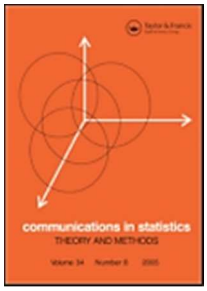


A Unified Approach to Multilevel Sample Selection Models

Journal:	*Communications in Statistics – Theory and Methods*
Manuscript ID:	LSTA-2013-0416.R2
Manuscript Type:	Original Paper
Date Submitted by the Author:	n/a
Complete List of Authors:	Ogundimu, Emmanuel; University of Warwick, Department of Statistics Hutton, Jane; University of Warwick, Department of Statistics
Keywords:	Unit and Item non-response, Closed skew-normal distribution, Selection distribution, Neck Disability Index
Abstract:	We propose a unified approach for multilevel sample selection models using a generalized result on skew distributions arising from selection. If the underlying distributional assumption is normal, then the resulting density for the outcome is the continuous component of the sample selection density, and has links with the closed skew-normal distribution (CSN). The CSN distribution provides a framework which simplifies the derivation of the conditional expectation of the observed data. This generalizes the Heckman's two-step method to a multilevel sample selection model. Finite sample performance of the maximum likelihood estimator of this model is studied through a Monte Carlo simulation.
Note: The following files were submitted by the author for peer review, but cannot be converted to PDF. You must view these files (e.g. movies) online.	
comm in stat.tex	

A Unified Approach to Multilevel Sample Selection Models

E. O. Ogundimu, and J. L. Hutton
{E.O.Ogundimu, J.L.Hutton}@warwick.ac.uk
University of Warwick, Department of Statistics
Coventry, CV4 7AL, UK.

January 17, 2014

Abstract

We propose a unified approach for multilevel sample selection models using a generalized result on skew distributions arising from selection. If the underlying distributional assumption is normal, then the resulting density for the outcome is the continuous component of the sample selection density, and has links with the closed skew-normal distribution (CSN). The CSN distribution provides a framework which simplifies the derivation of the conditional expectation of the observed data. This generalizes the Heckman's two-step method to a multilevel sample selection model. Finite sample performance of the maximum likelihood estimator of this model is studied through a Monte Carlo simulation.

Key Words: Unit and Item non-response; Closed skew-normal distribution; Selection distribution; Neck Disability Index.

1 Introduction

Scores derived from response to questionnaires are widely used in health and social studies to measure aspects of health and well being. This type of study is usually planned as a longitudinal study. Sometimes, the treatment effects at a measurement occasion may be desirable and a cross-sectional view of the data will make two missing data types inevitable- unit and item non-response. Unit non-response occurs when the whole questionnaire is missing for a patient and item non-response occurs where a response has not been provided for a question. The traditional practice is to use weighting adjustments for unit non-response and imputation methods for item non-response. Weighting adjustment means weights are assigned to sample respondents in order to compensate for their systematic differences relative to non-respondents, whereas imputation involves filling in missing values (singly or multiply) to produce a complete data set.

Although these methods have reached a high level of sophistication, they normally assume that the missing data mechanism is MAR (missing at random), an assumption that cannot be verified using the observed data alone. Apart from this, patients may refuse to answer sensitive

questions (e.g. underlying health issues, drug addiction) on a questionnaire for reasons related to the underlying true values for those questions. Thus, when we suspect that non-response may depend on missing values, then a proper analysis will be to model jointly the population of complete data and the non-response process. Sample selection models are therefore viable tools.

Sample selection models arise in practice as a result of the partial observability of the outcome of interest in a study. The 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 model was introduced by Heckman (1976) where he proposed a full maximum likelihood estimation under the assumption of normality. Although the model has its origin from the field of Economics, it has been applied extensively in other fields like Finance, Sociology and Political science, but sparingly in medical research. A prominent application to treatment allocation for patients and links with the skew-normal distribution was discussed by Copas and Li (1997).

The Heckman (1976) selection model, and by extension the Copas and Li (1997) model, was formulated with one-level selection equation. Sometimes, it is necessary to distinguish between the two forms of non-response. This implies that both unit and item non-response simultaneously affect the outcome of interest and both types of non-response are potentially correlated. This distinction can be used to study factors that affect the two non-responses independently and jointly.

Similar models have been discussed in the literature. Poirier (1980) investigated random utility models in which observed binary outcomes do not reflect the binary choice of a single decision-maker, but rather the joint unobserved binary choices of two decision-makers. This model was further developed by Ham (1982). A slight modification of this model was considered in De Luca and Peracchi (2006, 2012) in which an extension of Poirier (1980) model was used to jointly analyze items and unit non-response in a survey data. Further application of multilevel selection models in cross-sectional settings can be found in Bellio and Gori (2003) and Rosenman et. al. (2010).

A general selection distribution for a vector $\mathbf{Y} \in \mathbb{R}^p$ has a PDF (Probability Density Function) f_Y given by

$$f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{Y}^\star}(\mathbf{y}) \frac{P(\mathbf{S}^\star \in \mathbf{C} \,|\, \mathbf{Y}^\star = y)}{P(\mathbf{S}^\star \in \mathbf{C})}, \tag{1}$$

where $\mathbf{S}^\star \in \mathbb{R}^q$ and $\mathbf{Y}^\star \in \mathbb{R}^p$ are two random vectors, and $\mathbf{C}$ is a measurable subset of $\mathbb{R}^q$ (see Arellano-Valle et al. (2006)). Selection distributions depend on the subset $\mathbf{C}$ of $\mathbb{R}^q$. The usual selection subset is defined by

$$\mathbf{C}(\beta) = \{\mathbf{s} \in \mathbb{R}^q \,|\, \mathbf{s} > \beta\},$$

where β is a vector of truncation levels. In particular, equation (1) links selection distributions and skew distributions, and hidden truncation model can be considered a special case of these models. The use of selection distributions for making inference about the population characteristics (without appropriate corrections) in a sample selection framework is not recommended. This can lead to inflated type 1 error, where parameters in the model become significant, when in fact they are not; see for example (Little and Rubin, 2002; Carpenter et al., 2002).

Although the result in (1) is not new, it does not define a complete sample selection density. A sample selection density consists of a continuous component and a discrete component. The main goal of this article is to show that the continuous component of the classical Heckman

selection model and its extensions belong to the closed skew-normal (CSN) distribution and to propose the use of full information maximum likelihood, which is rarely used, for parameter estimation in multilevel sample selection settings. Since the CSN distribution is a well established distribution, moment based estimators and maximum likelihood estimators for any number of selection equations and one outcome equation can readily be defined. This provides a unified method for studying sample selection problems with more than two levels; current econometric literature is restricted to two levels.

We initially address two levels sample selection models, and then indicate the generalization. The article is organized as follows. In section 2, we describe the classical sample selection model and its multilevel extensions. Finite sample performance of the model is studied via Monte Carlo simulation in section 3. We also study the coverage attributes of the sample selection parameters in this section. Application of the model to a real dataset and model diagnostics are given in section 4. An extension of the model with an underlying skew-normal process is shown to have an extended version of the CSN distribution in section 5 and conclusions are given in section 6. The Appendix contains the derivation of three-level sample selection model via a hidden truncation framework.

2 Sample Selection Models

We first describe the classical one-level Heckman sample selection model in this section and its moment based estimator. The continuous component of the sample selection density of the model is linked with the CSN distribution. This link is used to formulate multilevel sample selection model in a straightforward way.

2.1 Copas and Li (1997) Sample selection model

Copas and Li's (1997) paper is probably the first instance where the link between sample selection models and selection (skew) distributions was established. Consider a univariate case of the model given in equation (1) but with the selection subset $C(0)$, which has a simple selection distribution. That is, let $Y_i^\star$ be the outcome variable of interest, assumed linearly related to covariates x_i through the standard multiple regression

$$Y_i^\star = \beta' x_i + \varepsilon_{1i}, \quad i = 1, \ldots, N. \tag{2}$$

Suppose the main model is supplemented by a selection (missingness) equation

$$S_{1i}^\star = \gamma' x_i + \varepsilon_{2i}, \quad i = 1, \ldots, N, \tag{3}$$

where β and γ are unknown parameters and x_i are fixed observed characteristics not subject to missingness, the variance of $\varepsilon_{2i} = 1$ because the variance is not identifiable from sign alone and the variance of $\varepsilon_{1i} = \sigma^2$. Selection is modeled by observing $Y_i^\star$ only when $S_{1i}^\star > 0$ (the 0 threshold is arbitrary since no symmetry is assumed), i.e. we observe $S_i = I(S_{1i}^\star > 0)$ and $Y_i = Y_i^\star S_i$ for $n = \sum_{i=1}^N S_i$ of N individuals. Thus an observation has the conditional density

$$f(y|x, S = 1) = \frac{f(y, S = 1|x)}{P(S = 1|x)} = \frac{f(y|x)P(S = 1|y, x)}{P(S = 1|x)}. \tag{4}$$

The quantity $f(y|x)$ is a proper PDF, with a skewing function $P(S=1|y,x)$, and a normalizing function $P(S=1|x)$. It is straightforward to show that under the additional assumption

