CHAPTER 5

Some General Considerations in Controlling Bias

Before presenting the various methods for controlling bias, we will raise several important caveats. The basic theme of this chapter is that the validity of any statistical adjustment rests on a set of assumptions which may be difficult to verify. In addition to an understanding of technical details, judgment is required in order to apply these techniques properly. Since a certain amount of practical experience is necessary to develop good judgment, we can offer no simple formulas. However, we can point out the major problems that arise in practice and some general approaches which are helpful in dealing with them.

5.1 OMITTED CONFOUNDING VARIABLES

In order to obtain a valid estimate of the treatment effect, the analyst must be sure that the variables used for adjustment include all important confounding

factors. In Chapter 2 we defined a confounding factor as a variable that has the following properties:

1. Is statistically associated with the risk factor.
2. Directly affects the outcome.

The main problem is to verify part 2 of this definition. The judgment that a particular variable exerts a direct causal influence on the outcome cannot be based on statistical considerations; it requires a logical argument or evidence from other investigations.

For example, suppose that we are investigating the effectiveness of an educational program aimed at improving the reading ability of elementary school children. Two classes are being compared, one receiving the new program and one utilizing the standard curriculum. The children have been rated on scales indicating the level of parent education and family economic circumstances. Suppose that the class receiving the new program contains a higher proportion of poor children. Then, if poverty is thought to have a direct influence on reading ability, it can be considered a confounding factor. But poverty may be closely linked to parent education in a complex causal relationship. Although some of the effect of parent education may be attributable to economic circumstances per se, there may be an independent component related to education itself. So even if we compared two equally poor children receiving identical treatments, we would still expect differences in parents' education to result in different expected reading abilities. That is, conditional on economic status, parent education still constitutes a confounding factor.

Now it might seem that including either education or economic status as adjustment variables would be reasonable, even though using both would be better. Moreover, if there exist other unmeasured variables mediating the effects of these two variables in combination, failure to include them would not seem very serious. In randomized studies that is in fact the case. Omitting a relevant variable results in less precise estimation, but the estimate of effect is unbiased. In nonrandomized studies, however, serious problems can result.

To see more clearly the nature of these problems, let us consider a hypothetical example. Suppose that in reality there are only two confounding factors, X_1 and X_2. Tables 5.1 and 5.2 display the joint frequency distribution of X_1 and X_2 and the average outcome values given X_1 and X_2 under the treatment and control conditions. From the calculations shown in Table 5.2, it is clear that if we do not adjust at all, we will estimate the treatment effect as

$$\text{Estimate of effect} = 51.25 - 61.25 = -10.$$

However, we can see that for each possible combination of X_1 and X_2 values,

Table 5.1 Joint and Marginal Frequency Distributions of X_1 and X_2[a]

Treatment Group

X_1 \ X_2	0	1	Total
0	10	90	100
1	90	10	100
Total	100	100	200

Comparison Group

X_1 \ X_2	0	1	Total
0	90	10	100
1	10	90	100
Total	100	100	200

[a] Each factor has two levels, denoted by 0 and 1.

Table 5.2 Average Outcome Values

Treatment Group

X_1 \ X_2	0	1
0	100	50
1	50	25

Comparison

X_1 \ X_2	0	1
0	100	50
1	50	25

Estimated average outcomes ignoring X_1 and X_2:

Treatment group: $$\frac{100(10) + 50(90) + 50(90) + 25(10)}{200} = 51.25$$

Comparison group: $$\frac{100(90) + 50(10) + 50(10) + 25(90)}{200} = 61.25$$

Estimated average outcomes adjusting for X_1 only:

Treatment group: $X_1 = 0$ $$\frac{100(10) + 50(90)}{100} = 55.0$$

$X_1 = 1$ $$\frac{50(90) + 25(10)}{100} = 47.5$$

Comparison group: $X_1 = 0$ $$\frac{100(90) + 50(10)}{100} = 95.0$$

$X_1 = 1$ $$\frac{50(10) + 25(90)}{100} = 27.5$$

the average outcome is equal for the two groups. That is, the treatment really has no effect, and our estimate is therefore incorrect.

Now let us adjust this estimate, by conditioning on values of X_1 alone. This simple method for correcting bias constitutes a special case of stratification, which is described in detail in Chapter 7. From the information at the bottom of Table 5.2, we obtain

$$\begin{aligned}\text{Estimate of effect} &= (\text{estimate given } X_1 = 0)P(X_1 = 0) \\ &\quad + (\text{estimate given } X_1 = 1)P(X_1 = 1) \\ &= (-40)\frac{100}{200} + (20)\frac{100}{200} = -10.\end{aligned}$$

In this example, adjustment by X_1 alone results in no reduction of bias. Similarly adjustment by X_2 alone removes no bias. Together, however, the two variables eliminate bias completely. More complex examples can be developed in which the bias when adjusting by each variable individually actually increases, although using both together eliminates all bias. Fortunately, such extreme situations are rare.

Of course, there is never a guarantee that all important variables have been considered. It is the analyst's responsibility to present evidence that for individuals who are equal in terms of variables included, there is no variable that still satisfies conditions 1 and 2 given at the beginning of this section. More precisely, if a variable does satisfy these conditions, its marginal effect must be very small.

