


Assignment to Treatment Group on the Basis of a Covariate

Author(s): Donald B. Rubin

Source: *Journal of Educational Statistics*, Spring, 1977, Vol. 2, No. 1 (Spring, 1977), pp. 1-26

Published by: American Educational Research Association and American Statistical Association

Stable URL: https://www.jstor.org/stable/1164933

Journal of Educational Statistics
Spring 1977, Volume 2, Number 1, Pp. 1-26

ASSIGNMENT TO TREATMENT GROUP ON THE BASIS OF A COVARIATE

Donald B. Rubin

Educational Testing Service

Key words: Non-Randomized Studies; Observational Studies; Covariance Adjustment; Causal Inference; Experimental Design; Treatment Assignment; Average Treatment Effects

ABSTRACT

When assignment to treatment group is made solely on the basis of the value of a covariate, X, effort should be concentrated on estimating the conditional expectations of the dependent variable Y given X in the treatment and control groups. One then averages the difference between these conditional expectations over the distribution of X in the relevant population. There is no need for concern about "other" sources of bias, e.g., unreliability of X, unmeasured background variables. If the conditional expectations are parallel and linear, the proper regression adjustment is the simple covariance adjustment. However, since the quality of the resulting estimates may be sensitive to the adequacy of the underlying model, it is wise to search for nonparallelism and nonlinearity in these conditional expectations. Blocking on the values of X is also appropriate, although the quality of the resulting estimates may be sensitive to the coarseness of the blocking employed. In order for these techniques to be useful in practice, there must be either substantial overlap in the distribution of X in the treatment groups or strong prior information.

1. INTRODUCTION

In some studies, the experimental units are divided

into two treatment groups solely on the basis of a covariate, X. By this we mean that if two units have the same value of X either they both must receive the same treatment or they must be randomly assigned (not necessarily with probability 0.5) to treatments. For example, those units (children) with high scores on X (a reading test) receive the experimental treatment (a compensatory reading program); those with low scores on X receive the standard control treatment; perhaps those with intermediate scores on X are randomly assigned with equal probability to the treatments. The critical point is that the probability that an experimental unit is exposed to Treatment 1 rather than Treatment 2 is a function only of the values of X in the sample. After exposure to treatments, a dependent variable Y (a second reading test) is recorded in both treatment groups.

The central question is: what is the average effect on Y of Treatment 1 vs. Treatment 2 for the relevant population? For simplicity of discussion, we will usually assume the relevant population is the one from which all the units being studied are considered a random sample, say P. The associated effect is called τ. Some researchers might wonder whether to use gain scores, simple posttest scores, covariance adjusted scores (possibly adjusted for reliability), or some other device to estimate τ.

We will show that the appropriate estimate of τ is the average value of the difference between the estimated conditional expectations of Y on X in the two treatment groups, the average being taken over all units in the study if the relevant population is P. The conditional expectations (regressions) can be estimated using least squares, robust techniques, blocking, or matching methods. Neither gain scores nor scores adjusted for the reliability of X are generally appropriate (no matter how unreliable X may be).

In the special case of parallel linear regressions of Y on X in the two groups and least squares estimation, the average difference between the estimated regressions in the two treatment groups corresponds to the simple covariance adjusted estimator. There are previous references to the appropriateness of the covariance adjusted estimator in versions of this special case; see, for example, Cox (1951, 1957), Finney (1957), Goldberger (1972a, 1972b), Greenberg (1953), Kenney (1975), Snedecor and Cochran (1967, pp. 438-439).

However, the results presented here are general and emphasize (1) recording the variable X used to make assignment decisions, (2) estimating the conditional expectations of Y given X in each treatment group, and (3) averaging the difference between the estimated conditional expectations over the estimated distribution of X in the relevant population. These three steps are essential in order to estimate causal effects of treatments in those studies which are not classical randomized designs. Bayesian analogues for these results are presented in Rubin (1977).

In the following development we use unbiasedness as the criterion indicating the appropriateness of estimators. We do so only to show that the estimator tends to estimate the correct quantity without further adjustment. We do not mean to suggest that all biased estimators are unacceptable (a biased estimator with small mean squared error may of course be preferable to an unbiased estimator whose variance is large).

2. DEFINING THE EFFECTS OF TREATMENTS: ASSUMPTIONS AND NOTATION

The definition of the effect of Treatment 1 vs. Treatment 2 that we will use is standard in the sense that if the population P is essentially infinite, then the average treatment difference in very large randomized experiments on random samples from P will estimate the effect with negligible variance. However, we will explicitly present the assumptions in order to avoid ambiguity. The definition follows that given in Rubin (1974), the basic idea being that for each unit in P there is a value of Y that we could observe if the unit had been exposed to Treatment 1 and another value that we could observe if the unit had been exposed to Treatment 2; an important assumption is that these values of Y do not change as the other units in P receive different treatments. It is also assumed that the values of X are the same no matter which treatments the units received (i.e., X is a proper covariate).

More precisely, first suppose that all units in P were exposed to Treatment 1; let μ_1 be the resulting average value of Y for all units in P, and let $\mu_1(x)$ be the resulting average value of Y for all those units in P with score x on variable X. Second, suppose that all units in P were exposed to Treatment 2; let μ_2 be the resulting average value of Y for all units in P, and let

$\mu_2(x)$ be the resulting average value of Y for all those units in P with score x on variable X. Letting $\underset{x\varepsilon P}{\text{ave}}\,[\,.\,]$ denote the average value of the quantity in brackets over the distribution of X in P, we have that

$$\mu_i = \underset{x\varepsilon P}{\text{ave}}\,[\,\mu_i(x)\,] \quad .$$

Assume that X is unaffected by the treatments so that a unit's score on X will be the same no matter how treatments are assigned; this will be the case when X is recorded before treatments are assigned. Also assume "no 'interference' between different units" (Cox, 1958, p. 19) so that a unit's Y value given Treatment i is unaffected by which treatments the other units in P received. Without this assumption, even if P were infinite, different infinitely large randomized experiments would estimate different effects, in the sense that the variance of the average treatment difference over all such randomized experiments generally would not be negligible. There are weaker assumptions under which one can estimate causal effects, but we do not consider them here. Note that the usual null hypothesis of no treatment effect assumes that Y given Treatment 1 equals Y given Treatment 2 for all units and assignments of treatments, a very special form of no interference.

