

Mixed-outcome data

Forthcoming in Encyclopedia of Biopharmaceutical Statistics

A. R. de Leon^{1,*} & K. Carrière Chough²

¹*Department of Mathematics & Statistics, University of Calgary*

²*Department of Mathematical & Statistical Sciences, University of Alberta*

Key words: *Copulas; Poisson outcomes; General location model; Logistic regression; Maximum likelihood; Probit regression; Pseudo-likelihood; Random-effects models.*

Introduction

Mixed outcomes are ubiquitous in applications in health and medicine, and joint analysis of such outcomes entails specification of models flexible enough to accommodate them. Such joint models are potentially advantageous in several statistical and practical respects. For example, intrinsically multivariate questions concerning relationships between outcomes and the joint influence of covariates on them may be easily answered by fully exploiting the multivariate nature of the data through joint models. From a statistical standpoint, joint analysis avoids multiple testing and naturally leads to global tests, thus resulting in increased power and better control of Type I error rates. Significant efficiency gains over separate univariate analyses have also been reported, specially in settings where there are missing data (e.g., Ref.^[11]). The following are examples of studies where mixed outcomes are of interest.

Example 1 (*Macular degeneration study*) *Longitudinal data are obtained from a two-armed randomized multi-center clinical trial on visual acuity of patients suffering from macular degeneration. In the trial, a patient's visual acuity was assessed at different time points through his ability to read lines of letters on standardized vision charts displaying line letters of decreasing size that he 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 data consist of a true count endpoint, corresponding to the total number of letters correctly read, and a surrogate binary endpoint, defined as the loss of at least three lines of vision at 1 year compared with their baseline performance. The study aim is to investigate the relationship between the mixed endpoints. See also Molenberghs and Verbeke^[17] and Pinto and Normand^[20].*

Example 2 (*Irwin's toxicity study*) *The data come from a three-day repeated dose-toxicity study on neurofunctional effects of a psychotropic drug on rats. The data consist of several*

*Correspondence to: A. R. de Leon, Department of Mathematics and Statistics, University of Calgary, Calgary, AB, Canada T2N 1N4 (E-mail: adeleon@math.ucalgary.ca)

binary outcomes relating to behaviour, including locomotor activity—characterized by the animal’s abnormal biting, restlessness, writhing—and positional passivity—defined as an animal’s struggle response to sequential handling—and a number of continuous outcomes, such as body temperature, pupil size, etc. The goal of the study is to determine treatment effects as well as the association between certain mixed outcomes. Details are given in Faes et al.^[8].

Example 3 (*Veterans’ Health Administration health performance monitoring study*) The study collected monitoring data on health performance of several geographically defined service networks in the U.S. with the objective of characterizing the quality of care of different service networks. The observed data consist of multilevel continuous and discrete outcomes collected repeatedly within clusters over time, and consist of mixed continuous and binary variables on visits, days between visits, readmissions, days between readmissions, etc., from a sample of veterans. The goal of the study is to characterize trends in quality of individual service networks on the basis of the mixed outcomes and to identify problems in quality for specific networks. Details are found in Daniels and Normand^[3].

In analyzing such mixed-outcome data, emphasis is usually placed on determining the joint distribution of the mixed outcomes, from which are obtained specific aspects of their relationships, such as marginal and conditional distributions, and associations. Most research dealing with statistical problems associated with joint analysis of mixed outcomes has appeared only recently due to difficulties in specifying a mixed-outcome joint distribution and the relative lack of standard models. This is further compounded in longitudinal and cluster data settings in medical and health studies, where associations between mixed outcomes at various levels need to be meaningfully delineated in the analysis.

In this contribution, we provide an up-to-date, comprehensive, and unified overview of available strategies for modelling and analyzing mixed outcomes. We highlight connections between various approaches and research threads in the literature, paying particular attention to their advantages and disadvantages. Several examples provide background material on the assorted challenges that arise in the analysis of mixed outcomes.

Joint mixed-outcome models

A number of joint modelling strategies for mixed outcomes have been studied in the literature. The general approach first specifies a model for the joint distribution of mixed outcomes, then fits the model to data at hand, and finally uses the model to draw inferences. The challenge is that such multivariate distributions are uncommon.

Consider a mixed-data set-up represented by the vectors \mathbf{X}_i and \mathbf{Y}_i denoting discrete (e.g., nominal or ordinal categorical variables, counts) and continuous outcomes, respectively, observed from subject $i = 1, \dots, n$. Joint analysis of \mathbf{X}_i and \mathbf{Y}_i requires either direct or indirect specification of the joint density $f_{\mathbf{X}_i, \mathbf{Y}_i}(\mathbf{x}, \mathbf{y})$. Models for the analysis of continuous outcomes have been well studied; however, only recently have researchers begun tackling the corresponding problems associated with discrete outcomes. Examples of earlier work in this area include the multivariate probit model and the so-called Dale model (see ^[17] for more details). While this has paved the way for advances in mixed-outcome data analysis^[1,10], there is no standard model similar to the multivariate normal distribution for continuous data or the multinomial distribution for discrete variables, resulting in a dearth of models. A taxonomy of currently available modelling approaches for mixed outcomes is shown in Figure 1.

