Published in final edited form as:

Stat Med. 2009 June 15; 28(13): 1753–1773. doi:10.1002/sim.3588.

Correlated Bivariate Continuous and Binary Outcomes: Issues and Applications

Armando Teixeira-Pinto^{1,2,*†} and Sharon-Lise T. Normand^{2,3}

¹Department of Biostatistics and Medical Informatics, Faculty of Medicine, University of Porto, Porto, Portugal

²Department of Biostatistics, Harvard School of Public Health, Boston, U.S.A.

³Department of Health Care Policy, Harvard Medical School, Boston, U.S.A.

SUMMARY

Increasingly multiple outcomes are collected in order to characterize treatment effectiveness or to evaluate the impact of large policy initiatives. Often the multiple outcomes are non-commensurate, e.g., measured on different scales. The common approach to inference is to model each outcome separately ignoring the potential correlation among the responses. We describe and contrast several full likelihood and quasi-likelihood multivariate methods for non-commensurate outcomes. We present a new multivariate model to analyze binary and continuous correlated outcomes using a latent variable. We study the efficiency gains of the multivariate methods relative to the univariate approach. For complete data, all approaches yield consistent parameter estimates. When the mean structure of all outcomes depends on the same set of covariates, efficiency gains by adopting a multivariate approach are negligible. In contrast, when the mean outcomes depend on different covariate sets large efficiency gains are realized. Three real examples illustrate the different approaches.

1. Introduction

Often multiple outcomes are collected in health-related studies in order to characterize treatment effectiveness or associations with covariates. This observation is particularly true in psychiatric studies where the primary outcome is an abstract construct that cannot be measured directly. Instead, several variables are measured as proxies of the underlying outcome of interest. For example, in evaluating the effectiveness of a new anti-psychotic, researchers will examine several outcomes such as the positive and negative syndrome scale (PANSS) score, symptom relapse, and quality of life.

Typically the multiple outcomes are *non-commensurate*, i.e., they are measured on different scales such as continuous and binary responses. Although there has been some development of multivariate methods for non-commensurate outcomes, the usual modeling strategy is to consider each outcome separately in a univariate framework. This strategy is less efficient in the sense that such an approach ignores the extra information contained in the correlation among the outcomes. Other advantages of a multivariate setting include better control over the type I error rates in multiple tests and the ability to answer intrinsically multivariate questions. For example, we might be interested in assessing the impact of a policy change on the quality

of care (underlying outcome) rather than its impact on each outcome measured as a proxy of quality of care.

The challenge for multivariate methods is the nonexistence of obvious multivariate distributions for non-commensurate variables. Two general likelihood-based multivariate approaches have been proposed to avoid the direct specification of the joint distribution of the outcomes: factorizing the joint distribution of the outcomes and introducing an unobserved (latent) variable to model the correlation among the multiple outcomes.

The main idea of the factorization method is to write the likelihood as the product of the marginal distribution of one outcome and the conditional distribution of the second outcome given the previous outcome. Cox and Wermuth [1] discussed two possible factorizations for modeling a continuous and a binary outcome as functions of covariates. Fitzmaurice and Laird [2], and Catalano and Ryan [3] extended this approach to situations of clustered data.

Several models using latent variables have been proposed to analyze multiple non-commensurate outcomes as functions of covariates. Sammel et al. [4] discussed a model where the outcomes are assumed to be a physical manifestation of a latent variable and conditional on this latent variable, the observed outcomes follow a one-parameter exponential family model. The observed outcomes are modeled as functions of fixed covariates and a subject-specific latent variable. A drawback of this model that was later addressed by the authors, is its non-robustness to misspecification of the covariance because the mean parameters depend heavily on the covariance parameters. For example, if the outcomes are not correlated, the estimates of the covariate effects may be biased [5].

Arminger and Kusters [6] considered each outcome as a manifestation of an underlying continuous latent variable that is normally distributed. Dunson [7] extended this approach to accommodate non-normal latent variables, clustered data, non-linear relationships between the observed outcome and the underlying variables, multiple latent variables for each outcome type and covariate-dependent modifications of the relationship between the latent and underlying variables. Similar approaches were presented in the context of toxicity studies where longitudinal measurements are taken regarding multiple outcomes [8;9]. Although very general, Dunson's approach [7] produces a non-identifiable model for the case of a bivariate, binary or continuous outcome. This fact is well known in factor analysis (see for example Reilly [10]) where each factor needs to be a combination of three or more outcomes in order for the model to be identifiable, otherwise the parameter space has to be reduced. Often this is achieved by fixing some parameters to a constant. However, in Dunson's model, it is not clear how to constrain the parameters to make the model identifiable without misspecifying the model for the mean or covariance. Lin et al. [11] addressed a similar identifiability problem in the context of models for multiple continuous outcomes by scaling the outcomes to have the same variance.

A quasi-likelihood approach was also proposed for non-commensurate outcomes. The generalized estimating equations (GEE) described by Liang and Zeger [12] were extended by Prentice and Zhao [13], and Zhao et al. [14] for mixtures of continuous and discrete outcomes. In their approach separate equations are used for each outcome and a working correlation matrix is used to induce the correlation among the outcomes. A sandwich-type variance can then be computed for the model parameters which is robust to the misspecification of the working correlation matrix. Despite the attractive properties of this approach, we are unaware of its use in practice.

In this paper we review the different approaches to model a binary and a continuous outcome. We introduce a new latent variable model by constraining the parameters of the latent model proposed by Dunson [7] for identifiability without restrictions on the correlation. We show that this latent model is equivalent to the factorization model presented by Catalano and Ryan

[3] by demonstrating that they are only different parameterizations of the same model. We also implement the GEE approach proposed by Prentice and Zhao [13]. Simulation studies are used to compare consistency, efficiency, and coverage of the multivariate approach with the univariate approach. Section 2 describes the usual univariate approach, and both likelihood-based and quasi-likelihood multivariate methods to model a continuous and a binary outcome. In Section 3 we compare estimates obtained from the latent variable model, the factorization model, and the GEE with those from the univariate approach in terms of bias and efficiency. Finally in Section 4 three real data sets illustrate our methods.

2. Models for Bivariate Binary and Continuous Outcomes

Let y_{bi} denote a binary outcome, y_{ci} denote continuous outcome for the i th of n patients, and $\mathbf{x}_{bi}$ and $\mathbf{x}_{ci}$ denote $r_b \times 1$ and $r_c \times 1$ vectors of covariates associated with each outcome. We use subindex k to denote a particular covariate, x_{bk} or x_{ck} . We use a *probit* link for the binary outcome and the *identity* link for the continuous outcome. In some models these link functions arise naturally from construction and in other models, the links are used for illustration only, although other links could be used.

2.1. Univariate Models

One common approach to model multiple outcomes as functions of covariates is to ignore the correlation between the outcomes and fit a separate model to each response variable. In this setting we use a probit regression for the binary response and a linear regression for the continuous response,

$$\begin{aligned} \text{probit}(E(y_{bi}|\mathbf{x}_{bi})) &= \text{probit}(\mu_{bi}) = \mathbf{x}_{bi}^T \beta_b \\ y_{ci}|\mathbf{x}_{ci} &= \mathbf{x}_{ci}^T \beta_c + \epsilon_i \end{aligned} \quad (1)$$

where $\beta_b = (\beta_{b1}, \dots, \beta_{br_b})$, $\beta_c = (\beta_{c1}, \dots, \beta_{cr_c})$, and $\epsilon_i \sim N(0, \sigma_c^2)$. The interpretation of the regression parameters for these models is the usual interpretation in univariate generalized linear regression models: β_{bk} is the change in the *probit* of the expected value of y_{bi} for an increase of one unit in the covariate x_{bk} and β_{ck} is the change in expected value of y_{ci} for one unit increase in the k th-covariate. Estimates for the regression parameters can be obtained by maximizing the likelihood.

