Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies

Author(s): Donald B. Rubin

Source: *Journal of the American Statistical Association*, Vol. 74, No. 366 (Jun., 1979), pp. 318-328

Published by: Taylor & Francis, Ltd. on behalf of the American Statistical Association

Stable URL: https://www.jstor.org/stable/2286330

Accessed: 22-06-2025 00:20 UTC

Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies

DONALD B. RUBIN*

Monte Carlo methods are used to study the efficacy of multivariate matched sampling and regression adjustment for controlling bias due to specific matching variables **X** when dependent variables are moderately nonlinear in **X**. The general conclusion is that nearest available Mahalanobis metric matching in combination with regression adjustment on matched pair differences is a highly effective plan for controlling bias due to **X**.

KEY WORDS: Covariance adjustment; Nonrandomized studies; Quasi-experiments.

1. INTRODUCTION

Our objective is to study the utility of matched sampling and regression adjustment (covariance adjustment) for controlling specific matching variables in observational studies. This introduction is brief; we assume that the reader is familiar with the literature on matching and covariance adjustment in observational studies (e.g., Althauser and Rubin 1970); Billewicz 1964, 1965; Campbell and Erlebacher 1970; Cochran 1953, 1968; Cochran and Rubin 1973; Gilbert, Light, and Mosteller 1975; Greenberg 1953; Lord 1960; McKinlay 1974, 1975a,b; and Rubin 1974, 1977, 1978a). In particular, this work is a natural extension of earlier Monte Carlo work on one matching variable (Rubin 1973a,b) and theoretical work on multivariate matching methods (Rubin 1976a,b).

Matched sampling refers to the selection of treatment units (e.g., smokers) and control units (e.g., nonsmokers) that have similar values of matching variables, **X** (e.g., age, weight), whereas regression adjustment refers to a statistical procedure that adjusts estimates of the treatment effects by estimating the relationship between the dependent variable Y (e.g., blood pressure) and **X** in each treatment group. Hence, matched sampling and regression adjustment may be used alone or in combination, that is, samples may be random or matched, and regression adjustment may or may not be performed. Our use of the term *matching* excludes methods that discard units with Y recorded; thus, our matching methods should be thought of as choosing units on which to record Y when Y can be recorded only on a limited number of units (e.g., Y is obtained by an expensive medical examination in the matched samples).

A major problem with matching methods is that in practice it is rare that enough matched pairs of treatment and control units with identical values of **X** can be found, and then the matching does not perfectly control for **X**. A major problem with regression adjustment is that the linear model relating Y to **X** may be wrong, and then the adjustment being applied may not be entirely appropriate. We study cases with imperfect matches and Y moderately nonlinear in **X**.

Cochran and Rubin (1973) summarize work on the efficacy of univariate matching and regression adjustment with quantitative Y and X. The general conclusions of these univariate investigations are that (a) a very simple and easy-to-use pair-matching method known as nearest available pair matching (order the treatment units and sequentially choose as a match for each treatment unit the nearest unmatched control unit) seems to be an excellent matching method; and (b) the combination of regression adjustment on matched samples is usually superior to either method alone.

We extend this work with quantitative Y and X to the case of bivariate **X**. The two main questions to be addressed are (a) Which of two multivariate nearest available pair-matching methods (discriminant, Mahalanobis metric) is preferable? and (b) Which of three regression adjustments (no adjustment, pooled estimate, estimate based on matched pair differences) is preferable? Section 2 introduces terminology and notation, and Section 3 defines the conditions of our Monte Carlo study. Section 4 presents results on the ability of regression adjustment to control bias in random samples. Section 5 presents results for matched samples, with and without regression adjustment. The broad conclusion is that nearest available Mahalanobis metric pair-matching coupled with regression adjustment on the matched pairs is a quite effective general plan for controlling the bias due to matching variables, and this combination is clearly superior to regression adjustment on random samples.

* Donald B. Rubin is Chairman, Statistical Research Group, Educational Testing Service, Princeton, NJ 08541. Research was partially supported by the National Institutes of Education, NIE-C-74-0126, and facilitated by a Guggenheim Fellowship. The author would like to thank D.T. Thayer for excellent and extensive programming support and editors and referees for helpful editorial suggestions.

© Journal of the American Statistical Association
June 1979, Volume 74, Number 366
Theory and Methods Section

2. TERMINOLOGY AND NOTATION

Let P_1 be a population of treatment units and let P_2 be a population of control units. Random samples are obtained from P_1 and P_2; these samples, G_1 and G_2, consist of N and rN units ($r \geq 1$) with recorded values of the matching variables, $\mathbf{X}$. Matched samples are created by assigning to each G_1 unit a G_2 unit having similar values of $\mathbf{X}$; the algorithm used to make the assignments is the matching method. The dependent variable Y is then recorded on all $2N$ units in the matched samples, and the effect of the treatment is estimated. Regression adjustments may be performed by fitting a linear model to the conditional expectation of Y given $\mathbf{X}$. These regressions are estimated from the matched samples and not the random samples because Y is only recorded in the matched samples. Of course, if $r = 1$, the matched samples are simply random samples of size N with pairing of G_1 and G_2 units.

2.1 Matching Methods to Be Studied

We will study two matching methods: nearest available pair matching on the estimated best linear discriminant and nearest available pair matching using the Mahalanobis metric to define distance. Nearest available pair-matching methods first order the G_1 units and then have each G_1 unit choose in turn the closest match from the yet unmatched G_2 units; that is, the first G_1 unit selects the closest G_2 unit, the second G_1 unit selects the closest G_2 unit from the $rN - 1$ not yet matched, and so on, until all G_1 units are matched. These matching methods are fully defined once we specify the order for matching the G_1 units and the precise meaning of closest. Since previous univariate work (Rubin 1973a) indicated that random ordering is usually satisfactory, we will study random order, nearest available matching methods. Closest is clearly defined for one matching variable but not for more than one.

Let $\mathbf{x}_i$ be the $N_i \times p$ data matrix of $\mathbf{X}$ in G_i (where $N_1 = N$, $N_2 = rN$), let $\bar{\mathbf{x}}_i$ be the $1 \times p$ sample mean vector in G_i, and let

$$\mathbf{S} = [(\mathbf{x}_1{}^T\mathbf{x}_1 - N_1\bar{\mathbf{x}}_1{}^T\bar{\mathbf{x}}_1) + (\mathbf{x}_2{}^T\mathbf{x}_2 - N_2\bar{\mathbf{x}}_2{}^T\bar{\mathbf{x}}_2)]/(N_1 + N_2 - 2)$$

be the pooled within-sample covariance matrix of $\mathbf{X}$ based on the random samples G_1 and G_2. Mahalanobis metric matching calculates the distance between a G_1 unit with score $\mathbf{X}_1$ and a G_2 unit with score $\mathbf{X}_2$ as

$$(\mathbf{X}_1 - \mathbf{X}_2)\mathbf{S}^{-1}(\mathbf{X}_1 - \mathbf{X}_2)^T . \tag{2.1}$$

Discriminant matching calculates each unit's score on the estimated discriminant as $\mathbf{X}\mathbf{D}^T$ where $\mathbf{D} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)\mathbf{S}^{-1}$, and then matches on this variable; equivalently, it defines the distance between a G_1 unit with score $\mathbf{X}_1$ and G_2 unit with score $\mathbf{X}_2$ as

$$(\mathbf{X}_1 - \mathbf{X}_2)\mathbf{D}^T\mathbf{D}(\mathbf{X}_1 - \mathbf{X}_2)^T . \tag{2.2}$$

We study these two particular matching methods because both matching methods are easy to implement using commonly available computer programs for sorting and calculating a pooled covariance matrix and because both matching methods have an appealing statistical property discussed in Rubin (1976a,b; 1978b).

