BIOMETRICS *29*, 185–203
March 1973

THE USE OF MATCHED SAMPLING AND REGRESSION ADJUSTMENT TO REMOVE BIAS IN OBSERVATIONAL STUDIES

DONALD B. RUBIN[1]

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

SUMMARY

The ability of matched sampling and linear regression adjustment to reduce the bias of an estimate of the treatment effect in two sample observational studies is investigated for a simple matching method and five simple estimates. Monte Carlo results are given for moderately linear exponential response surfaces and analytic results are presented for quadratic response surfaces. The conclusions are (1) in general both matched sampling and regression adjustment can be expected to reduce bias, (2) in some cases when the variance of the matching variable differs in the two populations both matching and regression adjustment can increase bias, (3) when the variance of the matching variable is the same in the two populations and the distributions of the matching variable are symmetric the usual covariance adjusted estimate based on random samples is almost unbiased, and (4) the combination of regression adjustment in matched samples generally produces the least biased estimate.

1. INTRODUCTION

This paper is an extension of Rubin [1973] to include regression adjusted estimates and parallel nonlinear response surfaces. The reader is referred to sections 1 and 2 of that paper for the statement of the general problem and an introduction to the notation.

After presenting the estimates of the treatment effect to be considered in the remainder of section 1, we go on in section 2 to present Monte Carlo results for the expected bias of the estimates assuming four exponential response surfaces, normally distributed X, and the random order, nearest available matching method. Section 3 is an attempt to understand the Monte Carlo results in a more general context by examining the bias of the estimates for quadratic response surfaces. Section 4 is a summary of the results.

1.1 *The five estimates of τ to be considered here*

We assume that the objective is to estimate the constant difference, τ, between parallel univariate response surfaces in two populations P_1 and P_2 :

$$\tau = R_1(x) - R_2(x) \text{ for all } x,$$

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

where $R_i(x)$ is the conditional expectation in P_i of the dependent variable Y given the matching variable $X = x$. Equivalently, we can write

$$R_i(x) = \alpha_i + V(x) \quad i = 1, 2 \tag{1.1.1}$$

where $V(0) = 0$ and $\alpha_1 - \alpha_2 = \tau$.

We often refer to the function $V(x)$ as "the response surface".

We assume G_1 is a random sample from P_1 of size N and G_2 is a random sample from P_2 of size rN, $r \geq 1$. τ will be estimated from G_1 and G_{2*}, an N size subsample of G_2 "matched" to G_1. For the jth matched pair of subjects in G_1 and G_{2*} with scores y_{ij} and x_{ij} on Y and X we write

$$y_{ij} = \alpha_i + v_{ij} + e_{ij} \tag{1.1.2}$$

where $v_{ij} = V(x_{ij})$, $E_c(e_{ij}) = 0$, $i = 1, 2$, and E_c is the expectation conditionally given the x_{ij}.

The simplest estimate of τ is average Y difference in G_1 and G_{2*}

$$\hat{\tau}_0 = \bar{y}_{1.} - \bar{y}_{2.}.$$

The other four estimates of τ we will consider here use an adjustment based on the assumption of a linear model, $V(x) = \beta x$ for a regression coefficient β, which we will temporarily assume to be correct. It is simple to show that the bias of $\hat{\tau}_0$ under this model is $\beta(\bar{x}_{1.} - \bar{x}_{2.})$, and hence if we knew β, the estimate $\bar{y}_{1.} - \bar{y}_{2.} - \beta(\bar{x}_{1.} - \bar{x}_{2.})$ would be unbiased. We can obtain an estimate of β, say $\hat{\beta}$, that is conditionally unbiased, $E_c(\hat{\beta}) = \beta$, by fitting a regression model. Thus the estimate $\bar{y}_{1.} - \bar{y}_{2.} - \hat{\beta}(\bar{x}_{1.} - \bar{x}_{2.})$ would be an unbiased estimate of τ under the linear model whether we have matched or not.

Probably the most common estimate of β, at least when dealing with random samples, comes from fitting the parallel linear response surface model by least squares. After fitting the means to each group the data are pooled and a pooled estimate of β, $\hat{\beta}_p$ is found. The estimate of τ is then

$$\hat{\tau}_p = (\bar{y}_{1.} - \bar{y}_{2.}) - \hat{\beta}_p(\bar{x}_{1.} - \bar{x}_{2.}).$$

This method is the standard approach of the analysis of covariance for two groups, and the estimate is of course unbiased under this model of parallel linear response surfaces.

Two more estimates of the regression coefficient are easily found. Assuming that the parallel linear response surface model is correct, the least squares estimate of β found from the G_1 sample, $\hat{\beta}$, is an unbiased estimate of β, as is the estimate found from the G_2 sample, $\hat{\beta}_2$. Hence, we have two more unbiased estimates of the regression coefficient, one estimated from the G_1 data and the other estimated from the G_{2*} data, and so two regression adjusted estimates of τ,

$$\hat{\tau}_1 = (\bar{y}_{1.} - \bar{y}_{2.}) - \hat{\beta}_1(\bar{x}_{1.} - \bar{x}_{2.})$$

$$\hat{\tau}_2 = (\bar{y}_{1.} - \bar{y}_{2.}) - \hat{\beta}_2(\bar{x}_{1.} - \bar{x}_{2.}).$$

These estimates of τ using within group estimates of β are most appropriate when the response surfaces are thought to be non-parallel and an average difference over the G_1 or G_2 sample is desired. See Cochran [1969] and Belsen [1956]. If the response surfaces are parallel and linear, these estimates will have larger variances than the pooled estimate, $\hat{\tau}_p$, because one is not using half of the data relevant to estimating β.

The last estimate of β to be considered is in some sense the most natural one when dealing with pair-matched data. Forming matched pair differences $y_{di} = y_{1i} - y_{2i}$ and $x_{di} = x_{1i} - x_{2i}$, $\hat{\beta}_d$ is the estimate of β found from the regression of y_{di} on x_{di} . Equivalently $\hat{\beta}_d$ is the estimate of β found from a two-way analysis of covariance, groups by matched pairs. It is easy to show that $\hat{\beta}_d$ is an unbiased estimate of β under the linear response surface model. The associated estimate of τ,

$$\hat{\tau}_d = (\bar{y}_{1.} - \bar{y}_{2.}) - \hat{\beta}_d(\bar{x}_{1.} - \bar{x}_{2.}),$$

is the constant in the linear regression on matched pair differences.

We have considered five estimates of the difference between parallel response surfaces, τ, all of the form

$$\hat{\tau} = (\bar{y}_{1.} - \bar{y}_{2.}) - \hat{\beta}(\bar{x}_{1.} - \bar{x}_{2.}). \tag{1.1.3}$$

The differences between the estimates are thus confined to estimating the regression coefficient and are summarized below in Table 1. Note that $\hat{\tau}_d$ is the only estimate that requires $G_1 - G_2$ pairs assigned in the final samples.

