11-2-2010

On the Estimation of Selection Models when Participation is Endogenous and Misclassified

Ian M. McCarthy
University of North Texas, ian.mccarthy@fticonsulting.com

Rusty Tchernis
Georgia State University

W. J. Usery Workplace Research Group Paper Series

Working Paper 2010-11-2
November 2010

On the Estimation of Selection When Participation is Endogenous and Misclassified

Ian M. McCarthy
University of North Texas

Rusty Tchernis
Georgia State University

Georgia State University
Andrew Young School of Policy Studies

On the Estimation of Selection Models when Participation is Endogenous and Misclassified

Ian M. McCarthy*
FTI Consulting
and University of North Texas

Rusty Tchernis
Georgia State University
and NBER

November 2010

Abstract

This paper presents a Bayesian analysis of the endogenous treatment model with misclassified treatment participation. Our estimation procedure utilizes a combination of data augmentation, Gibbs sampling, and Metropolis-Hastings to obtain estimates of the misclassification probabilities and the treatment effect. Simulations demonstrate that the proposed Bayesian estimator accurately estimates the treatment effect in light of misclassification and endogeneity.

JEL Classifications: C31, C35, C51, C52

Keywords: Misclassification, Misreporting, Endogeneity, Selection on Observables, Measurement Error, Bayesian Econometrics

*Correspondence to: Ian McCarthy, 2001 Ross Avenue, Suite 400, Dallas, TX 75201. E-mail: ian.mccarthy@fticonsulting.com

1 Introduction

Misclassification (or misreporting) occurs when subjects incorrectly report their status or participation in a particular program. For example, in job tenure surveys, a job change may be incorrectly recorded or respondents may incorrectly classify a promotion as a job change (Brown and Light, 1992), leading to misclassification of the job change indicator. Similar evidence of misclassification arises in a wide range of studies. Bollinger and David (1997), among others, find evidence of misreporting in food stamp program participation; Bitler, Currie, and Scholz (2003) in their study of WIC eligibility and participation from the Current Population Survey and the Survey of Income and Program Participation; Barron, Berger, and Black (1997) in their analysis of job training; Card (1996) in his study of union coverage; and Black, Sanders, and Taylor (2003) in their analysis of group health insurance eligibility. In general, there is substantial evidence that misclassification is at least somewhat prevalent in a variety of situations in which individuals self-report.[1]

The empirical problems resulting from misclassification have been well-documented.[2] For misclassified binary dependent variables, Bollinger and David (1997) and Hausman, Abrevaya, and Scott-Morton (1998) show that ignoring misclassification yields biased and inconsistent coefficient estimates. For misclassified binary covariates, Kreider (2010) notes that even a slight amount of misclassification can generate estimated treatment effects whose confidence intervals do not overlap those from the the true data. Additional empirical problems arise if the misclassified treatment is also endogenous (Kreider *et al.*, 2009). In this paper, we propose a Bayesian estimator that allows for both misclassified and endogenous treatment in a single framework. Through a series of simulations, we demonstrate the important impact misclassification and endogeneity can have if ignored. We show that: 1) our proposed estimator corrects for the bias introduced by misclassification and endogeneity under certain parametric assumptions; 2) in order to accurately estimate the treatment effect, accounting for endogeneity may be more important than accounting for misclassification in certain situations; and 3) with more information (e.g., more individuals in the data) accounting for both misclassification and endogeneity becomes more critical.

Separately, misclassification and endogenous treatment have each been well-studied in the literature. For the purposes of this paper, it is useful to classify the relevant literature into one of three approaches: 1) structural; 2) program evaluation; and 3) Bayesian. Heckman (2003, p. 361) differentiates between

[1] See Bound, Brown, and Mathiowetz (2001) for a review of misclassification and, more generally, measurement error in survey data.

[2] See Millimet (2010) for a thorough review of the literature on measurement error.

the program evaluation and structural approaches as follows: "The goal of the structural econometrics literature, like the goal of all science, is to understand the causal mechanisms producing effects so that one can use empirical versions of models to forecast the effects of interventions never previously experienced, to calculate a variety of policy counterfactuals and to use theory to guide choices of estimators to interpret evidence and to cumulate evidence across studies. These activities require models for understanding 'causes of effects' in contrast to the program evaluation literature that focuses only on the 'effects of causes' (Holland 1986)."[3] Heckman's definition of the structural approached appears to encompass the Bayesian approach; however, for the purposes of this paper, we limit the "structural" umbrella to a frequentist approach only.

Among the structural approach, endogenous treatment can be considered part of a larger class of simultaneous equation models with limited dependent variables (SLDV models). This class of models was first studied in detail by Amemiya (1978) and Heckman (1978). Numerous authors have since applied this class of models in a variety of settings and with a variety of estimators, including two-stage least squares, maximum likelihood, semi-parametric, and method of moments. Examples include Rivers and Vuong (1988), Blundell and Smith (1989, 1994), Vella (1993), Cameron and Heckman (1998), Carneiro, Hansen, and Heckman (2003), and Lewbel (2006), among others.[4]

Regarding misclassification, Aigner (1973), Bollinger (1996), Black, Berger, and Scott (2000), Fazis and Lowenstein (2003), Imbens and Manski (2004), Hu (2006), and Molinari (2008), among others, examine the estimation of models with misclassified binary covariates. Absent additional information, a common solution is to bound the unidentified parameter associated with the misclassified regressor. Authors have also proposed refinements to these bounds, and in some cases point-identification, through the use of auxiliary data, repeated measurement, or instrumental variables. For misclassified dependent variables, Hausman, Abrevaya, and Scott-Morton (1998) propose an adjusted maximum likelihood that can estimate the extent of misclassification and consistently estimate the coefficients of the latent utility specification. Sullivan (2007) adopts a similar approach in his study of occupational choice. Other studies of misclassified, discrete dependent variables include Li, Trivedi, and Guo (2003) and Abrevaya and Hausman (2004).

In the program evaluation literature, studies of endogeneity (i.e., selection on unobservables) include

[3]Heckman is slightly more critical of the program evaluation approach, citing Holland (1986) and others' statements that the program evaluation approach cannot estimate causal effects for variables that cannot be randomly assigned, such as gender.

[4]See Manski (1994) and Vella (1998) for surveys on sample selection in general.

Manski (1990), Imbens and Angrist (1994), Agrist, Imbens, and Rubin (1996), Klein and Vella (2009), and Millimet and Tchernis (2010). These approaches tend to adopt a more robust bounds analysis or rely on instrumental variables to identify the treatment effect. Studies of misclassification in the program evaluation literature include Lewbel (2007), Battistin and Sianesi (2007), and Kreider *et al.* (2009). Battistin and Sianesi (2007) adopt a bounding technique to estimate the average treatment effect on the treated (ATT) with misclassified treatment status. In addition to offering an extension for multiple treatments, Battistin and Sianesi's approach is novel in that they permit heterogeneous misclassification probabilities and focus specifically on the ATT. Kreider *et al.* (2009) provide one of the few techniques to analyze both misclassification and endogeneity in the same framework. Using data on food stamp program participation, the authors estimate a potential outcomes model that allows for both selection and misclassification. Similar to Battistin and Sianesi, the authors adopt partial identification bounding methods in order to estimate a range of the estimated treatment effect, and they show that the commonly found negative relationship between food stamp participation and health outcomes is largely reversed after accounting for endogeneity and misclassification. Although the allowance of both misclassification and endogeneity in a potential outcomes framework is clearly an important advancement, these estimators do not permit point-identification of the treatment effect nor do they permit point-identification of the misclassification probabilities.

Authors have also developed Bayesian estimators robust to endogeneity and, more recently, misclassified dependent variables. Li (1998), in an extension of Albert and Chip (1993), develops an estimator that permits endogeneity in the probit specification. Li adopts a Bayesian approach to the SLDV model and uses a sampling technique based on a combination of data augmentation and Gibbs sampling; however, Li does not allow for misclassification of the endogenous variable(s).

Balcombe *et al.* (2007) introduce misclassification in a conditional logit model. Balcombe and Fraser (2009) also apply Bayesian methods to estimate a model similar to Hausman, Abrevaya, and Scott-Morton (1998), adopting a probit specification rather than a logit. Other Bayesian studies of misclassified data, often focusing on medical applications, include Gaba and Winkler (1992), Evans *et al.* (1996), Prescott and Garthwaite (2002), Swartz *et al.* (2004), Gerlach and Stamey (2007), and Perez *et al.* (2007).

Despite advances in the structural, program evaluation, and Bayesian approaches, we are not aware of any estimator that currently permits point-estimation of treatment effects and misclassification probabilities in light of both misclassified and endogenous treatment participation. Leaning on the work of Hausman, Abrevaya, and Scott-Morton (1998), Li (1998), and Balcombe and Fraser (2009), our pro-

posed estimator and sampling algorithm adopts a combination of data augmentation, Gibbs sampling, and Metropolis-Hastings to obtain estimates of the misclassification probabilities and the treatment effect when participation is endogenous and misclassified.

Our Bayesian approach to misclassified or mismeasured data has some important advantages over the structural and program evaluation approaches. As Heckman (2003) discusses, perhaps the most important distinction from the program evaluation approach is the ability to formally model underlying behavioral mechanisms and focus on the specific coefficients of interest. This has important implications for policy analysis, particularly with regard to out-of-sample inferences and estimating the effects of various alternative policies under consideration.

