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

The major concern in making causal inferences from comparative studies is that a proper standard of comparison be used. A proper standard of comparison (see Chapter 1) requires that the performance of the comparison group be an adequate proxy for the performance of the treatment group if they had not received the treatment. One approach to obtaining such a standard is to choose study groups that are comparable with respect to all important factors except for the specific treatment (i.e., the only difference between the two groups is the treatment). Matching attempts to achieve comparability on the important potential confounding factor(s) at the design stage of the study. This is done by appropriately selecting the study subjects to form groups which are as alike as is possible with respect to the potential confounding variable(s). Thus the goal of the matching approach is to have no relationship between the risk and the potential confounding variables in the study sample. Therefore, these potential confounding variables will not satisfy part 1 of the definition of a confounding variable given at the beginning of Chapter 2, and thereby will not be confounding variables in the final study sample. This strategy of matching is in contrast to the strategy of adjustment, which attempts to correct for differences in the two groups at the analysis stage.

We stated that matching "attempts to achieve comparability" because it is seldom possible to achieve exact comparability between the two study groups. This is especially true in the case of several confounding variables. To judge how effective the various matching procedures can be in achieving comparability and thus reducing bias in the estimate of the treatment effect, it is necessary to model the relationship between the outcome or response variable and the confounding variable(s) in the two treatment groups. Since much of the research has been done assuming a numerical outcome variable that is linearly related to the confounding variable, we will tend to emphasize this type of relationship. The reader should not believe, however, that matching is applicable only in this case. There are matching techniques which are relatively effective in achieving comparability and reducing bias in the case of nonlinear relationships.

Before presenting the various matching techniques, we shall illustrate in Section 6.1 how making the two treatment groups comparable on an important confounding variable will eliminate the bias due to that variable in the estimate

of the treatment effect. Section 6.1 expands on material presented in Section 3.2.

The degree to which the two groups can be made comparable depends on (*a*) how different the distributions of the confounding variable are in the treatment and comparison groups, and (*b*) the size of the comparison population from which one samples. These factors influence the amount of bias reduction possible using any of the matching techniques, and are discussed in Section 6.2.

In the last introductory section of this chapter, Section 6.3, we list and discuss the conditions under which the results for the various matching techniques are applicable. Although these conditions are somewhat overly restrictive, they are necessary for a clear understanding of the concepts behind the various techniques.

Finally, the main emphasis of this chapter is on the reduction of the bias due to confounding. The other two sources of bias, bias due to model misspecification and estimation bias, however, can also be present. See Sections 5.4 and 5.5 for a discussion of these other sources of bias. All of the theoretical results that we present are for the case of no model misspecification. This should be kept in mind when applying the results to any study.

6.1 EFFECT OF NONCOMPARABILITY

For the sake of illustration, reconsider the example introduced in Chapter 3, the study of the association between cigarette smoking and high blood pressure. Recall that cigarette smoking is the risk variable and age is an important confounding variable. This last assumption implies that the age distributions of the smokers and nonsmokers must differ: otherwise, age would not be related to the risk variable (i.e., the groups would be comparable with respect to age). We shall further assume that the smokers are generally older (see Figure 3.3) and that the average blood pressure increases with age at the same rate for both smokers and nonsmokers (see Figure 3.4). Let X denote age in years and Y denote diastolic blood pressure in millimeters of mercury (mm Hg). The effect of the risk factor, cigarette smoking, can be measured by the difference in average blood pressure for any specific age, and because of the second assumption, this effect will be the same for all ages.

These two assumptions can be visualized in Figure 6.1. Suppose that we were to draw large random samples of smokers and nonsmokers from the populations shown in Figure 3.3. The sample frequency distributions would then be as illustrated in Figure 6.1 by the histograms. The smokers in the sample tend to be older than the nonsmokers. In particular, the mean age of the smokers is larger than that of the nonsmokers, $\bar{X}_S > \bar{X}_{NS}$. (Notice that the Y axis in Figure 6.1 does not correspond to the ordinate of the frequency distributions.)

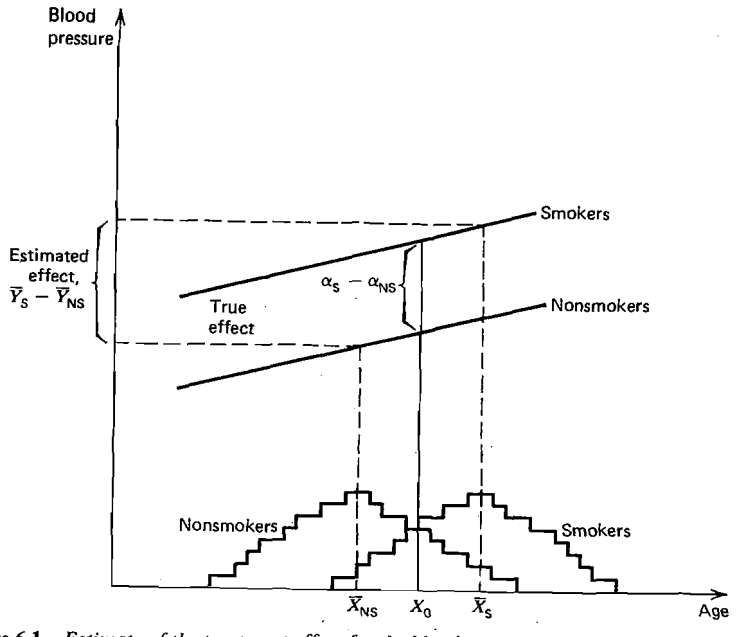

Figure 6.1 *Estimate of the treatment effect for the blood pressure–smoking example.*

The second assumption, specifying that the relationship between age and diastolic blood pressure in both groups is linear, is represented by the lines labeled "Smokers" and "Nonsmokers" (as in Figure 3.4). Algebraically, these relationships are:

$$Y_S = \alpha_S + \beta X \qquad \text{for smokers}$$
$$Y_{NS} = \alpha_{NS} + \beta X \qquad \text{for nonsmokers,} \tag{6.1}$$

where Y_S and Y_{NS} represent the average blood pressure levels among persons of age X, and β is the rate at which Y, blood pressure, changes for each 1-year change in X. [Note that for simplicity of presentation, random fluctuations or errors (Section 2.2) will be ignored for now.] For a specified age, X_0, therefore, the effect of the risk factor is

$$\begin{aligned} Y_S - Y_{NS} &= \alpha_S - \alpha_{NS} + \beta(X_0 - X_0) \\ &= \alpha_S - \alpha_{NS} \end{aligned} \tag{6.2}$$

(see Figure 6.1).

Let us first consider the simplest situation, where there is only one subject in each group and where each subject is age X_0. We will then have two groups that are exactly comparable with respect to age. The estimate of the treatment effect is the difference between the blood pressures of the two subjects. Since the blood pressures of these two subjects are as given in (6.1) with $X = X_0$, the estimated treatment effect will be as given in (6.2). Thus exact comparability has led to an unbiased estimate of the treatment effect. (Note that the same result would also hold for any nonlinear relationship between X and Y.)

Next consider the estimate of the treatment effect based on all subjects in the two samples. The estimate is found by averaging over all the values of Y in both groups and calculating the difference between these averages:

$$\bar{Y}_S - \bar{Y}_{NS} = \alpha_S - \alpha_{NS} + \beta(\bar{X}_S - \bar{X}_{NS}). \tag{6.3}$$

Thus because of the noncomparability of the two groups with respect to age, the estimate of the risk effect is distorted or biased by the amount $\beta(\bar{X}_S - \bar{X}_{NS})$. Since we do not know β, we cannot adjust for this bias. (An adjustment procedure based on estimating β is analysis of covariance; see Chapter 8.) Notice, however, that if we could equalize the two sample age distributions, or in the case considered here of a linear relationship, restrict the sampling so that the two sample means were equal, we would then obtain an unbiased estimate of the treatment effect. By making the groups comparable, one would be assured of averaging over the same values of X.

There are two basic approaches to forming matches to reduce bias due to confounding. These are referred to as pair and nonpair matching. *Pair matching* methods find a specific match (comparison-subject) for each treatment subject. It is clear that if we restrict the choice of subjects in the two groups such that for every treatment subject with age X_0 there is a comparison subject with exactly the same age, then by (6.2) the difference in blood pressures between each matched pair is an unbiased estimate of the treatment effect. Hence the average difference will also be unbiased.

Because of difficulties in finding comparison subjects with exactly the same value of a confounding variable as a treatment subject, various pair matching methods have been developed. For example, if the confounding variable is numerical, it is practically impossible to obtain exact matches for all treatment subjects. An alternative method, caliper matching, matches two subjects if their values of X differ by only a small tolerance (Section 6.4). In the case of a categorical confounding variable, one can use a pair matching method called stratified matching (Section 6.6). However, these methods cannot always guarantee the desired sample size, so another pair matching method, called nearest available pair matching (Section 6.5), was developed by Rubin (1973a).

In the second approach to matching, *nonpair matching,* no attempt is made to find a specific comparison subject for each treatment subject. Thus there are no identifiable pairs of subjects. There are two nonpair matching methods: frequency and mean matching. In frequency matching, Section 6.7, the distri-

bution of the confounding variable in the treatment group is stratified and one attempts to equalize the two distributions by equalizing the number of treatment and comparison subjects in each stratum. Mean matching, Section 6.8, attempts to reduce the amount of bias by equating just the sample means rather than attempting to equalize the two distributions as in the previous methods. The comparison group, which is of the same size as the treatment group, thus consists of those subjects whose group mean is closest to the mean of the treatment group.

6.2 FACTORS INFLUENCING BIAS REDUCTION

None of the matching methods requires the fitting of a specific model for the relationship between the response and the confounding variables. The effectiveness of a matching procedure, however, will depend on the form of the relationship between the response and the confounding variables. In addition, the effectiveness depends on the following three factors: (*a*) the difference between the means of the treatment and comparison distributions of a confounding variable, (*b*) the ratio of the population variances, and (*c*) the size of the control sample from which the investigator forms a comparison group. These three factors will now be discussed in detail.

To understand how these three factors influence the researcher's ability to form close matches and hence to achieve the maximum bias reduction, consider the slightly exaggerated distributions of a confounding variable, X, in the treatment and comparison populations shown in Figure 6.2. Both distributions are normal with a variance of 2.25. The mean of the comparison population is 3, and the mean of the treatment population is 0.

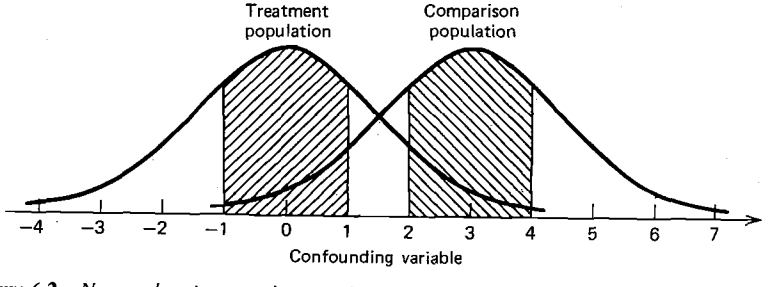

Figure 6.2 *Nonoverlapping samples, equal variances.*

Suppose that we have small random samples from both the treatment and the comparison populations and we wish to find a matched comparison group. Because of the assumed distribution, the treatment group is most likely to have values between −1 and +1, the middle 50% of the distribution (shaded area on

the left in Figure 6.2). The sample from the comparison population is called the *comparison* or *control reservoir;* it is the group of subjects from which one finds matches for the treatment group. Based on the assumed distribution of the confounding variable, the comparison reservoir is most likely to consist of subjects whose values of the confounding variable lie between 2 and 4 (shaded area on the right in Figure 6.2). Thus there would be little overlap between these two samples.