Note also that in our example, the marginal distributions of X_1 and X_2 are identical in the two groups. So it might appear that they are not confounding variables according to the definition given in Chapter 2. But if we apply the definition to the joint distribution, we see that it does apply to the pair (X_1, X_2), which together constitute a confounding factor.

The dilemma posed by statistical adjustments is that no matter what variables we include in the analysis (X_1 in our example), there may be an omitted variable (X_2 in our example) that together with the included variables constitutes a confounding variable. Moreover, it is not enough to demonstrate that all plausible confounding variables excluded have similar distributions across groups. As with our example, such a variable may still be important in combination with others. So the analyst must be fairly certain that no variable has been left out which mediates the causal effect of those variables included.

It is clear that judgment and experience are necessary in selecting variables. Also, close collaboration between statisticians and scientists in both the design and analysis of a study is highly desirable. The problems in selecting variables are primarily substantive and not statistical, although there are some statistical guidelines that may often prove useful.

Cochran (1965) suggests that the background variables be divided into three classes:

1. A small number of major variables for which some kind of matching or adjustment is considered essential. These are usually determined by knowledge of the specific subject matter and review of the literature.
2. Variables that *may* require matching or adjustment.
3. Variables that are believed to be unimportant or for which data are not available.

Decisions regarding the variables that fall in category 2 can be very difficult. The problem is similar to that of model specification in the context of multiple regression (see Cox and Snell, 1974; Mosteller and Tukey, 1977, Chap. 16). In regression analysis, we want to include enough relevant variables to ensure that the resulting model is a correct description of the relationship between an outcome variable and a set of input variables. A commonly used criterion for the importance of a particular variable, given a set of other variables, is the decrease in the proportion of explained variation when that variable is excluded. Since this number depends on which other variables are also included in the analysis, no unique measure of "importance" can be defined. However, by calculating this quantity for each variable in a proposed set and trying various plausible sets, it is often possible to get a sense of which variables play the most important causal roles.

In choosing variables for statistical adjustment, a similar idea can be applied. for each variable of a given set, the change in the adjustment that would result from omitting it can be calculated. By examining various possible combinations, we can sometimes get a good sense of which variables are the confounding factors. For example, suppose that one particular variable consistently makes a large difference in the estimated effect, regardless of which other variables are included, while all other variables have smaller effects that depend strongly on the composition of the whole variables set. In such a situation we would be satisfied to use only this one variable in our analysis.

While part 2 of the definition given at the beginning of this section is hardest to verify, part 1 is also important. A variable that is strongly related to outcome is confounding only of its distribution differs appreciably across the treatment groups. So before tackling the more difficult task described above, the analyst may want to reduce the number of potential factors by eliminating those variables with similar distributions across groups. However, as noted above, joint distributions as well as those of each variable separately must be considered.

In this context, an appreciable difference among treatment groups is not necessarily the same as a statistically significant difference. Significance tests place the burden of proof on the rejection of the null hypothesis. As we indicated in Chapter 2, a large difference will not be statistically significant if it has an even larger standard error. In small studies, associations between background variables and the treatment which are large enough to dictate the estimate of

treatment effect may not be statistically significant. The opposite problem can occur in large studies. Weak association between treatment and background variables may be statistically significant and yet be too small to affect the estimate of treatment effect.

A systematic method for examining the joint distributions of background variables is discriminant analysis. The *discriminant function* is defined as that linear combination of the background variables which maximizes the ratio of the "between-group" component of variance to the "within-group" component. Among all linear combinations of the original variables, the discriminant is the one which best separates the two groups. A thorough discussion of discriminant analysis is given by Lachenbruch (1975).

Having obtained the discriminant function, we can examine the joint distributions of those variables which enter into it most prominently. Alternatively, we can take the discriminant function itself as a single new confounding factor. Since the discriminant will generally include small contributions from many relatively unimportant background variables, we may wish to screen out some variables at the outset.

Further discussion of the variable selection problem in the context of discriminant analysis is given by Cochran (1964). He considers whether the effect of including specific variables in the discriminant functions can be assessed from the discriminating power of those variables considered individually. Although standard statistical theory warns that it cannot, an examination of 12 well-known numerical examples from the statistical literature revealed the following:

1. Most correlations (among background variables) are positive.
2. It is usually safe to exclude from a discriminant, before computing it, a group of variables whose individual discriminatory powers are poor, except for any such variate that has negative correlations with most of the individually good discriminators.
3. The performance of the discriminant function can be predicted satisfactorily from a knowledge of the performance of the individual variables as discriminators and of the average correlation coefficient among the variables.

We close this section with a brief discussion of those variables for which data are not available (Cochran's class 3). As noted above, failure to collect data on an important confounding variable can put the results of the study in serious question, particularly when the magnitude of the estimated treatment effect is small (even if it is statistically significant). However, when the magnitude of the treatment effect is large, one can often say that, even if an important confounding factor had been overlooked, it could not have accounted for the size of the observed effect. Bross (1966, 1967) has devised a quantification of this argument which he calls the "size rule." The basic idea behind the size rule is

to specify, for a given observed association between treatment and outcome, how large the associations between treatment and confounding factor and between confounding factor and outcome must be to explain away the observed treatment effect. However, as noted by McKinlay (1975), the derivation of Bross's rules requires assumptions that limit the applicability of his results.

