A Unified Approach to Multilevel Sample Selection Models

Emmanuel O. Ogundimu* and Jane L. Hutton

October 20, 2012

Abstract

Scores derived from response to questionnaires are widely used in health and social studies to measure aspects of health and well being. Respondent to studies do not always complete all questions, which result in two levels of missing data. If a subject declines participation, we have unit non-response; if questions are skipped, we have item non-response. We can regard the observed outcomes as the result of a two level selection process. We propose a unified approach for multilevel sample selection models in a parametric framework by treating the outcome variable as the non-truncated marginal of a truncated multivariate normal distribution. The resulting density for the outcome is the continuous component of the sample selection density, and has links with the closed skew-normal distribution. The closed skew-normal distribution provides a framework which simplifies the derivation of the conditional expectation and variance of the observed data. We use this to generalize the Heckman's two-step method to a multilevel sample selection model. Finite sample performance of the full information maximum likelihood estimator of this model is studied through a Monte Carlo simulation and issues about local identifiability of model parameters are discussed in the context of non-singularity of the information matrix. The method is applied to a clinical trial data of neck injuries with unit and item non-response.

*Correspondence to: E.O. Ogundimu, Department of Statistics, University of Warwick, Coventry CV4 7AL, UK. Tel.:+44 7539084180 fax: +44 2476524532. E-mail address: E.O.Ogundimu@warwick.ac.uk.

KEY WORDS: Unit and Item non-response; Closed skew-normal distribution; Hidden truncation; 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 type inevitable- unit and item non-response. Unit non-response occur when the whole questionnaire is missing for a patient and item non-response occur where a response has not been provided for a question. The traditional practice is to use weighting adjustment 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 complete data set.

Although these methods have reached a high level of sophistication, they normally assume that the missing data mechanism is MAR, 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. In multivariate settings with arbitrary patterns of non-response, imputation, and hence the MAR assumption, is convenient computationally, but it is often implausible (Robins and Gill, 1997). In this setting, MAR means that a patient's probabilities of responding to items may depend only on his or her own set of observed items, which is an unreasonable assumption. Specifically, the use of mean imputation is justifiable if items within the scale are strongly correlated with each other but correlation with external factors is low relative to within-scale correlations. This cannot be readily established in practice. 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 tool.

Sample selection models, also referred to as models with incidental (hidden) truncation, 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-response 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 Luca and Peracchi (2006) 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), Arendt and Holm (2006) and Rosenman et al. (2010).

The approach we want to consider here differs from previous studies in the following regards. First, we show that the sample selection models formulated using the hidden truncation approach of Arnold et al. (1993) has the unified framework of skew-distributions arising from selections of Arellano-Valle et al. (2006) as a special case. This link is generalized to multilevel selection problems where we establish that the continuous component of (multilevel) sample selection density belongs to the closed skew-normal (CSN) distribution. Second, interest may be in the underlying variance in the (multilevel) selection process. A new method that generalizes the method described on p. 784 of Greene (2003) for extracting consistent variance from the model is therefore needed. Third, the model formulation showed that the complete cases follow a well established distribution (CSN), and as such likelihood based methods can be used for parameter estimation in this model.

Without loss of generality, we will restrict attention to two-level sample selection model. The article is organized as follows. In section 2, we describe two alternative routes to the formulation of the Copas and Li (1997) selection model and show that

the continuous component of its log-likelihood function belongs to a wider class of the CSN distribution. Section 3 extends the Copas and Li (1997) model to multilevel selection problems and its properties are derived using well established results of the CSN distribution. Finite sample performance of the model is studied via Monte Carlo simulation in section 4. The model is applied to a real life data in section 5 and conclusion given in section 6.

2 Alternative routes to formulation of the Copas and Li (1997) model

In this section, we formulate a link between the continuous component of sample selection density and the extended skew-normal distribution using Arnold et al. (1993) and Arellano-Valle et al. (2006) methods. We also show that the resulting extended skew-normal distribution is a special case of a wider class of the closed skew-normal family. Since the closed skew-normal distribution is a well established distribution, its properties was shown to unify methods for the derivation of the mean and the variance of the distribution of the observed data in a straightforward way.

2.1 The Copas and Li (1997) model

Copas and Li (1997) present a model for missing data described as follows. Let $Y^\star$ be the outcome variable of interest, assumed linearly related to covariates x_i through the standard multiple regression

$$Y_i^\star = \beta' x_i + \sigma\varepsilon_{1i}, \quad i = 1, \ldots, N. \tag{2.1}$$

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

$$S_i^\star = \gamma' x_i + \varepsilon_{2i}, \quad i = 1, \ldots, N \tag{2.2}$$

where β and γ are unknown parameters and x_i are fixed observed characteristics not subject to missingness. Suppose further that

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

Note that the variance of $S_i^\star$ is posited as 1 because only its sign is observed and the variance is not identifiable in the model. It is assumed that $Y_i^\star$ and $S_i^\star$ are correlated with parameter ρ in the underlying process. The parameter $\rho \in$ [-1,1] determines the severity of the selection process. Due to the selection, when $S_i^\star > 0$ (the 0 threshold is arbitrary since no symmetry is assumed), we observe Y_i with n observations out of N from $Y_i^\star$ i.e. $S_i = I(S_i^\star > 0)$ and $Y_i = Y_i^\star S_i$.

The density of the observed data was given by:

$$f(y|x, S = 1) = \frac{\frac{1}{\sigma}\phi\left(\frac{y-\beta' x}{\sigma}\right)\Phi\left(\frac{\gamma' x + \rho\left(\frac{y-\beta' x}{\sigma}\right)}{\sqrt{1-\rho^2}}\right)}{\Phi(\gamma' x)}, \tag{2.3}$$

where ϕ and Φ denote standard normal PDF (probability density function) and corresponding CDF (cumulative distribution function) respectively.