TABLE 1

ESTIMATES OF THE RESPONSE SURFACE DIFFERENCE: $\hat{\tau} = \bar{y}_{1.} - \bar{y}_{2.} - \hat{\beta}(\bar{x}_{1.} - \bar{x}_{2.})$

Estimate of τ: $\hat{\tau}$	Estimate of β: $\hat{\beta}$
$\hat{\tau}_o$	$\hat{\beta}_o \equiv 0$
$\hat{\tau}_1$	$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$; $S_{xu} = \sum_j (x_{1j} - \bar{x}_{1.})\, u_{1j}$
$\hat{\tau}_2$	$\hat{\beta}_2 = \frac{S_{xy}}{S_{xx}}$; $S_{xu} = \sum_j (x_{2j} - \bar{x}_{2.})\, u_{2j}$
$\hat{\tau}_p$	$\hat{\beta}_p = \frac{S_{xy}}{S_{xx}}$; $S_{xu} = \sum_{i=1}^{2} \sum_{j=1}^{N} (x_{ij} - \bar{x}_{i.})\, u_{ij}$
$\hat{\tau}_d$	$\hat{\beta}_d = \frac{S_{xy}}{S_{xx}}$; $S_{xu} = \sum_{i=1}^{2} \sum_{j=1}^{N} (x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x}_{..})\, u_{ij}$

1.2 *The percent reduction in bias due to matched sampling and regression adjustment*

We now find the bias of the five estimates presented above. By Table 1 and (1.1.2) we have

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{S_{xv}}{S_{xx}} + \frac{S_{xe}}{S_{xx}} \left(\text{for} \quad \hat{\tau}_0 \,, \frac{S_{xv}}{S_{xx}} = \frac{S_{xe}}{S_{xx}} \equiv 0 \right)$$

and from (1.1.3)

$$\hat{\tau} = \alpha_1 - \alpha_2 + \bar{v}_{1.} - \bar{v}_{2.} + \bar{e}_{1.} - \bar{e}_{2.} - \frac{S_{xv}}{S_{xx}} (\bar{x}_{1.} - \bar{x}_{2.}) - \frac{S_{xe}}{S_{xx}} (\bar{x}_{1.} - \bar{x}_{2.}).$$

Thus the conditional bias of $\hat{\tau}$ given the x_{ij} is

$$E_c(\hat{\tau} - \tau) = \bar{v}_{1.} - \bar{v}_{2.} - \frac{S_{xv}}{S_{xx}} (\bar{x}_{1.} - \bar{x}_{2.}).$$

If the response surface is linear $(\bar{v}_{1.} - \bar{v}_{2.}) = \beta(\bar{x}_{1.} - \bar{x}_{2.})$ and $S_{xv}/S_{xx} = \beta$ for all estimates except $\hat{\tau}_0$ for which $S_{xv}/S_{xx} = 0$. Hence given parallel linear response surfaces, all estimates except $\hat{\tau}_0$ are unbiased and $\hat{\tau}_0$ will be unbiased if $\bar{x}_{1.} = \bar{x}_{2.}$. However, if the response surface is nonlinear all estimates are in general biased even if $\bar{x}_{1.} = \bar{x}_{2.}$. Thus a mean-matching method or a procedure that concludes unbiased estimates will result if $\bar{x}_{1.} = \bar{x}_{2.}$ is not necessarily appropriate if the response surface is non-linear. See Cochran [1970] and Rubin [1973] for examples of such procedures.

The expected bias of $\hat{\tau}$ over the matched sampling plan is $E\{\bar{v}_{1.} - \bar{v}_{2.} - S_{xv}/S_{xx}(\bar{x}_{1.} - \bar{x}_{2.})\}$ where E is the expectation over the distributions of X in matched samples. Given $r = 1$ (random samples), the expected bias of $\hat{\tau}_0$ is $E_1(\bar{v}_{1.}) - E_2(\bar{v}_{2.})$ where $E_i(\ )$ is the expectation over the distribution of X in P_i . It follows that the percent reduction in expected bias due to matching and/or regression adjustment is

$$100\left[1 - \frac{E\left[\bar{v}_{1.} - \bar{v}_{2.} - \frac{S_{xv}}{S_{xx}} (\bar{x}_{1.} - \bar{x}_{2.})\right]}{E_1(\bar{v}_{1.}) - E_2(\bar{v}_{2.})}\right]. \tag{1.2.1}$$

If the matches were exact, $x_{1j} = x_{2j}$, $j = 1, \cdots, N$ implying that $\bar{x}_{1.} = \bar{x}_{2.}$ and $\bar{v}_{1.} = \bar{v}_{2.}$; hence, the percent reduction in expected bias would be 100% for any response surface. In section 2 we present Monte Carlo values of the percent reduction in bias due to matching and/or regression adjustment for the estimates $\hat{\tau}_0$, $\hat{\tau}_1$, $\hat{\tau}_2$, $\hat{\tau}_p$, and $\hat{\tau}_d$ for some moderately non-linear response surfaces and imperfectly matched samples.

2. MONTE CARLO PERCENT REDUCTIONS IN BIAS

When dealing with finite matched samples, the expectations required to calculate the percent reductions in bias are usually analytically intractable. Hence, we will turn to Monte Carlo methods in order to obtain numerical

values for the percent reduction in bias of the different estimates in "typical" situations. These numerical values will be used to compare and evaluate the different estimators of τ. After specifying the conditions for the Monte Carlo investigations in section 2.1, we will present the Monte Carlo results in sections 2.2, 2.3, and 2.4.

2.1 *Conditions of the Monte Carlo investigation*

There are four conditions that must be specified in order to obtain Monte Carlo percent reductions in bias for the five estimates of τ.

1. the distribution of X in P_1 and P_2
2. the sample sizes N and rN
3. the matching method
4. the response surface $V(x)$.

We will assume that in P_i , $X \sim$ Normal (η_i , σ_i^2) $i = 1, 2$. Without loss of generality we can assume $\eta_1 = -\eta_2 \geq 0$ and $(\sigma_1^2 + \sigma_2^2)/2 = 1$. Then $B = 2\eta_1$ is the number of standard deviations $(\sqrt{(\sigma_1^2 + \sigma_2^2)/2})$ between the means of X. The choice of X as normal is obvious but restrictive; generalizing the Monte Carlo results to other distributions of X will be considered in section 3. Some limited experience indicates that the values $B = \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1$ and $\sigma_1^2/\sigma_2^2 = \frac{1}{2}, 1, 2$ are representative of the range that might occur in practice and so will be used.

With respect to sample sizes, we will assume $N = 50$ and $r = 2, 3, 4$. Previous work on matching, Rubin [1973], and preliminary results indicated very consistent trends for moderate and large N so that additional N's were judged unnecessary. Values of r were chosen to represent typical values that might be used in practice.

The matching method must assign to each subject in G_1 a distinct subject in G_2 as a match so that there are N matched pairs. We will assume the random order, nearest available pair-matching method investigated in Rubin [1973]. First randomly order the G_1 subjects, then for each G_1 subject choose in turn the closest match from those G_2 subjects not yet assigned as matches. This method was chosen for two basic reasons: (1) previous results indicate that it is a relatively intelligent pair-matching method that might be used in practice; and (2) the matching procedure is very fast to perform by computer. Since our study requires extensive Monte Carlo runs, the second point was of practical importance.

Some basic results for this matching method are given in Table 2 and Table 3, for the values of N and r and distribution of X specified above. Table 2 gives Monte Carlo percent reductions in bias for $\hat{\tau}_0$ assuming linear response surfaces, and Table 3 gives Monte Carlo values of the ratio of the expected variance of X in matched G_{2*} samples, $E(s_2^2)$, to the expected variance of X in random G_2 samples, σ_2^2 . Note that $E(s_2^2)$ is for these conditions always less than σ_1^2 , although in "easy conditions" (those in which the percent reduction in the bias of $X > 90\%$) $E(s_2^2)$ is close to σ_1^2 .