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

$$f(y|x,S=1;\Theta) = \frac{\frac{1}{\sigma}\phi\left(\frac{y-\beta'x}{\sigma}\right)\Phi\left(\frac{\gamma'x+\rho\left(\frac{y-\beta'x}{\sigma}\right)}{\sqrt{1-\rho^2}}\right)}{\Phi(\gamma'x)}, \tag{5}$$

(see Copas and Li (1997)), where $\Theta = (\beta, \sigma, \gamma, \rho)$. The parameter $\rho \in$ [-1,1] determines the correlation of $Y_i^\star$ and $S_{1i}^\star$, and hence the severity of the selection process. This equation is not the full sample selection density. The density of the sample selection model is composed of a continuous component corresponding to the conditional density $f(y|x,S=1;\Theta)$ and a discrete component given by $P(S=1|x)$. The marginal distribution of the selection equation determines the nature of the model to be fitted to the discrete component. In Copas and Li (1997) (and Heckman (1976)), a probit model $P(S=s) = \{\Phi(\gamma'x)\}^s\{1-\Phi(\gamma'x)\}^{1-s}$ was used. The log-likelihood function is therefore

$$l(\Theta) = \sum_{i=1}^{n} S_i\Big(\ln f(y_i|x_i,S_i=1;\Theta)\Big) + \sum_{i=1}^{n} S_i(\ln\Phi(\gamma'x_i)) + \sum_{i=1}^{n}(1-S_i)\ln\Phi(-\gamma'x_i). \tag{6}$$

The maximum likelihood estimation based on (6), which is equivalent to equation (14) of Copas and Li's (1997) model, is not robust to deviations from the normality assumption. This prompted Heckman (1979) to develop the two-step estimator (TS). The TS estimator is derived from the conditional expectation of the observed data, and is given by

$$E(Y|x,S^\star>0) = \beta'x + \sigma\rho\Lambda(\gamma'x), \tag{7}$$

where Λ is the inverse Mills ratio. Details of this model, including its sensitivity to collinearity among covariates, can be found in Heckman (1979) and Puhani (2000).

As expected, from Arellano-Valle et al. (2006), equation (5) belongs to the extended skew-normal (ESN) distribution family. To see this, we let $\mu = \beta'x$, $\lambda_0 = \gamma'x/\sqrt{1-\rho^2} \in \mathbb{R}$ and $\lambda_1 = \rho/\sqrt{1-\rho^2} \in \mathbb{R}$ in (5); we then have the PDF written in the four-parameter ESN form given by

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

where λ_0 & λ_1 are shift and shape parameters respectively (see (Azzalini, 1985; Capitanio et al., 2003)). The ESN distribution is a special case of the closed skew-normal (CSN) distribution, which is defined below.

Definition 1 *Consider $p \geq 1$, $q \geq 1$, $\boldsymbol{\mu} \in \mathbb{R}^p$, $\boldsymbol{\nu} \in \mathbb{R}^q$, D an arbitrary $q \times p$ matrix, Σ and Δ positive definite matrices of dimensions $p \times p$ and $q \times q$, respectively. Then the PDF of the CSN distribution is given by:*

$$f_{p,q}(\mathbf{y}) = C\phi_p(\mathbf{y};\boldsymbol{\mu},\Sigma)\Phi_q(D(\mathbf{y}-\boldsymbol{\mu});\boldsymbol{\nu},\Delta), \quad \mathbf{y}\in\mathbb{R}^p, \tag{8}$$

with:

$$C^{-1} = \Phi_q(\mathbf{0}; \boldsymbol{\nu}, \Delta + D\Sigma D'),$$

where $\phi_p(.;\boldsymbol{\eta},\Psi)$, $\Phi_p(.;\boldsymbol{\eta},\Psi)$ are the PDF and CDF (Cumulative Distribution Function) of a p-dimensional normal distribution with mean $\boldsymbol{\eta} \in \mathbb{R}^p$ and $p \times p$ covariance matrix Ψ. We write $\mathbf{Y} \sim CSN_{p,q}(\boldsymbol{\mu}, \Sigma, D, \boldsymbol{\nu}, \Delta)$, if $\mathbf{y} \in \mathbb{R}^p$ is distributed as CSN distribution with parameters $q, \boldsymbol{\mu}, D, \Sigma, \boldsymbol{\nu}, \Delta$. The special case of $\boldsymbol{\nu} = 0$ in (8), gives,

$$f_{p,q}(\mathbf{y}) \quad = 2^q \phi_p(\mathbf{y}; \boldsymbol{\mu}, \Sigma)\Phi_q(D(\mathbf{y} - \boldsymbol{\mu}); \mathbf{0}, \Delta),$$

which is the multivariate skew-normal distribution discussed in Azzalini and Valle (1996). When $q = 1$ and $\boldsymbol{\nu} \neq 0$ in (8), we obtain the multivariate ESN distribution. If $p = 2$ and $q = 1$, a bivariate skew-normal distribution is derived. It is straightforward to see that the PDF in (8) includes the normal distribution as a special case when D and $\boldsymbol{\nu}$ = 0. The properties of CSN distribution that simplify the formulation of multilevel sample selection models include its moment generation function and derivatives of multinormal integrals, and can be found in Gonzalez-Farias et. al. (2004) and Dominguez-Molina et. al. (2004).

Given equation (5), the continuous component of the Heckman (1976) sample selection density, which is essentially an ESN distribution, can be written as

$$f(y|x, S = 1) = \frac{\phi\Big(y; \beta'x, \sigma^2\Big)\Phi\Big(\frac{\rho}{\sigma}\Big(y - \beta'x\Big); -\gamma'x, 1 - \rho^2\Big)}{\Phi\Big(0; -\gamma'x, 1\Big)},$$

which has CSN form

$$(Y|x, S = 1) \sim CSN_{1,1}\Big(\beta'x, \sigma^2, \frac{\rho}{\sigma}, -\gamma'x, 1 - \rho^2\Big).$$

2.2 Multilevel Sample Selection Models

Multilevel sample selection arises when more than one selection process affects the outcome of interest in a study. These models have been discussed in the literature in various forms. Suppose (2) is supplemented with n possible selection processes (not necessarily hierarchical) given as

$$\begin{cases} S^\star_{1i} & = \alpha_1' x_i + \varepsilon_{2i} \\ S^\star_{2i} & = \alpha_2' x_i + \varepsilon_{3i} \\ \vdots & \\ S^\star_{ni} & = \alpha_n' x_i + \varepsilon_{(n+1)i}, \end{cases}$$

where $S_{1i} = I(S^\star_{1i} > 0)$, $S_{2i} = I(S^\star_{2i} > 0)$,..., $S_{ni} = I(S^\star_{ni} > 0)$. The usable observations are the $Y_i = Y_i^\star * S_{1i} * S_{2i} \cdots * S_{ni}$ with density $f(y_i|x_i, S_{1i} = 1, S_{2i} = 1, \ldots, S_{ni} = 1)$. This density is the continuous component of the multilevel sample selection density. The discrete component is determined by the marginal distribution of the selection mechanisms. Unlike in the single selection process, the binary regression is determined by the nature of the selection process.

When multilevel selection models are mentioned in the literature (econometric literature in particular), what usually comes to mind is a two-level selection process. This has an outcome

equation (binary (e.g. Poirier (1980)) or continuous (e.g. Ham (1982))) and two selection equations with trivariate Gaussian error distribution assumption. At the end, a two-level extension of the Heckman two-step method is derived and used to analyze the observed data. However, there are cases where more than two selection processes can affect the outcome of interest. In some of these cases, the selection mechanisms are combined to make the model more manageable and the complicated algebra required to write more than two-level Heckman selection method is avoided in the process.

The combination of selection equations may result in information loss and inability to answer pertinent research questions. Consequently, we propose a link between the continuous component of the multilevel sample selection density and the CSN density (or its extension when the underlying assumption is a symmetric or skew-symmetric distribution) using equation (1). This link is used to generalize the moment based Heckman's two-step method to a multilevel sample selection model. The discrete component is a multivariate probit model whose nature is determined by its level of observability. In addition, since the distribution of the observed cases follow the CSN distribution, full information maximum likelihood can be used for parameter estimation. Without loss of generality, we use a two-level sample selection model to illustrate the unification of multilevel sample selection problems into a distributional framework.

3 Two-level Selection models and Monte-carlo simulation

In this section, we show that the continuous component of a multilevel sample selection density is also a CSN density. This simplifies the derivation of the conditional expectation and variance of the observed data. A simulation study is used to study finite sample properties of the MLEs for the two-level model.

