Open access • Journal Article • DOI:10.1002/BIMJ.201400049

# Statistical evaluation of surrogate endpoints with examples from cancer clinical trials



Marc Buyse, Geert Molenberghs, Geert Molenberghs, Xavier Paoletti ...+5 more authors

**Institutions:** University of Hasselt, Katholieke Universiteit Leuven, Curie Institute, University of Tokyo ...+1 more institutions

**Published on:** 01 Jan 2016 - Biometrical Journal (Biom J)

**Topics:** Surrogate endpoint and Clinical endpoint

# CATEGORICAL DATA, MARGINAL MODELS FOR

**Abstract.** We briefly review two building blocks (generalized linear models and linear mixed models) for models for repeated categorical data. Three families of models for repeated categorical data are introduces: marginal models, conditional models, and random-effects models. A number of marginal models are presented for binary and for ordinal data in turn. Major differences between marginal models and conditional models (such as loglinear models) are discussed.

**Keywords and Phrases.** Binary data; conditional model; generalized linear mixed model; linear mixed model; logistic regression; marginal model; odds ratio; ordinal data.

## UNIVARIATE DATA

For the analysis of binary response variables, one of the most commonly used tools is logistic regression[2]. There are at least three obvious reasons for this. First, it is considered an extension of linear regression. Second, it fits within the theory of generalized linear models. Third, especially in a biometrical context, the interpretation of its parameters in terms of odds ratios is considered convenient. When the latter is less of a concern, such as in econometric applications, one frequently encounters probit regression.

Consider a response variable $Y_i$, measured on subjects $i = 1, \ldots, N$, together with covariates $\boldsymbol{x}_i$. A generalized linear model minimally specifies the mean $E(Y_i) = \mu_i$ and links it to a linear predictor in the covariates $\eta(\mu_i) = \boldsymbol{x}_i^T \boldsymbol{\beta}$, where $\eta(.)$ is the so-called

link function. Further, the variance of $Y_i$ is then link to the mean model by means of the mean-variance link $\text{Var}(Y_i) = \phi v(\mu_i)$, where $v(.)$ is a known variance function and $\phi$ is a scale or overdispersion parameter. Such a specification is sufficient to implement moment-based estimation methods, such as iteratively reweighted least squares or quasi likelihood[21]. In case full likelihood is envisaged, the above framework can be seen to be derived from the general exponential family definition

$$f(y|\theta_i, \phi) = \exp\left\{\phi^{-1}[y\theta_i - \psi(\theta_i)] + c(y, \phi)\right\} \tag{1}$$

with $\theta_i$ the natural parameter and $\psi(.)$ a function satisfying $\mu_i = \psi'(\theta_i)$ and $v(\mu_i) = \psi''(\theta_i)$. Hence, the previous results are recovered but extended. From (1) it immediately follows that the corresponding log-likelihood is linear in the statistics $\theta_i$, simplifying the form of the score equations,

$$S(\boldsymbol{\beta}) \;=\; \sum_i \frac{\partial \mu_i}{\partial \boldsymbol{\beta}}\, v_i^{-1}\, (y_i - \mu_i) \;=\; 0,$$

log-likelihood maximization and corresponding statistical inference.

For example, in the case of a binary outcome $Y_i$, the model can be written as

$$f(y_i|\theta_i, \phi) = \mu_i^{y_i}(1-\mu_i)^{1-y_i} = \exp\left\{y_i \ln\left(\frac{\mu_i}{1-\mu_i}\right) + \ln(1-\mu_i)\right\}$$

and hence the Bernoulli model and, by extension, logistic regression, fits within this framework. In particular,

$$\theta_i = \text{logit}(\mu_i) = \mu_i/(1-\mu_i) = \text{logit}[P(Y_i = 1|\boldsymbol{x}_i)], \tag{2}$$

$\mu = e^\theta/(1+e^\theta)$ and $v(\mu) = \mu(1-\mu)$.

In case one opts for a probit link, the logit in (2) is replaced by the inverse of the standard normal distribution $\Phi^{-1}$, i.e., the probit function. This model cannot be put within the exponential family context. Hence, the choice for logistic regression is often based on the mathematical convenience entailed by the exponential family

framework. Now, it has been repeatedly shown[2] that the logit and probit link functions behave very similarly, in the sense that for probabilities other than extreme ones (say, outside of the interval $[0.2; 0.8]$) both forms of binary regression provide approximately the same parameter estimates, up to a scaling factor equal to $\pi/\sqrt{3}$, the ratio of the standard deviations of a logistic and a standard normal variable.