2.2 The Treatment Effect

For the expected value of Y given $\mathbf{X}$ in P_i we write $\alpha_i + W_i(\mathbf{X})$. This expectation is often called the response surface for Y in P_i. The difference in expected values of Y for P_1 and P_2 units with the same value of $\mathbf{X}$ is thus $\alpha_1 - \alpha_2 + W_1(\mathbf{X}) - W_2(\mathbf{X})$; when P_1 and P_2 represent two treatment populations such that the variables in $\mathbf{X}$ are the only ones that affect Y and have different distributions in P_1 and P_2, then this difference is the effect of the treatment at $\mathbf{X}$. If $W_1(\mathbf{X}) = W_2(\mathbf{X}) = W(\mathbf{X})$ for all $\mathbf{X}$, the response surfaces are called parallel, and $\alpha_1 - \alpha_2$ is the effect of the treatment for all values of the matching variables $\mathbf{X}$.

Nonparallel response surfaces are not studied here in order to limit the number of conditions in the Monte Carlo experiment and because a straightforward argument suggests that matching must have beneficial effects when the response surfaces are nonparallel and the average treatment effect in P_1 is desired. The expected treatment effect over population P_1 is

$$E_1[\alpha_1 - \alpha_2 + W_1(\mathbf{X}) - W_2(\mathbf{X})] = E_1[\alpha_1 + W_1(\mathbf{X})] - E_1[\alpha_2 + W_2(\mathbf{X})] \tag{2.3}$$

where E_1 is the expectation over the distribution of $\mathbf{X}$ in P_1. An unbiased estimate of the first expectation in (2.3) is simply $\bar{y}_{1\cdot}$, the average observed Y in G_1. If we knew the P_2 response surface, an unbiased estimate of the second term in (2.3) would be

$$\sum_{j=1}^{N} [\alpha_2 + W_2(\mathbf{x}_{1j})]/N , \tag{2.4}$$

that is, the average value of the P_2 response surface across the values of $\mathbf{X}$ in the G_1 sample. Expression (2.4) implies that in order to estimate the expected treatment effect in P_1, we must extrapolate the P_2 response surface, $W_2(\cdot)$, into the region of G_1 data. When the P_2 response surface is estimated from G_2 data, this extrapolation can be subject to great error unless the sample from P_2 used to estimate $W_2(\cdot)$ has values of $\mathbf{X}$ similar to values in G_1, that is, unless the sample from P_2 is matched to G_1. See Billewicz (1965) and Rubin (1973a, 1977) for further discussion of nonparallel response surfaces.

Henceforth, we will assume $W_1(\mathbf{X}) = W_2(\mathbf{X}) = W(\mathbf{X})$ so that the treatment effect is $\tau = \alpha_1 - \alpha_2$.

2.3 Estimators of τ

We will consider three estimators of τ, all of the form

$$\hat{\tau} = (\bar{y}_{1\cdot} - \bar{y}_{2\cdot}) - (\bar{\mathbf{x}}_{1\cdot} - \bar{\mathbf{x}}_{2\cdot})\hat{\boldsymbol{\beta}}$$

where $\bar{y}_{i\cdot}$ and $\bar{\mathbf{x}}_{i\cdot}$ are the means of Y and $\mathbf{X}$ in the matched samples, and $\hat{\boldsymbol{\beta}}$ is an estimated regression coefficient of

1. Estimators of the Response Surface Difference: $\hat{\tau} = \bar{y}_{1\cdot} - \bar{y}_{2\cdot} - (\bar{\mathbf{x}}_{1\cdot} - \bar{\mathbf{x}}_{2\cdot})\hat{\boldsymbol{\beta}}$

Estimators of τ: $\hat{\tau}$	Estimators of $\boldsymbol{\beta}$: $\hat{\boldsymbol{\beta}}$
$\hat{\tau}_o$	$\hat{\boldsymbol{\beta}}_o \equiv 0$
$\hat{\tau}_p$	$\hat{\boldsymbol{\beta}}_p = \mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}$; $\mathbf{S}_{xu} = \sum_{i=1}^{2}\sum_{j=1}^{N}(\mathbf{x}_{ij} - \bar{\mathbf{x}}_{i\cdot})^T\mathbf{u}_{ij}$
$\hat{\tau}_d$	$\hat{\boldsymbol{\beta}}_d = \mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}$; $\mathbf{S}_{xu} = \sum_{i=1}^{2}\sum_{j=1}^{N}(\mathbf{x}_{ij} - \bar{\mathbf{x}}_{i\cdot} - \bar{\mathbf{x}}_{\cdot j} + \bar{\mathbf{x}}_{\cdot\cdot})^T\mathbf{u}_{ij}$

Y on $\mathbf{X}$. The differences between the estimators are thus confined to estimating the regression coefficient and are summarized in Table 1: $\hat{\tau}_o$ is simply the difference of Y means in the matched samples, $\hat{\tau}_p$ is the analysis of covariance estimator of τ from the two-group design ignoring the paired structure of the matched samples, and $\hat{\tau}_d$ is the analysis of covariance estimator of τ using the two group by N matched-pair structure of the matched samples (equivalently forming matched-pair differences, $y_{dj} = y_{1j} - y_{2j}$ and $\mathbf{x}_{dj} = \mathbf{x}_{1j} - \mathbf{x}_{2j}$, $\hat{\boldsymbol{\beta}}_d$ is the estimate of $\boldsymbol{\beta}$ found from regressing y_{dj} on $\mathbf{x}_{dj}$). Note that $\hat{\tau}_d$ is the only estimator that requires matched pairs to be assigned in the matched samples.

Simple algebra shows that the conditional bias of τ given the $\mathbf{x}_{ij}$ is

$$\bar{w}_{1\cdot} - \bar{w}_{2\cdot} - (\bar{\mathbf{x}}_{1\cdot} - \bar{\mathbf{x}}_{2\cdot})(\mathbf{S}_{xx}^{-1}\mathbf{S}_{xw}) \tag{2.5}$$

where $w_{ij} = W(\mathbf{x}_{ij})$, $i = 1, 2$, $j = 1, \ldots, N$, and for $\hat{\tau}_o$, $\mathbf{S}_{xw} \equiv \mathbf{0}$. With multivariate $\mathbf{X}$ and moderate N, the variance of this conditional bias can be substantial, and then the expected value of the conditional bias may not be a good indicator of the utility of a procedure. Hence, we will use the expected value of the squared conditional bias to measure the utility of a procedure:

$$E_*[\bar{w}_{1\cdot} - \bar{w}_{2\cdot} - (\bar{\mathbf{x}}_{1\cdot} - \bar{\mathbf{x}}_{2\cdot})(\mathbf{S}_{xx}^{-1}\mathbf{S}_{xw})]^2 \tag{2.6}$$

where E_* is the expectation over the distribution of $\mathbf{X}$ in matched samples. When $r = 1$, the expected squared bias of $\hat{\tau}_o$ is

$$[E_1(\bar{w}_{1\cdot}) - E_2(\bar{w}_{1\cdot})]^2 + [V_1(\bar{w}_{1\cdot}) + V_2(\bar{w}_{1\cdot})] = [E_1(W(\mathbf{X})) - E_2(W(\mathbf{X}))]^2 + [V_1W(\mathbf{X}) + V_2W(\mathbf{X})]/N$$

where E_i is the expectation and V_i the variance over the distribution of X in P_i. It follows that the percentage reduction in expected squared bias resulting from matching and/or regression adjustment is

$$100\left\{1 - \frac{E_*[\bar{w}_{1\cdot} - \bar{w}_{2\cdot} - (\bar{\mathbf{x}}_{1\cdot} - \bar{\mathbf{x}}_{2\cdot})(\mathbf{S}_{xx}^{-1}\mathbf{S}_{xw})]^2}{[E_1(W(\mathbf{x})) - E_2(W(\mathbf{x}))]^2 + [V_1W(\mathbf{x}) + V_2W(\mathbf{x})]/N}\right\}.$$

If the matches were always perfect ($\mathbf{x}_{1j} = \mathbf{x}_{2j}$, $j = 1, \ldots, N$), then $\bar{\mathbf{x}}_{1\cdot} = \bar{\mathbf{x}}_{2\cdot}$ and $\bar{w}_{1\cdot} = \bar{w}_{2\cdot}$; hence, the percentage reduction in expected squared bias would be 100 for any response surface and all of our estimators. If the response surfaces were parallel and linear, then the percentage reduction in expected squared bias would be 100 for the regression adjusted estimates ($\hat{\tau}_p$ and $\hat{\tau}_d$) whether random or matched samples were used. But in general with imperfect matches and nonlinear response surfaces, the percentage reduction in expected squared bias will be less than 100.

