

Bias-Reducing Estimation of Treatment Effects in the Presence of Partially Mismeasured Data^{*}

Stephan A. Wiehler [†]

January 30, 2007

Labor market policy evaluation studies often rely on a merged database from different administrative entities. Suppose that one observes *inter alia* a variable of dubious quality for the entire population and the correct value of the same variable for the treated subgroup from an extra source. This paper introduces a bias-reducing estimator of average treatment effects based on the propensity score, as a widespread tool in this area. Validation data are employed in order to control for mismeasurements of the non-validation units when treatment and validation status are binary and coincide. A Monte Carlo simulation reveals its dominance under realistic calibrations compared to naive parametric propensity score based approaches. An application to widely used German administrative data underlines its relevance.

Keywords: Measurement error, propensity score, treatment effects

JEL classification: C14, C15.

7313 words

^{*} Special thanks to Michael Lechner, Bo Honoré, Markus Frölich and Blaise Melly for helpful comments.

[†] Swiss Institute for International Economics and Applied Research, University of St. Gallen, Switzerland, email: stephan.wiehler@unisg.ch

1. Introduction

Labor market policy evaluation studies often rely on a merged database from different administrative entities, such as social insurance records, public employment service records and program registers. Depending on the purpose of the data, some information might be archived with varying preciseness. Since caseworkers base their program allocation decision on personal information of the unemployed person, the program register is usually assessed to be the best source of personal characteristics. Assuming selection on observables as the identifying assumption for average treatment effects and focussing on the propensity score as a central tool in the treatment evaluation literature, this paper analyzes how additional reliable validation data on personal characteristics, that are only available for participants, e.g. the program register, can be used to control for mismeasurements of personal characteristics in other administrative sources that affect participants and nonparticipants.

Using results from the measurement error literature in the limited dependent variable context, but in contrast to the commonly used assumption of random validation data, the validation status is allowed to equal the binary treatment status. It will be shown how the first step propensity score estimation can be improved. Furthermore, the concept of expected propensity scores will be introduced leading to a bias-reduced estimation of average treatment effects. A Monte Carlo study reveals that the new estimator performs better in terms of bias and mean squared error compared to the case of naive parametric propensity score models, either using or ignoring the validation data for the participants. An application to administrative data in Germany, that were widely used in evaluation studies, shows its practical relevance.

Partially mismeasured data often occur in applied work. Researchers evaluating data

often possess more detailed information about a subgroup of observations as a result of additional data, replicate measurements or a closer inspection. Greenlees, Reece, and Zieschang (1982) combine U.S. data from the Current Population Survey with data from Social Security benefit and earnings records and from federal income tax records to test the implications of observing item nonresponse that depends on the level of the underlying variable. Okner (1972), for instance, analyzed the 1967 Survey of Economic Opportunity and additionally used the 1966 Tax File because the income measures in the former were misreported. Hu and Ridder (2005) use U.S. data from the Survey of Income and Program Participation (SIPP) in combination with a random sample of mothers participated in the Aid to Families with Dependent Children Program (AFDC QC) in order to correct for misreported income in the SIPP. Bound, Brown, and Mathiowetz (2001) survey measurement error constellations in studies that use (merged) administrative data. In the field of treatment evaluation Lechner, Miquel, and Wunsch (2004, 2005) use a merged database to evaluate training programs in Germany. The information on education they use from the social insurance records (SIR) is assessed to be of bad quality since it is reported by the employers who do not have any direct utility from SIR and therefore report those data with less care. In addition, they possess good information on education for training participants from program records, archived by the caseworkers, who base their allocation decision on the true level of education.¹ Fitzenberger, Osikominu, and Völter (2006b) use the same data and develop imputation rules to correct for the measurement errors for the nonparticipants.²

¹ They use a set of assumptions to correct for the nonparticipants extensively described in Bender, Bergemann, Fitzenberger, Lechner, Miquel, and Wunsch (2005). This issue will be picked up again in the application at the end of the paper.

² Other studies based on this data source are Fitzenberger and Speckesser (2005) and Fitzenberger, Osikominu, and Völter (2006a). For measurement error issues in related fields see for example the recent contributions of Black and Smith (2005) and Bollinger (2003) or D'Agostino and Rubin (2000) in the context of estimating the PS in the presence of partially missing data. Battistin and Sianesi (2005) investigate misreporting on the treatment status.

The central role of the propensity score³ (PS) in the paradigm of the potential outcome approach to causality of Roy (1951) and Rubin (1974) is widely discussed in the literature. Nonparametric estimation techniques use its balancing property and the reduction of multidimensional individual characteristics into one measure. Matching, subclassification, regression on the PS and weighting by the inverse of the PS are the dominating approaches⁴. In the majority of cases, the PS is not observable and has to be estimated. Literature underlines that using the estimated PS instead of the true PS tends to improve the control of imbalances of the covariates between the different treatment groups and efficiency.⁵ Predominantly, the PS is modeled by parametric probit or logit models. Linking this to the strand of research on measurement errors in the maximum likelihood context, one strategy to tackle this problem was introduced by Carroll and Wand (1991) and Pepe and Fleming (1991). Both characterize the measurement error nonparametrically for a random subsample captured in the validation data. Carroll and Wand (1991) impute the likelihood contribution of the non-validation units by means of kernel regression techniques. Pepe and Fleming (1991) fill in the missing likelihood contribution by expected likelihood contributions.⁶ The primary concern of both papers is consistency of the underlying parameters. But, as D’Agostino and Rubin (2000) point out in their context of partially missing data, parameter estimation of a latent model’s index in the binary choice context is only an intermediate step in the evaluation context for a further one, the estimation of treatment probabilities and finally average treatment effects.

³ First proposed by Rosenbaum and Rubin (1983).

⁴ Gerfin and Lechner (2002); Heckman, Ichimura, Smith, and Todd (1996); Heckman, Ichimura, and Todd (1997); Imbens (2000); Lechner (1999); Lechner, Miquel, and Wunsch (2005); Rubin and Thomas (2000); Rosenbaum and Rubin (1984, 1985); Black and Smith (2004). See also two comprehensive surveys *inter alia* dealing with the propensity score by Heckman, LaLonde, and Smith (1999) and Imbens (2004).

⁵ Rosenbaum and Rubin (1984, 1985); Rosenbaum (1987); Rubin and Thomas (1996); Hirano, Imbens, and Ridder (2003); Hahn (1998).

⁶ Other related papers dealing with errors in variable models in a nonlinear context are Carroll and Stefanski (1990), who develop an asymptotic theory for the estimated coefficients in the latent model, and Lee and Sepanski (1995) in a non-linear least square framework. The latter replace the distorted non-validation part by means of linear projections.

The paper is therefore organized as follows. Section 2 deals with identification issues in the current setting of partially mismeasured data. It briefly summarizes previous work by Battistin and Chesher (2004), dealing with identification of average treatment effects in the presence of an entirely mismeasured covariate. Also assuming selection on observables, similar results will be presented in the current context. Section 3.1 presents the basic estimation problem. 3.2 introduces the underlying methodological fundament of the paper first proposed by Pepe and Fleming (1991). Because of its intuitive appeal it will be shown how this methodology can be used when treatment and validation status coincide. Second, treatment probabilities are estimated by means of expected propensity scores. A theoretical result emphasizing the relative dominance of the new estimator is presented. Section 3.3 and 4 illustrate the theoretical findings and practical relevance by means of a Monte Carlo simulation and an application to evaluate German training programs. Section 5 concludes.

