The Impact of Measurement Error on Evaluation Methods Based on Strong Ignorability*

Erich Battistin
Institute for Fiscal Studies

Andrew Chesher
University College London and Cemmap

13th February 2004

Abstract

When selection bias can purely be attributed to *observables*, several estimators have been discussed in the literature to estimate the average effect of a binary treatment or policy on a scalar outcome. Identification typically exploits the *unconfoundedness* of the treatment, which is verified if the participation status is independent of potential outcomes conditional on observable covariates. Assuming unconfoundedness, the average effect of the treatment can be estimated by differencing within subpopulation averages of treated and untreated units, or by propensity score methods under an additional condition on the support of the covariates exploited. The latter condition, together with unconfoundedness, makes participation into the treatment group *strongly ignorable*, as defined by Rosenbaum and Rubin (1983). This paper studies the impact of covariate *measurement error* on commonly used evaluation methods based on strong ignorability. An approximate expression for the measurement error bias is derived, and conditions are discussed for this to be zero. A bias correction procedure is also presented, which uses non-parametric estimates of functionals of the distribution of observed covariates.

Keywords: potential outcomes, small sigma asymptotics, treatment effects

*Preliminary and incomplete. This paper benefited from discussions with Martin Browning, Hide Ichimura, Andrea Ichino, Tobi Klein, Enrico Rettore and Barbara Sianesi and from comments by audiences at CAM (Copenhagen, November 2003), Cemmap (London, December 2003), "Brucchi Luchino" Workshop (Milan, December 2003) and Tinbergen Institute (February 2004). Address for correspondence: Erich Battistin, Institute for Fiscal Studies, 7 Ridgmount Street, London WC1E 7AE - UK. E-mail: erich_b@ifs.org.uk.

Contents

1 Introduction **3**

2 Identification of treatment effects in the absence of measurement error **4**
- 2.1 Parameters of interest . . . 4
- 2.2 Ignorable assignment . . . 4
- 2.3 Identification results . . . 5
  - 2.3.1 Effect on the population . . . 6
  - 2.3.2 Effect on the treated . . . 7
  - 2.3.3 Alternative estimation strategies . . . 7
- 2.4 A parametric example . . . 8

3 Covariate measurement error **9**
- 3.1 Approximate distributions . . . 9
- 3.2 Approximate expectations . . . 10
- 3.3 Remarks . . . 11

4 The effect of using mismeasured regressors **11**
- 4.1 Effect on the population . . . 11
- 4.2 Effect on the treated . . . 12
- 4.3 A parametric example (continued) . . . 13

5 A bias correction procedure **14**
- 5.1 Effect on the treated . . . 14
- 5.2 A parametric example (continued) . . . 15

6 More than one covariate, just one with error **15**

7 Example **15**
- 7.1 Approximation to the bias . . . 16
- 7.2 Exact expression for the bias . . . 16
- 7.3 Bias correction . . . 17

8 Conclusions **18**

1 Introduction

When evaluating the effect of a programme it is common to impose the restriction that, conditional on a set of observable variables, potential outcomes and a participation indicator are independently distributed. Under this restriction and a support condition which together constitute the *strong ignorability* restriction of Rosenbaum and Rubin (1983), the average effect of treatment on the treated and the average treatment effect are identified. Estimation typically proceeds by propensity score matching or by comparing weighted averages of outcomes for participants and nonparticipants.

In practice the conditioning variables, X, with respect to which strong ignorability are maintained may be observed with error, that is, instead of realisations of X one observes realisations of $Z \equiv g(X, U)$ where U is a vector of measurement errors. This paper explores the impact of such covariate measurement error on commonly used programme evaluation methods such as propensity score matching. The strategy we employ is as follows.

When the strong ignorability restriction holds there are correspondences which identify parameters of interest (e.g. the average effect of treatment on the treated) as functionals of the distribution of observable outcomes and covariates. Let F_{YX} denote this distribution. In the absence of measurement error data are informative about F_{YX}. A parameter θ is identified by a correspondence, $\theta \twoheadleftarrow \mathcal{H}(F_{YX})$ and $\mathcal{H}$ is termed an identifying functional. Matching, and other estimators employed in practice, $\hat{\theta}$, are analogue estimators obtained by applying identifying functionals to an estimate of the distribution of observable outcomes and covariates, that is $\hat{\theta} \equiv \mathcal{H}(\hat{F}_{YX})$.

When measurement error is present data are informative about the distribution of observable outcomes and *measurement error contaminated* covariates. Let F_{YZ} denote this distribution. If the presence of measurement error is ignored, or not perceived, then parameters of interest are estimated using realizations of (Y, Z) as if they were realizations of (Y, X), that is $\hat{\theta} \equiv \mathcal{H}(\hat{F}_{YZ})$. Under quite weak conditions $\hat{\theta} \xrightarrow{p} \mathcal{H}(F_{YZ})$.

We study the properties of $\mathcal{H}(F_{YZ})$ and its relationship to $\mathcal{H}(F_{YX})$, in particular $\Delta \equiv \mathcal{H}(F_{YZ}) - \mathcal{H}(F_{YX})$. The value of Δ depends on details of the features of the distribution of Y, X and U and a case by case analysis is required if exact results are to be obtained. We are interested in the generic impacts of measurement error and obtain information about these by considering the local effects of measurement error, that is by considering the value of Δ when $Z = g(X, \sigma U)$ and σ is small.

We consider the case in which $Z = X + \sigma U$ and U and X are independently distributed. Under conditions to be stated, for functionals $\mathcal{H}$ of interest,

$$\mathcal{H}(F_{YZ}) = \mathcal{H}(F_{YX}) + \sigma^2 \mathcal{H}^*(F_{YX}) + o(\sigma^2)$$

where $\lim_{\sigma \to 0}(\sigma^{-2} o(\sigma^2)) = 0$. The functional $\mathcal{H}^*$ is obtained using the method employed in Chesher (1991). Properties of this functional are explored to shed light on the "first order" impact of measurement error and the way in which this depends upon features of F_{YX}.

