TECHNICAL
REPORT

0531

# CONDITIONAL INDEPENDENCE OF MULTIVARIATE BINARY DATA WITH AN APPLICATION IN CARIES RESEARCH

M.J. GARCIA-ZATTERA, A. JARA, E. LESAFFRE AND D. DECLERCK

IAP STATISTICS
NETWORK

INTERUNIVERSITY ATTRACTION POLE

http://www.stat.ucl.ac.be/IAP

# Conditional independence of multivariate binary data with an application in caries research

María José García-Zattera*, Alejandro Jara*,
Emmanuel Lesaffre* and Dominique Declerk**

* Biostatistical Centre, Katholieke Universiteit Leuven, Belgium
** School of Dentistry, Katholieke Universiteit Leuven, Belgium

Jun, 27th, 2005

**Abstract**

For the analysis of caries experience in seven year old children we explored the association between the presence or absence of caries experience among different deciduous molars within each child. Some of the observed high associations have an etiological basis (e.g., between symmetrically opponent molars), while others (diagonally opponent molars) are assumed to be the result of the transitivity of association and hence are believed to disappear once conditioned on the caries experience status of the other deciduous molars. However, using discrete models for multivariate binary data, conditioning on the caries experience of the other teeth present in the mouth and on the (un)known subject-specific characteristics did not remove the latter type of association. When the association was explored on a latent scale, say by a multivariate probit model, then the partial correlation matrix indicated conditional independence. This contrast was confirmed when using other models on the (observed) binary scale and on the latent scale. While it seems logical that conditional dependence partly depends on the chosen model, our example shows that the results and conclusions can be markedly different. The explanation for this surprising result is exemplified mathematically and illustrated using dental data from the Signal Tandmobiel® study.

**Key words:** Conditional independence, Multivariate binary data, Latent variable representation, Multivariate probit model

## 1 Introduction

In oral health research it is of interest to assess the association of caries experience among different teeth. The knowledge that caries development on one tooth is related to caries development on another tooth can help the dentist in optimizing his/her clinical examination of the patient and directs preventive and restorative approaches. Further, the exploration of caries experience patterns in the mouth can also help in further refining the understanding of the etiology of the disease. Indeed, it is still not established whether

[1] Correspondence to: Biostatistical Centre, Katholieke Universiteit Leuven, Kapucijnenvoer 35, B-3000 Leuven, Belgium. E-mail : Emmanuel.Lesaffre@med.kuleuven.be

caries is a spatially local disease or not and the answer to that question might be related to the variety of factors determining caries activity (see, *e.g.*, Hujoel, Lamont, DeRouen, Davis and Lerouxi 1994, and references therein).

Based on data obtained in seven-year old children recruited in the Signal Tandmobiel® study, we examined the association between the presence/absence of caries experience on the eight deciduous molars and found a high association between symmetrically opponent molars, matching molars from the maxilla and the mandible (vertically opponent teeth) and diagonally opponent teeth. The first association is known and relatively easy to explain (Psoter, Zhang, Pendrys, Morse and Mayne 2003). The second association is somewhat more difficult to understand. However, the high association between diagonally opponent teeth is believed to be the result of the (assumed) transitivity of the associations, i.e. due to the high association between symmetrically opponent deciduous molars on the one hand and the association between vertically opponent molars on the other hand. This was verified by fitting a random effects logistic regression model (with subject as random effect) explaining the occurrence of caries experience on a deciduous molar by the caries experience on the other molars and subject specific characteristics. However, this model was not able to remove this high association, and the same was true for all other considered discrete models for multivariate binary vectors. In contrast, when the association was explored on a latent scale, say by a multivariate probit model, then the partial correlation matrix indicated conditional independence.

While we acknowledge that conclusions can change when different statistical models are used, we were surprised to see such a major difference when switching from one class of models (on the observed binary scale) to another class of models (on the latent continuous scale). In this paper we will highlight a possible reason why conditional independence is not invariant to the scale used for the analysis.

To illustrate the markedly different conclusions that can be obtained from different statistical models for multivariate binary responses, we analyzed the same caries experience data with two different models. First, the conditionally specified logistic regression model (CSLRM) as suggested by Joe and Liu (1996) was applied. Like the log-linear model (LLM) the model acts on the observed binary scale, but the model also allows the inclusion of covariates. Further, the model is intimately related to logistic regression. In the CSLRM, the association is measured by the odds ratio of a pair of binary responses conditional on the observed values of the remaining binary responses and the covariates. Consequently, the estimated odds ratios automatically express conditional (in)dependence. Secondly, the multivariate probit model (MPM) was applied, see, *e.g.*, Ashford and Sowden (1970), Lesaffre and Molenberghs (1991) or Chib and Greenberg (1998) for various implementations and examples. This model expresses the association between the binary responses via the correlation matrix of a multivariate normal latent random vector. Conditional (in)dependence can be evaluated by the partial correlation matrix.

In Section 2, independence and conditional independence are reviewed. In Section 3, we briefly review the Conditional Logistic Regression Model and the Multivariate Probit model. An application to oral health data from the Signal Tandmobiel® study is shown in Section 4. Finally, Section 5 gives some concluding remarks.

# 2 Independence and Conditional Independence

Suppose that $\mathbf{V}$ is a $m$-dimensional normally distributed random vector and that a random sample of $n$ individuals is available yielding vectors $\mathbf{V}_i$ $(i = 1, \ldots, n)$. However, we assume that $\mathbf{V}$ is not observed but latent and that either $\mathbf{Y}$ or $\mathbf{Z}$ is observed. The first observed random vector is continuous, namely $\mathbf{Z}_i = \mathbf{V}_i + \boldsymbol{\varepsilon}_i$ $(i = 1, \ldots, n)$, where $\boldsymbol{\varepsilon}_i$ is normally distributed and independent of $\mathbf{V}_i$. On the other hand, $\mathbf{Y}_i$ is a random multivariate binary response vector defined as $Y_{ij} = I(V_{ij} > c_j)$, where $c_j$ $(j = 1, \ldots, m)$ are specific cut off points.

