



DALARNA
UNIVERSITY

Working papers in transport, tourism, information technology and microdata analysis

Likelihood estimate of treatment effects under selection bias

**Md. Moudud Alam
Maengseok Noh
Youngjo Lee**

Nr: 2012:06

Editor: Hasan Fleyeh

Working papers in transport, tourism, information technology and microdata analysis

ISSN: 1650-5581

© Authors

Likelihood estimate of treatment effects under selection bias

MD. MOUDUD ALAM* , MAENGSEOK NOH AND YOUNGJO LEE

We consider methods for estimating causal effects of treatment in the situation where the individuals in the treatment and the control group are self selected, i.e., the selection mechanism is not randomized. In this case, simple comparison of treated and control outcomes will not generally yield valid estimates of casual effect. The propensity score method is frequently used for the evaluation of treatment effect. However, this method is based on some strong assumptions, which are not directly testable. In this paper, we present an alternative modeling approach to draw causal inference by using share random-effect model and the computational algorithm to draw likelihood based inference with such a model. With small numerical studies and a real data analysis, we show that our approach gives not only more efficient estimates but also less sensitive to model misspecifications, which we consider, than the existing methods.

AMS 2000 SUBJECT CLASSIFICATIONS: Primary 62P20, 60K35; secondary 62J12.

KEYWORDS AND PHRASES: causal inference, h-likelihood, shared random-effect model, summer job effect.

1. INTRODUCTION

The econometric evaluation studies often have to deal with the situation where the individuals in the treatment and the control group are self selected [4], in other words selection mechanism is not randomized [35]. Often it is reasonable to believe that the individuals with better (or worse) potentiality (which may be partly unobservable) might be selected into the treatment group (or vice versa). This is known as the situation of selection (selectivity) bias [8, 18]. In addition to deal with the selection bias, an evaluation study also has to deal with the identification of the causal effects of the treatment [35].

The causal effect of a treatment (or an intervention or action) is often defined as the difference between the potential outcomes (responses) under the treatment and under control (or non-treatment) environment [33]. However, a practical difficulty with the “potential outcomes” approach is that an individual can either be treated or not treated but one can not belong to the both groups. As a consequence we

can observe one of the potential outcomes but not the both. Hence, it becomes a challenging task to identify the causal effect from the observed data.

In the literature of econometric evaluations, considering individual’s self selection into the treatment in one hand and the relevant economic theory and the policy questions of interest on the other hand [14], structural econometric modeling approach [15], which is also known as the latent index approach [18] and selectivity bias approach [8, 18], is commonly suggested. Recent applied papers on the latter approach includes [7, 25]. However, these methods are criticized for their highly sensitivity distributional assumptions [16] (hereafter HTV).

In general, the propensity score method (PSM) [33] is frequently used for the evaluation of treatment effect. Other rather less popular approaches include the full Bayesian method [25, 35], Graphical method [30] and Structural Equations approach [1].

The PSM is constructed on some strong assumptions. The most critical one being the so called strong ignorability assumption (see Assumption-1; see also [17] for detailed discussion). The strong ignorability assumption states that, given a set of observable (background) covariates, the treatment allocation is unconfounded (or ignorable) with the potential outcomes. However, a direct test of the above assumption is not offered in the literature [30]. In stead, several indirect tests ([20] and the references cited therein) and sensitivity analyses are suggested [33].

Based on the concept of potential outcomes (observed and unobservable) we present an alternative modeling approach to draw causal inference by using shared random effects [23]. Our modelling framework is closely related to those presented in [8, 16] yet more flexible and parsimonious. We also present the computational algorithm to draw likelihood based inference with such a model. The use of shared random effects approach [23] enables us to provide close-form solution of the marginal likelihood and greater flexibility than the classical probit-normal models [7, 8]. The likelihood framework enables us to draw inference on the interest parameters and model selection in a straightforward way while HTV requires parametric bootstrap or delta-method to compute standard error of the parameters.

By using simulation we show that the proposed method produces reasonable estimate of the average treatment effect (ATE). We also show that under our model assumptions, which is reasonable from econometric point of view

*Corresponding author.

[14], the propensity score method can not produce reasonable estimate of the ATE while HTV's and our method can. For probit-normal model [16] with no covariate in the response model, our approach performs almost the same as the HTV's. However, as the number of covariate in the response model increases our method becomes more efficient (for moderate sample size). Our approach is also found to be less sensitive to certain model assumptions, compared with HTV's.

The rest of the paper is organized as follows. Section 2 presents the proposed causal model and outlines its computation. Section 3 presents a short overview of the propensity score method and its extensions and discusses the similarities and differences between our approach and PSM approach. Section 4 presents simulation study to compare the performance of our approach with the already existing alternatives. Section 5 presents a real data application of proposed models and methods by estimating the causal effect of Swedish teenagers' summer job experience on their income at later age.

2. CAUSAL MODELS AND H-LIKELIHOOD INFERENCE

In this section we present two alternative ways to formulate a causal model. We also provide the explanations of the model components and relate them to the components of already known approaches. The derivation of the h-likelihood is also presented and computational methods are outlined. In this paper we present only the situation of a binary treatment allocation.

2.1 Observed data h-likelihood

Let us denote the outcome (response) of an individual i ($i = 1, 2, \dots, n$) with $y_i^{(0)}$ when individual i belongs to the control group (non-treated) and $y_i^{(1)}$ when individual i belongs to the treatment group. Suppose that the parameter of interest is

$$\tau = E\left(y_i^{(1)} - y_i^{(0)} | X_i\right).$$

Let S denote the treatment indicator with $S_i = 1$ when individual i is treatment and $S_i = 0$ when in non-treated. However, one can not be both treated and non-treated at the same time. Therefore, we can observe either $y_i^{(0)}$ or $y_i^{(1)}$ for an individual i but not the both.

Suppose that given individual characteristics (observed confounding variables), X_i , and the random effect (latent unobservable individual potentiality), u_i , the response process satisfies for $j = 0, 1$

$$(1) \quad E\left(y_i^{(j)} | X_i, u_i\right) = \begin{cases} \alpha + X_i\beta + u_i & \text{if } j = 0 \\ \alpha + \tau + X_i\beta + \omega u_i & \text{if } j = 1 \end{cases}$$

where $u_i \sim N(0, \sigma_u^2)$ and $Var(y_i^{(j)} | X_i, u_i) = \phi$. The parameter τ represents the additive treatment effect, which

is the only interesting parameter, any other parameter in the model is a nuisance parameter. Let $y_{i,o}$ be the observed response, defined by

$$(2) \quad y_{i,o} = S_i y_i^{(1)} + (1 - S_i) y_i^{(0)} = y_i^{(S_i)}$$

Here the random-effect model method cannot be used because $Var(u_i)$ and ϕ are not separable. Note here that one individual can only be either treated or not treated leaving $y_i^{(S_i)}$ observable but not $y_i^{(1-S_i)}$.

Model (1) leads to the following model for observed response, given (S_i, X_i, u_i)

$$(3) \quad E(y_{i,o} | S_i, X_i, u_i) = \alpha + X_i\beta + \tau S_i + (1 - (1 - \omega) S_i) u_i$$

and $Var(y_{i,o} | X_i, u_i, S_i) = \phi$. This model gives

$$(4) \quad E(y_{i,o} | S_i, X_i) = \alpha + X_i\beta + \tau S_i$$

and $Var(y_{i,o} | X_i, S_i) = \phi + (1 - (1 - \omega) S_i)^2 \sigma_u^2 = \kappa_{S_i}$.

In order to complete the model specification we further assume $Pr(S_i = 1 | u_i) = p_i$ with

$$(5) \quad g(p_i) = \gamma + Z_i\delta + \rho u_i$$

where Z_i is a set of covariates which may share some columns in X_i and $g(\cdot)$ is a monotonic link function.