Good experimental design often reflects the assumption of no interference between different units. For example, consider a time-consuming compensatory reading treatment. First, suppose each student is a unit with the compensatory and regular reading treatments assigned to different students in the same classroom. In this case the no interference assumption may be suspect because of social interactions among the students and competition for the teacher's time (the effect of the compensatory reading treatment on a student being different when only a few students receive the compensatory reading treatment than when all the students in the class receive the compensatory reading treatment). Now suppose instead classrooms in different schools were the units, and each classroom was assigned either to the regular or compensatory treatment condition (perhaps all students in a classroom receiving the compensatory reading instruction, or a randomly chosen group of ten, or only those in need--these reflect three different compensatory reading treatments being applied to the classroom). With the choice of

classrooms as units, the no interference assumption seems quite plausible.

We are now ready to define the average causal effect of Treatment 1 vs. Treatment 2. Consider a unit randomly drawn from P and then exposed to Treatment i (i.e., each unit in P was equally likely to be chosen). Because of the assumption of no interference between units, the average value of Y for such a unit (i.e., averaging over all random draws of one unit from P) is μ_i, no matter what treatments the other units in P received. Hence $\tau = \mu_1 - \mu_2$ is called the average or expected effect of Treatment 1 vs. Treatment 2 on Y in the population P.

Now consider a unit randomly chosen from those units in P with $X = x$ and then exposed to Treatment i (i.e., each unit in P with $X = x$ was equally likely to be chosen). Because of the assumption of no interference between units, the average value of Y for such a unit exposed to Treatment i (i.e., averaging over all random draws of one unit from P with $X = x$) is $\mu_i(x)$ no matter what treatments the other units in P received. Hence $\mu_1(x) - \mu_2(x)$ is called the effect of Treatment 1 vs. Treatment 2 on Y at $X = x$ in P. (See Figure 1.) If $\mu_1(x) - \mu_2(x)$ is constant for all x, $\mu_1(x)$ and $\mu_2(x)$ are called parallel, and the effect of Treatment 1 vs. Treatment 2 is the same for each value of X. Generally, however, the relative effect of the treatments will depend on the value of X.

The $\mu_i(x)$ are called the conditional expectations of Y given X and treatment condition, or the "response functions of Y given X" or the "regressions of Y on X." Often the $\mu_i(x)$ are assumed to be linear in x, but this restriction is not needed for the general results presented here.

It follows from the above definitions that the average effect of Treatment 1 vs. Treatment 2 on Y in P, $\tau = \mu_1 - \mu_2$, is $\mu_1(x) - \mu_2(x)$ averaged over the distribution of X in P:

$$\tau = \underset{x \in P}{\text{ave}} \; [\mu_1(x) - \mu_2(x)] \quad . \tag{1}$$

This simple relationship is exploited to estimate τ in non-randomized studies. In Figure 1, τ is calculated by taking the vertical difference between the $\mu_1(x)$ and $\mu_2(x)$ curves at each x, and finding the average value of this difference weighted by the distribution of X in P.

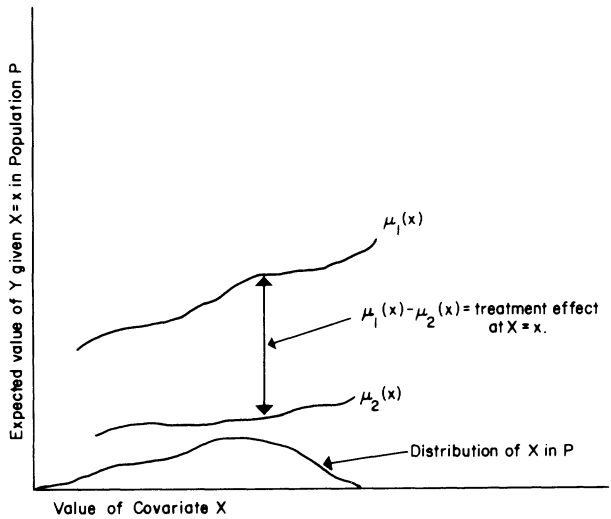

FIG. 1

The Treatment Effect in Population P:

$$\tau = \underset{x \epsilon P}{\text{Ave}} \, [\mu_1(x) - \mu_2(x)]$$

3. PRELIMINARY RESULTS

Throughout the rest of the paper we will assume the following sampling situation. A random sample of size $n_1 + n_2$ from P is divided into two groups of sizes n_1 and n_2 solely on the basis of the values of X and perhaps some randomization. That is, the assignment decisions are such that either all sampled units with the same value of X are in the same treatment group or are randomly divided (not necessarily with equal probability) into the two treatment groups. The first group is exposed to Treatment 1 and the second group is exposed to Treatment 2. Let x_{ij}, y_{ij} $i = 1,2$; $j = 1,\ldots,n_i$ be the values of X and Y in the two samples.

Since the x_{ij} are a random sample from P, Result 1 is immediate from equation (1).

Result 1: The quantity

$$\frac{1}{n_1 + n_2} \sum_{i=1}^{2} \sum_{j=1}^{n_i} \left[\mu_1(x_{ij}) - \mu_2(x_{ij})\right] \qquad (2)$$

is unbiased for τ.

Notice that the notation $\mu_1(x_{ij})$ and $\mu_2(x_{ij})$ in expression (2) means that the functions $\mu_1(x)$ and $\mu_2(x)$ are to be evaluated at the observed values x_{ij}, and that by the phrase "unbiased for τ" we mean that the average value of expression (2) over all random samples of size $n_1 + n_2$ from P is τ.

