REVSTAT – Statistical Journal
Volume 9, Number 1, March 2011, 37–55

MISSING DATA IN REGRESSION MODELS FOR NON-COMMENSURATE MULTIPLE OUTCOMES

Authors: Armando Teixeira-Pinto
– Serviço de Bioestatística e Informática Médica,
CINTESIS, Faculdade de Medicina, Universidade do Porto,
Portugal
tpinto@post.harvard.edu

Sharon-Lise Normand
– Department of Health Care Policy, Harvard Medical School,
and Department of Biostatistics, Harvard School of Public Health,
Boston, USA
normand@hcp.med.harvard.edu

Abstract:

- Biomedical research often involves the measurement of multiple outcomes in different scales (continuous, binary and ordinal). A common approach for the analysis of such data is to ignore the potential correlation among the outcomes and model each outcome separately. This can lead not only to loss of efficiency but also to biased estimates in the presence of missing data. We address the problem of missing data in the context of multiple non-commensurate outcomes. The consequences of missing data when using likelihood and quasi-likelihood methods are described, and an extension of these methods to the situation of missing observations in the outcomes is proposed. Two real data examples illustrate the methodology.

Key-Words:

- *mixed outcomes; multivariate; latent variable; non-commensurate; missing data; maximum likelihood; direct maximization; weighted generalized estimating equations.*

AMS Subject Classification:

- 62J12, 62H99.

1. INTRODUCTION

Many biomedical studies involve measurements of multiple outcomes on each subject. When the outcomes are commensurate, i.e., are measured on the same scale and are measuring the same underlying variable, classical tools of multivariate statistics can be used. However, multivariate methods to analyze outcomes measured on different scales or measuring different underlying variables, i.e., non-commensurate outcomes, are less common and rarely used in data analysis. A common solution used in the presence of non-commensurate outcomes is to analyze each outcome separately, ignoring the potential correlation between the outcomes. There are several disadvantages of this approach. First, there might be a loss of efficiency by ignoring the extra information contained in the correlation between the outcomes. Second, with separate analysis it is harder to answer intrinsic multivariate questions such as the existence of a covariate effect on the underlying outcome. Third, if some outcomes are missing for some individuals, different samples of individuals will be included in the analysis of the effect of exposure on different outcomes. Finally, the situation of missing data may also produce biased results if the missing data depends on the other outcomes.

The main difficulty of modeling non-commensurate variables is that there is no obvious multivariate distribution. Mainly, three approaches to model non-commensurate outcomes have been described in the literature. The first has its roots in the general location model ([10]) and has been extended to accommodate covariates ([2]) and clustered data ([6], [12]). The key idea is to factorize the likelihood as the product of marginal and conditional distributions, and model each term of the product. However, this approach does not generalize easily when the number of outcomes is increased. The second approach uses latent variables to induce the correlation between the outcomes and assumes that conditional on these latent quantities, the outcomes are independent ([14], [17], [5]). The third approach extends the framework of generalized estimating equations (GEE) to multivariate discrete and continuous outcomes ([11], [20], [19]). The main advantages of the GEE over likelihood methods is the lack of assumptions regarding the distribution of the data and its robustness to misspecification of the correlation between the outcomes. Naturally this will lead to less efficient but more robust estimates (see Teixeira-Pinto and Normand ([19]) for a summary of these and other approaches).

With the measurement of multiple outcomes there is a higher risk of missing data. Few authors have addressed the problem of missing data in the setting of non-commensurate outcomes. Fitzmaurice and Laird ([7]) proposed the use of the EM-algorithm ([3]) to fit the extension of the general location model in the presence of missing data. Shafer ([15]) described likelihood-based data

augmentation approaches to missing data assuming a general location model. In Little and Rubin's ([9]) nomenclature, the missing data is defined as missing at random (MAR) if it only depends on the observed data. If the missing data does not depend on the observed or unobserved data, the missing data is designated as missing completely at random (MCAR). In contrast, if the the missing data depends on unobserved data, the missing mechanism is said to be missing not at random (MNAR). The GEE gives consistent estimates in the presence of missing data only if the data are MCAR. This would also apply to the GEE extension proposed by several authors ([11], [20] and, [19]). However, Robins *et al.* ([13]) extended the common GEE methodology to situations of MAR by weighting each observation by its inverse probability of being observed.

In this paper we describe the properties of the latent variable model under missing data and extend the weighted GEE (WGEE) to multiple non-commensurate outcomes for MAR data. A study investigating the association between participation in a managed behavioral health care carve-out and quality of health care measured using bivariate mixed outcomes ([4]), and a study evaluating health-related quality of life after discharge from an intensive care unit using the Euroqol-5d instrument([8]), illustrate our methods.

2. LATENT VARIABLE MODEL FOR MULTIPLE CONTINUOUS AND BINARY OUTCOMES

Let $(y_{1i}, ..., y_{qi})$ represent a multivariate outcome for the i^{th}-individual $(i = 1, ..., n)$. We will use the symbol $\cdot$ in the subscript of $y_{k\cdot}$ to designate all the observations for outcome k or $y_{\cdot i}$ to indicate all the outcomes for the individual i. Let $\boldsymbol{x}_{ji}$ represent a vector of covariates for the i^{th}-individual associated with the j^{th}-outcome. We allow each outcome to be associated with its own set of covariates. Let R_{ji} be an indicator variable with value 1 if y_{ji} is observed and 0 otherwise. The superscript 'obs' is used to denote *observed* data. We assume throughout that the covariates are fixed and completely observed, and thus will be suppressed when writing the conditional distributions.

