The Consquences of Adjustment for a Concomitant Variable That Has Been Affected by the Treatment

Paul R. Rosenbaum

Journal of the Royal Statistical Society. Series A (General), Vol. 147, No. 5 (1984), 656-666.

Stable URL:
http://links.jstor.org/sici?sici=0035-9238%281984%29147%3A5%3C656%3ATCOAFA%3E2.0.CO%3B2-F

Journal of the Royal Statistical Society. Series A (General) is currently published by Royal Statistical Society.

J. R. Statist. Soc. A (1984)
147, *Part* 5, *pp.* 656–666

The Consequences of Adjustment for a Concomitant Variable That Has Been Affected by the Treatment

By PAUL R. ROSENBAUM

Educational Testing Service, USA

SUMMARY

Adjustments for bias in observational studies are not always confined to variables that were measured prior to treatment. Estimators that adjust for a concomitant variable that has been affected by the treatment are generally biased. The bias may be written as the sum of two easily interpreted components: one component is present only in observational studies; the other is common to both observational studies and randomized experiments. The first component of bias will be zero when the affected posttreatment concomitant variable is, in a certain sense, a surrogate for an unobserved pretreatment variable. The second component of bias can often be addressed by an appropriate sensitivity analysis.

Keywords: OBSERVATIONAL STUDIES; RANDOMIZED EXPERIMENTS; COVARIANCE ADJUSTMENT; MATCHED SAMPLING; SUBCLASSIFICATION; CAUSAL EFFECTS OF TREATMENTS

1. INTRODUCTION

1.1. *Adjustment for Bias in Observational Studies*

As defined by Cochran (1965, p. 234), an observational study is an attempt to "elucidate cause-and-effect relationships [. . .in which it] is not feasible to use controlled experimentation, in the sense of being able to impose the procedures or treatments whose effects it is desired to discover, or to assign subjects at random to different procedures". As a consequence of the nonrandom assignment of treatments, the difference in response of treated and control groups may reflect, not effects caused by the treatment, but rather inherent differences between the subjects assigned to treatment and control. Adjustment for pretreatment differences is, therefore, a central concern in observational studies (e.g., Cochran and Rubin, 1973; Rubin, 1977).

In practice, adjustments for bias are not always confined to pretreatment variables, for several reasons. First, it may be believed that a quantity is unchanging, so that its posttreatment and pretreatment values are the same; this is certainly plausible for an attribute such as sex (i.e. gender). Second, although a quantity may change with time, it may be believed to be unaffected by the treatment, so that a difference between treated and control groups in a posttreatment measure of this quantity would indicate some form of bias rather than a treatment effect; see Cox (1956, p. 48–49) and Rosenbaum (1984a, Section 3.2.c) for related discussion. Third, although a quantity may be affected by the treatment, there may be reason to believe the effects are slight compared to the effects of the treatment on the response. Finally, pretreatment measures of an important quantity may simply be unavailable though posttreatment values are available; this is particularly common in studies based in part or in whole on past records.

1.2. *Example: Cognitive Outcomes in Public and Private Schools*

In a widely discussed and controversial study, Coleman, Hoffer and Kilgore (1982) compared public† and private high schools with respect to student's scores on standardized tests of reading,

† Throughout this paper, "public" is used in its American meaning of "state-run" schools.

Present address: Research Statistics Group, Educational Testing Service, Princeton, NJ 08541, USA.

 0035–9238/84/147656 $2.00

vocabulary and mathematics. They concluded, "The three types of analysis carried out in this paper provide strong evidence that there is, in vocabulary and mathematics, higher achievement for comparable students in Catholic and other-private schools than in public; the results are less consistent in reading." Commenting on this study, Goldberger and Cain (1982) wrote:

> CHK use 17 measured background variables to control initial differences among the students entering [. . .public and private schools] . . . we doubt that all 17 variables together can substitute for direct initial measures of cognitive achievement, such as would be provided by accurate [reading, verbal and mathematics] scores obtained just prior to entering high school. . .(p109)

Coleman, Hoffer and Kilgore (1982) and Goldberger and Cain (1982) perform adjustments of senior test scores for sophomore test scores. (These adjustments are complicated by certain limitations of the data that do not concern us here.) As both studies note, sophomore test scores may have been affected by differences between private and public high schools: the concomitant variable may have been affected by the treatment. In this instance, the posttreatment variable—"sophomore test scores"—is used in place of an unmeasured pretreatment variable—"test scores before high school".

Both studies note that several other concomitant variables may also have been affected by private *vs* public schooling, although Goldberger and Cain (1982, p. 110) argue that the effects are probably fairly small. (The examples in Sections 1.2 and 1.3 are included to clarify the notation and discussion in later sections; no attempt will be made to comment specifically on the diverging opinions about the Coleman, Hoffer and Kilgore report, since such comment would necessarily involve technical issues beyond the scope of this paper.)

1.3. *Example: Prenatal Exposure to Barbiturates and Psychological Development*

A second example is from an ongoing study of prenatal exposures to barbiturates and their effects on subsequent psychological development (e.g., Rosenbaum and Rubin 1985). This study involved adjustment for several pretreatment variables including (a) mother's marital status, (b) mother's socioeconomic status, (c) mother's education, (d) mother's height, (e) mother's age, (f) year of child's birth, and (h) child's sex (which is, of course, determined at conception and unaltered by treatment, although it is first observed at birth after treatment).

There was, however, a strong suspicion, probably justified, that mothers treated with barbiturates had greater access to medical care and were in poorer health than untreated mothers before the start of treatment; i.e. that the treated and control groups differed with respect to relevant unobserved pretreatment variables. To control for these suspected but unmeasured pretreatment differences, adjustments were also made for several posttreatment variables, including (a) hormone treatment during pregnancy, (b) antihistamine use during pregnancy, (c) length of gestation, (d) frequency of cigarette smoking during pregnancy and (e) an index of pregnancy complications. The exposed and unexposed groups differed substantially with respect to several of these variables, but these differences could reflect either effects of the treatment, or inherent unobserved pretreatment differences between the exposed and unexposed groups, or a combination of the two. The expectation was, however, that barbiturate treatment would not greatly affect these variables, so the observed differences tended to confirm the suspicion that the treated and control groups differed with respect to unobserved pretreatment characteristics (c.f. Rosenbaum 1984a, Section 3).

