DIW Diskussionspapiere
Discussion Papers

Discussion Paper No. 245

Estimating Causal Effects with Matching Methods in the Presence and Absence of Bias Cancellation

by
Thomas A. DiPrete[1)] and Henriette Engelhardt[2)]

[1)] Duke University and Research Professor at DIW Berlin
[2)] Max Planck Institute for Demographic Research, Rostock

Berlin, February 2001

Deutsches Institut für Wirtschaftsforschung, Berlin
Königin-Luise-Str. 5, 14195 Berlin
Phone: +49-30-89789- 0
Fax: +49-30-89789- 200
Internet: http://www.diw.de
ISSN 1433-0210

Estimating Causal Effects with Matching Methods in the Presence and Absence of Bias Cancellation

Thomas A. DiPrete
Duke University
German Institute for Economic Research (DIW Berlin)

Henriette Engelhardt
Max Planck Institute for Demographic Research,
Rostock, Germany

October 31, 2000

An earlier version of this paper to be presented at the August, 2000 annual meetings of the American Sociological Association. This research was supported in part by the Max Planck Institute for Human Development, by the Max Planck Institute for Demographic Research, and by Duke University. We would like to thank Norman Braun and Patricia A. McManus for helpful comments on an earlier version.

Abstract

This paper explores the implications of possible bias cancellation using Rubin-style matching methods with complete and incomplete data. After reviewing the naïve causal estimator and the approaches of Heckman and Rubin to the causal estimation problem, we show how missing data can complicate the estimation of average causal effects in different ways, depending upon the nature of the missing mechanism. While – contrary to published assertions in the literature – bias cancellation does not generally occur when the multivariate distribution of the errors is symmetric, bias cancellation has been observed to occur for the case where selection into training is the treatment variable, and earnings is the outcome variable. A substantive rationale for bias cancellation is offered, which conceptualizes bias cancellation as the result of a mixture process based on two distinct individual-level decision-making models. While the general properties are unknown, the existence of bias cancellation appears to reduce the average bias in both OLS and matching methods relative to the symmetric distribution case. Analysis of simulated data under a set of difference scenarios suggests that matching methods do better than OLS in reducing that portion of bias that comes purely from the error distribution (i.e., from "selection on unobservables"). This advantage is often found also for the incomplete data case. Matching appears to offer no advantage over OLS in reducing the impact of bias due purely to selection on unobservable variables when the error variables are generated by standard multivariate normal distributions, which lack the bias-cancellation property.

Estimating Causal Effects with Matching Methods in the Presence and Absence of Bias Cancellation

In recent years, sociology and the other social sciences have paid increased attention to the problem of causality and the estimation of causal effects. In the experimental setup, the difference in the average outcome for otherwise statistically identical treatment and control groups forms the basis for estimating the causal effect of the treatment in question. Social scientists have long recognized the problems in constructing a similarly valid estimator based on observational data. However, a developing literature stimulated largely by the research of James Heckman and Donald Rubin has produced both new approaches to the problem and a deeper appreciation of the potential and the limitations of standard regression analysis, matching methods, instrumental variable techniques, econometric models of selection bias, or the method of "difference-in-differences" when applied to observational data (e.g., Heckman 1979; Rosenbaum and Rubin 1984; Holland 1986; Sobel 1995; Rubin and Thomas 1996; Angrist, Imbens, and Rubin 1996; Smith 1997; Winship and Morgan 1999; Heckman et al. 1998).

This paper explores the implications for causal estimation that arise both when, in the words of Heckman and Robb (1985), the selection is based on unobservable variables, and when the structure of bias is such that pointwise biases are offsetting at the individual level. The matching methods developed by Rubin are designed to handle situations where selection into treatment arises from an observable mechanism, but where the analyst, who is generally ignorant about the true structural model, risks a specification

bias when attempting to estimate treatment effects using standard regression-based techniques. As Heckman has noted in several publications, Rubin's method cannot eliminate "point-by-point" bias (i.e., bias for the effect of treatment, conditional on the observed covariates) when the selection process involves unobservable variables. Note, however, that scientists are typically concerned more with the *average* effect of treatment, and with reducing the bias in estimates of this average, than they are with the point-by-point treatment estimates, or the biases in these point-by-point treatment estimates. Heckman et al. (1998) present new evidence that point-by-point biases are sometimes offsetting (i.e., positive at some points, and negative at other points). In such situations, the average bias (or equivalently, the bias in the estimate of the average treatment effect) might be relatively small because the point-by-point biases partially cancel each other.

This paper investigates this possibility in greater depth, and explores its implications for estimating treatment effects with complete data as well as with data where the outcome variable is missing for a fraction of the cases. Building on the empirical results of Heckman et al. (1998) that report evidence of "bias cancellation" in evaluation data for the effects of job training on earnings, we analyze the statistical basis for this phenomenon, and offer a substantive rationale for bias cancellation as the outgrowth of a mixture process, in which individuals make decisions based on one or another of two distinct and partially offsetting decision models. We then investigate the empirical properties of matching estimators based on Rubin's propensity score method in the situation where bias cancellation does and does not occur. We conduct these empirical investigations using partially simulated data from the German Socioeconomic

Panel, in which further training as an adult worker ("Weiterbildung") is the "treatment" variable, and earnings is the outcome variable.

Bias in the Estimation of Causal Effects: General Considerations

As shown by Winship and Morgan (1999), the relationship between the observed difference in the outcome and the average causal treatment effect $\bar{d}$ can be expressed as follows:

$$\bar{Y}^{t}_{i\in T} - \bar{Y}^{c}_{i\in C} = \bar{d} + (\bar{Y}^{c}_{i\in T} - \bar{Y}^{c}_{i\in C}) + (1-p)(\bar{d}_{i\in T} - \bar{d}_{i\in C}), \quad (1)$$

where *T* indicates that an individual is assigned to (or has self-assigned to) the "treatment group" (e.g., job training, summer school, the use of a particular form of birth control, etc.), and *C* indicates that an individual is in the not-treated comparison group, which is often called the "control" group in an experimental context. The "*t*" superscript indicates the outcome that would potentially occur if the individual is treated, while the "*c*" superscript indicates the outcome that would potentially occur when the individual is not treated (this notation distinguishes between real and counterfactual outcomes; all individuals are considered to have both potential outcomes, only one of which is realized). Finally, p is the proportion of the population (sample) in the treatment group.

The term on the left hand side of equation (1), which represents the difference in the mean outcome for individuals in the treatment group and individuals in the comparison group might be thought of as the "naïve" estimator of the treatment effect, $\bar{d}$. The modifier "average" is important. In the standard regression model (in contrast to more sophisticated models such as random effects models), one typically assumes that each effect is fixed. But the standard experimental setup makes no such assumption.

Instead, the difference in the average outcome for the treatment group and the control group is taken to be the effect of treatment, and this effect is an average effect for those cases which are randomly assigned to the treatment group. The possibility that the effect varies across individuals is recognized explicitly in equation (1).

The left side of equation (1) is not in general equal to the average treatment effect because of the presence of the 2nd and 3rd terms on the right hand side. The second term on the right hand side equals the difference in the average outcome for members of the treatment and the control groups if none of these individuals were treated. These differences, which have nothing to do with the treatment effect per se, will nonetheless affect the naïve estimator and, when nonzero, will cause the naïve estimator to deviate from the average treatment effect. The third term on the right hand side of equation (1) is non-zero to the extent that the average treatment effect in the treatment group differs from what this average would be in the control group if all members of the control group were treated.

This equation can be extended to reveal the additional biases in the estimation of treatment effects that are potentially created by missing data on the outcome variable, which make even the "naïve" estimator above impossible to calculate directly. Suppose that a fraction of the control group is missing (we designate this fraction as p_{MC}), and also that a fraction of the treatment group is missing (we designate this fraction as p_{MT}). Using this terminology, we can re-express the expression on the left side of equation (1) as follows:

$$
\begin{aligned}
\bar{Y}^{t}_{i\in T} - \bar{Y}^{c}_{i\in C} &= \frac{O^{t}_{T} + M^{t}_{T}}{n_{OT} + n_{MT}} - \frac{O^{c}_{C} + M^{c}_{C}}{n_{OC} + n_{MC}} \\
&= \frac{O^{t}_{T}}{n_{OT}} - \frac{O^{c}_{C}}{n_{OC}} - p_{MT}\left(\frac{O^{t}_{T}}{n_{OT}} - \frac{M^{t}_{T}}{n_{MT}}\right) + p_{MC}\left(\frac{O^{c}_{C}}{n_{OC}} - \frac{M^{c}_{C}}{n_{MC}}\right) \\
&= \bar{Y}^{t}_{i\in OT} - \bar{Y}^{c}_{i\in OC} - p_{MT}\left(\bar{Y}^{t}_{i\in OT} - \bar{Y}^{t}_{i\in MT}\right) + p_{MC}\left(\bar{Y}^{c}_{i\in OC} - \bar{Y}^{c}_{i\in MC}\right), \qquad (2) \\
&= \bar{Y}^{t}_{i\in OT} - \bar{Y}^{c}_{i\in OC} - \bar{\Delta}^{t}_{T} p_{MT} + \bar{\Delta}^{c}_{C} p_{MC}
\end{aligned}
$$

where

$$
O^{t}_{T} = \sum_{i\in T} Y^{t}_{i\in T} I_i, \quad M^{t}_{T} = \sum_{i\in T} Y^{t}_{i\in T}\left(1 - I_i\right),
$$

and where

$$
S_i = \begin{cases} 1 \text{ if observed} \\ 0 \text{ if missing.} \end{cases}
$$

O and *M* are defined similarly for other subscript/superscript combinations. The denominators in equation (2) (n_{OT}, n_{MT}, n_{OC}, and n_{MC}) represent the number of cases that are observed in the treatment group, missing in the treatment group, observed in the control group, and missing in the control group, respectively. Equation (2) shows that there are two additional potential sources of bias, having to do with the difference in the means for observed and missing treated cases, and for observed and missing control cases.

