Journal of the American Statistical Association
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/uasa20

A Heckman Selection-t Model

Yulia V. Marchenko [a] & Marc G. Genton [b]
[a] Biostatistics, StataCorp LP, College Station, TX, 77845
[b] Department of Statistics, Texas A&M University, College Station, TX, 77843-3143

Available online: 31 Jan 2012

To cite this article: Yulia V. Marchenko & Marc G. Genton (2012): A Heckman Selection-t Model, Journal of the American Statistical Association, 107:497, 304-317

To link to this article: http://dx.doi.org/10.1080/01621459.2012.656011

A Heckman Selection-*t* Model

Yulia V. MARCHENKO and Marc G. GENTON

Sample selection arises often in practice as a result of the partial observability of the outcome of interest in a study. In the presence of sample selection, the observed data do not represent a random sample from the population, even after controlling for explanatory variables. That is, data are missing not at random. Thus, standard analysis using only complete cases will lead to biased results. Heckman introduced a sample selection model to analyze such data and proposed a full maximum likelihood estimation method under the assumption of normality. The method was criticized in the literature because of its sensitivity to the normality assumption. In practice, data, such as income or expenditure data, often violate the normality assumption because of heavier tails. We first establish a new link between sample selection models and recently studied families of extended skew-elliptical distributions. Then, this allows us to introduce a selection-*t* (SLt) model, which models the error distribution using a Student's *t* distribution. We study its properties and investigate the finite-sample performance of the maximum likelihood estimators for this model. We compare the performance of the SLt model to the conventional Heckman selection-normal (SLN) model and apply it to analyze ambulatory expenditures. Unlike the SLN model, our analysis using the SLt model provides statistical evidence for the existence of sample selection bias in these data. We also investigate the performance of the test for sample selection bias based on the SLt model and compare it with the performances of several tests used with the SLN model. Our findings indicate that the latter tests can be misleading in the presence of heavy-tailed data.

KEY WORDS: Heavy tails; Heckman model; Missing not at random; Sample selection; Skew-normal; Skew-*t*; Selection-normal; Selection-*t*; Two-step.

1. INTRODUCTION

Sample selection arises frequently in applications in many fields, including economics, biostatistics, finance, sociology, and political science, to name a few. Sample selection is a special case of a more general concept known in the econometrics literature as limited dependent variables—variables observed over a limited range of their support.

Let $Y^* \in \mathbb{R}$ be our outcome of interest. Suppose that we observe $Y = Y^*$ only when some unobserved random variable $U^* \in \mathbb{R}$ belongs to a subset $C \in \mathbb{R}$ of its support such that $0 < \Pr(U^* \in C) < 1$. That is, Y is subject to hidden truncation (or simply truncation when $U^* = Y^*$). Model parameters underlying Y^* are then estimated from the observed Y using the conditional density $f(Y|U^* \in C)$. In practice, truncation arises when the collected sample represents only a subset of a full population, for example, a sample of individuals with incomes below or above some threshold. Sometimes, the collected sample does represent a full population but because of some hidden truncation $U^* \in \mathbb{R}$, the outcome of interest Y^* is not observed for all of the participants. In this case, Y^* is subject to incidental truncation or sample selection (e.g., Greene 2008). The problem of sample selection or, more specifically, sample selection bias, arises when Y^* and U^* are correlated and, thus, must be modeled jointly. That is, inference based on only observed Y would not be valid. This problem is also known as data missing not at random (MNAR; Rubin 1976). For example, in a study of incomes, people with high (or low) income may be less likely to report it than people with average income. In the presence of sample selection, we observe an indicator $U = I(U^* \in C)$ and other explanatory variables for all of the sample and the outcome $Y = Y^*$ for part of the sample identified by $U = 1$. Thus, the sample selection model is composed of the continuous component $f(Y|U = 1)$ and the discrete component $\Pr(U)$.

The classical sample selection model was introduced by Heckman (1974) in the mid-1970s when he proposed a parametric approach to the estimation under the assumption of bivariate normality between Y^* and U^*. The main criticism of the proposed method was the sensitivity of the parameter estimates to the assumption of normality, often violated in practice, which led to his developing a more robust estimation procedure in the late 1970s, known as Heckman's two-step (TS) estimator (Heckman 1979). Both estimation methods were found to be sensitive to high correlation between variables of the outcome and selection equations, as is often encountered in practice (e.g., see Puhani 2000 and the references therein). Various other robust-to-normality methods have been proposed over the years for the analysis of a sample selection model, including a number of semiparametric (e.g., Ahn and Powell 1993; Powell 1994; Newey 1999) and nonparametric (e.g., Das, Newey, and Vella 2003) methods, relaxing the distributional assumption in general. For example, see Vella (1998), Matzkin (1994), and Li and Racine (2007, p. 315) for a general overview of methods for selection models.

In this article, we develop and study the properties of yet another estimation approach. Our approach maintains the original parametric framework of the model but considers a bivariate Student's t distribution as the underlying joint distribution of (Y^*, U^*) and estimates the parameters via maximum likelihood. Among various departures from normality occurring in practice, one of the most common is when the distribution of the data has heavier tails than in the normal distribution, such as the distribution of (log) incomes in the population. This makes it natural

Yulia V. Marchenko is Director of Biostatistics, StataCorp LP, College Station, TX 77845 (E-mail: *ymarchenko@stata.com*). Marc G. Genton is Professor, Department of Statistics, Texas A&M University, College Station, TX 77843-3143 (E-mail: *genton@stat.tamu.edu*). The authors thank the editor, the associate editor, and two referees for a careful review of the manuscript and valuable suggestions. The authors also thank Adelchi Azzalini and Reinaldo B. Arellano-Valle for helpful comments on an earlier version of the article, and David Drukker for useful discussions about selection models. Genton's research was partially supported by the National Science Foundation (NSF) grant DMS-1007504 and by award number KUS-C1-016-04 made by King Abdullah University of Science and Technology (KAUST).

Journal of the American Statistical Association
March 2012, Vol. 107, No. 497, Theory and Methods
DOI: 10.1080/01621459.2012.656011

to consider a Student's t distribution as an underlying joint distribution for the selection model. The robustness properties of the Student's t distribution (Lange, Little, and Taylor 1989; Azzalini and Genton 2008; DiCiccio and Monti 2009) also make it an attractive parametric alternative to the normal distribution. For example, this distribution has been used recently to relax the assumption of normality in various statistical models, such as censored regression (Arellano-Valle, Castro, González-Farías, and Muñoz-Gajardo 2011), treatment models (Chib and Hamilton 2000; Heckman, Tobias, and Vytlacil 2003), and switching regression (Scruggs 2007), to name a few. Lee (1982) considered the Student's t distribution within a transformation-based approach for the correction of the selection bias and estimated parameters using a two-stage method. Our approach uses full maximum likelihood estimation, which is fully efficient under the correct model specification and also allows simultaneous estimation of the degrees-of-freedom parameter. The selection bias test based on the selection-normal (SLN) model can be affected by heavy tails, as we demonstrate in our simulation, and our method addresses this issue. Our motivation for considering the Student's t distribution is also prompted by the existence of a link between the continuous part of the selection model and an extended skew-t distribution, studied extensively in the recent literature (Arellano-Valle and Genton 2010a).

The article is organized as follows. Section 2 describes the classical sample selection model and introduces the selection-t (SLt) model. The finite-sample performance of the SLt model is evaluated numerically and compared with that of the classical SLN model in Section 3. A numerical application of the SLt model is presented in Section 4. The article concludes with a discussion in Section 5. The relevant analytical results are given in the Appendices.

2. SELECTION MODELS

In this section, we first describe the classical sample selection model and its two commonly used estimation methods, maximum likelihood and TS. Next, we comment on the link between sample selection models and a family of skew-elliptical distributions. Finally, we formulate the SLt model and study its properties.

2.1 Classical Heckman Sample Selection Model

Suppose that the regression model of primary interest is

$$y_i^* = \mathbf{x}_i^\top \boldsymbol{\beta} + \epsilon_i, \quad i = 1, \ldots, N. \tag{1}$$

However, due to a certain selection mechanism,

$$u_i^* = \mathbf{w}_i^\top \boldsymbol{\gamma} + \eta_i, \quad i = 1, \ldots, N, \tag{2}$$

we observe only N_1 out of N observations y_i^* for which $u_i^* > 0$:

$$\begin{aligned} u_i &= I(u_i^* > 0), \\ y_i &= y_i^* u_i. \end{aligned} \tag{3}$$

Latent variables $y_i^* \in \mathbb{R}$ and $u_i^* \in \mathbb{R}$ are associated with primary and selection regressions, respectively; y_i is the observed counterpart of y_i^* and u_i is an indicator of whether the primary dependent variable is observed. The vectors $\boldsymbol{\beta} \in \mathbb{R}^p$ and $\boldsymbol{\gamma} \in \mathbb{R}^q$ are unknown parameters; the vectors $\mathbf{x}_i \in \mathbb{R}^p$ and $\mathbf{w}_i \in \mathbb{R}^q$ are observed characteristics; and ϵ_i and η_i are error terms from a bivariate normal distribution:

$$\begin{pmatrix} \epsilon_i \\ \eta_i \end{pmatrix} \sim N_2 \left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \rho\sigma \\ \rho\sigma & 1 \end{pmatrix} \right\}. \tag{4}$$

The SLN model (1)–(3) is known as "Type 2 tobit model" in the econometrics literature (Amemiya 1985, p. 385) and is sometimes also referred to as the "Heckman model."

Because we only observe the sign of u_i^* in (3), its variance is nonidentifiable and, without loss of generality, is set to 1 in (4). The parameter $\rho \in (-1, 1)$ governs the "selection bias" that arises when $\rho \neq 0$ and the standard ordinary least squares (OLS) regression is used to estimate $\boldsymbol{\beta}$ in (1). The zero threshold in (3) is arbitrary; any other constant threshold $c \neq 0$ would be absorbed by the intercept in (2).

As we mentioned in the introduction, the density of the sample selection model is composed of a continuous component corresponding to the conditional density $f(y|U = 1)$ and a discrete component $\Pr(U)$. The discrete component is described by the probit model $\Pr(U = u) = \{\Phi(\mathbf{w}^\top \boldsymbol{\gamma})\}^u \{\Phi(-\mathbf{w}^\top \boldsymbol{\gamma})\}^{1-u}$, where $\Phi(\cdot)$ is the standard normal cumulative distribution function. The conditional density is

$$\begin{aligned} &f(y|U = 1) \\ &\quad = \frac{1}{\sigma} \phi\left(\frac{y - \mathbf{x}^\top \boldsymbol{\beta}}{\sigma} \right) \frac{\Phi\left\{ \frac{\rho}{\sqrt{1-\rho^2}} \left(\frac{y - \mathbf{x}^\top \boldsymbol{\beta}}{\sigma} \right) + \frac{\mathbf{w}^\top \boldsymbol{\gamma}}{\sqrt{1-\rho^2}} \right\}}{\Phi(\mathbf{w}^\top \boldsymbol{\gamma})}, \end{aligned} \tag{5}$$

where $\phi(\cdot)$ denotes the standard normal density.

Then, the log-likelihood function for this model based on a single pair of observations (y, u) can be written as

$$\begin{aligned} l(\boldsymbol{\beta}, \boldsymbol{\gamma}, \rho, \sigma; y, u) &= u \ln f(y|U = 1) + u \ln \Phi(\mathbf{w}^\top \boldsymbol{\gamma}) \\ &\quad + (1 - u) \ln \Phi(-\mathbf{w}^\top \boldsymbol{\gamma}) \\ &= u \left[\ln \Phi \left\{ \frac{\mathbf{w}^\top \boldsymbol{\gamma} + \rho(y - \mathbf{x}^\top \boldsymbol{\beta})/\sigma}{\sqrt{1 - \rho^2}} \right\} \right. \\ &\quad \left. - \frac{1}{2} \left(\frac{y - \mathbf{x}^\top \boldsymbol{\beta}}{\sigma} \right)^2 - \ln(\sqrt{2\pi}\sigma) \right] \\ &\quad + (1 - u) \ln \Phi(-\mathbf{w}^\top \boldsymbol{\gamma}). \end{aligned} \tag{6}$$

Heckman (1974) showed that the maximum likelihood estimators (MLEs) obtained from the maximization of (6) have the same properties as the conditional MLEs, although (6) does not correspond to a true conditional density. When the distributional assumption of bivariate normality is correct, MLEs are fully efficient.