Comparing the structural and Bayesian approaches, Balcombe *et al.* (2007) show that their proposed Bayesian estimator outperforms that of Hausman, Abrevaya, and Scott-Morton (1998) based on an analysis of Bayes factors. Bollinger and Hasselt (2009) also show that the use of priors with misclassified data help narrow the bounds on unidentified coefficients relative to the bounds estimated in a classical regression context. This result from Bollinger and Hasselt mirrors the sentiment of Winkler and Gaba (1990), who showed that non-identifiability in a likelihood analysis can be avoided with a Bayesian estimator. Balcombe and Fraser (2009) also note that the Bayesian approach more easily incorporates various parameter constraints, such as restrictions on the extent of misclassification.

A Bayesian estimator also has advantages in accounting for endogeneity. For example, most classical applications to the SLDV model do not apply a full-information maximum likelihood (FIML), but rather adopt computationally simpler techniques such as simulated ML or alternative two-step algorithms (Blundell and Smith, 1994). But as Li (1998) notes, a Bayesian approach avoids the direct evaluation of the likelihood while still providing draws from the exact posterior distribution of the SLDV model. Li also notes that the Bayesian approach provides more precise small sample results compared to the classical treatment of the SLDV model.

In general, although it may be theoretically possible to implement our proposed estimator with simulated likelihood techniques, our Bayesian estimator avoids theoretical problems with maximization (e.g., convergence) and does not rely on additional asymptotic assumptions to estimate the standard errors. The remainder of this paper is organized as follows. Section 2 presents the model in question and our proposed estimator; Section 3 presents the sampling methodology to allow draws from the relevant posterior distributions; Section 4 presents the simulated data used to assess the performance of our proposed estimator and the estimation results for each respective dataset; and Section 5 concludes.

2 The Model and Estimator

We consider the following simultaneous equation system (the Roy model):

$$\begin{aligned} y_{1i} &= \mathbf{X}_{1i}'\gamma + \delta_i\alpha + \epsilon_{1i} \\ y_{2i}^* &= \mathbf{X}_{2i}'\beta + \epsilon_{2i}, \end{aligned} \tag{1}$$

where $\delta_i = \mathbf{1}(y_{2i}^* > 0)$,

$$\begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix} \sim N(\mathbf{0}, \mathbf{\Sigma}), \tag{2}$$

and

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & 1 \end{pmatrix}. \tag{3}$$

In this system, y_{2i}^* represents the latent treatment participation variable (i.e., the latent utility) with δ_i an indicator for whether person i received treatment. y_{1i} represents the outcome, and the effect of treatment on the outcome is measured by the coefficient α. The covariates, $\mathbf{X}_1$ and $\mathbf{X}_2$, are of dimensions $K_1 \times N$ and $K_2 \times N$, respectively. For identification of the selection equation coefficients, we assume at least one covariate in the selection equation is excluded from the outcome equation ($K_2 > K_1$). We allow for endogenous treatment with the condition $\sigma_{12} \neq 0$. As indicated in equation (3), we also normalize the $(2,2)$ element of $\mathbf{\Sigma}$ to one in order to identify the remaining coefficients in the model.[5] For ease of notation, we express the system in equations (1)-(3) as a standard seemingly unrelated regression (SUR) model,

$$\begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2^* \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1 & \delta & \\ & & \mathbf{X}_2 \end{pmatrix} \begin{pmatrix} \gamma \\ \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix}. \tag{4}$$

Or, more compactly,

$$\begin{aligned} \mathbf{Y} &= \mathbf{X}\mathbf{\Lambda} + \epsilon, \text{ where} \\ \epsilon &\sim N(\mathbf{0}, \mathbf{\Sigma}). \end{aligned}$$

[5] As discussed in more detail in Section 3, this restriction is not sufficient to point-identify σ_{12}, provided there is some amount of misclassification.

One of the main contributions of our proposed estimator is the allowance of misclassified participation in the above SLDV model. To incorporate potential misclassification, we adopt the specification and notation in Balcombe and Fraser (2009). This requires a distinction between the observed treatment indicator y_{2i} and the actual treatment indicator δ_i, where individual i misreports their treatment participation if $y_{2i} \neq \delta_i$. We denote by $\theta_{y_{2i}|\delta_i}$ the conditional probability of the observed treatment given the actual treatment, so that:

$$\begin{aligned}
\theta_{1|0} &= P\left(y_{2i} = 1|\delta_i = 0\right); \\
\theta_{0|1} &= P\left(y_{2i} = 0|\delta_i = 1\right); \\
\theta_{0|0} &= P\left(y_{2i} = 0|\delta_i = 0\right) = 1 - \theta_{1|0}; \text{and} \\
\theta_{1|1} &= P\left(y_{2i} = 1|\delta_i = 1\right) = 1 - \theta_{0|1}.
\end{aligned}$$

$\theta_{1|0}$ therefore represents the conditional probability of any given individual reporting that they did receive treatment when in fact they did not receive treatment. Conversely, $\theta_{0|1}$ represents the conditional probability of any given individual reporting that they did not receive treatment when in fact they did receive treatment. $\theta_{1|1}$ and $\theta_{0|0}$ represent the probability of the individual correctly reporting that they did and did not receive treatment, respectively. We also denote by $\Psi_{\delta_i|y_{2i}}$ the conditional probability of the actual treatment given the observed treatment, and we denote the probability of person i receiving treatment by $\Phi_i = P(y_{2i}^* > 0)$. In this case, unlike the standard probit model where Φ_i represents the value of the standard normal CDF evaluated at $\mathbf{X}_{2i}\beta$, Φ_i now represents the value of the *conditional* distribution of $\epsilon_2|\epsilon_1$ evaluated at $\mathbf{X}_{2i}'\beta$. From equation (2), $\epsilon_2|\epsilon_1$ follows a normal distribution with mean $\epsilon_1\frac{\sigma_{12}}{\sigma_{11}}$ and variance $1 - \frac{\sigma_{12}^2}{\sigma_{11}}$.

Similar to Hausman, Abrevaya, and Scott-Morton (1998) in their estimation with symmetric misclassification probabilities and Bollinger and Hasselt (2009), we assume that $\theta_{1|0} \leq 0.5$ and $\theta_{0|1} \leq 0.5$. This amounts to the requirement that individuals are on average more likely to correctly report their participation than to misreport. In other words, no matter the value of the observed covariates for a given individual, the researcher always places more weight on the reported participation decision than on the alternative. Although Balcombe and Fraser (2009) indicate that this assumption may not be necessary for formal identification of the misclassification probabilities, this is intuitively an important assumption if one is to make any meaningful, practical interpretation of the final results.

Based on the observed y_{2i} and $\mathbf{X}_{2i}$, we want to estimate the actual treatment indicator δ_i. From

Bayes' Theorem, it follows that

$$\Psi_{\delta_i|y_{2i}} = \frac{P(y_{2i}|\delta_i, \mathbf{X}_{2i})P(\delta_i|\mathbf{X}_{2i})}{\sum_{\delta=0}^{1} P(y_{2i}|\delta_i, \mathbf{X}_{2i})P(\delta_i|\mathbf{X}_{2i})} = \frac{\theta_{y_{2i}|\delta_i}}{\theta_{y_{2i}|\delta_i=0}(1-\Phi_i) + \theta_{y_{2i}|\delta_i=1}\Phi_i}. \tag{5}$$

Our estimation procedure described in the next Section 3 relies heavily on the relationship in (5) in order to estimate the true δ_i based on the observed y_{2i} and $\mathbf{X}_{2i}$. Intuitively, equation (5) indicates that as Φ_i increases (i.e., it becomes more likely that person i's latent utility is positive), $\Psi_{\delta_i=1|y_{2i}=1}$ increases while $\Psi_{\delta_i=1|y_{2i}=0}$ decreases. The observed data underlying the latent utility function therefore informs the researcher, along with the reported treatment participation, as to the probability of misclassification.

The resulting likelihood can be factored as follows:

$$\begin{aligned} f(\mathbf{y}_1, \mathbf{y}_2^*, \delta, \mathbf{y}_2|\mathbf{\Lambda}, \mathbf{\Sigma}, \Theta) \propto & f(\mathbf{y}_1|\mathbf{y}_2^*, \delta, \mathbf{y}_2, \mathbf{\Lambda}, \mathbf{\Sigma}, \Theta) \times f(\mathbf{y}_2^*|\delta, \mathbf{y}_2, \mathbf{\Lambda}, \mathbf{\Sigma}, \Theta) \\ & \times f(\delta|\mathbf{y}_2, \mathbf{\Lambda}, \mathbf{\Sigma}, \Theta) \times f(\mathbf{y}_2|\mathbf{\Lambda}, \mathbf{\Sigma}, \Theta), \end{aligned}$$

where $f(\mathbf{y}_1|\cdot)$ and $f(\mathbf{y}_2^*|\cdot)$ are each normal distributions with means and variances discussed in Section 3. Note that $f(\delta|\cdot) = \Psi_{\delta_i|y_{2i}}$, and $f(\mathbf{y}_2|\cdot)$ is the same likelihood derived in Hausman, Abrevaya, and Scott-Morton (1998):

$$f(\mathbf{y}_2|\cdot) = \prod_{i=1}^{N} \left\{ \left[\theta_{y_{2i}|0}(1-\Phi_i) + \theta_{y_{2i}|1}\Phi_i\right]^{y_{2i}} \left[\theta_{y_{2i}|0}(1-\Phi_i) + \theta_{y_{2i}|1}\Phi_i\right]^{1-y_{2i}} \right\}. \tag{6}$$

