CHAPTER 8

Analysis of Covariance

In this chapter we consider an adjustment strategy that is appropriate for cohort studies with a numerical outcome factor, a categorical treatment (or risk) factor, and a numerical confounding factor. Under these conditions, the *general linear model* can be applied to the problem of estimating treatment effects. The *analysis of covariance* (ANCOVA) represents the main application of the linear model for this purpose.

8.1 BACKGROUND

The general linear model represents the outcome value as a linear combination (weighted sum) of measured variables. Generally speaking, when these variables are all numerical, the linear model is called a *regression model*. When the variables are all categorical, we refer to the *analysis of variance* (ANOVA).

While both regression and analysis of variance can be formally subsumed under the general linear model, the two techniques have traditionally been treated as distinct. This historical separation occurred for two reasons. First, before high-speed computers were in general use, computational aspects of statistical techniques were of much interest. The most efficient computational procedures for regression and ANOVA were quite different. Second, the two methods tended to be applied to different sorts of problems.

The analysis of variance is usually thought of as a technique for comparing the means of two or more populations on the basis of samples from each. In practice, these populations often correspond to different treatment groups, so that differences in population means may be evidence for corresponding differences in treatment effects.

The ANOVA calculations involve a division of the total sample variance into within-group and between-group components. The within-group component provides an estimate of error variance, while the between-group component estimates error variance plus a function of the differences among treatment means. The ratio of between- to within-group variance provides a test of the null hypothesis that all means are equal. Moreover, the differences among group means provide unbiased estimates of the corresponding population mean differences, and standard errors based on the within-group variance provide confidence intervals for these differences and tests of their significance.

Regression analysis, on the other hand, is primarily used to model relationships between variables. With it, we can estimate the form of a relationship between a response variable and a number of inputs. We can try to find that combination of variables which is most strongly related to the variation in the response.

The analysis of covariance represents a marriage of these two techniques. Its first use in the literature was by R. A. Fisher (1932), who viewed the technique as one that "combines the advantages and reconciles the requirements of the two very widely applicable procedures known as regression and analysis of variance."

Combining regression and ANOVA provides the powerful advantage of making possible comparisons among treatment groups differing prior to treatment. Suppose we can identify a variable X that is related to the outcome, Y, and on which treatment groups have different means. We shall assume for simplicity that X is the only variable on which the groups differ. Then, if we knew the relationship between Y and X, we could appropriately adjust the observed differences on Y to take account of the differences on X.

8.2 EXAMPLE: NUTRITION STUDY COMPARING URBAN AND RURAL CHILDREN

Greenberg (1953) described a nutrition study designed to compare growth of children in an urban environment with that of rural children. Data were ob-

tained on the heights of children in the two samples: one from an urban private school and one from a rural public school. Differences in growth between these groups might be the result of the different environmental influences operating on the children. In particular, the rural children might be experiencing some nutritional deprivation relative to their urban counterparts. In the terminology of this book, height would be the response or outcome factor and nutrition the risk factor of interest.

The data are shown in Table 8.1. An analysis of variance conducted on the height data reveals that the observed difference between the groups (2.8 cm) is not statistically significant. So it might be concluded that there is no evidence here for a difference in nourishment between the urban and rural school children.

Table 8.1 Height and Age of Private and Rural School Children in a Study in North Carolina in 1948

	Private School		Rural School	
Students	Age (months)	Height (cm)	Age (months)	Height (cm)
1	109	137.6	121	139.0
2	113	147.8	121	140.9
3	115	136.8	128	134.9
4	116	140.7	129	149.5
5	119	132.7	131	148.7
6	120	145.4	132	131.0
7	121	135.0	133	142.3
8	124	133.0	134	139.9
9	126	148.5	138	142.9
10	129	148.3	138	147.7
11	130	147.5	138	147.7
12	133	148.8	140	134.6
13	134	133.2	140	135.8
14	135	148.7	140	148.5
15	137	152.0		
16	139	150.6		
17	141	165.3		
18	142	149.9		
Mean	126.8	144.5	133.1	141.7

Reprinted, by permission, from Greenberg (1953), Table 1.

Before reaching this conclusion, however, we should consider whether there are likely to be confounding factors. One variable that comes immediately to mind is age. The data on age are also presented in Table 8.1. The mean age for

the rural children is 6.3 months greater than that of the urban children. In a sense, then, the rural children have an "unfair advantage" conferred by their greater average age. Thus we might expect that if the age distributions were the same, the difference in average height between the groups would be even larger than the observed 2.8 cm. The analysis of covariance allows us to adjust the 2.8-cm difference to obtain a better (less-biased) estimate of the difference between groups that *would* have been observed had the mean ages in the two groups been equal. As we shall see in Section 8.3, ANCOVA produces an estimated difference of 5.5 cm, which is significant at the .05 level.

In addition to the bias reduction described above, another benefit results from the combination of regression analysis and ANOVA. Suppose that *within* treatment groups, a substantial proportion of the variance in Y can be explained by variation in X. In carrying out an ANOVA, we would like the within-group variance to reflect only random error. Regression analysis can be used to remove that part of the error attributable to X and thereby to increase the precision of group comparisons.

The Greenberg (1953) example mentioned above can be used to illustrate this point as well. It is clear from Table 8.1 that a substantial proportion of the variation in height is attributable to variation in age. Put differently, if all children in a group were of the same age, the variation in heights within that group would be substantially reduced. Since the relationship between height and age over this range is quite linear, we can estimate the pure error variation by taking residuals* around the regression line relating the two variables. In effect, this is what ANCOVA does, and when a high proportion of within-group variance is explained by the covariate, a large increase in precision results.