2. Identification

In the absence of validation data Battistin and Chesher (2004) discuss identification of various treatment effects for the case of a measurement error that affects a covariate of the whole population. Assuming selection on observables, they come to the conclusion that the true average treatment effect (on the treated) is identified given one of the following three conditions. First, the outcome variable is independent of the covariates. Second, the mean outcome does not change for varying values of the true value X given its distorted value $\tilde{X}$ and third, the participation status is independent of the covariates. They coevally claim that theses conditions are hardly relevant since they are unlikely to be fulfilled in non-experimental applications. So given the result of Battistin and Chesher (2004), the estimated treatment effect is most likely to be biased. A condition similar to the latter

will be found in the present framework.

As mentioned afore, in the current setting treatment status and validation data membership coincide so that validation data exist only for the treated observations. Assume for convenience one (partially distorted) covariate. For now let $X = (X_t, X_{nt})$ denote the covariates when we observe the truth for treated (t) and non-treated (nt) units and let $\tilde{X} = (X_t, \tilde{X}_{nt})$ denote the constellation with the true covariate for the treated and the distorted level for the non-treated. In generale, the average treatment effect on the treated (ATET) is defined as

$$\begin{aligned}\theta \equiv E[Y^1 - Y^0|D = 1] &= E[Y^1|D = 1] - E[Y^0|D = 1] \\ &= E[Y^1|D = 1] - \int E[Y^0|X = x, D = 1]f_{X|D=1}(x)dx \quad (1)\end{aligned}$$

where $D=1$ ($D=0$) denotes that an observation is treated (not treated) and Y^1 (Y^0) is the outcome measure in a post-program state having (not) received the treatment. The first term on the right hand side is directly observable from the data. The second term cannot be observed and is therefore called the unknown counterfactual. Literature has shown that assuming selection on observables tantamount to conditional independence (CIA), i.e. $Y^0, Y^1 \perp\!\!\!\perp D|X$, this counterfactual and hence the ATET and the average treatment effect (ATE) are identified.⁷ Under the CIA the counterfactual can be written as⁸

$$\begin{aligned}\int E[Y^0|X, D = 1]f_{X|D=1}(x)dx &= \int E[Y^0|X = x, D = 0]f_{X|D=1}(x)dx \\ &= \int E[Y(1 - D)|X = x] \frac{P(D = 1|X)}{[1 - P(D = 1|X)]P(D = 1)} f(x)dx\end{aligned}$$

⁷ see Barnow, Cain, and Goldberger (1981), Rosenbaum and Rubin (1983). This CIA claims that conditional on X , the treatment probability is independent of the potential outcomes. It allows to replace the unknown counterfactual $E[Y^0|X, D = 1]$ by the observable $E[Y^0|X, D = 0]$ and hence to identify θ .

⁸ See also the appendix.

Define $P(D = 1|X) \equiv P_{1|X}$, $\hat{P}(D = 1|X) \equiv p_{1|X}$ as an estimate of $P_{1|X}$ and N_T as the number of treated. Consequently, θ can be consistently estimated by $\hat{\theta}(X) \equiv \hat{E}(Y^1|D = 1) - \hat{E}(Y^0|D = 1)$, i.e.

$$\hat{\theta}_X = \frac{1}{N_T} \sum_{i=1}^N d_i y_i - \frac{1}{N_T} \sum_{i=1}^N (1 - d_i) y_i \frac{p_{i,1|X}}{1 - p_{i,1|X}}, \quad (2)$$

where it is assumed that $P_{1|X}$ is unknown and has to be estimated.⁹ Assuming now that one observes $\tilde{X} = (X_t, \tilde{X}_{nt})$ instead of $X = (X_t, X_{nt})$ it is possible to immediately figure out the bias of the estimated $\hat{\theta}$ as a function of the estimated PS.

$$B_\theta = \hat{\theta}_X - \hat{\theta}_{\tilde{X}} = \frac{1}{N_T} \sum_{i=1}^N (1 - d_i) y_i \left(\frac{p_{i,1|\tilde{X}}}{1 - p_{i,1|\tilde{X}}} - \frac{p_{i,1|X}}{1 - p_{i,1|X}} \right) \quad (3)$$

Given at least one observation that has a nonzero non-treatment outcome, the true effect is only identified if $p_{i,1|\tilde{X}} = p_{i,1|X}$, i.e. B is only zero when the estimated PS is not affected by the measurement error, which might be true but is unrealistic in most applications.¹⁰ This is the analogue to the third identification condition of Battistin and Chesher (2004). It also follows that one possibility to reduce the bias is to minimise the last term in brackets of (3). Following similar steps as above, the average treatment effect $\gamma = E(Y^1 - Y^0)$ can be consistently estimated by

$$\hat{\gamma} = \frac{1}{N} \sum_{i=1}^N \frac{d_i y_i}{p_{i,1|X}} - \frac{1}{N} \sum_{i=1}^N \frac{(1 - d_i) y_i}{1 - p_{i,1|X}}. \quad (4)$$

Calculating the corresponding bias for the estimated ATE in this case yields after some

⁹ The expressions in equation (2) and (4) are well-known formulas, e.g. the inverse probability estimator in Hirano, Imbens, and Ridder (2003) or in Dehejia and Wahba (1997).

¹⁰ We ignore the case that the measurement error and specific values of Y compensate.

transformations

$$B_\gamma = \frac{1}{N} \sum_{i=1}^N d_i y_i \left(\frac{1}{p_{i,1|X}} - \frac{1}{p_{i,1|\tilde{X}}} \right) + \frac{N_T}{N} B_\theta. \quad (5)$$

The bias of the ATE can be expressed as a function of B_θ and as one can see immediately the third identification result of Battistin and Chesher (2004) also holds in the current context, i.e. γ can only be estimated consistently if the propensity score is not affected by the distortion. Finally, one can derive a similar expression for the $\text{ATE}|X = x$ as is shown in the appendix with one important difference. So far N_T and N in equation (3) and (5) could be determined. As long as the conditioning variable is not the distorted one it is still possible to determine the number of observations in this particular class $N_{X=x}$. However, if the conditioning set is distorted the true conditional average treatment effect is never identified.

The next section illustrates the basic estimation problem in this set up, introduces the Pepe and Fleming approach, and modifies the latter for the case that validation and treatment status coincide in order to reduce the bias of the estimated PS.

3. Bias-Reducing Estimation of Propensity Scores

3.1. The Basic Problem of Estimating the PS

Consider a model with a binary outcome variable D that is observed following the rule

$$D = 1\{D^* \geq 0\}, \quad \text{with} \quad D^* = H(X_c, \beta) + \varepsilon \quad (6)$$

D^* is a latent model that indicates $D = 1$ when the threshold 0 is exceeded and $D = 0$ otherwise. $H(X_c, \beta)$ is predominantly modeled linearly. $X_c = \{X_{c1}, \dots, X_{cK}\}$ is the correct vector of characteristics and β the corresponding parameter vector.. In the absence of any

distortion the coefficient vector β can be consistently estimated by maximum likelihood techniques (ML). Technically, under the assumption that ε is i.i.d. the log likelihood takes the well-known form

$$\mathcal{L}(D, X_c; \beta) = \prod_{i=1}^N G(X_{c,i}, \beta)^{d_i} (1 - G(X_{c,i}, \beta))^{1-d_i}, \quad (7)$$

where $G(\cdot)$ is the c.d.f. of ε which is usually assumed to be the normal or the logistic. The first term in curly brackets is the likelihood contribution of the treated and the second term is the likelihood contribution of the non-treated. In the absence of a distortion both terms are evaluated at the true value of X .