3 Sampling Algorithm

To estimate the coefficients in the selection and outcome equations, we adopt a Bayesian Markov Chain Monte Carlo (MCMC) approach. As Li (1998) discusses, there is an equivalent representation of (1)-(3) that simplifies the MCMC sampling process. The equivalent system is based on a decomposition of the joint distribution of (ϵ_1, ϵ_2) into the product of the conditional distribution of $\epsilon_1|\epsilon_2$ and marginal distribution of ϵ_2, which yields

$$\begin{aligned} y_{1i} &= \mathbf{X}_{1i}'\gamma + \delta_i\alpha + \overbrace{\sigma_{12}\epsilon_{2i} + \nu_i}^{\epsilon_{1i}} \\ y_{2i}^* &= \mathbf{X}_{2,i}'\beta + \epsilon_{2i} \end{aligned} \tag{7}$$

where $\delta_i = \mathbf{1}(y_{2i}^* > 0)$, $\nu_i \sim N(0, \sigma^2)$, $\epsilon_{2i} \sim N(0,1)$, with ν_i independent of ϵ_{2i}. From $\epsilon_{1i} = \sigma_{12}\epsilon_{2i} + \nu_i$ and the joint normality assumption (equation (2) above), it follows that $\sigma^2 = \sigma_{11} - \sigma_{12}^2$. Based on this representation, we can rewrite $\mathbf{\Sigma}$ as

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma^2 + \sigma_{12}^2 & \sigma_{12} \\ \sigma_{12} & 1 \end{pmatrix}. \tag{8}$$

We specify prior distributions for β, α, γ, Θ, σ^2, and σ_{12} in order to obtain draws from the posterior distributions. Our assumed prior distributions are

$$\begin{aligned} &f(\mathbf{\Lambda}) \sim \mathbf{N}(\mathbf{0}, \mathbf{V}_\Lambda), \\ &f(\sigma_{12}) \sim N(0, V_{\sigma_{12}}), \\ &f(\sigma^2) \sim IG(a, b^{-1}), \text{ and} \\ &f(\theta_{0|1}) = f(\theta_{1|0}) \sim U[0, 0.5]. \end{aligned}$$

Our sampling algorithm derives from a combination of data augmentation, Gibbs sampling, and Metropolis-Hastings, and draws heavily from Gelfand and Smith (1990), Chib (1992), Rossi, McCulloch and Allenby (1996), Li (1998), and Balcombe and Fraser (2009). The idea is to draw the latent δ_i (the true treatment indicator) using observed data, with which we can draw the latent y_{2i}^* from the univariate truncated normal distribution. The model then reduces to a standard SUR model, where we can draw $\mathbf{\Lambda}$, σ_{12}, and σ^2. We obtain draws of the misclassification probabilities $\theta_{0|1}$ and $\theta_{1|0}$ from Metropolis-Hastings based on the likelihood for $\Theta = \left(\theta_{0|1}, \theta_{1|0}, \theta_{1|1}, \theta_{0|0}\right)'$. This algorithm is presented in more detail below:

1. Begin with initial values of all coefficients and denote by s the current iteration in the MCMC. Given Θ^{s-1}, the probability of the "true" participation conditional on the observed participation is given directly from equation (5):

$$\Psi^s_{\delta_i=1|y_{2i}} = \frac{\theta^{s-1}_{y_{2i}|1}\Phi_i^{s-1}}{\theta^{s-1}_{y_{2i}|1}\Phi_i^{s-1} + \theta^{s-1}_{y_{2i}|0}(1-\Phi_i^{s-1})},$$

   where Φ_i is the CDF of the standard normal distribution evaluated at $\mathbf{X}_{1i}'\gamma$. Note that we need only form Ψ_i for $\delta_i = 1$ *or* $\delta_i = 0$ since $\Psi_{\delta_i=1|y_{2i}} = 1 - \Psi_{\delta_i=0|y_{2i}}$.

2. Draw the "true" treatment participation variable based on the probabilities from step 1 above: $\delta_i^s = \mathbf{1}\left(r_i \leq \Psi_i^s\right)$, where r_i is a draw from a uniform distribution with support $[0,1]$. The intuition behind

this deserves some detail. Ψ_i represents the conditional probability of $\delta_i = 1$, given the observed choice y_{2i} and the underlying data $\mathbf{X}_{2i}$. As $\mathbf{X}_{2i}^{'}\beta$ increases, the probability of $\delta_i = 0$ decreases. If $\mathbf{X}_{2i}^{'}\beta$ is a large positive number *and* $y_{2i} = 1$, then the probability of misclassification will be minimal; however, if $\mathbf{X}_{2i}^{'}\beta$ is a large positive number and $y_{2i} = 0$, the probability of misclassification will be higher.

3. Draw latent treatment variables y_{2i}^{*} for $i = 1, ..., N$. Conditional on $\mathbf{\Lambda}$, $\mathbf{\Sigma}$, and δ, y_{2i}^{*} follows a truncated normal distribution, truncated below by 0 if $\delta_i = 1$ and truncated above by 0 if $\delta_i = 1$:

$$
\begin{aligned}
&y_{2i}^{*,s}|\mathbf{X}_{2i}, \delta_i^s = 1, \mathbf{\Lambda}^{s-1} \\
&\quad \sim \mathrm{N}_{[0,\infty)}\left(\mathbf{X}_{2i}^{'}\beta^{s-1} + \frac{\sigma_{12}^{s-1}}{\sigma^{2,s-1} + \sigma_{12}^{2,s-1}}\left(y_{1i} - \mathbf{X}_{1i}^{'}\gamma^{s-1} - \alpha^s\right), 1 - \frac{\sigma_{12}^{2,s-1}}{\sigma^{2,s-1} + \sigma_{12}^{2,s-1}}\right) \forall i \text{ and} \\
&y_{2i}^{*,s}|\mathbf{X}_{2i}, \delta_i^s = 0, \mathbf{\Lambda}^{s-1} \\
&\quad \sim \mathrm{N}_{(-\infty,0]}\left(\mathbf{X}_{2i}^{'}\beta^{s-1} + \frac{\sigma_{12}^{s-1}}{\sigma^{2,s-1} + \sigma_{12}^{2,s-1}}\left(y_{1i} - \mathbf{X}_{1i}^{'}\gamma^{s-1}\right), 1 - \frac{\sigma_{12}^{2,s-1}}{\sigma^{2,s-1} + \sigma_{12}^{2,s-1}}\right) \forall i.
\end{aligned}
$$

   We adopt the inverse transform method in order to obtain draws of the latent y_{2i}^{*} from the univariate truncated normal distribution.

4. Draw the outcome and selection equation coefficients, $\mathbf{\Lambda}$. Conditional on $\mathbf{y}_2^{*}$ and $\mathbf{\Sigma}$, and with a normal prior distribution, $\mathbf{\Lambda}$ follows a normal distribution:

$$
\begin{aligned}
&\mathbf{\Lambda}^s|\mathbf{y}_2^{*,s}, \mathbf{\Sigma}^s \sim \mathrm{N}\left(\mathbf{AB}, \mathbf{A}\right), \text{ where} \\
&\mathbf{A} = \left(\sum_{i=1}^{N} \mathbf{X}_i^{'}\mathbf{\Sigma}^{-1,s-1}\mathbf{X}_i + \mathbf{V}_\Lambda\right)^{-1} \text{ and} \\
&\mathbf{B} = \sum_{i=1}^{N} \mathbf{X}_i^{'}\mathbf{\Sigma}^{-1,s-1}\mathbf{Y}_i.
\end{aligned}
$$

5. Draw the covariance, σ_{12}. Conditional on $\mathbf{\Lambda}$, y_{1i}^{*}, δ, and σ^2, and with a normal prior distribution,

σ_{12} also follows a normal distribution:

$$\begin{aligned}
&\sigma_{12}^{s}|\mathbf{\Lambda}^{s}, \mathbf{y}_{2}^{*,s}, \delta^{s}, \sigma^{2,s-1} \sim \mathrm{N}\left(\mathbf{Dd}, \mathbf{D}\right) \text{ , where} \\
&\mathbf{D} = \left(\frac{\hat{\epsilon_2}^{'}\hat{\epsilon_2}}{\sigma^{2,s-1}} + V_{\sigma_{12}}^{-1}\right)^{-1}, \\
&\mathbf{d} = \frac{\hat{\epsilon_2}^{'}\hat{\epsilon_1}}{\sigma^{2,s-1}}, \\
&\hat{\epsilon_1} = \mathbf{y}_1 - \mathbf{X}_1^{'}\gamma^{s} - \delta^{s}\alpha^{s}, \text{ and} \\
&\hat{\epsilon_2} = \mathbf{y}_2^{*,s} - \mathbf{X}_2^{'}\beta^{s}.
\end{aligned}$$