In this paper the purpose of using random effects u_i is twofold. Here we assume the covariates X_i for the mean and Z_i for missingness indicator are known. Availability of any such background information is an advantage but it may not always be possible. In absence of them (or a subset of them), the effect of the omitted covariates is captured by u_i [6]. Therefore, the random-effect u_i removes any hidden bias [24, 33]. The random effects ensure that the two potential outcomes for the same person are correlated i.e. $Cor(y_i^{(0)}, y_i^{(1)}) \neq 0$ for $\sigma_u^2 \neq 0$. For $\omega \neq 0$ the random effects also imply that $Cor(y_i^{(1)} - y_i^{(0)}, S_i) \neq 0$ meaning that individuals might be selected into the treatment according to unobservable potential gains. The above correlations are often a matter of concern for the observational studies where the individuals with better (worse) potentiality can possibly be allocated to a certain (treatment/control) group due to the lack of randomized treatment allocation [15]. Consequently, models (4)–(5) can capture a possible correlation $Cov(y_{i,o}, S_i | X_i)$; parameter, ρ , incorporates a correlation between the treatment allocation (S_i) and observed outcome $y_{i,o}$.

When $\rho = 0$ then the response process is uncorrelated with the treatment allocation mechanism, in other words the treatment allocation is ignorable [35]. Under this situation τ can be estimated only by using the observed data (4). This is the case when the treatment allocation is completely randomized. However, in the lack of controlled randomization, τ estimated from the observed data (4) only may not

necessarily represent the causal effect [35]. If $\rho \neq 0$ then the treatment allocation mechanism is non-ignorable [35]. Therefore an ordinary least square (OLS) estimator of τ from (4) is inconsistent. For the later case Rubin [34, 35] suggested Bayesian method but we present a likelihood solution.

Let $h_{i,O}$ be the joint log-density of $(y_{i,o}, u_i, S_i)$. Based on the observed data $y_{i,o}$, the (log) h-likelihood [23] of $\psi = (\alpha, \beta, \tau, \omega, \phi, \gamma, \delta, \rho, \sigma_u^2)$ and $\mathbf{u} = (u_1, u_2, \dots, u_n)^T$ is given

$$\begin{aligned}
h_O &= \sum_{i=1}^n h_{i,O} = \sum_{i=1}^n \{ \log f_{\tau, \alpha, \phi, \sigma_u^2}(y_{i,o} | u_i, S_i) \\
&\quad + \log f_{\gamma, \rho, \sigma_u^2}(S_i | u_i) + \log f_{\sigma_u^2}(u_i) \} \\
&= -2n \log(2\pi\phi) \\
&\quad - \frac{1}{2\phi} \sum_{i=1}^n \{ y - \alpha - X_i\beta - \tau S_i - (1 - (1 - \omega) S_i) u_i \}^2 \\
&\quad + \sum_{i=1}^n \{ S_i \log(p_i) + (1 - S_i) \log(1 - p_i) \} \\
(6) \quad &- 2n \log(2\pi\sigma_u^2) - \frac{1}{2\sigma_u^2} \sum_{i=1}^n u_i^2.
\end{aligned}$$

For presentational simplicity we start with the case of $\omega = 1$ and then we extend it for $\omega \neq 1$. In order to estimate the model parameters, we integrate out the random effects, u_i 's, from (6) and obtain the following marginal log-likelihood (see Appendix A.1 for detailed derivation)

$$\begin{aligned}
\ell_o(\theta) &= -\frac{n}{2} \log(\kappa) - \frac{1}{2\kappa} \sum_{i=1}^n (y_i - \alpha - X_i\beta - S_i\tau)^2 \\
(7) \quad &+ \sum_{i=1}^n S_i \log(p_i^*) + \sum_{i=1}^n (1 - S_i) \log(1 - p_i^*)
\end{aligned}$$

where $\theta = (\tau, \alpha, \beta, \lambda, \omega, \kappa, \gamma, \delta)$, $g(p_i^*) = \eta_i^* = \frac{1}{C}(\gamma + Z_i\delta) + (y_i - \alpha - X_i\beta - S_i\tau)$, $C = \sqrt{1 + c^2 \left(\frac{\rho^2 \sigma_u^2 \phi}{\sigma_u^2 + \phi} \right)}$ with $c = \left(\frac{16\sqrt{3}}{15\pi} \right)$ for logit link and $c = 1$ for probit link, $\lambda = \frac{\rho\sigma_u^2}{\kappa C}$ and $\kappa = \phi + \sigma_u^2$. Here we notice that the individual parameters in (7) are not estimable since C always comes as a product with γ and δ and the variance components, ϕ and σ_u^2 , in $\kappa = \phi + \sigma_u^2$ are not separable. We will come back to this issue after extending the likelihood function for $\omega \neq 1$.

For $\omega \neq 1$ the above marginal likelihood becomes

$$\begin{aligned}
\ell_o(\theta^*) &= -\frac{1}{2} \sum_{i=1}^n \log(\kappa_{S_i}) \\
&\quad - \sum_{i=1}^n \frac{1}{2\kappa_{S_i}} (y_i - \alpha - X_i\beta - S_i\tau)^2 \\
(8) \quad &+ \sum_{i=1}^n S_i \log(p_i^*) + \sum_{i=1}^n (1 - S_i) \log(1 - p_i^*)
\end{aligned}$$

where $\theta^* = (\tau, \alpha, \beta, \lambda_0, \lambda_1, \omega, \kappa_0, \kappa_1, \gamma, \delta)$, $g(p_i^*) = \eta_i^* = \frac{1}{C_{S_i}}(\gamma + Z_i\delta) + \lambda_{S_i}(y_i - \alpha - X_i\beta - S_i\tau)$, $C_{S_i} = \sqrt{1 + c^2 \left(\frac{\rho^2 \sigma_u^2 \phi (1 - (1 - \omega) S_i)^2}{(1 - (1 - \omega) S_i)^2 \sigma_u^2 + \phi} \right)}$ with $c = \left(\frac{16\sqrt{3}}{15\pi} \right)$ for logit link and $c = 1$ for probit link, $\lambda_0 = \frac{\rho\sigma_u^2}{\kappa_0 C_{S_i=0}}$, $\lambda_1 = \frac{\rho\sigma_u^2}{\kappa_1 C_{S_i=1}}$ and $\kappa_{S_i} = \phi + \sigma_u^2 (1 - (1 - \omega) S_i)^2$. Again, all the parameters in (8) are not estimable for the same reason as it was for $\omega = 1$.

In order to estimate the parameters, as many as possible, we propose a Pseudo-likelihood approach [9]. From (7) and (8) we notice that if $\theta_0 = (\gamma, \delta)$ are known then the rest of the parameters in θ i.e. $\theta_1 = (\tau, \alpha, \beta, \lambda_0, \lambda_1, \omega_0, \omega_1, \kappa_0, \kappa_1)$, with $\theta = \theta_0 \cup \theta_1$, are estimable. Again, we see that after integrating out the random effects, the marginal model for selection process, S_i , follows a binary model with same functional form if g is a logit or a probit link [27]. Thus, if we estimate the selection model separately by using ML procedure, we can estimate the θ_0 up to a proportionality scale where the exact value of the proportionality constant depends on the variance of the random effects and the link function [27].

Notice that if we multiply θ_0 with a constant and divide C_{S_i} (or C , as applies) with the same constant the likelihood function does not change. Hence, we can replace θ_0 in (7) and (8) with its maximum likelihood estimate (MLE), $\hat{\theta}_0$, obtained only from the marginal selection model and estimate the remaining parameters by maximizing (7) or (8). The resulting estimate of $\hat{\theta}_1$ is a pseudo maximum likelihood (PL) estimate (since $\hat{\theta}_0$ is not the MLE of θ) but PL is known to have similar properties like a full MLE [9, 29]. Therefore using PL with the reparameterization in θ we can estimate the treatment effects parameters though the original model is not estimable.