The complete density of a sample selection model is comprised of a continuous component (the conditional density given by (2.3)), 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 process. In Copas and Li (1997), the model $P(S = s) = \{\Phi(\gamma' x)\}^s\{1 - \Phi(\gamma' x)\}^{1-s}$ (i.e a probit model) was used. The log-likelihood function is therefore

$$l(\beta, \sigma, \gamma, \rho) = \sum_{i=1}^{N} S_i\Big(\ln f(y_i|x_i, S_i = 1)\Big) + \sum_{i=1}^{N} S_i(\ln \Phi(\gamma' x_i)) + \sum_{i=1}^{N}(1 - S_i)\ln \Phi(-\gamma' x_i) \tag{2.4}$$

The conditional expectation of the observed data, with density given by (2.3) is

$$E(Y|x, S^\star > 0) = \beta' x + \sigma\rho\Lambda(\gamma' x), \tag{2.5}$$

where Λ is the inverse Mills ratio ϕ/Φ. This equation is the basis of Heckman's two-step method (Heckman, 1979). A standard probit model is fitted to the cases where $S = 1$ and γ is estimated. The resulting estimate of γ is used to form $\Lambda(\gamma' x)$ for each of the cases with $S = 1$. This quantity is then taken as an additional covariate in equation (2.5) and fitted by least squares. The coefficient of the additional covariate now gives an estimate of $\sigma\rho$. The conditional variance is given by

$$\text{var}(Y|x, S^\star > 0) = \sigma^2[1 - \rho^2\lambda(\gamma' x)\{\gamma' x + \lambda(\gamma' x)\}] \tag{2.6}$$

The estimates of both ρ and σ can be obtained by equating the average value on the RHS (right-hand side) of equation (2.6) to the observed residual variance of the second-step regression.

Two alternative formulations of the continuous component of (2.4) and their link with the ESN distribution (and by extension to CSN family) are shown next.

2.1.1 Hidden Truncation

Arnold et al. (1993) considered inference for the non-truncated marginal of a truncated bivariate normal distribution. The *hidden truncation* formulation can be described as follows: Let (Y, S) be a bivariate normal random variable with mean (μ_y, μ_s), variances (σ_y^2, σ_s^2) and correlation ρ. Assume that the values of Y are available (or selected) only if $S > c$ while the values of Y are not available otherwise. Clearly, $(Y, S > c)$ has a truncated bivariate normal distribution. By direct integration, the density of $W \equiv Y|S > c$, the non-truncated marginal, is given as

$$\begin{aligned} f_W(w) &= \frac{1}{\sigma_y}\phi\big(\frac{w-\mu_y}{\sigma_y}\big)\Phi\left(\frac{\frac{\mu_s - c}{\sigma_s} + \rho(\frac{w-\mu_y}{\sigma_y})}{\sqrt{1-\rho^2}}\right)\Bigg/\Big(1 - \Phi\big(\frac{c-\mu_s}{\sigma_s}\big)\Big) \\ &= \phi\big(\frac{w-\mu}{\sigma}\big)\Phi\big(\lambda_0 + \lambda_1(\frac{w-\mu}{\sigma})\big)\Big/\sigma\Phi\left(\frac{\lambda_0}{\sqrt{1+\lambda_1^2}}\right), \end{aligned} \tag{2.7}$$

where $\lambda_0 = (\mu_s - c)/\sigma_s\sqrt{1-\rho^2}$, $\lambda_1 = \rho/\sqrt{1-\rho^2}$, $\sigma = \sigma_y$, and $\mu = \mu_y$.

The model given by (2.7) is a ESN density (see Capitanio et al. (2003)). We show in the next section that this density equivalent to (2.3).

2.1.2 Skew distributions arising from selection

Arellano-Valle et al. (2006) unified a broad class of selection distributions arising from selection mechanism of the form $\mathbf{Y}|\mathbf{S} \in C$, where $\mathbf{Y}^\star \in \mathbb{R}^{d_1}$, $\mathbf{S}^\star \in \mathbb{R}^{d_2}$, and C

is a measurable subset in $\mathbb{R}^{d_2}$ such that $0 < \Pr(\mathbf{S}^\star \in C) < 1$. Suppose $Y^\star$ and $S^\star$ are as defined in (2.1) and (2.2) respectively, this concept expresses the continuous component of sample selection density as

$$f(y|x, S^\star > 0) = \frac{f(y|x)P(S^\star > 0|y, x)}{P(S^\star > 0|x)}. \tag{2.8}$$

Using the conditioning and marginalization properties of the bivariate normal distribution in (2.8), we have (2.3).

The link between Arnold et al. (1993) and Arellano-Valle et al. (2006) is readily established. If the regression parametrization $\mu_y = \beta' x$ and $\mu_s = \gamma' x$ are used in (2.7), and we put c and σ_s equal 0 and 1 respectively, equation (2.3) is recovered. Thus (2.3) is a special case of (2.7).

After some algebra, one can re-write (2.3) as

$$f(y|x, S = 1) = \frac{\phi\Big(y; \beta' x, \sigma^2\Big)\Phi\Big(\frac{\rho}{\sigma}\Big(y - \beta' x\Big); -\gamma' x, 1 - \rho^2\Big)}{\Phi\Big(0; -\gamma' x, 1\Big)}, \tag{2.9}$$

where in general, $\Phi(y; \mu, \sigma^2) = \Phi(\frac{y-\mu}{\sigma})$. Equation (2.9) corresponds to a form of the closed skew-normal density.

2.2 The closed skew-normal distribution

The CSN family is constructed in the multivariate framework because it is a generalization of the multivariate skew-normal distribution such that some important properties of the normal distribution are preserved (Gonzalez-Farias et al., 2004).

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:

$$f_{p,q}(\mathbf{y}) = C\phi_p(\mathbf{y}; \boldsymbol{\mu}, \Sigma)\Phi_q(D(\mathbf{y} - \boldsymbol{\mu}); \boldsymbol{\nu}, \Delta), \quad \mathbf{y} \in \mathbb{R}^p, \tag{2.10}$$

with:

$$C^{-1} = \Phi_q(\mathbf{0}; \boldsymbol{\nu}, \Delta + D\Sigma D'), \tag{2.11}$$

where $\phi_p(.;\boldsymbol{\eta},\Psi)$, $\Phi_p(.;\boldsymbol{\eta},\Psi)$ are the PDF and CDF 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 (2.10), gives,

$$f_{p,q}(\mathbf{y}) \quad = 2^q \phi_p(\mathbf{y};\boldsymbol{\mu},\Sigma)\Phi_q(D(\mathbf{y}-\boldsymbol{\mu});\mathbf{0},\Delta), \tag{2.12}$$

which is the multivariate skew normal distribution discussed in Azzalini and Valle (1996). It is straightforward to see that the PDF in (2.10) includes the normal distribution as a special case when D and $\boldsymbol{\nu}=0$.