If we had conditionally unbiased estimates of the values $\mu_1(x_{ij})$ and $\mu_2(x_{ij})$, $i = 1,2$; $j = 1,\ldots,n_i$, we could substitute them into expression (2) to obtain an unbiased estimate of τ. By conditionally unbiased we mean unbiased given the values x_{ij} that occur in the sample (i.e., averaging over all random draws from P that yield the same values for the x_{ij} as observed in our sample). Only the values of x_{ij} in the sample and the conditional expectations of Y given X under the two treatments are needed in order to obtain an unbiased estimate of τ. No matter how "unreliable" X is, no reliability correction is relevant; nor does it matter what functional forms $\mu_1(x)$ and $\mu_2(x)$ take.

Result 2 is the key to obtaining unbiased estimators of τ since it gives us conditionally unbiased estimates of some $\mu_1(x)$ and $\mu_2(x)$ values.

Result 2: The value y_{1j} is a conditionally unbiased estimate of $\mu_1(x_{1j})$ $j = 1,\ldots,n_1$, and the value y_{2j} is a conditionally unbiased estimate of $\mu_2(x_{2j})$ $j = 1,\ldots,n_2$.

In order to prove Result 2, first note that sampled

units with $X = x$ were randomly sampled from those units in P with $X = x$. Next note that since assignment was on the basis of X, sampled units with $X = x$ are either (a) always assigned the same treatment or (b) randomly assigned treatments (not necessarily with probability 0.5 of receiving each treatment). In either case, sampled units with $X = x$ who were assigned Treatment i were randomly chosen from those units in P with $X = x$. Now by the definition of $\mu_i(x)$ and the assumption of no interference between units, the average value of Y for a unit randomly drawn from those units in P with $X = x$ and then assigned Treatment i is $\mu_i(x)$ no matter what the other sampled values of X or the other treatment assignments. Therefore, y_{ij} (the observed values of Y for a sampled unit with $X = x_{ij}$ given exposure to Treatment i) is a conditionally unbiased estimate of $\mu_i(x_{ij})$.

Note the crucial role in this proof of assignment on the basis of X. If assignment depended on some variable Z other than X, sampled units with $X = x$ who were assigned to Treatment i were not randomly sampled from those units in P with $X = x$, but rather from those units in P with (a) $X = x$ and (b) Z satisfying the conditions that determined the assignment to Treatment i.

By Result 2, we have conditionally unbiased estimates of points on the Treatment 1-part of $\mu_1(x)$ (i.e., $\mu_1(x_{1j})$ $j = 1,\ldots,n_1$) and points on the Treatment 2-part of $\mu_2(x)$ (i.e., $\mu_2(x_{2j})$ $j = 1,\ldots,n_2$). However, we still lack conditionally unbiased estimates of points on the Treatment 2-part of $\mu_1(x)$ (i.e., $\mu_1(x_{2j})$ $j = 1,\ldots,n_2$) and points on the Treatment 1-part of $\mu_2(x)$ (i.e., $\mu_2(x_{1j})$ $j = 1,\ldots,n_1$). And we need these estimates in order to use Result 1 to obtain an unbiased estimate of τ.

We will discuss two general methods for obtaining conditionally unbiased estimates of these quantities: (a) fitting a model to the data to obtain estimates of the functions $\mu_1(x)$ and $\mu_2(x)$, and (b) grouping Treatment 1 and Treatment 2 units with similar values of X to obtain estimates of the difference $\mu_1(x) - \mu_2(x)$ at particular X values that are representative of the distribution of X in P.

4. ESTIMATING $\mu_1(x_{ij})$ AND $\mu_2(x_{ij})$ BY MODEL FITTING

One method for estimating the values of $\mu_1(x_{2j})$ and $\mu_2(x_{1j})$ is via a model for the functions $\mu_1(x)$ and $\mu_2(x)$. This is most appropriate when X takes on many values (e.g., age, height). Obviously the accuracy of the resulting estimates will be somewhat dependent on the accuracy of the model chosen.

The modelling of $\mu_1(x)$ and $\mu_2(x)$ will be illustrated in the simple case when we assume both are linear in x. The usual least squares estimates are:

$$\hat{\mu}_1(x) = \bar{y}_1 + \hat{\beta}_1(x - \bar{x}_1) \text{ , and} \tag{3}$$

$$\hat{\mu}_2(x) = \bar{y}_2 + \hat{\beta}_2(x - \bar{x}_2) \text{ ,} \tag{4}$$

where

$$\hat{\beta}_i = \frac{\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)(x_{ij} - \bar{x}_i)}{\sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2} \text{ , } \quad i = 1,2 \text{ .} \tag{5}$$

Result 3: If both $\mu_1(x)$ and $\mu_2(x)$ are of the form $\mu_i(x) = \alpha + \beta_i x$, the estimator

$$\bar{y}_1 - \bar{y}_2 - (\bar{x}_1 - \bar{x}_2)\frac{n_1\hat{\beta}_2 + n_2\hat{\beta}_1}{n_1 + n_2} \tag{6}$$

is unbiased for τ.

In order to prove Result 3, first note that expression (6) equals

$$\frac{1}{n_1 + n_2} \sum_{i=1}^{2} \sum_{j=1}^{n_i} \left[\hat{\mu}_1(x_{ij}) - \hat{\mu}_2(x_{ij})\right] \text{ ,} \tag{7}$$

where $\hat{\mu}_1(x_{ij})$, $\hat{\mu}_2(x_{ij})$, and $\hat{\beta}_i$ are given by equations (3), (4), and (5). By Result 2, the conditional expectation of $\hat{\mu}_i(x)$ is $\mu_i(x)$: (a) the conditional expectation of $\bar{y}_i$ is $\alpha_i + \beta_i \bar{x}_i$, and (b) the conditional expectation of

$$\hat{\beta}_i \text{ is } \sum_{j=1}^{n_i} (\mu_i(x_{ij}) - \overline{\mu_i(x_{ij})})(x_{ij} - \bar{x}_i) / \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 =$$

β_i . Consequently, $\hat{\mu}_1(x_{ij}) - \hat{\mu}_2(x_{ij})$ is a conditionally unbiased estimate of $\mu_1(x_{ij}) - \mu_2(x_{ij})$ (for $j = 1,\ldots,n_i$; $i = 1,2$) . Thus by Result 1, expression (7) (and thus (6)) is unbiased for τ .