Following [28] we can derive the asymptotic variance of $\tilde{\theta}_1$ as

$$\begin{aligned}
Asy.Var(\tilde{\theta}_1) &= \mathbf{R}_2^{-1} + \mathbf{R}_2^{-1} (\mathbf{R}_3^T \mathbf{R}_1^{-1} \mathbf{R}_3 \\
(9) \quad &\quad - \mathbf{R}_4^T \mathbf{R}_1^{-1} \mathbf{R}_3 - \mathbf{R}_3^T \mathbf{R}_1^{-1} \mathbf{R}_4) \mathbf{R}_2^{-1}
\end{aligned}$$

where $\mathbf{R}_1^{-1} = Asy.Var(\tilde{\theta}_0)$ based on the marginal likelihood of θ_0 only, $\mathbf{R}_2 = Asy.Var(\tilde{\theta}_1 | \theta_0 = \tilde{\theta}_0)$, $\mathbf{R}_3 = E \left(\frac{\partial \ell_o}{\partial \theta_0} \left(\frac{\partial \ell_o}{\partial \theta_1} \right)^T \right)$ and $\mathbf{R}_4 = E \left(\frac{\partial l_{\theta_0}}{\partial \theta_0} \left(\frac{\partial \ell_o}{\partial \theta_1} \right)^T \right)$ with l_{θ_0} being the log-likelihood function of the selection model, Model

(5), after ignoring the random effects. In practical application, \mathbf{R}_k 's, $k = 1, \dots, 4$, are replaced with their observed data counterparts (see e.g. [11], pp. 508–512).

In this paper, we also propose an approximate likelihood method based on the h-likelihood, which almost offers no computational difficulties. The idea is essentially to use Laplace approximation to approximate the marginal likelihood $\ell_o(\theta^*)$. It produces approximate MLEs of the mean parameters, restricted MLEs of the variance-covariance or dispersion parameters, and approximate standard errors based on an approximate Hessian matrix. Note that the h-likelihood h_O is equivalent to the Bayesian posterior with uniform prior $\pi(\psi) = 1$, whose joint modes may not work well. So Lee and Nelder [22] proposed various adjusted profile h-likelihoods (APHLs) for estimation of fixed parameters. Let \tilde{u}_i be the solution of equation $\partial h_{i,O}/\partial u_i = 0$, and let $D_i(h_{i,O}, u_i) = -\partial^2 h_{i,O}/\partial u_i^2$. Then, the APHL

$$\begin{aligned} p_u(h_O) &= \sum_{i=1}^n p_{u_i}(h_{i,O}) \\ (10) \quad &\equiv \sum_{i=1}^n \left[h_{i,O} - \frac{1}{2} \log \det \{ D_i(h_{i,O}, u_i) / (2\pi) \} \right] \Big|_{u_i = \tilde{u}_i} \end{aligned}$$

can be shown to be the first-order Laplace approximation to the marginal likelihood $\ell_o(\theta^*)$ by integrating out the random effects u_i .

Lee and Nelder [22] showed that the Laplace approximation (10) is identical to the Cox and Reid [5] adjusted profile likelihood to eliminate fixed parameters. Thus, we can use this form for the APHL to eliminate both fixed and random parameters simultaneously, by eliminating fixed parameters by conditioning on their MLEs and random parameters by integration. This allows a generalization of the restricted MLEs. Note that if we write $\psi = (\psi_1, \psi_2)$ where ψ_1 contains the mean parameters and ψ_2 contains the dispersion parameters. Then we typically use the Laplace approximation $p_u(h_O)$ to the marginal likelihood $\ell_o(\psi)$ for inference about the mean parameters ψ_1 . For inferences about dispersion parameters ψ_2 , we use the restricted MLEs by using $p_{\psi_1, u}(h_O) = \sum_{i=1}^N p_{\psi_1, u_i}(h_{i,O})$, which is defined similarly as in (10). Thus $\sum_{i=1}^N p_{\psi_1, u_i}(h_{i,O})$ gives an extension of the restricted log-likelihood by eliminating the mean parameters and missing covariates and random effects: see [23] for comparisons between various marginal posteriors and APHLs. Penalized quasi-likelihood and marginal quasi-likelihood methods [3] have been proposed, but they can yield serious biases. The first-order Laplace approximation often reduces bias substantially even for extreme binary data [23] and for small cluster size [12].

Because all dispersion parameters are not estimable, we set the value of $\phi = 1$. For the first-order approximation HL(1), we use $p_u(h_O)$ for ψ_1 and $p_{\psi_1, u}(h_O)$ for ψ_2 . Yun and Lee [38] have found that HL(1) estimates for ψ_1 works well, provided that HL(1) estimates for ψ_2 does not have bias.

However, HL(1) can give non-negligible biased estimators in binary data with small cluster sizes and large between-cluster variance components. In model considered here, we have the binary data with cluster size being equal to 1. For dispersion estimation Lee and Nelder [22] proposed to use the second-order Laplace approximation

$$s_{\psi_1, u}(h_O) = p_{\psi_1, u}(h_O) - \sum_{i=1}^n F_i / 24,$$

where $F_i = [-3(\partial^4 h_{i,O} / \partial u_i^4) / D_i^2(h_{i,O}, u_i) - 5(\partial^3 h_{i,O} / \partial u_i^3) / D_i^3(h_{i,O}, u_i)] \Big|_{u_i = \tilde{u}_i}$. Higher-order adjustment is useful in reducing bias, but is computationally extensive because of large number of extra terms. So it is advisable to use a lower-order adjustment, unless it gives non-ignorable biases. For HL(2) estimates, we use $p_u(h_O)$ for ψ_1 and $s_{\psi_1, u}(h_O)$ for ψ_2 .

2.2 Complete data h-likelihood

Let $\mathbf{y}_{\text{com}} = (\mathbf{y}_o, \mathbf{y}_m)$ be complete data, where $\mathbf{y}_o = (y_{1,o}, y_{2,o}, \dots, y_{n,o})^T$ is observed data and $\mathbf{y}_m = (y_{1,m}, y_{2,m}, \dots, y_{n,m})^T$ with $y_{i,m} = y_i^{(1-S_i)}$ is missing data. Let $\mathbf{y}_{\text{com}}^{(1)} = \{S_i y_i^{(1)} + (1-S_i) y_{i,m}\}$ and $\mathbf{y}_{\text{com}}^{(0)} = \{S_i y_{i,m} + (1-S_i) y_i^{(0)}\}$, then $\mathbf{y}_{\text{com}} = (\mathbf{y}_{\text{com}}^{(0)T}, \mathbf{y}_{\text{com}}^{(1)T})^T$. Now, causal inference becomes a missing data problem where 50% of the data are missing, possibly at random (MAR). Here, the complete data h-likelihood is given as

$$\begin{aligned} h_C &= -2n \log(\phi) - \frac{1}{2\phi} \sum_{i=1}^n \left(y_{i,\text{com}}^{(1)} - \alpha - X_i \beta - \tau S_i - u_i \right)^2 \\ &\quad - \frac{1}{2\phi} \sum_{i=1}^n \left(y_{i,\text{com}}^{(0)} - \alpha - X_i \beta - u_i \right)^2 \\ (11) \quad &+ \sum_{i=1}^n \log f_{\gamma, \rho, \phi_1, \sigma_u^2}(S_i | u_i) + \sum_{i=1}^n \log f_{\sigma_u^2}(u_i). \end{aligned}$$

It can be shown that $\exp[h_O] \propto \int \exp[h_C] d\mathbf{y}_m$ (see Appendix A.1 for proof). Therefore, any inferences about model parameters, based upon (6) and (11), are identical. Because σ_u^2 is not identifiable the h-likelihood procedure cannot be applied to obtain the estimator of τ . However, in the following we show that the h-likelihood estimator provides interesting insight into the meaning of the causal parameter.

