ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“

ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Том 104

ANNUAL OF SOFIA UNIVERSITY „ST. KLIMENT OHRIDSKI“

FACULTY OF MATHEMATICS AND INFORMATICS
Volume 104

EM ALGORITHM FOR MAXIMUM LIKELIHOOD ESTIMATION OF CORRELATED PROBIT MODEL FOR TWO LONGITUDINAL ORDINAL OUTCOMES

DENITSA GRIGOROVA

Correlated probit models (CPMs) are widely used for modeling of ordinal data or joint analyses of ordinal and continuous data which are common outcomes in medical studies. When we have clustered or longitudinal data CPMs with random effects are used to take into account the dependence between clustered measurements. When the dimension of the random effects is large, finding of the maximum likelihood estimates (MLEs) of the model parameters via standard numerical approximations is computationally cumbersome or in some cases impossible. EM algorithms for CPM for one ordinal longitudinal variable [13] and a joint CPM for one ordinal and one continuous longitudinal variable [14] are recently developed. ECM algorithm for ML estimation of the parameters of a joint CPM for two longitudinal ordinal variables will be presented. The algorithm is applied to estimation of CPM for the longitudinal ordinal outcomes self-rated health and categorized body mass index from the Health and Retirement Study (http://hrsonline.isr.umich.edu/, HRS). Results from fitting the model to the data and also results from some simulation studies will be reported.

Keywords: correlated probit model, EM algorithm, ordinal data, random effects.

2000 Math. Subject Classification: 62H12, 62J10, 62P10.

1. INTRODUCTION

Bliss [3, 4] and Gaduum [11] were the first to introduce the probit models for binary data. These models are also suitable for ordinal data as Aitchison and Silvey [1] proposed. The main characteristic of the probit models is the assumption of a latent normally distributed variable behind the observed ordinal outcome. The density of the latent variable is divided into as many pieces as the number of the levels of the ordinal response through thresholds. The area of each density piece represents the probability of observing the corresponding level of the ordinal outcome. The usefulness of the model is not affected when the existence of the latent variable does not seem natural.

Ashford and Sowden [2] introduced a multivariate extension of the probit model based on an underlying multivariate normal distribution. Ochi and Prentice [24] first introduced a correlated probit model but only for exchangeable binary data. Extensions of this model were proposed by Hedeker and Gibbons [12], Catalano [6], Grilli and Rampichini [15], Gueorguieva and Sanacora [17] among others. Gueorguieva [16] has a detailed overview on correlated probit models. Correlated probit models are widely used for modelling of multiple categorical variables or clustered/longitudinal ordinal outcomes for these models have two main advantages. They are easy for interpretation and they allow rich correlation structure of the latent variables via random effects and/or correlated errors. That allows to take into account the natural dependence of the measurements on the same subject or within cluster.

The correlated probit model does not have closed form expression for the likelihood function. Approximations need to be used in order to obtain estimates of the unknown parameters. There are several methods of statistical inference based on numerical, stochastic or analytical approximations. Most popular appear to be extensions of numerical approximations such as Gauss-Hermite quadrature ([10] pp. 306-307) or adaptive Gaussian Quadrature [19]. Another approach is based on analytical approximations (Breslow and Clayton [5], Wolfinger and O'Connell [29]) but it has been shown to produce bias in the parameter estimates especially for binary data or ordinal data with few categories. A third approach is the Expectation-Maximization (EM) algorithm [8]. An extension of the EM algorithm is the Expectation/Conditional Maximization (ECM) algorithm [23] which is used in cases of complicated M-step.

Ruud [26] is the first to apply the EM algorithm for the estimation of the parameters of probit models. Kawakatsu and Largey [18] extend Ruud's work to a joint model of a single ordinal and multivariate normal outcomes. Chan and Kuk [7] consider a correlated model for a clustered binary variable and propose an ECM algorithm for parameter estimation.

Our algorithm is a modification of the algorithm of Chan and Kuk [7] and Grigorova and Gueorguieva [13] to estimation of a joint model for two longitudinal ordinal outcomes by using the parameter transformation proposed by Ruud [26] for estimation of the threshold parameters.