TABLE 2

PERCENT REDUCTION IN BIAS OF X FOR RANDOM ORDER, NEAREST AVAILABLE PAIR-MATCHING: X NORMAL, $N = 50$

	$\sigma_1^2/\sigma_2^2 = \frac{1}{2}$			$\sigma_1^2/\sigma_2^2 = 1$			$\sigma_1^2/\sigma_2^2 = 2$		
r =	2	3	4	2	3	4	2	3	4
B = 1/4	99	100	100	92	96	98	66	79	86
1/2	98	99	100	87	95	97	59	75	81
3/4	93	99	100	78	91	94	53	69	75
1	84	97	99	69	84	89	51	63	71

TABLE 3

$(E(s_2^2)/\sigma_2^2) \times 100$ FOR RANDOM ORDER, NEAREST AVAILABLE PAIR-MATCHING: X NORMAL, $N = 50$

	$\sigma_1^2/\sigma_2^2 = \frac{1}{2}$			$\sigma_1^2/\sigma_2^2 = 1$			$\sigma_1^2/\sigma_2^2 = 2$		
r =	2	3	4	2	3	4	2	3	4
B = 1/4	49	48	48	92	92	96	134	149	157
1/2	49	49	48	82	90	94	116	132	149
3/4	48	48	48	69	83	87	99	114	125
1	45	45	47	59	68	76	76	91	100

The last condition to be specified is the range of nonlinear response surfaces, the functions $V(x)$ to be used. Since we are investigating regression adjusted estimates based on the model that the response surface $V(x)$ is linear, we will require the response surface to be moderately linear in the range of interest on the assumption that an alert investigator should be able to detect violent nonlinearity and thus use a more appropriate model for adjusting the estimate (e.g. add a quadratic term).

In Figure 1 we have plotted $V(x) = \exp(x/2)$ and $V(x) = \exp(x)$,[2] for $-3 < x < 3$ corresponding roughly to the range of X to be used in the Monte Carlo samples. Even when disguised by random error, it might reasonably be argued that the nonlinear aspects of $\exp(x)$ often would be quite apparent. For this range of X we consider the response surface $\exp(x)$

[2] Strictly by equation (1.1.1), $V(0) = 0$ so that we really should set $V(x) = \exp(x) - 1$. Since the constant has no effect on results it will be simpler to ignore it in discussion.

to be an extreme example of what might be termed moderately nonlinear and exp $(x/2)$ a more reasonable example of a moderately nonlinear response surface.

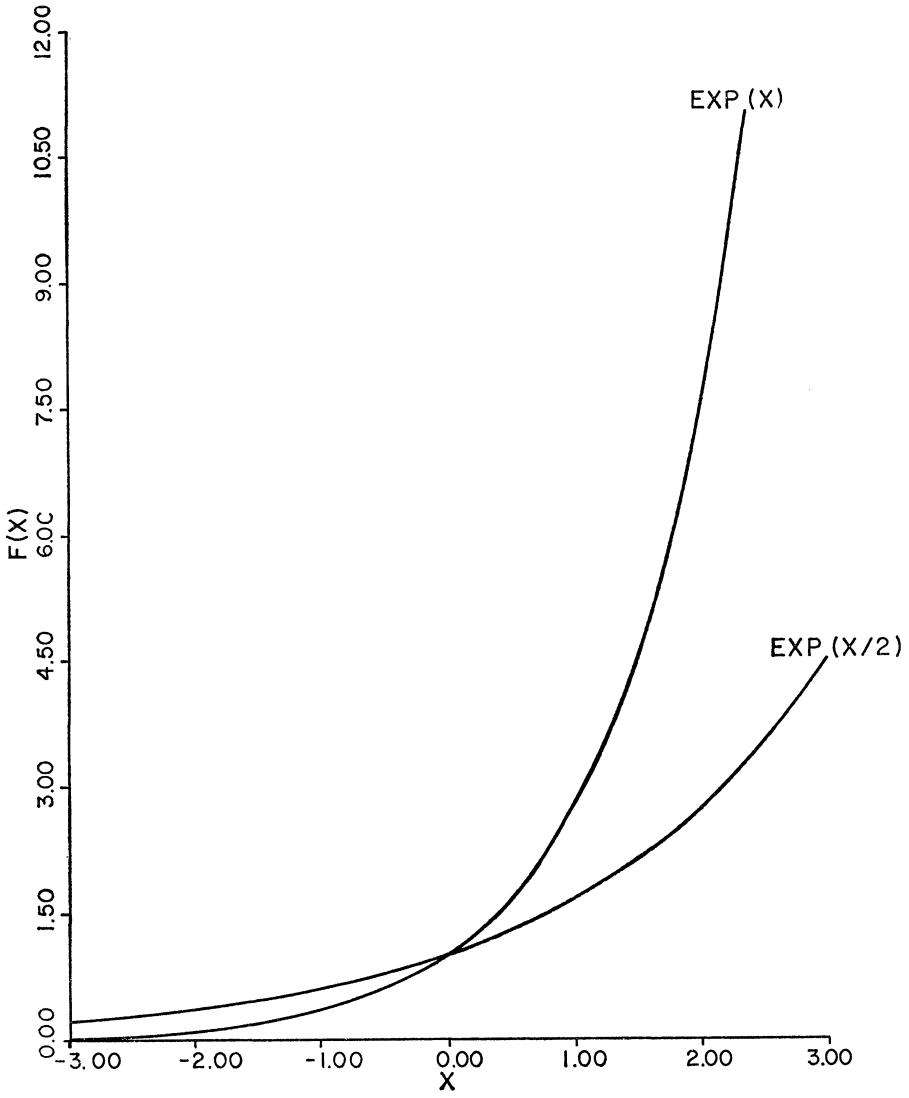

FIGURE 1

TWO EXPONENTIAL RESPONSE SURFACES

Since $\eta_1 \geq \eta_2$, the response surfaces exp (x) and exp $(x/2)$ are "unfavorable to matching" because they are increasing most rapidly for large X which is where close matches are difficult to find. The response surface exp (x) is, in this sense, more unfavorable to matching than exp $(x/2)$. Also, the response surfaces exp $(-x)$ and exp $(-x/2)$ are "favorable to matching" because

they are changing very slowly where the matches are poor. The four exponential response surfaces exp (ax), $a = \pm\frac{1}{2}, \pm 1$, were investigated in detail. We consider these to be representative of a range of moderately nonlinear response surfaces that are both favorable and unfavorable to matching.

Since an investigator when deciding whether or not to obtain matched samples or when deciding which estimate of β to use generally has some knowledge of the distribution of the matching variable in P_1 and P_2, it seems logical to present the results of the investigation classified by the distribution of the matching variable. We first present the results when the distribution of X is favorable to matching: $\sigma_1^2/\sigma_2^2 = \frac{1}{2}$, then when only the mean of X differs in P_1 and P_2, $\sigma_1^2 = \sigma_2^2$, and finally when the distribution of X is unfavorable to matching, $\sigma_1^2/\sigma_2^2 = 2$. "Favorable" refers to the previous work on ability to remove the bias of the matching variable X.

2.2 *Case I-distribution of X favorable to matching*—$\sigma_1^2/\sigma_2^2 = \frac{1}{2}$

The results for case I in which the distribution of X is favorable to matching ($\sigma_1^2/\sigma_2^2 = \frac{1}{2}$) are given in Table 4. First consider the results for $\hat{\tau}_0$ (no regression adjustment). $\hat{\tau}_0$ based on samples matched with $r \geq 2$ usually removes most of the bias (90%–110%) of $\hat{\tau}_0$ based on random samples.[3] With the largest initial bias it may remove quite a bit less (e.g. $B = 1, r = 2$, 76% for exp $(x/2)$ and 69% for exp (x)). For practical purposes, the previous results in Table 2 for linear response surfaces can be considered slightly optimistic typical values (in general, optimistic by less than 5%) for moderately nonlinear response surfaces) when $\sigma_1^2/\sigma_2^2 = \frac{1}{2}$.