Now a partial measurement error is introduced in this model. In general let $X = (\tilde{X}, X_K)$, where $\tilde{X} = \{X_1, \dots, X_{K-1}, \tilde{X}_K\}$ denotes the covariates that are observable for validation and non-validation observations with $K-1$ correctly observed and for simplicity one mismeasured covariate $\tilde{X}_K$. X_K denotes the true value of $\tilde{X}_K$ only observable for the validation units. Assume $X \perp\!\!\!\perp \varepsilon$, i.e. the orthogonality assumption of all regressors w.r.t. ε holds.¹¹ Let $V = 1$ ($V = 0$) denote that an observation is (not) in the validation sample. Recall that in the current setting focus is put on $D = V$. For illustration purpose assume now a linear specification of $H(\cdot)$. Suppose that the distortion is corrected for the validation units, i.e. the treated. Then the latent model takes the following form.

$$D^* = \sum_{k < K} \beta_k X_k + \beta_K (DX_K + (1 - D)\tilde{X}_K) + \tilde{\varepsilon}$$

$\tilde{\varepsilon}$ accounts for the fact that we are no longer in the true model of equation (6). Since $\tilde{\varepsilon}$ is now a function of D , we face an endogeneity problem. Consistency of $\hat{\beta}$ is only provided for $\beta_K = 0$ or $\tilde{X}_K = X_K$. For $\beta_K \neq 0$ the exogeneity condition is not satisfied hence,

¹¹ Hence, implicitly the measurement error is also orthogonal to ε .

leading to biased estimates of the latent model coefficients.¹²

3.2. Likelihood Adjustments

This section briefly introduces the methodological fundament, first proposed by Pepe and Fleming (1991), that will be modified later on. In their work they focus on estimating parameters in a maximum likelihood framework that includes information gained from a random validation sample. For $V \perp\!\!\!\perp D$ and given the data at hand, they formulate the general likelihood function.

$$\mathcal{L}(D, X; \beta) = \prod_{V=1} F_{\beta}(D|X_K, \tilde{X}) \prod_{V=0} F_{\beta}(D|\tilde{X}), \quad (8)$$

where F is the probability function of the outcome variable D given $\tilde{X}$ and X_K , respectively. The likelihood contributions of the validation and non-validation units differ in X_K which is only available for the validation units. Rewriting the second part of equation (8) in terms of X_K for the non-validation units yields

$$F_{\beta}(D|\tilde{X}) = \int F_{\beta}(D|X_K, \tilde{X}) f_{X_K|\tilde{X}}(x_K) dx_K \quad (9)$$

$f_{X_K|\tilde{X}}$ is not observable for $V = 0$. But it can be estimated non-parametrically for the validation units and applied to the non-validation units by the following assumption. For illustration purposes we consider the smallest nonempty conditioning set $\tilde{X}_K$.

Common Conditional Distribution Assumption (CCDA)

$$f_{X_K|\tilde{X}_K, V=1} = f_{X_K|\tilde{X}_K, V=0} = f_{X_K|\tilde{X}_K}. \quad (10)$$

It states that the conditional distribution of X_K determined for the validation units

¹² Not correcting at all leads to a bias as well for obvious reasons as long as $\beta_K \neq 0$.

would have also been determined for the non-validation units if validation data existed for this subgroup.¹³ The interpretation is that there is no systematic relation between the measurement error and the validation status V . Pepe and Fleming (1991) prove consistency of $\hat{\beta}$ and show that the asymptotic variance is the sum of the usual ML variance plus an additional term capturing the variation from the non-parametric estimate of $f_{X_K|\tilde{X}_K,D=0}$.

We now provide a condition for the applicability of the latter approach if $V = D$, i.e. the observations in the validation and treated sample coincide. Remember that in the current setting V and D are binary. Pepe and Fleming (1991) used the CCDA in order to extract $f_{X_K|\tilde{X}_K,V=0}$ by replacing it with $f_{X_K|\tilde{X}_K,V=1}$. Simply replicating the CCDA here

$$f_{X_K|\tilde{X}_K,D=1} = f_{X_K|\tilde{X}_K,D=0} = f_{X_K|\tilde{X}_K} \quad (11)$$

leads to severe doubts. It states that given the distorted level $\tilde{X}_K$, there is no systematic relation between the true X_K and D so that using $\tilde{X}_K$ instead of X_K resolves all problems and leads to unbiased estimates of the propensity score. However, taking labor market programs, allocation into treatment is considerably driven by upfront face-to-face interviews where X_K is reported by the unemployed person, so that the treatment probability is determined by X_K rather than by $\tilde{X}_K$. Hence, the CCDA might be very hard to justify for $V = D$.

It shall now be shown how the CCDA can be avoided still being able to recover $f_{X_K|\tilde{X}_K,D=0}$ from the data by using the unconditional distribution of X_K . Being aware of

¹³ This assumption is also used in Chen, Hong, and Tamer (2005). Example: Data from public authorities often fulfill this implicit restriction since they capture information about say contributors to the social insurance system, i.e. about all those persons who are employed within a certain period independent of their potential treatment status in the future, say a labor market program in case of unemployment.

the pitfalls of the CCDA for $V = D$, the reverse assumption might well be acceptable.

$$f_{\tilde{X}_K|X_K,D=1} = f_{\tilde{X}_K|X_K,D=0} = f_{\tilde{X}_K|X_K}. \quad (12)$$

It states that given the true level X_K the distorted $\tilde{X}_K$ has no influence on the treatment status. The following transformations are useful.

$$f_{X_K|\tilde{X}_K,D=0} = \frac{f_{\tilde{X}_K|X_K,D=0}f_{X_K|D=0}}{f_{\tilde{X}_K|D=0}} = \frac{f_{\tilde{X}_K|X_K,D=1}f_{X_K|D=0}}{f_{\tilde{X}_K|D=0}}, \quad (13)$$

where the first term in the numerator of the last fraction is replaced using the assumption in equation (12). By means of this, only the second term in the numerator is not feasible since X_K is only observable for $D = 1$. Since we face a discrete X_K , we can recover $f_{X_K|D=0}$ by applying the law of total probability and end up with

$$P(X_K|D=0) = \frac{P(X_K) - P(X_K|D=1)P(D=1)}{P(D=0)}. \quad (14)$$

Except for $P(X_K)$ all terms can be observed in the data. However, for $P(X_K)$ one might have access to an additional source that can be used to close this gap. Take for example education or age. Public authorities may provide general statistics from an independent census that can be used to extract $P(X_K)$. Thus, by using equation (13) and (14) it is possible to determine $f_{X_K|\tilde{X}_K,D=0}$ for $V = D$ without using the CCDA. Define $\mathcal{X}_m$ as the support of the conditional distribution $f_{X_K|\tilde{X}_K=m,D=0}$. The weighted likelihood in the current setting is therefore

$$\begin{aligned} \mathcal{L}^{wt}(D, X; \beta) &= \prod_{D=1} G(w_i) \prod_{D=0} \int_{\mathcal{X}_m} G(w_i) f_{X_K|\tilde{X}_K=m,D=0}(x_K) dx_K \\ w_i &= \sum_{k < K} \beta_k x_{ki} + \beta_K (d_i x_{Ki} + (1 - d_i) x_K). \end{aligned} \quad (15)$$

$G(w_i)$ is the CDF of a standard normal or a logistic distribution with a linear index

specification for illustration. The likelihood contribution of the treated is the same as in equation (7) since the true level is known for all covariates. The likelihood contribution for every non-treated is now the integral of $G()$ over $\mathcal{X}_m$, i.e. the potential states in which the distorted $\tilde{X}_K$ might had been in the absence of a distortion. Hence, following Pepe and Fleming (1991), consistency is provided for β . Carroll and Wand (1991) propose a very similar likelihood function, but use the validation information to estimate the likelihood contribution of the non-validation units by means of kernel regression methods. Another example of using the information from a validation sample in form of conditional densities of the true X_K given the distorted value $\tilde{X}_K$ in a GMM context is Chen, Hong, and Tamer (2005).