As Winship and Morgan point out, most research into the problem of causal estimation either assumes that the average treatment effect is the same in the treatment and the control group (i.e., $\bar{d}_{i\in T} = \bar{d}_{i\in C}$) or explicitly focuses attention on the average effect of the treatment for those in the treatment group (e.g., Heckman et al. 1998). In our investigation into the problem of bias cancellation below, we will also focus attention on this latter estimator.

Heckman et al. (1998) (see also Heckman, LaLonde and Smith [forthcoming] for a parallel discussion) have proposed the following structure for understanding causal effects and associated biases. Let

$$Y_{1i} = g_1(X_i) + U_{1i} \tag{3}$$

be the equation if one is treated, and

$$Y_{0i} = g_0(X_i) + U_{0i} \tag{4}$$

be the equation if one is not treated. Let $D = 1$ be the indicator of whether one is treated. Assume the estimator of interest is the effect of treatment for the treatment group. Then (after suppressing the index "*i*" for simplicity), the expected treatment effect for those who are treated, conditional on observed covariates, equals

$$E(d \mid X, D = 1) = g_1(X) - g_0(X) + E(U_1 - U_0 \mid X, D = 1). \tag{5}$$

Note that this estimator is different from what is sometimes called the "treatment effect" in an experimental context (see, e.g., the discussion in Smith [1997]). The "treatment effect" in an experimental context would be

$$E(d \mid X) = g_1(X) - g_0(X).$$

In contrast, the expectation, conditional on *D*, contains the additional third term involving the disturbance variables U_0 and U_1. In general, the left side of equation (5) cannot be directly estimated. Because one cannot observe $E(Y_0 \mid X, D = 1) = g_0(X) + E(U_0 \mid D = 1)$, one does not see the outcome for the treatment group in the case of no treatment. Therefore, one must estimate the treatment effect by contrasting the outcomes for those in the treatment group with the outcomes for the non-treated group. The naïve estimator, in other words, is

$$E(\tilde{d} \mid X, D=1) = g_1(X) - g_0(X) + E(U_1 \mid X, D=1) - E(U_0 \mid X, D=0). \quad (6)$$

The difference between the left side of equation (5) and equation (6) is

$$E(\tilde{d} \mid X, D=1) - E(d \mid X, D=1) = \{E(U_1 \mid X, D=1) - E(U_0 \mid X, D=0)\} - \{E(U_1 \mid X, D=1) - E(U_0 \mid X, D=1)\}.$$

The bias of the naïve treatment effect as an estimator for the effect of treatment for those assigned or self-assigned to treatment[1], conditional on observed covariates (this is also referred to as the "point-by-point" bias), becomes

$$B(X) = E(U_0 \mid X, D=1) - E(U_0 \mid X, D=0). \quad (7)$$

When this quantity is equal to zero we can say that the assignment mechanism is "ignorable" (cf. equation 3 in Heckman et al. 1998). Furthermore, as Rosenbaum and Rubin (1984) show, the bias under ignorable assignment is also zero if one conditions on the propensity score, i.e.,

$$B(P(X)) = E(U_0 \mid P(X), D=1) - E(U_0 \mid P(X), D=0) = 0, \quad (8)$$

where $P(X)$ is the probability that the case is assigned to (or selects) treatment, as a function of the covariates *X*. The case of ignorable assignment is equivalent to what Heckman and Robb (1985) refer to as "selection on observables."

In the case of "selection on unobservables," the bias is no longer zero at every point. One can still reduce the bias, however, if one has access to variables that predict assignment but that have no structural effect on the outcome. Again following Heckman et al. [1998]), assume that an index function

[1] See Heckman (1997) for further details, and for the corresponding bias when the naïve estimator is used to estimate the average treatment effect for the entire population.

$$I_1^* = H_1(Z_1) - n_1$$

determines participation in the treatment group where $D = 1$ when $I_1^* > 0$ and $D = 0$ otherwise, that Z_1 is observable (it may include *X*), and that (U_0, U_1) is potentially correlated with n_1. Under these assumptions,

$$\begin{aligned} E(U_0 \mid Z_1, X, D = 1) &= E(U_0 \mid n_1 < H_1(Z_1)), \\ E(U_0 \mid Z_1, X, D = 0) &= E(U_0 \mid n_1 \geq H_1(Z_1)). \end{aligned} \tag{9}$$

Heckman et al. (1998) note that this bias can be expressed under reasonable assumptions as

$$B(P_1(Z_1)) = E(U_0 \mid P_1(Z_1), D = 1) - E(U_0 \mid P_1(Z_1), D = 0), \tag{10}$$

where $P_1(Z_1)$ is the probability that $D = 1$, given Z_1. As long as U_1 and n_1 are correlated, the bias in (10) is not pointwise equal to zero in general, and matching methods will not generally provide unbiased estimates.

Theoretical and Empirical Foundations for the Presence of Bias Cancellation

While pointwise bias in equation (10) is not equal to zero, our primary goal is to recover accurate estimates of the *average* treatment effect, not the point-by-point treatment effect. Estimates of the average effect will typically involve some possibly weighted sum of the estimates of the point-by-point estimates (i.e., the treatment effect, conditional on observed covariates). If the bias itself varies point-by-point, then the extent of bias in the estimate of the *average* effect will depend both on the way the average is computed, and on the nature of the association between bias and covariates.

In particular, the average bias can be substantially reduced in situations where the *sign* of the bias depends upon covariates. In this situation, bias cancellation will occur when an average effect is computed. As an example, suppose that the true effect of summer school was an increase of 10 points on some benchmark test, and that this effect was constant in the population. Suppose that the naïve estimate over-predicted the effect by 5 points in any subgroup of students. Therefore, the overall naïve prediction of the treatment effect in the population as a whole would be too high by five points, i.e., it would appear as if the average student gained 15 points, when in fact only a 10 point gain could be attributed to summer school. Suppose instead, however, that the size of the bias varied by SES, perhaps being 8 points among low SES students and 2 points among high SES students. In this scenario, the overall average bias would lie somewhere between these values. But suppose further that the bias was not always the same sign, for example, suppose that the bias was –5 points among low SES students (i.e., it appeared that the low SES students gained only 5 points from summer school instead of the "true" 10 point average gain), while the bias was +5 points for the high SES students (it appeared that the high SES students gained 15 points on average, when the true average gain was 10 points). In this example, bias cancellation occurs in the computation of an overall average treatment effect. The overall average estimated effect would be closer to the true value of 10 points than would be the estimated effects for the two subgroups. Table 1, which is described below, gives an empirically determined illustration of bias cancellation in the context of job training.

Heckman et al. (1998) and Heckman, LaLonde, and Smith (forthcoming) have argued that bias reduction is especially large in the case of symmetric distributions, such

as the standard multivariate normal distribution. They argued that the average bias equals zero over intervals around $P = ½$ under the conditions that the latent variables n_1 and U_0 are symmetrically distributed around zero. Under these circumstances, they argued that the absolute magnitude of $B(P(Z))$ is symmetrically distributed around $P = ½$, while the sign reverses around $P = ½$. Consequently, if P itself is symmetrically distributed around ½, then the average bias would equal zero in these symmetric intervals. They further argued that even if P was not symmetrically distributed, the average bias would still be zero within symmetric intervals under an appropriate matching process .

This claim turns out to be incorrect. As we show in Appendix A, the sign of $B(P(Z))$ always remains the same in a symmetric distribution. Contrary to Heckman et al.'s assertion, therefore, the bias does not cancel in symmetric intervals around $P = ½$ in the model that is most commonly assumed by analysts attempting to correct for selection bias, namely the bivariate normal case.[2]

Nonetheless, Heckman et al.'s paper is very revealing about the empirical variation in bias as a function of observed covariates. The Department of Labor had previously collected experimental data in order to evaluate four training centers participating in the Job Training Partnership Act. In this experiment, a random group of individuals who volunteered to undergo job training were refused participation, and data on these individuals (including their subsequent earnings) were used to estimate the average effect of training on wage change. In addition, data was also collected on a non-experimental comparison group of individuals who were not trained. Heckman et al.