Now, one can write the expression given by (2.9) as a closed skew-normal density. If $p=q=1$, $\mu=\beta' x$, $\Sigma=\sigma^2$, $D=\rho/\sigma$, $\nu=-\gamma' x$, and $\Delta=1-\rho^2$, we have

$$(Y|x,S=1) \sim CSN_{1,1}\Big(\beta' x,\sigma^2,\frac{\rho}{\sigma},-\gamma' x,1-\rho^2\Big). \tag{2.13}$$

The moment generation function (mgf) of the CSN with density given in (2.10) is

$$M_{\mathbf{Y}}(\mathbf{t}) = \frac{\Phi_q(D\Sigma\mathbf{t};\boldsymbol{\nu},\Delta+D\Sigma D')}{\Phi_q(\mathbf{0};\boldsymbol{\nu},\Delta+D\Sigma D')} e^{\mathbf{t}'\boldsymbol{\mu}+\frac{1}{2}\mathbf{t}'\Sigma\mathbf{t}}, \quad \mathbf{t} \in \mathbb{R}^p. \tag{2.14}$$

To compute the first and second moments of CSN distribution, the derivatives of (2.14) may be considered. The first derivative of the mgf evaluated at $t=0$ is

$$E(\mathbf{Y}) = \frac{\partial}{\partial t} M_{\mathbf{Y}}(t)\Big|_{t=0} = \boldsymbol{\mu} + \Sigma D'\psi, \tag{2.15}$$

where $\psi = \frac{\Phi_q^\star(0;\boldsymbol{\nu},\Delta+D\Sigma D')}{\Phi_q(0;\boldsymbol{\nu},\Delta+D\Sigma D')}$, and for any positive definite matrix $\Omega_{q\times q}$ with elements $\omega_{i,j}$,

$$\Phi_q^\star(\mathbf{s};\boldsymbol{\nu},\Omega) = \frac{\partial}{\partial \mathbf{s}_i}\Phi_q(\mathbf{s};\boldsymbol{\nu},\Omega). \tag{2.16}$$

Simplification of the expression on the RHS (right-hand side) of (2.16) involves evaluation of derivatives of multi-normal integral. Dominguez-Molina et al. (2004)

gave the expression as

$$\frac{\partial}{\partial \mathbf{s}_i}\Phi_q(\mathbf{s};\boldsymbol{\nu},\Omega) = \phi(S_i;\nu_i,\omega_{ii})\Phi_{q-1}(\mathbf{s}_{\neg i};\boldsymbol{\nu}_{\neg i}+\Omega_{i\neg i}\omega_{ii}^{-1}(S_i-\nu_i),\Omega_{\neg i\neg i}-\omega_{ii}^{-1}\Omega_{i\neg i}\Omega'_{i\neg i}), \tag{2.17}$$

where $\mathbf{s}_{\neg i} = (s_1,\ldots,s_{i-1},s_{i+1},\ldots,s_p)'$, $\Omega_{\neg i\neg i}$ is the $q-1\times q-1$ matrix derived from $\Omega_{q\times q}$ by eliminating its i-th row and its i-th column and $\Omega_{i\neg i}$ is the $q-1$ vector derived from the i-th column of Ω by removing the i-th row term.

The variance is given as

$$\begin{aligned}\text{var}(Y) =& \frac{\partial^2}{\partial t\partial t'}M_{\mathbf{Y}}(t)\bigg|_{t=0} - E(Y)E(Y') \\ =& \Sigma + \boldsymbol{\mu}\boldsymbol{\mu}' + \boldsymbol{\mu}\psi' D\Sigma + \Sigma D\psi\boldsymbol{\mu}' + \Sigma D'\Lambda D\Sigma - E(Y)E(Y'),\end{aligned} \tag{2.18}$$

where $\Lambda = \frac{\Phi_q^{\star\star}(0;\boldsymbol{\nu},\Delta+D\Sigma D')}{\Phi_q(0;\boldsymbol{\nu},\Delta+D\Sigma D')}$, and

$$\Phi_q^{\star\star}(\mathbf{s};\boldsymbol{\nu},\Omega) = \frac{\partial^2}{\partial \mathbf{s}_i\partial \mathbf{s}_j}\Phi_q(\mathbf{s};\boldsymbol{\nu},\Omega). \tag{2.19}$$

The RHS of (2.19) is expressed as

$$\begin{aligned}\frac{\partial^2}{\partial \mathbf{s}_i\partial \mathbf{s}_j}\Phi_q(\mathbf{s};\boldsymbol{\nu},\Omega) =& \phi_2(s_{[i,j]};\boldsymbol{\nu}_{[i,j]},\Omega_{[i,j]})\Phi_{q-2}\Big(\mathbf{s}_{\neg[i,j]} - \Omega_{[i,j]\neg[i,j]}\Omega_{[i,j]}^{-1}(\mathbf{s}_{[i,j]}-\boldsymbol{\nu}_{[i,j]}); \\ & \boldsymbol{\nu}_{\neg[i,j]},\Omega_{\neg[i,j]\neg[i,j]} - \Omega_{[i,j]\neg[i,j]}\Omega_{[i,j]}^{-1}\Omega'_{[i,j]\neg[i,j]}\Big),\end{aligned} \tag{2.20}$$

where the definition is as before but with the components (i,j) taken simultaneously and $\phi_2(.;.,.)$ denotes the PDF of a standard bivariate normal distribution. By convention, $\Phi_0 = 1$

It is straightforward to see that the continuous component of Copas and Li (1997) model gives the conditional mean and variance in equations (2.5) and (2.6) when its CSN representation (2.9) is used in (2.15) and (2.18) respectively.

Although the derivation of the conditional expectation and variance in Copas and

Li (1997) model is straightforward, the expression becomes more complicated when the dimension of CDF of the normal distribution is greater than 1. Without loss of generality, we will illustrate this with two-level selection problem in the next section after showing that the continuous component of the two-level selection model can be derived as the non-truncated marginal of a truncated trivariate normal distribution.

3 Mathematical formulation of the Model

We showed in section 2 that the continuous part $(Y|x, S = 1)$ of Copas and Li (1997) selection model can arise from hidden truncation and that it belongs to the extended skew-normal family. Here, we quantify the statistical bias in two-level selection model and show that the Arnold et al. (1993) hidden truncation formulation of the Copas and Li (1997) model can be extended to multilevel selection problems. In particular, we establish that the continuous component of the model is a straightforward generalization of the extended skew-normal distribution to the closed skew-normal distribution.

3.1 Statistical bias in two-level sample selection problem

In this section, we present an expression that quantifies the overall non-response bias in two-level sample selection model. The non-response mechanism under which the bias vanishes is also described in a manner similar to the one discussed in Luca and Peracchi (2006). The model is developed by assuming the two-level selection equations correspond to unit and item non-response. We begin by extending the Copas and Li (1997) model with an additional selection equation,

$$S_{2i}^{\star} = \alpha' x_i + \varepsilon_{3i}, \quad i = 1, \ldots, N \tag{3.21}$$

Since we have two selection equations, we can take $S_{1i} = I(S_{1i}^{\star} > 0)$ and $S_{2i} = I(S_{2i}^{\star} > 0)$. The usable observations are on $Y_i = Y_i^{\star} S_{1i} S_{2i}$.