3.1 Two-level Selection models

Suppose equations (2) and (3) are supplemented with an additional selection equation

$$S_{2i}^{\star} = \alpha' x_i + \varepsilon_{3i}, \quad i = 1, \ldots, N. \tag{9}$$

If the regression errors follow a trivariate Gaussian distribution, and we normalize the variances of ε_{2i} and ε_{3i} to one in order to identify the coefficients of the binary response equations, then

$$\begin{pmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \\ \varepsilon_{3i} \end{pmatrix} \sim N_3 \left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \begin{pmatrix} \sigma^2 & \sigma\rho_{12} & \sigma\rho_{13} \\ \sigma\rho_{12} & 1 & \rho_{23} \\ \sigma\rho_{13} & \rho_{23} & 1 \end{pmatrix} \right).$$

Now, (4) can be generalized to a two-level selection model as

$$f(y|x, S_1 = 1, S_2 = 1)) = \frac{f(y|x)P(S_1 = 1, S_2 = 1|y, x)}{P(S_1 = 1, S_2 = 1)}. \tag{10}$$

The marginal distribution of Y is $f(y|x) = \phi(y; \beta' x, \sigma^2)$. Similarly,

$$P(S_1 = 1, S_2 = 1) = 1 - \Phi_2(-\gamma' x, -\alpha' x; \rho_{23}) = \Phi_2(\gamma' x, \alpha' x; \rho_{23}).$$

Using the conditional distribution properties of the normal distribution, $P(S_1 = 1, S_2 = 1|y, x)$ becomes

$$\Phi_2\left(D(y - \beta' x); \begin{pmatrix} -\gamma' x \\ -\alpha' x \end{pmatrix}, \Sigma_2^\star\right),$$

where $D = (\rho_{12}/\sigma, \rho_{13}/\sigma)'$ and $\Sigma_2^\star = \begin{pmatrix} 1 - \rho_{12}^2 & \rho_{23} - \rho_{12}\rho_{13} \\ \rho_{23} - \rho_{12}\rho_{13} & 1 - \rho_{13}^2 \end{pmatrix}$. When appropriate substitutions are made in equation (10) and the model is standardized, the resulting density becomes:

$$\frac{\phi(y; \beta' x, \sigma^2)\Phi_2\left(\frac{\gamma' x + \rho_{12}(\frac{y-\beta' x}{\sigma})}{\sqrt{1-\rho_{12}^2}}, \frac{\alpha' x + \rho_{13}(\frac{y-\beta' x}{\sigma})}{\sqrt{1-\rho_{13}^2}}; \frac{\rho_{23}-\rho_{12}\rho_{13}}{\sqrt{1-\rho_{12}^2}\sqrt{1-\rho_{13}^2}}\right)}{\Phi_2(\gamma' x, \alpha' x; \rho_{23})}. \tag{11}$$

Model (11) includes the three missing data mechanisms discussed in the missing data literature (see for example, Rubin (1976), Diggle and Kenward (1994) and Little and Rubin (2002)). If the non-intercept terms in γ and α, as well as ρ_{12} and ρ_{13} are zero in (11), the data are MCAR (missing completely at random). That is, the non-response processes for both the unit and item are independent both of observed data and of unobservable parameters of interest. If ρ_{12} and ρ_{13} are zero in (11) the data are MAR, and valid inference about the conditional distribution of Y given x can be made when adjustment for missing data is made using covariates on complete cases. The difference between MCAR and MAR missing data mechanism is that there are no predictors of missingness in the former, since the realized sample is a random sample from the fully responding units. If ρ_{12} or ρ_{13} is different from zero in (5), then the missing data are MNAR. In this case, the missing data process is said to be informative or non-ignorable, as valid inference depends on adequate adjustment for the selection process.

Equation (11) is equivalent to equation (10) given in Ahn (1992). It has a CSN density representation given by:

$$\frac{\phi(y; \mu_1, \sigma^2)\Phi_2(D(y - \mu_1); \boldsymbol{\nu}, \Sigma_2^\star)}{\Phi_2(\mathbf{0}; \boldsymbol{\nu}, \Sigma_2)}, \tag{12}$$

where $\mathbf{0} = (0, 0)'$, $\boldsymbol{\nu} = (-\mu_2, -\mu_3)'$, $\mu_1 = \beta' x$, $\mu_2 = \gamma' x$ and $\mu_3 = \alpha' x$. It is easy to see that $\Sigma_2 = \Sigma_2^\star + D\sigma^2 D'$.

Insert Figure 1 about here:

A plot of the PDF given by (12) is shown in Figure 1. The 'CSN(Normal)' represents the normal distribution as a special case of the CSN distribution. The parameters are $\mu_1 = 1, \sigma = 1$, $D = (0, 0)'$, $\boldsymbol{\nu} = (0, 0)'$, and $\Sigma_2^\star = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. The 'CSN(Skew-normal)' is a skew-normal equivalence of CSN distribution with $D = (1, 2)'$, and other parameters kept as in the normal case. The more general form of the CSN is marked as 'CSN(General)' with $\boldsymbol{\nu} = (-2, 4)'$ and other parameters kept as in the skew-normal. The more general CSN can be more or less skew depending on its parameters. This general form is the structure of the sample selection model.

In general, the continuous component of a multilevel sample selection density is a CSN density. In the bivariate case, it is given by equation (11). The discrete component of the log-likelihood function can be described by a bivariate probit model since the marginal distribution

of the selection equation is a bivariate normal distribution. Roughly speaking, the normalizing constant of the continuous component will turn out to be the observed component of the discrete process, which is $\Phi_2(\gamma'x, \alpha'x; \rho_{23})$ in this case. There are various bivariate models that fit into this framework depending on the assumption about the observability of S_1 and S_2. This ranges from separate observability of both S_1 and S_2 to observability of S_1S_2 only (see Meng and Schmidt (1985)).

The fact that sample selection density consists of two 'separate' components, and the continuous component belongs to an established family of skew distributions, makes the extension of the two-level sample selection problem into a multilevel sample selection problem straightforward. For instance, in the three-level sample selection problem, the continuous component of the sample selection density is a CSN density with dimensions p=1 and q=3 (see the derivation in the Appendix). This separation ensures that the continuous component of the density has no link with the hierarchical nature of the selection process. If the selection process is hierarchical, there is only one possible model for the discrete component regardless of the number of selection equations. This model is a multivariate probit model with full observability. As the number of selection equations increases however, the number of possible non-hierarchical multivariate probit models with sample selection increases. In practice, subject matter expertise can help narrow down the models to a manageable size when applied to a specific data problem. Extension of the continuous component to multivariate measurements is also straightforward.

3.2 Moments and Maximum Likelihood estimator for multilevel selection model

The fact that the continuous component of the two-level (hence multilevel) sample selection density is from a well established CSN family results in a straightforward formula for its mean and variance. These models turn out to be generalizations of Heckman's two-step method.

The mean is then given by:

$$E(Y|x, S_1^\star > 0, S_2^\star > 0) = \beta'x + \sigma\rho_{12}\Lambda_1(\theta) + \sigma\rho_{13}\Lambda_2(\theta), \tag{13}$$

where

$$\Lambda_1(\theta) = \frac{\phi(\gamma'x)\Phi\left(\frac{\alpha'x-\rho_{23}\gamma'x}{\sqrt{1-\rho_{23}^2}}\right)}{\Phi_2(\gamma'x, \alpha'x; \rho_{23})} \quad \text{and} \quad \Lambda_2(\theta) = \frac{\phi(\alpha'x)\Phi\left(\frac{\gamma'x-\rho_{23}\alpha'x}{\sqrt{1-\rho_{23}^2}}\right)}{\Phi_2(\gamma'x, \alpha'x; \rho_{23})}.$$

$\Lambda_1(\theta)$ and $\Lambda_2(\theta)$ are the bivariate inverse Mills ratio. This equation extends Heckman's two-step method (see equation (7)) to two-level selection problems. A standard bivariate probit model is fitted depending on what is assumed about the observability of S_1 and S_2 and γ & α are estimated. These are used to construct $\Lambda_1(\hat{\theta})$ and $\Lambda_2(\hat{\theta})$ for cases with S_1 and S_2 greater than zero. These quantities are taken as additional covariates in (13) and fitted by least squares. The coefficient of the additional covariates give estimates of $\sigma\rho_{12}$ and $\sigma\rho_{13}$ respectively.

To visualize the impact of the correlation (ρ_{23}) between the selection equations on the outcome model, we plot the second component of the expectation $(E(Y|x, S_1^\star > 0, S_2^\star > 0)$-$\beta'x)$ as a function of $\rho_{12}\gamma'x + \rho_{13}\alpha'x$, the combined mean of the selection variables. We fix ρ_{12} and ρ_{13} to be 0.7 and 0.5 respectively for values of ρ_{23}= {0, 0.3, 0.5 and 0.7}. The standard deviation, σ, simply scales the correction factors.

From Figure 2a, the conditional expectation will be overestimated if ρ_{23} is greater than zero and we assume it to be zero, for negative values of the combined selection predictors ($\rho_{12}\gamma'x +$

$\rho_{13}\alpha' x$). The difference diminishes as the linear predictors become positive. In addition, Figure 2a shows that the bivariate inverse Mills ratio can be linear over a wide range of its support. In practice, an exclusion restriction is usually imposed whereby there are variables in the selection equations that does not appear in the outcome model or vice-versa. This ensures that we do not rely on the non-linearity of the bivariate inverse Mills ratio for model identifiability.