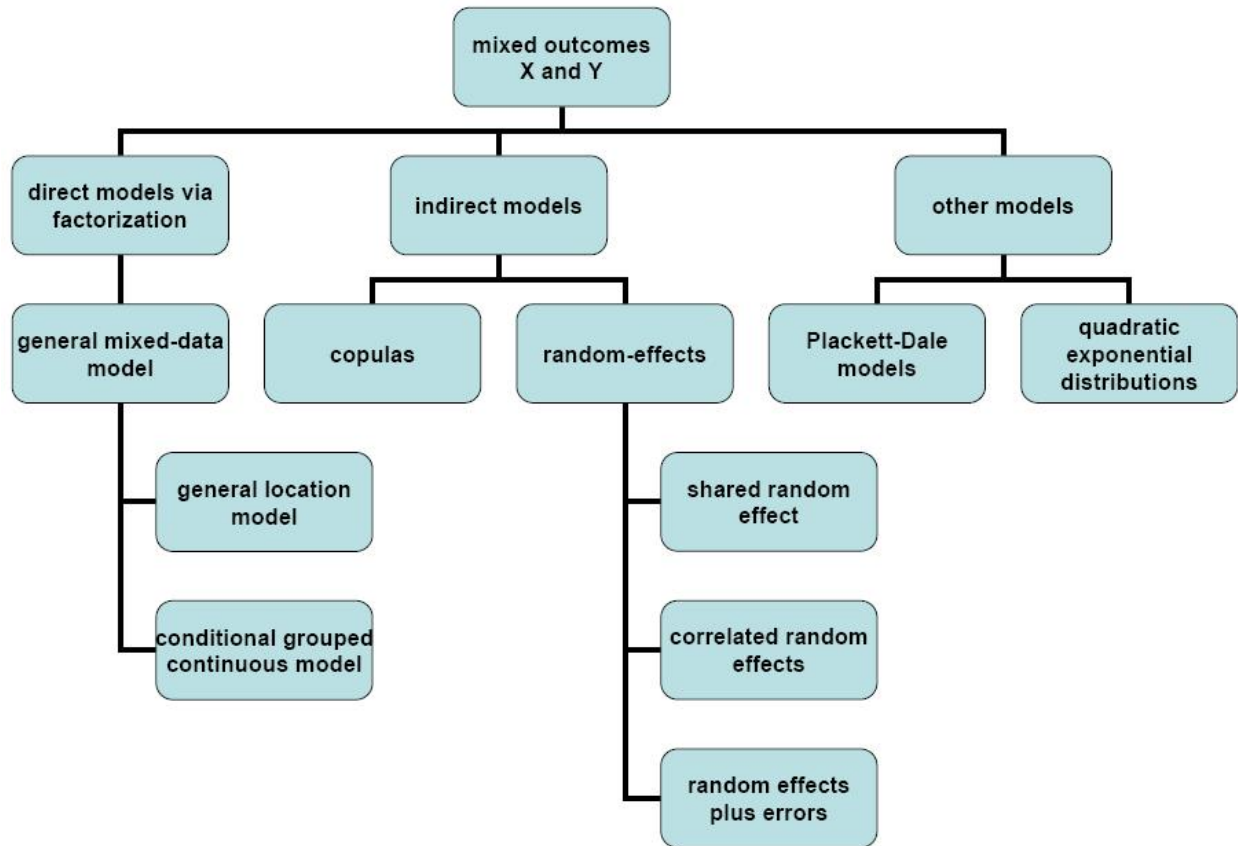


Figure 1: Taxonomy of models for mixed outcomes X (discrete) and Y (continuous).

In the next section, we survey various models for $f_{\mathbf{X}_i, \mathbf{Y}_i}(\mathbf{x}, \mathbf{y})$, many developed for specific applications and mostly based on some factorization of it into marginal and conditional components. A number of issues concerning their application to joint regression modelling of mixed-outcome data are discussed. These issues involve model specification as well as ensuing inference based on such models.

Factorization models

One of the earliest proposals of directly specifying the joint distribution factorizes it into a conditional distribution of one set of outcomes and a marginal distribution of the other set. This suggested two formulations of mixed-outcome joint distributions: (1) a marginal distribution for discrete outcomes and a conditional distribution for continuous outcomes, given discrete outcomes, and (2) a marginal distribution for continuous outcomes and a conditional distribution for discrete outcomes, given continuous outcomes.

Formulation (1) has received much attention in mixed-data literature, the most popular such approach being Olkin and Tate's^[19] so-called general location model (GLOM). The model is a special case of conditional Gaussian distributions used in graphical association models of mixed data, which assumes different multivariate normal distributions for con-

tinuous outcomes given discrete outcomes, and a marginal multinomial distribution for the latter.

