cemmap

centre for microdata methods and practice

Treatment effect estimation with covariate measurement error

Erich Battistin
Andrew Chesher

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP25/09

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

An ESRC Research Centre

Treatment Effect Estimation with Covariate Measurement Error*

Erich Battistin
University of Padova and IRVAPP

Andrew Chesher
University College London and Cemmap

September 7, 2009

Abstract

This paper investigates the effect that covariate measurement error has on a conventional treatment effect analysis built on an unconfoundedness restriction that embodies conditional independence restrictions in which there is conditioning on error free covariates. The approach uses small parameter asymptotic methods to obtain the approximate generic effects of measurement error. The approximations can be estimated using data on observed outcomes, the treatment indicator and error contaminated covariates providing an indication of the nature and size of measurement error effects. The approximations can be used in a sensitivity analysis to probe the potential effects of measurement error on the evaluation of treatment effects.

Keywords: measurement error, potential outcomes, small sigma asymptotics, treatment effects

*Original draft February 2004. This paper benefited from discussions with Martin Browning, Sergio Firpo, Markus Frölich, Hide Ichimura, Andrea Ichino, Tobi Klein, Michael Lechner, Enrico Rettore, Don Rubin, Barbara Sianesi and from comments by audiences at CAM (Copenhagen, November 2003), Cemmap (London, December 2003), Brucchi Luchino Workshop (Milan, December 2003), Tinbergen Institute (Amsterdam, February 2004), Microeconometrics of Spatial and Grouped Data (Banff, April 2004), North American Summer Meeting of the Econometric Society (Providence, June 2004), Econometric Study Group (Bristol, July 2004), European Meeting of the Econometric Society (Madrid, August 2004) and 2^{nd} Conference on Evaluation Research (Manheim, October 2004). Address for correspondence: Erich Battistin, Dipartimento di Scienze Statistiche, Via Cesare Battisti 241, 35123 Padova - I. E-mail: erich.battistin@unipd.it.

1 Introduction

In the absence of randomized assignment to treatment a commonly used strategy for identifying the causal effect of participation relies on an independence restriction requiring that counterfactual outcomes under "treatment" and "no treatment" are independent of treatment status conditional on a list of covariates. Different versions of this assumption are referred to as *ignorability* (Rubin, 1974, and Rosenbaum and Rubin, 1983), *selection on observables* (Heckman and Robb, 1985, and Heckman *et al.*, 1998), *conditional independence* (Lechner, 2001) or *unconfoundedness* (Imbens, 2004). The analysis is thus conditional on covariates that capture aspects of individuals' characteristics and environment which predispose individuals towards assignment rather than non-assignment or participation rather than non-participation.

This paper investigates the effect that covariate measurement error has on the conclusions drawn from a conventional treatment effect analysis that exploits independence restrictions that are conditional on error free covariates. The relevance of the problem for empirical applications rests upon the evidence on the extent of measurement errors in survey reports provided by several studies in the literature (see, for example, Bound *et al.*, 2001). We show that in most interesting scenarios the identifying restriction which holds when conditioning is on error free covariates fails to hold once conditioning is on error contaminated covariates. This is the case even when measurement error is simple in form and classical.

It thus follows that commonly employed procedures for estimating causal parameters employing propensity score matching or re-weighting using cross section or panel data may give misleading results when there is covariate measurement error. The precise effect of measurement error depends on detailed aspects of the data generating process about which there is little information in practice. This

paper thus focuses on approximations which are informative about the *generic* effects of measurement error in treatment effect analysis. The approximations provide the basis for sensitivity analysis which can highlight cases where the impact of measurement may be severe.

The main results of the paper can be summarized as follows. First, we show that measurement error does not necessarily imply attenuation effects for the causal parameters of interest. The net effect of measurement error varies in sign and magnitude from case to case, resulting from a complex combination of effects on the propensity score, on the densities of observed covariates and on the regression equations relating counterfactual outcomes to covariates and the treatment indicator. Since the estimation of treatment effects based on conditional independence assumptions generally involves complex functionals of the data, much of the simplicity of having classical errors is lost. However, we provide a simple rule to sign measurement error bias that is valid if the propensity score for error-free data follows a logit model and the no-treatment outcome equation is approximatively linear in the covariates needed to achieve identification.

Second, we show that there is information in the distributions of observable outcomes and measurement error contaminated covariates that can be used to obtain an indication of the direction and size of the measurement error bias in any particular case. This can be used to produce a partial correction which can readily be implemented using existing software, thus making our results operational.

Much of the recent work has considered identification and estimation of treatment effects in a general non-parametric setting (see Heckman *et al.*, 1999, and Imbens, 2004, for a review). However, measurement error in the treatment effect context has been little studied. For example, in the review paper by Bound *et al.* (2001) there is no mention of the problem. While there are recent results regarding misclassification of treatment indicators (see, for example, Mahajan, 2006, Lewbel, 2007, Hu, 2008, Molinari, 2008, and Battistin and Sianesi, 2009), to the best of our knowledge the only paper

to study covariate measurement error within a programme evaluation context is Cochran and Rubin (1973).[1] In reviewing the effectiveness of regression adjustment to control for confounding variables in observational studies, Cochran and Rubin (1973) exploit parametric (linear) regression equations relating potential outcomes to covariates and deal with the complication of having covariates affected by (not necessarily classical) measurement error. Their result can be obtained as a special case of the results presented in this paper. In fact, the expression for the bias derived by Cochran and Rubin (1973, page 431) coincides with the expression for the bias that we derive for a simple parametric case that we will use throughout this paper to set the methods developed in a familiar context.

Measures of causal effects such as the average treatment effect (ATE) and the average effect of treatment on the treated (ATT) are related to distributions of observables in a rather complex fashion which involves essential non-linearities. Recent results in the statistical literature on measurement error in non-linear models are surveyed in Carrol *et al.* (2006) but there are no results there on the treatment effect problem. In the econometrics literature the focus has mainly been on measurement error effects in the context of estimation of regression functions; see for example Hausman *et al.* (1991), Hausman *et al.* (1998), Li (2002), Hong and Tamer (2003), and Schennach (2004). Chesher (1991) gives results on the approximate effects of measurement error in a wide variety of contexts. Chesher and Schluter (2002) use these results in a study of the impact of measurement error on inequality and poverty measurement. This paper makes use of these results to study the impact of measurement error on treatment effect analysis. The strategy employed is briefly outlined in the next section.

1.1 The strategy

The notation employed in the *potential outcome* approach to causal inference is used throughout.[2] Y_1 and Y_0 are scalar random variables denoting the potential outcomes from respectively receiving and

[1] Heckman and Robb (1985) mention the problem very briefly.

[2] For reviews of the evaluation problem see Heckman *et al.* (1999), Heckman (2000) and Imbens (2004).

not receiving treatment. Binary $D \in \{0,1\}$ indicates treatment status, with $D = 1$ for treated units and $D = 0$ otherwise. X is a vector of characteristics which may be observed after contamination by measurement error U. The vector $Z \equiv X + \sigma U$ denotes the error contaminated X and $\sigma \geq 0$ is a scalar determining the magnitude of measurement error.

The treatment effect $\beta \equiv Y_1 - Y_0$ is not observable because realisations of Y_1, respectively Y_0, are only observable when D is 1, respectively 0, and so without further restriction causal parameters such as the ATE: $\beta_e \equiv E_{Y_1-Y_0}(Y_1 - Y_0)$, and the ATT: $\beta_t \equiv E_{Y_1-Y_0|D}(Y_1 - Y_0|1)$ are not identified.[3] Models incorporating the *strong ignorability* restriction by Rosenbaum and Rubin (1983) identify these causal parameters. This restriction comprises the conditional independence condition: $(Y_0, Y_1) \perp D|X$ and the support condition that for all x: $P(D = 1|X = x) \in (0,1)$. When this condition holds values of causal parameters can be uniquely determined from F_{YDX}, the joint distribution function of the observable outcome $Y \equiv DY_1 + (1 - D)Y_0$, D and X, through correspondences of the form $\theta = \mathcal{H}(F_{YDX})$ where θ denotes a causal parameter and $\mathcal{H}$ is a point identifying functional, that is a functional delivering a unique value.[4]

Various estimators follow on particular applications of the analogue principle. For example β_e could be estimated by calculating non- or semi-parametric estimates of the regressions of Y on X for the treated and of Y on X for the untreated and averaging differences in their predictions across the values of X. Other estimators, for example propensity score based procedures, are prompted by alternative ways of writing the identifying functional $\mathcal{H}$ and estimating its components using the

[3] The notation $E_{A|B}(g(A,B)|b)$ indicates the conditional expectation of $g(A,B)$ given $B = b$.

[4] There is for example the correspondence:

$$\beta_e = E_X\{E_{Y|DX}(Y|1,x) - E_{Y|DX}(Y|0,x)\},$$

for the ATE, and the correspondence:

$$\beta_t = E_{X|D}\{E_{Y|DX}(Y|1,x) - E_{Y|DX}(Y|0,x)|1\},$$

for the ATT (see Section 2).

properties of the propensity score (Rosenbaum and Rubin, 1983).

When realisations of measurement error contaminated Z are used in place of X, analogue estimators will typically estimate $\theta_Z \equiv \mathcal{H}(F_{YDZ})$ rather than the desired $\theta \equiv \mathcal{H}(F_{YDX})$, F_{YDZ} being the joint distribution function of Y, D and Z. In order to gain understanding of the impact of covariate measurement error on treatment effect estimators this paper studies the measurement error "bias": $\Delta \equiv \theta_Z - \theta$. The value of Δ depends on features of the distribution of Y, D and Z and a case by case analysis is required if exact results are to be obtained.[5] The focus here is on the generic effect of measurement error and information about this is obtained by considering its approximate effect as it comes to be a significant element, that is by considering the value of Δ when U has classical form and σ is small (note that when σ is zero there is no measurement error).

