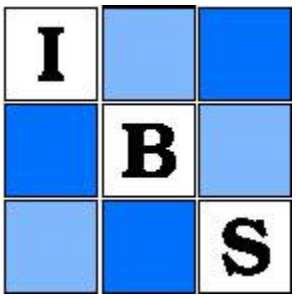

BIOMETRICS *29*, 159–183
March 1973

MATCHING TO REMOVE BIAS IN OBSERVATIONAL STUDIES

DONALD B. RUBIN[1]

Department of Statistics, Harvard University, Cambridge, Massachusetts 02138, USA

SUMMARY

Several matching methods that match all of one sample from another larger sample on a continuous matching variable are compared with respect to their ability to remove the bias of the matching variable. One method is a simple mean-matching method and three are nearest available pair-matching methods. The methods' abilities to remove bias are also compared with the theoretical maximum given fixed distributions and fixed sample sizes. A summary of advice to an investigator is included.

1. INTRODUCTION

Matched sampling is a method of data collection and organization designed to reduce bias and increase precision in observational studies, i.e. in those studies in which the random assignment of treatments to units (subjects) is absent. Although there are examples of observational studies which could have been conducted as properly randomized experiments, in many other cases the investigator could not randomly assign treatments to subjects. For example, consider the Kihlberg and Robinson [1968] study comparing severity of injury in automobile accidents for motorists using and not using seatbelts. One would not want to randomly assign subjects to "seatbelt" and "no seatbelt" treatments and then have them collide at varying speeds, angles of impact, etc. Neither, however, would one want to simply compare the severity of injury in "random" samples of motorists in accidents using and not using seatbelts; important variables such as "speed of automobile at time of accident" may be differently distributed in the two groups (i.e. seatbelted motorists are generally more cautious and therefore tend to drive more slowly). Hence, in observational studies, methods such as matched sampling or covariance adjustment are often needed to control bias due to specific variables.

We will investigate matched sampling on one continuous matching variable X (e.g., speed of automobile at time of accident) and two treatment populations, P_1 and P_2 (e.g., motorists in accidents using and not using seatbelts). Several articles have previously considered this situation. However, most of these have assumed that the average difference in the dependent variable

[1] Present Address: Educational Testing Service, Princeton, New Jersey 08540.

between the matched samples is an unbiased estimate of the effect of the treatment and thus were interested in the ability of matching to increase the precision of this estimate. See, for example, Wilks [1932], Cochran [1953], Greenberg [1953], and Billiwicz [1965]. Here, we will investigate the ability of matched sampling to reduce the bias of this estimate due to a matching variable whose distribution differs in P_1 and P_2 (e.g., to reduce the bias due to "speed at time of accident").

We assume that there is a random sample of size N from P_1 , say G_1 , and a larger random sample of size rN, $r \geq 1$, from P_2 , say G_2 . All subjects in G_1 and G_2 are assumed to have recorded scores on the matching variable X. Using these scores, a subsample of G_2 of size N will be chosen according to some "matching method"; we call this subsample G_{2*} . The effect of the treatment will then be estimated from the G_1 and G_{2*} samples both of size N. If r is one, G_{2*} would be a random sample from P_2 , and matching could not remove any bias due to X; if r is infinite, perfect matches could always be obtained, and all of the bias due to X could be removed. We will study moderate ratios of sample sizes, basically $r = 2, 3, 4$, although some results are given for $r = 6, 8, 10$.

Following Cochran [1968], we will use "the percent reduction in the bias of X due to matched sampling" as the measure of the ability of a matching method to reduce the bias of the estimated effect of the treatment; justification for this choice is given in section 2. Then section 3 states and proves a theorem giving the maximum obtainable percent reduction in bias given fixed distributions of X in P_1 and P_2 and fixed samples sizes N and rN. In section 4, the ability of a simple mean-matching method to reduce bias will be compared with the theoretical maximum. In section 5, we compare three "nearest available" pair-matching methods with respect to their ability to reduce bias. Section 6 serves to present practical advice to an investigator.

2. TERMINOLOGY; PERCENT REDUCTION IN BIAS

Suppose that we want to determine the effect of a dichotomous treatment variable on a continuous dependent variable, Y, given that the effect of a continuous matching variable, X, has been removed.[2] The dichotomous treatment variable is used to form two populations P_1 and P_2 . In P_1 and P_2 X and Y have joint distributions which in general differ from P_1 to P_2 . In P_i the conditional expectation of the dependent variable Y given a particular value of X is called the response surface for Y in P_i , and at $X = x$ is denoted $R_i(x)$.

The difference in response surfaces at $X = x$, $R_1(x) - R_2(x)$, is the effect of the treatment variable at $X = x$. If this difference between response surfaces is constant and so independent of the values of the matching variable,

[2] As Cochran [1968] points out, if the matching variable X is causally affected by the treatment variable, some of the real effect of the treatment variable will be removed in the adjustment process.

the response surfaces are called parallel, and the objective of the study is the estimation of the constant difference between them. See Figure 1. For linear response surfaces, "parallel response surfaces" is equivalent to "having the same slope".

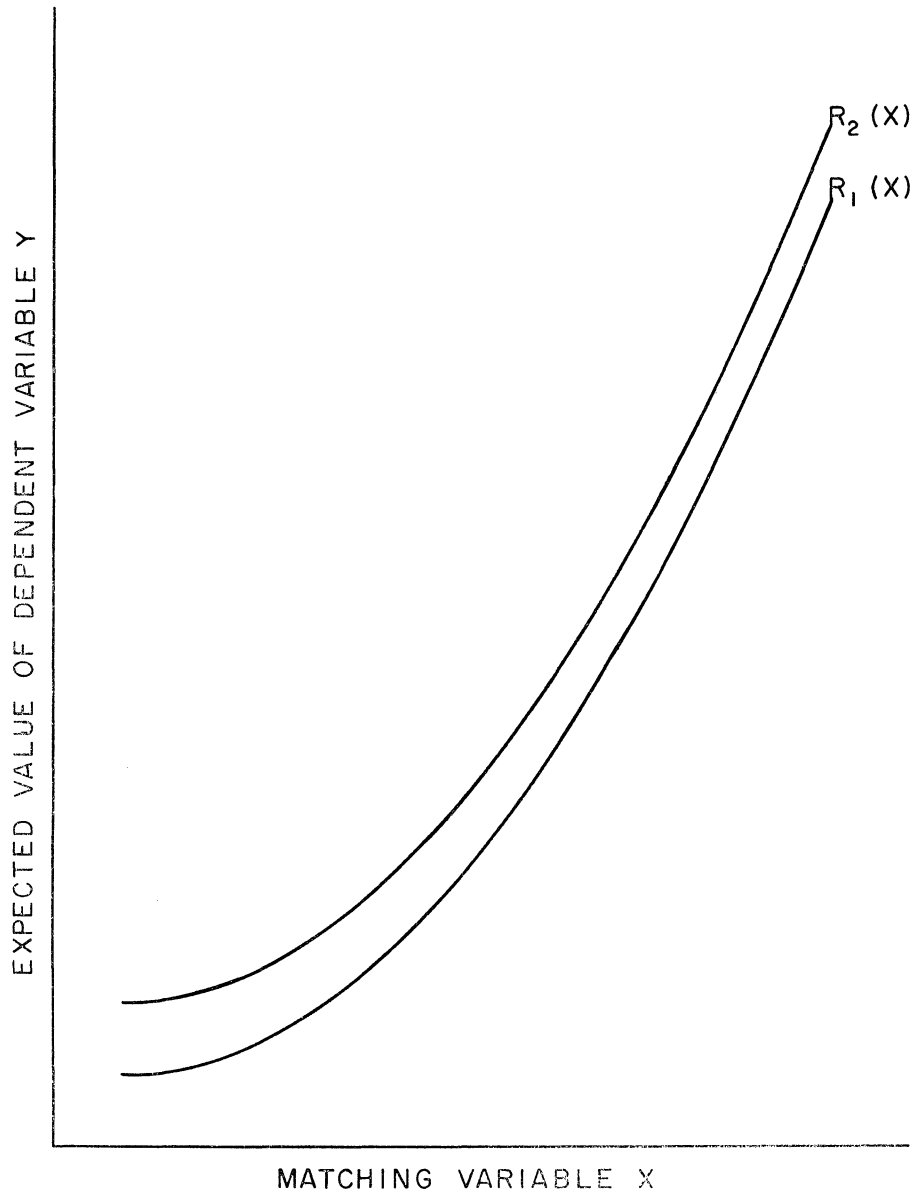

FIGURE 1

PARALLEL UNIVARIATE RESPONSE SURFACES

If $R_1(x) - R_2(x)$ depends on x, the response surfaces are non-parallel and there is no single parameter that completely summarizes the effect of the treatment variable. In this case we will assume that the average effect of the treatment variable (the average difference between the response surfaces) over the P_1 population is desired. Such a summary is often of interest, especially when P_1 consists of subjects exposed to an agent and P_2 consists