Regression adjusted estimates based on random samples seem quite sensitive to even moderate departures from linearity. The extremely wild values for percent reduction in bias (e.g. -288, 306) for exp (x) and exp $(x/2)$ when $B = \frac{1}{4}$ and for exp (x) when $B = \frac{1}{2}$ indicate that in these cases the regression adjusted estimates based on random samples are substantially more biased than $\hat{\tau}_0$ based on random samples. Since exp (x) and especially exp $(x/2)$ are not violently nonlinear these results are somewhat surprising.

The explanations seem to be as follows. Since exp (ax) $(a > 0)$ is monotonically increasing (implying that generally $\hat{\beta} > 0$) and $\eta_1 > \eta_2$, the adjustment $-\hat{\beta}(\eta_1 - \eta_2)$ is usually negative in this case. But if $\sigma_1^2 < (\sigma_2^2 - 2B/a)$ we have $E_1\{\exp(ax)\} < E_2\{\exp(ax)\}$, so that a *positive* adjustment is needed. Hence for $a = 1$, $\sigma_1^2 = \frac{2}{3}$, $\sigma_2^2 = \frac{4}{3}$, and $B = \frac{1}{4}$, the regression adjustment often greatly increases the original bias. Further, for $a = \frac{1}{2}$, $B = \frac{1}{4}$ as well as for $a = 1$, $B = \frac{1}{2}$, the regression adjustment is a gross over-adjustment, since the expected bias of $\hat{\tau}_0$ based on random samples is much less than suggested by the difference in means and the "average slope" of exp $(x/2)$. There is no one regression adjusted estimate based on random samples which is always best, although $\hat{\tau}_d$ appears to be more consistent than any other estimate. Also, $\hat{\tau}_1$ is always very poor, which is not surprising since the

[3] Values greater than 100% indicate that the expected value of $V(x)$ in matched G_{2*} samples is greater than the expected value of $V(x)$ in P_1, while values less than 0% indicate that the expected value of $V(x)$ in matched G_{2*} samples is less than the expected value of $V(x)$ in P_2.

TABLE 4

PERCENT REDUCTION IN BIAS, X NORMAL, $N = 50$, CASE I: $\sigma_1^2/\sigma_2^2 = \frac{1}{2}$

		$r = 1$ (random)				$r = 2$				$r = 3$				$r = 4$			
	Response Surface	exp(x)	exp(x/2)	exp(-x/2)	exp(-x)	exp(x)	exp(x/2)	exp(-x/2)	exp(-x)	exp(x)	exp(x/2)	exp(-x/2)	exp(-x)	exp(x)	exp(x/2)	exp(-x/2)	exp(-x)
	Estimate																
	$\hat{\tau}_0$	00	00	.00	00	106	95	99	100	103	99	100	100	102	99	100	100
	$\hat{\tau}_1$	-288	306	54	32	101	100	100	100	102	99	100	100	101	100	100	100
$B = \frac{1}{4}$	$\hat{\tau}_2$	-313	294	65	57	102	100	100	100	102	99	100	100	101	100	100	100
	$\hat{\tau}_p$	-304	298	62	48	101	100	100	100	102	99	100	100	101	100	100	100
	$\hat{\tau}_d$	-228	248	79	77	96	102	100	100	101	100	100	100	100	100	100	100
	$\hat{\tau}_0$	00	00	00	00	93	96	99	99	94	98	100	100	97	99	100	100
	$\hat{\tau}_1$	326	163	63	38	101	100	100	100	96	99	100	100	98	100	100	100
$B = \frac{1}{2}$	$\hat{\tau}_2$	276	138	88	88	101	100	100	100	95	99	100	100	98	100	100	100
	$\hat{\tau}_p$	292	146	80	72	101	100	100	100	96	99	100	100	98	100	100	100
	$\hat{\tau}_d$	226	125	96	104	108	101	100	100	101	100	100	100	100	100	100	100
	$\hat{\tau}_0$	00	00	00	00	85	89	96	98	94	97	100	100	97	99	100	100
	$\hat{\tau}_1$	220	148	65	38	102	100	101	101	97	99	100	100	98	100	100	100
$B = \frac{3}{4}$	$\hat{\tau}_2$	145	111	102	113	103	100	101	101	96	99	100	100	98	100	100	100
	$\hat{\tau}_p$	170	123	90	88	103	100	101	101	97	99	100	100	98	100	100	100
	$\hat{\tau}_d$	137	109	103	114	113	103	99	100	102	100	100	100	101	100	100	100
	$\hat{\tau}_0$	00	00	00	00	69	76	91	96	90	94	98	99	94	97	99	100
	$\hat{\tau}_1$	206	147	63	36	106	101	101	101	98	99	101	101	97	99	100	100
$B = 1$	$\hat{\tau}_2$	106	97	113	136	102	100	101	102	97	99	101	101	97	99	100	100
	$\hat{\tau}_p$	139	113	96	102	105	101	101	101	98	101	101	101	97	99	100	100
	$\hat{\tau}_d$	118	103	108	126	118	105	99	99	105	99	99	100	102	101	100	100

range of X in G_1 is limited compared to the range in random G_1 and G_2 samples, implying that $\hat{\tau}_1$ may give a poor linear approximation to the response surface.

Regression adjusted estimates based on samples matched with $r \geq 2$ are far superior to those based on random samples, the difference being more striking when the bias in X is small. In all conditions all of the regression adjusted estimates based on matched samples ($r \geq 2$) can be expected to remove all of the bias (98%–102%). However, there is some tendency for $\hat{\tau}_d$ to be inferior if the response surface is very unfavorable to matching (i.e. exp (x)). Except for $\hat{\tau}_d$ given the response surface exp (x), there is almost no improvement in using greater than a 2:1 ratio of sample sizes.

2.3 *Case II*—$\sigma_1^2 = \sigma_2^2$

The percent reductions in bias for case II ($\sigma_1^2 = \sigma_2^2$) are given in Table 5. First consider $\hat{\tau}_0$. Given the response surface, B, and $r > 1$, $\hat{\tau}_0$ is more biased

TABLE 5

PERCENT REDUCTION IN BIAS, X NORMAL, $N = 50$, CASE II: $\sigma_1^2/\sigma_2^2 = 1$