A method more robust to the normality assumption was proposed by Heckman (1979) and is known as the two-step estimation. The motivation for the TS estimation is based on the fact that the conditional expectation of the observed data is

$$E(Y|U^* > 0, \mathbf{x}, \mathbf{w}) = \mathbf{x}^\top \boldsymbol{\beta} + \rho\sigma\lambda(\mathbf{w}^\top \boldsymbol{\gamma}), \tag{7}$$

where $\lambda(\cdot) = \phi(\cdot)/\Phi(\cdot)$ denotes the inverse Mills ratio. The inconsistency of the classical OLS estimator from regression of y on $\mathbf{x}$ is explained by the existence of an extra term $\rho\sigma\lambda(\mathbf{w}^\top \boldsymbol{\gamma})$ when $\rho \neq 0$ in the regression function (7), which is omitted from the regression. The TS method is designed to correct the OLS regression for the omitted term and consists of two stages. At the first stage, the probit model $\Pr(U|\mathbf{w})$ is

fit to the data and the MLEs of $\boldsymbol{\gamma}$ are obtained. At the second stage, $\boldsymbol{\beta}$ and $\beta_\lambda = \rho\sigma$ are estimated by least squares regression of y on $\mathbf{x}$ and $\hat{\lambda}$, where $\hat{\lambda} = \phi(\mathbf{w}^\top\hat{\boldsymbol{\gamma}})/\Phi(\mathbf{w}^\top\hat{\boldsymbol{\gamma}})$. The consistent estimates of ρ and σ can then be obtained from $\hat{\beta}_\lambda$, least squares residual variance, and average-predicted probabilities from the probit model (e.g., Greene 2008, p. 886). The advantage of the TS method over maximum likelihood is that no distributional assumption is required for the error term of the second-stage equation for the consistency of the estimator. However, this method suffers from possible collinearity problems when $\mathbf{w}$ includes some of the covariates from the primary regression because of the linearity of the inverse Mills ratio $\lambda(\cdot)$ in a wide range of its support. In fact, both methods tend to perform poorly in the presence of high correlation between error terms and high collinearity among regressors in the primary and selection equations.

The sensitivity of the two methods to the collinearity among regressors in the primary and selection equations has been shown to be even more of an issue in practical applications than the misspecified error distribution (Puhani 2000). When $\mathbf{w} = \mathbf{x}$, which is not an unusual assumption in many practical settings, the identifiability of the parameters relies heavily on the functional form of the distribution. In particular, for the SLN model, the identification of the regression parameters is achieved through the nonlinearity of the inverse Mills ratio. The problem arises because of the linearity of the inverse Mills ratio $\lambda(\cdot)$ in a wide range of its support (see Figure 1). To alleviate this problem, the econometrics literature suggests to impose an exclusion restriction according to which at least one extra variable that is a good predictor of u_i^* is included in the selection equation and does not appear in the primary regression. In practice, however, it can be difficult to find such variables (e.g., Puhani 2000) because strong predictors of the selection equation are usually also strong predictors of the primary equation and thus should be included in the primary regression as well.

2.2 Link to the Family of Extended Skew-Elliptical Distributions

The continuous component of the sample selection density, the conditional density (5), corresponds to the extended skew-normal distribution (Azzalini 1985), studied in more detail recently by Arellano-Valle and Genton (2010a),

$$f_{\mathrm{ESN}}(y;\mu,\sigma^2,\alpha,\tau) = \frac{1}{\sigma}\phi\left(\frac{y-\mu}{\sigma}\right)\frac{\Phi\left(\alpha\frac{y-\mu}{\sigma}+\tau\right)}{\Phi(\tau/\sqrt{1+\alpha^2})}, \quad y\in\mathbb{R}, \tag{8}$$

with parameterization $\mu = \mathbf{x}^\top\boldsymbol{\beta}\in\mathbb{R}, \alpha = \rho/\sqrt{1-\rho^2}\in\mathbb{R}$, and $\tau = \mathbf{w}^\top\boldsymbol{\gamma}/\sqrt{1-\rho^2}\in\mathbb{R}$. The parameterization of the mean μ is simply the conventional parameterization used in the regression setting. The parameterization of the shape parameter α corresponds to the so-called "δ-parameterization" ($\delta = \rho$ in this case) arising from the stochastic representation of a skew-normal random variable (Azzalini 2005), also discussed below. The key distinction here is the parameterization of the shift parameter τ. In the sample selection setting, similar to the mean μ, τ is parameterized as a linear function of the predictors $\mathbf{w}$ from the selection equation. Thus, the model includes a $q\times 1$ vector of unknown coefficients $\boldsymbol{\gamma}$ rather than a single shift parameter τ.

Azzalini (1985) and Copas and Li (1997), among others, note that the distribution of $Y^*|U^*>0$, arising from hidden truncation when Y^* and U^* are jointly normal and the marginal distribution of U^* is standard normal, belongs to the family of skew-normal distributions (Azzalini 1985, 2005; Genton 2004), which is a special case of (8) with $\tau = 0$. In fact, the selection mechanism $Y^*|U^*>0$ corresponds to one of the stochastic representations of a skew-normal random variate when the location of U^* is zero, and an extended skew-normal random variate when the location of U^* is different from zero. Based on this link, it would be natural to use the skew-normal distribution to model data arising from hidden truncation under the assumption of the underlying bivariate normal distribution. For example, we can fit the skew-normal model, available in statistical packages R (Azzalini 2006) and Stata (Marchenko and Genton 2010), to the observed data to account for asymmetry in the distribution often induced by hidden truncation. Even in the sample selection setting, if we fit the extended skew-normal model (8) with regression parameterization of τ and μ to the observed data only, we will obtain consistent results, unlike the OLS regression.

We can also use this link to help study the properties of the sample selection model. From (6), the log-likelihood for the SLN model consists of two components. The first component is the log-likelihood of the extended skew-normal model for the observed data, and the second component is the probit log-likelihood describing the selection process. It is known that the profile log-likelihood for α (or ρ in the "δ-parameterization") of the skew-normal and the extended skew-normal models has a stationary point at $\alpha = 0$ (or $\rho = 0$), which leads to the singularity of the Fisher information and observed information matrices at that point. The singularity is caused by the chosen parameterization, as opposed to the unidentifiability of the model parameters in general. The sample selection model uses a similar parameterization. Also, in the sample selection framework, the hypothesis H_o: $\rho = 0$ is important for testing the existence of the sample selection bias. Thus, the stationarity of the likelihood at $\rho = 0$ would create difficulty in testing this hypothesis within the likelihood framework. As it turns out, the SLN model does not exhibit this property. The stationarity issue for the extended skew-normal model arises because the scores of the parameters are linearly dependent at $\alpha = 0$ (or $\rho = 0$) (Arellano-Valle and Genton 2010a). The scores for $\boldsymbol{\beta}$, σ, and ρ for the SLN model are linearly related. However, the score for $\boldsymbol{\gamma}$ is not zero, unlike the score for τ for the extended skew-normal model, and is not linearly dependent on any of the other scores. Hence, the observed information is not singular at $\rho = 0$. See Appendix C for details.

More generally, Arellano-Valle, Branco, and Genton (2006) unify all of the distributions arising from selection mechanisms $\mathbf{Y}^*|\mathbf{U}^*\in C$, where $\mathbf{Y}^*\in\mathbb{R}^{d_1}$, $\mathbf{U}^*\in\mathbb{R}^{d_2}$, and C is a measurable subset in $\mathbb{R}^{d_2}$ such that $0<\Pr(\mathbf{U}^*\in C)<1$, in a broad class of selection distributions. For example, if (Y^*, U^*) follow a bivariate elliptically contoured distribution (e.g., Fang, Kotz, and Ng 1990), then $Y^*|U^*>0$ has an extended skew-elliptical distribution (Arellano-Valle and Azzalini 2006; Arellano-Valle and Genton 2010b). Specifically, let $\mathrm{EC}_2(\boldsymbol{\xi}, \boldsymbol{\Omega}, g^{(2)})$ denote a family

of bivariate elliptically contoured distributions (with existing density) with a generator function $g^{(2)}(\cdot)$ defining a spherical bivariate density, a location column vector $\boldsymbol{\xi} \in \mathbb{R}^2$, and a positive definite scale matrix $\boldsymbol{\Omega} \in \mathbb{R}^{2\times 2}$. If $\mathbf{X} \sim \mathrm{EC}_2(\boldsymbol{\xi}, \boldsymbol{\Omega}, g^{(2)})$, then its density is $f_2(\mathbf{x}; \boldsymbol{\xi}, \boldsymbol{\Omega}, g^{(2)}) = |\boldsymbol{\Omega}|^{-1/2} g^{(2)}(Q_x)$, where $Q_x = (\mathbf{x}-\boldsymbol{\xi})^\top \boldsymbol{\Omega}^{-1}(\mathbf{x}-\boldsymbol{\xi})$ and $\mathbf{x} \in \mathbb{R}^2$ (Fang et al. 1990, p. 46). If

$$\begin{pmatrix} Y^* \\ U^* \end{pmatrix} \sim \mathrm{EC}_2 \left\{ \boldsymbol{\xi} = \begin{pmatrix} \mu \\ \mu_u \end{pmatrix}, \ \boldsymbol{\Omega} = \begin{pmatrix} \sigma^2 & \rho\sigma \\ \rho\sigma & 1 \end{pmatrix}, g^{(2)} \right\}, \quad (9)$$

then $Y^*|U^* > 0 \sim \mathrm{ESE}(\mu, \sigma^2, \alpha, \tau, g_z)$, an extended skew-elliptical distribution with the location parameter μ, the scale parameter σ^2, the shape parameter $\alpha = \rho/\sqrt{1-\rho^2}$, and the shift parameter $\tau = \mu_u/\sqrt{1-\rho^2}$. The corresponding extended skew-elliptical density is

$$f_{\mathrm{ESE}}(y; \mu, \sigma^2, \alpha, \tau, g_z) = \frac{1}{\sigma} f(z; g^{(1)}) \frac{F(\alpha z + \tau; g_z)}{F(\tau/\sqrt{1+\alpha^2}; g^{(1)})}, \ y \in \mathbb{R}, \quad (10)$$

where $z = (y-\mu)/\sigma$, $g_z(v) = g^{(2)}(v+z^2)/g^{(1)}(v)$, and $f(\cdot; g^{(1)})$ and $F(\cdot; g^{(1)})$ are the density and cumulative distribution function, respectively, of a univariate standard elliptical distribution with a generator function $g^{(1)}(\cdot)$, and $F(\cdot; g_z)$ is the cumulative distribution function of a univariate standard elliptical distribution with a generator function $g_z(\cdot)$. The extended skew-normal density (8) is a special case of (10) with the normal generator function $g^{(2)}(v) = \frac{1}{2\pi} e^{-v/2}$. When $\mu_u = 0$, (10) reduces to the family of skew-elliptical distributions studied by Branco and Dey (2001).

We can use this more general link to build more flexible parametric sample selection models, relaxing the classical assumption of underlying normality, as we demonstrate in the next subsection, using the Student's t distribution. If we consider an underlying bivariate elliptically contoured distribution (9), the continuous component of the resulting sample selection model will correspond to the extended skew-elliptical distribution (10). Following (6), the likelihood function will include the likelihood for the extended skew-elliptical model and the likelihood for the corresponding binary elliptical model. In the spirit of the SLN model, we can investigate the properties of such flexible sample selection models using some properties established for extended skew-elliptical distributions.

2.3 Selection-t Model

Using the link between sample selection models and extended skew-elliptical distributions, discussed in the previous subsection, we relax the assumption of bivariate normality and consider the case when the underlying error distribution is a bivariate Student's t distribution. That is, a SLt model is defined by (1)–(3) with bivariate Student's t error distribution:

$$\begin{pmatrix} \epsilon_i \\ \eta_i \end{pmatrix} \sim t_2 \left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \rho\sigma \\ \rho\sigma & 1 \end{pmatrix}, \nu \right\}, \quad (11)$$

where $t_2(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Omega}, \nu) = \frac{1}{2\pi} |\boldsymbol{\Omega}|^{-1/2} \{1 + \frac{(\mathbf{y}-\boldsymbol{\mu})^\top \boldsymbol{\Omega}^{-1}(\mathbf{y}-\boldsymbol{\mu})}{\nu}\}^{-(\nu+2)/2}$ is the density of a bivariate Student's t distribution. In this case, $\rho = 0$ does not imply independence of the primary and selection equations, as for the SLN model, unless $\nu = \infty$.