Result 4: If $\mu_1(x)$ and $\mu_2(x)$ are both linear in x and parallel, then the simple analysis of covariance estimator

$$\bar{y}_1 - \bar{y}_2 - (\bar{x}_1 - \bar{x}_2)\hat{\beta} \tag{8}$$

where $$\hat{\beta} = \frac{\sum_{i=1}^{2} \sum_{i=1}^{n_i} (y_{ij} - \bar{y}_i)(x_{ij} - \bar{x}_i)}{\sum_{i=1}^{2} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2}$$

is unbiased for τ .

The proof of Result 4 is essentially the same as the proof of Result 3 with the change that now $\beta_1 = \beta_2 = \beta$ and both $\hat{\beta}_1$ and $\hat{\beta}_2$ are replaced by $\hat{\beta}$ which is a conditionally unbiased estimate of β .

Results analogous to Results 3 and 4 follow when $\mu_1(x)$ and $\mu_2(x)$ are polynomial in x or any linear combination of specified functions of x (e.g., e^x). The only change is in the method of estimating the parameters, i.e., by a multiple least squares regression. Methods more robust than least squares (c.f., Beaton & Tukey, 1974) might be appropriate for estimating the conditional expectations $\mu_1(x)$ and

$\mu_2(x)$, especially when there is the possibility of outliers or long-tailed distributions.

Of course we never know whether the functions $\mu_1(x)$ and $\mu_2(x)$ are linear in x (or linear in some specified functions of x). But the unbiasedness of the estimators given by (6) and (8) is dependent upon the accuracy of the linear model. There is evidence (Rubin, 1973b) that in some cases the linear approximation is adequate to remove most of the bias present in the simple estimator $\bar{y}_1 - \bar{y}_2$ but that in other cases it is inadequate even when $\mu_1(x)$ and $\mu_2(x)$ are smooth monotone functions. The troublesome cases are basically those having quite different variances of X in the treatment groups.

If the observed values of X in the treatment groups are similar, it may be possible to check that both $\hat{\mu}_1(x)$ and $\hat{\mu}_2(x)$ are reasonable approximations to $\mu_1(x)$ and $\mu_2(x)$ for the full range of observed values of X . This checking is important because we must average $\hat{\mu}_1(x) - \hat{\mu}_2(x)$ over the full range of observed X values, and therefore must have confidence in both models for most of the values of X that occur in the sample.

If the X values in the two samples do not overlap (e.g., as in the regression discontinuity design, Campbell and Stanley, 1963, pp. 61-64) it is impossible to check the accuracy of either $\hat{\mu}_1(x)$ or $\hat{\mu}_2(x)$ for the full range of observed X values, and we must rely on our a priori assumptions. Consequently, in order for the model-fitting efforts described above to be useful in practice, we must either have samples that overlap or strong a priori information about the functional forms of the $\mu_i(x)$.

5. ESTIMATING $\mu_1(x_{ij})$ AND $\mu_2(x_{ij})$ BY BLOCKING ON X

When the assignment to treatment group allows the distribution of X in the two treatment groups to overlap substantially, it may be possible to obtain conditionally unbiased estimates of $\mu_1(x_{2j})$ and $\mu_2(x_{1j})$ without fitting a model. The obvious but crucial point is that if $x_{1j} = x_{2j}$, then

y_{1j} is conditionally unbiased for $\mu_1(x_{2j}) = \mu_1(x_{1j})$ and
y_{2j} is conditionally unbiased for $\mu_2(x_{1j}) = \mu_2(x_{2j})$.

Suppose that in the samples there are only K distinct values of X , say $x_1,\ldots,x_K$, where n_{1k} Treatment 1 units and n_{2k} Treatment 2 units have X values equal to x_k , $k = 1,\ldots,K$. Let $\bar{y}_{1k}$ be the average Y value for the n_{1k} Treatment 1 units whose X value equals x_k ; similarly let $\bar{y}_{2k}$ be the average Y value for the n_{2k} Treatment 2 units whose X value equals x_k . If $n_{ik} = 0$ for some i and k , then the corresponding $\bar{y}_{ik}$ is not defined.

Result 5: If $n_{1k} > 0$ and $n_{2k} > 0$ for all $k = 1,\ldots,K$, then the estimator

$$\frac{1}{n_1 + n_2}\left[\sum_{k=1}^{K}(n_{1k} + n_{2k})(\bar{y}_{1k} - \bar{y}_{2k})\right]$$

$$= \frac{1}{n_1 + n_2}\left[n_1\bar{y}_1 - n_2\bar{y}_2 + \sum_{k=1}^{K} n_{2k}\bar{y}_{1k} - \sum_{k=1}^{K} n_{1k}\bar{y}_{2k}\right] \quad (9)$$

is unbiased for τ .

Result 5 follows because by Result 2 $\bar{y}_{1k}$ is an unbiased estimate of $\mu_1(x_k)$, $\bar{y}_{2k}$ is an unbiased estimate of $\mu_2(x_k)$, and so $\bar{y}_{1k} - \bar{y}_{2k}$ is an unbiased estimate of $\mu_1(x_k) - \mu_2(x_k)$, $k = 1,\ldots,K$. That is, the difference between the Y mean for those Treatment 1 units whose X value is x_k and the Y mean for those Treatment 2 units whose X value is x_k is an unbiased estimate of the Treatment 1 vs. Treatment 2 effect at x_k . Hence, from Result 1 we have Result 5.

The advantage of the estimator given by (9) is that it does not depend on the accuracy of some underlying model for

its unbiasedness. The disadvantage of the estimator is that if X takes on many values, some n_{ik} may be zero and then the estimator is not defined; this occurrence is not unusual in practice.