		r = 1 (random)				r = 2				r = 3				r = 4			
Response Surface		exp(x)	exp(x/2)	exp(-x/2)	exp(-x)	exp(x)	exp(x/2)	exp(-x/2)	exp(-x)	exp(x)	exp(x/2)	exp(-x/2)	exp(-x)	exp(x)	exp(x/2)	exp(-x/2)	exp(-x)
Estimate																	
	$\hat{\tau}_0$	00	00	00	00	70	83	99	106	79	90	101	104	87	94	101	103
	$\hat{\tau}_1$	113	106	94	88	80	91	107	112	84	94	103	106	90	97	102	104
$B = \frac{1}{4}$	$\hat{\tau}_2$	88	94	106	113	77	91	107	113	82	94	104	107	90	96	102	105
	$\hat{\tau}_p$	101	100	100	101	79	91	107	113	83	94	104	107	89	96	102	105
	$\hat{\tau}_d$	109	101	107	125	100	99	103	108	100	100	101	103	100	100	101	102
	$\hat{\tau}_0$	00	00	00	00	60	74	94	98	75	87	98	100	84	92	99	100
	$\hat{\tau}_1$	127	113	88	77	80	91	105	108	83	94	103	104	88	96	102	102
$B = \frac{1}{2}$	$\hat{\tau}_2$	77	88	113	127	74	89	107	110	81	93	103	104	87	96	102	103
	$\hat{\tau}_p$	102	101	101	102	77	90	106	109	82	94	103	104	88	96	102	103
	$\hat{\tau}_d$	109	102	105	116	106	102	100	101	102	100	100	101	101	100	100	101
	$\hat{\tau}_0$	00	00	00	00	47	62	87	94	68	81	96	99	76	87	98	100
	$\hat{\tau}_1$	142	120	82	67	83	92	106	109	88	93	104	105	86	95	102	103
$B = \frac{3}{4}$	$\hat{\tau}_2$	67	82	120	142	71	87	109	112	78	91	105	106	83	94	103	104
	$\hat{\tau}_p$	104	101	101	104	79	90	107	110	81	92	104	105	84	94	103	103
	$\hat{\tau}_d$	111	102	104	114	110	103	99	100	105	102	99	100	103	101	100	100
	$\hat{\tau}_0$	00	00	00	00	39	53	82	91	55	70	92	97	65	79	96	99
	$\hat{\tau}_1$	158	127	77	58	88	93	106	109	82	92	105	106	82	93	103	104
$B = 1$	$\hat{\tau}_2$	58	77	127	158	68	85	110	114	72	88	107	109	75	90	105	105
	$\hat{\tau}_p$	108	102	102	108	81	90	107	111	79	91	105	107	79	92	104	105
	$\hat{\tau}_d$	112	102	106	120	113	104	99	99	109	103	100	100	106	102	100	99

in case II ($\sigma_1^2 = \sigma_2^2$) than in case I ($\sigma_1^2 = \sigma_2^2/2$). This result is expected since in case II the distribution of X is less favorable to matching than in case I. The values given in Table 2 for linear response surfaces can be considered mildly optimistic typical values (optimistic by 5%–10%) for these non-linear response surfaces when $\sigma_1^2 = \sigma_2^2$.

The results for the regression adjusted estimates in case II are surprising when compared with the results in case I. First, the regression adjusted

estimates based on random samples are much better in all conditions in case II than in case I. In fact, $\hat{\tau}_p$ based on random samples generally removes almost all of the bias (98%–102%). A possible explanation for this result is that since the variances and higher moments of X about the mean are equal in random samples from the two populations, all of the bias of $\hat{\tau}_0$ is due to the difference in the means of X which linear regression should be good at removing. This comment implies that regression adjusted estimates should be approximately unbiased for nonlinear response surfaces (that can be approximated by a polynomial) whenever the distributions of X in G_1 and G_2 are the same except for a difference in means. In section 3 we will see that this comment is not completely accurate and that the results we are discussing are somewhat dependent upon the symmetry of the normal distribution.

In samples matched with $r \geq 2$ $\hat{\tau}_d$ is the best estimate, in general removing most of the bias (99%–105%). The other regression adjusted estimates based on samples matched with $r \geq 2$ often have values outside the range 90%–110%. $\hat{\tau}_2$ is especially poor for matched samples possibly because of the small range of X in G_{2*} on which to base the regression (see Table 2). $\hat{\tau}_1$ and $\hat{\tau}_2$ are somewhat better when based on matched samples than when based on random samples. Surprisingly, $\hat{\tau}_p$ is often worse when $r \geq 2$ than when $r = 1$.

As would be expected, the estimates generally become slightly less biased as r increases from two to four.

2.4 *Case III—distribution of matching variable unfavorable to matching—* $\sigma_1^2/\sigma_2^2 = 2$

The results for case III in which the distribution of X is unfavorable to matching ($\sigma_1^2/\sigma_2^2 = 2$) are given in Table 6. First consider the results for $\hat{\tau}_0$. As expected from the results for linear response surfaces, $\hat{\tau}_0$ based on matched samples is more biased when $\sigma_1^2 > \sigma_2^2$ than when $\sigma_1^2 \leq \sigma_2^2$. In fact, for $\exp(-x)$ and $B = \frac{1}{4}$, $\hat{\tau}_0$ based on samples matched with $4 > r \geq 2$ can be more biased than $\hat{\tau}_0$ based on random samples. This strange result is due to the same circumstances as mentioned previously in case I when discussing regression adjusted estimates based on random samples. Even though $\exp(-x)$ is monotonically decreasing, in case III when $B = \frac{1}{4}$, $E_1\{\exp(-x)\} > E_2\{\exp(-x)\}$. Matching with $r < 4$ decreases the difference in means in G_1 and G_{2*} more than the difference in variances which tends to increase the bias of $\hat{\tau}_0$. In general, however, the values for percent reduction in bias of $\hat{\tau}_0$ for linear response surfaces given in Table 2 can be considered representative optimistic values (optimistic by about 10%) for moderately nonlinear response surfaces when $\sigma_1^2/\sigma_2^2 = 2$.

As already observed in case I, regression adjusted estimates based on random samples are quite biased when $\sigma_1^2 \neq \sigma_2^2$ even for moderately linear response surfaces. Also, as in case I, given the response surface and B, regression adjusted estimates based on samples matched with $r \geq 2$ are less biased than when based on random samples. However, the estimates are more

TABLE 6

PERCENT REDUCTION IN BIAS, X NORMAL, $N = 50$, CASE III: $\sigma_1^2/\sigma_2^2 = 2$

		r = 1 (random)				r = 2				r = 3				r = 4			
Response Surface		exp(x)	exp(x/2)	exp(-x/2)	exp(-x)	exp(x)	exp(x/2)	exp(-x/2)	exp(-x)	exp(x)	exp(x/2)	exp(-x/2)	exp(-x)	exp(x)	exp(x/2)	exp(-x/2)	exp(-x)
Estimate																	
	$\hat{\tau}_0$	00	00	00	00	35	48	121	-50	51	66	139	-48	55	70	120	01
	$\hat{\tau}_1$	57	65	294	-313	57	73	216	-141	64	79	184	-85	67	82	164	-40
B = $\frac{1}{4}$	$\hat{\tau}_2$	32	54	306	-288	47	68	227	-149	57	76	192	-96	61	80	170	-45
	$\hat{\tau}_p$	48	62	298	-304	53	71	220	-144	61	78	187	-89	65	81	167	-42
	$\hat{\tau}_d$	84	83	240	-218	90	90	177	-99	92	93	149	-29	94	95	140	-05
	$\hat{\tau}_0$	00	00	00	00	30	45	81	123	43	60	89	118	48	65	94	126
	$\hat{\tau}_1$	88	88	138	276	66	80	132	220	68	83	123	181	68	84	121	177
B = $\frac{1}{2}$	$\hat{\tau}_2$	38	63	163	326	47	71	140	237	55	77	128	191	58	79	124	184
	$\hat{\tau}_p$	72	80	146	292	60	77	135	226	63	81	125	185	64	82	122	180
	$\hat{\tau}_d$	108	98	126	240	107	100	111	171	105	100	108	147	104	100	107	146
	$\hat{\tau}_0$	00	00	00	00	23	38	72	90	39	55	85	98	42	60	89	100
	$\hat{\tau}_1$	113	102	111	145	76	85	121	152	72	85	116	138	71	85	114	131
B = $\frac{3}{4}$	$\hat{\tau}_2$	38	65	148	220	44	70	133	173	54	76	123	150	54	78	119	140
	$\hat{\tau}_p$	88	90	123	170	66	80	125	159	66	82	119	142	65	83	116	109
	$\hat{\tau}_d$	123	106	106	132	120	106	102	115	111	103	102	112	111	103	101	97
	$\hat{\tau}_0$	00	00	00	00	16	31	67	83	28	45	79	92	29	50	84	94
	$\hat{\tau}_1$	136	113	97	106	88	90	114	132	78	87	114	129	74	86	113	124
B = 1	$\hat{\tau}_2$	36	63	147	206	40	69	130	158	46	73	125	146	46	74	121	138
	$\hat{\tau}_p$	102	96	113	139	76	85	118	139	69	83	117	134	66	82	115	128
	$\hat{\tau}_d$	133	111	101	114	127	109	99	104	119	106	100	104	119	105	99	102