2.1. Latent variable model with outcome data MAR

One approach to model non-commensurate outcomes in a multivariate framework is to introduce latent variables, $\boldsymbol{u}_i = (u_{1i}, ..., u_{pi};\, p < q)$, to induce the correlation between the outcomes. Conditional on the latent variables $\boldsymbol{u}$ the outcomes are assumed to be independent ([5]). We assume that one of the outcomes, y_{1i}, has some missing observations and that these observations are MAR,

i.e., $P\big(R_{1i}=1 \mid y_{\cdot i}, \boldsymbol{x}_{ji}\big)$ depends on the observed data, for example, $y_{2i}, ..., y_{qi}$ and $\boldsymbol{x}_{1i}$. Let θ be the vector of parameters associated with the distribution of $y_{\cdot i} \,|\, \boldsymbol{x}_{ji}$. The log-likelihood for the observed data is given by

$$\begin{aligned} \log L\big(\theta; y_{1\cdot}^{\text{obs}}, ..., y_{q\cdot}, R_{1\cdot}, \boldsymbol{x}_{ji}\big) \;\propto\; & \log \prod_{i=1}^{n} \Big(f(y_{\cdot i} \,|\, \boldsymbol{x}_{ji})\, P\big(R_{1i}=1 \,|\, y_{\cdot i}, \boldsymbol{x}_{ji}\big) \Big)^{R_{1i}} \\ & \times \left(\int f(y_{\cdot i} \,|\, \boldsymbol{x}_{ji})\, P\big(R_{1i}=0 \,|\, y_{\cdot i}, \boldsymbol{x}_{ji}\big)\, \partial y_{1i} \right)^{(1-R_{1i})} . \end{aligned} \tag{2.1}$$

With some algebraic manipulation and using the fact that R_{1i} does not depend on y_{1i} we can re-write (2.1) as

$$\begin{aligned} = & \sum_{i=1}^{n} \Big(R_{1i} \log f(y_{\cdot i} \,|\, \boldsymbol{x}_{ji}) + (1-R_{1i}) \log f(y_{2i}, ..., y_{qi} \,|\, \boldsymbol{x}_{ji}) \Big) \\ & + \sum_{i=1}^{n} \Bigg(R_{1i} \log \Big(P\big(R_{1i}=1 \,|\, y_{2i}, ..., y_{qi}, \boldsymbol{x}_{ji}\big) \Big) \\ & + (1-R_{1i}) \log \Big(P\big(R_{1i}=0 \,|\, y_{2i}, ..., y_{qi}, \boldsymbol{x}_{ji}\big) \Big) \Bigg). \end{aligned} \tag{2.2}$$

The terms in the log-likelihood involving the missingness mechanism $P\big(R_{1i} \,|\, y_{2i}, ..., y_{qi}, \boldsymbol{x}_{ji}\big)$ will not involve the parameters θ associated with the distribution of $y_{\cdot i} \,|\, \boldsymbol{x}_{ji}$. These terms will not contribute for the estimation of θ and for this reason they can be ignored. Therefore, the log-likelihood can be written as the sum of terms associated with the distribution for complete observations and terms associated with the distribution for incomplete observations. Thus, the presence of missing data does not add extra difficulty to the maximization of the likelihood. In this case we say that the likelihood can be directly maximized because it does not require a more complex method, such as the EM-algorithm nor multiple imputation, to compute the maximum likelihood estimates.

Consider the case of a binary outcome, $y_{1\cdot}$, and a continuous outcome, $y_{2\cdot}$, where some entries of $y_{1\cdot}$ are missing. In this case $q=2$ and $p=1$. We assume the following model for the outcomes:

$$\begin{aligned} \text{probit}\big(E(y_{1i} \,|\, \boldsymbol{x}_{1i}, u_i)\big) \;&=\; \boldsymbol{\beta}_1^{*\text{T}} \boldsymbol{x}_{1i} + u_i \;, \\ y_{2i} \,|\, \boldsymbol{x}_{1i}, u_i \;&=\; \boldsymbol{\beta}_2^{\text{T}} \boldsymbol{x}_{2i} + \sigma_2\, u_i + \epsilon_{2i} \;, \end{aligned} \tag{2.3}$$

where $\epsilon_{2i} \sim N(0, \sigma_2^2)$ and u_i is a latent variable with $u_i \sim N(0, \sigma_u^2)$. The latent variable u_i in the model induces the correlation between the outcomes and the parameter σ_2 that multiplies the latent variable is introduced to standardize the different scales of the two outcomes. For more details see Teixeira-Pinto and Normand ([19]).

The log-likelihood for the observed data can be written as

$$\log L\big(\theta; y_{1\cdot}^{\text{obs}}, y_{2\cdot}, R_{1\cdot}, \boldsymbol{x}_{1i}, \boldsymbol{x}_{2i}\big) \;\propto \tag{2.4}$$

$$\propto \sum_{i=1}^{n}\Big(R_{1i}\log f(y_{1i}, y_{2i}\,|\,\boldsymbol{x}_{1i}, \boldsymbol{x}_{2i}) + (1-R_{1i})\log f(y_{2i}\,|\,\boldsymbol{x}_{2i})\Big)$$

$$= \sum_{i=1}^{n}\Big(R_{1i}\log\int f(y_{1i}\,|\,\boldsymbol{x}_{1i}, u_i)\, f(y_{2i}\,|\,\boldsymbol{x}_{2i}, u_i)\, f(u_i)\,\partial u_i \;+\; (1-R_{1i})\log f(y_{2i}\,|\,\boldsymbol{x}_{2i})\Big).$$

Depending on the link functions used for each outcome it might be possible to have a closed-form representation for the marginal distribution of each outcome. Using the identity link for the continuous outcome and the probit link for the binary as in (2.3), the model for the marginal means of each outcome can be written as

$$\text{probit}\big(P(y_{1i}=1\,|\,\boldsymbol{x}_{1i})\big) \;=\; \text{probit}\Big(\int P(y_{1i}=1\,|\,\boldsymbol{x}_{1i}, u_i)\, f(u_i)\, du_i\Big) \;=\; \frac{\boldsymbol{\beta}_1^{*\mathrm{T}}\boldsymbol{x}_{1i}}{\sqrt{1+\sigma_u^2}}\;, \tag{2.5}$$

$$y_{2i}\,|\,\boldsymbol{x}_{2i} = \boldsymbol{\beta}_2^{\mathrm{T}}\boldsymbol{x}_{2i} + \epsilon_{2i}^{*}\,, \qquad \text{where}\;\; \epsilon_{2i}^{*} \sim N\big(0, \sigma_2^2(1+\sigma_u^2)\big)\,.$$

If instead we choose a logit link for the binary outcome in equation (2.3), the model for the marginal mean does not have a closed-form representation.