Now, the interest is to estimate the conditional mean function of a random outcome using data from a clinical study. Suppose at each time points, N patients are expected to respond, then unit non-response may reduce the number of patients to $N_1 < N$ responding units. Further, non-response at item level may reduce effective number of observations to $N_2 < N_1$. The loss of information due to missing observations results

in efficiency loss relative to the ideal situation of complete response.

It is logical to consider at each time point a sequential framework where patients are first observed before they decide to answer specific item of the questionnaire. Let the indicator of unit response be S_1, which is always observed, while the indicator of item response be S_2 which is observed conditional on S_1 being present. Since the observations are present when the indicators are greater than zero, we can then describe the response process by $\pi_0 = \Pr\{S_1 = 0\}$ and $\pi_{0|1} = \Pr\{S_2 = 0|S_1 = 1\}$ representing the probability of unit non-response and the probability of item non-response conditional on unit response respectively. Since Y is the outcome of interest, we have using, the law of iterated expectations,

$$E(Y|S_1 = 1) - E(Y) = \pi_0[E(Y|S_1 = 1) - E(Y|S_1 = 0)] \tag{3.22}$$

In addition,

$$E(Y|S_1 = 1) = E(Y|S_1 = 1, S_2 = 1) + \pi_{0|1}[E(Y|S_1 = 1, S_2 = 0) - E(Y|S_1 = 1, S_2 = 1)]. \tag{3.23}$$

The difference between the conditional mean of Y for the fully responding patients and the unconditional mean of Y for the complete response is the overall non-response bias and is given as $E(Y|S_1 = 1, S_2 = 1)$ - $E(Y)$. Substituting (3.23) into (3.22) and rearranging gives

$$\begin{aligned} E(Y|S_1 = 1, S_2 = 1) - E(Y) = \pi_0[E(Y|S_1 = 1) - E(Y|S_1 = 0)] + \\ \pi_{0|1}[E(Y|S_1 = 1, S_2 = 1) - E(Y|S_1 = 1, S_2 = 0)]. \end{aligned}$$

Generally, the overall bias has two separate components that are proportional to the probabilities of unit and item non-response respectively. There are 3 ways by which the bias can be zero in the above equation. If there is neither unit nor item non-response ($\pi_0 = \pi_{0|1} = 0$), if both unit and item non-response are MAR ($E(Y|S_1 = 1)= E(Y|S_1 = 0)$ and $E(Y|S_1 = 1, S_2 = 1)= E(Y|S_1 = 1, S_2 = 0)$), and when the bias terms due to unit and item non-response have opposite sign and offset

each other. Next, we consider a model that removes this bias.

3.2 Two-level selection models

In section 2.1.1, hidden truncation problem was used to derive the continuous component of the sample selection density. We use similar approach here by deriving the continuous component of two-level sample selection model from the general hidden truncation framework before restricting the truncation point to zero and using regression parametrization to link it to the Arellano-Valle et al. (2006) unified framework. The derivation of the conditional mean and variance is complicated when there is more than one selection equation. A general result given in section 2.2 is used to simplify this and a two-level selection problem is used to illustrate it. Generalization to higher dimension is shown to be straightforward.

3.2.1 Hidden truncation method

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

$$\Sigma = \begin{pmatrix} \sigma^2 & \sigma\rho_{12} & \sigma\rho_{13} \\ \sigma\rho_{12} & 1 & \rho_{23} \\ \sigma\rho_{13} & \rho_{23} & 1 \end{pmatrix}. \tag{3.24}$$

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

$$\begin{cases} f(\mathbf{w}) & = \frac{1}{C}\frac{1}{\sqrt{(2\pi)^3|\Sigma|}} e^{-1/2(\mathbf{w}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{w}-\boldsymbol{\mu})}, \quad \mathbf{w} \in R \\ & = 0, \text{ otherwise} \end{cases}$$

where R is a rectangle in 3-space; R: $-\infty < y < \infty$, $c_{s_1} < s_1 < \infty$ and $c_{s_2} < s_2 < \infty$. C is a normalizing constant (necessary to ensure that the density function integrates to 1) given by

$$C = \int_R \frac{1}{C}\frac{1}{\sqrt{(2\pi)^3|\Sigma|}} e^{-1/2(\mathbf{w}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{w}-\boldsymbol{\mu})} \mathrm{d}\mathbf{w}.$$

This implies (Y, S_1, S_2) has a truncated trivariate normal distribution. S_1 and S_2 are truncated below at c_{s_1} and c_{s_2} 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,

$$f(y) = \frac{1}{C} e^{-1/2(\frac{y-\mu_1}{\sigma^2})^2} \int_{c_{s_1}}^{\infty} \int_{c_{s_2}}^{\infty} \frac{1}{\sqrt{(2\pi)^2 |A_{\neg y}^{-1}|}} e^{-1/2(\mathbf{w}_{\neg y} - \mathbf{m}(y))' A_{\neg y} (\mathbf{w}_{\neg y} - \mathbf{m}(y))} \mathrm{d}\mathbf{w}_{\neg y}, \quad (3.25)$$

where $\mathbf{w}_{\neg y} = (s_1, s_2)'$, $A_{\neg y}^{-1} = \Sigma_2^{\star} = \begin{pmatrix} 1-\rho_{12}^2 & \rho_{23}-\rho_{12}\rho_{13} \\ \rho_{23}-\rho_{12}\rho_{13} & 1-\rho_{13}^2 \end{pmatrix}$ (this is the inverse of the submatrix of the inverse of Σ when the row and column corresponding to y is deleted), and $\mathbf{m}(y)$ is defined as $\mathbf{m}(y) = \mu_{\neg 1} + (y - \mu_1/\sigma^2)\mathbf{k}$; with $\mu_{\neg 1} = (\mu_2, \mu_3)$, and $\mathbf{k} = (\sigma\rho_{12}, \sigma\rho_{13})'$. We determine C and the double integral in equation (3.25).

