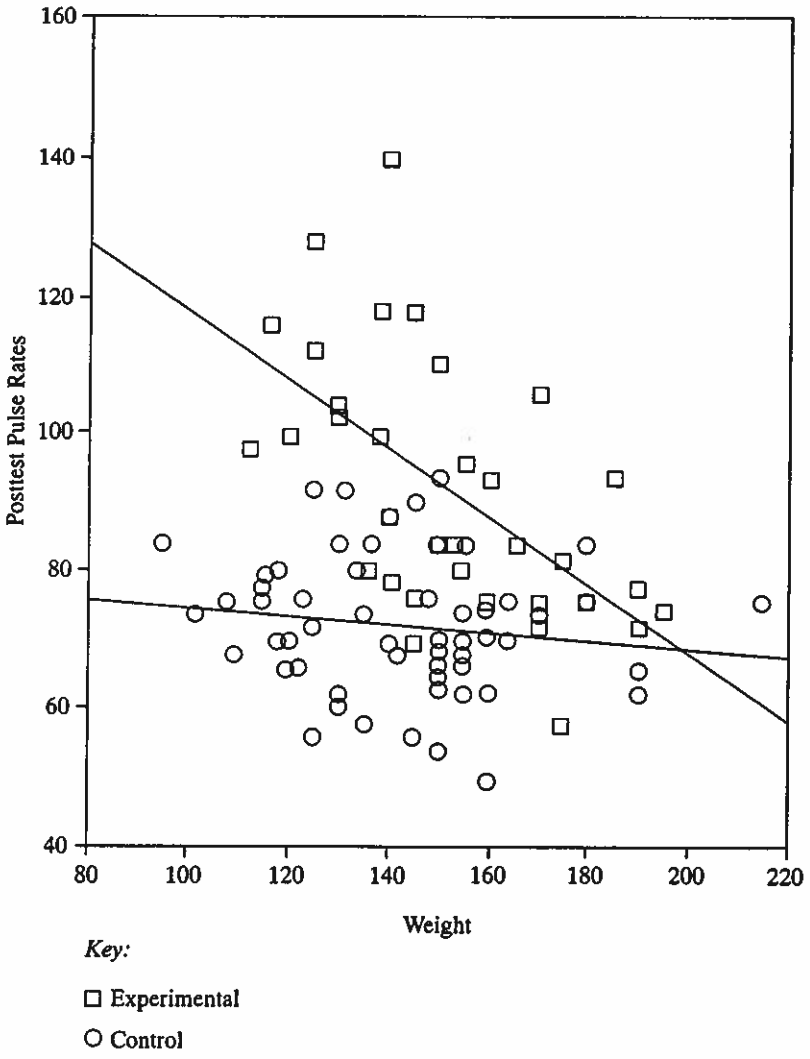
This interaction is probably due to a confounding between weight and gender. As further examination of the data reveals, the interaction arises because running in place has less effect on men (treatment effect estimate = 12.9) than women (treatment effect estimate = 34.4), and men tend to weigh

Figure 18.5. Data from the Pulse Study (Ryan, Joiner, and Ryan, 1985) Showing a Positive Treatment Effect and No Interaction Between the Treatment and the Pretest Pulse Rate.



more than women. In addition, the interaction between treatment and weight is not statistically significant when the data for each gender are analyzed separately. Thus the plot in Figure 18.6 could easily be misleading if one were not careful to examine the data in greater depth. In addition, if extrapolated beyond the reasonable range of the data, the regression lines in Figure 18.6 would suggest that running in place actually slows the pulse in very heavy individuals. The moral is that you need to keep your wits about you when analyzing data.

Figure 18.6. Data from the Pulse Study (Ryan, Joiner, and Ryan, 1985) Showing an Interaction Between the Treatment and Weight.



Concluding Comments on the Randomized Experiment

Although randomized experiments can be biased, especially by differential attrition, they are a potentially powerful tool for estimating treatment effects. When well implemented, a randomized experiment can produce results that are more credible and precise than the results from any other design. Evaluators implementing a randomized experiment in a field setting will find it useful to collect pretest measures that are highly correlated with the outcome measure so as to increase the precision of the estimate of the treatment effect. It is also useful to collect pretest measures to assess (1) the

integrity of the random-assignment process, (2) differential attrition, and (3) treatment-effect interactions.

Regression-Discontinuity Design

In the regression-discontinuity design, individuals are assigned to treatment conditions based on their scores on a quantitative pretest measure. Specifically, a cutoff score on the pretest is specified and individuals with pretest scores above the cutoff are assigned to one treatment condition while individuals with pretest scores below the cutoff are assigned to the other treatment condition. After individuals are assigned to treatment conditions, the different treatments are administered, and each individual is assessed on an outcome measure. Detailed discussion of the regression-discontinuity design is provided by Marcantonio and Cook in Chapter Seven of this book.

To estimate the average effect of the treatment, a separate regression line is fitted to the data on each side of the cutoff score. The vertical displacement between the two regression lines at the cutoff point is an estimate of the effect of the treatment for individuals near the cutoff point. For example, consider Figures 18.7 and 18.8. Both figures contain scatterplots of the outcome scores versus the pretest scores. The vertical line denotes the cutoff point on the pretest. The squares denote the scores of individuals who receive the treatment (individuals in the experimental group). These individuals all had scores on the pretest that fell below the cutoff. The circles denote the scores of individuals who do not receive the treatment (individuals in the control group). These individuals all had scores on the pretest that fell above the cutoff. Separate regression lines have been fitted to the data in each group and are plotted in the figures.

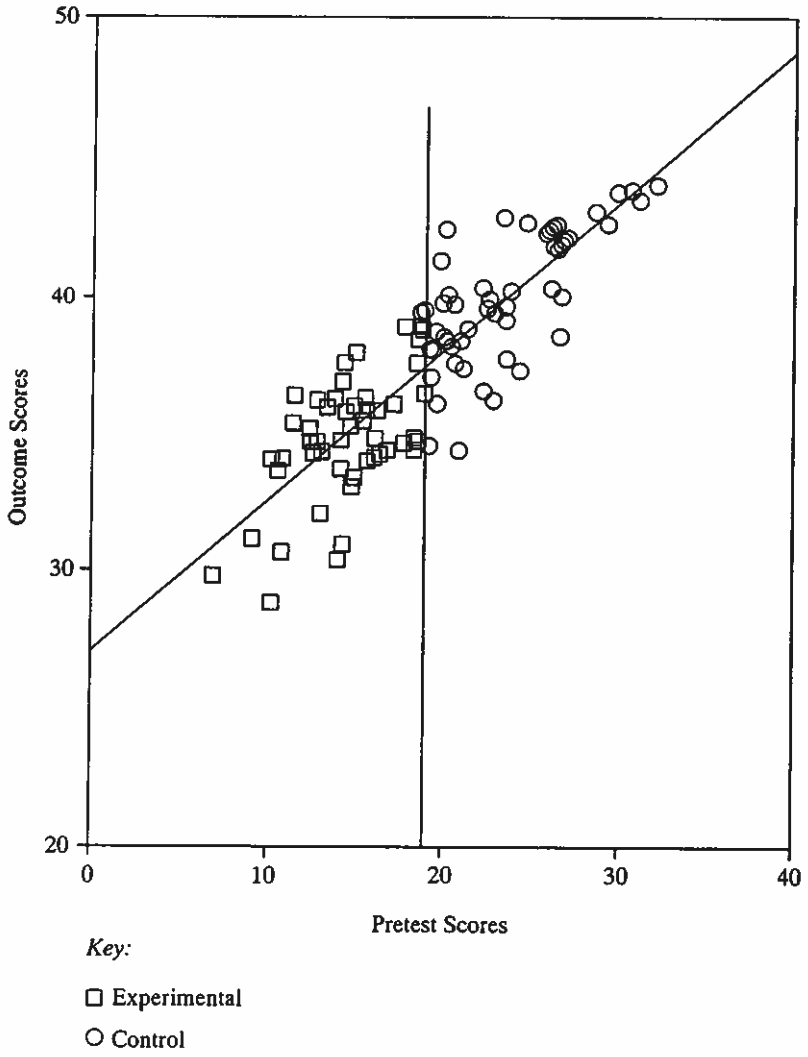
In Figure 18.7, there is no treatment effect. As a result, the two regression lines fall on top of one another, indicating that there is no vertical displacement or break between the lines at the cutoff point. In Figure 18.8, there is a positive treatment effect. As a result, the regression line for the experimental group is displaced above the regression line for the control group. The vertical displacement between the two lines is the estimate of the size of the treatment effect at the cutoff point.

Curvilinear Regression Lines

The regression lines must be properly fitted to the data in both the experimental and control conditions; otherwise a bias in the estimate of the treatment effect can occur. In particular, a bias can occur if the true relationship between outcome and pretest is curvilinear in one or both treatment groups but a straight regression line is fitted to the data.

