Statistics and Its Interface Volume 6 (2013) 349–359

Likelihood estimate of treatment effects under selection bias

Md. Moudud Alam, Maengseok Noh and Youngjo Lee*

We consider methods for estimating the 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, a 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 modelling approach to draw causal inferences by using a shared 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 is less sensitive to model misspecifications, which we consider, than existing methods.

AMS 2000 subject classifications: Primary 62P20, 60K35; secondary 62J12.
Keywords and phrases: Causal inference, Likelihood, Propensity score, Random-effect model.

1. INTRODUCTION

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 the 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 to the treatment group. This is known as the situation of selection (selectivity) bias [8, 19]. 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 a 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 both groups. As a consequence we can observe one of the potential outcomes but not both. Hence, it becomes a challenging task to identify the causal effect from the observed data.

In literature on econometric evaluations, considering an individual's self selection into the treatment on one hand and the relevant economic theory and the policy questions of interest on the other hand [14], structural econometric modelling approach [15], which is also known as the latent index approach [19], latent variable framework [16] (hereafter HTV) and selectivity bias approach [8, 19], is commonly suggested. Recent applied papers on this approach includes [7, 25]. However, these methods are criticized for their highly sensitive distributional assumptions [17].

In general, the propensity score method (PSM) [33] is frequently used for the evaluation of a 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 [18] 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]. Instead, several indirect tests ([21] 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 modelling approach to draw causal inferences by using shared random effects [24]. Our modelling framework is closely related to those presented in [8, 17] 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 [24] 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.

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

*Corresponding author.

view [14], the propensity score method cannot produce reasonable estimates of the ATE while HTV's and our method can. For a probit-normal model [17] with no covariate in the response model, our approach performs almost the same as the HTV's. However, as the number of covariates in the response model increases our method becomes more efficient (for a 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 the PSM approach. Section 4 presents simulation studies 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 a later age and Section 6 concludes.

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 ($i = 1, 2, \ldots, n$) 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

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

Let S denote the treatment indicator with $S_i = 1$ when individual i is in treatment group and $S_i = 0$ when in non-treated. However, one cannot 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 α, β, τ and ω are fixed parameters $u_i \sim N(0, \sigma_u^2)$ 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

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

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

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

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

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

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 outcome and Z_i for the missingness indicator are known. The availability of any such background information is an advantage but it may not always be possible. In absence of any background variables in $X_i \cup Z_i$, the effect of the omitted covariates is captured by u_i [6]. Therefore, the random-effect u_i removes any hidden bias [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 a 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 [24] of $\psi = (\alpha, \beta, \tau, \omega, \phi, \gamma, \delta, \rho, \sigma_u^2)$ and $\mathbf{u} = (u_1, u_2, \ldots, u_n)^T$ is given as

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

Here, since $f(y_{i,o}|u_i) = \sum_{s=0}^{1} f(y_{i,o}|u_i, S_i = s)P(S_i = s|u_i)$, it is clear that if $f(y_{i,o}|u_i, S_i)$ is normal then $f(y_{i,o}|u_i)$ is a mixture of two normal distributions. 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)

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

where $\theta = (\tau, \alpha, \beta, \lambda, \omega, \kappa, \gamma, \delta)$, $g(p_i^*) = \eta_i^* = \frac{1}{C}(\gamma + Z_i\delta) + (y_i - \alpha - X_i\beta - S_i\tau)$, $C = \sqrt{1 + c^2(\frac{\rho^2\sigma_u^2\phi}{\sigma_u^2+\phi})}$ with $c = (\frac{16\sqrt{3}}{15\pi})$ 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

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

where $\theta^* = (\tau, \alpha, \beta, \lambda_0, \lambda_1, \omega, \kappa_0, \kappa_1, \gamma, \delta)$, $g(p_i^*) = \eta_i^* = \frac{1}{C_{S_i}}(\gamma + Z_i\delta) + \lambda_{S_i}(y_i - \alpha - X_i\beta - S_i\tau)$, $C_{S_i} = \sqrt{1 + c^2(\frac{\rho^2\sigma_u^2\phi(1-(1-\omega)S_i)^2}{1-(1-\omega)S_i)^2\sigma_u^2+\phi})}$ with $c = (\frac{16\sqrt{3}}{15\pi})$ 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 they were 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 the selection process, S_i, follows a binary model with the same functional form if g is a logit or a probit link [27]. Thus, if we estimate the selection model separately by using an 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), $\tilde{\theta}_0$, obtained only from the marginal selection model and estimate the remaining parameters by maximizing (7) or (8). The resulting estimate of $\tilde{\theta}_1$ is a pseudo maximum likelihood (PL) estimate (since $\tilde{\theta}_0$ 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