Extensions for ordinal data include proportional odds logistic regression, baseline-category logit models, and continuation-ratio models[2].

# MODELS FOR REPEATED MEASURES

The linear mixed-effects model[15][26] is a commonly used tool for, among others, variance component models and for longitudinal data [CROSS REF].

Let $\boldsymbol{Y_i}$ denote the $n_i$-dimensional vector of measurements available for subject $i = 1, \ldots, N$. A general linear mixed model then assumes that $\boldsymbol{Y_i}$ satisfies

$$\boldsymbol{Y_i} = X_i\boldsymbol{\beta} + Z_i\boldsymbol{b_i} + \boldsymbol{\varepsilon_i}, \tag{3}$$

in which $\boldsymbol{\beta}$ is a vector of population-average regression coefficients called fixed effects, and where $\boldsymbol{b_i}$ is a vector of subject-specific regression coefficients. The $\boldsymbol{b_i}$ describe how the evolution of the $i$th subject deviates from the average evolution in the population. The matrices $X_i$ and $Z_i$ are $(n_i \times p)$ and $(n_i \times q)$ matrices of known covariates. The random effects $\boldsymbol{b_i}$ and residual components $\boldsymbol{\varepsilon_i}$ are assumed to be independent with distributions $N(\mathbf{0}, D)$, and $N(\mathbf{0}, \Sigma_i)$, respectively. Inference for linear mixed models is usually based on maximum likelihood or restricted maximum likelihood estimation under the marginal model for $\boldsymbol{Y_i}$, i.e., the multivariate normal model with mean $X_i\boldsymbol{\beta}$, and covariance $V_i = Z_iDZ_i' + \Sigma_i$[15][26]. Thus, we can adopt two *different* views on the linear mixed model. The fully *hierarchical* model is specified by

$$\begin{aligned} \boldsymbol{Y_i}|\boldsymbol{b}_i &\sim N_{n_i}(X_i\boldsymbol{\beta} + Z_i\boldsymbol{b}_i, \Sigma_i), \\ \boldsymbol{b}_i &\sim N(0, D), \end{aligned} \tag{4}$$

while the marginal model is given by

$$\boldsymbol{Y_i} \sim N_{n_i}(X_i\boldsymbol{\beta}, Z_iDZ_i' + \Sigma_i). \tag{5}$$

Even though they are often treated as equivalent, there are important differences between the hierarchical and marginal views on the model. Obviously, (4) requires the covariance matrices $\Sigma_i$ and $D$ to be positive definite, while in (5) it is sufficient for the resulting matrix $V_i$ to be positive definite. Different hierarchical models can produce the same marginal model. Some marginal models are not implied by hierarchical models.

When outcomes are of the categorical type, there is no such easy transition between marginal and random-effects model. More generally, Diggle, Heagerty, Liang and Zeger[7] and Aerts, Geys, Molenberghs, and Ryan[1] distinguish between three such families. Still focusing on continuous outcomes, a marginal model is characterized by the specification of a marginal mean function

$$E(Y_{ij}|\boldsymbol{x}_{ij}) = \boldsymbol{x}_{ij}'\boldsymbol{\beta}, \tag{6}$$

whereas in a random-effects model we focus on the expectation, conditional upon the random-effects vector:

$$E(Y_{ij}|\boldsymbol{b}_i, \boldsymbol{x}_{ij}) = \boldsymbol{x}_{ij}'\boldsymbol{\beta} + \boldsymbol{z}_{ij}'\boldsymbol{b}_i. \tag{7}$$

Finally, a third family of models conditions a particular outcome on the other responses or a subset thereof. In particular, a simple first-order stationary transition model focuses on expectations of the form

$$E(Y_{ij}|Y_{i,j-1}, \ldots, Y_{i1}, \boldsymbol{x}_{ij}) = \boldsymbol{x}_{ij}'\boldsymbol{\beta} + \alpha Y_{i,j-1}. \tag{8}$$

As we have seen before, random-effects models imply a simple marginal model in the linear mixed model case. This is due to the elegant properties of the multivariate

normal distribution. In particular, the expectation (6) follows from (7) by either (a) marginalizing over the random effects or by (b) by conditioning upon the random-effects vector $\boldsymbol{b_i} = \mathbf{0}$. Hence, the fixed-effects parameters $\boldsymbol{\beta}$ have both a marginal as well as a hierarchical model interpretation. Finally, when a conditional model is expressed in terms of residuals rather than outcomes directly, it also leads to particular forms of the general linear mixed effects model.

