Applications and Applied Mathematics: An International Journal (AAM)

Volume 5 | Issue 2 Article 17

12-2010

A Multivariate Latent Variable Model for Mixed – Data from Continuous and Ordinal Responses with Possibility of Missing Responses

Ehsan B. Samani
Shahid Beheshti University

M. Ganjali
Shahid Beheshti University

Available at
http://pvamu.edu/aam
Appl. Appl. Math.
ISSN: 1932-9466

Applications and Applied Mathematics:
An International Journal
(AAM)

Vol. 05, Issue 2 (December 2010), pp. 467 – 487
(Previously, Vol. 05, Issue 10, pp. 1564 – 1584)

A Multivariate Latent Variable Model for Mixed – Data from Continuous and Ordinal Responses with Possibility of Missing Responses

Ehsan Bahrami Samani
Department of Statistics
Shahid Beheshti University
P.O. BOX 1983963113
Tehran, Iran
ehsan_bahrami_samani@yahoo.com

M. Ganjali
Faculty of Mathematical Science
Shahid Beheshti University
P.O. BOX 1983963113
Tehran, Iran

Received: August 28, 2010; Accepted: November 29, 2010

Abstract

A joint model for multivariate mixed ordinal and continuous outcomes with potentially non-random missing values in both types of responses is proposed. A full likelihood-based approach is used to obtain maximum likelihood estimates of the model parameters. Some modified Pearson residuals are also introduced where the correlation between responses are taken into account. The joint modelling of responses with the possibility of missing values requires caution since the interpretation of the fitted model highly depends on the missing mechanism assumptions that are unexaminable in a fundamental sense. A common way to investigate the influence of perturbations of model components on the key results of the analysis is to compare the results derived from the original and perturbed models using an influence maximal normal curvatures. For This, influence of a small perturbation of elements of the covariance structure of the model on maximal normal curvature is also studied. To illustrate the utility of the proposed model, a large data set excerpted from the British Household Panel Survey (BHPS) is analyzed. For these data, the simultaneous effects of some covariates on life satisfaction, income and the amount of money spent on leisure activities per month as three mixed correlated responses are explored.

Keywords: Joint Modelling; Latent Variable; Maximum Likelihood; Missing Responses; Mixed Ordinal and Continuous Responses

MSC 2000 No.: 62J

1. Introduction

Outcomes related to mixed correlated ordinal and continuous data, are pervasive and research on analyzing them needs to be promoted. The data used in this paper is extracted from wave 15 of the British Household Panel Survey (BHPS); a survey of adult Britons, being carried out annually since 1991 by the ESRC UK Studies Center with the Institute for Social and Economical Research at the University of Essex. In these data, life satisfaction status (ordinal response), Income (continuous response) and the amount of money spent on leisure activities per month including money spent on entertainment and hobbies (ordinal response) are the mixed correlated responses and the effect of explanatory variables on these responses should be investigated simultaneously. Consequently, we need to consider a method in which these variables can be modelled jointly.

For joint modelling of responses, one method is to use the general location model of Olkin and Tate (1961), where the joint distribution of the continuous and categorical variables is decomposed into a marginal multinomial distribution for the categorical variables and a conditional multivariate normal distribution for the continuous variables, given the categorical variables [for a mixed poisson and continuous responses where Olkin and Tate's method is used see Yang et al. (2007) and for joint modelling of mixed outcomes using latent variables see McCulloch (2007)].

A second method for joint modelling is to decompose the joint distribution as a multivariate marginal distribution for the continuous responses and a conditional distribution for categorical variables given the continuous variables. Cox and Wermuth (1992) empirically examined the choice between these two methods. The third method uses simultaneous modelling of categorical and continuous variables to take into account the association between the responses by the correlation between errors in the model for responses. For more details of this approach see, for example, Heckman (1978) in which a general model for simultaneously analyzing two mixed correlated responses is introduced and Catalano and Ryan (1992) who extended and used the model for a cluster of discrete and continuous outcomes

Rubin (1976), Little and Rubin (2002) and Diggle and Kenward (1994) made important distinctions between the various types of missing mechanisms for each of the above mentioned patterns. They define the missing mechanism as missing completely at random (MCAR) if missingness is dependent neither on the observed responses nor on the missing responses, and missing at random (MAR) if, given the observed responses, it is not dependent on the missing responses. Missingness is defined as non-random if it depends on the unobserved responses. From a likelihood point of view MCAR and MAR are ignorable but not missing at random (NMAR) is non-ignorable.

For mixed data with missing outcomes, Little and Schuchter (1987) and Fitzmaurice and Laird (1997) used the general location model of Olkin and Tate (1961) with the assumption of missingness at random (MAR) to justify ignoring the missing data mechanism. This means that they used all available responses, without a model for missing mechanism, to obtain parameter estimates using the EM (Expectation Maximization) algorithm. A model for mixed continuous

AAM: Intern. J., Vol. 05, Issue 2 (December 2010) [Previously, Vol. 05, Issue 10, pp. 1564 – 1584]

and discrete binary responses with possibility of missing responses is introduced by Ganjali (2003). Ganjali and Shafie (2006) present a transition model for an ordered cluster of mixed continuous and discrete binary responses with non-monotone missingness.

The aim of this paper is to use and extend an approach similar to that of Heckman (1978), which jointly models a nominal and a continuous variable, for joint modelling of multivariate ordinal and continuous outcomes. The model is described in terms of a correlated multivariate normal distribution for the underlying latent variables of ordinal and continuous responses with potentially non-random missing values in both types of responses. Some modified Pearson residuals are also presented to detect outliers where for accounting them the correlations between responses are taken into account.

The likelihood and the modified residuals are presented in section 2. A sensitivity analysis of the model is provided in Section 3. Application of the model on BHPS data is presented in section 4. Finally, some concluding remarks are given in Section 5.

2. Model, Likelihood and Residuals

2.1. Model and Likelihood

We use Y_{ij} to denote jth ordinal response with c_j levels for the ith individual defined as,

$$Y_{ij} = \begin{cases} 1 & Y_{ij}^* < \theta_{1,j}, \\ k+1 & \theta_{k,j} \le Y_{ij}^* < \theta_{k+1,j}, \quad k = 1,\dots,c_j - 2 \\ c_j & Y_{ij}^* \ge \theta_{c_j-1,j}, \end{cases}$$

where $i = 1,\dots,n$, $j = 1,\dots,M_1$. $\theta_{1j},\dots,\theta_{c_j-1,j}$ are the cut-point parameters and Y_{ij}^* denotes the underlying latent variable for Y_{ij}. The ordinal response vector and the continuous response vector for the ith individual are denoted by $Y_i = (Y_{i1},\dots,Y_{iM_1})'$ and $Z_i = (Z_{i(M_1+1)},\dots,Z_{iM})'$, respectively. Typically, when missing data occur in an outcome, assume $R_{y_i} = (R_{y_{i1}},\dots,R_{y_{iM_1}})'$ as the indicator vector of responding to Y_i, define $R_{y_{ij}}$ as

$$R_{y_{ij}} = \begin{cases} 1 & R_{y_{ij}}^* > 0 \\ 0 & \textit{Otherwise}, \end{cases}$$

assumes $R_{z_i} = (R_{z_{i(M_1+1)}},\dots,R_{z_{iM}})'$ as the indicator vector for responding to Z_i and define $R_{z_{ij}}$ as

$$R_{z_{ij}} = \begin{cases} 1, & R^*_{z_{ij}} > 0 \\ 0, & Otherwise, \end{cases}$$

where $R^*_{y_{ij}}$ and $R^*_{z_{ij}}$ denote the underlying latent variables of the non-response mechanisms for the ordinal and continuous variables, respectively.