5.2 MEASUREMENT ERROR

Many observed variables really reflect two kinds of information. In part the value of the variable is governed by some stable individual characteristic that can be expected to relate to other characteristics in a systematic way. In part, however, it is determined by "random" fluctuations related to the particular circumstances under which the observation happened to be taken. This error component can vary across measurement situations even if the individual has not changed.

Measurement error is particularly troublesome in the fields of education and psychology, where the variables studied are often scores on psychometric tests. Many extraneous factors besides stable individual differences may influence test scores. Psychometricians have developed the concept of *reliability* as a way to quantify the amount of measurement error. Loosely speaking, the reliability represents the proportion of total variation comprised by variation in the underlying true score. The higher the reliability, the more confidence we can have that something real is being measured.

However, true scores are not directly observable. So various indirect methods must be used to assess the reliability of a variable measured with error, or *fallible* variable. For example, under certain assumptions, the correlation between scores of the same test given individuals at two different points in time can be used to estimate the reliability. For our purposes, the general concepts of measurement error and reliability will suffice. The reader interested in more detail on these concepts is referred to Lord and Novick (1968). We now consider the effects of measurement error on statistical adjustments.

One way to describe the effects of measurement error is in terms of omitted confounding variables. The presence of error in the observed variable means that there exists, in effect, an additional variable (error) that ought to be included along with the observed score as a confounding factor. To see this more clearly, suppose that there exists a dichotomous confounding variable T (for true score) which can have values 0 and 1.

The frequencies of the two possible values of T in the two groups are given in Table 5.3 together with the average outcome conditional on each T value. For simplicity, we assume that the real treatment effect is 0. Then if we knew T for each individual, we could calculate separate estimates of the treatment effect

within the two groups ($T = 0$; $T = 1$). Except for sampling fluctuation, the correct value of 0 would result.

Table 5.3 Frequencies and Average Outcomes for $T = 0$ and $T = 1$

		Frequencies	
		Treatment	Comparison
T	0	100	25
	1	100	175
	Total	200	200
		Average Outcomes	
T	0	50	
	1	100	

Now assume that T cannot be observed directly, but we can measure a variable X that reflects both T and measurement error E, where

$$E = \begin{cases} 0 & \text{if } X = T \\ 1 & \text{if } X = 1 - T \end{cases}$$

Then there is a joint distribution of X and E in each treatment group. For example, consider the distribution shown in Table 5.4. The relationship among T, X, and E can be expressed as in Table 5.5, and the average outcome values are as given in Table 5.6. From Table 5.6 it is clear that if we could obtain information on E as well as that for X, the pair (X, E) would constitute a confounding factor. Using X alone corresponds to the use of X_1 in the example of Section 5.1, and E plays the role of X_2.

Table 5.4 Joint Frequency Distributions of X and E[a]

Treatment Group

		E = 0	1	Total
X	0	80	20	100
	1	80	20	100
	Total	160	40	200

Comparison Group

		E = 0	1	Total
X	0	20	35	55
	1	140	5	145
	Total	160	40	200

[a] Each factor has two levels, denoted by 0 and 1.

Table 5.5 Relationship among T, X, and E

		E = 0	E = 1
X = 0		T = 0	T = 1
X = 1		T = 1	T = 0

Table 5.6 Mean Outcome as a Function of X and E

Treatment Group

	E = 0	E = 1
X = 0	50	100
X = 1	100	50

Comparison Group

	E = 0	E = 1
X = 0	50	100
X = 1	100	50

The main point of this section is that measurement error constitutes one special form in which an omitted confounding factor can arise. By adjusting on the basis of a fallible variable, we are ignoring the variable E, which is the discrepancy between X and the true score T. If we knew both X and E, we would know T and could adjust on it.

Of course, there is an implicit assumption here that adjustment on T would eliminate all bias. If this is not the case, the relationship between adjustment on the basis of X versus T is more complicated. The reader interested in more details is referred to Weisberg (1979).

5.3 THE REGRESSION EFFECT

Measurement error represents one very common example of omitted confounding variables. Another is the phenomenon of regression effects (see Thorndike, 1942). Mathematically, regression effects can be easily explained, but heuristic interpretations are often confusing. Rather than attempt a general exposition, we will discuss regression effects in the context of a concrete example.

Suppose that a remedial program is given to a group of children in a particular

school. The aim is to improve their reading ability. A pretest is given prior to the intervention and a posttest just after the program. A hypothetical data set is presented in Table 5.7. From these data we can calculate the mean score at the two testing points.

$$\text{Pretest mean} = 10.0$$
$$\text{Posttest mean} = 13.0$$

So the children have gained 3.0 points during the course of the program. But this 3.0 points represents the sum of a treatment effect plus any natural maturation that might have occurred anyway. In Chapter 12 we consider this particular kind of confounding in more detail. Our purpose here is simply to illustrate how regression effects can occur.

Table 5.7 Hypothetical Data on Treatment Group to Illustrate Regression Effect[a]

Pretest Score	Posttest Score 8	9	10	11	12	13	14	15	16
13						1	1	1	1
12					1	1	2	1	1
11				1	2	3	3	2	1
10			1	1	3	4	3	1	1
9			1	2	3	3	2	1	
8			1	1	2	1	1		
7			1	1	1	1			
6									
5									

Reprinted, by permission from Campbell and Stanley (1966), Fig. IA, copyright 1966, American Educational Research Association, Washington, D.C.