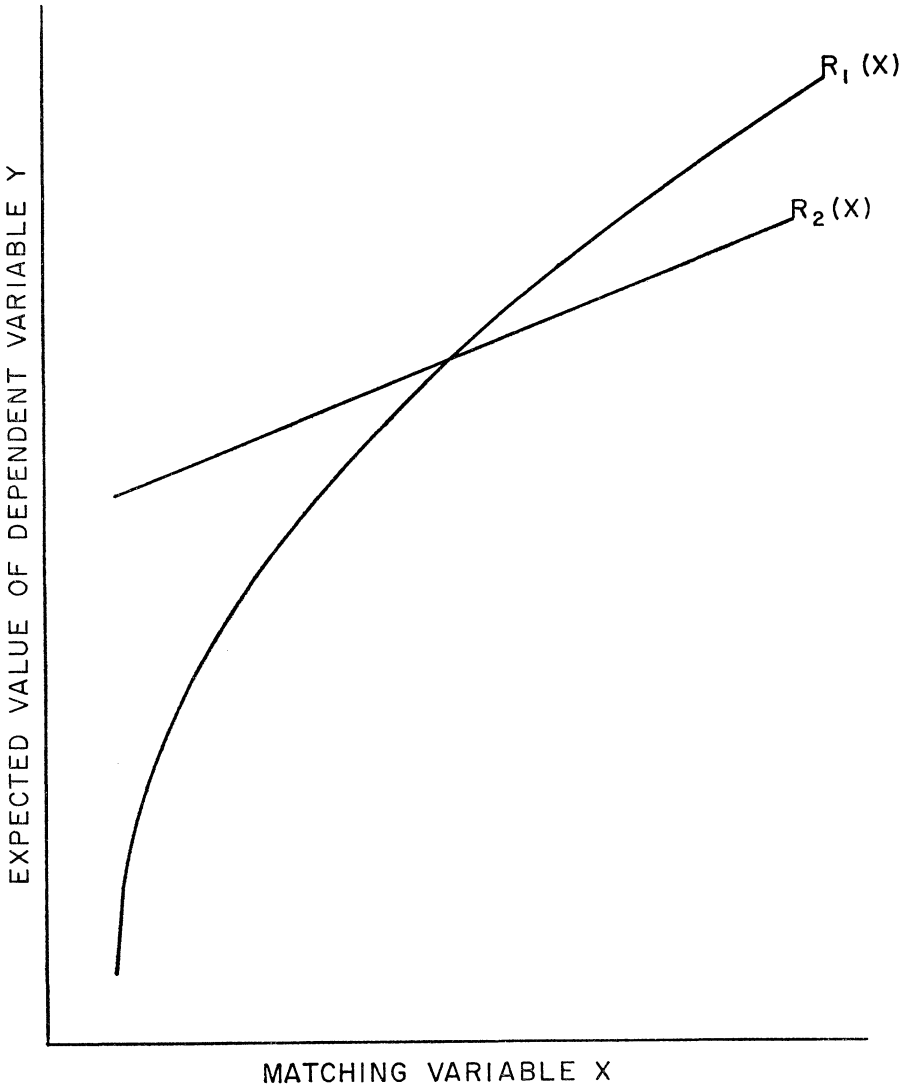

FIGURE 2

NONPARALLEL UNIVARIATE RESPONSE SURFACES

of controls not exposed to the agent; see for example Belsen's [1956] study of the effect of an educational television program.[3]

The average difference between non-parallel response surfaces over the P_1 population or the constant difference between parallel response surfaces will be called the (average) effect of the treatment variable or more simply "the treatment effect" and will be designated τ:

$$\tau = E_1\{R_1(x) - R_2(x)\}, \tag{2.1}$$

where E_1 is the expectation over the distribution of X in P_1.

Let y_{1j} and x_{1j} be the values of Y and X for the jth subject in G_1, and similarily let y_{2j} and x_{2j} be the values of Y and X for the jth subject in G_{2*}, $j = 1, \cdots, N$. Using the response surface notation we can write

$$y_{ij} = R_i(x_{ij}) + e_{ij} \quad i = 1, 2; \quad j = 1, \cdots, N \tag{2.2}$$

where $E_c(e_{ij}) = 0$ and E_c is the conditional expectation given the x_{ij}.

We assume that the difference between dependent variable averages in G_1 and G_{2*} will be used to estimate τ:

$$\hat{\tau}_0 = \frac{1}{N}\sum y_{1j} - \frac{1}{N}\sum y_{2j} = \bar{y}_{1.} - \bar{y}_{2.} \tag{2.3}$$

Let E represent the expectation over the distributions of X in matched samples and E_{2*} represent the expectation over the distribution of X in matched G_{2*} samples. Then using (2.3) and (2.1) we have that the expected bias of $\hat{\tau}_0$ over the matched sampling plan is

$$EE_c(\hat{\tau}_0 - \tau) = E_1R_2(x) - E_{2*}R_2(x) \tag{2.4}$$

since $EE_c(\bar{y}_{2.}) = E_{2*}R_2(x)$ and $EE_c(\bar{y}_{1.}) = E_1R_1(x)$. If the distribution of X in matched G_{2*} samples is identical to that in random G_1 samples then $E_1R_2(x) = E_{2*}R_2(x)$ and $\hat{\tau}_0$ has zero expected bias. If $r = 1$, that is if the G_{2*} sample is a random sample from P_2, then the expected bias of $\hat{\tau}_0$ is $E_1R_2(x) - E_2R_2(x)$ where E_2 is the expectation over the distribution of X in P_2.

In order to indicate how much less biased $\hat{\tau}_0$ based on matched samples is than $\hat{\tau}_0$ based on random samples, Cochran [1968] uses "the percent reduction in bias" or more precisely "the percent reduction in expected bias": $100 \times (1 -$ expected bias for matched samples/expected bias for random samples) which is from (2.4)

[3] In other cases, however, this average difference may not be of primary interest. Consider for example the previously mentioned study of the efficacy of seatbelts. Assume that if automobile speed is high seatbelts reduce the severity of injury, while if automobile speed is low seatbelts increase the severity of injury. (See Figure 2, where P_1 = motorists using seatbelts, P_2 = motorists not using seatbelts, X = automobile speed, and Y = severity of injury.) A report of this result would be more interesting than a report that there was no effect of seatbelts on severity of injury when averaged over the seatbelt wearer population. Since such a report may be of little interest if the response surfaces are markedly nonparallel, the reader should generally assume "nonparallel" to mean "moderately nonparallel." If the response surfaces are markedly nonparallel and the investigator wants to estimate the effect of the treatment variable averaged over P_2 (the population from which he has the larger sample), the methods and results presented here are not relevant and a more complex method such as covariance analysis would be more appropriate than simple matching. (See Cochran [1969] for a discussion of covariance analysis in observational studies.)

$$100\left\{1 - \frac{E_1R_2(x) - E_{2*}R_2(x)}{E_1R_2(x) - E_2R_2(x)}\right\} = 100\,\frac{E_{2*}R_2(x) - E_2R_2(x)}{E_1R_2(x) - E_2R_2(x)}. \qquad (2.5)$$

Notice that the percent reduction in bias due to matched sampling depends only on the distribution of X in P_1 , P_2 and matched G_{2*} samples, and the response surface in P_2 . If the response surface in P_2 is linear,

$$R_2(x) = \mu_2 + \beta_2(x - \eta_2)$$

where

$$\mu_2 = \text{mean of } Y \text{ in } P_2$$

$$\eta_i = \text{mean of } X \text{ in } P_i$$

and

$$\beta_2 = \text{regression coefficient of } Y \text{ on } X \text{ in } P_2 ,$$

we have for the denominator of (2.5), $\beta_2(\eta_1 - \eta_2)$ and for the numerator of (2.5) $\beta_2(\eta_{2*} - \eta_2)$ where η_{2*} is the expected value of X in matched G_{2*} samples, $E_{2*}(x)$ (equivalently, η_{2*} is the expected average X in G_{2*} samples, $E(\bar{x}_{2.})$).