$$\begin{aligned} Asy.Var\left(\tilde{\theta}_1\right) &= \mathbf{R_2^{-1}} + \mathbf{R_2^{-1}}(\mathbf{R_3^T R_1^{-1} R_3} \\ &\quad - \mathbf{R_4^T R_1^{-1} R_3} - \mathbf{R_3^T R_1^{-1} R_4})\mathbf{R_2^{-1}} \end{aligned} \tag{9}$$

where $\mathbf{R_1}^{-1} = Asy.Var(\tilde{\theta}_0)$ base on the marginal likelihood of θ_0 only, $\mathbf{R_2} = Asy.Var(\tilde{\theta}_1|\theta_0 = \tilde{\theta}_0)$, $\mathbf{R_3} = E(\frac{\partial\ell_o}{\partial\theta_0}(\frac{\partial\ell_o}{\partial\theta_1})^T)$ and $\mathbf{R_4} = E(\frac{\partial l_{\theta_0}}{\partial\theta_0}(\frac{\partial\ell_o}{\partial\theta_1})^T)$ 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, $k = 1, \ldots, 4$, are replaced by 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 [23] 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

$$p_u(h_O) = \sum_{i=1}^{n} p_{u_i}(h_{i,O})$$

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

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 [23] 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 $\sum_{i=1}^{N} p_{\psi_1,u_i}(h_i)$ gives an extension of the restricted log-likelihood by eliminating the mean parameters and missing covariates and random effects: see [24] 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 [24] 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 [39] 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 [23] proposed 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 $F_i = [-3(\partial^4 h_{i,O}/\partial {u_i}^4)/D_i^2(h_{i,O}, u_i) - 5(\partial^3 h_{i,O}/\partial {u_i}^3)/D_i^3(h_{i,O}, u_i)]|_{u_i=\tilde{u}_i}$. 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 $\mathbf{y_{com}} = (\mathbf{y_o}, \mathbf{y_m})$ be complete data, where $\mathbf{y_o} = (y_{1,o}, y_{2,o}, \ldots, y_{n,o})^T$ is observed data and $\mathbf{y_m} = (y_{1,m}, y_{2,m}, \ldots, y_{n,m})^T$ with $y_{i,m} = y_i^{(1-S_i)}$ is missing data. Let $\mathbf{y_{com}^{(1)}} = \{S_i y_i^{(1)} + (1 - S_i) y_{i,m}\}$ and $\mathbf{y_{com}^{(0)}} = \{S_i y_{i,m} + (1 - S_i) y_i^{(0)}\}$, then $\mathbf{y_{com}} = (\mathbf{y_{com}}^{(0)T}, \mathbf{y_{com}}^{(1)T})^T$. 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

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

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 an interesting insight into the meaning of the causal parameter.

Since

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

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

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

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's are not estimable.

3. COMPARISON WITH THE PROPENSITY SCORE METHODS

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

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

which is called the propensity score. In order to identify the treatment effect $\tau(x)$, the 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 $E(y_i^{(s)}|X_i, S_i = s) = E_{u_i}(E(y_i^{(s)}|X_i, u_i, S_i = s)) = E(y_i^{(s)}|X_i)\ \forall s = 0, 1$. Strong ignorability assumption implies the conditional independence i.e., $(y_i^{(0)}, y_i^{(1)}) \perp S_i|(X_i, Z_i)$ [18]. 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 $\tau(x)$ is estimated in the following way.

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

Since

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

due to the strong ignorability assumption (Assumption-1), we have the following unbiased estimating equation for $\tau(x)$

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

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, 20].

The PSM estimator from (15) avoids the unidentifiable missing estimation $\hat{y}_i^{(1-S_i)}$ in (14). However, it requires us 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 [20].