We apply the model to data on 12543 individuals from the Health and Retirement Study (http://hrsonline.isr.umich.edu/, HRS). HRS is a longitudinal survey among American citizens born between 1931 and 1941 and their spouses that assesses changes in labor force participation and health status over the transition period from working to retirement and the years after. The launch of the study was in 1992 and data were collected at intervals of two years. The study provides a wealth of information to address important questions about aging. In our work the goal was to assess gender-related differences and the effects of smoking on measures of physical health in this representative sample of individuals over 50 years of age. We considered two repeatedly measured dependent variables: categorized body mass index (CBMI) and self-rated health (SRH). CBMI was selected because values different from normal weight might be predictive of a variety of health problems. CBMI is also easy to measure and is objective. CBMI has four levels: underweight (BMI<18.5, coded as 1), normal (18.5<BMI<25, coded as 2), overweight (25<BMI<30, coded as 3), obese (BMI>30, coded as 4). SRH is an ordinal measure that takes the following possible values: excellent (coded as 1), very good (2), good (3), fair (4) and poor (5). This is a more direct measure of health but is based on self-report and is more subjective. The two measures are expected to be positively correlated and joint modeling would allow to estimate this correlation cross-sectionally and over time and to test for overall effects of smoking and gender on these measures over time.

The paper is organized as follows. Section 2 defines the correlated probit model and outlines the estimation of the parameters and of their standard errors. Section 3 describes the simulation studies that were performed in order to examine the performance of the proposed algorithm. An application of the model to the data from the first seven waves of the HRS is included in Section 4. Section 5 contains concluding remarks and discussion about possible extensions of the algorithm.

2. MODEL

From now on bold typeset is used for vectors and matrices.

Let y^*_{1ij} is the measurement of the first ordinal variable with m_1 levels on the ith subject at time j and y^*_{2ij} is the observation on second ordinal outcome with m_2 levels on the same subject at the same time, $j = 1, \dots, n_i, i = 1, \dots, n$. We assume that there are two latent normal variables y_{1ij} and y_{2ij} that generated the observed ordinal variables. We consider the following random effects model:

$$\begin{aligned} y_{1ij} &= \boldsymbol{x'_{1ij}\beta_1} + \boldsymbol{z'_{1ij}b_{1i}} + \epsilon_{1ij}, \\ y_{2ij} &= \boldsymbol{x'_{2ij}\beta_2} + \boldsymbol{z'_{2ij}b_{2i}} + \epsilon_{2ij}. \end{aligned} \tag{2.1}$$

The relation between the observed ordinal variable and the latent normal variable is the following:

$$y^*_{kij} \quad = \quad \begin{cases} 1, \ y_{kij} \le \alpha_{k,1}; \\ l, \ \alpha_{k,l-1} < y_{kij} \le \alpha_{k,l}, \quad l = 2, \dots, m_k - 1; \\ m_k, \ y_{kij} > \alpha_{k,m_k-1}; \end{cases} \tag{2.2}$$

for some unknown thresholds $\alpha_{k,1}, \dots, \alpha_{k,m_k-1}, \quad k = 1, 2.$

We assume a normal distribution of the q-dimensional vector of the random effects $\boldsymbol{b_i} = (\boldsymbol{b'_{1i}}, \boldsymbol{b'_{2i}})' \sim N(\mathbf{0_q}, \boldsymbol{\Sigma})$. The covariance matrix $\boldsymbol{\Sigma}$ is a quadratic $q \times q$ positive semi-definite matrix:

$$\boldsymbol{\Sigma} = Var\begin{pmatrix} \boldsymbol{b_{1i}} \\ \boldsymbol{b_{2i}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Sigma_{11}} & \boldsymbol{\Sigma_{12}} \\ \boldsymbol{\Sigma_{21}} & \boldsymbol{\Sigma_{22}} \end{pmatrix}.$$

The error terms on the same subject at the same time are not necessarily assumed independent $(\epsilon_{1ij}, \epsilon_{2ij})' \sim N(\mathbf{0_2}, \boldsymbol{\Sigma_\epsilon})$, where

$$\boldsymbol{\Sigma_\epsilon} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}.$$

The error terms among individuals and different time points are assumed to be independent. We also assume that the random effects and the error terms are independent of each other.