Thus, if G_1 is a random sample and the response surface in P_2 is linear, the percent reduction in bias due to matched sampling is

$$\Theta = 100\,\frac{\eta_{2*} - \eta_2}{\eta_1 - \eta_2}, \qquad (2.6)$$

which is the same as the percent reduction in bias of the matching variable X. Even though only an approximation if the P_2 response surface is not linear, we will use Θ, the percent reduction in the bias of the matching variable, to measure the ability of a matching method to remove bias.

3. THE MAXIMUM PERCENT REDUCTION IN BIAS GIVEN FIXED DISTRIBUTIONS AND FIXED SAMPLE SIZES

Assume that in P_i X has mean η_i (without loss of generality let $\eta_1 > \eta_2$), variance σ_i^2 and $(X - \eta_i)/\sigma_i \sim f_i$, $i = 1, 2$. Define the initial bias in X to be

$$B = \frac{\eta_1 - \eta_2}{\sqrt{\frac{\sigma_1^2 + \sigma_2^2}{2}}} > 0,$$

which if $\sigma_1^2 = \sigma_2^2$ is simply the number of standard deviations between the means of X in P_1 and P_2 .

Then if Θ is the percent reduction in bias of X due to some matching method that selects a matched sample, G_{2*} , of N subjects from a random sample, G_2 , of rN P_2 subjects, we have

$$\Theta \leq \Theta_{\max} = 100\,\frac{\Omega_2(r, N)}{B\sqrt{\frac{1 + \sigma_1^2/\sigma_2^2}{2}}}, \qquad (3.1)$$

where $\Omega_2(r, N)$ = the expected value of the average of the N largest observations from a sample of size rN from f_2 .

Since a reduction in bias greater than 100% is clearly less desirable than 100% reduction in bias, if B, σ_1^2/σ_2^2 , and $\Omega_2(r, N)$ are such that $\Theta_{\max} \geq 100$ this should be interpreted as implying the existence of a matching method that obtains 100% reduction in expected bias.[4]

This result follows immediately from (2.6): since $\eta_1 > \eta_2$, Θ is the largest when η_{2*} (i.e. $E(\bar{x}_{2\cdot})$) is largest, which is clearly achieved when the N subjects in G_2 with the largest X values are always chosen as matches. The expected value of these N largest values from a sample of rN is $\eta_2 + \sigma_2\Omega_2(r, N)$. Hence, the maximum value of Θ is

$$\Theta_{\max} = 100\,\frac{\sigma_2\Omega_2(r, N)}{\eta_1 - \eta_2} = 100\,\frac{\Omega_2(r, N)}{B\sqrt{\dfrac{1 + \sigma_1^2/\sigma_2^2}{2}}}.$$

The result in (3.1) is of interest here for two reasons. First, for fixed distributions and sample sizes and given a particular matching method, a comparison of Θ and min $\{100, \Theta_{\max}\}$ clearly gives an indication of how well that matching method does at obtaining a G_{2*} sample whose expected X mean is close to η_1 . In addition, the expression for $\Theta_{\max}$ will be used to help explain trends in Monte Carlo results. When investigating matching methods that might be used in practice to match finite samples, properties such as percent reduction in bias are generally analytically intractable. Hence, Monte Carlo methods must be used on specific cases. From such Monte Carlo investigations it is often difficult to generalize to other cases or explain trends with much confidence unless there is some analytic or intuitive reason for believing the trends will remain somewhat consistent. It seems clear that if $\Theta_{\max}$ is quite small (e.g. 20) no matching method will do very well, while if $\Theta_{\max}$ is large (e.g. 200) most reasonable matching methods should do moderately well. Hence, we will use trends in $\Theta_{\max}$ to help explain trends in the Monte Carlo results that follow.

Two trends for $\Theta_{\max}$ are immediately obvious from (3.1).

(1) Given fixed r, N, f_2 and σ_1^2/σ_2^2 , $\Theta_{\max}$ decreases as B increases.
(2) Given fixed r, N, f_2 and B, $\Theta_{\max}$ decreases as σ_1^2/σ_2^2 increases.

Given fixed f_2 , B, and σ_1^2/σ_2^2 two other trends are derivable from simple properties of the order statistics and the fact that $\Theta_{\max}$ is directly proportional to $\Omega_2(r, N)$ (see Appendix A for proofs).

(3) Given fixed B, σ_1^2/σ_2^2 , f_2 and N, $\Theta_{\max}$ increases as r increases: $\Omega_2(r, N) \leq \Omega_2(r + a, N)$, $a \geq 0$; N, rN, aN integers.

[4] A matching method that has as its percent reduction in expected bias min $\{100, \Theta_{\max}\}$ may be of little practical interest. For example, consider the following matching method. With probability $P = \min\{1, 1/\Theta_{\max}\}$ choose the N G_2 subjects with the largest observations as the G_{2*} sample and with probability $1 - P$ choose a random sample of size N as the G_{2*} sample. It is easily checked that the percent reduction in expected bias using this method is min $\{100, \Theta_{\max}\}$.

(4) Given fixed B, σ_1^2/σ_2^2, f_2 and r, Θ_{max} increases as N increases: $\Omega_2(r, N) \leq \Omega_2(r, N + b)$, $b \geq 0$; N, rN, rb integers.

From the fourth trend, we have $\Omega(r, 1) \leq \Omega(r, N) \leq \Omega(r, \infty)$. Values of $\Omega(r, 1)$ have been tabulated in Sarhan and Greenberg [1962] for several distributions as the expected value of the largest of r observations. $\Omega(r, \infty)$ can easily be calculated by using the asymptotic result

$$\Omega(r, \infty) = r \int_w^\infty zf(z), \quad \text{where} \quad \int_w^\infty f(z) = 1/r.$$

Values of $\Omega(r, 1)$ and $\Omega(r, \infty)$ are given in Table 3.1 for $X \sim \pm\chi_\nu^2$ ($\nu = 2(2)10$) and $X \sim$ Normal, and for $r = 2, 3, 4, 6, 8, 10$.

Table 3.1 can be summarized as follows.

(a) For fixed r and ν, the results for $+\chi_\nu^2$ are more similar to those for the normal than are those for $-\chi_\nu^2$. This result is expected since the largest N observations come from the right tail of the distribution and the right tail of $+\chi_\nu^2$ is more normal than the right tail of $-\chi_\nu^2$ which is finite.

(b) Given a fixed distribution, as r gets larger the results differ more from those for the normal especially for $-\chi_\nu^2$. Again this is not surprising because the tails of low degree of freedom χ^2 are not very normal, especially the finite tail.

(c) For $r = 2, 3, 4$, and moderately normal distributions ($\pm\chi_\nu^2$, $\nu \geq 8$) the results for the normal can be considered somewhat representative. This conclusion is used to help justify the Monte Carlo investigations of a normally distributed matching variable in the remainder of this article.

(d) Given a fixed distribution and fixed r, the values for $\Omega(r, 1)$ are generally within 20% of those for $\Omega(r, \infty)$, suggesting that when dealing with moderate sample sizes as might commonly occur in practice, we would expect the fourth trend (Θ_{max} increasing function of N) to be rather weak.

In Table 3.2 values of $\Omega(r, N)$ are given assuming f normal, the same values of r as in Table 3.1, and $N = 1, 2, 5, 10, 100, \infty$. Values were found with the aid of Harter [1960]. For fixed r, the values of $\Omega(r, N)$ for $N \geq 10$ are very close to the asymptotic value $\Omega(r, \infty)$, especially when $r > 2$. Even $\Omega(2, 10)$ is within about 3% of $\Omega(2, \infty)$. These results indicate that the values for $\Omega(r, \infty)$ given in Table 3.1 may be quite appropriate for moderate sample sizes.