3. WEIGHTED GENERALIZED ESTIMATING EQUATIONS FOR NON-COMMENSURATE OUTCOMES

3.1. WGEE with data MAR

Suppose we are in the same setting as in the previous section with a binary and a continuous outcome to motivate the WGEE. We adapt the WGEE proposed by Robins *et al.* ([13]) to the situation of multiple non-commensurate outcomes.

The generalization to multiple outcomes is relatively straightforward but some remarks will be made.

Let $y_{\cdot i} = (y_{1i}, y_{2i})^{\mathrm{T}}$ be a vector of a binary and a continuous outcome with the following marginal model for the outcomes:

$$\mu_{ji} \;=\; g_j^{-1}(\boldsymbol{\beta}_j^{\mathrm{T}}\boldsymbol{x}_{ji})\;, \tag{3.1}$$

where $\mu_{ji} = E(y_{ji}\,|\,\boldsymbol{x}_{ji})$, $j = (1, 2)$, g_j is the probit link for $j = 1$ and the identity link for $j = 2$. If both outcomes are completely observed, the estimating equation

is

$$\sum_{i=1}^{n} D_i^{\mathrm{T}} V_i^{-1} (y_{\cdot i} - \mu_{\cdot i}) \;=\; 0 \tag{3.2}$$

and has a solution that is a consistent and asymptotically normal estimator for $\boldsymbol{\beta}_j$ ([20], [19]) with variance $\Gamma^{-1}\,\Omega\,\Gamma^{-1}$, where $D_i = \left(\frac{\partial \mu_{\cdot i}}{\partial \boldsymbol{\beta}}\right)_j$, V_i is a 'working' covariance matrix for y_{1i} and y_{2i}, $\Gamma = E(D_i^{\mathrm{T}} V^{-1} D_i)$ and $\Omega = E\big(D_i^{\mathrm{T}} V^{-1} (y_{\cdot i} - \mu_{\cdot i}) \cdot \cdot (y_{\cdot i} - \mu_{\cdot i})^{\mathrm{T}} V^{-1} D_i\big)$. Typically, D_i is a block-diagonal matrix because the equations for each outcome do not share the regression parameters. The solution for the estimating equation is a consistent estimator of $\boldsymbol{\beta}$ even if V_i is misspecified. In the case of missing data, this result holds if the data are MCAR but not for MAR.

Suppose that some observations of y_{1i} are missing and the missing mechanism depends on y_{2i} and x_{ji}. If the variables y_{2i} and x_{ji} are always observed then y_{1i} is MAR. In this case $E(y_{1i}^{\mathrm{obs}} \,|\, \boldsymbol{x}_{ji}) \neq \mu_{1i}$ because

$$\begin{aligned} E(y_{1i}^{\mathrm{obs}} \,|\, \boldsymbol{x}_{ji}) \;&=\; E(R_{1i}\, y_{1i} \,|\, \boldsymbol{x}_{ji}) \;=\; E\big(E(R_{1i}\, y_{1i} \,|\, y_{1i}, y_{2i}, \boldsymbol{x}_{ji})\big) \\ &=\; E\big(y_{1i}\, E(R_{1i} \,|\, y_{2i}, y_{1i}, \boldsymbol{x}_{ji})\big)\,. \end{aligned} \tag{3.3}$$

R_{1i} does not depend on y_{1i} because the data are MAR, and $E\big(y_{1i}\, E(R_{1i} \,|\, y_{2i}, y_{1i}, \boldsymbol{x}_{ji})\big)$ simplifies to $E\big(y_{1i}\, P(R_{1i} = 1 \,|\, y_{2i}, \boldsymbol{x}_{ji})\big)$. Therefore, this expectation is not equal to μ_{1i} so the solution for the equation (3.2) is no longer a consistent estimate of $\boldsymbol{\beta}_1$. However, if we weight y_{1i} by its inverse probability of being observed $\pi_{1i} = P(R_{1i} \,|\, y_{2i}, \boldsymbol{x}_{ji})$, we have:

$$E\left(\frac{R_{1i}}{\pi_{1i}}\,(y_{1i} - \mu_{1i}) \mid \boldsymbol{x}_{1i}\right) \;=\; E\left(E\left(\frac{R_{1i}}{\pi_{1i}}\,(y_{1i} - \mu_{1i}) \mid y_{1i}, y_{2i}, \boldsymbol{x}_{ji}\right) \,\middle|\, \boldsymbol{x}_{1i}\right)$$

and, because $E(R_{1i} \,|\, y_{2i}, \boldsymbol{x}_{ji}) = \pi_{1i}$,

$$=\; E\big(y_{1i} - \mu_{1i} \,|\, \boldsymbol{x}_{1i}\big) \;=\; 0\;.$$

This motivates the following weighted estimating equation:

$$\sum_{i=1}^{n} D_i^{\mathrm{T}} V_i^{-1} \Delta_i (y_{\cdot i} - \mu_{\cdot i}) \;=\; 0 \tag{3.4}$$

and

$$\Delta_i \;=\; \begin{pmatrix} R_{1i}\, \pi_{1i}^{-1} & 0 \\ 0 & 1 \end{pmatrix}. \tag{3.5}$$

The estimating equation (3.4) has a solution $\hat{\boldsymbol{\beta}}$ which is a consistent estimate of $\boldsymbol{\beta}$ and has an asymptotic normal distribution with a consistent estimator

of its variance given by $\hat{\Gamma}^{-1}\left(\sum_{i=1}^{n}\hat{C}_i\,\hat{C}_i^{\mathrm{T}}\right)\Gamma^{-1^{\mathrm{T}}}$ where $\hat{\Gamma}=\sum_{i=1}^{n}(D_i^{\mathrm{T}}V^{-1}\Delta_i D_i)$, $\hat{C}_i=D_i^{\mathrm{T}}V^{-1}(y_{\cdot i}-\mu_{\cdot i})-\left(\sum_{i=1}^{n}D_i^{\mathrm{T}}V^{-1}(y_{\cdot i}-\mu_{\cdot i})\,S_i^{\mathrm{T}}\right)\left(\sum_{i=1}^{n}S_i\,S_i^{\mathrm{T}}\right)S_i$ and S_i is the score component for the i^{th}-individual from the model for π_{1i}.