[2] We discovered this error in the process of writing this paper. Heckman has agreed in email correspondence with us that his published claim is in error.

combined the data on the non-experimental comparison group with the data on trainees so that they could estimate a training effect using various selection-bias correcting methods with non-experimental data. Their goal was to compare the performance of these estimators against the presumably unbiased estimate of the average treatment effect from the experimental data. These data also allowed Heckman et al. to estimate the relationship between the bias in the estimated training effect using non-experimental data and covariates. Table 1, which is reproduced from table 7 in Heckman et al. (1998), shows that as an empirical matter, the bias in the estimated treatment effect varied with the estimated probability that an individual would choose to be treated. Table 1 also shows that the bias changed sign in the training evaluation data, being *positive* at relatively high probabilities of choosing treatment (i.e., the naïve predictor gave too large an estimate of the treatment effect for those with high probabilities of being treated), and being *negative* at low probabilities. Interestingly, the sign reversal took place around the median probability.[3] This empirical evidence of bias cancellation recurs throughout the Heckman et al. (1998) paper and is arguably one of the central empirical findings of their paper.[4]

[3] In Heckman et al.'s data, the overall proportion of sample members who chose to be trained was much lower than 0.5, and the point at which the bias changed direction in Heckman's data, at about $P = 0.04$, was far below the 0.5 probability level that figured in Heckman et al's theoretical justification for the bias-cancellation phenomenon.

[4] To quote from Heckman et al. (1998) "Our evidence of substantial pointwise bias that averages out to small bias over certain intervals is reminiscent of what can occur in the classical section bias model, as noted in the discussion surrounding Figure 1 [as we noted in the man text, Heckman et al.'s discussion about the classical selection model is incorrect]. Moreover, it is inconsistent with the identifying assumption used to justify matching. This empirical regularity occurs in the other models estimated below and is a central empirical finding of this paper." (Heckman et al. 1998, p. 1044). Heckman et al.'s assertion that bias cancellation is inconsistent with the identifying assumptions underlying matching is certainly true in that these assumptions assume that the *pointwise* bias would be zero. The focus in our paper concerns the average bias rather than the pointwise bias.

The empirical pattern found in the Heckman et al. data suggests that the distribution of disturbance variables in the JTPA training data was far from the bivariate normal distribution typically assumed in sample selection models. We speculate that this observed pattern occurred because the choice to be trained may have been generated by a *mixture* of assignment mechanisms, rather than the single assignment mechanism that is commonly assumed to operate in nonexperimental data. If the mixture process had the property that the correlation structure for those with low probabilities of receiving training was different than from the correlation structure for those with high probabilities of receiving training, the result would be similar to what Heckman et al. observed in their data.

To motivate the theoretical plausibility of such a mixing distribution, it is useful to consider the substantive meaning of the correlation between U_0 (the disturbance in the not-treated structural equation) and n_1 (the error in the assignment equation) that gives rise to the bias (see equation (10)). For simplicity, let us assume that individuals know their U_0, in other words, that they can perfectly predict outcomes in the situation where they chose not to be treated (the assumption of perfect knowledge is unnecessary – we could instead assume that their estimate of U_0 is highly correlated with the true U_0). We make the further assumption that individuals make their decisions to participate (i.e., their choice of n_1) in response to their knowledge of U_0. We then consider the following two decision-making scenarios as plausibly operating at the same time. In statistical terms, these scenarios differ in the assumed correlation between U_0 and n_1. In behavioral terms, these scenarios differ in the nature of the decision-making process.

Scenario A: Consider the case where U_0 and n_1 are positively correlated. From the definition of the selection function, the probability of selection into the treatment *increases* as n_1 *decreases*. A positive correlation means that (after controlling for measured variables) the higher the earnings would be in the absence of treatment, the lower is the probability of treatment. A negative correlation means that (after controlling for measured variables) the higher the earnings would be in the absence of treatment, the higher is the probability for selection into treatment. Recalling that the gain from treatment equals the structural effect of treatment plus the difference between the disturbance in the treatment and the non-treatment equations (equation (5)), we also need to acknowledge the possible impact of the correlation between U_0 and U_1 on our interpretation. Let us for simplicity assume that the structural effect of treatment is fixed in the population. If U_0 equals U_1, then everyone would get exactly the same benefit from treatment. In the more plausible scenario, however, U_0 and U_1 are positively correlated, but not perfectly so. In this latter scenario, individuals with high U_0 tend to benefit less from the treatment than do individuals with low U_0. Thus, a positive correlation between U_0 and n_1 means that (after controlling for measured variables) individuals who would experience lower than average gains from treatment are less likely to select themselves into treatment than are individuals who would experience higher than average gains from treatment.

As equation (7) shows, a positive correlation between U_0 and n_1 corresponds to a negative bias in the naïve estimate of the treatment effect. Those in the treatment group would have an average value of U_0 that is less than zero, while those in the control group

would have an average value of U_0 that is greater than zero. The naïve estimator of the treatment effect, which essentially differences the observed outcomes for the group selected into treatment from the observed outcomes for the group selected into not-treatment, would not take into account the selection effect, and thus would arrive at a downwardly biased estimate of the treatment effect.

Scenario B: Suppose, that the correlation between U_0 and n_1 is *negative*. In this scenario (after controlling for measured variables), the higher the earnings would be in the absence of treatment, the higher is the probability for selection into the treatment. In this situation, the group of individuals who select themselves into treatment actually would experience a smaller average impact of the treatment than would the group of individuals who do not select themselves into treatment. This behavior might seem to be irrational, but it could be interpreted as consistent with what Kahneman and Tversky (1979) referred to as "prospect theory." Prospect theory argues that the negative effects of loses on a utility function are greater than the positive effects of equivalent-size gains, and that most people therefore exhibit what they referred to as "loss aversion." If the correlation between U_0 and U_1 is positive (and this seems very likely), then those individuals who would have higher than predicted earnings if they were not treated would also have higher than predicted earnings if they were treated. If individuals with higher than predicted earnings developed a strong interest in "preserving their gains," then they would be especially likely to choose to be treated. For these individuals, a naïve estimator would overstate the "true" effect of treatment.

These scenarios can be further motivated via the following real-world evidence that one of us has observed in teaching a particular class. The first quiz in this class was

relatively difficult, and the grading distribution included C's and D's as well as A's and B's (standard statistical theory predicts that students with the low grades have an average U that is negative, while the students with A's have an average U that is positive). To offset the impact of this quiz, the instructor later (after the students knew their grades on the first quiz) offered students an optional "extra credit" essay assignment that was worth up to 10% of the total first quiz grade, and that also could be applied to a future quiz in the course. The "naïve" prediction was that students whose grades were C and lower would be most likely to turn in the optional essay. While many of these low-performing students did in fact use this opportunity to improve their grade, the instructor found that the students who received A's on the first quiz were most likely to do the extra-credit assignment, even though their predicted returns to this work were low. One interpretation of the student's behavior is that the high-performing students had anticipated the possibility of getting an A in the course on the basis of their performance on the first quiz. At the same time, they were aware of the fact that they couldn't be assured of A's on future quizzes (in effect, the students who received A's realized that their U was on average positive). These students were therefore especially interested in "preserving" their anticipated outcome by taking advantage of the insurance offered by the extra credit assignment.

The two forms of behavior described above in scenario A and scenario B produce offsetting biases. The pattern observed in Table 1 would result if the individuals who behaved "rationally" tended to be clustered at the lower values of the linear predictor for treatment, while those who behaved according to the "prospect theory" model were those clustered at the higher values of the linear predictor for treatment. If the distribution of

the disturbance variables was perfectly "mirror symmetrical" around some value of $H_1(Z_1)$ (in the sense that the correlations between U_0 and n_1 were equal and of opposite sign at values of $H_1(Z_1)$ that were equidistant from some specific reference point), then the complete bias cancellation written about by Heckman et al. (1998) would occur, and the average bias would be zero if the data were perfectly balanced around this point.[5] If the offsetting biases did not perfectly balance, then the average bias would not be eliminated, but it would be smaller than in the case where all point-by-point biases had the same sign. The extent of bias cancellation, or whether the bias even changes sign as a function of $H_1(Z_1)$, are empirical questions just as is the question of whether an error distribution is multivariate normal. The data presented by Heckman et al. (1998) however, suggests that in at least one important context bias cancellation does occur, while our theoretical argument above suggests that this case might not be an isolated instance.