In summary, then, ANCOVA combines the advantages of regression and ANOVA in comparing treatments by providing two important benefits. First, by estimating the form of the relationship between outcome and covariate, an appropriate adjustment can be made to remove biases resulting from group differences on the covariate. This advantage is of importance primarily in nonrandomized studies, where such group differences are likely to occur. Second, by reducing the variation within groups, the precision of estimates and tests used to compare groups can be increased. This advantage may be valuable in both randomized and nonrandomized studies.

By combining the advantages of ANOVA and regression, ANCOVA provides a powerful tool for estimating treatment effects. As noted by Fisher, however, the technique also "reconciles the requirements" of the techniques. Thus, to be valid, the ANCOVA must be used in situations satisfying the requirements for

* The *residual* corresponding to a given observation is defined as the difference between the actual observed Y and the value predicted by substituting the corresponding X value into the regression equation.

both techniques. Put differently, the usefulness of the ANCOVA rests on the validity of a certain mathematical model for the generation of data, which in turn rests on a set of assumptions. To obtain the advantages of both regression and ANOVA, we must be willing to assert that a somewhat restrictive model is valid.

In the remainder of this chapter, we will attempt to provide enough understanding of the rationale and assumptions underlying ANCOVA to enable the reader to understand when ANCOVA can be used and how to interpret the results generated. Since the actual calculations involved in carrying out the analysis are complex, they are almost always performed by a computer, and it would be unnecessarily confusing to present the formulas here. For the reader interested in more detail, a technical appendix containing some basic formulas is included at the end of this chapter. More extensive discussions can be found in Cochran (1957) and Winer (1971, Chap. 10).

8.3 THE GENERAL ANCOVA MODEL AND METHOD

To understand the rationale underlying the use of ANCOVA in nonrandomized studies, it is helpful to begin with a somewhat idealized situation. Suppose that on the basis of extensive prior research, the relationship between an outcome and confounding factor can be specified. For example, it might be known that for rural school children, the relationship between height and age over the age range being considered can be expressed as

$$\text{Average height} = 75 + 0.5\,(\text{age}).$$

Suppose that a particular group of rural children have been exposed to some special treatment, such as a dietary supplement. At the time they are measured this group has a mean age of 132 months and a mean height of 147 cm. Suppose further that another group has been exposed to a different treatment and is measured when the children are 120 months old on the average. The average height of this group is 133 cm.

Since the groups differ on mean age, it is not obvious which treatment has been more effective. To make a fair comparison, we must remove the effect of the confounding variable age. However, using the relationship specified above we know that the expected height for the two groups without any special treatment is given by:

Group 1: Average height = $75 + 0.5(132) = 141$ cm
Group 2: Average height = $75 + 0.5(120) = 135$ cm.

Therefore, the effects of the treatments are:

Group 1: Effect = observed − expected = $147 - 141 = 6$ cm
Group 2: Effect = observed − expected = $133 - 131 = -2$ cm.

and the difference between them is 8 cm.

Alternatively, we can say that because the groups differ by 12 months in age, the relationship predicts that they will differ by 6 cm. So we could effectively "adjust" the comparison between the two groups by subtracting 6 cm from the difference between them. Since the observed difference is 14 cm, this would leave 8 cm attributable to the difference in treatments received..

Because we are assuming in this example a known baseline relationship against which to measure performance under the treatments, we can obtain an absolute measure of effect for each treatment (6 cm and −2 cm). In most practical situations, we do not have available such an external standard, and we must use only data obtained during the study. Thus an absolute measure of effect for each group is impossible. On the other hand, it may still be possible to obtain from the data an estimate of the coefficient (0.5 cm/month in our example) relating outcome level to confounding variable. So it may be possible to adjust the observed difference to remove the effect of age from the comparison. In effect, this is how ANCOVA is used to estimate treatment effects in nonrandomized comparative studies.

The basic model underlying the use of the standard analysis of covariance asserts that there is a linear relationship between the outcome Y and the covariate X with identical slopes in the two groups, but possibly different intercepts. With two treatment groups, we can write the basic model as*

$$\begin{aligned} Y &= \alpha_1 + \beta X + e \quad \text{in group 1 (treatment)} \\ Y &= \alpha_0 + \beta X + e \quad \text{in group 0 (control),} \end{aligned} \tag{8.1}$$

where

α_1 = expected value of Y when $X = 0$ for group 1,

α_0 = expected value of Y when $X = 0$ for group 0

e = random variable representing error (expectation 0 for any given X).

Let $\overline{X}$ represent the sample mean of all the X observations in both groups, $\overline{X}_1$, the mean for group 1, and $\overline{X}_0$ the mean for group 0. Figure 8.1 illustrates this situation. Note that the direct comparison of $\overline{Y}_1$ and $\overline{Y}_0$ will be biased since $\overline{X}_1$

* For the reader familiar with regression analysis, this model can be represented as a two-variable regression model with variables X and a dummy variable taking the value 1 in group 1 and 0 in group 0.

$\neq \bar{X}_0$. In fact, taking means in (8.1) yields

$$\bar{Y}_1 = \alpha_1 + \beta\bar{X}_1 + \bar{e}_1$$
$$\bar{Y}_0 = \alpha_0 + \beta\bar{X}_0 + \bar{e}_0,$$

so that

$$E(\bar{Y}_1 - \bar{Y}_0) = \alpha_1 - \alpha_0 + \beta(\bar{X}_1 - \bar{X}_0). \tag{8.2}$$

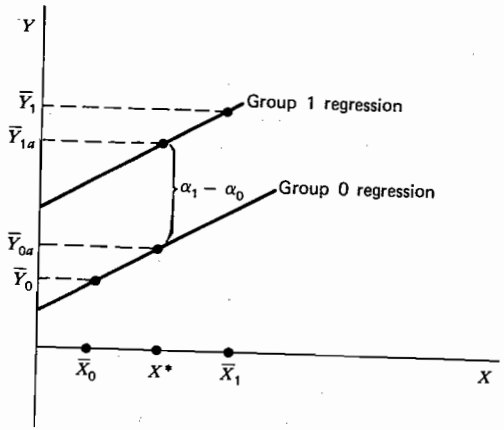


Figure 8.1. Standard ANCOVA assumptions.