A common practical method used when some $n_{ik} = 0$ is to aggregate values of the original X variable to define a new variable X^* for which all $n_{ik} > 0$. However, since the assignment process was on the basis of X (not X^*), the estimator given by (9) based on X^* is no longer necessarily unbiased for τ. If X^* takes on many values, the bias might be small. For a similar situation, Cochran (1968) concluded that in many cases blocking on an aggregated version of X with as few as 5 or 6 values was adequate to remove over 90% of the bias present in the simple estimator $\bar{y}_1 - \bar{y}_2$.

Of particular interest is the case in which X^* is chosen with minimum aggregation (i.e., K is maximized subject to the constraint that each $n_{ik} > 0$). It would be of practical importance to investigate the bias of the estimate (9) based on this X^* under (a) various underlying distributions of X in P, (b) different assignment processes based on X, and (c) several response functions $\mu_1(x)$ and $\mu_2(x)$.

Another method for handling cases in which some $n_{ik} = 0$ is to discard units. Result 6 is immediate from Result 1.

Result 6: If $\mu_1(x)$ and $\mu_2(x)$ are parallel, then

$$\sum_{k=1}^{K} \delta_k(\bar{y}_{1k} - \bar{y}_{2k}) \Big/ \sum_{k=1}^{K} \delta_k \tag{10}$$

where $$\delta_k = \begin{cases} 0 \text{ if } n_{1k} \times n_{2k} = 0 \\ \left(n_{1k}^{-1} + n_{2k}^{-1}\right)^{-1} \text{ otherwise} \end{cases}$$

is unbiased for τ.

Since $\mu_1(x)$ and $\mu_2(x)$ are parallel, there are many other choices of δ_k in (10) that will yield unbiased estimates of τ but they generally yield different precisions. The choice of optimal δ_k depends on conditions we have not discussed, those in (10) being optimal when the conditional variance of Y given X is constant. For further discussion, see for example, Kalton (1968) or Maxwell and Jones (1976).

Notice that the estimator given by (10) essentially discards those units whose X values are not the same as the X value of some unit who was exposed to the other treatment. This procedure, known as matching on the values of X, makes a lot of sense in some cases. Suppose X has been recorded and there is an additional cost in recording Y even though the treatments have already been given to all of the units. For example, the regular and compensatory reading programs have been given, background variables have been recorded, but there is an additional expense in giving and recording a battery of detailed posttests to each student. In these situations it is appropriate to ask how to choose the units on which to record Y. However, it may not be appropriate to assume the regressions are parallel, so that the estimator given by (10) may not be useful for estimating τ. Matching is more applicable when a subpopulation of P is of primary interest, that is, when the parameter of primary interest is the average treatment effect in a subpopulation.

6. GENERALIZING TO A SUBPOPULATION OF P DEFINED BY X

At times, the relevant population will not be P, but rather a subpopulation of P, say P_x defined by values of the covariate X (perhaps supplemented by some randomization). For example, the units exposed to Treatment 1 may be considered to be a random sample from the relevant population, perhaps those in need of extra treatment because of low values of X.

In such cases, all the results presented here generalize to estimating $\tau_x = \underset{x \epsilon P_x}{\text{ave}} \left[\mu_1(x) - \mu_2(x)\right]$. The quantity τ_x is the treatment effect in the population P_x because

$\mu_1(x) - \mu_2(x)$ is the treatment effect in P_x at $X = x$ as well as in P at $X = x$. That is, the conditional expectation of Y given (a) Treatment i, (b) $X = x$, and (c) X satisfies some criterion that defines membership in P_x is simply the conditional expectation of Y given (a) Treatment i and (b) $X = x$, which is defined to be $\mu_i(x)$.

Hence, Result 1 generalizes to estimating τ_x if the average over all X values in the sample is replaced by the average over X values that are representative of P_x. Result 2 is true as stated for P_x. In Result 3, the corresponding estimator of τ_x is now given by expression (7) with the averaging over all units replaced by averaging over units representative of P_x. For example, if the units exposed to Treatment 1 are considered a random sample from P_x, this averaging of expression (7) leads to

$$(\bar{y}_1 - \bar{y}_2) - (\bar{x}_1 - \bar{x}_2)\,\hat{\beta}_2 \tag{11}$$

as the unbiased estimator of τ_x. This estimator given by expression (11) is discussed in some detail by Belsen (1956) and Cochran (1970).

If $\mu_1(x)$ and $\mu_2(x)$ are parallel, $\tau = \tau_x$ so that Result 4 as well as Result 6 apply for obtaining unbiased estimates of the treatment effect for any subpopulation P_x.

The extension of Result 5 to the subpopulation P_x is somewhat more interesting, although equally straightforward. For example, again suppose the units exposed to Treatment 1 are a random sample from P_x; then if $n_{2k} > 0$ whenever $n_{1k} > 0$, the estimator

$$\frac{1}{n_1} \sum_{k=1}^{K} n_{1k} (\bar{y}_{1k} - \bar{y}_{2k}) \tag{12}$$

is unbiased for τ_x. This estimator discards those units exposed to Treatment 2 whose X values are not found among the units exposed to Treatment 1. Finding for each unit

exposed to Treatment 1 a unit exposed to Treatment 2 with the same X value and forming the estimate (12) has been called matched sampling (Rubin, 1973a). As discussed at the end of Section 5, estimators that discard data are most appropriate when one must decide for which units the value of Y should be recorded.

7. A SIMPLE EXAMPLE

Table I presents the raw data from an evaluation of a computer-aided program designed to teach mathematics to children in fourth grade. There were 25 children in Program 1 (the computer-aided program) and 47 children in Program 2 (the regular program). All children took a Pretest and Posttest, each test consisting of 20 problems, a child's score being the number of problems correctly solved. These data will be used to illustrate the estimation methods discussed in Sections 4, 5, and 6. We do not attempt a complete statistical analysis nor do we question the assumption of no interference between units.