4. MEAN-MATCHING

Thus far we have not specified any particular matching method. Under the usual linear model "mean-matching" or "balancing" (Greenberg [1953]) methods are quite reasonable but appear to be discussed rarely in the literature. In this section we will obtain Monte Carlo percent reductions in bias

TABLE 3.1
$\Omega(r, N)$

	r = 2		3		4		6		8		10	
	N = 1	∞	1	∞	1	∞	1	∞	1	∞	1	∞
$+\chi^2_2$	0.50	0.69	0.83	1.10	1.08	1.38	1.45	1.79	1.72	2.08	1.93	2.30
$+\chi^2_4$	0.53	0.74	0.86	1.13	1.10	1.39	1.43	1.75	1.67	2.00	1.85	2.19
$+\chi^2_6$	0.54	0.76	0.86	1.13	1.09	1.38	1.41	1.72	1.64	1.95	1.81	2.12
$+\chi^2_8$	0.54	0.77	0.86	1.13	1.09	1.37	1.40	1.69	1.61	1.92	1.78	2.08
$+\chi^2_{10}$	0.55	0.78	0.86	1.13	1.08	1.36	1.39	1.68	1.60	1.89	1.76	2.05
Normal	0.56	0.80	0.85	1.09	1.03	1.27	1.27	1.50	1.42	1.65	1.54	1.75
$-\chi^2_{10}$	0.55	0.78	0.78	0.99	0.92	1.11	1.09	1.25	1.19	1.34	1.26	1.39
$-\chi^2_8$	0.54	0.77	0.77	0.98	0.91	1.09	1.06	1.22	1.16	1.30	1.22	1.35
$-\chi^2_6$	0.54	0.76	0.76	0.95	0.88	1.06	1.02	1.17	1.11	1.24	1.17	1.28
$-\chi^2_4$	0.53	0.74	0.73	0.91	0.84	1.00	0.97	1.09	1.04	1.14	1.08	1.17
$-\chi^2_2$	0.50	0.69	0.67	0.81	0.75	0.86	0.83	0.91	0.87	0.93	0.90	0.95

TABLE 3.2

$\Omega(r, N)$; f NORMAL

	r = 2	3	4	6	8	10
N = 1	0.56	0.85	1.03	1.27	1.42	1.54
2	0.66	0.96	1.14	1.38	1.53	1.64
5	0.74	1.03	1.22	1.45	1.60	1.70
10	0.77	1.06	1.24	1.47	1.62	1.72
25	0.78	1.08	1.26	1.49	1.64	1.74
50	0.79	1.08	1.27	1.50	1.65	1.75
100	0.80	1.09	1.27	1.50	1.65	1.75
∞	0.80	1.09	1.27	1.50	1.65	1.75

for a simple mean-matching method and compare these with the theoretical maximums given by (3.1).

Assuming linear response surfaces it is simple to show from (2.3) that the bias of $\hat{\tau}_0$ for estimating τ is $\beta_2(\eta_1 - \bar{x}_{2\cdot}) + \beta_1(\bar{x}_{1\cdot} - \eta_1)$, where β_i is the regression coefficient of Y on X in P_i and $\bar{x}_{i\cdot}$ is the average X in the matched samples. Using $\bar{x}_{1\cdot}$ to estimate η_1 or assuming parallel response surfaces ($\beta_1 = \beta_2$) one would minimize the estimated bias of $\hat{\tau}_0$ by choosing the N G_2 subjects such that $|\bar{x}_{1\cdot} - \bar{x}_{2\cdot}|$ is minimized. A practical argument against using this mean-matching method is that finding such a subset requires the use of some time consuming algorithm designed to solve the transportation problem. Many compromise algorithms can of course be defined that approximate this best mean-match.

We will present Monte Carlo percent reductions in bias only for the following very simple mean-matching method. At the kth step, $k = 1, \cdots, N$, choose the G_2 subject such that the mean of the current G_{2*} sample of k subjects is closest to $\bar{x}_{1\cdot}$. Thus, at step 1 choose the G_2 subject closest to $\bar{x}_{1\cdot}$; at step 2 choose the G_2 subject such that the average of the first G_{2*} subject and the additional G_2 is closest to $\bar{x}_{1\cdot}$; continue until N G_2 subjects are chosen.

In Table 4.1 we present Monte Carlo values of Θ_{MN}, the percent reduction in bias for this simple mean-matching method.[5] We assume X normal, $B = \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1$; $\sigma_1^2/\sigma_2^2 = \frac{1}{2}, 1, 2$; $N = 25, 50, 100$; and $r = 2, 3, 4$. Some limited experience indicates that these values are typical of those that might occur in practice. In addition, values of r and N were chosen with the results of

[5] The standard errors for all Monte Carlo values given in Tables 4.1, 5.1, 5.2, and 5.3 are generally less than 0.5% and rarely greater than 1%.

TABLE 4.1

Θ_{MN}: PERCENT REDUCTION IN BIAS FOR A SIMPLE MEAN-MATCHING METHOD, X NORMAL

		$\sigma_1^2/\sigma_2^2 = \frac{1}{2}$				$\sigma_1^2/\sigma_2^2 = 1$				$\sigma_1^2/\sigma_2^2 = 2$			
	B =	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{3}{4}$	1	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{3}{4}$	1	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{3}{4}$	1
r =													
N = 25	2	99	98	95	84	99	97	89	77	97	90	75	63
	3	100	100	100	98	100	100	98	93	100	98	94	81
	4	100	100	100	100	100	100	100	98	100	100	97	91
N = 50	2	100	100	98	87	100	99	91	77	100	95	82	67
	3	100	100	100	100	100	100	99	96	100	100	97	84
	4	100	100	100	100	100	100	100	100	100	100	100	95
N = 100	2	100	100	100	88	100	100	96	80	100	98	85	64
	3	100	100	100	100	100	100	100	98	100	100	99	87
	4	100	100	100	100	100	100	100	100	100	100	100	96

Tables 3.1 and 3.2 in mind—values for percent reduction in bias may be moderately applicable for nonnormal distributions, especially when $r = 2$, and values given when $N = 100$ may be quite representative for $N > 100$. Θ_{MN} exhibits the four trends given in section 3 for $\Theta_{\max}$.

(1) Given fixed N, r, and σ_1^2/σ_2^2, Θ_{MN} decreases as B increases.
(2) Given fixed N, r, and B, Θ_{MN} decreases as σ_1^2/σ_2^2 increases.
(3) Given fixed B, σ_1^2/σ_2^2 and N, Θ_{MN} increases as r increases.
(4) Given fixed B, σ_1^2/σ_2^2 and r, except for one value (67% for $N = 50$, $\sigma_1^2/\sigma_2^2 = 2$, $r = 2$, $B = 1$) Θ_{MN} increases as N increases.

In Table 4.2 we present values of min $\{100, \Theta_{\max}\}$ for the same range of N, B and σ_1^2/σ_2^2 as in Table 4.1. Note first that the 67% for $N = 50$, $\sigma_1^2/\sigma_2^2 = 2$, $r = 2$, $B = 1$ mentioned above is larger than the theoretical maximum and thus suspect. Comparing the corresponding entries in Table 4.1 and Table 4.2 we see that the values for $N = 100$ always attain at least 96% of min $\{100, \Theta_{\max}\}$, while the values for $N = 50$ always attain at least 91% of min $\{100, \Theta_{\max}\}$, and those for $N = 25$ always attain at least 87% of min $\{100, \Theta_{\max}\}$. Hence this simple method appears to be a very reasonable mean-matching method, especially for large samples.

5. PAIR-MATCHING

Even though a simple mean-matching method can be quite successful at removing the bias of X, matched samples are generally not mean-matched. Usually matched samples are "individually" (Greenwood [1945]), "precision" (Chapin [1947]), or "pair" (Cochran [1953]) matched, subject by subject. The main reason is probably some intuitive feeling on the part of investigators that pair-matched samples are superior. One theoretical justification is that $\hat{\tau}_0$ based on exactly mean-matched samples has zero expected bias only if the P_2 response surface really is linear, while $\hat{\tau}_0$ based on exactly pair-matched samples has zero expected bias no matter what the form of the response surface. Since an investigator rarely knows for sure that the P_2 response surface is linear, if the choice is between exactly pair-matched samples and exactly mean-matched samples of the same size, obviously he would choose the exactly pair-matched samples.

The ease of constructing confidence limits and tests of significance is a second reason for using a pair-matching method rather than a mean-matching method. Significance tests and confidence limits that take advantage of the increased precision in matched samples are easily constructed with pair-matched data by using matched pair differences, while such tests and limits for mean-matched data must be obtained by an analysis of covariance (Greenberg [1953]).