This paper contributes to the literatures on treatment effects and on measurement error by deriving two results of theoretical and practical relevance. First, we will show that under sufficient smoothness there is the approximation:

$$\Delta \equiv \mathcal{H}(F_{YDZ}) - \mathcal{H}(F_{YDX}) = \sigma^2 \mathcal{H}^*(F_{YDX}) + o(\sigma^2), \tag{1}$$

where the term indicated as $o(\sigma^2)$ has the property $\lim_{\sigma \to 0} o(\sigma^2)/\sigma^2 = 0$ and the functional $\mathcal{H}^*$ is *known*. Properties of $\mathcal{H}^*$ are explored to shed light on the "first order"impact of measurement error and the way in which this depends upon features of F_{YDX}. This characterization represents the first contribution of this paper. Second, we will show that, since F_{YDZ} differs from F_{YDX} by at most $O(\sigma)$, $\mathcal{H}^*(F_{YDX})$ can be replaced by $\mathcal{H}^*(F_{YDZ})$ in equation (1) without disturbing the order of the approximation error leading to the following approximation:

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

[5] Alternatively, under certain conditions on the distribution of the measurement error Δ could be obtained via simulation using SIMEX (see, for example, Carrol *et al.*, 2006).

Since the available data on Y, D and Z are informative about F_{YDZ}, one can estimate $\mathcal{H}^*(F_{YDZ})$ and so gain a view of the likely first order effect of measurement error at conjectured values of σ and calculate a range of "bias corrected"estimates for a range of plausible values of σ. This measurement error bias correction represents the second contribution of this paper.

1.2 Plan of the paper

The remainder of the paper is organised as follows. Section 2 provides alternative expressions for the causal parameters of interest in terms of the joint distribution of Y, D and X when the strong ignorability restriction is maintained. These motivate a variety of estimators of the parameters β_e and β_t. Section 3 sets out the measurement error model considered here and presents small-variance approximations to the "large sample measurement error bias" Δ. In Section 4 a procedure for assessing the potential impact of measurement error is proposed. Throughout the paper we study a particular, simple, case in which expressions for the exact and approximate effects of measurement error for the parameters β_e and β_t can be derived. These are helpful in setting the methods developed here in a familiar context. Section 5 builds upon this example and presents numerical calculations of the exact effects of measurement error and of the errors incurred using the approximations proposed here. In Section 6 an empirical application of the method proposed is presented to estimate the returns to educational qualification for the UK, while Section 7 concludes.

2 Identification in the absence of measurement error

This section sets out identifying correspondences for the ATE and the ATT when there is a strong ignorability restriction with respect to variables X, and X is observed *without* measurement error. This prepares the way for the study of the effect of measurement error.

Recall that the observable random variables are: the binary treatment status indicator, D, the

covariates X and the outcome $Y \equiv DY_1 + (1-D)Y_0$. Let $e_X(x)$ denote the *propensity score*:

$$e_X(x) \equiv P[D = 1|X = x],$$

that is the probability of receiving treatment conditional on having values of X equal to x. The *strong ignorability restriction* with respect to X, which we will refer to by I_X, comprises the following two conditions which hold for all values x:

$$(Y_0, Y_1) \perp D|X = x, \tag{2}$$

$$e_X(x) \in (0, 1). \tag{3}$$

The former condition states that potential outcomes are conditionally independent of the treatment status given observable characteristics X, whereas the latter condition ensures that treated are observed at all values x.[6] If I_X holds, then for all $(d, d') \in \{0, 1\}$ there is:

$$E_{Y_d|X}(Y_d|x) = E_{Y_d|DX}(Y_d|d', x),$$

so that by defining:

$$\Lambda^1_X \equiv \int \left(E_{Y|DX}(Y|1, x) - E_{Y|DX}(Y|0, x)\right) f_X(x)dx, \tag{4}$$

$$\Gamma^1_X \equiv \int \left(E_{Y|DX}(Y|1, x) - E_{Y|DX}(Y|0, x)\right) f_{X|D}(x|1)dx, \tag{5}$$

the following identifying correspondences hold:[7]

$$\Lambda^1_X = \beta_e,$$

$$\Gamma^1_X = \beta_t.$$

[6] Throughout this paper we will not consider the case of *conditional mean independence*, which is weaker than (2) and is sufficient to identify the ATE and the ATT. The key results on the effects of measurement error that are presented in what follows would however hold under suitably defined mean independence restrictions.

[7] Here and later integrals are definite over the full support of the variables of integration. Note also that the correspondence $\Gamma^1_X = \beta_t$ only requires $e_X(x) \in [0, 1)$ and $Y_0 \perp D|X = x$.

Analogue estimators of the ATE and the ATT can be obtained by considering the empirical counterparts of (4) and (5), respectively.

Alternative equivalent representations of Λ_X^1 and Γ_X^1 lead to alternative analogue estimators of β_e and β_t. For example, provided that the support condition in (3) holds, for all values x there is:

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

that can be used in (4) and (5) to write:

$$\begin{aligned} \Lambda_X^2 &= \int \left(\frac{E_{YD|X}(YD|x)}{e_X(x)} - \frac{E_{YD|X}(Y(1-D)|x)}{1-e_X(x)} \right) f_X(x)dx, \\ \Gamma_X^2 &= \int \left(\frac{E_{YD|X}(YD|x)}{P(D=1)} - \frac{E_{YD|X}(Y(1-D)|x)}{1-e_X(x)} \frac{e_X(x)}{P(D=1)} \right) f_X(x)dx, \end{aligned}$$

so that $\Lambda_X^2 = \Lambda_X^1$ and $\Gamma_X^2 = \Gamma_X^1$. An alternative representation can be obtained by using the *balancing property* of the propensity score. Theorem 3 by Rosenbaum and Rubin (1983) states that under the restriction I_X for all $(d, d') \in \{0,1\}$ there is:

$$E_{Y_d|e_X}(Y_d|\eta) = E_{Y_d|De_X}(Y_d|d',\eta),$$

that is if treatment assignment is strongly ignorable given x, then it is also strongly ignorable given the propensity score e_X(x). Thus by defining:

$$\begin{aligned} \Lambda_X^3 &\equiv \int \left(E_{Y|De_X}(Y|1,\eta) - E_{Y|De_X}(Y|0,\eta) \right) f_{e_X}(\eta)d\eta, \\ \Gamma_X^3 &\equiv \int \left(E_{Y|De_X}(Y|1,\eta) - E_{Y|De_X}(Y|0,\eta) \right) f_{e_X|D}(\eta|1)d\eta, \end{aligned}$$

and by using the fact that treatment assignment and covariates are conditionally independent given the propensity score (see Theorem 2 by Rosenbaum and Rubin, 1983) we also have $\Lambda_X^3 = \Lambda_X^1$ and $\Gamma_X^3 = \Gamma_X^1$.

Λ^1_X and Λ^3_X motivate "matching" estimators for the ATE which average differences of values of the outcome across the treated and untreated at common values of, respectively, the vector X and the scalar propensity score. Λ^2_X motivates estimators which difference weighted averages of treated and untreated outcomes. Similarly, Γ^1_X and Γ^3_X motivate "matching" estimators for the ATT, and Γ^2_X motivates weighted estimators for the ATT.[8] The three representations for the ATE and the ATT are equivalent provided that the support condition in (3) is satisfied, and the analogue estimators variously based on them will converge to the same limit but, unless I_X holds, that limit will not in general be the causal parameter of interest.

3 The effect of measurement error

We now study the effect of conditioning on measurement error affected covariates when their error free counterparts satisfy a strong ignorability restriction.

First, we show that if the strong ignorability restriction I_X holds when X is error-free, it does not hold when some elements of X are error contaminated. This result implies that there are measurement-error-induced inconsistencies in matching or similar type of estimators of the ATE or the ATT. Second, we characterise the bias induced by measurement error on the estimation of the ATE (Proposition 1) and of the ATT (Proposition 2). As explained in Section 3.2 the results rest upon the assumption of classical measurement error in covariates and are approximations designed to be accurate for small values of the measurement error variance.

For both the ATT and the ATE the impact of measurement error is small when a variable measured with error has little effect either on the outcomes or on the propensity score but in this situation the variable is of little help in identifying the ATT and ATE. When the variable susceptible to measurement

[8] See for example Horvitz and Thompson (1952), Rosenbaum (1987), Dehejia and Wahba (1995), Hahn (1998), and Hirano *et al.* (2003).

error has strong identifying power the impact of measurement error is greatest. The sign of the effects depends on the directions of the effect of the variable susceptible to measurement error on the propensity score and in the regressions of Y_0 and Y_1 on X. As a result it is not possible to sign the bias induced by measurement error without information about the case in hand. Estimates of the approximations we develop deliver information about the sign of measurement-error-induced bias and its magnitude at specified values of the measurement error variance.

3.1 The general problem

When error contaminated data are used, that is when realisations of Z are employed instead of realisations of X, the various analogue estimators of the ATE and the ATT presented in Section 2 can be regarded as estimators of the parameters obtained when, in the definitions above, the probability law for (Y, D, Z) is employed in place of the probability law for (Y, D, X).

Define the propensity score with respect to Z as:

$$e_Z(z) \equiv P[D = 1|Z = z],$$

and define Λ_Z^i, Γ_Z^i, $i \in \{1, 2, 3\}$, by analogy with the definitions given in the previous section, where for example there is:

$$\Lambda_Z^1 \equiv \int \left(E_{Y|DZ}(Y|1, z) - E_{Y|DZ}(Y|0, z)\right) f_Z(z)dz.$$

Using arguments similar to those employed above, if for all z we have $e_Z(z) \in (0, 1)$, then $\Lambda_Z^1 = \Lambda_Z^2 = \Lambda_Z^3 \equiv \Lambda_Z$ and $\Gamma_Z^1 = \Gamma_Z^2 = \Gamma_Z^3 \equiv \Gamma_Z$ (the proof of this result is reported in the Appendix). The various analogue estimators of the ATE and ATT using error contaminated Z in place of error free X will be consistent estimators of respectively Γ_Z and Λ_Z which will generally deviate from β_e and β_t. In what follows we give approximations to Λ_Z and Γ_Z, thereby shedding light on the inconsistency induced by the presence of measurement error.

3.2 The measurement error model

Particular attention is given to the case in which one scalar element, X_* of $X \equiv [X_*, X^*]$ is observed with error, the remaining elements X^* being observed without error.[9] The observable measurement error contaminated variable Z_* is defined as $Z_* = X_* + \sigma U$ where $E[U] = 0$, $Var[U]$ exists and is normalised to one, and U is distributed independently of (Y_0, Y_1, D, X).[10] The vector $Z \equiv (Z_*, X^*)$ is thus observable.