TABLE I

Raw Data for 25 Program 1 Children and 47 Program 2 Children

Pretest Scores	Posttest Scores	
	Program 1	Program 2
10	15	6,7
9	16	7,11,12
8	12	5,6,9,12
7	8,11,12	6,6,6,6,7,8
6	9,10,11,13,20	5,5,6,6,6,6,6,6,6,8,8,8,9,10
5	5,6,7,16	3,5,5,6,6,7,8
4	5,6,6,12	4,4,4,5,7,11
3	4,7,8,9,12	0,5,7
2	4	4
1	-	-
0	-	7

7.1 Assignment on the Basis of Pretest

Suppose first that assignment to Program 1 or Program 2

was on the basis of Pretest, so that children with the same Pretest score were randomly assigned to programs with the probability of assignment to Program 1 increasing with lower Pretest scores. Hence, Pretest is the covariate X in the discussion of the previous sections. Posttest is the dependent variable Y. Figure 2 plots Posttest on Pretest for the treatment groups. Notice that although the Program 1 children scored somewhat lower on the Pretest than the Program 2 children (the Pretest means being 5.24 and 5.85 respectively) the Program 1 children scored higher on the Posttest than the Program 2 children (the Posttest means being 9.76 and 6.53 respectively). Consequently, we expect estimates of τ to be positive. Furthermore, notice that the distributions of X in the groups overlap substantially; hence, the methods described in the previous sections are appropriate for estimating treatment effects.

First consider estimating τ without assuming $\mu_1(x)$ and $\mu_2(x)$ are parallel. Using the least squares model-fitting methods described in Section 4, we fit separate linear conditional expectations in the two groups and obtain $\hat{\beta}_1 = 1.22$ and $\hat{\beta}_2 = 0.46$; from equation (6), the estimate of τ is 3.81. Fitting a quadratic conditional expectation in each group by least squares

$$\hat{\mu}_i(x) = \bar{y}_i + (x - \bar{x}_i)\,\hat{\gamma}_{1i} + (x^2 - \overline{x_i^2})\,\hat{\gamma}_{2i}\,, \quad i = 1,2, \tag{13}$$

we have that $\overline{x_1^2} = 31.40$, $\overline{x_2^2} = 38.28$, $\hat{\gamma}_{11} = 1.40$, $\hat{\gamma}_{12} = 0.06$, $\hat{\gamma}_{21} = -0.02$, and $\hat{\gamma}_{22} = 0.04$; since the average difference between $\hat{\mu}_1(x)$ and $\hat{\mu}_2(x)$ over the values of X that occur in the sample is

$$(\bar{y}_1 - \bar{y}_2) - (\bar{x}_1 - \bar{x}_2)\,\frac{n_1\hat{\gamma}_{12} + n_2\hat{\gamma}_{11}}{n_1 + n_2} - (\overline{x_1^2} - \overline{x_2^2})\,\frac{n_1\hat{\gamma}_{22} + n_2\hat{\gamma}_{21}}{n_1 + n_2}\,, \tag{14}$$

the resulting estimate of τ under the quadratic model is 3.81. The blocking methods of Section 5 may also be used to estimate τ. After pooling the one child who scored "0" on the Pretest with the two children that scored "2" on the Pretest, we use equation (9) and obtain 3.98 for the estimate of τ.

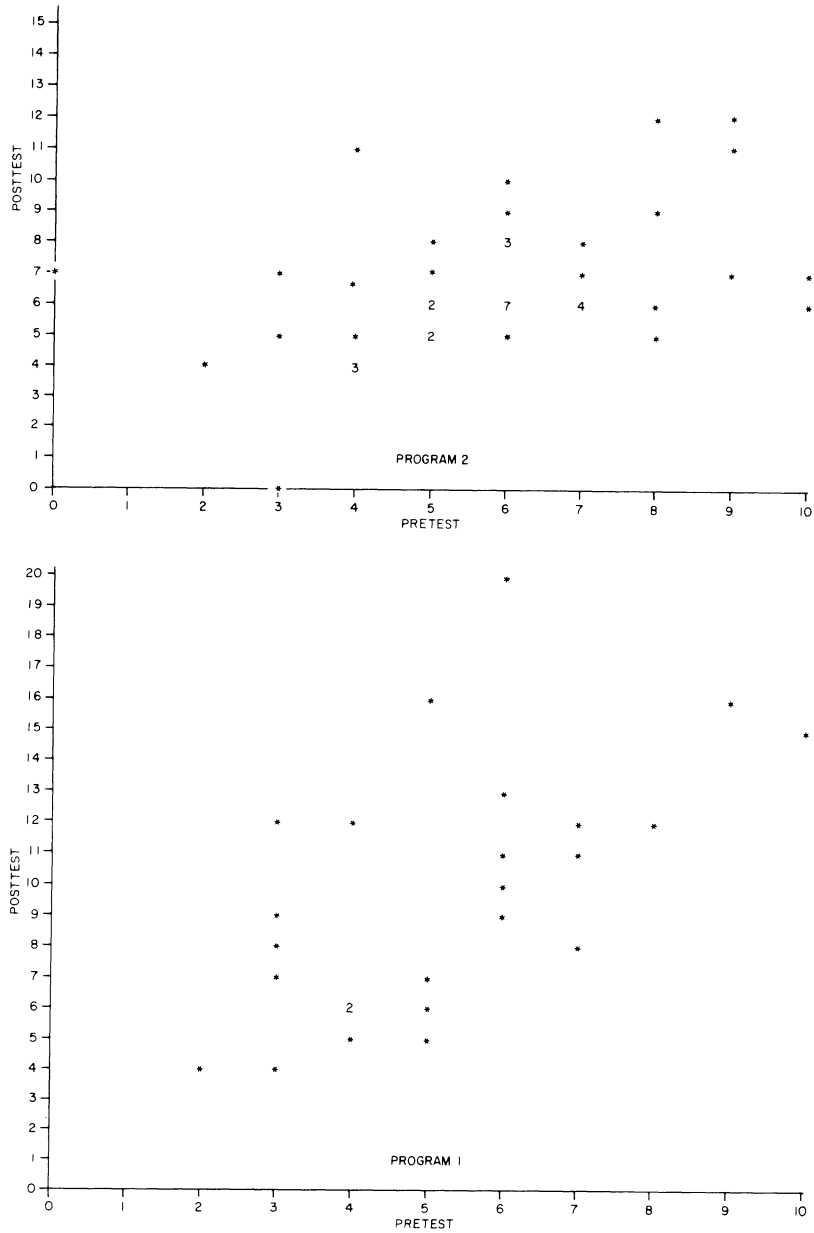

FIG. 2

Posttest vs. Pretest

Now, as discussed in Section 6, let us estimate the treatment effect for the population P_x, where the Program 1 children are considered a random sample from P_x. Assuming $\mu_2(x)$ is linear in x and using least squares estimation, equation (11) is appropriate and yields 3.51 for the estimate of τ_x. Using a quadratic model for $\mu_2(x)$ and least squares estimation, the appropriate estimate of τ_x is given by