The regression parameters for the fixed effects in model (2.1) are denoted by the vectors $\boldsymbol{\beta_k}, k = 1, 2$. The vectors of predictors for the fixed effects are $\boldsymbol{x_{kij}}, k = 1, 2$ and the predictors for the random effects are $\boldsymbol{z_{kij}}, k = 1, 2$.

From the observed data it is not possible to uniquely estimate all of the unknown parameters, so we pose the following restrictions: the first thresholds $\alpha_{k,1}, k = 1, 2$ are set to zero, the variance of the first normal error term σ_{11} is set to 1 and the variance of the second error term given the first error term is also 1, i.e. $\sigma_{2|1} = \sigma_{22} - \sigma_{12}^2/\sigma_{11} = 1$. Some other restrictions and reparameterisations are possible.

2.1. EM ALGORITHM FOR MLE

We propose an EM algorithm [8] for estimation of the unknown model parameters in (2.1) and thresholds in (2.2).

The EM algorithm is an iterative procedure for obtaining maximum likelihood estimates for models that depend on unobserved data. In our model the unobserved data are the latent variables and the random effects. Each iteration of the EM algorithm consists of two steps: E-step (Expectation step) and M-step (Maximisation step). Let us denote with $\mathbf{X}$ the observed data, with $\mathbf{Z}$ the unobserved data and with $\boldsymbol{\Gamma}$ the vector of unknown parameters of the model. The two steps at the $(k+1)$-st iteration of the algorithm are:

- E-step: $Q(\boldsymbol{\Gamma}|\boldsymbol{\Gamma}^{(\mathbf{k})}) = \mathrm{E}_{\mathbf{Z}|\mathbf{X},\boldsymbol{\Gamma}^{(\mathbf{k})}}[\ln L(\boldsymbol{\Gamma}; \mathbf{X}, \mathbf{Z})]$, where the 'complete data' likelihood function is $L(\boldsymbol{\Gamma}; \mathbf{X}, \mathbf{Z}) = p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\Gamma})$, where $p(.)$ is a density function,
- M-step: $\boldsymbol{\Gamma}^{(\mathbf{k+1})} = \arg\max_{\boldsymbol{\Gamma}} Q(\boldsymbol{\Gamma}|\boldsymbol{\Gamma}^{(\mathbf{k})})$.

The algorithm starts with initial values for the unknown parameters $\mathbf{\Gamma^{(0)}}$, iterates between the E-step and the M-step and stops when a converging criterion is met. Our choice for converging criterion is when $|\mathbf{\Gamma^{(k+1)}} - \mathbf{\Gamma^{(k)}}| < \epsilon$ for each element of the vector, where ϵ is a preselected small number.

The first difficulty in applying the EM algorithm to our model is the introduction of the thresholds in the complete data likelihood. We adopt the approach by Kawakatsu and Largey [18] who extend Ruud's work [26]. According to their method, we define the differences between consecutive thresholds with $\delta_{k,i} = \alpha_{k,i} - \alpha_{k,i-1},\ i = 2, \ldots, m_k - 1, k = 1, 2$ (we define additionally $\delta_{k,1} = \delta_{k,m_k} = 1$). It follows the connection $\alpha_{k,i} = \sum_{j=2}^{i} \delta_{k,j},\ k = 1, 2,\ i = 2, \ldots, m_k - 1$. Then we consider new variables, which are a linear transformation of the latent variables. The new variables are denoted by $y_{kij_{new}} = (y_{kij} - \alpha_{k,y^*_{kij}-1})/\delta_{k,y^*_{kij}}, k = 1, 2$, where $\alpha_{k,0} = 0, k = 1, 2$ and $\boldsymbol{y_{ki_{new}}} = (y_{ki1_{new}}, y_{ki2_{new}}, \ldots, y_{kin_{i_{new}}})'$.