6. Draw the variance, σ^2. Conditional on $\mathbf{\Lambda}$, σ_{12}, $\mathbf{y}_2^*$, and δ, and with an inverse gamma prior, σ^2 follows an inverse gamma distribution:

$$\sigma^{2,s}|\mathbf{\Lambda}^{s}, \sigma_{12}^{s}, \mathbf{y}_{2}^{*,s}, \delta^{s} \sim IG\left(\frac{N}{2} + a, \left[\frac{1}{b} + \frac{1}{2}\sum_{i=1}^{N}\left(y_{1i} - \mathbf{X}_{1i}^{'}\gamma^{s} - \delta_i^{s}\alpha^{s} - \sigma_{12}^{s}\hat{\epsilon_2}\right)^2\right]^{-1}\right).$$

7. Draw the misclassification probabilities, $\Theta = \left(\theta_{1|0}, \theta_{0|1}, \theta_{0|0}, \theta_{1|1}\right)$. To obtain draws from the posterior distribution of Θ, first recall the conditional likelihood function of the observed treatment indicator given in equation 6 above, and denote the natural log of equation 6 by $L\left(\Theta\right)$. We follow the standard Metropolis-Hastings algorithm to obtain draws from the posterior distribution. First, we obtain candidate draws

$$\begin{aligned}
\theta_{1|0}^{c} &= \theta_{1|0}^{s-1} + \rho_1 \text{ and} \\
\theta_{0|1}^{c} &= \theta_{0|1}^{s-1} + \rho_2,
\end{aligned}$$

where ρ_1 and ρ_2 are independent draws from normal distributions with mean zero and variance t_1^2 and t_2^2, respectively. In practice, the t_1 and t_2 represent the tuning parameters, which we adjust (if necessary) every 10 iterations to ensure the probability of accepting the candidate draws remains at acceptable levels. In our case, we adjust the tuning parameter up or down to maintain an acceptance probability of between 20% and 60%. We accept the candidate draw Θ^c with probability $\min\{1, p\}$, where $p = \exp\left\{L\left(\Theta^{c}\right) - L\left(\Theta^{s-1}\right)\right\}$.

4 MCMC Simulations

4.1 Simulated Data

To test our estimation procedure, we present a variety of simulations consistent with the model presented in equations (1)-(3). Our "baseline" simulation consists of $N = 5,000$ people. To simulate covariates in the outcome equation, we generate random draws for $\mathbf{X}_1 = \{\mathbf{x}_{1,1}, \mathbf{x}_{1,2}\}$, where

$$\mathbf{x}_{1,1} \sim \text{LN}(-1, 1) \text{ and}$$
$$\mathbf{x}_{1,2} \sim \text{N}(-1, 3),$$

where $\text{LN}(-1, 1)$ denotes a log-normal distribution with a mean of e^{-1} and standard deviation of e, and $\text{N}(-1, 3)$ denotes a normal distribution with a mean of -1 and standard deviation of 3. For the selection equation covariates, we set $\mathbf{x}_{2,1} = \mathbf{x}_{1,1}$, $\mathbf{x}_{2,2} = \mathbf{x}_{1,2}$, and $\mathbf{x}_{2,3} \sim U[-2, 1]$. We therefore identify the selection equation coefficients by excluding $\mathbf{x}_{2,3}$ from the outcome equation.

The observed outcome and latent participation variables are then simulated as follows:

$$y_{1i} = \mathbf{X}_{1i}'\gamma + \delta_i\alpha + \epsilon_{1i}$$
$$y_{2i}^* = \mathbf{X}_{2i}'\beta + \epsilon_{2i}, \text{ where}$$

$$\gamma = \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \alpha = 1.5, \beta = \begin{pmatrix} -1 \\ 2 \\ 1/2 \end{pmatrix}, \text{ and } \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix} \sim \text{N}\left(\mathbf{0}, \begin{pmatrix} 2 & 0.1 \\ 0.1 & 1 \end{pmatrix}\right).$$

We also set the misclassification probabilities $\theta_{1|0} = 0.3$ and $\theta_{0|1} = 0.1$.

Given Θ and the actual participation $\delta_i = \mathbf{1}\{y_{2i}^* > 0\}$, we simulate the observed participation y_{2i} as follows. We first generate a random number $r \sim U[0, 1]$. If $\delta_i = 1$, then $y_{2i} = \delta_i$ with probability $\theta_{1|1} = 1 - \theta_{0|1} = 0.9$, which we simulate by setting $y_{2i} = 1$ if $r \leq \theta_{1|1} = 0.9$ (or equivalently $r > \theta_{0|1} = 0.1$). Otherwise, we set $y_{2i} = 0$. Alternatively, if $\delta_i = 0$, then $y_{2i} = \delta_i$ with probability $\theta_{0|0} = 1 - \theta_{1|0} = 0.7$, which we simulate by setting $y_{2i} = 1$ if $r \leq \theta_{1|0} = 0.3$. Otherwise, $y_{2i} = 0$.

We refer to the above simulation as the baseline data, and consider alternative specifications as follows: 1) the number of individuals analyzed ($N = 5,000$ and $N = 500$); 2) the magnitude of the variance-covariance matrix Σ ($\sigma_{11} = 2$, $\sigma_{12} = 0.1$; $\sigma_{11} = 1$, $\sigma_{12} = -0.9$; $\sigma_{11} = 9$, $\sigma_{12} = -1.5$; and $\sigma_{11} = 2$, $\sigma_{12} = 0$);

3) values of the misclassification probabilities ($\theta_{1|0} = 0.3$, $\theta_{0|1} = 0.1$; $\theta_{1|0} = \theta_{0|1} = 0.4$; and $\theta_{1|0} = \theta_{0|1} = 0$); 4) exclusion of relevant variables in the selection equation; and 5) non-normal residuals. This yields a total of ten different simulated datasets. In each case, we are interested in the effect of misclassification and endogeneity on the coefficient estimates, particularly the treatment effect α. We therefore estimate each of the ten simulated datasets under four alternative specifications of the data generating process: 1) a correctly specified model accounting for both endogeneity and misclassification ("Model A"); 2) a model accounting for endogeneity but not misclassification ("Model B"); 3) a model accounting for misclassification but not endogeneity ("Model C"); and 4) a model ignoring both misclassification and endogeneity ("Model D"). Models B-D are therefore intentionally misspecified models designed to assess the impact of failing to account for certain aspects of the underlying data.

4.2 MCMC Results

We follow the steps in Section 3 above for 7,000 draws and drop the first 2,000. The posterior means and standard deviations for the baseline data are summarized in Table 4.1 below.

As Table 4.1 shows, the estimator proposed in this paper accurately estimates the coefficients in the selection and outcome equations (β_1, β_2, β_3, γ_1, and γ_2). More importantly, our approach accurately estimates the treatment effect (α) as well as the misclassification probabilities ($\theta_{0|1}$ and $\theta_{1|0}$) using no additional information than what is used in the standard treatment model with selection on observables. We also note that, consistent with Bollinger (1996) and others, σ_{12} is only partially identified in our model. Simulation results also show that σ_{12} is point-identified under certain restrictions on the misclassification process. However, the focus of our proposed estimator is on the accurate estimation of the misclassification probabilities and the treatment effect. As such, additional identification considerations regarding σ_{12} are beyond the scope of this paper.

Consistent with the general misclassification literature, Table 4.1 shows that failing to account for misclassification leads to biased coefficient estimates in the selection equation, where the posterior means for β_1 and β_2 are each at least approximately 30 standard deviations away from their respective true values. Failing to account for misclassification can also introduce bias in the estimate of the treatment effect. In this case, the estimated treatment effect of 1.299 is over 4 standard deviations away from the true treatment effect of 1.5.[6]

[6]These results are consistent across each of the ten data generating processes considered. The corresponding tables of posterior means and standard deviations for the remaining eight datasets are presented in the attached Appendix A.

