Scandinavian Journal of Statistics Theory and Applications

A Sample selection model with Skew-normal distribution

Journal:	*Scandinavian Journal of Statistics*
Manuscript ID:	SJS-13-200.R4
Manuscript Type:	Paper
Date Submitted by the Author:	n/a
Complete List of Authors:	Ogundimu, Emmanuel; University of Oxford, Centre for Statistics in Medicine Hutton, Jane; University of Warwick, Department of Statistics

A Sample selection model with Skew-normal distribution

Emmanuel O. Ogundimu* and Jane L. Hutton†

Abstract

Non-random sampling is a source of bias in empirical research. It is common for the outcomes of interest (e.g. wage distribution) to be skew in the source population. Sometimes the outcomes are further subjected to sample selection, which is a type of missing data, resulting in partial observability. Thus, methods based on complete cases for skew data are inadequate for the analysis of such data and a general sample selection model is required. Heckman proposed a full maximum likelihood estimation method under the normality assumption for sample selection problems, and parametric and non-parametric extensions have been proposed. We generalize Heckman (1976, 1979) to allow for underlying skew-normal distributions. Finite sample performance of the maximum likelihood estimator of the model is studied via simulation. Applications illustrate the strength of the model in capturing spurious skewness in bounded scores, and in modelling data where logarithm transformation could not mitigate the effect of inherent skewness in the outcome variable.

Key words: non-random sample, generalized skew-normal distribution, missing data, sample selection.

1 Introduction

Sample selection arises when the outcome of interest can only be observed in a subset of the population under study. The realised data are missing not at random (MNAR) because the observed data do not represent a random sample from the population, even after controlling for covariates. The use of linear regression models with normal errors on non-random samples may yield estimates that are biased and inconsistent, and misleading conclusions. Suppose we are interested in the effect of factors, which we assume are always observed, on the wages of all women who might work in some population. Observations on wages can only be provided by women who are working, but we wish to generalise the effects to all women in the population. The difference in the distribution of characteristics among workers and non-workers determines whether the problem of selection bias arises. If the decision to work is related to a factor, perhaps age, and also affects wages, a linear regression of the observed wages on the covariates is likely to give biased and inconsistent estimates.

*Correspondence to: Dr. E.O. Ogundimu, CSM, The University of Oxford, OXFORD, OX3 7LD, UK. emmanuel.ogundimu@csm.ox.ac.uk

†Statistics, University of Warwick, Coventry, CV4 7AL, UK

In order to overcome such problems, Heckman (1976) introduced a model to describe selection of women into the labor force; he proposed full information maximum likelihood estimation (FIML) under an assumption of bivariate normality. An alternative method called the two-step estimator (TS) was later developed, where sample selection was treated as a model misspecification problem due to an omitted variable (Heckman, 1979). The TS method, which is not as efficient as the FIML method, was meant for exploratory work, and to provide good starting values for FIML estimation. It can be readily derived as the conditional expectation of the observed data and, like the MLE, it is sensitive to collinearity among covariates (Puhani, 2000; Leung & Yu, 2000). To avoid this problem, the so-called exclusion restriction, which requires at least one effective prediction variable in the selection equation to be excluded from the outcome equation, is imposed. Some merits of the TS method, which contribute to its popularity, are given in Leung & Yu (2000).

Sample selection models can reduce bias when the model is correctly specified (Leung & Yu, 1996). Model misspecification includes the omission of important variables and interaction effects in the missingness process (Enders, 2010, p296), multicollinearity (Puhani, 2000) and the use of incorrect distributional assumptions for the residuals (Marchenko & Genton, 2012). Various generalizations have been proposed to mitigate the effect of distributional misspecification. These include, in the parametric framework, the use of student-t distributions (Marchenko & Genton, 2012) and a Bayesian extension of this model (Ding, 2014). In addition, copula-based transformation approaches have been discussed (Lee, 1983). Semi-parametric approaches to sample selection can be found in Ahn & Powell (1993). For non-parametric sample selection models, we refer the reader to Das et al. (2003). Vella (1998) provides an earlier review of sample selection models.

The ability of a parametric sample selection model to identify the outcome model intercept is an important advantage over semi-parametric and non-parametric alternatives, especially when prediction is the goal of the analysis. Availability of the model intercept, in our earlier example on determinants of wage distribution for women, allows the prediction of the wage of a woman who is not in paid work. Parametric and semi-parametric estimation of the intercept parameter have been proposed for both the semi-parametric and non-parametric sample selection models: see Andrews & Schafgans (1998) and references therein. Although the estimator may be consistent, its asymptotic distribution is unknown, which may adversely affect inference. Other proposals that involve semi-parametric and non-parametric first stage estimation (Newey, 2009; Ahn & Powell, 1993) in the TS method allow an intercept to be estimated, but they are not superior to parametric models in terms of efficiency in high dimensional data. In addition, parameters in parametric sample selection models have clear interpretations and roles: the theoretical mean in the unselected population can be estimated for covariate values of interest.

Problems in real data that can lead to deviation from normality include outliers, multimodality, mixtures, heavier tails than the normal distribution and skewness. In dealing with heavy tails in sample selection, Marchenko & Genton (2012) derived a model using links between hidden truncation and sample selection with an underlying bivariate-t error distribution. They demonstrate, in a simulation study, that their model is robust against outliers arising from normal mixtures. Other robust estimators have been proposed in the presence of outliers from various sources, including outliers in the design space (see, for example, Zhelonkin et. al., 2012). Transformations have been used to mitigate the effects of skewness, but parametric sample selection models for skew outcomes which retain the original scale are not available in the literature. Hence, we propose a model which is more efficient than approaches based on semi-parametric and non-parametric methods within a class of deviations from normality due to skewness.

The distribution of outcomes such as women's wages are likely to be skew in the source population, before selection. This consideration prompted Marchenko & Genton (2012) to log-transform the outcome variable (expenditure) used in their example to reduce the burden of skewness before applying their model. In general, the appropriate transformation to use is not always clear, and transformations are not always appropriate in modeling data resulting from selectively reported samples because interest is in inference for the unselected population. An additional disadvantage is not working on the original scale familiar to the applied researchers. On the other hand, distributional assumptions in a sample selection framework are generally untestable for the same reasons as the uncertainty in choice of transformations. The usual recommendation is to use a range of plausible parametric representations, especially those having the normal distribution as a special case. This is the path taken in Marchenko & Genton (2012). We follow a similar path and propose the use of the Azzalini skew-normal distribution (Azzalini, 1985) for modeling asymmetry in a sample selection framework.

The article is organized as follows. In section 2, we describe the conventional sample selection model and relate it to the general hidden truncation model formulation of Arnold & Beaver (2002). In section 3, a new model is derived using the general formulation of skew distributions arising from selection and linked with a hidden truncation formulation of the model. Two types of skewness in sample selection models are distinguished in this section. Section 4 reports a study of the finite sample properties of the proposed model. The model is applied to two data sets in section 5 and conclusions given in section 6.

2 Classical sample selection model

In this section, we review the classical selection normal model (SNM) and establish a link with the hidden truncation formulation given by Arnold & Beaver (2002).

2.1 Selection normal model (SNM)

Let $Y_i^\star$ be the outcome variable of interest, assumed to be linearly related to covariates x_i through a multiple regression model

$$Y_i^\star = \beta' x_i + \sigma\varepsilon_{1i}, \quad i = 1, \ldots, N;$$

which is supplemented by a selection (missingness) equation

$$S_i^\star = \gamma' w_i + \varepsilon_{2i}, \quad i = 1, \ldots, N,$$

where β, γ and σ are unknown parameters; x_i and w_i, which can overlap, are fixed observed characteristics not subject to missingness; and $(\varepsilon_{1i}, \varepsilon_{2i})$ are random errors with means zero, variances one and correlation ρ. If we observe $S_i = I(S_i^\star > 0)$ and $Y_i = Y_i^\star S_i$ for $n = \sum_{i=1}^N S_i$ of N individuals, the sample $Y_i, i = 1, \ldots, n$ is a selection from the N individuals. The variance of $S_i^\star$ is fixed at 1, because we only observe the sign of $S^\star$, which is insufficient information to estimate its variance. Thus an observation has, using Bayes' rule, the conditional density

$$f(y|x, S^\star > 0) = \frac{f(y, S^\star > 0|x, w)}{P(S^\star > 0|w)} = \frac{f(y|x)P(S^\star > 0|y, w)}{P(S^\star > 0|w)}. \qquad (1)$$

Equation (1) is the basis of the unification of selection problems as skew distributions given by Arellano-Valle et al. (2006). The quantity $f(y|x)$ is a proper probability density function (PDF), with a skewing function $P(S^\star > 0|y, w)$, and a normalizing function $P(S^\star > 0|w)$. Under the additional assumption of bivariate normal errors,

$$\begin{pmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \end{pmatrix} \sim N_2 \left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right\},$$

it is straightforward to show that the PDF is

$$f(y|x, S = 1; \Theta) = \frac{1}{\sigma}\phi(\frac{y - \beta' x}{\sigma})\Phi\left(\frac{\gamma' w + \rho\left(\frac{y-\beta' x}{\sigma}\right)}{\sqrt{1-\rho^2}}\right)\Big/\Phi(\gamma' w), \quad (2)$$

where $\Theta = (\beta, \sigma, \gamma, \rho)$. The parameter $\rho \in$ (-1,1) determines the correlation between $Y_i^\star$ and $S_i^\star$, and hence the nature and severity of the selection process. Model (2) includes three missing data mechanisms (see Copas & Li, 1997 for details). The complete density of classical sample selection model has a continuous component, with conditional density given by (2), and a discrete component given by $P(S = 1|w)$. The marginal distribution of the selection equation determines the model for the discrete process. In Copas & Li (1997) and Heckman (1976), a probit model $P(S = s) = \{\Phi(\gamma' w)\}^s\{1 - \Phi(\gamma' w)\}^{1-s}$ was used. The log-likelihood function is therefore

$$l(\Theta) = \sum_{i=1}^{n} S_i\Big(\ln f(y_i|x_i, S_i = 1; \Theta)\Big) + \sum_{i=1}^{n} S_i(\ln \Phi(\gamma' w_i)) + \sum_{i=1}^{n}(1 - S_i) \ln \Phi(-\gamma' w_i).$$