[a] Numbers indicate how many children received the particular combination of pretest and posttest scores.

Instead of looking at the entire group of children, let us focus on those who are farthest from the mean. Children with scores of 7 on the pretest receive an average score of 11.5 on the posttest. Although they start out 3 points below the mean, they end up only 1.5 points below the posttest mean. Those scoring 13 on the pretest (3 points above the mean) end up with an average posttest score of 14.5, which is only 1.5 points above the mean.

In general, any group of children selected on the basis of their pretest scores will (on the average) have posttest scores closer to the mean. This phenomenon is known as *regression toward the mean.* It results from imperfect correlation between pretest and posttest.

To understand the effect of regression toward the mean on methods for controlling bias, imagine that a comparison group has been selected from a nearby school. The data on these comparison children are shown in Table 5.8. For this comparison group

$$\text{Comparison pretest mean} = 8.0$$
$$\text{Comparison posttest mean} = 10.0$$

Because the groups started out at different levels (10.0 for treatment vs. 8.0 for comparison), a straightforward comparison of the posttest scores may be biased.

Table 5.8 Hypothetical Data on Comparison Group from a Different School[a]

	Posttest								
Pretest	7	8	9	10	11	12	13	14	15
13									
12									
11				1	1	1	1		
10			1	1	2	1	1		
9		1	2	3	3	2	1		
8	1	1	3	4	3	1	1		
7	1	2	3	3	2	1			
6	1	1	2	1	1				
5	1	1	1	1					

[a] Numbers indicate how many children received the particular combination of pretest and posttest scores.

One common approach in such situations is to match individuals with identical scores in the two groups. (Matching is discussed in detail in Chapter 6.) For example, we could compare the average scores of individuals with pretest scores of 7. Then, because these individuals all start out equal, we might expect the comparison to yield an unbiased estimate of the treatment effect. To see what actually happens in this situation, suppose that the true treatment effect is really zero. That is, the changes between pretest and posttest are entirely the result of natural growth.

Now consider what happens when we compare across groups. We have already seen that the 4 children scoring 7 on the pretest obtain an average of 11.5. What about the 12 children scoring 7 in the comparison school? These have an average posttest score of only 9.5. They have regressed toward the mean of their own population, which is 10.0 rather than 13.0. As a result, the estimated effect is

$$\hat{\alpha} = 11.5 - 9.5 = 2.0.$$

Because of differential regression to the mean in the two groups, a *regression effect* is generated. Even though there is no treatment effect, the treatment group appears to be doing better than the comparison group. Controlling for the pretest score in this manner does not eliminate bias completely.

We have presented this example at length because it represents the kind of explanation that is often given for biased estimates of effect after matching or statistical adjustment. Moreover, unlike our previous example, it illustrates the problem in the context of numerical confounding variables. However, the crux of the problem posed by the regression effect is simply that the variables used in carrying out the adjustment (e.g., a pretest or test on a related skill) represent an incomplete description of the differences between groups. That is, we have omitted some important confounding factors. The different joint distributions of pretest and posttest in the two groups represent another way to describe the fact that, conditional on the pretest, there still exist confounding factors that can bias the treatment comparison. Two children with identical pretest scores, but in different schools, do not have the same expectation on the posttest. For example, one school may already have a remedial reading program for younger children that tends to inflate pretest performance. Unless we include a variable that reflects the effect of this remedial program, the analysis can be seriously biased.

5.4 SPECIFYING A MATHEMATICAL MODEL

So far we have discussed possible problems relating to the variables used for adjustment, but have not focused on the particular method of analysis. For illustrative purposes, we have introduced simple forms of matching, or stratification, because they allow the basic issues to be seen clearly. However, the adequacy of statistical adjustment in an actual situation depends not only on using the correct variables, but also on applying a technique whose assumptions are valid.

Most of the methods we present in this book assume a particular mathematical form for the relationship among outcomes, risk variables, and covariates. These mathematical *models* will be discussed in detail in later chapters. Our purpose in this section is to discuss the general issue of proper model specification.

For simplicity we assume that only a single variable X is needed for adjustment. While in general the problems of model error and incomplete covariates are intertwined, we wish here to isolate the modeling problems. Let us define

Y = outcome variable (numerical)

α_i = treatment effect for individual i

Now in general the treatment effect may vary across individuals and may even

be systematically related to X (in Chapters 2 and 3 we discussed this issue of *interaction*). We shall assume here that the effect is the same for every subject.

Further, to highlight the main issues, we assume that there is no treatment effect; that is $\alpha_i = 0$ for all individuals. Figure 5.1 is an illustration of a typical relationship between Y and X in this situation. Then there exists some mathematical function, $g(X)$, relating X and the expected value of Y. That is, the average value of Y is given by $g(X)$. Now, in general, this function may differ in the two treatment groups. But with no treatment effect this would mean that the groups differed on some additional factor besides X. So because we are assuming X to be the only confounding variable, this function must be the same in both groups.