Extending this approach to the needs of treatment evaluation the concept of expected PS is now introduced. Since the first-step estimation in equation leads to consistent estimates of β , we are able to recover unbiased treatment probabilities for the treated. However, the true level X_K is not observable for the non-treated. A natural extension of the line of argumentation is to use the expected PS.

$$p_{i,1|X} = \left\{ \begin{array}{ll} G\left(\sum_{k=1}^K \hat{\beta}_k x_{k,i}\right) & \text{for } D = 1 \\ \int_{\mathcal{X}_m} G\left(\sum_{k < K} \hat{\beta}_k x_{k,i} + \hat{\beta}_K x_K\right) f_{X_K|\tilde{X}_K=m, D=0}(x_K) dx_K & \text{for } D = 0 \end{array} \right\}, \quad (16)$$

The expected propensity score for an observation with $\tilde{x}_{k,i} = m$ and $D = 0$ is a weighted sum of the propensity scores given the potential states on the respective support $\mathcal{X}_m$. To get a notion of the bias that still occurs for the non-treated the following transformations are useful. Asymptotically $\hat{\beta} = \beta$ and the bias of the expected propensity score for a non-treated individual i can be written as

$$\begin{aligned} B_i &= \int_{\mathcal{X}_m} G\left(\sum_{k < K} \hat{\beta}_k x_{k,i} + \hat{\beta}_K x_K\right) f_{X_K|\tilde{X}_K=m, D=0}(x_K) dx_K - G\left(\sum_{k < K} \beta_k x_{k,i} + \beta_K x_{Ki}\right) \\ &= \int_{\mathcal{X}_m} \left[G\left(\sum_{k < K} \beta_k x_{k,i} + \beta_K x_K\right) - G\left(\sum_{k < K} \beta_k x_{k,i} + \beta_K x_{Ki}\right) \right] f_{X_K|\tilde{X}_K=m, D=0}(x_K) dx_K \end{aligned}$$

Linearizing the expression in square brackets by a second order Taylor expansion in the neighborhood of the true value x_{Ki} helps to determine the asymptotic bias. After some transformations¹⁴, we can write the individual bias as a function of conditional moments of $f_{X_K|\tilde{X}_K=m,D=0}$.

$$B_i \approx G' \left(\sum_{k < K} \beta_k x_{k,i} + \beta_K x_{Ki} \right) \beta_K \mu_m + \frac{1}{2} G'' \left(\sum_{k < K} \beta_k x_{k,i} + \beta_K x_{Ki} \right) \beta_K^2 [\sigma_m^2 + \mu_m^2]$$

with $\mu_m = E(X_K - X_{Ki} | \tilde{X}_k = m)$ $\sigma_m^2 = V(X_K | \tilde{X}_k = m)$ (17)

G' and G'' are the partial derivatives of G w.r.t. the distorted variable. The following three conditions lead to a zero or small individual B_i . First, the individual bias is small if $G'(\cdot)$ and $G''(\cdot)$ are near zero. Specifying G' as the Gaussian or the logistic density implies that the index has to be extremely small or extremely large, i.e. B_i is small if the individual treatment probability is either rather small or large. Second, $B_i = 0$ if β_K is zero which is a straightforward result. Both findings also hold for the case of naive modeling, i.e. using an unadjusted likelihood function. Third, B_i is zero if the condition

$$\frac{\mu_m}{\sigma_m^2 + \mu_m^2} = -\frac{1}{2} \frac{G''(\cdot)}{G'(\cdot)} \beta_K \quad (18)$$

is satisfied. If this condition is fulfilled $\forall i$ the weighted estimator estimates approximately unbiased treatment effects.¹⁵ Despite the strength of the condition in equation (18) it is still less restrictive and a clear advantage compared to the case of naive modeling where unbiased estimation of the average treatment effect (on the treated) can only be achieved if $\tilde{X}_{Ki} = X_{Ki} \forall i$ or by an offsetting effect of the measurement error on X_K and the bias in $\hat{\beta}$.

¹⁴ see the appendix for details

¹⁵ Remember this result only holds in a neighborhood of X_{Ki} .

For the case that none of the three conditions hold, the following Monte Carlo sheds some light into the relative performance under different settings and data constellations.

3.3. Monte Carlo Simulation

For the following Monte Carlo study the true data generating process takes the form:

$$Y_i^1 = \sum_{j=1}^5 \phi_j X_{ji} + u_i, \quad Y_i^0 = \left[\sum_{j=1}^5 \omega_j X_{ji} \right] + \omega_6 X_{4i}^2 + \omega_7 X_{5i}^2 + \xi_i \quad (19)$$

$$D_i^* = \sum_{j=1}^5 \beta_j X_{ji} + \varepsilon_i, \quad Y_i = D_i Y_i^1 + (1 - D_i) Y_i^0, \quad (20)$$

where ε , u , and ξ are mutually independent draws from a Gaussian. ε is assumed to be NIID so that the binary choice model is a probit. Again, the observation rule $D_i = 1\{D_i^* \geq 0\}$ applies. The treatment selection is based on the true X . Y^1 and Y^0 are the treatment and non-treatment outcomes respectively. They are modeled differently by choosing different coefficient vectors ω and ϕ and different functionals in X . X_1 is a constant. X_2 is a binary transformation of a uniform variable, X_3 is drawn from a Poisson, X_4 is constructed by a draw from a standard normal. X_5 is designed to represent education, analogous to the application in section 4. It takes three values 1, 2, 3 with a skewed distribution with probabilities 0.470, 0.465, 0.065. This calibration of X_5 and the measurement error added to X_5 is assumed to have the form of the empirical analogues we actually found in German administrative data as described in table 5.

Remember that for the treated both levels $\tilde{X}_5$ and X_5 can be observed. Hence, the conditional distributions $f_{X_5|\tilde{X}_5=m, D=1}$ and $f_{X_5|\tilde{X}_5=m, D=0}$ can be estimated using equation (13) and (14). All necessary information can be determined and one can now apply the weighted likelihood function of equation (15). Realistically, X_3 and X_4 are allowed to be correlated with X_5 (ρ_{35} , ρ_{45}) in order to allow for an effect of the distortion on the

corresponding estimated coefficients $\hat{\beta}_3$ and $\hat{\beta}_4$. The parameter vector β is set to imply 10% treated and 90% non-treated, again analogous to the application. As a benchmark the naive (*na*) probit approach with an unadjusted likelihood is applied to the partially corrected data, i.e. correcting X_5 for the treated units, to show its shortcomings and the improvements of the weighted (*wt*) loglikelihood $\mathcal{L}^{wt}$. Additionally, the probit estimates based on the raw data, i.e. completely ignoring the validation data are displayed (*ig*).

Table 1 contains the estimation results of the selection probits. In general, the variance for every single estimate of $\hat{\beta}_{k,wt}$ is the largest. The reason is that the estimation of $\hat{\beta}_{k,wt}$ incorporates more variation caused by the first-step estimation of $f_{X_5|\bar{X}_5,D=0}$. Starting with $N = 2000$ the improvement of the coefficient $\hat{\beta}_{5,wt}$ can be seen very clearly. $\hat{\beta}_{5,wt}$ is much closer to the true value than $\hat{\beta}_{5,na}$ or $\hat{\beta}_{5,ig}$. The constants $\hat{\beta}_{1,na}$ and $\hat{\beta}_{1,ig}$ even have the wrong sign. Increasing the sample size, one can observe that $\hat{\beta}_{wt}$ converges closer to its true value whereas $\hat{\beta}_{na}$ and $\hat{\beta}_{ig}$ remain almost unchanged. Asterisks denote that hypothesis $\hat{\beta}_{k,j} = \beta_k$ for $j = ig, na$ and $N = 6000$ can be rejected for β_1 , β_4 , and β_5 at the 1% level. The hypothesis $\hat{\beta}_{wt} = \beta$ cannot be rejected for all elements.