Bias in the Estimation of Treatment Effects in the Presence of Missing Data

The above discussion assumes that that there is no missing data on the dependent variable (i.e., the only "problem" is the selection mechanism in the assignment process). In the presence of missing data on the dependent variable, the above setup becomes more complicated. If we let $S = 1$ designate that Y is observed, and $S = 0$ designate that Y is not observed because of missing data, then the fact of missing data changes the bias to the extent that

$$B(X \mid S=1) = E\left(U_0 \mid X, D=1, S=1\right) - E\left(U_0 \mid X, D=0, S=1\right) \neq B(X \mid S=0). \quad (11)$$

[5] Mathematical justification of this assertion is available upon request.

To understand the complications arising from bias of the form shown in equation (11), we assume (in parallel with the above discussion) that there is a second index function

$$I_2^* = H_2(Z_2) - n_2,$$

which determines whether *Y* is observed or missing, in addition to the first index function that determines whether a case is or is not assigned to the treatment. With respect to this second index function, $S = 1$ if *Y* is observed and $S = 0$ if *Y* is missing, where Z_2 is observable (it may overlap with *X*, and it may overlap with or be identical with Z_1), and where $(U_0, U_1), n_1$ and n_2 are potentially correlated with each other. Under these assumptions, the bias is expressible by elaborating equation (7):

$$B(X \mid S = 1) = E(U_0 \mid X, D = 1, S = 1) - E(U_0 \mid X, D = 0, S = 1). \quad (12)$$

This expression also gives the condition when missing data are "ignorable" for the causal effects problem. Missing data are ignorable when

$$B(X) = B(X \mid S = 1), \quad (13)$$

or, in other words, when

$$E(U_0 \mid X, D = 1, S = 1) - E(U_0 \mid X, D = 0, S = 1) = E(U_0 \mid X, D = 1) - E(U_0 \mid X, D = 0).$$

Under this condition, the results obtained above under the assumption of complete data are not altered when the expressions are conditioned upon observable data. In particular, the bias remains zero under ignorable assignment, or equivalently, selection on observables (see equation (8)).

Alternatively, the missing data are not "ignorable" when

$$B(P(X)) \neq B(P(X) \mid S = 1).$$

To better understand this bias, we elaborate the conditions in (9) to obtain

$$\begin{aligned}
E(U_0 \mid Z, X, D = 1, S = 1) &= E(U_0 \mid n_1 < H_1(Z_1), n_2 < H_2(Z_2)), \\
E(U_0 \mid Z, X, D = 1, S = 0) &= E(U_0 \mid n_1 < H_1(Z_1), n_2 \geq H_2(Z_2)), \\
E(U_0 \mid Z, X, D = 0, S = 1) &= E(U_0 \mid n_1 \geq H_1(Z_1), n_2 < H_2(Z_2)), \\
E(U_0 \mid Z, X, D = 0, S = 0) &= E(U_0 \mid n_1 \geq H_1(Z_1), n_2 \geq H_2(Z_2)).
\end{aligned}$$

If we set $H_1(Z_1) = F_{n_1}^{-1}(P(D = 1 \mid Z_1))$, and $H_2(Z_2) = F_{n_2}^{-1}(P(S = 1 \mid Z_2))$, where F is assumed to be strictly monotonic almost everywhere, the bias becomes

$$\begin{aligned}
B(P_1(Z_1), P_2(Z_2)) = E(U_0 \mid P_1(Z_1), P_2(Z_2), D = 1, S = 1) \\
- E(U_0 \mid P_1(Z_1), P_2(Z_2), D = 0, S = 1)
\end{aligned}. \qquad (14)$$

The question of interest again concerns the extent to which controlling for Z_1 (or the propensity score for assignment to treatment) as well as Z_2 (or the propensity score for missing) yields a condition under which the difference in the expectations becomes zero.

If U_0 is independent of v_2 after controlling for n_1, then estimators which are consistent under the assumptions of complete data remain consistent in the presence of missing data, because (controlling for *X*) the probability of missing is unrelated to the value of *Y*.[6] But if U_0 covaries with both v_2 and n_1, then the situation with incomplete data differs from that with complete data. In particular, if the conditional bias reverses sign as a function of the propensity score, the properties of bias cancellation will differ for the complete and incomplete data case. However, it is nonetheless possible that substantial

[6] In particular, if the probability of missing is a function of Z_2, and if Z_2 contains elements that are not in *X* and also are not in Z_1, there is no practical use to be made of these elements in the estimator of the treatment effect, because they provide no information about the values that *Y* take if it were not missing.

cancellation would occur in the missing data case just as in the complete data case, and matching estimates might therefore reduce the *average* bias relative to OLS.

The Role of U_1 in the Estimation of Treatment Effects

The above discussion concerned the possibility of bias that arises out of possible correlation between U_0 and n_1 in the case of complete data and from possible correlation between U_0, n_1, and n_2 in the case of incomplete data. It is also important to consider the implications for bias that arise from possible correlation between U_1 and (U_0, n_1, n_2). For the complete data case, U_1 technically plays no role in the structure of the point-by-point bias for the estimate of the treatment effect for those assigned or self-assigned to treatment. This can be seen in equation (7) or in equation (10).[8] Nonetheless, U_1 can play an important indirect role in the estimation process, because, as we suggested in our proposed justification for bias cancellation, the extent and nature of bias cancellation can depend upon the relationship between U_0 and U_1. Consequently, even if $B(X)$ is unaffected by U_1, the *average* of $B(X)$ across the range of *X* might be influenced by U_1 and by its correlation with the other disturbance variables. Furthermore, missing data bias is directly affected by the existence of correlation between U_1 and n_2. As for the case of complete data, however, the dominant impact of U_1 may work indirectly through its effect on bias cancellation rather than through its correlation with n_2.

---

[8] As Heckman (1997) shows, the bias does depends on U_1 as well as U_0 when the quantity to be estimated is the effect of treatment for a randomly selected sample from the population. This dependency on U_1 disappears however when the quantity to be estimated is the effect of treatment on the sample selected into treatment.

Data, and Methods

We illustrate the potential impact of matching estimates in the presence and absence of bias cancellation using panel data on additional job training. The use of real data for studying the extent of bias in alternative estimators has disadvantages that arise from our inability to test the underlying exclusion restrictions concerning Z_1 and Z_2, and from our inability to directly observe the shape of the (U_0, U_1, ν_1, ν_2) distribution. To overcome this limitation, we use simulated data in which the observed treatment variable, the observed dependent variable, and the observed pattern of missing data on the dependent variable are replaced by measures simulated from a predetermined model. The advantage of simulated data is that we know the model that generates the data, and we can use this knowledge to evaluate our estimation methods. We used actual survey data to the extent possible in order to create realistic examples; the treatment, outcome, and disturbance variables are simulated, but the control variables have the actual values found in the survey. We perform simulations with a model where the selection for the treatment effect is moderately strong, and where the pattern of missing data produces moderate bias in estimators that naively assume ignorable missing.

The starting point for our simulations was a subset of data drawn from the German Socioeconomic Panel (GSOEP). Starting in 1984, individuals aged 16 and above in nearly 6000 households have been interviewed on a yearly basis. In addition to the core questions on demographics and household composition, employment information, education, various types of income etc., modules on special topics are often

incorporated into the yearly interview.[9] In this paper we use the 1989 and 1993 special modules on any continuous training undertaken by sample members in the previous three years. We combined demographic, employment, income, and wage information from the 1985 to 1995 waves for respondents between the ages of 18 and 60 who were living in West Germany in 1989 or 1993 with the information on continuous training in order to obtain the data that formed the basis for our simulations. Our working dataset consisted of 9259 observations that were nonmissing on our predictor and dependent variables. In the sample for whom we had complete information, 12.5% reported further training.

The first step in the simulation involved the stochastic disturbances. We simulate data for several important cases, and show the resulting correlation matrices in Table 2. Recall from the above discussion that there four stochastic disturbances of interest in the general problem. The first two (U_0 and U_1) are the errors in the structural model for those who are not treated, and for those who are treated, respectively. The third error, n_1 affects selection into the treatment group. The fourth error, n_2, affects whether the outcome variable is missing or not. Our goal was to generate data in which the disturbance variables were correlated according to one of two distinct patterns. The first pattern is that produced by a standard symmetric distribution (we use the multivariate normal distribution). The second was to reproduce the pattern of bias cancellation that Heckman et al. (1998) found to exist in the JTPA evaluation data. While we believe that the substantive force producing bias cancellation is a mixture process corresponding to the two scenarios described above, we use the more expedient simulation strategy of

[9] For a detailed description of the GSOEP see Haisken-DeNew and Frick (1998).

assuming that the correlations between U_0 and n_1, and between U_0 and n_2, reversed beyond a specific point in the distribution, as described below.