Now, C can be written as a noncentral normal integral

$$\Phi_3 \left(\begin{pmatrix} -\infty \\ c_{s_1} \\ c_{s_2} \end{pmatrix}, \begin{pmatrix} \infty \\ \infty \\ \infty \end{pmatrix}, \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix}; \Sigma \right)$$

When the above is centralized, we have

$$\Phi_3 \left(\begin{pmatrix} -\infty \\ c_{s_1} - \mu_2 \\ c_{s_2} - \mu_3 \end{pmatrix}, \begin{pmatrix} \infty \\ \infty \\ \infty \end{pmatrix}; \Sigma \right) = \Phi_2 \left(\begin{pmatrix} 0 \\ 0, \end{pmatrix}, \begin{pmatrix} c_{s_1} - \mu_2 \\ c_{s_2} - \mu_3 \end{pmatrix}; \Sigma_2 \right), \quad (3.26)$$

where $\Sigma_2 = \begin{pmatrix} 1 & \rho_{23} \\ \rho_{23} & 1 \end{pmatrix}$. Using properties of multivariate normal cumulative distribution function and the definition of $\mathbf{m}(y)$, the double integral reduces to

$$\Phi_2 \left(\begin{pmatrix} \sigma\rho_{12} \\ \sigma\rho_{13} \end{pmatrix} \left(\frac{y - \mu_1}{\sigma^2}\right); \begin{pmatrix} c_{s_1} - \mu_2 \\ c_{s_2} - \mu_3 \end{pmatrix}, \Sigma_2^{\star} \right) \quad (3.27)$$

The required density is derived when equations (3.26) and (3.27) are substituted in equation (3.25). The PDF is

$$\frac{\phi(y; \mu_1, \sigma^2) \Phi_2(D(y - \mu_1); \boldsymbol{\nu}, \Sigma_2^{\star})}{\Phi_2(\mathbf{0}; \boldsymbol{\nu}, \Sigma_2)}, \quad (3.28)$$

where $\mathbf{0} = (0, 0)'$, $D = (\rho_{12}/\sigma, \rho_{13}/\sigma)'$, and $\boldsymbol{\nu} = (c_{s_1} - \mu_2, c_{s_2} - \mu_3)'$. It is easy to

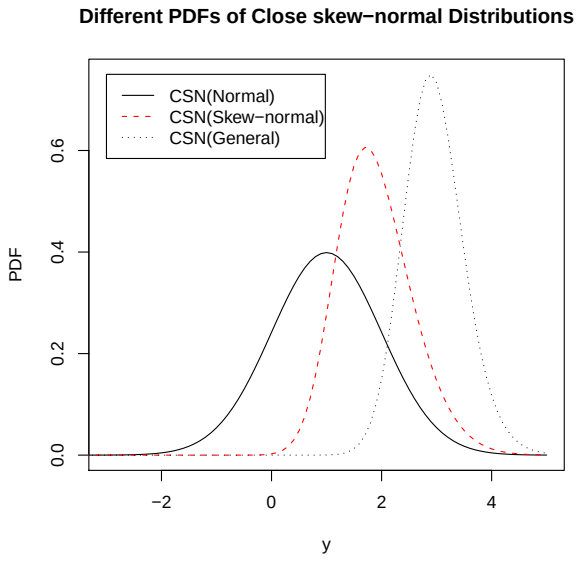

Figure 1: Comparison of Close skew-normal densities

see that $\Sigma_2 = \Sigma_2^\star + D\sigma^2 D'$, and thus (3.28) belongs to the closed skew-normal (CSN) family (see section 2.2).

A plot of the PDF given by (3.28) 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. Thus, the need for model formulation in the general CSN family.

Arellano-Valle et al. (2006) equivalence of (3.28) can be obtained by restricting c_{s_1} & c_{s_2} to be zero, and using regression parametrization $\mu_1 = \beta' x$, $\mu_2 = \gamma' x$ and $\mu_3 = \alpha' x$. We then obtain,

$$\frac{\phi(y;\beta'x,\sigma^2)\Phi_2\left(D(y-\beta'x);\begin{pmatrix}-\gamma'x\\-\alpha'x\end{pmatrix},\Sigma_2^\star\right)}{\Phi_2\left(\begin{pmatrix}0\\0\end{pmatrix};\begin{pmatrix}-\gamma'x\\-\alpha'x\end{pmatrix},\Sigma_2\right)}. \tag{3.29}$$

The mathematical rigor in the derivation of (3.29) can be avoided if one realizes the link between hidden truncation and skew distributions arising from selection. This link does not depend on the dimension of selection equations.

3.2.2 Skew distributions arising from selection method

The multivariate generalization of Copas and Li (1997) which turn out to be a closed skew-normal distribution can be used to derive (3.29). The idea is based on conditioning in the multivariate normal distribution. Suppose we consider (2.1), the outcome equation and (2.2) and (3.21), the selection equations such that the error terms are distributed normally with means zero and covariance matrix given by (3.24). Then,

$$\begin{pmatrix}Y\\S_1\\S_2\end{pmatrix} \sim N_3\left(\begin{pmatrix}\beta'x\\\gamma'x\\\alpha'x\end{pmatrix}\begin{pmatrix}\sigma^2 & \sigma\rho_{12} & \sigma\rho_{13}\\\sigma\rho_{12} & 1 & \rho_{23}\\\sigma\rho_{13} & \rho_{23} & 1\end{pmatrix}\right).$$

Now, (2.8) can be generalized to a two-level selection model as

$$f(y|x,S_1=1,S_2=1)) = \frac{f(y|x)P(S_1=1,S_2=1|y,x)}{P(S_1=1,S_2=1)}. \tag{3.30}$$

The quantity $f(y|x)$ is a proper PDF with a skewing function $P(S_1=1,S_2=1|y,x)$ and a normalizing function $P(S_1=1,S_2=1)$ to ensure that the LHS (left-hand side) integrates to 1. 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 and $\Sigma_2^\star$ are as defined in section 3.2.1. When appropriate substitutions are made in equation (3.30), the resulting density becomes:

$$\frac{\phi(y;\beta'x,\sigma^2)\Phi_2\left(\frac{\gamma'x+\rho_{12}(\frac{y-\beta'x}{\sigma})}{\sqrt{1-\rho_{12}^2}}, \frac{\alpha'x+\rho_{13}(\frac{y-\beta'x}{\sigma})}{\sqrt{1-\rho_{13}^2}}; \frac{\rho_{23}-\rho_{12}\rho_{13}}{\sqrt{1-\rho_{12}^2}\sqrt{1-\rho_{13}^2}}\right)}{\Phi_2(\gamma'x, \alpha'x; \rho_{23})} \tag{3.31}$$

which is the standardized version of (3.29). Equation (3.31) is equivalent to equation 10 given in Ahn (1992).

In general, the CSN distribution is the continuous component of the multilevel sample selection density. In the bivariate case, it is given by equation (3.31). 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 extension of this result to more than two-level selection problem is straightforward. For instance, in the three-level selection model, the continuous component of the sample selection density is a closed skew-normal distribution with dimensions p=1 and q=3. The normalizing constant of this density turns out to be the completely observed part of the discrete component, which is a trivariate probit model with level of observability determined by context.