The correlation matrix $\mathbf{R} \equiv (\rho_{jk})_{jk}$ corresponding to $\mathbf{V}$ describes the association structure of the latent vector and conditional independence is seen from the elements of the standardized concentration matrix $\mathbf{C} \equiv (c_{jk})_{jk}$, obtained from appropriately standardizing $\mathbf{R}^{-1}$. Namely, $V_j$ is conditionally independent of $V_k$ conditional on the other $V_m$ for $m \neq k, j$ when $c_{jk} = 0$. Observe that this property does not hold for other multivariate distributions and hence in these cases a partial correlation equal to zero does not imply conditional independence.

In this paper we are interested in the relationship between the association structure on the latent scale (of $\mathbf{V}$) and the observed scale (of $\mathbf{Y}$ and $\mathbf{Z}$) especially with respect to conditional independence. Clearly, if $\mathbf{R}$ is the identity matrix, then also the components of $\mathbf{Y}$ and $\mathbf{Z}$ are statistically independent. Further, the association structure of $\mathbf{Z}$ depends on the magnitude of the measurement error component defined by $\boldsymbol{\varepsilon}$. For instance, even when the components of $\mathbf{V}$ are perfectly related, the components of $\mathbf{Z}$ could show a poor correlation if the variability of $\boldsymbol{\varepsilon}$ is quite high. Furthermore, conditional independence can not be expected for $\mathbf{Z}$ even when it holds for $\mathbf{V}$. For the binary case, $\rho_{jk} = 0$, $j \neq k$ implies independence of $Y_j$ and $Y_k$. But again, conditional independence for $\mathbf{V}$ does not imply conditional independence for $\mathbf{Y}$ and this will be illustrated now.

Consider the random vector $\mathbf{V} \sim N_3(\mu, \mathbf{R})$, with,

$$\mu = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \qquad \mathbf{R} = \begin{pmatrix} 1.00 & 0.64 & 0.80 \\ 0.64 & 1.00 & 0.80 \\ 0.80 & 0.80 & 1.00 \end{pmatrix}$$

and the categorical variables $Y_j$, $(j = 1, 2, 3)$ defined as above. The standardized concentration matrix then becomes

$$\mathbf{C} = \begin{pmatrix} 1.00 & 0 & 0.62 \\ 0 & 1.00 & 0.62 \\ 0.62 & 0.62 & 1.00 \end{pmatrix}.$$

Since $c_{12} = 0$ the partial correlation coefficient $\rho_{V_1,V_2.V_3} = 0$ and thus $V_1 \perp\!\!\!\perp V_2 | V_3$. However, the probability of $Y_1$ and $Y_2$ given $Y_3$ is,

$$P(Y_1 = 1, Y_2 = 1 | Y_3 = 1) = 0.6557$$

while, $P(Y_1 = 1 | Y_3 = 1) = P(Y_2 = 1 | Y_3 = 1) = 0.7952$, and,

$$P(Y_1 = 1 | Y_3 = 1)\, P(Y_2 = 1 | Y_3 = 1) = 0.6323$$

Hence, we have shown numerically that conditional independence of variables $V_1$ and $V_2$ given $V_3$ does not imply $Y_1 \perp\!\!\!\perp Y_2|Y_3$. The evaluation of these expressions involves the computation of multivariate normal probabilities which was carried out using the methodology described in Genz (1992) and Genz (1993). A theoretical proof of this result is shown in the Appendix.

# 3 Two models for the analysis of multivariate binary responses

In this section we will describe two models that were used to illustrate the difference between analyzing the multivariate binary response on the observed binary scale and on the latent continuous scale. The models are just representative for their class of models since the above mentioned contrast remains when these models are replaced by other models, as indicated below.

## 3.1 The Conditionally Specified Logistic Regression Model: a model on the observed binary scale

Let $\mathbf{Y_i}$ be defined as before and let $\mathbf{X}_{ij}$ be the correspondent covariate vector. Joe and Liu (1996), discuss a model for multivariate binary responses with covariates. The conditional distribution of each binary response $Y_{ij}$ given the other binary responses $Y_{ik} = y_{ik}$, $k \neq j$ and the covariates $\mathbf{X}_{ij}$ is equivalent to a logistic regression with parameter vector $\boldsymbol{\beta}_j$ and parameters $\gamma_{jk}$, $k \neq j$. That is, for $j = 1, ..., m$,

$$\text{logit} P(Y_{ij} = 1|Y_{ik} = y_{ik}, k \neq j, \mathbf{X}_{ij}) = \mathbf{X}_{ij}\boldsymbol{\beta}_j + \sum_{k \neq j} \gamma_{jk} y_{ik}. \tag{1}$$

Joe and Liu (1996) showed that a necessary and sufficient condition for compatibility of conditional distributions is that $\gamma_{jk} = \gamma_{kj}$, $j \neq k$, and that the joint distribution is given by,

$$p(\mathbf{Y}|\mathbf{X}) = \prod_{i=1}^{n} \left[ c(\mathbf{X}_i, \boldsymbol{\beta}, \boldsymbol{\gamma})^{-1} \exp\left\{ \sum_{j=1}^{m} (\mathbf{X}_{ij}\boldsymbol{\beta}_j) y_{ij} + \sum_{1 \leq j < k \leq m} \gamma_{jk}\ y_{ij} y_{ik} \right\} \right] \tag{2}$$

with normalizing constant,

$$c(\mathbf{X}_i, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \sum_{y_1=0}^{1} \cdots \sum_{y_m=0}^{1} \exp\left\{ \sum_{j=1}^{m} (\mathbf{X}_{ij}\boldsymbol{\beta}_j) y_j + \sum_{1 \leq j < k \leq m} \gamma_{jk}\ y_j y_k \right\} \tag{3}$$