2.2. Factorization Models

Fitzmaurice and Laird [2] proposed a model for a correlated binary and a continuous outcome based on the factorization of the joint distribution of the outcomes, $f(y_b, y_c) = f(y_b)f(y_c | y_b)$. The expected values of the outcomes are related to the covariates $\mathbf{x}_b$ and $\mathbf{x}_c$, for example,

$$\begin{aligned} \text{probit}(E(y_{bi}|\mathbf{x}_{bi})) &= \text{probit}(\mu_{bi}) = \mathbf{x}_{bi}^T \beta_b \\ y_{ci}|y_{bi}, \mathbf{x}_{ci}, \mathbf{x}_{bi} &= \mathbf{x}_{ci}^T \beta_c + \tau(y_{bi} - \mu_{bi}) + \epsilon_{ci} \end{aligned} \quad (2)$$

where $\epsilon_{ci} \sim N(0, \sigma_c^2)$ and τ is the parameter for the regression of y_{ci} on y_{bi} . Large absolute values of τ indicate a strong correlation between the two outcomes. If $\tau = 0$ the two outcomes are independent given the covariate(s). The correlation that results from this model is

$$\text{Corr}(y_{bi}, y_{ci} | \mathbf{x}_{bi}, \mathbf{x}_{ci}) = \begin{cases} \frac{\text{sign}(\tau)}{\sqrt{1 + \frac{\sigma_c^2}{\tau^2 \text{Var}(y_{bi} | \mathbf{x}_{bi})}}}, & \text{if } \tau \neq 0 \\ 0, & \text{if } \tau = 0 \end{cases} \quad (3)$$

This factorization of the joint distribution has the convenient property that the model parameters have a marginal interpretation. β_{bk} is change of the *probit* expected value of y_{bi} for a one unit increase in the k th-covariate and because the term $(y_{bi} - \mu_{bi})$ has mean 0, β_{ck} is the change in the expected value of $y_{ci} | \mathbf{x}_{ci}, \mathbf{x}_{bi}$ for an increase of one unit in the covariate x_{ck} . Another characteristic of this model that makes it different from other approaches is the assumption regarding the distribution of y_{ci} . Conditional on y_{bi} and the covariates y_{ci} is assumed to be normally distributed, implying that the marginal distribution of y_{ci} is a mixture of two normals. For a high correlation between the two response variables, the marginal distribution of $y_{ci} | \mathbf{x}_{ci}, \mathbf{x}_{bi}$ will in fact be bimodal. Therefore the covariance of $y_{ci} | \mathbf{x}_{ci}, \mathbf{x}_{bi}$ depends on $\mathbf{x}_{bi}$, i.e., $\text{Var}(y_{ci} | \mathbf{x}_{ci}, \mathbf{x}_{bi}) = \tau^2 \Phi(\mathbf{x}_{bi}^T \beta_b) (1 - \Phi(\mathbf{x}_{bi}^T \beta_b)) + \sigma_c^2$.

Maximum likelihood estimates for the regression parameters of the factorization method can be obtained with commonly used algorithms for maximizing the likelihood. The log-likelihood function under the factorization model (2) is

$$\begin{aligned} l(y_b, y_c) &= \log \prod_{i=1}^n f(y_{bi}, y_{ci} | \mathbf{x}_{bi}, \mathbf{x}_{ci}) = \log \prod_{i=1}^n f(y_{ci} | y_{bi}, \mathbf{x}_{ci}, \mathbf{x}_{bi}) f(y_{bi} | \mathbf{x}_{bi}) \\ &= \sum_{i=1}^n \left(-\frac{1}{2} \log(2\pi\sigma_c^2) - \frac{1}{2\sigma_c^2} (y_{ci} - \mu_{ci} - \tau(y_{bi} - \Phi(\mu_{bi})))^2 \right) + \\ &\quad + \sum_{i=1}^n (y_{bi} \log(\Phi(\mu_{bi})) + (1 - y_{bi}) \log(\Phi(1 - \mu_{bi}))) \end{aligned} \quad (4)$$

where $\mu_{bi} = \mathbf{x}_{bi}^T \beta_b$, $\mu_{ci} = \mathbf{x}_{ci}^T \beta_c$ and $\Phi(\cdot)$ represents the *cdf* of the standard normal distribution.

The factorization of the joint distribution of y_{bi} and y_{ci} can also be considered in the reverse order: $f(y_b, y_c) = f(y_c) f(y_b | y_c)$ [1]. The model for the two outcomes is written as:

$$\text{probit}(E(y_{bi} | \mathbf{x}_{ci}, \mathbf{x}_{bi})) = \text{probit}(\mu_{bi}) = \mathbf{x}_{bi}^T \beta_b + \tau' \frac{y_{ci} - E(y_{ci} | \mathbf{x}_{ci})}{y_{ci} | \mathbf{x}_{ci} = \mathbf{x}_{ci}^T \beta_c + \epsilon_{ci}} \quad (5)$$

where $\epsilon_{ci} \sim N(0, \sigma_c^2)$ and τ' is the parameter for the regression of y_{bi} on y_{ci} . In this case the interpretation of the regression parameters for the binary outcome is conditional on the continuous outcome. To obtain the marginal effects we have to average over y_{ci} . The marginal

effect of the covariates on the binary outcome is then $\frac{\beta_b}{\sqrt{1 + \tau^2 \sigma_c^2}}$.

2.3. Latent Variable Models

Sammel et al. [4] presented a latent variable model where it is assumed that the observed outcomes are physical manifestations of a latent variable. Conditional on this latent variable, the outcomes are assumed to be independent, and are modeled as functions of fixed covariates and a subject-specific latent variable. The effect of the covariate(s) is modeled through the latent variable. Let u_i denote the latent variable and x_{ik} the covariate of interest, such as treatment. Then, $u_i | x_{ik} = \gamma x_{ik} + \delta_i$, with $\delta_i \sim N(0, 1)$. The parameter γ represents the association

between the covariate and the unobserved latent variable. The outcomes are modeled as functions of the latent variable:

$$\begin{aligned} \text{probit}(E(y_{bi}|u_i)) &= \beta_{b1} + \beta_{b2}u_i \\ y_{ci}|u_i &= \beta_{c1} + \beta_{c2}u_i + \epsilon_{ci} \end{aligned} \quad (6)$$

and $\epsilon_{ci} \sim N(0, \sigma_c^2)$. Here, β_{b2} and β_{c2} indicate the strength of the association between the observed outcomes and the latent variable. Conceptually this model is very appealing because it translates the idea that the outcomes are measuring an underlying construct. A drawback however is that some covariance parameters are also present in the mean. For example, because $E(y_{ci}|x_{ik}) = \beta_{c1} + \beta_{c2}\gamma_{x_{ik}}$ and $\text{Var}(y_{ci}|x_{ik}) = \beta_{c2}^2 + \sigma_c^2$ the model is very sensitive to misspecification of the correlation structure [5].

Another approach based on latent variables was proposed by Dunson [7]. A major difference between this approach and Sammel's approach relates to the association between the responses and the covariates. In Dunson's approach the covariates are not included in the model through the latent variable but rather introduced separately. For the case of a binary and a continuous outcome, Dunson's model would be written as

$$\begin{aligned} \text{probit}(E(y_{bi}|\mathbf{x}_{bi}, u_i)) &= \mathbf{x}_{bi}^T \beta_b + \lambda_b u_i \\ y_{ci}|\mathbf{x}_{ci}, u_i &= \mathbf{x}_{ci}^T \beta_c + \lambda_c u_i + \epsilon_{ci} \end{aligned} \quad (7)$$

where $\epsilon_{ci} \sim N(0, \sigma_c^2)$ and $u_i \sim N(0, \sigma_u^2)$ is a subject-specific latent variable. The means and covariance structure are modeled through difference parameters. The latent variable shared by both outcomes induces the correlation and it is assumed that given the latent variable, the two outcomes are independent. However, λ_b , λ_c , σ_u and σ_c are not identifiable. Fixing these parameters to any constant will result in a misspecification of the correlation between the outcomes. To better understand this argument, consider a similar model for two correlated continuous outcomes, y_1 and y_2 ,

$$\begin{aligned} y_{1i}|\mathbf{x}_{1i}, u_i &= \mathbf{x}_{1i}^T \beta_1 + \lambda_1 u_i + \epsilon_{1i} \\ y_{2i}|\mathbf{x}_{2i}, u_i &= \mathbf{x}_{2i}^T \beta_2 + \lambda_2 u_i + \epsilon_{2i} \end{aligned} \quad (8)$$

where $\epsilon_{1i} \sim N(0, \sigma_1^2)$, $\epsilon_{2i} \sim N(0, \sigma_2^2)$ and $u_i \sim N(0, \sigma_u^2)$. The parameters associated with the variance components of the outcomes (λ_1 , λ_2 , σ_u , σ_1 and σ_2) are not identifiable. There are five parameters to be estimated but only information from the $\text{Var}(y_1)$, $\text{Var}(y_2)$ and $\text{Cov}(y_1, y_2)$. We have to restrict at least two parameters to obtain an identifiable model. The correlation

induced by the model is given by $\frac{\lambda_1 \lambda_2 \sigma_u^2}{\sqrt{(\lambda_1^2 \sigma_u^2 + \sigma_1^2)(\lambda_2^2 \sigma_u^2 + \sigma_2^2)}}$. If we constrain the parameters λ_1

and λ_2 to be 1, for example, the correlation becomes $\frac{\sigma_u^2}{\sqrt{(\sigma_u^2 + \sigma_1^2)(\sigma_u^2 + \sigma_2^2)}}$. It is easy to build a case where such model fails to induce the correct correlation. Suppose that the