The TS estimator is derived from the conditional expectation of the observed data, which is given by

$$E(Y|x, S^\star > 0) = \beta' x + \sigma\rho\Lambda(\gamma' w), \quad (3)$$

where $\Lambda(\cdot) = \phi(\cdot)/\Phi(\cdot)$ is the inverse Mills ratio. To use (3) in practice, a standard probit model for S provides an estimate of $\hat{\gamma}$. The quantity $\Lambda(\hat{\gamma}' w)$ is then taken as an additional covariate in equation (3), and the least squares coefficient of $\Lambda(\hat{\gamma}' w)$ gives an estimate of $\sigma\rho$. When $\rho = 0$ in (3), the selection component of the model $(\sigma\rho\Lambda(\gamma' w))$ vanishes and the model for Y simplifies to linear regression with normal errors. As expected from Arellano-Valle et al. (2006), equation 2 defines the density of a distribution which belongs to the extended skew-normal (ESN) distribution family. To see this, we let $\mu = \beta' x$, $\lambda_0 = \gamma' w/\sqrt{1-\rho^2} \in \mathbb{R}$ and $\lambda = \rho/\sqrt{1-\rho^2} \in \mathbb{R}$ in (2); we then have the PDF written in the four-parameter ESN form given by

$$f_{ESN}(y; \mu, \sigma^2, \lambda_0, \lambda) = \frac{1}{\sigma}\phi(\frac{y-\mu}{\sigma})\Phi\left(\lambda_0 + \lambda(\frac{y-\mu}{\sigma})\right)\Big/\Phi\left(\frac{\lambda_0}{\sqrt{1+\lambda^2}}\right), \quad (4)$$

where λ_0 & λ are shift and shape parameters respectively (Azzalini, 1985).

The ESN distribution (2) & (4) is not identifiable when ρ, and hence λ are 0, but the model becomes identifiable in the sample selection framework due to the additional information from the selection process which is introduced through a probit model. The price to pay for the identifiability is the possibility of model misspecification. Although sensitivity analysis on the model parameters is justifiable, the use of a range of plausible parametric representations, particularly those having the normal distribution as a special case, is preferred.

Further, the attribution of non-normality of errors to the presence of selection bias is common in tests for selection in classical sample selection models. The model we present in the next section can produce correct tests for selection in the presence of non-normality due to skewness, just as the selection-t model of Marchenko & Genton (2012) produced adequate tests for selection when non-normality is due to heavy tails. It has been noted that with the assumption of normal residuals in the outcome equation, information about the selection process comes from the skewness in the observed residuals. However, if the population distribution of the residuals is skewed, but due to selection the observed residuals are symmetric, then the classical sample selection model will lead to the inference that there is no selection bias (Fitzmaurice comment, Copas & Li, 1997, p83). We introduce in the next section a sample selection model using a bivariate skew-normal distribution. The properties of the closed skew-normal (CSN) distribution (González-Farías et. al., 2004, p23) are used to develop the model.

3 Selection Skew-normal model (SSNM)

In this section, we assume that the errors follow a bivariate skew-normal distribution. We show that the resulting model has a link with models arising from hidden truncation. This link is used to derive a Heckman-type two-step estimation method under the skew-normal distribution. We also distinguish between population residual skewness and skewness due to selection.

3.1 Derivation of the SSNM model

The skew-normal distribution arising from selection defined in section 2.1 was derived using the conditional distribution properties of a bivariate normal distribution. We now relax the assumption to a bivariate skew-normal distribution for the errors:

$$\begin{pmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \end{pmatrix} \sim SN_2 \left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}, \begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix} \right\},$$

where λ_1 and λ_2 are the skewness parameters for $Y_i^\star$ and $S_i^\star$ respectively. The joint distribution of the outcomes and the selection process can be written as a closed skew-normal (CSN) distribution (González-Farías et. al., 2004, p23):

$$\begin{pmatrix} Y^\star \\ S^\star \end{pmatrix} \sim CSN_{2,1} \left\{ \boldsymbol{\mu} = (\beta' x, \gamma' w), \Sigma = \begin{pmatrix} \sigma^2 & \rho\sigma \\ \rho\sigma & 1 \end{pmatrix}, D = (\lambda_1/\sigma, \lambda_2), \nu = 0, \Delta = 1 \right\}.$$

Using the conditional distribution and marginalization properties of CSN, the relevant selection equation is

$$f(y|x, S=1; \Xi) = \frac{f(y|x) F_{ESN} \left\{ \gamma' w + \rho \left(\frac{y - \beta' x}{\sigma} \right); 0, 1-\rho^2, \frac{-\lambda_2}{\sqrt{1-\rho^2}}, -(\lambda_1 + \lambda_2) \left(\frac{y-\beta' x}{\sigma} \right) \right\}}{F_{SN} \left(\gamma' w; \frac{-(\lambda_2 + \lambda_1 \rho)}{\sqrt{1 + \lambda_1^2 - \lambda_1^2 \rho^2}} \right)}, \tag{5}$$

where F_{SN} and F_{ESN} are the cummulative distribution functions (CDFs) of the SN and the ESN distributions respectively, and $\Xi = (\beta, \sigma, \gamma, \rho, \lambda_1, \lambda_2)$ is the parameter vector of a distribution with density given by (5).

If λ_1 and λ_2 are set equal to zero in (5), the Copas & Li (1997) model, given by (2), is recovered. The model in (5) is identifiable in the sense that for any pair of model parameters $\Xi_1 \neq \Xi_2$, $f(y|\Xi_1) \not\equiv f(y|\Xi_2)$. Further, the observed information matrix is non-singular with probability 1 (see Appendix for the elements of observed information matrix of the reduced model).

We now restrict attention to a special case of the model given in (5). If λ_2 is set equal to zero, i.e. the selection random variable is normal, we get a simpler model:

$$f(y|x, S=1; \Omega) = \frac{\frac{2}{\sigma}\phi\left(\frac{y-\beta' x}{\sigma}\right)\Phi\left(\frac{\lambda_1(y-\beta' x)}{\sigma}\right)\Phi\left(\frac{\gamma' w+\rho\left(\frac{y-\beta' x}{\sigma}\right)}{\sqrt{1-\rho^2}}\right)}{F_{SN}\left(\gamma' w; \frac{-\lambda_1\rho}{\sqrt{1+\lambda_1^2-\lambda_1^2\rho^2}}\right)}, \tag{6}$$

where $\Omega = (\beta, \sigma, \gamma, \rho, \lambda_1)$. This situation arises when the underlying mechanism governing selection is not skewed. Equation (6) describes another class of skew-normal distribution. A substitution similar to that made in section 2.1 is used: we put $\mu = \beta' x$, $\lambda_0 = \gamma' w/\sqrt{1-\rho^2} \in \mathbb{R}$ and $\lambda = \rho/\sqrt{1-\rho^2} \in \mathbb{R}$ in (6), to obtain

$$f(y; \mu, \sigma^2, \lambda_0, \lambda_1, \lambda) = \frac{\frac{2}{\sigma}\phi\left(\frac{y-\mu}{\sigma}\right)\Phi\left(\frac{\lambda_1(y-\mu)}{\sigma}\right)\Phi\left(\lambda_0+\lambda\left(\frac{y-\mu}{\sigma}\right)\right)}{F_{SN}\left(\frac{\lambda_0}{\sqrt{1+\lambda^2}}; \frac{-\lambda_1\lambda}{\sqrt{1+\lambda_1^2+\lambda^2}}\right)}. \tag{7}$$

This model is the hidden truncation equivalent of model (6). If $\lambda_0 = 0$ in (7), then,

$$F_{SN}\left(0; \frac{-\lambda_1\lambda}{\sqrt{1+\lambda_1^2+\lambda^2}}\right) = \pi^{-1}\cos^{-1}\left(\frac{-\lambda_1\lambda}{\sqrt{1+\lambda_1^2}\sqrt{1+\lambda^2}}\right).$$

Thus, the density defined by equation (7) extends the model proposed in Jamalizadeh et. al. (2008). Indeed, when $\mu = 0$ and $\sigma = 1$ in equation (7), it can be regarded as an *extended two-parameter generalized skew-normal distribution*, denoted as $GSN(\lambda_0, \lambda_1, \lambda)$, since it extends the two parameter generalized skew-normal distribution of Jamalizadeh et. al. (2008) with an extra parameter, λ_0.

3.2 Inherent skewness and skewness due to selection

As mentioned, wage distribution may be skewed in the population. The samples observed can be a biased selection of the target population, if the decisions to work are related to the process that determines wages. We refer to the former as the population or inherent skewness. The latter is the induced skewness due to selection into the labour force. To capture the two types of skewness in real data, a population model with skew residuals, as described in section 3.1, is essential.

The link between equation (6) and its hidden truncation counterpart, equation (7), shows that there is a relationship between skewness and selection. Parameter λ_1 is the inherent skewness, ρ determines λ, the skewness arising from selection and λ_0, which is a function of the selection parameter, is a shift parameter. The shift parameter has the additional role of regulating kurtosis (Arellano-Valle & Genton, 2010). Note that ρ retains its role in characterizing the missing data mechanism. When $\rho = 0$, and $\lambda = 0$, the distribution simplifies to the Azzalini (1985) model.

We emphasize that the type of tail behavior of the asymmetric distribution, as characterized by Jones (2014), will not affect the interpretation of both the inherent skewness and skewness due to selection. Briefly, asymmetric distributions with the same tail behavior in each direction (e.g. the family of two-piece distributions) exhibit main-body-skewness while asymmetric distributions with different tail behavior in each direction show tail-skewness (e.g. skew distributions arising from hidden truncation). The Azzalini (1985) skew-normal distribution exhibits tail-skewness and is a natural model in contexts where data is generated by hidden truncation. We present next moment and maximum likelihood estimators of the SSNM model.