Note that from (8.1), we can interpret $\alpha_1 - \alpha_0$ as the expected difference between the outcomes of the two individuals with the same value of X but in two different groups. This difference will represent the differential effect of the two treatments unless there is some other variable related to Y which distinguishes the two subjects. To estimate $\alpha_1 - \alpha_0$, we cannot simply subtract $\bar{Y}_0$ from $\bar{Y}_1$, but must adjust each of these to move them, in effect, to a common X value, say X^*. Let us define the "adjusted" mean of Y for group 1 as

$$\bar{Y}_{1a} = \bar{Y}_1 - \beta(\bar{X}_1 - X^*).$$

$\bar{Y}_{1a}$ may be interpreted as an estimate of the mean outcome for members of group 1 whose X value is X^*. Similarly,

$$\bar{Y}_{0a} = \bar{Y}_0 - \beta(\bar{X}_0 - X^*)$$

estimates the mean outcome for members of group 0 whose X value is X^*. To estimate the difference between the means of the two groups at the same value of X (in this case X^*), we can simply take the difference of these two adjusted

means:

$$\begin{aligned}\bar{Y}_{1a} - \bar{Y}_{0a} &= \bar{Y}_1 - \beta(\bar{X}_1 - X^*) - [\bar{Y}_0 - \beta(\bar{X}_0 - X^*)] \\ &= \bar{Y}_1 - \bar{Y}_0 - \beta(\bar{X}_1 - \bar{X}_0).\end{aligned} \tag{8.3}$$

This adjusted group mean difference is an unbiased estimator of $\alpha_1 - \alpha_0$.

For simplicity, we have not discussed how the value of β necessary to perform the adjustments is actually obtained. In practice, we rarely have any a priori theoretical basis for determining the value of β and must therefore use the data to obtain an estimate, $\hat{\beta}$. The ANCOVA calculations provide us with an unbiased estimator based on the relationship between Y and X within the two groups. Thus the adjusted difference is of the form

$$\bar{Y}_{1a} - \bar{Y}_{0a} = \bar{Y}_1 - \bar{Y}_0 - \hat{\beta}(\bar{X}_1 - \bar{X}_0). \tag{8.4}$$

It can be shown that the substitution of an unbiased estimate $\hat{\beta}$ for the unknown true value β still yields an unbiased estimate of $\alpha_1 - \alpha_0$ under the model specified by (8.1).

In Appendix 8A we present the formula usually used to compute $\hat{\beta}$. It is called a pooled within-group regression coefficient, because it combines data on the relationship between Y and X in both groups. This combination of data provides high precision and is valid under our assumption that the regression lines are parallel.

We should mention in passing that this pooled coefficient is not found by calculating a regression coefficient from the data on both groups taken together as a single group, as is sometimes proposed. This latter approach may be viewed as comparing the mean residuals for the two groups around the *overall* regression line fitted to the entire sample. It is incorrect, however, in the sense that it does not yield an unbiased estimate of β or of the effect $\alpha_1 - \alpha_0$ under the model given by (8.1).

Using the standard ANCOVA calculations (see Appendix 8A), we obtain for the Greenberg (1953) example:

$$\hat{\beta} = 0.42 \text{ cm/month},$$

and because

$$\bar{X}_1 = 126.8 \text{ months}$$

and

$$\bar{X}_0 = 133.1 \text{ months},$$

the adjusted difference is

$$\bar{Y}_{1a} - \bar{Y}_{0a} = 2.8 - 0.42(126.8 - 133.1) = 5.5 \text{ cm}.$$

The initial difference of 2.8 cm in favor of the urban children has, after adjustment, been nearly doubled.

We may ask at this point whether this adjusted difference is statistically significant. To answer this question, we can look at the standard error provided as part of the ANCOVA calculations. This standard error can be used to perform a t test of

$$H_0 : \alpha_1 = \alpha_0.$$

More generally, when there are more than two treatment groups (say K groups), ANCOVA provides an F test of

$$H_0 = \alpha_1 = \alpha_2 = \alpha_3 = \cdots = \alpha_K.$$

If this test proves significant, we can reject the null hypothesis that all treatment groups have the same intercept. In this case we must conclude either that the treatments are differentially effective or that there is some unmeasured variable related to outcomes on which the groups vary (i.e., another confounding factor). In the Greenberg (1953) example, a t test for the difference of adjusted means results in a t value of 2.12, which is significant at the .05 level. So when age is taken into account, there appears to be a significant difference in height between the two samples.

8.4 ASSUMPTIONS UNDERLYING THE USE OF ANCOVA

In Section 8.3 we presented the basic model underlying the use of ANCOVA in the simple situation with two treatment groups and one covariate. This model is summarized by (8.1). While this statement of the model appears simple, it implies a large number of conditions that must be satisfied. Since the user must verify that these conditions hold, we present in this section a listing of the assumptions. With each of these, we indicate the consequences of failures to satisfy the assumption and how these can be detected in practice. The next section considers some ways of reducing the biases introduced by such failures.

Like any mathematical model attempting to represent reality, the ANCOVA model is never perfectly true. It is only a more-or-less accurate abstraction. So, although we may for simplicity discuss whether or not a particular condition holds in a particular situation, it should be remembered that such statements are only approximate. The real question is whether the ANCOVA model is good enough not to result in misleading results. With this caveat in mind, we now proceed to list the ANCOVA assumptions.