In (1) to (3), the parameters $\gamma_{jk}$ are interpreted as conditional log-odds ratios, since,

$$\begin{aligned} \exp\{\gamma_{jk}\} \quad &= \quad \frac{P(Y_{ij} = 1, Y_{ik} = 1|\mathbf{X}_{ij}, \mathbf{X}_{ik}, Y_{il} = y_{il}, l \neq j, k)}{P(Y_{ij} = 1, Y_{ik} = 0|\mathbf{X}_{ij}, \mathbf{X}_{ik}, Y_{il} = y_{il}, l \neq j, k)} \times \\ & \quad\quad \frac{P(Y_{ij} = 0, Y_{ik} = 0|\mathbf{X}_{ij}, \mathbf{X}_{ik}, Y_{il} = y_{il}, l \neq j, k)}{P(Y_{ij} = 0, Y_{ik} = 1|\mathbf{X}_{ij}, \mathbf{X}_{ik}, Y_{il} = y_{il}, l \neq j, k)} \end{aligned} \tag{4}$$

Note that for $m = 2$, there are no $Y_{il}$'s so that $\gamma_{12}$ is also the unconditional log-odds ratio and it is constant over the covariates. For $m \geq 3$, it is straightforward to show that the bivariate marginal distributions from (2), and the log-odds ratios depend on the covariates. Note also that the exponential family in (2) is not closed under margins and can be easily extended if interaction terms are needed.

In the absence of covariates, it is popular to analyze conditional independence in the observed binary scale with a log-linear model. For $Y_1$, $Y_2$ and $Y_3$ a LMM up to two-way interactions is given by

$$log(\mu_{jkl}) = \lambda + \lambda_j^{Y_1} + \lambda_k^{Y_2} + \lambda_l^{Y_3} + \lambda_{jk}^{Y_1Y_2} + \lambda_{jl}^{Y_1Y_3} + \lambda_{kl}^{Y_2Y_3}, \tag{5}$$

where $\lambda$ is the overall mean of the natural logarithm of the expected frequencies, $\lambda_j^{Y_1}$, $\lambda_k^{Y_2}$, $\lambda_l^{Y_3}$ represent the main effects for variables $Y_1$, $Y_2$ and $Y_3$, respectively; and $\lambda_{jk}^{Y_1Y_2}$, $\lambda_{jl}^{Y_1Y_3}$, and $\lambda_{kl}^{Y_2Y_3}$ represent the respective interaction effects. In this case, the null hypothesis of conditional independence between two variables given the other one, for instance $Y_1$ and $Y_2$ given $Y_3$, is $H_0 : \lambda_{jk}^{Y_1Y_2} = 0, \forall\ j, k$. We applied also the LLM to the dental example but with basically the same results.

A program for maximum likelihood analysis using the *R*-software (R Development Core Team 2004) was written for the analysis of the dental data with the CLRM. FORTRAN subroutines are called to speed up the computations. The function was implemented in the *R*-package *cslogistic* and is available from CRAN or upon request from the authors. This package also contain functions for Bayesian analysis under the CLRM.

## 3.2 The Multivariate Probit Model: a model on the latent continuous scale

A common used alternative modelling strategy for multivariate categorical data involves the introduction of latent variables (see *e.g.* Ashford and Sowden 1970; Lesaffre and Molenberghs 1991; Albert and Chib 1993; Chib and Greenberg 1998). The typical representation consists of considering the binary variables as a discrete version of underlying continuous data divided by a threshold in two categories. Indeed, the key idea is to introduce a $m$-dimensional latent variable vector $\mathbf{V}_i = (V_{i1}, ..., V_{im})$, such that,

$$p\left(\mathbf{Y_i}|\mathbf{V_i}\right) = \Sigma_{y_{i1},...,y_{im}} \left\{ I\left(Y_{i1} = y_{i1}, ..., Y_{im} = y_{im}\right) I \left\{ \bigcap_{j=1}^{m} V_{ij} \in A_{y_{ij}} \right\} \right\}, \tag{6}$$

where $A_{y_{ij}} = (0, +\infty)$ if $y_{ij} = 1$ and $A_{y_{ij}} = (-\infty, 0)$ if $y_{ij} = 0$, $j = 1, ..., m$. A common distributional assumption, leading to the Multivariate Probit Model, is $\mathbf{V_i} \sim N_m\left(\mathbf{X_i}\beta, \mathbf{R}\right)$, where $\mathbf{X_i}$ is a matrix of covariates associated to the regression parameters vector $\beta$ and, for identifiability reasons the matrix $\mathbf{R}$ must be in correlation form (Chib and Greenberg (1998)). The correlations $\rho_{jk} = corr\left(V_j, V_k\right)$ are known as the *tetrachoric correlation coefficients.*

For convenience, i.e., to avoid laborious first and second derivatives, a Bayesian approach was taken to analyze the dental data with the MPM. Noninformative prior distributions were given for all parameters of the model. Posterior distributions of the parameters were estimated using Markov Chain Monte Carlo techniques and the Metropolized hit-and-run algorithm proposed by Chen and Schmeiser (1993) was used to generate correlation matrices. The Markov chain was initialized with all the regression coefficients, except the intercepts, equal to zero. The first 10,000 samples were discarded as burn-in and an additional 400,000 iterations were used to compute posterior summaries (posterior mean and 95% Highest Posterior Density intervals using the method of Chen and Shao 1999). Convergence was checked using standard criteria (Cowles and Carlin 1996) as implemented in the BOA package (Smith 2005).