3.2.3 Moments and Maximum Likelihood estimator for multilevel selection model

The fact that the continuous component of the 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 generalization of Heckman's two-step method. Without loss of generality, we illustrate this with a two-level selection problem.

To derive the conditional mean and variance in this case, we make use of equations (2.15) and (2.18) respectively. The mean is given as:

$$E(Y|x, S_1^\star > 0, S_2^\star > 0) = \beta' x + \sigma\rho_{12}\Lambda_1(\theta) + \sigma\rho_{13}\Lambda_2(\theta) \tag{3.32}$$

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 (2.5)) 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 (3.32) and fitted by least squares. The coefficient of the additional covariates give estimates of $\sigma\rho_{12}$ and $\sigma\rho_{13}$ respectively.

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

$$\begin{aligned}
\text{var}(Y|x, S_1^\star > 0, S_2^\star > 0) =& \sigma^2 - \sigma^2\rho_{12}^2(\gamma' x)\Lambda_1(\theta) - \sigma^2\rho_{13}^2(\alpha' x)\Lambda_2(\theta) \\
&+ \frac{\phi_2(\gamma' x, \alpha' x; \rho_{23})}{\Phi_2(\gamma' x, \alpha' x; \rho_{23})}\left[2\sigma\rho_{12}\sigma\rho_{13} - \rho_{23}(\sigma^2\rho_{12}^2 + \sigma^2\rho_{13}^2)\right] \\
&- \Big(\sigma\rho_{12}\Lambda_1(\theta) + \sigma\rho_{13}\Lambda_2(\theta)\Big)^2 \\
=& \sigma^2 + \upsilon
\end{aligned} \tag{3.33}$$

The error terms of the selected sample are heteroscedastic. A generalization of

Heckman's estimator for σ^2 given by

$$\sigma^2 = (S - \sum \hat{\upsilon}_i)/N_2, \tag{3.34}$$

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:

$$\begin{aligned} l(\beta, \sigma, \gamma, \alpha, \rho_{12}, \rho_{13}, \rho_{23}) = \sum_{i=1}^{N} \Bigg(& S_{1i}S_{2i}\Big[\ln f(y_i|x_i, S_{1i}=1, S_{2i}=1)\Big] + S_{1i}S_{2i}\Big[\ln \Phi_2(\gamma' x_i, \alpha' x_i; \rho_{23})\Big] \\ & + S_{1i}(1-S_{2i})\Big[\ln \Phi_2(\gamma' x_i, -\alpha' x_i; -\rho_{23})\Big] + (1-S_{1i})S_{2i}\Big[\ln \Phi_2(-\gamma' x_i, \alpha' x_i; -\rho_{23})\Big] \\ & + (1-S_{1i})(1-S_{2i})\Big[\ln \Phi_2(-\gamma' x_i, -\alpha' x_i; \rho_{23})\Big] \Bigg). \end{aligned} \tag{3.35}$$

4 Monte Carlo simulation

The finite-sample performance of the models in section 3.2.3 are studied in two parts-the moment based estimator (3.32) and the maximum likelihood estimator (3.35). 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 $\Sigma = \begin{pmatrix} 1 & 0.7 & 0.5 \\ 0.7 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{pmatrix}$. This construction implies that the variance of the outcome model is 1.

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. Roughly 70% censored observations was generated by changing the intercept terms in the selection equations $S_{1i}^\star$ and $S_{2i}^\star$ to -0.1 and -0.2 respectively. In both cases, we allow for full observability in the bivariate process. A pilot simulation results show that there is very little gain in imposing exclusion restriction between the two selection equations, (although this is recommended in practice due to the linearity of the bivariate inverse

Table 1: Simulation results for the moment based estimator of two-level selection model.

		Bias				MSE			
		2TS	2TS0	TS	OLS	2TS	2TS0	TS	OLS
$m = 30\%$	β_0	0.0018	0.0159	0.0212	0.2842	0.0656	0.0898	0.0074	0.0819
	β_1	-0.0002	-0.0019	-0.0052	-0.1188	0.0091	0.0121	0.0027	0.0155
	σ	0.0370	0.4326	-0.0165	-0.1667	0.0123	0.5440	0.0026	0.0300
	ρ_{12}	-0.2305	-0.3471			0.6947	1.0817		
	ρ_{13}	-0.0489	-0.1999			0.0813	0.2676		
	ρ_{23}	0.0014	0.0034						
$m = 70\%$	β_0	0.0193	0.0464	0.0742	0.6857	0.5943	1.0811	0.0391	0.4733
	β_1	-0.0045	-0.0044	-0.0150	-0.1821	0.0301	0.0540	0.0055	0.0358
	σ	0.1481	0.8934	-0.0383	-0.2711	0.1051	1.9577	0.0067	0.0771
	ρ_{12}	-0.2923	-0.4247			0.7079	0.9836		
	ρ_{13}	-0.0899	-0.2682			0.0704	0.2774		
	ρ_{23}	0.0023	0.0022						

Mills ratio on a wide range of its support) and as such we did not impose this criteria. Since the moment based method is very common for modeling multilevel sample selection models in practice, we consider four alternative models in this case

- 2TS: Model that generalizes Heckman selection model and accounts for selectivity induced by the selection equations and further impose correlation of the error terms in the selection equations (3.32)
- 2TS0: Model that accounts for selection bias generated by the selection equations, but assumes that the errors in the selection equations are independent
- TS: Classical Heckman two-step method where the selection equations are collapsed to a single indicator for missingness (2.5)
- OLS: Ordinary least square regression using complete cases.

The use of full information maximum likelihood approach to multilevel sample selection problems is not common in the literature. This is due in part to the robustness of the moment based estimator (3.32), to deviation from normality. Nonetheless, we investigate its performance when the underlying normal assumption holds in a simulation study under three model specification. The model labelled 2SNM is the maximum likelihood counterpart of 2TS where correlation is imposed on the error terms of the two selection equations. The SNM1 and SNM2 models are the classical

Table 2: Simulation results for the likelihood based estimator of two-level selection model.