Such a close connection between the model families does not exist when outcomes are of a non-normal type, such as binary, categorical, or discrete. Choosing a model family ought to be done in terms of the scientific question to be answered. For example, opting for a marginal model renders answering conditional or subject-specific questions difficult if not impossible.

# MARGINAL MODELS

In marginal models, the parameters characterize the marginal probabilities of a subset of the outcomes, without conditioning on the other outcomes. Advantages and disadvantages of conditional and marginal modeling have been discussed in Diggle, Heagerty, Liang and Zeger[7], and Fahrmeir and Tutz[9]. The specific context of clustered binary data has received treatment in Aerts, Geys, Molenberghs, and Ryan[1]. Apart from full likelihood approaches, non-likelihood approaches, such as generalized estimating equations[18] or pseudo-likelihood[17][11] have been considered.

Bahadur[4] proposed a marginal model, accounting for the association via marginal correlations. Ekholm[8] proposed a so-called success probabilities approach. George and Bowman[10] proposed a model for the particular case of exchangeable binary data. Ashford and Sowden[3] considered the multivariate probit model, for repeated ordinal data, thereby extending univariate probit regression. Molenberghs and Lesaffre[22] and Lang and Agresti[14] have proposed models which parameter-

ize the association in terms of marginal odds ratios. Dale[5] defined the bivariate global odds ratio model, based on a bivariate Plackett distribution[24]. Molenberghs and Lesaffre[22][23] extended this model to multivariate ordinal outcomes. They generalize the bivariate Plackett distribution in order to establish the multivariate cell probabilities. Their 1994 method involves solving polynomials of high degree and computing the derivatives thereof, while in 1999 generalized linear models theory is exploited, together with the use of an adaption of the iterative proportional fitting algorithm. Lang and Agresti[14] exploit the equivalence between direct modeling and imposing restrictions on the multinomial probabilities, using undetermined Lagrange multipliers. Alternatively, the cell probabilities can be fitted using a Newton iteration scheme, as suggested by Glonek and McCullagh[12]. We will consider some of these models in turn.

## Some Marginal Models for Binary Data

Let the binary response $Y_{ij}$ indicate outcome $j$ for individual $i$. Let

$$\varepsilon_{ij} = \frac{Y_{ij} - \mu_{ij}}{\sqrt{\mu_{ij}(1-\mu_{ij})}} \quad \text{and} \quad e_{ij} = \frac{y_{ij} - \mu_{ij}}{\sqrt{\mu_{ij}(1-\mu_{ij})}},$$

where $y_{ij}$ is an actual value of the binary response variable $Y_{ij}$. Further, let $\rho_{ijk} = E(\varepsilon_{ij}\varepsilon_{ik})$, $\rho_{ijk\ell} = E(\varepsilon_{ij}\varepsilon_{ik}\varepsilon_{i\ell})$, $\ldots$, $\rho_{i12\ldots n_i} = E(\varepsilon_{i1}\varepsilon_{i2}\ldots\varepsilon_{in_i})$. The parameters $\rho_{ijk}$ are classical Pearson type correlation coefficients. The general Bahadur model can be represented by the expression $f(\boldsymbol{y}_i) = f_1(\boldsymbol{y}_i)c(\boldsymbol{y}_i)$, where

$$f_1(\boldsymbol{y}_i) = \prod_{j=1}^{n_i} \pi_{ij}^{y_{ij}}(1-\pi_{ij})^{1-y_{ij}}$$

and

$$c(\boldsymbol{y}_i) = 1 + \sum_{j<k} \rho_{ijk}e_{ij}e_{ik} + \sum_{j<k<\ell} \rho_{ijk\ell}e_{ij}e_{ik}e_{i\ell} + \ldots + \rho_{i12\ldots n_i}e_{i1}e_{i2}\ldots e_{in_i}.$$

Thus, the probability mass function is the product of the independence model $f_1(\boldsymbol{y}_i)$ (combining $n_i$ logistic regressions) and the correction factor $c(\boldsymbol{y}_i)$. The factor $c(\boldsymbol{y}_i)$ can be viewed as a model for overdispersion.