The joint model takes the form:

$$\begin{aligned} Y^*_{ij} &= \beta_j{}'X_i + \varepsilon^{(1)}_{ij}, && j = 1,\ldots,M_1, \\ Z_{ij} &= \beta_{j'}X_i + \varepsilon^{(2)}_{ij}, && j = M_1+1,\ldots,M, \\ R^*_{y_{ij}} &= \alpha_{j'}X_i + \varepsilon^{(3)}_{ij}, && j = 1,\ldots,M_1, \\ R^*_{z_{ij}} &= \alpha_{j'}X_i + \varepsilon^{(4)}_{ij}, && j = M_1+1,\ldots,M, \end{aligned} \tag{1}$$

where X_i is the design matrix for the ith individual. In the model presented above, the vector of parameters $\eta = (\beta,\theta)'$ where $\beta = (\beta_1,\ldots,\beta_M,\alpha_1,\ldots,\alpha_M)'$ and $\theta = (\theta_{1,1},\ldots,\theta_{c_{M_1}-1,M_1})'$ should be estimated. The vector, β_j for $j = M_1+1,\ldots,M$, includes an intercept parameter but β_j, for $j = 1,\ldots,M_1$, due to having cutpoint parameters, are assumed not to include any intercept.

Let

$$\varepsilon_i = (\varepsilon^{(1)'}_i, \varepsilon^{(2)'}_i, \varepsilon^{(3)'}_i, \varepsilon^{(4)'}_i)' \overset{iid}{\sim} MVN(0,\Sigma_\varepsilon),$$

where

$$\varepsilon^{(u)}_i = (\varepsilon^{(u)}_{i1},\ldots,\varepsilon^{(u)}_{iM_1})', \text{ for } u = 1,3, \quad \varepsilon^{(u)}_i = (\varepsilon^{(u)}_{i(M_1+1)},\ldots,\varepsilon^{(u)}_{iM})', \text{ for } u = 2,4$$

and

$$\Sigma_\varepsilon = \begin{pmatrix} \Sigma^\varepsilon_{11} & \Sigma^\varepsilon_{12} & \Sigma^\varepsilon_{13} & \Sigma^\varepsilon_{14} \\ \Sigma^\varepsilon_{21} & \Sigma^\varepsilon_{22} & \Sigma^\varepsilon_{23} & \Sigma^\varepsilon_{24} \\ \Sigma^\varepsilon_{31} & \Sigma^\varepsilon_{32} & \Sigma^\varepsilon_{33} & \Sigma^\varepsilon_{34} \\ \Sigma^\varepsilon_{41} & \Sigma^\varepsilon_{42} & \Sigma^\varepsilon_{43} & \Sigma^\varepsilon_{44} \end{pmatrix},$$

where $\Sigma^\varepsilon_{uu} = Var(\varepsilon^{(u)}_i)$, for $u = 1,2,3,4$ and $\Sigma^\varepsilon_{uv} = Cov(\varepsilon^{(u)}_i, \varepsilon^{(v)}_i)$, $u < v$, $u,v = 1,2,3,4$ and $\Sigma^\varepsilon_{uv} = \Sigma^\varepsilon_{vu}{}'$. Because of identifiability problem we have to assume $Var(Y^*_{ij}) = Var(R^*_{y_{ij}}) = Var(R^*_{y_{ij}}) = 1$, so $\Sigma^\varepsilon_{jj} = diag\{1,\ldots1\}$ for $j = 1,3,4$.

AAM: Intern. J., Vol. 05, Issue 2 (December 2010) [Previously, Vol. 05, Issue 10, pp. 1564 – 1584]

Note, if one of the matrixes $\Sigma_{13}^{\varepsilon}, \Sigma_{14}^{\varepsilon}, \Sigma_{23}^{\varepsilon}, \Sigma_{24}^{\varepsilon}$ is not zero, then the missing mechanism of response is not at random.

For example when there is not any missing value for continuous response, let

$$Y_j^* = \beta_j{}'X + \varepsilon_j^{(1)}, \qquad j = 1, 2,$$
$$Z_3 = \beta_{3'}X + \varepsilon_3^{(2)},$$
$$R_{y_j}^* = \alpha_{j'}X + \varepsilon_j^{(3)}, \qquad j = 1, 2.$$

For this model the covariance matrix takes the form,

$$\Sigma = \begin{pmatrix} 1 & \rho_{12} & \sigma\rho_{13} & \rho_{14} & \rho_{15} \\ \rho_{21} & 1 & \sigma\rho_{23} & \rho_{24} & \rho_{25} \\ \sigma\rho_{31} & \sigma\rho_{32} & \sigma^2 & \sigma\rho_{34} & \sigma\rho_{35} \\ \rho_{14} & \rho_{24} & \sigma\rho_{34} & 1 & \rho_{45} \\ \rho_{15} & \rho_{25} & \sigma\rho_{35} & \rho_{45} & 1 \end{pmatrix},$$

where $\rho_{12} = corr(Y_1^*, Y_2^*)$, $\rho_{13} = corr(Y_1^*, Z)$, $\rho_{14} = corr(Y_1^*, R_{Y_1}^*)$, $\rho_{15} = corr(Y_1^*, R_{Y_2}^*)$, $\rho_{23} = corr(Y_2^*, Z)$, $\rho_{24} = corr(Y_2^*, R_{Y_1}^*)$, $\rho_{25} = corr(Y_2^*, R_{Y_2}^*)$, $\rho_{34} = corr(Z, R_{Y_1}^*)$, $\rho_{35} = corr(Z, R_{Y_2}^*)$ and $\rho_{45} = corr(R_{Y_1}^*, R_{Y_2}^*)$.

The likelihood of the model for $y_i = (y_{i1}, ..., y_{iM_1})'$, $z_i = (z_{i(M_1+1)}, ..., z_{iM})'$ and $x_i = (x_{i1}, ..., x_{ip})'$, where p is the number of explanatory variables for the ith individual (the number of components in this vector may also be dependent on the chosen variable, i.e x_i be x_{ij} and p be p_j, here, we ignore this for simplicity), $z = (z_1', ..., z_{n'})'$, $y = (y_{1'}, ..., y_n')'$, $x = (x_{1'}, ..., x_{n'})', J_{obs}^y = \{j : y_{ij} \text{ is observed}\}, J_{Mis}^y = (J_{obs}^y)^C$ $\text{ and } J_{obs}^z = \{j : z_{ij} \text{ is observed}\}, J_{Mis}^z = (J_{obs}^z)^C$, is:

$$L = \prod_{i=1}^{n} f(Y_{i,obs}, Z_{i,obs}, R_{y_i}, R_{z_i} \mid X_i)$$
$$= \prod_{i=1}^{n} f(Y_{i,obs}, R_{y_{i,obs}}, R_{z_{i,obs}}, C_{Mis}^* \mid Z_{i,obs}, X_i) f(Z_{i,obs} \mid X_i)$$
$$= \prod_{i=1}^{n} P(Y_{i,obs}^*, R_{y_{i,obs}}^*, R_{z_{i,obs}}^*, C_{Mis}^* \mid Z_{i,obs}, X_i) f(Z_{i,obs} \mid X_i),$$