Arguing as in Chesher and Schluter (2002) in the cases studied here $\mathcal{H}^*(F_{YX}) = \mathcal{H}^*(F_{YZ}) + o(\sigma^2)$ and so there is

$$\mathcal{H}(F_{YZ}) = \mathcal{H}(F_{YX}) + \sigma^2 \mathcal{H}^*(F_{YZ}) + o(\sigma^2).$$

Since data are informative about F_{YZ} it may be possible to estimate $\mathcal{H}^*(F_{YZ})$ and so gain a view of the likely first order effect of measurement error at conjectured values of the measurement error variance σ^2.

The method is applied in a set of simple cases in which the exact impact of measurement error can be calculated and the quality of the "small σ" approximation is investigated.

2 Identification of treatment effects in the absence of measurement error

Let (Y_1, Y_0) be the potential outcomes from participating and not participating, respectively, and let D be the participation status. The causal effect of the program is then defined as the difference between the two potential outcomes, $\beta = Y_1 - Y_0$, which is *not* observable since being exposed to (denied) the program reveals Y_1 (Y_0) but conceals the other potential outcome (Holland, 1986).

2.1 Parameters of interest

Average effect of the treatment in the population (β_p) and average effect of the treatment on the treated (β_t)

$$\begin{aligned} \beta_p &= E_{Y_1}(Y_1) - E_{Y_0}(Y_0), \\ \beta_t &= E_{Y_1|D}(Y_1|1) - E_{Y_0|D}(Y_0|1). \end{aligned}$$

The latter parameter is of interest if one wishes to evaluate the effect of the treatment on the population that is likely to take up the treatment (Heckman *et al.*, 1999).

2.2 Ignorable assignment

Selection bias is caused by the fact that program participants ($D = 1$) differ from non-participants ($D = 0$) with respect to characteristics that affect potential outcomes. It follows that, because of differences in the composition, the two groups would *not* have the same outcomes even in the absence of the program (see Heckman *et al.*, 1999).

When differences in the composition of participants and non-participants can purely be attributed to *observable* characteristics, one can control for the selection bias by including in the model the appropriate conditioning variables. Under these circumstances, identification of the mean impact rests on the existence of an observable vector of individual characteristics X such that *strong ignorability* with respect to X (SI-X) holds true (Rosenbaum and Rubin, 1983). This corresponds to say that the following *two* conditions are *jointly* satisfied

$$(Y_0, Y_1) \perp D|X, \tag{1}$$

$$Var(D|X) > 0. \tag{2}$$

According to (1), it is as if individuals were randomly assigned to the treatment with a probability depending on X provided that such probability is non-degenerate at each value of these variables.[1] In a randomized experiment the latter condition is satisfied by design, since each individual has a positive probability of being randomized into or out of the program. In the case of observational studies, the *common support* assumption (2) is instead required (see Heckman *et al.*, 1998, and Lechner, 2001).

Since units presenting the same characteristics X have a common probability to enter the program, then an operational rule to obtain an *ex post* experimental-like data set is to match participants to non-participants on such probability (the so called *propensity score*), whose dimension is *invariant* with respect to the dimension of X. In fact, it can be proved (Theorem 3 by Rosenbaum and Rubin, 1983) that if SI-X is satisfied, then the treatment assignment is strongly ignorable also given the propensity score.

In terms of distribution functions, SI-X implies

$$F_{Y_i|DX}(y_i|d,x) = F_{Y_i|X}(y_i|x), \qquad i = 0,1$$

where $d \in \{0,1\}$. Condition (1) is actually stronger than required to get identification of causal effects, since as discussed in the next section the following *mean independence* condition

$$E_{Y_i|DX}(Y_i|d,x) = E_{Y_i|X}(Y_i|x), \qquad i = 0,1$$

would be sufficient.[2]

2.3 Identification results

Identification results for the parameters of interest are reviewed in what follows (see Heckman *et al.*, 1999, and Imbens, 2004). Throughout this section, $\stackrel{a}{=}$ will imply that SI-X (or mean independence together with the common support condition) is required for the result to hold.

Assuming SI-X, the average effect of the treatment can be estimated by matching, differencing within subpopulation averages of treated and untreated units, or by propensity score methods. It is shown below that the asymptotic behavior of these estimators can be studied by looking at the quantities (3) and (4) if the target parameter is β_p, or (5) if the target parameter is β_t.

[1] Assumption (1) is often referred to in the literature as *unconfoundedness* of the treatment given X.

[2] In practise, seldom a convincing case is made for mean independence without the case being equally strong for (1). Moreover, under mean independence one can not identify average treatment effects on transformations of the original outcome.

2.3.1 Effect on the population

Let $Y = Y_0 + D\beta$ be the observed outcome and let $e_X(x) = E_{D|X}(D|x)$. It follows that

$$\begin{aligned} E_{Y_1}(Y_1) &= \int E_{Y_1|X}(Y_1|x) f_X(x) dx, \\ &\stackrel{a}{=} \int E_{Y_1|DX}(Y_1|1,x) f_X(x) dx, \\ &= \int \frac{E_{YD|X}(YD|x)}{e_X(x)} f_X(x) dx, \end{aligned} \tag{3}$$

and

$$\begin{aligned} E_{Y_0}(Y_0) &= \int E_{Y_0|X}(Y_0|x) f_X(x) dx, \\ &\stackrel{a}{=} \int E_{Y_0|DX}(Y_0|0,x) f_X(x) dx, \\ &= \int \frac{E_{YD|X}(Y[1-D]|x)}{1-e_X(x)} f_X(x) dx, \end{aligned} \tag{4}$$

with the last equalities of each expression following from

$$\begin{aligned} E_{YD|X}(YD|x) &= E_{Y_1|DX}(Y_1|1,x) e_X(x), \\ E_{YD|X}(Y[1-D]|x) &= E_{Y_0|DX}(Y_0|0,x)[1-e_X(x)]. \end{aligned}$$

The quantities above can be consistently estimated by their sample analogues (see Horvitz and Thompson, 1952, Rosenbaum, 1987, Hahn, 1998, and Hirano *et al.*, 2003)

$$\begin{aligned} \hat{E}_{Y_1}(Y_1) &= \frac{1}{n} \sum_{i=1}^{n} \frac{d_i}{e_X(x_i)} y_i, \\ \hat{E}_{Y_0}(Y_0) &= \frac{1}{n} \sum_{i=1}^{n} \frac{1-d_i}{1-e_X(x_i)} y_i, \end{aligned}$$

so that

$$\hat{\beta}_p = \hat{E}_{Y_1}(Y_1) - \hat{E}_{Y_0}(Y_0).$$

The quantity $e_X(x)$ represents the conditional probability of participation given the observed characteristics X, which is often referred to in the literature as the *propensity score* (Rosenbaum and Rubin, 1983). The interpretation of the weighting procedure is appealing: participants and non-participants are given more (less) weight depending on whether they are under (over) represented in the population with respect to their characteristics X. Regardless of the number of X variables, weights can be defined using the propensity score which is always a scalar.