In both examples, adjustments for posttreatment variables are performed in lieu of adjustments for unmeasured pretreatment variables. The consequences of these adjustments are discussed in Section 3.

2. ADJUSTMENT FOR PRETREATMENT VARIABLES: A BRIEF REVIEW

2.1. *Estimating Treatment Effects in Randomized Experiments*

To motivate the discussion in Section 3 of the consequences of adjustment for affected concomitant variables, this section briefly reviews the rationale for adjustment for pretreatment

variables. Throughout, it is assumed that the subjects under study are a simple random sample from the relevant population.

Each subject has, in principle, two potential responses, R_1 and R_0, that would have been observed had the treatment been, respectively, applied (1) or withheld (0). In Section 1.2 for instance, R_1 is the senior test score of a specific child if sent to private school, and R_0 is the senior test score of the same child if sent to public school. Since each subject receives only one treatment, either R_1 or R_0 is observed, but not both. The effect of the treatment on a particular subject is defined as a comparison of R_1 and R_0, such as $R_1 - R_0$. This is the definition of treatment effects that has been used in the traditional literature on experimental design—e.g., Fisher (1935), Kempthorne, (1952, Section 8), and Cox (1958); the definition has been used in observational studies by Rubin (1974, 1977, 1978), Hamilton (1979), Holland (1979), and Holland and Rubin (1983), Rosenbaum and Rubin (1983a, b; 1984a) and Rosenbaum (1984a, b). The average effect of the treatment in a specified population of subjects is then

$$\tau = E(R_1 - R_0) = E(R_1) - E(R_0), \tag{2.1}$$

where $E(.)$ denotes expectation in the population.

For each subject, let the binary variable Z indicate whether treatment has been applied ($Z = 1$) or withheld ($Z = 0$). The response to the treatment, that is R_1, is observed only for subjects who received the treatment, that is subjects with $Z = 1$; similarly, R_0 is observed only for control subjects, that is subjects with $Z = 0$. Therefore, the difference in sample mean responses in the treated ($Z = 1$) and control ($Z = 0$) groups estimates

$$\Delta = E(R_1 \mid Z = 1) - E(R_0 \mid Z = 0), \tag{2.2}$$

which will be called the expected treatment difference. In randomized experiments, treatment assignment Z is independent of the response (R_1, R_0)—or in Dawid's (1979) notation, $(R_1, R_0) \perp\!\!\!\perp Z$—so that Δ, which can be directly estimated, equals the average treatment effect, τ. In observational studies, however, treatment assignment is generally not independent of the response, so Δ does not generally equal τ.

2.2. *Adjustment for Pretreatment Covariates in Observational Studies*

If $\mathbf{X}$ is a vector of covariates—that is, a vector of variables measurable prior to the assignment of treatments—then the expected $\mathbf{X}$-adjusted treatment difference, $\bar{\Delta}$, is defined as

$$\bar{\Delta} = E_{\mathbf{X}}\{E(R_1 \mid Z = 1, \mathbf{X}) - E(R_0 \mid Z = 0, \mathbf{X})\}, \tag{2.3}$$

where $E_{\mathbf{X}}(.)$ denotes expectation with respect to the distribution of $\mathbf{X}$ in the population. To estimate $\bar{\Delta}$, the three basic methods of adjustment described by Cochran (1965)—that is, subclassification, matched sampling, and covariance adjustment—may be used, either alone or in combinations; see Rubin (1977) and Rosenbaum and Rubin (1983a; 1984a, b, 1985) for discussion and examples. For instance, one straightforward approach to estimating $\bar{\Delta}$ is the following matching-differencing-averaging procedure: (a) randomly sample a value of $\mathbf{X}$ from the population; (b) randomly sample a treated and a control subject with this value of $\mathbf{X}$; (c) subtract the response of the sampled control subject from the response of the sample treated subject; (d) repeat steps (a)-(c), and average the differences to obtain an unbiased estimate of $\bar{\Delta}$.

2.3. *When is it Sufficient to Adjust for the Observed Covariates? Strongly Ignorable Treatment Assignment*

Although $\bar{\Delta}$ can be directly estimated, it does not generally equal the average treatment effect, τ. However, $\bar{\Delta}$ does equal τ when treatment assignment is strongly ignorable for (R_1, R_0) given $\mathbf{X}$, as defined by Rosenbaum and Rubin (1983a), that is when

$$(R_1, R_0) \perp\!\!\!\perp Z \mid \mathbf{X}, \tag{2.4a}$$

and

$$0 < \Pr(Z = 1 \mid \mathbf{X}) < 1 \text{ for all } \mathbf{X}, \tag{2.4b}$$

where in Dawid's (1979) notation $A \perp\!\!\!\perp B \mid C$ states A and B are conditionally independent given C. Treatment assignment is strongly ignorable in completely randomized experiments in which treatments are assigned by the flip of a fair coin, and more generally in randomized experiments in which treatments are assigned by the flip of a biased coin, where the bias is a (possibly unknown) function of **X** alone (Rubin, 1977). Under certain conditions, treatment assignment can be strongly ignorable if assignments are a deterministic function of **X** and certain irrelevant unobserved covariates (Rosenbaum, 1984a, Section 2.3). Condition (2.4b) ensures that, in an infinite population, there will be both treated and control subjects at each value of **X**, so that for example, appropriate matches may be found. In short, if treatment assignment is strongly ignorable for (R_1, R_0) given the pretreatment variables in **X**, then appropriate adjustment for **X** is sufficient to directly estimate the average treatment effect, τ.

Ignorable treatment assignment is a fairly tenuous assumption in many observational studies. It is often possible to assess the sensitivity of conclusions to plausible violations of this assumption (e.g., Rosenbaum and Rubin, 1983b). Moreover, the assumption can often be tested by contrasting observed results with the predictions of a causal theory (Rosenbaum, 1984a).