where

$$
\begin{aligned}
Z_{i,obs} &= \{Z_{ij}, \forall j \in J^{z}_{obs}\},\\
Y_{i,obs} &= \{Y_{ij}, \forall j \in J^{y}_{obs}\},\\
Y^{*}_{i,obs} &= \{\theta_{y_{ij}-1} \le Y^{*}_{ij} \le \theta_{y_{ij}}, \forall j \in J^{y}_{obs}\},\\
R_{z_{i,obs}} &= \{R_{z_{ij}} = 1, \forall j \in J^{z}_{obs}\},\\
R_{y_{i,obs}} &= \{R_{y_{ij}} = 1, \forall j \in J^{y}_{obs}\},\\
R^{*}_{z_{i,obs}} &= \{R^{*}_{z_{ij}} > 0, \forall j \in J^{z}_{obs}\},\\
R^{*}_{z_{i,Mis}} &= (R^{*}_{z_{i,obs}})^{C},\\
R^{*}_{y_{i,obs}} &= \{R^{*}_{y_{ij}} > 0, \forall j \in J^{y}_{obs}\},\\
R^{*}_{y_{i,Mis}} &= (R^{*}_{y_{i,obs}})^{C},\\
R_{y_i} &= (R_{y_{i1}}, ..., R_{y_{iM_1}})',\\
R_{z_i} &= (R_{z_{i(M_1+1)}}, ..., R_{z_{iM}})'\\
C^{*}_{Mis} &= \{R_{y_{ij}} = 0; \forall j \in J^{y}_{Mis}, R_{z_{ij}} = 0; \forall j \in J^{z}_{Mis}\},
\end{aligned}
$$

and

$$P(Y^{*}_{i,obs}, R^{*}_{y_{i,obs}}, R^{*}_{z_{i,obs}}, C^{*}_{Mis} \mid Z_{i,obs}, X_i) = \gamma^{*}_{i} - \Gamma^{*}_{y_i} - \Gamma^{*}_{z_i} + \Gamma^{*}_{y_i z_i},$$

where

$$\gamma^{*}_{i} = P(Y^{*}_{i,obs}, C^{*}_{Mis} \mid Z_{i,obs}, X_i),$$

$$\Gamma^{*}_{z_i} = P(Y^{*}_{i,obs}, C^{*}_{Mis}, R^{*}_{z_{i,Mis}} \mid Z_{i,obs}, X_i),$$

$$\Gamma^{*}_{y_i} = P(Y^{*}_{i,obs}, C^{*}_{Mis}, R^{*}_{y_{i,Mis}} \mid Z_{i,obs}, X_i),$$

$$\Gamma^{*}_{y_i z_i} = P(Y^{*}_{i,obs}, C^{*}_{Mis}, R^{*}_{y_{i,Mis}}, R^{*}_{z_{i,Mis}} \mid Z_{i,obs}, X_i).$$

Details of extracting the above likelihood is given in Appendix A. This likelihood can be maximized by function “**nlminb**” in software ***R***. This function may be used for minimization of a function of parameters. For maximization of a likelihood function one may minimize minus log likelihood function. The function “**nlminb**” uses optimization method of port routine which is given in “*http://netlib.bell-labs.com/cm/cs/cstr/153.pdf*.”. The function “**nlminb**” uses a sequential quadratic programming (SQP) method to minimize the requested function. The details

AAM: Intern. J., Vol. 05, Issue 2 (December 2010) [Previously, Vol. 05, Issue 10, pp. 1564 – 1584]

of this method can be found in Fletcher (2000). The observed Hessian matrix may be obtained by "**nlminb**" function or may be provided by function "**fdHess**".

2.2. Residuals

The missing values of responses create problems for the usual residual diagnostics, [see Hirano et al. (1998)]. The Pearson residuals for continuous response, assuming MCAR, can take the form

$$r_{z_{ij}} = \frac{Z_{ij} - E(Z_{ij} \mid x_i)}{(Var(Z_{ij} \mid x_i))^{1/2}}. \tag{2}$$

These unconditional residuals will be misleading if missingness is MAR or NMAR as in these cases the values of the continuous variable come from a conditional or truncated conditional distribution. We can examine the residuals of the responses conditional on being observed. Let start with using theoretical form, involving $E(Z_{ij} \mid R_{zij} = 1)$, $E(Y_{ij} \mid R_{y_{ij}} = 1)$, $Var(Z_{ij} \mid R_{z_{ij}} = 1)$ and $Var(Y_{ij} \mid R_{y_{ij}} = 1)$ rather than their predicted values to define residuals. The Pearson residuals for continuous response can take the form

$$r_{z_{ij}} = \frac{Z_{ij} - E(Z_{ij} \mid R_{z_{ij}} = 1)}{(Var(Z_{ij} \mid R_{z_{ij}} = 1))^{1/2}}, \tag{2a}$$

and that for ordinal response can take the form

$$r_{y_{ij}} = \frac{Y_{ij} - E(Y_{ij} \mid R_{y_{ij}} = 1)}{[Var(Y_{ij} \mid R_{y_{ij}} = 1)]^{1/2}}. \tag{2b}$$

For example, when $M_1 = 1$, $c = 3$ and $M = 2$, we have

$$E(Z \mid R_Z = 1) = \beta_2' X_2 + \sigma\rho_{24} Q,$$
$$Var(Z \mid R_z = 1) = \sigma^2[(1-\rho_{24}^2) + \rho_{24}^2(1 - Q\alpha_2' X - Q^2)],$$
$$E(Y \mid R_Y = 1) = 3 + \frac{Q_{12} + Q_{11} - (Q_2 + Q_1)}{\Phi(\alpha_1' X)},$$
$$Var(Y \mid R_Y = 1) = \frac{15(Q_{12} + Q_{11} - (Q_2 + Q_1))}{\Phi(\alpha_1' X)} - [\frac{Q_{12} + Q_{11} - (Q_2 + Q_1)}{\Phi(\alpha_1' X)}]^2 - \frac{2Q_{12} - Q_2}{\Phi(\alpha_1' X)},$$

where $\phi(.)$, $\Phi(.)$ and $\Phi_{12}(.)$ are, respectively, the density function, cumulative univariate and cumulative bivariate normal distributions and $Q = \frac{\phi(-\alpha_2' X)}{\Phi(\alpha_2' X)}$, $Q_1 = \Phi(\theta_1 - \beta_{1'} X)$, $Q_2 = \Phi(\theta_2 - \beta_{1'} X)$, $Q_{11} = \Phi_{12}(\theta_1 - \beta_{1'} X, -\alpha_1' X; \rho_{13})$, $Q_{12} = \Phi_{12}(\theta_2 - \beta_{1'} X, -\alpha_1' X; \rho_{13})$ and $\rho_{13} = Corr(Y^*, R_y^*)$.

To define some appropriate modified residuals, let

$$\Sigma = \begin{pmatrix} \Sigma_{11,obs} & \Sigma_{12,obs} \\ \Sigma_{21,obs} & \Sigma_{22,obs} \end{pmatrix},$$

where

$$\Sigma_{11,obs} = Var(Y_i \mid R_{y_{iobs}}), \Sigma_{22,obs} = Var(Z_i \mid R_{z_{iobs}}) \text{ and } \Sigma_{12,obs} = \Sigma_{21,obs}' = Cov(Z_i, Y_i \mid R_{y_{iobs}}, R_{z_{iobs}})$$

are theoretical parameters, rather than their predicted values to define residuals, where

$$R_{z_{i,obs}} = \{R_{z_{ij}} = 1, \forall j \in J^z_{obs}\}, \quad R_{y_{i,obs}} = \{R_{y_{ij}} = 1, \forall j \in J^y_{obs}\}, \; J^y_{obs} = \{j : y_{ij} \text{ is observed}\},$$

and

$$J^z_{obs} = \{j : z_{ij} \text{ is observed}\}.$$

The Pearson residuals for the vector of continuous responses, assuming MAR and NMAR, can take the form

$$r^P_{iz} = \Sigma_{22,obs}^{-\frac{1}{2}}[Z_i - E(Z_i \mid R_{z_{iobs}})],$$

and that for the vector of ordinal response can take the form

$$r^P_{iy} = \Sigma_{11,obs}^{-\frac{1}{2}}[Y_i - E(Y_i \mid R_{y_{iobs}})]. \tag{2c}$$

The estimated Pearson residuals can be found by using the maximum likelihood estimates of the parameters, obtained by system (1), in (2a) and (2b). However, the residuals in equations (2a) and (2b) do not take into account the correlation between responses. These unconditional residuals will be misleading if missingness is MAR or NMAR as in these cases the values of the continuous variable come from a conditional or truncated conditional distribution. We can examine the residuals of the responses conditional on being observed. The estimated Pearson residuals $(\hat{r}_z, \hat{r}_y)$ can be found by using the maximum likelihood estimates of the parameters, obtained by system (1), in appendix B .