PSM is very easy to understand and implement. This is why it has received so much attention. However, the strong ignorability assumption 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 [37]. Moreover, it works without requiring the strong ignorability condition.

In application, the PSM estimator often turns out to be implausible, in that the estimator exceeds any reasonable bound, if some observations have extreme (close to 0 or 1) propensity scores. In those cases the matching method [33] and normalization of the inverse propensity weight (NAIPW) [21], by restricting the sum of the weights to one, are proposed in the literature. The normalization approach gives the following estimator

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

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

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

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 $S_i|(X_i,Z_i)$, $y^{(0)}|(X_i,S_i=0)$ and $y^{(1)}|(X_i,S_i=1)$. 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 estimates of $p(Z_i)$ and that Assumption-1 holds. The response model can only improve the estimation if it has some predictive power, but it 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 model's 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 a small sample.

Our method does not need the strong igoreability 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 [34, 35, 25] is that we do not 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 conducted a series of simulation studies. Since our method is closely related to the HTV's we started with a simulation study consisting of the setting similar to HTV and then we examined situations with more covariates and functional misspecification of the selection model.

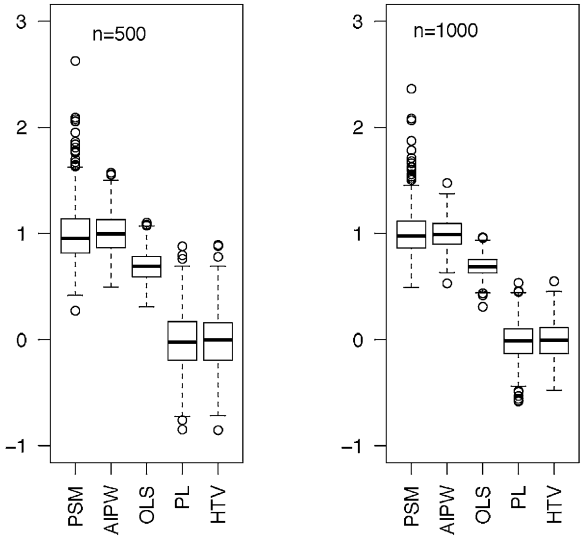

Figure 1. Bias and variation in ATE estimates by different methods.

For the first simulation study, we generated data with $y_{i,o}|S,u_i \sim N(\alpha+\tau S_i+(1-(1-\omega)S_i)u_i, 1.5)$, $\Phi^{-1}(p_i) = \gamma+\delta_1 X_{1,i}+\rho u_i$, $X_i \sim N(0,1)$, $u_i \sim N(0,0.75)$, $(\alpha,\tau,\gamma,\beta) = (2,1,0,1)$ were chosen according to [17] 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. $Cor(y_i^{(0)},y_i^{(1)}) = 0$ in HTV). At each Monte Carlo iteration we simulated $n = 500$ and 1,000 observations and we used 500 Monte Carlo replications. The results are summarized in terms of $\hat{\tau}-\tau$ in Figure 1.

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

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 are 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 overcome any bias. Since there are not many covariates, closeness between HTV and PL is also understandable. This situation might change as the number of covariates in the response models increases.

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.*

For the second simulation study, we generated data with sample sizes $n = 250$, 500 and 1,000, 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 the simulation, we considered $E(y_{i,o}|X_i, S_i, u_i) = \alpha + \beta_1 X_{1i} + \beta_2 X_{2,i} + \tau S_i + u_i$ and $\Phi^{-1}(p_i) = \gamma + \delta_1 X_{1,i} + \delta_2 X_{2,i} + \rho u_i$ where Φ represents a standard normal CDF. For the PSM method we estimated the propensity score with a probit model with both X_1 and X_2 as the covariates. We implemented improved IPW estimate (AIPW) [10] with both the covariates through a spline model for the means of y_i and S_i. We also carried out a simple regression (OLS) estimation of τ by regressing S, X_1 and X_2 on y_o.

The results showed the same pattern as the one in Figure 1. Hence, we do not report all the detailed results. The only difference in the results of the second simulation was that the HTV turned out to be inefficient, in terms of higher mean squared error (MSE; see MSE and Bias in Table 1) compared to PL, HL(1) and HL(2). Other methods (AIPW and OLS) were found biased, as expected.