Clearly, the Bahadur model has a very tractable form, a clear advantage over other, more implicitly defined, models. However, a practical drawback is the fact that the correlation between two responses is highly constrained when the higher order correlations are removed. Such a decision is often made to keep the computations as well as the modeling exercise within reasonable limits. Even when higher order parameters are included, the parameter space of marginal parameters and correlations is known to be of a very peculiar shape. Bahadur[4] discusses restrictions on the correlation parameters. When there are no higher order association parameters, and the outcomes are exchangeable, it can be deduced that the lower bound for $\rho_{i(2)}$ approaches zero as the cluster size increases. However, it is important to notice that also the upper bound for this correlation parameter is constrained. Indeed, even though it is one for clusters of size two, the upper bound varies between $1/(n_i - 1)$ and $2/(n_i - 1)$ for larger clusters. Taking a cluster of size 12, the upper bound is in the range $(0.09; 0.18)$. Kupper and Haseman[13] present numerical values for the constraints on $\rho_{i(2)}$ for choices of the marginal probability and $n_i$. Restrictions for a specific version where a third order association parameter is included as well are studied by Prentice[25]. Declerck, Aerts, and Molenberghs[6] give a thorough treatment of the restrictions on the parameter space of the Bahadur model. In conclusion, this model is very appealing at first sight, since it combines logistic regression for the univariate marginal distributions with at first sight very interpretable correlation coefficients. However, our intuition about correlation coefficients largely comes from the normal distribution, where there is a total separation between the mean parameters and the dependence parameters. The only constraint on the correlation coefficients

in the multivariate normal is exactly the same as in a purely moment-based setting, i.e., they should produce a positive definite correlation matrix. In this case, they are heavily constrained, not only by themselves but also by the marginal parameters. This renders not only computations but also interpretation an extremely different task. Thus, while the correlation is undoubtedly a meaningful parameter in the case of normally distributed outcomes, it can be highly questionable in the context of binary outcomes.

Ekholm[8] extended the idea of logistic regressions from the univariate marginal distributions to pairwise and higher-order probabilities. Defining $\mu_{ijk} = P[Y_{ij} = 1, Y_{ik} = 1|\boldsymbol{x}_i]$ to be the pairwise success probability, he considered

$$\eta_{ijk} = \text{logit}(\mu_{ijk}) = \ln(\mu_{ijk}) - \ln(1 - \mu_{ijk}),$$

with similar definition for higher orders. While such an approach is appealing at first sight since there is an apparent symmetry between univariate and higher-order logits, this is only seemingly so. For example, while both $P(Y_{ij} = 1|\boldsymbol{x}_i)$ and $P(Y_{ij} = 0|\boldsymbol{x}_i)$ are linear in the covariates on the logit scale, this is not true for neither $P[Y_{ij} = 0, Y_{ik} = 1|\boldsymbol{x}_i]$, $P[Y_{ij} = 1, Y_{ik} = 0|\boldsymbol{x}_i]$, nor $P[Y_{ij} = 0, Y_{ik} = 0|\boldsymbol{x}_i]$, in spite of $P[Y_{ij} = 1, Y_{ik} = 1|\boldsymbol{x}_i]$ being modeled linearly on the logit scale. Moreover, the range of $\mu_{ijk}$ is restricted by the values for the univariate probabilities, the so-called Fréchet bounds:

$$\max(0, \mu_{ij} + \mu_{ik} - 1) \leq \mu_{ijk} \leq \min(\mu_{ij}, \mu_{ik}),$$

implying complicated restrictions on the parameters entering the linear predictor $\eta_{ijk}$ as coefficients to the covariates. Even when valid probabilities are obtained, interpretation of such coefficients is problematic. This is therefore another instance of a model that is appealing at first sight but engenders a lot of practical and interpretational difficulties. And therefore, this model does not provide a meaningful parameterization.