Since the new variables are a linear transformation of the latent variables, they are also normally distributed. But given the observed ordinal variables, the transformed variables have truncated multivariate normal distribution with boundaries of truncation independent of the unknown threshold parameters. For example, if we observe the first level of y^*_{1ij}, the new variable $y_{1ij_{new}}$ is truncated at $(-\infty, 0]$, if y^*_{1ij} is between the first and the last level, the new variable is truncated at $(0, 1]$, and if we observe the last level of y^*_{1ij}, the new variable is truncated at $(0, \infty)$.

We use the approach by Chan and Kuk [7] in order to find closed form expressions for the unknown parameters $\mathbf{\Gamma} = (\boldsymbol{\beta_1'}, \boldsymbol{\beta_2'}, \boldsymbol{\delta_1'}, \boldsymbol{\delta_2'}, \boldsymbol{vect(\Sigma)'}, \lambda)'$, where $\boldsymbol{\delta_k} = (\delta_{k,2}, \ldots, \delta_{k,m_k-1}),\ k = 1, 2,\ \boldsymbol{vect(\Sigma)}$ is the vector of unique elements in the covariance matrix $\mathbf{\Sigma}$ and $\lambda = \sigma_{12}$.

2.1.1. COMPLETE DATA LOG-LIKELIHOOD

Complete data log-likelihood is the joint density of the transformed latent variables and the random effects. It has the following form:

$$\begin{aligned} \ln L &= \ln f(\boldsymbol{b}, \boldsymbol{y_{1_{new}}}, \boldsymbol{y_{2_{new}}}) = \sum_{i=1}^{n} \ln f(\boldsymbol{b_i}) f(\boldsymbol{y_{1i_{new}}}|\boldsymbol{b_i}) f(\boldsymbol{y_{2i_{new}}}|\boldsymbol{b_i}, \boldsymbol{y_{1i_{new}}}) \\ &= \sum_{i=1}^{n} \ln[f(\boldsymbol{b_i}) \prod_{j=1}^{n_i} f(y_{1ij_{new}}|\boldsymbol{b_i}) f(y_{2ij_{new}}|\boldsymbol{b_i}, y_{1ij_{new}})], \end{aligned}$$

where $f(.)$ denotes a normal density function.

From the model definition and the assumption for the distribution of the random effects it follows that apart from the constants the log-likelihood is:

$$\begin{aligned}\ln L = &- 0.5\sum_{i=1}^{n}\ln|\boldsymbol{\Sigma}| - 0.5\sum_{i=1}^{n}\boldsymbol{b_i'}\boldsymbol{\Sigma^{-1}}\boldsymbol{b_i} - 0.5\sum_{i=1}^{n}\sum_{j=1}^{n_i}\ln\sigma_{11} + \sum_{i=1}^{n}\sum_{j=1}^{n_i}\ln\delta_{1,y^*_{1ij}}\\ &- \frac{1}{2\sigma_{11}}\sum_{i=1}^{n}\sum_{j=1}^{n_i}(\delta_{1,y^*_{1ij}}y_{1ij_{new}} - \mu_{1ij_{new}})^2 - 0.5\sum_{i=1}^{n}\sum_{j=1}^{n_i}\ln\sigma_{2|1} + \sum_{i=1}^{n}\sum_{j=1}^{n_i}\ln\delta_{2,y^*_{2ij}}\\ &- \frac{1}{2\sigma_{2|1}}\sum_{i=1}^{n}\sum_{j=1}^{n_i}[\delta_{2,y^*_{2ij}}y_{2ij_{new}} - \mu_{2ij_{new}} - \lambda(\delta_{1,y^*_{1ij}}y_{1ij_{new}} - \mu_{1ij_{new}})]^2,\end{aligned}$$

where

$$\begin{aligned}\mu_{1ij_{new}} &= \boldsymbol{x'_{1ij}}\boldsymbol{\beta_1} + \boldsymbol{z'_{1ij}}\boldsymbol{b_{1i}} - \alpha_{1,y^*_{1ij}-1},\\ \mu_{2ij_{new}} &= \boldsymbol{x'_{2ij}}\boldsymbol{\beta_2} + \boldsymbol{z'_{2ij}}\boldsymbol{b_{2i}} - \alpha_{2,y^*_{2ij}-1}.\end{aligned}$$