The last entry in the matrix Δ_i is 1 because only $y_{1\cdot}$ is missing for some subjects and $y_{2\cdot}$ is always observed. The weights π_{ji} are unknown and have to be estimated. We can use, for example, a logistic regression to estimate $\pi_{1i}=P(R_{1i}=1\,|\,y_{2i},\boldsymbol{x}_{ji})$ as in (3.6) and plug in the estimates in equation (3.4).

$$\text{logit}(\pi_{1i}) \;=\; \zeta_0+\zeta_1\,y_{2i}+\zeta_2\,\boldsymbol{x}_{ji}\;. \tag{3.6}$$

The assumption of MAR implies that if R_{ji} depends on the other outcomes, then only one outcome can have missing observations. However, if there are missing observations in y_{2i} or in one of the covariates involved in the model (3.6), we no longer have a case of MAR and we are not able to estimate all the weights π_{1i}.

3.2. Estimation of the Covariance Parameters

Although we are mainly interested in the estimation of the parameters $\boldsymbol{\beta}_j$, consistent estimators for the parameters in $V_i=\begin{pmatrix}\sigma_1^2 & \rho\,\sigma_2\,\sigma_1\\ \rho\,\sigma_2\,\sigma_1 & \sigma_2^2\end{pmatrix}$ are needed in equation (3.4). One way of obtaining these estimators is to add estimating equations for these parameters. Because we are not concerned about estimating σ_1, σ_2 and ρ efficiently, we can use the following unbiased equations based on the method of moments:

$$\sum\frac{R_{1i}}{\pi_{1i}}\left(\sigma_1-\sqrt{\frac{\sum(y_{1i}-\mu_{1i})^2}{n}}\right)\;=\;0\;, \tag{3.7}$$

$$\sum\left(\sigma_2-\sqrt{\frac{\sum(y_{2i}-\mu_{2i})^2}{n}}\right)\;=\;0\;, \tag{3.8}$$

$$\sum\frac{R_{1i}}{\pi_{1i}}\left(\rho-\frac{\sum(y_{1i}-\mu_{1i})\,(y_{2i}-\mu_{2i})}{\sigma_2\,\sqrt{n\sum(y_{1i}-\mu_{1i})^2}}\right)\;=\;0\;.$$

Equations (3.4) and (3.7) can be solved jointly to obtain estimates for all the parameters. If instead of missing observations in y_{1i} we had missing observations in y_{2i}, then the terms in equation for σ_2 would also require to be weighted in order to obtain an unbiased estimator for σ_2.

This entire approach can be applied to more than two outcomes. However, the assumption of MAR implies that missingness mechanism has to depend only in completely observed outcomes. If this is not the case the data are MNAR.

4. SIMULATION STUDY

Data were generated using the model

$$(4.1)\quad (y_{1i}^*, y_{2i}) \,|\, (x_i, z_{1i}, z_{2i}) \;\sim\; MVN\left(\begin{pmatrix} .5 + 2\,x_i + 2\,z_{1i} \\ 5 + 10\,x_i + 2\,z_{2i} \end{pmatrix}, \begin{pmatrix} 1 & 6\times 1\times .8 \\ & 36 \end{pmatrix}\right),$$

with x_i generated from a Bernoulli(.5), z_{1i} generated from a Uniform$(-1, 0)$ and z_{2i} from $N(1, 4)$. Then, y_{1i}^* was categorized in the following way:

$$(4.2)\qquad y_{1i} = \begin{cases} 0\,, & \text{if } y_{1i}^* \le 0\ , \\ 1\,, & \text{if } y_{1i}^* > 0\ . \end{cases}$$

By using a probit link to model y_{1i} as $\text{probit}\big(P(y_{1i} = 1 \,|\, x_i, z_i)\big) = \alpha_1 + \beta_1 x_i + \gamma_1 z_i$, we have $P(y_{1i} = 1 \,|\, x_i, z_i) = P(y_{1i}^* > 0 \,|\, x_i, z_i) = \Phi\big(\frac{.5 + 2\,x_i + 2\,z_i}{\sigma_1}\big)$. By construction $\sigma_1 = 1$ thus, the true parameters for the probit regression maintain the same value as in (4.1), i.e., $\alpha_1 = .5$, $\beta_1 = 2$ and $\gamma_1 = 2$.

We generated 1000 datasets with 400 bivariate observations each. Some observations for the continuous outcome were deleted according to the model $\text{logit}\big(P(R_{2i} = 1 \,|\, y_{1i}, x_i)\big) = .5 - 3.5\,y_{1i} - x_i$. The parameters were chosen to obtain approximately 25% of missing observations (about 40% of missing y_{2i} when $x_i = 0$ and 5% when $x_i = 1$).