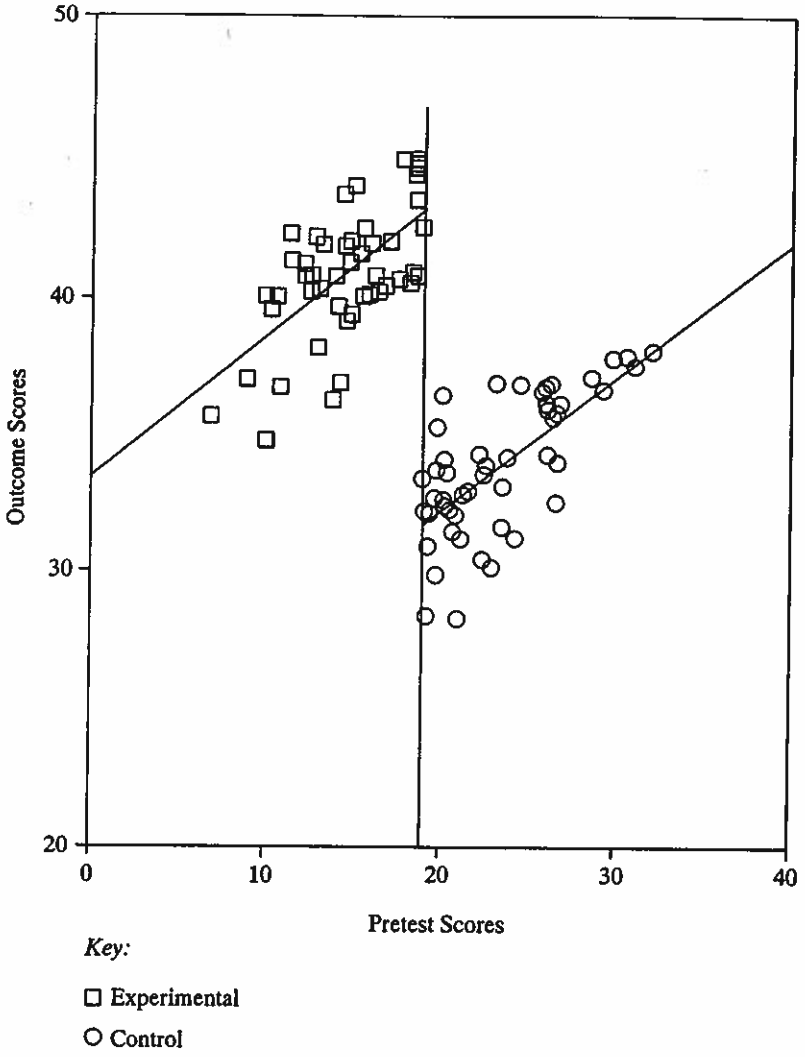
The biasing effect of curvilinearity is demonstrated in Figure 18.9. In this figure, there is no scatter in the data around the true (curvilinear) regression line, meaning that the outcome scores can be perfectly predicted

Figure 18.7. Data from a Hypothetical Regression-Discontinuity Design Where the Treatment Has No Effect.



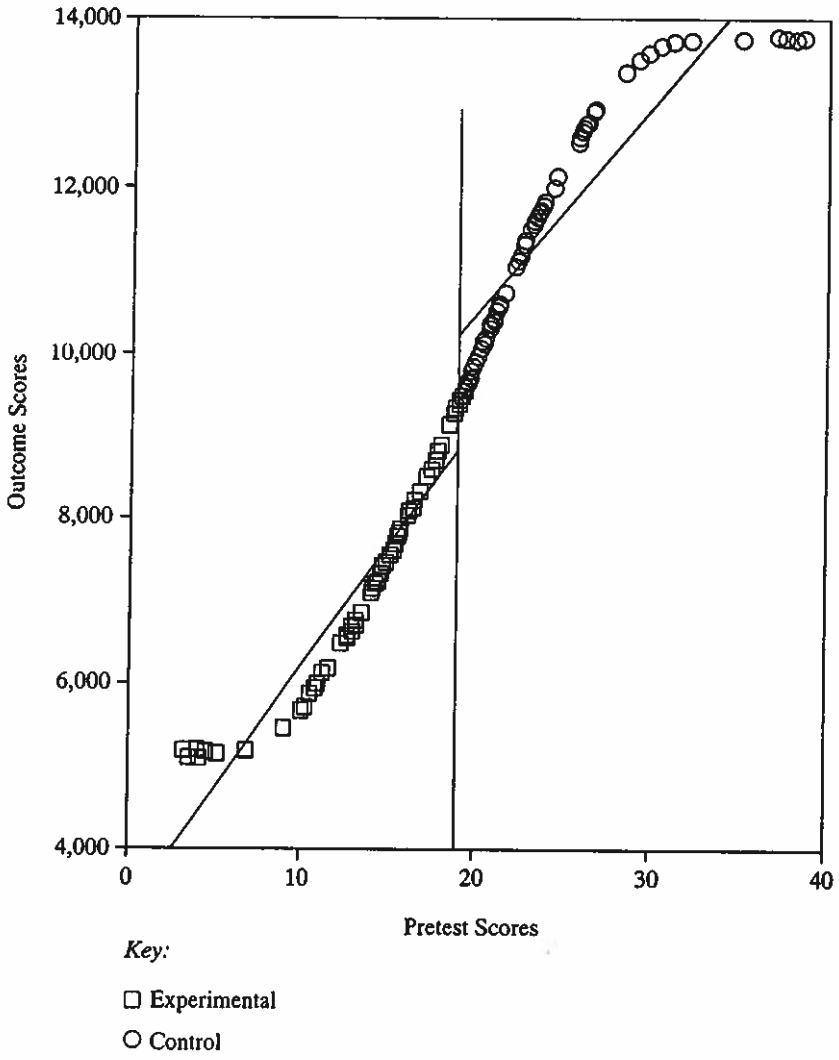
from the pretest scores. This is unrealistic, but to include scatter might make the main point of the example less clear. Also notice that the true relationship between the pretest and outcome scores is curvilinear. Since there is no break in the data at the cutoff point, there is no treatment effect. However, if linear regression lines were fitted to the data as shown in the figure, there would be a break between the lines at the cutoff point. This means that an analysis of the data using linear regression lines would find a treatment effect when in fact there is none. Only if the curvilinearity in the data were properly modeled would the analysis reach the correct conclusion about the absence of a treatment effect.

Figure 18.8. Data from a Hypothetical Regression-Discontinuity Design Where the Treatment Has a Positive Effect.



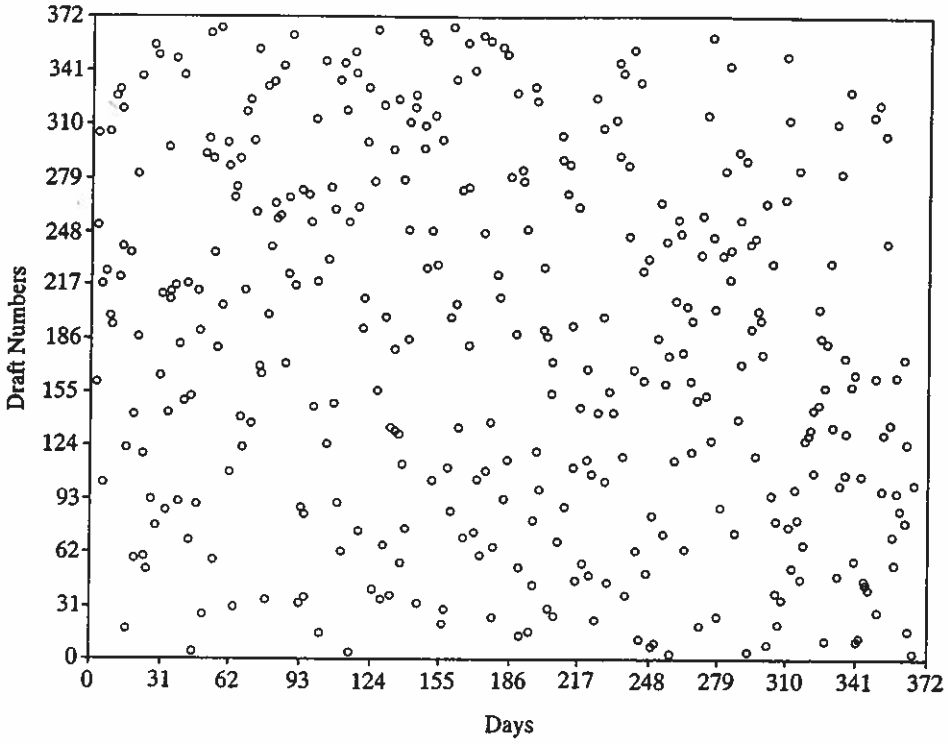
To determine the shape of the true relationship between the outcome and pretest, it helps to plot the data. If the data have a lot of scatter around the true regression lines (unlike the plot in Figure 18.9), sometimes the shape of the relationship can be more easily discerned by adding a *median trace.* A median trace is created by dividing the scatterplot into vertical columns (either of equal width or of equal numbers of data points) and calculating the median of the outcome scores within each column. These medians are then plotted on top of the scatterplot. Often a median trace reveals the nature of the relationship more clearly than the scatterplot of the original data alone.

Figure 18.9. Data from a Hypothetical Regression-Discontinuity Design Where the True Regression Line is Curvilinear but Linear Regression Lines Have Been Fitted to the Data.



An example of a median trace is presented in Figures 18.10 and 18.11 (Moore and McCabe, 1989). In the early 1970s, American men were subject to a military draft conducted by lottery. Priority numbers were supposed to be assigned at random according to the day of birth. Controversy arose over the randomness of the assignment of priority numbers in 1970. Figure 18.10 plots the draft priority number on the vertical axis and the day of birth on the horizontal axis. A quick look at this plot suggests that no relationship exists between the two variables, as would be the case if the lottery were random. However, the median trace plotted by month in Figure

Figure 18.10. Plot of the Selective Service Draft Priority Numbers Versus Day of Birth for 1970.