1. *Equality of regression slopes.* ANCOVA assumes that the relationship between Y and X in each group differs only in terms of the intercept (α_1) but

not the slope (β). This assumption is essential if we are to have the *possibility* of interpreting the difference between the lines ($\alpha_1 - \alpha_0$) as a measure of treatment effect. The problem of nonparallel regressions in different treatment groups is discussed in Section 3.3 and is a general problem involved in all adjustment strategies. The nature of the difficulty is illustrated in Figure 8.2. The expected difference between two individuals in different groups with identical X values depends on X. Thus there is no unique summary value which can be interpreted as *the* treatment effect.

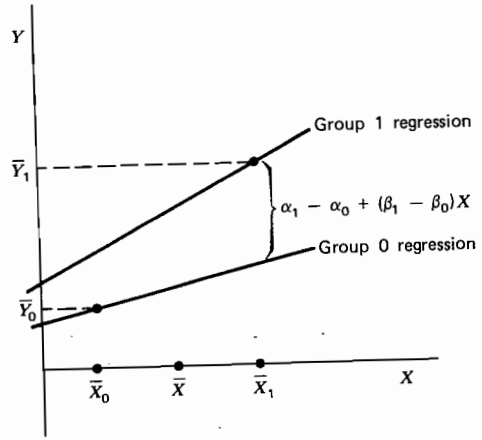


Figure 8.2. Nonparallel linear regressions in two groups.

In such a situation we say there is an *interaction* between the treatment effect and the covariate. If an interaction is suspected, it is worthwhile to examine carefully the graph of Y versus X in the two groups. Visual inspection will usually be adequate to detect serious departures from parallelism.

A formal statistical test for the equality of slopes can also be conducted. If such a test is carried out, and the null hypothesis of slopes rejected, we cannot apply ANCOVA. If, on the other hand, the null hypothesis is not rejected, we still cannot be sure that the slopes are identical. This is a general property of statistical tests. Our ability to assert that the null hypothesis in fact holds if it is not rejected is related to the "power" of the test, which is difficult to compute. Generally speaking, however, the power increases with the sample size. So a statistical test can provide evidence on whether the slopes are equal, but no certainty unless the sample sizes are very large.

2. *Linearity.* The ANCOVA assumes a *linear* relationship between Y and X. The simplest, and usually adequate, test of linearity is to plot a graph of Y

versus X in each group. Formal statistical tests of linearity are available if there is any doubt. The simplest involves calculating the regression line in each group and examining the residuals. Standard texts on regression analysis (e.g., Draper and Smith, 1966; Chatterjee and Price, 1977; Mosteller and Tukey, 1977) provide more detail.

3. *Covariate measured without error.* In some situations, the variable thought to be linearly related to Y cannot be measured directly, and an imperfect substitute containing some measurement error must be used. In Section 5.2 we discussed the issues of measurement error and reliability in some detail. When the observed X, consisting in part of error, is used in the ANCOVA model, both estimates and tests may be affected. In both randomized and nonrandomized studies, the precision of the estimated effect and the power of statistical tests will generally decrease as the reliability decreases. Further, in nonrandomized studies, measurement error will introduce bias in situations where using the true X yields an unbiased estimate (see Cochran, 1968; Elashoff, 1969). When even the true X would result in bias, the effect of measurement error is more complex. Sometimes a fallible variable may even be preferable to a corresponding true score (Weisberg, 1979), although such situations are extremely rare in practice. As a general rule, it is desirable whenever possible to use variables with high reliability.

4. *No unmeasured confounding variables.* The existence of unmeasured variables which are related to the outcome and have unequal distributions in the treatment groups is a general problem in the analysis of nonrandomized studies (Section 5.1). Let us consider what happens when an ANCOVA is performed which does not consider such a variable. Suppose that there exists a variable Z with means $\overline{Z}_1$ and $\overline{Z}_0$ for the groups. Then, instead of (8.1), the true model might be described by

$$Y = \mu = \alpha_i + \beta X + \gamma Z + e \qquad i = 0, 1. \tag{8.5}$$

In this case, the appropriate adjustment becomes

$$\overline{Y}_{ia} = \overline{Y}_i - \beta(\overline{X}_i - \overline{X}) - \gamma(\overline{Z}_i - \overline{Z}) \qquad i = 0, 1.$$

Thus if we adjust using X only as a covariate, and if $\overline{Z}_1 \neq \overline{Z}_0$, we have adjusted for only part of the differences between groups which is related to Y. Further discussion of this issue can be found in Cochran and Rubin (1973), Cronbach et al. (1977), and Weisberg (1979).

5. *Errors independent of each other.* The error terms (e) in the model are random variables which are assumed to be probabilistically independent of one another. This means that the value of the error term corresponding to any observation has no systematic relationship to that of any other error term.

Nonindependence of errors can affect the validity of tests used to compare treatment groups, but will not introduce bias into the estimates of treatment effects. Nonindependence is difficult to detect empirically, and there is usually no reason to suspect its occurrence. However, in some situations there may be theoretical considerations suggesting nonindependence. Suppose, for instance, that the rural children in our example actually came from a small number of families. Then we might expect high correlations between the error terms corresponding to children in the same family. Roughly speaking, the effect of such intercorrelations is to reduce the effective sample size on which inferences are based. That is, the precision is lower than would be expected on the basis of the sample size used.

6. *Equality of error variance.* Ordinarily, as in most applications of linear models, it is assumed that all error terms have the same variance. In an ANCOVA situation, it is possible that the treatment grops have different error variances. The estimates of treatment effects will still be unbiased in this case, but the validity of tests may be affected. If there is some reason to suspect this inequality of error variances, the residuals from the fitted lines in the two groups can be compared. If the variances of these residuals differ greatly, caution in the interpretation of test results is advised (see Glass et al., 1972).