2.3.2 Effect on the treated

Along the same lines of what discussed in the previous section,[3] it follows that

$$\begin{aligned} E_{Y_0|D}(Y_0|1) &= \int E_{Y_0|DX}(Y_0|1,x) f_{X|D}(x|1)dx, \\ &\stackrel{a}{=} \int E_{Y_0|DX}(Y_0|0,x) f_{X|D}(x|1)dx, \\ &= \int E_{Y_0|DX}(Y_0|0,x) \frac{e_X(x) f_X(x)}{P(D=1)}dx, \\ &= \int \frac{E_{YD|X}(Y[1-D]|x)}{1-e_X(x)} \frac{e_X(x)}{P(D=1)} f_X(x)dx. \end{aligned} \tag{5}$$

Therefore, a consistent estimate of the treatment effect can be obtained from

$$\begin{aligned} \hat{E}_{Y_1|D}(Y_1|1) &= \frac{1}{n_1}\sum_{i=1}^{n} d_i y_i, \\ \hat{E}_{Y_0|D}(Y_0|1) &= \frac{1}{n_1}\sum_{i=1}^{n} \frac{(1-d_i)e_X(x_i)}{1-e_X(x_i)} y_i, \end{aligned}$$

and

$$\hat{\beta}_t = \hat{E}_{Y_1|D}(Y_1|1) - \hat{E}_{Y_0|D}(Y_0|1).$$

2.3.3 Alternative estimation strategies

Estimation strategies alternative to the ones presented above can be obtained by using the empirical analogues of the distributions $f_X(x)$ and $f_{X|D}(x|1)$ combined with an estimator of the conditional expectation $E_{Y_d|DX}(Y_d|d,x)$, $d \in \{0,1\}$. This yields the *generalized matching* estimators

$$\begin{aligned} \hat{E}_{Y_1}(Y_1) &= \frac{1}{n}\sum_{i=1}^{n} \hat{E}_{Y_1|DX}(Y_1|1,x_i), \\ \hat{E}_{Y_0}(Y_0) &= \frac{1}{n}\sum_{i=1}^{n} \hat{E}_{Y_0|DX}(Y_0|0,x_i), \\ \hat{E}_{Y_0|D}(Y_0|1) &= \frac{1}{n_1}\sum_{i=1}^{n_1} \hat{E}_{Y_0|DX}(Y_0|0,x_i), \end{aligned}$$

for the quantities in (3), (4) and (5), respectively. Conditional expectations in the previous expressions can be estimated semi-non-parametrically following one of the several methods suggested in the literature (see Imbens, 2004, for a review).

It is worth noting that any "X-adjusted" estimator is asymptotically equivalent to an "$e_X(x)$-adjusted" estimator. This result straightforwardly follows

[3] Note that, throughout this section, only conditional (or mean) independence of Y_0 from D given X is required, as the Y_1 outcome does not enter the equations below.

from the fact that $X \perp D | e_X(x)$, that is from the fact that the propensity score is a *balancing score* for X (see Theorem 2 by Rosenbaum and Rubin, 1983, and Frölich, 2003). For example, by using this property and the law of iterated expectations one would get

$$\begin{aligned} & \int E_{Y_0|De_X}(Y_0|0,e) f_{e_X|D}(e|1) de, \qquad (6) \\ = & \int\int E_{Y_0|DX}(Y_0|0,x) f_{X|De_X}(x|0,e) f_{e_X|D}(e|1) dx de, \\ = & \int\int E_{Y_0|DX}(Y_0|0,x) f_{X|De_X}(x|1,e) f_{e_X|D}(e|1) dx de, \\ = & \int E_{Y_0|DX}(Y_0|0,x) f_{X|D}(x|1) dx, \end{aligned}$$

which corresponds to (5). The empirical analogue of (6) defines the *propensity score matching* estimator of β_t (see, for example, Heckman *et al.*, 1999). It follows that this class of estimators is also covered by our discussion.

2.4 A parametric example

To fix ideas, consider the following parametric regression

$$y_i = \alpha + \beta d_i + \delta x_i + \varepsilon_i \qquad (7)$$

for the case of homogeneous returns to the treatment ($\beta_i = \beta$) and $E(\varepsilon_i | d_i, x_i) = 0$. If participation is SI-X, then ordinary least squares provide a consistent estimate of β.

By partialing out the effect of D from (7)

$$E(y_i|d_i) = \alpha + \beta d_i + \delta E(x_i|d_i),$$

it follows that

$$\widetilde{y}_i = \delta \widetilde{x}_i + \varepsilon_i,$$

where $\widetilde{y}_i = y_i - E(y_i|d_i)$ and $\widetilde{x}_i = x_i - E(x_i|d_i)$. A consistent estimate of δ can be obtained from the last regression, and identification of β follows from

$$\beta = [E(y_i|1) - E(y_i|0)] - \delta[E(x_i|1) - E(x_i|0)].$$

Accordingly, the effect β is identified by the raw difference of mean outcomes *net of* the composition difference with respect to X scaled by δ.[4]

[4]Note that, in a fully controlled experiment, the distribution of X is the same for treated and controls, so that the last term in the previous expression is zero regardless of the value of δ.

3 Covariate measurement error

In what follows identification results for β_p and β_t are discussed when the sample analogues of the expressions in (3), (4) and (5) are computed unknowingly observing Z in place of X. Let $Z = X + U$ with $U \perp (X, D, Y)$ and $E[U] = 0$, $E[U^2] = \sigma^2$. For the moment regard X as *scalar continuously* distributed on the real line.

Two things are worth noting. First, measurement error U is such that Z and X have the same support, and this coincides with the real line. Second, the common support of the Z distributions is *not* modified by the measurement error and coincides with the common support of the X distributions (i.e. the real line). If (2) is verified, then $Var(D|Z) > 0$.