With virtually no overlap between our samples, it is impossible to form matched groups which are comparable. Using any of the pair matching techniques, we could not expect to find many comparison subjects with values of X closer than 1 unit to any treatment subject. Similarly for the nonpair matching methods, regardless of the way one stratifies the treatment frequency distribution, there will not be enough comparison subjects in each stratum. In addition, the means of the two groups would be about 3 units apart. Any attempt to match in this situation would be unwise, since only a small proportion of the two groups could be made reasonably comparable.

Continuing this example, suppose that another, much larger sample is drawn from the comparison population such that the values of X in the reservoir lie between zero and 6. The treatment group remains fixed with values of X between +1 and −1. The resulting overlap of the two samples is shown in Figure 6.3 against the background of the underlying population distributions. Notice that by increasing the size of the comparison sample, we are more likely to have members of the comparison reservoir which have the same or similar values of X as members of the treatment group. The number and closeness of the possible pair matches has improved; for frequency matching we should be able to find more comparison subjects falling in the strata based on the treatment group; and the difference in the sample means, after mean matching, should be less than the previous value of 3. Again, as was the case with nonoverlapping samples, we may still be unable to find adequate matches for all treatment group subjects. This "throwing away" of unmatchable subjects is a waste of information which results in a lower precision of the estimated treatment effect.

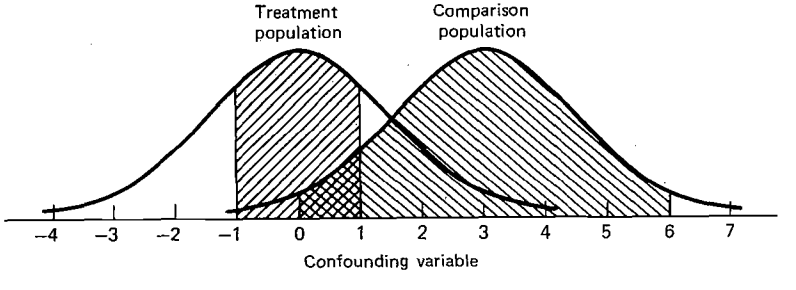

Figure 6.3 *Overlapping samples, equal variances.*

Now consider what would happen if the population variances of the confounding variable were not equal. In particular, suppose that the variance of the treatment population, σ_1^2, remains at 2.25, while the variance of the comparison population, σ_0^2, is 9.0. (Again, this is a slightly exaggerated example but is useful to illustrate our point.) With the treatment sample fixed, random sampling from the comparison population would most likely result in a sample as shown by the shading in Figure 6.4. Notice the amount of overlap that now exists between the treatment group and the comparison reservoir. There are clearly more subjects in the comparison reservoir, with values of the confounding variable between +1 or −1, than in the previous example (Figure 6.3).

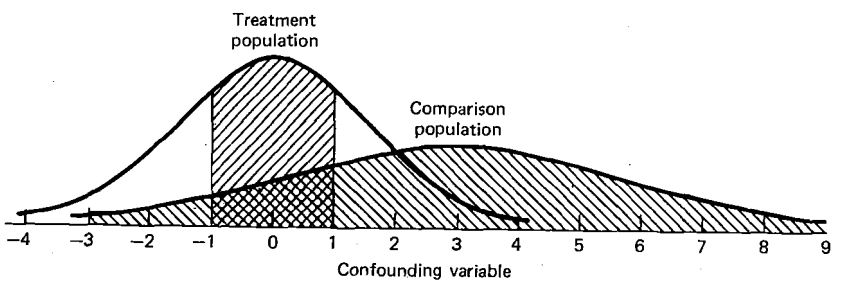

Figure 6.4 *Overlapping samples, unequal variances.*

After comparing these examples, the relationship among the three factors—the difference between the population means of the two distributions, the ratio of the population variances, and the size of the comparison reservoir—should be clear. The farther apart the two population means are, the larger the comparison reservoir must be to find close matches, unless the variances are such that the two population distributions overlap substantially.

To determine numerically the bias reduction possible for a particular matching technique, it is necessary to quantify these three factors. Cochran and Rubin (1973) chose to measure the difference between the population means by a quantity referred to as the *initial difference*. This measure, B_X, may be viewed as a standardized distance measure between two distributions and is defined as

$$B_X = \frac{\eta_1 - \eta_0}{\sqrt{(\sigma_1^2 + \sigma_0^2)/2}}. \tag{6.4}$$

The eta terms, η_1 and η_0, denote the means of the treatment and the comparison populations, respectively. Similarly, σ_1^2 and σ_0^2 represent the respective population variances.

In the first example of this section, the initial difference was equal to 2.0. With the variance of the comparison population increased to 9, however, the initial difference was equal to 1.3 and the two distributions overlapped more.

The ratio of the treatment variance to the comparison variance, σ_1^2/σ_0^2, is the second important factor in determining the number of close matches that can be formed, and hence the bias reduction possible. Generally, the smaller the ratio, the easier it will be to find close matches.

The last factor is the size of the comparison reservoir from which one finds matches. In the previous examples we assumed that the random sample from the treatment population was fixed. That is, we wanted to find a match for every subject in that sample and the subjects in the treatment group could not be changed in order to find matches. Removal of a treatment subject was the only allowable change if a suitable match could not be found. This idea of a fixed treatment group is used in the theoretical work we cite and is perhaps also the most realistic approach in determining the bias reduction possible. An alternative and less restrictive approach assumes that there exists a treatment reservoir from which a smaller group will be drawn to form the treatment group. Such an approach would allow for more flexibility in finding close matches.

In the following methodological sections the size of the comparison reservoir is stated relative to the size of the fixed treatment group. Thus a comparison reservoir of size r means that the comparison reservoir is r times larger than the treatment group. Generally, r is taken to be greater than 1.

6.3 ASSUMPTIONS

In discussing the various matching procedures, we shall make the following assumptions:

1. There is one confounding variable.
2. The risk variable in cohort studies or the outcome variable in case-control studies is dichotomous.
3. The treatment effect is constant for all values of the confounding variable. (This is the no interaction assumption of Section 3.3.)
4. For cohort studies we wish to form treatment and comparison groups of equal size. (For case-control studies, we would construct case and control groups of equal size.)
5. The treatment group (or case group) is fixed.

The assumption of only one confounding variable is made for expository purposes. In Section 6.10 we will discuss matching in the case of multiple confounding variables. The second assumption corresponds to the most common situation where matching is used. Matching cannot be used if the risk variable

in cohort studies or the outcome variable in case-control studies is numerical. The third assumption of no interaction, or parallelism, is crucial for estimating the treatment effect. Researchers should always be aware that implicitly they are making this assumption and when possible they should attempt to verify it. For example, in Section 5.2, we discuss how the assumption of parallelism may be unjustified when one is dealing with fallible measurements. If this assumption is not satisfied, the researcher will have to reconsider the advisability of doing the study or else to report the study findings over the region for which the assumption holds. The fourth assumption, that the treatment and comparison groups are of equal size, is also made for expository purposes. In addition, the efficiency of matching is increased with equal sample sizes for a given total sample size. In Section 6.11 we consider the case of multiple comparison subjects per treatment subject. The last assumption of a fixed treatment group is one of the assumptions under which most of the theoretical work is done. A fixed treatment group is typically the situation in retrospective studies where the group to be studied, either case or exposed, is clearly defined.

While the type of study has no effect on the technique of matching, the forms of the outcome and confounding variables do. The various matching techniques can be used in either case-control or cohort studies. The only difference is that in a cohort study one matches the groups determined by the risk or exposure factor, whereas in a case-control study, the groups are determined by the outcome variable. Throughout this chapter, any discussion of a cohort study applies also to a case-control study, with the roles of the risk and outcome variables reversed.

The form of the risk or outcome variable and the confounding variables (i.e., numerical or categorical) determines the appropriate matching procedure and whether matching is even possible. If the confounding variable is of the unordered categorical form, such as religion, there is little difficulty in forming exact matches. We shall, therefore, make only passing reference to this type of confounding variable. Instead, we shall emphasize numerical and ordered categorical confounding variables, where the latter may be viewed as having an underlying numerical distribution. Numerical confounding variables are of particular importance because exact matching is very difficult in this situation. Most of the theoretical work concerning matching has been done for a numerical confounding variable and dichotomous risk variable (cohort study).

6.4 CALIPER MATCHING

Caliper matching is a pair matching technique that attempts to achieve comparability of the treatment and comparison groups by defining two subjects to be a match if they differ on the value of the numerical confounding variable,

X, by no more than a small tolerance, ϵ. That is, a matched pair must have the property that

$$|X_1 - X_0| \leq \epsilon.$$

The subscript 1 denotes treatment group and 0 denotes comparison group. By selecting a small-enough tolerance ϵ, the bias can in principle be reduced to any desired level. However, the smaller the tolerance, the fewer matches will be possible, and in general, the larger must be the reservoir of potential comparison subjects.

Exact matching corresponds to caliper matching with a tolerance of zero. In general, though, exact matching is only possible with unordered categorical confounding variables. Sometimes, however, the number of strata in a categorical variable is so large that they must be combined into a smaller number of strata. In such cases or in the case of ordered categorical variables, the appropriate pair matching technique is stratified matching (Section 6.6).

6.4.1 Methodology

To illustrate caliper matching we shall consider the cohort study of the association of blood pressure and cigarette smoking.

Example 6.1 Blood pressure and cigarette smoking: Suppose that a tolerance of 2 years is specified and that the ages in the smokers group are 37, 38, 40, 45, and 50 years. (We shall assume that the smoker and nonsmoker groups are comparable on all other important variables.) A comparison reservoir twice the size ($r = 2$) of the smoking group consists of nonsmokers of ages 25, 27, 32, 36, 38, 40, 42, 43, 49, and 53 years. The estimated means of the two groups are 42.0 and 38.5 years, respectively. The ratio of the estimated variances, s_S^2/s_{NS}^2, is $0.37 = 29.50/79.78$.

The first step in forming the matches is to list the smokers and determine the corresponding comparison subjects who are within the 2-year tolerance from each smoker. For our example, this results in the possible pairing given in Table 6.1*a*.

Table 6.1a Potential Caliper Matches for Example 6.1

Smokers	Nonsmokers
37	36, 38
38	38, 40
40	40, 42
45	43
50	49

It is clearly desirable to form matches for all the treatment subjects that are as close as possible. Thus the matched pairs shown in Table 6.1*b* would be formed.

Table 6.1b Caliper-Matched Pairs

Smokers	Nonsmokers
37	36
38	38
40	40
45	43
50	49

Notice that if the 49-year-old nonsmoking subject had not been in the reservoir, we would not have been able to match all five smokers. We might then have decided to keep the first four matches and drop the 50-year-old smoker from the study. This results in a loss of precision, because the effective sample size is reduced. Alternatively, the tolerance could be increased to 3 years and the 50-year-old smoker matched with the 53-year-old nonsmoker. The latter approach does not result in lower precision, but the amount of bias may increase. Finally, had there been two or more comparison subjects with the same value of X, the match subject should be chosen randomly.

In Example 6.1 we knew the composition of the comparison reservoir before the start of the study. Often, however, this is not the case. Consider, for example, a study of the effect of specially trained nurses aids on patient recovery in a hospital. Such a study would require that the patients be matched on important confounding variables as they entered the hospital. When the comparison reservoir is unknown, the choice of a tolerance value that will result in a sufficient number of matched pairs can be difficult. The researcher cannot scan the reservoir, as we did in the example, and discover that the choice of ϵ is too small. For this reason, using caliper matching in a study where the comparison reservoir is unknown can result in matched sample sizes that are too small. In such a situation, the researcher can sometimes attempt to get a picture of the potential comparison population through records (i.e., historical data).