The Bayesian Multivariate Logistic Model (MLM) of O'Brien and Dunson (2004) was also fitted to the dental data. Since the posterior distribution of regression coefficients, and marginal and partial correlation coefficients were basically the same as for the MPM, they are not shown. However, is important to note that in the MLM framework the partial correlation matrix does not represent conditional independence.

# 4 Analysis of the Oral Health Example

## 4.1 The Oral Health Question

The Signal-Tandmobiel® study is a longitudinal prospective oral health screening study conducted in Flanders (Belgium) between 1996 and 2001. For this project, 4468 children were examined on a yearly basis during their primary school time by one of sixteen trained and calibrated dental examiners. Data on oral hygiene and dietary habits were obtained through structured questionnaires, completed by the parents. For a more detailed description of the Signal-Tandmobiel® study we refer to Vanobbergen et al. (2000).

Based on the first year oral health data, we examined the association pattern of caries experience in the mouth. It is well known that a strong association between neighboring teeth exists. However, it is also of interest to know whether other relationships exist. The knowledge of these relationships could be important for dental practice because it can trigger the dentist to examine the mouth in a more focused manner. Also, specific associations can direct preventive and restorative approaches. Furthermore, a thorough exploration of the caries experience pattern is also of dental theoretical interest because it can suggest further refinement of the knowledge on the etiology of caries.

Here, caries experience of the 8 deciduous molars was analyzed using a conditionally specified logistic regression model and a multivariate probit model. For ease of exposition, the European notation to indicate the location of a deciduous tooth in the mouth is shown in Figure 1. Covariates included in the models were age (in years) (Age), gender (boys versus girls) (Gender), age at start of brushing (in years) (Startbr), regular use of fluoridated supplements (yes versus no) (Sysfl), daily use of sugar containing drinks (no versus yes) (Drinks), number of between-meal snacks (two or less than two a day versus more than two a day) (Meals) and frequency of tooth brushing (once or more a day versus less than once a day) (Freqbrus). Except for the intercept, it was assumed that the covariates have a common effect on the probabilities of caries experience for all teeth.

Table 1 shows the unconditional odds ratios expressing the association of caries experience in the eight different molars. The table shows that adjacent (e.g., molars 54 and 55), homologous (e.g., molars 54 and 64) and vertically opponent teeth (e.g., molars 54 and 84) have a high association. However, also the association between diagonally opponent teeth (e.g., molars 54 and 74) seems to be high. Observe that in this analysis no correction

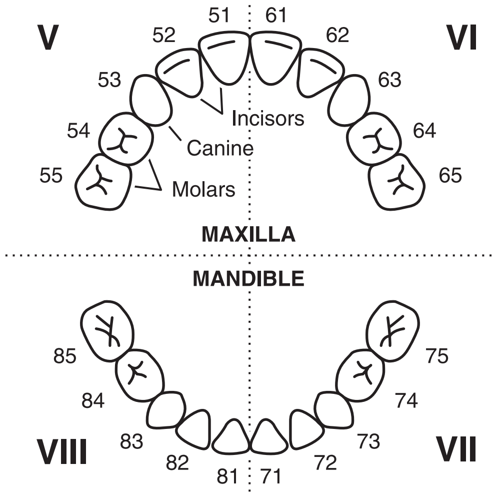

Figure 1: European notation to indicate the location of the deciduous teeth in the mouth.

for covariates was achieved nor did we take into account the caries experience pattern of other teeth in the mouth. The dentists speculated that the high association between the diagonally opponent teeth was due to the high association between the homologous teeth and the high association between the opponent teeth. It was hoped that a conditional analysis could demonstrate this.

Table 1: Signal-Tandmobiel® Study: unconditional odds ratios (95% CI) for caries experience in deciduous molars.

| | Tooth | | | | | | |
|---|---|---|---|---|---|---|---|
| **Tooth** | **64** | **74** | **84** | **55** | **65** | **75** | **85** |
| **54** | 16.54<br>(13.40 ; 20.34) | 8.59<br>(7.08 ; 10.43) | 7.87<br>(6.49 ; 9.55) | 11.00<br>(9.03 ; 13.41) | 7.08<br>(5.85 ; 8.56) | 5.68<br>(4.71 ; 6.85) | 5.60<br>(4.65 ; 6.75) |
| **64** | | 8.33<br>(6.89 ; 10.07) | 7.48<br>(6.20 ; 9.03) | 7.05<br>(5.85 ; 8.50) | 11.84<br>(9.74 ; 14.41) | 5.29<br>(4.40 ; 6.35) | 5.22<br>(4.35 ; 6.26) |
| **74** | | | 24.18<br>(19.88;29.40) | 6.64<br>(5.58 ; 7.91) | 6.19<br>(5.20 ; 7.36) | 9.46<br>(7.93 ; 11.29) | 7.58<br>(6.38 ; 9.01) |
| **84** | | | | 6.48<br>(5.44 ; 7.71) | 6.46<br>(5.43 ; 7.68) | 8.27<br>(6.95 ; 9.84) | 8.88<br>(7.46 ; 10.58) |
| **55** | | | | | 14.69<br>(12.12 ; 17.79) | 8.89<br>(7.42 ; 10.65) | 8.61<br>(7.19 ; 10.31) |
| **65** | | | | | | 7.79<br>(6.52 ; 9.30) | 8.13<br>(6.80 ; 9.72) |
| **75** | | | | | | | 20.31<br>(16.70 ; 24.70) |

## 4.2 Conditionally Specified Logistic Regression

Table 2 presents the regression coefficients of the conditional logistic regression model. The results indicate clear differences in caries experience with respect to age of the child, age at start of brushing, regular use of fluoridated supplements, daily use of sugar containing drinks and number of between-meal snacks.