Example 4 (*Logistic-normal model*) Consider a binary outcome X_i and a continuous outcome Y_i . Given outcome-specific covariates \mathbf{z}_{x_i} and \mathbf{z}_{y_i} , a marginal regression model for X_i and a conditional regression model for Y_i given X_i are specified as

$$\log\left(\frac{\mu_{x_i}}{1-\mu_{x_i}}\right) = \mathbf{z}_{x_i}^\top \boldsymbol{\beta}_x \quad \text{and} \quad \mu_{y_i|x_i} = \mathbf{z}_{y_i}^\top \boldsymbol{\beta}_y + \gamma(X_i - \mu_{x_i}),$$

where $\mu_{x_i} = P(X_i = 1)$ and $\mu_{y_i|x_i} = E(Y_i|X_i)$. With $Y_i|X_i \sim \text{normal}(\mu_{y_i|x_i}, \sigma^2)$ and $X_i \sim \text{bernoulli}(\mu_{x_i})$, the joint density of X_i and Y_i is then

$$f_{X_i, Y_i}(x, y) = f_{X_i}(x) f_{Y_i|X_i}(y|x) = \frac{1}{\sigma} \phi\left(\frac{y - \mu_{y_i|x_i}}{\sigma}\right) \exp\{\eta_{x_i} x - \log(1 + e^{\eta_{x_i}})\},$$

where $\eta_{x_i} = \log\{\mu_{x_i}/(1 - \mu_{x_i})\}$ and $\phi(\cdot)$ is the standard normal density. The association between X_i and Y_i is induced by γ as follows:

$$\text{corr}(X_i, Y_i) = \frac{\gamma}{\sqrt{\gamma^2 + \frac{\sigma^2}{\mu_{x_i}(1-\mu_{x_i})}}}.$$

Note that X_i and Y_i are uncorrelated whenever $\gamma = 0$.

The logistic-normal model extends GLOM to incorporate marginal regression models for binary and continuous outcomes. A number of refinements and extensions of the logistic-normal model have since been studied by several authors. Fitzmaurice and Laird [9] adapt this model to mixed data with missing responses.

An example of formulation (2) is the so-called conditional grouped continuous model (CGCM), which assumes a logistic or probit conditional distribution for binary given continuous outcomes and a marginal multivariate normal distribution for the latter. In such models, a continuous latent vector \mathbf{Y}_i^* underlying \mathbf{X}_i is supposed to exist and a joint distribution $f_{\mathbf{Y}_i^*, \mathbf{Y}_i}(\mathbf{y}^*, \mathbf{y})$ is assumed for \mathbf{Y}_i^* and \mathbf{Y}_i . A threshold model is then postulated for \mathbf{X}_i and its corresponding latent vector \mathbf{Y}_i^* , so that $f_{\mathbf{X}_i, \mathbf{Y}_i}(\mathbf{x}, \mathbf{y}) = f_{\mathbf{Y}_i}(\mathbf{y}) f_{\mathbf{X}_i|\mathbf{Y}_i}(\mathbf{x}|\mathbf{y})$ is specified through $f_{\mathbf{Y}_i^*, \mathbf{Y}_i}(\mathbf{y}^*, \mathbf{y}) = f_{\mathbf{Y}_i}(\mathbf{y}) f_{\mathbf{Y}_i^*|\mathbf{Y}_i}(\mathbf{y}^*|\mathbf{y})$. While $f_{\mathbf{Y}_i^*, \mathbf{Y}_i}(\mathbf{y}^*, \mathbf{y})$ is completely arbitrary, modelling it as multivariate normal is convenient because of the latter's convenient marginal and conditional distributional properties.

Example 5 (*Probit-normal model*) Consider again binary and continuous outcomes X_i and Y_i in Example 4. To model the joint distribution of X_i and Y_i , let $Y_i \sim \text{normal}(\mu_{y_i}, \sigma^2)$ and assume, underlying X_i , a continuous latent variable Y_i^* such that $X_i = I\{Y_i^* > 0\}$, where Y_i and Y_i^* are jointly normal with correlation ρ . With $\mu_{y_i} = \mathbf{z}_{y_i}^\top \boldsymbol{\beta}_y$ and $\mu_{y_i^*} = \mathbf{z}_{x_i}^\top \boldsymbol{\beta}_{x|y}$, the joint density of X_i and Y_i is

$$f_{X_i, Y_i}(x, y) = \left\{ \Phi\left(\frac{\mathbf{z}_{x_i}^\top \boldsymbol{\beta}_{x|y}}{\sqrt{1-\rho^2}} + \gamma(y - \mu_{y_i})\right) \right\}^x \left\{ 1 - \Phi\left(\frac{\mathbf{z}_{x_i}^\top \boldsymbol{\beta}_{x|y}}{\sqrt{1-\rho^2}} + \gamma(y - \mu_{y_i})\right) \right\}^{1-x} \\ \times \frac{1}{\sigma} \phi\left(\frac{y - \mu_{y_i}}{\sigma}\right),$$

where $\gamma = \rho/(\sigma\sqrt{1-\rho^2})$ and $\Phi(\cdot)$ is the standard normal distribution function. We also get

$$\text{corr}(X_i, Y_i) = \frac{\text{cov}\left(\Phi\left(\frac{\mathbf{z}_{x_i}^\top \boldsymbol{\beta}_{x|y}}{\sqrt{1-\rho^2}} + \gamma(Y_i - \mu_{y_i})\right), Y_i\right)}{\sigma\sqrt{\Phi\left(\frac{\mu_{y_i}^* + \gamma\mu_{y_i}}{\sqrt{1+\gamma^2\sigma^2}}\right)\left\{1 - \Phi\left(\frac{\mu_{y_i}^* + \gamma\mu_{y_i}}{\sqrt{1+\gamma^2\sigma^2}}\right)\right\}}}$$

For identifiability, it is necessary that Y_i^* has unit variance. In addition, $\boldsymbol{\beta}_{x|y}$ must contain an intercept because the cutpoint was arbitrarily taken as 0. Note that X_i and Y_i are uncorrelated whenever $\gamma = 0$.