Table 4.1: MCMC Results for Baseline Model

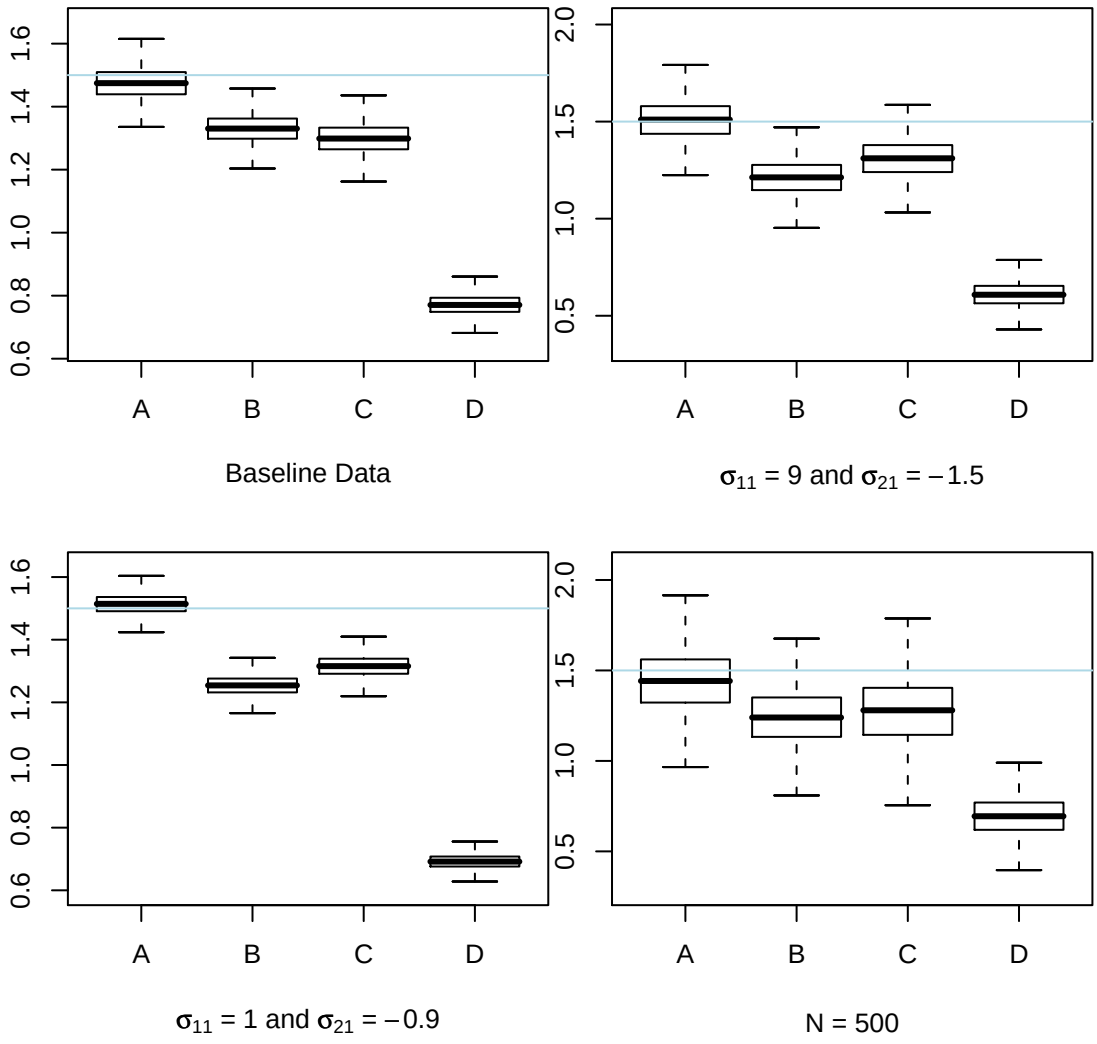
	True Value	**Posterior Means**[a]			
		Model A	Model B	Model C	Model D
γ_1	-1	-1.004	-1.077	-0.978	-0.916
		(0.024)	(0.026)	(0.024)	(0.023)
γ_2	1	1.002	1.089	1.014	1.107
		(0.008)	(0.007)	(0.008)	(0.007)
α	1.5	1.475	1.299	1.331	0.771
		(0.052)	(0.051)	(0.047)	(0.033)
β_1	-1	-0.994	0.058	-0.835	0.058
		(0.122)	(0.019)	(0.123)	(0.019)
β_2	2	2.093	0.192	2.002	0.187
		(0.109)	(0.007)	(0.111)	(0.007)
β_3	0.5	0.396	-0.059	0.376	-0.064
		(0.078)	(0.019)	(0.097)	(0.020)
σ_{11}	2	2.066	2.380	2.105	2.265
		(0.042)	(0.053)	(0.044)	(0.045)
σ_{12}	0.1	-0.295	-0.568	0	0
		(0.051)	(0.042)		
$\theta_{1\|0}$	0.3	0.313	0	0.310	0
		(0.009)		(0.008)	
$\theta_{0\|1}$	0.1	0.105	0.000	0.106	0.000
		(0.008)		(0.009)	

[a]Standard deviations are in parenthesis.

One of the primary interests of this paper is the accurate estimation of the treatment effect in light of misclassification and endogenous treatment. From Table 4.1, it is clear that our proposed estimator can accurately estimate the treatment effect under at least some circumstances; however, in order to assess the robustness of our proposed estimator, we simulated additional datasets and looked for any changes in our estimate of α (relative to the true value) under different data generating processes. Figure 4.1 presents the densities for α under four of the simulated datasets discussed previously, and for each dataset, we present densities for all four models under consideration.

The densities plotted in Figure 4.1 illustrate some important points. First, the posterior distributions for α under the correctly specified models (the solid line) most accurately cover the true values in each of the four datasets considered, while the distributions under the most misspecified model (the dash-dotted line) never cover the true value. Second, there is evidence that accounting for misclassification and endogeneity becomes more important as more information is considered, as illustrated by the increased

Figure 4.1: Box Plots for α^a

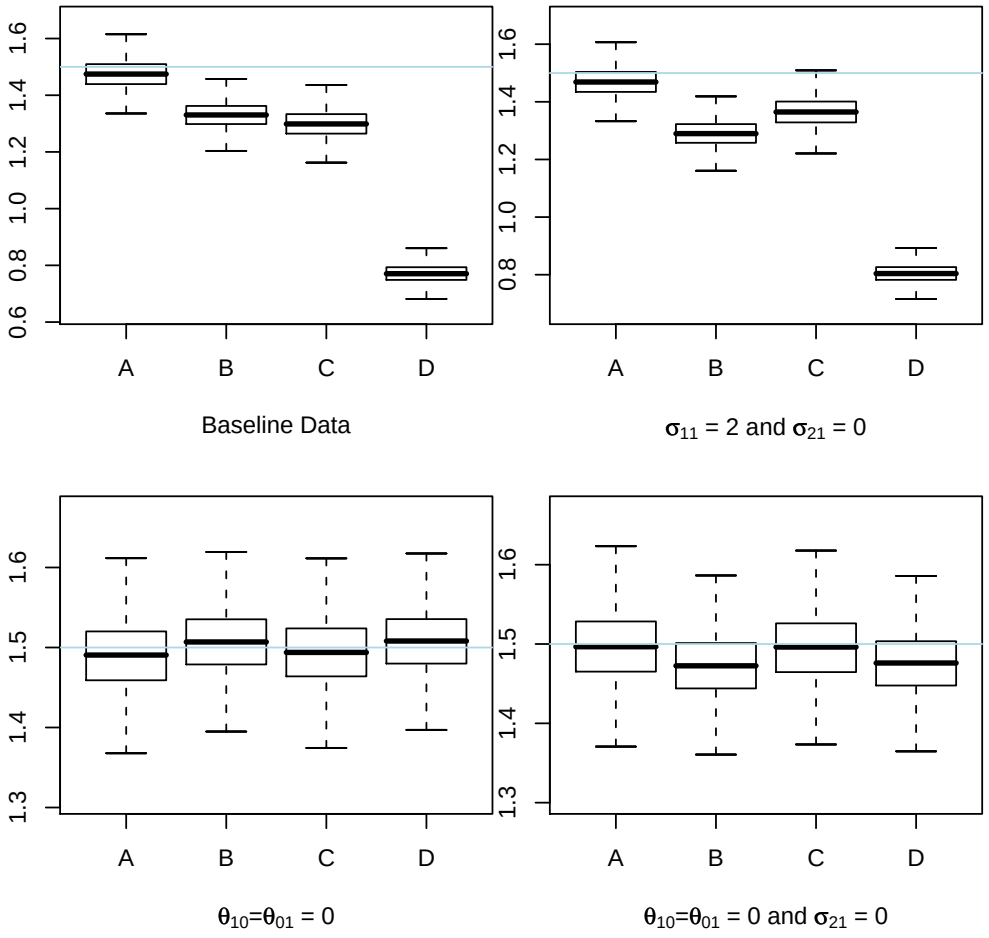


[a]Box plots for Model A (correctly specified model), Model B (ignoring misclassification), Model C (ignoring endogeneity), and Model D (ignoring endogeneity and misclassification).

bias for $N = 5,000$ relative to $N = 500$.

We summarize additional simulations as follows: Figure 4.2 presents densities for the baseline data, data without misclassification but with endogeneity ($\theta_{1|0} = \theta_{0|1} = 0$), data with misclassification but without endogeneity, and data without either misclassification or endogeneity. Figure 4.3 presents densities for the baseline data, data with an excluded relevant variable in the selection equation, data with high misclassification probabilities, and data with non-normal errors.