In Section 1.2, one interpretation of the comments by Goldberger and Cain is that they judged the assumption of ignorable treatment assignment given the 17 measured background variables to be untenable in the Coleman, Hoffer, Kilgore study. They seemed to imply, however, that they would have judged this assumption to be more plausible if the unmeasured "test scores before high school" could have been included in **X**.

3. THE CONSEQUENCES OF ADJUSTMENT FOR A CONCOMITANT VARIABLE THAT HAS BEEN AFFECTED BY THE TREATMENT

3.1. *Overview: Should We Supplement Adjustments For Pretreatment Variables With Additional Adjustments for Posttreatment Variables?*

This section examines the large sample bias of estimators that adjust for both an affected posttreatment concomitant variable and the unaffected pretreatment covariates in **X**. In Section 3.2, which parallels Section 2.2, an expression for the large sample expectation of the estimator is derived. The bias of the estimator is most conveniently studied in relation to an additional quantity—the net treatment difference—which is discussed in Sections 3.3 and 3.4. In Sections 3.5 and 3.6, properties of the bias are examined under a variety of assumptions. In particular, in Section 3.5 it is observed that if it is sufficient to adjust for the observed pretreatment covariates—that is, if treatment assignment is ignorable given **X**—then additional adjustments for a posttreatment variable can introduce an avoidable bias. On the other hand, it observed in Section 3.6 that if it is not sufficient to adjust for the observed pretreatment covariates—in particular, if treatment assignment is not ignorable given **X** alone, but is ignorable given (**X**, **U**) where **U** is a vector of unobserved pretreatment covariates—then adjustment for **X** alone will often yield biased estimates, while adjustment for **X** and a posttreatment variable can, under certain fairly restrictive assumptions, yield unbiased estimates. Alternative methods of analysis are discussed in Section 4, including procedures for studying the sensitivity of conclusions to violations of the restrictive assumptions in Section 3.6 that concern the effect of the treatment on the posttreatment concomitant variable.

3.2. *The Expected Difference in Responses After Adjustment for an Affected Concomitant Variable*

For each subject, let $(\mathbf{S}_1, \mathbf{S}_0)$ be the pair of potentially observable values of a posttreatment concomitant variable, where $\mathbf{S}_1$ and $\mathbf{S}_0$ are, respectively, the values that would have been observed had the treatment been applied or withheld. Note that there are two versions of the posttreatment variable—$\mathbf{S}_1$ and $\mathbf{S}_0$—whereas there is only a single version of the pretreatment covariate **X**, which could not have been affected by a treatment that has not yet been applied (c.f. Holland and Rubin 1983, Section 2.2). In Section 1.2 for instance, S_1 is the sophomore test score of a specific

child if sent to private school, and S_0 is the sophomore test score of the same child if sent to public school. If Coleman, Hoffer, and Kilgore (1982) were correct in their claim that private schools outperform public schools, we might expect the same pattern in both the sophomore and senior year, so that $S_1 > S_0$ and $R_1 > R_0$ for most if not all children.

We now examine the expected difference in the response (R_1, R_0) in the treated and control groups after adjustment for both the observed pretreatment variables $\mathbf{X}$ and the observed value, $\mathbf{S}_Z$, of the posttreatment concomitant variable $(\mathbf{S}_1, \mathbf{S}_0)$. Suppose, for example, the following matching-differencing-averaging procedure is used to make the adjustments: a value of $(\mathbf{X}, \mathbf{S}_Z)$ is randomly sampled from the population, and then a treated subject and a control subject are each randomly sampled from among the treated subjects and the control subjects with this value of $(\mathbf{X}, \mathbf{S}_Z)$. Conditional on the sampled value, $(\mathbf{x}, \mathbf{s})$, of $(\mathbf{X}, \mathbf{S}_Z)$, the expected difference in the responses of the sampled treated ($Z = 1$) and control ($Z = 0$) subjects is

$$\begin{aligned}\Delta(\mathbf{x}, \mathbf{s}) &= E(R_1 \mid Z = 1, \mathbf{S}_Z = \mathbf{s}, \mathbf{X} = \mathbf{x}) - E(R_0 \mid Z = 0, \mathbf{S}_Z = \mathbf{s}, \mathbf{X} = \mathbf{x}) \\ &= E(R_1 \mid Z = 1, \mathbf{S}_1 = \mathbf{s}, \mathbf{X} = \mathbf{x}) - E(R_0 \mid Z = 0, \mathbf{S}_0 = \mathbf{s}, \mathbf{X} = \mathbf{x}) \qquad (3.1)\end{aligned}$$

so the expected difference in responses after adjustment for $(\mathbf{X}, \mathbf{S}_Z)$ is

$$\tilde{\Delta} = E\{\Delta(\mathbf{X}, \mathbf{S}_Z)\}. \qquad (3.2)$$

The quantity $\tilde{\Delta}$ will be called the expected $(\mathbf{X}, \mathbf{S}_Z)$-adjusted treatment difference. As with the expected $\mathbf{X}$-adjusted treatment difference, $\bar{\Delta}$, we can estimate $\tilde{\Delta}$ not just by matched sampling, but alternatively by covariance adjustment or subclassification.

3.3. *The Net Treatment Difference and its Role in Studying the Treatment Mechanism*

The bias, $\tilde{\Delta} - \tau$, of an estimator that adjusts for $(\mathbf{X}, \mathbf{S}_Z)$ can most easily be studied using an additional quantity, the net treatment difference at $(\mathbf{x}, \mathbf{s})$, defined by

$$\nu(\mathbf{x}, \mathbf{s}) = E(R_1 \mid \mathbf{S}_1 = \mathbf{s}, \mathbf{X} = \mathbf{x}) - E(R_0 \mid \mathbf{S}_0 = \mathbf{s}, \mathbf{X} = \mathbf{x}), \qquad (3.3)$$

and its expectation over the distribution of $(\mathbf{X}, \mathbf{S}_Z)$, namely,