Since the coefficients themselves are not the primary objects of interest, the focus is now put on $P_{1|X}$ again. For every replication the estimated PS's are calculated for the *na*- and *wt*-probits, denoted by $p_{1|X}^{na}$ and $p_{1|X}^{wt}$. Those estimated PS's are then compared to the true $P_{1|X}$ and to the estimated PS in the absence of a distortion $p_{1|X}^{nodi}$. This distinction is done since Rosenbaum (1987), Rosenbaum and Rubin (1984, 1985) point out that using the estimated instead of the true PS performs better in terms of balancing the covariates and efficiency.¹⁶ The mean squared prediction error (MSPE) is calculated for every pair.

Table 2 reports the average MSPE.¹⁷

¹⁶ Hirano, Imbens, and Ridder (2003) show that weighting by the inverse of a non-parametrically estimated propensity score leads to efficient estimates of the average treatment effect. ? proves that the distinction between knowing the PS or not is irrelevant for the asymptotic variance bound of the estimated average treatment effect, but not for the average treatment effect on the treated.

¹⁷ The comparison for the *ig*-case is not reported for clarity reasons. The average MSPE for this case is

Table 1: Estimation Results of the Selection Probit

β	N=2000			N=4000			N=6000		
	$\hat{\beta}_{na}$	$\hat{\beta}_{ig}$	$\hat{\beta}_{wt}$	$\hat{\beta}_{na}$	$\hat{\beta}_{ig}$	$\hat{\beta}_{wt}$	$\hat{\beta}_{na}$	$\hat{\beta}_{ig}$	$\hat{\beta}_{wt}$
$\beta_1 = 0.2$	-0.36 (0.12)	-0.45 (0.12)	0.20 (0.43)	-0.37 (0.08)	-0.46 (0.09)	0.17 (0.23)	-0.37* (0.07)	-0.45* (0.07)	0.18 (0.10)
$\beta_2 = -0.4$	-0.38 (0.08)	-0.38 (0.08)	-0.40 (0.09)	-0.38 (0.05)	-0.38 (0.05)	-0.40 (0.06)	-0.38 (0.04)	-0.35 (0.04)	-0.40 (0.05)
$\beta_3 = 0.1$	0.08 (0.03)	0.07 (0.03)	0.08 (0.03)	0.08 (0.02)	0.07 (0.02)	0.08 (0.02)	0.08 (0.02)	0.07 (0.02)	0.09 (0.02)
$\beta_4 = 0.4$	0.31 (0.04)	0.29 (0.04)	0.34 (0.05)	0.31 (0.03)	0.29 (0.03)	0.35 (0.04)	0.31* (0.03)	0.29* (0.03)	0.36 (0.03)
$\beta_5 = -0.8$	-0.45 (0.07)	-0.39 (0.07)	-0.75 (0.12)	-0.45 (0.04)	-0.38 (0.05)	-0.77 (0.07)	-0.45* (0.04)	-0.38* (0.04)	-0.78 (0.06)

The true values of the parameters are given in the first column. The values of β were chosen to imply a treated/non-treated ratio of 1/9. The Monte Carlo includes 500 replications. The first two columns for each sample size denote the estimates for the naive probit approach incorporating ($\hat{\beta}_{na}$) validation data or not ($\hat{\beta}_{ig}$). Standard errors in parenthesis. $\rho_{35} = 0.2$, $\rho_{45} = 0.3$. The starting values for the maximization of the log likelihood are the OLS estimates. To account for global concavity of $\mathcal{L}^m$ different starting values were used. The results do not change. (*) denotes that the hypothesis $\hat{\beta}_j = \beta$ for $j = ig, na$ and $N = 6000$ can be rejected at the 1% level. $\hat{\beta}_{wt} = \beta$ cannot be rejected even on the 5% level.

Table 2: Average MSPE of the predicted propensity score for 1000 replications

	N=2000	N=4000	N=6000
$P_{1 X} - p_{1 X}^{na}$	0.0055	0.0054	0.0054
$P_{1 X} - p_{1 X}^{wt}$	0.0051	0.0049	0.0048
$p_{1 X}^{nodi} - p_{1 X}^{na}$	0.0053	0.0052	0.0053
$p_{1 X}^{nodi} - p_{1 X}^{wt}$	0.0049	0.0047	0.0047

Average mean squared prediction error of treatment probabilities for the *na*- and *wt*-probits compared to the true PS $P_{1|X}$ and to the estimated PS in the absence of a distortion $p_{1|X}^{nodi}$.

It can be seen that the average MSPE for *wt* is smaller for N=2000 and decreases faster in relative terms compared to the *na*-case. This holds for both comparisons, with the true PS and with the estimated PS in the absence of a distortion.

For further insights the average treatment effect is actually estimated following formula (4). The columns in table 3 present the absolute value of the relative bias, standard deviation, and mean squared error of the estimated ATE in the *na*, *ig* and *wt*-case. Again, applying the weighted likelihood together with expected propensity scores clearly cuts

worse than the naive approach.

down the relative bias considerably for all N . Clearly, this gain comes to the expense of an increased variance, which is twice as high for wt compared to an . Overall the weighting estimator cuts down the MSE to 19% of the na -approach and to approximately 9% of the ig -approach for $N=6000$. One can also observe that $\left| \frac{\gamma - \hat{\gamma}}{\gamma} \right|$ for na and ig decreases only to a small extent as N becomes larger.

Table 3: Estimated average treatment effect γ

	N=2000			N=4000			N=6000		
	$\hat{\gamma}_{na}$	$\hat{\gamma}_{ig}$	$\hat{\gamma}_{wt}$	$\hat{\gamma}_{na}$	$\hat{\gamma}_{ig}$	$\hat{\gamma}_{wt}$	$\hat{\gamma}_{na}$	$\hat{\gamma}_{ig}$	$\hat{\gamma}_{wt}$
$\left \frac{\gamma - \hat{\gamma}}{\gamma} \right $	6.24	9.32	0.64	5.46	8.21	0.43	5.26	7.93	0.30
std.	0.06	0.08	0.12	0.04	0.05	0.08	0.04	0.04	0.07
MSE	.028	.060	.015	.025	.056	.006	.026	.058	.005

500 replications; The table reports the absolute bias, standard deviation and the mean squared error for the weighting (wt) and naive (na) estimator as well as for the case of completely ignoring the validation data (ig).