3. MONTE CARLO COMPARISONS OF PROCEDURES — CONDITIONS

Except for the cases noted at the end of Section 2.2, the computations of percentage reductions in expected squared bias in matched samples appear to be analytically intractable. Hence we turn to Monte Carlo techniques. Our study can be compactly described as a $2 \times 3 \times 6 \times 4 \times 3 \times 3 \times 8$ factorial study with one summary value (percentage reduction in expected squared bias) per cell. This summary value was in fact obtained by a covariance adjustment on 100 replications using the first, second, and third moments of $\mathbf{X}$ in each of the random samples G_1, G_2 as nine covariates. The resultant precision of the value is roughly equivalent to that obtained with 300 replications and yields standard errors usually less than 1 percent, although larger in cases with smaller percentage reductions in expected squared bias. The Appendix provides details of the design.

The factors in this study are

Factor 1: matching method: metric matching, discriminant matching.
Factor 2: regression adjustment: $\hat{\tau}_o$, $\hat{\tau}_p$, $\hat{\tau}_d$.
Factor 3: ratio of sample sizes, r: 1, 2, 3, 4, 6, 9 ($N = 50$ for all conditions).
Factor 4: bias along discriminant, B: $\frac{1}{4}$, $\frac{1}{2}$, $\frac{3}{4}$, 1.
Factor 5: ratio of variances along discriminant, σ^2: $\frac{1}{2}$, 1, 2.
Factor 6: ratio of variances orthogonal to discriminant, ξ^2: $\frac{1}{2}$, 1, 2.
Factor 7: response surfaces $W(\mathbf{X})$; curvature along and orthogonal to discriminant: $++$, $+0$, $+-$, $0+$, $0-$, $-+$, -0, $--$; see Equation (3.2).

The first three factors define the procedures that we study. The next three factors specify the distributions of the matching variables in P_1 and P_2; we assume that $\mathbf{X}$ has the following normal distributions:

$$\begin{aligned} \text{In } P_1\ \mathbf{X} &\sim N\left(\begin{pmatrix}\eta\\0\end{pmatrix}, \begin{bmatrix}\sigma^2 & 0\\0 & \xi^2\end{bmatrix}\right), \\ \text{In } P_2\ \mathbf{X} &\sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\right), \end{aligned} \tag{3.1}$$

where

$$B = \eta\Big/\left(\frac{1+\sigma^2}{2}\right)^{\frac{1}{2}} = \tfrac{1}{4}, \tfrac{1}{2}, \tfrac{3}{4}, 1;$$

$$\sigma^2 = \tfrac{1}{2}, 1, 2; \quad \text{and} \quad \xi^2 = \tfrac{1}{2}, 1, 2\ .$$

The last factor defines the nonlinear response surface $W(\mathbf{X})$. The $++$ notation for the eight levels of $W(\mathbf{X})$

refers to nonlinearity along the discriminant and nonlinearity orthogonal to the discriminant. Specifically, we let

$$W(\mathbf{X}) = W(u, v) = \exp\left[a\left(\frac{2}{1+\sigma^2}\right)^{\frac{1}{2}}(u - \eta/2) + b\left(\frac{2}{1+\xi^2}\right)^{\frac{1}{2}} v\right] \quad (3.2)$$

where $(a, b) = (+\frac{1}{2}, +\frac{1}{2}), (+\frac{1}{2}, 0), (+\frac{1}{2}, -\frac{1}{2}), (+.1, +\frac{1}{2}), (-.1, -\frac{1}{2}), (-\frac{1}{2}, +\frac{1}{2}), (-\frac{1}{2}, 0), (-\frac{1}{2}, -\frac{1}{2})$. Values of a were set to $\pm.1$ instead of 0 in order to avoid cases in which $\hat{\tau}_o$ is unbiased in random samples; η, σ^2, and ξ^2 appear in (3.2) so that the response surfaces (3.2) with the distributions of $\mathbf{X}$ given by (3.1) are equivalent to the response surfaces $W(u, v) = \exp(au + bv)$ with the distributions of $\mathbf{X}$ standardized so that $\mathbf{E}_1(\mathbf{X}) + \mathbf{E}_2(\mathbf{X}) = \mathbf{0}$ and $\frac{1}{2}[\mathbf{V}_1(\mathbf{X}) + \mathbf{V}_2(\mathbf{X})] = \mathbf{I}$.

The response surfaces given by (3.2) are moderately nonlinear for the distributions given by (3.1). In order to justify the use of the phrase *moderately nonlinear* to describe these response surfaces, we calculate the percentage of the variance of $W(\mathbf{X})$ that can be attributed to the linear regression on $\mathbf{X}$ in P_i:

$$R_i^2 = \mathbf{C}_i(\mathbf{X}, W(\mathbf{X}))^T\mathbf{V}_i(\mathbf{X})^{-1}\mathbf{C}_i(\mathbf{X}, W(\mathbf{X}))/\mathbf{V}_i(W(\mathbf{X})) \ ,$$

where $\mathbf{C}_i(\cdot, \cdot)$ is the covariance in P_i. Straightforward algebra using (3.1), (3.2), and the fact that if $t \sim N(\theta, \phi)$, then

$$E[\exp(\gamma t)] = \exp[\theta\gamma + \gamma^2\phi/2]$$

and

$$E[t \cdot \exp(\gamma t)] = (\theta + \gamma\phi)E[\exp(\gamma t)]$$

shows that

$$R_i^2 = A_i/[\exp(A_i) - 1]$$

where

$$A_i = 2a^2[1 + \sigma^{2(2i-3)}] + 2b^2[1 + \xi^{2(2i-3)}] \ .$$

It follows that the percentage of variance of $W(\mathbf{X})$ that can be attributed to the linear regression on $\mathbf{X}$ varies between 70 and 92 percent across all conditions of this Monte Carlo study.

4. REGRESSION ADJUSTMENTS WITH RANDOM SAMPLES

In practice, it is not uncommon for researchers to conduct observational studies without any matching. Random samples from P_1 and P_2 are chosen and regression adjustments are used to control the $\mathbf{X}$ variables. We begin our study of the Monte Carlo results by considering estimators with $r = 1$.

When $r = 1$, $\hat{\tau}_o$ yields no reduction in squared bias with either matching procedure, and $\hat{\tau}_p$ is the same for both matching procedures because it is the usual analysis of covariance estimator with two groups and no blocking; therefore, when $r = 1$, only three of the six possible estimators defined by the first two factors are of interest. Although when $r = 1$ $\hat{\tau}_d$ metric matched and $\hat{\tau}_d$ discriminant matched use the same units they in general yield different percentage reductions in expected squared bias because they pair the units in different ways before performing the regression adjustment on matched pair differences.