biased in case III than in case I, presumably because the matching is poorer than in case I. As might be expected $\hat{\tau}_2$ is generally the worst regression adjusted estimate based on matched samples since the range of X in the matched G_{2*} samples is small compared to the range of X over both samples. $\hat{\tau}_d$ is the best estimate in case III, with $r \geq 2$ generally removing most (90%–110%) of the bias. In those conditions in which $\hat{\tau}_d$ does poorly the only better estimate is $\hat{\tau}_0$ which in general is not very satisfactory. The advantages of matching with $r = 4$ rather than with $r = 2$ are greater in case III than in cases I or II but still are not substantial.

3. LINEAR REGRESSION ADJUSTMENT AND QUADRATIC RESPONSE SURFACES

Somewhat surprisingly there appears to be little literature on the use of linear regression to remove bias when the response surfaces are not exactly

linear. A theoretical study of this situation would be especially valuable for interpreting results such as those presented in section 2. The following discussion is an attempt to understand the preceding Monte Carlo results within a more general framework than provided by normal distributions and exponential response surfaces. It is not intended to be a complete study of linear regression and nonlinear response surfaces. Assuming that the response surface is actually quadratic, we derive expressions for the bias of $\hat{\tau}_1$, $\hat{\tau}_2$ and $\hat{\tau}_p$ in section 3.1 and for the bias of $\hat{\tau}_d$ in section 3.2. Since exp (ax) may be approximated by such a response surface in the range of conditions considered, these expressions will be used to help interpret the Monte Carlo results of section 2.

3.1 *Bias of* $\hat{\tau}_0$, $\hat{\tau}_1$, $\hat{\tau}_2$ *and* $\hat{\tau}_p$

Consider two samples of size N with means $\bar{x}_{i.}$, $i = 1, 2$, sample variances $s_i^2 = \Sigma(x_{ij} - \bar{x}_{i.})^2/N$, $i = 1, 2$ and sample skewness $k_i = \Sigma(x_{ij} - \bar{x}_{i.})^3/N$, $i = 1, 2$. We will assume that the true response surface can be accurately approximated by a quadratic response surface: $V(x) = \beta x + \delta x^2$. Hence, for the samples

$$y_{ij} = \alpha_i + \beta x_{ij} + \delta x_{ij}^2 + e_{ij} \tag{3.1.1}$$

where $E_c(e_{ij}) = 0$, and $E_c(\ )$ is the expectation conditionally given the x_{ij} . Using the simple results

$$\begin{aligned} \frac{1}{N}\sum x_{ij}^2 &= \bar{x}_{i.}^2 + s_i^2 \\ \frac{1}{N} E_c \sum[(x_{ij} - \bar{x}_{i.})y_{ij}] &= \beta s_i^2 + \delta(2\bar{x}_{i.}s_i^2 + k_i) \end{aligned} \tag{3.1.2}$$

we can calculate the bias of $\hat{\tau}_0$, $\hat{\tau}_1$, $\hat{\tau}_2$ and $\hat{\tau}_p$ as follows.

$$E_c(\hat{\tau}_0 - \tau) = \beta(\bar{x}_{1.} - \bar{x}_{2.}) + \delta(\bar{x}_{1.}^2 - \bar{x}_{2.}^2) + \delta(s_1^2 - s_2^2), \tag{3.1.3}$$

$$E_c(\hat{\tau}_i - \tau) = \delta\left\{(s_1^2 - s_2^2) \mp (\bar{x}_{1.} - \bar{x}_{2.})^2 - (\bar{x}_{1.} - \bar{x}_{2.})\frac{k_i}{s_i^2}\right\} \tag{3.1.4}$$

where the $-$ holds for $\hat{\tau}_1$ and the $+$ for $\hat{\tau}_2$, and

$$\begin{aligned} E_c(\hat{\tau}_p - \tau) = \delta\Bigg[(s_1^2 - s_2^2) + (\bar{x}_{1.} - \bar{x}_{2.})\left\{(\bar{x}_{1.} + \bar{x}_{2.}) - \frac{2(\bar{x}_{1.}s_1^2 + \bar{x}_{2.}s_2^2)}{s_1^2 + s_2^2}\right\} \\ - \frac{\bar{x}_{1.} - \bar{x}_{2.}}{s_1^2 + s_2^2}(k_1 + k_2)\Bigg]. \end{aligned} \tag{3.1.5}$$

We now use expressions (3.1.4) and (3.1.5) to interpret the Monte Carlo results of section 2. First consider random samples. For the results presented $E(\bar{x}_{1.}) = B/2$, $E(\bar{x}_{2.}) = -B/2$, $E(s_i^2) \doteq \sigma_i^2$, $i = 1, 2$, $E(s_1^2 + s_2^2)/2 \doteq 1$, $E(k_1) = 0$ and $E(k_2) = 0$ since the normal is symmetric. Thus for the random samples considered in section 3 the expected bias of $\hat{\tau}_i$ $i = 1, 2$ is approximately

$$\delta\{(\sigma_1^2 - \sigma_2^2) \mp B^2\}, \tag{3.1.6}$$

and for $\hat{\tau}_p$ the expected bias is approximately

$$\delta(\sigma_1^2 - \sigma_2^2)(1 + B^2/2). \tag{3.1.7}$$

Hence in case II when $\sigma_1^2 = \sigma_2^2$, $\hat{\tau}_p$ should be approximately unbiased, while both $\hat{\tau}_1$ and $\hat{\tau}_2$ should be biased by an amount $\mp\delta B^2$. This claim is substantiated by the Monte Carlo results presented in section 3. In case I, $\sigma_1^2 = \frac{2}{3}$, $\sigma_2^2 = \frac{4}{3}$. Since $1 \geq B > 0$, from (3.1.6) and (3.1.7) the least biased estimate should be $\hat{\tau}_2$ followed by $\hat{\tau}_p$ and then $\hat{\tau}_1$. The Monte Carlo results again substantiate this claim. In case III $\sigma_1^2 = \frac{4}{3}$, $\sigma_2^2 = \frac{2}{3}$ and the ordering implied by (3.1.6) and (3.1.7) is $\hat{\tau}_2$, $\hat{\tau}_p$, $\hat{\tau}_1$, which is again in agreement with previous results.