Sometimes, the marginal effect of the covariates (x_i) on the outcome Y_i in the observed sample may be of interest. For the extended Heckman two-step method given by (13), the effect consists of three components- the direct effect of the covariates on the mean of Y_i which is captured by β and the indirect effects of the covariates in the two selection equations. The marginal effect is given by:

$$\begin{aligned}\frac{\partial}{\partial x_i}E(Y|x, S_1^\star > 0, S_2^\star > 0) =&\beta' - \sigma\rho_{12}\Big[\Big(\frac{\gamma' - \rho_{23}\alpha'}{\sqrt{1-\rho_{23}^2}}\Lambda(\theta)\Big) - \alpha'(\alpha' x)\Lambda_2(\theta)\Big] \\ &-\sigma\rho_{13}\Big[\gamma'\Lambda_1(\theta)\Lambda_2(\theta) + \big(\Lambda_2\Big(\theta\Big)\big)^2\Big],\end{aligned} \tag{14}$$

where $\Lambda(\theta) = \phi(\gamma' x)\phi\Big(\frac{\alpha' x - \rho_{23}\gamma' x}{\sqrt{1-\rho_{23}^2}}\Big)\Big/\Phi_2(\gamma' x, \alpha' x; \rho_{23})$. Figure 2b shows that the conditional marginal effect of x_i on Y_i will be underestimated if ρ_{23} is greater than zero and we assume it to be zero, for negative values of the combined selection predictors ($\rho_{12}\gamma' x + \rho_{13}\alpha' x$). These effects also dies out as the predictors becomes positive.

Insert Figure 2 about here:

A consistent estimate of the variance can be derived from the conditional variance given by:

$$\begin{aligned}\text{var}(Y|x, S_1^\star > 0, S_2^\star > 0) =&\sigma^2 - \sigma^2\rho_{12}^2(\gamma' x)\Lambda_1(\theta) - \sigma^2\rho_{13}^2(\alpha' x)\Lambda_2(\theta) \\ &+\frac{\phi_2(\gamma' x, \alpha' x; \rho_{23})}{\Phi_2(\gamma' x, \alpha' x; \rho_{23})}\Big[2\sigma\rho_{12}\sigma\rho_{13} - \rho_{23}(\sigma^2\rho_{12}^2 + \sigma^2\rho_{13}^2)\Big] \\ &-\Big(\sigma\rho_{12}\Lambda_1(\theta) + \sigma\rho_{13}\Lambda_2(\theta)\Big)^2 \\ =&\sigma^2 + \upsilon.\end{aligned} \tag{15}$$

The error terms of the selected sample are heteroscedastic. A generalization of Heckman's estimator for σ^2 given by

$$\sigma^2 = (S - \sum \hat{\upsilon}_i)/N_2,$$

where S is the sum of squared residuals from the second-step regression, N_2 is the size of the complete cases, and υ_i equals $\hat{\upsilon}_i$ after parameter estimates have been substituted for their true values, can be used to get consistent estimator for σ^2.

The log-likelihood function takes the form:

$$\begin{aligned}l(\Omega) = \sum_{i=1}^{N}\Bigg(&S_{1i}S_{2i}\Big[\ln f(y_i|x_i, S_{1i} = 1, S_{2i} = 1)\Big] + S_{1i}S_{2i}\Big[\ln\Phi_2(\gamma' x_i, \alpha' x_i; \rho_{23})\Big] \\ &+S_{1i}(1 - S_{2i})\Big[\ln\Phi_2(\gamma' x_i, -\alpha' x_i; -\rho_{23})\Big] + (1 - S_{1i})S_{2i}\Big[\ln\Phi_2(-\gamma' x_i, \alpha' x_i; -\rho_{23})\Big] \\ &+(1 - S_{1i})(1 - S_{2i})\Big[\ln\Phi_2(-\gamma' x_i, -\alpha' x_i; \rho_{23})\Big]\Bigg),\end{aligned} \tag{16}$$

where $\Omega = (\beta, \sigma, \gamma, \alpha, \rho_{12}, \rho_{13}, \rho_{23})$.

3.3 Monte Carlo Simulation

The finite-sample performance of the models in section 3.2 are studied here. We consider maximum likelihood estimator ((16)) which we referred to as the multilevel selection normal model (MSNM). The moment based multilevel two-step (MTS) estimator given by (13) and multiple imputation (MI) of missing data are also considered. The outcome equation is $Y_i^\star = 0.5 + 1.5x_i + \varepsilon_{1i}$, where $x_i \overset{iid}{\sim} N(0,1)$ and $i = 1, \ldots, N = 1000$. The two-level selection equations are given as $S_{1i}^\star = 1 + 0.4x_i + 0.3w_i + \varepsilon_{2i}$ and $S_{2i}^\star = 1 + 0.6x_i + 0.7w_i + \varepsilon_{3i}$, where $w_i \overset{iid}{\sim} N(0,1)$. The error terms are generated from a trivariate normal distribution with covariance matrix $\Sigma = \begin{pmatrix} 1 & 0.7 & 0.5 \\ 0.7 & 1 & \rho_{23} \\ 0.5 & \rho_{23} & 1 \end{pmatrix}$. This construction implies that the variance of the outcome model is 1. We take $\rho_{23} = \{0, 0.3, 0.5, 0.7\}$ to assess its effect on the parameters of the outcome model, and 1000 replications are used in all the simulation.

We only observe values of $Y_i^\star$ when both $S_{1i}^\star$ and $S_{2i}^\star$ are greater than zero. With this representation, roughly 30% of the observations were censored, and we allow for full observability in the bivariate process. For the MI, x_1 and x_2 are used as covariates for the imputation model and the regression model of interest only includes x_1. We use 10 imputations for each of the samples generated.

Insert Tables 1 and 2 about here:

Tables 1 and 2 show the results of the simulation when the correlation between the selection equations are low and moderately high. The MSNM preforms better than the MTS as expected, although at a higher computational cost. When $\rho_{23} = 0$, 45% of the simulation results were discarded under the MTS method because the values of ρ_{12} and ρ_{13} are outside their admissible range. Roughly, 8% and 1.9% of the simulation results are outside the admissible range when ρ_{23} = 0.3 and 0.5 respectively. However, all the values of ρ_{12} and ρ_{13} are in the interval [-1,1] when $\rho_{23} = 0.7$. A noticeable impact of the values of ρ_{23} is that the intercept of both the MSNM and the MTS models consistently have lower bias as ρ_{23} increases. The reverse is the case with MI, and the procedure is clearly inappropriate for multilevel sample selection models.

Arguably, the log-likelihood function of the MSNM models are not globally concave. The initial values have to be chosen carefully for the model to converge, and not to converge to a local maximum. This was accomplished in the simulation by choosing the initial values in the neighborhood of the values used for the data generation. In addition, the correlation parameters (ρ_{12} and ρ_{13}) were constrained to be within [-1,1]. This results in further flexibility which cannot be achieved under the MTS method since the correlations are tied to the variance, which in turn is heteroscedastic.

We also assess the coverage probabilities of Wald test and LRT for the hypothesis of selection bias, that is the hypothesis $H_0 : \rho_{12} = \rho_{13} = 0$. The data are simulated as described earlier with 1000 replications but with sample sizes N = 500 and 1000. We take 0.05 as the nominal significance level and the multivariate Wald test is as described on p.78 of Enders (2010).

Insert Table 3 about here:

Table 3 shows the results of the coverage probabilities for fixed values of ρ_{23}. The LRT maintains the nominal coverage when $\rho_{23} = 0$ and N =1000. The performance of the LRT is better than the Wald test both in the medium and large sample sizes.

4 Application to MINT Trial

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

The main goal of the study is to determine if there is any meaningful difference in the treatments. 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 assess pain-related activity restrictions in 10 areas including personal care, lifting, sleeping, driving, concentration, reading and work and result in a score between 0 and 50. It was developed in 1989 by Howard Vernon as a modification of the Oswestry Low Back Pain Disability Index. The NDI has been shown to be reliable and valid (Vernon and Mior, 1991), hence its use as a standard instrument for measuring self-rated disability due to neck pain by clinicians and researchers.

There are 599 patients with a total of 1934 measurements and 372 (62%) patients have complete observations (i.e. scores at all measurement occasions). Patients were allocated equally to the treatment of interest, physiotherapy, and the control, 'usual advice' contained in the Whiplash Book (Burton et. al., 2001)). The mean age is approximately 41 years with range 18 to 78 years. The fact that the responses were derived from the use of a 10-item questionnaire posed several challenges. One of the challenges is item and unit non-response and dropout with time. The question on driving has the highest number of missingness among the items on the NDI scale. A possible explanation for this is that some of the patients are not driving, and the question is not open-ended to avoid skipping it when it is not relevant. Regardless of this possibility, clinicians are interested in knowing if the question motivated the respondents to fail to answer other questions on the scale. We focus on the measurement at months 12, where more data are missing on this question than other measurement occasions, and use the two-level sample selection model to jointly analyze the non-response processes.