For the third simulation, we generated data in the same way as the second simulation except that we used 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 the probit case, the errors in response and selection models follow multivariate normal distribution. However, in logit cases we do not know the joint distribution of the error terms. Therefore, we computed the ATE for HTV by treating the error distribution of the selection models to be both normal or logistic (neither is correct). We also examined the effect of the misspecification of the link function for the PL, HL(1) and HL(2), too. Since we already knew, from the second simulation study, that the PSM, AIPW and OLS were biased we did 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 rest from the third simulation (True Link: Logit). When the true link is probit, the HTV estimator is unbiased but the biases of PL, HTV and HL(2) decrease rapidly as the sample size increases. 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 a 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 the youths keeps growing. Since, one of the causes for high youth unemployment rate would be the difficulty in getting their first job. A remedy to provide the school leavers with work experience is often suggested (see e.g. [36] and the references cited therein). Therefore, it is interesting to examine the causal effect of early summer job experience on the future unemployment of high school graduates. To check this hypothesis, we created a data set by merging several data bases maintained by Statistics Sweden (SCB). Main source was the LOUISE data base (renamed as LISA since 2004; see http://www.scb.se/) which contains longitudinal information on earnings and demographics. We used a random sample of 5% individuals from the LOUISE database between 1995 and 2002. We defined a teenager (aged between 16 and 19 years) as a summer jobber if (s)he had any gainful employment during June–July. We added further information on a teenager's middle and high school grades and parents' social and economic status from other

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

databases maintained by SCB. We used OLS, PL, AIPW and HL methods to estimate the average summer-job effect. The outcome variable was $\log(1 + income)$ with the observed income being adjusted for inflation using the year 2002 as the base. We fitted 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 the 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. The 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 persistent long term effect on future income. The results from PL might look rather striking but it is consistent with the results of [38] though they used experimental data from only one municipality in Sweden.

6. CONCLUSION

In this paper, we present a shared random effects models approach for drawing inferences under selection bias. The proposed models include the Gaussian or "textbook selection" model [16] as a special case. We provide a close-form solution for the model parameters and their asymptotic variance estimation by using the h-likelihood method. For computation, we offer three algorithms: PL, HL(1) and HL(2). The PL is computationally fast but the simulation results indicate that it may not be as precise as the other alternatives. The HL(1) turns out to be the most efficient (in terms of MSE) though it is computationally slower than the PL, but faster than HL(2). Both the simulation study and the real data application show that the proposed approach can be a better alternative than the already existing ones in drawing inferences on treatment effects under non-ignorable treatment allocation.

APPENDIX A

A.1 Derivarion of equation (7)

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

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

Secondly,

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

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

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

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 $u_i \sim N(\frac{\sigma_u^2}{(\phi+\sigma_u^2)}(y_i^{(S_i)} -$

$\mu_i), \frac{\phi\sigma_u^2}{(\phi+\sigma_u^2)})$. 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

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

where

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

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)

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

For logit link, i.e. when $g(p_i) = \log(\frac{p_i}{1-p_i}) = \eta_i$, using the relation between standard normal and standard logistic CDF [22], 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

$$E_{u_i}(p_i) = \Phi\left(\frac{\gamma + Z_i\delta + \frac{\rho\sigma_u^2}{(\phi+\sigma_u^2)}\left(y_i^{(S_i)} - \mu_i\right)}{\sqrt{\left(\left(\frac{15\pi}{16\sqrt{3}}\right)^2 + \frac{\rho^2\phi\sigma_u^2}{(\phi+\sigma_u^2)}\right)}}\right)$$

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

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 [40] for further discussion).

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

Let,

$$\begin{aligned} y^{(1)} &= X\beta^{(1)} + U^{(1)}, \\ y^{(0)} &= X\beta^{(0)} + U^{(0)}, \\ D_i^* &= Z\delta + U^D \end{aligned} \tag{21}$$

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 tryvariate distribution. Denote $Var(U^{(0)}) = \sigma_0^1$, $Var(U^{(1)}) = \sigma_1^1$, $Cor(U^{(0)}, U^D) = \varrho_0$, $Cor(U^{(1)}, U^D) = \varrho_1$ 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)

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

and

(23)

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

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

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

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 $\widehat{\delta}$.
3. Run treatment-specific regression for groups $S = 1$ and $S = 0$ separately.
4. Given, $\hat{\beta}^{(0)}$, $\widehat{\beta}^{(1)}$, $\widehat{\varrho_1\sigma_1}$ and $\widehat{\varrho_0\sigma_0}$ from step 3 we estimate the ATE as

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

This is the ATE of HTV implemented in the simulation. Asymptotic variance estimate of this ATE estimator is given in [16].