When errors are nonnormally distributed, Lee (1982, 1983) proposed a general estimation method transforming the error components to bivariate normality. He applied it to the case when the error distribution is bivariate t and estimated parameters using a two-stage method. We focus on the full maximum likelihood estimation of the SLt model (1)–(3) and (11), which is fully efficient under the bivariate-t assumption.

The bivariate Student's t distribution from (11) corresponds to the bivariate elliptical distribution (9) with the generator function $g^{(2)}(v) = \frac{1}{2\pi} \nu^{(\nu+2)/2} (1+v)^{-(\nu+2)/2}$. Then, from (10), the distribution of $Y^*|U^* > 0$ is extended skew-t with density:

$$f_{\mathrm{EST}}(y; \mu, \sigma^2, \alpha, \tau, \nu) = \frac{1}{\sigma} t(z; \nu) \times \frac{T\left\{(\alpha z + \tau)\left(\frac{\nu+1}{\nu+z^2}\right)^{1/2}; \nu+1\right\}}{T(\tau/\sqrt{1+\alpha^2}; \nu)}, \quad y \in \mathbb{R}, \quad (12)$$

where $z = (y-\mu)/\sigma$, and $t(\cdot; \nu)$ and $T(\cdot; \nu)$ are the density and the cumulative distribution function, respectively, of a univariate Student's t distribution with ν degrees of freedom. Thus, the density $f(y|U=1)$, corresponding to the continuous component of the SLt model, is described by (12) with parameterization $\mu = \mathbf{x}^\top \boldsymbol{\beta}$, $\alpha = \rho/\sqrt{1-\rho^2}$, and $\tau = \mathbf{w}^\top \boldsymbol{\gamma}/\sqrt{1-\rho^2}$. The extended skew-$t$ distribution was studied in detail in Arellano-Valle and Genton (2010a).

For the SLt model, the conditional expectation of the observed data is

$$E(y|U^* > 0, \mathbf{x}, \mathbf{w}) = \mathbf{x}^\top \boldsymbol{\beta} + \rho\sigma \lambda_\nu(\mathbf{w}^\top \boldsymbol{\gamma}), \quad \nu > 1, \quad (13)$$

where $\lambda_\nu(v) = \frac{\nu+v^2}{\nu-1} \frac{t(v;\nu)}{T(v;\nu)}$ (Lee 1983; Arellano-Valle and Genton 2010a). We can see that similar to the SLN model, the conventional OLS regression of y on $\mathbf{x}$ will produce inconsistent results when $\rho \neq 0$. We can visualize the impact of using the SLN model to model the regression function (13) by plotting functions $\lambda(\cdot)$ from (7) and $\lambda_\nu(\cdot)$ in Figure 1.

From Figure 1, we can see that for negative values of the selection linear predictor $\mathbf{w}^\top \boldsymbol{\gamma}$, the conditional expectation will be underestimated under the SLN model for moderate values of ν degrees of freedom. The difference diminishes as the degrees of freedom increase.

Marginal effects of the predictors on y in the observed sample are often of interest in practice. Suppose that $w_k = x_k$, then the conditional marginal effect of x_k on y under the SLt model is

$$\frac{\partial E(y|U^* > 0, \mathbf{x}, \mathbf{w})}{\partial x_k} = \beta_k + \rho\sigma\gamma_k \lambda'_\nu(\mathbf{w}^\top \boldsymbol{\gamma}), \quad \nu > 1,$$

where $\lambda'_\nu(v) = \frac{\partial \lambda_\nu(v)}{\partial v} = -\lambda_\nu(v)\{v\frac{\nu+1}{\nu+v^2} + \lambda_\nu(v)\}$. Figure 2 depicts functions $\lambda'(\cdot)$ and $\lambda'_\nu(\cdot)$ for several degrees of freedom. From the graph, the conditional marginal effect of x_k on y will be overestimated by the SLN model for negative values of $\mathbf{w}^\top \boldsymbol{\gamma}$ and moderate values of ν degrees of freedom. Similar to the SLN model, the log-likelihood function of the SLt model can be decomposed into the log-likelihood of the extended skew-t distribution and the log-likelihood for the binary t model. From (12), the log-likelihood for the SLt model based on a single pair of observations (y, u) is

$$l(\boldsymbol{\beta}, \boldsymbol{\gamma}, \rho, \sigma, \nu; y, u) = u \ln f(y \mid U = 1) + u \ln T(\mathbf{w}^\top \boldsymbol{\gamma}; \nu) + (1-u) \ln T(-\mathbf{w}^\top \boldsymbol{\gamma}; \nu)$$

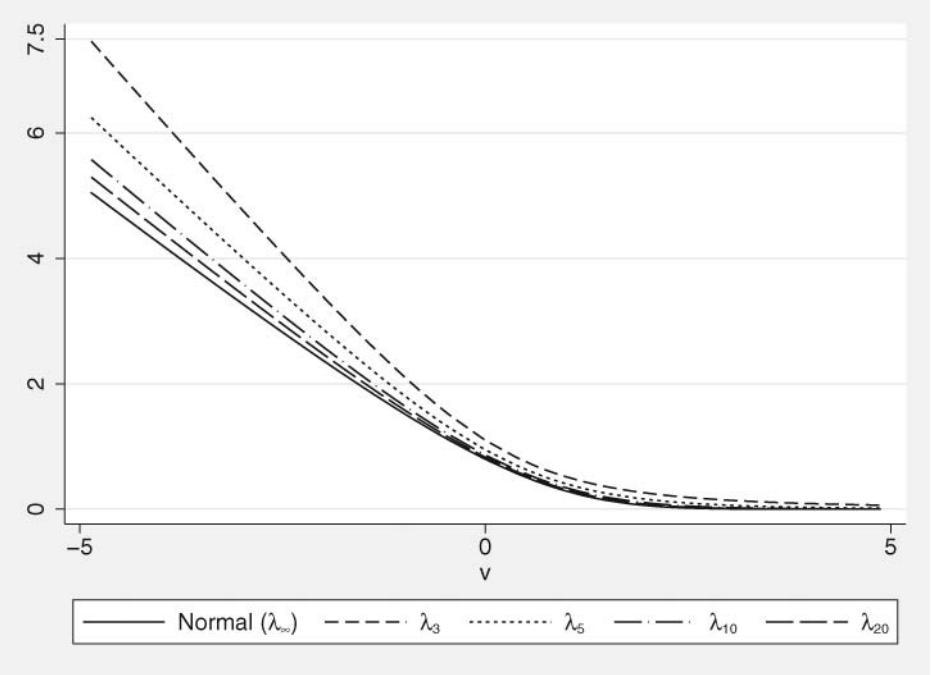

Figure 1. Plot of $\lambda_\nu(\cdot)$ for different values of ν with $\lambda_\infty(\cdot) = \lambda(\cdot)$, corresponding to the normal case.

$$= u \ln t(z;\nu) - u \ln \sigma + (1-u) \ln T(-\mathbf{w}^\top \boldsymbol{\gamma};\nu) + u \ln T\left\{\left(\frac{\nu+1}{\nu+z^2}\right)^{1/2} \frac{\rho z + \mathbf{w}^\top \boldsymbol{\gamma}}{\sqrt{1-\rho^2}}; \nu+1\right\}, \quad (14)$$

where $z = (y - \mathbf{x}^\top \boldsymbol{\beta})/\sigma$. There are no closed-form expressions for the MLEs of the parameters in (14). Thus, the MLEs are obtained numerically. The scores and the Hessian matrix corresponding to (14) are given in Appendices A and B, respectively.

3. MONTE CARLO SIMULATIONS

3.1 Finite-Sample Properties of the MLEs

To study finite-sample properties of the MLEs for the SLt model, we consider several simulation scenarios. The primary regression is $y_i^* = 0.5 + 1.5x_i + \epsilon_i$, where $x_i \overset{\text{iid}}{\sim} N(0,1)$ and $i = 1, \ldots, N = 1000$. We consider two types of selection regressions: $u_i^* = 1 + x_i + 1.5w_i + \eta_i$, with the exclusion restriction $w_i \overset{\text{iid}}{\sim} N(0,1)$, and $u_i^* = 1 + x_i + \eta_i$, without the exclusion restriction. The covariates x_i and w_i are independent and are

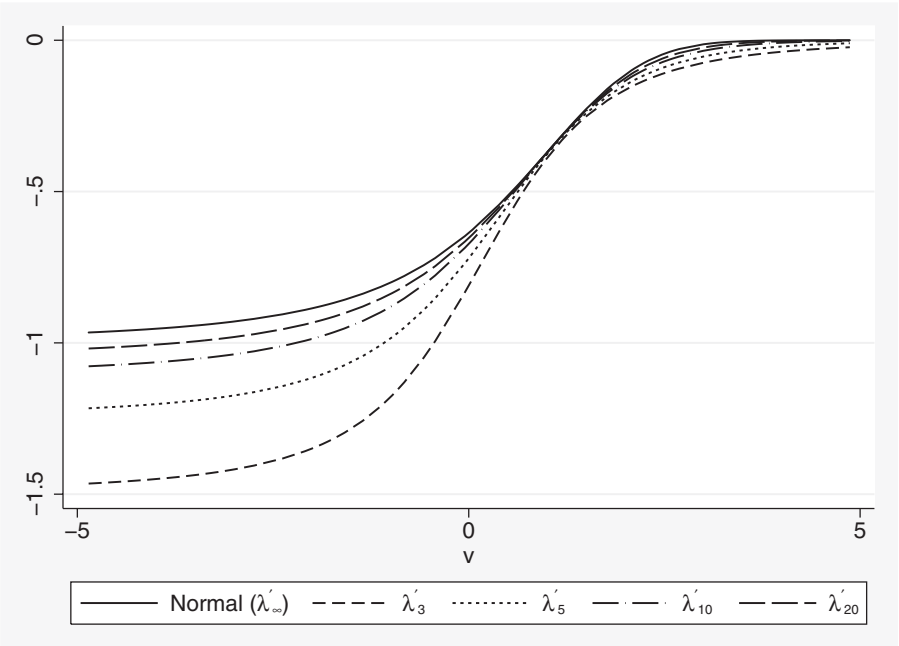

Figure 2. Plot of $\lambda'_\nu(\cdot)$ for different values of ν with $\lambda'_\infty(\cdot) = \lambda'(\cdot)$, corresponding to the normal case.