Since

$$\begin{aligned} \frac{\partial h_C}{\partial \alpha} = 0 &\iff \sum_{i=1}^n \left(y_{i,\text{com}}^{(1)} - \alpha - X_i \beta - \tau S_i - u_i \right) \\ (12) \quad &+ \sum_{i=1}^n \left(y_{i,\text{com}}^{(0)} - \alpha - X_i \beta - u_i \right) = 0 \end{aligned}$$

$$(13) \quad \frac{\partial h_C}{\partial \tau} = 0 \iff \sum_{i=1}^n \left(y_{i,com}^{(1)} - \alpha - X_i \beta - \tau - u_i \right) = 0,$$

we have

$$\sum_{i=1}^n \left(y_{i,com}^{(0)} - \alpha - X_i \beta - u_i \right) = 0.$$

Substituting from (13) we have

$$\sum_{i=1}^n \left(y_{i,com}^{(1)} - y_{i,com}^{(0)} - \tau \right) = 0.$$

However, half of the complete data are missing, so that we need to impute the missing data by $\hat{y}_i^{(1-S_i)}$, the solution of $\partial h_O / \partial y_i^{(1-S_i)} = 0$;

$$\hat{y}_i^{(0)} = \alpha + X_i \beta + u_i \quad \text{and} \quad \hat{y}_i^{(1)} = \alpha + X_i \beta + \tau + u_i.$$

This gives the estimating equation for τ

$$(14) \quad \sum_{i=1}^n \left(S_i y_i^{(1)} + (1 - S_i) \hat{y}_i^{(1)} - (1 - S_i) y_i^{(0)} - S_i \hat{y}_i^{(0)} - \tau \right) = 0,$$

where Equation (14) is the missing data h-likelihood estimating equation for the causal parameter, τ . Equation (14) essentially says that the average treatment effect is simply the mean of difference between the two potential outcomes with the missing potential outcome being replaced by an imputed value of it from the h-likelihood. Anyway this estimating equation cannot be used because u_i are not estimable.

3. COMPARISON WITH THE PROPENSITY SCORE METHODS

In the model (5), the marginal probability of getting the treatment becomes

$$p(Z_i) = Pr(S_i = 1 | Z_i) = \int p_i f_{\sigma_u^2}(u_i) du_i,$$

which is called the propensity score. In order to identify the treatment effect τ , PSM approach adopts the so called ‘‘strong ignorability assumption’’ (see Assumption-1).

Assumption-1: (Strong ignorability assumption) Given a set of background information, Z_i , the potential outcomes are unconfounded with the actual treatment assignment i.e.

$$E(y_i^{(0)} | Z_i, X_i, S_i = 0) = E(y_i^{(0)} | X_i)$$

and

$$E(y_i^{(1)} | Z_i, X_i, S_i = 1) = E(y_i^{(1)} | X_i).$$

Assumption-1 is also known as ‘‘conditional independence’’ and ‘‘unconfoundedness’’ assumption. It is worth noting that a part of Assumption-1 is always implicit in the model specification (2)–(5) in that $E(y_i^{(s)} | X_i, S_i = s) = E_{u_i} \left(E(y_i^{(s)} | X_i, u_i, S_i = s) \right) = E(y_i^{(s)} | X_i) \quad \forall s = 0, 1$. Strong ignorability assumption implies the conditional independence i.e., $(y_i^{(0)}, y_i^{(1)}) \perp S_i | (X_i, Z_i)$ [17]. A direct consequence of the conditional independence is that, $E(S_i | y_i^{(S_i)}, X_i, Z_i) = E(S_i | Z_i)$. This does not hold for the model presented in Section 2, since $E(y_i^{(S_i)} | S_i, X_i, u_i)$ and $E(S_i | Z_i, u_i)$ shares the same random effect u_i .

Under Assumption-1 τ is estimated in the following way.

$$E(y_i^{(1)} S_i | X_i, Z_i) = E(y_i^{(1)} | S_i = 1) p(Z_i)$$

$$E(y_i^{(0)} (1 - S_i) | X_i, Z_i) = E(y_i^{(0)} | X_i, S_i = 0) (1 - p(Z_i))$$

Since

$$\begin{aligned} & E(y_i^{(1)} | X_i, S_i = 1) - E(y_i^{(0)} | X_i, S_i = 0) \\ &= E(y_i^{(1)} | X_i) - E(y_i^{(0)} | X_i) \\ &= E(y_i^{(1)} - y_i^{(0)} | X_i) \\ &= \tau \end{aligned}$$

due to the strong ignorability assumption (Assumption-1), we have the following unbiased estimating equation for τ

$$(15) \quad \sum_{i=1}^n \left(\frac{y_i^{(1)} S_i}{p(Z_i)} - \frac{(1 - S_i) y_i^{(0)}}{1 - p(Z_i)} - \tau \right) = 0$$

The resulting PSM estimator is also known as the inverse propensity weighted (IPW) estimator [10, 32]. Standard error estimates of PSM estimator are given in [2, 19].

The PSM estimator from (15) avoids the unidentifiable missing estimation $\hat{y}_i^{(1-S_i)}$ in (14). However, it requires to know $p(Z_i)$. If $p(Z_i)$ is unknown and is estimated from the data, PSM estimating equation (15) does not necessarily give an unbiased estimator, but the estimator is still efficient and asymptotically normally distributed, under some reasonable assumptions, along with the strong ignorability assumption [19].

PSM is very easy to understand and implement. This is why it has received so much attention. However, the strong ignorability condition is often criticized [30]. There is no statistical tool to assess the validity of the strong ignorability assumption. It is also found that the PSM can be biased if we observe only $X_{i,o}$ such that $X_{i,o} \subset X_i$ while S_i depends on the whole set of covariates in X_i [36]. Moreover, it works without requiring the strong ignorability condition.

In application, PSM estimator often turns out to be unplausible, in that the estimator exceeds any reasonable

bound, if some observations have extreme (closed to 0 or 1) propensity score. In those cases the matching method [33] and normalization of the inverse propensity weight (NAIPW) [20], by restricting the sum of the weights to one, are proposed in the literature. The normalization approach gives the following estimator

$$\hat{\tau}_{NIPW} = \left(\sum_{i=1}^n \frac{S_i}{\hat{p}(X_i)} \right)^{-1} \sum_{i=1}^n \left(\frac{y_i^{(1)} S_i}{\hat{p}(Z_i)} \right) - \left(\sum_{i=1}^n \frac{1 - S_i}{1 - \hat{p}(X_i)} \right)^{-1} \sum_{i=1}^n \left(\frac{(1 - S_i) y_i^{(0)}}{1 - \hat{p}(Z_i)} \right)$$

Glynn and Quinn [10] suggested further improvement to τ_{NIPW} by using the predictive information in X_i about y_i . They [10] call their estimator as the Augmented Inverse Propensity Weighted (AIPW) estimator which is given as

$$\begin{aligned} \hat{\tau}_{AIPW} = & \frac{1}{n} \sum_{i=1}^n \left[\left(\frac{y_i^{(1)} S_i}{\hat{p}(Z_i)} - \frac{(1 - S_i) y_i^{(0)}}{1 - \hat{p}(Z_i)} \right) - \frac{S_i - \hat{p}(Z_i)}{\hat{p}(Z_i) \{1 - \hat{p}(Z_i)\}} \right. \\ & \left. \left(\{1 - \hat{p}(Z_i)\} \hat{E} \left(y_i^{(1)} | X_i, S_i = 1 \right) \right. \right. \\ (16) & \left. \left. + \hat{p}(Z_i) \hat{E} \left(y_i^{(0)} | X_i, S_i = 0 \right) \right) \right] \end{aligned}$$

In application, we might be unsure about the true X_i and Z_i . Therefore, Glynn and Quinn [10] suggested functional models (e.g. generalized additive models (GAM) [13]) for modelling $S_i | (X_i, Z_i)$, $y_i^{(0)} | (X_i, S_i = 0)$ and $y_i^{(1)} | (X_i, S_i = 1)$. Note that, for AIPW, the response and the selection model do not have to be the true model. Suffice it to have a selection model that ensures consistent estimate of $p(Z_i)$ and that Assumption-1 holds. The response model can only improve the estimation if it has some predictive power, does not matter how much.