6.4.2 Appropriate Conditions

Caliper matching is appropriate regardless of the form of the relationship between the confounding and outcome variables (or risk variable in case-control studies). In this section we demonstrate how caliper matching is effective in reducing bias in both the linear and nonlinear cases.

Linear Case. To understand how caliper matching works in the linear case, let us consider the estimate of the treatment effect or risk effect of smoking on blood pressure based on a 45-year-old smoker and a 43-year-old nonsmoker in Example 6.1. Assume that blood pressure is linearly related to age and that the relationships are the same for both groups, with the exception of the intercept values. Figure 6.5 represents this situation.

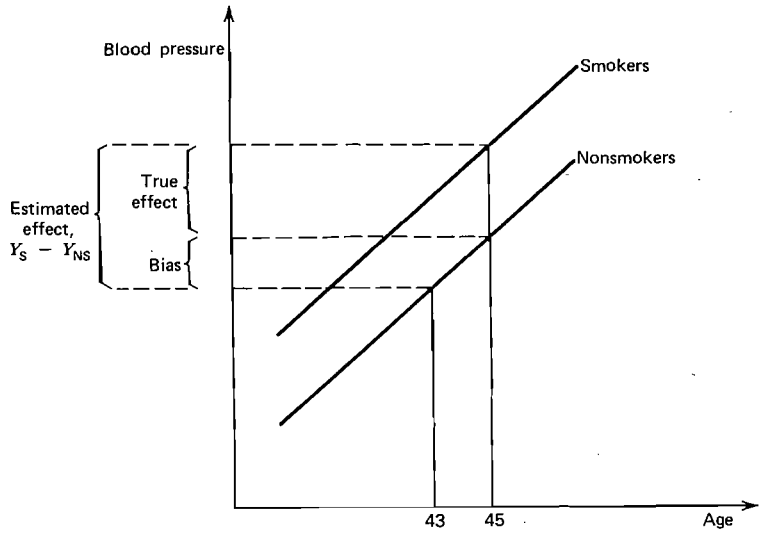

Figure 6.5 *Estimate of treatment effect—linear relationship.*

The estimate of the risk effect is shown by the large brace to the left in Figure 6.5. The amount of bias or distortion is shown by the small brace labeled "Bias." Relating this example to (6.3), we see that the bias is equal to the unknown regression coefficient β, multiplied by the difference in the values of the confounding variable. In the case of these two subjects, the bias is 2β. This is the maximum bias allowable under the specified tolerance for each individual estimate of the treatment effect, and consequently for the estimated treatment effect, based on the entire matched comparison group.

When we average the ages in Table 6.1*b*, we find that the mean age of the smokers is 42.0 years; of the nonsmokers comparison group, 41.2 years; and for the comparison reservoir, 38.5 years. Caliper matching thus reduced the difference in means from 3.5 (= 42.0 − 38.5) to 0.8 (= 42.0 − 41.2). In general, the extent to which the bias after matching, 0.8β in this case, is less than the maximum possible bias, 2β in this case, will depend on the quantities discussed in Section 6.2: the difference between the means of the two populations, the ratio of the variances and the size of the comparison reservoir as well as the tolerance.

Nonlinear Case. Let us now consider the case where the response and the confounding variable are related in a nonlinear fashion. To illustrate the effect of caliper matching in this situation, we shall assume that blood pressure is related to age squared. Algebraically this relationship between the response Y and

age, X, can be written as

$$Y = \alpha + \beta X^2$$

where α is the intercept. The estimate of the treatment effect assuming that Y is numerical is

$$\bar{Y}_S - \bar{Y}_{NS} = \alpha_S - \alpha_{NS} + \beta\,(\bar{X}_S^2 - \bar{X}_{NS}^2). \tag{6.5}$$

Hence any bias is a function of the difference in the means of age squared. Note that the means of the squared ages are different from the squares of the mean ages. Again let us visualize this relationship in Figure 6.6.

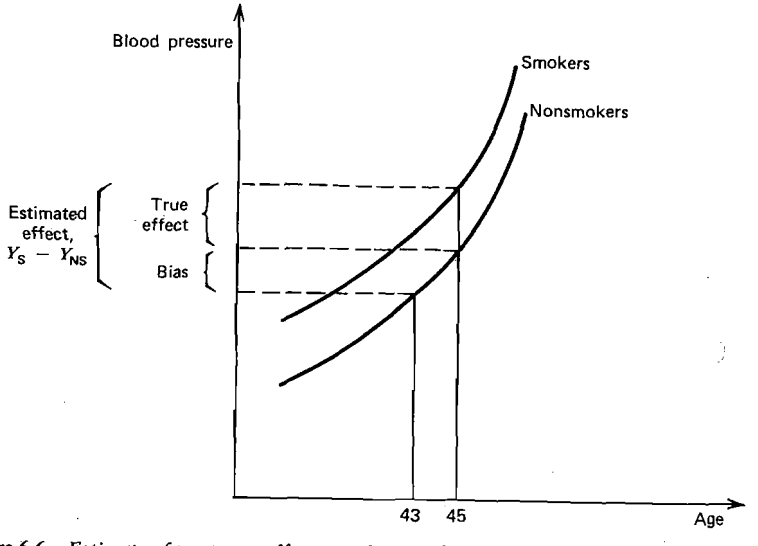

Figure 6.6 *Estimate of treatment effect—nonlinear relationship.*

The individual estimate of the treatment effect determined from the matched pair of a 45-year-old smoker and a 43-year-old nonsmoker is shown in Figure 6.6 by the large brace to the left. This estimate can be compared to the true treatment effect shown by the topmost smaller brace. The bias is the difference between the two and is indicated by the second small brace. From (6.5) we obtain the bias as

$$\beta(45^2 - 43^2) = \beta(176).$$

If we were using the matched groups from Example 6.1, upon averaging over the two groups we would find that the estimate of the treatment effect would

be biased by the amount

$$\beta\,(\bar{X}_S^2 - \bar{X}_{NS}^2) = \beta(69.6).$$

It is important to realize that equality of the means of the two groups is not enough to ensure an unbiased estimate of the treatment effect if the relationship between the response and the confounding variable is nonlinear. Equality of the means yields unbiased estimates only in the linear case.

6.4.3 Evaluation of Bias Reduction

So far we have only demonstrated how caliper matching can reduce the bias due to confounding. In this section we present theoretical results concerning the bias reduction one can expect using caliper matching in the linear case. The estimator of the treatment effect is the mean difference in response. The effectiveness of caliper matching and all other matching techniques is examined relative to estimating the treatment effect from random samples, where the confounding variable is not taken into account. (For a definition of the measure of effectiveness, the expected percent reduction in bias, see Cochran and Rubin, 1973.)

Table 6.2 gives an indication of the expected percent bias reduction for different tolerance values. The results are independent of the sample size and reservoir size. They were derived assuming that the initial difference between the two populations is less than 0.5 (i.e., $B_X < 0.5$), that the distributions of the confounding variable are normal, and that the outcome is linearly related to the confounding variable. Notice that the tolerance is specified in terms of a proportion, a, of a standard deviation. It appears that tight caliper matching (i.e., $a = 0.2$) can be expected to remove nearly all the bias in the treatment effect relative to random sampling. It also appears that the ratio of the variances (i.e., σ_1^2/σ_0^2) has a negligible effect on the percent reduction in bias.

Table 6.2 Percent Bias Reduction for Caliper Matching*

a	$\sigma_1^2/\sigma_0^2 = 1/2$	$\sigma_1^2/\sigma_0^2 = 1$	$\sigma_1^2/\sigma_0^2 = 2$
0.2	99	99	98
0.4	96	95	93
0.6	91	89	86
0.8	86	82	77
1.0	79	74	69

Reprinted, by permission of the Statistical Publishing Society, from Cochran and Rubin (1973), Table 2.3.1.

* Tolerance $\epsilon = a\sqrt{(\sigma_1^2 + \sigma_0^2)/2}$.

One can use this table to get some indication of bias reduction to be expected for different tolerances if the values or estimates of the population variances are known and if $B_X < 0.5$. Suppose we knew that $\sigma_1^2/\sigma_0^2 = 1/2$, where $\sigma_1^2 = 4$; then if we took $a = 0.8$, we could expect about 86% of the bias to be removed. The tolerance would be $0.8\sqrt{(4+8)/2} = 1.96$. If we used $a = 0.4$, we could expect to remove 96% of the bias over random sampling and the tolerance would be 0.98.

As we have mentioned previously, the major disadvantage of caliper matching is the need for the comparison reservoir to be large. In their theoretical work, Cochran and Rubin did not take into account the possibility that the desired number of matches would not be found from the comparison reservoir, although the probability of this occurrence is nonnegligible. Nor are the results known for distributions other than normal. Most likely, the results presented are applicable to symmetric distributions, but the case of skew distributions has not been investigated for caliper matching.

6.5 NEAREST AVAILABLE MATCHING

In some situations when caliper matching is performed with a small tolerance, there is a nonnegligible probability that some individuals cannot be matched. To avoid this problem, Rubin (1973a, b) developed a method known as nearest available pair matching. We shall refer to this matching procedure as *nearest available matching*. This method ensures that the desired number of matches are obtained by being less restrictive in deciding what a match is. A match is formed by finding the closest possible comparison subject for each individual in the treatment group from the yet-unmatched individuals in the comparison reservoir. Since nearest available matching does not use a fixed tolerance as does caliper matching, the reservoir does not have to be larger than the treatment group. However, the matches are not guaranteed to be as close as those found under caliper matching.

6.5.1 Methodology

There are three variants of nearest available matching, each based on a particular ordering of the subjects in the treatment group with respect to the confounding variable. The specification of the ordering completely defines the pair matching method. In one variant of the method, referred to as random-order nearest available matching, the N treatment subjects are randomly ordered on the values of the confounding variable, X. Let us denote these ordered values by X_{11} to X_{1N}. Starting with X_{11}, a match is defined as that subject from the comparison reservoir whose value X_{0j} is nearest X_{11}. The matches are therefore assigned to minimize $|X_{11} - X_{0j}|$ for all subjects in the comparison reservoir.

If there are ties (i.e., two or more comparison subjects for whom $|X_{11} - X_{0j}|$ is a minimum), the match is formed randomly. The nearest available partner for the next treatment subject with value X_{12} is then found from the remaining subjects in the reservoir. The matching procedure continues in this fashion until matches have been found for all N treatment subjects.

The other two variants of nearest available matching result from ranking the members of the treatment group on confounding variable values from the highest to the lowest (HL) value or from the lowest to the highest (LH) value. Matches are then sought starting with the first ranked treatment subject, as for the random-order version.

Example 6.2 Nearest available matching: Suppose that in a blood pressure study, there are three smokers with ages 40, 45, and 50, and five nonsmokers in the reservoir with ages 30, 32, 46, 49, and 55. In addition, suppose that the randomized order of the smokers' ages is 40, 50, and 45. Then the random-order nearest available matching technique will match the 40-year-old smoker with the 46-year-old nonsmoker, the 50-year-old smoker with the 49-year-old nonsmoker, and the 45-year-old smoker with the 55-year-old nonsmoker.

In the case of the other two variants, the following matches would be made: for HL, the 50-year-old with the 49-year-old, the 45-year-old with the 46-year-old, and the 40-year-old with the 32-year-old; for LH, the 40-year-old with the 46-year-old, the 45-year-old with the 49-year-old, and the 50-year-old with the 55-year-old. Notice in this example that each variant resulted in different matched pairs.