Figure 4.2: Box Plots for α^a

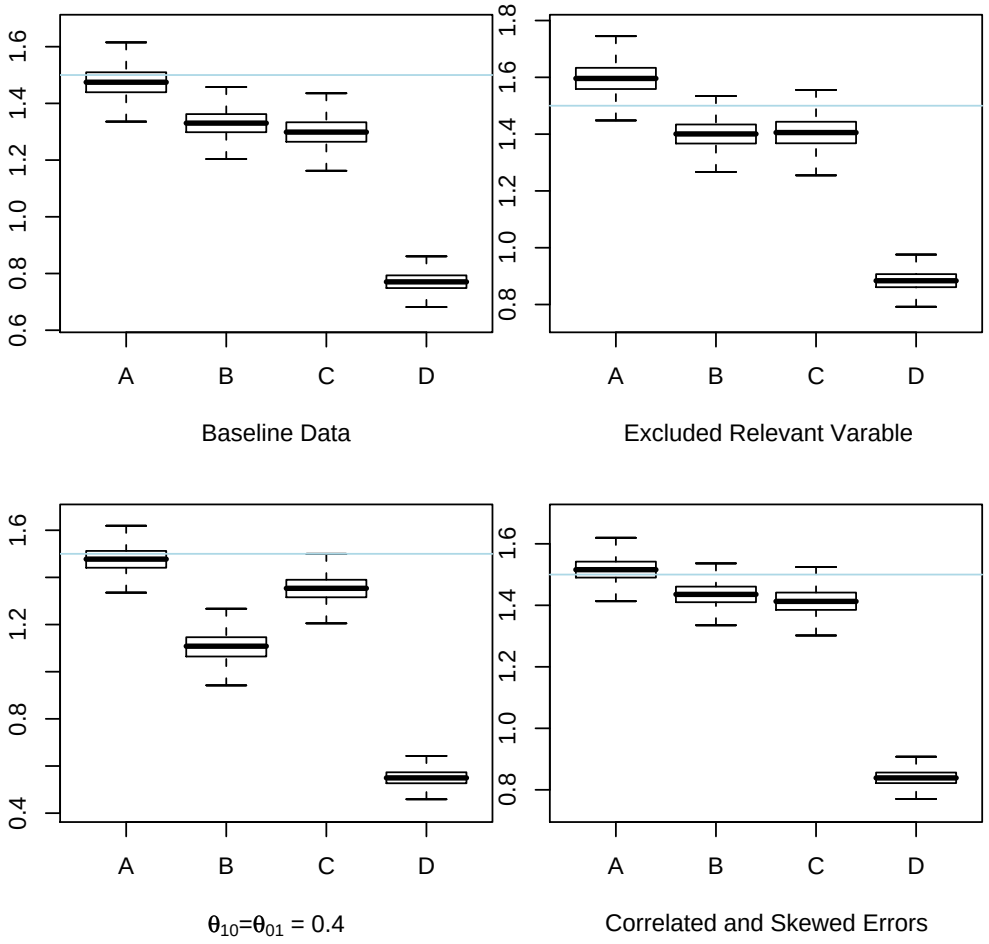


[a]Box plots for Model A (correctly specified model), Model B (ignoring misclassification), Model C (ignoring endogeneity), and Model D (ignoring endogeneity and misclassification).

Two points to note from Figure 4.2 are: 1) as long as there is no misclassification in the true data generating process, a model that assumes $\sigma_{12} = 0$ (Model C) still performs relatively well; and 2) if the true data generating process is subject to misclassification but not endogeneity, a model that ignores misclassification but allows for endogeneity can still improve the estimate of the treatment effect relative alternative specifications. Point (2) is somewhat surprising at first glance. Intuitively, the justification for this result is that even with *ex ante* independence between ϵ_1 and ϵ_2, the outcome equation is still subject to a misclassified binary covariate. And as discussed in Kreider (2010) and others, such a misclassification yields a negative correlation between the true and observed participation. Therefore, the presence of

misclassification introduces endogeneity into the model, and models that ignore potential endogeneity may not accurately estimate the treatment effect. This is also clear in Table A.5, where the posterior mean for σ_{12} is approximately -0.55 despite the true $\sigma_{12} = 0$.

Figure 4.3: Box Plots for α^a



[a]Box plots for Model A (correctly specified model), Model B (ignoring misclassification), Model C (ignoring endogeneity), and Model D (ignoring endogeneity and misclassification).

We are also interested in the performance of our estimator with non-normal errors, excluded relevant variables, and high misclassification probabilities. Figure 4.3 therefore illustrates the draws for α under each of these considerations. The lower-left quadrant presents the posterior distributions for α based on data with high misclassification probabilities ($\theta_{1|0} = \theta_{0|1} = 0.4$), the upper-right quadrant presents draws where a relevant variable has been intentionally excluded from the selection equation, and the lower-left

presents draws with non-normal errors.[7] In all cases, Figure 4.3 illustrates that our proposed estimator accurately estimates the treatment effect in light of non-normal errors, high misclassification probabilities, or excluded relevant variables. Figure 4.3 also illustrates that, with high misclassification probabilities, a model that allows for endogeneity and ignores misclassification largely improves the estimate of α relative to a model that accounts for misclassification but ignores endogeneity. This is consistent with the discussion above regarding *ex post* correlation in the presence of misclassification.

5 Summary and Conclusions

This paper contributes to the growing literature on misclassified and endogenous treatment participation. We propose a Bayesian estimator that can incorporate both misclassified and endogenous treatment in a single framework. MCMC simulations demonstrate that, without accounting for misclassification or endogeniety of treatment assignment, estimates of the treatment effect and other coefficients are inconsistent. We show that our proposed Bayesian estimation procedure and sampling algorithm accurately estimate the treatment effect in light of misclassified and endogenous treatment.

Our work has a number of important and natural extensions to be addressed. First, the misclassification probability could be modeled on an individual level, particularly if there data that can help model these probabilities. For example, the personal misclassification probability could be modeled as a function of both personal data, either included in the selection model, or excluded from it, as well as data from a higher level of aggregation, e.g. state or county level data. Second, a panel data version of this model, where misclassification probabilities could be a function of both past and future participation, could prove to be useful.

[7] Our process for simulating non-normal errors is summarized in Appendix B.

A Summary Tables for Additional Simulations

Table A.1: MCMC Results for $\sigma_{11} = 1$ and $\sigma_{21} = -0.9$

	True Value	**Posterior Means**[a]			
		Model A	Model B	Model C	Model D
γ_1	-1	-1.003	-1.075	-0.951	-0.888
		(0.016)	(0.019)	(0.016)	(0.017)
γ_2	1	1.000	1.090	1.023	1.110
		(0.005)	(0.005)	(0.005)	(0.005)
α	1.5	1.514	1.315	1.254	0.692
		(0.033)	(0.035)	(0.033)	(0.024)
β_1	-1	-0.928	0.052	-0.883	0.050
		(0.084)	(0.019)	(0.116)	(0.019)
β_2	2	1.748	0.190	1.832	0.184
		(0.081)	(0.007)	(0.118)	(0.007)
β_3	0.5	0.413	-0.023	0.537	-0.043
		(0.053)	(0.018)	(0.100)	(0.020)
σ_{11}	1	0.939	1.303	0.987	1.144
		(0.020)	(0.032)	(0.021)	(0.023)
σ_{12}	-0.9	-0.586	-0.666	0	0
		(0.028)	(0.029)		
$\theta_{1\|0}$	0.3	0.321	0	0.310	0
		(0.009)		(0.008)	
$\theta_{0\|1}$	0.1	0.127	0	0.106	0
		(0.010)		(0.009)	

[a]Standard deviations are in parenthesis.

Table A.2: MCMC Results for $\sigma_{11} = 9$ and $\sigma_{21} = -1.5$

	True Value	**Posterior Means**[a]			
		Model A	Model B	Model C	Model D
γ_1	-1	-1.019	-1.083	-0.960	-0.869
		(0.048)	(0.053)	(0.048)	(0.047)
γ_2	1	0.996	1.087	1.022	1.111
		(0.016)	(0.014)	(0.016)	(0.014)
α	1.5	1.509	1.309	1.214	0.609
		(0.108)	(0.103)	(0.097)	(0.067)
β_1	-1	-0.893	0.063	-0.596	0.063
		(0.131)	(0.019)	(0.113)	(0.019)
β_2	2	2.072	0.188	1.697	0.187
		(0.119)	(0.007)	(0.130)	(0.007)
β_3	0.5	0.447	-0.054	0.404	-0.052
		(0.089)	(0.020)	(0.081)	(0.021)
σ_{11}	9	8.760	9.216	8.832	9.011
		(0.179)	(0.194)	(0.178)	(0.181)
σ_{12}	-1.5	-0.577	-0.750	0	0
		(0.108)	(0.087)		
$\theta_{1\|0}$	0.3	0.311	0	0.304	0
		(0.009)		(0.008)	
$\theta_{0\|1}$	0.1	0.114	0	0.107	0
		(0.009)		(0.009)	

[a]Standard deviations are in parenthesis.

Table A.3: MCMC Results for $N = 500$

	True Value	**Posterior Means**[a]			
		Model A	Model B	Model C	Model D
γ_1	-1	-0.975	-1.068	-0.918	-0.840
		(0.093)	(0.111)	(0.092)	(0.092)
γ_2	1	0.982	1.062	1.001	1.086
		(0.027)	(0.025)	(0.027)	(0.024)
α	1.5	1.440	1.276	1.241	0.695
		(0.174)	(0.189)	(0.162)	(0.114)
β_1	-1	-0.847	0.136	-0.808	0.153
		(0.339)	(0.069)	(0.427)	(0.070)
β_2	2	1.937	0.191	2.108	0.191
		(0.289)	(0.022)	(0.587)	(0.021)
β_3	0.5	0.179	0.024	0.313	0.039
		(0.225)	(0.060)	(0.302)	(0.062)
σ_{11}	2	2.239	2.577	2.277	2.430
		(0.147)	(0.187)	(0.147)	(0.154)
σ_{12}	0.1	-0.349	-0.582	0	0
		(0.156)	(0.151)		
$\theta_{1\|0}$	0.3	0.291	0	0.292	0
		(0.026)		(0.027)	
$\theta_{0\|1}$	0.1	0.105	0	0.100	0
		(0.029)		(0.031)	

[a]Standard deviations are in parenthesis.