$\text{Var}(y_1|\mathbf{x}_{1i}) = \sigma_u^2 + \sigma_1^2 = .5$, $\text{Var}(y_2|\mathbf{x}_{2i}) = \sigma_u^2 + \sigma_2^2 = 5$ and $\text{Corr}(y_1, y_2 | \mathbf{x}_{1i}; \mathbf{x}_{2i}) = .8$. So, $\sigma_u^2 < .5$ and

the correlation induced by the model (7) becomes $\text{Corr}(y_1, y_2 | \mathbf{x}_{1i}, \mathbf{x}_{2i}) < \frac{.5}{\sqrt{.5 \times 5}} \approx .32$ which is incorrect. Fixing the variances of the error terms or the latent variable will lead to similar

inconsistencies. A similar argument can be given for the model in (7). Although there is one less parameter than (8), there is less information to estimate the parameters because $Var(y_b | x_{bi})$ is fully determined by the $E(y_b | x_{bi})$.

2.3.1. A New Latent Variable Model—To determine appropriate constraints to the parameters in (7) we use an idea similar to the scaled multivariate mixed model proposed by Lin and Ryan [11]. Let y_1 and y_2 be two continuous normally distributed outcomes associated with covariates x_1 and x_2 , respectively. Given the covariates, we assume that the two outcomes are correlated. We define $y_1^* = \frac{y_1}{\sigma_1}$ and $y_2^* = \frac{y_2}{\sigma_2}$ where σ_1 and σ_2 are scaling parameters such that

$$\begin{aligned} y_{1i}^* | \mathbf{x}_{1i}, u_i &= \frac{y_1}{\sigma_1} | \mathbf{x}_{1i}, \mathbf{x}_{2i}, u_i = \mathbf{x}_{1i}^T \beta_1^* + u_i + \epsilon_{1i}^* \\ y_{2i}^* | \mathbf{x}_{2i}, u_i &= \frac{y_2}{\sigma_2} | \mathbf{x}_{1i}, \mathbf{x}_{2i}, u_i = \mathbf{x}_{2i}^T \beta_2^* + u_i + \epsilon_{2i}^* \end{aligned} \quad (9)$$

where $\epsilon_{1i}^* \sim N(0, 1)$, $\epsilon_{2i}^* \sim N(0, 1)$ and $u_i \sim N(0, \sigma_u^2)$ is a latent variable that induces the correlation between the two variables y_{1i}^* and y_{2i}^* . We can rewrite (9) and obtain the final expression for a latent model for two continuous outcomes:

$$\begin{aligned} y_{1i} | \mathbf{x}_{1i}, u_i &= \mathbf{x}_{1i}^T \beta_1 + \sigma_1 u_i + \epsilon_{1i} \\ y_{2i} | \mathbf{x}_{2i}, u_i &= \mathbf{x}_{2i}^T \beta_2 + \sigma_2 u_i + \epsilon_{2i} \end{aligned} \quad (10)$$

where $\beta_1 = \sigma_1 \beta_1^*$, $\beta_2 = \sigma_2 \beta_2^*$, $\epsilon_{1i} \sim N(0, \sigma_1^2)$, $\epsilon_{2i} \sim N(0, \sigma_2^2)$ and $u_i \sim N(0, \sigma_u^2)$. The correlation between

the two outcomes induced by the model is $Corr(y_1, y_2 | \mathbf{x}_1, \mathbf{x}_2) = \frac{\sigma_u^2}{1 + \sigma_u^2}$. So, the range of correlations that we can model is $[0; 1)$ which requires that the outcomes are positively correlated. In many practical situations the researcher can anticipate the sign of the correlation. If the outcomes are expected to be negatively correlated a possible solution is to invert the coding of the binary outcome or to multiply the continuous outcomes by -1 and this way reverse the sign of the correlation.

These considerations motivate the constraints for the model (7) as follows. Let y_b and y_c be a binary and a continuous variable associated with covariates x_b and x_c , respectively. We want to develop a multivariate model that takes into account the potential correlation between y_b and y_c . The variable y_c is assumed to be normally distributed given the covariates x_c . Suppose there is an underlying variable y_{bi}^* , normally distributed given the covariates x_{bi} , that is associated with the binary outcome, y_{bi} , in the following way:

$$y_{bi} = \begin{cases} 0, & \text{if } y_{bi}^* \leq 0 \\ 1, & \text{if } y_{bi}^* > 0 \end{cases} \quad (11)$$

Define $y_{ci}^* = \frac{y_{ci}}{\sigma_c}$ where σ_c is a scale parameter for the continuous outcome. The regression equations for the two variables can be written as:

$$\begin{aligned} y_{bi}^* | \mathbf{x}_{bi}, u_i &= \mathbf{x}_{bi}^T \beta_b^* + u_i + \epsilon_{bi}^* \\ y_{ci}^* | \mathbf{x}_{ci}, u_i &= \mathbf{x}_{ci}^T \beta_c^* + u_i + \epsilon_{ci}^* \end{aligned} \quad (12)$$

with $\epsilon_{bi}^* \sim N(0, 1)$, $\epsilon_{ci}^* \sim N(0, 1)$ and $u_i \sim N(0, \sigma_u^2)$. The variances of the error terms are fixed at 1 by design. This is just a convenient standardization to obtain a common variance and does not represent a restriction of the model. Any other standardization would work as well. The latent variable u_i is introduced in both equations to induce the correlation between the outcomes. It is assumed that given u_i , y_{bi}^* and y_{ci}^* are independent and consequently y_{bi} and y_{ci} are also independent given u_i .

Because $E(y_{ci}) = \sigma_c E(y_{ci}^*)$, we can write the equation for the continuous outcome as $y_{ci} | \mathbf{x}_{ci}, u_i = \mathbf{x}_{ci}^T \beta_c + \sigma_c u_i + \epsilon_{ci}$, where $\beta_c = \sigma_c \beta_c^*$ and $\epsilon_{ci} \sim N(0, \sigma_c^2)$. The correlation between y_{bi}^* and y_{ci} is a function only of σ_u and is given by $\frac{\sigma_u^2}{1 + \sigma_u^2}$. However, y_{bi}^* is not observed. We can write the regression equation for the binary outcome, y_{bi} , as

$P(y_{bi} = 1 | \mathbf{x}_{bi}, u_i) = P(y_{bi}^* > 0 | \mathbf{x}_{bi}, u_i) = \Phi(\mathbf{x}_{bi}^T \beta_b^* + u_i)$. The final model is then,

$$\begin{aligned} \text{probit}(P(y_{bi} = 1 | \mathbf{x}_{bi}, u_i)) &= \mathbf{x}_{bi}^T \beta_b^* + u_i \\ y_{ci} | \mathbf{x}_{ci}, u_i &= \mathbf{x}_{ci}^T \beta_c + \sigma_c u_i + \epsilon_{ci} \end{aligned} \quad (13)$$

The correlation between y_{bi} and y_{ci} that results from this model can be calculated as:

$$\text{Corr}(y_{bi}, y_{ci} | \mathbf{x}_{bi}, \mathbf{x}_{ci}) = \frac{\sigma_u^2}{(1 + \sigma_u^2)} \frac{\phi\left(\frac{\mathbf{x}_{bi}^T \beta_b^*}{\sqrt{\sigma_u^2 + 1}}\right)}{\sqrt{\Phi\left(\frac{\mathbf{x}_{bi}^T \beta_b^*}{\sqrt{\sigma_u^2 + 1}}\right) \left(1 - \Phi\left(\frac{\mathbf{x}_{bi}^T \beta_b^*}{\sqrt{\sigma_u^2 + 1}}\right)\right)}} \quad (14)$$

where $\phi(\cdot)$ is the standard normal density.