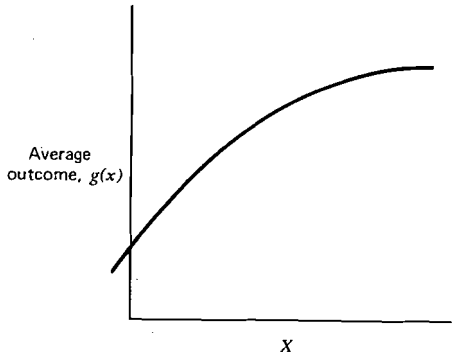

Figure 5.1 *Typical relationship between average outcome and confounding factor.*

Now, $g(X)$ is itself a variable. Moreover, because the distribution of X can differ in the two treatment groups, so can the distribution of $g(X)$. Let

$$E_1[g(X)] = \text{expected value of } g(X) \text{ in treatment group}$$
$$E_0[g(X)] = \text{expected value of } g(X) \text{ in comparison group}$$

Then we can define the bias in estimating the (zero) treatment effect as

$$E(\bar{Y}_1 - \bar{Y}_0) = E_1[g(X)] - E_0[g(X)] = \eta.$$

That is, on the average, the difference between the group means depends on the distributions of X in the two groups *and* the functional form of the relationship between outcome and covariate.

Because X can be measured in the two groups, its distribution can be determined. So if the mathematical form of g can be specified, the amount of bias

can be estimated and subtracted from the raw mean difference. How, then, might this function be found?

Recall that so far we have been assuming that the treatment effect is zero, so that the same functional form holds for both groups. Now, let us suppose that the treatment has an unknown effect we wish to estimate but that it is constant across individuals. Then

$$Y = g(X) \qquad \text{for comparison group}$$
$$Y = \alpha + g(X) \qquad \text{for treatment group}$$

and

$$E(\bar{Y}_1 - \bar{Y}_0) = \alpha + \eta.$$

So we want to divide the total mean difference into two components, a part (α) attributable to the treatment and a part (η) resulting from differences between groups on the distribution of X. There are two possible ways to accomplish this: (*a*) use the comparison group data only to estimate g and then calculate α, and (*b*) fit a model including both $g(X)$ and α directly, using all the data on both groups.

Although in general it is possible to estimate any functional form, there is one class of mathematical functions that is particularly convenient: the linear functions. With only one X, a linear relationship has the form

$$g(X) = \mu + \beta X.$$

The graph of such a function is a straight line. A useful property of linear functions is that the average value of a function of X is the function of the average value of X. This means that

$$\begin{aligned} E(\bar{Y}_1 - \bar{Y}_0) &= \alpha + \mu + \beta\bar{X}_1 - (\mu + \beta\bar{X}_0) \\ &= \alpha + \beta(\bar{X}_1 - \bar{X}_0) = \alpha + \eta \end{aligned}$$

that is,

$$\eta = \beta(\bar{X}_1 - \bar{X}_0).$$

Therefore, we can form the estimate

$$\hat{\alpha} = \bar{Y}_1 - \bar{Y}_0 - \beta(\bar{X}_1 - \bar{X}_0),$$

and we will have

$$E(\hat{\alpha}) = \alpha + \eta - \eta = \alpha.$$

The details of this approach will be elaborated upon in Chapter 8, where we refer to it as the analysis of covariance.

The assumption of linearity greatly facilitates the analysis of data from nonrandomized studies. This assumption is at the heart of several techniques

discussed in this book. Although it may seem that linearity is a very strong condition, it still allows a certain amount of flexibility when used in conjunction with transformation of the data. Even though the relationship between Y and X may not be linear, it may be possible to rescale either or both to bring about linearity. For example, if the relationship between Y and X is exponential,

$$g(X) = e^{\mu+\beta X},$$

Then

$$\log g(X) = \mu + \beta X.$$

So by using the logarithm of Y as the outcome measure, a linear model analysis is possible.

The estimation of $g(X)$ as the basis for statistical adjustment allows a much more efficient use of the data than do such approaches as matching or stratification, which do not depend on a model. In matching, for instance, it may be difficult to find a large enough number of close matches to allow precise estimation. This issue is discussed in detail in Chapter 6. By assuming a mathematical structure, we may be able to estimate α precisely using relatively small sample sizes.

On the other hand, if the model used turns out to be incorrect, our results may be misleading. Suppose, for example, that we are using the comparison group data to estimate $g(X)$ in the absence of the treatment, and that $g(X)$ actually has the nonlinear form illustrated in Figure 5.2. The numbers at the bottom of Figure 5.2 represent the X values for comparison (0) and treatment group (1) subjects. Suppose that we estimate a linear model based on comparison group

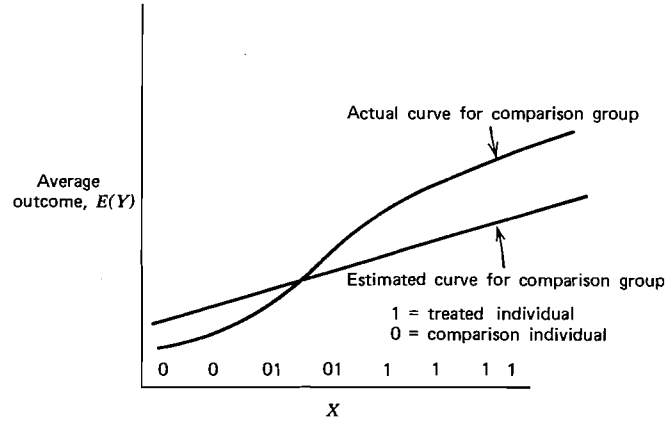

Figure 5.2 *Model misspecification.*

data. Then for individuals with high X values, the linear function underestimates their expected outcome. For those with low X values, the expected outcome is overestimated. If the treatment group tends to lie near the high end, as shown in Figure 5.2, and the control group near the low end, the actual outcome difference η produced by this difference will be much larger than that estimated on the basis of linearity. The estimate of α will be correspondingly biased.