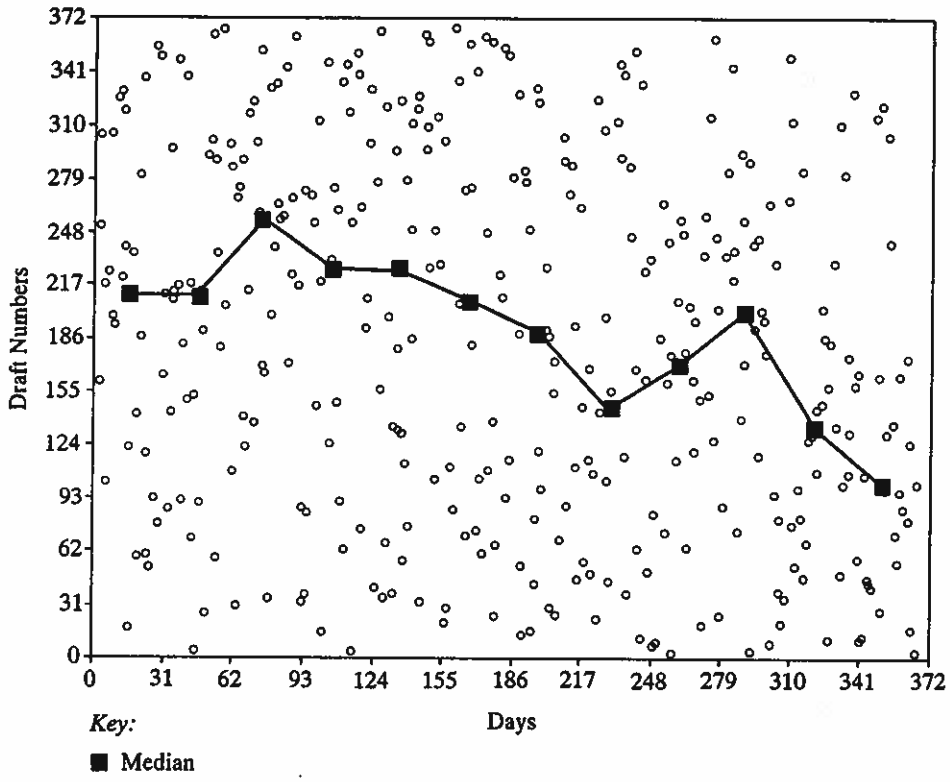
Source: From INTRODUCTION TO THE PRACTICE OF STATISTICS, by David Moore and George McCabe. Copyright © 1989 by W. H. Freeman and Co. Reprinted by permission. Data from Feinberg, 1971.

18.11 shows a clear downward slope, which reveals the true relationship that exists in the data (men born later in the year tend to have lower priority numbers) and provides confirming evidence of the faulty randomization procedure.

Another technique for assessing the shape of a regression line is to smooth the data in the scatterplot using a moving average (or moving median). A moving average of length five, for example, is generated by taking the average of the outcome scores for the individuals with the lowest five pretest scores. This average is plotted on the scatterplot above the third lowest pretest scores. The lowest pretest score is then dropped and the average of the outcome scores for the individuals with the five next lowest pretest scores is calculated and plotted above the fourth lowest pretest score, and so on. A moving average can also be calculated for other lengths to determine which one is the most revealing.

If curvilinearity is detected, curvilinear rather than straight regression lines need to be fitted. This change can be accomplished by transforming

Figure 18.11. A Plot of the Selective Service Draft Priority Numbers Versus Day of Birth for 1970 with a Median Trace by Month.



Source: From INTRODUCTION TO THE PRACTICE OF STATISTICS by David Moore and George McCabe. Copyright © 1989 by W. H. Freeman and Co. Reprinted by permission. Data from Feinberg, 1971.

the data or by using polynomial regression. In either case, a statistical consultant might prove helpful. Usually the fitting process involves a good bit of trial and error. After each trial, it is recommended that the residuals from the regression analysis be plotted against the pretest scores. The residuals are the discrepancies between the data points and the regression line. Plotting them often can help reveal where the regression line fails to fit the data.

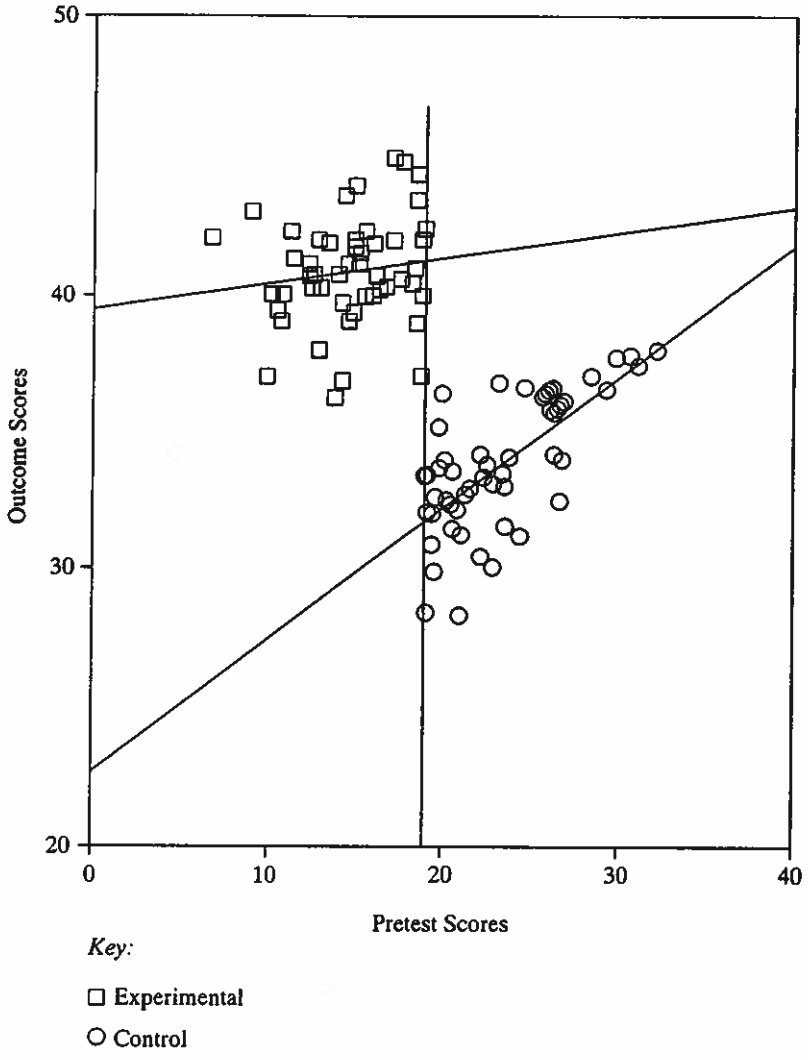
Treatment-Effect Interactions

It is possible that the effect of the treatment is different for individuals with different scores on the pretest. In other words, there may be a treatment-effect interaction with the pretest. An example is given in Figure 18.12. In this figure there is a large displacement between the regression lines at the cutoff point revealing that the treatment has a positive effect for individuals near the cutoff point. The size of the treatment effect varies, however, de-

pending on the individual's pretest score. If both regression lines are extrapolated onto the other side of the cutoff point, as is done in the figure, the treatment effect for individuals with low pretest scores is shown to be much larger than the treatment effect for individuals with high pretest scores.

There is a potential problem here: drawing the conclusion that the treatment effect differs for individuals with different pretest scores requires extrapolating the regression lines as described above, and the further the lines are extrapolated, the greater is the chance for error. The possibility for an increase in error arises because the regression lines are being extrapolated into regions in which there are no data. The regression line for the

Figure 18.12. Data from a Hypothetical Regression-Discontinuity Design with a Treatment-Effect Interaction.



experimental group is being extrapolated onto the other side of the cutoff point where none of the individuals received the treatment. The converse holds for the regression line in the control group. Researchers should place more confidence in the estimate of the treatment effect at the cutoff point than at any other point on the pretest because the estimate at this point involves the least amount of extrapolation and therefore is the most credible and precise.

Nonetheless, it is important that an interaction between the treatment and the pretest be taken into account when fitting the regression lines. To ignore an interaction when one is present (that is, fitting parallel regression lines when the lines are not parallel) can bias the estimate of the treatment even at the cutoff point. Therefore, regression lines must be fitted to take account of an interaction if one is present, but conclusions about the effect of the treatment for individuals with pretest scores other than at the cutoff should be drawn with caution.