After generating four disturbance variables for each sample member, our second step was to replace the observed treatment variable with a simulated treatment variable. We estimated a probit equation for further training as a function of covariates (the results of this probit estimation can be found in appendix Table A1). We used this probit equation to simulate a further training variable, where our simulated variable was constructed by combining the prediction from the probit model with the simulated stochastic disturbance n_1. This simulated treatment variable by construction has the same binary distribution as the original variable. We then produced a second simulated further training variable by adding unity to all the linear predictors from the probit equation in order to get a variable whose probability distribution was closer to being symmetrical around 0.5. With this modification, the proportion of the sample with further training on this second simulated variable rose to 41.6%. By combining our simulated training and disturbance variables in different ways, we obtained the following six cases:

Case 1: U_0, n_1, and n_2 follow a multivariate normal distribution, but we specify independence between U_1 and the other three variables.[10] Because this is a symmetric distribution, there is no bias cancellation. In this example, we specified U_0 and n_1 to have a positive correlation. This corresponds to Scenario A above. In addition, we specified

[10] The observed correlation between U_1 and the other variables is not zero in Table 3, but this occurs because of sampling fluctuations. In the model that generated these data, the correlation between U_1 and the other variables was set to zero.

v_1 and n_2 to be negatively correlated with each other. (This latter specification has no effect, of course, in our "complete data" examples below, because for these examples, the value of n_2 has no impact on the data structure). Recall that in the index models specified above, the binary outcome was *more* likely when the stochastic disturbance was smaller. For the case of incomplete data, therefore, a negative correlation implies that individuals who were more likely to choose training were more likely to have missing data on the outcome variable. The assumption that U_0 and U_1 are uncorrelated would be completely unreasonable if there was some unmeasured person-specific fixed effect on the outcome variable. In our examples, however, we use the *change* in the gross monthly wage as the dependent variable, and therefore the effect of unmeasured permanent individual characteristics on wage levels are differenced out (in this respect, our models correspond to what are sometimes called conditional difference-in-difference estimators in econometrics [Heckman et al. 1998]). Any remaining correlation between U_0 and U_1 would arise from unmeasured factors that vary over time.

Case 2a: U_0, n_1, and n_2 follow a multivariate normal distribution, where the correlation structure is the same as in case 2 for the 75% of cases with the lowest propensity scores, but then reverses, so that U_0 and n_1 have a negative correlation above the 75% point of the propensity score distribution. This corresponds to Scenario A above for the lower ¾ of the distribution, and to Scenario B for the upper ¼ of the distribution. The values presented in the table were computed from simulated data where the proportion trained in the simulated data equals the proportion trained in the actual GSOEP subsample.

Case 2b: U_0, n_1, and n_2 follow a "mirror symmetric" multivariate normal distribution, where the correlation structure is the same as in case 2 for $P(\text{train}) \leq ½$, but reverses for higher probabilities. Case 3b pertains to the simulated data where the proportion trained is enhanced to about 40% of the sample (as opposed to the 12.5 % who were trained in the actual subsample).

Case 3: U_0, U_1, n_1, and n_2 are symmetrically distributed. There is no bias cancellation in this case.

Case 4a: U_0, n_1, and n_2 follow a "mirror symmetric" multivariate normal distribution around $P(\text{train}) = ½$, and in addition, n_1 and n_2 are correlated with U_1. Proportion trained equals the observed proportion.

Case 4b: U_0, n_1, and n_2 follow a "mirror symmetric" multivariate normal distribution around $P(\text{train}) = ½$, and in addition, n_1 and n_2 are correlated with U_1. Proportion trained is enhanced as discussed in case 2b.

The third step was to simulate the outcome variable. We estimated a wage equation using OLS, where the two-year difference in the natural logarithm of the wage was the dependent variable, and the right hand side included a set of standard covariates (see appendix Table A1), specifically including further training as a predictor. We then replaced the product of observed further training multiplied by its estimated effect with our simulated training variable multiplied by an effect that we specified (we enhanced the effect of training to equal to three times the estimated effect in the OLS equation). We added the simulated stochastic disturbance to the estimated log(wage) in order to obtain a simulated outcome variable.

In the fourth step, we estimated a probit equation in which the outcome variables was an indicator variable equal to one if that the wage was observed in our data, and equal to zero otherwise. We then used the results of this probit equation in combination with our simulated value for n_2 to select a subset of the complete observations where we *a priori* set the outcome variable to missing in our simulated incomplete data. Our procedure generated missing outcome data for 37% of the cases.

In our examples, we assume that the "true" structural effect of training is fixed in the population. The "true" total effect of training varies across the sample, however, because the total effect includes an error component in addition to the structural effect, and this error component is not fixed in the population (see equation (5)).

Note that in Heckman et al.'s data, the probability value at which the sign of the bias reverses is far lower than 0.5; indeed, it appears to be roughly in the center of the data. This is a highly fortuitous event, because it implies that substantial cancellation would occur by averaging over the entire sample. In our simulations, we instead specified that the sign reversal occur at the point where the probability of receiving the treatment is 0.5 for the enhanced data, where 41.5% of sample members were trained. For the simulated data where the proportion trained equals the observed proportion, we specified that the sign reversal occur at the 75^{th} percentile of the propensity scores.

Because we know the underlying structure of the simulated data, we can compute the average bias for the entire sample or for any subsample. This allows us to compare the pattern of bias in our data with the pattern that Heckman et al. observed in the training data. Table 3 shows the relationship between the bias and the propensity score in our simulated data. As can be seen, the simulated data based on symmetric error distributions

looks quite different from the empirical pattern observed in the Heckman et al. data. In contrast, the patterns in the data where the correlation reverses at a specific value of the propensity score (i.e., a specific value of P(train)) are qualitatively similar to the empirical pattern found in the Heckman et al. data.

We analyzed these simulated data using two methods. The first was OLS. The second involved matching data between those in the treatment group with respondents who were not treated. Our matching procedure is a variant of the propensity/Mahalanobis metric matching method proposed by Rosenbaum and Rubin (1985) and Rubin (1991) and applied by Lechner (1999). The steps of the matching procedure are described in appendix B.

For the training-enhanced data in which bias cancellation was (induced to be) present, we also included estimates based on a two-step matching procedure. In the first step, we focused on the group of respondents who had been treated, and matched cases whose probability of training was greater than 0.5 with cases whose propensity score was as close in magnitude as possible but opposite in sign, in order to increase the extent of bias cancellation. The resulting matching yielded a distribution of treated cases that was approximately symmetrically distributed around the probability 0.5. We then matched each of this subset of treated cases with its most closely matching counterpart in the control sample.

For the incomplete data, we followed a slightly different procedure. A matching based solely on propensity scores would be unsatisfactory, because it would often be true that one or the other cases in the match would have missing data on the outcome variable, and thus many complete cases would be lost because they were matched to incomplete

cases. In order to increase the usable sample size in our matching procedure, we first dropped all cases that were missing on the dependent variable. Then, we matched as before.

Results

Tables 4 and 5 contain the results of our analyses of the simulated data described above. Table 4 contains results for the situation where U_1 is presumed to be independent of the other three error variables (cases 1, 2a, and 2b in Table 2). Table 5 contains results for the cases where all four error variables are correlated (cases 3, 4a, and 4b in Table 2). The rows in these tables contain the following information:

Row 1: The "true" experimental effect of training. This effect is equal to the average ß for all members of the sample. In our examples, here, we artificially set ß to be three times as great as the OLS estimate for the training effect in the GSOEP subsample with the simple specification that we report in Table A1. We set ß to the same value for all observations in the simulated data.

Row 2: The "true" average effect for the self-selected sample. This value, which is intrinsically unobservable in real data, equals the average of equation (5) for sample members.

Row 3: The "naïve" estimate of the causal effect, which is shown in equation (6), where we substitute sample averages for expectations.

Row 4: The OLS estimate of the treatment effect, using complete data. Because the primary focus of this paper concerns the character of the unobservable variables, we avoid getting into the issue of specification bias, and instead use the same specification in

our OLS estimation that was used to generate the outcome variable. The only remaining source of bias possible in our calculations, therefore, comes from the failure of the OLS assumptions about the error variable to be correct.

Row 5: The OLS estimate of the treatment effect, using incomplete data.

Row 6: The estimate from matching, using complete data.

Row 7: The estimate from matching using incomplete data, where the cases with missing data on the outcome variable are first dropped before matching is done.

Row 8: The estimate for the complete data from the two-step matching procedure described above.

Row 9: The estimate for incomplete data from the two-step matching procedure described above.

Row 10: The fraction of the bias from OLS that is eliminated in the matching estimate (based on row 8 if relevant, otherwise row 5).

Row 11: The fraction of the bias from OLS that is eliminated in the matching estimate (based on row 9 if relevant, otherwise row 6).