The variable measured with error and the measurement error are continuously distributed both with unbounded support. The approximation method of Chesher (1991) employed here requires this, and also requires that various moments of distributions, and derivatives of functions, up to third order exist. Most importantly, the continuous distribution and support restrictions ensure that the common support condition is satisfied for error contaminated Z if it is satisfied for error free X. If the distributions of X conditional on $D = 1$ and $D = 0$ share the same support and the distribution of U is independent of D, then the values x in the two groups are contaminated by realisations drawn from the same continuous distribution. It thus follows that the distributions of Z conditional on $D = 1$ and $D = 0$ must preserve the support property.

The approximations exploited below involve derivatives of distribution (F), density (f) and log

[9] Extension to cases with more than one variable observed with error is straightforward, but notationally more demanding.

[10] The independence restriction could be relaxed, for example by allowing the variance of measurement error to depend on D or X_*. This would introduce additional (unidentifiable) parameters whose values would need to be specified when investigating the potential impact of measurement error.

density (g) functions for which there is the following notation with $d \in \{0,1\}$ and $i \in \{1,2\}$:

$$\begin{aligned} F^{(i)}_{Y_d|X}(y|x) &\equiv \frac{\partial^i}{\partial x_*^i} F_{Y_d|X}(y|x), \\ f^{(i)}_{X|D}(x|d) &\equiv \frac{\partial^i}{\partial x_*^i} f_{X|D}(x|d), \\ f^{(i)}_{X}(x) &\equiv \frac{\partial^i}{\partial x_*^i} f_{X}(x), \\ g_{X|D}(x|d) &\equiv \frac{\partial}{\partial x_*} \log f_{X|D}(x|d) = \frac{f^{(1)}_{X|D}(x|d)}{f_{X|D}(x|d)}. \end{aligned}$$

Note that all partial derivatives are with respect to the variable subject to measurement error, that the distribution function derivatives, unlike the density function derivatives, are with respect to a conditioning argument, and that the function $g_{X|D}(x|d)$ can be written as the X_*-derivative of the log *conditional* density of X_* given D and X^*.

3.3 Example

This example is simple and convenient, in that it allows us to derive *analytical* expressions for the *exact* bias and the *approximations* discussed below. We will return to the case dealt with here to set the methods presented in a familiar context. Moreover, the example considered allows us to establish a link with the results by Cochran and Rubin (1973). To ease notation, suppose that X consists of just *one* variable and that the regression functions of Y on X for the groups $D = 0$ and $D = 1$ are *linear*, as follows:

$$\begin{aligned} E_{Y_0|DX}(Y_0|0,x) &= \alpha_0 + \gamma_0 x, \\ E_{Y_1|DX}(Y_1|1,x) &= \alpha_1 + \gamma_1 x. \end{aligned}$$

The extension to the case of multidimensional X and (or) polynomial regression functions proceeds along the same lines, but is notationally more demanding.[11] Assume that Z is observed in place of

[11] Cochran and Rubin (1973) consider the case of one continuous covariate X and regressions of Y on X linear and parallel for the population of the treated and the untreated (i.e. $\gamma_0 = \gamma_1$).

X and that, conditional on D, (X, U) are *normally* distributed random variables with group specific parameters, implying that for $d \in \{0, 1\}$ we have:

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

and:

$$E_{X|DZ}(X|d, z) = \mu_d + \frac{\lambda_d^2}{\lambda_d^2 + \sigma^2} (z - \mu_d).$$

Define $p \equiv P[D = 1]$. The following expressions follow under I_X when X is measured *without* error.

$$\begin{aligned} E_{Y_0}(Y_0) &= \alpha_0 + \gamma_0[(1-p)\mu_0 + p\mu_1], \\ E_{Y_1}(Y_1) &= \alpha_1 + \gamma_1[(1-p)\mu_0 + p\mu_1], \\ E_{Y_0|D}(Y_0|1) &= \alpha_0 + \gamma_0\mu_1. \end{aligned}$$

These lead to the following expressions for the ATE and the ATT.

$$\begin{aligned} \beta_e &= (\alpha_1 - \alpha_0) + (\gamma_1 - \gamma_0)\left((1-p)\mu_0 + p\mu_1\right), \\ \beta_t &= (\alpha_1 - \alpha_0) + (\gamma_1 - \gamma_0)\mu_1. \end{aligned}$$

3.4 The effect on the strong ignorability condition

First consider the effect of measurement error on the strong ignorability condition. The general question we will provide an answer to is the following: *does strong ignorability with respect to X ensures strong ignorability with respect to error ridden covariates Z?* Applying the approximation of Chesher (1991) for conditional distribution functions with covariates measured with error gives the following approximation for the distribution functions of potential outcomes Y_d conditional on D and Z:

$$F_{Y_d|DZ}(y|d', z) \simeq F_{Y_d|X}(y|z) + \sigma^2 F^{(1)}_{Y_d|X}(y|z) g_{X|D}(z|d') + \frac{\sigma^2}{2} F^{(2)}_{Y_d|X}(y|z), \tag{6}$$

where $(d, d') \in \{0, 1\}$ and, here and later, $A \simeq B$ indicates $A = B + o(\sigma^2)$. Strong ignorability with respect to X has been exploited in producing this approximation. The first term on the right hand side of (6) is the conditional distribution of Y_d given X *and* D evaluated at $X = z$ which, by virtue of the strong ignorability restriction, is equal to the conditional distribution of Y_d given X alone evaluated at $X = z$. Strong ignorability also makes conditioning on D irrelevant in the first and second derivative terms.

Dependence on D can arise through the second term of the approximation to $F_{Y_d|DZ}$ because of the appearance of the log density derivative $g_{X|D}$. It follows that, by considering only terms of order $o(\sigma^2)$, there is *not* strong ignorability with respect to Z local to $\sigma^2 = 0$ unless for all z and y we have:

$$F^{(1)}_{Y_d|X}(y|z) \left[g_{X|D}(z|1) - g_{X|D}(z|0) \right] = 0, \qquad d \in \{0, 1\}$$

for which a sufficient condition is that either:

$$F^{(1)}_{Y_d|X}(y|z) = 0, \qquad d \in \{0, 1\} \tag{7}$$

for all z and y, or:

$$g_{X|D}(z|1) = g_{X|D}(z|0), \tag{8}$$

for all z.

The condition (7) virtually requires potential outcomes to be independent of X_*, while the condition (8) requires X_* to be independent of D in which case the propensity score does not depend on X_*. In neither case is X_* influential in identifying the causal parameters of interest. We conclude that identifying power vested in a variable X_* by virtue of a strong ignorability restriction is lost when X_* is measured with error of the simple form studied here.

3.5 The effect on regression and density functions

Now consider the effect of measurement error on the regressions of Y_0 and Y_1 on D and X which appear in the identifying correspondences for the ATE and ATT. The same approach that leads to (6) gives the following approximation for the regression functions of Y_d, $d \in \{0,1\}$, when conditioning is on error contaminated Z rather than error free X:

$$E_{Y_d|DZ}(Y_d|d',z) \simeq E_{Y_d|X}(Y_d|z) + \sigma^2 E^{(1)}_{Y_d|X}(Y_d|z) g_{X|D}(z|d') + \frac{\sigma^2}{2} E^{(2)}_{Y_d|X}(Y_d|z). \tag{9}$$

Here $d' \in \{0,1\}$, and the terms $E^{(1)}$ and $E^{(2)}$ are first and second derivatives of the regression functions with respect to X_*, as follows:

$$E^{(i)}_{Y_d|X}(Y_d|x) \equiv \frac{\partial^i}{\partial x_*^i} E_{Y_d|X}(Y_d|x), \qquad d \in \{0,1\}, \quad i \in \{1,2\}.$$

The second term in (9) captures local *attenuation* effects of measurement error while the third term captures the local *smoothing* induced by measurement error (Chesher, 1991). The strong ignorability restriction pertaining to error free X has been exploited in producing this approximation, removing dependence on D from the conditional expectation and its two derivatives on the right hand side of (9). Outside the sort of special cases discussed in the previous section, the outcomes Y_0 and Y_1 are locally *dependent* on D given error contaminated Z even though they are *independent* of D given error free X, this dependence arising *via* the log density derivative $g_{X|D}$.[12]

The marginal density function of X and its conditional density given D also appear in the identifying correspondences for the ATT and ATE set out in Section 2 and they always differ from their

[12] Along the lines of what discussed in the previous section, it thus follows that the conditioning on error ridden variables also invalidates causal inference based on mean independence restrictions.

counterparts involving Z. Define the following second partial derivatives of density functions:

$$\begin{aligned} f_X^{(2)}(x) &\equiv \frac{\partial^2}{\partial x_*^2} f_X(x), \\ f_{X|D}^{(2)}(x|d) &\equiv \frac{\partial^2}{\partial x_*^2} f_{X|D}(x|d), \quad d \in \{0,1\}. \end{aligned}$$

There are the following approximations (see Chesher, 1991) which capture the general spreading effect of measurement error, lowering (raising) the density functions where they are concave (convex):

$$\begin{aligned} f_Z(z) &\simeq f_X(z) + \frac{\sigma^2}{2} f_X^{(2)}(z), \\ f_{Z|D}(z|d) &\simeq f_{X|D}(z|d) + \frac{\sigma^2}{2} f_{X|D}^{(2)}(z|d), \quad d \in \{0,1\}. \end{aligned}$$

It is clear that measurement error in conditioning variables perturbs the identifying correspondences for the ATE and ATT which lie at the heart of matching and other procedures built on the foundation of a strong ignorability restriction. In general analogue estimation using error contaminated Z instead of error free X will result in inconsistent estimation. The magnitude of, and the influences on this inconsistency are studied in the next section using the approximations derived for the regressions and for the density functions.

3.6 The effect on the identification of the causal parameters of interest

This section gives approximations to $\Gamma_Z - \beta_e$ and $\Lambda_Z - \beta_t$. To this end we develop approximations to the following objects:

$$\begin{aligned} \Delta_0 &\equiv \int E_{Y|DZ}(Y|0,z) f_Z(z) dz - E_{Y_0}[Y_0], \\ \Delta_1 &\equiv \int E_{Y|DZ}(Y|1,z) f_Z(z) dz - E_{Y_1}[Y_1], \\ \Delta_{0|1} &\equiv \int E_{Y|DZ}(Y|0,z) f_{Z|D}(z|1) dz - E_{Y_0|D}[Y_0|1], \end{aligned}$$

from which the desired results follow on noting that for the ATE: $\Lambda_Z - \beta_e = \Delta_1 - \Delta_0$ and for the ATT: $\Gamma_Z - \beta_t = -\Delta_{0|1}$. The approximations involve the first partial derivative of the propensity score

with respect to error contaminated X_*:

$$e_X^{(1)}(x) \equiv \frac{\partial}{\partial x_*} e_X(x).$$

Proposition 1 and 2 give results for the ATE and ATT, respectively.