Adaptations of the probit-normal model to clustered data settings are discussed in Refs.^[1,21].

de Leon and Carrière^[4] describe the general mixed-data model (GMDM) as a hybrid of GLOM and CGCM and provide a unified treatment of these two standard mixed-data models. An attractive feature of GMDM is that it incorporates associations and correlations between different outcomes and accounts for their measurement levels. The model can serve as a platform for extending conventional multivariate methods to the case of mixed data with discrete and continuous outcomes.

Asymmetry in treatment of outcomes

It is common to rely on the ‘arrow of time’ to order outcomes and decide the direction of conditioning. However, many mixed-outcome models use a structural approach to classify them into continuous or discrete. Factorization models induce a hierarchy in outcomes, with the conditioning outcome treated as intermediate variable, and the conditioned outcome as the ultimate response.

Example 6 (*Probit-normal model*) Consider Example 5 again. It is easy to see that

$$\mu_{x_i|y_i} = P(X_i = 1|Y_i = y) = \Phi\left(\frac{\mathbf{z}_{x_i}^\top \boldsymbol{\beta}_{x|y}}{\sqrt{1-\rho^2}} + \gamma(y - \mu_{y_i})\right),$$

so that the marginal model becomes

$$\mu_{x_i} = P(X_i = 1) = \Phi\left(\frac{\mu_{y_i}^* + \gamma\mu_{y_i}}{\sqrt{1+\gamma^2\sigma^2}}\right) = \Phi(\mathbf{z}_{x_i}^\top \boldsymbol{\beta}_x + \mathbf{z}_{y_i}^\top \bar{\boldsymbol{\beta}}_y),$$

where $\boldsymbol{\beta}_x = \boldsymbol{\beta}_{x|y}/\sqrt{1+\gamma^2\sigma^2}$. Note that the marginalized probit model for X_i includes the regression coefficient $\bar{\boldsymbol{\beta}}_y = \gamma\boldsymbol{\beta}_y/\sqrt{1+\gamma^2\sigma^2} = \rho\boldsymbol{\beta}_y/\sigma$.

Observe that the probit-normal model reverses the factorization in the logistic-normal model. Conditioning on continuous outcomes in this case suggests a predictive model where discrete outcomes are the responses of true interest with continuous outcomes serving as ‘explanatory’ variables. This may not be appropriate in many practical applications, where a more symmetrical treatment of outcomes is needed.

Direction of conditioning

Factorization models are not invariant to the direction of conditioning taken and the factorization adopted. While ideally dictated by subject-matter considerations, the choice of the direction of conditioning is mainly for statistical convenience. In toxicological studies, where such models were first developed, the biological mechanism is not well understood, variations of logistic-normal and probit-normal models have thus been studied. The resulting models are not comparable, as parameters have different interpretations depending on the factorization used.

Example 7 (*Logistic-normal vs. probit-normal models*) From the logistic-normal model, $\mu_{y_i} = E(Y_i) = \mathbf{z}_{y_i}^\top \boldsymbol{\beta}_y$, so that regression parameters $\boldsymbol{\beta}_x$ and $\boldsymbol{\beta}_y$ have marginal interpretations. This is not true of the probit-normal model. For example, the interpretation of $\boldsymbol{\beta}_{x|y}$ is conditional on continuous outcome Y_i ; the marginal covariate effect $\boldsymbol{\beta}_x$ is obtained by averaging over Y_i . The logistic-normal model also suggests a normal-mixture marginal distribution for Y_i with

$$\text{var}(Y_i) = \gamma^2 \mu_{x_i} (1 - \mu_{x_i}) + \sigma^2,$$

implying that the variance of Y_i depends on covariate \mathbf{z}_{x_i} . This contrasts with the homogeneity of Y_i in the probit-normal model.

Another drawback of factorization models is that they do not easily extend to the setting of multiple mixed outcomes. No easy expressions can be obtained for associations between outcomes as well. It is also possible for models to yield very different estimates of the associations^[21].

Random-effects models

Indirect approaches to specifying mixed-outcome joint distributions have also been studied. One approach that has found widespread adoption in practice introduces shared or correlated random effects to incorporate correlations between mixed outcomes in the resulting joint model. Daniels and Normand^[3] use this strategy, albeit in a Bayesian framework, to profile health care units. Applications to high-dimensional mixed data analysis are provided by Faes et al.^[8].

The basic idea in this approach is to use random effects, either shared or correlated, to build in correlation between mixed outcomes. The approach does not resort to factorization, and thus yields a symmetrical treatment of mixed outcomes. Its hierarchical structure allows for considerable flexibility in accounting for different measurement levels, delineation of various associations, incorporation of covariate effects, and extension to longitudinal and clustered data settings. Details can be found in Refs.^[15,17].

Shared random effects

While random effects provide a general and versatile route for incorporating in the model various correlations between outcomes and differences in measurement levels, they are not without their shortcomings. McCulloch^[15] illustrates a problem for the case of one binary outcome and one continuous outcome, with respective conditional Bernoulli and normal

distributions, given a shared normal random effect, which induces correlation between outcomes.