Tables 4 and 5 demonstrate an assertion made by Heckman and Robb (1985) that is still often under-appreciated in the social sciences, namely that the "true" experimental effect differs from the "true" treatment effect for individuals who are selected (or self-selected) through a non-random assignment process. The effect in row 2 is "true" in the sense that it is the average difference between the outcome that *was* experienced by the treated population, and the outcome they *would have* experienced if they had not been treated (one can only have "access" to this information in a simulated world; in the "real"

world, such information is intrinsically missing). This outcome is the sum of the "true" experimental effect of the treatment, and the difference in the error in the "treated" and the "not-treated" equation.

Table 4 also shows that – under the assumptions of cases 1, 2a, and 2b – the true average effect for the self-selected sample is greater than the experimental effect. This relationship arises from the imposed positive correlation between U_0 and n_1 (which implies that individuals who anticipate negative shocks if they proceed along their current path are more likely to choose the alternative path of treatment) coupled with the imposed independence between U_0 and U_1. Furthermore, at least for our examples, the increment in the total effect that comes from the difference between U_1 and U_0 is much greater in the symmetric case than in the bias cancellation case.

In our examples, the OLS estimate (row 4) was (not surprisingly) quite close to the "naïve" estimate that arises from substituting the control group's outcome in the absence of treatment for the outcome that the treatment group would have experienced if they were not treated. In our example, the OLS bias slightly increased in size when the incomplete data were generated by a bias-canceling process, while the OLS bias was more significantly enhanced for data generated from a symmetric distribution. We note again that the OLS specification was "perfect" in that it exactly matched the specification that was used to generate the data

Rows 5 and 6 show the results we obtained using the matching procedure based on the probit propensity score for selection into further training. When the data were generated by a symmetric distribution, the matching method slightly under-performed the

OLS estimates. When the data were generated by a bias-canceling distribution, however, matching outperformed OLS. The advantage for matching was greater for the analysis of the incomplete data than for the analysis of the complete data. Finally, the two-step matching (rows 8 and 9) led to a still greater reduction in bias, relative to the OLS estimator.

Table 5 shows results for cases 3, 4a, and 4b from Table 2, in which (unlike for Table 4) we imposed a positive correlation between U_0 and U_1. This change has two major consequences. The first major consequence is to reduce the size of the "true" average effect of selecting training for the sample that was non-randomly selected into training (row 2). The second major consequence is to increase the downward bias of the OLS estimate and of the corresponding matching estimates in the symmetric distribution cases (row 4, columns 1 and 3, and rows 6 and 7). As in Table 4, the matching estimate was inferior to the OLS estimate in the symmetric distribution case (again, however, recall that the OLS equation has no misspecification bias – it exactly matches the equation used to simulate the dependent variable). Also, as in Table 4, the matching estimators outperformed the OLS estimator in both of the complete data cases. For the enhanced training data, the matching estimator also outperformed the OLS estimator in the incomplete data case. However, in the simulated data where the proportion trained equaled the observed fraction trained, the matching estimator was inferior to the OLS estimator, which in this case was actually rather close to the "true" value.

Discussion and Conclusions

Taking as a starting point Heckman et al.'s (1998) discovery that bias cancellation sometimes occurs in the data-generating process for treatment via non-random selection, we have made three contributions to the literature on causal estimation. First, we have shown that Heckman et al.'s theoretical justification for bias cancellation (the assertion that bias cancellation arises naturally from symmetrically distributed stochastic disturbances in the structural and the assignment equations) is incorrect; in fact, bias cancellation never occurs when the errors are symmetrically distributed. Second, we have suggested an alternative mechanism by which a bias-canceling data generating process might occur. Third, using the GSOEP as a starting point, we have generated and analyzed simulated data on further training and wage changes, and we have shown how bias emerges from the set of correlations between the errors in the structural equations, the error in the assignment to treatment equation, and the error in the missing data equation. Using correlation structures that appear to correspond reasonably well with the observed bias pattern that Heckman et al. found for evaluation data of the JTPA, we find that bias cancellation in the error distribution reduces the bias even for the OLS estimator. Generally speaking (thought not in all instances), a simple matching estimator does an even better job than OLS, even when the OLS estimate derives from an equation that was perfectly specified. We further found that a two-step balancing procedure (first on the cases treated and then a matching between treated and control cases) produces superior estimates when the axis for matching in the first step roughly matches the point at which bias reversal occurs in the data.

Because bias cancellation is a property of the distribution of an unobservable variable, it is difficult to know whether Heckman et al. discovered a common pattern or

an unusual one. It is also difficult to know whether their finding that bias reversal took place near the midpoint of distribution is common or unusual. In our opinion, the answers to these questions require first an understanding of *why* bias cancellation occurs. We have provided an initial plausible answer in this paper. It is our hope that further theoretical refinement, and further empirical analysis of cases where experimental and non-experimental data can be directly compared, are the best hope for greater understanding of this phenomenon, and for the development of more accurate estimators for causal effects in the social sciences.

One final observation should be made. The standard Heckman two-step correction for sample selection bias typically assumes that the structural and assignment errors follow a bivariate normal distribution. This distribution is very different from a distribution that would produce bias cancellation. If bias cancellation is common, then the applicability of the Heckman two-step correction becomes more questionable. Again, further research is needed to establish the robustness of various estimators of causal effects across the range of plausible distributions in order to better understand the potential implications of bias cancellation for the estimation of causal effects using non-experimental data.

Appendix A: Bias Cancellation and Symmetric Distributions

Assume that the distribution of n_1 and U_0 is symmetrical. Assume further that for a group of individuals indexed by "*i*," $H_1(Z_{1i}) = a$, while for a second group of individuals indexed by i', $H_1(Z_{1i'}) = -a$. Then for group *i*, equation (10) can be re-expressed as

$$B(P_1(Z_1) \mid H_1(Z_1) = a) = E(U_0 \mid n_1 < a) - E(U_0 \mid n_1 \geq a), \tag{15}$$

while for group i',

$$B(P_1(Z_1) \mid H(Z_1) = -a) = E(U_0 \mid n_1 < -a) - E(U_0 \mid n_1 \geq -a).$$

But in a bivariate symmetric distribution,

$$E(U_0 \mid v_1 < a) = -E(U_0 \mid v_1 \geq -a).$$

Therefore,

$$\begin{aligned} B(P_1(Z_1) \mid H_1(Z) = a) + B(P_1(Z_1) \mid H(Z_1) = -a) &= E(U_0 \mid n_1 < a) - E(U_0 \mid n_1 \geq a) \\ &\quad + E(U_0 \mid n_1 < -a) - E(U_0 \mid n_1 \geq -a) \\ &= 2 \cdot E(U_0 \mid v_1 < a) - 2 \cdot E(U_0 \mid n_1 \geq a) \\ &= 2 \cdot B(P_1(Z_1) \mid H_1(Z_1) = a) \\ &\neq 0. \end{aligned}$$

To see this another way, note (e.g., Heckman et al. 1998, footnote 25), that

$$\begin{aligned} E(U_0 \mid P(X)) &= E(U_0 \mid P(X), D = 1)P(D = 1) + E(U_0 \mid P(X), D = 0)P(D = 0) \\ &= E(U_0 \mid P(X), D = 1)P(X) + E(U_0 \mid P(X), D = 0)(1 - P(X)), \end{aligned}$$

where $D = 1$ when an individual is treated, and $D = 0$ when an individual is not treated. $P(X)$, the propensity score, gives the probability that an individual is treated. If we assume that $E(U_0 \mid P(X)) = 0$, then, by rearranging terms, we obtain

$$E(U_0 \mid P(X), D=1) = -\frac{1-P(X)}{P(X)} E(U_0 \mid P(X), D=0). \quad (16)$$

Recalling that

$$B(X) = E\left(U_0 \mid X, D=1\right) - E\left(U_0 \mid X, D=0\right),$$

it follows from (16) that

$$B(P(X)) = \frac{-1}{P(X)} E(U_0 \mid P(X),\ D=0). \quad (17)$$

This expression can only change sign if $E(U_0 \mid P(X),\ D=0)$ changes sign for some value of $P(X)$, but this never happens in a symmetric distribution.