The parameters β_b^* in (13) are interpreted conditional on u_i . Given u_i , β_{bk}^* is the change on the *probit* of the expected value of y_{bi} for an increase of one unit in the covariate x_{bk} . For this reason the parameters β_b^* of the latent model cannot be directly compared with the regression parameters of the marginal models such as (1) and (2). To obtain the marginal effects that can be compared with the other models we have to average over the u_i 's (see eq. 15).

$$P(y_{bi} | \mathbf{x}_{bi}) = \int P(y_{bi} | \mathbf{x}_{bi}, u_i) f(u_i) du_i = \Phi\left(\frac{\mathbf{x}_{bi}^T \beta_b^*}{\sqrt{1 + \sigma_u^2}}\right).$$

So, $\beta_b = \frac{\beta_b^*}{\sqrt{1 + \sigma_u^2}}$ are the marginal effects associated with the covariates. For the continuous outcome, β_c is interpreted as conditional or marginal effects of the covariates.

The log likelihood for the model is written as:

$$\begin{aligned}
l(y_b, y_c) &= \log \prod_{i=1}^n f(y_{bi}, y_{ci} | \mathbf{x}_{bi}, \mathbf{x}_{ci}) \\
&= \log \prod_{i=1}^n \int f(y_{bi} | u_i, \mathbf{x}_{bi}) f(y_{ci} | u_i, \mathbf{x}_{ci}) f(u_i) du_i \\
&= \log \prod_{i=1}^n \int_{-\infty}^{\infty} [\Phi(\mu_{bi} + u_i)]^{y_{bi}} [1 - \Phi(\mu_{bi} + u_i)]^{(1-y_{bi})} \frac{\exp\left(-\frac{(y_{ci} - \mu_{ci} - \sigma_c u_i)^2}{2\sigma_c^2}\right)}{\sqrt{2\pi\sigma_c^2}} \frac{\exp\left(-\frac{u_i^2}{2\sigma_u^2}\right)}{\sqrt{2\pi\sigma_u^2}} du_i
\end{aligned} \tag{15}$$

where $\mu_{bi} = \mathbf{x}_{bi}^T \beta_b^*$ and $\mu_{ci} = \mathbf{x}_{ci}^T \beta_c$. Estimates for the marginal effects $\widehat{\beta}_b$ are obtained using

$\frac{\widehat{\beta}_b^*}{\sqrt{1 + \widehat{\sigma}_u^2}}$. The estimated standard errors for $\widehat{\beta}_b$ can be approximated using the Delta method.

The properties of the *probit* link allow a simplification of the likelihood for the latent variable model. The integral in (15) has a closed-form solution and solving this integral (Appendix 1) we obtain the same model as the reverse factorization (5) but with a different parameterization:

$$\begin{aligned}
l(y_b, y_c) &= \log \prod_{i=1}^n \left[\Phi(\mathbf{x}_{bi}^T \beta_b + \tau(y_{ci} - \mathbf{x}_{ci}^T \beta_c)) \right]^{y_{bi}} \\
&\quad \times \left[1 - \Phi(\mathbf{x}_{bi}^T \beta_b + \tau(y_{ci} - \mathbf{x}_{ci}^T \beta_c)) \right]^{1-y_{bi}} \phi\left(\frac{y_{ci} - \mathbf{x}_{ci}^T \beta_c}{\sqrt{\sigma_c^2(\sigma_u^2 + 1)}}\right)
\end{aligned} \tag{16}$$

$$\text{where } \beta_b = \beta_b^* \sqrt{\frac{\sigma_u^2 + 1}{2\sigma_u^2 + 1}} \text{ and } \tau = \frac{\sigma_u^2}{\sqrt{2\sigma_u^2 + 1}}.$$

2.4. Generalized Estimating Equations

Liang and Zeger [12] introduced the methodology of generalized estimating equations (GEE) in the context of longitudinal data. In this methodology, the correlation among measurements on the same individual (or in the same cluster) is treated as a nuisance parameter. A 'working' correlation matrix is plugged in the equations to obtain estimates for the regression parameters. These estimators are consistent even if the 'working' correlation matrix is misspecified. The variances of the parameters estimators are obtained by correcting the 'working' correlation matrix resulting in what became known as the sandwich estimator. The main advantage of the GEE method is this robustness to misspecification of the covariance. Prentice and Zhao[13], and Zhao et al. [14] also proposed an estimation approach for mixed continuous and discrete using the quadratic exponential and the partly exponential families, respectively. For the case of a binary and a continuous outcomes if we assume the following model for the means of the two outcomes,

$$\begin{aligned}
\text{probit}(E(y_{bi} | \mathbf{x}_{bi})) &= \text{probit}(\mu_{bi}) = \mathbf{x}_{bi}^T \beta_b \\
E(y_{ci} | \mathbf{x}_{ci}) &= \mu_{ci} = \mathbf{x}_{ci}^T \beta_c
\end{aligned} \tag{17}$$

then the estimating equation

$$\sum_{i=1}^n D_i^T V_i^{-1} \begin{pmatrix} y_{bi} - \mu_{bi} \\ y_{ci} - \mu_{ci} \end{pmatrix} = 0 \tag{18}$$

has a solution that is a consistent and asymptotically normal estimator for β_b and β_c [14] with variance $\Gamma^{-1}\Omega^{-1}$, where V_i is a 'working' covariance matrix for y_{bi} and y_{ci} , ρ is the correlation between the outcomes, $\Gamma = E(D_i^T V^{-1} D_i)$ and

$$D_i = \begin{pmatrix} \frac{\partial \mu_{bi}}{\partial \beta_b} & \frac{\partial \mu_{bi}}{\partial \beta_c} \\ \frac{\partial \mu_{ci}}{\partial \beta_b} & \frac{\partial \mu_{ci}}{\partial \beta_c} \end{pmatrix} \quad V_i = \begin{pmatrix} \sigma_b^2 & \sigma_b \sigma_c \rho \\ \sigma_b \sigma_c \rho & \sigma_c^2 \end{pmatrix}$$

$$\Omega = E \left(D_i^T V^{-1} \begin{pmatrix} y_{bi} - \mu_{bi} \\ y_{ci} - \mu_{ci} \end{pmatrix} \begin{pmatrix} y_{bi} - \mu_{bi} \\ y_{ci} - \mu_{ci} \end{pmatrix}^T V^{-1} D_i \right) \quad (19)$$

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 β_b and β_c even if V_i is misspecified. There are several strategies to obtain estimates for the parameters in the covariance matrix. A simple solution is to use the method of moments to estimates σ_b , σ_c and ρ ,

$$\hat{\sigma}_b^2 = \frac{\sum_{i=1}^n (y_{bi} - \hat{\mu}_{bi})^2}{n} \quad \hat{\sigma}_c^2 = \frac{\sum_{i=1}^n (y_{ci} - \hat{\mu}_{ci})^2}{n} \quad \hat{\rho} = \frac{\sum_{i=1}^n (y_{ci} - \hat{\mu}_{ci})(y_{bi} - \hat{\mu}_{bi})}{\sqrt{\sum_{i=1}^n (y_{bi} - \hat{\mu}_{bi})^2 \sum_{i=1}^n (y_{ci} - \hat{\mu}_{ci})^2}} \quad (20)$$

where $\hat{\mu}_{bi}$ can be obtained by, for example, running a *probit* and a linear regression as in (1). Any other consistent estimates could be used instead of (20), for example $\hat{\sigma}_b^2 = \hat{\mu}_{bi}(1 - \hat{\mu}_{bi})$.

3. Simulation Study

3.1. Simulation settings

We performed a Monte Carlo simulation study to investigate consistency, efficiency and coverage of 95% confidence intervals for estimates obtained by the univariate model, factorization model, latent variable model and GEE. Two different sets of simulations were considered. In the first set, two outcomes associated with a common covariate (exposure) were simulated. Different effect sizes of the covariate on the outcomes were used to simulate no effect, small effect and large effect. Data were generated from a bivariate normal distribution,

$$\begin{pmatrix} y_{bi}^* \\ y_{ci} \end{pmatrix} \sim MVN \left(\begin{pmatrix} -0.5 + \beta_{b1} x_i \\ 5 + \beta_{c1} x_i \end{pmatrix}, \begin{pmatrix} 1 & 6\rho \\ 6\rho & 36 \end{pmatrix} \right) \quad (21)$$

with x_i generated from a Bernoulli(.5). In the first simulation, the vector of coefficients associated with the covariate was chosen as $(\beta_{b1}, \beta_{c1}) = (0, 2)$, representing no effect of x on y_b and a small effect (defined as 1/3 of a standard deviation) on y_c . For the second simulation, the vector of coefficients was chosen as $(\beta_{b1}, \beta_{c1}) = (0.2, 2)$ representing a small effect on both outcomes (1/5 and 1/3 of a standard deviation respectfully). Finally, $(\beta_{b1}, \beta_{c1}) = (1, 6)$ representing a large effect (1 standard deviation) of x on y_b and y_c .