AIPW has the following favourable properties. Firstly, it preserves the so called doubly-robust property in that it is consistent only if one of the processes, either the response y or the selection S , is correctly modelled. Secondly, it overcomes the problems due to extreme $\hat{p}(Z_i)$, to some extent [10]. Thirdly, possible non-linearity in any of the models components can be captured by using a suitable GAM. Finally, under the same conditions necessary for the validity of the PSM estimator, the AIPW estimator is asymptotically normally distributed and attains the so called non-parametric efficiency bound [10]. However, if the propensity score estimates are highly variable, the AIPW can be inefficient in small sample.

Our method do not need the strong ignoreability assumption i.e., it works under $E(S_i | y_{i,o}, Z_i) \neq E(S_i | Z_i)$. The ML estimator is efficient if the assumed model is true while the PSM estimator is robust. The advantage of our method over the full Bayesian method [25, 34, 35] is that we do not

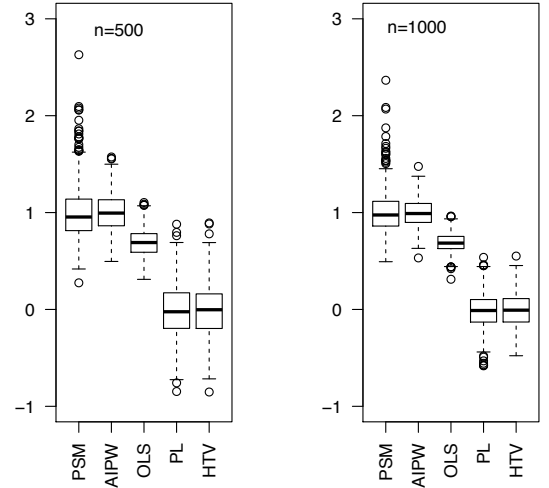


Figure 1. Bias and variation in ATE estimates by different methods

need any subjective prior and we can estimate the parameters, and their standard errors, analytically while a Bayesian might have to rely on the time consuming Markov-chain Monte-Carlo (MCMC) simulation [25, 26].

4. SIMULATION STUDY

In order to assess the performances of the different estimators of the causal effects, presented in Section 2 and Section 3, we conduct a series of simulation studies. Since our method is closely related to the HTV's we conduct a simulation study with a setting similar to HTV and then we examine situations with more covariates and functional misspecification of the selection model.

For the first simulation study we generate data with $y_{i,o} | S, u_i \sim N(\alpha + \tau S_i + (1 - (1 - \omega) S_i) u_i, 1.5)$, $\Phi^{-1}(p_i) = \gamma + \delta_1 X_{1,i} + \rho u_i$, $X_i \sim N(0, 1)$, $u_i \sim N(0, 0.75)$, $(\alpha, \tau, \gamma, \beta) = (2, 1, 0, 1)$ are chosen according to HTV and $\omega = 1.5$. This gives the same construction as HTV (see Appendix A.2) except for the specific restrictions on the variance and covariances (e.g. $Cov(y_i^{(0)}, y_i^{(1)}) = 0$ in HTV). At each Monte-Carlo iteration we simulate $n = 500$ and 1000 observations and we use 500 Monte-Carlo replication. The results are summarized in terms of $\hat{\tau} - \tau$ in Figure 1.

All the simulations are conducted in R [31]. We use our own programme to calculate the PL, PSM and HTV estimate. AIPW is implemented via `estimate.ATE` function from `CausalGAM` library [10]. We use default spline model for response and selection model in AIPW.

Figure 1 shows that the ATE estimate of PSM, AIPW and OLS are highly biased. The other two methods HTV

and PL are unbiased and their performances are not distinguishable. The results are well-understandable. Since, the strong ignorability condition does not hold, AIPW and PSM is not expected to be unbiased. The OLS is also supposed to be biased as the covariate S_i in the linear model $y_{i.o} = a + bS_i + \varepsilon_i$ is correlated with the error term. However, it is nice to see that both the PL and the HTV overcomes any bias. Since there are not many covariate, closeness between HTV and PL is also understandable. This situation might change as the number of covariate in the response model increases.

For the second simulation study, we generate data with sample sizes $n = 250, 500$ and 1000 , two independent background variables, $X_j \sim N(0, 1); \forall j = 1, 2$, $\alpha = 1.8$, $\beta = (-1.6, 0.8)^T$, $\tau = 1.9$, $\phi = 1.5$, $\sigma_u^2 = 0.75$, $\gamma = -0.7$, $\delta = (1.4, -0.9)^T$ and $\rho = -0.7$. In simulation we consider $E(y_{i.o}|X_i, S_i, u_i) = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \tau S_i + u_i$ and $\Phi^{-1}(p_i) = \gamma + \delta_1 X_{1i} + \delta_2 X_{2i} + \rho u_i$ where Φ represents a standard normal CDF. For PSM method we estimate the propensity score with a probit model with both X_1 and X_2 as the covariate. We implement improved IPW estimate (AIPW) [10] with both the covariates through a Spline model for the means of the y_i and S_i . We also carry out a simple regression (OLS) estimator of τ by regressing S , X_1 and X_2 on y_o .

The results show the same pattern as the one in Figure 1. Hence, we do not report all the detailed results. Only difference in the results of the second simulation was that the HTV turned out to be inefficient (see MSE and Bias in Table 1) compared to PL, HL(1) and HL(2). Other methods (AIPW and OLS) are biased, as expected.

For the third simulation, we generate data in the same way as the second simulation except that we use a logit link for the selection model where as a probit link was used in second simulation. It is worth noting that HTV requires the error-distributions of both the response and the selection models to be known. For probit case the errors in response and selection models follow multivariate normal distribution. However, in logit case we do not know the joint distribution of the error terms. Therefore, we compute the ATE for HTV by treating the error distribution of the selection models are both normal or logistic (neither is correct). We also examine the effect of the miss-specification of the link function for the PL, HL(1) and HL(2), too. Since we already know that the PSM, AIPW and OLS are biased we do not include them in the third simulation. We only report the bias and the mean squared errors of PL, HTV, HL(1) and HL(2) under different specification of the link function (Table 1).

In Table 1, the first case is from the second simulation (True Link: Probit) and the rests from the third simulation (True Link: Logit). When the true link is probit, HTV estimator is unbiased but the biases of PL, HTV and HL(2) decrease rapidly as sample size increase. HTV has the largest MSE, while HL(1) tends to have the smallest. When the link

is misspecified bias of HTV does not decrease as n grows and has the largest MSE. HL(1) has generally good statistical property with computational efficiency.

5. A REAL DATA APPLICATION: ESTIMATION OF THE SUMMER JOB EFFECTS

In most European countries, the unemployment rate of youths keeps growing. Since, one of the causes for high youth unemployment rate would be the difficulty in getting the first job, a remedy to provide the school leavers with work experience is often suggested. Therefore, it is interesting to examine the causal effect of early summer job experience on the future unemployment of the high school graduates. To check this hypothesis, we create the data set by merging several data bases maintained by Statistics Sweden (SCB). Main source is LOUISE data base (renamed as LISA since 2004; see <http://www.scb.se/>) which contain longitudinal information on earnings and demographics. We use a random sample of 5% individuals from LOUISE database between 1995 and 2002. A teenager (age between 16 and 19 years) is defined as a summer jobber if (s)he had any gainful employment during June-July. We add further information on teenagers' middle and high school grades and parents' social and economic status from other databases maintained by SCB. We use OLS, PL, AIPW and HL methods to estimate average summer-job effect. The response is $\log(1 + income)$ with the observed income being adjusted for inflation using year 2002 as the base. We fit a logistic model with 8 covariates (measured at age 19) to obtain the probability to join the summer job. Outcome model under OLS, PL, AIPW and HL used 12 covariates (OLS also included a dummy for summer job).