Appendix B: Methodology for Matching

The steps of the matching procedure used in this paper, which are similar to the procedure used by Lechner (1999) are as follows:

1. Split the observations into two pools, a treatment group T (further training) and a comparison group C (no further training). Estimate a probit model for participation in the treatment group.
2. Based on the estimated probit model compute the propensity score $\hat{b}'X_T$ and the variance $\text{var}(\hat{b}'X_T)$ for all treated persons T. Construct for all treated persons the interval $\hat{b}'X_T \pm w\sqrt{\text{var}(\hat{b}'X_T)}$, and choose w such that one obtains a confidence interval of the desired size around $\hat{b}'X_T$.[11]
3. Randomly select a treated person from the treatment group n_T.
4. Find observations in the control group that obey $\hat{b}'X_C \in \hat{b}'X_T \pm w\sqrt{\text{var}(\hat{b}'X_T)}$.
   A) If there *no* observation of the control group lying between the given limits of the confidence interval, the selected person will not be considered further and step 3 has to be repeated.
   B) If there is *only one* observation between the given limits of the confidence interval, go to step 6.
   C) If there is *more than one* observation in the confidence interval proceed as follows: Compute additional match variables in relation to the start date of observation n_T and a subset of variables already included in the estimation of the propensity score. Denote these variables a_T and a_C. Evaluate the distance $d(T,C) = (\hat{b}'X_T, a_T)' - (\hat{b}'X_C, a_C)'$ between each non-treated and treated. Choose those non-trainee who is the 'closest neighbor' of the trainee T in terms of the Mahalanobis distance, defined as: $md = d(T,C)'\text{cov}^{-1}d(T,C)$, where cov is the estimated sample covariance matrix of $(\hat{b}'X, a)'$ in the group of non-trainees.
5. Remove the treated and non-treated (now matched control) observations from their respective groups.
6. If there are any observations left in the trainee group, begin again with step 3.

In the illustrative cases reported in this paper, we matched on the propensity score by itself – no additional covariates were used.

[11] Rubin (1991), for example defined $w = 0.25$, while Lechner (1999) defined $w = 1.65$. We, like Rubin, set $w = 0.25$.

References

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association* 91: 444-455.

Beran, Rudolf. 1979. "Testing for Ellipsoidal Symmetry of a Multivariate Density." *The Annals of Statistics* 7: 150-162.

Dempster, A.P. 1969. *Elements of Continuous Multivariate Analysis*. Reading, MA: Addison-Wesley.

Dempster, A.P., N.M. Laird, and Donald B. Rubin. 1977. "Maximum Likelihood from Incomplete Data via the EM Algorithm (with Discussion)." *Journal of the Royal Statistical Society* B39: 1-38.

Greene, William H. 1999. *Econometric Analysis*, Volume 4. Upper Saddle River, New Jersey: Prentice Hall.

Hagenaars, Jacques A. 1990. *Categorical Longitudinal Data: Log-Linear Panel, Trend, and Cohort Analysis*. Newbury Park: Sage.

Haisken-DeNew, John P., and Joachim R. Frick (Eds.). 1998. *Desktop Companion to the German Socio-Economic Panel Study (GSOEP)*, Version 2.2. Berlin: German Institute for Economic Research (DIW).

Heckman, James J. 1979. "Sample Selection Bias as a Specification Error." *Econometrica* 47: 153-161.

Heckman, James J., and Richard Robb Jr. 1985. "Evaluating the Impact of Interventions." Pp. 156-245 in: James J. Heckman, and Burton Singer (eds.): *Longitudinal Analysis of Labor Market Data*. Cambridge: Cambridge University Press.

Heckman, James. 1997. "Instrumental Variables: A Study of Implicit Behavioral Assumptions Used in Making Program Evaluations." *Journal of Human Resources* 32: 441-462.

Heckman, James J., Hidehiko Ichimura, Jeffrey Smith, and Petra Todd. 1998. "Characterizing Selection Bias Using Experimental Data." *Econometrica* 66: 1017-1098.

Heckman, James J., Robert J. LaLonde, and Jeffrey A. Smith. 2000. "The Economics and Econometrics of Active Labor Market Programs." In *Handbook of Labor Economics, Vol.5*, edited by Orley Ashenfelter and David Card. Amsterdam: North Holland.

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81: 945-960.

Kahneman, Daniel, and Amos Tversky. 1979. "Prospective Theory: An Analysis of Decision under Risk." *Econometrica* 47: 263-291

Lechner, Michael. 1999. "Earnings and Employment Effects of Continuous Off-the-Job Training in East Germany After Unification." *Journal of Business & Economic Statistics* 17/1: 74-90.

Little, Roderick J.A., and Donald B. Rubin. 1987. *Statistical Analysis with Missing Data*. New York: Wiley.

Rosenbaum, Paul R., and Donald B. Rubin. 1984. "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score." *Journal of the American Statistical Association* 79: 516-524.

Rosenbaum, Paul R., and Donald B. Rubin. 1985. “Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score.” *The American Statistician* 39: 33-38.

Rubin, Donald B. 1976. “Inference and Missing Data.” *Biometrika* 63: 581-592.

Rubin, Donald B. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Rubin, Donald B. 1991. “Practical Implications of Models of Statistical Inference for Causal Effects and the Critical Role of the Assignment Mechanism. *Biometrics* 47: 1213-1234.

Rubin, Donald B., and Neal Thomas”. 1996. “Matching Using Estimated Propensity Scores: Relating Theory to Practice.” *Biometrics* 52: 249-261.

Smith, Herbert L. 1997. “Matching with Multiple Controls to Estimate Treatment Effects in Observational Studies.” Pp. 325-353 in *Sociological Methodology* 1997, edited by Adrian E. Raftery. Oxford: Basil Blackwell.

Sobel, Michael E. 1995. “Causal Inference in the Social and Behavioral Sciences.” Pp. 1-138 in Gerhard Arminger, Clifford C. Clogg, and Michael E. Sobel (eds.), *Handbook of Statistical Modeling for the Social and Behavioral Sciences.* New York: Plenum Press.

Winship, Christopher, and Robert D. Mare. 1989. “Loglinear Models with Missing Data: A Latent Class Approach.” *Sociological Methodology* 19: 331-367.

Winship, Christopher, and Stephen L. Morgan. 1999. “The Estimation of Causal Effects from Observational Data.” *Annual Review of Sociology* 25: 659-706.

Table 1: Estimated Selection Bias (in Dollars/Month) in Heckman et al.'s (1998) Comparison of Experimental and Non-Experimental Evaluation Data for Four Training Centers Participating in the Job Training Partnership Act

	Decile of the Empirical Distribution for the Probability of Choosing to be Trained								
	1	2	3	4	5	6	7	8	9
	[.0002, .0023)	[.0023, .0087)	[.0087, .0152)	[.0152, .0269)	[.0269, .0410)	[.0410, .0822)	[.0822, .0983)	[.0983, .1337)	[.1337, .2534]
Average Bias across 6 quarters	-282 (116)	-188 (91)	-118 (81)	-63 (79)	3 (98)	168 (130)	169 (117)	81 (147)	488 (281)

Source: Heckman et al. (1998: 1049), Table 7.

Table 2: Correlation Structures used in Simulations

	U_0	U_1	n_1	n_2
Case 1	Symmetric distribution, U_1 independent of U_0, n_1, and n_2.			
U_0	1			
U_1	-0.0079	1		
n_1	0.5984	-0.0116	1	
n_2	-0.7989	0.0067	-0.7999	1
Case 2a	Bias Cancellation, U_1 independent of U_0, n_1, and n_2, proportion trained in simulation equals proportion trained in observed data.			
U_0	1			
U_1	0.0063	1		
n_1	0.2999	0.009	1	
n_2	-0.395	-0.0063	-0.7978	1
Case 3b	Bias Cancellation, U_1 independent of U_0, n_1, and n_2, proportion trained in simulation is enhanced.			
U_0	1			
U_1	0.0012	1		
n_1	0.2814	-0.0062	1	
n_2	-0.3768	0.0054	-0.7982	1
Case 3	Symmetric distribution, U_1 is correlated with U_0, n_1, and n_2.			
U_0	1			
U_1	0.6949	1		
n_1	0.5928	0.4905	1	
n_2	-0.793	-0.4863	-0.7977	1
Case 4a	Bias Cancellation, U_1 is correlated with U_0, n_1, and n_2, proportion trained in simulation equals proportion trained in observed data.			
U_0	1			
U_1	0.695	1		
n_1	0.3029	0.244	1	
n_2	-0.406	-0.2462	-0.7977	1
Case 4b	Bias Cancellation, U_1 is correlated with U_0, n_1, and n_2, proportion trained in simulation is enhanced.			
U_0	1			
U_1	0.6949	1		
n_1	0.2817	0.2275	1	
n_2	-0.3764	-0.2302	-0.7977	1

Note: Case 2a has the same correlation structure as Case 1 if $P(\text{train}) < 0.5$. Correlations between U_0 and the other variables reverse if $P(\text{train}) > 0.5$. The simulation model assumes an enhanced proportion of the sample are trained. Case 2b is the same as Case 2a, except that the simulation model is based on a proportion trained that matches the actual GSOEP data, and the point of sign reversal is the midpoint of the propensity scores. Case 4a has the same correlation structure as Case 3 if $P(\text{train}) < 0.5$. Correlations between U_0, U_1 and the other variables reverse if $P(\text{train}) > 0.5$ (correlations between U_0 and U_1 remain the same regardless of $P(\text{train})$). The simulation model assumes an enhanced proportion of the sample are trained.