As residuals in appendix B are defined by conditioning on observing the responses, they differ from those of Ten Have et al. (1998), who found the expectation and variance of responses, and consequently the residuals, unconditional on the fact that responses should be observed. Then, to calculate the residuals they assumed no link between response and nonresponse mechanisms. This gives biased estimates of the means and variances of responses if missingness is not at random. However, the residuals in appendix B do not take into account the correlation between responses. Some modified Pearson residuals, defined in appendix B, are conditioned on the indicator of responses, and can take into account the correlation between responses.

3. Sensitivity Analysis

We used sensitivity analysis to study model output variation with changes in model inputs. Cook (1886) presented general methods for assessing the local influence of minor perturbations of a statistical model. Generally, one introduces perturbations into the model through the $q \times 1$ vector ω which is restricted to some open subset Ω of R^q. Let $L(\eta \mid \omega)$ denote the log- likelihood function corresponding to the perturbed model for a given ω in Ω. For a given set of observed data, where η is a $p \times 1$ vector of unknown parameters, we assume that there is an ω_0 in Ω such that $L(\eta) = L(\eta \mid \omega_0)$ for all η. Finally, Let $\hat{\eta}$ and $\hat{\eta}_\omega$ denote the maximum likelihood estimators under $L(\eta)$ and $L(\eta \mid \omega)$, respectively. To assess the influence of varying ω throughout Ω, we consider the Likelihood displacement defined as: $LD(\omega) = 2[L(\hat{\eta}) - L(\hat{\eta}_\omega)]$.

A graph of $LD(\omega)$ versus ω contains essential information on the influence of the pertubation scheme in questions. It is useful to view this graph as the geometric surface formed by the values of the $(q+1) \times 1$ vector $\alpha(\omega) = (\alpha_1, \alpha_2)' = (\omega, LD(\omega))'$ as ω varies throught Ω. When $q = 1$, the curvature of such plane curves at ω_0 is,

$$C = \frac{|\dot{\alpha}_1 \ddot{\alpha}_2 - \dot{\alpha}_2 \ddot{\alpha}_1|}{(\dot{\alpha}_1^2 - \dot{\alpha}_2^2)^{3/2}}, \tag{3a}$$

where the first and second derivations $\dot{\alpha}_i$ and $\ddot{\alpha}_i$ are evaluated at ω_0. Since $\dot{\alpha}_1 = 1$ and $\dot{\alpha}_2 = \ddot{\alpha}_1 = 0$, C redues to $C = \ddot{\alpha}_2 = \ddot{L}D(\omega_0)$. When $q > 1$, an influence graph is a surface in R^{q+1}. The normal curvature C_l of the lifted line in the driection l can now be obtained by applying (3.a) to the plan curve $(a, LD(\omega(a)))$, where $\omega(a) = \omega_0 + al$, $a \in R$, and l is a fixed nonzero vector of unit length in R^q. Cook (1986) proposed looking at local influences, i.e., at the normal curvatures C_l of $\alpha(\omega)$ in ω_0, in the direction of some q-dimensional vector l of unit length. Let Δ_i be the p -dimensional vector defined by $\Delta_i = \frac{\partial^2 l_i(\eta \mid \omega_i)}{\partial \omega_i \partial \eta}\Big|_{\eta=\hat{\eta},\omega_i=0}$ and define Δ as the $p \times n$ matrix with Δ_i as its ith column. Further, let $\ddot{L}$ denote the $p \times p$ matrix of second-order derivatives of $l(\eta \mid \omega_0)$ with respect to η, also evaluated at $\eta = \hat{\eta}$. Cook (1986) has then shown that C_l can be easily calculated by

$$C_l = 2\left| l^T \Delta^T (\ddot{L})^{-1} \Delta l \right|. \tag{3b}$$

Obviously, C_l can be calculated for any direction l. One evident choice is the vector l_i containing one in the $i\,th$ position and zero elsewhere, corresponding to the perturbation of the ith weight only.

For finding the condition for MAR, let $W = (Z,Y) = (W_{obs}, W_{mis})$, $W^* = (Z,Y^*) = (W^*_{obs}, W^*_{mis})$ and $R^* = (R^*_z, R^*_Y)$ where W^*_{obs} is the vector of latent variables related to the observed part of $W = (Z,Y)$, W^*_{mis} is the vector of latent variables related to the missing part of $W = (Z,Y)$.

According to our joint model, the vector of responses along with the missing indicators $(W^*, R^*) = (W^*_{obs}, W^*_{mis}, R^*)$ have a multivariate normal distribution with the following covariance structure,

$$\Sigma = \begin{pmatrix} \Sigma_{o,o} & \Sigma_{o,m} & \Sigma_{o,R^*} \\ \Sigma_{m,o} & \Sigma_{m,m} & \Sigma_{m,R^*} \\ \Sigma_{R^*,o} & \Sigma_{R^*,m} & \Sigma_{R^*,R^*} \end{pmatrix},$$

where

$$\Sigma_{o,o} = cov(W^*_{obs}, W^*_{obs}),$$
$$\Sigma_{m,m} = cov(W^*_{mis}, W^*_{mis}),$$
$$\Sigma_{o,m} = cov(W^*_{obs}, W^*_{mis}),$$
$$\Sigma_{o,R^*} = cov(W^*_{obs}, R^*),$$
$$\Sigma_{R^*,R^*} = cov(R^*, R^*).$$

The joint density function can also be partitioned as

$$f(W^*, R^*) = f(W^*_{mis}, R^* \big| W^*_{obs})\, f(W^*_{obs}),$$

AAM: Intern. J., Vol. 05, Issue 2 (December 2010) [Previously, Vol. 05, Issue 10, pp. 1564 – 1584]

where $f(W_{mis}^*, R^* | W_{obs}^*)$ and $f(W_{obs}^*)$ have, respectively, a conditional and a marginal normal distribution. According to the missing mechanism definitions, to have a MAR mechanism the covariance matrix of the above mentioned conditional normal distribution,

$$\Sigma_{m,R^*|o} = \text{cov}(W_{mis}{}^*, R^* | W_{obs}{}^*)$$
$$= \begin{pmatrix} \Sigma_{m,m} & \Sigma_{m,R^*} \\ \Sigma_{R^*,m} & \Sigma_{R^*,R^*} \end{pmatrix} - \begin{pmatrix} \Sigma_{m,o} & \Sigma_{R^*,o} \end{pmatrix}' \Sigma_{o,o}{}^{-1} \begin{pmatrix} \Sigma_{m,o} & \Sigma_{R^*,o} \end{pmatrix}$$
$$= \begin{pmatrix} \Sigma_{m,m} - \Sigma_{m,o}{}'\Sigma_{o,o}{}^{-1}\Sigma_{m,o} & \Sigma_{m,R^*} - \Sigma_{m,o}{}'\Sigma_{o,o}{}^{-1}\Sigma_{R^*,o} \\ \Sigma_{R^*,m} - \Sigma_{R^*,o}{}'\Sigma_{o,o}{}^{-1}\Sigma_{m,o} & \Sigma_{R^*,R^*} - \Sigma_{R^*,o}{}'\Sigma_{o,o}{}^{-1}\Sigma_{R^*,o} \end{pmatrix}$$

should satisfy the following constraint,

$$\Sigma_{m,R^*} - \Sigma_{m,o}^{'}\Sigma_{o,o}^{-1}\Sigma_{R^*,o} = 0. \tag{3c}$$