$$(\bar{y}_1 - \bar{y}_2) - (\bar{x}_1 - \bar{x}_2)\,\hat{\gamma}_{12} - (\overline{x_1^2} - \overline{x_2^2})\,\hat{\gamma}_{22} \qquad (15)$$

which equals 3.54 for our data. And using the matching estimate given by (12) (discarding the data from the Program 2 child who scored "0" on the Pretest), we obtain 3.84 for the estimate of τ_x.

Finally, suppose that we assume $\mu_1(x)$ and $\mu_2(x)$ are parallel. Least squares estimation of the linear model gives $\hat{\beta} = 0.7180$, and thus from equation (8), 3.67 for the estimate of $\tau = \tau_x$. Using least squares to estimate a parallel quadratic fit,

$$\hat{\mu}_i(x) = \bar{y}_i + (x - \bar{x}_i)\,\hat{\gamma}_1 + (x^2 - \overline{x_i^2})\,\hat{\gamma}_2\,, \quad i = 1,2, \qquad (16)$$

we find $\hat{\gamma}_1 = 0.3855$ and $\hat{\gamma}_2 = 0.0293$; since τ is estimated under this quadratic model by

$$(\bar{y}_1 - \bar{y}_2) - (\bar{x}_1 - \bar{x}_2)\,\hat{\gamma}_1 - (\overline{x_1^2} - \overline{x_2^2})\,\hat{\gamma}_2\,, \qquad (17)$$

we obtain 3.66 for the estimate of $\tau = \tau_x$. The blocking estimate of $\tau = \tau_x$ found by substituting into equation (10) is 3.97.

The nine estimates presented for this example are summarized in Table II. The pattern of values for these estimates suggests that $\mu_1(x)$ and $\mu_2(x)$ may not be parallel, since the effect of Program 1 vs. Program 2 appears smaller for the lower values of X that occur more frequently in the Program 1 group. The implication is that the children who scored higher on the Pretest tended to profit more from

Program 1. However, the estimates displayed in Table II exhibit little variability, ranging between 3.5 and 4.0.

TABLE II

Estimates of Treatment Effect for Data in Table I
(Relevant Expression Numbers in Parentheses)

Parameter Being Estimated	Method of Estimation		
	Model-fitting		Blocking
	Linear	Quadratic	
τ	(6) 3.81	(14) 3.81	(9) 3.98[a]
τ_x with Program 1 units a random sample from P_x	(11) 3.51	(15) 3.54	(12) 3.84
$\tau = \tau_x$ assuming $\mu_1(x)$ and $\mu_2(x)$ parallel	(8) 3.67	(17) 3.67	(10) 3.97

[a]The unit with Pretest score of "0" is blocked with those units with Pretest score "2."

Of course in practice, one should be concerned not only with the variability of the estimated treatment effects across different models, but also with the variability of the estimated treatment effect given a particular model (i.e., the standard error of the estimate under the model). For details on calculating standard errors of estimators like these under the normal error model see, for example, Snedecor and Cochran (1967, pp. 268-270, 423).

7.2 Assignment Not on the Basis of Pretest

Suppose now that instead of assignment to treatment group on the basis of Pretest, all Program 1 children came from one school and all Program 2 children came from a different school. That is, suppose that School 1 children had been assigned to Program 1 and School 2 children had been assigned to Program 2. The same plot of Posttest on Pretest as given in Figure 2 would be observed, but the estimates given above would not be unbiased for the effect of Program 1 vs. Program 2 because the covariate that was used to assign treatments was not Pretest but School. Now School must be included as a covariate in order to apply the results of this paper. However, the plot of Y vs. School looks like Figure 3: each treatment group has only one value of the covariate X. We now cannot estimate the response functions $\mu_1(x)$ or $\mu_2(x)$ using the methods discussed in Sections 4, 5, 6 because there is no range of X values in each group. Nor can we block Program 1 and Program 2 children with similar values of X, because there are no such children.

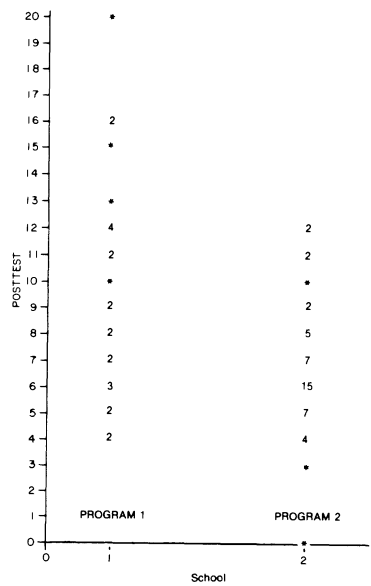

FIG. 3

Posttest vs. School

Thus, if assignment to treatment group is on the basis of School, the methods we have presented cannot be directly applied because there is no overlap in the distribution of the covariate School in the two treatment groups. Using the estimates presented in Section 7.1 with Pretest as the covariate makes the implicit assumption that in each treatment group, the expected value of Y given Pretest and School is the same as the expected value of Y given just Pretest. Whether this assumption is reasonable depends, for instance, on how children were assigned to schools.