Other Sources of Bias

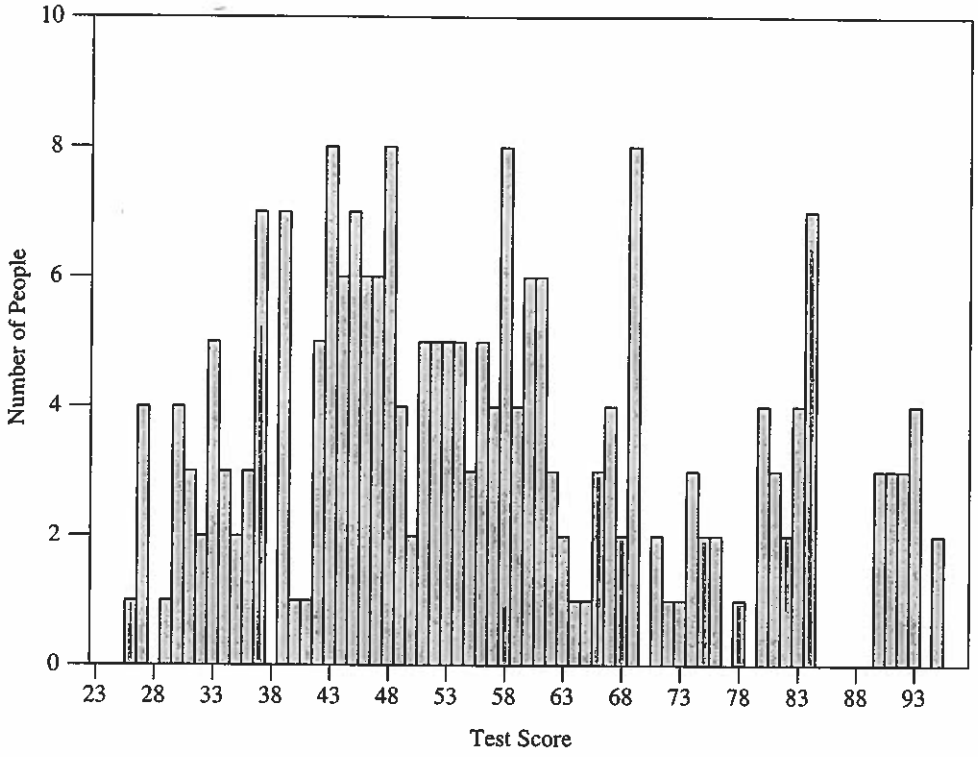
Any source of discontinuity in the regression of the outcome on the pretest scores that is not due to the treatment is a potential source of bias. A bias could be introduced if more individuals in one group drop out of the study than do individuals in the other group. For such reasons it is important to assess the nature and degree of any differential attrition.

A bias also can be introduced if the assignment to treatment conditions is not based on the cutoff score as is assumed. For example, a bias could be introduced if individuals with pretest scores on the "wrong" side of the cutoff are able to alter or lie about their pretest scores so as to be admitted to the treatment group. Evidence that this has occurred might be obtained by plotting the distribution of the pretest scores and looking for gaps or dips in the distribution near the cutoff score. Such manipulation of cutoff scores was alleged to have occurred in a civil service examination for engineering positions in Chicago in 1966 (Freedman, Pisani, Purves, and Adhikari, 1991, p. 51). There were fifteen job openings and 223 applications. A plot of the distribution of the examination scores is given in Figure 18.13. The substantial gap between the highest fifteen scores and the rest of the scores in the distribution suggests that some were altered.

An Example

The study, described earlier, of the effect on subjects' pulse rate of running in place was a randomized experiment. However, this study could be turned into a regression-discontinuity design simply by deleting data based on a cutoff score. Suppose that individuals had been assigned to treatment conditions using a cutoff score based on their initial pulse rate. In particular, suppose that all individuals with a pretest pulse rate higher than 70 had been

Figure 18.13. The Distribution of Scores on a Civil-Service Examination in Chicago in 1966.

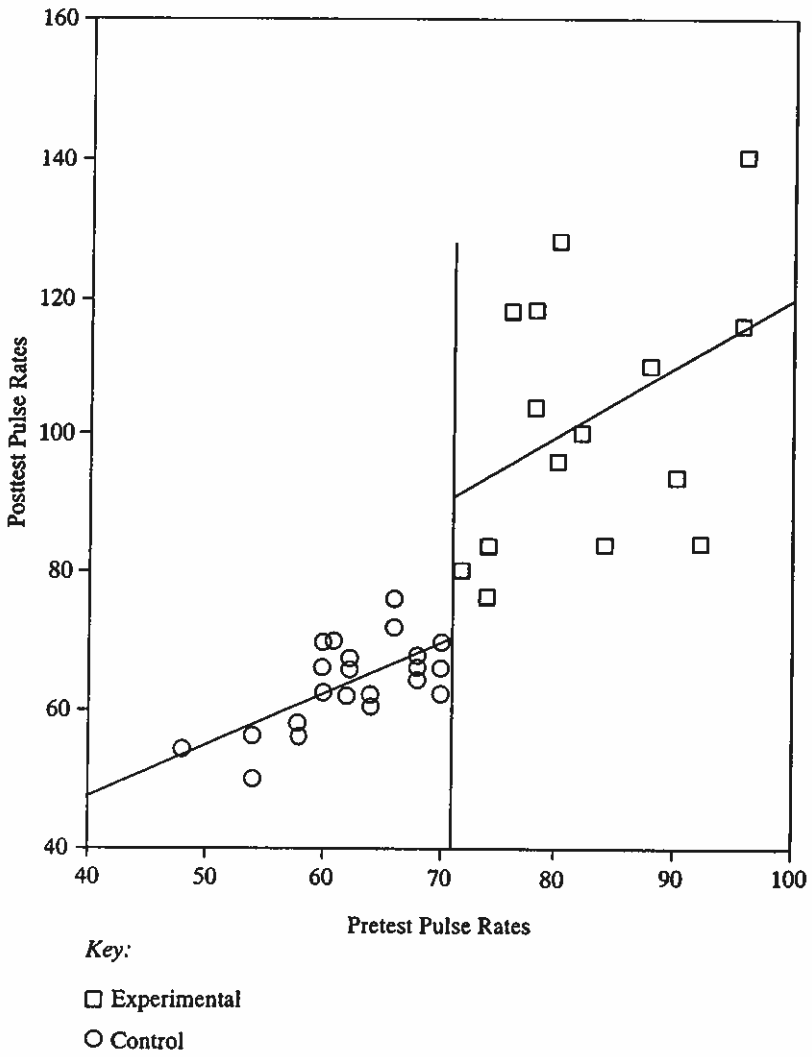


Source: Freedman, Pisani, Purves, and Adhikari, 1991, p. 51. Reprinted by permission of W. W. Norton and Company.

assigned to the experimental condition; as a result, the data from individuals in the experimental condition who had a pretest pulse of *70 or below* will be ignored. Conversely, suppose that all individuals with a pretest pulse rate of 70 or below were assigned to the control condition; in this case, the data from individuals in the control condition who had a pretest pulse *above 70* will be ignored. This assignment produces the data in Figure 18.14 which, for all intents and purposes, is a regression-discontinuity design.

With these data, the effect of the treatment can be estimated by the discrepancy between the regression lines at the cutoff point. This estimate is 20.72 (95% confidence interval = 7.4 to 34.0), which is statistically significant ($t = 3.16$, df = 38, $p = .003$). The result agrees well with the results produced when the data from the study were analyzed as a randomized experiment. Note, however, that the estimate of the treatment effect from the regression-discontinuity design is less precise—that is, the confidence interval is wider—than the estimate from the randomized experiment, partly because of the loss of data.

Figure 18.14. Data from the Pulse Study (Ryan, Joiner, and Ryan, 1985) in the Form of a Regression-Discontinuity Design.



Concluding Comments on the Regression-Discontinuity Design

The regression-discontinuity design is particularly well suited for studying treatments that are assigned on the basis of need or merit. In these cases, a quantitative assessment of need or merit can be used as the pretest. As a result, the design can sometimes be implemented when the random assignment of individuals to treatment conditions is not possible.

Nonetheless, the randomized experiment has at least three advantages compared to the regression-discontinuity design. First, the randomized experiment can accommodate some misfitting of the regression model and still

produce a reasonable estimate of the effectiveness of the treatment. This is much less true for the regression-discontinuity design. For example, if the relationship between the pretest and the outcome is curvilinear and this curvilinearity is not correctly modeled, the estimate of the treatment effect at the cutoff point in the regression-discontinuity design can be biased. However, in the randomized experiment, incorrectly modeling curvilinearity can reduce precision and power, but it will not bias the estimate of the average effect of the treatment.

Second, while it is more important that the correct regression model be fitted in the regression-discontinuity design than in the randomized experiment (for the reason just noted), doing so is usually more difficult because data are missing compared to a randomized experiment. A regression-discontinuity design is essentially a randomized experiment with missing data. A comparison of Figures 18.2 and 18.8 shows that the data either above or below the cutoff point are missing for each treatment group in the regression-discontinuity design, unlike the data for the randomized experiment. Because of these missing data, it is far more difficult to be confident about correctly modeling both curvilinearity and treatment-effect interactions in the regression-discontinuity design than in the randomized experiment.

Third, even if the correct regression model is fitted to the data, the estimate of the treatment effect in the regression-discontinuity design will be less precise (and the statistical significance test will be less powerful) than the estimate of the treatment effect in the randomized experiment. Even under ideal conditions, more than two-and-a-half times as many subjects are required in the regression-discontinuity design to have the same degree of precision and power as in the randomized experiment (Goldberger, 1972).

Nonequivalent Comparison Group Design

Unlike the randomized experiment, in the nonequivalent comparison group design, individuals are not assigned to treatment conditions at random. Nor are individuals assigned to treatment conditions according to a cutoff score on a pretest, as in the regression-discontinuity design. Rather, in the nonequivalent comparison group design, individuals are assigned to the treatment conditions in some other, nonrandom fashion. They might self-select themselves into treatment conditions, or researchers might assign the treatments to preexisting groups, such as schools, that were formed previously in a nonrandom fashion. As a result, the nonequivalent comparison group design is often used to study the effects of a disability, or the effects of treatments to which random assignment would be unethical, such as the results of dropping out of school. Further discussion of the nonequivalent comparison group design is provided in Chapter Six of this volume by Rog.

Without random assignment to conditions, the individuals in the different treatment groups can, and usually will, differ in substantial ways. These differences are called *selection differences* and can masquerade as a treatment

effect. Even in the absence of a true treatment effect, the outcome scores in the treatment groups are likely to differ substantially because of initial selection differences. As a result, selection differences are a threat to validity and must be taken into account when data from a nonequivalent comparison group design are analyzed.