As a further step sensitivity checks were conducted to test for robustness of the results with respect to certain components of the simulation. Clearly, with only five covariates, X_5 has a strong impact on the estimation results. However, adding more covariates, partially correlated with X_5 , does not change the qualitative result, i.e. the dominance of the weighted estimator. Increasing the treated/non-treated ratio, it shows up that the relative bias of $\hat{\gamma}_{na}$ and $\hat{\gamma}_{ig}$ decreases. However, the corresponding MSE is still higher compared to wt . Increasing the distortion results in a lower speed of convergence of the propensity score model consequently increasing the bias and variance of the estimated average treatment effect. All those mutations do not change the qualitative content of the results. But there is one sensitivity check that is worth mentioning. So far, the underlying distribution was skewed with probabilities 0.47, 0.465, 0.065. Changing this to 1/3 for each category and adding a symmetric measurement error, in the sense that $f_{X_K=n|\tilde{X}_K=m} = f_{X_K=m|\tilde{X}_K=n} \forall n, m$, one can observe that the na - and ig -approach catch up in terms of bias and MSE. The reason is that this artificial setting allows the distortion

to cancel out in the unweighted likelihood function leading to better estimates of $\hat{\beta}_{na}$ and $\hat{\beta}_{ig}$ and the respective $\hat{\gamma}_{na}$ and $\hat{\gamma}_{ig}$. Hence, the relative reduction of the MSE for the *wt*-case decreases, but is still existent.¹⁸

4. Application: Effects of Training Programs in Germany

4.1. Data

The data that are used to show the practical relevance of the weighted estimator proposed above, are merged records from different administrative entities in Germany. It is a combination of data from the social insurance records (SIR) on employment, data on benefit receipt during times of unemployment (BRR), and information on program participants (PPR), the latter two from the public employment service. Those data have been previously used by Lechner, Miquel, and Wunsch (2004, 2005), Lechner and Wunsch (2006), Fitzenberger, Osikominu, and Völter (2006a,b). For a detailed description of the data, the reader is referred to the respective articles. Those data comprise inter alia information on education from the SIR that is archived for all individuals who are subject to social insurance contributions between 1980 and 2003. This variable is reported by the employer. Some of the individuals in the SIR subsequently become unemployed and take part in a training program. For them we also observe information on education, archived by the caseworker in the process of program allocation. The latter information is assessed to be more reliable since caseworkers usually base their program allocation decision, among others, on education, whereas employer have no direct utility from SIR and therefore report education with less care.¹⁹ Being aware of this problem, Lechner, Miquel, and Wunsch (2004, 2005) impose a set of assumptions and correct the information for the nonpartic-

¹⁸ Results of the sensitivity analysis are available from the author on request.

¹⁹ Quite the contrary, this obligation to report such data invokes displeasure, and employer do not care about the quality of their reports, except for the salary paid.

ipants upfront.²⁰ Here, the raw information on education from the two sources SIR and PPR is used to demonstrate the impact of the estimator. Lechner and Wunsch (2006) use the data in a different context and compare the effects of participating in a training program versus nonparticipation in different phases of the German business cycle. They aggregate short, long and retraining programs into one category *training* which is suitable in the current context.²¹

Based on this data, we select a participation window and define a participant as an unemployed person, who takes part in a training program between 1993 and 1995. We only consider the first participation in that window. A nonparticipant is in principle also eligible, but not allocated to a training program between 1993 and 1995. Doing so, we end up with a sample of 2'466 participants in training and 25'678 nonparticipants. Table 4 reports descriptive statistics of the two groups.

In the group of participants we observe less women, foreigners and less married people. Participants have a lower (higher) fraction of (un-)employment 6 years before the entry into unemployment. The remaining benefit claim for nonparticipants is with 8.5 months more than twice as long as for participants. The length of the previous employment is longer for nonparticipants. In addition, participants have spent more time in previous labor market programs.

Looking at education extracted from SIR and coded as 1 for *no vocational degree*, 2 for *vocational degree*, and 3 for *academic degree*, we observe that participants and nonparticipants do not differ in means. Transforming education into dummies, we observe small differences of the coefficients on the three categories. However, looking at education from the PPR and comparing it to the SIR in levels, we observe that education is on average

²⁰ The underlying correction procedure are reported in Bender, Bergemann, Fitzenberger, Lechner, Miquel, and Wunsch (2005). The application strongly hinges on a set of other preparatory steps to define the final sample and participants and nonparticipants respectively. The provision of the data by Conny Wunsch is gratefully acknowledged.

²¹ The fractions for short, long, and retraining are 46, 34, and 20%.

Table 4: Descriptive Statistics of Nonparticipants and Participants

	Non.-P.	Partic.
# observations	25'678	2'466
(1) female	46.69	41.40
(2) age	34.43	34.91
(3) foreigner	9.51	7.46
(4) married	50.25	36.50
(5) at least one child	35.34	33.25
(6) remaining benefit claim in month at program entry	8.47	4.01
(7) fraction of empl. 72 months before entry into UE	59.89	47.09
(8) fraction of unempl. 72 months before entry into UE	15.73	31.73
(9) total months in program before entry into UE	0.99	1.57
(10) duration last empl. in months	38.54	31.11
(11) mean duration in empl. 48 months before entry into UE	28.11	20.97
(12) mean duration in unempl. 48 months before entry into UE	13.36	20.49
(13) unempl. rate	7.77	8.00
(14) residence in city>250.000 inhabitants	27.26	29.72
education SIR (<i>level 1,2,3</i>) <i>dummies</i>	1.72	1.72
no vocational degree	32.18	33.54
vocational degree	63.99	61.31
academic degree)	3.83	5.15
education PPR (<i>level 1,2,3</i>) <i>dummies</i>	n.a.	1.67
no vocational degree	n.a.	39.09
vocational degree	n.a.	54.54
academic degree	n.a.	6.37

Note: All numbers in percent if not stated otherwise. *SIR*: social insurance records, *PPR*: program participants records. Education levels: 1 for *no vocational degree*, 2 for *vocational degree*, and 3 for *academic degree*.

overreported in the SIR data. We observe 5.5 percent more participants with the lowest education dummy and 7 percent less participants in the medium category.

To get a an impression of the measurement error, it is useful to look at the empirical distribution of the participants. Table 5 shows that overreporting is an issue especially for persons without a vocational degree following PPR. Almost 38 percent of them are archived as having a vocational degree in the SIR. 16.4 percent of those who have a vocational degree in PPR are reported as having no vocational degree in the SIR. Even 5.6 percent who have an university degree following PPR are reported to be without a vocational degree in the SIR.

As shown in section 3.2, the applicability of the estimator hinges on the existence of

Table 5: Empirical Distribution of the Measurement Error

cells in %		$\tilde{X}_K$		
		1	2	3
X_K	1	62.0	16.4	5.7
	2	37.9	82.1	26.8
	3	0.1	1.5	67.5
# obs.		964	1'345	157

Note: X_K is again the true value of education from the program participant register (PPR) and $\tilde{X}_K$ the distorted value of education from the social insurance records (SIR).

an independent census that captures the unconditional distribution of education for the population under inspection, here the population of unemployed between 1993 and 1995 who are eligible for training programs. Such a census is available from the yearly statistic of the Federal Employment Agency of Germany²². This statistic is collected independently of the sources SIR and PPR and therefore fulfils the requirements for the estimation. We use the average fraction of the years '93, '94, and '95 of all unemployed without vocational degree (47%) with vocational degree (46.5%), and academic degree (6.5%) and plug it into the estimation.

Table 6 shows the results of the participation probit and the estimated average treatment effect of training on earnings. For clarity reasons, we only report covariates that are sensitive to the applied methodology w.r.t. magnitude, sign and/or significance. The other covariates in the probit models cover all important fields of personal and regional characteristics as well as labor market history, as listed in table (4). We use a linear specification for education.²³

Looking at the estimated coefficients of the first four covariates, it shows up that the coefficients of age and the fraction of time in employment 72 months before the entry into unemployment react slightly in size. The coefficient of foreigner status also exhibits

²² Bundesanstalt für Arbeit (1996)

²³ A number of specification test were performed to allow for more flexible specifications of education, for instance including dummies. All likelihood ratio test could not reject the linear specification.