Another reason for the use of pair-matching methods is that each matched pair could be considered a study in itself. Thus, the investigator might assume the response surfaces are nonparallel and use the difference $y_{1i} - y_{2i}$ to estimate the response surface difference at x_{1i}. It follows from (2.2) that the

TABLE 4.2

MIN {100, θ_{max}}; X NORMAL

		$\sigma_1^2/\sigma_2^2 = \frac{1}{2}$			$\sigma_1^2/\sigma_2^2 = 1$				$\sigma_1^2/\sigma_2^2 = 2$			
	B =	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{3}{4}$	1	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{3}{4}$	1	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{3}{4}$	1
	r =												
N = 25	2	118	105	95	82	98	87	76	67	67	62	57	48
	3	106	103	101	97	99	94	88	81	80	73	68	62
	4	103	102	101	99	98	97	93	89	83	80	76	68
N = 50	2	113	107	99	86	100	91	79	69	69	60	54	51
	3	103	102	102	99	100	97	92	85	81	75	69	63
	4	102	101	101	100	100	98	95	89	87	82	76	71
N = 100	2	108	106	101	86	100	92	81	69	69	60	55	49
	3	102	101	102	100	101	98	92	86	82	76	70	64
	4	101	101	101	100	99	99	96	90	85	81	76	71

bias of $y_{1i} - y_{2i}$ for estimating $R_1(x_{1i}) - R_2(x_{1i})$ is $R_2(x_{1i}) - R_2(x_{2i})$. Assuming this bias to be some unknown increasing function of $|x_{1i} - x_{2i}|$, one minimizes the bias of each estimate, $y_{1i} - y_{2i}$, by minimizing each $|x_{1i} - x_{2i}|$ rather than $|\bar{x}_{1.} - \bar{x}_{2.}|$.

If each G_1 subject is closest to a different G_2 subject, assigning matches to minimize each $|x_{1i} - x_{2i}|$ is easily done. However, if two or more G_1 subjects are closest to the same G_2 subject, the best way to assign individual matches is not obvious, unless the investigator decides upon some criterion to be minimized, such as one proportional to the average squared bias of the N individual estimates assuming parallel linear response surfaces, $1/N \sum (x_{1i} - x_{2i})^2$. As was already mentioned, in order to find the G_{2*} sample that minimizes any such quantity, some rather time consuming algorithm designed to solve the transportation problem must be used.

Even though more complex pair-matching methods often may be superior, we will investigate three simple "nearest available" pair-matching methods. A nearest available pair-matching method assigns the closest match for $g_1 \in G_1$ from the yet unmatched G_2 subjects and thus is completely defined if the order for matching the G_1 subjects is specified. The three orderings of the G_1 subjects to be considered here are: (1) the subjects are randomly ordered (random), (2) the subject not yet matched with the lowest score on X is matched next (low-high), and (3) the subject not yet matched with the highest score on X is matched next (high-low). The results will depend on our assumption that $\eta_1 > \eta_2$, for if η_1 were less than η_2 , the values for the low-high and high-low pair-matching methods would be interchanged.

In Tables 5.1, 5.2, and 5.3 we present Monte Carlo values for the percent reduction in bias for random ordering (Θ_{RD}), low-high ordering (Θ_{LH}) and high-low ordering (Θ_{HL}). We assume the same range of conditions as given in Table 4.1 for Θ_{MN}. Θ_{RD} and Θ_{HL} exhibit the four trends given in section 3 for $\Theta_{\max}$ and exhibited in Table 4.1 for Θ_{MN}.

(1) Given fixed N, r, and σ_1^2/σ_2^2 , Θ_{RD} and Θ_{HL} decrease as B increases.
(2) Given fixed N, r, and B, Θ_{RD} and Θ_{HL} decrease as σ_1^2/σ_2^2 increases.
(3) Given fixed B, σ_1^2/σ_2^2 , and N, Θ_{RD} and Θ_{HL} increase as r increases.
(4) Given fixed B, σ_1^2/σ_2^2 and r, Θ_{RD} and Θ_{HL} generally increase as N increases.

These same four trends hold for all orderings if "Θ_{RD} and Θ_{HL} increase" is replaced by "Θ_{RD} , Θ_{HL} , and Θ_{LH} get closer to 100%". Values of Θ greater than 100% indicate that $\eta_{2*} > \eta_1$ which is of course not as desirable as $\eta_{2*} \simeq \eta_1$ which implies $\Theta \simeq 100$.

Comparing across Tables 5.1, 5.2, and 5.3 we see that given fixed B, σ_1^2/σ_2^2 , r and N, $\Theta_{LH} \geq \Theta_{RD} \geq \Theta_{HL}$. This result is not surprising for the following reason. The high-low ordering will have a tendency not to use those G_2 subjects with scores above the highest G_1 score while the low-high ordering will have a tendency not to use those G_2 subjects with scores below the lowest G_1 scores. Since we are assuming $B > 0$ ($\eta_1 > \eta_2$), the low-high

TABLE 5.1

θ_{RN}: PERCENT REDUCTION IN BIAS FOR RANDOM ORDER, NEAREST AVAILABLE MATCHING; X NORMAL

		$\sigma_1^2/\sigma_2^2 = \frac{1}{2}$				$\sigma_1^2/\sigma_2^2 = 1$				$\sigma_1^2/\sigma_1^2 = 2$			
	B =	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{3}{4}$	1	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{3}{4}$	1	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{3}{4}$	1
	r =												
N = 25	2	100	100	100	91	100	100	100	78	100	100	85	64
	3	100	100	100	100	100	100	100	100	100	100	100	88
	4	100	100	100	100	100	100	100	100	100	100	100	100
N = 50	2	100	100	100	91	100	100	100	79	100	100	87	65
	3	100	100	100	100	100	100	100	100	100	100	100	89
	4	100	100	100	100	100	100	100	100	100	100	100	100
N = 100	2	100	100	100	92	100	100	100	80	100	100	87	65
	3	100	100	100	100	100	100	100	100	100	100	100	89
	4	100	100	100	100	100	100	100	100	100	100	100	100

TABLE 5.2

θ_{HL}: PERCENT REDUCTION IN BIAS FOR HIGH-LOW ORDER, NEAREST AVAILABLE MATCHING; X NORMAL

		$\sigma_1^2/\sigma_2^2 = \frac{1}{2}$				$\sigma_1^2/\sigma_2^2 = 1$				$\sigma_1^2/\sigma_2^2 = 2$			
	B =	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{3}{4}$	1	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{3}{4}$	1	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{3}{4}$	1
	r =												
N = 25	2	97	94	89	80	87	82	75	66	63	60	56	48
	3	99	98	97	93	94	91	86	81	77	72	67	61
	4	99	99	99	97	95	95	92	88	81	79	76	68
N = 50	2	99	98	93	84	92	87	78	69	66	59	53	51
	3	100	99	99	97	96	95	91	84	79	75	69	63
	4	100	100	100	99	98	97	94	89	86	81	75	71
N = 100	2	100	99	96	86	95	90	81	69	67	59	55	49
	3	100	100	99	98	99	96	91	86	81	75	70	64
	4	100	100	100	99	99	98	96	90	85	81	76	71

TABLE 5.3

Θ_{LH}: PERCENT REDUCTION IN BIAS FOR LOW-HIGH ORDER, NEAREST AVAILABLE MATCHING; X NORMAL

		$\sigma_1^2/\sigma_2^2 = \frac{1}{2}$				$\sigma_1^2/\sigma_2^2 = 1$				$\sigma_1^2/\sigma_2^2 = 2$			
	B =	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{3}{4}$	1	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{3}{4}$	1	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{3}{4}$	1
	r =												
N = 25	2	78	83	81	75	77	77	72	65	59	58	55	48
	3	92	94	93	90	89	88	84	79	75	71	67	61
	4	96	97	97	95	93	93	90	87	79	78	75	68
N = 50	2	86	90	86	79	85	84	76	68	63	58	53	51
	3	96	97	96	94	93	93	89	84	77	74	69	63
	4	98	99	98	97	96	96	93	88	84	81	75	71
N = 100	2	93	94	90	82	90	87	79	69	65	59	55	49
	3	98	98	98	96	96	95	90	85	80	75	70	64
	4	99	99	99	98	98	97	95	90	84	81	76	71

ordering should yield the most positive $\bar{x}_{2\cdot}$ followed by the random ordering and then the high-low ordering. When $\sigma_1^2/\sigma_2^2 = \frac{1}{2}$ and $B \leq \frac{1}{2}$, Θ_{LH} can be somewhat greater than 100 (e.g. 113) while $100 \geq \Theta_{RD} \geq 94$. In all other cases ($\sigma_1^2/\sigma_2^2 > \frac{1}{2}$ or $\sigma_1^2/\sigma_2^2 = \frac{1}{2}$ and $B \leq \frac{1}{2}$), Θ_{LH} is closer to 100% than Θ_{RD} or Θ_{HL}. In general the results for Θ_{RD}, Θ_{LH}, and Θ_{HL} are quite similar for the conditions considered.