The two simplest and most commonly used statistical procedures for taking account of selection differences are analysis of covariance and gain-score analysis. Both procedures use the pretest scores to control the biasing effects of selection differences. Which, if either, procedure is appropriate depends on the circumstances.

Analysis of Covariance

Suppose the two treatment groups differ on the pretest because individuals who have high scores on the pretest tend to be in one group more than the other. Further, suppose that individuals with high scores on the pretest tend to have high scores on the outcome or posttest. Because of these initial selection differences, the groups will differ on the posttest even in the absence of a treatment effect.

Analysis of covariance takes account of the effects of the selection differences by statistically matching individuals on their pretest scores before drawing comparisons between the groups on the outcome scores. In particular, analysis of covariance estimates the average effect of the treatment as the mean difference in outcome scores between individuals from the two treatment groups who are statistically matched on their pretest scores.

This procedure serves to remove selection differences as measured by the pretest. Nonetheless, there are two potential inadequacies in this approach. First, if there are selection differences between the groups that are not measured by the pretest but that influence the posttest, these will not be controlled for by the analysis of covariance and therefore, will still bias the estimate of the treatment effect. The more highly correlated the pretest is with the posttest, the less room there is for selection differences that are not measured by the pretest and that influence the posttest. Therefore, the best pretests to use for removing selection differences are generally those that are highly correlated with the posttest. This usually means using a pretest that is operationally identical to the posttest (Campbell and Boruch, 1975; Cronbach, 1982, points out this will not always be true, however). In addition, the analyst can use more than one pretest in the analysis of covariance. In this case, the analysis of covariance will match on all the pretests that are included in the analysis before drawing comparisons of the outcome scores. But no matter how many variables are included in the analysis, in most instances a reasonable suspicion will remain that not all the sources of selection differences have been taken into account. If the suspicion is correct, the analysis will remain biased by selection differences.

Second, selection differences will not be properly controlled for if any of the pretests that are included in the analysis are measured with

error. Unfortunately, measurement error is ubiquitous in the social sciences. However, procedures have been devised for taking account of measurement error in the pretests (or covariates) in the analysis of covariance. If there is only a single covariate in the regression analysis, all that is required is an estimate of the reliability of the covariate (Campbell and Boruch, 1975; Reichardt, 1979). If multiple pretests are included in the analysis, the most widely used correction procedure requires multiple measures of each covariate and performs the analysis using a structural equation modeling program such as LISREL (Jöreskog and Sörbom, 1988) or EQS (Bentler, 1989). In either case, assistance from a statistical consultant may be required.

Gain-Score Analysis

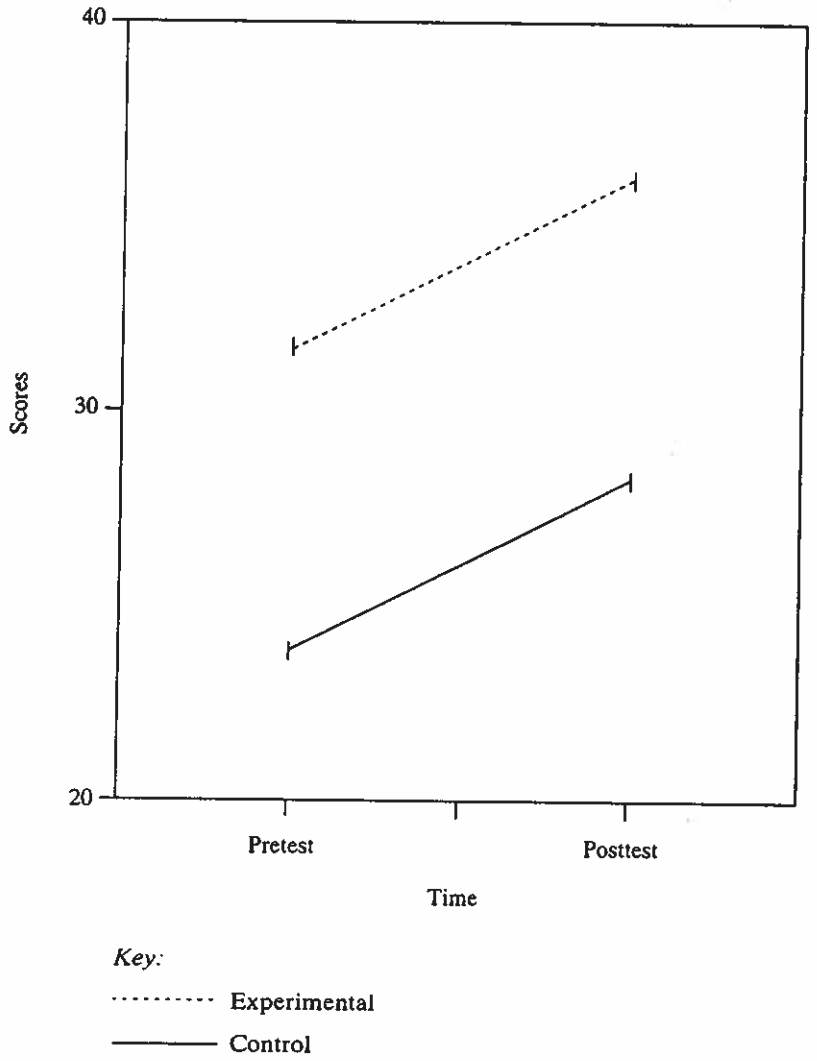
Gain-score analysis requires that the pretest be operationally identical to the outcome or posttest variable. In this case, the pretest can be subtracted from the posttest to create a gain score for each individual. The average effect of the treatment is then estimated as the mean difference in gain scores between the treatment groups.

To understand how gain-score analysis takes account of selection differences, consider Figures 18.15 and 18.16. In both these figures, the mean pretest and posttest scores for the experimental and control groups are plotted. Gain-score analysis assumes that if there is no effect of the treatment on average, the line connecting the pretest mean to the posttest mean in the experimental group will be parallel to the line connecting the pretest mean to the posttest mean in the control group, as in Figure 18.15. In other words, gain-score analysis assumes the treatment has no effect on average if the average gain from pretest to posttest in the experimental group is the same as the average gain from pretest to posttest in the control group.

A treatment effect is present on average only if these two lines are not parallel, as in Figure 18.16. In this case, the average effect of the treatment is what accounts for the difference in the slopes of the two lines. In other words, the treatment effect is the average gain from pretest to posttest in the experimental group minus the average gain from pretest to posttest in the control group.

In both figures, selection differences account for the mean difference between the groups on the pretest. Gain-score analysis assumes that in the absence of a treatment effect, the size of these selection differences will remain the same at the time of the posttest. If, in the absence of a treatment effect, the groups would remain as far apart at the time of the posttest as they were at the time of the pretest, gain-score analysis provides an unbiased estimate of the average treatment effect. On the other hand, if in the absence of a treatment effect the groups would either grow farther apart (for example, the rich getting richer and the poor getting poorer) or come closer together (for example, due to regression toward the mean), gain-score analysis will be biased.

Figure 18.15. Pretest and Posttest Means Showing No Treatment Effect in a Gain-Score Analysis.

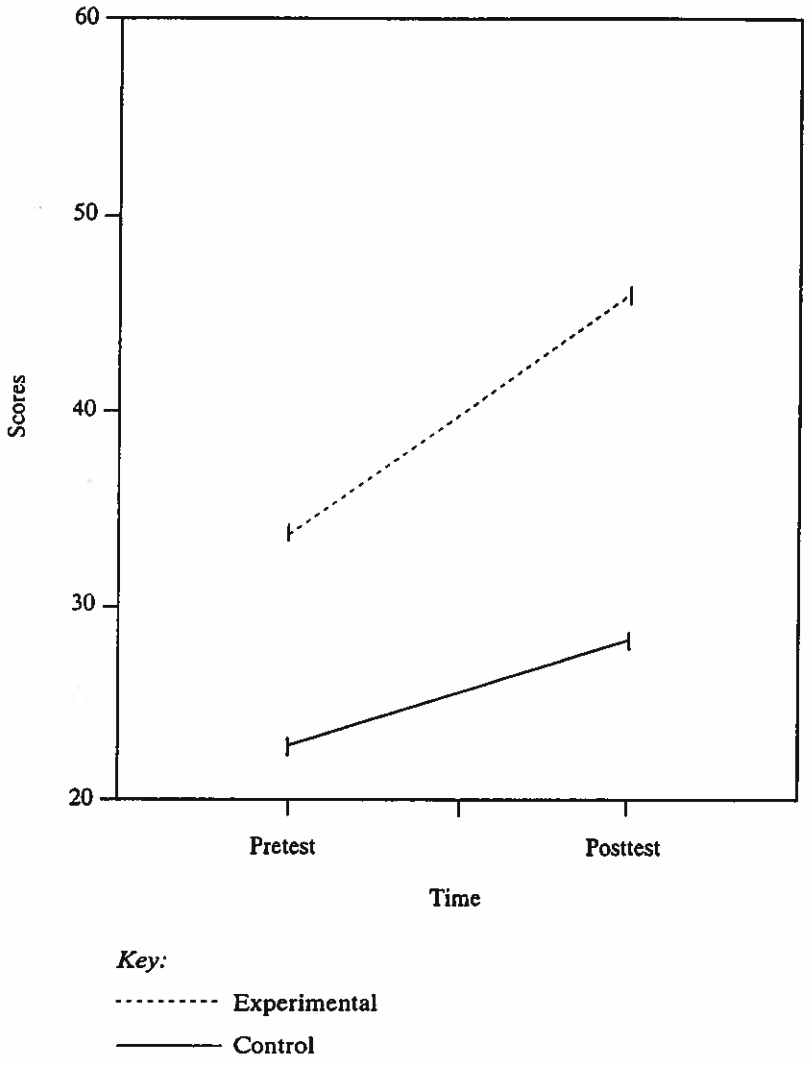


Additional pretest variables can be added to the gain-score analysis as covariates so as to adjust for any initial selection differences on these variables via statistical matching and to assess treatment-effect interactions. However, if the pretest that was used to create the gain score is added as a covariate, the result is the same as would be obtained by using analysis of covariance rather than gain-score analysis.