Table 6: Estimation Results of the Participation Probit and Average Treatment Effects γ

$\hat{\beta}$	wt	na	ig
constant	-1.664** (.0713)	-1.337** (.0682)	-1.475** (.0714)
age	0.003* (.0014)	0.004** (.0013)	0.003* (.0014)
foreigner	-0.055 (.0408)	-0.098* (.0416)	-0.069 (.0416)
fraction of empl. 72 months before entry into UE + other covariates	0.268** (.0755) n.r.	0.281** (.0753)	0.267** (.0754)
education (<i>level 1,2,3</i>)	0.208** (.0212)	0.001 (.0221)	0.086** (.0215)
$\hat{\gamma}$ in Euro/month			
after 6 months	-127** (15.43)	-118** (16.09)	-120** (15.92)
after 36 months	84** (18.47)	113** (20.11)	103** (19.07)

Note: The three columns are estimated using a linear specification of education in the probit model. Significance is denoted by (*) for 5% and (**) for 1%. Standard errors are estimated using bootstrapping with 250 replications, where sampling is done with replacement, $M = N$. (n.r.) For clarity reasons we only report variables with either a change in magnitude, sign and/or significance. The other coefficients can be obtained from the author on request.

a strong variation. We observe a decrease in the (*na*)-case, which leads to a significant negative coefficient. For the *wt*-case we find a negative but not significant impact of the foreigner status on the selection into training programs.²⁴ Not surprisingly, we find a significant correlation of -0.2 with the education variable from the SIR, indicating that coefficients of correlated variables are also affected by the measurement error, which was one finding of the simulation.

For education the picture is quite different. It shows up that education has no or a very small positive impact on the probability of participating in training programs for the *na*- and *ig*-case. However, applying the weighting estimator one observes that the coefficient rises significantly up to 0.21, which is plausible since 20 percent of the training programs are retraining, which requires participants to have at least a vocational degree, that can actually be retrained. Hence, it can be stated that the choice of the weighting

²⁴ This is consistent with the descriptive statistics, where observe only a small difference in the fraction of foreigners in both groups.

estimator has a clear impact on the first step estimation results and on the corresponding interpretation of the selection mechanism into training.

Using (expected) propensity score of the first step and inverse probability weighting, we estimate the average treatment effect of training programs on earnings 6 and 36 months after program entry. In the lower part of table 6 it can be observed that the negative effect 6 months after the program, which is usually labeled as *lock-in* effect, as in van Ours (2004), is larger in the *wt*-case compared to *na* or *ig*. After 36 months we observe that the estimated average treatment effects on earnings is lower in the *wt*-case compared to the others, but still positive.

Overall, it can be stated that the weighted estimator together with expected propensity score leads to a clear change of the estimated coefficients of the latent model in the selection probits and, finally, of the estimated average treatment effects. We do not find a qualitative change of the interpretation of the effects, though in the size of the effects.

5. Conclusion

This paper investigated a widespread problem of labor market policy evaluation using merged data from different administrative sources. A covariate of dubious quality is observed in one source for the entire population, where the same covariate is observed without error in another source only for a subpopulation, here the treated units. Identification conditions of the average treatment effect (on the treated) are discussed. Assuming selection on observables as the identifying assumption and focussing on the propensity score as a central tool in the treatment evaluation literature, this paper employs results from the strand of literature on measurement errors in the maximum likelihood context by Pepe and Fleming (1991) and adjusts it to the current setting where validation and the binary treatment status coincide. Introducing expected propensity scores leads to a bias-reduced

estimation of the participation probabilities and finally of the estimated average treatment effect. A Monte Carlo reveals that given a realistic data generating processes with a calibration taken from actual administrative data from Germany this new estimator outperforms naive parametric propensity score models, either using or ignoring the validation data, by far. Applying this new estimator in an evaluation of German training programs shows that it has a clear impact on the interpretation of the allocation process into training and that it changes the size of the estimated average treatment effects of training on subsequent earnings considerably.

A. Appendix

A.1. Unknown counterfactual as a function of the Propensity Score

Similar steps are done in Battistin and Chesher (2004). Using the CIA the unknown counterfactual can be written as

$$\begin{aligned} E[Y^0|X, D=0] &= \int Y^0 f(Y^0|X, D=0) dY^0 = \int Y^0 \frac{f(Y^0, D=0|X)}{P(D=0|X)} dY^0 \\ &= \frac{\int Y(1-D)f(Y(1-D)|X) dY}{1-P(D=1|X)} = \frac{E(Y(1-D)|X)}{1-P(D=1|X)} \end{aligned}$$

Using $f(X|D=1) = \frac{f(X, D=1)}{P(D=1)} = \frac{P(D=1|X)f(X)}{P(D=1)}$ and putting it together yields the expression in section 2:

$$\int E[Y^0|X, D=0] f(X|D=1) dx = \int E(Y(1-D)|X=x) \frac{P(D=1|X)}{[1-P(D=1|X)]P(D=1)} f(x) dx$$

A.2. Bias for the Conditional Average Treatment Effect given $X = x$

$$\begin{aligned} B_{\gamma|X=x} &= \frac{1}{N_{X=x}} \sum_{i \in \{X=x\}} d_i y_i \left(\frac{1}{p_{i,1|X}} - \frac{1}{p_{i,1|\tilde{X}}} \right) \\ &+ \frac{1}{N_{X=x}} \sum_{i \in \{X=x\}} (1-d_i) y_i \left(\frac{p_{i,1|\tilde{X}}}{1-p_{i,1|\tilde{X}}} - \frac{p_{i,1|X}}{1-p_{i,1|X}} \right) \end{aligned}$$

A.3. Individual Bias for the Expected Propensity Score

$$B_i = \int_{\mathcal{X}_m} \left[G \left(\sum_{k < K} \beta_k x_{k,i} + \beta_K x_K \right) - G \left(\sum_{k < K} \beta_k x_{k,i} + \beta_K x_{Ki} \right) \right] f_{X_K|\tilde{X}_K=m, D=0}(x_K) dx_K$$

Taylor expanding the first term in squared brackets around the true value x_{Ki} vanishes $G(\sum_{k < K} \beta_k x_{k,i} + \beta_K x_{Ki})$ and leads to

$$\begin{aligned} B_i &\approx \int_{\mathcal{X}_m} G' \left(\sum_{k < K} \beta_k x_{k,i} + \beta_K x_{Ki} \right) \beta_K (x_K - x_{Ki}) f_{X_K | \tilde{X}_K=m, D=0}(x_K) dx_K \\ &+ \frac{1}{2} \int_{\mathcal{X}_m} G'' \left(\sum_{k < K} \beta_k x_{k,i} + \beta_K x_{Ki} \right) \beta_K^2 (x_K - x_{Ki})^2 f_{X_K | \tilde{X}_K=m, D=0}(x_K) dx_K \end{aligned}$$

Given that G is predominantly modeled by the normal or the logistic distribution it is reasonable to stop after the second order because derivatives of higher order than G'' are almost zero on the entire support. Reformulating yields

$$\begin{aligned} B_i &\approx G' \left(\sum_{k < K} \beta_k x_{k,i} + \beta_K x_{Ki} \right) \beta_K E(X_K - X_{Ki} | \tilde{X}_K = m) \\ &+ \frac{1}{2} G'' \left(\sum_{k < K} \beta_k x_{k,i} + \beta_K x_{Ki} \right) \beta_K^2 E([X_K - X_{Ki}]^2 | \tilde{X}_K = m) \end{aligned}$$

Using $V(a) = E(a^2) - E(a)^2$ we end up in

$$\begin{aligned} B_i &\approx G' \left(\sum_{k < K} \beta_k x_{k,i} + \beta_K x_{Ki} \right) \beta_K E(X_K - X_{Ki} | \tilde{X}_K = m) \\ &+ \frac{1}{2} G'' \left(\sum_{k < K} \beta_k x_{k,i} + \beta_K x_{Ki} \right) \beta_K^2 \left[V(X_K - X_{Ki} | \tilde{X}_K = m) + E(X_K - X_{Ki} | \tilde{X}_K = m)^2 \right], \end{aligned}$$

which is the result of equation (17).