In line with the study design, 599 patients are expected to return the questionnaire. After removing covariates with missing values, the sample size consists of 567 patients. Of the patients, 79 patients returned the questionnaire blank (genuine unit non-response). Vernon recommended that patients with only 2 missed items should be considered complete, with mean imputation used for adjustment. There are 45 patients in this category, making a total of 101 unit non-respondents. The unit is first observed before the question of interest, the driving question, is answered. We have 61 patients who did not respond to this question. Of course, unit non-respondents are also item non-respondents, making patients with item non-response to be effectively 185 patients.

The questions to answer are whether unit and item non-response are related and whether both are related to the outcome of interest. To answer the first question, we consider a bivariate probit model with sample selection for unit and item non-response and estimate the correlation parameter. This model is also used to identify possible predictors of non-response in the unit and item equations. Unlike the discrete component of (16), the log-likelihood function for a

bivariate probit sample selection model is

$$l(\gamma,\alpha,\rho_{23}) = \sum_{i=1}^{N}\Bigg(S_{1i}S_{2i}\Big[\ln\Phi_2(\gamma' x_i, \alpha' x_i;\rho_{23})\Big] + S_{1i}(1-S_{2i})\Big[\ln\Phi_2(\gamma' x_i, -\alpha' x_i;-\rho_{23})\Big] + (1-S_{1i})\Big[\ln\Phi(-\gamma' x_i)\Big]\Bigg). \tag{17}$$

A simulation study (not reported here) showed that if the selection model (17) is correctly specified, correct specification includes imposing exclusion restrictions on the covariates in the two equations of the unit and item, the model parameters are consistent. In addition, one can test the hypothesis of conditional independence between unit and item non-response using a Wald test or likelihood ratio test. To fit the two-step method to a two-level selection problem with sample selection between unit and item non-response, the probit model needed in the bivariate inverse Mills ratio is the one given by equation (17). Patients may feel that the treatment they received is of no benefit, and thereby discontinue treatment. This will lead to unit non-response rather than item non-response. We therefore include treatment indicator ('physio') as a possible predictor of unit non-response. Other baseline variables in the data, e.g. ethnicity, employment status and region could not be included in the data analysis because they are subjected to various degrees of missingness.

Insert Table 4 about here:

The results in Table 4 show that there is conditional independence between unit and item non-response for the scores using Wald test. This was further affirmed by the likelihood ratio test that compares the maximized values of the log-likelihood function in Table 4 with the sum of the log-likelihoods for two simple probit models for unit and item non-response separately (P-value=0.262). Increasing age is associated with increase in both types of non-response. The driving question is less likely to be answered by women. Fewer older women than older men have driving licences.

Insert Table 5 about here:

Table 5 contains the results of the MSNM models with $\rho_{23} \neq 0$ & $\rho_{23} = 0$, and the extended Heckman's 1979 model (MTS). The 'wad' variable stands for Whiplash Associated Disorder (Whiplash describes both the mechanism of injury and the symptoms caused by that injury). It is a categorical variable with grade 3 the most severe neck disability and grade 1 the least, recorded before the patient enters the study. The 'baseline' are the NDI scores of patients at the start of the study. We include 'wad' and 'baseline' variables in the outcome model to assess their relationship with the NDI scores at month 12. Age and baseline are positively associated with the outcome.

Insert Figure 3 about here:

The results of MSNM model with $\rho_{23} \neq 0$ are reported for completeness sake. This result also strengthen the earlier conclusion about conditional independence of unit and item non-response reported in Table 4 (p-value = 0.243 for the LRT). Under the model with conditional independence ($\rho_{23} = 0$), separate probit models are used for unit and item non-response for the discrete components of the log-likelihood function given in (16). The standard errors for the parameter estimates in the MTS method are obtained using 200 non-parametric bootstrap samples. The MTS estimates show a significant effect of sex on NDI score, indicating the inadequacy of the method.

We examine the normalized profile log-likelihood of the correlation parameters from our preferred model (MSNM with $\rho_{23} = 0$). Figure 3 shows the plots of these profiles, with the profile for ρ_{23} obtained from the bivariate probit model in Table 4. The profiles for ρ_{13} and ρ_{23} are well behaved, but the profile of ρ_{12} is asymmetric and does not decay to zero. Consequently, the use of standard errors to produce interval inference can sometimes be misleading. In the case of the NDI scores, the Wald standard errors and the profile interval resulted in the same conclusion for ρ_{12} and we retained the former in Table 5.

In practice, sample selection models are used for explanation purposes rather than prediction. However, it is possible to validate the model by the validation of its components. That is, the outcome and the selection components are validated separately. The approach we use for validation is by plugging in the parameter estimates from our preferred model into equation (13) to obtain fitted scores. Figure 4 shows the residual plot and the histogram of the residuals. The parameter estimates are unbiased but the errors are heteroscedastic. Factors that could be responsible for this include omission of variables that can possibly interact with the variables included in the model. We are unable to include these covariates because of severe missingness. In addition, the scores are bounded and we recommend transformation of the outcome using, for example, the logistic transformation before the application of the multilevel model.

Insert Figure 4 about here:

5 Multilevel Skew-normal Selection model

Multilevel sample selection models are generally identifiable but the price to pay for the identifiability is possibility of model misspecification. Although sensitivity analysis on the model parameters is justifiable, the use of a range of plausible parametric representations, especially those having the normal distribution as special case, is preferred. The two most common deviations from normality are nonnormal peakedness and asymmetry. As noted by Mudholkar and Hutson (2000), the effects of asymmetry on the normal theory methods are generally more serious than those of the nonnormal peakedness. A similar model to the one given by equation (11) can be constructed with an underlying multivariate skew-normal distribution (Azzalini and Valle, 1996; Azzalini and Capitanio, 1999) as follows.

Suppose we have a joint process where the outcome Y is skewed and the two selection models have skewness parameters equals to zero. The joint distribution can be written in a CSN form. That is,

$$\begin{pmatrix} Y \\ S_1 \\ S_2 \end{pmatrix} \sim CSN_{3,1}\Big\{\boldsymbol{\mu} = (\beta'x, \gamma'x, \alpha'x), \Sigma = \begin{pmatrix} \sigma^2 & \sigma\rho_{12} & \sigma\rho_{13} \\ \sigma\rho_{12} & 1 & \rho_{23} \\ \sigma\rho_{13} & \rho_{23} & 1 \end{pmatrix}, D = (\lambda/\sigma, 0, 0), \nu = 0, \Delta = 1\Big\}.$$

The conditional probability $P(S_1 = 1, S_2 = 1|y, x)$ is

$$\begin{aligned} CSN_{2,1}\Big\{\boldsymbol{\mu} = &\Big[\gamma'x + \rho_{12}\Big(\frac{y-\beta'x}{\sigma}\Big), \alpha'x + \rho_{13}\Big(\frac{y-\beta'x}{\sigma}\Big)\Big]', \Sigma = \Sigma_2^{\star}, D^{\star} = (0,0)', \\ &\nu = \lambda\Big(\frac{y-\beta'x}{\sigma}\Big), \Delta = 1\Big\}, \end{aligned} \tag{18}$$

where $\Sigma_2^{\star}$ is as defined in section 3.1. Since the skewness parameters in (18) are zero, it reduces to the normal distribution given in equation (11). Similarly, the marginal selection process

$P(S_1 = 1, S_2 = 1)$ has a bivariate skew-normal distribution

$$SN_2\left\{\begin{pmatrix}\gamma'x\\ \alpha'x\end{pmatrix}, \begin{pmatrix}1 & \rho_{23}\\ \rho_{23} & 1\end{pmatrix}, \begin{pmatrix}\frac{-\lambda(\rho_{12}-\rho_{13}\rho_{23})}{(1-\rho_{23}^2+\lambda[\rho_{12}^2+\rho_{13}^2-2\rho_{12}\rho_{13}\rho_{23}])}\\ \frac{-\lambda(\rho_{13}-\rho_{12}\rho_{23})}{(1-\rho_{23}^2+\lambda[\rho_{12}^2+\rho_{13}^2-2\rho_{12}\rho_{13}\rho_{23}])}\end{pmatrix}\right\}.$$

The continuous component of this model has density

$$\frac{\frac{2}{\sigma}\phi\left(\frac{y-\beta'x}{\sigma}\right)\Phi\left(\frac{\lambda(y-\beta'x)}{\sigma}\right)\Phi_2\left(\frac{\gamma'x+\rho_{12}(\frac{y-\beta'x}{\sigma})}{\sqrt{1-\rho_{12}^2}}, \frac{\alpha'x+\rho_{13}(\frac{y-\beta'x}{\sigma})}{\sqrt{1-\rho_{13}^2}}; \frac{\rho_{23}-\rho_{12}\rho_{13}}{\sqrt{1-\rho_{12}^2}\sqrt{1-\rho_{13}^2}}\right)}{P(S_1 = 1, S_2 = 1)}. \quad (19)$$

The normalizing constant $P(S_1 = 1, S_2 = 1)$ determines the nature of the binary regression model for the discrete process, which is a bivariate binary regression model with skew-normal link. The correlations and the skewness parameter λ contribute to the skewness in the model.