$$\tilde{\nu} = E\{\nu(\mathbf{X}, \mathbf{S}_Z)\}. \qquad (3.4)$$

The net treatment difference is a quantity that is most often discussed in an attempt to provide insight into the mechanism or process by which the treatment produces its effects. Cochran (1957, Section 2.3), for example, considers an experiment in which soil fumigants are used to increase crop yields. Here R_1 and R_0 are, respectively, the oat yields on a plot that would have been observed in the presence (1) or absence (0) of the fumigant; S_1 and S_0 are the corresponding numbers of eelworms found on the plots. The fumigant is thought to affect both (R_1, R_0) and (S_1, S_0). If the fumigant produces a substantial increase in oat yield (i.e., if $\tau \gg 0$), then this effect might be due entirely to control of the damage done by eelworms; in this case we might expect $\nu(\mathbf{x}, \mathbf{s}) = 0$ for all $(\mathbf{x}, \mathbf{s})$, so that a fumigated plot and an unfumigated plot that happened to have the same number of eelworms would have the same expected oat yield. If $\tau \gg 0$ and $\tilde{\nu} > 0$, then the fumigant appears to have beneficial effects besides the control of eelworms. If $\tau \gg 0$ but $\tilde{\nu} < 0$, then the fumigant has a positive overall affect, but a comparison of the oat yields of a fumigated plot and an unfumigated plot which happened to have equal numbers of eelworms would favour the unfumigated plot. In this last case, the fumigant appears to produce a positive effect, perhaps by controlling the eelworm population, but the treatment also appears to have smaller detrimental side effects. Cox (1958, Section 4), Yates (1960, Section 9.11) and Holland and Rubin (1983, Section 3.3) discuss other examples in which the net treatment difference provides insight into the treatment mechanism.

Although the net treatment difference is often a parameter of considerable interest, it is not generally a treatment effect; that is, generally, there may be no actual treatment whose average effect is $\tilde{\nu}$. For example, if fumigants are to be used to kill eelworms, thereby increasing overall oat yield, it may be necessary to accept some minor detrimental side effects of the fumigants:

there may be no actual treatment that produces either effect without also producing the other. See Box (1966) and Lord (1969) for closely related discussion.

3.4. *The Bias of Estimators That Adjust for an Affected Concomitant Variable*

It is convenient to write the bias of an estimator that adjusts for $(\mathbf{X}, \mathbf{S}_Z)$ as

$$\tilde{\Delta} - \tau = (\tilde{\Delta} - \tilde{\nu}) + (\tilde{\nu} - \tau). \tag{3.5}$$

The $(\tilde{\nu} - \tau)$ component of the bias is present in both randomized experiments and observational studies, whereas, the $(\tilde{\Delta} - \tilde{\nu})$ component is peculiar to observational studies.

Comparison of (3.1) and (3.3) shows that $(\tilde{\Delta} - \tilde{\nu})$ measures the difference between the regression of R_1 on $\mathbf{S}_1$ and $\mathbf{X}$ in the treated group, $E(R_1 \mid Z = 1, \mathbf{S}_1, \mathbf{X})$, and the corresponding regression in the population, $E(R_1 \mid \mathbf{S}_1, \mathbf{X})$, as well as difference between the regression, $E(\mathbf{R}_0 \mid Z = 0, \mathbf{S}_0, \mathbf{X})$, of R_0 on $\mathbf{S}_0$ and $\mathbf{X}$ in the control group and the corresponding regression, $E(R_0 \mid \mathbf{S}_0, \mathbf{X})$, in the population. In other words, $(\tilde{\Delta} - \tilde{\nu})$ measures the degree to which the nonrandom assignment of treatments (Z) alters certain population regressions.

Unlike $(\tilde{\Delta} - \tilde{\nu})$ which measures the bias in estimating a conditional expectation, the $(\tilde{\nu} - \tau)$ component of bias measures the consequences of the inappropriate averaging of a conditional expectation to obtain a marginal expectation. More specifically, $(\tilde{\nu} - \tau)$ measures the consequences of averaging both $E(R_1 \mid \mathbf{S}_1, \mathbf{X})$ and $E(R_0 \mid \mathbf{S}_0, \mathbf{X})$ over the distribution $\Pr(\mathbf{S}_Z \mid \mathbf{X})$ in (3.4) rather than the appropriate distributions: $\Pr(\mathbf{S}_1 \mid \mathbf{X})$ and $\Pr(\mathbf{S}_0 \mid \mathbf{X})$, respectively.

3.5. *Some Properties of the Bias Under Various Assumptions*

(i) *Unaffected concomitant variables.* If the treatment has no effect on the posttreatment variables $(\mathbf{S}_1, \mathbf{S}_0)$, so that

$$\mathbf{S}_1 = \mathbf{S}_0 = \mathbf{S}_Z, \text{ for all subjects in the population,} \tag{3.6}$$

then, from (3.3) and (3.4), the average net treatment difference equals the average treatment effect, $\tilde{\nu} = \tau$. Condition (3.6) is not, however, sufficient to ensure that the expected $(\mathbf{X}, \mathbf{S}_Z)$-adjusted treatment difference, $\tilde{\Delta}$, equals the average treatment effect, τ.

(ii) *Ignorable treatment assignment.* If treatment assignment is strongly ignorable for $\{(R_1, \mathbf{S}_1), (R_0, \mathbf{S}_0)\}$ given $\mathbf{X}$ – that is, if

$$(R_1, \mathbf{S}_1, R_0, \mathbf{S}_0) \perp\!\!\!\perp Z \mid \mathbf{X} \tag{3.7}$$

and

$$0 < \Pr(Z = 1 \mid \mathbf{X}) < 1 \text{ for all } \mathbf{X} \tag{3.8}$$

– then the $(\mathbf{X}, \mathbf{S}_Z)$-adjusted difference $\tilde{\Delta}$ equals the average net treatment difference $\tilde{\nu}$. To see this, note that (3.7) implies

$$R_t \perp\!\!\!\perp Z \mid \mathbf{S}_t, \mathbf{X} \text{ for } t = 0, 1 \tag{3.9}$$

by familiar properties of conditional independence, or Lemma 4 of Dawid (1979), hence, $\Delta(\mathbf{x}, \mathbf{s}) = \nu(\mathbf{x}, \mathbf{s})$ in (3.1) and (3.3), so $\tilde{\Delta} = \tilde{\nu}$ in (3.2) and (3.4). Informally, if adjustment for $\mathbf{X}$ is sufficient to remove bias in both (R_1, R_0) and $(\mathbf{S}_0, \mathbf{S}_1)$, then the $(\mathbf{X}, \mathbf{S}_Z)$-adjusted difference $\tilde{\Delta}$ equals the average net treatment difference $\tilde{\nu}$. Note in particular that in conventional randomized experiments, (3.7) and (3.8) hold, so $\tilde{\Delta} = \tilde{\nu}$.