3.3 Moments and Maximum Likelihood Estimators of the Selection skew-normal model

The conditional expectation of the random variable defined by the density (6) can readily be derived using its hidden truncation equivalent, model (7). Let $Z_{\lambda_0,\lambda_1,\lambda} \sim GSN(\lambda_0, \lambda_1, \lambda)$ have density function

$$f_Z(z) = k(\lambda_0, \lambda_1, \lambda)\phi(z)\Phi(\lambda_1 z)\Phi(\lambda_0 + \lambda z), \quad z \in \mathbb{R}, \tag{8}$$

where $k(\lambda_0, \lambda_1, \lambda)^{-1} = \frac{1}{2} F_{SN}\left(\frac{\lambda_0}{\sqrt{1+\lambda^2}}; \frac{-\lambda_1\lambda}{\sqrt{1+\lambda_1^2+\lambda^2}}\right)$.

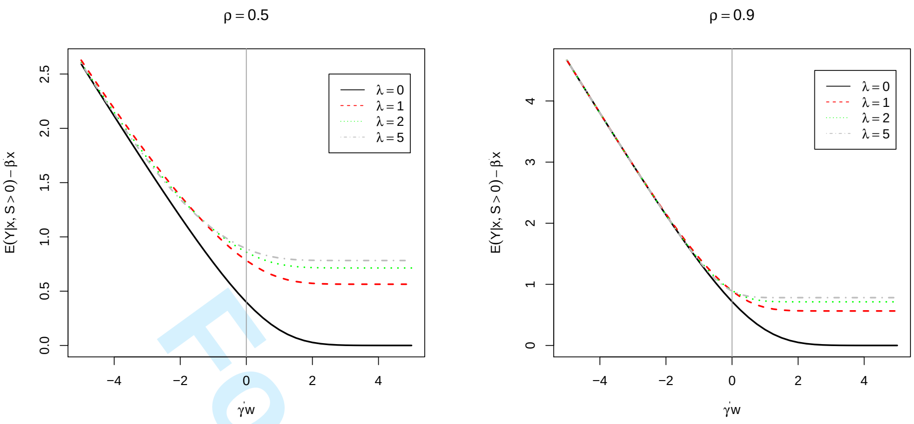
Theorem 1 *The moment generating function of* $Z_{\lambda_0,\lambda_1,\lambda} \sim GSN(\lambda_0, \lambda_1, \lambda)$ *is*

$$M(t; \lambda_0, \lambda_1, \lambda) = k(\lambda_0, \lambda_1, \lambda)e^{t^2/2}\Phi_2\left(\frac{\lambda_1 t}{\sqrt{1+\lambda_1^2}}, \frac{\lambda_0 + \lambda t}{\sqrt{1+\lambda^2}}; \frac{\lambda_1\lambda}{\sqrt{1+\lambda_1^2}\sqrt{1+\lambda^2}}\right), \tag{9}$$

where $k(\lambda_0, \lambda_1, \lambda)$ is as given in (8) and $\Phi_2(.,.,\rho)$ denotes the CDF of $N_2(0, 0, 1, 1, \rho)$. The derivation of equation (9) can be found in the Appendix. If we use an affine transformation and the regression parametrization of section 3.1, the conditional mean of Y can be derived from (9) as

$$\begin{aligned} E(Y|x, S^\star > 0) =& \beta' x + \sigma\left[\left(\frac{2}{F_{SN}\left(\gamma' w; \frac{-\lambda_1\rho}{\sqrt{1+\lambda_1^2-\lambda_1^2\rho^2}}\right)}\right)\left\{\frac{1}{\sqrt{2\pi}}\frac{\lambda_1}{\sqrt{1+\lambda_1^2}}\Phi\left(\frac{\gamma' w\sqrt{1+\lambda_1^2}}{\sqrt{1+\lambda_1^2-\lambda_1^2\rho^2}}\right)\right.\right. \\ &\left.\left.+\rho\phi(\gamma' w)\Phi\left(\frac{-\gamma' w\lambda_1\rho}{\sqrt{1+\lambda_1^2-\lambda_1^2\rho^2}}\right)\right\}\right]. \end{aligned} \tag{10}$$

When $\lambda_1 = 0$ in equation (10), we have the Heckman two-step model given in equation (3). To visualize the impact of using a selection-normal model when the correct model for the conditional mean is given by equation (10), we plot the second component of the expectation, i.e. the correction $E(Y|x, S^\star > 0)$- $\beta' x$, as a function of $\gamma' w$, the mean of the selection variable. We take $\rho = 0.5$ and 0.9 for values of λ_1 = 0, 1, 2 and 5, where $\lambda_1 = 0$ corresponds to the inverse Mills ratio correction in (3). The standard deviation, σ, simply scales the correction factor and ρ is the correlation between the outcome and the selection process.



(a) (b)

Figure 1: Correction factor for different values of skewness parameter with $\lambda = 0$ corresponding to the normal case: (a) $\rho = 0.5$; (b) $\rho = 0.9$.

Figure 1 shows that for positive values of the selection linear predictor $\gamma' w$, the conditional expectation will be underestimated under the usual selection-normal model. This underestimation increases as the skewness increases. However, for negative values of $\gamma' w$, the underestimation of the conditional expectation by the selection-normal model compared to selection skew-normal model decreases and the difference dies out as $\gamma' w$ becomes more negative and missingness increases. This observation is also true for other values of ρ (results not shown).

The complete density of the skew-normal selection model, like the selection normal model, has a continuous component given by (6) and a discrete component for $P(S|w)$. As stated earlier, the marginal distribution of the selection process determines the nature of the distribution of the binary variable, which in this case has probability

$$P(S = s|w) = \{F_{SN}(\gamma' w; \lambda^\star)\}^s \{1 - F_{SN}(\gamma' w; \lambda^\star)\}^{1-s},$$

where $\lambda^\star = -\lambda_1 \rho / \sqrt{1 + \lambda_1^2 - \lambda_1^2 \rho^2}$. This is a binary regression model with a skew-normal link. The log-likelihood function is therefore

$$l(\Omega) = \sum_{i=1}^{n} S_i \Big(\ln f(y_i|x_i, S_i = 1)\Big) + \sum_{i=1}^{n} S_i \Big(\ln F_{SN}(\gamma' w_i; \lambda^\star)\Big) + \sum_{i=1}^{n} (1 - S_i) \ln F_{SN}(-\gamma' w_i; -\lambda^\star).$$

4 Simulation Studies

4.1 Finite-sample properties of the SSNM estimator

In this section we study the finite-sample properties of the selection skew-normal model (SSNM). We compare its performance with the selection normal model (Heckman, 1976), and Heckman's two-step method (Heckman, 1979). We first consider simulation settings where the outcome model is skew but the selection model is normal. This setting is similar to that with wage determinants, described in the introduction. The outcome equation is $Y_i^\star = 0.5 + 1.5x_i + \varepsilon_{1i}$, where $x_i \overset{iid}{\sim} N(0,1)$ and $i = 1, \ldots, N = 1000$. Two selection equations are considered: with an exclusion restriction, $S_i^\star = 1 + x_i + 1.5w_i + \varepsilon_{2i}$, $w_i \overset{iid}{\sim} N(0,1)$, and without an exclusion restriction, $S_i^\star = 1 + x_i + \varepsilon_{2i}$. Hence, $\beta' = (0.5, 1.5)$, and $\gamma' = (1, 1, 1.5)$ and (1, 1) for selection with and without the exclusion restriction, respectively. The covariates x_i and w_i are independent and are also independent of the error terms $\varepsilon_i' = (\varepsilon_{1i}, \varepsilon_{2i})$. The error terms are generated from bivariate skew-normal distributions with λ_1 = 0, 0.5, 1, and 2. The covariance matrix is $\Sigma = \begin{pmatrix} \sigma^2 & \rho\sigma \\ \rho\sigma & 1 \end{pmatrix}$, where $\sigma = 1$ and the correlation $\rho = 0.5$. The $\lambda_1 = 0$ case corresponds to an underlying bivariate normal process. We only observe values of $Y_i^\star$ when $S_i^\star > 0$. About 30% of values are not observed with this exclusion restriction, and about 20% without it. Similar settings have been used for the evaluation of robustness of the t distribution in a sample selection framework (Marchenko & Genton, 2012). As x_i increases, P($S_i^* > 0$) increases, so Y_i is more likely to be observed and larger values of x will be over-represented. For ρ near one (-1), larger errors, ε_i, increase (decrease) the chance that Y_i is observed. A pilot simulation showed that the choice of the parameters used for the simulation have no effect on the results. Simulation results are based on 1000 replications. For brevity, we summarize the results only for $\lambda_1 = 0$ and $\lambda_1 = 1$ in Tables 1 and 2, as the pattern is similar for the other skewness values.

The results with the exclusion restriction show that SSNM is generally less biased for $\lambda_1 = 1$ (Table 1). The bias in SNM and TS generally increases with λ_1. Even under the normality assumption, $\lambda_1 = 0$, the performance of SSNM is comparable to SNM and TS. Although SNM and TS have slightly less biased estimates for the intercept of the outcome model, the intercept of the selection equation is less biased for SSNM than SNM and TS. In terms of variance, SNM and TS are generally more efficient. Other parameters are comparable across the three models. Hence, SNM and TS do not show great advantages overall even with underlying normal assumptions.

In the absence of an exclusion restriction, with a normal underlying process, the SSNM intercept has a lower bias than SNM but higher than TS (Table 2). For the regression parameters of interest, the three models are comparable, so again the SSNM model appears useful even when the underlying process is normal. When λ_1 increases, the bias and variance of SSNM decrease. The variances of the outcome and selection models intercepts are consistently larger for the SSNM model. The loss of efficiency is perhaps due to estimation of the extra skewness parameter.

The SSNM estimates are generally better than the SNM and TS models for σ and ρ when $\lambda_1 \geq 1$ both in the presence and absence of the exclusion restriction (Tables 1 and 2). Since the variance σ^2 describes the variability of the outcomes Y_i, correct prediction intervals for new observations will be obtained under SSNM model. Further, in applied settings (e.g. the MINT Trial, section 5.1), interest may be in patients who do not return their questionnaire.

This requires a correct model for the selection process. The SSNM gave consistently smaller bias compared to SNM and TS models for the selection equation when $\lambda_1 \geq 1$, even under normality assumptions, with or without the exclusion restriction. The results with varying underlying correlation, in the presence of exclusion restriction, for $\lambda_1 = 1$ and 2 are similar to those for $\rho = 0.5$.