6.5.2 Appropriate Conditions

Nearest available matching is similar to caliper matching except that there is no fixed tolerance. Based on the prior discussion of caliper matching and some theoretical results, it follows that nearest available matching is effective in removing bias due to confounding if the relationship between the response and confounding variables is linear. For nonlinear relationships, no results are available.

The main difficulty in discussing what conditions are most appropriate for using nearest available matching is the fact that the reduction in bias is so strongly influenced by the closeness of the distributions of the treatment group and the comparison reservoir. If there is a large overlap between the two groups of subjects, nearest available matching will be very similar to caliper matching with a suitably large tolerance. If, however, there is a moderate to small amount of overlap, the desired number of matches will be found, but the final amount of bias in the estimate of the treatment effect may be large.

6.5.3 Evaluation of Bias Reduction

In selecting a particular nearest available matching procedure, an investigator may want to base his or her choice primarily on the percent reduction in bias

obtainable. Assuming a linear relationship between the response and confounding variables, Cochran and Rubin (1973) performed a simulation study to determine which of the three nearest available matching estimators was least biased. They also assumed that the confounding variable was normally distributed with the mean of the treatment population greater than the mean in the comparison population ($\eta_1 > \eta_0$). Their results showed that the percent reduction in bias was largest for the low–high nearest available matching and smallest for the high–low variant.

Because nearest available matching does not guarantee as close matches as are possible with caliper matching, Rubin (1973a) also compared the closeness of the matches obtained by the three procedures as measured by the average of the squared error $(X_1 - X_0)^2$ within pairs. When the procedures were judged by this criterion, the order of performance was reversed. The HL nearest available matching had the lowest average squared error and the LH had the largest. This result is not too surprising, considering the relationship between the population means ($\eta_1 > \eta_0$). The HL procedure would start with the treatment subject who is likely to be the most difficult to match: namely, the one with the largest value of X. This would tend to minimize the squared within-pair difference.

Since the differences between the three matching procedures are small on both criteria, random-order nearest available matching appears to be a reasonable compromise. In Table 6.3, from Cochran and Rubin (1973), results of the percent reduction in bias are summarized for random-order nearest available matching as a function of the initial difference, the values of the ratio of the population variances, and sizes of the reservoir. Results for the number of matches $N = 25$ and $N = 100$ (not shown) differ only slightly from those for $N = 50$.

Table 6.3 Percent Bias Reduction for Random-Order Nearest Available Matching: X Normal; N = 50*

B_X / r	$\sigma_1^2/\sigma_0^2 = 1/2$			$\sigma_1^2/\sigma_0^2 = 1$			$\sigma_1^2/\sigma_0^2 = 2$		
	1/4	1/2	1	1/4	1/2	1	1/4	1/2	1
2	99	98	84	92	87	69	66	59	51
3	100	99	97	96	95	84	79	75	63
4	100	100	99	98	97	89	86	81	71

Reprinted, by permission of the Statistical Publishing Society, from Cochran and Rubin (1973), Table 2.4.1.

* X = confounding variable; B_X = initial difference; r = ratio of the size of the comparison reservoir and the treatment group; σ_1^2 = variance of confounding variable in the treatment population; σ_0^2 = variance of confounding variable in the comparison population.

With this method, the percent reduction in bias decreases steadily as the initial difference between the normal distributions of the confounding variable increases from 1/4 to 1. In contrast with results reported in Table 6.2 for caliper matching, the percent reduction in bias does depend on the ratio of the population variances. Based on Table 6.3, random-order nearest available matching does best when $\sigma_1^2/\sigma_0^2 = 1/2$. When $\eta_1 > \eta_0$ and $\sigma_0^2 > \sigma_1^2$, large values of the confounding variable in the treatment group, the ones most likely to cause bias, will receive closer partners out of the comparison reservoir than if $\sigma_0^2 < \sigma_1^2$.

Investigators planning to use random-order nearest available matching can use Table 6.3 to obtain an estimate of the expected percent bias reduction. Suppose an estimate of the initial difference B_X is 1/2, with $\sigma_1^2/\sigma_0^2 = 1$, and it is known that the reservoir size is 3 times larger than the treatment group ($r = 3$). It follows that random-order nearest available matching results in an expected 95% reduction in bias.

6.6 STRATIFIED MATCHING

Stratified matching is an appropriate pair matching procedure for categorical confounding variables. If, like sex or religious preference, the variable is truly categorical, with no underlying numerical distribution, the matches are exact and no bias will result. Often, however, the confounding variable is numerical but the investigator may choose to work with the variable in its categorical form. Suppose, for example, that in the study of smoking and blood pressure, all the subjects were employed and that job anxiety is an important confounding variable. The investigator has measured job anxiety by a set of 20 true–false questions so that each subject can have a score from 0 to 20. Such a factor is very difficult to measure, however, and the investigator may decide that it is more realistic and more easily interpretable to simply stratify the range of scores into low anxiety, moderate anxiety, and high anxiety. Having formed these three strata, the investigator can now randomly form individual pair matches within each stratum. An example of this procedure in the case of multiple confounding variables is given in Section 6.11.

The only theoretical paper discussing the bias reduction properties of stratified matching is that of McKinlay (1975). She compared stratified matching to various stratification estimators (Section 7.6) for a numerical confounding variable converted to a categorical variable. She considered various numbers of categories and a dichotomous outcome. She found that the estimator of the odds ratio from stratified matched samples had a larger mean squared error and, in some of the cases considered, a larger bias than did the crude estimator, which ignores the confounding variable. (Stratified matching is compared with

stratification in Section 13.2.2.) The mean squared error results are due in part to the loss of precision caused by an inability to find matches for all the treatment subjects. This point is considered further in Section 13.2.

6.7 FREQUENCY MATCHING

Frequency matching involves stratifying the distribution of the confounding variable in the treatment group and then finding comparison subjects so that the number of treatment and comparison subjects is the same within each stratum. This is not a pair matching method, and the number of subjects may differ across strata.

For the sake of illustration we shall concentrate on the case of a numerical response. This will allow us to demonstrate more easily how frequency matching helps to reduce the bias. Because frequency matching is equivalent to stratification with equal numbers of comparison and treatment subjects within each stratum, we leave the discussion of the various choices of estimators in the case of a dichotomous response to Chapter 7.

6.7.1 Methodology

Frequency matching is most useful when one does not want to deal with pair matching on a numerical confounding variable or an ordinal measure of an underlying numerical confounding variable. An example of the latter situation is initial health care status, where the categories reflect an underlying continuum of possible statuses. In either case, the underlying distribution must be stratified. Samples are then drawn either randomly or by stratified sampling from the comparison reservoir in such a way that there is an equal number of treatment and comparison subjects within each stratum. Criteria for choosing the strata are discussed in Section 6.7.3 after we have presented the estimator of the treatment effect.

Example 6.3 ***Frequency matching:*** Let us consider the use of frequency matching in the smoking and blood pressure study. Suppose that the age distribution of the smokers was stratified into 10-year intervals as shown on the first line of Table 6.4, and that 100 smokers were distributed across the strata as shown on the second line of the table. The third line of the table represents the results of a random sample of 100 nonsmokers from the comparison reservoir. Notice that since frequency matching requires the sample sizes to be equal within each stratum, the investigator needs to draw more nonsmokers in all strata except for ages 51 to 60 and 71 to 80. In these two strata the additional number of nonsmokers would be dropped from the study on a random basis. (Note that stratified sampling, if possible, would have avoided the problem of too few or too many persons in a stratum.)

Table 6.4 Smokers and Nonsmokers Stratified by Age

Age	11–20	21–30	31–40	41–50	51–60	61–70	71–80	Total
Smokers	1	3	10	21	30	25	10	100
Nonsmokers	0	2	8	20	32	20	18	100

6.7.2 Appropriate Conditions

Frequency matching is relatively effective in reducing bias in the parallel linear response situation provided that enough strata are used. We shall explain this by means of simple formulas for the estimator of the treatment effect assuming a numerical response.

Recall from Section 6.1 that we can represent the linear relationship between the response Y and the confounding variable X by

$$\begin{aligned} Y_1 &= \alpha_1 + \beta X_1 \qquad \text{in the treatment group} \\ Y_0 &= \alpha_0 + \beta X_0 \qquad \text{in the comparison group.} \end{aligned} \tag{6.6}$$

In general, the estimator of the treatment effect in the kth stratum is

$$\bar{Y}_{1k} - \bar{Y}_{0k} = (\alpha_1 - \alpha_0) + \beta(\bar{X}_{1k} - \bar{X}_{0k}), \tag{6.7}$$

where a bar above the variables indicates the mean calculated for the kth stratum. The bias in the kth stratum is $\beta(\bar{X}_{1k} - \bar{X}_{0k})$.

Clearly, the maximum amount of distortion in the estimate from the kth stratum occurs when $\bar{X}_{1k} - \bar{X}_{0k}$ is maximized. The maximum value is then β times the width of the kth stratum.

One overall estimate of the treatment effect is the weighted combination of the individual strata differences in the response means:

$$\bar{Y}_1 - \bar{Y}_0 = \frac{1}{N}\sum_{k=1}^{K} n_k(\bar{Y}_{1k} - \bar{Y}_{0k}), \tag{6.8}$$

where n_k is the number of treatment or comparison subjects in the kth stratum ($k = 1, 2, \ldots, K$) and N is the total number of treatment subjects. Rewriting (6.8) in terms of treatment effect and regression coefficients, we obtain, using (6.7),

$$\begin{aligned} \bar{Y}_1 - \bar{Y}_0 &= \frac{1}{N}\sum_{k=1}^{K} n_k[\alpha_1 - \alpha_0 + \beta(\bar{X}_{1k} - \bar{X}_{0k})] \\ &= (\alpha_1 - \alpha_0) + \frac{1}{N}\sum_{k=1}^{K} n_k\beta(\bar{X}_{1k} - \bar{X}_{0k}). \end{aligned} \tag{6.9}$$

From (6.9) we see that the amount of bias reduction possible using frequency

matching is determined by the difference in the distributions of the two groups within each stratum. This, in turn, is a function of the manner in which the strata were determined. The more similar the distributions of the treatment and comparison populations are within each stratum, the less biased the individual estimates of the treatment effect will be.

6.7.3 Evaluation of Bias Reduction

Assuming that both distributions of the confounding variable are normal with equal variances but the mean of the treatment population is zero and the mean of the comparison population is small but nonzero, Cox (1957) derived the percent reduction in bias for strata with equal number of subjects. Cochran and Rubin (1973) extended Table 1 of Cox, and these results are given in Table 6.5. The strata are based on the distribution of the treatment group.

Table 6.5 Percent Bias Reduction with Equal-Sized Strata in Treatment Population: X Normal

Number of strata:	2	3	4	5	6	8	10
% reduction in bias:	64	79	86	90	92	94	96

Reprinted, by permission of the Statistical Publishing Society, from Cochran and Rubin (1973), Table 4.2.1.

These percentages are at most 2% lower than the maximum amount of bias reduction possible using strata with an unequal number of subjects. Cochran (1968) extended these calculations for some nonnormal distributions: the chi-square, the t, and the beta, and he concluded that the results given in Table 6.5 can be used as a guide to the best boundary choices even when the confounding variable is not normally distributed.

From this information we can conclude that if the distributions of the confounding variable are approximately normal and differ only slightly in terms of the mean, and based on the distribution of the treatment group we form four strata with equal numbers of subjects, we can expect to reduce the amount of bias in the estimate of the treatment effect by 86%.