For our application (see section of application), we have missing values only for our ordinal variable and we may have $W_{mis}^* = (Y_1, Y_2)'$ and $W_{obs}^* = Z$, (Y_1 and Y_2 are ordinal responses and Z is continuous response.). For missing mechanism we only need to define $R^* = (R_{y_1}^*, R_{y_2}^*)'$, as we do not have any missing value for our continuous response, and

$$\Sigma_{m,m} = 1,\ \Sigma_{m,R^*} = \rho_{23},\ \Sigma_{R^*,R^*} = 1,$$
$$\Sigma_{o,o} = \sigma^2,\ \Sigma_{m,o} = \sigma\rho_{12},\ \Sigma_{R^*,o} = \sigma\rho_{13},$$

so that the above constraint will be reduced to,

$$\Sigma_{m,R^*} - \Sigma_{m,o}{}'\Sigma_{o,o}{}^{-1}\Sigma_{R^*,o} = \rho_{23} - \rho_{12}\rho_{13} = 0.$$

The vector $\omega = (\omega_1, \omega_2, \omega_3, \omega_4)'$, where $\omega_1 = \rho_{14} - \rho_{13}\rho_{34}$, $\omega_2 = \rho_{24} - \rho_{13}\rho_{35}$, $\omega_3 = \rho_{15} - \rho_{23}\rho_{34}$ and $\omega_4 = \rho_{25} - \rho_{23}\rho_{35}$, of weights defining the perturbation of the MAR model. If ω equal to $\omega_0 = (0,0,0,0)'$, the log-likelihood function would be the log-likelihood function of a MAR model. This reflects the influence of the condition for MAR of the responses. The corresponding local influence measure, denoted by C_i, then becomes $C_i = 2|\Delta_i^T \ddot{L}^{-1}\Delta_i|$. Another important direction is the direction l_{max} of maximal normal curvature C_{max}. It shows how to perturb the condition for MAR of the responses to obtain the largest local changes in C_{max} . C_{max} is the largest eigenvalue of $\Delta_i^T \ddot{L}^{-1}\Delta_i$ and l_{max} is the corresponding eigenvector.

An assessment of the influence of minor perturbations of the model is important (vide, cook, 1986). Likelihood displacement, global influence and maximal normal curvature C_{max} will be used for measuring the influence of a perturbation of the covariance structure of the model in system (1), on deviation from the assumption of MAR. For sensitivity analysis we calculate maximal normal curvature C_{max}. This curvature does not indicate extreme local sensitivity. Hence, final results of the model are not highly sensitive to the elements of the covariance structure.

4. Application

4.1. Data

The data used in this paper is excerpted from the 15th wave (2005) of the British Household Panel Survey (BHPS); a longitudinal survey of adult Britons, being carried out annually since 1991 by the ESRC UK Longitudinal Studies Center with the Institute for Social and Economical Research at the University of Essex. These data are recorded for 11251 individuals. The selected variables which will be used in this application are explained in the following.

One of the responses is the life Satisfaction (LS), [where the related question is *QA: "How dissatisfied or satisfied are you with your life overall?"*], which is measured by directly asking the level of an individual's satisfaction with life overall, resulting in a three categories of ordinal variable [1: Not satisfied at all (10.300%). 2: Not satis/dissat (45.400%) and 3: Completely satisfied (44.300 %)].

The amount of money spent on leisure activities per month including money spent on entertainment and hobbies (AM) is also measured [where the related question is *QB: "Please look at this Response categories/range and tell me about how much you personally spend in an average month on leisure activities, and entertainment and hobbies, other than eating out?"*] as an ordinal response with three categories, [0: Nothing (17.515 %). 1: Under 50 Pound (53.449%) and 2: 50 Pound or over (29.036%.)].

In our application, the percentage of missing values of LS and AM are 5.000% and 4.000%, respectively. Moreover, the exact amount of an individuals annual income (INC) in the past year in thousand pounds, considered here in the logarithmic scale, is also excerpted as a continuous response variable (mean: 4.068). As some values of annual income in thousand pounds are between 0 and 1, some of the logarithms of incomes are less than 0.

These three responses, LS, AM and logarithm of income are endogenous correlated variables and should be modelled as a multivariate vector of responses.

Socio-demographic characteristics, namely: Gender (male: 44.200% and female: 55.800%), Marital Status (MS) [married or living as couple: 68.500%, widowed: 8.300%, divorced or separated: 8.400% and never married: 14.800%], Age (mean: 49.180) and Highest Educational Qualification (HEQ) [higher or first degree: 15.100%, other higher QF: 64.600%, other QF: 2.000% and no qualification: 18.300%] are also included in the model as covaiates.

AAM: Intern. J., Vol. 05, Issue 2 (December 2010) [Previously, Vol. 05, Issue 10, pp. 1564 – 1584]

The vector of explanatory variables is $X = (Gender, Age, MS_1, MS_2, MS_3, HEQ_1, HEQ_2, HEQ_3)$ where MS_1, MS_2 and MS_3 are dummy variables for married or living as couple, widowed and divorced or separated, respectively, and HEQ_1, HEQ_2, HEQ_3 are dummy variables for higher or first degree, other higher QF and other QF, respectively. A Kolmogorov - Smirnov test does not lead to rejection of the assumption of normality of log (INC) (P-value=0.777). A direct acyclic graph of modeling three correlated responses, given the vector of exogenous covariates, is given in figure 1. The main approach of analysing these data is a multivariate multiple regression with some missing response values.

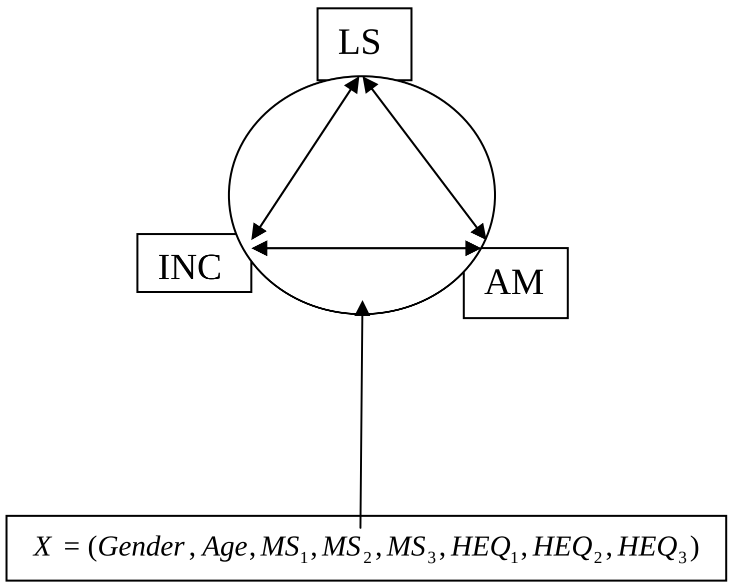

Figure 1: Direct acyclic graph of modeling LS= Life satisfaction, AM= The amount of money spent on leisure and INC= Income as three correlated responses, X =(Gender, AGE, HEQ (HEQ1, HEQ2, HEQ3) , MS(MS1, MS2, MS3)) is the vector of covariates.