		Bias			MSE		
		2SNM	SNM1	SNM2	2SNM	SNM1	SNM2
$m = 30\%$	β_0	0.0028	0.0062	0.0062	0.0061	0.0044	0.0044
	β_1	-0.0094	0.0007	0.0007	0.0023	0.0021	0.0021
	σ	-0.0037	-0.0115	-0.0115	0.0017	0.0018	0.0018
	γ_0	0.0047	-0.3717		0.0029	0.1405	
	γ_1	0.0003	0.1442		0.0030	0.0235	
	γ_2	0.0021	0.2456		0.0025	0.0630	
	α_0	0.0038		-0.3717	0.0033		0.1405
	α_1	0.0020		-0.0558	0.0032		0.0058
	α_2	0.0056		-0.1544	0.0036		0.0265
	ρ_{12}	-0.1352	0.0729		0.1013	0.0171	
	ρ_{13}	0.0099		0.1271	0.0253		0.0279
	ρ_{23}	0.0009			0.0032		
$m = 70\%$	β_0	0.0701	0.0544	0.0544	0.0397	0.0309	0.0309
	β_1	-0.0156	-0.0099	-0.0099	0.0050	0.0049	0.0049
	σ	-0.0049	-0.0331	-0.0331	0.0052	0.0057	0.0057
	γ_0	-0.0002	-0.5212		0.0017	0.2739	
	γ_1	0.0010	0.1415		0.0021	0.0226	
	γ_2	0.0090	0.2359		0.0019	0.0584	
	α_0	-0.0002		-0.4212	0.0022		0.1797
	α_1	0.0021		-0.0585	0.0027		0.0060
	α_2	0.0077		-0.1641	0.0030		0.0297
	ρ_{12}	-0.1183	-0.0818		0.0917	0.0206	
	ρ_{13}	-0.0038		0.1182	0.0247		0.0279
	ρ_{23}	0.0001			0.0022		

Heckman selection where the two selection equations are collapsed into a single indicator for missingness. Since we have two selection equations, and we do not know the true underlying equation out of the two, the SNM1 model is assumed when the first selection equation is the correct model ($S_{1i}^{\star} = 1 + 0.4x_i + 0.3w_i + \varepsilon_{2i}$) and SNM2 model is assumed for the second selection equation ($S_{2i}^{\star} = 1 + 0.6x_i + 0.7w_i + \varepsilon_{3i}$).

Table 2 is the results of the simulation when the likelihood based estimator is used. When interest is not in the selection process, the results shows that collapsing the indicator for missingness and the use of classical Heckman model (SNM1 & SNM2) gives consistent parameter estimates for the outcome as well as the 2SNM model. However, correct specification of the selection model may be difficult in the classical Heckman model since more than one equation now governs the selection process and different covariates might feature in the equations. In addition, it is known that high degree of censoring usually leads to efficiency loss as compared to full data. This however, does not affect the consistency of the parameters as long as the model is correctly specified. In fact, the consistency of model parameters under 70% censored observation, as shown in our simulation result, does not appear worse than the 30% case (although there is increase in the variance of the former). However, results for moment based estimates (see Table 1) showed that a high level of censoring might affect the consistency of the estimates. Parameter estimates from OLS are not consistent.

5 Application to MINT Trials

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 two 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 measurements occasion). Further, approximately 50% of the patients are in the two treatment groups resulting from balanced randomisation. 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. We focus on the measurement at months 8 and use the two-level sample selection model to jointly analyze this two two non-response process.

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. Out of this, 77 patients returned the questionnaire blank (genuine unit non-response). Vernon (2009) recommended that patients with only 2 missed items should be considered complete, with mean imputation used for adjustment. Rather than discarding this patients, we categorize them as item non-respondents with 43 patients falling into this category. Of course, unit non-respondents are also item non-respondents, making patients with item non-response to be effectively 120. The fully responding units (complete cases) are 447 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 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 (3.35), the log-likelihood function for a bivariate probit sample selection model is

$$l(\gamma,\alpha,\rho_{23}) = \sum_{i=1}^{N}\Big(S_{1i}S_{2i}\Big[\ln\Phi_2(\gamma' x_i, \alpha' x_i;\rho_{23})\Big] + S_{1i}(1-S_{2i})\Big[\ln\Phi_2(\gamma' x_i, -\alpha' x_i;-\rho_{23})\Big] + (1-S_{1i})\Big[\ln\Phi(-\gamma' x_i)\Big]\Big). \tag{5.36}$$

A simulation study (not reported here) showed that if model (5.36) is correctly specified, correct specification includes imposing exclusion restriction 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 Wald test or the likelihood ratio test. To fit the two step model (3.32) 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 (5.36). This approach was taken by Luca and Peracchi (2006). We consider the maximum likelihood approach to this problem using the NDI scores.

Table 3: Probit model for dropout at months 8

	Missing at 8 months					
	Bivariate Probit			Individual Probit		
	Estimate	S.E.	p-value	Estimate	S.E.	p-value
int(unit)	1.085	0.005	0.000	1.019	0.124	0.000
age	0.002	0.000	0.000	0.017	0.005	0.002
sex(female)	0.015	0.006	0.011	0.117	0.138	0.398
trt(physio)	0.008	0.006	0.161	0.067	0.134	0.616
int(item)	1.599	0.045	0.000	0.841	0.100	0.000
age	-0.020	0.000	0.000	0.001	0.005	0.914
sex(female)	-0.302	0.008	0.000	-0.062	0.124	0.616
ρ_{23}	0.078	0.147	0.595			

The results in Table 3 show that there is conditional independence between unit and item non-response for the scores. This was further affirmed by the likelihood ratio test that compares the maximized values of the log-likelihood in (3) with the sum of the log-likelihoods for two simple probits models for unit and item non-response separately.

Table 4: Fit of Two-level selection models ($\rho_{23} \neq 0$) & $\rho_{23} = 0$), and Heckman selection model to the NDI scores at 8 months.

	2SNM ($\rho_{23} \neq 0$)			2SNM($\rho_{23} = 0$)			Heckman Selection		
	Estimate	S.E.	p-value	Estimate	S.E	p-value	Estimate	S.E	p-value
				Selection Equations					
int(unit)	0.872	0.005	0.000	0.872	0.005	0.000	0.804	0.115	0.000
age	-0.008	0.000	0.000	-0.008	0.000	0.000	0.001	0.004	0.938
sex(female)	-0.129	0.005	0.000	-0.127	0.005	0.000	-0.069	0.125	0.578
trt(physio)	0.044	0.005	0.000	0.042	0.005	0.000	0.085	0.122	0.489
int(item)	4.263	0.482	0.000	4.333	0.273	0.000			
age	0.012	0.035	0.745	0.004	0.018	0.830			
sex(female)	1.584	11.032	0.878	1.984	30.090	0.949			
ρ_{23}	0.656	0.810	0.419						
				Outcome Equation					
int	-0.294	0.055	0.000	-0.342	0.061	0.000	-0.260	1.498	0.862
age	0.096	0.001	0.000	0.094	0.001	0.000	0.082	0.027	0.003
sex(female)	0.658	0.031	0.000	0.641	0.031	0.000	0.571	0.722	0.429
trt(physio)	-0.354	0.030	0.000	-0.354	0.030	0.000	-0.418	0.716	0.560
baseline	0.628	0.002	0.000	0.626	0.002	0.000	0.626	0.052	0.000
wad2	-0.072	0.041	0.081	-0.107	0.041	0.009	-0.101	0.976	0.918
wad3	-0.487	0.056	0.000	-0.524	0.056	0.000	-0.517	1.343	0.701
σ	7.453	0.031	0.000	7.377	0.036	0.000	7.388	0.850	0.000
ρ_{12}	-0.500	0.016	0.000	-0.456	0.022	0.000	-0.460	0.503	0.361
ρ_{13}	-0.055	7.554	0.994	0.289	0.767	0.707			