4.1. Univariate analysis

We fit separate regressions for each outcome, ignoring the missingness mechanism and the correlation between the outcomes. We used a probit regression for the binary outcome (4.3) and a linear regression for the continuous (4.4):

$$(4.3)\qquad \text{probit}\big(E(y_{1i} \,|\, x_i, z_{1i})\big) \;=\; \alpha_1 + \beta_1 x_i + \gamma_1 z_{1i}\ ,$$

$$(4.4)\qquad E(y_{2i} \,|\, x_i, z_{2i}) \;=\; \alpha_2 + \beta_2 x_i + \gamma_2 z_{2i}\ .$$

4.2. Latent variable model

We fit the latent variable model,

$$(4.5)\qquad \text{probit}\big(E(y_{1i} \,|\, x_i, z_{1i}, u_i)\big) \;=\; \alpha_1^* + \beta_1^* x_i + \gamma_1^* z_{1i} + u_i\ ,$$

$$(4.6)\qquad E(y_{2i} \,|\, x_i, z_{2i}, u_i) \;=\; \alpha_2 + \beta_2 x_i + \gamma_2 z_{2i} + \sigma_2 u_i\ .$$

It can shown that the above model is the correct model for the data generation process. To obtain marginal effects of the covariates as in the other models we have to average over the latent variable u_i. In this case the marginal effects can be obtained by dividing the parameters by $\sqrt{1+\sigma_u^2}$, for example, the marginal effect of x on y_1 is $\beta_1 = \frac{\beta_1^*}{\sqrt{1+\sigma_u^2}}$. We used PROC NLMIXED from SAS to fit the latent variable model. The initial parameters were obtained by fitting separate regressions for each outcome (univariate analysis). The initial value for the correlation parameter was set to be 0.5.

4.3. Weighted generalized estimating equations

We assumed the following model for the means of the outcomes:

$$\text{probit}\big(E(y_{1i}\,|\,x_i, z_{1i})\big) \;=\; \text{probit}(\mu_{1i}) \;=\; \alpha_1 + \beta_1 x_i + \gamma_1 z_{1i}\ , \tag{4.7}$$

$$E(y_{2i}\,|\,x_i, z_{2i}) \;=\; \mu_{2i} \;=\; \alpha_1 + \beta_1 x_i + \gamma_1 z_{2i}\ . \tag{4.8}$$

We solved the WGEE:

$$\sum_{i=1}^{n} \begin{pmatrix} -\phi(A_i) & 0 \\ -x_i\,\phi(A_i) & 0 \\ -z_{1i}\,\phi(A_i) & 0 \\ 0 & 1 \\ 0 & x_i \\ 0 & z_{2i} \end{pmatrix} \begin{pmatrix} \sigma_1^2 & \rho\,\sigma_2\,\sigma_1 \\ \rho\,\sigma_2\,\sigma_1 & \sigma_2^2 \end{pmatrix}^{-1} \begin{pmatrix} 1 & 0 \\ 0 & \frac{R_{2i}}{\hat{\pi}_{2i}} \end{pmatrix} \begin{pmatrix} y_{1i}-\mu_{1i} \\ y_{2i}-\mu_{2i} \end{pmatrix} = 0 \tag{4.9}$$

with $A_i = \alpha_1 + \beta_1 x_i + \gamma_1 z_{1i}$ and $\sigma_1 = \sqrt{\Phi(A_i)\,(1-\Phi(A_i))}$. The weights $\hat{\pi}_{2i}$ were estimated using the logistic regression

$$\text{logit}\big(R_{2i}=1\,|\,y_{1i}, x_i\big) \;=\; \text{logit}(\pi_{2i}) \;=\; \zeta_0 + \zeta_1\,y_{1i} + \zeta_2\,x_i\ . \tag{4.10}$$

Two additional equations were added to the system of equations (4.9) to obtain estimates of the unknown parameters σ_2 and ρ:

$$\begin{aligned} &\sum \frac{R_{2i}}{\pi_{2i}} \left(\sigma_2 - \sqrt{\frac{\sum (y_{2i}-\mu_{2i})^2}{n}} \right) = 0\ , \\ &\sum \frac{R_{2i}}{\pi_{2i}} \left(\rho - \frac{\sum (y_{1i}-\mu_{1i})\,(y_{2i}-\mu_{2i})}{\sigma_2 \sqrt{n \sum (y_{1i}-\mu_{1i})^2}} \right) = 0\ . \end{aligned} \tag{4.11}$$

The WGEE were solved using a program developed in SAS with PROC IML.

4.4. Results

The results of the simulations are summarized in Tables 1 and 2. Overall, the latent variable model performed better than the univariate approach and the WGEE. The estimates of the parameters associated with the continuous outcome, $\hat{\alpha}_2$ and $\hat{\beta}_2$, were biased for the univariate model, and the mean square errors (MSE) were about 4 and 6 times higher than the corresponding MSE estimates obtained from the latent variable model. The remaining estimates for the univariate approach were not biased but they had slightly higher standard errors than the latent model. This is explained by the fact that the latent variable model uses the additional information of the correlation between the outcomes as described by Teixeira-Pinto and Normand ([19]).

Table 1: Estimates and standard errors averaged over the results of 1000 simulated datasets with sample size equal to 400. About 25% data were deleted for the continuous outcome using a model for the missingness mechanism that depends on the binary outcome.

Estimates (true value)	Univariate		Latent		WGEE	
	Mean	(SE)	Mean	(SE)	Mean	(SE)
Binary outcome						
$\hat{\alpha}_1$ $(\alpha_1 = .5)$	0.521	(0.167)	0.521	(0.148)	0.519	(0.159)
$\hat{\beta}_1$ $(\beta_1 = 2)$	2.025	(0.181)	2.025	(0.172)	2.019	(0.181)
$\hat{\gamma}_1$ $(\gamma_1 = 2)$	2.045	(0.305)	2.044	(0.257)	2.035	(0.288)
Continuous outcome						
$\hat{\alpha}_2$ $(\alpha_2 = 5)$	6.523	(0.581)	5.009	(0.556)	5.033	(0.601)
$\hat{\beta}_2$ $(\beta_2 = 10)$	8.737	(0.702)	9.980	(0.685)	9.944	(0.737)
$\hat{\gamma}_2$ $(\gamma_2 = 2)$	2.001	(0.170)	1.999	(0.145)	1.999	(0.171)

Table 2: Mean square error (MSE) and relative bias (estimate/true value) averaged over the results of 1000 simulated datasets with sample size equal to 400. About 25% data were deleted for the contin uous outcome using a model for the missingness mechanism that depends on the binary outcome.