In the second set of simulations, a different covariate was added to each outcome and data were generated from

$$\begin{pmatrix} y_{bi}^* \\ y_{ci} \end{pmatrix} \sim MVN \left(\begin{pmatrix} -1 + \beta_{b1}x_i + \beta_{b2}x_{bi} \\ 5 + \beta_{c1}x_i + \beta_{c2}x_{ci} \end{pmatrix}, \begin{pmatrix} 1 & 6\rho \\ 6\rho & 36 \end{pmatrix} \right) \quad (22)$$

with x_i generated from a Bernoulli(.5), x_{bi} generated from $N(0,1)$ and x_{ci} generated from a $N(0,4)$. For this set of simulations, the vector of coefficients (β_{b1}, β_{c1}) was chosen as in the first set, combining different situations of no effect, small effect and large effect of the covariate x on the two outcomes.

For each simulation, we used (11) to create the binary variable y_{bi} from y_{bi}^* . The covariates x_i , x_{bi} and x_{ci} , were chosen so that the simulation would include binary and continuous covariates with some ad-hoc distribution. The estimation of the parameters that define the mean structure is expected to be identical for the different models used. Hence, the key parameter for our simulation study is the correlation between the two underlying variables y_{bi}^* and y_{ci} because it is the parameter that should have an impact on the standard errors of the estimates obtained by the different approaches. We thus generated datasets with different levels of correlation ($\rho = 0, .3, .6, .9$). For each level of correlation we generated 1000 independent samples with 200 subjects each. However, the correlation between the outcomes y_{bi} and y_{ci} depends on the covariate values. For $x_i = 0$ (and for $x_{bi} = x_{ci} = 0$) the correlations between the outcomes y_{bi} and y_{ci} corresponding to $(\rho = 0, .3, .6, .9)$ are $(0, .2, .5, .7)$, respectively.

The data generated from (22) were modeled using the following:

1. Univariate approach (ignoring the correlation between the outcomes)

$$\begin{aligned} \text{probit} (P(y_{bi}=1|x_i, x_{bi})) &= \alpha_b + \beta_{b1}x_i + \beta_{b2}x_{bi} \\ y_{ci}|x_i, x_{ci} &= \alpha_c + \beta_{c1}x_i + \beta_{c2}x_{ci} + \epsilon_{ci}, \\ &\epsilon_{ci} \sim N(0, \sigma_c^2) \end{aligned} \quad (23)$$

2. Factorization approach

$$\begin{aligned} \text{probit} (P(y_{bi}=1|x_i, x_{bi})) &= \alpha_b + \beta_{b1}x_i + \beta_{b2}x_{bi} \\ y_{ci}|x_i, x_{ci}, x_{bi} &= \alpha_c + \beta_{c1}x_i + \beta_{c2}x_{ci} + \tau(y_{bi} - E(y_{bi}|x_i, x_{bi})) + \epsilon_{ci}, \\ &\epsilon_{ci} \sim N(0, \sigma_c^2) \end{aligned} \quad (24)$$

3. Latent variable approach

$$\begin{aligned} \text{probit} (P(y_{bi}=1|x_i, x_{bi}, u_i)) &= \alpha_b^* + \beta_{b1}^*x_i + \beta_{b2}^*x_{bi} + u_i \\ y_{ci}|x_i, x_{ci}, u_i &= \alpha_c + \beta_{c1}x_i + \beta_{c2}x_{ci} + \sigma_c u_i + \epsilon_{ci}, \\ u_i &\sim N(0, \sigma_u^2) \text{ and } \epsilon_{ci} \sim N(0, \sigma_c^2) \end{aligned} \quad (25)$$

4. Generalized estimating equations

$$\begin{aligned} \text{probit} (\mu_{bi}) &= \alpha_b + \beta_{b1}x_i + \beta_{b2}x_{bi} \\ \mu_{ci} &= \alpha_c + \beta_{c1}x_i + \beta_{c2}x_{ci} \end{aligned} \quad (26)$$

and the estimating equation as described in section 2.4.

For data generated from (21) similar models were used but without the terms associated with

x_{bi} and x_{ci} . For the latent variable model, the parameters $\alpha_b = \frac{\alpha_b^*}{\sqrt{1+\sigma_u^2}}$, $\beta_{b1} = \frac{\beta_b^*}{\sqrt{1+\sigma_u^2}}$ and $\beta_{b2} = \frac{\beta_{b2}^*}{\sqrt{1+\sigma_u^2}}$ corresponding to the marginal effects were computed and used for comparison with the regression parameters in the other models. The latent variable model (25) is the correct model given our data generation process. Both the univariate and factorization models have the correct structure for the means but not for the covariance (except the univariate models when $\rho = 0$). The univariate approach (23) assumes the outcomes are independent and the factorization model (24) assumes that the variance of $y_{ci} | x_i, x_{ci}, x_{bi}$ depends on the covariates x_i and x_{bi} .

The models (23), (24), and (26) were fitted using PROC NLMIXED from SAS to assure the same numerical algorithms were used to maximize the likelihoods. An example of the SAS code to fit the latent variable model is presented in the Appendix. The 95% confidence intervals for the parameter estimates were computed as $\hat{\gamma} + t_{.975} \widehat{SE}(\hat{\gamma})$, where $\hat{\gamma}$ represents the maximum likelihood estimate for parameter of interest. The GEE were solved using a program written in PROC IML from SAS. The nonlinear optimization algorithm by Nelder-Mead simplex method implemented in PROC IML was used because it was the most successful in converging to the solutions in the simulated datasets. Estimates of the parameters of the covariance matrix were obtained using the *probit* and linear regression from (23). The same estimates were used as initial values for the optimization algorithm.

3.2. Simulation results

All the settings produced identical point estimates (MLEs) of the parameters, despite the model used to fit the data, the effect size of the covariate or the correlation level between the two outcomes. This indicates that all the models produce consistent estimates of the regression parameters. Coverage of the confidence intervals were also close to the nominal value (95%) in all simulations. The only difference observed between the models was found on the standard errors of the estimates for some settings.

Because the MLEs were identical across all models, the differences in the mean square errors (MSE) observed in some settings are mostly due to the differences on the standard error of the estimates. Tables I, II and III present the ratio of the MSE of the multivariate models (factorization model, latent variable model and GEE) to the univariate model in different settings depending on the effect size of the shared covariate and correlation level between the outcomes.

The results are summarized as follows. For the estimates of the parameters associated with the covariate shared by the two outcomes, the multivariate models produced estimates with MSE identical to the univariate model. The only exceptions were observed for the β_{b1} estimates when the correlation between the outcomes was large. In this case the multivariate models had lower MSE than the univariate model. The latent variable model, in this situation, produced the estimates with lowest MSE. When the true model involved different covariates associated with each outcome, the estimates of the parameters associated with the unshared covariates had a lower MSE for the multivariate models if the outcomes were correlated. For example, the latent variable model produced some estimates with approximately half the MSE than the univariate model, for a high correlation between the outcomes.

4. Applications

The first Example 4.1 illustrates the similar performances of the approaches when the outcomes share the same covariates and the correlation between the outcomes is low. Example 4.2 illustrates a similar situation to Example 4.1 but with strong correlation between the outcomes. Example 4.3 illustrates how inferences can change with a multivariate approach if the outcomes are associated with different covariates.

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

Dickey et al. [15] conducted a prospective observational study of 420 adults with schizophrenia who sought care for a psychiatric crisis. The main study objective was to compare care for patients who were and were not enrolled in managed care. 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. Higher values for the self-reported quality represent higher quality. The means (SD) age of patients were 40 (8.5) and 41 (7.9) in the managed care and not managed care groups respectively. 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 means (SD) self-reported quality of interpersonal interactions between patient and clinician appeared similar, 3.20 (0.67) for the managed care group and 3.21 (0.65) for the not managed group. We used the univariate (1), the factorization model(2), the latent variable model (13) and the GEE (as described in section 2.4) to estimate the marginal association of managed care and outcomes. No other covariates were used in the models. Only patients with complete data were included (n=394).