7. *Normality of errors.* For the ANCOVA tests to be strictly valid, it must be assumed that the errors follow a normal distribution. Departures from normality may affect statistical tests and the properties of estimators in a variety of ways, depending on the actual form of the error distribution. The normality assumption can be tested by examining the distribution of residuals. While severe departures from normality may affect the properties of tests, ANCOVA appears to be generally rather robust (see Elashoff, 1969; Glass et al., 1972). Thus most researchers assume that the normality assumption is not critical.

8.5 DEALING WITH DEPARTURES FROM THE ASSUMPTIONS

As indicated in Section 8.4, several assumptions underlie the use of ANCOVA. Departures from these assumptions may result in biased effect estimates and/or a loss of precision in statistical tests and estimates. While the precision of a statistical procedure is important when the sample size is not large, our primary emphasis in this book has been on the reduction of bias in nonrandomized studies.

In this section we consider what can be done when various departures from the standard ANCOVA assumptions are suspected. Of the seven conditions discussed in Section 8.4, only four bear seriously on the possibility of bias: linearity of the relationship between Y and X, same slopes for regression lines in the two groups, absence of measurement error in the covariate, and absence of other unmeasured covariates.

8.5.1 Nonparallel Regressions

As in Section 8.4, we consider first the case of linear but nonparallel regressions. This is the situation illustrated in Figure 8.2. Since the slopes of the lines, as well as their intercepts, differ in the two groups, the basic model becomes

$$Y + \alpha_i + \beta_i X + e \qquad i = 0, 1. \tag{8.6}$$

From (8.6) the difference between the expected outcomes of the individuals with the same X but in different groups is given by

$$\alpha_1 - \alpha_0 + (\beta_1 - \beta_0)X.$$

That is, the treatment effect is a linear function of X. To estimate this function, we can compute estimates of $\hat{\beta}_1$ and $\hat{\beta}_0$ separately from the two treatment groups and form

$$\overline{Y}_{1a} = \overline{Y}_1 - \hat{\beta}_1(\overline{X}_1 - X)$$
$$\overline{Y}_{0a} = \overline{Y}_0 - \hat{\beta}_0(\overline{X}_0 - X),$$

the treatment means adjusted to an arbitrary point X. Taking the difference yields an unbiased estimated of the treatment effect for any X:

$$\overline{Y}_{1a} - \overline{Y}_{0a} = \overline{Y}_1 - \overline{Y}_0 - \hat{\beta}_1(\overline{X}_1 - X) + \hat{\beta}_0(\overline{X}_0 - X). \tag{8.7}$$

If a single summary value is desired, some "typical" value of X must be inserted in this expression. This might be $\overline{X}$, the mean of X in the two groups together, or the mean from some other standard population. The choice of an X value at which to estimate the treatment effect must be guided by logical rather than statistical considerations. The value should be one that is of practical importance. For example, if we know that the treatment will be applied in the future to individuals with an average value that is at least approximately known, we may whish to estimate the effect at this value.

In many situations, the individuals to receive treatment in the future are expected to be similar to those receiving treatment during the study. So we might wish to choose $X = \overline{X}_1$ in (8.7). We then obtain

$$\overline{Y}_{1a} - \overline{Y}_{0a} = \overline{Y}_1 - \overline{Y}_0 - \hat{\beta}_0(\overline{X}_1 - \overline{X}_0). \tag{8.8}$$

This is of the same form as the standard ANCOVA estimate of the treatment mean difference except that $\hat{\beta}_0$, the estimate based on control group data only has replaced $\hat{\beta}$, the estimate based on pooling the data from both groups. This estimate, first suggested by Belson (1956) and later analyzed by Cochran (1969), is not widely known but offers advantages over the usual estimate in some situations.

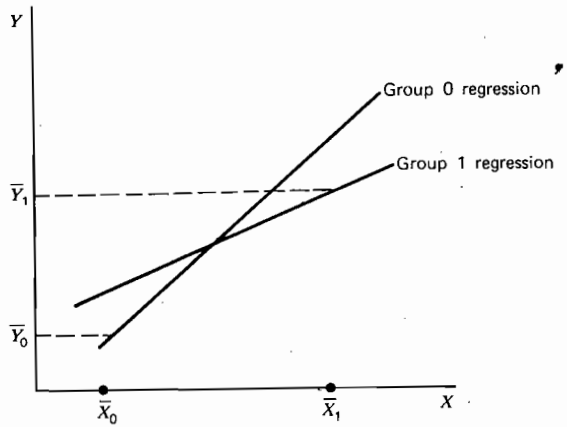
Suppose first that the usual ANCOVA model (8.1) holds. In this case the Belson estimate is unbiased but somewhat less precise (larger variance) than

he usual estimate. On the other hand, particularly if the control group receives ı traditional treatment modality, there may be outside evidence and/or a large ample available to estimate $\hat{\beta}_0$. These factors may outweigh the loss of data from he treatment group.

If the true slopes in the two groups are different, the Belson estimate still has a meaningful interpretation. As noted above, it represents an estimate of the difference in outcomes for individuals with an X value of $\overline{X}_1$. That is, it estimates the effect for a typical individual in the group that received the treatment.

Note that in one sense, (8.7) is more general than the usual ANCOVA model. The usual model represents the special case when $\beta_1 = \beta_0$. On the other hand, unless $\beta_1 = \beta_0$, we cannot use the pooled estimate of β, based on combined data from the two groups. Estimating separate coefficients, as in (8.7), entails the use of smaller samples for each estimated coefficient. For modest sample sizes, this may lead to a slight decrease in precision.