The conditional odds ratios for caries experience in deciduous molars are shown in Table 3. The table shows that adjacent, homologous and vertically opponent teeth have

Table 2: Signal-Tandmobiel® Study: Conditionally specified logistic regression analysis for caries experience in eight deciduous molars.

| Covariate | Estimate | 95% CI |
|---|---|---|
| Age(yrs) | 0.067 | ( 0.031 ; 0.103) |
| Gender(girls) | 0.013 | (-0.016 ; 0.042) |
| Startbr(yrs) | 0.039 | ( 0.025 ; 0.053) |
| Sysfl(no) | 0.109 | ( 0.078 ; 0.139) |
| Drinks(yes) | 0.098 | ( 0.067 ; 0.130) |
| Meals (> 2/day) | 0.041 | ( 0.010 ; 0.072) |
| Freqbrus(< 1/day) | 0.039 | (-0.001 ; 0.079) |

a high association. However, all the associations between diagonally opponent teeth remained highly positive and significant.

Table 3: Signal-Tandmobiel® Study: Conditional odds ratios (95% CI) for caries experience in deciduous molars.

| | Tooth | | | | | | |
|---|---|---|---|---|---|---|---|
| **Tooth** | **64** | **74** | **84** | **55** | **65** | **75** | **85** |
| **54** | 5.59<br>(4.46 ; 7.00) | 1.93<br>(1.44 ; 2.58) | 1.68<br>(1.25 ; 2.26) | 4.04<br>(3.14 ; 5.18) | 0.91<br>(0.68 ; 1.20) | 1.02<br>(0.75 ; 1.39) | 1.12<br>(0.83 ; 1.51) |
| **64** | | 2.13<br>(1.60 ; 2.85) | 1.58<br>(1.18 ; 2.13) | 1.01<br>(0.76 ; 1.33) | 5.07<br>(3.98 ; 6.44) | 1.02<br>(0.75 ; 1.39) | 0.93<br>(0.69 ; 1.27) |
| **74** | | | 0.15<br>(0.13 ; 0.195) | 1.41<br>(1.04 ; 1.91) | 1.04<br>(0.76 ; 1.41) | 3.09<br>(2.35 ; 4.07) | 1.15<br>(0.85 ; 1.55) |
| **84** | | | | 1.14<br>(0.84 ; 1.55) | 1.70<br>(1.26 ; 2.28) | 1.51<br>(1.14 ; 2.02) | 3.10<br>(2.35 ; 4.09) |
| **55** | | | | | 6.00<br>(4.80 ; 7.51) | 2.17<br>(1.65 ; 2.84) | 2.14<br>(1.63 ; 2.81) |
| **65** | | | | | | 1.82<br>(1.38 ; 2.41) | 2.17<br>(1.65 ; 2.86) |
| **75** | | | | | | | 9.88<br>(8.06 ;12.10) |

## 4.3 Multivariate Probit Model

Based on the analysis using the MPM model, we can conclude that the association of caries experience in the mouth was high and significant between symmetrical and vertically opponent teeth but also important for diagonally opponent teeth. This is seen from the posterior summaries of the correlation matrix shown in Table 4. The analysis revealed that all correlation coefficients were significant and considerably high.

The posterior summaries of the regression coefficients in the model showed basically the same results as for the conditionally specified logistic regression model and are therefore omitted here.

From the estimated correlation one can calculate the partial correlation matrix. Here all partial correlations are smaller than the corresponding correlations, but the difference was the biggest for the diagonally opponent molars. For instance, Figures 2 and 3, show the posterior distributions of tetrachoric correlations and partial correlations for

Table 4: Signal-Tandmobiel® Study: Posterior mean (95% HPD) of latent marginal correlation matrix for caries experience.

| | Tooth | | | | | | |
|---|---|---|---|---|---|---|---|
| **Tooth** | **64** | **74** | **84** | **55** | **65** | **75** | **85** |
| **54** | 0.78<br>(0.74 ; 0.81) | 0.65<br>(0.62 ; 0.69) | 0.62<br>(0.58 ; 0.66) | 0.70<br>(0.65 ; 0.74) | 0.61<br>(0.56 ; 0.66) | 0.55<br>(0.50 ; 0.59) | 0.55<br>(0.49 ; 0.62) |
| **64** | | 0.64<br>(0.60 ; 0.68) | 0.61<br>(0.56 ; 0.65) | 0.61<br>(0.56 ; 0.66) | 0.72<br>(0.68 ; 0.76) | 0.53<br>(0.48 ; 0.58) | 0.53<br>(0.46 ; 0.58) |
| **74** | | | 0.85<br>(0.82 ; 0.87) | 0.61<br>(0.56 ; 0.66) | 0.60<br>(0.55 ; 0.64) | 0.69<br>(0.64 ; 0.73) | 0.64<br>(0.59 ; 0.69) |
| **84** | | | | 0.60<br>(0.55 ; 0.65) | 0.60<br>(0.56 ; 0.64) | 0.66<br>(0.61 ; 0.70) | 0.67<br>(0.62 ; 0.71) |
| **55** | | | | | 0.77<br>(0.73 ; 0.81) | 0.67<br>(0.63 ; 0.71) | 0.67<br>(0.62 ; 0.72) |
| **65** | | | | | | 0.64<br>(0.60 ; 0.68) | 0.66<br>(0.60 ; 0.70) |
| **75** | | | | | | | 0.82<br>(0.79 ; 0.85) |

homologous pairs 54 and 64, and diagonal opponent teeth 54 and 74, respectively. Clearly, the largest difference between the two correlation coefficients is seen for teeth 54 and 74.