We split the data according to age in years (19-23). The PL and HL results (see Table 2) show that the summer job experience has a short term positive effect on income but this effect is wiped out after age 22. OLS PL and HL estimates of summer job effect do not differ very much from each other except for the age group 22 years where PL shows a negative effect which again turns out positive but insignificant in the next age group. Therefore, we may conclude that summer job experience does not have a consistent log term effect on future income. The results from PL might look rather striking but it is consistent with the results of [37] though they used experimental data from only one municipality in Sweden.

ACKNOWLEDGEMENTS

First author's contribution to this work was initiated during his post-doctoral visit at Seoul National University. This work was partially supported by the National Research Foundation (NRF) of Korea grant funded by the Korean government (MEST) (No. 2012-0009183).

Table 1. Bias and MSE of PL and HTV under different misspecification of the link function in the selection model.

True link/Assumed Link	Method	Bias			MSE		
		n=250	n=500	n=1000	n=250	n=500	n=1000
Probit/Probit	PL	-0.15*	0.00	0.00	0.77	0.27	0.14
	HTV	-0.04	0.02	0.00	1.19	0.91	0.45
	HL(1)	-0.17*	-0.14*	-0.10*	0.11	0.079	0.041
	HL(2)	-0.07*	-0.05*	0.01	0.14	0.071	0.057
Logit/Probit	PL	-0.38*	-0.11*	-0.10*	1.37	0.77	0.47
	HTV	-0.20*	-0.15*	-0.20*	6.09	2.82	1.34
	HL(1)	-0.14*	-0.16*	-0.13*	0.079	0.046	0.042
	HL(2)	0.08*	0.07*	0.02	0.12	0.058	0.026
Logit/Logit	PL	-0.30*	-0.09*	-0.03	1.13	0.82	0.49
	HTV	-0.22*	-0.14*	-0.18*	6.50	2.58	1.24
	HL(1)	-0.07*	-0.05*	-0.04*	0.098	0.047	0.029
	HL(2)	0.02	0.01	-0.01	0.15	0.087	0.051

* Significantly different from 0 at 5% level.

Table 2. Estimated summer job effect.

Age	n	OLS	PL	AIPW	HL(1)	HL(2)
19	10625	0.62	0.60	0.61	0.62	0.62
		(0.046)	(0.058)	(0.049)	(0.049)	(0.049)
20	10593	0.25	0.19	0.24	0.25	0.25
		(0.045)	(0.056)	(0.046)	(0.047)	(0.047)
21	8027	0.24	0.32	0.24	0.26	0.25
		(0.050)	(0.065)	(0.052)	(0.052)	(0.052)
22	5340	0.16	-0.26	0.14	0.18	0.19
		(0.062)	(0.049)	(0.063)	(0.064)	(0.064)
23	2657	0.16	0.20	0.15	0.16	0.16
		(0.087)	(0.228)	(0.084)	(0.089)	(0.089)

Note: Values within the parenthesis show standard errors

REFERENCES

- BOLLEN, K. A. (1989). *Structural Equations with Latent Variables*, Wiley, New York. [MR0996025](#)
- BRAVO, F. and JACHO-CHAVEZ, D. T. (2011). Empirical likelihood for efficient semiparametric average treatment effects. *Econometric Rev.* **30** 1–24. [MR2747366](#)
- BRESLOW, N. E. and CLAYTON, D. (1993). Approximate inference in generalized linear mixed models. *J. Amer. Statist. Assoc.* **88** 9–25
- COBB-CLARK, D. A. and CROSSLEY, T. (2003). Econometrics for evaluations: an introduction to recent development. *Econ. Rec.* **79** 491–511.
- COX, D. R. and REID, N. (1987). Parameter orthogonality and approximate conditional inference. *J. Roy. Statist. Soc. Ser. B* **49** 1–39. [MR0893334](#)
- COX, D. R. and WONG, M. Y. (2010). A note on the sensitivity to assumptions of a generalised linear mixed model. *Biometrika* **97** 209–214. [MR2594428](#)
- DAGSVIK, J. K., HAEGELAND, T. and RAKNERUD, A. (2011). Estimating the returns to schooling: a likelihood approach based on normal mixture. *J. Appl. Econometrics* **26** 613–640. [MR2828969](#)
- GAREN, J. (1984). A selectivity bias approach with a continuous choice variable. *Econometrica* **52** 1199–1218.
- GONG, G. and SAMANIEGO, F. J. (1981). Pseudo maximum likelihood estimation: theory and applications. *Ann. Stat.* **9** 861–869.
- GLYNN, A. N. and QUINN, K. M. (2009). An introduction to the augmented inverse propensity weighted estimator. *Polit. Anal.* **18** 36–56.
- GREENE, W. H. (2003). *Econometric Analysis*, Pearson Education, Upper Saddle River, N.J.
- JOE, H. (2008). Accuracy of Laplace approximation for discrete response mixed models. *Comput. Statist. Data Anal.* **52** 5066–5074. [MR2526575](#)
- HASTIE, T. and TIBSHIRANI, R. (1986). Generalized Additive Models (with discussion). *Statist. Sci.* **1** 297–318. [MR0858512](#)
- HECKMAN, J. J. (2008). Econometric causality. *Int. Stat. Rev.* **76** 1–27.
- HECKMAN, J. J. and VYTLACIL, E. (2005). Structural equations, treatment effects, and econometric policy evaluation. *Econometrica* **73** 669–738. [MR2135141](#)
- HECKMAN, J. J., TOBIAS, J. L. and VYTLACIL, E. (2000). Simple estimator for treatment parameters in a latent variable framework with application to estimating the returns to schooling, Working Paper 7950, NBER Working Paper Series, National Bureau of Economic Research, Cambridge, MA.
- HECKMAN, J. J., ICHIMURA, H. and TODD, P. E. (1997). Matching as an econometric evaluation estimator: evidence from evaluating a job training programme. *Rev. Econom. Stud.* **16** 605–654. [MR1623713](#)
- HECKMAN, J. J. (1976). The common structure of statistical models of truncation, sample selection, and limited dependent variables and a simple estimator for such models. *Ann. Econ. Soc. Meas.* **5** 475–492.
- HIRANO, K., IMBENS, G. W. and RIDDER, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* **71** 1161–1189. [MR1995826](#)
- IMBENS, G. (2004). Nonparametric estimation of average treatment effects under exogeneity: a review. *Rev. Econ. Stat.* **86** 4–29.
- JOHNSON, N. L. and KOTZ, S. (1970). *Distributions in Statistics: Continuous Univariate Distributions-2*, Wiley, New York.
- LEE, Y. and NELDER, J. A. (2001). Hierarchical generalised linear models: a synthesis of generalised linear models, random-effect models and structured dispersions. *Biometrika* **88** 987–1006. [MR1872215](#)
- LEE, Y., NELDER, J. A. and PAWITAN, Y. (2006). *Generalized Linear Models with Random Effects: Unified Analysis via H-likelihood*, Chapman and Hall/CRC, Boca Raton, FL. [MR2259540](#)
- LEE, Y. and NOH, M. (2012). Parameterisation and sensitivity in hierarchical generalised linear models. Submitted for publication.
- LI, M. and TOBIAS, J. L. (2008). Modelling and evaluating treatment effects in econometrics. *Adv. Econometrics* **21** 57–91. [MR2556002](#)
- LI, M. and TOBIAS, J. L. (2011). Bayesian inference in a correlated random coefficients model: modeling causal effect heterogeneity with an application to heterogeneous returns to schooling. *J.*