The methods we have so far considered for comparing treatments when regression lines are nonparallel involve specifying a particular covariate value and estimating the effect conditional on this value. If we have some reason for focusing attention on a particular X value, or set of values, this approach will be useful. In some situations, however, we may be more interested in identifying the set of X values for which each treatment is preferable. Figure 8.2 illustrates a situation where for all X values of practical interest, treatment 1 is superior. In Figure 8.3, however, we have a case where treatment 1 is superior for small values of the covariate but inferior for large values. Knowing even approximately where the break-even point is located could have important practical implications.



gure 8.3. Crossing linear regressions in two groups.

Because our estimates of the regression coefficients β_1 and β_0 are subject to sampling variability, we cannot specify the crossing point exactly. However, it is possible to determine a region of X values where the treatment effect is significantly positive or significantly negative, at a specified level of statistical significance. For other values of X, we cannot make a useful statement about which treatment is superior. This approach is known as the Johnson–Neyman technique (Johnson and Neyman, 1936). A good exposition of the technical details and some refinements of the original procedure can be found in Potthoff (1964), and a less technical exposition in Walker and Lev (1959, Chap. 14).

8.5.2 Nonlinear Regressions

The second major threat to the validity of the ANCOVA is nonlinearity of the regressions of Y and X. There are essentially three cases to consider here. The first is illustrated by the solid lines in Figure 8.4: the regressions of Y on X are nonlinear but parallel in the two groups. The treatment effect is in principle clearly defined, but may be difficult to estimate in practice.

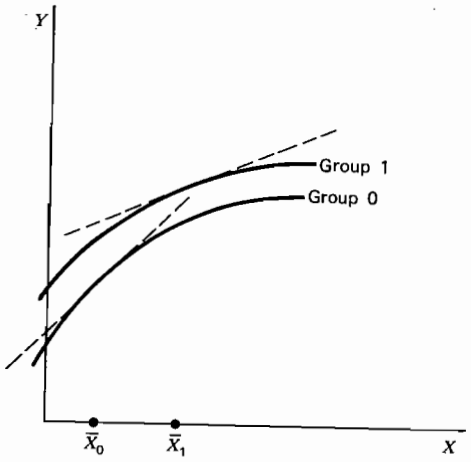


Figure 8.4. Parallel nonlinear regressions.

Let us consider first what happens when we carry out a standard ANCOVA in this situation. Loosely speaking, if $\bar{X}_1$ and $\bar{X}_0$ are not too far apart, and the curvature of the regression not too great, the fitted lines will be approximately parallel and not too misleading. The farther apart $\bar{X}_1$ and $\bar{X}_0$ are, the more different will be the slopes of the curve at the X values in the two groups, and the greater will be the difference in the estimated slopes of the two regression lines.

(The dashed lines in Figure 8.4 illustrate the two different linear regressions.) So, we will be faced with all the problems of nonparallel regression described above.

One way to handle suspected nonlinearity is to model the nonlinear regressions. By making a transformation of the X variable, we may obtain a much better fit to the observations. For example, we might find that a model of the form

$$Y = \alpha_i + \beta\sqrt{X} + e$$

adequately describes the data. A standard ANCOVA can then be carried out using $\sqrt{X}$ rather than X as the covariate.

The second case to consider is that of a nonlinear relationship between Y and X which is not necessarily parallel for the two groups, but which can be turned into a standard model by appropriate transformations of Y and/or X. For example, suppose that the model is given by

$$Y = \exp(\alpha_i + \beta X + e).$$

Then

$$\log Y = \alpha_i + \beta X + e$$

Thus using a logarithmic transformation on Y will allow the standard ANCOVA to be employed. Of course, it must be remembered after the analysis that the effect is defined in transformed (in this case, logarithmic) units. So it may be necessary to transform back to the original units in order to interpret the estimated effect. For example, suppose that an ANCOVA on the log scale produces an estimate

$$\alpha_1 - \alpha_0 = 3.$$

Then in terms of the original model, we have

$$Y = \begin{cases} \exp(\alpha_0 + \beta X + e) & \text{for treatment 0,} \\ \exp(3 + \alpha_0 + \beta X + e) & \text{for treatment 1.} \end{cases}$$

Note that

$$\exp(3 + \alpha_0 + \beta X + e) = \exp(3)\exp(\alpha_0 + \beta X + e)$$
$$= (20.1)\exp(\alpha_0 + \beta X + e).$$

So the estimated effect of changing from treatment 0 to treatment 1 is to *multiply* the response by a factor of about 20.

Finding the appropriate transformations is largely a matter of trial and error. Standard statistical texts offer some guidance (see, e.g., Chatterjee and Price, 1977; Draper and Smith, 1966; Mosteller and Tukey, 1977; Tukey, 1977).

The third case involves nonlinear, nonparallel regressions where no suitable transformation can be found. In this case, both interpretational and technical

problems become very difficult. Some recent research has been conducted on the comparison of quadratic regressions (Rogosa, 1977; Borich et al. 1976; Wunderlich and Borich, 1973), but in general the analyst can do no better than to fit separate regressions for the two groups.

8.5.3 Measurement Error

The third possible threat to the validity of ANCOVA is measurement error in the covariate. Classical measurement theory (see Lord and Novick, 1968) defines the *reliability* of a variable as the percentage of its variance attributable to variation in the true characteristic of interest. This notion is meaningful if we think of the observed score as the sum of true and error components. In Section 5.2 we discussed measurement error as a general issue in statistical adjustment.