4.2. Models for BHPS Data

We apply the model described in section 2 to evaluate the effect of Age, Gender, HEQ and MS simultaneously on LS, AM and Income. We shall also try to find answers for some questions, including (1) do male's LS, AM and income differ from female's? (2) How does HEQ affect the three responses? (3) do significant correlations exist between three responses? (4) what would be the consequence of not considering these correlations? (5) Is the missing data mechanism for LS and AM at random (MAR)?

For comparative purposes, two models are considered. The first model (model I) is a marginal model which does not consider the correlation between three responses and can be presented as,

$$LS^* = \beta_{11}\, MS_1 + \beta_{12}\, MS_2 + \beta_{13}\, MS_3 + \beta_{14}\, HEQ_1 + \beta_{15}\, HEQ_2 + \beta_{16}\, HEQ_3 + \beta_{17} Gender + \beta_{18} AGE + \varepsilon_1$$

$$AM^* = \beta_{21}\, MS_1 + \beta_{22}\, MS_2 + \beta_{23}\, MS_3 + \beta_{24}\, HEQ_1 + \beta_{25}\, HEQ_2 + \beta_{26}\, HEQ_3 + \beta_{27} Gender + \beta_{28} AGE + \varepsilon_2$$

$$\log(INC) = \beta_{30} + \beta_{31}\ MS_1 + \beta_{32}\ MS_2 + \beta_{33}\ MS_3 + \beta_{34}\ HEQ_1 + \beta_{35}\ HEQ_2 + \beta_{36}\ HEQ_3 + \beta_{37} Gender + \beta_{38} AGE + \varepsilon_3$$

$$R^*_{LS} = \alpha_{11}\ MS_1 + \alpha_{12}\ MS_2 + \alpha_{13}\ MS_3 + \alpha_{14}\ HEQ_1 + \alpha_{15}\ HEQ_2 + \alpha_{16}\ HEQ_3 + \alpha_{17} Gender + \alpha_{18} AGE + \varepsilon_4.$$

$$R^*_{AM} = \alpha_{21}\ MS_1 + \alpha_{22}\ MS_2 + \alpha_{23}\ MS_3 + \alpha_{24}\ HEQ_1 + \alpha_{25}\ HEQ_2 + \alpha_{26}\ HEQ_3 + \alpha_{27} Gender + \alpha_{28} AGE + \varepsilon_5.$$

Above, in the third equation the logarithm is taken in the base e. The second model (model II) uses model I and takes into account the correlations between five errors. Here, a multivariate normal distribution with correlation parameters $\rho_{12}, \rho_{13}, \rho_{14}, \rho_{15}, \rho_{23}, \rho_{24}, \rho_{25}$, ρ_{34}, ρ_{35} and ρ_{45} is assumed for the errors and these parameters should be also estimated.

4.3. Results

Results of using two models are presented in Table 1. Deviance for testing model (I) against model (II) is equal to 1636.09 with 10 degrees of freedom (P-value < 0.0001) which indicates that model (II) has a better fit to these data. As it can be seen, two correlation parameters ρ_{12} and ρ_{23} are strongly significant. They show a positive correlation between LS and AM (ρ_{12} = 0.137, P-value < 0.0001) and a positive correlation between log(INC) and AM (ρ_{23} = 0.134, P-value < 0.0001). Consideration of the responses associations yields more precise estimates as indicated by the smaller variance estimates and the smaller estimated variance of log(INC) in model (II). So, we restrict our interpretation to the results of model (II).

Model (II) shows a significant effect of age (the older the individual the more the life satisfaction), MS (married people are more satisfied than never married people and divorced or separated people are less satisfied than never married people), HEQ (the higher the qualification the higher the life satisfaction) and gender (males are more satisfied than females) on the life satisfaction status. All explanatory variables have significant effects on the ordinal response of amount of money spent on leisure activities. Never married people spend more on leisure time activities than other people.

The higher the education the more the leisure time activities. Females spend more amount of money than males for leisure time and older people spend less money than younger ones. Also the effect of all explanatory variables are significant on the logarithm of income. Parameter estimates indicate that as the degree of educational qualification increases log(INC) increases. Never married people have less logarithm of income than married people and divorced or separated people. Females have more logarithm of income than males and the older people earn less money than younger ones.

By these results we can conclude that the two responses are correlated and also the missing indicator for LS and AM are not related to three responses. This leads to have a missing completely at random mechanism. Using residuals defined in equation (Appendix B, B1) computed by parameter estimates of model II, we have not found any abnormal observation. To

AAM: Intern. J., Vol. 05, Issue 2 (December 2010) [Previously, Vol. 05, Issue 10, pp. 1564 – 1584] 481

perform the sensitivity analysis we find C_{max} . This is confirmed by the curvature $C_{max} = 1.02$ computed from (3.b). This curvature does not indicate extreme local sensivity.

Table 1. Results of using two models for BHPS data (LS: Life Satisfaction, AM: Amount of Money spent, MS: Marital Statue and HEQ: educational qualification, parameter estimates highlighted in **bold** are significant at 5 % level.)

	MODEL I		MODEL II	
parameter	Est.	S.D.	Est.	S.D.
LS				
MS (baseline: Never married)				
Married or Living as Couple	**0.333**	**0.053**	**0.201**	**0.030**
Widowed	0.071	0.096	0.033	0.038
Divorced or Separated	**-0.622**	**0.080**	**-0.357**	**0.044**
HEQ (baseline: No QF)				
Higher or First Degree	**- 0.090**	**0.378**	**-0.061**	**0.095**
Other higher QF	**-0.205**	**0.373**	**-0.132**	**0.095**
Other QF	**- 0.297**	**0.373**	**-0.211**	**0.100**
Gender (baseline: Female)				
Male	**0.069**	**0.036**	**0.040**	**0.024**
AGE	**0.007**	**0.001**	**0.008**	**0.001**
cutpoint 1	**-1.661**	**0.378**	**-0.987**	**0.101**
cutpoint 2	**0.785**	**0.379**	**0.450**	**0.118**
AM				
Married or Living as Couple	**-0.324**	**0.056**	**-0.186**	**0.032**
Widowed	**-0.400**	**0.095**	**-0.233**	**0.054**
Divorced or Separated	**-0.440**	**0.082**	**-0.257**	**0.047**
HEQ (baseline: No QF)				
Higher or First Degree	**0.921**	**0.387**	**0.579**	**0.093**
Other higher QF	**0.445**	**0.385**	**0.289**	**0.090**
Other QF	**-0.195**	**0.389**	**-0.077**	**0.088**
Male	**-0.780**	**0.038**	**-0.449**	**0.021**
AGE	**-0.024**	**0.001**	**-0.013**	**0.001**
cutpoint 1	**-4.052**	**0.394**	**-2.330**	**0.105**
cutpoint 2	**-1.290**	**0.394**	**-0.690**	**0.122**
logINC				
Constant	**4.245**	**0.078**	**4.245**	**0.039**
MS (baseline: Never married)				
Married or Living as Couple	**0.114**	**0.011**	**0.113**	**0.011**
Widowed	**0.217**	**0.019**	**0.216**	**0.019**
Divorced or Separated	**0.215**	**0.077**	**0.216**	**0.016**
HEQ (baseline: No QF)				
Higher or First Degree	**0.391**	**0.077**	**0.393**	**0.036**
Other higher QF	**0.177**	**0.077**	**0.178**	**0.035**
Other QF	0.031	0.077	0.033	0.035
Gender (baseline: Female)				
Male	**-0.227**	**0.007**	**-0.225**	**0.007**
AGE	**-0.002**	**0.001**	**-0.003**	**0.001**
σ^2	**0.180**	**0.002**	**0.155**	**0.002**

Table 1 (continue): Results of using two models for BHPS data (LS: Life Satisfaction, AM: Amount of Money spent, MS: Marital Statue and HEQ: educational qualification, parameter estimates highlighted in **bold** are significant at 5 % level.)