An Example

The study of the effect of running in place on pulse rate described previously was a randomized experiment. However, we can create a nonequivalent

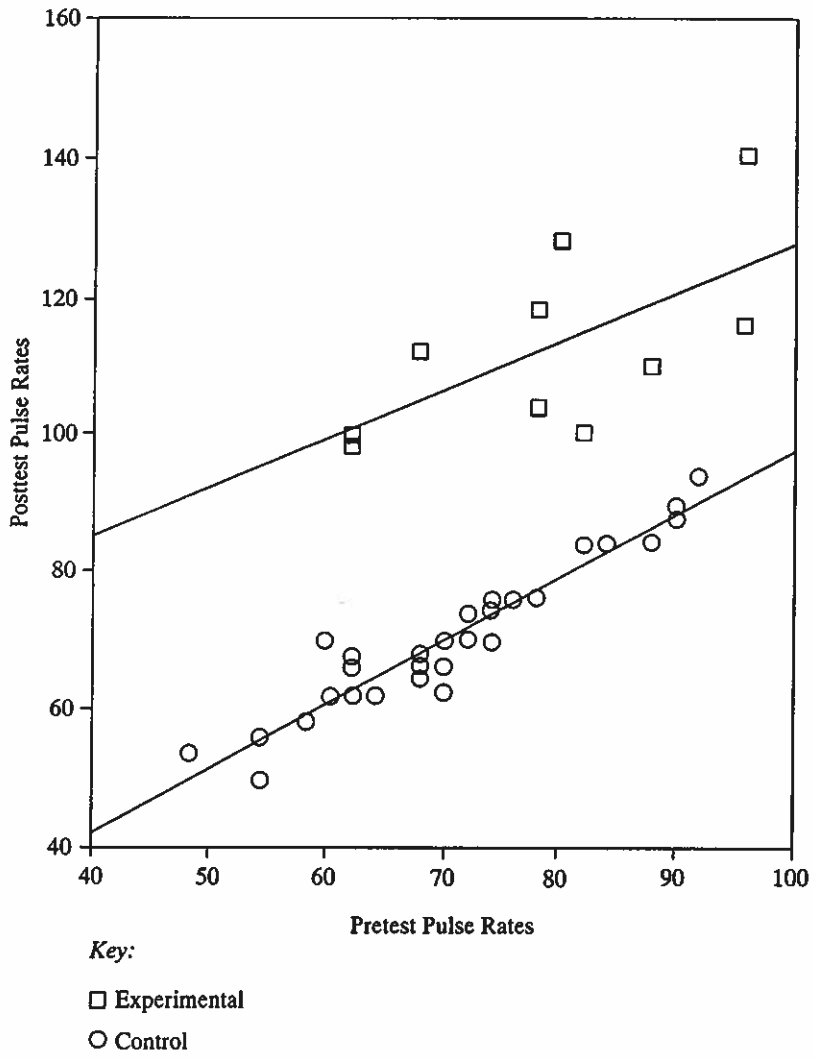
Figure 18.16. Pretest and Posttest Means Showing a Positive Treatment Effect in a Gain-Score Analysis.



comparison group design by imagining that data are available only for the women who were in the experimental group and only for the men who were in the control group. The plot of the pretest pulse rate versus the posttest pulse rate for these individuals is presented in Figure 18.17. These data represent a nonequivalent comparison group design, as evidenced by the horizontal displacement between the pretest scores in the two groups showing that there is an initial difference between the groups on the pretest pulse rate.

A gain-score analysis of the data in Figure 18.17 produces an estimate of the treatment effect of 33.6 (95% confidence interval = 29.20 to 38.00). The estimate of the treatment effect from the analysis of covariance

Figure 18.17. Data from the Pulse Study (Ryan, Joiner, and Ryan, 1985) in the Form of a Nonequivalent-Comparison-Group Design.



with the pretest pulse rate as the covariate is 34.86 (95% confidence interval = 30.34 to 39.37). The estimates from these two analyses are similar because the correlation between the pretest pulse rate and the posttest pulse rate under resting conditions was 0.92, which is very high. The results from the gain-score analysis and the analysis of covariance will not always be so similar.

These estimates of the treatment effect from the analysis of the data as a nonequivalent comparison group design are not very close to the estimate of the average effect of the treatment as derived from the randomized experiment (which was about 19). But the estimates from the nonequivalent

comparison group design are close to the estimate of the effect for women alone as derived from the randomized experiment (which was 34.4). This is probably what should be expected given the way in which this nonequivalent comparison group design was created (with only women in the experimental group and only men in the control group). Seldom will the nature of selection difference be so well known, however. In most practical circumstances, it will usually be more difficult to make sense of the results from nonequivalent comparison group designs.

Concluding Comments on the Nonequivalent Comparison Group Design

Other statistical models for taking account of selection differences (besides the analysis of covariance and gain-score analysis) are available. Many of these procedures, such as selection modeling and modeling propensity scores (see Rindskopf, 1986), are relatively complex statistically and probably require the help of a statistical consultant. Unfortunately, just as with the analysis of covariance and gain-score analysis, there is no guarantee that any of these statistical procedures will adequately account for selection differences. The problem is that properly implementing any of these methods requires information about the nature of selection differences that is usually not available. The reason is that assignment to treatments was not random as in a randomized experiment or was not determined by a known pretest as in the regression-discontinuity design.

Uncertainty about how properly to control for selection differences is the great weakness of the nonequivalent comparison group design. The only resolution for this uncertainty is to use a range of assumptions about the nature of the selection differences and thereby produce a range of estimates of the size of the treatment effect; even then caution must be used in interpreting results (Reichardt and Gollob, 1987). In other words, while researchers can report that a range of estimates derived from a variety of statistical analyses is their best guess about the size of the treatment effect, they should forthrightly acknowledge that this best guess could be far wrong. Otherwise researchers run the risk of misleading their audience.

Usually the best way to deal with initial selection differences is to try to make them as small as possible when the study is being designed and implemented. One way to do this is to forsake the nonequivalent comparison group design in favor of the randomized experiment.

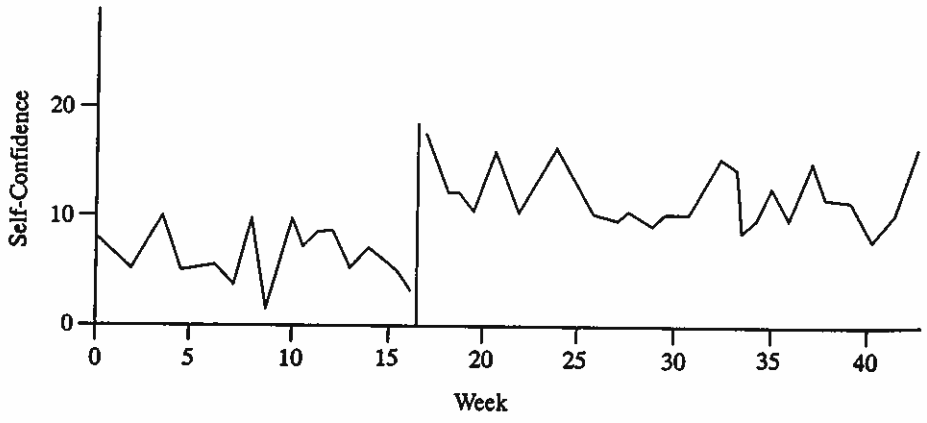
Interrupted Time Series Design

In the interrupted time series design, measurements are made repeatedly at regular intervals before the treatment is introduced, the treatment is then introduced, and measurements are again repeated at regular intervals (for a total of K time points). Further discussion of the interrupted time

series design is provided by Marcantonio and Cook in Chapter Seven of this handbook.

The term *unit* is used here to refer to the entity about which data are collected. The interrupted time series design can be implemented with a single unit ($N = 1$) or with multiple units ($N > 1$); in either case the units can be either individuals or groups of individuals. In one instance, Blose and Holder (1987) used the interrupted time series design to assess the effects of the liberalization of drinking laws in North Carolina. In this study, N was equal to one, and the unit was a community because data on traffic fatalities were collected at the level of the community. In contrast, Smith, Gabriel, Schoot, and Padia (1976) used the interrupted time series design to assess the effects of the Outward Bound program on participants' self-confidence using approximately $N = 200$ individuals as the units. The time series of these data are plotted in Figure 18.18. The vertical line just past week 15 indicates the point at which the individuals participated in the Outward Bound program. The scores plotted at each time point are average responses across a random sample of the two hundred participants.

Figure 18.18. Mean Levels of Self-Confidence Before and After Participation in an Outward Bound Program.



Sources: Smith, Gabriel, Schoot, and Padia, 1976. Copyright 1976 by Sage Publications. Reprinted by permission of Sage Publications, Inc. Also Glass, 1988. Copyright 1988 by the American Educational Research Association. Reprinted by permission of the publisher.