Table 1. Simulation results (in 1/10,000) in the presence of the exclusion restriction

	Bias			MSE			
	SLt	SLN	TS	SLt	SLN	TS	SE diff. SLt
$\rho = 0$							
β_1	24	14	12	**25**	32	32	1
β_0	5	25	28	**36**	49	54	6
γ_2	67	–1752	–1753	**91**	360	360	21
γ_1	104	–2619	–2623	**158**	771	773	50
γ_0	46	–1737	–1738	**81**	346	347	8
atanh ρ	35	21	14	**149**	157	183	0.1
$\ln(\sigma)$	–14	–188	–185	**21**	27	27	5
$\ln(\nu)$	306	—	—	418	—	—	51
$\rho = 0.2$							
β_1	4	–7	12	**27**	33	34	16
β_0	–39	–4	–46	**37**	49	53	2
γ_2	59	–1737	–1739	**91**	353	354	23
γ_1	33	–2651	–2655	**138**	775	776	22
γ_0	44	–1726	–1726	**79**	340	341	0.0
atanh ρ	83	–59	51	**147**	173	188	22
$\ln(\sigma)$	23	–187	–182	**20**	27	27	5
$\ln(\nu)$	394	—	—	402	—	—	18
$\rho = 0.5$							
β_1	9	41	48	**23**	32	31	3
β_0	24	23	8	**34**	48	55	23
γ_2	38	–1777	–1766	**87**	368	365	16
γ_1	79	–2684	–2646	**147**	806	785	25
γ_0	48	–1747	–1731	**73**	347	342	29
atanh ρ	49	–57	20	**176**	269	309	76
$\ln(\sigma)$	−15	–193	–188	**21**	27	26	0.0
$\ln(\nu)$	285	—	—	449	—	—	122
$\rho = 0.7$							
β_1	–10	75	40	**23**	30	31	7
β_0	–9	–90	–14	**26**	36	49	1
γ_2	32	–1831	–1763	**78**	382	358	9
γ_1	4	–2832	–2686	**141**	882	797	37
γ_0	–5	–1829	–1767	**80**	378	357	24
atanh ρ	125	214	74	**183**	300	549	43
$\ln(\sigma)$	21	–176	–188	**23**	29	29	10
$\ln(\nu)$	351	—	—	425	—	—	32
SD_{range}	453–1346	484–1720	485–2342	28–272	40–511	37–1397	—

NOTE: The results for $\ln(\sigma)$ for the SLN and TS methods are based on the true value of the variance $\nu/(\nu - 2)$. The smallest MSE is marked as bold. The standard deviations of biases for $\ln(\nu)$ ranged between 1967 and 2102, and of MSEs between 687 and 889.

also independent from the error terms ϵ_i and η_i. The error terms (ϵ_i, η_i) are generated from a bivariate Student's t distribution with $\nu = 5$ degrees of freedom and with the scale matrix

$$\mathbf{\Omega} = \begin{pmatrix} \sigma^2 & \rho\sigma \\ \rho\sigma & 1 \end{pmatrix},$$

where $\sigma = 1$. We consider several values of correlation $\rho \in \{0, 0.2, 0.5, 0.7\}$. We observe only values y_i for which $u_i^* > 0$, that is, $y_i = u_i y_i^*$, where $u_i = I(u_i^* > 0)$. In the considered scenarios, the degree of censoring corresponds to about 30% in the absence of the exclusion restriction and about 40% in the presence of the exclusion restriction.

We compare the performance of the SLt model to the SLN model and the Heckman's TS method when errors come from a bivariate Student's t distribution. Simulation results based on $R = 1000$ replications are presented in Tables 1 and 2. The results are presented in the estimation metric with the support $(-\infty, \infty)$, where atanh ρ is the inverse hyperbolic tangent of ρ, atanh $\rho = \ln\{(1 + \rho)/(1 - \rho)\}/2$. SE diff. reports the absolute difference between the standard deviation of the parameter estimates and the mean of their standard error estimates. SD_{range} reports the minimum and the maximum standard deviations of biases and mean squared errors (MSE) across parameters, except $\ln(\nu)$, and values of ρ.

Our simulations demonstrate good performance of the SLt MLEs in a finite sample under the correct specification of the model and error distribution. Biases and MSE of the parameter estimates are close to zero, and standard error estimates obtained using the inverse of the negative Hessian matrix, presented in

Table 2. Simulation results (in 1/10,000) in the absence of the exclusion restriction

	Bias			MSE			
	SLt	SLN	TS	SLt	SLN	TS	SE diff. SLt
$\rho = 0$							
β_1	13	–9	–29	**72**	153	244	23
β_0	–5	46	82	**116**	261	480	32
γ_1	21	–1573	–1565	**79**	289	286	49
γ_0	–1	–1430	–1419	**57**	234	230	11
atanh ρ	9	–89	–109	**547**	1109	2533	78
$\ln(\sigma)$	49	–146	–22	**20**	27	33	2
$\ln(\nu)$	197	—	—	398	—	—	22
$\rho = 0.2$							
β_1	–41	–56	72	**70**	168	233	33
β_0	74	134	–55	**106**	287	443	31
γ_1	88	–1521	–1498	**71**	273	264	2
γ_0	51	–1392	–1371	**56**	226	219	3
atanh ρ	–108	–213	350	**527**	1223	2343	96
$\ln(\sigma)$	45	–164	–58	**21**	35	41	4
$\ln(\nu)$	289	—	—	415	—	—	42
$\rho = 0.5$							
β_1	–75	178	131	**54**	143	170	45
β_0	74	–137	–94	**73**	221	312	58
γ_1	23	–1626	–1525	**75**	309	275	38
γ_0	58	–1417	–1342	**60**	233	212	27
atanh ρ	–103	466	900	**430**	1255	2936	149
$\ln(\sigma)$	27	–154	–123	**27**	44	54	30
$\ln(\nu)$	480	—	—	490	—	—	153
$\rho = 0.7$							
β_1	–26	342	38	**31**	94	141	25
β_0	44	–243	141	**39**	126	253	17
γ_1	8	–1693	–1485	**60**	329	260	32
γ_0	–12	–1543	–1378	**57**	271	222	13
atanh ρ	–78	800	586	**295**	984	3754	43
$\ln(\sigma)$	3	–178	–275	**27**	42	63	15
$\ln(\nu)$	520	—	—	492	—	—	118
SD_{range}	447–2340	498–3514	538–6103	30–766	60–1942	66–9526	—

NOTE: The results for $\ln(\sigma)$ for the SLN and TS methods are based on the true value of the variance $\nu/(\nu - 2)$. The smallest MSE is marked as bold. The standard deviations of biases for $\ln(\nu)$ ranged between 1987 and 2162, and of MSEs between 640 and 943.

the Appendices, adequately reflect variability in the parameter estimates.

Compared with other methods, SLt leads to smaller biases, in general, and smaller MSE of the parameter estimates. In the presence of the exclusion restriction (Table 1), the results for the primary regression coefficients are comparable across the three methods, with SLt being slightly more efficient. In the absence of the exclusion restriction (Table 2), the SLN and TS methods demonstrate some bias in the estimates of the primary regression coefficients and correlation as the correlation between errors increases, whereas the SLt estimates maintain low biases and MSE. In both cases, the estimates of the selection regression coefficients are severely biased under the SLN and TS methods.

We also repeated the above simulation scenarios (data not shown here) for $\nu = 3$, $\nu = 100$, and $\nu = \infty$, corresponding to the normally distributed errors. Our findings for $\nu = 3$ were similar to the above, with biases of the primary regression coefficients for the SLN and TS methods being even more prominent in the absence of the exclusion restriction. As the degrees of freedom increase, the Student's t distribution approaches the normal distribution, and as expected, the results from the SLt model were similar to those from the SLN model, with the latter being slightly more efficient for $\nu = \infty$.

To get some insight into the robustness of the SLt model, we also consider a simulation scenario where the errors come from a mixture of normal distributions modeling outliers. We consider a mixture, $(1 - p)N_2(\mathbf{0}, \mathbf{\Sigma}) + pN_2(\mathbf{0}, k\mathbf{\Sigma})$, of the two bivariate normal distributions with zero means and covariance matrices that differ by a scale factor $k > 0$. We present simulation results for the 90% mixture ($p = 0.1$), with $k = 12$ and

$$\mathbf{\Sigma} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

, where $\rho \in \{0, 0.2, 0.5, 0.7\}$. Let σ_ϵ be the variance of the error term from the primary equation, then $\sigma_\epsilon = \sigma\sqrt{1 - p + pk} = 1.45$ in this simulation scenario. Tables 3 and 4 contain simulation results for the scenario with and without the exclusion restriction, respectively, as described earlier. The estimate of

Table 3. Simulation results (in 1/10,000) for the normal mixture in the presence of the exclusion restriction

	Bias			MSE		
	SLt	SLN	TS	SLt	SLN	TS
$\rho = 0$						
β_1	7	12	15	**23**	37	39
β_0	–16	–21	–28	**37**	59	74
γ_2	1087	–2995	–3004	**298**	980	985
γ_1	693	–2010	–2015	**155**	457	459
γ_0	694	–1966	–1970	**140**	431	432
atanh ρ	64	64	78	**157**	187	257
$\ln(\sigma_\epsilon)$	–40	–358	–352	66	**48**	**48**
$\rho = 0.2$						
β_1	22	13	64	**25**	41	42
β_0	–23	90	–24	**35**	62	72
γ_2	1223	–2943	–2953	**333**	953	959
γ_1	812	–1965	–1972	**171**	437	440
γ_0	834	–1909	–1910	**164**	410	410
atanh ρ	101	–236	24	**155**	213	268
$\ln(\sigma_\epsilon)$	42	–324	–311	66	**45**	**45**
$\rho = 0.5$						
β_1	–35	29	23	**22**	39	38
β_0	13	10	22	**29**	57	67
γ_2	1140	–3083	–2994	**295**	1045	984
γ_1	757	–2031	–1996	**159**	469	455
γ_0	797	–1960	–1925	**167**	434	420
atanh ρ	115	–104	–91	**173**	375	434
$\ln(\sigma_\epsilon)$	20	–371	–371	69	51	**50**
$\rho = 0.7$						
β_1	2	191	120	**24**	44	42
β_0	10	–178	–34	**26**	47	68
γ_2	1071	–3340	–3039	**269**	1206	1005
γ_1	712	–2175	–2030	**148**	530	470
γ_0	733	–2097	–1955	**148**	484	428
atanh ρ	144	288	56	**186**	419	807
$\ln(\sigma_\epsilon)$	29	–357	–378	70	**51**	52
SD_{range}	467–1357	590–2027	593–2842	31–476	58–634	60–2834

NOTE: The smallest MSE is marked as bold. Three values of atanh ρ exceeding 3, for which ρ is close to its boundary value 1, were reset to 3 when computing MSE for the TS method and $\rho = 0.7$.

$\ln(\sigma_\epsilon)$ for the SLt method is computed using the estimated scale and degrees-of-freedom parameters.

As we anticipated, the SLt method outperforms the SLN and TS methods. It has the smallest MSE and smaller bias, in general, of the three considered methods. The bias of the coefficients from the selection equation is about 8% and is considerably higher than the bias of the coefficients from the primary equation. This may be explained by the greater sensitivity of the underlying binary model to model misspecification. Although the SLN method has a slightly smaller MSE for $\ln(\sigma_\epsilon)$ than the SLt method, it also has a large bias, so the SLt method is still better. Overall, the simulation results indicate some robustness of the SLt model to the outliers generated by a normal mixture.

3.2 Test of Selection Bias When the Error Distribution Is Bivariate *t*

In this section, we investigate the performance of three tests, commonly used for testing the presence of sample selection bias in the OLS regression when the error distribution is bivariate t. The tests under consideration are the SLN Wald test of H_o: atanh$\rho = 0$ (equivalent to H_o: $\rho = 0$) (SLN), the likelihood ratio test of independent equations under the SLN model (LRT), and the Wald test of H_o: $\beta_\lambda = \rho\sigma = 0$ under the TS estimation. We also compare the performance of these tests to the SLt Wald test of H_o: atanh$\rho = 0$ (SLt) obtained under the correct SLt model specification.

The data are simulated as described in Section 3.1. We consider the scenario in the presence of an exclusion restriction, with varying sample sizes ($N = 500, 1000$), varying degrees of freedom ($\nu = 3, 5, 100$), and varying values of ρ. We also compare performance of the tests in the case when the error distribution is a mixture of normal distributions, as described in the previous section. The results presented in Tables 5 and 6 are based on $R = 5000$ replications. We considered two nominal levels 0.01 and 0.05 and obtained similar findings. We present results for the nominal level 0.01.