In what follows we show that *measurement error* bias arises in the estimation of β_p and β_t since SI-X does *not* imply SI-Z. In other words, if participants and non-participants are balanced with respect to Z, the two groups are *not* balanced with respect to the distribution of X so that the condition $X \perp D|Z$ fails to hold.[5] In what follows, conditions are derived for the measurement bias to be zero (Conditions 1-3 below).

3.1 Approximate distributions

Consider $F_{Y|DZ}$. Direct application of the approximation for conditional distribution functions when covariates are measured with error, given in Chesher (1991), regarding D as measured *without* error and X as measured *with* error, and using the SI-X assumption, gives[6]

$$F_{Y|DZ}(y|d,z) \simeq F_{Y|X}(y|z) + \sigma^2 F'_{Y|X}(y|z)\left(\frac{f'_{X|D}(z|d)}{f_{X|D}(z|d)}\right) + \frac{\sigma^2}{2}F''_{Y|X}(y|z),$$

where recall $Y \equiv (Y_0, Y_1)$ and $y \equiv (y_0, y_1)$ and $A \simeq B$ indicates $A = B + o(\sigma^2)$.[7]

[5] Since the conditional distribution of X given D and Z can be written as

$$\begin{aligned} f_{X|DZ}(x|d,z) &= \frac{f_{D|X}(d|x)}{f_{D|Z}(d|z)} f_{X|Z}(x|z), \\ f_{D|Z}(d|z) &= \int f_{D|X}(d|x) f_{X|Z}(x|z) dx, \end{aligned}$$

it follows that

$$f_{X|DZ}(x|d,z) = f_{X|Z} f(x|z) \Leftrightarrow \frac{f_{D|X}(d|x)}{\int f_{D|X}(d|x) f_{X|Z}(x|z) dx} = 1,$$

which is satisfied if $X \perp D$.

[6] Throughout this paper, we will assume that the conditions stated in Chesher (1991) are satisfied.

[7] For vector X and using the Einsteinian summation convention (summation over repeated raised and lowered indices) there is

$$F_{Y|DZ}(y|d,z) \simeq F_{Y|X}(y|z) + \sigma_{ij} F^i_{Y|X}(y|z)\left(\frac{f^j_{X|D}(z|d)}{f_{X|D}(z|d)}\right) + \frac{\sigma_{ij}}{2}F^{ij}_{Y|X}(y|z),$$

where $Z_k = X_k + U_k$ and $E[U_i U_j] = \sigma_{ij}$.

Note all the above is for the *joint* distribution of Y_1 and Y_0. We have for the marginal distribution of Y_i, $i \in \{0,1\}$

$$F_{Y_i|DZ}(y_i|d,z) \simeq F_{Y_i|X}(y_i|z) + \sigma^2 F'_{Y_i|X}(y_i|z)\left(\frac{f'_{X|D}(z|d)}{f_{X|D}(z|d)}\right) + \frac{\sigma^2}{2}F''_{Y_i|X}(y_i|z).$$

Thus, *locally*, Y is SI-Z if

$$F'_{Y_i|X}(y_i|z)\left(\frac{f'_{X|D}(z|1)}{f_{X|D}(z|1)} - \frac{f'_{X|D}(z|0)}{f_{X|D}(z|0)}\right) = 0, \qquad i \in \{0,1\}$$

for which a sufficient condition is either of the following

Condition 1 $F'_{Y_i|X}(y_i|z) = 0$ *for all values of its arguments.*

Condition 2 *For all values of* z

$$\frac{f'_{X|D}(z|1)}{f_{X|D}(z|1)} = \frac{f'_{X|D}(z|0)}{f_{X|D}(z|0)}.$$

The former condition virtually requires Y to be independent of X, which is not an interesting case. The latter condition requires $X \perp D$ which is also uninteresting (the propensity score would be uninformative under this condition).[8]

3.2 Approximate expectations

Replacing F by f gives the approximation for density functions (if Y is continuously distributed), as follows (see Chesher, 1991)

$$f_{Y_i|DZ}(y_i|d,z) \simeq f_{Y_i|X}(y_i|z) + \sigma^2 f'_{Y_i|X}(y_i|z)\left(\frac{f'_{X|D}(z|d)}{f_{X|D}(z|d)}\right) + \frac{\sigma^2}{2}f''_{Y_i|X}(y_i|z).$$

Replacing F by E gives the result for *regression* functions, as follows

$$E_{Y_i|DZ}(Y_i|d,z) \simeq E_{Y_i|X}(Y_i|z) + \sigma^2 E'_{Y_i|X}(Y_i|z)\left(\frac{f'_{X|D}(z|d)}{f_{X|D}(z|d)}\right) + \frac{\sigma^2}{2}E''_{Y_i|X}(Y_i|z).$$

As above, *mean independence* given Z holds if

$$E'_{Y_i|X}(y_i|z)\left(\frac{f'_{X|D}(z|1)}{f_{X|D}(z|1)} - \frac{f'_{X|D}(z|0)}{f_{X|D}(z|0)}\right) = 0, \qquad i \in \{0,1\}.$$

Accordingly, either Condition 2 or the following

Condition 3 $E'_{Y_i|X}(y_i|z) = 0$ *for all values* z.

are sufficient for mean independence given Z to hold.[9]

[8] There is, for all x

$$\int_{-\infty}^{x} \nabla_x \log f_{X|D}(s|1)ds = \int_{-\infty}^{x} \nabla_x \log f_{X|D}(s|0)ds$$

which implies

$$\log f_{X|D}(s|1) = \log f_{X|D}(s|0) + \kappa$$

for all x and $\kappa = 0$ since both densities must integrate to 1.

[9] The development of all these approximations most elegantly *starts* with the approximation for regression functions. The approximate distribution function is then obtained by noting

3.3 Remarks

Results in this section point out that groups of individuals balanced with respect to the distribution of Z are *not* balanced with respect to the distribution of X, so that the condition $X \perp D|Z$ fails to hold. Along the same lines, it straightforwardly follows that the propensity score based on Z is *not* a balancing score for X, so that the condition $X \perp D|e_Z$ is not satisfied. Accordingly, by computing any propensity score adjustment unknowingly based on Z in place of X, one will get biased estimates of the treatment effect.