Proposition 1 ***(Effects on the ATE)*** *If (i) there is the strong ignorability restriction I_X and (ii) for $d \in \{0,1\}$ the following conditions hold:*

$$\lim_{z \to \pm\infty} E_{Y_d|X}(Y_d|z) f_X^{(1)}(z) = 0, \qquad \lim_{z \to \pm\infty} E_{Y_d|X}^{(1)}(Y_d|z) f_X(z) = 0, \tag{10}$$

then:

$$\begin{aligned} \Delta_0 &\simeq -\sigma^2 \int E_{Y_0|X}^{(1)}(Y_0|z) \frac{e_X^{(1)}(z)}{1 - e_X(z)} f_X(z) dz, \\ \Delta_1 &\simeq \sigma^2 \int E_{Y_1|X}^{(1)}(Y_1|z) \frac{e_X^{(1)}(z)}{e_X(z)} f_X(z) dz, \end{aligned}$$

so that:

$$\Lambda_Z - \beta_e \simeq \sigma^2 \int \left[\frac{E_{Y_1|X}^{(1)}(Y_1|z)}{e_X(z)} + \frac{E_{Y_0|X}^{(1)}(Y_0|z)}{1 - e_X(z)} \right] e_X^{(1)}(z) f_X(z) dz.$$

■

The proof of the proposition is reported in the Appendix. The approximations are obtained by substituting in the expressions for Δ_0 and Δ_1 the approximations to the regression functions $E_{Y|DZ}(Y|0,z)$ and $E_{Y|DZ}(Y|1,z)$ and the approximation to the density function $f_Z(z)$ given in Section 3.5, deleting terms of order $o(\sigma^2)$ and integrating with respect to z, exploiting the conditions (10) which place restrictions on the large X behaviour of the regression functions and the tail behaviour of the density of X.[13]

[13] These conditions will be satisfied if, for example, the regression function is a polynomial in Z and the tails of the density function decrease at an exponential rate (see Chesher, 1991). The same conditions are exploited in Proposition 2.

The measurement error inconsistency is larger the greater is the sensitivity of the propensity score and the regression functions of Y_0 and Y_1 on X to changes in X_*, the variable susceptible to measurement error, and of course the larger is the measurement error variance.

Proposition 2 ***(Effects on the ATT)*** *If (i) there is the strong ignorability restriction I_X and (ii) the following condition holds:*

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

then:

$$\Gamma_Z - \beta_t = -\Delta_{0|1} \simeq \sigma^2 \int E^{(1)}_{Y_0|X}(Y_0|z) \frac{e^{(1)}_X(z)}{e_X(z)[1 - e_X(z)]} f_{X|D}(z|1) dz.$$

■

The proof, which proceeds along the same lines as the proof of Proposition 1, is reported in the Appendix. It follows that the measurement error induced inconsistency of the ATT is larger the greater is the sensitivity of the propensity score and the regression of Y_0 on X to changes in X_*. Note that here and before the sign of the bias depends on the particular application at hand. However, it is worth noting that the following rule of thumb for the ATT can be derived that may turn out useful for empirical applications. If the propensity score follows a logit model and the regression of Y_0 on X is approximatively linear in X_*, then the approximation in Proposition 2 simplifies to $\sigma^2 \gamma_* \delta_*$ (see the Appendix), where γ_* is the coefficient of X_* in the regression of Y_0 on X and δ_* is the coefficient of X_* in the propensity score. This result may be helpful in signing the bias on the ATT induced by measurement error. Note also that an alternative expression for the approximation to the bias $\Gamma_Z - \beta_t$ is the following (see the Appendix):

$$\frac{\sigma^2}{P[D=1]} \int E^{(1)}_{Y_0|X}(Y_0|z) \frac{e^{(1)}_X(z)}{1 - e_X(z)} f_X(z) dz,$$

which can be obtained by applying the Bayes' theorem to $f_{X|D}(x|1)$, thus implying that the approximation to $\Gamma_Z - \beta_t$ equals the approximation to Δ_0 times $-\frac{1}{P[D=1]}$ (see Proposition 1).

3.7 Example (continued)

3.7.1 Exact expression for the bias induced by errors in covariates

Note that for the regression functions of Y on Z there is, for $d \in \{0, 1\}$:

$$\begin{aligned} E_{Y|DZ}(Y|d,z) &= \int E_{Y|DX}(Y|d,x) f_{X|DZ}(x|d,z)dx, \\ &= \alpha_d + \gamma_d E_{X|DZ}(X|d,z), \\ &= \alpha_d + \gamma_d \mu_d + \frac{\gamma_d \lambda_d^2}{\lambda_d^2 + \sigma^2}(z - \mu_d), \end{aligned}$$

which exhibits the usual attenuation. We therefore have the following expressions for the *exact* bias introduced by measurement error:

$$\begin{aligned} \Delta_0 &= -p\gamma_0(\mu_1 - \mu_0)\frac{\sigma^2}{\lambda_0^2 + \sigma^2}, \\ \Delta_1 &= (1-p)\gamma_1(\mu_1 - \mu_0)\frac{\sigma^2}{\lambda_1^2 + \sigma^2}, \\ \Delta_{0|1} &= -\gamma_0(\mu_1 - \mu_0)\frac{\sigma^2}{\lambda_0^2 + \sigma^2}. \end{aligned}$$

Note that all these expressions are zero when $\sigma^2 = 0$, which is as it should be. The expression for the bias induced by measurement error in Cochran and Rubin (1973, page 431) corresponds to the difference between Δ_1 and Δ_0 imposing classical measurement error and parallel regressions of Y on X for the population of the treated and the untreated.

3.7.2 Approximation to the bias using Propositions 1 and 2

We now derive an approximation to the bias in the example. This depends on functionals of the *unobserved* variable X and we use results from the previous section. First, note that under the distributional assumptions made in this example the regularity conditions in Proposition 1 and Proposition

2 are met, implying that the approximations for Δ_0, Δ_1 and $\Delta_{0|1}$ can be re-arranged to get:[14]

$$\Delta_i \simeq \sigma^2 \int \left[E^{(1)}_{Y_i|D,X}(Y_i|i,z) g_{X|D}(z|i) + E^{(2)}_{Y_i|D,X}(Y_i|i,z) \right] f_X(z) dz,$$

for $i \in \{0,1\}$, and:

$$\Delta_{0|1} \simeq \sigma^2 \int \left[E^{(1)}_{Y_0|D,X}(Y_0|0,z) g_{X|D}(z|0) + E^{(2)}_{Y_0|D,X}(Y_0|0,z) \right] f_{X|D}(z|1) dz. \tag{11}$$

Using this result and the linearity of the regression functions, it also follows that:

$$\begin{aligned} \Delta_i &\simeq \sigma^2 \gamma_i \int g_{X|D}(z|i) f_X(z) dz, \qquad i \in \{0,1\} \\ \Delta_{0|1} &\simeq \sigma^2 \gamma_0 \int g_{X|D}(z|0) f_{X|D}(z|1) dz. \end{aligned}$$

These expressions can be used to study the *accuracy* of the approximation, that is the ratio between either of the approximations to Δ_0, Δ_1 or $\Delta_{0|1}$ and the exact value of the bias that has been derived for this example. Since by using the properties of the normal distribution we have:

$$g_{X|D}(x|i) \equiv \frac{f^{(1)}_{X|D}(x|i)}{f_{X|D}(x|i)} = -\frac{x - \mu_i}{\lambda_i^2}, \qquad i \in \{0,1\}$$

the approximations to the bias derived in Proposition 1 and Proposition 2 can be written as:

$$\begin{aligned} \Delta_0 &\simeq -p\gamma_0(\mu_1 - \mu_0)\frac{\sigma^2}{\lambda_0^2}, \\ \Delta_1 &\simeq (1-p)\gamma_1(\mu_1 - \mu_0)\frac{\sigma^2}{\lambda_1^2}, \\ \Delta_{0|1} &\simeq -\gamma_0(\mu_1 - \mu_0)\frac{\sigma^2}{\lambda_0^2}. \end{aligned}$$

[14] The use of probit or logit specifications for the propensity score would require numerical integration for some of the steps that follow. For this reason, in the remainder of this example we will use alternative expressions for the bias in Proposition 1 and Proposition 2 that corresponds the approximations (16) and (17) in the Appendix.

It is worth noting that the approximation error is of order $O(\sigma^4)$, as the symmetric distribution of U causes $O(\sigma^3)$ terms to disappear (see Chesher, 1991). Moreover, we have that the terms:

$$\begin{aligned}
\frac{-p\gamma_0(\mu_1-\mu_0)\frac{\sigma^2}{\lambda_0^2}}{\Delta_0} &= 1+\frac{\sigma^2}{\lambda_0^2}, \\
\frac{(1-p)\gamma_1(\mu_1-\mu_0)\frac{\sigma^2}{\lambda_1^2}}{\Delta_1} &= 1+\frac{\sigma^2}{\lambda_1^2}, \\
\frac{-\gamma_0(\mu_1-\mu_0)\frac{\sigma^2}{\lambda_0^2}}{\Delta_{0|1}} &= 1+\frac{\sigma^2}{\lambda_0^2},
\end{aligned}$$

represent the accuracy of the approximations to Δ_0, Δ_1 and $\Delta_{0|1}$, respectively, which depends on the noise-to-signal ratio.

4 Accounting for measurement error

In this section 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, reducing the order of bias to terms which are $o(\sigma^2)$. Our strategy uses quantities constructed from non-parametric estimates of functionals of the probability law (Y, D, Z), and thus exploits *nothing but* the error contaminated data without requiring any functional assumptions on the regression of Y on D and X nor additional information (such as instrumental variables or validation data).[15] In Section 4.1 we show how the measurement error bias can be reduced using the approach suggested by Chesher and Schluter (2002). The general idea is then applied within the context of the running example in Section 4.2.