If the correlation between the outcome and selection equations is negative, as $-S^\star$ has correlation $-\rho$ with $Y^\star$, and $-S^\star > 0$ is equivalent to $S^\star < 0$, a simple transformation recovers the original model. The TS expression in equation (3) is replaced by $E(Y|x, S^\star < 0) = \beta' x - \sigma\rho\Lambda(-\gamma' w)$. The equivalent substitution in equation (10) can be used to impute the missing values in the data.

Table 1: Simulation results for estimators in the presence of exclusion restriction.

		Bias			Variance		
		SSNM	SNM	TS	SSNM	SNM	TS
$\lambda_1 = 0.0$	β_0	0.002	-0.000	0.000	0.011	0.002	0.003
	β_1	-0.000	-0.000	-0.001	0.002	0.002	0.002
	γ_0	0.006	0.007	0.007	0.007	0.005	0.005
	γ_1	0.004	0.005	0.006	0.006	0.006	0.006
	γ_2	0.008	0.010	0.011	0.009	0.009	0.009
	σ	0.003	-0.001	-0.001	0.002	0.000	0.001
	ρ	-0.001	-0.001	-0.002	0.008	0.008	0.011
	λ_1	-0.003	-	-	0.018	-	-
$\lambda_1 = 1.0$	β_0	0.045	0.562	0.562	0.034	0.002	0.002
	β_1	0.000	0.001	0.001	0.001	0.001	0.001
	γ_0	0.040	0.352	0.353	0.027	0.008	0.009
	γ_1	0.011	0.053	0.055	0.007	0.007	0.007
	γ_2	0.020	0.084	0.086	0.013	0.012	0.013
	σ	-0.011	-0.170	-0.170	0.007	0.000	0.000
	ρ	-0.072	-0.064	-0.066	0.013	0.012	0.014
	λ_1	-0.050	-	-	0.145	-	-

SSNM - Selection skew normal model; SNM - Selection normal model; TS - Two step estimators.

Classical sample selection models may attribute the presence of skewness in the distribution of residuals to selection bias. A case study for this scenario will be discussed in section 5.1.1, where the SNM model attributed the presence of (spurious) skewness to selection bias. To study this effect, we examine the influence of skewness on the standard tests for selection bias, $\rho = 0$, using 95% coverage properties of ρ. The effect of skewness on the power of the selection bias test is also examined when data are generated with $\rho = 0.1$, 0.3 and 0.5. We study the effects with $\lambda_1 = 0$, 0.5, 1 and 2. For the parameter combination $\rho = 0$ and $\lambda_1 = 0$, the SSNM model experienced numerical instability, with about 30% of runs not converging. This is because the model is essentially the Azzalini (1985) model when $\rho = 0$. Similar numerical instability has been noted with direct parametrization of the Azzalini (1985) model. To make the methods (SSNM and classical estimators) directly comparable, we replaced the SSNM estimator with the SNM estimator for the non-converging models.

Table 3 shows the coverage and power of the test of selection bias ($\rho = 0$), using Wald confidence intervals (CIs) for the estimators. For $\rho = 0$, the table shows that the coverage of the

Table 2: Simulation results in the absence of exclusion restriction.

		Bias			Variance		
		SSNM	SNM	TS	SSNM	SNM	TS
	β_0	0.014	0.015	0.005	0.061	0.008	0.012
$\lambda_1 = 0.0$	β_1	-0.012	-0.012	0.004	0.006	0.006	0.009
	γ_0	-0.000	0.007	0.007	0.017	0.004	0.004
	γ_1	0.003	0.010	0.010	0.006	0.005	0.005
	σ	0.023	-0.002	0.006	0.006	0.001	0.002
	ρ	-0.036	-0.043	-0.024	0.046	0.043	0.065
	λ_1	0.002	-	-	0.114	-	-
	β_0	0.064	0.558	0.560	0.034	0.004	0.005
$\lambda_1 = 1.0$	β_1	-0.008	0.005	0.003	0.004	0.004	0.004
	γ_0	0.076	0.526	0.534	0.047	0.007	0.007
	γ_1	0.009	0.043	0.049	0.007	0.007	0.006
	σ	-0.014	-0.164	-0.163	0.007	0.001	0.001
	ρ	-0.067	-0.060	-0.073	0.072	0.051	0.067
	λ_1	-0.076	-	-	0.145	-	-

SSNM - Selection skew normal model; SNM - Selection normal model; TS - Two step estimators.

test, the proportion of simulations in which the Wald intervals include $\rho = 0$, is always slightly less than 0.95. Coverage generally increases as λ_1 increases. Results are similar for the three methods. The power for $\rho = 0.1$ is low, but increases rapidly with ρ, for given λ_1. For fixed ρ the power decreases as the skewness (λ_1) increases. As λ_1 and ρ are not orthogonal, and the probability that $S^\star > 0$ decreases with both ρ and λ_1, power is lost because fewer outcomes are observed. In addition, the ratio of the model based standard errors and the empirical standard errors are close to 1 for all the models for the values of skewness (λ_1) considered.

4.2 Impact of first stage skewed error

In many econometric applications of sample selection models, a probit model is used to describe the first stage selection process. We study the impact of skewness in the first stage using the SSNM and classical sample selection estimators. Data were generated for the exclusion restriction case as in section 4.1, but with skewness only in the selection equation. We fix ρ at 0.5 and consider λ_2, the skewness in the selection equation, for 0.5 and 1. The case with $\lambda_2 = 0$ is omitted as it is equivalent to the result in Table 1 with normal errors. Figures 2 and 3 are used to summarize the results for the parameters in the models. The TS estimator results (omitted) are similar to the SNM results. Both the SSNM and the SNM estimators for the intercepts in the outcome and selection models (β_0 & γ_0) are biased, with marginal gains with the use of the SSNM estimator. The estimators of the coefficient of the covariate in the outcome model (β_1) is unaffected but the covariates in the selection process (γ_1 & γ_2) are biased. Both σ and the correlation between selection and outcome residuals (ρ) are under-estimated. As skewness increases, the under-estimation of σ and ρ increases.

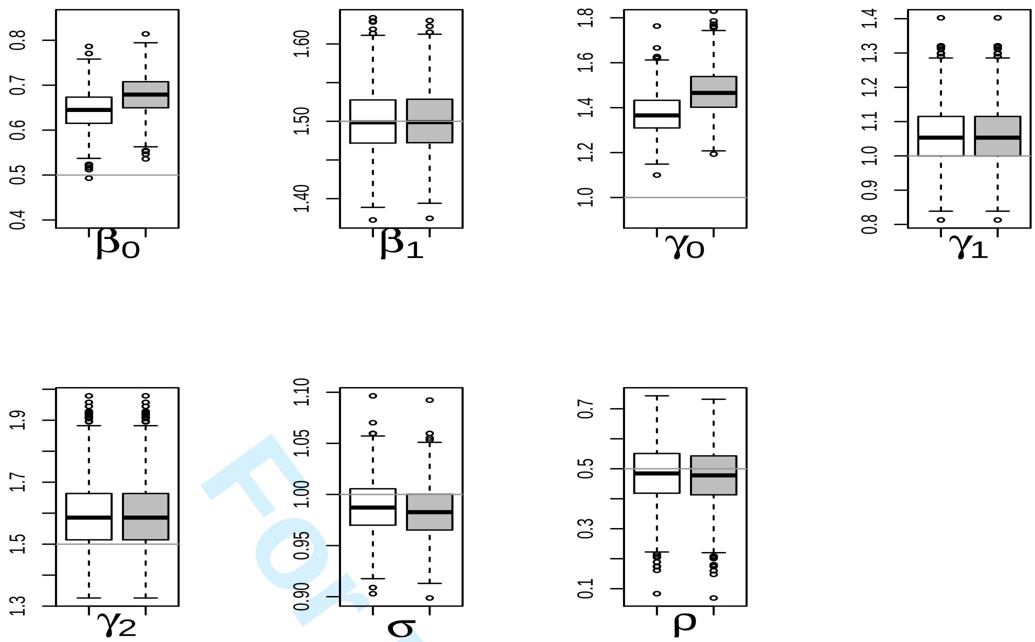


Figure 2: Box plot of model parameters (SSNM - white and SNM - gray) from data generated with skew selection model, $\lambda_2 = 0.5$.

Table 3: 95% Confidence interval coverage and power

(a) Test for selection bias

	λ_1	SSNM	SNM	TS
$\rho = 0.0$	0	93.0	93.4	93.4
	0.5	93.8	93.5	94.0
	1.0	93.9	94.4	94.4
	2.0	94.2	94.3	94.8

(b) Power to detect $\rho \neq 0$

	λ_1	SSNM	SNM	TS
$\rho = 0.1$	0	17.0	19.7	16.0
	0.5	15.5	15.4	14.8
	1.0	13.2	12.7	12.2
	2.0	10.3	9.3	8.8
$\rho = 0.3$	0	76.1	76.2	74.8
	0.5	73.1	72.4	69.5
	1.0	60.8	60.3	58.1
	2.0	48.4	47.2	45.4
$\rho = 0.5$	0	99.6	99.6	99.4
	0.5	98.5	98.3	98.1
	1.0	96.6	96.2	94.8
	2.0	89.8	88.7	86.0

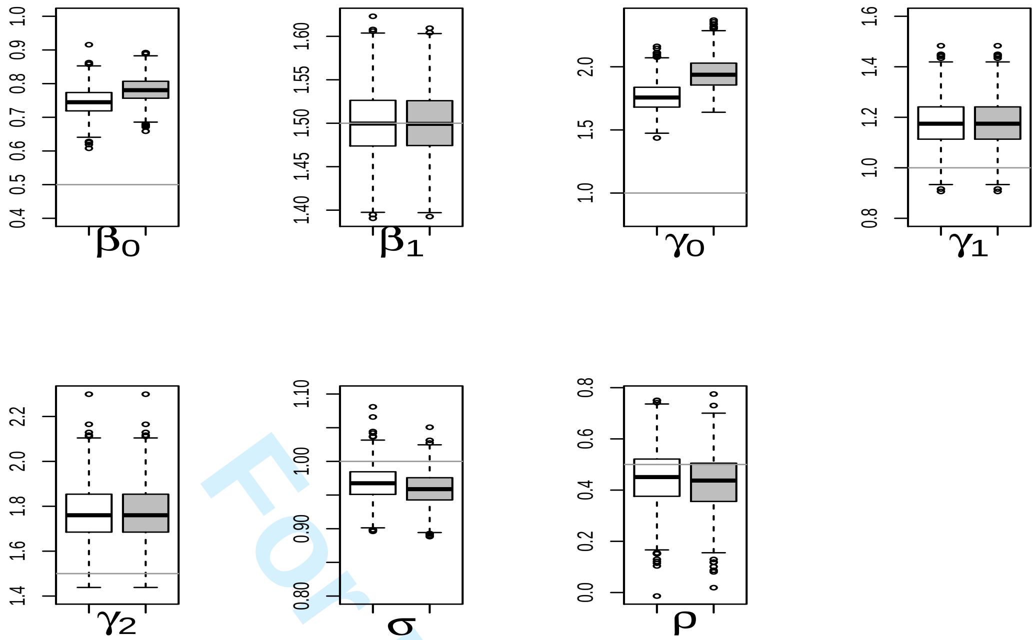


Figure 3: Box plot of model parameters (SSNM - white and SNM - gray) from data generated with skew selection model, $\lambda_2 = 1$.