The managed care estimates and the corresponding standard errors on patient/clinician relationship and anti-psychotic prescription were identical for all the models considered (Table IV). In this example, the marginal correlation between the two outcomes was low, 0.06. For the multivariate models, it is easy to test simultaneously for an overall effect of managed care exposure on the outcomes, i.e., $H_0 : \beta_b = \beta_c = 0$. This can be accomplished using a likelihood ratio test. The result for this test obtained through the latent variable model was p-value=0.97 ($\chi^2_2=.07$) indicating no evidence of a managed care effect on quality of care as measured by the two outcomes.

4.2. Example 2: Efficacy of Interferon- α on Vision for Macular Degeneration

The data arise from a randomized multi-center clinical trial comparing an experimental treatment (interferon- α) to a corresponding placebo in the treatment of patients with age-related macular degeneration. We focus on the comparison between placebo and the highest dose (6 million units daily) of interferon- α (Z). The full results of this trial have been reported elsewhere [16]. Patients with macular degeneration progressively lose vision. In the trial, a patient's visual acuity was assessed at different time points through their ability to read lines of letters on standardized vision charts. These charts display lines of letters of decreasing size which the patient must read from top (largest letters) to bottom (smallest letters). Each line with at least four letters correctly read is called one line of vision. The patient's visual acuity is the total number of lines correctly read. The primary endpoint of the trial was a binary outcome defined as the loss of at least three lines of vision at 1 year compared to their baseline performance. We also consider the difference between visual acuity at 6 months and baseline as a secondary endpoint (continuous outcome). We used the univariate, the factorization model, the latent variable model and the GEE to estimate the marginal effect of interferon- α treatment on visual performance. Treatment was the only covariate included in the models.

A total of 190 patients (87 in the treatment arm and 103 in the placebo arm) completed the study. The correlation between the two outcomes was 0.63. For patients who received the treatment, 54% lost at least three lines of vision at 1 year versus 38% in the placebo group. The mean (SD) loss of visual acuity at 6 months were 8.4 (11.9) letters for the treatment arm and 5.5 (13.7) for the placebo arm. The results of all approaches were identical despite the high correlation between the outcomes. However, the estimate of treatment effect for the binary outcome, loss of at least three lines of vision at 1 year, was smaller in the latent variable model (Table V). All models lead to the same conclusion regarding the poor performance of the interferon- α . The overall effect of treatment ($H_0: \beta_b = \beta_c = 0$) obtained by the latent model was not statistically significant ($\chi^2_2=4$, p-value=0.14).

4.3. Example 3: Restenosis Following Coronary Stenting Using Bare-Metal Stents

Coronary disease results from lesions of fatty plaque that build up within the arterial wall. These plaque lesions may either rupture, causing a heart attack, or gradually obstruct blood flow, causing angina. Coronary stents are thin expandable metallic tubes that are delivered within the coronary artery by a catheter and are then expanded precisely at the site of an obstructive lesion. Typically up to two primary endpoints (measures of restenosis) are measured after coronary stenting. We use data from one arm of a non-inferiority randomized trial of bare-metal coronary stents. The first endpoint obtained from all patients is the incidence of clinically-driven repeat revascularization, denoted the target lesion revascularization (TLR) rate (binary outcome). TLR is designated by a clinical events committee that have access to clinical and angiographic laboratory data. The second endpoint, proportion diameter stenosis (PDS), is the degree of vessel re-narrowing and is quantified by a computer-based system (continuous outcome). The PDS is obtained on a small randomly selected subset of patients. Both TLR and PDS are measured 9 months after coronary stenting. The goal is to estimate restenosis for diabetic patients taking into account potential confounders.

From the 313 patients, 105 had both PSD and TLR measured and included in the analysis. The overall rate of TLR was 14% and the mean (SD) of PDS was 0.43 (0.17). The correlation between the two outcomes was 0.58. Fourteen patients were diabetic. The overall mean (SD) for length of lesion was 12.8 (5.2). Using a univariate approach only history of diabetes mellitus (diabetes) was significantly associated with the outcomes. For the latent model, the lesion length was also associated with TLR but not with PDS (Table VI). Note that inference for diabetic patients is the same as in the univariate approach. If lesion length is included in the equation for the outcome TLR then both outcomes would share the same covariates and the results from the latent model become identical to the univariate model (the association of lesion length would not be significant in the latent model). This is in agreement with the simulations where efficiency gains were realized for estimates of the parameters associated with the 'non-shared' covariates.

5. Discussion

We presented different approaches to model correlated binary and continuous outcomes. We proposed a new multivariate latent variable model that overcomes the identifiability problems of Dunson's model and the sensitivity to misspecification of the covariance matrix of Sammel's model. We also implemented a quasi-likelihood approach based on a GEE. Simulation results suggest that the four approaches lead to consistent estimates of the regression parameters. Two findings are noteworthy. First, we demonstrated that if the two outcomes share the same covariates, the results of a multivariate approach are identical to that of a univariate approach that ignores the correlation between the outcomes. Although counterintuitive, this result is consistent with other situations of multivariate data. In the setting of seemingly unrelated regressions with normally distributed outcomes and for the particular case of common set of

covariates associated with the outcomes, the ordinary least squares estimate is still the best linear unbiased estimator (see for example Zellner [17] and Rotnitzky [18]), despite the correlation between the outcomes.

Second, we know that for binary outcomes jointly modeled with the same covariates, there is a small gain in efficiency by taking into account the correlation. This only occurs if the outcomes strongly associated. Our result for non-commensurate outcomes is a combination of these two properties. The estimates of the parameters associated with the continuous outcome have the same standard errors as the univariate approach. The estimates of the parameters associated with the binary outcome show a small gain in efficiency when compared with the univariate approach but only for high correlation between the outcomes.

Third, the efficiency gain is higher when the outcomes share a different set of covariates and with higher levels of correlation between the outcomes. This suggests that if one anticipates that different covariates maybe associated with the outcomes, the multivariate approach offers some advantages. Fitzmaurice and Laird [19] have previously shown higher gains in efficiency when compared with the univariate approach than those shown here. However, the efficiency gains observed by the authors were inflated as consequence of heteroscedasticity in the data. If data are generated under the factorization model, the variance depends on the covariate. In this case the univariate approach, assuming homoscedasticity, will lead to less efficient estimates due to misspecification of the variance.

The better performance of the latent variable model over the factorization model in our simulation study was expected because the data was generated from the latent model. Nonetheless, the factorization model was sometimes superior but never inferior to the univariate approach. This suggests that the misspecification of correlation between the outcomes will not be worse than the assumption of independence. In contrast to the factorization approach, the latent variable model presented is easily extended to several continuous and/or several binary outcomes by including additional latent variables as long as the outcomes are positively correlated. However, some of the assumptions of the model, such as the distribution of the latent variables, are not easily assessed. In the presence of missing observations in one of the outcomes, the factorization approach only uses the complete cases or it requires the EM-algorithm to include all the cases in the analysis [19]. This is not the case with the latent model. If the missing data is missing at random or missing completely at random [20] this situation can be easily accommodated due to the conditional independence of the outcomes given the latent variable. Furthermore, the latent variable model is easily fitted using standard software.

We focused on comparing the univariate and multivariate approaches using common operational characteristics such as MSE and coverage of the confidence intervals. We note that these characteristics may not fully capture the benefits of the multivariate models. Research to understand the advantage of adopting a multivariate model for joint inference of the parameters is an important next step. For example, when the outcomes represent an underlying construct and the primary research question relates to an exposure effect, joint inference may be a key task. Such situations occur in clinical trials with more than one primary endpoint or when there is simultaneous concern with safety and efficacy.

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. The schizophrenia managed care data were generously provided through the efforts of Barbara Dickey, Ph.D., Harvard Medical School, Boston, MA; the bare-metal stent data by Laura Mauri, M.D., M.Sc., Harvard Clinical Research Institute, Boston, MA; and the macular degeneration data by Geert Molenberghs, Ph.D., Hasselt University, Belgium.