Table 2 presents the Monte Carlo percentage reductions in expected squared bias for $\hat{\tau}_p$. Although $\hat{\tau}_p$ does quite well in many conditions, especially when $\sigma^2 = \xi^2 = 1$, in other conditions it does quite poorly, especially

2. Percentage Reduction in Expected Squared Bias; Monte Carlo Values for $\hat{\tau}_p$ in Random Samples of Size 50

Response Surface	B =	$\sigma^2 = ½$ ¼	½	¾	1	$\sigma^2 = 1$ ¼	½	¾	1	$\sigma^2 = 2$ ¼	½	¾	1
	++	−52	−55	17	54	43	65	80	87	87	93	95	96
	+0	40	81	93	97	96	99	99	100	87	96	98	99
	+−	−27	−13	47	75	51	74	87	92	87	94	97	97
$\xi^2 = ½$	0+	37	22	10	05	40	28	18	16	43	33	26	25
	0−	65	72	78	82	66	73	78	82	67	73	78	82
	−+	74	87	93	96	83	91	95	97	87	93	96	97
	−0	88	96	99	99	96	99	99	100	38	81	94	97
	−−	75	88	94	96	83	92	95	97	82	91	95	96
	++	35	60	77	86	84	92	95	96	85	94	97	98
	+0	40	81	93	97	96	99	99	100	87	96	98	99
	+−	48	73	87	93	84	93	96	98	82	93	97	98
$\xi^2 = 1$	0+	82	84	85	87	83	85	87	89	84	87	89	90
	0−	86	88	90	91	86	87	89	91	85	87	89	91
	−+	86	94	97	98	89	95	97	98	58	79	90	95
	−0	88	96	99	99	96	99	99	100	39	81	94	97
	−−	85	93	97	97	83	92	95	97	39	68	85	92
	++	80	89	93	95	82	92	95	97	74	88	94	97
	+0	40	81	93	97	96	99	99	100	87	96	98	99
	+−	82	91	95	97	79	90	95	97	70	86	93	97
$\xi^2 = 2$	0+	64	71	77	81	63	71	77	81	63	71	77	81
	0−	43	34	29	30	41	30	23	24	38	26	18	17
	−+	91	96	97	97	60	79	89	93	−12	02	55	79
	−0	88	96	99	99	96	99	99	100	39	81	94	97
	−−	85	93	95	96	42	69	84	90	−43	−26	42	72

3. Percentage Reduction in Expected Squared Bias for $\hat{\tau}_d$, Metric Matching With $r = 1$

Response Surface	$B =$	$\sigma^2 = \frac{1}{2}$				$\sigma^2 = 1$				$\sigma^2 = 2$			
		¼	½	¾	1	¼	½	¾	1	¼	½	¾	1
	++	−12	−43	24	66	54	49	87	86	91	92	93	93
	+0	53	87	96	99	97	99	99	99	92	98	99	98
	+−	−09	12	63	84	60	79	90	93	91	93	94	94
$\xi^2 = \frac{1}{2}$	0+	47	47	47	44	56	52	67	60	64	64	68	65
	0−	74	84	89	92	79	87	93	94	78	87	93	94
	−+	84	94	97	96	89	94	96	97	86	93	95	96
	−0	94	98	99	98	97	99	99	99	47	88	97	99
	−−	83	93	94	95	87	93	96	97	72	86	95	97
	++	47	70	77	86	88	91	95	95	91	96	94	95
	+0	60	89	96	98	96	99	99	99	92	98	99	98
	+−	60	80	92	95	86	95	97	97	91	97	98	97
$\xi^2 = 1$	0+	86	87	85	89	89	89	90	89	89	90	91	93
	0−	88	90	92	93	91	91	93	94	91	92	95	95
	−+	92	97	97	97	91	96	97	97	64	83	93	96
	−0	95	98	99	98	96	98	99	99	48	88	97	99
	−−	88	93	94	94	86	89	94	96	47	76	91	96
	++	87	89	88	93	88	93	96	93	86	94	94	94
	+0	57	89	96	98	96	98	99	99	92	98	99	98
	+−	84	93	96	97	83	94	97	97	81	94	97	97
$\xi^2 = 2$	0+	73	77	84	86	75	83	88	87	76	83	88	89
	0−	55	52	54	56	58	59	57	59	54	53	56	57
	−+	91	94	95	95	63	81	90	93	06	26	73	88
	−0	94	98	99	98	96	98	99	99	51	88	97	99
	−−	82	88	91	92	48	75	88	91	−19	07	65	87

when $\sigma^2 = \xi^2 = 2$ and $\sigma^2 = \xi^2 = \frac{1}{2}$. The negative values in Table 2 indicate that the regression adjustment actually increases bias. These results show that $\hat{\tau}_p$ based on random samples cannot be counted on to control the X-variables when the response surfaces are nonlinear.

Some insight into the problem with $\hat{\tau}_p$ can be achieved by considering the large sample case. In large samples, $\hat{\tau}_p$ is approximately $E_1(W(\mathbf{X})) - E_2(W(\mathbf{X})) - \eta c$ where c is the pooled slope of $W(\mathbf{X})$ on the discriminant, that is, the first component of $[\mathbf{C}_1(\mathbf{X}, W(\mathbf{X})) + \mathbf{C}_2(\mathbf{X}, W(\mathbf{X}))]/(1 + \sigma^2)$. From (3.1) and (3.2), we can write $\eta c = aB[E_1(W(\mathbf{X}))(\sigma^2/(1 + \sigma^2)) + E_2(W(\mathbf{X}))/(1 + \sigma^2)]$; because $W(\mathbf{X})$ and B are positive, ηc has the same sign as a. If $\sigma^2 = \xi^2 = 1$, then the initial bias, $E_1(W(\mathbf{X})) - E_2(W(\mathbf{X}))$, has the same sign as a, implying that the adjustment $-\eta c$ is in the correct direction. However, if $\sigma_1^2 \neq 1$ and/or $\xi^2 \neq 1$, then the adjustment may be in the wrong direction and actually increase bias.

Somewhat surprising, estimating the regression coefficient from matched-pair differences in the random samples can result in better estimates. Table 3 presents Monte Carlo values for the percentage reduction in expected squared bias for $\hat{\tau}_d$ metric matched with $r = 1$; $\hat{\tau}_d$ is superior to $\hat{\tau}_p$ in 209 of 288 cases, and in only two cases ($\sigma^2 = 1$, $\xi^2 = \frac{1}{2}$, $B = \frac{1}{2}$, ++ and $\sigma^2 = 2$, $\xi^2 = \frac{1}{2}$, $B = \frac{1}{4}$, −−) do the results favor $\hat{\tau}_p$ by more than 5 percent. The fact that $\hat{\tau}_d$ is usually better than $\hat{\tau}_p$ is consistent with Monte Carlo results and analytic considerations presented in Rubin (1973b) for the univariate case.

The results for $\hat{\tau}_d$ discriminant matched are similar but inferior to the results for $\hat{\tau}_d$ metric matched. The mean of the 288 metric minus discriminant differences in percentage reduction in expected squared bias is 5.4, the minimum difference is −15.6, the maximum difference is 58.1, 219 differences are positive, and only three differences are less than −10 ($\sigma^2 = \frac{1}{2}$, $\xi^2 = \frac{1}{2}$, $B = \frac{3}{4}$, ++; $\sigma^2 = 1$, $\xi^2 = \frac{1}{2}$, $B = \frac{1}{2}$, ++; $\sigma^2 = 2$, $\xi^2 = \frac{1}{2}$, $B = \frac{1}{4}$, −−).