George and Bowman[10] proposed a model for the analysis of exchangeable binary data. The probability mass function for the number of successes $Z_i = \sum_{j=1}^{n_i} Y_{ij}$ in cluster $i$ is presented as

$$f(z_i|\lambda_{i,z_i}, \lambda_{i,z_i+1}, \ldots, \lambda_{i,n_i}, n_i) = \binom{n_i}{z_i} \sum_{\ell=0}^{n_i-z_i} (-1)^\ell \binom{n_i - z_i}{\ell} \lambda_{i,z_i+\ell}, \tag{9}$$

in which

$$\lambda_{i,k} = \begin{cases} P(Y_{i1} = 1, Y_{i2} = 1, \ldots, Y_{ik} = 1) & \text{if } k = 1, \ldots, n_i, \\ 1 & \text{if } k = 0. \end{cases}$$

As a consequence, the parameter $\lambda_{i,k}$ can be interpreted as the probability that in cluster $i$, all individuals in a set of $k$ exhibit the event under consideration. The mean of the number of successes and the second order correlation between two responses of the same cluster can be expressed in terms of $\lambda_{i,k}$ parameters: $E(Z_i) = n_i\lambda_{i,1}$ and $\text{Corr}(Y_{ij}, Y_{ik}) = (\lambda_{i,2} - \lambda_{i,1}^2)/\lambda_{i,1}(1 - \lambda_{i,1})$. George and Bowman focus attention on the so-called *folded logistic* parameterization:

$$\lambda_{i,z_i+\ell}(\boldsymbol{\beta}) = \frac{2}{1 + \exp\left[-X_i\boldsymbol{\beta}\ln(z_i + \ell + 1)\right]}. \tag{10}$$

However, it turns out that the "specific" George-Bowman model with the folded logistic parameterization does not simplify to the binomial model in this case. In addition, the model is not coding invariant, i.e., if the 0/1 coding for successes and failures is swapped, the model changes and so do the maximum likelihood estimates. The first issue renders the model less desirable from an interpretational point of view, while the second one illustrates it is ill conceived. All of this is strongly reminiscent of the problems encountered with the continuation-ratio model.

## Some Marginal Models for Repeated Ordinal Data

While we already encountered problems with marginal models for repeated binary data, issues are magnified with ordinal outcomes. We will introduce a modeling

formalism in this section, then introduce conditional models in the next, after which we will be in a position to discuss the meaningfulness of one relative to the other.

The probit and Dale models have been proposed for multivariate and repeated ordered categorical outcomes, of which binary outcomes are a special case. In the case of the probit model, the ordinal outcome vector is assumed to arise from discretizing an underlying multivariate normal, whereas in the case of the Dale model an underlying Plackett distribution is assumed. In the first case, the association is captured by means of correlation coefficients, whereas in the second case global odds ratios are used to model the association.

The outcome for cluster $i$ is a series of measurements $Y_{ij}$ $(j = 1, \ldots, n_i)$. Assume that $Y_{ij}$ can take on $c_j$ distinct ordered values $k_j = 1, \ldots, c_j$. It is convenient to define so-called cumulative multi-indicator functions:

$$z_i(\boldsymbol{k}) = z_i(k_1, \ldots, k_{n_i}) = I(\boldsymbol{y}_i \leq \boldsymbol{k}).$$

The corresponding probability is denoted by $\mu_i(\boldsymbol{k})$. The choice to use cumulative indicators is in agreement with the ordinal nature of the outcomes. Setting one or more of the indices $k_j$ equal to their maximal value $c_j$ has the effect of marginalizing over the corresponding outcome. Doing this for all but one index results in the univariate indicators $z_{ijk} = I(y_{ij} \leq k)$ and their corresponding marginal probability $\mu_{ijk}$.

The ordering needed to stack the multi-indexed counts and probabilities into a vector will be assumed fixed. Several orderings of both $\boldsymbol{z}_i$ and $\boldsymbol{\mu}_i$ are possible. A natural choice is the lexicographic ordering, but this has the disadvantage of dispersing the univariate marginal counts and means over the entire vector. Therefore, we will group the elements first by dimensionality.

We can now complete the model by choosing appropriate link functions. For the vector of links $\boldsymbol{\eta}_i$ we consider a function, mapping the $C_i$-vector $\boldsymbol{\mu}_i$ $(C_i = c_1 \cdot c_2 \cdot \ldots \cdot c_{T_i})$

to

$$\boldsymbol{\eta}_i = \boldsymbol{\eta}_i(\boldsymbol{\mu}_i), \tag{11}$$

a $C_i'$-vector. Often, $C_i = C_i'$, and $\boldsymbol{\eta}_i$ and $\boldsymbol{\mu}_i$ have the same ordering. A counterexample is provided by the probit model, where the number of link functions is smaller than the number of mean components, as soon as $n_i > 2$.