Table 5 presents posterior summaries of the whole partial correlation matrix. Note that a zero entry in this matrix corresponds to conditional independence between corresponding latent variables. While, all the associations between neighboring and symmetrical teeth remained highly positive and significant, the association between opponent and diagonally opponent teeth were in most of the cases not significant, suggesting that the highly observed marginal association could be almost totally explained by the transitivity of the correlation structure. Further, compared to the results of the CLRM a total of ten discordant results were found. In eight of these cases, the conditional odds ratios are significant while the partial correlations are not.

Table 5: Signal-Tandmobiel® Study: Posterior mean (95% HPD) of latent partial correlation matrix for caries experience.

| | Tooth | | | | | | |
|---|---|---|---|---|---|---|---|
| **Tooth** | **64** | **74** | **84** | **55** | **65** | **75** | **85** |
| **54** | 0.57<br>(0.49 ; 0.65) | 0.10<br>(-0.02 ; 0.23) | 0.06<br>(-0.07 ; 0.18) | 0.42<br>(0.33 ; 0.51) | -0.23<br>(-0.37 ; -0.13) | -0.03<br>(-0.16 ; 0.09) | 0.02<br>(-0.09 ; 0.15) |
| **64** | | 0.14<br>(0.02 ; 0.26) | 0.01<br>(-0.11 ; 0.14) | -0.21<br>(-0.32 ; -0.10) | 0.49<br>(0.41 ; 0.58) | -0.03<br>(-0.16 ; 0.09) | -0.04<br>(-0.17 ; 0.08) |
| **74** | | | 0.65<br>(0.60 ; 0.70) | 0.03<br>(-0.08 ; 0.14) | -0.05<br>(-0.17 ; 0.06) | 0.26<br>(0.16 ; 0.36) | -0.08<br>(-0.19 ; 0.02) |
| **84** | | | | -0.01<br>(-0.130 ; 0.09) | 0.06<br>(-0.05 ; 0.18) | -0.05<br>(-0.16 ; 0.06) | 0.22<br>(0.11 ; 0.32) |
| **55** | | | | | 0.50<br>(0.42 ; 0.58) | 0.14<br>(0.03 ; 0.26) | 0.10<br>(-0.03 ; 0.21) |
| **65** | | | | | | 0.07<br>(-0.06 ; 0.19) | 0.13<br>(0.02 ; 0.26) |
| **75** | | | | | | | 0.58<br>(0.52 ; 0.65) |

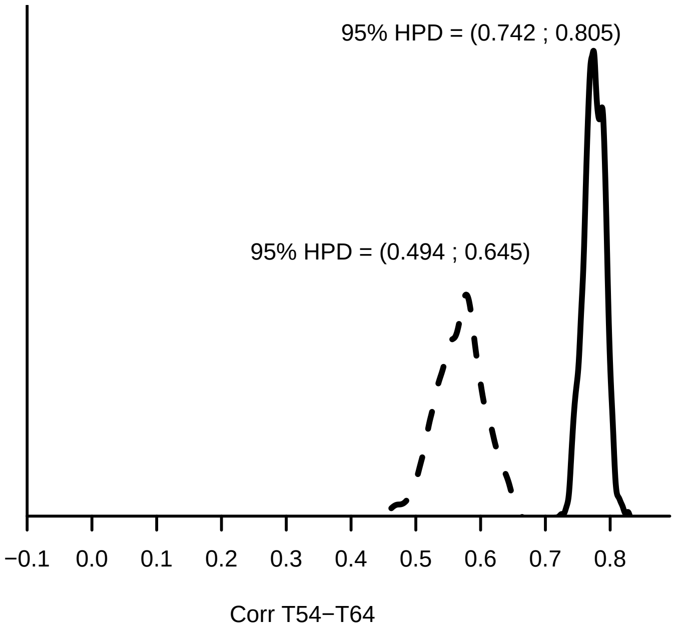

Figure 2: Signal-Tandmobiel® Study: Tetrachoric correlation coefficients for caries experience in tooth 54 and 64. The marginal and the partial correlations are shown in solid and dashed lines, respectively.

# 5 Concluding Remarks

Conditional independence is presently accepted as a fundamental concept not only in the theory of statistical inference (see, e.g., Dawid 1979; Nogales, Oyola, and Pérez 2000), but also in structural modelling (Pearl 1995). Model building most often deals with structural properties underlying a process generating latent as well as observed variables.

The multivariate probit model represents one of the strategies for the analysis of clustered multivariate binary data, which is described in terms of a correlated Gaussian distribution for underlying latent variables that are manifested as discrete variables through a threshold specification. Although latent variable modelling could be viewed as a dubious exercise fraught with unverifiable assumptions and naive inferences regarding causality, the multivariate probit model is a natural way of relating stimulus and response where such an interpretation for a threshold approach is readily available; examples include attitude measurement, assigning pass/fail gradings for examinations based on mark cut-off, and bioassay settings were the underlying continuous scale can be a lethal dose of a drug.

We showed that the association structure on the latent continuous scale is not transferable to the observed binary scale. In particular that conditional independence on the latent scale does not transfer to the observed scale. Which of the two scales provide us with the answer will depend on the problem and biological evidence there is. The important issue of this paper is to show that the two analyses can and often will yield two different interpretations. With regard to our oral health example on caries experience