5 Empirical Studies

We present two examples that illustrate the performance of the SSNM model. The first example illustrates the ability of the SSNM model to capture (spurious) skewness in bounded scores. In the second example, we consider the merits of the model in a data set where $\log$- transformation could not mitigate the effect of non-normality resulting from inherent skewness.

5.1 Application to MINT Trials

We examine data from a multi-center randomized controlled trial of treatments for Whiplash Associated Disorder, the Managing Injuries of the Neck Trial (MINT), in which two treatment regimes were compared: physiotherapy versus reinforcement of advice for patients with continuing symptoms three weeks after their initial visit to the Emergency Department (ED) (Lamb et. al., 2007). As with many patient-reported outcome or quality of life studies, data were collected using questionnaires at regular intervals, 4, 8 and 12 months, after patients' ED attendance.

The primary outcome of interest is return to normal function after the whiplash injury, and is measured using the Neck Disability Index (NDI). The NDI is a self-completed questionnaire which assesses pain-related activity restrictions in 10 areas including personal care, lifting, sleeping, driving, concentration, reading and work, leading to a score between 0 and 50. The NDI has been shown to be reliable and valid (Vernon & Mior, 1991), hence its use as a standard instrument for measuring self-rated disability due to neck pain. The fact that the responses were derived from the use of a 10-item questionnaire posed several challenges. These include the discreteness (Likert-scale type) of the scores, item and unit non-response and dropout with time, which might be responsible for the skewness present in the observed data.

There are 599 patients with a total of 1934 measurements, and 372 (62%) patients have complete observations (i.e. scores at all measurement occasions). The age range is 18 to 78 years. Vernon recommended that patients with only 2 missed items should be considered complete, with mean imputation used for adjustment. We follow this recommendation and any patient with more than 2 missing items is considered as unit missing. In effect, we have only unit non-response left in the data-set and we consider monotone dropout. We first identify predictors of dropout at each measurement occasion using probit regression. At month 8, age and sex of the patients are good predictors of missingness. We restricted attention to this measurement occasion to illustrate the new model.

5.1.1 Application of selection skew-normal model to the NDI scores

A preliminary fit of the SSNM, SNM and TS models to the NDI scores showed that sex did not have a significant effect on NDI, so it was removed from the outcome equation. This provided an exclusion restriction. The covariates in the selection equation are $w = (1, age, sex)$, while the outcome equation contains $x = (1, age, previous, physiotherapy)$, where $previous$ is the score at 4 weeks.

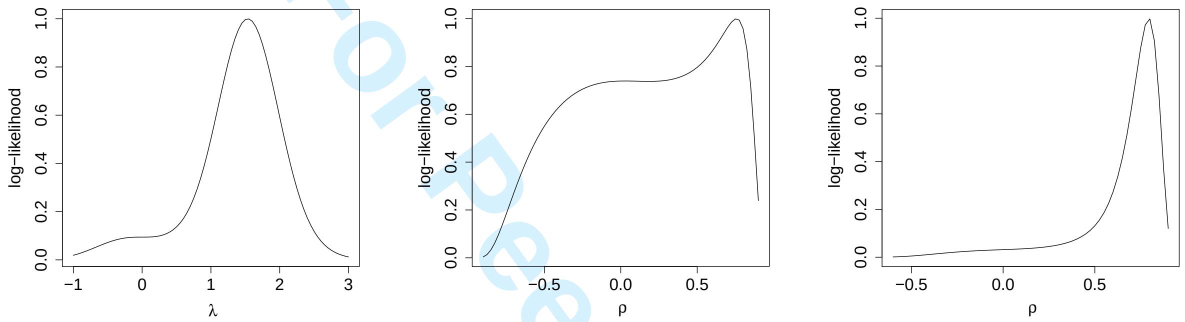
Table 4: Fit of selection skew-normal model, selection-normal model, and Heckman two-step model to the NDI scores at 8 months.

	Selection skew-normal		Selection-normal		Two-step	
	Estimate	95% PI	Estimate	95% PI	Estimate	95% CI
			Selection Equation			
(Intercept)	0.208	(-0.05,1.34)	0.835	(0.64,1.03)	0.818	(0.59,1.04)
age	0.021	(0.01,0.04)	0.024	(0.01,0.04)	0.025	(0.01,0.04)
sex(female)	0.309	(0.08,0.65)	0.335	(0.09,0.59)	0.383	(0.09,0.68)
			Outcome Equation			
(Intercept)	-3.769	(-5.29,-1.53)	0.799	(-0.43,2.01)	1.030	(-2.43,4.49)
age	0.074	(-0.00,0.12)	0.086	(0.04,0.13)	0.068	(-0.02,0.16)
previous	0.678	(0.61,0.75)	0.687	(0.62,0.75)	0.708	(0.64,0.78)
physiotherapy	0.766	(-0.27,1.81)	0.887	(-0.17,1.95)	1.007	(-0.07,2.08)
σ	7.723	(6.51,8.84)	6.166	(5.60,6.76)	5.703	(1.71,9.69)
ρ	0.758	(-0.77,0.91)	0.794	(0.52,0.90)	0.474	(-0.78,1.73)
λ_1	1.537	(0.54,2.45)	-	-	-	-
log-likelihood	-1452.672		-1455.033		-	

PI - profile likelihood interval; CI: Wald confidence interval.

Profile likelihood ratio intervals (PI) are used in Table 4 to quantify uncertainties in the parameters of the SSNM and the SNM models to allow for asymmetry, and Wald confidence intervals are used for the TS method. The SSNM skewness parameter ($\lambda_1 = 1.54$) is statistically significant. Both the Wald test and the PI under the SNM suggest that there is selection bias (i.e. $\rho \neq 0$). However, the TS method and the SSNM imply $\rho = 0$. This behavior is typical of the SNM where non-normality of the errors are attributed to the presence of selection bias (Marchenko & Genton, 2012). The SSNM model has a better fit to the NDI data than the SNM model (likelihood ratio test statistic 4.72, p<0.03).

The introduction of extra parameters gives a more flexible model, but can lead to difficulties in model identifiability. The profile log-likelihood for the shape parameter of a univariate skew-normal distribution always has a stationary point at $\lambda_1 = 0$. This problem is also visible in the SSNM model since it has the Azzalini's skew-normal distribution as its basis. To illustrate this, we examine the profile log-likelihood for the parameters λ_1 and ρ for the NDI scores. At $\lambda_1 = 0$ in Figure 4a, the profile log-likelihood has a stationary point. The profile log-likelihood for ρ under the SSNM (see Figure 4b) is flat in the neighborhood of zero, but less flat for SNM (see Figure 4c).



(a) (b) (c)

Figure 4: Normalized profile log-likelihoods of skewness and correlation parameters (NDI scores): (a) SSNM (λ_1); (b) SSNM(ρ); (c) SNM(ρ).

The results under the SSNM model are consistent with $\rho = 0$, i.e. MAR, with the skewness in the response variable being intrinsic to the measured outcome, and not due to selection. Similarly, the TS model (with bootstrap standard errors) also supports the MAR assumption. However, the SNM model suggests the data are MNAR; if the outcome variable is normal in the population, informative missingness is required to explain the observed result.

5.1.2 Application of SSNM model to transformed NDI scores

Although the hypothesis of normality was rejected for a test of size 5% for the NDI scores using likelihood ratio test, it cannot be rejected at 1%. This is perhaps due to (spurious) skewness in the data as a result of the ceiling and floor effects of the bounds. In addition, the scores are clustered at one end of the scale making the distribution J-shaped. We examine the impact of transformation on the NDI scores using a logistic transformation, which is commonly used for bounded scores. We transformed the NDI scores to $Z = \log\left(\frac{NDI}{50-NDI}\right)$ because none of the values of NDI were 50 or 0. The parameter estimates of the three models are very similar and the skewness parameter is not significantly different from zero; the logistic transformation captures the skewness of the bounded scores (Table 5). All three models show no selection for the transformed scores. Treatment effects are borderline significant for the transformed scores, but not significant for the untransformed scores. The interpretations of model parameters are different. For untransformed scores, the parameter estimates give estimates of the theoretical mean for the covariates in the unselected population (Arnold et al., 1993, p483). In Table 5 however, the interpretation is more complex as the SSNM and SNM models are based on the logit-skew normal and logit-normal distributions respectively.

Table 5: Fit of selection skew-normal model, Selection-normal model, and Heckman two-step model to the transformed NDI scores at 8 months.