However, it is worth noting that, *regardless of* the nature of the measurement error U, e_Z is a balancing score for Z, that is the condition $Z \perp D|e_Z$ is satisfied. This results holds whatever the nature of the error is and it is a straightforward implication of Theorem 2 by Rosenbaum and Rubin (1983). For example, along the same lines of what derived in (6), it can be shown that

$$\begin{aligned} & \int E_{Y_0|De_Z}(Y_0|0,e) f_{e_Z|D}(e|1)de, \\ = & \int E_{Y_0|DZ}(Y_0|0,z) f_{Z|D}(z|1)dz. \end{aligned}$$

In the next section, we will be interested in studying what happens to alternative estimators of the quantities (3), (4) and (5) when Z is used instead of X. The implication of $Z \perp D|e_Z$ stated in the last expression will allow us to develop an *unified* approach to studying the asymptotic behaviour of these estimators.

4 The effect of using mismeasured regressors

The measurement error bias is derived for β_p (Proposition 1) and β_t (Proposition 2). The proof of Proposition 1 is omitted because similar in spirit to the proof of Proposition 2, which is instead reported in the Appendix.[10]

4.1 Effect on the population

By using Z in place of X, one will obtain consistent estimators of

$$A_i = \int_{-\infty}^{\infty} E_{Y_i|DZ}(Y_i|i,z) f_Z(z)dz, \qquad i \in \{0,1\}$$

which correspond to (3) and (4) when Z is used instead of X. Limits of integration $(-\infty, \infty)$ will be suppressed in what follows.

---

that

$$F_{Y|DZ}(y|d,z) = E[1_{[Y_0 \leq y_0 \cap Y_1 \leq y_1]}|d,z],$$

and applying the formula for the approximation for regression functions. The approximation for density functions is obtained by differentiating the approximation for distribution functions.

[10] The regularity conditions required in these propositions are based on the assumption

$$\int f_{X|D}(x+\lambda|0)dz = 1, \qquad \forall \lambda : |\lambda| \leq \tau.$$

Proposition 1 *If SI-X holds and*

$$\lim_{z\to\pm\infty} E_{Y_i|X}(Y_i|z)f'_X(z) = 0,$$
$$\lim_{z\to\pm\infty} E'_{Y_i|X}(Y_i|z)f_X(z) = 0,$$

neglecting terms which are $o(\sigma^2)$ there is the following expression for A_i

$$\boxed{A_i \simeq E_{Y_i}[Y_i] + \sigma^2 B_i,}$$

where

$$\begin{aligned} B_i &= \int E'_{Y_i|X}(Y_i|z)\frac{f'_{X|D}(z|i)}{f_{X|D}(z|i)}f_X(z)dz \\ &+ \int E''_{Y_i|X}(Y_i|x)f_X(z)dz. \qquad \blacksquare \end{aligned}$$

Accordingly, the estimated effect in the population differs from the true effect (at the second order for σ) by means of the following factor

$$\begin{aligned} \Delta(\beta_p) &= \sigma^2(B_1 - B_0) \\ &= \int \left[E'_{Y_1|DX}(Y_1|1,z)\frac{f'_{X|D}(z|1)}{f_{X|D}(z|1)} - E'_{Y_0|DX}(Y_0|0,z)\frac{f'_{X|D}(z|0)}{f_{X|D}(z|0)}\right] f_X(z)dz \\ &+ \int \left[E''_{Y_1|DX}(Y_1|1,x) - E''_{Y_0|DX}(Y_0|0,x)\right] f_X(z)dz. \end{aligned}$$

4.2 Effect on the treated

Under SI-X there is

$$E_{Y_0|D}[Y_0|1] = \int E_{Y_0|DX}(Y_0|0,x)\frac{f_{X|D}(x|1)}{f_{X|D}(x|0)}f_{X|D}(x|0)dx.$$

Someone unknowingly observing Z in place of X and computing the sample analogue of this expression will obtain an estimator of

$$A = \int E_{Y_0|DZ}(Y_0|0,z)f_{Z|D}(z|1)dz.$$

Proposition 2 *If SI-X holds and*

$$\lim_{z\to\pm\infty} E_{Y_0|DX}(Y_0|0,z)f'_{X|D}(z|1) = 0,$$
$$\lim_{z\to\pm\infty} E'_{Y_0|DX}(Y_0|0,z)f_{X|D}(z|1) = 0,$$

neglecting terms which are $o(\sigma^2)$ there is the following expression for A

$$\boxed{A \simeq E_{Y_0|D}[Y_0|1] + \sigma^2 B,} \tag{8}$$

where

$$\begin{aligned} B &= \int E'_{Y_0|DX}(Y_0|0,z)\left(\frac{f'_{X|D}(z|0)}{f_{X|D}(z|0)}\right) f_{X|D}(z|1)dz \\ &\quad + \int E''_{Y_0|DX}(Y_0|0,z) f_{X|D}(z|1)dz. \qquad \blacksquare \end{aligned}$$

Accordingly, the estimated effect differs from the true effect in the population by means of the following term

$$\Delta(\beta_t) = \sigma^2 B.$$

Consider the case in which $f_{X|D}(z|1) = f_{X|D}(z|0)$. Then the first term in B becomes

$$\begin{aligned} \int E'_{Y_0|DX}(Y_0|0,z) f'_{X|D}(z|0)dz &= \int E'_{Y_0|DX}(Y_0|0,z) f'_{X|D}(z|1)dz, \\ &= -\int E''_{Y_0|DX}(Y_0|0,z) f_{X|D}(z|1)dz, \end{aligned}$$

the second line following on integrating by parts. Clearly in this case $B = 0$, which is as it should be.

4.3 A parametric example (continued)

Using the parametric example introduced above, it is easy to show that measurement error in X will make ordinary least squares estimates biased for β. In fact, classical measurement error in X implies that using Z as a proxy for X will partially, *but only partially*, control for the confounding effects of X on the estimation of β (Wickens, 1972). Measurement error in X biases not only δ (which is a nuisance parameter for the problem), but more importantly biases also β (unless D and X are not correlated, which is not an interesting case).