We stated at the beginning of Section 6.7.2 that frequency matching was relatively effective in reducing the bias in the linear parallel situation. No theoretical work has been done for the nonlinear parallel situation. Frequency matching does, however, have the advantage of allowing one to use the analysis of variance to test for interactions. One can test for parallelism as well as linearity, thus determining whether frequency matching was appropriate.

6.8 MEAN MATCHING

A simple way of attempting to equate the distributions of the confounding variable in the study samples is to equate their means. This is called *mean matching* or *balancing*. The members of the comparison group are selected so that $|\bar{X}_1 - \bar{X}_0|$ is as small as possible. Although mean matching is very simple to employ, it depends strongly on the assumption of a linear parallel response relationship and we therefore do not recommend its use. One can employ analysis of covariance (Chapter 8) in this case and achieve greater efficiency. We include the following discussion of mean matching so that the reader can understand the basis for our recommendation.

6.8.1 Methodology

There is more than one way to form matches in mean matching. However, the only algorithm which is guaranteed to find the comparison group that minimizes $|\bar{X}_1 - \bar{X}_0|$ is to calculate $\bar{X}_0$ for all possible groups of size N from the comparison reservoir. This is generally far too time-consuming. An easier algorithm uses partial means, and we shall demonstrate its use with the following example.

Example 6.4 Mean matching: Suppose that we decided to use mean matching on age in the blood pressure study, where we have three smokers, aged 40, 45, and 50 years. First, we would calculate the mean age of the smokers, which is 45 years ($\bar{X}_S = 45$). Next, we would select successive subjects from the nonsmokers such that the means of the nonsmokers ages, calculated after the selection of each subject (partial means), are as close as possible to 45. Suppose that the nonsmokers in the comparison reservoir have the following ages: 32, 35, 40, 41, 45, 47, and 55 years. The first nonsmoker selected as a match would be age 45; the second subject selected would be 47 years old, since the partial mean, $(45 + 47)/2 = 46$, is closest to 45. The last nonsmoker to be selected would be 41 years of age, again since the partial mean, $(2/3)(46) + (1/3)(41) = 44.3$, is closest to $\bar{X}_S$. Note that this algorithm did not minimize $|\bar{X}_S - \bar{X}_{NS}|$, since choosing the nonsmokers aged 35, 45, and 55 would give equality of the two sample mean ages $[(35 + 45 + 55)/3 = 45]$.

6.8.2 Appropriate Conditions

Mean matching can be very effective in reducing bias in the case of a parallel linear response relationship. Suppose in the blood pressure example that the population means η_S and η_{NS} for smokers and nonsmokers were 50 and 45, respectively. Then, for large enough random samples, we might expect to find that $\bar{X}_{NS} = 45$ and $\bar{X}_S = 50$.

From (6.3) it follows that the estimated treatment effect is biased by an

amount equal to $\beta(\bar{X}_S - \bar{X}_{NS}) = 5\beta$. However, if mean matching had been used to reduce $|\bar{X}_S - \bar{X}_{NS}|$ to, say, 0.7, as in Example 6.4, then the bias in $(\bar{Y}_S - \bar{Y}_{NS})$ would have been reduced by 86% (= 4.3/5.0). (The initial difference in the means due to random sampling is 5.0.)

Mean matching is *not* effective in removing bias in the case of a parallel nonlinear response relationship (see Figure 6.7). Assume that in another blood pressure study three smokers of ages 30, 35, and 40 years were mean-matched with three nonsmokers of ages 34, 35, and 36 years, respectively. Their blood pressures are denoted by × in Figure 6.7. Notice that unlike the previous linear situations, $\bar{Y}_S$ and $\bar{Y}_{NS}$ do not correspond to the mean ages $\bar{X}_S$ and $\bar{X}_{NS}$. They will both be greater than the values of Y which correspond to the means due to the nonlinearity. Here $(\bar{Y}_S - \bar{Y}_{NS})$ is an overestimate of the treatment effect. The estimate should be equal to the length of the vertical line, which represents the treatment effect. In general, the greater the nonlinearity, the greater the overestimation or bias will be, in general.

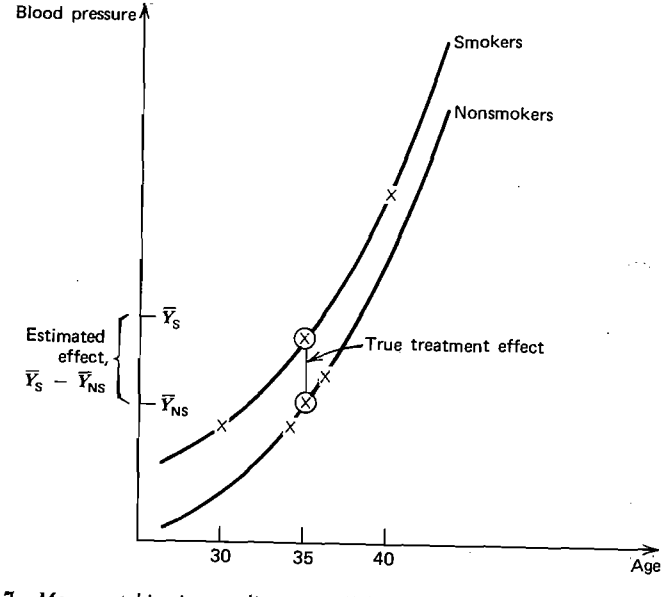

Figure 6.7 *Mean matching in a nonlinear parallel relationship.*
×, *blood pressure for a specific age;* ⊗, *blood pressure corresponding to mean age in either group.*

6.8.3 Evaluation of Bias Reduction

Cochran and Rubin (1973) have investigated the percentage of bias reduction possible using the partial mean algorithm presented in Section 6.8.1 under the assumptions of a linear parallel relationship, a normally distributed confounding variable, and a sample size of 50 in the treatment group. They found that, except in the cases where the initial difference $B_X = 1$, mean matching removes essentially all the bias. In addition, its effectiveness increases with the size of the comparison reservoir. The bias that results from improper use of mean matching (i.e., in nonlinear cases) has not been quantified.

6.9 ESTIMATION AND TESTS OF SIGNIFICANCE

In this section we indicate the appropriate tests of significance and estimators of the treatment effect for each matching technique. Because the choice of test and estimator depends on the form of the outcome variable, we begin with the numerical case followed by the dichotomous case. Also, in keeping with the general intent of this book, we do not give many details on the test statistics but rather cite references in which further discussion may be found. The tests and estimators for frequency-matched samples are the same as for stratification and are discussed in greater detail in Chapter 7.

In the case of a numerical outcome variable for which one of the pair matching methods (caliper, nearest available, or stratified) has been used, the correct test of significance for the null hypothesis of no treatment effect is the paired-t test (see Snedecor and Cochran, 1967, Chap. 4). This test statistic is the ratio of the mean difference, which is the estimate of the treatment effect, to its standard error. The difference between the paired-t test and the usual t test for independent (nonpaired) samples is in the calculation of the standard error.

If in the case of a numerical outcome variable, frequency matching has been used, the standard t test is appropriate, with the standard error determined by an analysis of variance. (See Snedecor and Cochran, 1967, Chap. 10, for a discussion of the analysis of variance.) The treatment effect is estimated by the mean difference. If, however, the within-stratum variances are not thought to be equal, then, as in the case of stratification, one should weight inversely to the variance (see Section 7.7 and Kalton, 1968). In the case of mean matching, the correct test is again the t test. The standard error, however, must be calculated from an analysis of covariance (see Greenberg, 1953).

When the outcome variable is dichotomous, as discussed in Chapter 3, the treatment effect may be measured by the difference in proportions, the relative risk, or the odds ratio. The estimator of the difference in rates is the difference between the sample proportions, $p_1 - p_0$. This is an unbiased estimator if the

matching is exact. For estimating the odds ratio, the stratification estimators appropriate for large numbers of strata are applicable (see Section 7.6.1), with each pair comprising a stratum. In this case the conditional maximum likelihood estimator is easy to calculate and is identical to the Mantel–Haenszel (1959) estimator. For each pair (stratum), a 2×2 table can be created. For the jth pair, we have four possible outcomes:

		Control Subject 1	Control Subject 0	
Treatment Subject	1	a_j	b_j	
	0	c_j	d_j	
				1

For example, $b_j = 1$ if, in the jth pair, the outcome for the control subject is 0 and for the treatment subject it is 1. The estimator of the odds ratio, ψ, is then $\hat{\psi} = \Sigma_j b_j / \Sigma_j c_j$. The estimator will be approximately unbiased if the matching is exact and the number of pairs is large.

Because of the relationship between these measures of the treatment effect (difference of proportions, relative risk, and odds ratio) under the null hypothesis of no treatment effect (Section 3.1), McNemar's test can be used in the case of pair-matched samples, regardless of the estimator (see Fleiss, 1973, Chap. 8). Similarly, when frequency matching is used, we have a choice of tests, such as Mantel–Haenszel's or Cochran's test, regardless of the estimator (see Fleiss, 1973, Chap. 10). Since the analysis of a frequency-matched sample is the same as an analysis by stratification, the reader is referred to Chapter 7 for a more detailed discussion.

6.10 MULTIVARIATE MATCHING

So far we have limited the discussion of matching to a single confounding variable. More commonly, however, one must control simultaneously for many confounding variables. To date, all research has been on multivariate pair matching methods. To be useful, a multivariate matching procedure should create close individual matches on all variables. In addition, ideally, as in the univariate case, the procedure should not result in the loss of many subjects because of a lack of suitable matches. The advantage of constructing close individual matches, as in the univariate case, is that with perfectly matched pairs the matching variables are perfectly controlled irrespective of the underlying model relating the outcome to the risk and confounding variables.

Discussions of multivariate matching methods in the literature are quite limited. References include Althauser and Rubin (1970), for a discussion of an applied problem; Cochran and Rubin (1973), for a more theoretical framework; Rubin (1976a, b), for a discussion of certain matching methods that are equal percent bias reducing (EPBR); Carpenter (1977), for a discussion of a modification of the Althauser–Rubin approach; and Rubin (1979), for a Monte Carlo study comparing several multivariate methods used alone or in combination with regression adjustment.

In the following sections we first discuss straightforward generalizations of univariate caliper and stratified matching methods to the case of multiple confounding variables. The methods included are multivariate caliper matching, and multivariate stratified matching. Then we discuss metric matching methods wherein the objective is to minimize the distance between the confounding variable measurements in the comparison and treatment samples. Several alternative distance definitions will be presented.

Next we discuss discriminant matching. This matching method reduces the multiple confounding variables to a single confounding variable by means of the linear discriminant function. Any univariate matching procedure can then be applied to the linear discriminant function.

In trying to rank the multivariate matching techniques according to their ability to reduce the bias, one is faced with the problem of how to combine the reduction in bias due to each confounding variable into a single measure so that the various methods can be compared. For example, the effectiveness of caliper matching depends, in part, on the magnitudes of all the tolerances that must be chosen.

To partially circumvent this problem of constructing a single measure of bias reduction, Rubin (1976a, b; 1979) introduced the notion of matching methods of the equal percent bias reducing (EPBR) type. For the linear case, Rubin showed that the percent bias reduction of a multivariate matching technique is related to the reduction in the differences of the means of each confounding variable, and that if the percent reduction is the same for each variable, that percentage is the percent reduction for the matching method as a whole. EPBR matching methods are techniques used to obtain equal percent reduction on each variable and, hence, guarantee a reduction in bias.

Discriminant matching and certain types of metric matching have the EPBR property, so that we can indicate which of these EPBR methods can be expected to perform best in reducing the treatment bias in the case of a linear response surface.

6.10.1 Multivariate Caliper Matching

Multivariate caliper matching, like its univariate counterpart, is effective in

reducing bias provided that the tolerances used for each confounding variable are small and the comparison reservoir is large, generally much larger than in the univariate case.