Even though the results for $\hat{\tau}_d$ are somewhat better than for $\hat{\tau}_p$, in cases in which $\hat{\tau}_p$ does poorly, so does $\hat{\tau}_d$. Of course, with real data we could try fitting higher-order (e.g., quadratic) terms, although the nonlinearity might be difficult to detect because these response surfaces are only moderately nonlinear. Our study does not include quadratic terms in the regression adjustment but does include matched sampling. Hence, we turn to results with $r > 1$ to see if matched sampling improves the estimation of τ.

5. RESULTS WITH $r > 1$

We now consider the utility of matched sampling, alone and in combination with regression adjustment. Because the analyses of the 7-factor Monte Carlo study are somewhat involved, we begin in Section 5.1 by presenting specific results for two procedures in order to convey the flavor of our conclusions. The remainder of Section 5 presents detailed analyses of the results of the Monte Carlo study with $r > 1$. Section 5.2 presents an analysis of variance (ANOVA) of the study. Section 5.3 shows that although the difference between metric and discriminant matching varies with the estimator, the distribution of $\mathbf{X}$, and the response surface, metric matching is clearly superior to discriminant matching. Section 5.4 focuses on the results for metric matching and concludes that $\hat{\tau}_d$ is the best regression adjustment procedure that

we have considered. Section 5.5 displays results for $\hat{\tau}_d$ metric matched that can be used to suggest ratios of sample sizes needed to remove nearly all of the bias for a variety of nonlinear response surfaces.

5.1 Two Specific Estimators

Tables 4 and 5 present percentage reductions in expected squared bias for metric matched samples with $r = 2$ for $\hat{\tau}_o$ and $\hat{\tau}_d$. By comparing Tables 3 and 5, we immediately see advantages to matching even with $r = 2$. The estimator $\hat{\tau}_d$ metric matched with $r = 2$ usually removes most of the squared bias and is clearly better than $\hat{\tau}_d$ (or $\hat{\tau}_p$) with $r = 1$. Of the 288 differences between the percentage reductions in expected squared bias for $\hat{\tau}_d$ metric matched with $r = 2$ and $\hat{\tau}_d$ metric matched with $r = 1$, only two are negative; of the 288 differences between the percentage reductions in expected squared bias for $\hat{\tau}_d$ metric matched with $r = 2$ and $\hat{\tau}_p$ metric matched with $r = 1$, only seven are negative; and the most negative of these nine differences is only -1. In the next sections we will see that in those cases that are difficult for matching (e.g., $\sigma^2 = \xi^2 = 2$), larger ratios result in even better estimates.

By comparing Tables 4 and 5 we see that in all cases except with $\sigma^2 = \xi^2 = 2$, there is an advantage to using regression adjustment on matched samples. When $\sigma^2 = \xi^2 = 2$, $\hat{\tau}_o$ is superior to $\hat{\tau}_d$ in a few cases, but $\hat{\tau}_d$ is usually better. Without knowledge of the response surface, $\hat{\tau}_d$ is to be preferred to $\hat{\tau}_o$; with such knowledge, a more appropriate regression adjustment can be used. We will see that $\hat{\tau}_d$ is in this sense always preferable to $\hat{\tau}_o$ (or $\hat{\tau}_p$).

5.2 An ANOVA of the Results When $r > 1$

Table 6 presents an ANOVA of the 7-factor study where factor 3 has five levels $r = 2, 3, 4, 6, 9$. For simplicity of display, the response surface factor and the three distribution of **X** factors have been collapsed into one "distribution" factor with 288 levels. In fact, little information was lost by collapsing the distributional factors because of numerous large higher-order interactions among the distributional factors and because larger interactions between procedure factors and distributional factors tended to involve higher-order interactions among the distributional factors. The purpose of this ANOVA is simply to see which are the large sources of variation.

Table 7 summarizes the procedure factors by the average value of the percentage reduction in expected squared bias over the 288 distributional conditions. If there were no interactions between procedures and conditions, then Table 7 would be an adequate summary for our Monte Carlo study; that is, there would be good and bad procedures and easy and hard distributional conditions, but comparisons between procedures would be the same in each distributional condition. However, there are nontrivial interactions between procedures and distributions in the sense that if we fit the procedure-plus-distribution additive model to the 7-factor study, we are left with some large residuals to explain. Although Table 7 is not an entirely adequate summary for our study, we will see in Sections 5.3 through 5.5 that most trends displayed there are not misleading. The major trends in Table 7 are that

1. Metric matching is superior to discriminant matching.

4. Percentage Reduction in Expected Squared Bias for $\hat{\tau}_o$, Metric Matching With $r = 2$

Response Surface	B =	$\sigma^2 = 1/2$				$\sigma^2 = 1$				$\sigma^2 = 2$			
		1/4	1/2	3/4	1	1/4	1/2	3/4	1	1/4	1/2	3/4	1
	++	95	88	88	82	83	77	73	67	50	49	49	46
	+0	95	96	94	87	89	85	79	71	65	61	57	51
	+−	95	91	92	87	87	82	77	70	60	54	52	48
$\xi^2 = 1/2$	0+	98	94	90	80	96	93	88	78	92	87	78	67
	0−	98	98	95	91	98	97	93	77	95	94	89	82
	−+	99	99	99	98	99	98	98	96	93	96	93	90
	−0	99	99	99	98	98	98	97	95	85	93	90	87
	−−	99	99	99	97	99	98	97	95	91	96	93	90
	++	80	83	84	80	71	74	71	65	50	53	50	45
	+0	96	96	94	88	89	87	81	72	67	64	58	51
	+−	91	90	89	84	82	80	76	68	63	59	54	48
$\xi^2 = 1$	0+	83	83	79	73	79	76	73	68	70	67	62	57
	0−	90	91	93	92	90	92	91	90	88	89	87	83
	−+	99	99	99	98	95	98	98	96	81	93	93	90
	−0	99	99	99	98	98	98	97	95	85	93	91	88
	−−	98	99	99	98	93	98	98	96	77	92	91	89
	++	56	70	75	71	60	64	63	57	45	45	43	39
	+0	95	95	94	88	89	86	80	72	67	63	57	50
	+−	71	78	79	75	69	71	67	61	54	53	48	43
$\xi^2 = 2$	0+	56	60	58	57	56	56	55	51	51	48	47	42
	0−	55	50	51	55	60	56	56	63	60	63	64	66
	−+	89	96	98	99	72	93	97	97	54	82	91	91
	−0	99	99	98	97	97	97	97	95	84	93	90	87
	−−	89	96	98	98	75	92	96	96	51	80	90	88

5. Percentage Reduction in Expected Squared Bias for $\hat{\tau}_d$, Metric Matching With r = 2

Response Surface	B =	$\sigma^2 = 1/2$				$\sigma^2 = 1$				$\sigma^2 = 2$			
		1/4	1/2	3/4	1	1/4	1/2	3/4	1	1/4	1/2	3/4	1
	++	99	97	96	97	98	96	98	98	95	97	97	97
	+0	99	100	99	100	99	100	100	100	96	99	99	99
	+−	98	97	97	96	97	97	98	97	95	97	97	96
$\xi^2 = 1/2$	0+	99	98	97	95	99	98	98	96	98	98	97	96
	0−	100	99	99	99	100	100	99	99	99	99	99	99
	−+	100	100	100	100	99	100	100	100	93	97	99	100
	−0	100	100	100	100	99	100	100	100	79	95	99	100
	−−	100	100	100	100	99	99	100	100	89	95	98	99
	++	98	98	98	97	98	98	99	99	95	98	96	98
	+0	99	100	100	100	99	100	100	100	96	99	99	99
	+−	98	99	99	98	98	99	99	98	95	98	97	97
$\xi^2 = 1$	0+	97	97	97	96	98	96	97	97	96	96	95	97
	0−	98	98	98	98	98	98	98	98	97	97	98	98
	−+	99	100	100	100	98	99	100	100	80	90	97	99
	−0	100	100	100	100	99	100	100	100	79	95	99	100
	−−	99	100	100	100	97	98	99	100	74	86	95	98
	++	93	96	98	97	93	97	98	98	92	97	96	98
	+0	98	99	100	100	99	99	100	100	96	99	99	99
	+−	92	96	98	98	93	97	98	98	91	96	98	98
$\xi^2 = 2$	0+	88	90	92	93	89	92	93	92	88	92	93	96
	0−	80	76	78	80	81	81	80	79	78	77	77	79
	−+	94	97	99	99	84	93	97	99	51	62	87	95
	−0	100	100	100	100	99	100	100	100	77	94	99	100
	−−	93	97	99	99	80	90	96	98	41	50	82	93