Econometrics **162** 345–361. [MR2795622](#)

- [27] LIANG, K. Y. and ZEGER, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73** 13–22. [MR0836430](#)
- [28] MURPHY, K. M. and TOPEL, R. H. (1985). Estimation and inference in two-step econometric models. *J. Bus. Econom. Statist.* **3** 370–379.
- [29] PARKE, W. R. (1986). Pseudo maximum likelihood estimation: the asymptotic distribution. *Ann. Statist.* **14** 355–357. [MR0829575](#)
- [30] PEARL, J. (2009). *Causality: Models, Reasoning and Inference*, Cambridge University Press, New York. [MR2548166](#)
- [31] R DEVELOPMENT CORE TEAM (2011). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org/>.
- [32] ROBINS, J. M., ROTNITZKY, A. and ZHAO, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed models. *J. Amer. Statist. Assoc.* **89** 846–866 [MR1294730](#)
- [33] ROSENBAUM, P. A. and RUBIN, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55. [MR0742974](#)
- [34] RUBIN, D. (2005). Causal inference using potential outcomes: design, modeling, decisions. *J. Amer. Statist. Assoc.* **100** 322–331. [MR2166071](#)
- [35] RUBIN, D. (1978). Bayesian inference for causal effect: the role of randomization. *Ann. Statist.* **6** 34–58. [MR1429927](#)
- [36] SMITH, J. A. and TODD, P. E. (2005). Does matching overcome LaLonde’s critique of nonexperimental estimators? *J. Econometrics* **125** 305–353. [MR214337](#)
- [37] WANG, Y., CARLING, K. and NÄÄS, O. (2006). High school students’ summer jobs and their ensuing labour market achievement, IFAU working paper 14:2006, Institute for Labour Market Policy Evaluation (IFAU), Uppsala, Sweden.
- [38] YUN, S. and LEE, Y. (2004). Comparison of hierarchical and marginal likelihood estimators for binary outcomes. *Comput. Statist. Data Anal.* **45** 639–650. [MR2055468](#)
- [39] ZEGER, S. L., LIANG, K. L. and ALBERT, P. S. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics* **44** 1049–1060. [MR0999450](#)

APPENDIX A

A.1 Derivation of equation (7):

Let $T_i = 1 - S_i$. First note that

$$\begin{aligned}
 \exp(h_O) &= \prod_{i=1}^n \int \exp(h_M) dy_i^{(T_i)} \\
 &= \prod_{i=1}^n \int f(y_{i,com}^{(0)} | S_i, u_i) f(y_{i,com}^{(1)} | S_i, u_i) f(S_i, u_i) dy_i^{(T_i)} \quad (19) \\
 &= \prod_{i=1}^n \int f(y_i^{(S_i)} | S_i, u_i) f(S_i, u_i) f(y_i^{(T_i)} | S_i, u_i) dy_i^{(T_i)} \\
 &= \prod_{i=1}^n f(y_i^{(S_i)} | S_i, u_i) f(S_i, u_i) \int f(y_i^{(T_i)} | S_i, u_i) dy_i^{(T_i)} \\
 &= \prod_{i=1}^n f(y_i^{(S_i)} | S_i, u_i) f(S_i | u_i) f(u_i).
 \end{aligned}$$

Secondly,

$$\begin{aligned}
 L &= \exp(\ell) = \int \exp h_O du \\
 (17) \quad &= \prod_{i=1}^n \int_{-\infty}^{\infty} f(y_i^{(S_i)} | S_i, u_i) f(S_i | u_i) f(u_i) du_i.
 \end{aligned}$$

Now, with $y_i^{(S_i)} | u_i, S_i \sim N(\mu_i + u_i, \phi)$, $\mu_i = \alpha + X_i \beta + \tau S_i$, $S_i | u_i \sim \text{Bin}(1, p_i)$, $g(p_i) = \eta_i = \gamma + Z_i \delta + \rho u_i$ and $u_i \sim N(0, \sigma_u^2)$, we have the following simplifications.

$$\begin{aligned}
 L &= L(\tau, \alpha, \phi, \sigma_u^2, \gamma, \rho) \\
 &= \prod_{i=1}^n \int \frac{1}{\sqrt{2\pi\phi}} \exp\left[-\frac{(y_i^{(S_i)} - \mu_i - u_i)^2}{2\phi}\right] \\
 &\quad \times \frac{1}{\sqrt{2\pi\sigma_u^2}} \exp\left[-\frac{u_i^2}{2\sigma_u^2}\right] f(S_i | u_i) du_i \\
 &= \prod_{i=1}^n \left(\frac{1}{\sqrt{2\pi(\phi + \sigma_u^2)}} \exp\left[-\frac{1}{2(\phi + \sigma_u^2)} (y_i^{(S_i)} - \mu_i)^2\right] \right) \\
 &\quad \times \prod_{i=1}^n \int \frac{\sqrt{(\phi + \sigma_u^2)}}{\sqrt{2\pi\phi\sigma_u^2}} \exp\left[-\frac{(\phi + \sigma_u^2)}{2\phi\sigma_u^2} \left(u_i - \frac{\sigma_u^2}{(\phi + \sigma_u^2)} (y_i^{(S_i)} - \mu_i)\right)^2\right] f(S_i | u_i) du_i.
 \end{aligned} \quad (18)$$

But, $f(S_i | u_i) = p_i$ if $S_i = 1$ and $f(S_i | u_i) = 1 - p_i$ if $S_i = 0$. Hence, the integral term in (18) is $(E_{u_i}(p_i))^{S_i} (1 - E_{u_i}(p_i))^{T_i}$ where the expectation is taken with respect to the new distribution of u_i being $u_i \sim N\left(\frac{\sigma_u^2}{(\phi + \sigma_u^2)} (y_i^{(S_i)} - \mu_i), \frac{\phi\sigma_u^2}{(\phi + \sigma_u^2)}\right)$. For probit link, i.e. $g(p_i) = \Phi^{-1}(p_i)$ with Φ being a standard normal distribution function, we can calculate the above expectation as

$$\begin{aligned}
 E_{u_i}(p_i) &= E_{u_i}(E(S_i | u_i)) \\
 &= E_{u_i}(\Phi(\gamma + Z_i \delta + \rho u_i)) \\
 &= E_{u_i}(Pr(\gamma + Z_i \delta + \rho u_i \geq \epsilon_i)) \\
 &= Pr(\gamma + Z_i \delta \geq \epsilon_i - \rho u_i) \\
 &= Pr\left(\gamma + Z_i \delta + \frac{\rho\sigma_u^2}{(\phi + \sigma_u^2)} (y_i^{(S_i)} - \mu_i) \geq \epsilon_i^*\right)
 \end{aligned} \quad (19)$$

where

$$\begin{aligned}
 \epsilon_i &\sim N(0, 1), \\
 u_i^* &= \left(u_i - \frac{\sigma_u^2}{(\phi + \sigma_u^2)} (y_i^{(S_i)} - \mu_i)\right) \sim N\left(0, \frac{\phi\sigma_u^2}{(\phi + \sigma_u^2)}\right)
 \end{aligned}$$

and

$$\epsilon_i^* = \epsilon_i - \rho u_i^* \sim N\left(0, 1 + \frac{\rho^2 \phi \sigma_u^2}{(\phi + \sigma_u^2)}\right).$$

Hence we obtain from (21)

$$(20) \quad E_{u_i}(p_i) = \Phi \left(\frac{\gamma + Z_i \delta + \frac{\rho \sigma_u^2}{(\phi + \sigma_u^2)} (y_i^{(S_i)} - \mu_i)}{\sqrt{\left(1 + \frac{\rho^2 \phi \sigma_u^2}{(\phi + \sigma_u^2)}\right)}} \right)$$