Suppose that there are L confounding variables. A comparison subject is considered to be a match for a treatment subject when the difference between their measured lth confounding variable ($l = 1, 2, \ldots, L$) is less than some specified tolerance, ϵ_l (i.e., $|X_{1l} - X_{0l}| \leq \epsilon_l$) for all l.

Example 6.5 Multivariate caliper matching: Consider a hypothetical study comparing two therapies effective in reducing blood pressure, where the investigators want to match on three variables: previously measured diastolic blood pressure, age, and sex. Such confounding variables can be divided into two types: categorical variables, such as sex, for which the investigators may insist on a perfect match ($\epsilon = 0$); and numerical variables, such as age and blood pressure, which require a specific value of the caliper tolerances. Let the blood pressure tolerance be specified as 5 mm Hg and the age tolerance as 5 years. Table 6.6 contains measurements of these three confounding variables. (The subjects are grouped by sex to make it easier to follow the example.)

Table 6.6 Hypothetical Measurements on Confounding Variables for Example 6.6

Treatment Group				Comparison Reservoir			
Subject Number	Diastolic Blood Pressure (mm Hg)	Age	Sex	Subject Number	Diastolic Blood Pressure (mm Hg)	Age	Sex
1	94	39	F	1	80	35	F
2	108	56	F	2	120	37	F
3	100	50	F	3	85	50	F
4	92	42	F	4	90	41	F
5	65	45	M	5	90	47	F
6	90	37	M	6	90	56	F
				7	108	53	F
				8	94	46	F
				9	78	32	F
				10	105	50	F
				11	88	43	F
				12	100	42	M
				13	110	56	M
				14	100	46	M
				15	100	54	M
				16	110	48	M
				17	85	60	M
				18	90	35	M
				19	70	50	M
				20	90	49	M

In this example there are 6 subjects in the treatment group and 20 subjects in the comparison reservoir. Given the specified caliper tolerances, the first subject in the treatment group is matched with the fourth subject in the comparison reservoir. The difference between their blood pressures is 4 units, their ages differ by 2 years, and both are females. We match the second treatment subject with the seventh comparison subject since their blood pressures and sex agree exactly and their ages differ by only 3 years. The remaining four treatment subjects, subjects 3, 4, 5, and 6, would be matched with comparison subjects 10, 8, 19, and 18, respectively. Notice that if the nineteenth comparison subject were not in the reservoir, the investigator would have to either relax the tolerance on blood pressure, say to 10 mm Hg, or discard the fifth treatment subject from the study.

Expected Bias Reduction. Table 6.2 gives the expected percent of bias reduction for different tolerances assuming a single, normally distributed confounding variable and a linear and parallel response relationship. Table 6.2 can also be used in the case of multiple confounding variables if these variables or some transformation of them are normally and independently distributed, and if the relationship between the outcome and confounding variables is linear and parallel. The expected percent of bias reduction is then a weighted average of the percent associated with each variable.

If the investigators know (a) the form of the linear relationship, (b) the population parameters of the distribution of each of the confounding variables, and (c) that the confounding variables or some transformation of them are independent and normally distributed, then the best set of tolerances in terms of largest expected treatment bias reduction in Y could theoretically be determined by evaluating equation (5.1.5) in Cochran and Rubin (1973) for several combinations of tolerances. In practice, this would be very difficult to do.

6.10.2 Multivariate Stratified Matching

The extension of univariate stratified matching to the case of multiple confounding variables is straightforward. Subclasses are formed for each confounding variable, and each member of the treatment group is matched with a comparison subject whose values lie in the same subclass on all confounding variables.

Example 6.6 Multivariate stratified matching: Consider again the blood pressure data presented in Table 6.6. Suppose that the numerical confounding variable, diastolic blood pressure, is categorized as ≤ 80, 81–94, 95–104, and ≥ 105, and age as 30–40, 41–50, and 51–60. Including the dichotomous variable, sex, there are in total ($4 \times 3 \times 2 =$) 24 possible subclasses into which a subject may be classified. In Table 6.7 we enumerate the 12 possible subclasses for males and females separately. Within each cell we have listed the subject numbers and indicated by the subscript t those belonging to the treatment group.

Table 6.7 Stratification of Subjects on Confounding Variables in Example 6.6[a]

Diastolic Blood Pressure	Age 30–40	Age 41–50	Age 51–60
		Males	
-80		5_t, 19	
81–94	6_t, 18	20	17
95–104		12, 14	15
105-		16	13
		Females	
-80	1, 9		
81–94	1_t	4_t, 3, 4, 5, 8, 11	6
95–104		3_t	
105-	2	10	2_t, 7

[a] Within each cell the subject number from Table 6.6 is given. Those with a subscript t are the treatment group subjects.

With this stratification, the second treatment subject is matched with the seventh comparison subject. The fifth treatment subject would be matched with the nineteenth comparison subject and the fourth treatment subject would be randomly matched with one of comparison subjects 3, 4, 5, 8, or 11. The last treatment subject would be matched with the eighteenth comparison subject. Subjects 1 and 3 in the treatment group do not have any matches in the comparison reservoir and must therefore be omitted from the study, or else the subclass boundaries must be modified.

It should be clear from this simple example that as the number of confounding variables increases, so does the number of possible subclasses, and hence the larger the comparison reservoir must be in order to find an adequate number of matches.

The expected number of matches for a given number of subclasses and given reservoir size r have been examined by McKinlay (1974) and Table 6.8 presents a summary of her results. The number of categories in Table 6.8 equals the product of the number of subclasses for each of the L confounding variables. In McKinlay's terminology we had 24 categories in Example 6.6. Her results are based on equal as well as markedly different joint distributions of the L confounding variables in the treatment and comparison populations (see McKinlay, 1974, Table 1, for the specific distributions). For example, in a study with 20 subjects in the treatment group and 20 in the comparison reservoir, stratified matching on 10 categories where the confounding variable distributions in the two populations are exactly the same will result in about 66 percent of the treatment group being matched (i.e., only 13 suitable comparison subjects would be expected to be found). Clearly, large reservoirs are required if multivariate

Table 6.8 Expected Percentages of Matches in Multivariate Stratified Matching

N_t, Size of Treatment Group	r	Same Distribution 10 Categories	Same Distribution 20 Categories	Different Distribution 10 Categories	Different Distribution 20 Categories
20	1	66.0	53.0	55.0	43.5
	2.5	94.0	84.5	84.5	72.5
	5	98.5	96.0	96.5	89.0
	10	100.0	99.0	99.5	96.5
50	1	78.0	68.6	62.4	55.2
	2	97.0	91.6	86.6	78.0
	4	99.8	98.8	98.0	92.8
	10	100.0	100.0	100.0	99.0
100	1	84.3	77.3	65.3	60.5
	2	99.1	96.8	90.3	83.7
	5	100.0	99.9	99.8	97.2

Adapted, by permission of the Royal Statistical Society, from McKinlay (1974), Tables 2 and 3.

stratified matching is to be used effectively. With 20 treatment subjects one would need more than 100 comparison subjects for matching with only negligible loss of treatment subjects.

No information is available on the bias reduction one can expect for a given reservoir size, r, and given population parameters of the joint distribution of the L confounding variables in the treatment and comparison populations.

6.10.3 Minimum Distance Matching.

Both multivariate caliper matching and stratified matching are straightforward extensions of univariate techniques in that a matching restriction exists for each variable. In this section we discuss *minimum distance matching* techniques that take all of the confounding variables into account at one time, thus reducing multiple matching restrictions to one. For two subjects to be a match, their confounding variable values must be close as defined by some distance measure. The matching can be done with a "fixed" tolerance, as in univariate caliper matching, or as nearest available matching. We begin with the fixed tolerance case. Because distance is defined by a distance function or metric, these techniques are also referred to as *metric matching*.

One distance function is Euclidean distance which is defined as

$$\sum_{l=1}^{L} (X_{1l} - X_{0l})^2, \tag{6.10}$$

where X_{il} is the value of the lth confounding variable for a subject in the treatment ($i = 1$) or the comparison ($i = 0$) group. A major problem with the use of Euclidean distance is that the measure (6.10) and hence choice of matched subjects strongly depend on the scale used for measuring the confounding variables. For example, measuring a variable in centimeters rather than in meters would increase that variable's contribution to the Euclidean distance 10,000-fold.

A common technique for eliminating this problem of choice of scale is to convert all variables to *standardized scores*. A standardized score (Z) is the observed value of a confounding variable (X), divided by that confounding variable's standard deviation (s): $Z = X/s$. Equation (6.10) then would become

$$\sum_{l=1}^{L} (Z_{1l} - Z_{0l})^2, \tag{6.11}$$

where Z_{il} is the standardized score of the lth confounding variable ($l = 1, \ldots, L$) for a subject in the treatment ($i = 1$) or the comparison ($i = 0$) group. Use of (6.11) as a matching criterion has been termed *circular matching* (Carpenter, 1977).

To better understand circular matching and its relation to multivariate caliper matching, consider the case of two confounding variables shown in Figure 6.8. Suppose that the two confounding variables have been transformed to standardized scores. Point A is a treatment subject with standardized scores of a_1 for the first confounding variable and a_2 for the second. If we were to use multivariate caliper matching with a common tolerance ϵ, we would search for a comparison subject with a standardized score of the first confounding variable in the interval $[a_1 - \epsilon, a_1 + \epsilon]$, and at the same time, a value of the second standardized score in the interval $[a_2 - \epsilon, a_2 + \epsilon]$. Thus the search is for a comparison subject like subject B, with confounding variable values within the *square* shown in Figure 6.8.

In circular matching with tolerance ϵ, the search is for a comparison subject whose confounding variable values satisfy

$$(Z_{11} - Z_{01})^2 + (Z_{12} - Z_{02})^2 \leq \epsilon,$$

that is, for values in the *circle* of radius ϵ centered at A. In Figure 6.8, subject B would not be a match for subject A if circular matching with tolerance ϵ were used.

There have been several suggestions for calculating the standard deviation to be used in the standardized scores. Cochran and Rubin (1973) suggest using

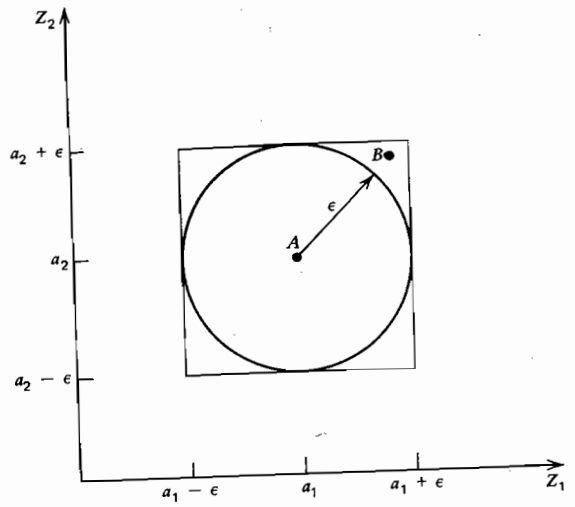

Figure 6.8 *Caliper matching on standardized scores.* ϵ = *tolerance.*

the standard deviations calculated from the comparison group only, while Smith et al. (1977) suggest using only the treatment group standard deviations. The advantage of the latter suggestion is that the standard deviations may be calculated before identifying the comparison subjects. Finally, a pooled estimate of the standard deviation can be used if one believes that the variances of the two groups are similar. All three suggestions suffer from the restriction that the measurements for calculating the standard deviations must be available prior to any matching.