	MODEL I		MODEL II	
parameter	**Est.**	**S.D.**	**Est.**	**S.D.**
R^*_{LS}				
MS (baseline: Never married)				
Married or Living as Couple	**0.017**	**0.004**	**0.016**	**0.003**
Widowed	**-0.003**	**0.008**	**-0.004**	**0.007**
Divorced or Separated	**0.020**	**0.007**	**0.002**	**0.016**
HEQ (baseline: No QF)				
Higher or First Degree	**0.056**	**0.027**	**0.054**	**0.018**
Other higher QF	**0.051**	**0.027**	**0.051**	**0.017**
Other QF	**0.024**	**0.027**	**0.023**	**0.018**
Gender (baseline: Female)				
Male	**-0.005**	**0.003**	**-0.004**	**0.018**
AGE	**-0.008**	**0.001**	**0.016**	**0.002**
R^*_{AM}				
MS (baseline: Never married)				
Married or Living as Couple	**0.019**	**0.003**	**0.016**	**0.002**
Widowed	**-0.005**	**0.009**	**-0.005**	**0.008**
Divorced or Separated	**0.031**	**0.07**	**0.020**	**0.021**
HEQ (baseline: No QF)				
Higher or First Degree	**0.065**	**0.021**	**0.043**	**0.013**
Other higher QF	**0.050**	**0.041**	**0.051**	**0.050**
Other QF	**0.025**	**0.030**	**0.021**	**0.019**
Gender (baseline: Female)				
Male	**-0.007**	**0.001**	**-0.005**	**0.018**
AGE	**-0.007**	**0.001**	**0.019**	**0.001**
Corr(LS^*,AM^*)	-	-	**0.137**	**0.012**
Corr (LS^*,INC)	-	-	**0.134**	**0.001**
Corr (LS^*, R^*_{LS})	-	-	**0.001**	**0.001**
Corr (LS^*, R^*_{AM})	-	-	**0.001**	**0.001**
Corr(AM^*,INC)	-	-	**0.013**	**0.012**
Corr(AM^*, R^*_{LS})	-	-	**0.001**	**f0.012**
Corr(AM^*, R^*_{AM})	-	-	**0.004**	**0.010**
Corr(INC, R^*_{LS})	-	-	**0.003**	**0.016**
Corr(INC, R^*_{AM})	-	-	**0.014**	**0.015**
Corr(R^*_{LS}, R^*_{AM},)	-	-	**0.009**	**0.022**
-Loglike	**35924.076**		**35106.031**	

5. Discussion

In this paper a multivariate latent variable model is presented for simultaneously modelling of ordinal and continuous correlated responses with potentially non-random missing values in both types of responses. Some modified residuals are also presented to detect outliers. We assume a multivariate normal distribution for errors in the model. However, any other multivariate distribution such as t or logistic can be also used. Binary responses are a special case of ordinal responses. So, our model can also be used for mixed binary and continuous responses. For correlated nominal, ordinal and continuous responses Deleon and Carriere (2007) have developed a model by extending general location model. Generalization of our model for nominal, ordinal and continuous responses is an ongoing research on our part.

REFERENCES

Ash, R. B. (2000). Probability and Measure Theory. Academic Press. London, UK.

Catalano, P. and Ryan, L. M. (1992). Bivariate latent variable models for clustered discrete and continuous outcoms. *Journal of the American Statistical Association*, **50**, 3, 1078-1095.

Cook, R. D. (1986). Assessment of local influence. *Royal Statistical Society* ,48, 133-169.

Cox, D. R, and Wermuth, N. (1992). Response models for mixed binary and quantitative variables. *Biometrika*, **79**(3): 441-461 .

De Leon, A. R. and Carrire, K. C. (2007). General mixed-data model: Extension of general location and grouped continuous models. *Canadian Journal of Statistics*, 35(4), 533-548.

Diggle, P.J. and Kenward, M. G. (1994). Informative Drop-out in longitudinal data analysis. *Journal of Applied Statistics*, **43**, 49-93.

Fletcher, R. (2000). Practical Methods of optimization. John Wiley.

Fitzmaurice, G. M. and Laird, N. M. (1997). Regression models for mixed discrete and continuous responses with potentially missing values. *Biometrics*, **53**, 110-122 .

Ganjali, M. (2003). A model for mixed continuous and discrete responses with possibilty of missing responses. *Journal of Sciences, Islamic Republic of Iran* . **14**(1): 53-60.

Ganjali, M. and Shafie, K. (2006). A transition model for an ordered cluster of mixed continuous and discrete responses with non- monotone missingness. *Journal of Applied Statistical Sciences* **15**(4): 331-343.

Heckman, J. J. D. (1978). Endogenous variable in a simutaneous Equations system. E*conometrica*, **46**(6): 931-959.

Hirano, K. Imbens, G. Ridder, G. and Rubin, D. (1998). Combining Panel data sets with attrition and refreshment samples. Technical working paper, National Bureau of Economic Research, Cambridge, Massachuset.

Little, R. J. and Rubin, D. (2002). Statistical analysis with missing data. Second edition. New york, Wiley.

Little, R. J. and Schluchter M. (1987). Maximum likelihood estimation for mixed continuous and categorical data with missing values. *Biometrika*, 72, 497-512.
McCulloch, C. (2007). Joint modelling of mixed outcome type using latend variables. S*tatistical methods in Medical Research*. ***17(***1), 53-73.
Olkin, L. and Tate, R. F. (1961). Multivariate correlation models with mixed discrete and continuous variables. *Annals of Mathematical Statistics*, **32**: 448-456.
Rubin, D. B. (1976). Inference and missing data. *BiometriKa*, 82, 669-710.
Ten Have, T. R. T., Kunselman, A. R., Pulkstenis, E. P., and Landis, J. R. (1998). Mixed effects logistic regression models for longitudinal binary response data with informative dropout. *Biometrics*, 54, 367-383.
Yang, Y., Kang, J., Mao, K. and Zhang, J. (2007), Regression models for mixed Poisson and continuous longitudinal data. *Statistics in Medicine*; **26**: 3782-3800.

Appendix A

Details of the likelihood For Mixed-Data from Ordinal and continuous Data With Missing Responses:

We apply the following definition and theorem for multivariate distributions to obtain the form of the likelihood.

Definition 2.2.1. If

$F(w_1,\cdots,w_{M_1}) = P(W_1^* \le w_1,...,W_{M_1}^* \le w_{M_1})$ is a distribution function, for $(a_j \le b_j)$ the operator $\Delta_{b_j a_j} F(w_1,\cdots,w_{M_1})$ is defined as
$F(w_1,\cdots,w_{(j-1)},b_j,w_{(j+1)},...,w_{M_1}) - F(w_1,\cdots,w_{(j-1)},a_j,w_{(j+1)},...,w_{M_1})$.