If $\lambda = 0$ in (19), then equation (11) is recovered. This model can be extended to more than two-level selection processes. The central model will be an extended version of the CSN distribution and an appropriate binary regression model depending on the degree of observability of the outcome. The required extended CSN distribution can be derived by adding p-dimensional random vector from the multivariate skew-normal distribution to an independent q-dimensional random vector from the truncated multivariate normal distribution.

6 Concluding Remarks

Classical sample selection models and their multilevel counterparts have been in the literature for some time. We have therefore, not claimed any originality in this proposal. What we have done however, is to unify two streams of literature on this matter and propose a framework for easy generalization to any number of selection equations in a straightforward manner, and which to the best of our knowledge has not been proposed elsewhere. We also demonstrated the advantage of the full information maximum likelihood and the power of the LRT in two-level sample selection models, which are rarely used in practice.

The econometric literature usually assumes a joint Gaussian error distribution for the outcome and the selection equations. By using properties of truncated normal distribution, the moment-based estimators of sample selection model is derived. On the other hand, the statistics literature contains studies on the closed skew-normal (CSN) distribution. Although the CSN distribution is elegant and a generalization of the Azzalini skew-normal distribution, its use is limited in likelihood based methods due to identifiability issues. When used in sample selection framework, the CSN becomes identifiable due to extra information from the selection process.

A simulation study was conducted to assess the performance of the likelihood (MSNM) and the moment (MTS) based estimators of the two-level sample selection model. The performance of multiple imputation (MI) was also investigated. The MSNM approach outperformed the MTS method and the variance in the MTS method are consistently larger. The impact of the correct specification of the association parameter (ρ_{23}) between the selection equations was also examined. As ρ_{23} becomes larger, the bias in σ also increases. The discarded results when $\rho_{23} = 0$ under the MTS method can suggest that likelihood estimators also have problems at the boundary of the parameter space. The MI approach failed in this setting and it is not

recommended generally for cross-sectional MNAR data. In addition, we showed that LR tests generally have better coverage than the multivariate Wald test for the MSNM model. The application demonstrated the value of the MSNM in separating factors associated with item and unit non-response, and in providing more accurate estimates of the outcome model.

On model identifiability, the Fisher information matrix for two selectivity criteria was derived in Ahn (1992) and was shown to be nonsingular. Even in the more than two-level cases, we expect the model to be identifiable. The continuous component (CSN) would necessarily be non-identifiable in general, and the use of a selection distribution in sample selection framework is incorrect. The model will become identifiable from the additional information from the discrete component. The central advantage of the CSN distribution in sample selection framework with symmetric and asymmetric distributions cannot be overemphasized. We have shown in (19) that the model is an extension of the CSN distribution.

The main advantage attributed to the moment estimator of sample selection model is its robustness to deviations from normality assumption. But the problem of multicollinearity outweigh this robustness gain as noted by Puhani (2000). As the number of selection equations increase, the construction of appropriate sets of covariates for exclusion restriction becomes difficult. Considering the simplicity of the CSN density, it is easier to construct a likelihood function and optimize it. Computational complexity in the likelihood framework is no longer a set back as R software, for instance, have packages that can evaluate multidimensional normal integrals up to 1000 components (Genz and Bretz , 2009). The generalization we proposed has better prospects in observational studies and surveys where multilevel selection processes need to be analyzed jointly and with information on likely variables that could potentially be responsible for a particular selection process included in the analysis.

Acknowledgments

This work was supported by an Engineering and Physical Sciences Research Council grant, for the Centre for Research in Statistical Methodology (CRiSM) EP/D002060/1, which provided a studentship to EO. We thank Prof. Sallie Lamb and the MINT trial team for the permission to use the Neck disability index data.

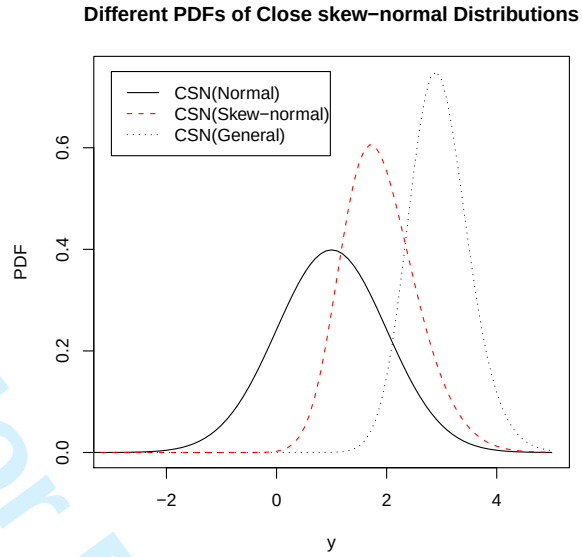
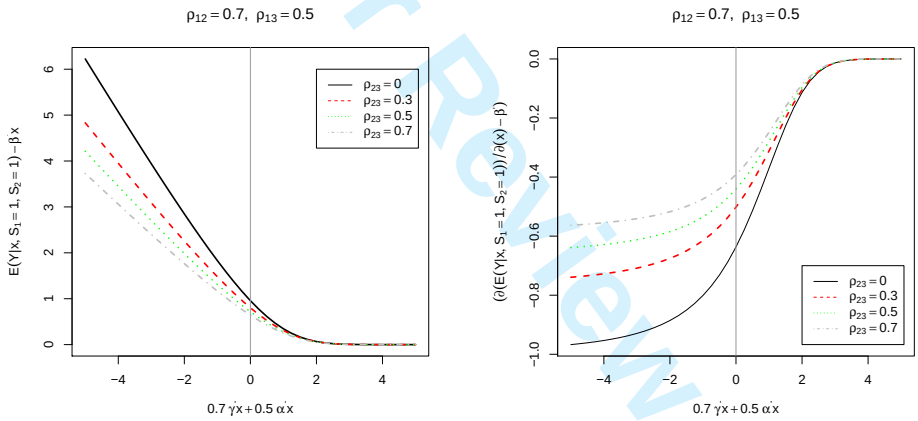


Figure 1: Comparison of Close skew-normal densities



(a) (b)

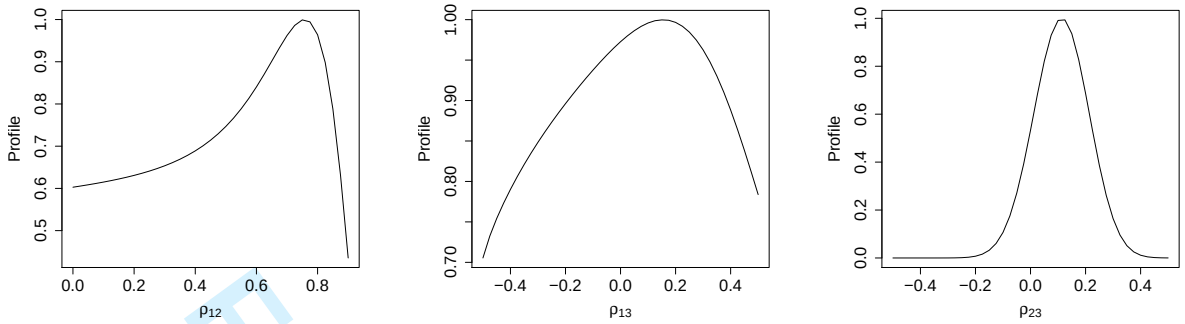
Figure 2: Plots of correction factor and marginal effect for different values of correlation between the selection equations ρ_{23}: (a) Correction factor; (b) Marginal effects.

Table 1: Simulation results (multiplied by 10,000) for zero and low correlation between the selection equations.

		Bias			MSE		
		MSNM	MTS	MI	MSNM	MTS	MI
$\rho_{23}=0$	β_0	204	948	2987	46	240	914
	β_1	-74	-337	-6059	20	49	3685
	σ	-24	242	3268	14	75	1084
	γ_0	42	64		26	28	
	γ_1	41	56		27	30	
	γ_2	19	-28		26	27	
	α_0	52	88		34	34	
	α_1	44	80		34	34	
	α_2	131	205		40	38	
	ρ_{12}	-823	-5412		678	5967	
	ρ_{13}	137	1915		316	834	
	ρ_{23}	-2	32		55	49	
$\rho_{23}=0.3$	β_0	79	103	3203	52	258	1048
	β_1	-80	-330	-6534	21	41	4283
	σ	-32	392	3520	15	121	1255
	γ_0	42	30		27	27	
	γ_1	35	30		28	29	
	γ_2	29	12		27	26	
	α_0	53	86		35	36	
	α_1	46	58		32	33	
	α_2	117	130		39	37	
	ρ_{12}	-923	-5812		725	6904	
	ρ_{13}	60	1442		269	628	
	ρ_{23}	13	26		43	39	

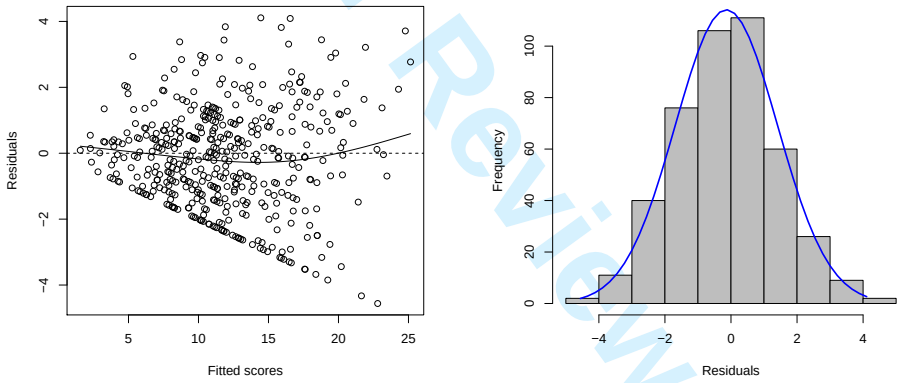
Table 2: Simulation results (multiplied by 10,000) for moderate to high correlation between the selection equations.