4.1 Measurement error bias correction

Since X can be replaced by Z in expressions multiplied by σ^2 without altering the order of the approximation error (see Chesher and Schluter, 2002), we can modify expressions in Proposition 1

[15] 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.*, 1998). 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.

and Proposition 2 to get:

$$\Delta_0 \simeq -\sigma^2 \int E^{(1)}_{Y_0|DZ}(Y_0|0,z)\frac{e^{(1)}_Z(z)}{1-e_Z(z)}f_Z(z)dz, \tag{12}$$

$$\Delta_1 \simeq \sigma^2 \int E^{(1)}_{Y_1|DZ}(Y_1|1,z)\frac{e^{(1)}_Z(z)}{e_Z(z)}f_Z(z)dz, \tag{13}$$

where for $i \in \{0,1\}$ and $j \in \{1,2\}$:

$$\begin{aligned} E^{(j)}_{Y_i|DZ}(Y_i|i,z) &\equiv \frac{\partial}{\partial z_*^j}E_{Y_i|DZ}(Y_i|i,z), \\ e^{(j)}_Z(z) &\equiv \frac{\partial}{\partial z_*^j}e_Z(z), \end{aligned}$$

and the approximation to $\Delta_{0|1}$ is obtained by multiplying the approximation to Δ_0 by $-\frac{1}{P[D=1]}$.

It follows that, for *known* values of the measurement error variance σ^2, the quantities above are identified from observed data neglecting terms which are $o(\sigma^2)$. This implies that approximately corrected causal effects can be obtained by estimating the bias for the ATE (i.e. $\Delta_1 - \Delta_0$) and the bias for the ATT (i.e. $-\Delta_{0|1}$) for σ^2 passing through a range of plausible values. Note that these approximations require knowledge of the first derivative of the regressions $E_{Y_1|DZ}(Y_0|0,z)$ and $E_{Y_0|DZ}(Y_0|0,z)$ as well as of the propensity score $e_Z(z)$. The quantities above are weighted averages (over the entire population or over the population of the treated) of first derivatives of regression functions, where weights are defined from the propensity score. The approximate impact of measurement error can be estimated for any candidate value of the measurement error variance using only error contaminated data, thus allowing investigation of the sensitivity of the causal parameter of interest to the presence of measurement error.[16]

[16] As the propensity score is a conditional expectation for which the approximation discussed in Section 3.5 applies, using the same approach as in Chesher and Schluter (2002) there is:

$$\tilde{e}_Z(z) \equiv e_Z(z) - \sigma^2 e^{(1)}_Z(z)g_Z(z) - \frac{\sigma^2}{2}e^{(2)}_Z(z),$$

where:

$$f^{(1)}_Z(z) \equiv \frac{\partial}{\partial z_*^i}f_Z(z), \qquad g_Z(z) \equiv \frac{\partial}{\partial z_*}lnf_Z(z) = \frac{f^{(1)}_Z(z)}{f_Z(z)}.$$

When non-parametric estimation is feasible, one could estimate derivatives with respect to Z_* of the regression of Y on Z for non-participants ($D = 0$) and for participants ($D = 1$) by local polynomials (see Fan and Gijbels, 1996). Alternatively, one could specify a fairly flexible parametric model for $E_{Y|DZ}(Y|i,z)$, for $i \in \{0,1\}$, from which the required derivatives are easily obtained. A similar argument applies to the estimation of the propensity score, though having a parametric model (e.g. a logit model) for the regression of D on Z can be rather convenient.

4.2 Example (continued)

The correction consists of two steps. First, the following approximations to the bias introduced by measurement error are considered:

$$\Delta_i \simeq \sigma^2 \int \left[E^{(1)}_{Y_i|DZ}(Y_i|i,z) g_{Z|D}(z|i) + E^{(2)}_{Y_i|DZ}(Y_i|i,z) \right] f_Z(z) dz,$$

for $i \in \{0,1\}$, and:

$$\Delta_{0|1} \simeq \sigma^2 \int \left[E^{(1)}_{Y_0|DZ}(Y_0|0,z) g_{Z|D}(z|0) + E^{(2)}_{Y_0|DZ}(Y_0|0,z) \right] f_{Z|D}(z|1) dz, \tag{14}$$

which are obtained by replacing X with Z in equations (16) and (17) in the Appendix. Second, the terms on the right hand side of these expressions, which are identified from observable variables, are subtracted from the expressions for $E_{Y_0}(Y_0)$, $E_{Y_1}(Y_1)$ and $E_{Y_0|D}(Y_0|1)$ obtained from raw data. The resulting expressions are still different from the quantities of interest, but such difference comes from terms which are of order $o(\sigma^4)$.

Since for $i \in \{0,1\}$ we have:

$$E^{(1)}_{Y_i|DZ}(Y_i|i,z) = \frac{\gamma_i \lambda_i^2}{\lambda_i^2 + \sigma^2}, \qquad E^{(2)}_{Y_i|DZ}(Y_i|0,z) = 0,$$

This result suggests that one could match treated to the untreated with respect to values of the pseudo propensity score $\tilde{e}_Z(z)$ to reduce the order of measurement error bias. We did not pursue further this idea, as its proof would involve stochastic expansions rather than moment expansions as those in Chesher (1991).

$$g_{Z|D}(z|i) \equiv \frac{f^{(1)}_{Z|D}(z|i)}{f_{Z|D}(z|i)} = -\frac{z-\mu_i}{\lambda_i^2+\sigma^2},$$

the approximation to the bias based on functionals of the observed variable Z results in the following expressions:

$$\Delta_0 \simeq -p\gamma_0(\mu_1-\mu_0)\frac{\sigma^2\lambda_0^2}{(\lambda_0^2+\sigma^2)^2}, \qquad \Delta_1 \simeq (1-p)\gamma_1(\mu_1-\mu_0)\frac{\sigma^2\lambda_1^2}{(\lambda_1^2+\sigma^2)^2},$$

$$\Delta_{0|1} \simeq -\gamma_0(\mu_1-\mu_0)\frac{\sigma^2\lambda_0^2}{(\lambda_0^2+\sigma^2)^2}.$$

It follows that, after our correction procedure, the $O(\sigma^2)$ biases generated by measurement error – whose exact expressions have been derived above – are replaced by $O(\sigma^4)$ biases as follows:

$$\text{for } E_{Y_0}(Y_0): \quad -p\gamma_0(\mu_1-\mu_0)\left(\frac{\sigma^2}{\lambda_0^2+\sigma^2}\right)^2,$$

$$\text{for } E_{Y_0}(Y_0): \quad (1-p)\gamma_1(\mu_1-\mu_0)\left(\frac{\sigma^2}{\lambda_1^2+\sigma^2}\right)^2,$$

$$\text{for } E_{Y_0|D}(Y_0|1): \quad -\gamma_0(\mu_1-\mu_0)\left(\frac{\sigma^2}{\lambda_0^2+\sigma^2}\right)^2.$$

5 Exact calculations

This section reports exact calculations designed to investigate the accuracy of the approximations proposed. We have already studied the exact values of the bias and of the approximation to such bias in the fully Gaussian case by deriving their analytical expressions within the context of the leading example that we considered throughout the paper. These expressions depend on features of the joint distribution of the covariates X and of the measurement error U, as well as on assumptions on the conditional expectation of Y given D and X, and cannot be solved analytically in general. In what follows we present results from numerical calculations obtained for the simple setup maintained in the leading example allowing for departures from normality for both the distribution of X and the distribution of U. The general setup is described in Section 5.1 and the results of the exercise are reported in Section 5.2.

5.1 Set up

To ease calculations parallel regressions of Y of X as in Cochran and Rubin (1973) are considered, that is we set:

$$\begin{aligned} E_{Y_1|DX}(Y_1|1,x) &= \alpha_1 + \gamma x, \\ E_{Y_0|DX}(Y_0|0,x) &= \alpha_0 + \gamma x, \end{aligned}$$

for the treated and for the untreated, respectively, with $\gamma = 1$. Under this specification the average outcome difference between the treated and the untreated does not change with X, so that the ATE and the ATT coincide and are equal to $\alpha_1 - \alpha_0$. Exact and approximate biases are invariant with respect to $\alpha_1 - \alpha_0$.

We obtain the *exact* value of the bias for the ATT:[17]

$$\Gamma_Z - \beta_t = \mu_1 - \int\int x f_{X|DZ}(x|0,z) f_{Z|D}(z|1) dx dz, \tag{15}$$

μ_1 being the mean of X for the treated. To this end, we use the exponential power (EP) family of distributions (see Box and Tiao, 1973) to model the distribution of X given D and the distribution of U, from which the density functions $f_{X|DZ}(x|0,z)$ and $f_{Z|D}(z|1)$ in (15) can be obtained. The EP family is a three parameter family of symmetric distributions. Its density function will be denoted by $EP(\mu, \lambda, \zeta)$. It has mean μ, variance λ^2 and $\zeta \in (-1, 1)$ is a shape parameter. Setting $\zeta = +1$ yields a Laplace (double exponential), high tailed density. Setting $\zeta = 0$ yields a normal density. Setting $\zeta = -1$ there is a uniform density on $(\mu - \sqrt{3}\lambda, \mu + \sqrt{3}\lambda)$.[18]

We assume that $X|D = d$ and U are distributed according to an $EP(\mu_d, \lambda_d, \zeta_x)$ distribution and

[17] The same numerical results are obtained using the expressions for the ATT and the ATE which have the same value in this example.

[18] Let $A \in (-\infty, \infty)$ have an exponential power distribution with parameters (μ, λ, ζ), $\zeta \in (-1, 1)$. Then the density function of A is as follows:

$$f_A(a) \equiv G \exp\left(-H \left|\frac{a-\mu}{\lambda}\right|^{\frac{2}{1+\zeta}}\right),$$

an $EP(0,1,\zeta_u)$ distribution, respectively, and we set $\mu_0 = 0$ and $\lambda_1^2 = \lambda_0^2 = 1$. The analytic expression for (15) in the $\zeta_x = \zeta_u = 0$ (everything normal) case have been already discussed in the example described above. In all remaining cases numerical integration is required, and accuracy of numerical computations can be checked against the exact result obtained for the $\zeta_x = \zeta_u = 0$ case.

The (infeasible) approximation to (15) based on the result in Proposition 2 is obtained using (11), so that there is:[19]

$$\Gamma_Z - \beta_t \simeq \sigma^2 \int g_{X|D}(x|0) f_{X|D}(x|1) dx.$$

Similarly, we use (14) to compute the value of the approximation to (15) from raw data.