Now consider case I and samples matched with $r \geq 2$: $E(s_1^2) \doteq E(s_2^2) \doteq \sigma_1^2$ (see Table 3). For $\hat{\tau}_p$ the expected bias is approximately

$$\delta\left\{-\frac{E(\bar{x}_{1.} - \bar{x}_{2.})^2}{2\sigma_1^2} E(k_1 + k_2)\right\}, \tag{3.1.8}$$

and for $\hat{\tau}_i$, $i = 1, 2$ the expected bias is approximately

$$\delta\left\{\mp E(\bar{x}_{1.} - x_{2.})^2 - \frac{E(\bar{x}_{1.} - \bar{x}_{2.})}{\sigma_1^2} E(k_i)\right\}. \tag{3.1.9}$$

In case I for matched samples with ($r \geq 2$), $E(\bar{x}_{1.} - \bar{x}_{2.}) \doteq 0$ (see Table 2); hence $\hat{\tau}_p$, $\hat{\tau}_1$ and $\hat{\tau}_2$ all have approximately zero expected bias. Notice however, that if $E(s_1^2) = E(s_2^2)$ but $E(\bar{x}_{1.} - \bar{x}_{2.})$ is large, $\hat{\tau}_p$ would have approximately zero expected bias only if $E(k_1 + k_2) \doteq 0$ (case II, random samples). Thus a situation in which there is a large bias in the mean of X but very similar higher moments ($s_1^2 \doteq s_2^2$, $k_1 \doteq k_2$) would not necessarily be favorable to using the estimate $\hat{\tau}_p$ unless the distribution of X is symmetric ($k_1 \doteq k_2 \doteq 0$).

In cases II and III for samples matched with $r \geq 2$, $\sigma_1^2 > E(s_2^2)$ (see Table 3) and $E(\bar{x}_{1.} - \bar{x}_{2.})$ is not trivial (see Table 2). Expressions (3.1.4) and (3.1.5) suggest that $\hat{\tau}_2$ should be the worst estimate since $E(s_1^2 - s_2^2) > 0$, $E(\bar{x}_{1.} - \bar{x}_{2.})^2 > 0$, and $E(k_2/s_2^2) < 0$ (results not presented indicate that in the conditions considered with $r > 1$, $E(k_2) < 0$). Also, $\hat{\tau}_1$ should be better than $\hat{\tau}_2$ or $\hat{\tau}_p$ in these cases because $E(k_1) \doteq 0$ and we are subtracting $E(\bar{x}_{1.} - \bar{x}_{2.})^2$ from $E(s_1^2 - s_2^2)$. The Monte Carlo results confirm these trends.

The above discussion has at least to some extent explained the Monte Carlo results in section 2 for $\hat{\tau}_1$, $\hat{\tau}_2$ and $\hat{\tau}_p$ and thus generated an understanding of the effect of these estimates which is not tied to normal distributions and exponential response surfaces.

3.2 *Bias of* $\hat{\tau}_d$

Thus far no explanation has been offered for the often superior performance of the regression adjusted estimate based on matched pair differences, $\hat{\tau}_d$. This omission was intentional because even the intuitive explanation given below is somewhat involved.

We begin by showing that for any set of N matched pairs there is some mth degree ($1 \leq m \leq N + 1$) non-trivial polynomial response surface for which the pairs are exactly matched. Let $P(x) = \sum_{k=1}^{m} a_k x^k$ be an mth degree

polynomial in X; then if all matched pairs are exactly matched with respect to P we have

$$\sum_{k=1}^{m} a_k(x_{1j}^k - x_{2j}^k) = 0, \qquad j = 1, \cdots, N.$$

Such a polynomial always exists if $m = N + 1$ because we can always find a non-trivial solution for $N + 1$ unknowns given N homogeneous linear equations. If the pairs are exactly matched, the minimum m is 1; the samples are exactly matched for linear response surfaces as well as for all higher degree response surfaces.

Intuitively, one might feel that as the matches become better the minimum degree of the response surface for which the pairs are exactly matched should decrease, or at least the degree of the response surface for which the pairs are almost exactly matched should decrease. In this discussion we will assume that the matched pairs are close enough so that they are almost exactly matched for some quadratic response surface:

$$a_1(x_{1j} - x_{2j}) + a_2(x_{1j}^2 - x_{2j}^2) = d_j \tag{3.2.1}$$

where d_j is small for all matched pairs and its average value over matched pairs is zero. Since we require the average value of d_j to be zero we have

$$0 = a_1(\bar{x}_{1.} - \bar{x}_{2.}) + a_2(\bar{x}_{1.}^2 + s_1^2 - \bar{x}_{2.}^2 - s_2^2).$$

If $\bar{x}_{1.} \equiv \bar{x}_{2.}$, all regression adjusted estimates considered are identical to $\hat{\tau}_0$. If $\bar{x}_{1.} \not\equiv \bar{x}_{2.}$ without loss of generality, let $a_2 = 1$, so that

$$a_1 = -\frac{(\bar{x}_{1.}^2 - \bar{x}_{2.}^2) + (s_1^2 - s_2^2)}{(\bar{x}_{1.} - \bar{x}_{2.})}. \tag{3.2.2}$$

The bias of $\hat{\tau}_d$ is

$$\begin{aligned} E_c(\hat{\tau}_d - \hat{\tau}) &= E_c\Bigg[(\bar{y}_{1.} - \bar{y}_{2.}) \\ &\quad - \frac{\sum \{x_{1j} - x_{2j} - (\bar{x}_{1.} - \bar{x}_{2.})\}(y_{1j} - y_{2j})}{\sum \{x_{1j} - x_{2j} - (\bar{x}_{1.} - \bar{x}_{2.})\}^2}(\bar{x}_{1.} - \bar{x}_{2.})\Bigg] \\ &= \delta\{(\bar{x}_{1.}^2 - \bar{x}_{2.}^2) + (s_1^2 - s_2^2)\} \\ &\quad - (\bar{x}_{1.} - \bar{x}_{2.})\frac{\delta \sum \{x_{1j} - x_{2j} - (\bar{x}_{1.} - \bar{x}_{2.})\}(x_{1j}^2 - x_{2j}^2)}{\sum \{x_{1j} - x_{2j} - (\bar{x}_{1.} - \bar{x}_{2.})\}^2}. \end{aligned} \tag{3.2.3}$$

But from (3.2.1) and (3.2.2) assuming $\bar{x}_{1.} \not\equiv \bar{x}_{2.}$, we have

$$(x_{1j}^2 - x_{2j}^2) = d_j + (x_{1j} - x_{2j})\frac{(\bar{x}_{1.}^2 - \bar{x}_{2.}^2) + (s_1^2 - s_2^2)}{(\bar{x}_{1.} - \bar{x}_{2.})}. \tag{3.2.4}$$

Inserting (3.2.4) into (3.2.3) and taking expectations over matched samples we have

$$E(\hat{\tau}_d - \tau) \doteq \delta A, \quad \text{where} \quad A = -\frac{E(\bar{x}_{1.} - \bar{x}_{2.})}{\text{Var}\,(x_{1j} - x_{2j})}\,\text{Cov}\,(x_{1j} - x_{2j}, d_j),$$

and Var ( ) and Cov ( ) are the variance and covariance over the distribution of x_{ij} in matched samples.

If the samples are well matched with respect to some quadratic response surface the d_i should be relatively small and relatively uncorrelated with the $x_{1i} - x_{2i}$, implying a small A. Hence, $\hat{\tau}_d$ should be approximately unbiased if the response surface is basically quadratic and the samples are well matched with respect to some other quadratic response surface. If the true response surface is exactly quadratic and the matches are exact with respect to any quadratic response ($d_i \equiv 0$), $\hat{\tau}_d$ will be conditionally unbiased. A might be large if the samples are very poorly matched as with random samples; thus, it is not surprising that for random samples in case II ($\sigma_1^2 = \sigma_2^2$), $\hat{\tau}_p$ is superior to $\hat{\tau}_d$. However, even in those situations in which $\hat{\tau}_d$ is not the best estimate, the expected bias of $\hat{\tau}_d$ is rarely substantially larger than that of the least biased estimate. In general it appears that for moderately linear response surfaces and moderately well matched samples, $\hat{\tau}_d$ will be the least biased estimate that we have considered.