APPENDIX 1: Likelihood for the latent variable model

We show that by solving the integral in the likelihood for the latent variable model (15) we get the likelihood of the reverse factorization model (5) but with a different parameterization:

$$l(y_b, y_c) = \log \prod_{i=1}^n \int f(y_{bi}|u_i, \mathbf{x}_{bi}) f(y_{ci}|u_i, \mathbf{x}_{ci}) f(u_i) du_i$$

$$\log \prod_{i=1}^n [\Phi(p_i)]^{y_{bi}} [1 - \Phi(p_i)]^{1-y_{bi}} \phi\left(\frac{y_{ci} - \mu_{ci}}{\sqrt{\sigma_c^2(\sigma_u^2 + 1)}}\right) \quad (27)$$

where,

$$p_i = \frac{\mu_{bi} + \frac{\sigma_u^2}{(\sigma_u^2 + 1)} \left(\frac{y_{ci} - \mu_{ci}}{\sigma_c} \right)}{\sqrt{\frac{2\sigma_u^2 + 1}{\sigma_u^2 + 1}}} \quad (28)$$

Letting $\beta_b = \beta_b^* \sqrt{\frac{\sigma_u^2 + 1}{2\sigma_u^2 + 1}}$ and $\tau = \frac{\sigma_u^2}{\sqrt{2\sigma_u^2 + 1}}$ we get

$$l(y_b, y_c) = \log \prod_{i=1}^n \left[\Phi(\mathbf{x}_{bi}^T \beta_b + \tau(y_{ci} - \mathbf{x}_{ci}^T \beta_c)) \right]^{y_{bi}} \quad (29)$$

$$\times \left[1 - \Phi(\mathbf{x}_{bi}^T \beta_b + \tau(y_{ci} - \mathbf{x}_{ci}^T \beta_c)) \right]^{1-y_{bi}} \phi\left(\frac{y_{ci} - \mathbf{x}_{ci}^T \beta_c}{\sqrt{\sigma_c^2(\sigma_u^2 + 1)}}\right) \quad (30)$$

$$= \log \prod_{i=1}^n f(y_{bi}|y_{ci}, \mathbf{x}_{bi}) f(y_{ci}|\mathbf{x}_{ci}) \quad (31)$$

This likelihood is the same likelihood for the reverse factorization model (5), i.e., both approaches are a different parameterization of the same model.

APPENDIX 2: SAS Code to Fit the Latent Variable Model

The SAS code below illustrates how to use the procedure PROC NLMIXED in SAS to fit the latent variable model (13) for a binary (y_1) and a continuous (y_2) outcomes associates with a common covariate (x_1).

```
proc nlmixed data=datasetname;
  parms a1=1 b1=1 a2=1 b2=1 sigmab=1 sigma2=1;
  bounds sigma2>0, sigmab>0;
  ll=y1*log(PROBNORM(a1+b1*x1+u))+(1-y1)*
    log(PROBNORM(-a1-b1*x1-u))-log(sigma2)-
    .5*1/(sigma2**2)*(y2-a2-b2*x1-u*sigmab)**2;
```



```

model y1 ~ general(11);
random u ~ normal(0,sigmas) subject=id;
estimate 'marginal effect of x1' b1/sqrt(1+sigmas);
run;

```

references

- [1]. Cox DR, Wermuth N. Response models for binary and quantitative variables. *Biometrika*. 1992; 79(3):441–461.
- [2]. Fitzmaurice GM, Laird NM. Regression models for a bivariate discrete and continuous outcome with clustering. *Journal of the American Statistical Association*. 1995; 90:845–852.
- [3]. Catalano PJ, Ryan LM. Bivariate latent variable models for clustered discrete and continuous outcomes. *Journal of the American Statistical Association*. 1992; 87:651–658.
- [4]. Sammel MD, Ryan LM, Legler JM. Latent variable models for mixed discrete and continuous outcomes. *Journal of the Royal Statistical Society, Series B: Methodological*. 1997; 59:667–678.
- [5]. Sammel M, Lin X, Ryan L. Multivariate linear mixed models for multiple outcomes. *Statistics in Medicine*. 1999; 18:2479–2492. [PubMed: 10474154]
- [6]. Arminger, G.; Küsters, U. Latent trait and latent class models. Plenum Press; New York, U.S.A.: 1988. chap. Latent trait models with indicators of mixed measurement level; p. 51-73.
- [7]. Dunson DB. Bayesian latent variable models for clustered mixed outcomes. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*. 2000; 62(2):355–366.
- [8]. Dunson DB, Chen Z, Harry J. A Bayesian approach for joint modeling of cluster size and subunit-specific outcomes. *Biometrics*. 2003; 59(3):521–530. [PubMed: 14601753]
- [9]. Gueorguieva RV, Agresti A. A correlated probit model for joint modeling of clustered binary and continuous responses. *Journal of the American Statistical Association*. 2001; 96(455):1102–1112.
- [10]. Reilly T. A necessary and sufficient condition for identification of confirmatory factor analysis models of factor complexity one. *Sociological Methods and Research*. 1995; 23(4):421–441.
- [11]. Lin X, Ryan L, Sammel M, Zhang D, Padungtod C, Xu X. A scaled linear mixed model for multiple outcomes. *Biometrics*. 2000; 56(2):593–601. [PubMed: 10877322]
- [12]. Liang K, Zeger S. Longitudinal data analysis using generalized linear models. *Biometrika*. 1986; 73(1):13–22.
- [13]. Prentice RL, Zhao LP. Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. *Biometrics*. 1991; 47(3):825–839. [PubMed: 1742441]
- [14]. Zhao LP, Prentice RL, Self SG. Multivariate mean parameter estimation by using a partly exponential model. *Journal of the Royal Statistical Society. Series B*. 1992; 54(3):805–811.
- [15]. Dickey B, Normand SLT, Hermann RC, Eisen SV, Cortes DE, Cleary PD, Ware N. Guideline recommendations for treatment of schizophrenia: the impact of managed care. *Arch Gen Psychiatry*. 2003; 60(4):340–8. [PubMed: 12695310]
- [16]. Pharmacological Therapy for Macular Degeneration Study Group. Interferon α -iia is ineffective for patients with choroidal neovascularization secondary to age-related macular degeneration: results of a prospective randomized placebo-controlled clinical trial. *Archives of Ophthalmology*. 1997; 115:865–872. [PubMed: 9230826]
- [17]. Zellner A. An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American Statistical Association*. 1962; 57(298):348–368.
- [18]. Rotnitzky A, Holcroft CA, Robins JM. Efficiency comparisons in multivariate multiple regression with missing outcomes. *Journal of Multivariate Analysis*. 1997; 61:102–128.
- [19]. Fitzmaurice GM, Laird NM. Regression models for mixed discrete and continuous responses with potentially missing values. *Biometrics*. 1997; 53:110–122. [PubMed: 9147588]
- [20]. Little, RJ.; Rubin, D. Statistical Analysis with Missing Data. John Wiley and Sons, Inc; Hoboken, New Jersey, U.S.A.: 2002.

Table 1

Mean square errors (MSE) from the simulation study with no effect of the shared covariate ($\beta_{b1} = 0$) on the binary outcome and a small effect on the continuous outcome ($\beta_{c1} = 2$; $1/3$ of a SD): ratio of the MSE of the multivariate models (factorization model(24), latent variable model(25) and GEE(18)) to the univariate model(23). Results obtained from 1000 samples of size 200 for each correlation level and for data generated with a common covariate for both outcomes (simulation I) and generated with a common covariate plus a different covariate for each outcome (simulation II).

	Simulation I ^a				Simulation II ^b			
	Correlation coefficient ^c				Correlation coefficient			
	0	.3	.6	.9	0	.3	.6	.9
$\beta_{b1} = 0$								
Factorization	1.00	1.00	1.00	1.00	1.00	1.00	0.93	0.91
Latent	1.00	1.00	0.97	0.86	1.00	1.00	0.93	0.76
GEE	1.00	1.00	1.00	1.00	1.00	1.01	0.97	0.95
$\beta_{b2} = 1$ (1 SD)								
Factorization					1.00	0.95	0.83	0.65
Latent					1.00	0.97	0.83	0.57
GEE					1.00	0.99	0.85	0.83
$\beta_{c1} = 2$ (1/3 SD)								
Factorization	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.05
Latent	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99
GEE	1.00	1.01	1.00	1.00	1.00	1.01	1.02	1.05
$\beta_{c2} = 1$ (1/6 SD)								
Factorization					1.00	0.98	0.84	0.69
Latent					1.00	0.98	0.82	0.60
GEE					1.00	0.98	0.84	0.69

^aData generating process described in (21)

^bData generating process described in (22)

^cThis coefficient refers to the correlation between the underlying variables. The correlation between the outcomes will depend on the covariates value. For $x_i = 0$, the correlation between the observed outcomes corresponding to (0, .3, .6, .9) are respectively (0, .2, .5, .7)

Table II

Mean square errors (MSE) from the simulation study with small effect of the shared covariate ($\beta_{b1} = 0.2$; 1/5 of a SD) on the binary outcome and a small effect on the continuous outcome ($\beta_{c1} = 2$; 1/3 of a SD): ratio of the MSE of the multivariate models (factorization model(24), latent variable model(25) and GEE(18)) to the univariate model(23). Results obtained from 1000 samples of size 200 for each correlation level and for data generated with a common covariate for both outcomes (simulation I) and generated with a different covariate plus a different covariate for each outcome (simulation II).