Suppose that if we knew the true covariate scores, an ANCOVA model using them would accurately describe the data. Sometimes the equations relating outcomes to true (but unmeasurable) covariates are known as *structural equations*. If we use our imperfectly reliable, but observable, covariate, the resulting $\hat{\beta}$ turns out to be a biased estimate of the β in the structural equation. A biased treatment comparison will result, with the nature of the bias depending upon the nature of the measurement error. An appropriate correction is possible if the reliability of the covariate is known or can be estimated (see Cochran, 1968; DeGracie and Fuller, 1972; Lord, 1960; Stroud, 1974). However, these methods are quite complex and heavily dependent on certain untestable assumptions. So it is probably wiser to focus attention on collecting reliable information rather than trying to assess precisely the degree of reliability and adjust for it in the analyses.

8.5.4 Multiple Confounding Variables

Finally, we discuss the situation when other differences between groups in addition to those related to our measured covariate are suspected. If we can identify and measure other confounding variables, we can adjust for several covariates at once. Suppose, for example, that the model described by (8.5) holds. Then it is possible to obtain unbiased estimates $\hat{\beta}$ and $\hat{\gamma}$ of both β and γ to use these in our adjustment. For example, the adjustment treatment means would be given by

$$\overline{Y}_i - \hat{\beta}(\overline{X}_i - \overline{X}) - \hat{\gamma}(\overline{Z}_i - \overline{Z}).$$

Combining the ability to use transformations of the data with the capability for multivariate adjustments allows great flexibility in fitting an appropriate model for the data. This flexibility must, however, be weighed against the need to verify that all assumptions are met in this more complex situation.

We discussed above the problems in the single covariate situation resulting from possible differences in regression slopes, nonlinearity, and measurement error. With multiple covariates these problems are compounded. When several covariates are involved, we cannot use simple graphical methods to help in assessing the validity of assumptions, and models for measurement error become extremely complex.

The data analyst is faced with a dilemma. To obtain a good fit to the data for each group and include all potential confounding factors, he or she is tempted to include several covariates. But the more covariates included, the greater the potential problems in meeting and verifying the basic ANCOVA assumptions.

Now it might be though that the analyst should simply include the one or two most important possible confounding factors, expecting to eliminate most of the bias and avoiding the complexity of multiple covariates. While this procedure may often work well, there are situations where it can be quite misleading. It may even result in an estimate of treatment effect that is more biased than the unadjusted difference of group means. An artificial example of this phenomenon was given in Section 5.1. As another example of how this might occur, suppose that in the Greenberg (1953) data the rural children were not only older, but also tended to have shorter parents in such a way that the effects of these two factors, age and heredity, were exactly counterbalanced. Then, by using ANCOVA to adjust for age differences between groups, we would unwittingly create an artificial difference between the groups.

This example illustrates the care which is necessary in drawing inferences on the basis of ANCOVA. While a preponderance of short parents in one group might be an obvious factor to take into account, a confounding variable may be much more difficult to identify in other practical problems. It would be nice to give some simple guidelines for dealing with this problem. Unfortunately, there is no way to guarantee that the ANCOVA model is correctly specified. As with other statistical adjustment strategies considered in this book, the investigator may be criticized for omitting a particular confounding variable thought by someone else to be important. The general discussion of this problem contained in Chapter 5 includes some broad guidelines on choosing an adequate covariate set. A more detailed discussion of the issues in the context of ANCOVA is presented by Weisberg (1979), and some practical guidelines are offered by Cochran (1965).

APPENDIX 8A FORMULAS FOR ANALYSIS-OF-COVARIANCE CALCULATIONS

We consider the general situation where K treatments are being compared. These will be indexed by $k = 1, 2, \ldots, K$. Let X_{ik} and Y_{ik} represent the covariate

and outcome values for individual i in group k. Let $\bar{X}_k$ and $\bar{Y}_k$ be the means for the n_k individuals in group k. Then we can define the between-group (treatment) sums of squares and cross-products by

$$T_{xx} = \sum_{k=1}^{K} n_k(\bar{X}_k - \bar{X})^2$$

$$T_{yy} = \sum_{k=1}^{K} n_k(\bar{Y}_k - \bar{Y})^2$$

$$T_{xy} = \sum_{k=1}^{K} n_k(\bar{X}_k - \bar{X})(\bar{Y}_k - \bar{Y}),$$

where $\bar{X}$ and $\bar{Y}$ are the grand means of X and Y across all groups. Similarly, we define within-group (error) sums of squares and cross-products:

$$E_{xx} = \sum_{k=1}^{K} \sum_{i} (X_{ik} - \bar{X}_k)^2$$

$$E_{yy} = \sum_{k=1}^{K} \sum_{i} (Y_{ik} - \bar{Y}_k)^2$$

$$E_{xy} = \sum_{k=1}^{K} \sum_{i} (X_{ik} - \bar{X}_k)(Y_{ik} - \bar{Y}_k),$$

where Σ_i indicates the sum over individuals within each group. We also define the quantity

$$f = \text{total number of subjects minus number of groups}$$

and, using the definitions above we define

$$S_{xx} = T_{xx} + E_{xx}$$

$$S_{yy} = T_{yy} + E_{yy}$$

$$S_{xy} = T_{xy} + E_{xy}.$$

Then we can calculate the residual mean squares for treatments and error:

$$s_e^2 = \left(E_{yy} - \frac{E_{xy}^2}{E_{xx}}\right) \Big/ (f - 1)$$

$$s_t^2 = \left(T_{yy} - \frac{S_{xy}^2}{S_{xx}} + \frac{E_{xy}^2}{E_{xx}}\right) \Big/ (K - 1).$$