In the case of uncorrelated errors ($\rho = 0$), only the SLt Wald test maintains correct nominal levels for small degrees of

Table 4. Simulation results (in 1/10,000) for the normal mixture in the absence of the exclusion restriction

	Bias			MSE		
	SLt	SLN	TS	SLt	SLN	TS
$\rho = 0$						
β_1	15	14	37	**66**	143	297
β_0	–40	–37	–70	**96**	242	609
γ_1	634	–1595	–1583	**123**	296	291
γ_0	615	–1407	–1391	**108**	230	225
atanh ρ	145	139	269	**485**	901	2729
$\ln(\sigma_\epsilon)$	–64	–422	–249	64	**55**	57
$\rho = 0.2$						
β_1	–67	82	194	**65**	182	282
β_0	49	–25	–198	**95**	298	557
γ_1	679	–1608	–1566	**131**	302	287
γ_0	635	–1432	–1392	**109**	238	226
atanh ρ	–92	24	638	**494**	1180	3186
$\ln(\sigma_\epsilon)$	–28	–378	–246	66	**55**	70
$\rho = 0.5$						
β_1	–88	446	485	**48**	192	306
β_0	86	–378	–489	**60**	259	588
γ_1	704	–1767	–1577	**136**	359	291
γ_0	689	–1547	–1381	**113**	272	221
atanh ρ	–21	897	2591	**378**	1390	7508
$\ln(\sigma_\epsilon)$	–42	–313	–95	74	**66**	147
$\rho = 0.7$						
β_1	–62	745	631	**34**	152	282
β_0	52	–627	–572	**36**	153	519
γ_1	703	–1954	–1559	**125**	427	285
γ_0	679	–1751	–1382	**108**	341	221
atanh ρ	–0.0	1686	3715	**276**	1174	10, 051
$\ln(\sigma_\epsilon)$	–31	–249	32	68	**54**	205
SD_{range}	580–2199	566–3620	544–9316	46–696	68–1768	98–16,743

NOTE: The smallest MSE is marked as bold. Values of atanh ρ exceeding 3 in absolute value, for which ρ is close to its boundary values ± 1, were reset to ± 3 when computing bias and MSE for the TS method.

freedom; see Table 5. The other tests are known to be sensitive to the normality assumption and, thus, they often attribute nonnormality of errors to the presence of selection bias ($\rho \neq 0$), which leads to an inflated Type I error. As the degrees of freedom increase to 100, the significance levels of all tests become similar and close to the nominal level. As such, we compare powers of all tests only for $\nu = 100$ (Table 6) and report powers of the SLt test for other degrees of freedom (Table 7). The powers of all tests are similar for $\nu = 100$ and increase with sample size N and correlation ρ. The SLN test has slightly higher powers than the SLt test, which is expected with greater degrees of freedom. The TS test has lowest powers among the considered tests, as is expected with a less efficient approach. As can be observed from Table 7, the powers of the SLt test tend to be slightly higher for $\nu = 3$ than for $\nu = 5$.

Table 8 contains powers of the SLt test of $\rho = 0$ in the presence of the exclusion restriction when the error distribution is a 90% mixture of normal distributions. We expect the power of the SLt test in this scenario to be similar to the power of the SLt test with Student's t errors with small degrees of freedom. The

Table 5. Empirical significance levels (as %) of the tests of selection bias for the nominal significance level 0.01

	$N = 500$				$N = 1000$			
	SLt	SLN	TS	LRT	SLt	SLN	TS	LRT
$\nu = 3$	1.1	5.8	4.7	4.5	1	4.8	4.4	3.9
$\nu = 5$	1.2	2.5	2.8	2	0.9	2	2.3	1.8
$\nu = 100$	1.2	1.3	1.3	1.2	1	1	1.1	1
Mixture	1.1	3.8	4.3	3.2	0.9	3.0	4.5	2.7

NOTE: Standard errors ranged between 0.13 (SLt) and 0.33 (SLN).

Table 6. Powers (as %) of the tests of selection bias for $\nu = 100$ for the nominal significance level 0.01

	$N = 500$				$N = 1000$			
ρ	SLt	SLN	TS	LRT	SLt	SLN	TS	LRT
0.1	2.4	2.5	2.4	2.4	4.6	4.7	4.8	4.7
0.2	8.7	8.9	8.5	8.8	19.3	19.8	19.6	19.5
0.3	22.9	23.4	22.5	23.2	51.2	52.0	51.1	51.8
0.4	45.3	46.3	43.5	46.0	81.2	81.5	80.0	81.4
0.5	72.2	72.9	68.4	72.7	95.2	95.2	94.4	95.2
0.6	89.4	89.4	86.0	89.4	97.0	97.0	96.8	97.0
0.7	95.1	95.2	93.9	95.1	97.5	97.5	97.5	97.5

NOTE: Standard errors ranged between 0.21 and 0.7.

Table 7. Powers (as %) of the SLt test of $\rho = 0$ for $\nu = 3$ and $\nu = 5$ for the nominal significance level 0.01

	$N = 500$		$N = 1000$	
ρ	$\nu = 3$	$\nu = 5$	$\nu = 3$	$\nu = 5$
0.1	2.5	2.6	4.9	4.2
0.2	8.7	7.8	19.6	18.1
0.3	22.3	21.8	50.6	48.4
0.4	47.0	45.1	83.1	80.8
0.5	73.4	73.0	97.5	96.7
0.6	91.4	91.4	99.9	99.8
0.7	97.3	96.8	99.8	99.9

NOTE: Standard errors ranged between 0.04 and 0.7.

Table 8. Powers (as %) of the SLt test of $\rho = 0$ for the normal mixture for the nominal significance level 0.01

ρ	$N = 500$	$N = 1000$
0.1	2.8	4.5
0.2	8.8	19.0
0.3	21.9	49.2
0.4	46.3	80.9
0.5	72.3	96.9
0.6	91.5	99.8
0.7	98.6	100

NOTE: Standard errors ranged between 0.08 and 0.7.

powers from Table 8 are indeed similar to those from Table 7 with small degrees of freedom.

4. NUMERICAL APPLICATION TO AMBULATORY EXPENDITURES

We consider the data on ambulatory expenditures from the 2001 Medical Expenditure Panel Survey analyzed by Cameron and Trivedi (2010). The data consist of 3328 observations, of which 526 (15.8%) correspond to zero values of expenditures. The dataset includes several explanatory variables, such as age, gender, education status, and others. The distribution of expenditures is highly skewed, so the analysis is performed using the log scale. Because the decision to spend is likely to be related to the spending amount, it is natural to consider a sample selection model for the analysis. Cameron and Trivedi (2010, p. 561), among others, used the classical Heckman sample selection model to analyze these data; see Appendix D for the results from Stata's `heckman` command (StataCorp 2011). The primary regression includes such factors as age, gender, ethnicity, education status, insurance status, and the number of chronic diseases. The selection equation additionally includes income, imposing the exclusion restriction on the model, although the use of income for this purpose is debatable. All of the considered factors are strong predictors of the decision to spend. All factors other than the insurance status are also strong predictors of the spending amount. The reported Wald test ($p = 0.380$) of no sample selection, H_o: $\rho = 0$ or, more precisely, H_o: $\operatorname{atanh}\rho = 0$, does not provide sufficient evidence to reject this hypothesis, implying that spending amount is unrelated to the decision to spend and can be analyzed separately using standard OLS regression. This conclusion seems implausible.

As noted by Cameron and Trivedi, the assumption of underlying normality is suspect for these data. Thus, we use the proposed SLt model to analyze these data; see the results from Stata's user-written command `heckt` in Appendix D. (This command is available from the authors upon request.)

We obtain results similar to those from the SLN model regarding the coefficients in the primary and selection equations. There is a significant change, however, in the inference about the existence of sample selection bias. A Z test of H_o: $\operatorname{atanh}\rho = 0$ provides sufficient evidence ($p = 0.009$) to reject the null hypothesis of no sample selection bias at a 1% significance level, in agreement with our intuition. The estimate of ρ reported by the SLN model is -0.131, whereas the SLt model reports an estimate of -0.322, which is also more in agreement with the TS estimate of ρ, -0.359. Our simulations showed that the estimate of ρ from the SLN model can be biased when data come from a nonnormal distribution, which is likely what we observe in this example. The estimated degrees of freedom are 13 with a 95% confidence interval of (8–20), indicating some deviation from normality.

5. CONCLUSION

We introduced the sample SLt model that extends the conventional sample selection model of Heckman (1974) to have a bivariate Student's t error distribution. This model provides a greater flexibility for modeling heavier-tailed data than the SLN model by introducing only one extra parameter, the degrees of freedom, controlling the tails of the distribution. We considered maximum likelihood estimation of the parameters. Monte Carlo simulations demonstrated good performance of the MLEs in finite samples. Monte Carlo simulations also showed that the SLt model performs better than the SLN model for heavier-tailed data and is also more robust to collinearity between the primary and the selection regressors for moderate values of degrees of freedom. The robustness to the collinearity is appealing because there is no need to impose exclusion restrictions, which are often difficult to formulate in practice. Our simulations demonstrated high sensitivity (inflated Type I error) of the selection bias test based on the SLN model to the presence of heavy tails, whereas the selection bias test based on the SLt model maintained nominal coverage. We also provided some insight into the power of this test.

Although the considered parametric SLt model is not as flexible in modeling various shapes of the distribution compared with semiparametric and nonparametric methods, it is useful to model heavy-tailed data, which occur rather often in practice. Its advantages include a gain of efficiency within a class of heavy-tailed Student's t distributions; an ability to identify an intercept, which, as was noted by Heckman, is an important parameter of interest in many economic applications; and also the relative simplicity and speed efficiency of the implementation.

In this article, we used a bivariate Student's t distribution to allow for heavier tails in the error distribution. It is straightforward to extend the presented results to some other parametric distributions. For example, one can use another version of a bivariate t distribution where a separate degrees-of-freedom parameter is considered for each dimension. From a practical

standpoint, it would be even more appealing to consider some flexible parametric distributions accommodating the two most common deviations from normality—the heavier tails and the asymmetry of the distribution. A bivariate skew-normal distribution (Azzalini and Dalla Valle 1996; Azzalini and Capitanio 1999; Azzalini 2005) and a bivariate skew-t distribution (Azzalini and Capitanio 2003) are two appealing candidates. Keeping in mind the link described in Section 2.2, we can justify the use of the skewed distributions to model errors as a natural way of modeling some hidden truncation already present in the population from which the data were sampled (see Little and Rubin 2002, p. 324, for an example of such a population).

APPENDIX A: SCORE EQUATIONS

Let $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\gamma}^\top, \rho, \sigma, \nu)$. The log-likelihood for the SLt model based on a single pair of observations (y, u) is

$$l(\boldsymbol{\theta}; y, u) = u \ln t(y; \mathbf{x}^\top\boldsymbol{\beta}, \sigma^2, \nu) + u \ln T(\eta_\theta; \nu + 1) + (1-u) \ln T(-\mathbf{w}^\top\boldsymbol{\gamma}; \nu), \tag{A.1}$$

where $u = 1$ if y is observed, and $u = 0$ if y is unobserved, $z = (y - \mathbf{x}^\top\boldsymbol{\beta})/\sigma$,

$$\ln t(y; \mathbf{x}^\top\boldsymbol{\beta}, \sigma^2, \nu) = c(\nu, \sigma) - \frac{\nu+1}{2} \ln\left(1 + \frac{z^2}{\nu}\right),$$

$$\eta_\theta = \sqrt{\frac{\nu+1}{\nu+z^2}} \frac{\rho z + \mathbf{w}^\top\boldsymbol{\gamma}}{\sqrt{1-\rho^2}},$$

and

$$c(\nu, \sigma) = \ln\Gamma\left(\frac{\nu+1}{2}\right) - \ln\Gamma\left(\frac{\nu}{2}\right) - 0.5\ln\pi - 0.5\ln\nu - \ln\sigma.$$

Let

$$Q_\nu = \sqrt{\frac{\nu+1}{\nu + \left(\frac{y-\mathbf{x}^\top\boldsymbol{\beta}}{\sigma}\right)^2}} = \sqrt{\frac{\nu+1}{\nu+z^2}},$$

$$\zeta_{\theta_{-\nu}} = \frac{\frac{\rho(y-\mathbf{x}^\top\boldsymbol{\beta})}{\sigma} + \mathbf{w}^\top\boldsymbol{\gamma}}{\sqrt{1-\rho^2}} = A_{\rho\rho} z + A_\rho \mathbf{w}^\top\boldsymbol{\gamma},$$

$$\eta_\theta = \sqrt{\frac{\nu+1}{\nu + \left(\frac{y-\mathbf{x}^\top\boldsymbol{\beta}}{\sigma}\right)^2}} \frac{\frac{\rho(y-\mathbf{x}^\top\boldsymbol{\beta})}{\sigma} + \mathbf{w}^\top\boldsymbol{\gamma}}{\sqrt{1-\rho^2}} = Q_\nu \zeta_{\theta_{-\nu}} = Q_\nu\left(A_{\rho\rho} z + A_\rho \mathbf{w}^\top\boldsymbol{\gamma}\right),$$

$$M_\nu(x) = \frac{t(x;\nu)}{T(x;\nu)} = \frac{\partial \ln T(x;\nu)}{\partial x},$$

where $A_\rho = 1/\sqrt{1-\rho^2}$ and $A_{\rho\rho} = \rho/\sqrt{1-\rho^2} = \rho A_\rho$.