Estimates	Mean square error			Relative bias		
	Univ.	Latent	WGEE	Univ.	Latent	WGEE
Binary outcome						
$\hat{\alpha}_1$	0.030	0.024	0.027	1.042	1.042	1.037
$\hat{\beta}_1$	0.033	0.031	0.038	1.013	1.013	1.009
$\hat{\gamma}_1$	0.097	0.071	0.089	1.023	1.022	1.018
Continuous outcome						
$\hat{\alpha}_2$	2.669	0.317	0.371	1.305	1.002	1.007
$\hat{\beta}_2$	2.108	0.494	0.857	0.874	0.998	0.994
$\hat{\gamma}_2$	0.028	0.021	0.031	1.001	1.000	1.000

The WGEE estimates had very similar bias to the latent variable model, although the MSEs for all estimates were higher in the WGEE due to higher variances for the estimates. This loss of efficiency is expected when compared to a full likelihood method such as the latent variable model. The sandwich estimator for the variance of the estimates is robust to the misspecification of the correlation between the outcomes and for this reason is more conservative.

5. EXAMPLES

5.1. Example 1: Managed Care and Quality of Care for Schizophrenia

Dickey *et al.* ([4]) conducted a prospective observational study of 420 adults with schizophrenia who sought care for a psychiatric crisis. The main objective of the study was to compare care for patients who were and were not enrolled in managed care because advocates for those with mental illness worried that patients who had their care managed may have worse care than those who did not. Two outcomes, one binary (whether the patient was prescribed an atypical anti-psychotic medication) and one continuous (self-reported quality of interpersonal interactions between patient and clinician) were measured for the 197 patients who had their care managed and the 223 patients whose care was not managed. The self-reported quality of interpersonal interactions between patient and clinician was missing for 26 patients (6%). The information regarding the prescription of an atypical anti-psychotic was available for all the subjects. There was a significant difference in the proportion of patients who were prescribed an atypical anti-psychotic medication between the group without data on the quality of interpersonal interactions between patient and clinician (50%) and the group of patients with data on this outcome (71%) ($\chi^2_2 = 5$, p-value $= 0.03$). This result suggests that the data are MAR. There was no statistical significant association between the missing indicator and the sociodemographic characteristics using a significance level of 0.05.

We used separate regression models for each outcome (the univariate approach) ignoring the correlation between the outcomes and the missing data (equations 4.3 and 4.4). We fit the latent variable model (4.5) and the WGEE (4.9). For the latent variable model we computed the marginal effects estimates of managed care on the outcomes by dividing the regression coefficients by $\sqrt{1+\sigma_u^2}$ as described in section 2.1. The weights for the WGEE were obtained from a logistic model for the probability of missing observation in the self-reported quality of interpersonal interactions between patient and clinician outcome using the prescription of an atypical anti-psychotic and managed care status as covariates. The estimates for the weights were given by the inverse of the estimated probabilities from the logistic model, $\text{logit}(\hat{\pi}_{2i}) = \text{logit}\big(\hat{P}(R_{2i} \mid y_{1i}, x_i)\big) = 2.23 + 0.88\, y_{1i} - 0.11\, x_i$.

The mean (SD) age of patients was 40 (8.5) and 41 (7.9) in the managed care and not managed care group, respectively. Other sociodemographic characteristics of the patients are described in Table 3. No significant differences were observed for the two outcomes analyzed regarding the sociodemographic characteristics. Seventy one percent of the patients in the managed care group received atypical anti-psychotic medication versus 68% in the not managed care group. The mean (SD) self-reported quality of interpersonal interactions between patient and clinician was 3.20 (0.67) for the managed care group and 3.21 (0.65) for the not managed group.

Table 3: Sociodemographic characteristics of 420 patients with schizophrenia.

Sociodemographic characteristics	Type of care		p-value
	Managed ($n = 197$)	Not Managed ($n = 223$)	
Age			
< 35 years	24	21	
35–44 years	46	44	0.338
45–54 years	21	29	
55–64 years	8	6	
Male sex	47	66	< 0.001
Race or Ethnicity			
White	51	66	
African American	31	22	0.005
Other	18	12	
Never married	64	68	0.364
High school education or less	74	59	0.002
Homeless	15	9	0.069
English speaking	90	93	0.277

The effect estimates of managed care on the outcomes were identical and not statistically significant at the 0.05 level for all the models (Table 4). This suggests no difference in the quality of care between the managed and not managed care groups. For the outcome with some missing observations, patient/clinician relationship outcome, the estimated effect of managed care was the same for the latent variable model and the WGEE ($\hat{\beta}_2 = -0.019$). The effect estimate for the univariate approach was slightly smaller ($\hat{\beta}_2 = -0.017$). Although this result is consistent with the simulation study, it is hard to argue that such a small difference is a consequence of ignoring the MAR mechanism rather than random variation. The WGEE provided identical standard errors of the estimators to the other two approaches. This can be explained by the low correlation between the outcomes (0.059 as estimated by the WGEE).

Table 4: Managed care effect on the two outcomes related to quality of care: "patient/clinician relationship" and "prescription of anti-psychotic medication". Data on 420 patients with schizophrenia but only 394 patients had information regarding patient/clinician's relationship.