Table 3: Average Bias (Percent Change in Wages) as a Function of the Probability of Receiving Training and the Correlation Structure of the Error Distribution, Simulated Data Derived from the German Socio-Economic Panel

		Probability of Receiving Training					
		=0.2	0.2-0.4	0.4-0.6	0.6-0.8	0.8-1.0	Total
Independent Errors	Observed Fraction Trained	-.0009085 (7,739)	.0010467 (1,520)				-.00049803
	Enhanced Fraction Trained	.0032581 (424)	-.0001875 (4,104)	-.0009074 (3784)	.0012711 (947)		-.00044939
Symmetric Error Distribution $(U_0 n_1 n_2)$ (U_1)	Observed Fraction Trained	-.0424048 (7,739)	-.0378478 (1,520)				-.04031706
	Enhanced Fraction Trained	-.0422625 (424)	-.0371024 (4,104)	-.0369201 (3,784)	-.0356539 (947)		-.03489765
	Enhanced Fraction Trained + Missing Data	.005406 (231)	-.0161984 (2251)	-.0162938 (2038)	-.0209719 (538)		-.01631066
Symmetric Error Distribution $(U_0 n_1 n_2\ U_1)$	Observed Fraction Trained	-.0424055 (7,739)	-.0371473 (1,520)				-.04018091
	Enhanced Fraction Trained	-.0440875 (424)	-.0372851 (4,104)	-.0361595 (3,784)	-.0380109 (947)		-.03489929
	Enhanced Fraction Trained + Missing Data	.0056368 (225)	-.0170164 (2242)	-.0152895 (2,045)	-.0231464 (594)		-.01658261
Bias Canceling Error Distribution $(U_0 n_1 n_2)$ (U_1)	Observed Fraction Trained	-.029372 (7,739)	.0401667 (1,520)				-.00909556
	Enhanced Fraction Trained	-.0422625 (424)	-.0371024 (4,104)	-.0082022 (3,784)	.0356539 (947)		-.01576551
	Enhanced Fraction Trained + Missing Data	.005406 (231)	-.0161984 (2,251)	-.007105 (2,038)	.0209719 (585)		-.01386696
Bias Canceling Error Distribution $(U_0 n_1 n_2\ U_1)$	Observed Fraction Trained	-.0275602 (7,739)	.0371473 (1,520)				-.00908321
	Enhanced Fraction Trained	-.0440875 (424)	-.0372851 (4,104)	-.0077371 (3,784)	.0380109 (947)		-.01545179
	Enhanced Fraction Trained + Missing Data	.0056368 (225)	-.0170164 (2,242)	-.0067509 (2,045)	.0231464 (594)		-.01338929

Note: Sample size is in parentheses. Sample size is zero for cells without entries.

Table 4: Results from Simulated "Complete" and "Incomplete" data; Standard Errors in Parentheses. Imposed correlations between U_0, n_1, and n_2 as in Panels 1, 2a, and 2b of Table 2

	Fraction Trained is as Observed		Fraction Trained is Enhanced	
	Symmetric Distribution (No Bias Cancellation)	Distribution-Produced Bias Cancellation[a]	Symmetric Distribution (No Bias Cancellation)	Distribution-Produced Bias Cancellation[b]
1. "True" Experimental Effect	.0598882	.0598882	.0598882	.0598882
2. "True" Average Effect for the Self-Selected Sample	.0965456	.0666063	.080498	.0686005
3. "Naïve" Estimate	.05622857	.05751075	.04560031	.05283497
4. OLS w/ Complete Data	.0553208 (.0012067)	.0589223 (.0012051)	.0443285 (.0007998)	.0551916 (.0008216)
5. OLS w/ Incomplete Data	.0398571 (.0024087)	.0583686 (.0023052)	.0361194 (.0009991)	.0541858 (.0010707)
6. Matching w/ Complete Data	.0526012 (.0017689)	.05972 (.0017515)	.0438123 (.0009681)	.0560871 (.0010086)
7. Matching w/ Incomplete Data	.0365883 (.0037172)	.0624105 (.0042309)	.0333029 (.0013417)	.0550488 (.0017387)
8. Matching w/ Complete Data, Balance on *P*(train)	N/A	N/A	N/A	.0580569 (.0011865)
9. Matching w/ Incomplete Data, Balance on *P*(train)	N/A	N/A	N/A	.0588391 (.001522)
10. Percent Bias Reduction vs. OLS w/ Complete Data	Negative	10%	Negative	21%[c]
11. Percent Bias Reduction vs. OLS w/ Incomplete Data	Negative	49%	Negative	32%[c]

Notes: [a] Correlation reversed at the 75th percentile. [b] Correlation reversed at *P*(train) = 0.50.
[c] Computations based on prior balancing on *P*(train).

Table 5: Results from Simulated "Complete" and "Incomplete" Data; Standard Errors in Parentheses. Imposed Correlations between U_0, U_1, n_1, and n_2 as in Panels 3, 4a, and 4b of Table 2

	Fraction Trained is as Observed		Fraction Trained is Enhanced	
	Symmetric Distribution (No Bias Cancellation)	Distribution-Produced Bias Cancellation[a]	Symmetric Distribution (No Bias Cancellation)	Distribution-Produced Bias Cancellation[b]
1. "True" Experimental Effect	.0598882	.0598882	.0598882	.0598882
2. "True" Average Effect for the Self-Selected Sample	.0650643	.0621441	.0633629	.0612021
3. "Naïve" Estimate	.05306085	.05447632	.04415599	.0457503
4. OLS Estimated Effect w/ Complete Data	.0237555 (.0011862)	.0529266 (.0012418)	.0263096 (.0007654)	.0452318 (.0008271)
5. OLS Estimated Effect w/ Incomplete Data	.0194808 (.0020761)	.06458 (.002264)	.0275309 (.0009437)	.0496346 (.0010442)
6. Matching w/ Complete Data	.0245459 (.0016305)	.0555769 (.0019284)	.0251452 (.0010919)	.0458716 (.0010732)
7. Matching w/ Incomplete Data	.0204226 (.0032822)	.070556 (.0042763)	.0243591 (.0016763)	.0507479 (.0017142)
8. Matching w/ Complete Data, Balance on *P*(train)	N/A	N/A	N/A	.0524417 (.0012716)
9. Matching w/ Incomplete Data, Balance on *P*(train)	N/A	N/A	N/A	.0554937 (.0015522)
10. Percent Reduction vs. OLS w/ Complete Data	2%	29%	Negative	45%[c]
11. Percent Reduction vs. OLS w/ Incomplete Data	2%	Negative	Negative	51%[c]

Notes: [a] Correlation reversed at the 75th percentile. [b] Correlation reversed at *P*(train) = 0.50.
[c] Computations based on prior balancing on *P*(train).

Appendix Table A1: Probit equation that predicted further training, and that was used to simulate the further training variable, and wage change equation in the actual GSOEP data

	Probit: further training (N=9259)		OLS: wage change (N=9288)	
Variable	Coeff.	Std. Err.	Coeff.	Std. Err.
Age (years)	-.0142518	.0021539	-.0021475	.0002207
Female (0/1)	-.1845155	.0485277	.0107384	.0043469
German born (0/1)	.3209961	.1077199		
Partner employed (0/1)	.0460505	.0446889		
Number of kids	.0266904	.0201011		
Single (0/1)	-.0996424	.0680448		
Widowed (0/1)	-.2127866	.285247		
Divorced (0/1)	-.0983756	.1007506		
Separated (0/1)	.1491953	.1522548		
Highest degree: Hauptschule (0/1)	-.2587919	.1971771	.0148508	.0247899
Highest degree: Realschule (0/1)	.1936957	.1968371	.0169327	.0248534
Highest degree: Fachhochschule (0/1)	.2403201	.2073403	.05421	.0263434
Highest degree: Abitur (0/1)	-.0382701	.2186963	.0307022	.0273629
Highest degree: other (0/1)	a		.0361914	.0436854
Highest degree: University (0/1)	.5229582	.2143258	.0548742	.0277211
Part time employed (0/1)	-.309015	.0685107		
Further training (0/1)	b		.0199627	.0062371
Federal land: Schleswig-Holstein (0/1)	-.3694068	.117805		
Federal land: Hamburg (0/1)	-.1811827	.1434875		
Federal land: Niedersachsen (0/1)	-.1870488	.0944307		
Federal land: Bremen (0/1)	-.7159365	.2686803		
Federal land: Nordrhein-Westfalen (0/1)	-.3407164	.088223		
Federal land: Hessen (0/1)	-.2843468	.09855		
Federal land: Rheinland-Pfalz (0/1)	-.3335949	.1074662		
Federal land: Baden-Wuerttemberg (0/1)	-.1990602	.0922416		
Federal land: Bayern (0/1)	-.2995077	.0917482		
Constant	-.3512989	.2640488	.1613154	.0255958

Notes: Reference categories: Male; not German born; partner not employed; married; highest degree: no degree; full time employed; no further training; federal land: Berlin. a) category dropped because it predicts failure perfectly; b) category dropped due to collinearity.