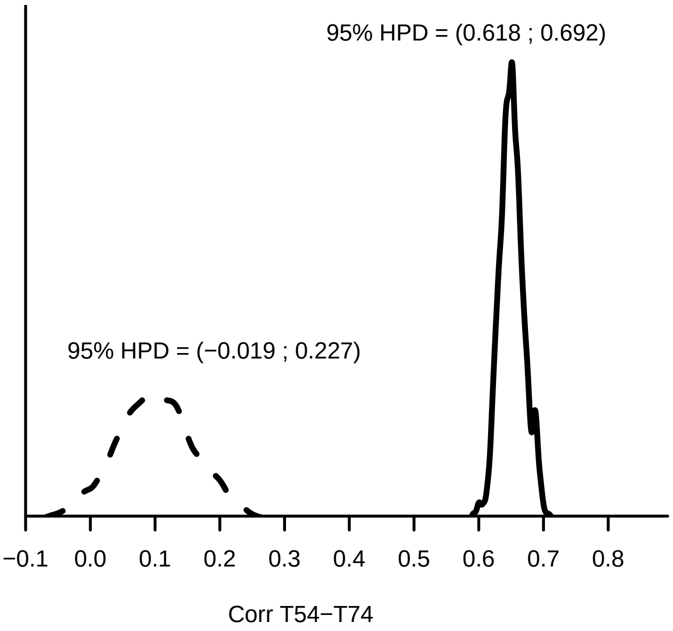

Figure 3: Signal-Tandmobiel$^{®}$ Study: Tetrachoric correlation coefficients for caries experience in tooth 54 and 74. The marginal and the partial correlations are shown in solid and dashed lines, respectively.

we conclude that opponent and diagonally opponent molars are indeed (conditionally) independent for caries experience. The basis for this conclusion is: (a) our findings with the MPM and (b) the absence of a biological explanation for a direct association of caries experience in diagonally opponent teeth. There is further dental evidence for our conclusion. Indeed, Veerkamp and Weerheijm (1995) pointed out that caries experience also very much depends on the eruption stage. Namely, that caries can only develop when the respective tooth has been exposed long enough. Now teeth in the maxilla emerge earlier than teeth in the mandible. Hence, symmetrically opponent molars have about the same emergence time while opponent and diagonally opponent teeth emerge at different ages providing extra evidence that these associations are not etiological.

Our findings are also of importance in model building exercises in general. In fact, the decision to increase the complexity of the model depends on whether the extra variate has a (significant) relationship with a particular response, conditional on the already included covariates and the remaining responses. In this context, Webb and Forster (2004) suggested a MPM, characterized by the structure of the inverse correlation matrix of the latent variables. Their model building exercise was based on tests for conditional dependence on the latent scale while the interpretations were done on the observed binary scale. Hence if their analysis were done on the observed scale, quite different models could be obtained implying a quite different interpretation.

Finally, it is worth mentioning that a similar phenomenon will occur when the actual data are continuous but discretized for the sake of the analysis, a practice that is often seen in medical papers. For the same reason as pointed out above, markedly different conclusions might be drawn from the analysis on the continuous scale and the analysis on the discretized scale.

# Appendix

## Proof that conditional independence on the latent scale does not imply conditional independence on the binary scale

Let $\mathbf{V} \sim N_3(\mu, \mathbf{R})$, where,

$$\mu = \begin{pmatrix} \mu_{V_1} \\ \mu_{V_2} \\ \mu_{V_3} \end{pmatrix}, \qquad \mathbf{R} = \begin{pmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{12} & 1 & \rho_{23} \\ \rho_{13} & \rho_{23} & 1 \end{pmatrix}$$

When $V_1 \perp\!\!\!\perp V_2 | V_3$ holds, then e.g.,

$$\begin{aligned} P(V_1 > 0, V_2 > 0 | V_3 > 0) &= \frac{\int_0^\infty \int_0^\infty \int_0^\infty f(v_1|v_3)\ f(v_2|v_3)\ f(v_3)\ dv_1\ dv_2\ dv_3}{\int_0^\infty f(v_3)\ dv_3} \\ &= \frac{\int_0^\infty \Phi\left(\frac{\mu_1^*}{\sigma_1^*}\right) \Phi\left(\frac{\mu_2^*}{\sigma_2^*}\right) f(v_3)\ dv_3}{\int_0^\infty f(v_3)\ dv_3} \\ &= \int_{\mathcal{A}} \Phi\left(\frac{\mu_1^*}{\sigma_1^*}\right) \Phi\left(\frac{\mu_2^*}{\sigma_2^*}\right)\ h(x)\ dx \end{aligned} \tag{7}$$

where, $\mathcal{A} = [0, \infty)$, $\mu_1^* = \mu_{V_1} + \rho_{13}(x - \mu_X)$, $\mu_2^* = \mu_{V_2} + \rho_{23}(x - \mu_X)$, $\sigma_1^{2*} = \sqrt{1 - \rho_{Z_1,Z_3}^2}$, $\sigma_2^{2*} = \sqrt{1 - \rho_{23}^2}$, and $h(x) \equiv TN_{(0,\infty)}(\mu_X, \sigma_X^2)$, which means Truncated Normal between zero and infinity with location $\mu_X$ and scale $\sigma_X^2$.

On the other hand, when $Y_1 \perp\!\!\!\perp Y_2 | Y_3$ holds,

$$\begin{aligned} P(Y_1 = 1, Y_2 = 1 | Y_3 = 1) &= \mathrm{P}(Y_1 = 1 | Y_3 = 1)\mathrm{P}(Y_2 = 1 | Y_3 = 1) \\ &= \mathrm{P}(V_1 > 0 | V_3 > 0)\mathrm{P}(V_2 > 0 | V_3 > 0) \\ &= \int_{\mathcal{A}} \Phi\left(\frac{\mu_1^*}{\sigma_1^*}\right) h(x) dx \\ &\quad \int_{\mathcal{A}} \Phi\left(\frac{\mu_2^*}{\sigma_2^*}\right) h(x) dx \end{aligned} \tag{8}$$

Expression (7) is larger than expression (8), because

$$E(g_1(X)\, g_2(X)) \geq E(g_1(X))\, E(g_2(X)) \tag{9}$$

holds for all real-valued functions $g_1$ and $g_2$ which are nondecreasing (in each component) and are such that the expectations in (9) exist. The equality holds iff $g_1(X) = c$ or $g_2(X) = c$ (a.s.) (see, e.g., Esary, Proschan and Walkup 1967). This shows that conditional independence on the latent scale does not imply conditional independence on the observed scale and vice versa.