Estimated effects	Model		
	Univariate	Latent	WGEE
	β (Std. Error) p-value	β (Std. Error) p-value	β (Std. Error) p-value
Binary: Prescription of anti-psychotic ($n = 420$)			
Intercept	0.549 (0.089) ≤ 0.001	0.548 (0.089) ≤ 0.001	0.549 (0.088) ≤ 0.001
Managed care	-0.081 (0.129) 0.530	-0.079 (0.129) 0.538	-0.081 (0.128) 0.527
Continuous: Patient/clinician relationship ($n = 394$)			
Intercept	3.213 (0.045) ≤ 0.001	3.213 (0.045) ≤ 0.001	3.213 (0.045) ≤ 0.001
Managed Care	-0.017 (0.066) 0.799	-0.019 (0.066) 0.775	-0.019 (0.067) 0.771
$\hat{\sigma}_2$	0.656	0.630	0.656
$\hat{\sigma}_u$	—	0.286	—
$\hat{\rho}$	—	—	0.059

5.2. Example 2: Quality of life after discharge from Intensive Care

Granja *et al.* ([8]) evaluated the health-related quality of life (HRQOL) of adult patients discharged from an intensive care unit (ICU) located in Portugal. The 485 patients who agreed to participate in the study were asked to complete the Euroqol 5D (EQ-5D) instrument to evaluate their HRQOL ([1]), 6 months after discharge from ICU. This instrument includes two main sections. The first contains five questions that measure functional dimensions of HRQOL (mobility, self-care, usual activities, pain/discomfort and anxiety/depression) and it is summarized by a general score designated as the EQ-5D index. The EQ-5D index varies from 0 to 100, where 100 indicates no disability in the 5 dimensions. The second part of the instrument is a visual analogue scale (VAS) in which patients mark their perception of their health state in a 0 to 100 scale (100 – best imaginable state, 0 – worst imaginable state). For the analysis the VAS scale was dichotomized using the middle point of its scale (less or equal to 50 and more than 50).

In this example we will focus on the impact of patient's severity when admitted to the ICU (measured by the Apache II score) on the HRQOL after discharge (measured by the EQ-5D index and dichotomized VAS). Some studies reported that most of the patients who survive ICU do not show significant decrease in physical ability but they report psychological problems ([18], [16]). This finding suggests that the effect of the severity of the episode that led to ICU admission may be different for functional HRQOL and for patient's perception of their HRQOL. If this is the case, both aspects of HRQOL should be reported in HRQOL studies.

The effect of patient's severity at ICU admission on HRQOL should be adjusted to age and previous health state (non-chronic disease, chronic disease with no disability and chronic disease with disability). All the patients completed the first part of the questionnaire involved in the calculation of the EQ-5D index, but only 366 completed the VAS question.

Table 5 summarizes some demographic and clinical information from the 485 patients. The mean (SD) age of the 485 patients was 55.2 (17.4) years old. Twenty eight percent (28%) of the patients reported that they had no chronic disease prior to admission to ICU and 21% reported they had chronic disease that caused some kind of disability. The remaining 51% indicated that they had chronic disease with no disability before admission to ICU. The mean (SD) Apache II score at admission was 13.0 (6.8). For the 366 patients who completed the VAS scale, 64% reported a value above 50. The mean (SD) for the EQ-5D index was 74.2 (17.4). The group of patients that completed both parts of the questionnaire had significantly higher EQ-5D index than those who did not completed the VAS question (77.9 vs. 52.6, p-value < 0.001).

Table 5: Demographic and clinical characteristics of 485 patients that participated in the study of HRQOL after ICU admission.

Demographic and clinical characteristics	($n = 485$)
Age (mean (SD))	55.2 (17.4)
Male sex (n (%))	275 (57)
Apache II score (mean (SD))	13.0 (6.8)
Previous health state (n (%))	
non-chronic disease	138 (28)
chronic disease with no disability	245 (51)
chronic disease with disability	102 (21)
ICU length of stay in days (median (IQR))	2 (1–6)

Similarly to example 1, we run separate models for each outcome (a linear regression for the EQ-5D index and a probit regression for the dichotomized VAS) and we fit the latent model and WGEE using the same link functions as

the univariate models. The effect of previous health state on both measures of HRQOL was linear for the three categories, so it entered the model as an interval variable with no need to create dummy variables for the categories. The weights for the WGEE were obtained from a logistic model for the probability of missing observation in the VAS question using the EQ-5D index, Apache II score, age and the previous health state as covariates. The estimates for the weights were given by the inverse of the estimated probabilities from the logistic model, $\text{logit}(\hat{\pi}_{1i}) = \text{logit}\big(\hat{P}(R_{1i} \,|\, y_{2i}, x_{1i}, x_{2i}, x_{3i})\big) = 0.85 + 0.04\, y_{1i} - 0.03\, x_{1i} - 0.04\, x_{2i} - 0.17\, x_{3i}$.

The results are summarized in Table 6. The HRQOL is associated with patient's age and the health state previous to admission. The severity at admission measured by Apache II is not associated with the functional aspect of HRQOL (p-value $= 0.999$). These results were consistent in all approaches.

Table 6: Effect of severity at admission to ICU (Apache II), adjusted to age and previous health state, on health-related quality of life measured (D-VAS and EQ-5D index), 6 months after discharge from an ICU. A total of 485 patients entered the study but only 366 completed the question regarding D-VAS.

Estimated effects	Model		
	Univariate	Latent	WGEE
	β (Std. Error) p-value	β (Std. Error) p-value	β (Std. Error) p-value
Binary: D-VAS ($n = 366$)			
Intercept	−2.069 (0.290) < 0.001	2.018 (0.280) < 0.001	2.027 (0.280) < 0.001
Age	−0.011 (0.004) 0.014	−0.009 (0.004) 0.029	−0.012 (0.004) 0.014
Previous health state	−0.460 (0.111) < 0.001	−0.494 (0.106) < 0.001	−0.442 (0.111) < 0.001
Apache II	−0.018 (0.011) 0.093	−0.028 (0.011) 0.009	−0.025 (0.012) 0.040
Continuous: EQ-5D ($n = 485$)			
Intercept	100.3 (3.902) < 0.001	100.3 (3.886) < 0.001	103.3 (3.489) < 0.001
Age	−0.244 (0.061) < 0.001	−0.244 (0.061) < 0.001	−0.244 (0.055) < 0.001
Previous health state	−8.116 (1.540) < 0.001	−8.116 (1.533) < 0.001	−8.115 (1.458) < 0.001
Apache II	≈ 0 (0.157) 0.999	≈ 0 (0.157) 0.999	≈ 0 (0.163) 0.999
$\hat{\sigma}_2$	21.94	14.86	21.94
$\hat{\sigma}_u$	—	1.086	—
$\hat{\rho}$	—	—	0.532

The major difference between the univariate and the multivariate methods is the result for the effect of Apache II on the dichotomized VAS. The estimate in the latent model and WGEE is higher than that in the univariate approach and it becomes statistically significant at the 0.05 level. This may indicate that the patient's perception about his or her own HRQOL is affected by the degree of severity of the episode leading to ICU admission. This fact would not be identified in the univariate analysis.