5.2 Results

Results are reported in Table 1 by row for different values of the parameter μ_1 ($\mu_1 = 1, 2, 3$), and by column for increasing values of the measurement error variance, expressed in the table as percentages (10%, 20% and 30%) relative to the variance of error free X.[20] The calculations are reported for different combinations of the shape parameters (ζ_x, ζ_u) and, since there are analytic results for the fully Gaussian case, we checked that numerical computations can reproduce those results for the $\zeta_x = \zeta_u = 0$ case. Since all the exact and approximate biases in the cases considered are negative we report their absolute values to improve the readability of the table. For the example considered,

where:

$$\begin{aligned} G &\equiv \frac{1}{\lambda} \times \frac{\Gamma(3(1+\zeta)/2)^{1/2}}{(1+\zeta)\Gamma((1+\zeta)/2)^{3/2}}, \\ H &\equiv \left(\frac{\Gamma(3(1+\zeta)/2)}{\Gamma((1+\zeta)/2)} \right)^{\frac{1}{1+\zeta}}. \end{aligned}$$

[19]The following expression involves a straightforward one dimensional numerical integration after noting that, if A has an $EP(\mu, \lambda, \zeta)$ distribution, the first derivative of its log density is:

$$-sign(a-\mu) \frac{2}{\lambda(1+\zeta)} H \left| \frac{a-\mu}{\lambda} \right|^{\frac{1-\zeta}{1+\zeta}}.$$

All calculations were done using Mathematica 7.0, Wolfram Research Inc., (2008). More details on the computational aspects of our exercise, on the accuracy of the computations as well as the Mathematica programmes that we used are available upon request.

[20]For example since $Var(X|D=d) = 1$ in the example for $d \in \{0,1\}$ in the columns headed 10% the measurement error variance is 0.1.

measurement error results in *under*-estimation of the ATT and ATE.

The measurement error bias can be substantial. For larger values of the measurement error variance the bias on the ATT increases but not quite as fast as linearly and this pattern is robust to characteristics of distributions X and U. The feasible approximation is generally closer in absolute terms to the exact value of the bias than the infeasible approximation even for large values of the measurement error variance. However, in many cases the former approximation slightly understates and the latter approximation slightly overstates the magnitude of the bias. As expected, the accuracy of both approximations tends to worsen as the measurement error variance increases. In most cases both approximations are quite accurate. The only really poor cases are in the top block of the table in which the error free covariate X has a relatively high tailed distribution ($\zeta_x = -0.5$).

6 Empirical application

In this section we present an application of the measurement error correction procedure to the estimation of the returns to educational qualifications in the UK using data from the National Child Development Survey (NCDS) and building on previous work by Blundell *et al.* (2005). As causal effects are defined as the difference between the outcome following from the realisation of a certain state of the world and the counterfactual outcome that would have resulted had the state been different, the assessment of differences in earnings arising from alternative educational choices fits well in the causal framework.

We will maintain the assumption that the information available from the NCDS is enough to correct for ability and omitted variables bias (see the discussion in Blundell *et al.*, 2005). This consists of individuals' gender, age and ethnicity, family background, mother's and father's age and education, father's social class, mother's employment status and number of siblings at age 16. Most importantly,

Table 1: Exact calculations for the example considered

		error variance σ^2			error variance σ^2			error variance σ^2		
		10%	20%	30%	10%	20%	30%	10%	20%	30%
	μ_1	$\zeta_x = -0.5, \zeta_u = -0.5$			$\zeta_x = -0.5, \zeta_u = 0$			$\zeta_x = -0.5, \zeta_u = +0.5$		
exact value of the bias	1	0.118	0.194	0.254	0.126	0.206	0.268	0.136	0.218	0.279
approximation using (Y, D, X)	1	0.183	0.366	0.548	0.183	0.366	0.548	0.183	0.366	0.548
approximation using (Y, D, Z)	1	0.107	0.164	0.202	0.095	0.138	0.167	0.084	0.123	0.150
exact value of the bias	2	0.280	0.434	0.549	0.344	0.515	0.636	0.418	0.591	0.708
approximation using (Y, D, X)	2	0.640	1.280	1.919	0.640	1.280	1.919	0.640	1.280	1.919
approximation using (Y, D, Z)	2	0.325	0.468	0.553	0.216	0.280	0.317	0.140	0.186	0.220
exact value of the bias	3	0.483	0.715	0.882	0.693	0.967	1.142	0.922	1.181	1.341
approximation using (Y, D, X)	3	1.645	3.290	4.935	1.645	3.290	4.935	1.645	3.290	4.935
approximation using (Y, D, Z)	3	0.783	1.088	1.245	0.365	0.420	0.446	0.155	0.186	0.210
	μ_1	$\zeta_x = 0, \zeta_u = -0.5$			$\zeta_x = 0, \zeta_u = 0$			$\zeta_x = 0, \zeta_u = +0.5$		
exact value of the bias	1	0.090	0.164	0.226	0.091	0.167	0.231	0.092	0.170	0.236
approximation using (Y, D, X)	1	0.100	0.200	0.300	0.100	0.200	0.300	0.100	0.200	0.300
approximation using (Y, D, Z)	1	0.083	0.141	0.182	0.083	0.139	0.178	0.082	0.137	0.173
exact value of the bias	2	0.175	0.314	0.429	0.182	0.333	0.462	0.193	0.363	0.507
approximation using (Y, D, X)	2	0.200	0.400	0.600	0.200	0.400	0.600	0.200	0.400	0.600
approximation using (Y, D, Z)	2	0.168	0.291	0.384	0.165	0.278	0.355	0.161	0.259	0.317
exact value of the bias	3	0.251	0.441	0.597	0.273	0.500	0.692	0.311	0.601	0.842
approximation using (Y, D, X)	3	0.300	0.600	0.900	0.300	0.600	0.900	0.300	0.600	0.900
approximation using (Y, D, Z)	3	0.256	0.453	0.614	0.248	0.417	0.533	0.234	0.353	0.410
	μ_1	$\zeta_x = +0.5, \zeta_u = -0.5$			$\zeta_x = +0.5, \zeta_u = 0$			$\zeta_x = +0.5, \zeta_u = +0.5$		
exact value of the bias	1	0.090	0.164	0.226	0.090	0.165	0.228	0.090	0.166	0.230
approximation using (Y, D, X)	1	0.099	0.197	0.295	0.099	0.197	0.295	0.099	0.197	0.295
approximation using (Y, D, Z)	1	0.082	0.138	0.295	0.082	0.137	0.176	0.082	0.137	0.175
exact value of the bias	2	0.142	0.265	0.375	0.144	0.275	0.392	0.148	0.288	0.416
approximation using (Y, D, X)	2	0.150	0.300	0.450	0.150	0.300	0.450	0.150	0.300	0.450
approximation using (Y, D, Z)	2	0.138	0.252	0.345	0.082	0.248	0.334	0.136	0.243	0.320
exact value of the bias	3	0.179	0.320	0.456	0.180	0.339	0.494	0.183	0.371	0.556
approximation using (Y, D, X)	3	0.179	0.358	0.537	0.179	0.358	0.537	0.179	0.358	0.537
approximation using (Y, D, Z)	3	0.170	0.324	0.463	0.170	0.320	0.449	0.169	0.312	0.424

Note. All values, excluding those for the $\zeta_x = \zeta_u = 0$ case, are obtained via numerical integration. The exact calculations for the $\zeta_x = \zeta_u = 0$ case are derived in Section 3.7 and Section 4.2. The exact value of the bias is obtained from (15), the infeasible approximation to the bias using (Y, D, X) is obtained from (11), and the feasible approximation to the bias using (Y, D, Z) is obtained from (14).

information is available on the types of schools attended by individuals as well as on scores at math and reading ability tests taken at age 7 and age 11.

The information used here comes from a sample of $2,682$ working males at age 33 for whom non-missing information on educational qualifications and test scores is available. Four *incremental* categories of education are considered: no qualifications, O Level qualifications, A Level qualifications and Higher education.[21] We focus on the estimation of the ATTs relative to these categories, which represent the average payoff to individuals' own educational choices. Specifically, we consider the estimation of *three* ATT parameters, defined by having O Level qualifications *vis-à-vis* having no qualification, having A Level qualifications *vis-à-vis* having O Level qualifications and moving to Higher education *vis-à-vis* stopping at A Level qualifications.

The set of regressors X controlled for in the analysis has been chosen to match closely the specification in Blundell *et al.* (2005), though we consider *raw scores* at verbal and math tests taken at age 7 and age 11 instead of *quartiles* of these scores (as in their application). This allows us to define a *continuous* measure of ability for individuals in the sample by combining results across all tests. First, verbal and math scores at both ages have been standardized so that they take values on the same scale (between 0 and 10). Second, the average verbal and average math scores have been computed from the two tests taken at age 7 and age 11. Finally, the logged sum of the two mean scores is considered. The resulting distribution of the ability score is reported in the first panel of Figure 1, while sample size by educational groups is in Table 2.[22] We investigate the sensitivity of point estimates of returns to measurement error in the above defined indicator of ability.

The top panel of Table 2 presents estimation results from raw data (i.e. not accounting for mea-

[21] A detailed description of NCDS data, the variables being used in our application and more details about the educational categories considered can be found in Blundell *et al.* (2005).

[22] A negligible fraction of individuals in each group have been dropped from the analysis to ensure common support with respect to individuals in the adjacent educational category (see the graphical evidence reported in Figure 1).

Table 2: Incremental returns to O Levels, A Levels and Higher Education and the effect of measurement error in the ability score

	O Level		A Level		HE	
Raw Estimates						
	ATT	**Std.Err.**	**ATT**	**Std.Err.**	**ATT**	**Std.Err.**
OLS	0.1567	0.0277	0.0781	0.0199	0.1951	0.0239
Matching	0.1732	0.0288	0.0809	0.0198	0.2002	0.0257
Weighting	0.1763	0.0286	0.0829	0.0203	0.1925	0.0280
Stratification	0.1846	0.0264	0.0827	0.0212	0.1984	0.0264
Approximation to measurement error bias						
Extent of error	**Bias**	**Std.Err.**	**Bias**	**Std.Err.**	**Bias**	**Std.Err.**
10%	0.0037	0.0015	0.0015	0.0005	0.0057	0.0011
20%	0.0073	0.0030	0.0030	0.0010	0.0114	0.0022
30%	0.0110	0.0045	0.0045	0.0015	0.0171	0.0033
Treated on support	96.45%		99.32%		99.09%	
Sample Size	732		737		768	
Sample size for the 'No qualifications' group: 445						

Note. Educational categories are defined in Section 6. The extent of measurement error is defined as the noise-to-signal ratio. The measurement error correction is obtained by considering the empirical analogue of the quantity in Proposition 2. Bootstrap standard errors based on 500 replications reported throughout.