		Bias			MSE		
		MSNM	MTS	MI	MSNM	MTS	MI
$\rho_{23} = 0.5$	β_0	28	18	3344	61	656	1140
	β_1	-94	-20	-6845	23	91	4699
	σ	-37	370	3666	17	123	1359
	γ_0	47	51		29	29	
	γ_1	3	11		30	32	
	γ_2	21	-26		25	26	
	α_0	38	69		33	32	
	α_1	20	46		32	33	
	α_2	56	69		36	36	
	ρ_{12}	-1352	-2305		1013	6947	
	ρ_{13}	99	-489		253	813	
	ρ_{23}	9	14		32	28	
$\rho_{23} = 0.7$	β_0	20	-7	3441	53	560	1206
	β_1	-74	-14	-7169	21	77	5154
	σ	-51	2099	3803	16	1366	1461
	γ_0	40	47		26	26	
	γ_1	49	52		29	29	
	γ_2	54	37		27	26	
	α_0	53	58		34	35	
	α_1	62	62		32	33	
	α_2	39	30		37	35	
	ρ_{12}	-830	-1890		565	7600	
	ρ_{13}	-150	-618		223	445	
	ρ_{23}	24	20		19	19	



(a) (b) (c)

Figure 3: Normalized Profile likelihoods of correlation parameters: (a) Unit non-response (ρ_{12}); (b) Item non-response (ρ_{13}); (c) Correlation between unit and item non-response (ρ_{23}).



(a) (b)

Figure 4: (a) Residual plot for the observed scores; (b) Histogram of residuals with Normal Curve.

Table 3: Empirical significance levels (as %) of the tests of selection bias for the nominal significance level $\alpha = 0.05$ in the MSNM model.

	$N = 500$		$N = 1000$	
ρ_{23}	Wald	LRT	Wald	LRT
0.0	4.7	5.4	5.2	5.0
0.3	4.2	5.3	4.6	4.9
0.5	3.6	4.4	4.4	4.9
0.7	3.5	4.6	4.3	4.7

Table 4: Probit model for dropout at months 12.

	Missing at 8 months					
	Bivariate Probit			Individual Probit		
	Estimate	S.E.	p-value	Estimate	S.E.	p-value
int(u)[a]	0.397	0.243	0.103	0.403	0.242	0.096
age	0.016	0.005	0.004	0.016	0.005	0.004
sex(f)	0.102	0.138	0.462	0.100	0.138	0.467
physiotherapy[b]	0.061	0.133	0.648	0.050	0.133	0.710
int(i)[c]	0.190	0.233	0.416	0.196	0.234	0.403
age	0.014	0.005	0.010	0.013	0.005	0.012
sex(f)	0.635	0.136	0.000	0.632	0.136	0.000
ρ_{23}	0.113	0.101	0.260			
Loglik	-439.321			-217.520[d]		-222.431[e]

[a]Intercept for unit non-response.
[b]Treatment received. 'Usual advice' was used as reference.
[c]Intercept for item non-response.
[d]item non-response.
[e]unit non-response.

Table 5: Fit of Two-level selection models ($\rho_{23} \neq 0$) & $\rho_{23} = 0$), and extended Heckman's two-step method to the NDI scores at 12 months.

	MSNM($\rho_{23} \neq 0$)			MSNM($\rho_{23} = 0$)			MTS[a]		
	Estimate	S.E.	p-value	Estimate	S.E	p-value	Estimate	S.E	p-value
				Selection Equations					
int(u)[b]	0.359	0.242	0.139	0.365	0.242	0.131	0.403	0.242	0.096
age	0.016	0.005	0.003	0.016	0.005	0.003	0.016	0.005	0.004
sex(f)	0.087	0.137	0.528	0.086	0.137	0.529	0.100	0.138	0.467
physiotherapy	0.129	0.138	0.352	0.114	0.138	0.406	0.050	0.133	0.710
int(i)[c]	0.180	0.235	0.443	0.188	0.235	0.424	0.196	0.234	0.403
age	0.014	0.005	0.010	0.014	0.005	0.011	0.013	0.005	0.012
sex(f)	0.636	0.136	0.000	0.633	0.136	0.000	0.632	0.136	0.000
ρ_{23}	0.118	0.101	0.241						
				Outcome Equation					
intercept	-7.690	3.067	0.013	-7.573	3.148	0.017	-5.960	1.784	0.001
age	0.141	0.035	0.000	0.141	0.036	0.000	0.122	0.028	0.000
sex(f)	1.500	1.294	0.247	1.437	1.320	0.277	1.958	0.781	0.006
physiotherapy	0.181	0.758	0.812	0.162	0.759	0.831	0.193	0.821	0.814
baseline	0.536	0.068	0.000	0.538	0.068	0.000	0.500	0.048	0.000
whiplash2[d]	-0.623	1.060	0.557	-0.620	1.061	0.559	-0.612	1.086	0.287
whiplash3[e]	-0.630	1.425	0.659	-0.636	1.425	0.656	-0.625	1.494	0.676
σ	8.079	0.573	0.000	8.058	0.557	0.000	7.772	0.321	0.000
ρ_{12}	0.767	0.169	0.000	0.757	0.178	0.000	0.741	0.012	0.000
ρ_{13}	0.223	0.571	0.697	0.154	0.584	0.792	0.353	0.715	0.622
Loglik	-1877.497			-1878.180					

[a]Extended two-step method where individual probit model is used for the inverse Mills ratio.
[b]Intercept for unit non-response.
[c]Intercept for item non-response.
[d]Whiplash Associated Disorder- grade2. Grade1 was used as reference.
[e]Whiplash Associated Disorder- grade3. Grade1 was used as reference.

References

Ahn, S. C. (1992). The Lagrangean multiplier test for a model with two selectivity criteria, *Economics Letters* 38: 9–15.

Arellano-Valle, R. B., Branco, M. D. and Genton, M. G. (2006). A unified view of skewed distributions arising from selections, *The Canadian Journal of Statistics* 34: 581–601.

Arnold, B. C. and Beaver, R. J. and Groeneveld, R. A. and 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. *Scandinavian Journal of Statistics* 12: 171–178.

Azzalini, A. and Capitanio, A. (1999). Statistical applications of multivariate skew-normal distribution, *Journal of the Royal Statistical Society: Series B* 61: 579–602.

Azzalini, A. and Valle, A. D. (1996). The Multivariate Skew-normal distribution, *Biometrika* 83: 715–726.

Bellio, R. and Gori, E. (2003). Impact evaluation of job training programmes: Selection bias in multilevel models, *Journal of Applied Statistics* 30: 893–907.

Burton, K. and McClune, T. and Waddell, G. (2001). *The Whiplash Book*, TSO, The Stationery Office, London.

Capitanio, A., Azzalini, A. and Stanghellini, E. (2003). Graphical models for skew-normal variates, *Scandinavian Journal of Statistics* 30: 129–144.

Carpenter, J. and Pocock, S. and Lamm, C. J. (2002). Coping with missing data in clinical trials: A model-based approach applied to asthma trials, *Statistics in Medicine* 21: 1043–1066.

Cartinhour, J. (1990). One-dimensional marginal density functions of a truncated multivariate normal density function, *Communications in Statistics- Theory and Methods* 19: 197–203.

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

De Luca, G. and Peracchi, F. (2006). A sample selection model for unit and item nonresponse in cross-sectional surveys. CEIS Tor Vergata -Research Paper Series 95.

De Luca, G. and Peracchi, F. (2012). Estimating Engel curves under unit and item nonresponse, *Journal of Applied Econometrics* 27: 1076-1099.

Diggle, P. and Kenward, M. G. (1994). Informative drop-out in longitudinal data analysis (with discussion), *Applied Statistics* 43: 49–93.

Dominguez-Molina, J. A. and Gonzalez-Farias, G. and Ramos-Quiroga, R. (2004). *Skew-normality in Stochastic Frontier Analysis. In M. G. Genton (Ed.), Skew-Elliptical Distributions and Their Applications: A Journey Beyond Normality*, Chapman & Hall, CRC, Florida.