2. $\hat{\tau}_d$ is superior to $\hat{\tau}_p$ especially with metric matched samples, and both $\hat{\tau}_d$ and $\hat{\tau}_p$ are superior to $\hat{\tau}_o$ especially for smaller r.
3. Larger ratios of sample sizes are better, although only modest benefits accrue when moving from $r = 2$ to larger ratios, the benefit being largest with $\hat{\tau}_o$ and smallest with $\hat{\tau}_d$.

Further analysis will show that the conclusion from 1. and 2., to the effect that τ_d metric matched is the best combination of matching method and regression adjustment, is correct. However, the conclusion from 3., to the effect that ratios larger than 2 are not needed, is not always true; with some combinations of distributions of **X** and response surfaces, using larger ratios for matching can result in substantial improvements.

6. Analysis of Variance of 7-Factor Monte Carlo Study With r > 1

Source	Degrees of Freedom	Mean Square
Matching Method (metric, discriminant)	1	36.615
Regression Adjustment ($\hat{\tau}_o$,$\hat{\tau}_p$,$\hat{\tau}_d$)	2	7.414
Ratio (r = 2,3,4,6,9)	4	.705
Distribution[a]	287	.442
M.M. × R.A.	2	.430
M.M. × Ratio	4	.214
M.M. × Distribution	287	.241
R.A. × Ratio	8	.114
R.A. × Distribution	574	.027
Ratio × Distribution	1148	.004
M.M. × R.A. × Ratio	8	.051
M.M. × R.A. × Distribution	514	.007
M.M. × Ratio × Distribution	1148	.002
R.A. × Ratio × Distribution	2296	.002
M.M. × R.A. × Ratio × Distribution	2296	.001

[a] Factors 4, 5, 6, and 7 defined in Section 3.

7. Percentage Reduction in Expected Squared Bias Averaging Over Distributional Conditions

Ratio	Discriminant Matching			Metric Matching		
	$\hat{\tau}_o$	$\hat{\tau}_p$	$\hat{\tau}_d$	$\hat{\tau}_o$	$\hat{\tau}_p$	$\hat{\tau}_d$
1	00[a]	78	78	00[a]	78	84
2	71	83	84	81	91	96
3	74	84	85	88	94	97
4	75	84	85	90	95	98
6	74	84	85	94	96	99
9	75	85	86	95	97	99

[a] Theoretical values

5.3 Metric Matching vs. Discriminant Matching

Table 6 indicates that the matching method factor and its interactions with distributional factors make a large contribution to the variation in the Monte Carlo study. Table 7 suggests that metric matching is on the average superior to discriminant matching. But these tables do not tell us whether the interaction between matching method and distribution is the result of a varying superiority of metric matching over discriminant matching or an occasional superiority of discriminant matching. If discriminant matching were better than metric matching for only some distributions of **X**, then we should decide which matching method to use on the basis of the observed distribution of **X** in G_1 and G_2. Our results clearly

8. Summary of Differences in Percentage Reduction in Expected Squared Bias for Each Estimator: Metric Matching Minus Discriminant Matching[a]

	Ratio $r =$	2	3	4	6	9
$\hat{\tau}_o$ (metric)	min	−19.9	−16.0	−13.0	−7.8	−6.1
$-\hat{\tau}_o$ (discrim)	max	124.5	120.2	137.6	139.3	175.6
	mean	9.9	13.3	15.4	19.2	20.6
	# > 0	194	219	232	245	254
$\hat{\tau}_p$ (metric)	min	−2.0	−.7	−.7	−.4	−.5
$-\hat{\tau}_p$ (discrim)	max	102.0	73.4	90.2	83.1	82.9
	mean	8.9	9.8	10.9	12.0	12.7
	# > 0	262	260	256	259	262
$\hat{\tau}_d$ (metric)	min	−1.2	−1.3	−.6	−.4	−.2
$-\hat{\tau}_d$ (discrim)	max	107.6	79.1	90.1	83.9	81.6
	mean	11.5	12.0	12.9	13.2	13.2
	# > 0	269	264	264	271	273

[a] 288 differences for each estimator, one for each distributional condition in Monte Carlo study.

show metric matching to be superior to discriminant matching for all distributions of $\mathbf{X}$ considered.

Focus on one estimator, that is, one adjustment ($\hat{\tau}_o$, $\hat{\tau}_d$, $\hat{\tau}_p$) and one ratio ($r = 2, 3, 4, 6, 9$); for each of the 288 distributional conditions, take the difference between the percentage reduction in expected squared bias obtained by metric matching and obtained by discriminant matching. If all 288 differences were positive, we would know that metric matching was superior to discriminant matching for that estimator (e.g., $\hat{\tau}_d$ with $r = 2$).

Table 8 summarizes the $3 \times 5 = 15$ metric minus discriminant sets of differences across the 288 distributional conditions. With $\hat{\tau}_d$ and $\hat{\tau}_p$ there are essentially no cases where discriminant matching is to be preferred to metric matching; the most negative difference is only −2 percent. With $\hat{\tau}_o$, metric matching is usually superior to discriminant matching, but further study of these differences is enlightening. All cases where the differences are ≤ -10 percent occur when $\sigma^2 = 2$ and $\xi^2 = \frac{1}{2}$: five cases when $r = 2$, three cases when $r = 3$, and two cases when $r = 4$. Table 9 displays the metric minus discriminant differences for $\hat{\tau}_o$ with $r = 2, 3, 4$ for $\sigma^2 = 2$ and $\xi^2 = \frac{1}{2}$. Clearly, even for these distributions of $\mathbf{X}$, metric matching is to be preferred unless exceptionally strong prior knowledge suggests a response surface that has curvature increasing as one moves from the P_2 range of $\mathbf{X}$ to the P_1 range of $\mathbf{X}$, and even then little can be lost by metric matching especially if $r \geq 4$.

This conclusion holds for $\hat{\tau}_o$ even when the response surface is linear in the discriminant (i.e., $W(\mathbf{X}) = \mathbf{X}(1, 0)^T$). Of the $3(\sigma^2) \times 3(\xi^2) \times 4(\mathrm{B}) \times 5(r) = 180$ differences in percentage reduction in expected squared bias for $\hat{\tau}_o$ (Mahalanobis metric matched) minus $\hat{\tau}_o$ (discriminant matched), only 26 were greater than 2 percent, and of these, 6 favored discriminant matching (five 3 percent differences, one 4 percent difference), whereas 20 favored Mahalanobis metric matching (four 3 percent differences, six 4 percent, two 5 percent, four 6 percent, three 7 percent, and one 9 percent difference).

Because there seems to be no reason to recommend discriminant matching over metric matching, we restrict further investigation of the Monte Carlo study to results obtained by metric matching.