Example 8 (*Probit-normal model with shared random effect*) Consider a binary outcome $X_i \sim \text{bernoulli}(\mu_{x_i})$ and a continuous outcome $Y_i \sim \text{normal}(\mu_{y_i}, \sigma^2)$. A random effect shared by X_i and Y_i is introduced as follows:

$$\Phi^{-1}(\mu_{x_i|u_i}) = \mathbf{z}_{x_i}^\top \boldsymbol{\beta}_{x|u} + U_i \quad \text{and} \quad \mu_{y_i|u_i} = \mathbf{z}_{y_i}^\top \boldsymbol{\beta}_{y|u} + \lambda U_i,$$

where $U_i \sim \text{normal}(0, \tau^2)$, with X_i and Y_i conditionally independent given U_i . The coefficient λ adjusts for difference in scales between the outcomes. The joint density is

$$f_{X_i, Y_i}(x, y) = \frac{1}{\sigma\tau} \int_{-\infty}^{\infty} \mu_{x_i|u}^x (1 - \mu_{x_i|u})^{1-x} \phi\left(\frac{y - \mu_{y_i|u}}{\sigma}\right) \phi(u) du.$$

We also get

$$\text{corr}(X_i, Y_i) = \sqrt{\frac{\lambda^2 \tau^2}{\sigma^2 + \lambda^2 \tau^2}} \frac{\text{cov}(\Phi(\mathbf{z}_{x_i}^\top \boldsymbol{\beta}_{x|u} + U_i), \lambda U_i)}{\sqrt{\Phi(\mathbf{z}_{x_i}^\top \boldsymbol{\beta}_x) \{1 - \Phi(\mathbf{z}_{x_i}^\top \boldsymbol{\beta}_x)\}}},$$

where $\boldsymbol{\beta}_x = \boldsymbol{\beta}_{x|u} / \sqrt{1 + \tau^2}$.

With a conditional probit link for binary outcome X_i given random effect U_i , the marginal regression parameter $\boldsymbol{\beta}_x$ is smaller than the conditional regression parameter $\boldsymbol{\beta}_{x|u}$ by a factor $\sqrt{1 + \tau^2}$. In practical applications where interest lies in marginal effects of covariates, this lack of a marginal interpretation may be considered unattractive. In addition, the model constrains the correlation between X_i and Y_i . McCulloch^[15] shows that $\text{corr}(X_i, Y_i) \rightarrow \sqrt{2/\pi} \approx 0.798$, as $\tau^2 \rightarrow \infty$ with σ^2 fixed.

The situation is somewhat better in the Poisson-normal case in that marginal and conditional regression coefficients, with the exception of intercepts, are identical; however, the correlation becomes even more problematic, being integrally tied to overdispersion in the Poisson outcome. Details are discussed in Ref.^[15].

Correlated random effects

A possible remedy to the shortcomings of the shared random effect approach is to allow for separate but correlated random effects. This results in models that more flexibly describe correlations between mixed outcomes. The following example involving count and continuous outcomes with correlated random effects is discussed in detail by McCulloch^[15].

Example 9 (*Poisson-normal model with correlated random effects*) Consider a count outcome $X_i \sim \text{poisson}(\mu_{x_i})$ and a continuous outcome $Y_{it} \sim \text{normal}(\mu_{y_i}, \sigma^2)$. It is assumed that the continuous outcome is repeatedly observed over time $t = 1, \dots, n_i$, to identify subject-to-subject variation. Correlated random effects for X_i and Y_{it} are introduced as follows:

$$\log(\mu_{x_i|u_i}) = \mathbf{z}_{x_i}^\top \boldsymbol{\beta}_{x|u} + U_{x_i} \quad \text{and} \quad \mu_{y_i|u_i} = \mathbf{z}_{y_i}^\top \boldsymbol{\beta}_{y|u} + \lambda U_{y_i},$$

where U_{x_i} and U_{y_i} are jointly normally distributed with $E(U_{x_i}) = E(U_{y_i}) = 0$, $\text{var}(U_{x_i}) = \tau_x^2$, $\text{var}(U_{y_i}) = \tau_y^2$, and $\text{corr}(U_{x_i}, U_{y_i}) = \rho$. It is straightforward to show that

$$\text{corr}(X_i, Y_{it}) = \frac{\rho \tau_x \tau_y \exp(\mathbf{z}_{x_i}^\top \boldsymbol{\beta}_{x|u} + \frac{1}{2} |\rho| \tau_x^2)}{(\sigma^2 + \tau_y^2) \{\mu_{x_i} + \mu_{x_i}^2 (e^{\tau_x^2} - 1)\}},$$

with $\mu_{x_i} = \exp(\mathbf{z}_{x_i}^\top \boldsymbol{\beta}_{x|u} + \tau_x^2/2)$.

In the Poisson-normal model in Example 9, correlated normal random effects allow for negative correlations between outcomes. It is also easier to disentangle overdispersion in the Poisson outcome, as this now depends on the covariance matrix of the random effects^[15]. However, allowing for correlated random effects in high-dimensional problems may not be computationally feasible^[8] in practice.