We consider particular choices of link functions. The univariate logit link becomes $\eta_{ijk} = \ln(\mu_{ijk}) - \ln(1 - \mu_{ijk}) = \text{logit}(\mu_{ijk})$. The probit link is $\eta_{ijk} = \Phi_1^{-1}(\mu_{ijk})$, with $\Phi_1$ the univariate standard normal distribution.

However, univariate links alone do not fully specify $\boldsymbol{\eta}_i$, and hence leave the joint distribution partly undetermined. Full specification of the association requires addressing the form of pairwise and higher-order probabilities. First, we will consider the pairwise associations. Let us denote the bivariate probabilities, pertaining to the $j_1$th and $j_2$th outcomes, by

$$\mu_{i,j_1j_2,k_1k_2} = \mu_i(c_1, \ldots, c_{j_1-1}, k_1, c_{j_1+1}, \ldots, c_{j_2-1}, k_2, c_{j_2+1}, \ldots, c_{n_i}).$$

The Dale model is based on the marginal global odds ratio defined by

$$\psi_{i,j_1j_2,k_1k_2} = \frac{(\mu_{i,j_1j_2,k_1k_2})(1 - \mu_{ij_1k_1} - \mu_{ij_2k_2} + \mu_{i,j_1j_2,k_1k_2})}{(\mu_{ij_2k_2} - \mu_{i,j_1j_2,k_1k_2})(\mu_{ij_1k_1} - \mu_{i,j_1j_2,k_1k_2})} \tag{12}$$

and usefully modeled on the log scale. Higher order global odds ratios are defined similarly[22] and are omitted here.

The multivariate probit model also fits within the class defined by (11). For three categorical outcome variables, the inverse link is specified by

$$\mu_{ijk} = \Phi_1(\eta_{ijk}), \tag{13}$$

$$\mu_{i,j_1j_2,k_1k_2} = \Phi_2(\eta_{ij_1k_1}, \eta_{ij_1k_2}, \eta_{i,j_1j_2,k_1k_2}), \tag{14}$$

$$\mu_{i,123,k_1k_2k_3} = \Phi_3(\eta_{i1k_1}, \eta_{i2k_3}, \eta_{i3k_3}, \eta_{i,12,k_1k_2}, \eta_{i,13,k_1k_3}, \eta_{i,23,k_2k_3}), \tag{15}$$

where the notation for the three-way probabilities is obvious. The association links $\eta_{i,ts,k\ell}$ represent any transform (e.g., Fisher's $z$-transform) of the polychoric correlation coefficient. It is common practice to keep each correlation constant throughout a table, rather than having it depend on the categories: $\eta_{i,j_1j_2,k_1k_2} \equiv \eta_{i,j_1j_2}$. Relaxing this requirement may still give a valid set of probabilities, but the correspondence between the categorical variables and a latent multivariate normal variable is lost. Finally, observe that univariate links and bivariate links (representing correlations) fully determine the joint distribution. This implies that the mean vector and the link vector will have a different length, except in the univariate and bivariate cases.

It is useful to contrast these models against conditional models. However, we would like to assert that both the global odds ratio model, building upon univariate logits, and the multivariate probit model, often provide meaningful models, where a choice of one versus the other, even though there are differences, is less pronounced. This will be rather different in the context of random-effects models.

## Generalized Estimating Equations

The main issue with full likelihood approaches is the computational complexity they entail. When we are mainly interested in first-order marginal mean parameters and pairwise interactions, a full likelihood procedure can be replaced by quasi-likelihood method[21]. In quasi-likelihood, the mean response is expressed as a parametric function of covariates; the variance is assumed to be a function of the mean up to possibly unknown scale parameters. Wedderburn[27] first noted that likelihood and quasi-likelihood theories coincide for exponential families and that the quasi-likelihood "estimating equations" provide consistent estimates of the regression parameters $\boldsymbol{\beta}$ in any generalized linear model, even for choices of link and variance functions that do not correspond to exponential families.