For logit link, i.e. when $g(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \eta_i$, using the relation between standard normal and standard logistic CDF [21], i.e. $(1 + \exp[-\pi x/\sqrt{3}])^{-1} \simeq \Phi(16x/15)$, we obtain

$$\begin{aligned} p_i &= \left(1 + \exp\left[-\frac{\pi}{\sqrt{3}} \left(\frac{\eta_i \sqrt{3}}{\pi}\right)\right]\right)^{-1} \\ &= \Phi\left(\frac{16}{15} \left(\frac{\eta_i \sqrt{3}}{\pi}\right)\right) \\ &= \Pr\left(\gamma + Z_i \delta + \rho u_i \geq \frac{15\pi}{16\sqrt{3}} \epsilon_i\right) \end{aligned}$$

so that

$$\begin{aligned} E_{u_i}(p_i) &= \Phi \left(\frac{\gamma + Z_i \delta + \frac{\rho \sigma_u^2}{(\phi + \sigma_u^2)} (y_i^{(S_i)} - \mu_i)}{\sqrt{\left(\left(\frac{15\pi}{16\sqrt{3}}\right)^2 + \frac{\rho^2 \phi \sigma_u^2}{(\phi + \sigma_u^2)}\right)}} \right) \\ &= \Phi \left(\frac{16}{15} \left(\frac{15}{16} \left\{ \frac{\gamma + Z_i \delta + \frac{\rho \sigma_u^2}{(\phi + \sigma_u^2)} (y_i^{(S_i)} - \mu_i)}{\sqrt{\left(\left(\frac{15\pi}{16\sqrt{3}}\right)^2 + \frac{\rho^2 \phi \sigma_u^2}{(\phi + \sigma_u^2)}\right)}} \right\} \right) \right) \\ &= \left(1 + \exp \left[-\frac{\pi}{\sqrt{3}} \frac{15}{16} \frac{\gamma + Z_i \delta + \frac{\rho \sigma_u^2}{(\phi + \sigma_u^2)} (y_i^{(S_i)} - \mu_i)}{\sqrt{\left(\left(\frac{15\pi}{16\sqrt{3}}\right)^2 + \frac{\rho^2 \phi \sigma_u^2}{(\phi + \sigma_u^2)}\right)}} \right] \right)^{-1} \\ &= \left(1 + \exp \left[-\frac{\gamma + Z_i \delta + \frac{\rho \sigma_u^2}{(\phi + \sigma_u^2)} (y_i^{(S_i)} - \mu_i)}{\sqrt{\left(1 + \left(\frac{16\sqrt{3}}{15\pi}\right)^2 \frac{\rho^2 \phi \sigma_u^2}{(\phi + \sigma_u^2)}\right)}} \right] \right)^{-1}. \end{aligned}$$

Substituting these results into (18) we obtain

$$L = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}(\phi + \sigma_u^2)} \exp \left[-\frac{(y_{i,o} - \mu_i)^2}{2(\phi + \sigma_u^2)} \right] (p_i^*)^{S_i} (1 - p_i^*)^{T_i}$$

where,

$$g(p_i^*) = \eta^* = \frac{1}{C} \left(\eta + \frac{\rho \sigma_u^2}{(\phi + \sigma_u^2)} (y_{i,o} - \mu_i) \right)$$

with $C = \sqrt{1 + c^2 \frac{\rho^2 \phi \sigma_u^2}{(\phi + \sigma_u^2)}}$ where $c = 1$ for probit link and $c = \frac{16\sqrt{3}}{15\pi}$ for logit link. The integral for p_i^* is exact for probit link but it is an approximation, yet very accurate, for logit link (see [39] for further discussion).

A.2 Heckman, Tobias and Vytlačil's [16] modeling framework and their estimation technique (HTV)

Let,

$$(21) \quad \begin{aligned} y^{(1)} &= X\beta^{(1)} + U^{(1)} \\ y^{(0)} &= X\beta^{(0)} + U^{(0)} \end{aligned}$$

$$D_i^* = Z\delta + U^D$$

where $\beta^{(0)}$, $\beta^{(1)}$ and δ are model parameters and $U^{(0)}$, $U^{(1)}$ and U^D are zero mean error terms which follow a joint trivariate distribution. Denote $Var(U^{(0)}) = \sigma_0^2$, $Var(U^{(1)}) = \sigma_1^2$, $Cor(U^{(0)}, U^D) = \varrho_0$, $Cor(U^{(1)}, U^D) = \varrho_1$ and $Cor(U^{(0)}, U^{(1)}) = \varrho_{10}$.

The selection rule assigns people to the treatment ($S_i = 1$) if $U_i^D \geq -Z_i \delta$. This is equivalent to setting $S_i = 1$ when $J(U_i^D) \geq J(-Z_i \delta)$ for some strictly increasing function J . Suppose that $U^D \sim F$, where F is an absolutely continuous distribution function which can be non-normal but its density function is symmetric about 0. Define $J_\Phi(U^D) \approx \tilde{U}^D$ and $J_\Phi(u) = \Phi^{-1}F(u)$. The this model (21) gives the following selection-correction conditional mean functions

$$(22) \quad E(y^{(1)} | S(Z) = 1, X = x, Z = z) = x\beta^{(1)} + \varrho_1 \sigma_1 \frac{\phi(J_\Phi(Z\delta))}{F(Z\delta)}$$

and

$$(23) \quad E(y^{(0)} | S(Z) = 1, X = x, Z = z) = x\beta^{(0)} + \varrho_0 \sigma_0 \frac{\phi(J_\Phi(Z\delta))}{F(Z\delta)}$$

where, ϕ is the standard normal pdf. For $F = \Phi$ we have $\phi(J_\Phi(Z\delta)) = \Phi(Z\delta)$. If we also assume the other two error terms to be normal, we have a tri-variate normal distribution of the error terms leading to the same model as the one in Section 2 with probit link.

The ATE is given as

$$(24) \quad ATE(x) = E(y^{(1)} - y^{(0)} | X = x) = x(\beta^{(1)} - \beta^{(0)})$$

Estimation of the model parameters is carried out in the following way.

1. Obtain $\hat{\delta}$ from a binary choice model using F as the distribution of U^D .
2. Compute appropriate selection correction terms by using (22) and (23) evaluated at $\hat{\delta}$.
3. Run treatment-specific regression for groups $S = 1$ and $S = 0$ separately.
4. Given, $\hat{\beta}^{(0)}$, $\hat{\beta}^{(1)}$, $\hat{\varrho}_1\hat{\sigma}_1$ and $\hat{\varrho}_0\hat{\sigma}_0$ from step 3 we estimate the ATE as

$$(25) \quad \widehat{ATE} = \frac{1}{n} \sum_{i=1}^n x_i \left(\hat{\beta}^{(1)} - \hat{\beta}^{(0)} \right)$$

This is the ATE of HTV implemented in the simulation.

Notice that HTV estimate of ATE quickly becomes inefficient as the number of columns in X increases with fixed sample size. It requires the marginal distribution of U^D to be known exactly in order to compute the selection corrected means in (22)-(23). Finally, if $\hat{\delta}$ is not an unbiased

estimate, which is the case if $\hat{\delta}$ is the MLE and the sample size is small, then (22) and (23) are not unbiased estimating equations.

Md. Moudud Alam
 School of Technology and Business Studies, Dalarna University,
 Falun, Sweden 791 80
 E-mail address: maa@du.se

Maengseok Noh
 Department of Statistics, Pukyong National Univeristy,
 Busan, Korea 608-737
 E-mail address: msnoh@pknu.ac.kr

Youngjo Lee
 Department of Statistics, Seoul National Univeristy,
 Seoul, Korea, 151-742
 E-mail address: youngjo@snu.ac.kr