	Simulation I ^a				Simulation II ^b			
	Correlation coefficient ^c				Correlation coefficient			
	0	.3	.6	.9	0	.3	.6	.9
$\beta_{b1} = 0.2$ (1/5 SD)								
Factorization	1.00	1.00	1.00	1.00	1.01	1.00	0.98	0.98
Latent	1.00	1.00	1.00	0.89	1.01	1.00	0.97	0.90
GEE	1.06	1.00	1.00	0.97	1.04	1.01	0.98	0.98
$\beta_{b2} = 1$ (1 SD)								
Factorization					1.00	0.96	0.82	0.71
Latent					1.00	0.96	0.82	0.58
GEE					1.00	0.96	0.83	0.72
$\beta_{c1} = 2$ (1/3 SD)								
Factorization	1.00	1.00	1.00	1.00	1.00	1.00	1.01	1.05
Latent	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
GEE	1.00	1.01	1.00	1.00	1.00	1.00	1.01	1.05
$\beta_{c2} = 1$ (1/6 SD)								
Factorization					1.02	0.94	0.82	0.62
Latent					1.00	0.94	0.82	0.51
GEE					1.02	0.94	0.82	0.62

^aData generating process described in (21)

^bData generating process described in (22)

^cThis coefficient refers to the correlation between the underlying variables. The correlation between the outcomes will depend on the covariates value. For $x_i = 0$, the correlation between the observed outcomes corresponding to (0, .3, .6, .9) are respectively (0, .2, .5, .7)

Mean square errors (MSE) from the simulation study with large effect of the shared covariate ($\beta_{b1} = 1$; 1 SD) on the binary outcome and a large effect on the continuous outcome ($\beta_{c1} = 6$; 1 SD): ratio of the MSE of the multivariate models (factorization model(24), latent variable model(25) and GEE(18)) to the univariate model(23). Results obtained from 1000 samples of size 200 for each correlation level and for data generated with a common covariate for both outcomes (simulation I) and generated with a common covariate plus a different covariate for each outcome (simulation II).

Table III

	Simulation I ^a				Simulation II ^b			
	Correlation coefficient ^c				Correlation coefficient			
	0	.3	.6	.9	0	.3	.6	.9
$\beta_{b1} = 1$ (1 SD)								
Factorization	1.00	1.00	1.00	1.00	1.00	1.00	0.96	0.92
Latent	1.00	1.00	1.00	0.95	1.00	1.00	0.94	0.79
GEE	1.00	1.00	1.00	1.00	1.35	1.26	1.25	1.15
$\beta_{b2} = 1$								
Factorization (1 SD)					1.00	0.95	0.87	0.62
Latent					1.00	0.95	0.85	0.48
GEE					1.69	1.45	1.39	0.83
$\beta_{c1} = 6$ (1 SD)								
Factorization	1.00	1.00	1.00	1.00	1.00	1.00	1.01	1.03
Latent	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
GEE	1.00	1.00	1.00	1.00	1.01	1.01	1.01	1.03
$\beta_{c2} = 1$ (1/6 SD)								
Factorization					1.00	0.98	0.81	0.66
Latent					1.00	0.98	0.81	0.55
GEE					1.00	0.96	0.81	0.66

^aData generating process described in (21)

^bData generating process described in (22)

^cThis coefficient refers to the correlation between the underlying variables. The correlation between the outcomes will depend on the covariates value. For $x_i = 0$, the correlation between the observed outcomes corresponding to (0, .3, .6, .9) are respectively (0, .2, .5, .7)

Table IV

Managed care effect on the two outcomes related to quality of care: “patient/clinician relationship” and “prescription of anti-psychotic medication”. Data on 394 patients with schizophrenia.

	Model			
	Univariate	Factorization	Latent	GEE
Binary: Prescription of atypical anti-psychotic				
Intercept (p-value)	0:552 ± 0:091 (<0.001)	0:552 ± 0:091 (<0.001)	0:551 ± 0:091 (<0.001)	0:552 ± 0:092 (<0.001)
Managed care (p-value)	−0.009 ± 0:134 (0.948)	−0.009 ± 0:134 (0.948)	−0.007 ± 0:134 (0.958)	−.009 ± 0:134 (0.948)
Continuous: Patient/clinician relationship				
Intercept (p-value)	3:213 ± 0:045 (<0.001)	3:213 ± 0:045 (<0.001)	3:213 ± 0:045 (<0.001)	3:213 ± 0:045 (<0.001)
Managed care (p-value)	−0.017 ± 0:066 (0.799)	−0.017 ± 0:066 (0.799)	−0.017 ± 0:066 (0.799)	−0.017 ± 0:066 (0.799)
$\hat{\sigma}_c$	0.656	0.655	0.631	0.0656
$\hat{\tau}$	-	0.084	-	-
$\hat{\sigma}_u$	-	-	0.281	-
$\hat{\rho}$	-	-	-	0.058

Table V

Effect of high dose of interferon- α (Z) in visual performance of patients. Visual performance was assessed by loss of at least three lines of vision at 1 year (binary outcome) and visual acuity at 6 months (continuous outcome). Data from 190 patients with macular degeneration.

	Model			
	Univariate	Factorization	Latent	GEE
Binary: Loss of at least three lines of vision				
Intercept (p-value)	-0.31±0.13 (0.015)	-0.31±0.13 (.015)	-0.21±0.12 (.091)	-0.31±0.13 (0.015)
Treatment (p-value)	0.41±0.18 (0.027)	0.41±0.18 (0.027)	0.35±0.18 (0.050)	0.43±0.18 (0.029)
Continuous: Visual acuity				
Intercept (p-value)	5.53±1.26 (<0.001)	5.53±1.26 (<0.001)	5.53±1.26 (<0.001)	5.57±1.34 (<0.001)
Treatment (p-value)	2.83±1.86 (0.130)	2.83±1.87 (0.131)	2.83±1.86 (0.130)	2.86±1.84 (0.126)
$\hat{\sigma}_2$	12.80	10.05	5.65	12.80
$\hat{\tau}$	-	16.14	-	-
$\hat{\sigma}_u$	-	-	2.03	
$\hat{\rho}$	-	-	-	-0.62

Table VI

Restenosis following coronary artery stenting. Target lesion revascularization at 9 months following stent deployment is a binary measure; proportion diameter stenosis is also measured at 9 months (continuous). Data on 105 patients who received a bare-metal stent.

	Model			
	Univariate	Factorization	Latent	GEE
Binary: Target lesion revascularization				
Intercept (p-value)	-1.28±0.18 (<0.001)	-1.28±0.18 (<0.001)	-1.26±0.18 (<0.001)	-1.28±0.18 (<0.001)
Diabetes (p-value)	0.85±0.38 (0.027)	0.85±0.38 (0.028)	0.82±0.37 (0.030)	0.85±0.38 (0.028)
Continuous : Proportion diameter stenosis				
Intercept (p-value)	34.26±4.39 (<0.001)	33.62±3.72 (<0.001)	33.80±3.95 (<0.001)	33.64±3.43 (<0.001)
Diabetes (p-value)	9.56±4.65 (0.042)	9.53±5.26 (0.073)	9.53±4.65 (0.043)	9.56±4.92 (0.055)
Length (p-value)	0.58±0.31 (0.067)	0.63±0.26 (0.016)	0.62±0.28 (0.027)	0.63±0.23 (0.007)
$\hat{\sigma}_c$	16.66	13.70	7.40	16.56
$\hat{\tau}$	-	28.78	-	-
$\hat{\sigma}_u$	-	-	2.02	-
$\hat{\rho}$	-	-	-	0.57