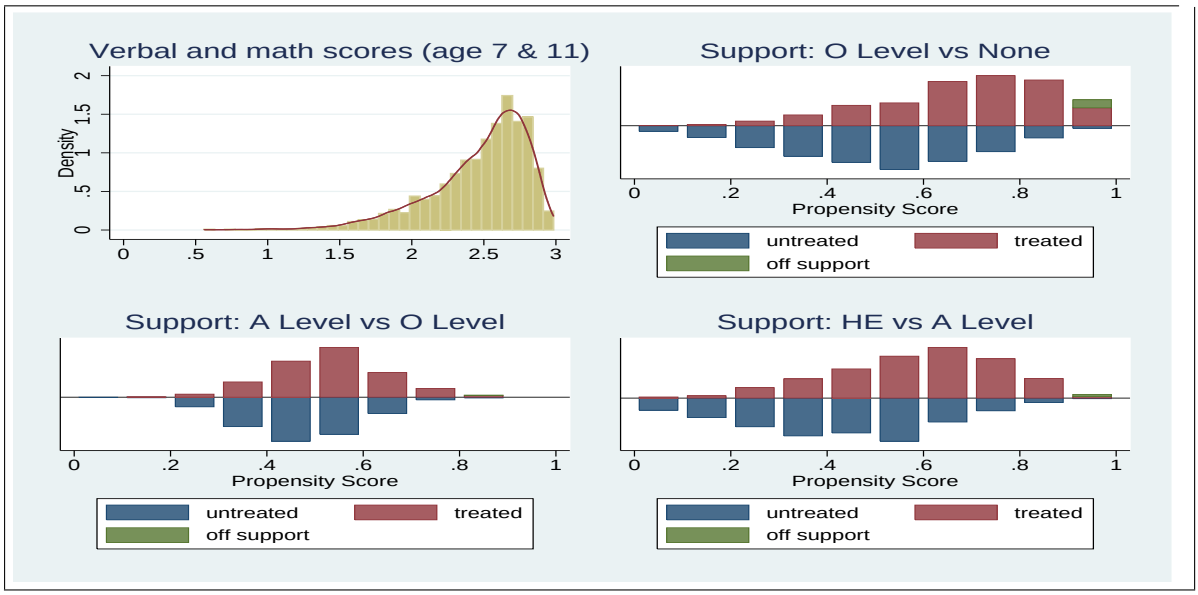

Figure 1: Distribution of the raw ability score and common support issues for the educational categories considered

surement error) for the incremental returns "O Level vs None", "A Level vs O Level" and "Higher Education vs A Level". *Four* estimation methods are considered. First, results from an OLS regression of wages on X and a dummy for the educational qualification of interest fully interacted with X are considered (*OLS*). Second, we report results from matching individuals on the propensity score predicted from a logit specification (*Matching*); the distance between individuals in adjacent educational categories has been defined using a normal kernel function. Third, results from a weighting procedure based on the estimated score are considered (*Weighting*). Finally, estimation based on stratification on the estimated score is considered (*Stratification*). Bootstrap standard errors based on 500 replications are reported throughout. Results appear to be rather robust with respect to the estimation method considered and in line with those in Blundell *et al.* (2005).

We then allow for classical measurement error in the raw indicator of ability and implement the correction procedures described in Section 4. The impact of measurement error is investigated by means of a sensitivity analysis with respect to *three* values of the measurement error variance σ^2, corresponding to 10, 20 and 30 percent of the variance of the raw ability indicator. In bottom panel of Table 2 we report an approximation to the bias estimated from the following quantity:

$$\sigma^2 \frac{1}{n_1} \sum_{i=1}^{n} \frac{d_i \hat{e}_Z^{(1)}(z_i)}{\hat{e}_Z(z_i)[1-\hat{e}_Z(z_i)]} \hat{E}^{(1)}_{Y_0|D,Z}(Y_0|0,z_i),$$

that is by considering the analogue estimator of the expression derived in Proposition 2 when X is replaced by Z. In particular, we assumed a *logit* specification for $e_Z(z)$ and that the relationship between the outcome Y and the regressors Z for the group $D=0$ is *linear*.[23] It turns out that the bias is always positive and small - about one percentage point in value - and statistically different from zero in all cases.

[23]The group $D=0$ comprises individuals with no qualification, O levels and A levels depending on the ATT parameter being estimated. Results from alternative specifications of the regression function and of $e_Z(z)$ can be derived along the same lines, but are not reported here as they proved qualitatively similar to those in Table 2.

7 Conclusions

There has been much theoretical and applied work focussed on the evaluation problem, that is on the measurement of the causal impact of a generic 'treatment' on outcomes of interest. This paper has proposed a method for bias reduction in estimation of treatment effects built on the assumption of ignorable assignment given a set of covariates when they may be affected by measurement error.

The method can be used to explore the sensitivity of results to the presence of measurement error after having estimated the returns to 'treatment' employing propensity score matching, stratification matching and conditional differences in differences estimators. The procedure exploits nothing but the error contaminated covariate data, and can be easily implemented using available software.

We show that measurement error in general invalidates restrictions that would be identifying were data error-free. This results in biased estimates of the causal parameters of interest such as the average treatment effect or the average treatment effect on the treated. As empirical applications typically make use of estimators that are defined from non-linear functionals of raw data (e.g. propensity score matching), this bias is difficult to sign. Our results provide a first order approximation to this bias for small values of the measurement error variance, and the evidence we provide indicates that the approximation is still valid when measurement error explains 30% of the variance of the error-ridden covariate.

References

[1] Battistin, E. and Sianesi, B. (2009), *Misreported Schooling and Returns to Education: Evidence from the UK*, forthcoming in the Review of Economics and Statistics

[2] Blundell, R., Dearden, L. and Sianesi, B. (2005), *Evaluating the effect of education on earnings: models, methods and results from the National Child Development Survey*, Journal of the Royal Statistical Society Series A, Vol. 168, No. 3, pp. 473-512

[3] Bound, J., Brown, C. and Mathiowetz, N. (2001), *Measurement error in survey data*, in J.J. Heckman and E. Leamer (eds.), *Handbook of Econometrics. Vol. 5*, Amsterdam: North-Holland, pp. 3705-3843

[4] Box, G.E.P. and Tiao, G.C. (1973), *Bayesian Inference in Statistical Analysis*, Addison-Wesley

[5] Carrol, R.J., Ruppert, D., Stefanski, L.A., and C. Crainiceanu (2006), *Measurement Error in Nonlinear Models: A Modern Perspective, Second Edition*, Chapman and Hall CRC Press

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

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

[8] Chesher, A. and Schluter, C. (2002), *Welfare Measurement and Measurement Error*, The Review of Economic Studies, Vol. 69, No. 2, pp. 357-378

[9] Cochran, W. G. and Rubin, D. B. (1973), *Controlling Bias in Observational Studies: a Review*, Sankhyā Series A, Vol. 35, No. 4, pp. 417-466

[10] Dehejia, R. and Wahba, S. (1995), *Causal Effects in Nonexperimental Studies*, unpublished manuscript, Harvard University

[11] Fan, J. and Gijbels, I. (1996), *Local polynomial modelling and its application*, London: Chapman and Hall

[12] 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

[13] Hausman, J., Ichimura, H., Newey W., and J. Powell (1991), *Measurement Error in Polynomial Regression Models*, Journal of Econometrics, Vol. 50, pp. 271-295

[14] 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

[15] Heckman, J.J., and Robb, R. Jr. (1985), *Alternative methods for evaluating the impact of interventions.* In J. Heckman and B. Singer (Eds.), Longitudinal Analysis of Labor Market Data (pp. 156-245). New York: Cambridge University Press

[16] Heckman, J.J., H. Ichimura, J. Smith, and P. Todd (1998), *Characterizing Selection Bias Using Experimental Data*, Econometrica, Vol. 66, pp. 1017-1098

[17] 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

[18] Heckman, J.J. (2000), *Causal Parameters and Policy Analysis in Econometrics: A Twentieth Century Retrospective*, Quarterly Journal of Economics, Vol. 115, pp. 45-97

[19] 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. 1161-1189

[20] Hong, H., and E. Tamer (2003), *A Simple Estimator for Nonlinear Errors in Variable Models*, Journal of Econometrics, 117, pp. 1-19

[21] 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

[22] Hu, Y. (2008), *Identification and Estimation of Nonlinear Models with Misclassification Error Using Instrumental Variables: A General Solution*, Journal of Econometrics, Vol. 144, No. 1, pp. 27-61

[23] Imbens, G.W. (2004), *Semiparametric Estimation of Average Treatment Effects under Exogeneity: A Review*, Review of Economics and Statistics, Vol. 86, pp. 4-29

[24] Lechner, M. (2001), *Identification and Estimation of Causal Effects of Multiple Treatments under the Conditional Independence Assumption*, in M. Lechner and F. Pfeiffer (Eds.), Econometric Evaluations of Active Labor Market Policies in Europe, Heidelberg: Physica Verlag

[25] Lewbel, A. (2007), *Estimation of Average Treatment Effects With Misclassification*, Econometrica, Vol. 75, pp. 537-551

[26] Li, T. (2002), *Robust and Consistent Estimation of Nonlinear Errors-in-Variables Models*, Journal of Econometrics, Vol. 110, pp. 1-26

[27] Mahajan, A. (2006), *Identification and Estimation of Regression Models with Misclassification*, Econometrica, Vol. 74, No. 3, pp. 631-665

[28] Molinari, F. (2008), *Partial Identification of Probability Distributions with Misclassified Data*, Journal of Econometrics, Vol. 144, No. 1, pp. 81-117

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

[30] 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, pp. 41-55

[31] Rubin, D.B. (1974), *Estimating Causal Effects of Treatments in Randomised and Non-randomised Studies*, Journal of Educational Psychology, 66, pp. 688-701

[32] Schennach, S. (2004), *Estimation of Nonlinear Models with Measurement Error*, Econometrica, 72, 33-75

[33] Wolfram Research Inc.,(2008), *Mathematica Edition: Version 7.0*, Champaign, Illinois: Wolfram Research, Inc.