It is hard to say how severe the departure from the assumed model must be to cause serious problems. Determining an adequate model requires judgment as well as a knowledge of particular statistical methods. In each of the subsequent chapters on individual techniques, more detail will be given on the model assumptions and how they can be verified.

Finally, we note that problems of variable selection, including measurement error and regression effects, are intertwined with those of model selection. A correctly specified model must include appropriate variables *and* have a proper mathematical form. When we transform a variable, we change both the variable and the functional form. What matters is whether the model and variables ultimately employed in the analysis accurately represent the underlying phenomenon.

5.5 SAMPLING ERROR

Throughout the previous discussion we have largely ignored the fact that analyses are often based on small or moderate sample sizes. We have focused on problems that will cause the estimated effect to deviate from the actual effect even with very large samples. We now discuss an additional source of error, that attributable to sample fluctuation.

For illustrative purposes, suppose that the true model underlying a set of data is given by the following equations:

Treatment: Average outcome = $5 + X$

Comparison: Average outcome = $2 + X$

This situation is illustrated in Figure 5.3. The treatment effect is 3 in this example.

Of course, in a real situation we will not know the exact relationship between X and the outcome. The problems in using the wrong mathematical model were discussed in Section 5.4. We showed, for example, that using a linear function when a nonlinear one is appropriate can lead to bias in estimating the treatment effect. Now let us assume that the functional form is in fact linear, but that we must estimate the slope and intercept from a given set of data. The sets of observed values may look as shown in Figure 5.4. The model states that *on the average,* for a given value of X, the treated individuals have a value $5 + X$ and

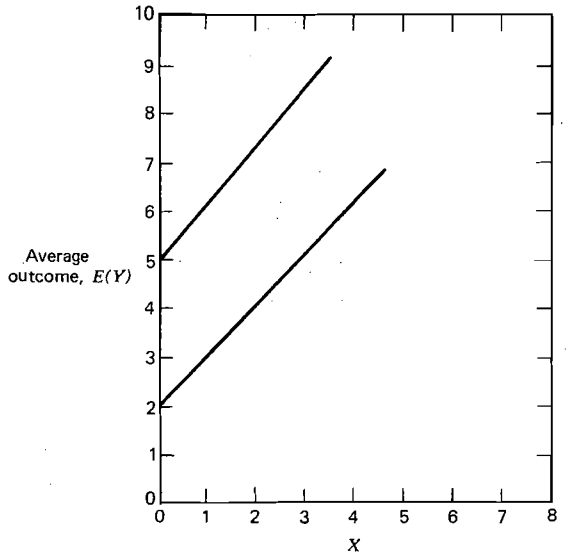

Figure 5.3 *Linear relationship between average outcome and confounding factor: constant treatment effect.*

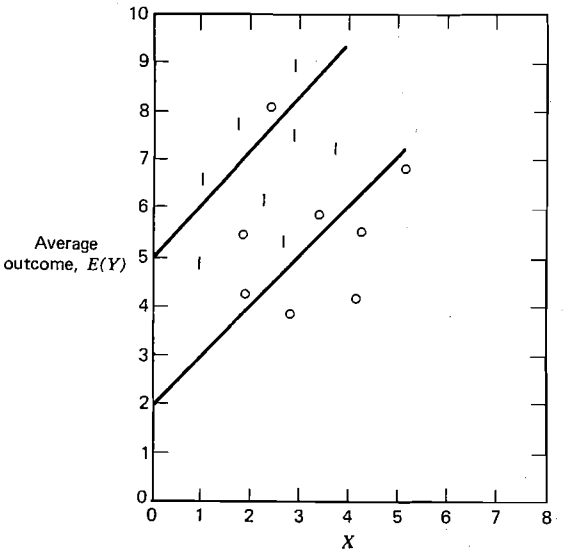

Figure 5.4 *Scatter of actual observations around expected value.*

the control subjects $2 + X$. However, the individual scores fluctuate randomly about these lines. If the number of subjects is very large, it is possible to estimate the true intercepts and slopes with great precision. However, with a sample of only 20 or 30 in each group, there can be substantial variation in these estimates from sample to sample.

Without going into detail on statistical techniques, let us imagine that we are estimating the slope using the data from the two groups combined. The difference between the estimated intercepts is then the estimator of the treatment effect. For 10 independent samples with 20 in each group, we would obtain results that vary around the true value of 3, but differ from sample to sample. The estimator may be correct on the average and therefore be what is called by statisticians an *unbiased* estimator, as mentioned in Chapter 2. However, for any particular sample there will be a *sampling error,* which may be substantial. The sampling error will generally become smaller and smaller as the sample size increases, although for some estimators it is not negligible, even for very large samples. A precise consideration of these matters would involve technicalities beyond the level of this book. Unless otherwise stated, we can assume that sampling error will disappear for a large-enough sample size.