This simple example brings out two critical points relevant to all nonrandomized studies. First, knowledge of the assignment process is critical to drawing inferences about the effect of the treatments; one cannot simply look at the plot of Posttest on Pretest and properly estimate treatment effects. Second, even when the assignment mechanism is fully understood, the most defensible analysis of the data requires the distribution of the covariate to overlap in the two groups; without overlap, the analysis relies on assumptions that cannot be checked using the data at hand.

An example similar to this one is discussed by Lord (1967), but is used to emphasize the benefits of randomization.

8. DISCUSSION OF NEEDED INVESTIGATIONS

In this paper we have stated the fact that if assignment to treatment group is on the basis of the value of a covariate, X , one must concentrate effort on the essential problem of estimating the conditional expectation of Y given X in each treatment group. One then averages the difference between these conditional expectations over the values of X that are representative of the population of interest.

Two general methods for estimating these expectations were discussed: model fitting and blocking on the values of X . Little relevant work has been done on how well these techniques are likely to do in practice, either alone or in combination. A relevant simulation would include several careful choices of:

(a) the sample size, $n_1 + n_2$
(b) the distribution of X in P
(c) the assignment mechanism
(d) the functional forms for the conditional expectations, $\mu_1(x)$ and $\mu_2(x)$.

One would then find the distribution of estimates resulting from using the model fitting and blocking methods discussed here.

The case of multivariate X is of real interest because in natural settings we may not know the assignment mechanism but may feel that it can be described reasonably well by a particular collection of variables that are recorded. For example, teachers deciding which students should receive compensatory reading treatments presumably use personal assessments of students in addition to background characteristics of the children and test scores (not "true scores"), but the assignment mechanism might be adequately approximated by some function of the recorded background variables and tests, personal assessments hopefully being largely determined by the recorded variables.

All the results presented here for univariate X generalize immediately (conceptually at least) to multivariate X (e.g., $\hat{\beta}$ is now a vector of regression coefficients). Some work on multivariate matching methods is given in Althauser and Rubin (1970), Cochran and Rubin (1973) and Rubin (1976a, 1976b), but has received little attention otherwise.

Certainly a serious effort on both the univariate case and the multivariate case is worthwhile, not only in order to improve the analysis of existing nonrandomized studies but also in order to study the possibility of finding designs that are tolerable given social constraints, not randomized in the usual sense, but still allow useful inferences for the effects of treatments.

ACKNOWLEDGMENTS

I wish to thank P. W. Holland, M. R. Novick and a referee for suggestions that substantially improved the presentation, and D. T. Thayer for performing the computations needed in Section 7.

REFERENCES

Althauser, R. P., & Rubin, D. B. The computerized construction of a matched sample. The American Journal of Sociology, 1970, 76, 325-346.

Beaton, A. E., & Tukey, J. W. The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. Technometrics, 1974, 16, 147-185.

Belsen, W. A. A technique for studying the effects of a television broadcast. Applied Statistics, 1956, 5, 195-202.

Campbell, D. T., & Stanley, J. C. Experimental and quasi-experimental designs for research. Chicago: Rand McNally, 1963.

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

Cochran, W. G. The use of covariance in observational studies. Applied Statistics, 1970, 18, 270-275.

Cochran, W. G., & Rubin, D. B. Controlling bias in observational studies: A review. Sankhyā-A, 1973, 35 (Pt. 4), 417-446.

Cox, D. R. Some systematic experimental designs. Biometrika, 1951, 38, 312-323.

Cox, D. R. The use of a concomitant variable in selecting an experimental design. Biometrika, 1957, 44, 150-158.

Cox, D. R. Planning of experiments. New York: Wiley, 1958.

Finney, D. J. Stratification, balance and covariance. Biometrics, 1957, 13, 373-386.

Goldberger, A. S. Selection bias in evaluating treatment effects: Some formal illustrations. Discussion paper. Univ. of Wisc. Institute for Research on Poverty, 1972. (a)

Goldberger, A. S. Selection bias in evaluating treatment effects: The case of interaction. Discussion paper. Univ. of Wisc. Institute for Research on Poverty, 1972. (b)

Greenberg, B. G. Use of covariance and balancing in analytic surveys. American Journal of Public Health, 1953, 43, 692-699.

Kalton, G. Standardization: A technique to control for extraneous variables. Applied Statistics, 1968, 16, 118-136.

Kenney, D. A. A quasi-experimental approach to assessing treatment effects in the nonequivalent control group design. Psychological Bulletin, 1975, 82(3), 345-362.

Lord, F. M. A paradox in the interpretation of group comparisons. Psychological Bulletin, 1967, 68, 304-305.

Maxwell, S. E., & Jones, L. V. Female and male admission to graduate school: An illustrative inquiry. Journal of Educational Statistics, 1976, 1(1), 1-37.

Rubin, D. B. Matching to remove bias in observation studies. Biometrics, 1973, 29, 159-183. (Correction note 30, p. 728.) (a)

Rubin, D. B. The use of matched sampling and regression adjustment to remove bias in observational studies. Biometrics, 1973, 29, 184-203. (b)

Rubin, D. B. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology, 1974, 66, 688-701.

Rubin, D. B. Multivariate matching methods that are equal percent bias reducing, I: Some examples. Biometrics, 1976, 32, 109-120. (Correction note p. 955.) (a)

Rubin, D. B. Multivariate matching methods that are equal percent bias reducing II: Maximums on bias reduction for fixed sample sizes. Biometrics, 1976, 32, 121-132. (Correction note p. 955.) (b)

Rubin, D. B. Bayesian inference for causal effects: The role of randomization. Annals of Statistics, 1977, in press.

Snedecor, G. W., & Cochran, W. G. Statistical methods. Ames: Iowa State Press, 1967.

AUTHOR

RUBIN, DONALD B. Address: Educational Testing Service, Princeton, N.J. 08540. Title: Chairman, Statistics. Degrees: A.B. Princeton University, M.S. Harvard University, Ph.D. Harvard University. Specialization: Statistical inference for causal effects and problems with missing values.

[*Manuscript received February 1976; revised January 1977.*]