Comparing the results in this section with those in section 4, it is easily checked that if $\sigma_1^2/\sigma_2^2 \leq 1$ the three pair-matching methods generally attain more than 85% of min $\{100, \Theta_{\max}\}$ in Table 4.2 indicating that they can be reasonable methods of matching the means of the samples. However, if $\sigma_1^2/\sigma_2^2 = 2$, the pair-matching methods often attain less than 70% of the corresponding Θ_{MN} in Table 4.1 indicating that when $\sigma_1^2/\sigma_2^2 > 1$ these pair-matching methods do not match the means very well compared to a single mean-matching method.

Remembering that pair-matching methods implicitly sacrifice closely matched means for good individual matches, we also calculated a measure of the quality of the individual matches. These results presented in Appendix C indicate that, in general, the high-low ordering yields the closest individual matches followed by the random ordering. This conclusion is consistent with the intuition to match the most difficult subjects first in order to obtain close individual matches.

6. ADVICE TO AN INVESTIGATOR

In review, we assume there are two populations, P_1 and P_2, defined by two levels of a treatment variable. There is a sample, G_1, of size N from P_1 and a larger sample, G_2, of size rN from P_2, both of which have recorded scores on the matching variable X. The objective of the study is to estimate τ, the average effect of the treatment variable on a dependent variable Y over the P_1 population. We assume that $\hat{\tau}_0 = \bar{y}_{1\cdot} - \bar{y}_{2\cdot}$ will be used to estimate τ where $\bar{y}_{1\cdot}$ is the average Y in the G_1 sample and $\bar{y}_{2\cdot}$ is the average Y in an N-size subsample of G_2 matched to G_1, G_{2*}.

Depending upon the particular study, the investigator may be able, within limits, to control three "parameters".

(a) N, the size of the smaller initial sample (G_1); equivalently, the size of each of the final samples.
(b) r, the ratio of the sizes of the larger initial sample (G_2) and the smaller initial sample (G_1).
(c) The matching rule used to obtain the G_{2*} sample of size N from the G_2 sample of size rN.

Below we present advice for choosing these "parameters" in the order first N, then r, and then the matching method.

(a) Choosing N

We will use a standard method for estimating N (Cochran [1963]) which

assumes that the investigator wants $\hat{\tau}_0$ to be within $\pm\Delta$ of τ with probability $1 - \alpha$: Prob $\{|\hat{\tau}_0 - \tau| > \Delta\} = \alpha$. Letting $s/\sqrt{N}$ be the estimated standard error of $\hat{\tau}_0$ we would choose

$$N = z^2 s^2/\Delta^2, \tag{6.1}$$

where z is the standard normal deviate corresponding to $1 - \alpha$ confidence limits (e.g. if $\alpha = 0.05$, $z \simeq 2$).[6] In order to use (6.1) we must have an estimate of the standard error of $\hat{\tau}_0$, $s/\sqrt{N}$.

Suppose that the response surfaces are linear with slopes β_1 and β_2 and that $\bar{x}_{2\cdot}$ will be exactly matched to $\bar{x}_{1\cdot}$ in the final samples by using one of the matching methods discussed in sections 4 and 5. Thus, $E\, E_c(\hat{\tau}_0) = \tau$, and it is easy to show that

$$\begin{aligned} s^2/N &= E\, E_c(\hat{\tau}_0 - \tau)^2 \\ &= E\, E_c[\beta_2(\eta_1 - \bar{x}_{2\cdot}) + \beta_1(\bar{x}_{1\cdot} - \eta_1) + \bar{e}_{1\cdot} - \bar{e}_{2\cdot}]^2. \end{aligned} \tag{6.2}$$

Setting $\bar{x}_{2\cdot} = \bar{x}_{1\cdot}$ and assuming the usual independent error model where $E_c(e_{ij}^2) = \sigma_{e_i}^2$, $i = 1, 2$, (6.2) becomes

$$s^2/N = \frac{\sigma_{e_1}^2}{N} + \frac{\sigma_{e_2}^2}{N} + \frac{\sigma_1^2}{N}(\beta_1 - \beta_2)^2. \tag{6.3}$$

Rarely in practice can one estimate the quantities in (6.3). Generally, however, the investigator has some rough estimate of an average variance of Y, say $\hat{\sigma}_y^2$, and of an average correlation between Y and X, say $\hat{\rho}$. Using these he can approximate $(1/N)\{\sigma_{e_1}^2 + \sigma_{e_2}^2\}$ by $(2/N)\hat{\sigma}_y^2(1 - \hat{\rho}^2)$.

Approximating $\sigma_1^2/N(\beta_1 - \beta_2)^2$ is quite difficult unless one has estimates of β_1 and β_2 . The following rough method may be useful when the response surfaces are at the worst moderately nonparallel. If the response surfaces are parallel $\sigma_1^2/N(\beta_1 - \beta_2)^2$ is zero and thus minimal. If the response surfaces are at most moderately nonparallel, one could assume $(\beta_1 - \beta_2)^2 \leq 2\beta_1^2$ in most uses.[7] Hence, in many practical situations one may find that $0 \leq \sigma_1^2/N(\beta_1 - \beta_2)^2 \leq 2(\sigma_1^2/N)\beta_1^2$, where the upper bound can be approximated by $2(\hat{\rho}^2\hat{\sigma}_y^2/N)$. Hence, a simple estimated range for s^2 is

$$2\hat{\sigma}_y^2(1 - \hat{\rho}^2) < s^2 < 2\hat{\sigma}_y^2 . \tag{6.4}$$

If the investigator believes that the response surfaces are parallel and linear, the value of s^2 to be used in (6.1) can be chosen to be near the minimum of this interval. Otherwise, a value of s^2 nearer the maximum would be appropriate.

(b) Choosing r

First assume that mean-matching is appropriate, i.e. assume an essentially linear response surface in P_2 , and that the sole objective is to estimate τ.

[6] Moderate samples ($N > 20$) are assumed. For small samples $N = t_{N-1}^2 s^2/\Delta^2$ where t_{N-1} is the student-deviate with $N - 1$ degrees of freedom corresponding to $1 - \alpha$ confidence limits.

[7] A less conservative assumption is $(\beta_1 - \beta_2)^2 \leq \beta_1^2$.

We will choose r large enough to expect 100% reduction in bias using the simple mean-matching method of section 4.

(1) Estimate $\gamma = B[(1 + \sigma_1^2/\sigma_2^2)/2]^{1/2}$ and the approximate shape of the distribution of X in P_2 . In order to compensate for the decreased ability of the mean-matching method to attain the theoretical maximum reduction in bias in small or moderate samples (see section 4), if N is small or moderate ($N \leq 100$) increase γ by 5 to 15% (e.g. 10% for $N = 50$, 5% for $N = 100$).

(2) Using Table 3.1 find the row corresponding to the approximate shape of the distribution of X in P_2 . Now find approximate values of r_1 and r_∞ such that $\Omega(r_1 , 1) \simeq \gamma$ and $\Omega(r_\infty , \infty) \simeq \gamma$. If N is very small ($N < 5$), r should be chosen to be close to r_1 ; otherwise, results in Table 3.2 suggest that r can be chosen to be much closer to r_∞ . r should probably be chosen to be greater than two and in most practical applications will be less than four.