(iii) *A sufficient condition for* $\tilde{\Delta} = \tau$. Combining (i) and (ii) above, we see that the expected $(\mathbf{X}, \mathbf{S}_Z)$-adjusted treatment difference, $\tilde{\Delta}$, will equal the average treatment effect, τ, when the treatment has no effect on the posttreatment concomitant variables $(\mathbf{S}_1, \mathbf{S}_0)$ and treatment assignment is strongly ignorable for $\{(R_1, \mathbf{S}_1), (R_0, \mathbf{S}_0)\}$ given $\mathbf{X}$. However, strong ignorability for $\{(R_1, \mathbf{S}_1), (R_0, \mathbf{S}_0)\}$ given $\mathbf{X}$ implies strong ignorability for (R_1, R_0) given $\mathbf{X}$, so that the expected $\mathbf{X}$-adjusted treatment difference, $\bar{\Delta}$, also equals the average treatment effect, τ. Informally, if the treatment has no effect on $(\mathbf{S}_1, \mathbf{S}_0)$, and if adjustment for $\mathbf{X}$ is sufficient to

remove bias in both (R_1, R_0) and $(\mathbf{S}_1, \mathbf{S}_0)$, then adjustment for $(\mathbf{X}, \mathbf{S}_Z)$ instead of $\mathbf{X}$ neither introduces nor removes bias.

Cox (1958, p. 49) discusses an example: a randomized experiment in a textile mill, in which (R_1, R_0) measures an aspect of the textiles produced, and (S_1, S_0) is the relative humidity in the mill measured during processing. Since the experiment is randomized, (3.7) and (3.8) hold, so $\tilde{\Delta} = \tilde{\nu}$. Moreover, the treatment is such that it could not possibly affect the relative humidity in the mill, so $S_1 = S_0 = S_Z$, and $\tilde{\nu} = \tau$. As a result, appropriate adjustment for the relative humidity during processing yields unbiased estimates of the average treatment effect τ, and moreover as Cox notes, it may increase the efficiency of estimation. This adjustment is not, however, necessary for the control of bias, since in randomized experiments the difference in sample mean responses is also unbiased for τ.

(iv) *The general case.* In general, $\tilde{\Delta}$ does not equal the average treatment effect, τ, nor the $\mathbf{X}$-adjusted treatment difference, $\bar{\Delta}$, nor the average net treatment difference, $\tilde{\nu}$. In particular, if treatment assignment is strongly ignorable for (R_1, R_0) given $\mathbf{X}$, then $\bar{\Delta} = \tau$ so adjustment for $\mathbf{X}$ alone is sufficient to remove bias, but $\tilde{\Delta}$ need not equal τ; i.e., adjustment for $(\mathbf{X}, \mathbf{S}_Z)$ can introduce a bias that could have been avoided by simply confining adjustments to pretreatment variables $\mathbf{X}$.

(v) *A parallel linear model.* To examine the $(\tilde{\nu} - \tau)$ component of bias, consider the following linear model:

$$E(R_1 \mid \mathbf{S}_1 = \mathbf{s}, \mathbf{X} = \mathbf{x}) = \alpha_1 + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x} + \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{s} \tag{3.13a}$$

$$E(R_0 \mid \mathbf{S}_0 = \mathbf{s}, \mathbf{X} = \mathbf{x}) = \alpha_0 + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x} + \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{s}, \tag{3.13b}$$

which resembles the models often used in covariance analysis. Then from (2.1)

$$\begin{aligned}\tau &= E\{(\alpha_1 + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X} + \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{S}_1) - (\alpha_0 + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X} + \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{S}_0)\} \\ &= \alpha_1 - \alpha_0 + \boldsymbol{\gamma}^{\mathrm{T}}\{E(\mathbf{S}_1) - E(\mathbf{S}_0)\},\end{aligned}$$

whereas from (3.3) and (3.4) we have,

$$\begin{aligned}\tilde{\nu} &= E\{\alpha_1 + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X} + \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{S}_Z) - (\alpha_0 + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X} + \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{S}_Z)\} \\ &= \alpha_1 - \alpha_0\end{aligned}$$

so that

$$\tilde{\nu} - \tau = -\boldsymbol{\gamma}^{\mathrm{T}}\{E(\mathbf{S}_1) - E(\mathbf{S}_0)\}. \tag{3.14}$$

Hence, $\tilde{\nu} - \tau$ is a linear function of the average effect of the treatment on $(\mathbf{S}_1, \mathbf{S}_0)$; that is, a linear function of $E(\mathbf{S}_1) - E(\mathbf{S}_0)$. From (3.14), the $\tilde{\nu} - \tau$ component of bias can be either positive or negative.

In summary, under the simple conditions considered in this section, adjustments for posttreatment variables—i.e., adjustments for $(\mathbf{X}, \mathbf{S}_Z)$ instead of adjustments for $\mathbf{X}$ alone—are justified only when they are unnecessary; that is the conditions that lead to $\tilde{\Delta} = \tau$ also imply $\bar{\Delta} = \tau$. The next section shows that, under certain conditions, adjustment for $(\mathbf{X}, \mathbf{S}_Z)$ can yield unbiased estimates of τ when adjustment for $\mathbf{X}$ alone is insufficient; however, this adjustment for $(\mathbf{X}, \mathbf{S}_Z)$ is appropriate only under strong assumptions both about the effect of the treatment on $(\mathbf{S}_1, \mathbf{S}_0)$ and about unobserved covariates.