2.1.2. CLOSED FORM EXPRESSIONS FOR THE ESTIMATORS

We obtain closed form expressions for the estimators of the unknown parameters by setting the first derivatives of the complete data log-likelihood to zero.

The estimator for the covariance matrix $\boldsymbol{\Sigma}$ of the random effects is:

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=0}^{n}\boldsymbol{b_i}\boldsymbol{b_i'}.$$

The regression parameters for the fixed effects for the first variable $\boldsymbol{\beta_1}$ satisfy the following system of equations:

$$\begin{aligned}(1+\lambda^2)\sum_{i=1}^{n}\sum_{j=1}^{n_i}\boldsymbol{x_{1ij}}\boldsymbol{x'_{1ij}}\boldsymbol{\beta_1} = &(1+\lambda^2)\sum_{i=1}^{n}\sum_{j=1}^{n_i}(\delta_{1,y^*_{1ij}}y_{1ij_{new}} - \boldsymbol{z'_{1ij}}\boldsymbol{b_{1i}} + \alpha_{1,y^*_{1ij}-1})\boldsymbol{x_{1ij}}\\ &-\lambda\sum_{i=1}^{n}\sum_{j=1}^{n_i}[\delta_{2,y^*_{2ij}}y_{2ij_{new}} - (\boldsymbol{x'_{2ij}}\boldsymbol{\beta_2} + \boldsymbol{z'_{2ij}}\boldsymbol{b_{2i}} - \alpha_{2,y^*_{2ij}-1})]\boldsymbol{x_{1ij}}.\end{aligned}$$

The regression parameters for the fixed effects for the second variable $\boldsymbol{\beta_2}$ satisfy the following system of equations:

$$\sum_{i=1}^{n}\sum_{j=1}^{n_i}\boldsymbol{x_{2ij}}\boldsymbol{x'_{2ij}}\boldsymbol{\beta_2} = \sum_{i=1}^{n}\sum_{j=1}^{n_i}[\delta_{2,y^*_{2ij}}y_{2ij_{new}} - \boldsymbol{z'_{2ij}}\boldsymbol{b_{2i}} + \alpha_{2,y^*_{2ij}-1} - \lambda(\delta_{1,y^*_{1ij}}y_{1ij_{new}} - \mu_{1ij_{new}})]\boldsymbol{x_{2ij}}.$$

It follows that the regression parameters $\boldsymbol{\beta_2}$ are a least square solution of the regression of $\tilde{y}_{2ij}$ on $\boldsymbol{x_{2ij}}$, where

$$\tilde{y}_{2ij} = \delta_{2,y^*_{2ij}}y_{2ij_{new}} - \boldsymbol{z'_{2ij}}\boldsymbol{b_{2i}} + \alpha_{2,y^*_{2ij}-1} - \lambda(\delta_{1,y^*_{1ij}}y_{1ij_{new}} - \mu_{1ij_{new}}).$$

The equation for λ is:

$$\lambda \sum_{i=1}^{n}\sum_{j=1}^{n_i}(\delta_{1,y^*_{1ij}} y_{1ij_{new}} - \mu_{1ij_{new}})^2$$
$$= \sum_{i=1}^{n}\sum_{j=1}^{n_i}(\delta_{2,y^*_{2ij}} y_{2ij_{new}} - \mu_{2ij_{new}})(\delta_{1,y^*_{1ij}} y_{1ij_{new}} - \mu_{1ij_{new}}).$$