Now assume pair-matches are desired, i.e. the response surfaces may be nonlinear, nonparallel and each $y_{1i} - y_{2i}$ may be used to estimate the treatment effect at x_{1i} . We will choose r large enough to expect 95% + reduction in bias using the random order-nearest available pair-matching method of section 5. Perform steps (1) and (2) as above for mean-matching. However, since in section 5 we found that if $\sigma_1^2/\sigma_2^2 > 1$ nearest available pair-matching did not match the means of the samples very well compared to the simple mean-matching method, r should be increased. The following is a rough estimate (based on Tables 5.1 and 4.1) of the necessary increase:

if $\sigma_1^2/\sigma_2^2 = \frac{1}{2}$, r remains unchanged
if $\sigma_1^2/\sigma_2^2 = 1$, increase r by about 50%
if $\sigma_1^2/\sigma_2^2 = 2$, at least double r.

(c) Choosing a Matching Method

We assume G_1 and G_2 (i.e. r and N) are fixed and the choice is one of a matching method. If the investigator knows the P_2 response surface is linear and wants only to estimate τ, the results in section 4 suggest that he can use the simple mean-matching method described in section 4 and be confident in many practical situations of removing most of the bias whenever $r > 2$.

If confidence in the linearity of the P_2 response surface is lacking and/or the investigator wants to use each matched pair to estimate the effect of the treatment variable at a particular value of X, he would want to obtain close individual matches as well as closely matched means. Results in section 5 indicate that in many practical situations the random order nearest available pair-matching method can be used to remove a large proportion of the bias in X while assigning close individual matches. The random order nearest available pair-matching is extremely easy to perform since the G_1 subjects do not have to be ordered; yet, it does not appear to be inferior to either high-low or low-high orderings and thus seems to be a reasonable choice in practice.

If a computer is available, a matching often superior to that obtained with the simple mean-matching or one random order nearest available

pair-matching may be easily obtained by performing the simple mean-matching and several nearest available pair-matchings (i.e. several random orderings, low-high ordering, high-low ordering) and choosing the "best" matching. There should be no great expense in performing several matchings. Using Fortran IV subroutines given in Appendix B for the simple mean-matching method and nearest available pair-matching methods, a matching of 100 G_1 subjects from 400 G_2 subjects takes about $1\frac{1}{2}$ seconds on an IBM 360/65.

In order to decide which matching is "best", record for all matched samples $\bar{d} = \bar{x}_{1.} - \bar{x}_{2.}$ and $\bar{d}^2 = 1/N \sum (x_{1i} - x_{2i})^2$. Pair-matches (and thus $\bar{d}^2$) for the mean-matched sample can be found by using a nearest available pair-matching method on the final samples. If several matchings give equally small values of $\bar{d}$, choose the matching that gives the smallest value of $\bar{d}^2$. If $\bar{d}$ for one matched sample is substantially smaller than for any of the other matched samples but $\bar{d}^2$ for that sample is quite large, the investigator must either (1) make a practical judgement as to whether closely matched means or close individual matches are more important for his study, or (2) attempt to find matches by a matching method more complex than the ones considered here.

Admittedly, the practical situations and methods of estimating τ covered above are quite limited. The following article extends this work to include regression (covariance) adjusted estimates of τ and nonlinear parallel response surfaces. Rubin [1970] includes extensions to the case of many matching variables. Althauser and Rubin [1970] give a nontechnical discussion of some problems that arise with many matching variables.

ACKNOWLEDGMENTS

This work was supported by the Office of Naval Research under contract N00014-67A-0298-0017, NR-042-097 at the Department of Statistics, Harvard University.

I wish to thank Professor William G. Cochran for many helpful suggestions and criticisms on earlier drafts of this article. I would also like to thank the referees for their helpful comments.

CROISEMENTS EN VUE D'EFFACER LES BIAIS DANS DES ETUDES

RESUME

Plusieurs méthodes de croisements qui croisent tous les niveaux d'un même échantillon tiré d'un autre échantillon plus grand sur une variable continue de croisement sont comparées en ce qui concerne leur pouvoir d'effacer le biais de la variable de croisement. Une des méthodes est une méthode croisant la moyenne et trois sont des méthodes croisant la paire la plus proche. Les possibilités des méthodes en ce qui concerne la suppression du biais sont aussi comparées au maximum theorique étant donné des distributions fixées et des tailles d'échantillon fixées. Un résumé pour aider le chercheur est inclus.

REFERENCES

Althauser, R. P. and Rubin, D. B. [1970]. The computerized construction of a matched sample. *Amer. J. Soc. 76*, 325–46.

Belsen, W. A. [1956]. A technique for studying the effects of a television broadcast. *Appl. Statist. 5*, 195–202.

Billewicz, W. Z. [1965]. The efficiency of matched samples: an empirical investigation. *Biometrics 21*, 623–43.

Chapin, F. S. [1947]. *Experimental Designs in Sociological Research.* Harper and Brothers, New York.

Cochran, W. G. [1953]. Matching in analytical studies. *Amer. J. Pub. Health 43*, 684–91.

Cochran, W. G. [1963]. *Sampling Techniques.* Wiley, New York.

Cochran, W. G. [1968]. The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics 24*, 295–313.

Cochran, W. G. [1969]. The use of covariance in observational studies. *Appl. Statist. 18*, 270–5.

Cox, D. R. [1957]. The use of a concomitant variable in selecting an experimental design. *Biometrics*, 150–8.

Greenberg, B. G. [1953]. The use of covariance and balancing in analytical surveys. *Amer. J. Pub. Health 43*, 692–9.

Greenberg, E. [1945]. *Experimental Sociology: A Study in Method.* Kings Crown Press, New York.

Harter, H. L. [1960]. Expected values of normal order statistics. Aeronautical Research Laboratories Technical Report, 60–292.

Kihlberg, J. K. and Robinson, S. J. [1968]. Seat belt use and injury patterns in automobile accidents. Cornell Aeronautical Laboratory Report No. VJ-1823-R30.

Peters, C. C. and Van Voorhis, W. R. [1940]. *Statistical Procedures and Their Mathematical Bases.* McGraw-Hill, New York.

Rubin, D. B. [1970]. The use of matched sampling and regression adjustment to observational studies. Statistics Department Report CP-4, Harvard University.

Sarhan, A. E. and Greenberg, B. G. [1962]. *Contributions to Order Statistics.* Wiley, New York.

Yinger, J., Milton, I. K., and Laycock, F. [1967]. Treating matching as a variable in a sociological experiment. *Amer. Soc. Review 32*, 801–12.

Wilks, S. S. [1932]. On the distributions of statistics in samples from a normal population of two variables with matched sampling of one variable. *Metron 9*, 87–126.

APPENDIX A: PROOFS OF TRENDS (3) AND (4) IN SECTION 3

We prove the intuitively obvious trend (3) by considering a random sample of size $(a + r)N$ from f. Call the order statistics $x_{(1)}, \cdots, x_{(N)}, \cdots, x_{(a+r)N}$ where $x_{(1)}$ is the largest observation. The average of the N largest observations from these $(a + r)N$ is $1/N \sum_1^N x_{(i)}$. By randomly discarding aN of the original observations, we have a random sample of size rN from f. But in any such subset the average of the N largest observations is less than or equal to $1/N \sum_1^N x_{(i)}$. Averaging over repeated random samples we have that $\Omega(r, N) \leq \Omega(r + a, N)$, N, rN, aN positive integers.

We prove trend (4) by a similar but more involved argument. Consider a random sample of size $r(N + b)$ and let $x_{(1)}, \cdots, x_{(N)}, \cdots, x_{(N+b)}, \cdots, x_{(r(N+b))}$ be the order statistics. The average of the $N + b$ largest from these $r(N + b)$ is $1/(N + b) \sum_{i=1}^{N+b} x_{(i)}$. Choosing a random rN-size subset of these

observations, we have that the expected value of the average of the N largest from such a subset is

$$\frac{1}{n}\sum_{S\in\mathbf{S}}\frac{1}{N}\text{ (total of largest } N \text{ observations from } S)$$

where

$\mathbf{S}$ = set of all distinct rN size subsets of original $r(N+b)$

$n = \binom{r(N+b)}{rN}$ = the number of elements in $\mathbf{S}$.

This expression can be rewritten as

$$\frac{1}{n}\frac{1}{N}\sum_{i=1}^{r(N+b)}\lambda_i x_{(i)}$$

where λ_i = number of elements of $\mathbf{S}$ in which $x_{(i)}$ is one of the N largest observations, $\sum\lambda_i = Nn$.