3.6. *Posttreatment Variables as Surrogates for Pretreatment Covariates*

Adjustment for a posttreatment concomitant variable, $(\mathbf{S}_1, \mathbf{S}_0)$, is often performed in lieu of adjustment for a pretreatment covariate, $\mathbf{U}$, that is simply unavailable; e.g., in Section 1.2 adjustment for observed "sophomore test scores" (i.e. $\mathbf{S}_Z$) instead of "test scores before high school" (i.e., $\mathbf{U}$). In this section, we examine conditions under which adjustment for $(\mathbf{X}, \mathbf{S}_Z)$ controls

bias due to $(\mathbf{X}, \mathbf{U})$.

Suppose that treatment assignment is strongly ignorable for $\{(R_1, \mathbf{S}_1), (R_0, \mathbf{S}))\}$ given $(\mathbf{X}, \mathbf{U})$ – that is, suppose

$$(R_1, \mathbf{S}_1, R_0, \mathbf{S}_0) \perp\!\!\!\perp Z \mid (\mathbf{X}, \mathbf{U}) \tag{3.15}$$

and

$$0 < \Pr(Z = 1 \mid \mathbf{X} = \mathbf{x}, \mathbf{U} = \mathbf{u}) < 1 \text{ for all } \mathbf{x}, \mathbf{u} \tag{3.16}$$

– so that, if $\mathbf{U}$ had been measured, appropriate adjustment for $(\mathbf{X}, \mathbf{U})$ would have been sufficient to estimate the effect of the treatment on both (R_1, R_0) and $(\mathbf{S}_1, \mathbf{S}_0)$. A similar assumption has been discussed by Rosenbaum and Rubin (1983b) and Rosenbaum (1984a).

A posttreatment concomitant $(\mathbf{S}_1, \mathbf{S}_0)$ will be called a *surrogate* for the unobserved pretreatment covariate, $\mathbf{U}$, if

$$R_t \perp\!\!\!\perp \mathbf{U} \mid (\mathbf{S}_t, \mathbf{X}) \text{ for } t = 0, 1. \tag{3.17}$$

If, as in Section 1.2, the variables $\mathbf{U}$, $(\mathbf{S}_1, \mathbf{S}_0)$ and (R_1, R_0) are measures of a similar quantity at three consecutive times, then (3.17) is a kind of Markov condition that applies to the potential responses under each treatment: informally, for each treatment, the late posttreatment response (R_t = senior test scores) and the pretreatment response ($\mathbf{U}$ = test scores before high school) are unrelated given the early posttreatment response ($\mathbf{S}_t$ = sophomore test scores). Assumption (3.17) may be plausible when the subjects are growing, learning, maturing, aging or succumbing to a chronic disease, and $\mathbf{U}$, $(\mathbf{S}_1, \mathbf{S}_0)$ and (R_1, R_0) are consecutive measures of progress, achievement, development or deterioration.

If $(\mathbf{S}_1, \mathbf{S}_0)$ is a surrogate for $\mathbf{U}$, then the $(\tilde{\Delta} - \tilde{\nu})$ component of bias is zero. More formally, *if treatment assignment is strongly ignorable for* $\{(R_1, \mathbf{S}_1), (R_0, \mathbf{S}_0)\}$ *given both* $\mathbf{X}$ *and the unobserved* $\mathbf{U}$, *and if* $(\mathbf{S}_1, \mathbf{S}_0)$ *is a surrogate for* $\mathbf{U}$, *then* $\tilde{\Delta} = \tilde{\nu}$. (To prove this, note that by Section 3.5 (ii), it is sufficient to show that (3.9) holds. Now (3.15) implies

$$R_t \perp\!\!\!\perp Z \mid (\mathbf{S}_t, \mathbf{X}, \mathbf{U}) \text{ for } t = 0, 1 \tag{3.18}$$

by familiar properties of conditional independence, or Lemma 4 of Dawid (1979). Together (3.17) and (3.18) imply

$$R_t \perp\!\!\!\perp (Z, \mathbf{U}) \mid (\mathbf{S}_t, \mathbf{X}) \text{ for } t = 0, 1, \tag{3.19}$$

again using familiar properties of conditional independence or Dawid's Lemma 4. Of course, (3.19) implies (3.9) as required.)

Combining this observation with Section 3.5 (i), we see that $\tilde{\Delta}$ *equals the average treatment effect* τ *if three conditions had simultaneously*:

(a) *treatment assignment is strongly ignorable for* $\{(R_1, \mathbf{S}_1), (R_0, \mathbf{S}_0)\}$ *given* $(\mathbf{X}, \mathbf{U})$,

(b) $(\mathbf{S}_1, \mathbf{S}_0)$ *is a surrogate for* $\mathbf{U}$, *and*

(c) $(\mathbf{S}_1, \mathbf{S}_0)$ *is unaffected by the treatment, so that* $\mathbf{S}_1 = \mathbf{S}_0$;

i.e., conditions (a)-(c) imply that adjustment for the observed $(\mathbf{X}, \mathbf{S}_Z)$ removes bias due to $(\mathbf{X}, \mathbf{U})$, even though $\mathbf{U}$ is not observed.

Note, however, that the assumption that $(\mathbf{S}_1, \mathbf{S}_0)$ is entirely unaffected is dubious in the example of Section 1.2, and is without strong support in the example of Section 1.3.

4. ALTERNATIVE METHODS OF ANALYSIS

4.1. *Avoid Adjustment for Posttreatment Variables*

The simplest approach is to avoid all adjustments for posttreatment variables. For the control of bias, this approach is clearly appropriate in most randomized experiments, and more generally whenever treatment assignment is strongly ignorable given $\mathbf{X}$ (see Section 3.5 iv). However, when treatment assignment is not strongly ignorable given $\mathbf{X}$, this approach can be unsatisfactory, especially when the posttreatment variable $(\mathbf{S}_1, \mathbf{S}_0)$ is thought to be closely related to an un-

measured pretreatment variable (**U**) that is relevant to both treatment assignment (Z) and response (R_1, R_0), as in Sections 1.2 and 1.3. In such cases, the **X**-adjusted treatment difference $\bar{\Delta}$ may be no better as an approximation to τ than the $(\mathbf{X}, \mathbf{S}_Z)$-adjusted difference $\tilde{\Delta}$.