References

- BARNOW, B., G. CAIN, AND A. GOLDBERGER (1981): “Selection on Observables,” *Evaluation Studies Review Annual*, 5, 43–59.
- BATTISTIN, E., AND A. CHESHER (2004): “The Impact of Measurement Error on Evaluation Methods Based on Strong Ignorability,” working paper, University College London.
- BATTISTIN, E., AND B. SIANESI (2005): “Misreported Schooling and Returns to Educa-

- tion: Evidence from the UK,” Working Paper, Institute for Fiscal Studies, London.
- BENDER, S., A. BERGEMANN, B. FITZENBERGER, M. LECHNER, R. MIQUEL, AND C. WUNSCH (2005): *Die Wirksamkeit von FuU-Massnahmen: Ein Evaluationsversuch mit prozessproduzierten Daten aus dem IAB*. Beiträge zur Arbeitsmarkt- und Berufsforschung.
- BLACK, D., AND J. SMITH (2004): “How Robust is the Evidence on the Effects of College Quality? Evidence from Matching,” *Journal of Econometrics*, 121, 99–124.
- (2005): “Estimating the Returns to College Quality with Multiple Proxies for Quality,” Working paper, University of Maryland.
- BOLLINGER, C. (2003): “Measurement Error in Human Capital and the Black-White Wage Gap,” *Review of Economics and Statistics*, 85, 578–585.
- BOUND, J., C. BROWN, AND N. MATHIOWETZ (2001): *Measurement Errors in Survey Data*, vol. IV of *Handbook of Econometrics*. North-Holland, Amsterdam.
- BUNDESANSTALT FÜR ARBEIT (1996): “Amtliche Nachrichten der Bundesanstalt für Arbeit - Arbeitsstatistik 1995,” Nürnberg.
- CARROLL, R., AND L. STEFANSKI (1990): “Approximate Quasi-likelihood Estimation in Models With Surrogate Predictors,” *Journal of the American Statistical Association*, 85, 652–663.
- CARROLL, R., AND M. WAND (1991): “Semiparametric Estimation on Logistic Measurement Error Models,” *Journal of the Royal Statistical Society*, 53, 573–585.
- CHEN, X., H. HONG, AND E. TAMER (2005): “Measurement Error Models with Auxiliary Data,” *Review of Economic Studies*, 72, 343–366.

- D'AGOSTINO, R., AND D. RUBIN (2000): "Estimating and Using Propensity Scores With Partially Missing Data," *Journal of the American Statistical Association*, 95, 749–759.
- DEHEJIA, R., AND S. WAHBA (1997): "Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programs," *Econometric Methods for Program Evaluation*, Ph.D. Dissertation, Harvard University.
- FITZENBERGER, B., R. OSIKOMINU, AND R. VÖLTER (2006a): "Get Training or Wait? Long Run Employment Effects of Training Programs for the Unemployed in West Germany," Working Paper.
- (2006b): "Imputation Rules to Improve the Education Variable in the IAB Employment Subsample," *Journal of the Applied Social Sciences*, forthcoming.
- FITZENBERGER, B., AND S. SPECKESSER (2005): "Employment Effects of the Provision of Specific Professional Skills and Techniques in Germany," Unpublished Manuscript, Goethe University of Frankfurt.
- GERFIN, M., AND M. LECHNER (2002): "A Microeconometric Evaluation of the Swiss Active Labour Market Policy," *The Economic Journal*, 112, 854–893.
- GREENLEES, J., W. REECE, AND K. ZIESCHANG (1982): "Imputation of Missing Values When the Probability of Response Depends On the Variable Being Imputed," *Journal of the American Statistical Association*, 77, 251–261.
- HAHN, J. (1998): "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica*, 66, 315–331.
- HECKMAN, J., H. ICHIMURA, J. SMITH, AND P. TODD (1996): "Sources of Selection Bias in Evaluating Social Programs. An Interpretation of Conventional Measures and Evi-

- dence on the Effectiveness of Matching as a Program Evaluation Method,” *Proceedings of the National Academy of Science of the United States of America*, 93, 13416–13420.
- HECKMAN, J., H. ICHIMURA, AND P. TODD (1997): “Matching as a Econometric Evaluation Estimator: Evidence from Evaluation of a Job Training Program,” *Review of Economic Studies*, 64, 605–654.
- HECKMAN, J., R. LALONDE, AND J. SMITH (1999): *The Economics and Econometrics of Active Labor Market Programs*, vol. III A of *Handbook of Economics*. North-Holland, Amsterdam.
- HIRANO, K., G. IMBENS, AND G. RIDDER (2003): “Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score,” *Econometrica*, 71, 1161–1189.
- HU, Y., AND G. RIDDER (2005): “Estimation of Nonlinear Models with Mismeasured Regressors Using Marginal Information,” IEPR Working Paper 05.39.
- IMBENS, G. (2000): “The Role of the Propensity Score in Estimating Dose-Response Functions,” *Biometrika*, 87, 706–710.
- (2004): “Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review,” *The Review of Economics and Statistics*, 86, 4–29.
- LECHNER, M. (1999): “Earnings and Employment Effects of Continuous Off-the-Job Training in East Germany After Unification,” *Journal of Business & Economic Statistics*, 17, 74–90.
- LECHNER, M., R. MIQUEL, AND C. WUNSCH (2004): “Long Run Effects of Public Sector Sponsored Training in West Germany,” Discussion Paper 2004-19, Department of Economics, University of St. Gallen.

- (2005): “The Curse and Blessing of Training the Unemployed in a Changing Economy: The Case of East Germany After Unification,” Discussion Paper 1684, Institute for the Study of Labor (IZA).
- LECHNER, M., AND C. WUNSCH (2006): “Are Training Programs More Effective When Unemployment is High,” Discussion Paper 2006-23, University of St. Gallen (SIAW).
- LEE, L., AND J. SEPANSKI (1995): “Estimation of Linear and Nonlinear Errors-in-Variables Models Using Validation Data,” *Journal of the American Statistical Association*, 90, 130–140.
- OKNER, B. (1972): “Constructing a New Data Base From Existing Microdata Sets: The 1966 Merge File,” *Annals of Economics and Social Measurement*, 1, 325–362.
- PEPE, M., AND T. FLEMING (1991): “A Nonparametric Method for Dealing With Mismeasured Covariate Data,” *Journal of the American Statistical Association*, 86, 108–113.
- ROSENBAUM, P. (1987): “Model-Based Direct Adjustments,” *Journal of the American Statistical Association*, 82, 387–394.
- ROSENBAUM, P., AND D. RUBIN (1983): “The Central Role of the Propensity Score in Observational Studies for Causal Effects,” *Biometrika*, 70, 41–55.
- (1984): “Reducing Bias in Observational Studies Using Subclassifications on the Propensity Score,” *Journal of the American Statistical Association*, 79, 516–524.
- (1985): “Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score,” *The American Statistician*, 39, 33–38.
- ROY, A. (1951): “Some Thoughts on the Distribution of Earnings,” *Oxford Economic Papers*, 3, 135–146.

- RUBIN, D. (1974): “Estimating Causal Effects of Treatments in Randomized and Non-randomized Studies,” *Journal of Educational Psychology*, 66, 668–701.
- RUBIN, D., AND N. THOMAS (1996): “Matching Using Estimated Propensity Scores: Relating Theory to Practice,” *Biometrics*, 52, 249–264.
- (2000): “Combining Propensity Score Matching With Additional Adjustments for Prognostic Covariates,” *Journal of the American Statistical Association*, 95, 573–585.
- VAN OURS, J. (2004): “The Locking-in Effect of Subsidized Jobs,” *Journal of Comparative Economics*, 32, 37–52.