Since $z_i = x_i + u_i$, the estimation of δ based on

$$\widetilde{y}_i = \delta \widetilde{z}_i + \upsilon_i$$

features the usual *attenuation bias*, so that the following parameter

$$\frac{\sigma_x^2}{\sigma_x^2 + \sigma^2}\delta$$

is estimated in place of δ. Accordingly

$$[E(y_i|1) - E(y_i|0)] - \delta[E(x_i|1) - E(x_i|0)]\frac{\sigma_x^2}{\sigma_x^2 + \sigma^2} \neq \beta.$$

Because of the the measurement error U, the difference in raw means for X is only *partially* 'washed out' from the difference in raw means for Y, resulting in biased estimates for the effect β. Note that Condition 3 here would be satisfied if $\delta = 0$.

5 A bias correction procedure

The most common solution to the bias introduced by the measurement error in linear regression models is to exploit instrumental variables. However, it is well known that they do not yield consistent estimators of the parameters of interest in non-linear models (see, for example, Hausman *et al.*, 1995).

This section is along the same lines of what discussed in Chesher (2000). A method is proposed for obtaining estimates of the treatment effects which are purged of the major part of the effect of the measurement error. The method uses a quantity constructed from non-parametric estimates of functionals of the distribution of observed covariates Z. It follows that our procedure exploits *nothing but* the error contaminated data and does not require any functional assumptions on the regression of Y on D and X nor additional information (such as instrumental variables or validation data).[11]

In what follows, we will discuss how our correction procedure works for β_t. In further work, we will also apply the same correction to β_p.

5.1 Effect on the treated

Since X can be replaced by Z in expressions (e.g. B) multiplied by σ^2 without altering the order of the approximation error we have

$$\boxed{A \simeq E_{Y_0|D}[Y_0|1] + \sigma^2 B^*,}$$

where

$$\begin{aligned} B^* &= \int E'_{Y_0|DZ}(Y_0|0,z)\left(\frac{f'_{Z|D}(z|0)}{f_{Z|D}(z|0)}\right) f_{Z|D}(z|1)dz \\ &\quad + \int E''_{Y_0|DZ}(Y_0|0,z) f_{Z|D}(z|1)dz. \end{aligned}$$

This corresponds to what derived in (8) when X is replaced by Z. As the last expression can be rearranged to get

$$\int \left[E'_{Y_0|DZ}(Y_0|0,z)\frac{d}{dz}\log f_{Z|D}(z|0) + E''_{Y_0|DZ}(Y_0|0,z)\right] f_{Z|D}(z|1)dz,$$

it follows that B^* can be estimated by

$$\begin{aligned} \hat{B}^* &= \frac{1}{n_1}\sum_{i=1}^{n}\frac{(1-d_i)e_Z(z_i)}{1-e_Z(z_i)}b(z_i), \\ b(z_i) &= E'_{Y_0|DZ}(Y_0|0,z_i)\frac{d}{dz}\log f_{Z|D}(z_i|0) - E''_{Y_0|DZ}(Y_0|0,z_i), \end{aligned}$$

from available data.

To estimate $E'_{Y_0|DZ}(Y_0|0,z)$ and $E''_{Y_0|DZ}(Y_0|0,z)$ do parametric or nonparametric estimation of the regression of Y_0 on Z for people with $D = 0$ and

[11] As pointed out by Chesher (2000), when the error free regression function of Y on X is linear in X, the method proposed here can be combined with conventional instrumental variables methods.

calculate first and second derivatives with respect to Z. To estimate the remaining elements one can do nonparametric density estimation for the $D = 0$ group (see the discussion in Chesher, 2000). Alternatively one might have a parametric model for D given X in which case one could estimate that and then do nonparametric density estimation of $f_Z(z)$ and then use, e.g.

$$\hat{f}_{Z|D}(z|0) = \frac{[1 - e_Z(z_i)]\hat{f}_Z(z)}{\hat{P}[D = 0]}.$$

5.2 A parametric example (continued)

It follows from (7) that

$$E(Y|d, z) = \beta d + \delta z - \delta E(U|d, z),$$

since $E(\varepsilon_i|d_i, x_i) = 0$. The last expression qualifies the bias induced by measurement error as an *omitted variable* problem. The regression of Y on D and Z fails to identify the parameter of interest β because the term $E(U|d, z)$ is omitted from the regression. Chesher (2000) shows that the following approximation holds

$$E(Y|d, z) \simeq \beta d + \delta z - \delta\sigma^2 g(d, z),$$

where $g(d, z)$ is a term that can be estimated from *observed* data (i.e. it is function of Z and D only). The augmented regression including the $g(d, z)$ term can be used to get a 'bias reduced' estimate of β. Note that, as long as $g(d, z)$ is *not* linear in Z (which would be true if U was normally distributed), then σ^2 could also be estimated from observed data.

6 More than one covariate, just one with error

In the expressions above, differentiation is with respect to the error contaminated covariate and the density $f_{X|D}$ becomes $f_{X^*|X_*D}$ where X^* is the error contaminated covariate and X_* contains the remaining covariates.

7 Example

This example is artificial, but rather convenient. Throughout this section normality will be assumed for the error U. Moreover, suppose that the regression function of Y on X for the $D = 0$ group is linear (as in Rubin, 1977)

$$E_{Y_0|DX}(Y_0|0, x) = \alpha_0 + \beta_0 x$$

and that

$$X|D = d \sim N(d\mu_1 + (1 - d)\mu_0, d\lambda_1^2 + (1 - d)\lambda_0^2),$$

for $d \in \{0,1\}$.

Assume that β_t is of interest to the analyst. According to what presented in the previous section, we wish to approximate

$$A = \int E_{Y_0|DZ}[Y_0|0,z] f_{Z|D}(z|1)dz,$$

which is what people will unwittingly estimate if they ignore measurement error. Three quantities are derived for the example considered in this section: the approximation to the measurement error bias in Proposition 2 is in (9); the *exact* expression for this bias (that is, the expression in terms of the unobserved X) is in (10); finally, the bias resulting from our correction procedure is in (11).