4.2. *Compare Adjusted and Unadjusted Estimates*

One straightforward method that is not infrequently used in practice (e.g., Coleman, Hoffer, and Kilgore, 1982) involves estimating both the **X**-adjusted difference $\bar{\Delta}$ and the $(\mathbf{X}, \mathbf{S}_Z)$-adjusted difference $\tilde{\Delta}$. The implicit assumption is that τ lies somewhere between $\bar{\Delta}$ and $\tilde{\Delta}$. If this assumption were correct, then qualitatively similar estimates of $\bar{\Delta}$ and $\tilde{\Delta}$—i.e., a small estimated value of $|\bar{\Delta} - \tilde{\Delta}|$—would suggest a narrow range of estimated values for τ. Unfortunately, the implicit assumption that τ lies between $\bar{\Delta}$ and $\tilde{\Delta}$ is not true without additional conditions. However, the comparison of estimates of $\bar{\Delta}$ and $\tilde{\Delta}$ is a particular instance of the more general sensitivity analysis described in the next section.

4.3. *Sensitivity Analysis Under Linear Regression Models*

Often, we can study the sensitivity of conclusions to assumptions about unobserved covariates (e.g., Rosenbaum and Rubin 1983b). For example, when the structure in Section 3.6 applies—that is, when treatment assignment is ignorable given observed and unobserved pretreatment covariates $(\mathbf{X}, \mathbf{U})$, and $(\mathbf{S}_1, \mathbf{S}_0)$ is a surrogate for **U**—then a relatively simple sensitivity analysis is often possible. Under these assumptions, $\tilde{\Delta} = \tilde{\nu}$, so the bias (3.5) of $\tilde{\Delta}$ is $\tilde{\nu} - \tau$. Moreover, these assumptions imply (3.19) and (3.9), so the regressions $E(R_1 \mid \mathbf{S}_1 = \mathbf{s}, \mathbf{X} = \mathbf{x})$ and $E(R_0 \mid \mathbf{S}_0 = \mathbf{s}, \mathbf{X} = \mathbf{x})$ can be directly estimated from the data (i.e., formally, for $t = 0, 1$, the directly estimable regression $E(R_t \mid Z = t, \mathbf{S}_t = \mathbf{s}, \mathbf{X} = \mathbf{x})$ equals the relevant population regression $E(R_t \mid \mathbf{S}_t = \mathbf{s}, \mathbf{X} = \mathbf{x})$). In particular, the parameters of the linear model in (3.13) can be directly estimated, for example by least squares regression of R_Z on $(Z, \mathbf{X}, \mathbf{S}_Z)$. Under the model (3.13) we have by (3.14) and the equality of $\tilde{\Delta}$ and $\tilde{\nu}$,

$$\tau = \tilde{\Delta} + \boldsymbol{\gamma}^{\mathrm{T}} \{E(\mathbf{S}_1 - \mathbf{S}_0)\} \tag{4.1}$$

where $\tilde{\Delta} = \alpha_1 - \alpha_0$ and $\boldsymbol{\gamma}$ are directly estimable. We cannot estimate $E(\mathbf{S}_1 - \mathbf{S}_0)$ directly, since this would require adjustment for **X** and the unobserved **U**. Expression (4.1) may, however, be used to examine the range of plausible values of τ as assumptions about $E(\mathbf{S}_1 - \mathbf{S}_0)$—the average effect of treatment on $(\mathbf{S}_1, \mathbf{S}_0)$—are varied. For example, if it is possible to determine bounds on the size of the effect of the treatment on $(\mathbf{S}_1, \mathbf{S}_0)$—i.e., bounds on $E(\mathbf{S}_1 - \mathbf{S}_0)$—perhaps using data from past studies or perhaps using logical arguments, then (4.1) may be used to estimate bounds on τ. Alternatively, the method described by Rosenbaum and Rubin (1983a) may be used to examine the range of plausible values of $E(\mathbf{S}_1 - \mathbf{S}_0)$ for various assumptions about **U**; this method is appropriate when the coordinates of $\mathbf{S}_t$ are binary.

When all adjustments are made by linear least squares covariance adjustment, the two estimates in Section 4.2 are particular instances of estimates obtained from (4.1), under different assumptions about $E(\mathbf{S}_1 - \mathbf{S}_0)$. Clearly, if it is assumed that the treatment has no effect on $(\mathbf{S}_1, \mathbf{S}_0)$, so $E(\mathbf{S}_1 - \mathbf{S}_0) = 0$, then $\tau = \tilde{\Delta}$ in (4.1), yielding one of the estimates in Section 4.2. Suppose instead that $E(\mathbf{S}_1 - \mathbf{S}_0)$ is estimated using linear least squares covariance adjustment for **X**, ignoring possible biases due to the unobserved **U**; that is, $E(\mathbf{S}_1 - \mathbf{S}_0)$ is estimated as the (vector) coefficient of Z in the (multivariate) regression of $\mathbf{S}_Z$ on $(Z, \mathbf{X})$. Then standard arguments (Seber, 1977, p. 66, Theorem 3.7) imply the estimate of τ obtained by substitution in (4.1) is the usual estimate of $\bar{\Delta}$ obtained from least squares covariance adjustment for **X** alone; that is, from linear regression of R_Z on $(Z, \mathbf{X})$. Whether either of these two assumptions about $E(\mathbf{S}_1 - \mathbf{S}_0)$ is appropriate will depend on the context.

4.4. *Sensitivity Analysis For Discrete* **X** *and Binary* $(\mathbf{S}_1, \mathbf{S}_0)$

If X is a discrete variable taking values $x = 1, 2, \ldots, K$, and if S_1 and S_0 are each binary valued, then τ may be written,

$$\tau = \sum_{x=1}^{K} \sum_{s=0}^{1} [E(R_1 \mid S_1 = s, X = x) \Pr(S_1 = s \mid X = x) -$$

$$E(R_0 \mid S_0 = s, X = x) \Pr(S_0 = s \mid X = x)] \Pr(X = x).$$

Since X is observed for all subjects, an unbiased and consistent estimate of $\Pr(X = x)$ is the sample proportion of subjects with $X = x$. When the structure in Section 3.6 applies—that is when treatment assignment is strongly ignorable given observed and unobserved pretreatment covariates $(\mathbf{X}, \mathbf{U})$, and (S_1, S_0) is a surrogate for $\mathbf{U}$—then by (3.19) a consistent estimate of $E(R_1 \mid S_1 = s, X = x)$ is the sample mean response of all treated $(Z = 1)$ subjects with $S_1 = s$ and $X = x$, and similarly a consistent estimate of $E(R_0 \mid S_0 = s, X = x)$ is the sample mean of all control $(Z = 0)$ subjects with $S_0 = s$ and $X = x$. Although $\Pr(S_1 = s \mid X = x)$ and $\Pr(S_0 = s \mid X = x)$ cannot be estimated directly without additional assumptions because of confounding due to $\mathbf{U}$, they may be studied under a range of plausible assumptions about $\mathbf{U}$ (Rosenbaum and Rubin 1983b, Appendix), thereby producing a range of plausible values for τ.