	Selection skew-normal		Selection-normal		Two-step	
	Estimate	95% PI	Estimate	95% PI	Estimate	95% CI
			Selection Equation			
(Intercept)	0.815	(0.54,1.59)	0.816	(0.60,1.04)	0.818	(0.60,1.04)
age	0.025	(0.01,0.04)	0.025	(0.01,0.04)	0.025	(0.01,0.04)
sex(female)	0.387	(0.09,0.68)	0.386	(0.10,0.67)	0.383	(0.10,0.67)
			Outcome Equation			
(Intercept)	-2.618	(-3.08,-1.26)	-2.613	(-2.80,-2.17)	-2.680	(-3.02,-2.34)
age	0.007	(-0.01,0.02)	0.007	(-0.01,0.01)	0.009	(-0.00,0.02)
previous	0.083	(0.08,0.10)	0.083	(0.08,0.09)	0.083	(0.08,0.09)
physiotherapy	0.138	(-0.00,0.45)	0.138	(-0.00,0.28)	0.138	(-0.00,0.28)
σ	0.727	(0.68,1.34)	0.727	(0.68,0.79)	0.748	(0.26,1.24)
ρ	0.102	(-0.91,0.90)	0.102	(-0.90,0.51)	0.414	(-0.96,1.79)
λ_1	0.009	(-1.15,1.12)	-	-	-	-
loglike	-639.537		-639.538		-	

5.2 Wage offer function of married women

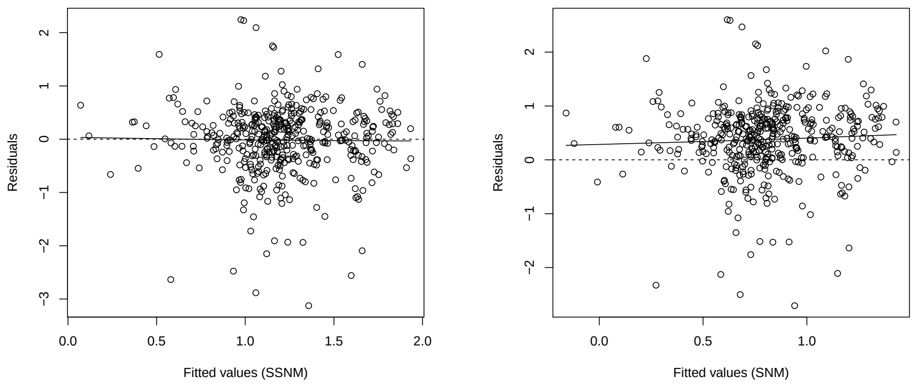
We re-analyze the data reported in Mroz (1987) on married women's labour force participation, from the University of Michigan Panel Study of Income Dynamics for the year 1975 (interview year was 1976). We estimate a simple female wage model which takes into account the labour force participation. This model specification was used to illustrate the R package *sampleSelection* (see Henningsen & Toomet, 2014). The outcome of interest, female wage, is missing for 325 (43%) of the 753 women in the sample. Logarithm of wage depends on education status and city, i.e. $x = (1, educ, city)$. The selection equation uses husband's wage, number of children 5 years old or younger, marginal tax rate of the wife and the wife's father's educational attainment as well as education status and city. That is, $w = (x, huswage, kid5, mtr, fatheduc, educ, city)$. Other model specifications for this example have been considered in the literature (see various model specifications in Henningsen & Toomet, 2014).

The results of fitting a SSNM model and its competitors underscore the benefits of the model even when the outcome is transformed to near normality (Table 6). Although the statistical significance of covariates in the models are similar, the skewness parameter is significant at a 1% level. In addition, the SSNM provides improved fit over the SNM as it captured inherent skewness (λ_1) in the outcome. The profile likelihoods are well behaved due to the logarithm transformation of the outcome variable.

Table 6: Fit of Selection skew-normal model, Selection-normal model, and Heckman two-step model to labor wage data.

	Selection skew-normal		Selection-normal		Two-step	
	Estimate	95% PI	Estimate	95% PI	Estimate	95% CI
			Selection Equation			
(Intercept)	2.916	(1.59,4.41)	3.802	(2.33,5.33)	3.374	(1.69,5.06)
huswage	-0.094	(-0.13,-0.07)	-0.103	(-0.14,-0.07)	-0.117	(-0.15,-0.08)
kids5	-0.384	(-0.57,-0.23)	-0.415	(-0.59,-0.25)	-0.550	(-0.74,-0.36)
mtr	-5.134	(-6.80,-3.66)	-5.782	(-7.48,-4.15)	-5.329	(-7.20,-3.46)
fatheduc	-0.016	(-0.04,0.01)	-0.020	(-0.05,0.01)	-0.005	(-0.04,0.03)
educ	0.095	(0.05,0.14)	0.112	(0.07,0.16)	0.122	(0.07,0.17)
city	-0.043	(-0.22,0.13)	-0.040	(-0.24,0.16)	-0.031	(-0.24,0.18)
			Outcome Equation			
(Intercept)	1.102	(0.63,1.56)	0.669	(0.25,1.10)	0.642	(0.07,1.21)
educ	0.075	(0.04,0.11)	0.066	(0.03,0.10)	0.067	(0.03,0.10)
city	0.120	(-0.02,0.26)	0.107	(-0.04,0.26)	0.100	(-0.05,0.25)
σ	0.964	(0.86,1.08)	0.800	(0.73,0.88)	0.797	(0.65,0.94)
ρ	-0.792	(-0.88,-0.55)	-0.780	(-0.86,-0.65)	-0.758	(-1.06,-0.45)
λ_1	-1.588	(-2.16,-1.05)	-	-	-	-
loglike	-874.608		-881.802		-	

Figure 5 shows the residual plots for the SSNM and the SNM models using equations (10) and (3) respectively for the labour wage data. The SSNM model residuals are centred on zero, but the SNM model residuals are positively biased. That is, the skewness is not correctly modelled by the SNM model and as such, the residuals have a mean other than zero ($E(X) \neq \mu$). In this case, the model will under predict new outcomes if the model is used for prediction.



(a) (b)

Figure 5: Residual plots for the SSNM and the SNM models using the labour wage data.

6 Concluding Remarks

We introduced a sample selection model with underlying bivariate skew-normal distribution which we called a selection skew-normal model (SSNM). This model is more flexible than the conventional sample selection model since it has an extra parameter that regulates skewness, with the conventional sample selection model as a special case. Its moment estimator was derived using the link between skew models arising from selection and hidden truncation formulations of skew models. The moment estimator was shown to extend Heckman's two-step method. Simulation was used to compare maximum likelihood estimator of the model with conventional sample selection models under moderate correlation ($\rho = 0.5$) and skewness varying from 0 to 2. We also considered the effect of varying the correlation ρ with λ_1 fixed at 1 and 2, and an exclusion restriction criteria (results not shown here). The SSNM model outperformed the conventional sample selection models for the skewness parameters considered. The conventional sample selection model has a negligible advantage when $\lambda_1 = 0$ with smaller bias in the intercept of outcome equation.

In addition, we studied the impact of non-normality, due to skewness, in the first stage of sample selection estimators. Our results indicated that the proposed model has marginal improvements over classical methods in terms of unbiasedness. If there are practical reasons to suspect skewness in the selection process, then the model given by equation (5) can be analysed as the parameters in the model have interpretable roles. Alternatively, λ_1 and λ_2 corresponding to the skewness parameters in the outcome and selection processes can be constrained as $\lambda_1 = \lambda_2 = \lambda$. The submodel of (5) analysed here was based on the constraint $\lambda_2 = 0$. This has practical justification. The distribution of income or wages, for example, is expected to be skewed in the source population. However, the decision to work so as to earn wages may not carry sufficient information about shape parameters. We also indicated that negative correlation between the outcome and the selection variables is equivalent to reversing the indicator for selection. The expression for the conditional mean can be used in an imputation model to assess the sensitivity of conclusions when data are suspected to be missing not at random.

We illustrated the strength of the SSNM model in two examples. For the NDI scores, where responses exhibited skewness, which might be due to the effect of bounds on the scores, our model successfully discriminates between selection effects and skewness. The SSNM detected the absence of skewness after a logistic transformation of the scores. However, the transformation requires specification of the bounds, so if estimation of the final recovery level is of interest, then this transformation cannot be used. The labor force participation data illustrated the merits of the SSNM model when a standard transformation could not adjust for extreme non-normality due to inherent skewness in the outcome variable. The SSNM model will be more efficient than semi-parametric and non-parametric approaches within a class of deviations from normality due to skewness. The ability of the model to identify intercepts allows evaluation of model fit using residual analysis. The model can easily be used for predictive purposes using its moment based expression.

The model presented here is very simple to use and the likelihood function can be easily coded in R software. Starting values can be obtained using the two-step method (TS). The optimization routine used was BFGS in R software but other numerical maximization algorithms can be used as well, although we do not recommend the use of Nelder-Mead optimization method which appears not to work well with the CDF of Azzalini's skew-normal distribution. We recommend the use of profile likelihood ratio interval for inference on the model parameters as the likelihood function is usually asymmetrical.

It is straightforward to extend these results to model skewness and heavy tails simultaneously. If we assume a bivariate skew-t distribution for the outcome and selection equations, but restrict the skewness parameter to zero in the selection equation, we can develop a new class of sample selection model. The continuous component of this model is a form of the univariate extended skew-t distribution, EST (see Arellano-Valle & Genton, 2010), but with an additional skewness parameter. Arellano-Valle & Genton (2010) gave analytic proof that the EST distribution, unlike the ESN, does not have a stationary point at $\lambda_1 = 0$. We expect the additional skewness parameter in the EST distribution not to induce stationarity at $\lambda_1 = 0$ in the model, and thus produce a more stable parameter estimates than the SSNM model. In addition, the selection equation is a binary regression with the skew-t link.

To apply this model in practice, we recommend that the model is fitted in conjunction with the conventional sample selection model, to assess the degree of departure from symmetry.

Acknowledgements

The authors would like to thank the editor, an associate editor and the anonymous referees for the constructive comments, which led to improvements in the manuscript. This work was carried out while EOO was a PhD student at Warwick Statistics, in receipt of support from CRiSM, EPSRC grant reference no. EP/D002060/1. We thank Professor S Lamb and the MINT team for access to the data.

References

Ahn, H. & Powell, J. L. (1993). Semi-parametric estimation of censored selection models with a nonparametric selection mechanism. *J. Econometrics* **58**, 3–29.

Andrews, D. W. & Schafgans, M. M. (1998). Semiparametric Estimation of the Intercept of a Sample Selection Model. *Rev. Econ. Stud.* **65**, 497–517.