Table A.4: MCMC Results for $\theta_{1|0} = \theta_{0|1} = 0$

	True Value	**Posterior Means**[a]			
		Model A	Model B	Model C	Model D
γ_1	-1	-1.001	-1.002	-1.005	-1.005
		(0.022)	(0.022)	(0.022)	(0.022)
γ_2	1	1.001	1.001	0.999	0.999
		(0.008)	(0.008)	(0.008)	(0.007)
α	1.5	1.490	1.493	1.507	1.507
		(0.045)	(0.045)	(0.042)	(0.042)
β_1	-1	-1.018	-1.018	-1.019	-1.021
		(0.073)	(0.064)	(0.067)	(0.064)
β_2	2	2.065	2.074	2.055	2.075
		(0.100)	(0.078)	(0.080)	(0.078)
β_3	0.5	0.481	0.491	0.478	0.490
		(0.046)	(0.050)	(0.054)	(0.050)
σ_{11}	2	1.996	1.995	1.995	1.994
		(0.040)	(0.040)	(0.040)	(0.040)
σ_{12}	0.1	0.065	0.053	0	0
		(0.067)	(0.065)		
$\theta_{1\|0}$	0	0.001	0	0.001	0
		(0.001)		(0.001)	
$\theta_{0\|1}$	0	0.001	0	0.001	0
		(0.001)		(0.001)	

[a]Standard deviations are in parenthesis.

Table A.5: MCMC Results for $\sigma_{21} = 0$

	True Value	**Posterior Means**[a]			
		Model A	Model B	Model C	Model D
γ_1	-1	-0.995	-1.087	-0.958	-0.919
		(0.024)	(0.027)	(0.023)	(0.023)
γ_2	1	1.003	1.087	1.019	1.106
		(0.008)	(0.007)	(0.008)	(0.007)
α	1.5	1.469	1.364	1.290	0.804
		(0.051)	(0.053)	(0.048)	(0.033)
β_1	-1	-0.992	0.051	-0.815	0.048
		(0.152)	(0.019)	(0.119)	(0.019)
β_2	2	2.077	0.185	1.699	0.185
		(0.156)	(0.007)	(0.109)	(0.007)
β_3	0.5	0.562	-0.030	0.414	-0.055
		(0.086)	(0.019)	(0.085)	(0.020)
σ_{11}	2	2.048	2.340	2.097	2.209
		(0.042)	(0.053)	(0.044)	(0.044)
σ_{12}	0	-0.328	-0.596	0	0
		(0.053)	(0.045)		
$\theta_{1\|0}$	0.3	0.315	0	0.309	0
		(0.009)		(0.009)	
$\theta_{0\|1}$	0.1	0.110	0	0.104	0
		(0.009)		(0.009)	

[a]Standard deviations are in parenthesis.

Table A.6: MCMC Results for $\sigma_{21} = 0$ and $\theta_{1|0} = \theta_{0|1} = 0$

	True Value	**Posterior Means**[a]			
		Model A	Model B	Model C	Model D
γ_1	-1	-0.996	-0.996	-0.991	-0.991
		(0.023)	(0.022)	(0.022)	(0.022)
γ_2	1	1.001	1.001	1.003	1.003
		(0.008)	(0.008)	(0.008)	(0.007)
α	1.5	1.496	1.495	1.473	1.475
		(0.047)	(0.045)	(0.042)	(0.042)
β_1	-1	-0.996	-0.998	-1.012	-1.000
		(0.057)	(0.063)	(0.071)	(0.063)
β_2	2	1.888	1.894	1.923	1.896
		(0.055)	(0.075)	(0.088)	(0.075)
β_3	0.5	0.456	0.467	0.470	0.465
		(0.048)	(0.049)	(0.046)	(0.049)
σ_{11}	2	1.989	1.988	1.990	1.987
		(0.040)	(0.040)	(0.040)	(0.040)
σ_{12}	0	-0.080	-0.072	0	0
		(0.065)	(0.062)		
$\theta_{1\|0}$	0	0	0	0	0
		(0.000)		(0.000)	
$\theta_{0\|1}$	0	0.002	0	0.002	0
		(0.001)		(0.002)	

[a]Standard deviations are in parenthesis.

Table A.7: MCMC Results for $\theta_{1|0} = \theta_{0|1} = 0.4$

	True Value	**Posterior Means**[a]			
		Model A	Model B	Model C	Model D
γ_1	-1	-0.995	-1.039	-0.992	-0.826
		(0.025)	(0.027)	(0.041)	(0.024)
γ_2	1	1.003	1.146	1.032	1.138
		(0.008)	(0.007)	(0.010)	(0.007)
α	1.5	1.477	1.353	1.105	0.550
		(0.053)	(0.053)	(0.061)	(0.035)
β_1	-1	-1.083	-0.035	0.017	-0.036
		(0.176)	(0.018)	(0.336)	(0.019)
β_2	2	1.933	0.057	1.371	0.057
		(0.153)	(0.006)	(0.232)	(0.006)
β_3	0.5	0.272	0.002	-0.277	0.004
		(0.105)	(0.018)	(0.213)	(0.019)
σ_{11}	2	2.083	2.645	2.201	2.391
		(0.044)	(0.063)	(0.048)	(0.048)
σ_{12}	0.1	-0.375	-0.806	0	0
		(0.046)	(0.042)		
$\theta_{1\|0}$	0.4	0.412	0	0.400	0
		(0.009)		(0.010)	
$\theta_{0\|1}$	0.4	0.412	0	0.423	0
		(0.013)		(0.014)	

[a]Standard deviations are in parenthesis.

Table A.8: MCMC Results for Excluded Relevant Variable

	True Value	**Posterior Means**[a]			
		Model A	Model B	Model C	Model D
γ_1	-1	-1.027	-1.096	-0.982	-0.940
		(0.024)	(0.027)	(0.025)	(0.024)
γ_2	1	0.981	1.076	0.999	1.093
		(0.008)	(0.007)	(0.008)	(0.007)
α	1.5	1.596	1.405	1.401	0.884
		(0.055)	(0.055)	(0.050)	(0.034)
β_1	-1	-0.671	0.052	-0.665	0.051
		(0.095)	(0.018)	(0.107)	(0.019)
β_2	2	1.563	0.184	1.472	0.181
		(0.119)	(0.006)	(0.122)	(0.006)
β_3	0.5	0.452	-0.014	0.428	-0.027
		(0.065)	(0.020)	(0.078)	(0.020)
σ_{11}	2	2.182	2.479	2.236	2.367
		(0.045)	(0.055)	(0.045)	(0.048)
σ_{12}	0.1	-0.312	-0.553	0	0
		(0.047)	(0.046)		
$\theta_{1\|0}$	0.3	0.303	0	0.301	0
		(0.009)		(0.009)	
$\theta_{0\|1}$	0.1	0.106	0	0.102	0
		(0.009)		(0.010)	

[a]Standard deviations are in parenthesis.

Table A.9: MCMC Results for Non-Normal, Correlated Errors

	True Value	**Posterior Means**[a]			
		Model A	Model B	Model C	Model D
γ_1	-1	-1.012	-1.129	-0.989	-0.936
		(0.019)	(0.022)	(0.019)	(0.019)
γ_2	1	1.002	1.088	1.010	1.108
		(0.006)	(0.006)	(0.006)	(0.005)
α	1.5	1.516	1.413	1.436	0.839
		(0.038)	(0.041)	(0.037)	(0.026)
β_1	-1	-1.333	0.068	-1.336	0.067
		(0.113)	(0.020)	(0.135)	(0.021)
β_2	2	2.555	0.183	2.422	0.181
		(0.155)	(0.006)	(0.190)	(0.006)
β_3	0.5	0.773	-0.009	0.787	-0.034
		(0.100)	(0.018)	(0.106)	(0.020)
σ_{11}	1	1.131	1.449	1.160	1.315
		(0.024)	(0.035)	(0.025)	(0.026)
σ_{12}	0.5	-0.177	-0.603	0	0
		(0.034)	(0.035)		
$\theta_{1\|0}$	0.3	0.315	0	0.315	0
		(0.008)		(0.008)	
$\theta_{0\|1}$	0.1	0.092	0	0.089	0
		(0.008)		(0.008)	

[a]Standard deviations are in parenthesis.

B Process for Simulating Non-Normal Errors

We simulated correlated and skewed errors as follows:

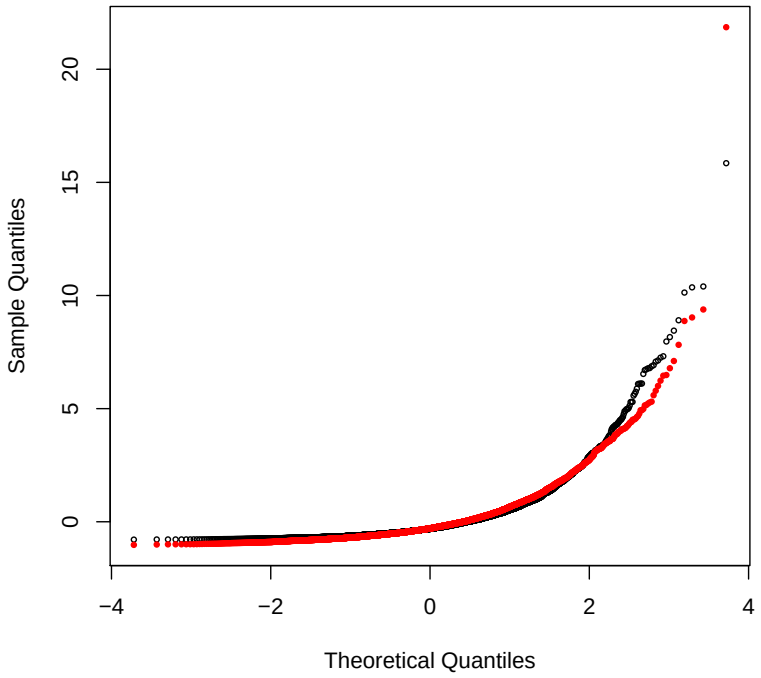
1. Obtain two $N \times 1$ vectors of standard normal errors, ($\epsilon \sim \mathrm{N}(0,1)$);
2. Transform (ϵ_1, ϵ_2) by setting $\tilde{\epsilon} = (\exp\{\epsilon_1\}, \exp\{\epsilon_2\})$;
3. Transform $\tilde{\epsilon}$ by setting $e = \tilde{\epsilon} \times \mathrm{Chol}\left(\mathbf{C}^{-1}\right) \times \mathrm{Chol}(\mathbf{S})$, where $\mathbf{C}$ is the 2×2 variance-covariance matrix of $\tilde{\epsilon}_1$ and $\tilde{\epsilon}_2$, and
$$\mathbf{S} = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix};$$
4. Demean the resulting errors to obtain mean-zero, skewed, and correlated errors.