4.5. *Redefine the Estimand*

Commonly, treatment effects are estimated to determine whether and when the treatment should be applied. Less commonly, treatment effects are estimated to shed light on some physical, biological or social process. In the latter case, it may not be entirely unreasonable to take the net treatment difference, $\tilde{\nu}$, as the estimand, keeping in mind, of course, that $\tilde{\nu}$ is not generally the average effect of any actual treatment. As noted in Section 3.6, $\tilde{\nu}$ can be estimated under weaker assumptions than are required to estimate τ. It is important to recall, however, that $\tilde{\nu}$ and τ need not have the same sign; see the discussion of the effects of fumigants in Section 3.3.

5. SUMMARY

In randomized experiments, adjustments for posttreatment concomitant variables should be avoided when estimating treatment effects, except in certain rather special circumstances (e.g., Section 3.5 (iii)), since the adjustments themselves can introduce a bias where none existed previously.

In observational studies, the situation is more complicated: no single course of action is appropriate for all such studies. Estimators that adjust for the pretreatment covariates $(\mathbf{X})$ alone and estimators that adjust for both pretreatment and posttreatment variables $(\mathbf{X}, \mathbf{S}_Z)$ can each lead to biased estimates, but for quite different reasons. Adjustment for a posttreatment variable may be advisable in either of two circumstances: (a) when, as in Section 1.2, a posttreatment variable (e.g., sophomore test scores) is a plausible surrogate for a clearly relevant but unobserved pretreatment variable (e.g., test scores before high school), or (b) when, as in Section 1.3, even after adjustment for $\mathbf{X}$, the treated and control groups differ substantially with respect to a posttreatment variable (e.g., hormone use) for which a large treatment effect was not expected. In both cases, it will typically be necessary to examine the sensitivity of conclusions to a range of plausible assumptions about the effect of the treatment on the posttreatment variable (e.g. Sections 4.3, and 4.4).

REFERENCES

Box, G. E. P. (1966) The use and abuse of regression. *Technometrics*, 8, 625–629.

Cochran, W. G. (1957) Analysis of covariance: Its nature and uses. *Biometrics*, **13**, 261–281.

——— (1965) The planning of observational studies of human populations. *J. R. Statist. Soc.* A, **182**, 234–255.

Cochran, W. G. and Rubin, D. (1973) Controlling bias in observational studies: A review. *Sankyā* A, **35**, 417–446.

Coleman, J., Hoffer, T. and Kilgore, S. (1982) Cognitive outcomes in public and private schools. *Sociology of Education*, **55**, 65–76.

Cox, D. R. (1958) *The Planning of Experiments*. New York: Wiley.

Dawid, A. P. (1979) Conditional independence in statistical theory (with Discussion). *J. R. Statist. Soc.* B, **41**, 1–31.

Dunn, R. and Wrigley, N. (1985) Beta-logistic models of urban shopping centre choice. *Geographical Analysis.* **17**(2).

Goldberger, A. S. and Cain, G. S. (1982) The causal analysis of cognitive outcomes in the Coleman, Hoffer and Kilgore report. *Sociology of Education*, **55**, 103–122.

Holland, P. W. (1979) The tyranny of continuous models in a world of discrete data. *IHS-Journal*, (Physica-Verlag), **3**, 29–42.

Holland, P. W. and Rubin, D. B. (1983) On Lord's Paradox. In *Principles of Modern Psychological Measurement: A Festschrift for Frederic M. Lord.* (H. Wainer and S. Messick, eds.), Hillsdale, NJ: Lawrence Erlbaum.

Kempthorne, O. (1952) *The Design and Analysis of Experiments.* New York: Wiley.

Lord, F. M. (1969) Statistical adjustments when comparing preexisting groups. *Psychol. Bull.*, **72**, 336–337.

Rosenbaum, P. R. (1984a) From association to causation in observational studies: The role of tests of strongly ignorable treatment assignment. *J. Amer. Statist. Ass.*, **79**, 41–48.

——— (1984b) Conditional permutation tests and the propensity score in observational studies. *J. Amer. Statist. Ass.*, **79**, 565–574.

Rosenbaum, P. R. and Rubin, D. B. (1983a) The central role of the propensity score in observational studies for causal effects. *Biometrika*, **70**, 41–55.

——— (1983b) Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *J. R. Statist. Soc.* B, **45**, 212–218.

——— (1984a) Estimating the effects caused by treatments: Discussion of a paper by Pratt and Schlaifer. *J. Amer. Statist. Ass.*, **79**, 26–28.

——— (1984b) Reducing bias in observational studies using subclassification on the propensity score. *J. Amer. Statist. Ass.*, **79**, No. 3 (in press).

——— (1985) Constructing a control group using multivariate matching methods that incorporate the propensity score. *American Statistician*, **39**(1), in press.

Rubin, D. B. (1974) Estimating causal effects of treatments in randomized and non-randomized studies. *J. Educ. Psychol.*, **66**, 688–701.

——— (1977) Assignment to treatment group on the basis of a covariate. *J. Educ. Statist.*, **2**, 1–26.

——— (1978) Bayesian inference for causal effects: The role of randomization. *Ann. Statist.*, **6**, 34–58.

Seber, G. A. F. (1977) *Linear Regression Analysis.* New York: Wiley.

Yates, F. (1960) *Sampling Methods for Censuses and Surveys.* London: Griffin.