Arellano-Valle, R. B., Branco, M. D. & Genton, M. G. (2006). A unified view of skewed distributions arising from selections. *Canad. J. Statist.* **34**, 581–601.

Arellano-Valle, R. B. & Genton, M. G. (2010). Multivariate extended skew-t distributions and related families. *Metron* **68**, 201–234.

Arnold, B. C. & Beaver, R. J. (2002). Skewed multivariate models related to hidden truncation and/or selective reporting (with discussion). *Test* **11**, 7–54.

Arnold, B. C., Beaver, R. J., Groeneveld, R. A. & Meeker, W. Q. (1993). The nontruncated marginal of a truncated bivariate normal distribution. *Psychometrika* **58**, 471–488.

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scand. J. Statist.* **12**, 171–178.

Copas, J. B. & Li, H. G. (1997). Inference for Non-random Samples. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **59**, 55–95.

Das, M., Newey, W. K. & Vella, F. (2003). Non-parametric estimation of sample selection models. *Rev. Econ. Stud.* **70**, 33–58.

Ding, P. (2014). Bayesian robust inference of sample selection using selection-t models. *J. Multivariate Anal.* **124**, 451–464.

Enders, C. K. (2010). *Applied Missing Data Analysis*. The Guildford Press, New York.

González-Farías, G., Domínguez-Molina, J. A. & Gupta, A. K. (2004). The closed skew-normal. In *Skew-elliptical distributions and their applications: a journey beyond normality* (ed. M. G. Genton), Chapman & Hall/CRC, Boca Raton, FL.

Heckman, J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Ann. Econ. Soc. Meas.* **5**, 475–492.

Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica* **47**, 153–161.

Henningsen, A. & Toomet O. (2014). Sample Selection Models. *R Package Reference Manual*. `http://cran.r-project.org/web/packages/sampleSelection/sampleSelection.pdf`.

Jamalizadeh, A., Behboodian, J. & Balakrishnan, N. (2008). A two-parameter generalized skew-normal distribution. *Statist. Probab. Lett.* **78**, 1722–1728.

Jones, M. C. (2014). On Families of Distributions With Shape Parameters. *Int. Stat. Rev.* (In Press).

Lamb, S. E., Gates, S., Underwood, M. R., Cooke, M. W., Ashby, D., Szczepura, A., Williams, M. A., Williamson, E. M., Withers, E. J., Isa, S. M & Gumber, A. (2007). Managing Injuries of the Neck Trial (MINT): design of a randomised controlled trial of treatments for whiplash associated disorders. *BMC Musculoskelet Disord.* **8**: 7.

Lee, L. (1983). Generalized econometric models with selectivity. *Econometrica* **51**, 507–512.

Leung, S. F. & Yu, S. (1996). On the choice between sample selection and two-part models. *J. Econometrics* **72**, 197–229.

Leung, S. F. & Yu, S. (2000). Collinearity and two-step estimation of sample selection models: problems, origins and remedies. *Comput. Econ.* **15**, 173–199.

Marchenko, Y. V. & Genton, M. G. (2012). A Heckman selection-t model. *J. Amer. Statist. Assoc.* **107**, 304–317.

Mroz, T. A. (1987). The Sensitivity of an Empirical Model of Married Women's Hours of Work to Economic and Statistical Assumptions. *Econometrica* **55**, 765–799.

Newey, W. K. (2009). Two-step series estimation of sample selection models. *Econometrics Journal* **12**, 217–229.

Puhani, A. P. (2000). The Heckman correction for sample selection and its critique. *J. Econ. Surv.* **14**, 53–68.

Vella, F. (1998). Estimating models with sample selection bias: a survey. *J. Hum. Resour.* **33**, 127–172.

Vernon, H. & Mior, S. (1991). The Neck Disability Index: a study of reliability and validity. *J. Manipulative Physiol. Ther.* **7**, 409–415.

Zhelonkin, M., Genton, M. G. & Ronchetti, E. (2012). Robust Inference in Sample Selection Models. *Manuscript*.

7 Appendix

Details of derivation of the SSNM model

As the joint distribution of the outcomes and the selection process can be written in a closed skew-normal (CSN) distribution form as

$$\begin{pmatrix} Y^\star \\ S^\star \end{pmatrix} \sim CSN_{2,1}\Big\{\boldsymbol{\mu} = (\beta' x, \gamma' w), \Sigma = \begin{pmatrix} \sigma^2 & \rho\sigma \\ \rho\sigma & 1 \end{pmatrix}, D = (\lambda_1/\sigma, \lambda_2), \nu = 0, \Delta = 1\Big\},$$

the distribution of $S|Y$, using the conditional distribution property of CSN, is

$$S|Y \sim CSN_{1,1}\Big\{\gamma' w + \rho\Big(\frac{y - \beta' x}{\sigma}\Big), 1 - \rho^2, \lambda_2, -(\lambda_1 + \lambda_2)\Big(\frac{y - \beta' x}{\sigma}\Big), 1\Big\},$$

and $P(S^\star > 0|y, x)$ is an ESN lower tail probability written as

$$F_{ESN}\Big\{\gamma' w + \rho\Big(\frac{y - \beta' x}{\sigma}\Big); 0, 1 - \rho^2, \frac{-\lambda_2}{\sqrt{1-\rho^2}}, -(\lambda_1 + \lambda_2)\Big(\frac{y - \beta' x}{\sigma}\Big)\Big\}. \tag{11}$$

To determine the expression $P(S^\star > 0|x)$ in equation (1) we need to extract its marginal distribution from the bivariate process. Using the property of marginalization of CSN, we have

$$P(S^\star > 0|w) = F_{SN}\Big(\gamma' w; \frac{-(\lambda_2 + \lambda_1 \rho)}{\sqrt{1 + \lambda_1^2 - \lambda_1^2 \rho^2}}\Big), \tag{12}$$

where F_{SN} denotes the CDF of a skew-normal random variable. The 'base' PDF of the outcome equation is

$$f(y|x) = \frac{2}{\sigma}\phi\Big(\frac{y - \beta' x}{\sigma}\Big)\Phi\Big\{\lambda_1\Big(\frac{y - \beta' x}{\sigma}\Big)\Big\}. \tag{13}$$

Substituting (11), (12) and (13) into the general sample selection equation (1) gives equation (5).

Derivation of Moment Generating function

$$\begin{aligned} E(e^{tZ}) =& k(\lambda_0, \lambda_1, \lambda)\int_{-\infty}^{\infty} e^{tz}\phi(z)\Phi(\lambda_1 z)\Phi(\lambda_0 + \lambda z)\, dz \\ =& k(\lambda_0, \lambda_1, \lambda)e^{t^2/2}\int_{-\infty}^{\infty} \phi(z - t)\Phi(\lambda_1 z)\Phi(\lambda_0 + \lambda z)\, dz \end{aligned}$$

Put $x = z - t$. Then,

$$\begin{aligned}
E(e^{tZ}) =& k(\lambda_0, \lambda_1, \lambda) e^{t^2/2} \int_{-\infty}^{\infty} \phi(x) \Phi(\lambda_1 x + \lambda_1 t) \Phi(\lambda_0 + \lambda x + \lambda t)\, dx \\
=& k(\lambda_0, \lambda_1, \lambda) e^{t^2/2} E(\Phi(\lambda_1 X + \lambda_1 t) \Phi(\lambda_0 + \lambda X + \lambda t)) \\
=& k(\lambda_0, \lambda_1, \lambda) e^{t^2/2} P(Y_1 - \lambda_1 X < \lambda_1 t, Y_2 - \lambda X < \lambda_0 + \lambda t) \\
=& k(\lambda_0, \lambda_1, \lambda) e^{t^2/2} \Phi_2\left(\frac{\lambda_1 t}{\sqrt{1+\lambda_1^2}}, \frac{\lambda_0 + \lambda t}{\sqrt{1+\lambda^2}}; \frac{\lambda_1 \lambda}{\sqrt{1+\lambda_1^2}\sqrt{1+\lambda^2}} \right)
\end{aligned}$$

where X, Y_1, Y_2 are iid $N(0,1)$, and

$$P(Y_1 - \lambda_1 X < \lambda_1 t, Y_2 - \lambda X < \lambda_0 + \lambda t) = \Phi_2\left(\frac{\lambda_1 t}{\sqrt{1+\lambda_1^2}}, \frac{\lambda_0 + \lambda t}{\sqrt{1+\lambda^2}}; \frac{\lambda_1 \lambda}{\sqrt{1+\lambda_1^2}\sqrt{1+\lambda^2}} \right).$$

Derivation of Gradients and Observed information matrix

The gradient of the selection skew-normal model log-likelihood can be derived as follows:

$$\begin{aligned}
\frac{\partial l}{\partial \beta} =& S_i\left(\sum_{i=1}^{n} \left\{ \frac{1}{\sigma} z_i - \frac{\lambda_1}{\sigma} K_3 - \frac{\rho}{\sigma r^{1/2}} K_1 \right\} x_i \right) \\
\frac{\partial l}{\partial \gamma} =& S_i\left(\frac{1}{r} \sum_{i=1}^{n} K_1 x_i \right) + (1 - S_i)\left(\sum_{i=1}^{n} (-2) K_2 x_i \right) \\
\frac{\partial l}{\partial \sigma} =& S_i\left(\sum_{i=1}^{n} \left\{ -\frac{n}{\sigma} + \frac{1}{\sigma} z_i^2 - \frac{\lambda_1}{\sigma} K_3 z_i - \frac{\rho}{\sigma r^{1/2}} K_1 z \right\} \right) \\
\frac{\partial l}{\partial \rho} =& S_i\left(\sum_{i=1}^{n} \frac{1}{r^{3/2}} K_1 \left(\rho \gamma' w_i + z_i \right) \right) + (1 - S_i)\left(\sum_{i=1}^{n} \frac{-2\lambda_1}{\sqrt{2\pi u}} K_4 \right) \\
\frac{\partial l}{\partial \lambda_1} =& S_i\left(\sum_{i=1}^{n} K_3 z_i \right) + (1 - S_i)\left(\sum_{i=1}^{n} \frac{-2\rho}{(1+\lambda_1^2)\sqrt{2\pi u}} K_4 \right),
\end{aligned}$$

where, $r = (1 - \rho^2)$, $z = (y - \beta' x_i)/\sigma$, $u = (1 + \lambda_1^2 - \lambda_1^2 \rho^2)$, and

$$\begin{cases}
\omega = \dfrac{\gamma' w_i + \rho\left(\frac{(y_i - \beta' x_i)}{\sigma} \right)}{\sqrt{1-\rho^2}} \ \&\ K_1 = \phi(\omega)/\Phi(\omega), \quad K_2 = \dfrac{\phi\left(\gamma' w_i\right) \Phi\left(\frac{-\gamma' w_i \lambda_1 \rho}{\sqrt{1+\lambda_1^2-\lambda_1^2\rho^2}} \right)}{F_{SN}\left(-\gamma' w_i; 0, 1, \frac{\lambda_1 \rho}{\sqrt{1+\lambda_1^2-\lambda_1^2\rho^2}} \right)} \\
\zeta = \lambda_1 \left(\frac{y_i - \beta' x_i}{\sigma} \right) \ \&\ K_3 = \phi(\zeta)/\Phi(\zeta), \qquad K_4 = \dfrac{\phi\left(\frac{\gamma' w_i \sqrt{1+\lambda_1^2}}{\sqrt{1+\lambda_1^2-\lambda_1^2\rho^2}} \right)}{F_{SN}\left(-\gamma' w_i; 0, 1, \frac{\lambda_1 \rho}{\sqrt{1+\lambda_1^2-\lambda_1^2\rho^2}} \right)}
\end{cases}$$