Equation (6.11) can be rewritten as

$$\sum_{l=1}^{L} (X_{1l} - X_{0l})^2 / s_l^2.$$

As can be seen from this, circular matching only takes the variances of the confounding variables into account and neglects possible correlations between these variables. An alternative metric matching technique which takes correlations into account is *Mahalanobis metric matching,* based on the following measure of distance in matrix notation:

$$(\boldsymbol{X}_1 - \boldsymbol{X}_0)' \boldsymbol{S}^{-1} (\boldsymbol{X}_1 - \boldsymbol{X}_0), \tag{6.12}$$

where $\boldsymbol{S}$ is the matrix of sample variances and covariances (specifically, the pooled within-sample covariance matrix) and $\boldsymbol{X}$ represents a column vector of values of the confounding variables. This distance function can be used in the

same way as Euclidean distance using a fixed tolerance to define a match.

Circular and Mahalanobis metric matching with tolerance ϵ can result in a loss of information, however, since there is no guarantee of matching all treatment subjects, even when the comparison reservoir is large. One approach to overcome this potential loss of treatment subjects is to generalize the method of nearest available matching (Section 6.5). Cochran and Rubin (1973) suggest randomly ordering the treatment subjects and then assigning as a match the comparison subject who is not yet matched and who is nearest as measured by some distance function, such as (6.10), (6.11), or (6.12). Such methods are called *nearest available metric matching* methods. Smith et al. (1977) proposed nearest available circular matching. Rubin (1979) compared the percent bias reduction of nearest available Mahalanobis metric matching with that of nearest available discriminant matching (Section 6.10.4). Rubin's study is discussed in Section 6.10.5.

6.10.4 Discriminant Matching

Another approach for dealing with multiple confounding variables is to transform the many variables to a single new variable and then to apply a univariate matching procedure to this single variable. One such transformation is the *linear discriminant function.* Basically, the linear discriminant function is a linear combination of the confounding variables that best predicts group membership. In a sense, it is the variable on which the groups differ the most.* By matching on this single variable, it is hoped to achieve the maximum amount of bias reduction. Using one of the univariate matching procedures described above on the linear discriminant function, those cases will be selected whose discriminant function values are the closest. For more detailed references on discriminant matching, see Cochran and Rubin (1973) and Rubin (1976a, b; 1979). Snedecor and Cochran (1967, Chap. 13) show how to use multiple regression to calculate the discriminant function.

6.10.5 Multivariate Matching with Linear Adjustment

Rubin (1979) empirically examined the nearest available Mahalanobis metric and nearest available discriminant matching methods, alone and in combination with regression adjustment on the matched pair differences for various sampling situations, and for various underlying models, both linear and nonlinear. (See Section 13.3.1 for a discussion of matching with regression adjustment.) Rubin selected these two methods because, under certain distributional assumptions,

*A good survey paper on discriminant analysis is that of Lachenbruch and Goldstein (1979). In this paper a discussion is given of discriminant analysis on numerical variables, categorical variables, and multivariate data containing both numerical and categorical variables.

they are equal percent bias reducing (EPBR); that is, they yield the same percent reduction in bias for each matching variable. As a result, this percent bias reduction is a straightforward criterion of how well the EPBR matching method has reduced bias in the estimate of the treatment effect.

The broad conclusion of Rubin (1979) is that nearest available pair matching using the Mahalanobis metric, together with regression adjustment on the matched pair differences, is an effective plan for controlling the bias due to the confounding variables, even for moderately nonlinear relationships.* Over a wide range of distributional conditions used in his Monte Carlo study, this metric matching method reduced the expected squared bias by an average of 12% more than did random sampling with no matching. (Notice that for univariate matching methods, the results given in Tables 6.2, 6.3, and 6.5 all relate to percent bias reduction and not to percent squared bias reduction.) This metric matching reduces more than 90% of the squared bias. Without regression adjustment, nearest available discriminant matching is equivalent to nearest available Mahalanobis metric matching, although Rubin finds that Mahalanobis metric matching is more robust against alternative model and distributional specifications.

6.11 MULTIPLE COMPARISON SUBJECTS

Occasionally, matched samples may be generated by matching each treatment subject with more than one comparison subject. Matching with multiple controls is especially advantageous when the number of potential comparison subjects is large relative to the number of available treatment subjects or when the unit cost for obtaining the comparison subjects is substantially lower than that of obtaining treatment subjects.

The present discussion concentrates on the dichotomous outcome case. Assume that each treated subject is matched with the same number, say q, of comparison subjects. The selection of a particular multivariate matching procedure should be based on the same principles explained in previous sections for a single comparison subject. For an example of pair matching using multiple controls, see Haddon et al. (1961).

Let the data from the jth matched group, $j = 1, 2, \ldots, N$, be represented in terms of a 2×2 frequency table, Table 6.9. Here $a_j = 1$ if the treatment subjects have the outcome factor present and $a_j = 0$ otherwise; b_j is the number of comparison subjects who have the outcome factor present.

* For univariate matching, an extreme example of a moderately nonlinear relationship is $Y = \exp(X)$, whereas $Y = \exp(X/2)$ is more reasonable. In multivariate matching, a similar statement can be made.

Table 6.9 Multiple Comparison Subjects Data from jth Matched Sample

Outcome	Treatment Subjects	Comparison Subjects	Total
Factor present (= 1)	a_j	b_j	$a_j + b_j$
Factor absent (= 0)	$1 - a_j$	$q - b_j$	$1 + q - (a_j + b_j)$
Total	1	q	$1 + q$

For simplicity, let us make the following definitions:

$$A = \sum_{j=1}^{N} a_j,$$

where A is the total number of treatment subjects who have the outcome factor present, and

$$B = \sum_{j=1}^{N} b_j,$$

where B is the total number of control subjects who have the outcome factor present. Therefore, the rate at which the outcome factor is present among the treatment group is $p_1 = A/N$, and the rate at which it is present among the comparison group is $p_0 = B/qN$. The difference in rates, as a measure of treatment effect, is then estimated by $p_1 - p_0$.

To estimate the odds ratio, each set of $q + 1$ subjects is considered a stratum, and estimators appropriate to stratified samples are applied. (See Section 7.6.1 for a more detailed discussion of estimators of the odds ratio that are appropriate when the number of strata becomes large.) Two such estimators, the conditional maximum likelihood and Mantel–Haenszel estimators, are given by Miettinen (1970). For $q \geq 3$, the conditional maximum likelihood estimator becomes difficult to use because it requires an iterative solution. For the case of exact matching, the conditional maximum likelihood estimator will be approximately unbiased for large N; Miettinen conjectures that the same is true for the Mantel–Haenszel estimator. No comparison of these two estimators as applied to multiple comparison subjects has been made. McKinlay's (1978) results for stratification (Section 7.6.2) imply that the Mantel–Haenszel estimator will be less biased than the conditional maximum likelihood estimator will be. For the case of a single comparison subject for each treatment subject ($q = 1$), the two estimators are identical (and are given in Section 6.9).

To test the null hypothesis of no treatment effect, we wish to consider the difference between p_1 and p_0. An appropriate test statistic is

$$T = \frac{p_1 - p_0}{\mathrm{SE}(p_1 - p_0)} = \frac{qA - B}{(q+1)(A+B) - \sum_{j=1}^{N} (a_j - b_j)^2},$$

where $\mathrm{SE}(p_1 - p_0)$ is the standard error of the difference. Miettinen (1969) has shown for large N that T has a standard normal distribution under the null hypothesis. He has also studied the power of the test and has given criteria, in terms of reducing cost, for deciding on an appropriate value of q, the number of comparison subjects per treatment subject.

When the outcome variable is continuous, one could compare the value for each treatment subject with the mean value of the corresponding controls, resulting in N differences. For a discussion of this approach, see Ury (1975).

Ury (1975) also presents an analysis of the statistical efficiency that can be gained by matching each case with several independent controls. For the dichotomous as well as the continuous outcome variables, the efficiency of using q controls versus a single control is approximately equal to $2q/(q + 1)$. For example, using 2 controls would increase the efficiency by about 33%; using 3 controls, by about 50%.

6.12 OTHER CONSIDERATIONS

This section includes three miscellaneous topics regarding matching. Sections 6.12.1 and 6.12.2 present results for matching that relate to general problems discussed in Sections 5.1 and 5.2, respectively: omitted confounding variables and measurement error. Some ideas regarding judging the quality of matches when exact matching is not possible are given in Section 6.12.3.

6.12.1 Omitted Confounding Variables

A common criticism investigators must face is that all the important confounding variables have not been taken into account. Unfortunately, with respect to matching, there are only very general indications of the effect of an omitted confounding variable.

Should the omitted confounding variable Z have a linear, parallel relationship with the included confounding variable X in the two populations, then matching solely on X removes only that part of the bias which can be attributed to the linear regression of Z on X. The amount of bias removed depends on the value of the regression coefficient of Z on X.

According to Cochran and Rubin (1973), if the regression of Z on X are nonlinear but parallel, then in large samples, matching solely on X will remove only that part of the bias due to Z that corresponds to the linear component of the regression of Z on X. These results generalize to the case of multiple confounding variables.

6.12.2 Errors of Measurement in Confounding Variables

If we assume that the response is linearly related to the correctly measured, or true, confounding variable in both populations, but that we can only match on values which are measured with error, then except under certain special conditions, the relationship between the response and fallible confounding variable will not be linear.

As an indication of the effect of measurement error on matching, consider the case where the response and the fallible confounding variable are linearly related. Then matching on the fallible variable has the effect of multiplying the expected percent reduction in bias by the ratio of β^*/β (Cochran and Rubin, 1973). In this ratio, β^* is the regression coefficient of the response on the fallible confounding variable and β is the regression coefficient of the response on the true confounding variable. Since this ratio is usually less than 1, matching on a confounding variable measured with error results in less bias reduction than does matching on the corresponding accurately measured confounding variable.

6.12.3 Quality of Pair Matches

In the case of pair matching the investigator can be lead to significant errors of interpretation if the quality of the matches is poor. Quality is judged by the magnitude of the differences between the values of the confounding variables for the comparison and treatment subjects. In this section we discuss a general approach that uses stratification to investigate any imperfect matching and its effect. We also discuss two approaches suggested by Yinger et al. (1967) for the case of a numerical outcome variable.

Perhaps the obvious first step in determining the overall quality of the pair matches obtained is to employ simple summary statistics such as the mean or median of the absolute differences between pairs for each particular confounding variable. Such statistics, however, do not give the investigator any indication of a relationship between the response and the closeness of matches. It is the existence of such a relationship which should be taken into account when interpreting the findings of a study. For example, if in a study on weight loss (numerical response) the pairs which show the greatest difference in weight loss were the pairs who were most imperfectly matched, the investigator should be suspicious of the apparent effect of the treatment.

How can investigators determine if there is any relationship between response and the quality of the matches? In the case of a categorical response, the investigators can take one of two approaches, depending on the number of confounding variables. If there are only a few variables, they can determine summary statistics for each response category. This may be viewed as analyzing the effect of possible imperfect matching by stratifying on the response. The summary statistics should be nearly equivalent for all response categories.

If there are several confounding variables, the investigators may instead wish to determine a single summary statistic of the quality of the matches for each response category. This can be done in a two-step procedure. First, for each confounding variable, the differences between matched pairs are categorized and weights are assigned to each category. For example, in the weight-loss study, if age is one of the confounding variables, a difference of 0 to 6 months may receive a weight of 0, and a difference of 7 to 11 months a weight of 1, while a difference of 12 months or more may receive a weight of 3. For the second step, the weights are summed across all confounding variables for each matched pair in a response strata and again we could either take the mean or the median as a summary statistic of the closeness of the matches. These numbers should agree across response strata. We wish to point out that these weights are arbitrary and are only meant to be used for within-study comparisons.