Let $S_\alpha = \frac{\partial l(\boldsymbol{\theta})}{\partial \alpha}$ be the score function corresponding to the parameter α. For $\alpha \in \{\boldsymbol{\beta}^\top, \boldsymbol{\gamma}^\top, \rho, \sigma\}$,

$$\begin{aligned} S_\alpha &= u \frac{\partial \ln t(y; \mathbf{x}^\top\boldsymbol{\beta}, \sigma^2, \nu)}{\partial \alpha} + u \frac{\partial \ln T(\eta_\theta; \nu+1)}{\partial \eta_\theta} \frac{\partial \eta_\theta}{\partial \alpha} \\ &\quad - (1-u) \frac{\partial \ln T(-\mathbf{w}^\top\boldsymbol{\gamma}; \nu)}{\partial(-\mathbf{w}^\top\boldsymbol{\gamma})} \frac{\partial \mathbf{w}^\top\boldsymbol{\gamma}}{\partial \alpha} \\ &= -u I(\alpha = \sigma) \frac{1}{\sigma} - u Q_\nu^2 z \frac{\partial z}{\partial \alpha} + u M_{\nu+1}(\eta_\theta) \frac{\partial \eta_\theta}{\partial \alpha} \\ &\quad - (1-u) I(\alpha = \gamma_k) w_k M_\nu(-\mathbf{w}^\top\boldsymbol{\gamma}). \end{aligned}$$

In particular, the scores are

$$\begin{aligned} S_{\beta_k} &= \frac{u x_k Q_\nu}{\sigma}\left[Q_\nu z + \left\{\zeta_{\theta_{-\nu}}(\nu+z^2)^{-1} z - A_{\rho\rho}\right\} M_{\nu+1}(\eta_\theta)\right], \\ &\qquad k = 1, \ldots, p, \\ S_{\gamma_k} &= u w_k A_\rho Q_\nu M_{\nu+1}(\eta_\theta) - (1-u) w_k M_\nu(-\mathbf{w}^\top\boldsymbol{\gamma}), \\ &\qquad k = 1, \ldots, q, \\ S_\rho &= u Q_\nu A_\rho^3 \left(z + \rho \mathbf{w}^\top\boldsymbol{\gamma}\right) M_{\nu+1}(\eta_\theta), \\ S_\sigma &= \frac{u}{\sigma}\left[-1 + Q_\nu^2 z^2 + Q_\nu z\left\{\zeta_{\theta_{-\nu}}(\nu+z^2)^{-1} z - A_{\rho\rho}\right\} M_{\nu+1}(\eta_\theta)\right], \\ S_\nu &= u\left\{\frac{\partial c(\nu,\sigma)}{\partial \nu} - \frac{1}{2}\ln\left(1 + \frac{z^2}{\nu}\right) + \frac{Q_\nu^2 z^2}{2\nu} + \frac{\partial \ln T(\eta_\theta; \nu+1)}{\partial \nu}\right\} \\ &\quad + (1-u_i) \frac{\partial \ln T(-\mathbf{w}^\top\boldsymbol{\gamma}; \nu)}{\partial \nu}, \end{aligned}$$

where $\frac{\partial c(\nu,\sigma)}{\partial \nu} = \frac{1}{2}\psi\left(\frac{\nu+1}{2}\right) - \frac{1}{2}\psi\left(\frac{\nu}{2}\right) - \frac{1}{2\nu}$, $\psi(\cdot)$ is the derivative of the log-gamma function, and $\frac{\partial \ln T(x;\nu)}{\partial \nu}$ must be computed numerically.

APPENDIX B: HESSIAN MATRIX

Let $\mathbf{s}_i(\hat{\boldsymbol{\theta}})$ be the score vector of the SLt model from Appendix A for observation i for $i = 1, \ldots, N$, evaluated at the MLE $\hat{\boldsymbol{\theta}}$. Then, under appropriate regularity conditions, the Hessian matrix H can be approximated using $H = \frac{1}{N}\sum_{i=1}^N \mathbf{s}_i(\hat{\boldsymbol{\theta}})\mathbf{s}_i(\hat{\boldsymbol{\theta}})^\top$.

We also provide direct computation of the Hessian matrix below. Let $S_{\alpha_1\alpha_2} = \frac{\partial^2 l(\boldsymbol{\theta})}{\partial\alpha_1 \partial\alpha_2}$ be the second-order partial derivative of $l(\boldsymbol{\theta})$ from (15) with respect to α_1 and α_2. The lower diagonal entries of the Hessian matrix are

$$\begin{aligned} S_{\beta_k\beta_l} &= \frac{u x_k x_l Q_\nu^2}{\sigma^2}\left[\left\{2z^2(\nu+z^2)^{-1} - 1\right\}\right. \\ &\quad \left. + \left\{\zeta_{\theta_{-\nu}} \frac{z}{\nu+z^2} - A_{\rho\rho}\right\}^2 M'_{\nu+1}(\eta_\theta)\right] + \frac{u x_k x_l}{\sigma^2} Q_\nu (\nu+z^2)^{-1} \\ &\quad \times \left[\zeta_{\theta_{-\nu}}\left\{3(\nu+z^2)^{-1} z^2 - 1\right\} - 2A_{\rho\rho} z\right] M_{\nu+1}(\eta_\theta), \\ S_{\beta_k\gamma_n} &= u \frac{x_k w_n Q_\nu A_\rho}{\sigma}\left[(\nu+z^2)^{-1} z\left\{M_{\nu+1}(\eta_\theta) + \zeta_{\theta_{-\nu}} M'_{\nu+1}(\eta_\theta)\right\}\right. \\ &\quad \left. - A_{\rho\rho} M'_{\nu+1}(\eta_\theta)\right], \\ S_{\beta_k\rho} &= u \frac{x_k Q_\nu A_\rho^3}{\sigma}\left[\frac{z(z+\rho\mathbf{w}^\top\boldsymbol{\gamma})}{\nu+z^2}\left\{M_{\nu+1}(\eta_\theta) + \eta_\theta M'_{\nu+1}(\eta_\theta)\right\}\right. \\ &\quad \left. - Q_\nu(z+\rho\mathbf{w}^\top\boldsymbol{\gamma}) A_{\rho\rho} M'_{\nu+1}(\eta_\theta) - M_{\nu+1}(\eta_\theta)\right], \\ S_{\beta_k\sigma} &= u \frac{2x_k Q_\nu^2 z}{\sigma^2}\left(\frac{z^2}{\nu+z^2} - 1\right) \\ &\quad + u \frac{x_k z M'_{\nu+1}(\eta_\theta)}{\sigma^2}\left(\eta_\theta \frac{z}{\nu+z^2} - Q_\nu A_{\rho\rho}\right)^2 \\ &\quad + u \frac{x_k M_{\nu+1}(\eta_\theta)}{\sigma^2}\left\{\eta_\theta \frac{z}{\nu+z^2}\left(3\frac{z^2}{\nu+z^2} - 2\right)\right. \\ &\quad \left. - 2Q_\nu A_{\rho\rho} \frac{z^2}{\nu+z^2} + Q_\nu A_{\rho\rho}\right\}, \\ S_{\beta_k\nu} &= \frac{u x_k Q_\nu^2 z}{\sigma}\left(\frac{1}{\nu+1} - \frac{1}{\nu+z^2}\right) \\ &\quad + \frac{u x_k Q_\nu}{\sigma}\left(\zeta_{\theta_{-\nu}} \frac{z}{\nu+z^2} - A_{\rho\rho}\right) \\ &\quad \times \left\{\frac{\partial M_{\nu+1}(\eta_\theta)}{\partial \nu} + \frac{1}{2} M_{\nu+1}(\eta_\theta)\left(\frac{1}{\nu+1} - \frac{1}{\nu+z^2}\right)\right\} \\ &\quad - \frac{u x_k}{\sigma} M_{\nu+1}(\eta_\theta)\eta_\theta \frac{z}{(\nu+z^2)^2}, \end{aligned}$$

$$
\begin{aligned}
S_{\gamma_m\gamma_n} &= uw_m w_n Q_\nu^2 A_\rho^2 M'_{\nu+1}(\eta_\theta) + (1-u)w_m w_n M'_{\nu+1}(-\mathbf{w}^\top\boldsymbol{\gamma}),\\
S_{\gamma_m\rho} &= uw_m Q_\nu A_\rho^3\{\rho M_{\nu+1}(\eta_\theta) + Q_\nu A_\rho(z+\rho\mathbf{w}^\top\boldsymbol{\gamma})M'_{\nu+1}(\eta_\theta)\},\\
S_{\gamma_m\sigma} &= \frac{uw_m A_\rho Q_\nu z}{\sigma}\Bigg[\frac{z}{\nu+z^2}\{\eta_\theta M'_{\nu+1}(\eta_\theta) + M_{\nu+1}(\eta_\theta)\}\\
&\quad - Q_\nu A_{\rho\rho} M'_{\nu+1}(\eta_\theta)\Bigg],\\
S_{\gamma_m\nu} &= uw_m A_\rho Q_\nu\left\{\frac{\partial M_{\nu+1}(\eta_\theta)}{\partial\nu} + \frac{1}{2}\left(\frac{1}{\nu+1} - \frac{1}{\nu+z^2}\right)\right\}\\
&\quad - (1-u)w_m\frac{\partial M_{\nu+1}(-\mathbf{w}^\top\boldsymbol{\gamma})}{\partial\nu},\\
S_{\rho^2} &= uQ_\nu^2 A_\rho^6(z+\rho\mathbf{w}^\top\boldsymbol{\gamma})^2 M'_{\nu+1}(\eta_\theta)\\
&\quad + uQ_\nu A_\rho^5\{3\rho z + (1+2\rho^2)\mathbf{w}^\top\boldsymbol{\gamma}\}M_{\nu+1}(\eta_\theta),\\
S_{\rho\sigma} &= \frac{uzQ_\nu A_\rho^3}{\sigma}\Bigg((z+\rho\mathbf{w}^\top\boldsymbol{\gamma})\Bigg[\frac{z}{\nu+z^2}\Big\{\eta_\theta M'_{\nu+1}(\eta_\theta)\\
&\quad + M_{\nu+1}(\eta_\theta)\Big\} - Q_\nu A_{\rho\rho} M'_{\nu+1}(\eta_\theta)\Bigg] - M_{\nu+1}\Bigg),\\
S_{\rho\nu} &= uQ_\nu A_\rho^3(z+\rho\mathbf{w}^\top\boldsymbol{\gamma})\Bigg\{\frac{\partial M_{\nu+1}(\eta_\theta)}{\partial\nu}\\
&\quad + \frac{1}{2}\left(\frac{1}{\nu+1} - \frac{1}{\nu+z^2}\right)M_{\nu+1}(\eta_\theta)\Bigg\},\\
S_{\sigma^2} &= \frac{u}{\sigma^2} - \frac{3uQ_\nu^2 z^2}{\sigma^2} + \frac{2uQ_\nu}{\sigma^2}\frac{z^2}{\nu+z^2}\Bigg[Q_\nu z^2\\
&\quad + \left\{\frac{3}{2}\zeta_{\theta_{-\nu}}\left(\frac{z^2}{\nu+z^2} - 1\right) - Q_\nu A_{\rho\rho}\right\}M_{\nu+1}(\eta_\theta)\Bigg]\\
&\quad + \frac{2uQ_\nu A_{\rho\rho} z}{\sigma^2}M_{\nu+1}(\eta_\theta)\\
&\quad + \frac{uz^2}{\sigma^2}\left(\eta_\theta\frac{z}{\nu+z^2} - Q_\nu A_{\rho\rho}\right)^2 M'_{\nu+1}(\eta_\theta),\\
S_{\sigma\nu} &= \frac{uQ_\nu z^2}{\sigma}\left\{Q_\nu\left(\frac{1}{\nu+1} - \frac{1}{\nu+z^2}\right) - \frac{M_{\nu+1}(\eta_\theta)}{(\nu+z^2)^2}\right\}\\
&\quad + \frac{uQ_\nu z}{\sigma}\Bigg\{\frac{\partial M_{\nu+1}(\eta_\theta)}{\partial\nu}\\
&\quad + \frac{1}{2}\left(\frac{1}{\nu+1} - \frac{1}{\nu+z^2}\right)M_{\nu+1}(\eta_\theta)\Bigg\},
\end{aligned}
$$

$$
\begin{aligned}
S_{\nu^2} &= u\frac{\partial^2 c(\nu,\sigma)}{\partial^2\nu} + \frac{u}{2\nu}\frac{z^2}{\nu+z^2}\\
&\quad + \frac{uQ_\nu^2 z^2}{2\nu}\left(\frac{1}{\nu+1} - \frac{1}{\nu+z^2}\right) - \frac{uz^4}{2\nu^2}\\
&\quad + u\frac{\partial^2\ln T(\eta_\theta;\nu+1)}{\partial^2\nu} + (1-u)\frac{\partial^2\ln T(-\mathbf{w}^\top\boldsymbol{\gamma};\nu)}{\partial^2\nu},
\end{aligned}
$$

where $k,l = 1,\ldots,p$ and $m,n = 1,\ldots,q$, $M'_\nu(x) = \frac{\partial M_\nu(x)}{\partial x} = -M_\nu(x)\left\{x\frac{\nu+1}{\nu+x^2} + M_\nu(x)\right\}$, and $\frac{\partial M_{\nu+1}(x)}{\partial\nu}$ must be computed numerically. Moreover,

$$
\frac{\partial^2 c(\nu,\sigma)}{\partial^2\nu} = \frac{1}{4}\psi'\left(\frac{\nu+1}{2}\right) - \frac{1}{4}\psi'\left(\frac{\nu}{2}\right) + \frac{1}{2}\frac{1}{\nu^2},
$$

where $\psi'(\cdot)$ is the tri-gamma function.