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

Genz, A. and Bretz, F. (2009). *Computation of Multivariate Normal and t Probabilities. Lecture Notes in Statistics, Vol. 195*, Springer-Verlag, Heidelberg.

Gonzalez-Farias, G. and Dominguez-Molina, J. A. and Gupta, A. K. (2004). *The closed skew-normal. In M. G. Genton (Ed.), Skew-Elliptical Distributions and Their Applications: A Journey Beyond Normality*, Chapman & Hall, CRC, Florida.

Ham, J. C. (1982). Estimation of a labour supply model with censoring due to unemployment and underemployment, *Review of Economic Studies* 49: 335–354.

Heckman, J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models, *Annals of Economic and Social Measurement* 5: 475–492

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

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 and Gumber, A. (2007). Managing Injuries of the Neck Trial (MINT): design of a randomised controlled trial of treatments for whiplash associated disorders, *BMC Muscloskeletal Disorder* 8: 7.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data (2 ed.)*, John Wiley & Sons Ltd, New Jersey.

Meng, C. and Schmidt, P. (1985). On the cost of partial observability in the bivariate probit model, *International Economic Review* 26: 71–85.

Mudholkar, G. S. and Hutson, A. D. (2000). The epsilon-skew normal distribution for analyzing near-normal data, *Journal of Statistical Planning and Inference* 83: 291–309.

Poirier, D. J. (1980). Partial observability in bivariate probit models, *Journal of Econometrics* 12: 209–217.

Puhani, A. P. (2000). The Heckman correction for sample selection and its critique, *Journal of Economic Surveys* 14: 53–68

Rosenman, R. and Mandal, B. and Tennekoon, V. and Hill, L. G.(2010). Estimating treatment effectiveness with sample selection. Working paper 2010-05.

Rubin, D. B. (1976). Inference and missing data, *Biometrika* 63: 581–592.

Vernon, H. and Mior, S. (1991). The Neck Disability Index: a study of reliabity and validity, *Journal of Manipulative and Physiological Therapeutics* 7: 409–415.

Appendix

Three-level sample selection model

Since hidden truncation models are special cases of selection distributions, a three-level selection model can be derived using an extension of the model of Arnold et al. (1993). We consider the non-truncated marginal of a truncated quadrivariate normal random variable and apply the result of Cartinhour (1990).

Suppose $f(y, s_1, s_2, s_3)$ is the density of a quadrivariate normal random variable with mean vector $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3, \mu_4)'$ and covariance matrix

$$\Sigma = \begin{pmatrix} \sigma^2 & \sigma\rho_{12} & \sigma\rho_{13} & \sigma\rho_{14} \\ \sigma\rho_{12} & 1 & \rho_{23} & \rho_{24} \\ \sigma\rho_{13} & \rho_{23} & 1 & \rho_{34} \\ \sigma\rho_{14} & \rho_{24} & \rho_{34} & 1 \end{pmatrix}.$$

Suppose further that $W = (Y, S_1, S_2, S_3)'$ has joint density

$$\begin{cases} f(\mathbf{w}) & = \frac{1}{C}\frac{1}{\sqrt{(2\pi)^4|\Sigma|}} e^{-1/2(\mathbf{w}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{w}-\boldsymbol{\mu})}, \quad \mathbf{w} \in R \\ & = 0, \text{ otherwise} \end{cases}$$

where R is a rectangle in 4-space; R: $-\infty < y < \infty$, $c_{s_1} < s_1 < \infty$, $c_{s_2} < s_2 < \infty$ and $c_{s_3} < s_3 < \infty$. C is a normalizing constant (necessary to ensure that the density function integrates to 1) given by

$$C = \int_R \frac{1}{C}\frac{1}{\sqrt{(2\pi)^4|\Sigma|}} e^{-1/2(\mathbf{w}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{w}-\boldsymbol{\mu})} \mathrm{d}\mathbf{w}.$$

This implies (Y, S_1, S_2, S_3) has a truncated quadrivariate normal distribution. S_1, S_2 and S_3 are truncated below at c_{s_1}, c_{s_2} and c_{s_3} respectively. We are interested in the marginal distribution of Y, which is the only non-truncated random variable in this formulation. Using Cartinhour (1990), we can write the required density as,

$$f(y) = \frac{1}{C} e^{-1/2(\frac{y-\mu_1}{\sigma^2})^2} \int_{c_{s_1}}^{\infty}\int_{c_{s_2}}^{\infty}\int_{c_{s_3}}^{\infty} \frac{1}{\sqrt{(2\pi)^3|A_{\neg y}^{-1}|}} e^{-1/2(\mathbf{w}_{\neg y}-\mathbf{m}(y))'A_{\neg y}(\mathbf{w}_{\neg y}-\mathbf{m}(y))} \mathrm{d}\mathbf{w}_{\neg y}, \quad (20)$$

where $\mathbf{w}_{\neg y} = (s_1, s_2, s_3)'$, $A_{\neg y}^{-1} = \Sigma_3^{\star} = \begin{pmatrix} 1-\rho_{12}^2 & \rho_{23}-\rho_{12}\rho_{13} & \rho_{24}-\rho_{12}\rho_{14} \\ \rho_{23}-\rho_{12}\rho_{13} & 1-\rho_{13}^2 & \rho_{34}-\rho_{13}\rho_{14} \\ \rho_{24}-\rho_{12}\rho_{14} & \rho_{34}-\rho_{13}\rho_{14} & 1-\rho_{14}^2 \end{pmatrix}$ (this is the inverse of the submatrix of the inverse of Σ when the row and column corresponding to y are deleted), and $\mathbf{m}(y)$ is defined as $\mathbf{m}(y) = \mu_{\neg 1} + (y - \mu_1/\sigma^2)\mathbf{k}$; with $\mu_{\neg 1} = (\mu_2, \mu_3, \mu_4)$, and $\mathbf{k} = (\sigma\rho_{12}, \sigma\rho_{13}, \sigma\rho_{14})'$. We determine C and the double integral in equation (20).

Now, C can be written as a centralized normal integral

$$\Phi_4\left(\begin{pmatrix} -\infty \\ c_{s_1}-\mu_2 \\ c_{s_2}-\mu_3 \\ c_{s_3}-\mu_4 \end{pmatrix}, \begin{pmatrix} \infty \\ \infty \\ \infty \\ \infty \end{pmatrix}; \Sigma\right) = \Phi_3\left(\begin{pmatrix} 0 \\ 0 \\ 0, \end{pmatrix}, \begin{pmatrix} c_{s_1}-\mu_2 \\ c_{s_2}-\mu_3 \\ c_{s_3}-\mu_4 \end{pmatrix}; \Sigma_3\right), \quad (21)$$

where $\Sigma_3 = \begin{pmatrix} 1 & \rho_{23} & \rho_{24} \\ \rho_{23} & 1 & \rho_{34} \\ \rho_{24} & \rho_{34} & 1 \end{pmatrix}$. Using properties of multivariate normal cumulative distribution function and the definition of $\mathbf{m}(y)$, the triple integral reduces to

$$\Phi_3 \left(\begin{pmatrix} \sigma\rho_{12} \\ \sigma\rho_{13} \\ \sigma\rho_{14} \end{pmatrix} \left(\frac{y-\mu_1}{\sigma^2}\right); \begin{pmatrix} c_{s_1} - \mu_2 \\ c_{s_2} - \mu_3 \\ c_{s_3} - \mu_4 \end{pmatrix}, \Sigma_3^\star \right) \tag{22}$$

The required density is derived when equations (21) and (22) are substituted in equation (20). The PDF is

$$\frac{\phi(y; \mu_1, \sigma^2)\Phi_3(D(y-\mu_1); \boldsymbol{\nu}, \Sigma_3^\star)}{\Phi_3(\mathbf{0}; \boldsymbol{\nu}, \Sigma_3)},$$

where $\mathbf{0} = (0,0,0)'$, $D = (\rho_{12}/\sigma, \rho_{13}/\sigma, \rho_{14}/\sigma)'$, and $\boldsymbol{\nu} = (c_{s_1} - \mu_2, c_{s_2} - \mu_3, c_{s_3} - \mu_4)'$. It is easy to see that $\Sigma_3 = \Sigma_3^\star + D\sigma^2 D'$.

The continuous component of the three-level sample selection model can be written as

$$\frac{\phi(y; \beta' x, \sigma^2)\Phi_3\left\{ D(y - \beta' x); \begin{pmatrix} -\alpha_1' x \\ -\alpha_2' x \\ -\alpha_3' x \end{pmatrix}, \Sigma_3^\star \right\}}{\Phi_3\left\{ \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}; \begin{pmatrix} -\alpha_1' x \\ -\alpha_2' x \\ -\alpha_3' x \end{pmatrix}, \Sigma_3 \right\}}.$$

where c_{s_1}, c_{s_2} & c_{s_3} are zero, and $\mu_1 = \beta' x$, $\mu_2 = \alpha_1' x$, $\mu_3 = \alpha_2' x$ and $\mu_4 = \alpha_3' x$.