The resulting errors simulated from this process are mean-zero, with a variance-covariance matrix given by

$$\hat{\mathbf{S}} = \begin{pmatrix} 1.000 & 0.5004 \\ 0.5004 & 1.000 \end{pmatrix},$$

and as indicated in the QQ-plot in Figure B.1, the errors are non-normal.

Figure B.1: Normal Q-Q Plots for Simulated Errors[a]



[a]The red and black plots represent $\tilde{\epsilon}_1$ and $\tilde{\epsilon}_2$, respectively.

References

[1] Abrevaya, J. and J. Hausman (2004). "Response Error in a Transformation Model with an Application to Earnings-Equation Estimation." *Econometrics Journal* 7:366-388.

[2] Aigner, D. (1973). "Regression with a Binary Independent Variable Subject to Errors of Observation." *Journal of Econometrics* 1:49-60.

[3] Albert, J. and S. Chib (1993). "Bayesian Analysis of Binary and Polychotomous Data." *Journal of the American Statistical Association* 88:669-679.

[4] Amemiya, T. (1978). "The Estimation of a Simultaneous Equation Generalized Probit Model." *Econometrica* 46:1193-1205.

[5] Angrist, J., G. Imbens, and D. Rubin (1996). "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association* 91:444-455.

[6] Balcombe, K. and I. Fraser (2009). "Dichotomous-Choice Contingent Valuation with 'Don't Know' Responses and Misreporting." *Journal of Applied Econometrics* 24:1137-1152.

[7] Barron, J., M. Berger, and D. Black (1997). "How Well Do We Measure Training?" *Journal of Labor Economics* 15:507-528.

[8] Bitler, M., J. Currie, and J. Scholz (2003). "WIC Eligibility and Participation." *Journal of Human Resources* 38:1139-1179.

[9] Black, D., M. Berger, and F. Scott (2000). "Bounding Parameter Estimates with Nonclassical Measurement Error." *Journal of the American Statistical Association* 95:739-748.

[10] Black, D., S. Sanders, and L. Taylor (2003). "Measurement of Higher Education in the Census and Current Population Survey." *Journal of the American Statistical Association* 95:545-554.

[11] Blundell, R. and R. Smith (1989). "Estimation in a Class of Simultaneous Equation Limited Dependent Variable Models." *Review of Economic Studies* 56:37-58.

[12] Blundell, R. and R. Smith (1994). "Coherency and Estimation in Simultaneous Models with Censoring and Qualitative Dependent Variables." *Journal of Econometrics* 64:355-373.

[13] Bollinger, C. (1996). "Bounding Mean Regressions when a Binary Regressor is Mismeasured." *Journal of Econometrics* 73:387-399.

[14] Bollinger, C. and M. David (1997). "Modeling Discrete Choice with Response Error: Food Stamp Participation." *Journal of the American Statistical Association* 92:827-835.

[15] Bollinger, C. and M. Hasselt (2009). "A Bayesian Analysis of Binary Misclassification: Inference in Partially Identified Models." Working Paper.

[16] Bound, J., C. Brown, and N. Mathiowetz (2001). "Measurement Error in Survey Data," in J. Heckman and E. Leamer (eds.) *Handbook of Econometrics*, Volume 5, 3705-3843.

[17] Brown, J. and A. Light (1992). "Interpreting Panel Data on Job Tenure." *Journal of Labor Economics* 10:219-257.

[18] Cameron, C. and J. Heckman (1998). "Life Cycle Schooling and Dynamic Selection Bias: Models and Evidence of Five Cohors of American Males." *Journal of Political Economy* 106:262-333.

[19] Carneiro, P., K. Hansen, and J. Heckman (2003). "Estimating Distributions of Treatment Effects With and Application to the Returns to Schooling." *International Economic Review* 44:361-422.

[20] Card, D. (1996). "The Effect of Unions on the Structure of Wages." *Econometrica* 64:957-979.

[21] Chib, S. (1992). "Bayesian Inference in the Tobit Censored Regression Model." *Journal of Econometrics* 51:79-90.

[22] Frazis, H. and M. Lowenstein (2003). "Estimating Linear Regressious with Mismeasured, Possibly Endogeneous, Binary Explanatory Variables." *Journal of Econometrics* 117:151-178.

[23] Gelfand, A., and A. Smith (1990). "Sampling Based Approaches to Calculating Marginal Densities." *Journal of the American Statistical Association* 85:398-409.

[24] Hausman, J., J. Abrevaya, and F. Scott-Morton (1998). "Misclassification of the Dependent Variable in a Discrete-Response Setting." *Journal of Econometrics* 87:239-269.

[25] Heckman, J. (1978). "Dummy Endogenous Variables in a Simultaneous Equation System." *Econometrica* 46:931-959.

[26] Heckman, J. (2003). "Building Bridges between Structural and Program Evaluation Approaches to Evaluating Policy." *Journal of Economic Literature* 48:356-398.

[27] Holland, P. (1986). "Statistics and Causal Inference." *Journal of the American Statistical Association* 81(396): 945-960.

[28] Hu, Y. (2006). "Bounding Parameters in a Linear Regression Model with a Mismeasured Regressor Using Additional Information." *Journal of Econometrics* 133:51-70.

[29] Imbens, G. and J. Angrist (1994). "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62:467-475.

[30] Imbens, G. and C. Manski (2004). "Confidence Intervals for Partially Identified Parameters." *Econometrica* 72:1845-1857.

[31] Imbens, G. and J. Wooldridge (2009). "Recent Developments in the Econometrics of Program Evaluation." *Journal of Economic Literature* 47:5-86.

[32] Klein, R. and F. Vella (2009). "A Semiparametric Model for Binary Response and Continuous Outcomes Under Index Heteroskedasticity." *Journal of Applied Econometrics* 24:735-762.

[33] Kreider, B., J. Pepper, C. Gunderson, and D. Jollife (2009). "Identifying the Effects of Food Stamps on Child Health Outcomes When Participation is Endogenous and Misreported." Staff General Research Papers 13124, Iowa State University, Department of Economics.

[34] Kreirder, B. (2010). "Regression Coefficient Identification Decay in the Presence of Infrequent Classification Errors." *Review of Economics and Statistics*, forthcoming.

[35] Lewbel, A. (2006). "Endogenous Selection or Treatment Model Estimation." Working Paper, Boston College.

[36] Lewbel, A. (2007). "Estimation of Average Treatment Effects with Misclassification." *Econometrica* 75:537-551.

[37] Li, K. (1998). "Bayesian Inference in a Simultaneous Equation Model with Limited Dependent Variables." *Journal of Econometrics* 85:387-400.

[38] Li, T., P. Trivedi, and J. Guo (2003). "Modeling Response Bias in Count: A Structural Approach with an Application to the National Crime Victimization Survey Data." *Sociological Methods & Research* 31:514-544.

[39] Manski, C. (1990). "Nonparametric Bounds on Treatment Effects." *American Economic Review* 80:319-323.

[40] Manski, C. (1994). "The Selection Problem," in C. Sims (ed.) *Advances in Econometrics*, Cambridge: Cambridge University Press.

[41] Millimet, D. (2010). "The Elephant in the Corner: A Cautionary Tale About Measurement Error in Treatment Effects Models," in D. Drukker (ed.) *Advances in Econometrics: Missing Data Methods.*

[42] Millimet, D. and R. Tchernis (2010. "Estimating Treatment Effects Without an Exclusion Resctriction: With an Application to the School Breakfast Program." NBER Working Paper No. 15539.

[43] Molinari, F. (2008). "Partial Identification of Probability Distributions with Misclassified Data." *Journal of Econometrics* 144:81-117.

[44] Rivers, D. and Q. Vuong (1988). "Limited Information Estimators and Exogeneity Tests for Simultaneous Probit Models." *Journal of Econometrics* 39:347-366.

[45] Rossi, P., R. McCulloch, and G. Allenby (1996). "The Value of Purchase History Data in Target Marketing." *Marketing Science* 15:321-340.

[46] Vella, F. (1993). "A Simple Estimator for Simultaneous Models with Censored Endogenous Regressors." *International Economic Review* 34:441-457.

[47] Vella, F. (1998). "Estimating Models With Sample Selection Bias: A Survey." *Journal of Human Resources* 33:127-169.

[48] Sullivan, P. (2007). "Estimation of an Occupational Choice Model when Occupations are Misclassified." *Journal of Human Resources* 44:494-535.