These can be used to calculate an F statistic to test the null hypothesis that all treatment effects are equal:

$$F = \frac{s_t^2}{s_e^2}.$$

Under the null hypothesis this ratio has an F distribution with $K - 1$ and $f - 1$ degrees of freedom. The estimated regression coefficient of Y on X is

$$\hat{\beta} = \frac{E_{xy}}{E_{xx}}.$$

From the definitions of E_{xy} and E_{xx} given above, it is clear why this is called a pooled within-group estimator. The estimated standard error for the adjusted difference between two group means (say group 0 and group 1) is given by

$$s_d = s_e \sqrt{\frac{1}{n_0} + \frac{1}{n_1} + \frac{(\bar{X}_1 - \bar{X}_0)^2}{E_{xx}}},$$

when n_0 and n_1 are the sample sizes of the two groups. A test of the null hypothesis that the adjusted difference is zero is provided by the statistic

$$t = \frac{\bar{Y}_1 - \bar{Y}_0 - \hat{\beta}(\bar{X}_1 - \bar{X}_0)}{s_d}.$$

Under the null hypothesis, it has a t distribution with $f - 1$ degrees of freedom.

REFERENCES

Belson, W. A. (1956), A Technique for Studying the Effects of a Television Broadcast, *Applied Statistics*, **5**, 195–202.

Borich, G.D., Godbout, R. C., and Wunderlich, K. W. (1976), *The Analysis of Aptitude-Treatment Interactions: Computer Programs and Calculations*, Austin, TX: Oasis Press.

Chatterjee, S., and Price, B. (1977), *Regression Analysis by Example*, New York: Wiley.

Cochran, W. G. (1957), Analysis of Covariance: Its Nature and Uses, *Biometrics*, **13**, 261–281.

Cochran, W. G. (1965). The Planning of Observational Studies of Human Populations, *Journal of the Royal Statistical Society, Series A*, **128**, 234–266.

Cochran, W. G. (1968), Errors of Measurement in Statistics, *Technometrics*, **10**(4), 637–666.

Cochran, W. G. (1969), The Use of Covariance in Observational Studies, *Applied Statistics*, **18**, 270–275.

Cochran, W. G., and Rubin, D. B. (1973). Controlling Bias in Observational Studies: A Review, *Sankhya, Series A*, **35**(4), 417–446.

Cronbach, L. J., Rogosa, D. R., Floden, R. E., and Price, G. G. (1977) Analysis of Covariance in Nonrandomized Experiments: Parameters Affecting Bias, occasional paper of the Stanford Evaluation Consortium.

DeGracie, J. S., and Fuller, W. A. (1972), Estimation of the Slope and Analysis of Covariance When the Concomitant Variable Is Measured with Error, *Journal of the American Statistical Association*, **67**, 930–937.

Draper, N. R., and Smith, H. (1966), *Applied Regression Analysis*, New York: Wiley.

Elashoff, J. D. (1969), Analysis of Covariance, a Delicate Instrument, *American Educational Research Journal*, **6**, 383–402.

Fisher, R. A. (1932), *Statistical Methods for Research Workers,* 4th ed., Edinburgh: Oliver & Boyd.

Glass, G. V., Peckham, P. D., and Sanders, J. R. (1972), Consequences of Failure to Meet Assumptions Underlying the Analysis of Variance and Covariance, *Review of Educational Research,* **42,** 237–288.

Greenberg, B. G. (1953), The Use of Analysis of Covariance and Balancing in Analytical Surveys, *American Journal of Public Health,* **43**(6), 692–699.

Johnson, P. O., and Neyman, J. (1936), Tests of Certain Linear Hypotheses and Their Applications to Some Educational Problems, *Statistical Research Memoirs,* **1,** 57–93.

Lord, F. M. (1960), Large-Scale Covariance Analysis When the Control Variable Is Fallible, *Journal of the American Statistical Association,* **55,** 307–321.

Lord, F. M., and Novick, M. R. (1968), *Statistical Theories of Mental Test Scores,* Reading, MA: Addison-Wesley.

Mosteller, F. C., and Tukey, J. W. (1977), *Data Analysis and Regression,* Reading, MA: Addison-Wesley.

Potthoff, R. F. (1964), On the Johnson–Neyman Technique and Some Extensions Thereof, *Psychometrika,* **29,** 241–256.

Rogosa, D. R. (1977), Some Results for the Johnson–Neyman Technique, unpublished doctoral dissertation, Stanford University.

Stroud, T. W. F. (1974), Comparing Regressions When Measurement Error Variances Are Known, *Psychometrika,* **39**(1), 53–68.

Tukey, J. W. (1977), *Exploratory Data Analysis,* Reading, MA: Addison-Wesley.

Walker, H. M., and Lev, J. (1959), *Elementary Statistical Methods,* New York: Holt, Rinehart and Winston.

Weisberg, H. I. (1979), Statistical Adjustments and Uncontrolled Studies, *Psychological Bulletin,* **86,** 1149–1164.

Winer, B. J. (1971), *Statistical Principles in Experimental Design,* 2nd ed., New York: McGraw-Hill.

Wunderlich, K. W., and Borich, G. D. (1973), Determining Interactions and Regions of Significance for Curvilinear Regressions, *Educational and Psychological Measurement,* **33,** 691–695.

CHAPTER 9

Logit Analysis

Logit analysis can be applied in comparative studies to estimate the effect of a risk factor on a dichotomous outcome factor as measured by the odds ratio. The usefulness of logit analysis is its ability to adjust for many confounding variables simultaneously. These confounding variables can be either categorical or numerical.

We will begin by motivating the mathematical model that underlies logit analysis (Section 9.1) and showing how logit analysis can be used to control for a confounding variable (Section 9.2). Details of various aspects of implementation are given in Sections 9.3 to 9.8 with some additional mathematical details in Appendix 9A. Initially, discussion is restricted to cohort studies; the differences applicable to case-control studies are presented in Section 9.6.