Appendix

Proof of Proposition 1

Using the approximations in Section 3.5 and neglecting terms which are $o(\sigma^2)$ we have the following expression for $i \in \{0, 1\}$:

$$\begin{aligned}\Delta_i &\simeq \frac{\sigma^2}{2}\int E_{Y_i|X}(Y_i|z)f_X^{(2)}(z)dz + \sigma^2\int E_{Y_i|X}^{(1)}(Y_i|z)g_{X|D}(z|i)f_X(z)dz \\ &+ \frac{\sigma^2}{2}\int E_{Y_i|X}^{(2)}(Y_i|z)f_X(z)dz.\end{aligned}$$

Use the assumptions:

$$\begin{aligned}\lim_{z\to\pm\infty} E_{Y_i|X}(Y_i|z)f_X^{(1)}(z) &= 0, \\ \lim_{z\to\pm\infty} E_{Y_i|X}^{(1)}(Y_i|z)f_X(z) &= 0,\end{aligned}$$

and integrate by parts to get:

$$\int E_{Y_i|X}(Y_i|z)f_X^{(2)}(z)dz = -\int E_{Y_i|X}^{(1)}(Y_i|z)f_X^{(1)}(z)dz = \int E_{Y_i|X}^{(2)}(Y_i|z)f_X(z)dz,$$

so that for $i \in \{0, 1\}$:

$$\Delta_i \simeq \sigma^2\int E_{Y_i|X}^{(1)}(Y_i|z)g_{X|D}(z|i)f_X(z)dz + \sigma^2\int E_{Y_i|X}^{(2)}(Y_i|z)f_X(z)dz. \tag{16}$$

Since by the Bayes' theorem one can write:

$$g_{X|D}(x|d) \equiv \frac{\partial}{\partial x_*}lnf_{X|D}(x|d) = \frac{\partial}{\partial x_*}lnf_{D|X}(i|x) + \frac{\partial}{\partial x_*}lnf_X(x),$$

it also follows that:

$$\begin{aligned}\Delta_i &\simeq \sigma^2\int E_{Y_i|X}^{(1)}(Y_i|z)\frac{\partial}{\partial x_*}lnf_{D|X}(i|z)f_X(z)dz + \sigma^2\int E_{Y_i|X}^{(1)}(Y_i|z)f_X^{(1)}(z)dz \\ &+ \sigma^2\int E_{Y_i|X}^{(2)}(Y_i|z)f_X(z)dz.\end{aligned}$$

Integrating by parts again we have:

$$\Delta_i \simeq \sigma^2 \int E^{(1)}_{Y_i|X}(Y_i|z)\frac{\partial}{\partial x_*} ln f_{D|X}(i|z) f_X(z) dz.$$

It therefore follows that the following expressions for Δ_0 and Δ_1 hold:

$$\begin{aligned}
\Delta_0 &\simeq -\sigma^2 \int E^{(1)}_{Y_0|X}(Y_0|z)\frac{e^{(1)}_X(z)}{1-e_X(z)} f_X(z) dz, \\
&= \sigma^2 \int E^{(1)}_{Y_0|X}(Y_0|z) g_{X|D}(z|0) f_X(z) dz + \sigma^2 \int E^{(2)}_{Y_0|X}(Y_0|z) f_X(z) dz, \\
\Delta_1 &\simeq \sigma^2 \int E^{(1)}_{Y_1|X}(Y_1|z)\frac{e^{(1)}_X(z)}{e_X(z)} f_X(z) dz, \\
&= \sigma^2 \int E^{(1)}_{Y_1|X}(Y_1|z) g_{X|D}(z|1) f_X(z) dz + \sigma^2 \int E^{(2)}_{Y_1|X}(Y_1|z) f_X(z) dz,
\end{aligned}$$

from which the expression for $\Lambda_Z - \beta_e$ is obtained. ■

Proof of Proposition 2

Use the same argument as in Proposition 1 to get:

$$\Delta_{0|1} \simeq \sigma^2 \int E^{(1)}_{Y_0|X}(Y_0|z) g_{X|D}(z|0) f_{X|D}(z|1) dz + \sigma^2 \int E^{(2)}_{Y_0|X}(Y_0|z) f_{X|D}(z|1) dz. \tag{17}$$

Use the assumption:

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

and integrate by parts to write:

$$\int E^{(2)}_{Y_0|X}(Y_0|z) f_{X|D}(z|1) dz = -\int E^{(1)}_{Y_0|X}(Y_0|z) g_{X|D}(z|1) f_{X|D}(z|1) dz.$$

As by using the Bayes' theorem we have:

$$\begin{aligned}
g_{X|D}(z|1) - g_{X|D}(z|0) &= \frac{\partial}{\partial x_*} ln \frac{f_{X|D}(z|1)}{f_{X|D}(z|0)}, \\
&= \frac{\partial}{\partial x_*} ln \frac{e_X(z)}{1-e_X(z)},
\end{aligned}$$

we can write:

$$\begin{aligned}\Delta_{0|1} &\simeq -\sigma^2 \int E^{(1)}_{Y_0|X}(Y_0|z)\frac{e^{(1)}_X(z)}{e_X(z)[1-e_X(z)]} f_{X|D}(z|1)dz, \\ &= \sigma^2 \int E^{(1)}_{Y_0|X}(Y_0|z) g_{X|D}(z|0) f_{X|D}(z|1)dz + \sigma^2 \int E^{(2)}_{Y_0|X}(Y_0|z) f_{X|D}(z|1)dz,\end{aligned}$$

from which the expression for $\Gamma_Z - \beta_t$ is obtained. ■

Show that $\Lambda^1_Z = \Lambda^2_Z = \Lambda^3_Z$ and $\Gamma^1_Z = \Gamma^2_Z = \Gamma^3_Z$

Start from the following definitions:

$$\begin{aligned}\Lambda^1_Z &\equiv \int \left(E_{Y|DZ}(Y|1,z) - E_{Y|DZ}(Y|0,z)\right) f_Z(z)dz, \\ \Gamma^1_Z &\equiv \int \left(E_{Y|DZ}(Y|1,z) - E_{Y|DZ}(Y|0,z)\right) f_{Z|D}(z|1)dz,\end{aligned}$$

and use:

$$\begin{aligned}E_{YD|Z}(YD|z) &= E_{Y|DZ}(Y|1,z)e_Z(z), \\ E_{YD|Z}(Y(1-D)|z) &= E_{Y|DZ}(Y|0,z)\left[1-e_Z(z)\right],\end{aligned}$$

to write:

$$\begin{aligned}\Lambda^1_Z = \Lambda^2_Z &= \int \left(\frac{E_{YD|Z}(YD|z)}{e_Z(z)} - \frac{E_{YD|Z}(Y\,(1-D)\,|z)}{1-e_Z(z)}\right) f_Z(z)dz, \\ \Gamma^1_Z = \Gamma^2_Z &= \int \left(\frac{E_{YD|Z}(YD|z)}{P(D=1)} - \frac{E_{YD|Z}(Y\,(1-D)\,|z)}{1-e_Z(z)}\frac{e_Z(z)}{P(D=1)}\right) f_Z(z)dz.\end{aligned}$$

Finally, use the *balancing property* of the propensity score $e_Z(z)$ (Rosenbaum and Rubin, 1983) as well as the fact that z is finer than $e_Z(z)$ to write:

$$\begin{aligned}
\Lambda_Z^3 &\equiv \int \left(E_{Y|De_Z}(Y|1,\eta) - E_{Y|De_Z}(Y|0,\eta)\right) f_{e_Z}(\eta)d\eta, \\
&= \int \left(\int E_{Y|DZe_Z}(Y|1,z,\eta) - E_{Y|DZe_Z}(Y|0,z,\eta) f_{Z|e_Z}(z|\eta)dz\right) f_{e_Z}(\eta)d\eta, \\
&= \int \left(\int E_{Y|DZ}(Y|1,z) - E_{Y|DZ}(Y|0,z) f_{Z|e_Z}(z|\eta)dz\right) f_{e_Z}(\eta)d\eta, \\
&= \int E_{Y|DZ}(Y|1,z) - E_{Y|DZ}(Y|0,z) \left(\int f_{Ze_Z}(z,\eta)d\eta\right) dz, \\
&= \int E_{Y|DZ}(Y|1,z) - E_{Y|DZ}(Y|0,z) f_Z(z)dz, \\
&= \Lambda_Z^1.
\end{aligned}$$

A similar argument applies to Γ_Z^3. ■

Proof of results in the discussion after Proposition 2

If the true propensity score $e_X(x)$ is a logit:

$$e_X(x) = \frac{e^{\delta x}}{1+e^{\delta x}},$$

there is:

$$e_X^{(1)}(x) = \delta_* e_X(x)[1-e_X(x)],$$

and thus:

$$\frac{e_X^{(1)}(z)}{e_X(z)[1-e_X(z)]} = \delta_*.$$

Note also that the expression for the bias can be rearranged to get:

$$\begin{aligned}
\Gamma_Z - \beta_t &\simeq \sigma^2 \int E_{Y_0|X}^{(1)}(Y_0|z) \frac{e_X^{(1)}(z)}{e_X(z)[1-e_X(z)]} f_{X|D}(z|1)dz, \\
&= \sigma^2 \int E_{Y_0|X}^{(1)}(Y_0|z) \frac{e_X^{(1)}(z)}{e_X(z)[1-e_X(z)]} \frac{e_X(z)f_X(z)}{P[D=1]}dz, \\
&= \frac{\sigma^2}{P[D=1]} \int E_{Y_0|X}^{(1)}(Y_0|z) \frac{e_X^{(1)}(z)}{1-e_X(z)} f_X(z)dz,
\end{aligned}$$

and:

$$-\frac{\Delta_0}{P[D=1]} \simeq \frac{\sigma^2}{P[D=1]} \int E^{(1)}_{Y_0|X}(Y_0|z) \frac{e^{(1)}_X(z)}{1-e_X(z)} f_X(z) dz,$$

the last expression following from Proposition 1. ■