5.6 SEPARATION OF GROUPS ON A CONFOUNDING FACTOR

In order for a confounding factor to create substantial bias in estimating a treatment effect, its distribution in the two treatment groups must differ significantly. However, if the groups are very widely separated on a confounding variable used in the analysis, certain problems mentioned in the previous sections become particularly severe. Figure 5.5 illustrates the situation where the groups are widely separated on a variable X. We have mentioned that one basic approach to bias control is the comparison of individuals with identical (or similar) values of X. This matching, if exact, will remove any bias attributable to X regardless of the functional form of the relationship between outcome and X. But it is clear that if the groups are completely separated, no matches can be found. More generally, if there is little overlap, matches may be found for only a small proportion of subjects. The feasibility of matching with different degrees of separation is discussed more completely in Chapter 6.

A second problem is that extreme separation may be an indicator that the two groups are quite different in character. So there are likely to be other variables on which they differ that are related to the outcome. It may be difficult to find a few variables that capture all the relevant variation. For example, suppose that the treatment group includes only individuals under 35 years of age and the control group contains only individuals over 35. Then the groups represent dif-

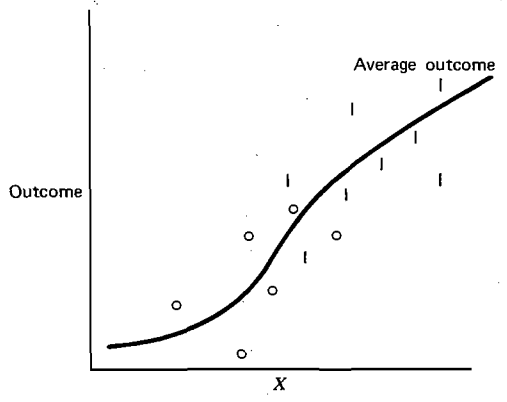

Figure 5.5 *Wide separation of groups on the confounding factor.*

ferent "generations," with quite different experiences at comparable ages and quite different life-styles. It may be meaningless to compare such groups.

A third problem is the loss of precision in estimating a model relating the expected outcome to the confounding variable. In Section 5.4 we mentioned that one approach is to estimate the function on the basis of the comparison group data only. However, when the groups are widely separated, there will be very few observations on comparison subjects in the range of X values occupied by the treatment group, and so it will be hard to obtain a precise estimate. In the case of complete separation, we must rely on extrapolation of the estimated function completely beyond the range of the data, a procedure that is always hazardous.

If we assume a known functional form and a constant treatment effect, we can, instead, estimate the treatment effect from the data on both groups. However, if there are very few observations in the range where the two distributions overlap, we must rely heavily on model assumptions, such as the assumption of no interaction. An incorrect model specification will be very difficult to detect.

The problem of complete, or near-complete, separation may sometimes arise from the desire to give a certain treatment to those who are thought to need it most. Thus there may be a conflict between research design criteria and ethical considerations. Sometimes this conflict can be resolved by an imaginatively designed study. Mather et al. (1971) report on a study of 1203 episodes of acute myocardial infarction (heart attacks). The purpose of the study was to compare home care by the family doctor with hospital treatment initially in an intensive care unit. Normally, such a comparison would be impossible—the less severely ill patients would be sent home, the emergency cases to the intensive care unit. We can imagine an index of "severity" being measured on each patient. This

would clearly satisfy our definition of a confounding factor, but the distribution of this factor within the home-care and hospital-care group would not have substantial overlap.

Here, however, there was agreement between various hospitals and doctors participating in the study that while some patients would clearly need hospitalization and others should clearly be treated at home, there were some patients for whom the decision was not clear-cut. For these patients, randomization was used to decide between home care and hospitalization. The decision on acceptability of a random assignment was made by the patient's own doctor, before he knew what the result of the randomization would be.

In all, 343 cases were allocated at random, and subsequent analysis confirmed that the randomized groups did not differ substantially in composition with respect to other background variables, such as age, past history of heart disease, and blood pressure when first examined. It was found that the randomized group treated at home had a 44% lower mortality than did the randomized group treated in the hospital. As might be expected, the experience of the other two groups was very different. The conclusion that home care is better than hospitalization had only been firmly established for the randomized group, although we might speculate that it would also hold for at least some other individuals.

As a final comment, we note that when there are several potential confounding variables, it is possible that the two groups are completely separated on these variables considered jointly, although the distributions of each variable individually do have substantial overlap. Consider the situation illustrated by Figure 5.6, where X_1 and X_2 represent two background variables: X_1 = age (decades),

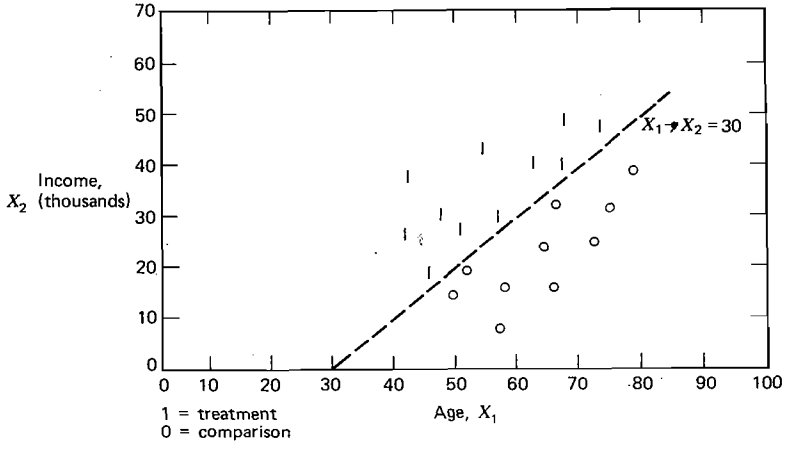

Figure 5.6 *Complete separation of joint distributions even though marginal distributions overlap.*