Note that the derivative of $F_{SN}\left(-\gamma' w_i; 0, 1, \frac{\lambda_1 \rho}{\sqrt{1+\lambda_1^2-\lambda_1^2\rho^2}} \right)$ w.r.t. γ follows the usual differentiation of CDF to get the PDF. However, the derivatives of ρ and γ in this expression are not a straightforward application of this principle. The approach we followed is to re-write the CDF above as a standard bivariate normal integral $\left(2\Phi_2\left(-\gamma' w_i, 0; -\lambda_1\rho/\sqrt{1+\lambda_1^2}\right)\right)$. We make use of the fact that, if $\Phi_2(.,.;\rho)$ and $\phi_2(.,.;\rho)$ are standard bivariate normal CDF and PDF respectively, then $\frac{d\Phi_2(.,.;\rho)}{d\rho} = \phi_2(.,.;\rho)$.

The elements of the observed information matrix are:

$$\begin{aligned}
\frac{-\partial^2 l}{\partial\beta^2} =& S_i\left(\sum_{i=1}^{n}\left\{\frac{1}{\sigma^2}+\frac{\lambda_1^2}{\sigma^2}\left[\zeta K_3+K_3^2\right]+\frac{\rho^2}{\sigma^2 r}\left[\omega K_1+K_1^2\right]\right\}x_i^2\right)\\
\frac{-\partial^2 l}{\partial\gamma^2} =& S_i\left(\sum_{i=1}^{n}\frac{1}{r}\left\{\omega K_1+K_1^2\right\}x_i^2\right)+(1-S_i)\left(\sum_{i=1}^{n}-\left\{2\gamma' w_i K_2-\frac{2\lambda_1\rho}{\sqrt{2\pi u}}K_4-4K_2^2\right\}x_i^2\right)\\
\frac{-\partial^2 l}{\partial\sigma^2} =& S_i\left(\sum_{i=1}^{n}\left\{-\frac{n}{\sigma^2}+\frac{3}{\sigma^2}z_i^2-\frac{1}{\sigma^2}\left[2\zeta K_3-\zeta^3K_3-\zeta^2K_3^2\right]\right.\right.\\
&\left.\left.-\frac{\rho}{\sigma^2 r^{1/2}}\left[2z_iK_1-\frac{\rho}{r^{1/2}}z_i^2\omega K_1-\frac{\rho}{r^{1/2}}z_i^2K_1^2\right]\right\}\right)\\
\frac{-\partial^2 l}{\partial\rho^2} =& S_i\left(\sum_{i=1}^{n}\left\{-\frac{3\rho}{r^2}\left(\rho\gamma' w_i+z_i\right)K_1-\frac{\gamma' w_i}{r}K_1+\frac{1}{r^{5/2}}\left(\rho\gamma' w_i+z_i\right)^2\omega K_1+\frac{1}{r^{5/2}}\left(\rho\gamma' w_i+z_i\right)^2K_1^2\right\}\right)\\
&+(1-S_i)\left(\sum_{i=1}^{n}\left\{\frac{2\lambda_1^3\rho}{\sqrt{2\pi u^3}}K_4-\frac{2\lambda_1^3\rho(1+\lambda_1^2)(\gamma' w_i)^2}{\sqrt{2\pi u^5}}K_4+\frac{4\lambda_1^2}{2\pi u}K_4^2\right\}\right)\\
\frac{-\partial^2 l}{\partial\lambda_1^2} =& S_i\left(\sum_{i=1}^{n}\left\{\zeta K_3+K_3^2\right\}z_i\right)+(1-S_i)\left(\sum_{i=1}^{n}\left\{\frac{-2\lambda_1\rho^3(\gamma' w_i)^2}{(1+\lambda_1^2)\sqrt{2\pi u^5}}K_4-\frac{2\lambda_1\rho(1-\rho^2)}{(1+\lambda_1^2)\sqrt{2\pi u^3}}K_4\right.\right.\\
&\left.\left.-\frac{4\lambda_1\rho}{(1+\lambda_1^2)^2\sqrt{2\pi u}}K_4+\frac{4\rho^2}{(1+\lambda_1^2)^2 2\pi u}K_4^2\right\}\right)\\
\frac{-\partial^2 l}{\partial\beta\partial\gamma} =& S_i\left(\sum_{i=1}^{n}\left\{-\frac{\rho}{\sigma r}\omega K_1-\frac{\rho}{\sigma r}K_1^2\right\}x_i^2\right)\\
\frac{-\partial^2 l}{\partial\beta\partial\sigma} =& S_i\left(\sum_{i=1}^{n}\left\{\frac{2}{\sigma^2}z_i+\frac{\lambda_1}{\sigma^2}\zeta^2K_3-\frac{\lambda_1}{\sigma^2}K_3+\frac{\lambda_1}{\sigma^2}\zeta K_3^2+\frac{\rho^2}{\sigma^2 r}z_i\omega K_1-\frac{\rho}{\sigma^2 r^{1/2}}K_1+\frac{\rho^2}{\sigma^2 r}z_iK_1^2\right\}x_i\right)\\
\frac{-\partial^2 l}{\partial\beta\partial\rho} =& S_i\left(\sum_{i=1}^{n}\left\{-\frac{\rho}{\sigma r^2}\left(\rho\gamma' w_i+z_i\right)\omega K_1+\frac{1}{\sigma r^{3/2}}K_1-\frac{\rho}{\sigma r^2}\left(\rho\gamma' w_i+z_i\right)K_1^2\right\}x_i\right)\\
\frac{-\partial^2 l}{\partial\beta\partial\lambda_1} =& S_i\left(\sum_{i=1}^{n}\left\{\frac{1}{\sigma}K_3-\frac{1}{\sigma}\zeta^2K_3-\frac{1}{\sigma}\zeta K_3^2\right\}x_i\right)\\
\frac{-\partial^2 l}{\partial\gamma\partial\sigma} =& S_i\left(\sum_{i=1}^{n}\left\{-\frac{\rho}{\sigma r}\omega K_1z_i-\frac{\rho}{\sigma r}K_1^2z_i\right\}x_i\right)\\
\frac{-\partial^2 l}{\partial\gamma\partial\rho} =& S_i\left(\sum_{i=1}^{n}\left\{\frac{1}{r^2}\left(\rho\gamma' w_i+z_i\right)\omega K_1-\frac{\rho}{r^{3/2}}K_1+\frac{1}{r^2}\left(\rho\gamma' w_i+z_i\right)K_1^2\right\}x_i\right)\\
&+(1-S_i)\left(\sum_{i=1}^{n}\left\{\frac{4\lambda_1}{\sqrt{2\pi u}}\phi\left(\frac{\gamma' w_i\sqrt{1+\lambda_1^2}}{u^{1/2}}\right)K_2-\frac{2\lambda_1(1+\lambda_1^2)}{\sqrt{2\pi u^3}}(\gamma' w_i)K_4\right\}x_i\right)
\end{aligned}$$

$$\frac{-\partial^2 l}{\partial\gamma\partial\lambda_1} =(1-S_i)\left(\sum_{i=1}^{n}\left\{\frac{4\rho}{(1+\lambda_1^2)\sqrt{2\pi u}}\phi\left(\frac{\gamma' w_i\sqrt{1+\lambda_1^2}}{u^{1/2}}\right)K_2-\frac{2\rho}{\sqrt{2\pi u^3}}(\gamma' w_i)K_4\right\}x_i\right)$$

$$\frac{-\partial^2 l}{\partial\sigma\partial\rho} =S_i\left(\sum_{i=1}^{n}\left\{\frac{1}{\sigma r^{3/2}}K_1 z_i-\frac{\rho}{\sigma r^2}\left(\rho\gamma' w_i+z_i\right)\omega K_1 z_i-\frac{\rho}{\sigma r^2}\left(\rho\gamma' w_i+z_i\right)K_1^2 z_i\right\}\right)$$

$$\frac{-\partial^2 l}{\partial\sigma\partial\lambda_1} =S_i\left(\sum_{i=1}^{n}\left\{\frac{1}{\sigma}K_3 z_i-\frac{1}{\sigma}\zeta^2 K_3 z_i-\frac{1}{\sigma}\zeta K_3^2 z_i\right\}\right)$$

$$\frac{-\partial^2 l}{\partial\rho\partial\lambda_1} =(1-S_i)\left(\sum_{i=1}^{n}\left\{\frac{2}{\sqrt{2\pi u^3}}K_4-\frac{2\lambda_1^2\rho^2}{\sqrt{2\pi u^5}}(\gamma' w_i)^2 K_4+\frac{4\lambda_1\rho}{2\pi u(1+\lambda_1^2)}K_4^2\right\}\right).$$

The arguments of the functions $K_i, i=1,2,3,4$ are suppressed.