For $i = 1, \cdots, N$, λ_i = the number of subsets in which $x_{(i)}$ occurs $= m = \binom{r(N+b)-1}{rN-1}$. For all $i > N$, $\lambda_i \leq m$. Consider the above summation as a weighted sum of the $x_{(i)}$ where the weights are ≥ 0 and add to 1 ($\sum\lambda_i/nN = 1$). Increasing the weights on the largest $x_{(i)}$ while keeping the sum of the weights the same cannot decrease the total value of the sum. Thus,

$$\begin{aligned}\frac{1}{n}\frac{1}{N}\sum_{1}^{r(N+b)}\lambda_i x_{(i)} &\leq \frac{1}{n}\frac{1}{N}\left\{m\sum_{1}^{N+b-1}x_{(i)} + (nN - m(N+b-1))x_{(N+b)}\right\}\\ &\leq \frac{m}{nN}\left\{\sum_{1}^{N+b-1}x_{(i)} + \left(\frac{nN}{m} - (N+b-1)\right)x_{(N+b)}\right\}\\ &\leq \frac{1}{N+b}\left\{\sum_{1}^{N+b}x_{(i)}\right\}\end{aligned}$$

since $(m/nN) = 1/(N+b)$.

Hence, the expected average of the top N from a random rN-size subset is less than or equal to the average of the top $b+N$ from the original $r(b+N)$; thus averaging over repeated random samples we have

$$\Omega(r, N) \leq \Omega(r, b+N),\ N,\ rN,\ r(b+N) \text{ positive integers.}$$

APPENDIX B

FORTRAN SUBROUTINES FOR NEAREST AVAILABLE PAIR MATCHING AND SIMPLE MEAN MATCHING

Notation used in the subroutines

$N1 = N$ = size of G_1

$N2$ = rN = size of initial G_2

$X1$ = vector of length N giving matching variable scores for G_1 sample, i.e. 1st entry is first G_1 subject's score

$X2$ = vector of length rN giving scores for G_2 on matching variable

$AV1$ = $\bar{x}_{1.}$

D = $\bar{x}_{1.} - \bar{x}_{2.}$; output for matched samples

$D2$ = $1/N \sum (x_{1i} - x_{2i})^2$; output for matched samples

$IG1$ = vector giving ordering of G_1 sample for nearest available matching (a permutation of $1 \cdots N1$)

$IG2$ = "current" ordering of G_2 sample. After each call to a matching

```
      SUBROUTINE NAMTCH(D,D2,IG2, IG1,N1,N2,X1,X2)
C     SUBROUTINE TO PERFORM NEAREST AVAILABLE MATCHING
C     NECESSARY INPUTS ARE IG2, IG1, N1, N2, X1, X2
      DIMENSION IG1(1),IG2(1),X1(1),X2(1)
      D=0.
      D2=0.
      DO 200 I=1,N1
      K=IG1(I)
  200 CALL MATCH(D,D2,IG2,X1(K),I,N2,X2)
      D=D/FLOAT(N1)
      D2=D2/FLOAT(N1)
      RETURN
      END


      SUBROUTINE MNMTCH(D,IG2,N1,N2,X2,AV1)
C     SUBROUTINE TO PERFORM SIMPLE MEAN MATCHING TO AV1
C     NECESSARY INPUTS ARE IG2, N1, N2, X2, AV1
      DIMENSION IG2(1),  X2(1)
      D=0.
      D2=0.
      DO 200 I=1,N1
      XX=AV1+D
  200 CALL MATCH(D,D2,IG2, XX,I,N2,X2)
      D=D/FLOAT(N1)
      RETURN
      END


      SUBROUTINE MATCH(D,D2,IG2,X1,K1,N2,X2)
C     SUBROUTINE PICKS G2 SUBJECT BETWEEN (INCLUSIVE) K1 AND N2 IN LIST
C     IG2 WHO HAS SCORE (IN X2) CLOSEST TO VALUE X1
C     HIS SUBJECT NUMBER IS PUT IN IG2(K1) AND PREVIOUS ENTRY IN IG2(K1)
C     IS MOVED BEYOND K1 ENTRY
      DIMENSION X2(1),IG2(1)
      LL=IG2(N2)
      IF (K1 .EQ. N2)  GO TO 410
      DMIN=ABS(X1-X2(LL))
      K2=N2-K1
      DO 400 LK=1,K2
      K=N2-LK
      L=IG2(K)
      IF (ABS(X1-X2(L)) .LT. DMIN)  GO TO 300
      IG2(K+1)=L
      GO TO 400
  300 IG2(K+1)=LL
      LL=L
      DMIN=ABS(X1-X2(LL))
  400 CONTINUE
  410 CONTINUE
      IG2(K1)=LL
      D=D+X1-X2(LL)
      D2=D2+(X1-X2(LL))**2
      RETURN
      END
```

subroutine, the G_1 subject having subject number $IG1(K)$ is matched to G_2 having subject number $IG2(K)$, $K = 1, \cdots, N1$. Subject number, of course, refers to order in vectors $X1$ and $X2$. Before the first call to a matching subroutine one should set $IG2(K) = K$, $K = 1, \cdots, N2$. After this initialization $IG2$ should be considered output of matching routines.

APPENDIX C

THE QUALITY OF INDIVIDUAL MATCHES:

$$100 \times \frac{E \sum (x_{1j} - x_{2j})^2/N}{(B^2/2 + 1)(\sigma_1 + \sigma_2)}, \; X \text{ NORMAL}^*$$

PAIR MATCHING METHOD

				$\sigma_1^2/\sigma_2^2 = \frac{1}{2}$				$\sigma_1^2/\sigma_2^2 = 1$				$\sigma_1^2/\sigma_2^2 = 2$			
			B =	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{3}{4}$	1	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{3}{4}$	1	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{3}{4}$	1
Random Order	N=25	r=1		47	58	72	81	35	50	63	77	35	48	63	76
		2		01	02	04	10	02	05	10	16	10	13	19	28
		3		00	00	01	02	01	02	05	08	06	09	13	19
		4		00	00	00	01	01	01	03	04	04	06	09	15
	N=50	1		42	55	70	80	30	47	64	76	31	42	57	73
		2		00	01	03	08	01	03	08	15	07	13	20	26
		3		00	00	00	01	00	01	03	06	04	08	12	18
		4		00	00	00	00	00	01	02	04	03	05	09	13
	N=100	1		37	56	69	80	25	46	63	76	25	41	60	73
		2		00	00	02	06	01	02	07	14	06	12	19	27
		3		00	00	00	01	00	01	02	05	03	07	11	17
		4		00	00	00	00	00	00	01	04	03	05	09	13
High-Low Order	N=25	1		55	42	42	47	30	26	31	40	27	28	34	45
		2		01	01	02	04	02	03	05	09	08	09	13	19
		3		00	00	01	01	01	01	03	05	05	06	09	13
		4		00	00	00	01	01	01	02	03	04	05	07	11
	N=50	1		51	39	38	43	21	19	27	37	20	22	29	40
		2		00	01	01	03	01	02	04	07	05	08	12	16
		3		00	00	00	01	00	01	02	03	03	05	08	11
		4		00	00	00	00	00	00	01	02	02	03	06	09
	N=100	1		49	39	35	41	15	16	24	35	16	19	28	39
		2		00	00	01	02	00	01	03	06	04	07	11	17
		3		00	00	00	00	00	00	01	02	02	03	07	11
		4		00	00	00	00	00	00	01	02	02	03	05	08
Low-High Order	N=25	1		93	108	122	126	71	92	108	119	55	77	96	110
		2		02	04	10	20	04	09	17	26	13	19	27	39
		3		00	01	02	05	01	03	08	13	07	12	18	26
		4		00	00	01	02	01	02	04	07	05	08	12	20
	N=50	1		95	110	122	127	69	92	110	120	52	71	91	108
		2		00	02	08	18	02	07	16	25	11	19	29	36
		3		00	00	01	03	01	02	05	11	06	11	18	24
		4		00	00	00	01	00	01	03	07	04	07	13	19
	N=100	1		92	111	122	128	62	91	109	120	44	69	92	109
		2		00	01	06	17	01	06	14	25	10	19	28	38
		3		00	00	00	02	00	01	05	10	05	10	17	24
		4		00	00	00	01	00	01	02	07	04	07	13	18

*If perfectly matched, equals 00. If randomly matched from random samples, equals 100.

Received January 1971, *Revised July* 1972

Key Words: Matching; Matched sampling; Observational studies; Quasi-experimental studies; Controlling bias; Removing bias; Blocking.