Table 4 contains the results of a two-level sample selection model with $\rho_{23} \neq 0$ & $\rho_{23} = 0$, and the classical full information Heckman sample selection model where a single indicator is used for unit and item non-response. The results in the columns 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 3. Under the model with conditional independence ($\rho_{23} = 0$), separate probit models are used for unit and item missingness for the discrete components of the log-likelihood function given in (3.35). In a situation where the model is correctly specified, the model gives a pointer to the fact that the selection process may not be important, that is, ρ_{12} and ρ_{13} are of opposite sign and the bias due to unit and item non-response offset each other (see section 3). In addition, the classical sample

selection model also adduce to the fact that the selectivity generated by unit and item non-response is not different from zero.

6 Conclusion

Classical sample selection models and its 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.

The econometric literature usually assume a joint Gaussian error distribution for the outcome and the selection equations. By using properties of truncated normal distribution, the moment based estimator 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.

We have shown in this article that the sample selection models can be constructed either through the use of hidden truncation approach or conditioning in the multivariate normal distribution, and that the latter is a special case of the former in sample selection framework. In addition, it was established that the resulting distribution is the CSN distribution. Using the properties of CSN distribution, moment based estimator for any number of selection equations and with one outcome equation can readily be defined. This gives a unified method for studying more than two-level selection problems which is the current practice in econometric literature. We also emphasize that the density of the sample selection is comprised of a continuous component (CSN) and a discrete component. The model fitted to the discrete component is determined by the marginal distribution of the selection equations. If the marginal distribution is normal, the degree of observability in the discrete process determines the probit model to be fitted and was shown to depend on context.

A simulation study was conducted to assess the performance of the moment and the likelihood based estimators under two-level selection process. Consistent parameter estimates for the outcome models were obtained under the two methods. For the moment based method, the degree of censoring is slightly important. However, the model with 70% censored observations is as consistent as the one with 30% censored observation under the likelihood method. In the likelihood method, collapsing

the selection process into a single non-response indicator gave consistent parameter estimates for the outcome model. The single selection model needs to be correctly specified (a daunting exercise in practice), and there should be no interest in the two selection equations for this to be a reasonable model. Of course, the results from the classical Heckman model using the collapsed single non-response indicator tends not to work well when the Gaussian assumption is violated.

The NDI scores was analyzed using a multilevel sample selection model in which unit and item non-response (is assumed to) simultaneously affect the outcome of interest. Initial analysis showed that the unit ad item non-response are conditionally independent ($\rho_{23} = 0$). A model based on this assumption showed that the dependence between the unit missingness and the outcome model (ρ_{12}), and the dependence between the item missingness and the outcome model (ρ_{13}) are of opposite signs. This offset each other, implying that there are no selection bias. This was affirmed by using Heckman two-step method, where indicators for the unit and item non-response were collapsed to a single non-response indicator.

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, but will become identifiable from the additional information from the discrete component. However, it is advisable that exclusion restriction is used in the model regardless of the level of observability of the discrete process. The model has better prospects in observational studies and surveys where multilevel selection process are needed to be analyzed jointly and with information on likely variable that could potentially be responsible for a particular selection process included in the analysis.

Funding

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.

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

Arendt, J. N. and A. Holm (2006). Probit models with binary endogenous regressors. *Department of Business and Economics, University of Southern Denmark. Paper No.* **4 – 2006**, 1.

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

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

Bellio, R. and E. Gori (2003). Impact evaluation of job training programmes: Selection bias in multilevel models. *Journal of Applied Statistics* **30** : **8**, 893–907.

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

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 H. Li (1997). Inference for non-random samples. *J. R. Statist. Soc. B* **59**, 55–95.

Dominguez-Molina, J. A., G. Gonzalez-Farias, and R. Ramos-Quiroga (2004). Skew-normality in stochastic frontier analysis. In M. G. Genton (Ed.), *Skew-Elliptical Distributions and Their Applications: A Journey Beyond Normality*, pp. 223–241. Boca Raton, Florida: Chapman & Hall, CRC.

Gonzalez-Farias, G., J. A. Dominguez-Molina, and A. K. Gupta (2004). The closed skew-normal. In M. G. Genton (Ed.), *Skew-Elliptical Distributions and Their Applications: A Journey Beyond Normality*, pp. 25–42. Boca Raton, Florida: Chapman & Hall, CRC.

Greene, W. H. (2003). *Econometric analysis* (5 ed.). Upper Saddle River, NJ: Prentice-Hall.

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. *Annals of Economic and Social Measurement* **47**, 153–161.

Lamb, S. E., S. Gates, M. R. Underwood, M. W. Cooke, D. Ashby, A. Szczepura, M. A. Williams, E. M. Williamson, E. J. Withers, S. M. Isa, and A. Gumber (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.

Luca, G. D. and F. Peracchi (2006). *A sample selection model for unit and item nonresponse in cross-sectional surveys.* Last accessed July 20, 2010 at `http://www.share-project.org/t3/share/uploads/tx_sharepublications/deLuca_Peracchi_06.pdf`.

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

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

Robins, J. M. and R. D. Gill (1997). Non-response models for the analysis of non-monotone ignorable missing data. *Statistics in Medicine* **16**, 39–56.

Rosenman, R., B. Mandal, V. Tennekoon, and L. G. Hilll (2010). *Estimating treatment effectiveness with sample selection.* Last accessed December 31, 2011 at `http://faculty.ses.wsu.edu/WorkingPapers/Rosenman/WP2010-5.pdf`.

Vernon, H. (2009). *The Neck Disability Index: An instrument for measuring self-rated disability due to neck pain or whiplash-assocaited disorder.* Last accessed Febuary 20, 2010 at `http://www.cmcc.ca/Portals/0/PDFs/Research_05_2009_NDI_Manual.pdf`.

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