To estimate the effect of the treatment, the first step is to model the trend in the data collected before the treatment was introduced. This trend is then projected forward in time and compared to the trend in the data collected after the treatment was introduced. The difference between the projected and actual trends is the estimate of the treatment effect. In Figure 18.18, for example, the trend in the self-confidence data before participation in Outward Bound is lower than the trend in the data after participa-

tion. As a result, it appears as if Outward Bound has a positive effect on self-confidence in the population of individuals in the study.

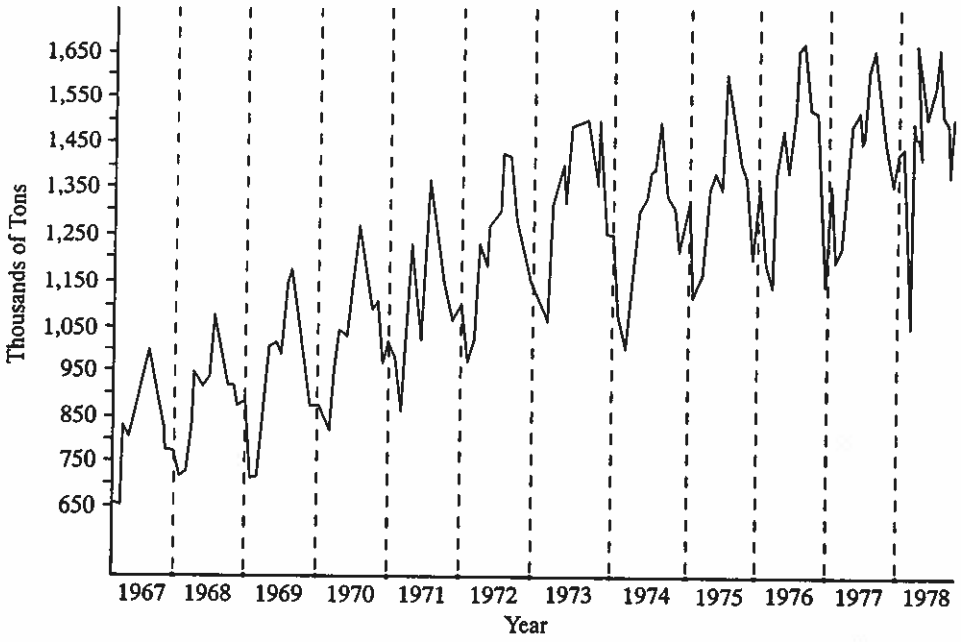
As Marcantonio and Cook emphasize in Chapter Seven of this handbook, the interrupted time series design is very similar to the regression-discontinuity design. The only difference is that in the regression-discontinuity design, assignment to treatment conditions is determined by a cutoff score on a pretest measure; in the interrupted time series design, the assignment to treatment condition is determined by a cutoff score based on chronological time. This distinction has an important implication that will be described below, but for the most part, the logic for the analysis of data from the interrupted time series design is similar to the logic for the analysis of data from the regression-discontinuity design.

Just as in the regression-discontinuity design, whether the estimate of the treatment effect is unbiased in the interrupted time series design depends on whether the trends in the data, both before and after the treatment is introduced, have been accurately modeled. To achieve this accuracy, the researcher must correctly model any curvilinearity. Curvilinearity can be modeled by either transforming the data or using polynomial regression. In time series analysis, a data transformation called first-order differencing can be used to remove linear trends, second-order differencing can be used to remove quadratic trends, and so on (Box and Jenkins, 1970; McCleary and Hay, 1980). A statistical consultant can be helpful here. Correctly modeling the trends in the data also means that treatment-effect interactions must be properly taken into account. In the context of the interrupted time series design, a treatment-effect interaction means that the treatment effect changes over time.

Just as in the regression-discontinuity design, smoothing the data (using either a median trace or a moving average as described above) can make it easier for the analyst to recognize both curvilinear trends and treatment-effect interactions. As an illustration, Figure 18.19 is a time series plot of shipments of oil to service stations in France (Hogarth, 1980). By looking at these data, can you describe the nature of the effect of the Arab oil embargo toward the end of 1973 and the effect of increases in the price of oil toward the beginning of 1976? The time series in Figure 18.20 is the same data after smoothing (see Makridakis and Wheelwright, 1978) and reveals the nature of these effects much more clearly. Notice how the effect of the oil embargo in 1973 is quite abrupt while the effect of price increases in 1976 is more gradual.

The one important distinction between the regression-discontinuity design and the interrupted time series design arises because of possible autocorrelation of data. Different from outcome data in the regression-discontinuity design, the outcome data in the interrupted time series design are likely to be correlated among themselves. That is, the observation at time 1 in the time series is likely to be correlated with the observation at time 2, which is likely to be correlated with the observation at time 3, and so

Figure 18.19. Time Series Plot of Shipments of Oil to Service Stations in France.



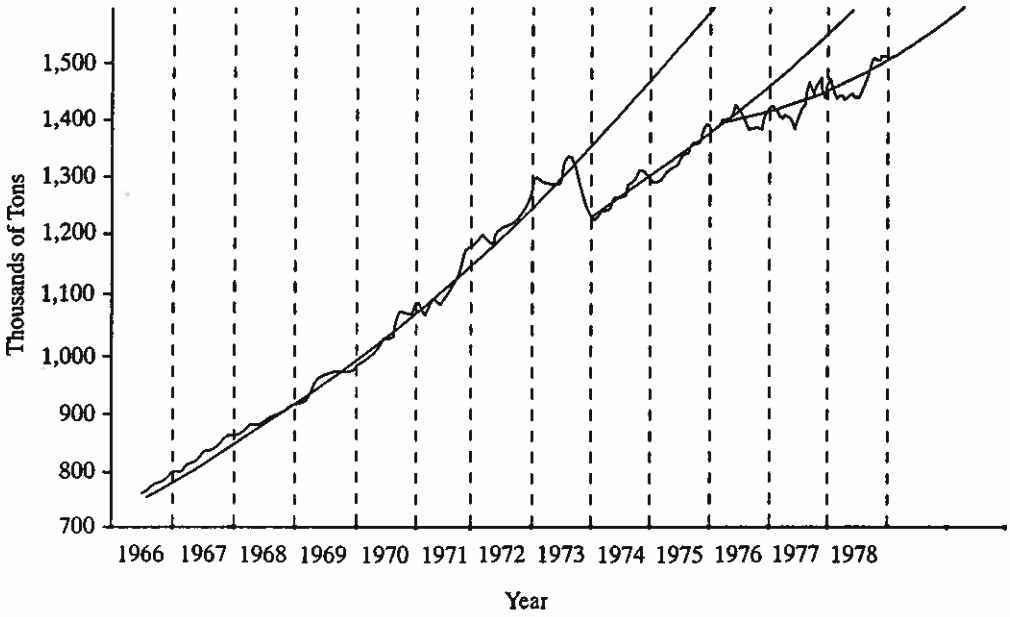
Source: From R. Hogarth, *Judgement and Choice.* Copyright 1980 by John Wiley & Sons. Original figure supplied by S. Makridakis and B. Majani. Reprinted by permission of John Wiley & Sons, Ltd., and Spyros Makridakis.

on. Such autocorrelation produces no bias in the estimate of the size of the treatment effect using standard regression procedures, but it does bias statistical significance tests and confidence intervals that are created by standard regression procedures. This bias occurs because standard regression procedures assume that there is no autocorrelation among the data points. To control for the effects of autocorrelation among the outcome scores, the regression analysis must be modified. Three different approaches for this are described below. Which one is most appropriate depends on the circumstances. In any case, seeking help from a statistical consultant is probably advisable.

ARIMA Modeling

The autoregressive, integrated, moving average (ARIMA) modeling approach assumes that the degree of autocorrelation in the observations is constant over time. ARIMA modeling uses the data to estimate the degree of autocorrelation and then adjusts the regression analysis accordingly (Box and Jenkins, 1970; McCain and McCleary, 1979; McCleary and Hay, 1980). One potential advantage is that ARIMA modeling can be used with $N = 1$. One potential drawback is that the number of time points (K) must usually

Figure 18.20. Time Series Plot of Shipments of Oil to Service Stations in France After Data Smoothing.



Source: From R. Hogarth, *Judgement and Choice.* Copyright 1980 by John Wiley & Sons. Original figure supplied by S. Makridakis and B. Majani. Reprinted by permission of John Wiley & Sons, Ltd., and Spyros Makridakis.

be relatively large. Some statisticians suggest that the number of repeated observations (K) must be at least fifty, but the minimum size of K depends on the variability in the data: If there is relatively little variation (which is more likely when the unit is a group of individuals such as a community or state than when the unit is an individual) the minimum value for K might be substantially smaller.

If N is greater than 1, ARIMA modeling could be applied to the data from each unit separately, or the data at each time point could be aggregated across the units (as in the Outward Bound study) and ARIMA modeling applied to the aggregated data. The first approach would allow the researcher to assess individual differences in the effectiveness of the treatment while the second might allow K to be smaller. Unfortunately, many software packages either do not offer ARIMA modeling or do not provide the options for using ARIMA modeling to estimate the effects of treatments. The BMDP program is one that allows both (Dixon, 1985).