APPENDIX C: FISHER AND OBSERVED INFORMATION MATRICES AT $\rho = 0$ FOR THE SELECTION-NORMAL MODEL

When $\nu \longrightarrow \infty$, $Q_\nu = 1$, $\eta_\theta = \zeta_{\theta_{-\nu}} = A_{\rho\rho}z + A_\rho\mathbf{w}^\top\boldsymbol{\gamma}$, $M_\nu(x) = M(x) = \frac{\phi(x)}{\Phi(x)} = \frac{\partial\ln\Phi(x)}{\partial x}$, $M'_\nu(x) = M'(x) = -M(x)\times\{x + M(x)\}$, $\frac{\partial c(\nu,\sigma)}{\partial\nu} = 0$, and $\frac{\partial T(x;\nu)}{\partial\nu} = 0$.

The scores are

$$
\begin{aligned}
S_{\beta_k} &= \frac{ux_k}{\sigma}\Big(z - A_{\rho\rho}M(\eta_\theta)\Big), \quad k = 1,\ldots,p\\
S_{\gamma_k} &= uw_k A_\rho M(\eta_\theta) - (1-u)w_k M(-\mathbf{w}^\top\boldsymbol{\gamma}), \quad k = 1,\ldots,q\\
S_\rho &= uA_\rho^3\Big(z + \rho\mathbf{w}^\top\boldsymbol{\gamma}\Big)M(\eta_\theta)\\
S_\sigma &= \frac{u}{\sigma}\Big\{-1 + z^2 - zA_{\rho\rho}M(\eta_\theta)\Big\}\\
S_\nu &= 0.
\end{aligned}
$$

At $\rho = 0$, $A_\rho = 1$, $A_{\rho\rho} = 0$, $\eta_\theta = \mathbf{w}^\top\boldsymbol{\gamma}$ and the scores are

$$
\begin{aligned}
S_{\beta_k} &= \frac{ux_k}{\sigma}z, \quad k = 1,\ldots,p,\\
S_{\gamma_k} &= uw_k M(\mathbf{w}^\top\boldsymbol{\gamma}) - (1-u)w_k M(-\mathbf{w}^\top\boldsymbol{\gamma}), \quad k = 1,\ldots,q,\\
S_\rho &= uzM(\mathbf{w}^\top\boldsymbol{\gamma}),\\
S_\sigma &= \frac{u}{\sigma}\Big\{-1 + z^2\Big\},\\
S_\nu &= 0.
\end{aligned}
$$

Without loss of generality, let $p = q = 2$, $x_{1i} = w_{1i} = 1$, $x_{2i} \neq 1$, and $w_{2i} \neq 1$ (we exclude the degenerate case of $p = q = 1$ and $x_{1i} = w_{1i} = 1$). Then the observed information for $(\beta_1,\beta_2,\gamma_1,\gamma_2,\rho,\sigma)$ for the sample of size N at $\rho = 0$ is

$$
\begin{pmatrix}
\sum_i\frac{u_i}{\sigma^2} & \sum_i\frac{u_i x_{2i}}{\sigma^2} & 0 & 0 & \sum_i\frac{u_i M_i}{\sigma} & \sum_i\frac{2u_i z_i}{\sigma^2}\\
\sum_i\frac{u_i x_{2i}}{\sigma^2} & \sum_i\frac{u_i x_{2i}^2}{\sigma^2} & 0 & 0 & \sum_i\frac{u_i x_{2i} M_i}{\sigma} & \sum_i\frac{2u_i z_i x_{2i}}{\sigma^2}\\
0 & 0 & \sum_i B_i & \sum_i w_{2i}B_i & -\sum_i u_i z_i M'_i & 0\\
0 & 0 & \sum_i w_{2i}B_i & \sum_i w_{2i}^2 B_i & -\sum_i u_i z_i w_{2i} M'_i & 0\\
\sum_i\frac{u_i M_i}{\sigma} & \sum_i\frac{u_i x_{2i} M_i}{\sigma} & -\sum_i u_i z_i M'_i & -\sum_i u_i z_i w_{2i} M'_i & -\sum_i u_i\left(z_i^2 M'_i + \mathbf{w}_i^\top\boldsymbol{\gamma} M_i\right) & \sum_i\frac{u_i z_i M_i}{\sigma}\\
\sum_i\frac{2u_i z_i}{\sigma^2} & \sum_i\frac{2u_i z_i x_{2i}}{\sigma^2} & 0 & 0 & \sum_i\frac{u_i z_i M_i}{\sigma} & -\sum_i\frac{u_i}{\sigma^2} + \sum_i\frac{3u_i z_i^2}{\sigma^2}
\end{pmatrix},
$$

where $M_i = M(\mathbf{w}_i^\top \boldsymbol{\gamma})$, $M_i' = M'(\mathbf{w}_i^\top \boldsymbol{\gamma})$, $B_i = B(u_i; \mathbf{w}_i^\top \boldsymbol{\gamma}) = -u_i M'(\mathbf{w}_i^\top \boldsymbol{\gamma}) - (1-u_i)M'(-\mathbf{w}_i^\top \boldsymbol{\gamma})$, and $\mathbf{w}_i^\top \boldsymbol{\gamma} = \gamma_1 + \gamma_2 w_{2i}$.

To compute the Fisher information, consider the following expectations:

$$
\begin{aligned}
E(U_i) &= \Phi(\mathbf{w}_i^\top \boldsymbol{\gamma}),\\
E(U_i Z_i) &= E\{U_i E(Z_i|U_i)\}\\
&= E\left[U_i\left\{E\left(\frac{Y_i^* - \mathbf{x}_i^\top \boldsymbol{\beta}}{\sigma}|U_i = 1\right) + 0\right\}\right]\\
&= \rho\frac{\phi(\mathbf{w}_i^\top \boldsymbol{\gamma})}{\Phi(\mathbf{w}_i^\top \boldsymbol{\gamma})}E(U_i) = \rho\phi(\mathbf{w}_i^\top \boldsymbol{\gamma})|_{\rho=0} = 0,\\
E(U_i Z_i^2) &= E\{U_i E(Z_i^2|U_i)\} = \left(1 - \rho^2 \mathbf{w}_i^\top \boldsymbol{\gamma}\frac{\phi(\mathbf{w}_i^\top \boldsymbol{\gamma})}{\Phi(\mathbf{w}_i^\top \boldsymbol{\gamma})}\right)E(U_i)\\
&= \Phi(\mathbf{w}_i^\top \boldsymbol{\gamma}) - \rho^2 \mathbf{w}_i^\top \boldsymbol{\gamma}\phi(\mathbf{w}_i^\top \boldsymbol{\gamma})|_{\rho=0} = \Phi(\mathbf{w}_i^\top \boldsymbol{\gamma}).
\end{aligned}
$$

Note that

$$
\begin{aligned}
E\{B(U_i; \mathbf{w}_i^\top \boldsymbol{\gamma})\} &= M'(\mathbf{w}_i^\top \boldsymbol{\gamma})E(U_i) + (1 - E(U_i))M'(-\mathbf{w}_i^\top \boldsymbol{\gamma})\\
&= \Phi(\mathbf{w}_i^\top \boldsymbol{\gamma})M'(\mathbf{w}_i^\top \boldsymbol{\gamma}) + \Phi(-\mathbf{w}_i^\top \boldsymbol{\gamma})M'(-\mathbf{w}_i^\top \boldsymbol{\gamma})\\
&= \frac{-\phi^2(\mathbf{w}_i^\top \boldsymbol{\gamma})}{\Phi(\mathbf{w}_i^\top \boldsymbol{\gamma})\Phi(-\mathbf{w}_i^\top \boldsymbol{\gamma})},
\end{aligned}
$$

$$
\begin{aligned}
&E\left\{-U_i Z_i^2 M'(\mathbf{w}_i^\top \boldsymbol{\gamma}) - U_i \mathbf{w}_i^\top \boldsymbol{\gamma} M(\mathbf{w}_i^\top \boldsymbol{\gamma})\right\}|_{\rho=0}\\
&\quad= -\Phi(\mathbf{w}_i^\top \boldsymbol{\gamma})M'(\mathbf{w}_i^\top \boldsymbol{\gamma}) - \mathbf{w}_i^\top \boldsymbol{\gamma}\Phi(\mathbf{w}_i^\top \boldsymbol{\gamma})M(\mathbf{w}_i^\top \boldsymbol{\gamma})\\
&\quad= \mathbf{w}_i^\top \boldsymbol{\gamma}\phi(\mathbf{w}_i^\top \boldsymbol{\gamma}) + \frac{\phi^2(\mathbf{w}_i^\top \boldsymbol{\gamma})}{\Phi(\mathbf{w}_i^\top \boldsymbol{\gamma})} - \mathbf{w}_i^\top \boldsymbol{\gamma}\phi(\mathbf{w}_i^\top \boldsymbol{\gamma}) = \frac{\phi^2(\mathbf{w}_i^\top \boldsymbol{\gamma})}{\Phi(\mathbf{w}_i^\top \boldsymbol{\gamma})}.
\end{aligned}
$$

Then, the Fisher information matrix is

$$
\begin{pmatrix}
\sum_i \frac{\Phi(\mathbf{w}_i^\top \boldsymbol{\gamma})}{\sigma^2} & \sum_i \frac{x_{2i}\Phi(\mathbf{w}_i^\top \boldsymbol{\gamma})}{\sigma^2} & 0 & 0 & \sum_i \frac{\phi(\mathbf{w}_i^\top \boldsymbol{\gamma})}{\sigma} & 0\\
\sum_i \frac{x_{2i}\Phi(\mathbf{w}_i^\top \boldsymbol{\gamma})}{\sigma^2} & \sum_i \frac{x_{2i}^2\Phi(\mathbf{w}_i^\top \boldsymbol{\gamma})}{\sigma^2} & 0 & 0 & \sum_i \frac{x_{2i}\phi(\mathbf{w}_i^\top \boldsymbol{\gamma})}{\sigma} & 0\\
0 & 0 & \sum_i \frac{\phi^2(\mathbf{w}_i^\top \boldsymbol{\gamma})}{\Phi(\mathbf{w}_i^\top \boldsymbol{\gamma})\Phi(-\mathbf{w}_i^\top \boldsymbol{\gamma})} & \sum_i \frac{w_{2i}\phi^2(\mathbf{w}_i^\top \boldsymbol{\gamma})}{\Phi(\mathbf{w}_i^\top \boldsymbol{\gamma})\Phi(-\mathbf{w}_i^\top \boldsymbol{\gamma})} & 0 & 0\\
0 & 0 & \sum_i \frac{w_{2i}\phi^2(\mathbf{w}_i^\top \boldsymbol{\gamma})}{\Phi(\mathbf{w}_i^\top \boldsymbol{\gamma})\Phi(-\mathbf{w}_i^\top \boldsymbol{\gamma})} & \sum_i \frac{w_{2i}^2\phi^2(\mathbf{w}_i^\top \boldsymbol{\gamma})}{\Phi(\mathbf{w}_i^\top \boldsymbol{\gamma})\Phi(-\mathbf{w}_i^\top \boldsymbol{\gamma})} & 0 & 0\\
\sum_i \frac{\phi(\mathbf{w}_i^\top \boldsymbol{\gamma})}{\sigma} & \sum_i \frac{x_{2i}\phi(\mathbf{w}_i^\top \boldsymbol{\gamma})}{\sigma} & 0 & 0 & \sum_i \frac{\phi^2(\mathbf{w}_i^\top \boldsymbol{\gamma})}{\Phi(\mathbf{w}_i^\top \boldsymbol{\gamma})} & 0\\
0 & 0 & 0 & 0 & 0 & \sum_i \frac{2\Phi(\mathbf{w}_i^\top \boldsymbol{\gamma})}{\sigma^2}
\end{pmatrix}.
$$

Unlike the case of the extended skew-normal model (Arellano-Valle and Genton 2010a, p. 17), the scores above, corresponding to the SLN model, are not linearly dependent (provided that at least one x_k or w_k is not equal to 1). Also, the observed information and Fisher information for $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\gamma}, \rho, \sigma)$ are nonsingular at $\rho = 0$, which is not the case for the extended skew-normal model.