Yinger et al. (1967) have two methods for studying the effects of imperfect matching in the case of a numerical outcome. The first of their methods consists of forming a rough measure of the equivalence of the treatment and comparison groups by the weighting method discussed above for a categorical outcome. They call this measure the index of congruence.

Consider a study of reading ability, where age, sex, and birth order are confounding variables. Table 6.10 illustrates the calculation of the index of congruence for such a study. Here the index of congruence can range from 0 to 8 points, where a score of 0 indicates close matching and a score of 8 indicates the maximum possible difference between a treatment and control subject. Again, these scores are arbitrary and only meant as descriptive measures for within-study comparisons.

To determine if there is any relationship between the response and the quality of the matches, we can either calculate the correlation coefficient between the estimated treatment effect and the index of congruence, or plot the relationship. Ideally, both the correlation coefficient and the slope of the plotted curve should be close to zero, indicating no relationship.

The index of congruence gives only a rough measure of the group equivalence, in part because it does not take into account any directional influences of the confounding variables. The second of the Yinger et al. methods forms a directional measure of congruence which takes this factor into account.

The investigator may have prior knowledge (e.g., from previous research) of the directional influence of the confounding variables on the outcome. Con-

Table 6.10 Index of Congruence Calculation

Confounding Variable	Score	Range of Possible Point Differences between Matched Pairs
Age difference		
0–6 months	= 0	
7–11 months	= 1	0–3
12+ months	= 3	
Sex		
Same	= 0	0–3
Different	= 3	
Birth order		
Both either firstborn or not firstborn	= 0	0–2
Otherwise	= 2	
Total		0–8

sider again the study of reading ability and suppose that increasing age had a positive influence while increasing rank of birth had a negative influence. In addition, suppose the matching on sex was exact, so that we need not consider the directional influence. For the directional measure of congruence we will use only scores -1, 0, and 1, where -1 indicates that the treatment subject has a value of the confounding variable implying that the response for the treatment subject is expected to be inferior to that of the comparison subject; 0 indicates that they are expected to be the same, and 1 indicates that the treatment subject response is expected to be superior. Then, the directional index of congruence for matching a 12-year-old firstborn treatment subject with a 14-year-old second-born comparison subject of the same sex would be zero, since the treatment subjects superiority due to a lower birth order would be offset by his or her inferiority due to a lower age. In contrast, the index of congruence based on Table 6.10 for such a match would be 5.

The investigator may also plot the estimated treatment effect versus the value of the directional measure of congruence for each pair. By considering the scatter diagram and regression curve, the investigator can judge to what degree the treatment effect is related to differences between the matched pairs. Ideally, the plot should again show no relationship.

6.13 Conclusions

The most practical of the pair matching methods is nearest available matching. It has the advantage that matches can always be found. However, because of the varying tolerance, it will not be as effective as caliper matching in reducing the bias in the estimation of the treatment effect.

The pair matching methods are the best methods to use when the relationship is nonlinear. Rubin (1973a) found that the percentage reduction in bias for random-order nearest available matching in the linear case (Table 6.2) was overestimated by less than 10 percent in most nonlinear cases. Pair matching methods do require a large control reservoir, however, and are therefore difficult to use in studies with a large treatment group or where it takes a long time to find comparison subjects. They seem to be the most effective when σ_1^2/σ_0^2 is approximately 1 and to be least effective when σ_1^2/σ_0^2 is approximately 2 or more, with $\eta_1 > \eta_0$. Rubin also concludes that matching with $r \geq 2$ generally improves the estimate of the treatment effect, especially if the variance of the confounding variable is greater in the comparison population than in the treatment population.

Nonpair matching methods—mean and frequency—are quicker than are pair matching methods. However, mean matching is not used often because of its strong dependence on the assumption of linearity. If the investigator feels very confident that the relationship is a linear parallel one, and if the treatment and comparison groups are about the same size, mean matching may be considered as a fast matching procedure which has the same precision as pair matching in such a situation.

APPENDIX 6A: SOME MATHEMATICAL DETAILS

6A.1 Matching Model

The general mathematical model used to analyze the effect of matching with a numerical outcome variable can be represented as

$$Y_{ij} = R_i(X_{ij}) + e_{ij} \qquad i = 1, 0; j = 1, 2, \ldots, n_i, \tag{6.13}$$

where $i = 1$ represents the treatment group, $i = 0$ the comparison group, and j is the jth observation in each group. Furthermore, Y_{ij} is the response variable and is a function of the confounding variable X_{ij}. The residual e_{ij} has mean zero and variance σ_i^2, and X_{ij} has mean η_i. We assume that the Y and X are numerical variables.

We now consider the specific forms of the response function R_i (·) that correspond to the linear parallel and quadratic parallel relationships.

6A.2 Parallel Linear Regression

For the parallel linear regression, the model (6.13) becomes

$$Y_{ij} = \mu_i + \beta(X_{ij} - \eta_i) + e_{ij} \qquad i = 1, 0; j = 1, 2, \ldots, n_i \qquad (6.14)$$

or

$$Y_{ij} = \alpha_i + \beta X_{ij} + e_{ij},$$

where

$$\alpha_i = \mu_i - \beta\eta_i.$$

Notice that the slope is the same for both the treatment and the comparison groups. In this case the *treatment effect* $\alpha_1 - \alpha_0$ is defined as the difference in the intercept terms:

$$\alpha_1 - \alpha_0 = (\mu_1 - \mu_0) - \beta(\eta_1 - \eta_0). \qquad (6.15)$$

Since the estimator of the treatment effect (6.15) is the difference in the mean responses between the two groups, $\bar{Y}_1 - \bar{Y}_0$, with mean response defined as $E(\bar{Y}_i|\bar{X}_i) = \alpha_i + \beta\bar{X}_i$, the expected value of the estimator is

$$E(\bar{Y}_1 - \bar{Y}_0|\bar{X}_1, \bar{X}_0) = \alpha_1 - \alpha_0 + \beta(\bar{X}_1 - \bar{X}_0). \qquad (6.16)$$

From (6.16) it follows that the estimator $(\bar{Y}_1 - \bar{Y}_0)$ is biased by an amount $\beta(\bar{X}_1 - \bar{X}_0)$. Thus a matching procedure that makes $|\bar{X}_1 - \bar{X}_0|$ as small as possible will be preferred.

6A.3 Parallel Nonlinear Regression

Consider a parallel quadratic relationship. Then the matching model can be written as

$$Y_{ij} = \mu_i + \beta(X_{ij} - \eta_i) + \delta X_{ij}^2 + e_{ij}.$$

It follows that

$$E(\bar{Y}_1 - \bar{Y}_0|\bar{X}_1, \bar{X}_0) = \mu_1 - \mu_0 - \beta(\eta_1 - \eta_0) + \beta(\bar{X}_1 - \bar{X}_0) + \delta(\overline{X_1^2} - \overline{X_0^2}) + \delta(s_1^2 - s_0^2), \qquad (6.17)$$

where $s_i^2 = \Sigma_{j=1}^{N} (X_{ij} - \bar{X}_i)^2/N$ for $i = 1, 0$. Comparing (6.17) with the treatment difference (6.15), we see that the bias equals

$$\beta(\bar{X}_1 - \bar{X}_0) + \delta(\overline{X_1^2} - \overline{X_0^2}) + \delta(s_1^2 - s_0^2). \qquad (6.18)$$

Clearly, in this nonlinear case, equality of the confounding variable means is not sufficient. (In particular, mean matching is not appropriate.) This emphasizes the motivation of tight pair matching. By choosing pairs so that each treatment subject is very closely matched to a comparison subject, any difference in the sample confounding variable distributions that may be important [such as means and variances in (6.18)] is made small.

REFERENCES

Althauser, R. P., and Rubin, D. B. (1970), The Computerized Construction of a Matched Sample, *American Journal of Sociology*, **76**, 325–346.

Carpenter, R. G. (1977), Matching When Covariables Are Normally Distributed, *Biometrika*, **64**(2), 299—307.

Cochran, W. G. (1968), The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies, *Biometrics*, **24**(2), 295–313.

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

Cox, D. R. (1957), Note on Grouping, *Journal of the American Statistical Association*, **52**(280), 543–547.

Fleiss, J. L. (1973), *Statistical Methods for Rates and Proportions*, New York: Wiley.

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

Haddon, W., Jr., Valien, P., McCarroll, J. R., and Umberger, C. J. (1961), A Controlled Investigation of the Characteristics of Adult Pedestrians Fatally Injured by Motor Vehicles in Manhattan, *Journal of Chronic Disease*, **14**, 655–678. Reprinted in E. R. Tufte, Ed., *The Quantitative Analysis of Social Problems*, Reading, MA: Addison-Wesley, 1970, pp. 126–152.

Kalton, G. (1968), Standardization: A Technique to Control for Extraneous Variables, *Journal of the Royal Statistical Society, Series C*, **17**, 118–136.

Lachenbruch, P. A., and Goldstein, M. (1979), Discriminant Analysis, *Biometrics*, **35**(1), 69-85.

McKinlay, S. M. (1974), The Expected Number of Matches and Its Variance for Matched-Pair Designs, *Journal of the Royal Statistical Society, Series C*, **23**(3), 372–383.

McKinlay, S. M. (1975), The Effect of Bias on Estimators of Relative Risk for Pair-Matched and Stratified Samples, *Journal of the American Statistical Association*, **70**(352), 859–864.

McKinlay, S. M. (1978), The Effect of Nonzero Second-Order Interaction on Combined Estimators of the Odds Ratio, *Biometrika*, **65**, 191–202.

Mantel, N., and Haenszel, W. (1959), Statistical Aspects of the the Analysis of Data from Retrospective Studies of Disease, *Journal of the National Cancer Institute*, **22**, 719–748.

Miettinen, O. S. (1969), Individual Matching with Multiple Controls in the Case of All-or-None Responses, *Biometrics*, **25**(2), 339–355.

Miettinen, O. S. (1970), Estimation of Relative Risk from Individually Matched Series, *Biometrics*, **26**(1), 75–86.

Rubin, D. B. (1973a), Matching to Remove Bias in Observational Studies, *Biometrics*, **29**(1), 159–183.

Rubin, D. B. (1973b), The Use of Matched Sampling and Regression Adjustment to Remove Bias in Observational Studies, *Biometrics*, **29**(1), 185–203.

Rubin, D. B. (1976a), Multivariate Matching Methods That Are Equal Percent Bias Reducing, I: Some Examples, *Biometrics*, **32**(1), 109–120; 955.

Rubin, D. B. (1976b), Multivariate Matching Methods That Are Equal Percent Bias Reducing, II: Maximums on Bias Reduction for Fixed Samples Sizes, *Biometrics*, **32**(1), 121–132, 955.

Rubin, D. B. (1979), Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies, *Journal of the American Statistical Association*, **74**, 318–328.

Smith, A. H., Kark, J. D., Cassel, J. C., and Spears, G. F. S. (1977), Analysis of Prospective Epidemiologic Studies by Minimum Distance Case-Control Matching, *American Journal of Epidemiology*, **105**(6), 567–574.

Snedecor, G. W., and Cochran, W. G. (1967), *Statistical Methods*, 6th ed., Ames, IA: Iowa State University Press.

Ury, H. K. (1975), Efficiency of Case-Control Studies with Multiple Controls per Case: Continuous or Dichotomous Data, *Biometrics*, **31**, 643–649.

Yinger, J. M., Ikeda, K., and Laycock, F. (1967), Treating Matching as a Variable in a Sociological Experiment, *American Sociological Review*, **23**, 801–812.

CHAPTER 7

Standardization and Stratification