X_2 = income. Individuals in group 1 tend to have a higher income and lower age than those in group 0, but the distributions of each variable in the two groups have substantial overlap. However, the particular age–income combinations that occur in one group do not occur in the other. So the groups are separated on X_1 and X_2 jointly, even though they are not separated on X_1 and X_2 individually.

This problem can be very hard to recognize and emphasizes the need for a multivariate exploration of the potential confounding factors. In this example, the discriminant method might suggest that

$$D = X_1 - X_2$$

can be used to distinguish the treatment and control subjects. Note that all individuals with values of D greater than 30 are in the control group, while those with D less than 30 are in the treatment group.

5.7 SUMMARY

In this chapter we have discussed the general problems that may affect statistical adjustment strategies applied to nonrandomized studies. These problems may be seen as operating at three levels:

1. Variable selection.
2. Specifying form of mathematical model.
3. Small-sample fluctuation.

Variable selection involves knowledge of the substantive area under investigation. The aim is to include enough information to ensure that after adjusting for the measured variables, there will be no bias in the estimate of treatment (risk factor) effect. There will be no bias if the only systematic difference between two individuals with identical measured values is directly caused by the treatment (risk factor). In attempting to verify this assumption, statistical methods may be helpful, but only in conjunction with a careful analysis of possible causal relationships. Ideally, statisticians and substantive researchers should work together to select a variable set that can be defended on both statistical and conceptual grounds.

Two particularly common problems are measurement error and regression effects. We have pointed out how these can be viewed as special cases of omitted confounding factors. As such, they do not pose different problems or require special solutions. If a proposed variable set includes fallible variables, or those subject to "regression," it simply means that we must be sure to include enough other variables so that the total set is adequate.

Having an adequate set of adjustment variables allows us *in principle* to obtain an unbiased estimate of the treatment effect. However, to obtain such an estimate we must employ one of the techniques described in the subsequent chapters of this book. Each analysis strategy is based on a particular set of assumptions about the mathematical form of relationships among variables. To the extent that these assumptions do not hold in a given situation, the results may be biased.

Finally, even with an adequate set of variables and a correctly specified model, we are subject to problems arising from finite samples. That is, the estimate obtained from a particular analysis may contain a component attributable to random fluctuations. For very large sample sizes, we would expect these errors to be negligible, but for small samples we can expect the estimate to deviate substantially from the true effect. Where possible, confidence bounds should be provided in addition to a point estimate.

In each of Chapters 6 to 11 we present the basic concepts and mechanics underlying one approach to statistical adjustment. Each of these techniques is vulnerable to the general problems described in this chapter, and we will not repeat in each chapter the general caveats given here. However, we will explain in some detail how these considerations apply to the particular technique, trying to indicate what problems are most likely to arise and how to deal with them.

After reading these chapters, the reader should have a clearer understanding of the issues raised in this chapter. In Chapter 14 we will review some of these issues and present additional areas related more specifically to the methods described in Chapters 6 to 11.

REFERENCES

Bross, I. D. J. (1966), Spurious Effects from an Extraneous Variable, *Journal of Chronic Disease,* **19,** 736–747

Bross, I. D. J. (1967), Pertinency of an Extraneous Variable, *Journal of Chronic Disease,* **20,** 487–495.

Campbell, D. T., and Stanley, J. G. (1966), *Experimental and Quasi-experimental Designs for Research,* Chicago: Rand McNally.

Cochran, W. G. (1964), On the Performance of the Linear Discriminant Function, *Technometrics,* **6,** 179–190.

Cochran, W. G. (1965), The Planning of Observational Studies of Human Populations, *Journal of the Royal Statistical Society, Series A,* **128,** 254–266.

Cox, D. R., and Snell, F. J. (1974), The Choice of Variables in Observational Studies, *Journal of the Royal Statistical Society, Series C,* **23,** 1, 51–59.

Lachenbruch, P. A. (1975), *Discriminant Analysis,* New York: Hafner.

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

McKinlay, S. (1975), The Design and Analysis of the Observational Study—A Review, *Journal of the American Statistical Association,* **70,** 503–520.

Mather, H. G., Pearson, N. G., Read, K. L. Q., Shaw, D. B., Steed, G. R., Thorne, M. G., Jones, S., Guerrier, C. J., Eraut, C. D., McHugh, P. M., Chowdhurg, N. R., Jafary, M. H., and Wallace, T. J. (1971), Acute Myocardial Infarction: Home and Hospital Treatment, *British Medical Journal,* **3,** 334–338.

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

Thorndike, F. L. (1972), Regression Fallacies in the Matched Group Experiment, *Psychometrika,* 7(2), 85–102.

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

CHAPTER 6

Matching