The equations for $\delta_{1,l},\ l = 2, \ldots, m_1 - 1$, are quadratic equations of the form: $a_1\delta_{1,l}^2 + b_1\delta_{1,l} + c_1 = 0$, which always have real roots and the bigger root is always positive. The constants a_1, b_1, c_1 are given as follows:

$$\begin{aligned}
a_1 =&(1+\lambda^2)\sum_{i,j}\sum_{y^*_{1ij}=l}(y^2_{1ij_{new}}) + (1+\lambda^2)(n_{1,l+1} + n_{1,l+2} + \ldots + n_{1,m_1}),\\
b_1 =&-\sum_{i,j}\sum_{y^*_{1ij}=l} y_{1ij_{new}}[\mu_{1ij_{new}} + \lambda(\delta_{2,y^*_{2ij}} y_{2ij_{new}} - \mu_{2ij_{new}} + \lambda\mu_{1ij_{new}})]\\
&+\sum_{i,j}\sum_{y^*_{1ij}>l}[\delta_{1,y^*_{1ij}} y_{1ij_{new}} - (\boldsymbol{x'_{1ij}\beta_1} + \boldsymbol{z'_{1ij}b_{1i}} - \alpha_{1,y^*_{1ij}-1,-l})]\\
&-\sum_{i,j}\sum_{y^*_{1ij}>l}\{\lambda[\delta_{2,y^*_{2ij}} y_{2ij_{new}} - \mu_{2ij_{new}} - \lambda(\boldsymbol{x'_{1ij}\beta_1} + \boldsymbol{z'_{1ij}b_{1i}} - \alpha_{1,y^*_{1ij}-1,-l})]\},\\
c_1 =&-n_{1,l},
\end{aligned}$$

where $n_{1,l}$ is the number of the observations of the categorical variable y_1^* at l-th level and $\alpha_{1,y^*_{1ij}-1,-l} = \delta_{1,1} + \cdots + \delta_{1,l-1} + \delta_{1,l+1} + \cdots + \delta_{1,y^*_{1ij}-1}$.

Analogously, the equations for $\delta_{2,l},\ l = 2, \ldots, m_2 - 1$, are quadratic equations of the form: $a_2\delta_{2,l}^2 + b_2\delta_{2,l} + c_2 = 0$, which always have real roots and the bigger root is always positive. The constants a_2, b_2, c_2 are given as follows:

$$\begin{aligned}
a_2 =&\sum_{i,j}\sum_{y^*_{2ij}=l}(y^2_{2ij_{new}}) + n_{2,l+1} + n_{2,l+2} + \ldots + n_{2,m_2},\\
b_2 =&-\sum_{i,j}\sum_{y^*_{2ij}=l} y_{2ij_{new}}[\mu_{2ij_{new}} + \lambda(\delta_{1,y^*_{1ij}} y_{1ij_{new}} - \mu_{1ij_{new}})]\\
&+\sum_{i,j}\sum_{y^*_{2ij}>l}[\delta_{2,y^*_{2ij}} y_{2ij_{new}} - (\boldsymbol{x'_{2ij}\beta_2} + \boldsymbol{z'_{2ij}b_{2i}} - \alpha_{2,y^*_{2ij}-1,-l})\\
&\qquad - \lambda(\delta_{1,y^*_{1ij}} y_{1ij_{new}} - \mu_{1ij_{new}})],\\
c_2 =&-n_{2,l},
\end{aligned}$$

where $n_{2,l}$ is the number of the observations of the categorical variable y_2^* at l-th level and $\alpha_{2,y^*_{2ij}-1,-l} = \delta_{2,1} + \cdots + \delta_{2,l-1} + \delta_{2,l+1} + \cdots + \delta_{2,y^*_{2ij}-1}$.

In order to update the new estimates of the parameters, we need to express the conditional expectations in the closed form expressions for the estimators. We will show that all of the conditional expectations depend only on the first two moments of truncated multivariate normal distribution.