Random effects plus correlated errors

Still another alternative is to add correlated error terms to models for the outcomes. To see how this can be done, assume \mathbf{X}_i and \mathbf{Y}_i to be conditionally independent, given an unobserved random effect \mathbf{U}_i . Generalized linear mixed models for \mathbf{X}_i and \mathbf{Y}_i are then specified, either via an exponential family or a marginal moments specification. Random effect \mathbf{U}_i is subject-specific and accounts for correlation patterns of unobserved heterogeneity; to account for residual correlations, marginal models are assumed for the outcomes, in addition to random effect \mathbf{U}_i . Specifically, we have

$$\begin{pmatrix} \mathbf{X}_i \\ \mathbf{Y}_i \end{pmatrix} = \begin{pmatrix} \mathbf{h}_{x_i}(\mathbf{U}_i, \mathbf{z}_{x_i}) \\ \mathbf{h}_{y_i}(\mathbf{U}_i, \mathbf{z}_{y_i}) \end{pmatrix} + \begin{pmatrix} \boldsymbol{\varepsilon}_{x_i} \\ \boldsymbol{\varepsilon}_{y_i} \end{pmatrix} = \mathbf{h}(\mathbf{U}_i, \mathbf{z}_i) + \boldsymbol{\varepsilon}_i,$$

where the inverse link functions in $\mathbf{h}(\cdot, \cdot)$ depend on outcome types, and \mathbf{z}_i is a vector of covariates. For discrete outcomes such as binary outcomes and counts, data approximation based on linearization may be adopted. Note that $\mathbf{h}(\cdot, \cdot)$ allows for a non-additive specification of fixed and random effects in the link function; it is also possible to incorporate latent variables for discrete outcomes. Correlations between \mathbf{X}_i and \mathbf{Y}_i are induced by covariance matrices of \mathbf{U}_i and $\boldsymbol{\varepsilon}_i$.

Example 10 (*Repeated measures logistic-normal model with random effects and errors*) Let X_{it} and Y_{it} be binary and continuous outcomes for subject $i = 1, \dots, n$, at time $t = 1, \dots, n_i$. The model is given by

$$\begin{pmatrix} X_{it} \\ Y_{it} \end{pmatrix} = \begin{pmatrix} \frac{e^{\mathbf{z}_{x_{it}}^\top \boldsymbol{\beta}_{x|u} + U_{x_i}}}{1 + e^{\mathbf{z}_{x_{it}}^\top \boldsymbol{\beta}_{x|u} + U_{x_i}}} \\ \mathbf{z}_{y_{it}}^\top \boldsymbol{\beta}_{y|u} + U_{y_i} \end{pmatrix} + \begin{pmatrix} \varepsilon_{x_{it}} \\ \varepsilon_{y_{it}} \end{pmatrix},$$

where U_{x_i} and U_{y_i} are as defined in Example 9, with independent errors $\varepsilon_{x_{it}}$ and $\varepsilon_{y_{it}}$. Faes et al^[8] derive a first-order approximation for the correlation between X_{it} and Y_{it} as

$$\text{corr}(X_{it}, Y_{it}) \approx \frac{\rho \tau_x \tau_y v_{it}}{\sqrt{v_{it}(1 + v_{it} \tau_x^2)(\tau_y^2 + \sigma^2)}},$$

with $\text{var}(\varepsilon_{xit}) \approx v_{it} = e^{\mathbf{z}_{xit}^\top \boldsymbol{\beta}_{x|u}} / (1 + e^{\mathbf{z}_{xit}^\top \boldsymbol{\beta}_{x|u}}) \left\{ 1 - e^{\mathbf{z}_{xit}^\top \boldsymbol{\beta}_{x|u}} / (1 + e^{\mathbf{z}_{xit}^\top \boldsymbol{\beta}_{x|u}}) \right\}$ and $\text{var}(\varepsilon_{yit}) = \sigma^2$.

The inclusion of random effects in the above model accounts for correlations between repeated measurements. In the case of independent random effects, the outcomes are approximately marginally uncorrelated. Incarnations of the above model, with varying levels of generality, are given by Faes et al.^[8], Gueorguieva and Sanacora^[11], and Gueorguieva and Agresti^[12]. Bayesian versions of the model have been studied by Daniels and Normand^[3] and Dunson^[6], among others. This approach is quite attractive; however, for Poisson outcomes, for example, it leads to the same problem of overdispersion being built into the model, as what happens in the shared random effect approach^[15].

Other models

A number of other alternative models have been proposed and studied by various authors. A Plackett-Dale distribution has been used by Molenberghs et al.^[18] as an alternative to multivariate normal distributions for latent variables. The quadratic exponential model, which has a general form that can accommodate mixed outcomes, has been adopted by Sammel et al.^[23] in clustered settings. The model considers conditional log-odds ratios as measures of associations and requires a normalizing constant, the computational demands of which can be prohibitive in certain cases. Regan and Catalano^[22] introduced mixed-probit models for binary and continuous outcomes that, unlike factorization-based models, treat outcomes in a more symmetrical fashion and have marginal interpretations for their regression parameters. However, these models were developed for specific applications, and thus, may lack generality and flexibility to be applied in other mixed-outcome data settings.

Finally, several authors have adopted copula functions to indirectly specify mixed-outcome joint distributions. This is only a recent phenomenon in modelling mixed outcomes, and unresolved issues, both methodological and practical, abound, specially as they apply to discrete data.