Theorem 2.2.2. If for $j = 1,...,M_1$, $a_j \le b_j$, then

$$P(a_1 < W_1^* \le b_1,...,a_{M_1} < W_{M_1}^* \le b_{M_1}) = \Delta_{b_1 a_1} \cdots \Delta_{b_{M_1} a_{M_1}} F(w_1,\cdots,w_{M_1}),$$

where

$$\Delta_{b_1 a_1} \cdots \Delta_{b_{M_1} a_{M_1}} F(w_1,\cdots,w_{M_1}) = F_0 - F_1 + F_2 - ... + (-1)^{M_1} F_{M_1};$$

and F_j is the sum of all $\binom{M_1}{j}$ terms of the from $F(g_1,...,g_{M_1})$ with $g_k = a_k$ for exactly j integers in $\{1,...,M_1\}$, and $g_k = b_k$ for the remaining $M_1 - j$ integers.

Proof:

See Ash (2000, p. 28).

AAM: Intern. J., Vol. 05, Issue 2 (December 2010) [Previously, Vol. 05, Issue 10, pp. 1564 – 1584] 485

Suppose the g_1- elements of Y_i and g_2- elements of Z_i are observed, so $J^y_{obs}=\{o_1,...,o_{g_1}\}$ and $J^z_{obs}=\{o_1,...,o_{g_2}\}$. Using the Theorem 2.2.2,

$$\gamma_i^*=\Delta_{b_{o_1}a_{o_1}}...\Delta_{b_{og_1}a_{og_1}}P(Y^*_{io_1}\le\omega_{io_1},...,Y^*_{iog_1}\le\omega_{iog_1},C^*_{Mis}\,|\,Z_{i,obs},X_i)$$
$$=F^{(1)}_{i0}-F^{(1)}_{i1}+F^{(1)}_{i2}-...+(-1)^{g_1}F^{(1)}_{ig_1}$$

$$\Gamma^*_{z_i}=\Delta_{b_{o_1}a_{o_1}}...\Delta_{b_{og_1}a_{og_1}}P(Y^*_{io_1}\le\omega_{io_1},...,Y^*_{iog_1}\le\omega_{iog_1},C^*_{Mis},R^*_{z_{i,Mis}}\,|\,Z_{i,obs},X_i)$$
$$=F^{(2)}_{i0}-F^{(2)}_{i1}+F^{(2)}_{i2}-...+(-1)^{g_1}F^{(2)}_{ig_1}$$

$$\Gamma^*_{y_i}=\Delta_{b_{o_1}a_{o_1}}...\Delta_{b_{og_1}a_{og_1}}P(Y^*_{io_1}\le\omega_{io_1},...,Y^*_{iog_1}\le\omega_{iog_1},C^*_{Mis},R^*_{y_{i,Mis}}\,|\,Z_{i,obs},X_i)$$
$$=F^{(3)}_{i0}-F^{(3)}_{i1}+F^{(3)}_{i2}-...+(-1)^{g_1}F^{(3)}_{ig_1}$$

$$\Gamma^*_{y_iz_i}=\Delta_{b_{o_1}a_{o_1}}...\Delta_{b_{og_1}a_{og_1}}P(Y^*_{io_1}\le\omega_{io_1},...,Y^*_{iog_1}\le\omega_{iog_1},C^*_{Mis},R^*_{y_{i,Mis}},R^*_{z_{i,Mis}}\,|\,Z_{i,obs},X_i)$$
$$=F^{(4)}_{i0}-F^{(4)}_{i1}+F^{(4)}_{i2}-...+(-1)^{g_1}F^{(4)}_{ig_1},$$

where

$b_{o_j}=\theta_{o_j y_{io_j}}$, $a_{o_j}=\theta_{o_j(y_{io_j}-1)}$ for $j=1,...,g_1$ and $F^{(1)}_{ij},F^{(2)}_{ij},F^{(3)}_{ij}$ and $F^{(4)}_{ij}$ are the sum of all $\binom{g}{j}$ terms of the from:

$$P(Y^*_{ic_1}\le c_1,...,Y^*_{ic_{g_1}}\le c_{g_1},C^*_{Mis}\,|\,Z_{i,obs},X_i),$$

$$P(Y^*_{ic_1}\le c_1,...,Y^*_{ic_{g_1}}\le c_{g_1},C^*_{Mis},R^*_{z_{i,Mis}}\,|\,Z_{i,obs},X_i),$$

$$P(Y^*_{ic_1}\le c_1,...,Y^*_{ic_{g_1}}\le c_{g_1},C^*_{Mis},R^*_{y_{i,Mis}}\,|\,Z_{i,obs},X_i),$$

$$P(Y^*_{ic_1}\le c_1,...,Y^*_{ic_{g_1}}\le c_{g_1},C^*_{Mis},R^*_{y_{i,Mis}},R^*_{z_{i,Mis}}\,|\,Z_{i,obs},X_i),$$

with $c_k=a_k$ for exactly j integers in $\{1,...,g_1\}$, and $c_k=b_k$ for the remaining g_1-j integers. With these the likelihood is given in Section 2.

Appendix B
Residuals

Suppose $M_1 = 2, c = 3$ and $M = 3$. To obtain the correlated modified residuals when y_{i2} and Z_{i3} responses are observed, let $Y_i = (Y_{i1}, Y_{i2})'$ and $Z_i = Z_{i3}$

$$r_i^P = \hat{\Sigma}^{-\frac{1}{2}}(K_i - \hat{\mu}_i), \tag{B1}$$

where

$$K_i = (Y_i, Z_i)' \ , \ \ \hat{\mu}_i = (\hat{E}(Y_i \mid R_{y_{i1}} = 1), \hat{E}(Z_i \mid R_{y_{i1}} = 1))'$$

and

$$\hat{\Sigma} = \begin{pmatrix} \hat{V}ar(Y_i \mid R_{y_{i1}} = 1) & \hat{C}ov(Y_i, Z_i \mid R_{y_{i1}} = 1) \\ \hat{C}ov(Y_i, Z_i \mid R_{y_{i1}} = 1) & \hat{V}ar(z_i \mid R_{y_{i1}} = 1) \end{pmatrix},$$

where

$$\begin{aligned} Cov(Y_i, Z_i \mid R_{y_{i1}} = 1) &= E(Y_i Z_i \mid R_{y_{i1}} = 1) - E(Z_i \mid R_{y_{i1}} = 1)E(Y_i \mid R_{y_{i1}} = 1) \\ &= E(Z_i E(Y_i \mid z_i, R_{y_{i1}} = 1)) - E(Z_i \mid R_{y_{i1}} = 1)E(Y_i \mid R_{y_{i1}} = 1) \\ &= \int z_i \ E(Y_i \mid z_i, R_{y_{i1}} = 1) \ f_{Z_i}(z_i \mid R_{y_{i1}} = 1) dz_i - E(Z_i \mid R_{y_{i1}} = 1)E(Y_i \mid R_{y_{i1}} = 1) \\ &= \int [z_i (\sum_{y_i} y_i \ P(Y_i = y_i \mid z_i, R_{y_{i1}} = 1))) \ f_{Z_i}(z_i \mid R_{y_{i1}} = 1) dz_i] \\ &\quad -[\int z_i f_{z_i}(z_i \mid R_{y_{i1}} = 1) \ dz_i][\sum_{y_i} y_i \ P(Y_i = y_i \mid R_{y_{i1}} = 1))]. \end{aligned} \tag{B2}$$

The Pearson residuals for the *ith* observation is based on the Pearson goodness-of-fit statistics

$$\chi^2 = \sum_{i=1}^{n} \chi_p^2(K_i, \hat{\mu}_i)$$

with the following *ith* component

$$\chi_p^2(K_i, \hat{\mu}_i) = (K_i - \hat{\mu}_i)' \hat{\Sigma}^{-1} (K_i - \hat{\mu}_i).$$

A cholesky decomposition is used for finding the square root of $\hat{\Sigma}_\rho$ in (B1) and the function integrate in ***R*** is used to numerically calculate the integral given in (B2).