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 estimator, 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.

ACKNOWLEDGEMENTS

This work was supported by the National Research Foundation (NRF) of Korea grant funded by the Korean government (MEST) (No. 2011-0030810, No. 2009-0065135).

Received 26 November 2012

REFERENCES

[1] Bollen, K. A. (1989). *Structural Equations with Latent Variables*, Wiley, New York. MR0996025

[2] Bravo, F. and Jacho-Chavez, D. T. (2011). Empirical likelihood for efficient semiparametric average treatment effects. *Econometric Rev.* **30** 1–24. MR2747366

[3] Breslow, N. E. and Clayton, D. (1993). Approximate inference in generalized linear mixed models. *J. Amer. Statist. Assoc.* **88** 9–25.

[4] Cobb-Clark, D. A. and Crossley, T. (2003). Econometrics for evaluations: An introduction to recent development. *Econ. Rec.* **79** 491–511.

[5] Cox, D. R. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference. *J. Roy. Statist. Soc. Ser.* B **49** 1–39. MR0893334

[6] 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

[7] 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

[8] Garen, J. (1984). A selectivity bias approach with a continuous choice variable. *Econometrica* **52** 1199–1218.

[9] Gong, G. and Samaniego, F. J. (1981). Pseudo maximum likelihood estimation: Theory and applications. *Ann. Stat.* **9** 861–869. MR0619289

[10] Glynn, A. N. and Quinn, K. M. (2009). An introduction to the augmented inverse propensity weighted estimator. *Polit. Anal.* **18** 36–56.

[11] Greene, W. H. (2003). *Econometric Analysis*, Pearson Education, Upper Saddle River, N.J.

[12] Joe, H. (2008). Accuracy of Laplace approximation for discrete response mixed models. *Comput. Statist. Data Anal.* **52** 5066–5074. MR2526575

[13] Hastie, T. and Tibshirani, R. (1986). Generalized Additive Models (with discussion). *Statist. Sci.* **1** 297–318. MR0858512

[14] Heckman, J. J. (2008). Econometric causality. *Int. Stat. Rev.* **76** 1–27.

[15] Heckman, J. J. and Vytlacil, E. (2005). Structural equations, treatment effects, and econometric policy evaluation. *Econometrica* **73** 669–738. MR2135141

[16] Heckman, J. J., Tobias, J. L. and Vytlacil, E. (2003). Simple estimator for treatment parameters in a latent-variable framework. *Rev. Econ. Statist.* **85** 748–755

[17] 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.

[18] 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

[19] 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.

[20] 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

[21] Imbens, G. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Rev. Econ. Stat.* **86** 4–29.

[22] Johnson, N. L. and Kotz, S. (1970). *Distributions in Statistics: Continuous Univariate Distributions–2*, Wiley, New York. MR0270476

[23] 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

[24] 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

[25] Li, M. and Tobias, J. L. (2008). Modelling and evaluating treatment effects in econometrics. *Adv. Econometrics* **21** 57–91. MR2544064

[26] 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. MR1940632

[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. MR0472152

[36] Ruhm, C. J. (1997). Is high school employment consumption or investment? *J. Labor Econ.* **15** 735–776.

[37] Smith, J. A. and Todd, P. E. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators? *J. Econometrics* **125** 305–353. MR2143379

[38] 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.

[39] Yun, S. and Lee, Y. (2004). Comparison of hierarchical and marginal likelihood estimators for binary outcomes. *Comput. Statist. Data Anal.* **45** 639–650. MR2055468

[40] Zeger, S. L., Liang, K. L. and Albert, P. S. (1988). Models for longitudinal data: A generalized estimating equation approach. *Biometrics* **44** 1049–1060. MR0980999

Md. Moudud Alam
School of Technology and Business Studies
Dalarna University
Falun, 791 88
Sweden
E-mail address: maa@du.se

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

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