7.1 Approximation to the bias

The approximation as derived above, that is the right hand side of (8), is as follows

$$A_X^a \equiv \alpha_0 + \beta_0\mu_1 + \sigma^2 \int \left[E'_{Y_0|X}(Y_0|z) \left(\frac{f'_{X|D}(z|0)}{f_{X|D}(z|0)} \right) + E''_{Y_0|X}(Y_0|z) \right] f_{X|D}(z|1)dz,$$

where we stress the dependence from distributions and expectations involving X by writing A_X^a. Since

$$\begin{aligned} E'_{Y_0|DX}(Y_0|0,z) &= \beta_0 \\ E''_{Y_0|DX}(Y_0|0,z) &= 0 \\ \frac{f'_{X|D}(z|0)}{f_{X|D}(z|0)} &= -\frac{1}{\lambda_0^2}(z-\mu_0), \end{aligned}$$

we have

$$A_X^a = \alpha_0 + \beta_0\mu_1 - \beta_0 \frac{\sigma^2}{\lambda_0^2}(\mu_1 - \mu_0),$$

so that

$$\boxed{bias(A_X^a) = -\beta_0 (\mu_1 - \mu_0) \tfrac{\sigma^2}{\lambda_0^2}.} \tag{9}$$

Although the approximation A_X^a is not exact, the approximation error is of order $O(\sigma^4)$.[12]

7.2 Exact expression for the bias

The *exact* expression for A is as follows. First consider the expectation in the expression for A. We have, conditional on $D = 0$

$$\begin{bmatrix} X \\ Z \end{bmatrix} | D = 0 \sim N\left(\begin{bmatrix} \mu_0 \\ \mu_0 \end{bmatrix}, \begin{bmatrix} \lambda_0^2 & \lambda_0^2 \\ \lambda_0^2 & \lambda_0^2 + \sigma^2 \end{bmatrix} \right),$$

and so

$$X | (Z \cap D = 0) \sim N\left(\mu_0 + \frac{\lambda_0^2}{\lambda_0^2 + \sigma^2}(z - \mu_0), \lambda_0^2 - \frac{\lambda_0^4}{\lambda_0^2 + \sigma^2} \right).$$

[12] It is the symmetric distribution of U which causes $O(\sigma^3)$ terms to disappear.

Therefore, for the expectation appearing in A there is (remember that $Y_0 \perp Z|X$)

$$\begin{aligned} E_{Y_0|DZ}(Y_0|0,z) &= \int E_{Y_0|DZX}(Y_0|0,z,x) f_{X|ZD}(x|z,0)dx, \\ &= \int (\alpha_0 + \beta_0 x)\, f_{X|ZD}(x|z,0)dx, \\ &= \alpha_0 + \beta_0\mu_0 + \frac{\beta_0\lambda_0^2}{\lambda_0^2+\sigma^2}(z-\mu_0), \end{aligned}$$

which exhibits the usual attenuation, and since $Z|D=1 \sim N(\mu_1, \lambda_1^2+\sigma^2)$

$$\begin{aligned} A &= \alpha_0 + \beta_0\mu_0 + \frac{\beta_0\lambda_0^2}{\lambda_0^2+\sigma^2}(\mu_1-\mu_0), \\ &= \alpha_0 + \beta_0\mu_1 - \beta_0(\mu_1-\mu_0)\frac{\sigma^2}{\lambda_0^2+\sigma^2}. \end{aligned}$$

The final term gives the *exact* bias caused by measurement error[13]

$$\boxed{bias(A) = -\beta_0(\mu_1-\mu_0)\left(\tfrac{\sigma^2}{\lambda_0^2+\sigma^2}\right).} \tag{10}$$

The accuracy of the approximation is understood by considering

$$A - A_X^a = \beta_0(\mu_1-\mu_0)\frac{\sigma^4}{\lambda_0^2(\lambda_0^2+\sigma^2)}.$$

7.3 Bias correction

Our bias correction procedure proposes subtracting from a consistent estimator of A a consistent estimator of $\sigma^2 B^*$, where B^* is defined as follows

$$B^* = \int \left[E'_{Y_0|DZ}(Y_0|0,z)\left(\frac{f'_{Z|D}(z|0)}{f_{Z|D}(z|0)}\right) + E''_{Y_0|DZ}(Y_0|0,z)\right] f_{Z|D}(z|1)dz.$$

The value of B^* is now derived for this example. Since

$$\begin{aligned} E'_{Y_0|DZ}(Y_0|0,z) &= \frac{\beta_0\lambda_0^2}{\lambda_0^2+\sigma^2}, \\ E''_{Y_0|DZ}(Y_0|0,z) &= 0, \\ \left(\frac{f'_{Z|D}(z|0)}{f_{Z|D}(z|0)}\right) &= -\frac{1}{\lambda_0^2+\sigma^2}(z-\mu_0), \end{aligned}$$

it follows that

$$B^* = -\frac{\beta_0\lambda_0^2}{(\lambda_0^2+\sigma^2)^2}(\mu_1-\mu_0).$$

Using our proposed procedure produces a consistent estimator of

$$A^{cor} \equiv A - \sigma^2 B_Z = \alpha_0 + \beta_0\mu_1 - \beta_0(\mu_1-\mu_0)\frac{\sigma^4}{(\lambda_0^2+\sigma^2)^2}.$$

[13] Note, just to check, that when $\sigma^2 = 0$ (that is when $Z = X$) this reduces to $A = \alpha_0 + \beta_0\mu_1 = E_{Y_0|D}[Y_0|1]$.

So, after our correction procedure, the bias in (10) is replaced by a bias equal to

$$\boxed{bias(A^{cor}) = -\beta_0(\mu_1 - \mu_0)\left(\frac{\sigma^2}{\lambda_0^2+\sigma^2}\right)^2.} \tag{11}$$

8 Conclusions

This paper proposes a method for bias reduction in estimation of treatment effects based on ignorable assignment given a set of covariates, with one covariate subject to measurement error. Our procedure exploits nothing but the error contaminated covariate data.

In further work, we will look at exact calculations designed to investigate the performance of the proposed procedure. Moreover, we will apply the approach described here to real data.

References

[1] Chesher, A. (1991), *The Effect of Measurement Error*, Biometrika, Vol. 78, No. 3, pp. 451-462

[2] Chesher, A. (2000), *Measurement Error Bias Reduction*, unpublished manuscript, University College London

[3] Chesher, A. and Schluter, C. (2002), *Welfare Measurement and Measurement Error*, Review of Economic Studies, Vol. , No. , pp. ??-??