5.4 Comparing Regression Adjustments

Table 10 compares the regression adjustments (based on metric matched samples) for each ratio across the 288 distributional conditions. The comparison of $\hat{\tau}_d$ with $\hat{\tau}_p$ shows that although there is usually not much difference between the adjustments, $\hat{\tau}_d$ is clearly superior to $\hat{\tau}_p$, the most negative difference in percentage reductions in expected squared bias being −2.2 percent when $r = 2$.

The comparison of $\hat{\tau}_d$ with $\hat{\tau}_o$ in Table 10 shows that although $\hat{\tau}_d$ is usually substantially better than $\hat{\tau}_o$, $\hat{\tau}_o$ is better than $\hat{\tau}_d$ in a few cases. All five cases having differences in percentage reductions in expected squared bias favoring $\hat{\tau}_o$ by 10 percent or more occur when $\sigma^2 = \xi^2 = 2$, three with $r = 2$ and two with $r = 3$. Tables 4 and 5 provide the results for $\hat{\tau}_o$ and $\hat{\tau}_d$ when $r = 2$ and show that without strong prior knowledge of the response surface, τ_d is better than τ_o even when $\sigma^2 = \xi^2 = 2$; a researcher having rather specific knowledge of the response surface should be fitting a model relevant to that response surface and should not be using a linear approximation ($\hat{\tau}_d$ or $\hat{\tau}_p$) or no adjustments ($\hat{\tau}_o$).

The conclusion thus far is simple: Use $\hat{\tau}_d$ based on metric matched samples. It remains for us to summarize the efficacy of using different ratios for the matching.

9. Differences in Percentage Reduction in Expected Squared Bias: $\hat{\tau}_o$ (Metric Matched) Minus $\hat{\tau}_o$ (Discriminant Matched) When Distributions of $\mathbf{X}$ Are Relatively Favorable to Discriminant Matching ($\sigma^2 = 2$, $\xi^2 = 1/2$)

	$r = 2$				$r = 3$				$r = 4$			
$B =$	1/4	1/2	3/4	1	1/4	1/2	3/4	1	1/4	1/2	3/4	1
++	−08	−15	−18	−12	−03	−16	−15	−10	05	−05	−13	−12
+0	−01	−04	−03	−01	−03	−04	−03	−02	01	−03	−03	−02
+−	−04	−20	−13	−08	−01	−07	−09	−08	16	−01	−08	−06
0+	63	71	35	24	64	64	57	57	74	94	88	61
0−	52	49	30	20	55	73	36	30	61	45	31	29
−+	18	16	10	11	17	16	08	07	21	13	09	06
−0	11	−01	−01	−01	24	01	−01	−01	20	00	−01	−00
−−	32	26	10	06	27	26	11	06	29	15	09	06

10. Summary of Differences in Percentage Reduction in Expected Squared Bias for Estimators Based on Metric Matched Samples

	Ratio $r =$	2	3	4	6	9
$\hat{\tau}_d$ (metric)	min	−2.2	−.6	−.0	−.1	−.7
$-\hat{\tau}_p$ (metric)	max	34.1	26.3	20.9	16.2	12.6
	mean	4.2	3.4	3.0	2.2	1.8
	# > 0	275	281	285	284	283
$\hat{\tau}_d$ (metric)	min	−29.8	−16.1	−9.5	−9.3	−6.0
$-\hat{\tau}_o$ (metric)	max	59.7	45.0	37.2	26.9	21.6
	mean	14.9	9.2	7.5	5.0	3.8
	# > 0	270	259	270	271	270

5.5 The Effect of Using Different Ratios for Matching

Tables 2 through 5 give specific results for the estimators $\hat{\tau}_p(r = 1)$, $\hat{\tau}_d(r = 1, \text{metric})$, $\hat{\tau}_o(r = 2, \text{metric})$, and $\hat{\tau}_d(r = 2, \text{metric})$. These tables show that in some cases any of these estimators can remove nearly all of the squared bias, whereas in other cases even the best of them, $\hat{\tau}_d$ with $r = 2$, removes less than 50 percent of squared bias. The results are sensitive to the distribution of $\mathbf{X}$ and the response surface, especially when $r = 1$. Table 5 shows that if $\sigma^2 \leq 1$ and $\xi^2 \leq 1$, $\hat{\tau}_d$ metric matched with $r = 2$ can be counted on to remove most of the bias, and in most cases with either $\sigma^2 > 1$ or $\xi^2 > 1$ usually removes more than 90 percent of the bias. The clearest need for improved estimation occurs when both $\sigma^2 > 1$ and $\xi^2 > 1$. Increasing r yields better estimates.

In order to indicate the advantages of increasing r, we present Table 11, which gives "pessimistic" results for $\hat{\tau}_d$ metric matched. Pessimistic means that for each ratio and each distribution of $\mathbf{X}$ defined by a value of (B, σ^2, ξ^2), we have produced the minimum percentage reduction in squared bias over the eight response surfaces. Hence, within the context of our Monte Carlo study, these represent the worst results that can be obtained by using $\hat{\tau}_d$ metric matched. We see that increases in r result in improved estimation. When $r = 4$, $\hat{\tau}_d$ metric matched removes 73 percent of the expected squared bias in the worst case and usually removes more than 90 percent even when $\sigma^2 = \xi^2 = 2$. The most difficult cases are those with small initial bias. Using a ratio equal to $2\sigma^2\xi^2$ for matching usually removes most of the bias.

6. DISCUSSION

The general conclusion from our analyses of the Monte Carlo study is that the best procedure we have considered is $\hat{\tau}_d$ (the regression-adjusted estimator based on matched pair differences) using large r (large ratio of size of reservoir to size of matched sample) and metric matched samples (specifically, using the Mahalanobis metric, (2.1)). Obtaining G_2 with large r can be expensive in practice, however, and we have seen that the improvements that accrue from using $r = 9$ rather than $r = 2$ are modest, except when the spread of the $\mathbf{X}$ distribution is larger in P_1 than P_2. Tentative advice would be to metric match using a ratio of $2\sigma^2\xi^2$ (i.e., perhaps twice the determinant of $\mathbf{\Sigma}_1 \mathbf{\Sigma}_2^{-1}$ where $\mathbf{\Sigma}_i$ is the covariance of $\mathbf{X}$ in P_i) and perform regression adjustments on matched pair differences. Our results demonstrate quite clearly that matching can dramatically improve estimation.

Of course, a realistic criticism of this work is that we have not considered other procedures, ones that carefully try to search for nonlinear components in the response surfaces or try to perform sophisticated Bayesian or empirical Bayesian analyses that average over a variety of nonlinear models for the response surfaces. Our reaction to this criticism is that although we hope that in any real data analysis such techniques would be applied, con-

11. Pessimistic Percentage Reductions in Expected Squared Bias for $\hat{\tau}_d$, Metric Matched

			$\sigma^2 = ½$				$\sigma^2 = 1$				$\sigma^2 = 2$			
r	ξ^2	$B =$	¼	½	¾	1	¼	½	¾	1	¼	½	¾	1
1	½		−12	−43	24	44	54	49	67	60	47	64	68	65
	1		47	70	77	86	86	89	90	89	47	76	91	93
	2		55	52	54	56	48	59	57	59	−19	07	56	57
2	½		98	97	96	95	97	96	98	96	79	95	97	96
	1		97	97	97	96	97	96	97	97	74	86	95	97
	2		80	76	78	80	80	81	80	79	41	50	77	79
3	½		99	99	99	98	97	98	98	98	85	96	98	97
	1		99	99	99	99	94	98	98	98	82	88	96	97
	2		88	84	83	83	87	85	85	85	61	65	82	85
4	½		100	99	100	99	98	98	99	99	92	97	98	98
	1		99	99	99	99	98	98	98	99	88	92	97	97
	2		92	89	90	88	90	92	91	89	73	77	89	89
6	½		99	100	100	100	99	99	100	100	94	98	99	98
	1		99	99	99	99	98	99	99	99	94	94	98	97
	2		94	94	94	94	94	94	94	92	80	81	92	92
9	½		100	100	99	100	100	99	100	100	97	99	99	99
	1		99	99	100	100	99	97	99	100	95	99	99	99
	2		96	96	96	96	95	96	96	96	85	85	95	95

sidering all of them in a Monte Carlo study is impossible because good data-analytic techniques must be, by nature, conditional on the observed data. Presumably, such sophisticated estimators would do at least as well as $\hat{\tau}_d$, and so the results for $\hat{\tau}_d$ (metric matched) can be thought of as minimums for methods that try to estimate the response surface by more than a linear fit. We feel that the general benefits from obtaining matched samples would hold for more sophisticated estimators because with matched samples the sensitivity of the regression adjustment to model specification is reduced.