4. SUMMARY

We now summarize the Monte Carlo results of section 2 and the analytic results of section 3 in the form of advice to an investigator who wants to estimate the constant difference between parallel univariate response surfaces in two populations P_1 and P_2 . This constant difference is called the treatment effect and designated τ.

In review, we assume G_1 is a random sample from P_1 of size N and G_2 is a random sample from P_2 of size rN, $r \geq 1$. τ will be estimated using G_1 and a N-size subsample of G_2 matched to G_1 on the matching variable X, G_{2*} . If $r = 1$ G_{2*} is a random sample from P_2 .

Five estimates of τ are considered (see Table 1 for explicit definitions).

(1) $\hat{\tau}_0$—the average difference across matched pairs.

(2) $\hat{\tau}_p$—the covariance adjusted estimate using the pooled estimate of the regression coefficient from a 1-way analysis of variance. This estimate is the natural regression adjusted estimate when dealing with random samples.

(3) $\hat{\tau}_d$—the regression adjusted estimate using matched pair differences, or equivalently a two-way analysis of variance. This regression adjusted estimate is the natural one when dealing with matched samples.

(4) $\hat{\tau}_1$—the regression adjusted estimate using only the G_1 sample to estimate the regression coefficient.

(5) $\hat{\tau}_2$—the regression adjusted estimate using only the matched G_{2*} sample to estimate the regression coefficient.

$\hat{\tau}_1$ and $\hat{\tau}_2$ are most natural when average differences between non-parallel response surfaces are desired.

4.1 *No regression adjustments*—$\hat{\tau}_0$

If no regression adjustments are to be performed, random order, nearest available pair-matching with a ratio of sample sizes $r \geq 2$ generally reduces the bias of the estimate, $\hat{\tau}_0 = \bar{y}_{1.} - \bar{y}_{2.}$, of the response surface difference, especially if the variance of the matching variable is greater in P_2 than in P_1 . The values of percent reduction in bias for linear response surfaces given in Table 2 are optimistic by less than 10% in most nonlinear cases considered here. However, in some "unfavorable" cases, the values of percent reduction in bias for moderately nonlinear response surfaces are much less than suggested by the values for linear response surfaces. Detailed advice on $\hat{\tau}_0$ for linear response surfaces is given in Rubin [1973].

4.2 *Regression adjusted estimates*—$\hat{\tau}_1$, $\hat{\tau}_2$, $\hat{\tau}_p$, $\hat{\tau}_d$

A. $\hat{\tau}_d$ vs $\hat{\tau}_p$—Variances of X Approximately Equal and Distribution of X Symmetric

When the variances of X are approximately equal in P_1 and P_2 and the distribution of X is symmetric in both P_1 and P_2 the Monte Carlo results of section 2 and the analytic results of section 3 suggest the following conclusions:

(1) The estimate $\hat{\tau}_p$ based on random samples is approximately unbiased when the response surfaces are approximately linear or quadratic; hence, for these simple and often assumed distributions of X there may be little gain in obtaining matched samples.
(2) If matched samples have been obtained with $4 \geq r \geq 2$, $\hat{\tau}_d$ will be the least biased estimate but generally not less biased than $\hat{\tau}_p$ based on random samples, and hence probably less preferred because of the fewer degrees of freedom used to estimate the regression coefficient.

B. $\hat{\tau}_d$ vs $\hat{\tau}_p$—Variances of X Unequal and/or Distributions of X Non-Symmetric

When the variances of X are different in P_1 and P_2 and/or the distributions of X are not symmetric in P_1 and P_2 , results in sections 2 and 3 suggest the following conclusions:

(1) Matching with $r \geq 2$ and using the estimate $\hat{\tau}_d$ based on matched pair differences should in most cases be the least biased procedure we have considered, removing most (90–110%) of the original bias of $\hat{\tau}_0$ based on random samples.
(2) Even with random samples ($r = 1$), assigning matches and using $\hat{\tau}_d$ in these cases may often be superior to the usual pooled estimate $\hat{\tau}_p$.
(3) If in the final matched samples the variances of X are approximately equal and the distributions of X appear symmetric, $\hat{\tau}_p$ may be slightly less biased than $\hat{\tau}_d$.
(4) In general, the decrease in bias of a regression adjusted estimate from matching with $r = 4$ rather than $r = 2$ is minor.

(5) If the response surface is linear, all regression adjusted estimates are unbiased, but $\hat{\tau}_p$ will be superior to $\hat{\tau}_d$ because it uses all of the data to estimate the regression coefficient and thus has smaller variance.

C. $\hat{\tau}_1$ and $\hat{\tau}_2$

In general, for the cases considered in sections 2 and 3, $\hat{\tau}_1$ and $\hat{\tau}_2$ are inferior to either $\hat{\tau}_d$ or $\hat{\tau}_p$ and their use should be avoided when the response surfaces are parallel.

4.3 *Other estimates of* τ

It could be argued in those cases in which $\hat{\tau}_d$ was the least biased estimate that the extra $N - 1$ degrees of freedom (D.F.) used to estimate parameters for the N matched pairs when forming matched pair differences could be better used on the pooled data to estimate the response surface more accurately. Thus a reasonable suggestion would be to obtain matched or random samples and whenever the response surfaces are thought to be even slightly nonlinear use the pooled data to estimate a quadratic (or higher order) term in X assuming parallel response surfaces. A possible criticism of this method is that in a multivariate case one may not have a large enough sample to estimate all quadratic terms. The generalization of this work to the multivariate case is currently being studied.

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 two referees for their helpful comments.

L'UTILISATION D'UN ECHANTILLONNAGE AVEC APPARIEMENT ET D'UN AJUSTEMENT PAR REGRESSION POUR SUPPRIMER LES BIAIS DANS DES ENQUÊTES D'OBSERVATION

RESUME

On étudie dans cet article la capacité d'un échantillonnage avec stratification et de l'ajustement par régression linéaire à réduire les biais de l'estimation de l'effet d'un traitement dans deux enquêtes d'observation et ceci par une méthode d'appariement simple et par cinq estimations simples. On donne les résultats, par la méthode de Monte Carlo, pour des surfaces de résponses modérément linéaires exponentielles et des résultats analytiques pour des surfaces de réponses modérément linéaires exponentielles et aussi quadratiques. Les conclusions sont:

(1) on peut s'attendre à ce que en général à la fois l'échantillonnage avec appariement et l'ajustement par régression réduisent les biais

(2) dans quelques cas, lorsque la variance de la variable d'appariement diffère dans les deux populations, à la fois l'appariement et l'ajustement par la régression peuvent accroître le biais

(3) quand la variance de la variable d'appariement est la même dans les deux populations et que les distributions de la variable d'appariement sont symétriques l'estimation habituelle ajustée par covariance, fondée sur des échantillons aléatoires, est presque non biaisée

(4) la combinaison d'ajustement par régression dans des échantillons appariés donne en général l'estimateur le moins biaisé.

REFERENCES

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

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

Cochran, W. G. [1970]. Performance of a preliminary test of comparability in observational studies. *ONR Technical Report No. 29*, Harvard University.

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

Rubin, D. B. [1973]. Matching to remove bias in observational studies. *Biometrics 29*, 159–183.

Received August 1971, *Revised July* 1972

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