# Acknowledgements

The first two authors are supported by the Research Grant OT/00/35, Catholic University Leuven; data collection was supported by Unilever, Belgium. The Signal Tandmobiel® study comprises following partners: D. Declerck (Dental School, Catholic University Leuven), L. Martens (Dental School, University Ghent), J. Vanobbergen (Oral Health Promotion and Prevention, Flemish Dental Association), P. Bottenberg (Dental School, University Brussels), E. Lesaffre (Biostatistical Centre, Catholic University Leuven) and K. Hoppenbrouwers (Youth Health Department, Catholic University Leuven; Flemish Association for Youth Health Care). We also acknowledge support from the Interuniversity Attraction Poles Program P5/24 – Belgian State – Federal Office for Scientific, Technical and Cultural Affairs.

# References

Agresti, A.: 1990, *Categorical Data Analysis*, John Wiley, New York, USA.

Albert, J. H. and Chib, S.: 1993, Bayesian analysis of binary and polychotomous response data, *Journal of the American Statistical Association* **88**, 669–679.

Ashford, J. R. and Sowden, R. R.: 1970, Multi-variate probit analysis, *Biometrics* **26**, 535–546.

Chen, M. H. and Dey, D.: 1998, Bayesian modelling of correlated binary responses via scale mixture of multivariate normal link functions, *Sankhyā* **60**, 322–343.

Chen, M. H. and Schmeiser, B. W.: 1993, Performance of the gibbs, hit-and-run, and metropolis samplers, *Journal of Computational and Graphical Statistics* **2**, 251–272.

Chen, M. H. and Shao, Q. M.: 1999, Monte carlo estimation of bayesian credible and hpd intervals, *Journal of Computational and Graphical Statistics* **8 (1)**, 69–92.

Chib, S.: 2000, Bayesian methods for correlated binary data, *in* D. Dey, S. Ghosh and B. Mallick (eds), *Generalized linear models: A Bayesian perspective*, Marcel Dekker, pp. 113–131.

Chib, S. and Greenberg, E.: 1998, Analysis of multivariate probit models, *Biometrika* **85**, 347–361.

Cowles, M. and Carlin, B.: 1996, Markov chain monte carlo convergence diagnostics: a comparative study, *Journal of the American Statistical Association* **91**, 883–904.

Dale, J.: 1986, Global cross-ratio models for bivariate, discrete, ordered responses, *Biometrics* **42**, 909–917.

Dawid, A. P.: 1979, Conditional independence in statistical theory (with discussion), *Journal of the Royal Statistical Society, Series B* **41**, 1–31.

Esary, J. D., Proschan, F. and Walkup, D. W.: 1967, Association of random variables, with applications, *Annals of Mathematical Statistics* **38**, 1466–1474.

Genz, A.: 1992, Numerical computation of multivariate normal probabilities, *Journal of Computational and Graphical Statistics* **1**, 141–150.

Genz, A.: 1993, Comparison of methods for the computation of multivariate normal probabilities, *Computing Science and Statistics* **25**, 400–405.

Hujoel, P. P., Lamont, R. J., DeRouen, T. A., Davis, S. and Lerouxi, B. G.: 1994, Within-subject coronal caries distribution patterns: An evaluation of randomness with respect to the midline, *Journal of Dental Research* **73 (9)**, 1575–1580.

Joe, H. and Liu, Y.: 1996, A model for a multivariate binary response with covariates based on compatible conditionally specified logistic regressions, *Statistics and Probability Letters* **31**, 113–120.

Lesaffre, E. and Molenberghs, G.: 1991, Multivariate probit analysis: A neglected procedure in medical statistics, *Statistics in Medicine* **10**, 1391–1403.

Liang, K. Y. and Zeger, S. L.: 1986, Longitudinal data analysis using generalized linear models, *Biometrika* **73**, 13–22.

Molenberghs, G. and Lesaffre, E.: 1994, Marginal modeling of correlated ordinal data using a multivariate plackett distribution, *Journal of the American Statistical Association* **89**, 633–644.

Nogales, A. G., Oyola, J. A., and Pérez, P.: 2000, On conditional independence and the relationship between sufficiency and invariance under the bayesian point of view, *Statistics and Probability Letters* **46**, 75–84.

O'Brien, S. and Dunson, D.: 2004, Bayesian multivariate logistic regression, *Biometrics* **60**, 739–746.

Pearl, J.: 1995, Causal diagrams for empirical research (with discussion), *Biometrika* **82**, 669–710.

Psoter, W. J., Zhang, H., Pendrys, D. G., Morse, D. E. and Mayne, S. T.: 2003, Classification of dental caries patterns in the primary dentition: a multidimensional scaling analysis, *Community Dent Oral Epidemiol* **31**, 231–238.

R Development Core Team: 2004, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-00-3.
**URL:** *http://www.R-project.org*

Smith, B. J.: 2005, *Bayesian Output Analysis Program (BOA) for MCMC*, College of Public Health, University of Iowa, Iowa, USA.
**URL:** *http://www.public-health.uiowa.edu/boa*

Vanobbergen, J., Martens, L., Lesaffre, E. and Declerck, D.: 2000, The signal-tandmobiel project, a longitudinal intervention health promotion study in flanders (belgium): baseline and first year results, *European Journal of Paediatric Dentistry* **2**, 87–96.

Veerkamp, J. S. and Weerheijm, K. L.: 1995, Nursing-bottle caries: the importance of a developmentperspective, *J Dent Child* **62(6)**, 381–386.

Webb, E. L. and Forster, J. J.: 2004, Bayesian model determination for multivariate ordinal and binary data, *Technical report*, University of Southampton, School of Mathematics.