In conclusion, because we feel that results similar to ours would be obtained for more than two matching variables, for nonnormal matching variables, and other nonlinear response surfaces, we feel that an effective plan for controlling matching variables in observational studies is to perform regression adjustments on matched samples obtained by nearest available Mahalanobis metric pair matching. Of course this summary of advice assumes the sampling situation described in Section 2, with dependent variables recorded in the matched samples but not recorded in the initial random samples.

APPENDIX: DISCUSSION OF MONTE CARLO STUDY

The Monte Carlo study is a 7-factor study as described in Section 3. The first three factors define 36 procedures, and the last four factors define 288 distributional conditions. For the moment, focus on one condition and one procedure: Over repeated matched samples we wish to know the expected value of BIAS2, expression (2.6).

Letting NSIM be the number of sampling replications in each condition, we could simply perform NSIM replications in each condition, drawing samples independently across the conditions and independently for each procedure. Such a sampling scheme would be more expensive than necessary for several reasons. First, it would generate many more random numbers than needed. Second, it would not provide efficient comparisons of procedures within conditions or a procedure across conditions because the independent sampling would not have created correlated estimates of BIAS2. Third, no attempt would have been made to increase the precision of the study by using simulation covariates, quantities that are defined for each procedure and condition and correlated with BIAS2, but whose expectations we know from analytic considerations, for example, the moments of $\mathbf{X}$ in the random samples G_1 and G_2.

First consider the third point, using simulation covariates to increase precision. In our study with normal random variables, we know the expectations of all sample moments in G_1 and G_2, and these should be related to the ease of obtaining well-matched samples and the utility of regression adjustment. For example, G_1 and G_2 samples with means farther apart than usual should imply that the resultant matched samples will have means farther apart than usual. We can let the data estimate these relationships between sample moments in G_1 and G_2 and BIAS2. Pilot studies indicated that using the first three moments in G_1 and G_2 as simulation covariates resulted in the most cost-effective plan. More covariates would have been more expensive because extra storage would have been required in core, and fewer covariates would have been more expensive because additional simulations would have been needed to obtain the same precision. Roughly speaking, our use of simulation covariates allowed us to reduce the number of simulations by a factor of 3. Typically, the squared multiple correlation between BIAS2 and the simulation covariates was about .6 and higher when the standard errors were higher; hence, without the covariance adjustment we would have needed roughly NSIM = 300 in order to obtain the precision that was obtained using NSIM = 100 and simulation covariates.

The issues of minimizing the number of random numbers generated in order to save cost and correlating the estimates across conditions to increase precision of comparisons are really handled in the same manner. First, focus on one distributional condition and consider a fixed ratio r and a fixed matching method. Then we want to compare the three adjustments $\hat{\tau}_o$, $\hat{\tau}_p$, $\hat{\tau}_d$ with respect to BIAS2. By calculating all adjustments on the same matched sample, we make the estimates of BIAS2 correlated across adjustments and hence make comparisons more precise. Now let us consider different matching methods with the same ratio; we cannot use the same matched samples from G_2 but we can and do use the same random samples (G_1, G_2) for matching. For example, BIAS2 for $\hat{\tau}_o$ $r = 2$ metric and $\hat{\tau}_o$ $r = 2$ discriminant are calculated in matched samples obtained from the same random samples. Using the same random samples correlates the metric and discriminant results, increasing the precision when comparing metric matching with discriminant matching for each estimator. Finally, consider different ratios for matching; we cannot use the same random samples G_2, but we can and do use the same random samples G_1 and overlapping random samples G_2 (i.e., the $r = 2$ G_2 sample includes the $r = 1$ G_2 sample, the $r = 3$ G_2 sample includes the $r = 2$ G_2 sample, etc.). Using overlapping random samples correlates the results for different ratios and hence increases the precision of comparisons between estimators using different ratios.

Furthermore, we can correlate results across response surfaces and distributions of $\mathbf{X}$, thereby increasing precision of comparisons of procedures between distributional conditions and reducing computational costs by reducing the number of random deviates that must be generated. Correlating results across response surfaces is trivial because all nine response surfaces can be studied from the same matched sample. Correlating results across distributions of $\mathbf{X}$ is done by having the $4 \times 3 \times 3$ cases use the same $N(0, 1)$ deviates. Only the G_1 sample needs to be linearly transformed in accordance with equation (3.1).

The $N(0, 1)$ deviates were generated by Marsaglia's rectangle-wedge-tail method described in Knuth (1969).

[*Received January 1978. Revised December 1978.*]

REFERENCES

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

Billewicz, W.Z. (1964), "Matched Samples in Medical Investigations," *British Journal of Preventative Social Medicine*, 18, 167–173.

Billewicz, W.Z. (1965), "The Efficiency of Matched Samples: An Empirical Investigation," *Biometrics*, 21, 623–643.

Campbell, D.T., and Erlebacher, A. (1970), "How Regression Artifacts in Quasi-Experimental Evaluations Can Mistakenly Make Compensatory Education Look Harmful," in *The Disadvantaged Child, Volume 3, Compensatory Education: A National Debate*, ed. J. Hellmuth, New York: Brunner/Mazel.

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

Cochran, W.G. (1953), "Matching in Analytical Studies," *American Journal of Public Health*, 43, 684–691.

Cochran, W.G. (1963), *Sampling Techniques*, New York: John Wiley & Sons.

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

———, and Rubin, D.B. (1973), "Controlling Bias in Observational Studies: A Review," *Sankhya-A*, 35, 417–446.

Gilbert, J.P., Light, R.J., and Mosteller, F. (1975), "Assessing Social Innovation: An Empirical Base for Policy," in *Evaluation and Experiment: Some Critical Issues in Assessing Social Programs*, ed. A.R. Lumsdaine and C.A. Bennett, New York: Academic Press.

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

Knuth, D.E. (1969), *Seminumerical Algorithms*, Volume 2, Reading, Mass.: Addison-Wesley.

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

McKinlay, S.M. (1974), "The Expected Number of Matches and Its Variance for Matched-Pair Designs," *Applied Statistics*, 23, 372–383.

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

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

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

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

Rubin, D.B. (1974), "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66, 688–701.

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

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

Rubin, D.B. (1977), "Assignment to Treatment Group on the Basis of a Covariate," *Journal of Educational Statistics*, 2, 1–26.

Rubin, D.B. (1978a), "Bayesian Inference for Causal Effects: The Role of Randomization," *The Annals of Statistics*, 7, 34–58.

Rubin, D.B. (1978b), *Bias Reduction Using Mahalanobis Metric Matching*, Research Bulletin 78–17, Princeton, N.J.: Educational Testing Service.