For clustered and repeated data, Liang and Zeger [18] proposed so-called *generalized estimating equations* (GEE or GEE1) which require only the correct specification of the univariate marginal distributions provided one is willing to adopt "working" assumptions about the association structure. They estimate the parameters associated with the expected value of an individual's vector of binary responses and phrase the working assumptions about the association between pairs of outcomes in terms of marginal correlations. The method combines estimating equations for the regression parameters $\boldsymbol{\beta}$ with moment-based estimating for the correlation parameters entering the working assumptions.

Prentice[25] extended their results to allow joint estimation of probabilities and pairwise correlations. Lipsitz, Laird and Harrington[20] modified the estimating equations of Prentice[25] to allow modeling of the association through marginal odds ratios rather than marginal correlations. When adopting GEE1 one does not use information of the association structure to estimate the main effect parameters. As a result, it can be shown that GEE1 yields consistent main effect estimators, even when the association structure is misspecified. However, severe misspecification may seriously affect the efficiency of the GEE1 estimators. In addition, GEE1 should be avoided when some scientific interest is placed on the association parameters.

A second order extension of these estimating equations (GEE2) that include the marginal pairwise association as well has been studied by Liang, Zeger and Qaqish[19]. They note that GEE2 is nearly fully efficient though bias may occur in the estimation of the main effect parameters when the association structure is misspecified.

# MARGINAL VERSUS CONDITIONAL MODELS AND GLOBAL ODDS RATIOS

Fitting a marginal model is typically more involved than fitting conditional models, such as loglinear models [CROSSREF]. Among the marginal models considered, the Bahadur and George-Bowman models are easier to fit but they are less desirable, the first one because of a extremely heavily constrained parameter space, posing computational and interpretational issues, the second one because the folded logistic parameterization fails to reduce to logistic regression in the case of independent variables. It should be clear that most marginal models have constrained parameter spaces. This is often quoted as an interpretational disadvantage. However, the same is true for the multivariate normal model since the covariance matrix has to be positive definite. Exactly the same constraint applies to the multivariate probit model and similar but less tractable constraints apply to the multivariate Dale model. In contrast, the parameters of a loglinear model can take on any value in the Euclidean space whilst still producing valid probabilities.

It should be noted that, in this respect, marginal models differ one from the other. While in the Bahadur model the association parameter is restricted even when $n_i = n = 2$, this is not the case in the Dale model, where the odds ratio can range over the entire parameter space $[0, +\infty]$. Restrictions in the higher dimensional case, while in existence, are much weaker.

One of the main interpretational advantages of marginal models is their upward compatibility or reproducibility[19]. This means that when a marginal model (e.g., the Dale, probit, or Bahadur model) is used to model a response vector, then the appropriate sub-model applies to any subvector of the response vector. Precisely, such a sub-vector still follows a model of the same structure, with as parameter

vector the corresponding sub-vector. In particular, the univariate margins of the marginal models discussed above are typically of the logistic type, the probit model the obvious exception. The George-Bowman model is not upward compatible. Neither is the loglinear model upward compatible. This reduces the meaningfulness of such models in a number of contexts.

Marginal models should be chosen whenever there are marginal research questions, e.g., pertaining to one or a few occasions, or the evolution between them. They are also useful when not only the strength of association between occasions, but also a quantification of this association is of interest. Of course, when the number of measurement occasions within a subject grows, such models become intractable from a likelihood perspective. One can then resort to alternative approaches, such as generalized estimating equations or pseudo-likelihood.

Arguably, the odds ratio is a meaningful measure of association between repeated categorical outcomes. Of course, these can be defined in several ways. The loglinear model is based on conditional odds ratios, whereas the multivariate Dale model is based on marginal odds ratios. Apart from a marginal-conditional dimension, there is also a local-global dimension to the discussion. The association parameters in a standard log-linear model for contingency tables[2] are local, conditional odds ratios. The Dale model for ordinal outcomes uses global, marginal odds ratios. Lapp, Molenberghs, and Lesaffre[16] provide some support for the use of global odds ratios rather than local ones for cross-classified ordinal data.

**Acknowledgements.** We gratefully acknowledge support from Belgian IUAP/PAI network "Statistical Techniques and Modeling for Complex Substantive Questions with Complex Data".

Geert Molenberghs

Center for Satistics
Limburgs Universitair Centrum
Universitaire Campus, Building D
B3590 Diepenbeek, Belgium

geert.molenberghs@luc.ac.be

# References

[1] Aerts, M., Geys, H., Molenberghs, G., and Ryan, L.M. (2002). *Topics in Modelling of Clustered Binary Data*. London: Chapman & Hall.