Example 11 (*Robit-normal model*) To model the joint distribution of binary outcome X_i and continuous outcome Y_i , we let $Y_i \sim N(\mu_{y_i}, \sigma^2)$ and assume, underlying X_i , a continuous latent variable $Y_i^* \sim t_1(\mu_{y_i^*}, 1, \nu)$, i.e., the univariate (generalized) t -distribution with mean $\mu_{y_i^*}$, unit scale, and degrees of freedom ν , such that $X_i = I\{Y_i^* > 0\}$.

With $\mu_{y_i} = \mathbf{z}_{y_i}^\top \boldsymbol{\beta}_y$ and $\mu_{y_i^*} = \mathbf{z}_{x_i}^\top \boldsymbol{\beta}_x$, and using the Gaussian copula, de Leon and Wu^[5] derive the joint density of X_i and Y_i as

$$f_{X_i, Y_i}(x, y) = \left\{ \Phi \left(-\frac{\Phi^{-1}(w_i) - \rho \left(\frac{y - \mu_{y_i}}{\sigma} \right)}{\sqrt{1 - \rho^2}} \right) \right\}^x \left\{ 1 - \Phi \left(-\frac{\Phi^{-1}(w_i) - \rho \left(\frac{y - \mu_{y_i}}{\sigma} \right)}{\sqrt{1 - \rho^2}} \right) \right\}^{1-x} \times \frac{1}{\sigma} \phi \left(\frac{y - \mu_{y_i}}{\sigma} \right),$$

where $w_i = P(X_i = 1) = P(Y_i^* > 0)$ and ρ is the Pearson correlation between so-called normal scores $\Phi^{-1}\{P(Y_i^* \leq 0)\}$ and $\Phi^{-1}\{P(Y_i \leq 0)\}$, a ‘proxy’ for the polyserial correlation

between X_i and Y_i .

Liu^[14] introduces the term robit regression based on a t -latent distribution as a robust alternative to and extension of logit and probit regression models. Liu^[14] shows that robit models approximate both logit (with $\nu \approx 7$) and probit (with large ν) regressions, and thus provide a general approach to binary regression modelling.

Copula functions, which are common in actuarial and financial applications, have proved useful in practice when the joint distribution of interest is either not available or difficult to specify but marginal distributions can be specified with confidence, like in mixed-outcome data settings. The approach entails specifying marginal distributions $f_{\mathbf{X}_i}(\mathbf{x})$ and $f_{\mathbf{Y}_i}(\mathbf{y})$, and combining them to form a joint distribution via a suitable copula function. Note that, similar to random effects approaches, copula-based models treat mixed outcomes symmetrically; however, unlike factorization and random effects models, their regression parameters are marginally meaningful. Recent applications of copulas to mixed outcomes are discussed in Song^[24,25] and Hoff^[13].

Model estimation

A variety of approaches have been considered in the literature for model estimation. Given a full specification of the model, a likelihood-based approach may be considered. However, evaluation and direct maximization of the full likelihood may be computationally prohibitive in practice. In addition, a full likelihood approach raises questions about the robustness of the likelihood specification quite apart from any computational difficulties. To circumvent these, the full likelihood may be replaced by a more computationally tractable function, a so-called pseudo-likelihood function. One such function is obtained by compositing pairwise likelihoods over the data. Faes et al.^[7] adopts this strategy in a high-dimensional mixed-outcome analysis via mixed models implemented using standard software, such as SAS procedures `NLMIXED` and `glimmix`. Working with such a modification of the likelihood function generally yields consistent estimates of model parameters, including correlations under a range of possible models for higher-order dependency as captured by the full joint distribution. Computational and statistical performance of these methods has been shown to range from acceptably good to excellent.

Several alternative estimation strategies based on the EM and Monte Carlo methods have also been proposed in the literature^[23]. However, as with conventional full likelihood estimation, these methods have the disadvantage of excessive computational requirements. Non-likelihood based approaches such as generalized estimating equations (GEE)^[20] may also be adapted to this context. These may prove more useful and computationally more feasible than likelihood-based methods in practice.

Conclusion

This contribution provides an up-to-date survey of joint analysis of mixed outcomes. Both direct and indirect approaches to joint model specification are discussed along with their advantages and disadvantages. Because correlations between the mixed outcomes are usually of practical interest, particular attention is given to differences in how such correlations are incorporated in the models. Technical and practical issues relating to model estimation are also briefly discussed, with focus on pseudo-likelihood methods.

The use of copulas in modelling mixed outcomes is a recent phenomenon. Because of this, unresolved issues, both methodological and practical, abound. One is the paucity, if not total lack, of copulas that can be used for mixed outcomes; of theoretical, but of lesser practical, concern is the non-uniqueness of copulas for discrete outcomes. Another involves the interpretation of the dependence parameter of copula functions, a serious problem for discrete distributions. de Leon and Wu^[5] develop copula-based regression models for mixed outcomes by adopting a latent-variable formulation of the discrete outcomes and using Gaussian copulas to “glue” mixed-outcome marginal regression models. Work on generalizations of their approach to high-dimensional settings with possibility of incorporating random effects would be worthwhile. The challenge here lies in defining models that allow for different levels of association among outcomes, as in longitudinal studies.

Incomplete data are ubiquitous in studies involving mixed outcomes^[9]. This is an area where little research has been done. Adaptations of the models discussed in this contribution to handle missing data would be of great importance in practice.