APPENDIX D: STATA OUTPUT FOR THE NUMERICAL RESULTS

Heckman Selection-Normal Model

```
. heckman lny age female edu blhisp totchr ins, sel(dy=age female educ blhisp totchr ins income)
Heckman selection model                         Number of obs      =      3328
(regression model with sample selection)        Censored obs       =       526
                                                Uncensored obs     =      2802
                                                Wald chi2(6)       =    288.88
Log likelihood = -5836.219                      Prob > chi2        =    0.0000
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
lny						
age	0.2120	0.0230	9.21	0.000	0.1669	0.2571
female	0.3481	0.0601	5.79	0.000	0.2303	0.4660
educ	0.0187	0.0105	1.77	0.076	-0.0020	0.0394
blhisp	-0.2186	0.0597	-3.66	0.000	-0.3355	-0.1016
totchr	0.5399	0.0393	13.73	0.000	0.4628	0.6170
ins	-0.0300	0.0511	-0.59	0.557	-0.1301	0.0701
_cons	5.0441	0.2281	22.11	0.000	4.5969	5.4912
dy						
age	0.0879	0.0274	3.21	0.001	0.0342	0.1417
female	0.6627	0.0609	10.87	0.000	0.5432	0.7821
educ	0.0619	0.0120	5.15	0.000	0.0384	0.0855
blhisp	-0.3639	0.0619	-5.88	0.000	-0.4852	-0.2427
totchr	0.7970	0.0711	11.20	0.000	0.6575	0.9364
ins	0.1701	0.0629	2.71	0.007	0.0469	0.2934
income	0.0027	0.0013	2.06	0.040	0.0001	0.0053
_cons	-0.6761	0.1940	-3.48	0.000	-1.0563	-0.2958
/athrho	-0.1313	0.1496	-0.88	0.380	-0.4246	0.1619
/lnsigma	0.2398	0.0145	16.59	0.000	0.2115	0.2682
rho	-0.1306	0.1471			-0.4008	0.1605
sigma	1.2710	0.0184			1.2355	1.3076
lambda	-0.1660	0.1879			-0.5342	0.2022

```
LR test of indep. eqns. (rho = 0):   chi2(1) =     0.91   Prob > chi2 = 0.3406
```

Heckman Selection-*t* Model

```
. heckt lny age female edu blhisp totchr ins, sel(dy=age female educ blhisp totchr ins income)
Heckman-t selection model                       Number of obs      =      3328
(regression model with sample selection)        Censored obs       =       526
                                                Uncensored obs     =      2802
                                                Wald chi2(6)       =    325.58
Log likelihood = -5822.076                      Prob > chi2        =    0.0000
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
lny						
age	0.2068	0.0226	9.16	0.000	0.1626	0.2511
female	0.3065	0.0562	5.45	0.000	0.1963	0.4168
educ	0.0173	0.0102	1.69	0.091	-0.0028	0.0374
blhisp	-0.1930	0.0577	-3.35	0.001	-0.3060	-0.0799
totchr	0.5127	0.0357	14.36	0.000	0.4427	0.5827
ins	-0.0525	0.0505	-1.04	0.298	-0.1514	0.0464
_cons	5.2058	0.2088	24.93	0.000	4.7966	5.6151
dy						
age	0.0986	0.0297	3.31	0.001	0.0403	0.1569
female	0.7249	0.0685	10.58	0.000	0.5905	0.8592
educ	0.0648	0.0128	5.06	0.000	0.0397	0.0899
blhisp	-0.3936	0.0665	-5.92	0.000	-0.5240	-0.2632
totchr	0.8901	0.0872	10.21	0.000	0.7192	1.0610
ins	0.1800	0.0680	2.65	0.008	0.0467	0.3133
income	0.0030	0.0014	2.06	0.040	0.0001	0.0058
_cons	-0.7480	0.2077	-3.60	0.000	-1.1551	-0.3409
/athrho	-0.3338	0.1278	-2.61	0.009	-0.5844	-0.0833
/lnsigma	0.1780	0.0215	8.29	0.000	0.1359	0.2201
/lndf	2.5594	0.2205	11.61	0.000	2.1272	2.9916
rho	-0.3220	0.1146			-0.5258	-0.0831
sigma	1.1948	0.0257			1.1456	1.2462
df	12.9279	2.8508			8.3912	19.9173
lambda	-0.3847	0.1402			-0.6594	-0.1100

[Received June 2010. Revised July 2011.]

REFERENCES

Ahn, H., and Powell, J. L. (1993), "Semiparametric Estimation of Censored Selection Models With a Nonparametric Selection Mechanism," *Journal of Econometrics*, 58, 3–29. [304]

Amemiya, T. (1985), *Advanced Econometrics*, Cambridge MA: Harvard University Press. [305]

Arellano-Valle, R. B., and Azzalini, A. (2006), "On the Unification of Families of Skew-Normal Distributions," *Scandinavian Journal of Statistics*, 33, 561–574. [306]

Arellano-Valle, R. B., Branco, M. D., and Genton, M. G. (2006), "A Unified View on Skewed Distributions Arising From Selections," *The Canadian Journal of Statistics*, 34, 581–601. [306]

Arellano-Valle, R. B. Castro, L. M., González-Farías, G., and Muñoz-Gajardo, K. A. (2011), "Student-*t* Censored Regression Model: Properties and Inference," Manuscript. [304]

Arellano-Valle, R. B., and Genton, M. G. (2010a), "Multivariate Extended Skew-*t* Distributions and Related Families," *Metron*, 68, 201–234. [304,306,307,316]

——— (2010b), "Multivariate Unified Skew-Elliptical Distributions," *Chilean Journal of Statistics*, 1, 17–33. [306]

Azzalini, A. (1985), "A Class of Distributions Which Includes the Normal Ones," *Scandinavian Journal of Statistics*, 12, 171–178. [306]

——— (2005), "The Skew-Normal Distribution and Related Multivariate Families" (with discussion by Marc G. Genton and a rejoinder by the author), *Scandinavian Journal of Statistics*, 32, 159–200. [306,313]

——— (2006), R Package sn: *The Skew-Normal and Skew*-t *Distributions*, (version 0.4-6), available at *http://azzalini.stat.unipd.it/SN*. [306]

Azzalini, A., and Capitanio, A. (1999), "Statistical Applications of the Multivariate Skew-Normal Distribution," *Journal of the Royal Statistical Society*, Series B, 61, 579–602. [313]

——— (2003), "Distributions Generated by Perturbation of Symmetry With Emphasis on a Multivariate Skew *t* Distribution," *Journal of the Royal Statistical Society*, Series B, 65, 367–389. [313]

Azzalini, A., and Dalla Valle, A. (1996), "The Multivariate Skew-Normal Distribution," *Biometrika*, 83, 715–726. [313]

Azzalini, A., and Genton, M. G. (2008), "Robust Likelihood Methods Based on the Skew-*t* and Related Distributions," *International Statistical Review*, 76, 106–129. [304]

Branco, M. D., and Dey, D. K. (2001), "A General Class of Multivariate Skew-Elliptical Distributions," *Journal of Multivariate Analysis*, 79, 99–113. [307]

Cameron, A. C., and Trivedi, P. K. (2010), *Microeconometrics Using Stata* (Revised ed.), College Station, TX: Stata Press. [313]

Chib, S., and Hamilton, B. H. (2000), "Bayesian Analysis of Cross-Section and Clustered Data Treatment Models," *Journal of Econometrics*, 97, 25–50. [304]

Copas, J. B., and Li, H. G. (1997), "Inference for Non-Random Samples," *Journal of the Royal Statistical Society*, Series B, 59, 55–95. [306]

Das, M., Newey, W. K., and Vella, F. (2003), "Nonparametric Estimation of Sample Selection Models," *Review of Economics Studies*, 70, 33–58. [304]

DiCiccio, T., and Monti, A. (2009), "Inferential Aspects of the Skew-*t* Distribution," *Manuscript in preparation*. [304]

Fang, K.-T., Kotz, S., and Ng, K. W. (1990), *Symmetric Multivariate and Related Distributions*, New York: Chapman & Hall. [306]

Genton, M. G. (2004), *Skew-Elliptical Distributions and Their Applications: A Journey Beyond Normality* (Edited volume), Boca Raton, FL: Chapman & Hall/CRC. [306]

Greene, W. H. (2008), *Econometric Analysis* (6th ed.), Upper Saddle River, NJ: Prentice–Hall. [304,305]

Heckman, J. J. (1974), "Shadow Prices, Market Wages, and Labor Supply," *Econometrica*, 42, 679–694. [304,305,313]

——— (1979), "Sample Selection Bias as a Specification Error," *Econometrica*, 47, 153–161. [304,305]

Heckman, J. J., Tobias, J. L., and Vytlacil, E. J. (2003), "Simple Estimators for Treatment Parameters in a Latent Variable Framework," *Review of Economics and Statistics*, 85, 748–754. [304]

Lange, K. L., Little, R. J. A., and Taylor, J. M. G. (1989), "Robust Statistical Modeling Using the *t* Distribution," *Journal of the American Statistical Association*, 84, 881–896. [304]

Lee, L. F. (1982), "Some Approaches to the Correction of Selectivity Bias," *Review of Economic Studies*, 49, 355–372. [304,307]

——— (1983), "Generalized Econometric Models With Selectivity," *Econometrica*, 51, 507–512. [307]

Li, Q., and Racine, J. S. (2007), *Nonparametric Econometrics*, Princeton, NJ: Princeton University Press. [304]

Little, R. J. A., and Rubin, D. B. (2002), *Statistical Analysis With Missing Data* (2nd ed.), Hoboken, NJ: Wiley. [313]

Marchenko, Y. V., and Genton, M. G. (2010), "A Suite of Commands for Fitting the Skew-Normal and Skew-*t* Models," *Stata Journal*, 10, 507–539. [306]

Matzkin, R. L. (1994), "Nonparametric Identification," in *Handbook of Econometrics*, Volume 4, eds. R. Engle and D. McFadden, Amsterdam: Elsevier, pp. 2443–2521. [304]

Newey, W. K. (1999), "Two-Step Series Estimation of Sample Selection Models," Working Paper no. 99-04, Cambridge, MA: Massachusetts Institute of Technology. [304]

Powell, J. L. (1994), "Estimation of Semiparametric Models," in *Handbook of Econometrics*, eds. J. J. Heckman and E. Leamer, Amsterdam: Elsevier, pp. 5307–5368. [304]

Puhani, P. A. (2000), "The Heckman Correction for Sample Selection and Its Critique," *Journal of Economic Surveys*, 14, 53–68. [304,306]

Rubin, D. B. (1976), "Inference and Missing Data," *Biometrika*, 63, 581–592. [304]

Scruggs, J. T. (2007), "Estimating the Cross-Sectional Market Response to an Endogenous Event: Naked vs. Underwritten Calls of Convertible Bonds," *Journal of Empirical Finance*, 14, 220–247. [304]

StataCorp. (2011), Stata Statistical Software: Release 12, College Station, TX: StataCorp LP. [313]

Vella, F. (1998), "Estimating Models With Sample Selection Bias: A Survey," *Journal of Human Resources*, 33, 127–172. [304]