[2] Agresti, A. (1990). *Categorical Data Analysis*. New York: John Wiley.

[3] Ashford, J.R. and Sowden, R.R. (1970). Multivariate probit analysis. *Biometrics*, **26**, 535–546.

[4] Bahadur, R.R. (1961). A representation of the joint distribution of responses of $n$ dichotomous items. In: *Studies in item analysis and prediction*, H. Solomon (Ed.), Stanford Mathematical Studies in the Social Sciences VI. Stanford, California, Stanford University Press.

[5] Dale, J.R. (1986). Global cross-ratio models for bivariate, discrete, ordered responses. *Biometrics*, **42**, 909–917.

[6] Declerck, L., Aerts, M., and Molenberghs, G. (1998). Behaviour of the likelihood ratio test statistic under a Bahadur model for exchangeable binary data. *Journal of Statistical Computations and Simulations*, **61**, 15–38.

[7] Diggle, P.J., Heagerty, P., Liang, K-Y., and Zeger, S.L. (2002). *Analysis of Longitudinal Data*. New York: Oxford University Press.

[8] Ekholm, A. (1991). Fitting regression models to a multivariate binary response. In *A Spectrum of Statistical Thought: Essays in Statistical Theory, Economics, and Population Genetics in Honour of Johan Fellman,* G. Rosenqvist, K. Juselius, K. Nordström, J. Palmgren, (eds.). Helsinki: Swedish School of Economics and Business Administration, pp. 19–32.

[9] Fahrmeir, L. and Tutz, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models.* Heidelberg: Springer-Verlag.

[10] George, E.O. and Bowman, D. (1995). A full likelihood procedure for analysing exchangeable binary data. *Biometrics*, **51**, 512–523.

[11] Geys, H., Molenberghs, G., and Lipsitz, S.R. (1998). A note on the comparison of pseudo-likelihood and generalized estimating equations for marginal odds ratio models. *Journal of Statistical Computation and Simulation*, **62**, 45–72.

[12] Glonek, G.F.V. and McCullagh, P. (1995). Multivariate logistic models. *Journal of the Royal Statistical Society, Series B*, **57**, 533–546.

[13] Kupper, L.L. and Haseman, J.K. (1978). The use of a correlated binomial model for the analysis of certain toxicology experiments. *Biometrics*, **34**, 69-76.

[14] Lang, J.B. and Agresti, A. (1994). Simultaneously modeling joint and marginal distributions of multivariate categorical responses. *Journal of the American Statistical Association*, **89**, 625–632.

[15] Laird, N.M. and Ware, J.H. (1982). Random effects models for longitudinal data. *Biometrics* **38**, 963–974.

[16] Lapp, K., Molenberghs, G., and Lesaffre, E. (1998). Local and global cross ratios to model the association between ordinal variables. *Computational Statistics and Data Analysis*, **28**, 387–411.

[17] le Cessie, S. and van Houwelingen, J.C. (1995). Testing the fit of a regression model via score tests in random effects models. *Biometrics* **51**, 600–614.

[18] Liang, K.-Y. and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13–22.

[19] Liang, K.Y., Zeger, S.L., and Qaqish, B. (1992). Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society, Series B*, **54**, 3–40.

[20] Lipsitz, S.R., Laird, N.M., and Harrington, D.P. (1991). Generalized estimating equations for correlated binary data: using the odds ratio as a measure of association. *Biometrika*, **78**, 153–160.

[21] McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models*. London: Chapman & Hall.

[22] Molenberghs, G. and Lesaffre, E. (1994). Marginal modelling of correlated ordinal data using a multivariate Plackett distribution. *Journal of the American Statistical Association*, **89**, 633–644.

[23] Molenberghs, G. and Lesaffre, E. (1999). Marginal modelling of multivariate categorical data. *Statistics in Medicine*, **18**, 2237–2255.

[24] Plackett, R.L. (1965). A class of bivariate distributions. *Journal of the American Statistical Association*, **60**, 516–522.

[25] Prentice, R.L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics*, **44**, 1033-1048.

[26] Verbeke, G. and Molenberghs, G. (2000). *Linear mixed models for longitudinal data*, Springer Series in Statistics, Springer-Verlag, New-York.

[27] Wedderburn, R.W.M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika*, **61**, 439–447.

## Related Entries: XXX