Further reading

The last decade has seen many remarkable advances in statistical methodology for analyzing mixed data and we give here only a few selected references. Regan and Catalano^[22] provide an excellent early survey of mixed-outcome analysis, with particular focus on toxicological applications. Molenberghs and Verbeke^[17] provide a good introduction to generalized linear mixed models for mixed outcomes and include a number of practical examples. Faes et al.^[7] update Molenberghs and Verbeke^[17] and include recent advances in analysis of longitudinal mixed outcomes in high-dimensions. Bayesian approaches including methods for handling incomplete mixed-outcome data, are described in Daniels and Hogan^[2]. McCulloch et al.^[16] devote a chapter on random-effects models for mixed outcomes. The use of copulas for modelling mixed data is discussed in Song^[24].

Acknowledgement

This work was supported by grants from the Alberta Heritage Foundation for Medical Research and the Natural Sciences and Engineering Research Council of Canada.

References

- [1] Catalano, P. J. and Ryan, L. M. (1992). Bivariate latent variable models for clustered discrete and continuous outcomes. *Journal of the American Statistical Association*, **87**, 651-658.
- [2] Daniels, M. J. and Hogan, J. W. (2008). *Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis*, Chapman & Hall/CRC.
- [3] Daniels, M. J. and Normand, S.-L. T. (2006). Longitudinal profiling of health care units based on continuous and discrete patient outcomes. *Biostatistics*, **7**, 1-15.
- [4] de Leon, A. R. and Carrière, K. C. (2007). General mixed-data model: extension of general location and grouped continuous models. *Canadian Journal of Statistics*, **35**, 533-548.

- [5] de Leon, A. R. and Wu, B. (2009). Copula-based regression models for a bivariate mixed discrete and continuous outcome. Submitted.
- [6] Dunson, D. B. (2000). Bayesian latent variable models for clustered mixed outcomes. *Journal of the Royal Statistical Society-B*, **62**, 355-366.
- [7] Faes, C., Geys, H., and Catalano, P. J. (2009). Joint models for continuous and discrete longitudinal data. In *Longitudinal Data Analysis*, Fitzmaurice, G., Davidian, Verbeke, G., and Molenberghs, G. (eds.), Chapman & Hall/CRC.
- [8] Faes, C., Aerts, M., Molenberghs, G., Geys, H., Teuns, G., and Bijmens, L. (2008). A high-dimensional joint model for longitudinal outcomes of different nature. *Statistics in Medicine*, **27**, 4408-4427.
- [9] Fitzmaurice, G. M. and Laird, N. M. (1997). Regression models for mixed discrete and continuous responses with potentially missing values. *Biometrics*, **53**, 110-122.
- [10] 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.
- [11] Gueorguieva, R. V. and Sanacora, G. (2006). Joint analysis of repeatedly observed continuous and ordinal measures of disease severity. *Statistics in Medicine*, **25**, 1307-1322.
- [12] Gueorguieva, R. V. and Agresti, A. (2001). A correlated probit model for joint modeling of clustered binary and continuous responses. *Journal of the American Statistical Association*, **96**, 1102-1112.
- [13] Hoff, P. D. (2007). Extending the rank likelihood for semiparametric copula estimation. *Annals of Applied Statistics*, **1**, 265-283.
- [14] Liu, C. (2004). Robit regression: a simple robust alternative to logistic and probit regression. In *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, Gelman, A. and Meng, X.-L. (eds), Wiley.
- [15] McCulloch, C. (2007). Joint modeling of mixed outcome types using latent variables. *Statistical Methods in Medical Research*, **17**, 1-21.
- [16] McCulloch, C. E., Searle, S. R., and Neuhaus, J. M. (2008). *Generalized, Linear, and Mixed Models*, 2nd edition, Wiley.
- [17] Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*, Springer.
- [18] Molenberghs, G., Geys, H., and Buyse, M. (2001). Evaluation of surrogate endpoints in randomized experiments with mixed discrete and continuous outcomes. *Statistics in Medicine*, **20**, 3023-3038.
- [19] Olkin, I. and Tate, R. F. (1961). Multivariate correlation models with mixed discrete and continuous variables. *Annals of Mathematical Statistics*, **32**, 448-465 (correction in **36**, 343-344).

- [20] Pinto, A. T. and Normand, S.-L. T. (2009). Correlated bivariate continuous and binary outcomes: Issues and applications. *Statistics in Medicine*, **28**, 1753-1773.
- [21] Regan, M. M. and Catalano, P. J. (2002). Combined continuous and discrete outcomes. In *Topics in Modelling of Clustered Data*, Aerts, M., Geys, H., Molenberghs, G., and Ryan, L. (eds.), Chapman & Hall.
- [22] Regan, M. M. and Catalano, P. J. (1999). Likelihood models for clustered binary and continuous outcomes: Application to developmental toxicology. *Biometrics*, **55**, 760-768.
- [23] 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-B*, **59**, 667-678.
- [24] Song, P. X.-K. (2007). *Correlated Data Analysis: Modeling, Analytics, and Applications*, Springer.
- [25] Song, P. X.-K., Li, M., and Yuan, Y. (2009). Joint regression analysis of correlated data using Gaussian copulas. *Biometrics*, **65**, 60-68.