[4] Frölich, M. (2003), *Programme Evaluation and Treatment Choice*, Lecture Notes in Economics and Mathematical Systems, Berlin: Spriger-Verlag

[5] Hahn, (1998), *On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects*, Econometrica, Vol. 66, No. 2, pp. 315-331

[6] Hausman, J.A. Newey, W.K. and Powell, J.L. (1998), *Nonlinera Errors in Variables Estimation of Some Engel Curves*, Journal of Econometrics, Vol. 66, No. 5, pp. 1017-1098

[7] Heckman, J.J. Ichimura, H. Smith, J. and Todd, P. (1998), *Characterizing Selection Bias Using Experimental Data*, Econometrica, Vol. 65, No. , pp. 205-233

[8] Heckman, J.J. Lalonde, R. and Smith, J. (1999), *The Economics and Econometrics of Active Labor Market Programs*, Handbook of Labor Economics, Volume 3, Ashenfelter, A. and Card, D. (eds.), Amsterdam: Elsevier Science

[9] Hirano, K. Imbens, G. and Ridder, G. (2003), *Efficient Estimation of Average Treatment Effects using the Estimated Propensity Score*, Econometrica, Vol. 71, No. 4, pp. ???

[10] Holland, P. (1986), *Statistics and Causal Inference*, Journal of the American Statistical Association, Vol. 81, No. 396, pp. 945-970

[11] Horvitz, D.G. and Thompson, D.J. (1952), *A Generalization of Sampling Without Replacement From a Finite Universe*, Journal of the American Statistical Association, Vol. 47, No. 260, pp. 663-685

[12] Imbens, G.W. (2004), *Semiparametric Estimation of Average Treatment Effects under Exogeneity: a Review*, Review of Economics and Statistics, forthcoming

[13] Lechner, M. (2001), *A note on the common support problem in applied evaluation studies*, Discussion Paper 2001-01, Department of Economics, University of St. Gallen

[14] Rosenbaum, P.R. (1987), *Model-Based Direct Adjustment*, Journal of the American Statistical Association, Vol. 82, No. 398, pp. 387-394

[15] Rosenbaum, P.R. and Rubin, D.B. (1983), *The central role of the propensity score in observational studies for causal effects*, Biometrika, Vol. 70, No. 1, 41-55

[16] Rubin, D.B. (1977), *Assignment to Treatment Group on the Basis of a Covariate*, Journal of Educational Statistics, Vol. 2, 4-58

[17] Wickens, M.R. (1972), *A Note on the Use of Proxy Variables*, Econometrica, Vol. 40, No. 4, pp. 759-761

Appendix

Proof of Proposition 2

Proof. Using the approximation to $E_{Y_0|DZ}(Y_0|0,z)$ and the approximation

$$f_{Z|D}(z|1) \simeq f_{X|D}(z|1) + \frac{\sigma^2}{2} f''_{X|D}(z|1)$$

gives

$$\begin{aligned} A &\simeq \int \left(E_{Y_0|X}(Y_0|z) + \sigma^2 E'_{Y_0|X}(Y_0|z) \left(\frac{f'_{X|D}(z|0)}{f_{X|D}(z|0)} \right) + \frac{\sigma^2}{2} E''_{Y_0|X}(Y_0|z) \right) \\ &\quad \times \left(f_{X|D}(z|1) + \frac{\sigma^2}{2} f''_{X|D}(z|1) \right) dz \end{aligned}$$

and neglecting terms which are $o(\sigma^2)$ there is the following expression for A:

$$A \simeq E_{Y_0|D}[Y_0|1] + \sigma^2 B$$

where

$$\begin{aligned} B &= \int E'_{Y_0|DX}(Y_0|0,z) \left(\frac{f'_{X|D}(z|0)}{f_{X|D}(z|0)} \right) f_{X|D}(z|1) dz \\ &\quad + \frac{1}{2} \int E''_{Y_0|DX}(Y_0|0,z) f_{X|D}(z|1) dz \\ &\quad + \frac{1}{2} \int E_{Y_0|DX}(Y_0|0,z) f''_{X|D}(z|1) dz. \end{aligned}$$

Consider the final term in this expression. On integrating by parts once we have

$$\begin{aligned} \int_{-\infty}^{\infty} E_{Y_0|DX}(Y_0|0,z) f''_{X|D}(z|1) dz &= \Big[E_{Y_0|DX}(Y_0|0,z) f'_{X|D}(z|1) \Big]_{-\infty}^{\infty} \\ &\quad - \int_{-\infty}^{\infty} E'_{Y_0|DX}(Y_0|0,z) f'_{X|D}(z|1) dz \end{aligned}$$

and if[14]

$$\lim_{z \to \pm\infty} E_{Y_0|DX}(Y_0|0,z) f'_{X|D}(z|1) = 0$$

there is

$$\int_{-\infty}^{\infty} E_{Y_0|DX}(Y_0|0,z) f''_{X|D}(z|1) dz = - \int_{-\infty}^{\infty} E'_{Y_0|DX}(Y_0|0,z) f'_{X|D}(z|1) dz.$$

Integrating by parts a second time gives

$$\begin{aligned} - \int_{-\infty}^{\infty} E'_{Y_0|DX}(Y_0|0,z) f'_{X|D}(z|1) dz &= - \Big[E'_{Y_0|DX}(Y_0|0,z) f_{X|D}(z|1) \Big]_{-\infty}^{\infty} \\ &\quad + \int_{-\infty}^{\infty} E''_{Y_0|DX}(Y_0|0,z) f_{X|D}(z|1) dz \end{aligned}$$

[14] This condition will be satisfied if for example $E_{Y_0|DX}(Y_0|0,z)$ is a polynomial function of z and the tails of $f_{X|D}(z|1)$ decrease at an exponential rate.

and if

$$\lim_{z \to \pm\infty} E'_{Y_0|DX}(Y_0|0,z) f_{X|D}(z|1) = 0$$

there is

$$\int_{-\infty}^{\infty} E_{Y_0|DX}(Y_0|0,z) f''_{X|D}(z|1) dz = \int_{-\infty}^{\infty} E''_{Y_0|DX}(Y_0|0,z) f_{X|D}(z|1) dz$$

and then

$$\begin{aligned} B &= \int E'_{Y_0|DX}(Y_0|0,z) \left(\frac{f'_{X|D}(z|0)}{f_{X|D}(z|0)} \right) f_{X|D}(z|1) dz \\ &\quad + \int E''_{Y_0|DX}(Y_0|0,z) f_{X|D}(z|1) dz. \end{aligned}$$

■