Multivariate Analysis of Variance

The multivariate analysis of variance (MANOVA) approach allows the autocorrelations among observations to have any constant or nonconstant

pattern over time (Swaminathan and Algina, 1977). The MANOVA approach uses the data to estimate the autocorrelations at each time point and then adjusts the regression analysis accordingly. By relaxing the assumption made by ARIMA that the degree of autocorrelation is constant over time, MANOVA gains the advantage that K can be quite small. The disadvantage is that by relaxing this assumption, N must be substantially larger than K. The MANOVA analysis fits a common (aggregate) trend to the data from all N units and estimates the average effect of the treatment across all N units.

Hierarchical Linear Modeling

The hierarchical linear modeling (HLM) approach requires that N be substantially greater than 1 but, unlike the MANOVA approach, does not require that N be greater than K or even that observations be collected at the same time points on the different units (Bryk and Raudenbush, 1987, 1992). The HLM approach fits regression models at two different levels. At the first level, HLM fits a regression model and estimates the effects of the treatment for each unit individually. At the second level, HLM fits a regression model to the estimates of the treatment effects from the first level allowing for the inclusion of additional covariates. The model at the second level provides an estimate of the average treatment effect and estimates of interactions of the treatment with any of the covariates that are included. By using two hierarchical levels of analysis, the HLM approach circumvents the need to model the nature of the autocorrelation among the observations.

Concluding Comments on the Interrupted Time Series Design

The interrupted time series design can be biased by "history" or other threats to validity (see Chapter Seven of this volume). One way to remove these biases is by adding a "control" time series of observations that is susceptible to the same biases as the experimental series but is not given the treatment. The data from the control series can be analyzed just like the data from the experimental series. The treatment effect is then estimated as the difference between the discontinuity in the experimental series at the point of the intervention and the discontinuity in the control series at the same time point.

One advantage of the interrupted time series design, compared to the other three designs described in this chapter, is that the interrupted time series design allows the researcher to study the time course of the treatment effect. For example, the researcher can assess whether the treatment effect occurs abruptly or is delayed, and whether it increases, decreases, or remains the same over time. The models for studying the time course of the treatment have been especially well developed within the ARIMA modeling approach (Box and Tiao, 1975).

Conclusions

When estimating the effects of treatments using any of the four designs described above, we recommend the following practices.

Draw pictures of the data. Pictures can help you decide which statistical analyses are appropriate and can help you interpret the results of statistical analyses.

Watch for improper fits in the statistical analysis such as a linear regression line being fitted to curvilinear data. Fitting the wrong model can bias the estimates of treatment effects.

Assess treatment-effect interactions. Treatment-effect interactions reveal how a treatment effect varies either across individuals or time. Understanding how the effect of a treatment varies is as important, if not more so, than estimating the average effect of the treatment.

Ask yourself if there are any hidden biases in the statistical analyses. For example, it is usually impossible to determine from the data alone whether the analysis of covariance properly takes account of selection differences in the nonequivalent comparison group design. You also have to understand the logic of what the analysis of covariance does and, using your (imperfect) substantive knowledge of the study, decide whether that logic fits the circumstances.

Report the degree of uncertainty forthrightly in the results. Biases cannot all be removed with complete certainty. As a result, there will always be uncertainty about the size of treatment effects. Researchers should make sure that readers are not misled into believing that the results are more certain than is warranted. Proper presentation of results includes using both confidence intervals and multiple analyses when you are not sure which single analysis is correct.

References

Ball, S., and Bogatz, G. A. *The First Year of "Sesame Street": An Evaluation.* Princeton, N.J.: Educational Testing Service, 1970.

Bentler, P. M. *EQS Structural Equations Program Manual.* Los Angeles: BMDP Statistical Software, 1989.

Blose, J. O., and Holder, H. D. "Liquor-by-the-Drink and Alcohol-Related Traffic Crashes: A Natural Experiment Using Time-Series Analysis." *Journal of Studies on Alcohol,* 1987, *48,* 52–60.

Borenstein, M., and Cohen, J. *Statistical Power Analysis: A Computer Program.* Hillsdale, N.J.: Erlbaum, 1988.

Boruch, R. F., and Wothke, W. "Seven Kinds of Randomization Plans for Designing Field Experiments." In R. F. Boruch and W. Wothke (eds.), *Randomization and Field Experimentation.* New Directions for Program Evaluation, no. 28. San Francisco: Jossey-Bass, 1985.

Box, G.E.P., and Jenkins, G. M. *Time-Series Analysis: Forecasting and Control.* San Francisco: Holden-Day, 1970.

Box, G.E.P., and Tiao, G. C. "Intervention Analysis with Application to Economic and Environmental Problems." *Journal of the American Statistical Association,* 1975, *70,* 70–92.

Braucht, G. N., and Reichardt, C. S. "A Computerized Approach to Trickle-Process, Random Assignment." *Evaluation Review,* 1993, *17,* 79–90.

Bryk, A. S., and Raudenbush, S. W. "Application of Hierarchical Linear Models to Assessing Change." *Psychological Bulletin,* 1987, *101,* 147–158.

Bryk, A. S., and Raudenbush, S. W. *Hierarchical Linear Models: Applications and Data Analysis Methods.* Newbury Park, Calif.: Sage, 1992.

Campbell, D. T., and Boruch, R. F. "Making the Case for Randomized Assignment to Treatments by Considering the Alternatives: Six Ways in Which Quasi-Experimental Evaluations in Compensatory Education Tend to Underestimate Effects." In C. A. Bennett and A. A. Lumsdaine (eds.), *Evaluation and Experiment: Some Critical Issues in Assessing Social Programs.* New York: Academic Press, 1975.

Cohen, J. *Statistical Power Analysis for the Behavioral Sciences.* (Rev. ed.) New York: Academic Press, 1977.

Conner, R. F. "Selecting a Control Group: An Analysis of the Randomization Process in Twelve Social Reform Programs." *Evaluation Quarterly,* 1977, *1,* 195–243.

Cronbach, L. J. *Designing Evaluations of Educational and Social Programs.* San Francisco: Jossey-Bass, 1982.

Dixon, W. J. *BMDP Statistical Software.* Berkeley: University of California Press, 1985.

Draper, N. R., and Smith, H. *Applied Regression Analysis.* New York: Wiley, 1981.

Feinberg, S. E. "Randomization and Social Affairs: The 1970 Draft Lottery." *Science,* 1971, *171,* 255–261.

Freedman, D., Pisani, R., Purves, R., and Adhikari, A. *Statistics.* (2nd ed.) New York: Norton, 1991.

Glass, G. V. "Quasi-Experiments: The Case of Interrupted Time Series." In R. M. Jaeger (ed.), *Complementary Methods for Research in Education.* Washington, D.C.: American Educational Research Association, 1988.

Goldberger, A. S. *Selection Bias in Evaluating Treatment Effects: Some Formal Illustrations.* Discussion Paper 123-72, Madison, University of Wisconsin, Institute for Research on Poverty, 1972.

Hamilton, L. C. *Regression with Graphics: A Second Course in Applied Statistics.* Pacific Grove, Calif.: Brooks/Cole, 1992.

Hogarth, R. *Judgement and Choice.* New York: Wiley, 1980.

Jöreskog, K. G., and Sörbom, D. *LISREL 7: A Guide to the Program and Applications.* Chicago: SPSS, 1988.

Kraemer, H. C., and Thiemann, S. *How Many Subjects? Statistical Power Analysis in Research.* Newbury Park, Calif.: Sage, 1987.

Lipsey, M. W. *Design Sensitivity: Statistical Power for Experimental Research.* Newbury Park, Calif.: Sage, 1990.

McCain, L. J., and McCleary, R. "The Statistical Analysis of Simple Interrupted Time-Series Quasi-Experiments." In T. D. Cook and D. T. Campbell, *Quasi-Experimentation: Design and Analysis Issues for Field Settings.* Chicago: Rand McNally, 1979.

McCleary, R., and Hay, R. A., Jr. *Applied Time Series Analysis for the Social Sciences.* Newbury Park, Calif.: Sage, 1980.

Makridakis, S., and Wheelwright, S. C. *Interactive Forecasting: Univariate and Multivariate Methods.* (2nd ed.) San Francisco: Holden-Day, 1978.

Moore, D. S., and McCabe, G. P. *Introduction to the Practice of Statistics.* New York: Freeman, 1989.

Reichardt, C. S. "The Statistical Analysis of Data from Nonequivalent Group Designs." In T. D. Cook and D. T. Campbell, *Quasi-Experimentation: Design and Analysis Issues for Field Settings.* Chicago: Rand McNally, 1979.

Reichardt, C. S., and Gollob, H. F. "Taking Uncertainty into Account When Estimating Effects." In M. M. Mark and R. L. Shotland (eds.), *Multiple Methods for Program Evaluation.* New Directions for Program Evaluation, no. 35. San Francisco: Jossey-Bass, 1987.

Rindskopf, D. "New Developments in Selection Modeling for Quasi-Experimentation." In W.M.K. Trochim (ed.), *Advances in Quasi-Experimental Design and Analysis.* New Directions for Program Evaluation, no. 31. San Francisco: Jossey-Bass, 1986.

Ryan, B. F., Joiner, B. L., and Ryan, T. A., Jr. *Minitab Handbook.* (2nd ed.) Boston: Duxbury, 1985.

Smith, M. L., Gabriel, R., Schoot, J., and Padia, W. L. "Evaluation of the Effects of Outward Bound." In G. V. Glass (ed.), *Evaluation Studies Review Annual: Volume 1.* Newbury Park, Calif.: Sage, 1976.

Swaminathan, H., and Algina, J. "Analysis of quasi-experimental time-series designs." *Journal of Multivariate Behavioral Research,* 1977, *12,* 111-131.