6. CONCLUSION

We developed likelihood and quasi-likelihood methods to analyze multiple non-commensurate outcomes in the presence of missing data. Although this type of data is common in biomedical studies, the usual approach is to analyze each outcome separately ignoring the correlation among the outcomes. This can lead to loss of efficiency and biased estimates in the case of MAR. The WGEE has the advantage of being robust to the misspecification of the correlation between the outcomes and MAR while the latent variable model is a full likelihood approach which typically gives more efficient estimates but assumes that the mean and covariance models are correctly specified. Another alternative to WGEE is to use the multiple imputation methodology. We could assume a model to impute values for the missing observations and repeat the process to create several complete datasets. Then we could solve a regular GEE for each dataset and obtain the estimates of the regression parameters as the mean over the estimates obtained in each complete dataset.

We have shown both in simulations and in real data analysis that the estimation of associations can be biased in situations of MAR in the outcomes. The bias can be substantially reduced by jointly model the outcomes in a multivariate framework.

ACKNOWLEDGMENTS

This work was supported by Grant R01-MH54693 (Teixeira-Pinto and Normand) and R01-MH61434 (Normand), both from the National Institute of Mental Health and PTDC/SAU-ESA/100841/2008 from Fundação para a Ciência e Tecnologia. The schizophrenia managed care data were generously provided through the efforts of Barbara Dickey, Ph.D., Harvard Medical School, Boston, MA. The health-related quality of life data was generously provided by Cristina Granja, MD, Ph.D., Hospital Pedro Hispano, Porto, Portugal.

REFERENCES

[1] Brooks, R. and the EuroQol Group (1996). EuroQol: the current state of play, *Health Policy*, **37**, 53–72.

[2] Cox, D.R. and Wermuth, N. (1992). Response models for binary and quantitative variables, *Biometrika*, **79**, 441–461.

[3] Dempster, A.; Laird, N. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of Royal Statistical Society, Series B*, **39**, 1–38.

[4] Dickey, B.; Normand, S.-L.T.; Hermann, R.C.; Eisen, S.V.; Cortes, D.E.; Cleary, P.D. and Ware, N. (2003). Guideline recommendations for treatment of schizophrenia: the impact of managed care, *Arch. Gen. Psychiatry*, **60**, 340–8.

[5] Dunson, D.B. (2000). Bayesian latent variable models for clustered mixed outcomes, *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, **62**, 355–366.

[6] Fitzmaurice, G.M. and Laird, N.M. (1995). Regression models for a bivariate discrete and continuous outcome with clustering, *Journal of the American Statistical Association*, **90**, 845–852.

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

[8] Granja, C.; Teixeira-Pinto, A. and Costa-Pereira, A. (2002). Quality of life after intensive care: evaluation with EQ-5D questionnaire, *Intensive Care Medicine*, **28**, 898–907.

[9] Little, R.J. and Rubin, D. (2002). *Statistical Analysis with Missing Data*, John Wiley and Sons, Inc., Hoboken, New Jersey, U.S.A.

[10] Olkin, I. and Tate, R. (1961). Multivariate correlation models with mixed discrete and continuous variables, *The Annals of Mathematical Statistics*, **32**, 448–465.

[11] Prentice, R.L. and Zhao, L.P. (1991). Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses, *Biometrics*, **47**, 825–839.

[12] Regan, M.M. and Catalano, P.J. (1999). Likelihood models for clustered binary and continuous outcomes: application to developmental toxicology, *Biometrics*, **55**, 760–768.

[13] Robins, J.; Rotnitzky, A. and Zhao, L. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data, *Journal of the American Statistical Association*, **90**, 106–121.

[14] Sammel, M.D.; Ryan, L.M. and Legler, J.M. (1997). Latent variable models for mixed discrete and continuous outcomes, *Journal of the Royal Statistical Society, Series B: Methodological*, **59**, 667–678.

[15] Schafer, J. (1997). *Analysis of Incomplete Multivariate Data (Chapter 9)*, Chapman and Hall / CRC Monographs on Statistics and Applied Probability, 72, New York, NY.

[16] Schelling, G.; Stoll, C.; Haller, M.; Briegel, J.; Manert, W.; Hummel, T.; Lenhart, A.; Heyduck, M.; Polasek, J.; Meier, M.; Preuss, U.; Bullinger, M.; Schffel, W. and Peter, K. (1998). Health-related quality of life and posttraumatic stress disorder in survivors of the acute respiratory distress syndrome, *Critical Care Medicine*, **26**, 634–635.

[17] Shi, J. and Lee, S. (2000). Latent variable models with mixed continuous and polytomous data, *Journal of the Royal Statistical Association, Series B*, **62**, 77–87.

[18] Sukantarat, K.; Greer, S.; Brett, S. and Williamson, R. (2007). Physical and psychological sequelae of critical illness, *British Journal of Health Psychology*, **12**, 65–74.

[19] Teixeira-Pinto, A. and Normand, S.-L.T. (2009). Correlated bivariate continuous and binary outcomes: issues and applications, *Statistics in Medicine*, **28**, 1753–73.

[20] Zhao, L.P.; Prentice, R.L. and Self, S.G. (1992). Multivariate mean parameter estimation by using a partly exponential model, *Journal of the Royal Statistical Society, Series B*, **54**, 805–811.