Let us introduce the following notation:

$$\boldsymbol{X_{ki}} = \begin{pmatrix} \boldsymbol{x'_{ki1}} \\ \boldsymbol{x'_{ki2}} \\ \vdots \\ \boldsymbol{x'_{kin_i}} \end{pmatrix}, \quad \boldsymbol{Z_{ki}} = \begin{pmatrix} \boldsymbol{z'_{ki1}} \\ \boldsymbol{z'_{ki2}} \\ \vdots \\ \boldsymbol{z'_{kin_i}} \end{pmatrix}, \quad \boldsymbol{\beta_k} = \begin{pmatrix} \beta_{k1} \\ \beta_{k2} \\ \vdots \\ \beta_{kp_k} \end{pmatrix}, \qquad k = 1,\, 2,$$

$$\boldsymbol{\alpha_{k,i}} = \begin{pmatrix} \alpha_{k,y^*_{ki1}-1} \\ \alpha_{k,y^*_{ki2}-1} \\ \vdots \\ \alpha_{k,y^*_{kin_i}-1} \end{pmatrix}, \qquad \boldsymbol{\delta^{-1}_{k,i}} = \begin{pmatrix} 1/\delta_{k,y^*_{ki1}} \\ 1/\delta_{k,y^*_{ki2}} \\ \vdots \\ 1/\delta_{k,y^*_{kin_i}} \end{pmatrix}, \qquad k = 1,\, 2,$$

$$\boldsymbol{\delta^{-1}_{i}} = \begin{pmatrix} \boldsymbol{\delta^{-1}_{1,i}} \\ \boldsymbol{\delta^{-1}_{2,i}} \end{pmatrix}, \quad \boldsymbol{\alpha_i} = \begin{pmatrix} \boldsymbol{\alpha_{1,i}} \\ \boldsymbol{\alpha_{2,i}} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta_1} \\ \boldsymbol{\beta_2} \end{pmatrix},$$

$$\boldsymbol{y_{i_{new}}} = \begin{pmatrix} \boldsymbol{y_{1i_{new}}} \\ \boldsymbol{y_{2i_{new}}} \end{pmatrix}, \quad \boldsymbol{Z_i} = \begin{pmatrix} \boldsymbol{Z_{1i}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{Z_{2i}} \end{pmatrix}, \quad \boldsymbol{X_i} = \begin{pmatrix} \boldsymbol{X_{1i}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{X_{2i}} \end{pmatrix}.$$

Then the joint distribution of $\boldsymbol{y_{1i_{new}}}$, $\boldsymbol{y_{2i_{new}}}$ and $\boldsymbol{b_i}$ is multivariate normal:

$$\begin{pmatrix} \boldsymbol{y_{1i_{new}}} \\ \boldsymbol{y_{2i_{new}}} \\ \boldsymbol{b_i} \end{pmatrix} \sim N \left[\begin{pmatrix} (\boldsymbol{X_{1i}\beta_1} - \boldsymbol{\alpha_{1,i}}) \circ \boldsymbol{\delta^{-1}_{1,i}} \\ (\boldsymbol{X_{2i}\beta_2} - \boldsymbol{\alpha_{2,i}}) \circ \boldsymbol{\delta^{-1}_{2,i}} \\ \boldsymbol{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} & \boldsymbol{C} \\ \boldsymbol{B'} & \boldsymbol{D} & \boldsymbol{E} \\ \boldsymbol{C'} & \boldsymbol{E'} & \boldsymbol{\Sigma} \end{pmatrix} \right],$$

where $\circ$ is the Hadamard (element-wise) product, the elements of the covariance matrix are:

$$\begin{aligned} \boldsymbol{A} &= (\boldsymbol{Z_{1i}\Sigma_{11}Z'_{1i}} + \sigma_{11}\boldsymbol{I_{n_i}}) \circ \boldsymbol{\delta^{-1}_{1,i}\delta^{-1'}_{1,i}}, \\ \boldsymbol{B} &= (\boldsymbol{Z_{1i}\Sigma_{12}Z'_{2i}} + \sigma_{12}\boldsymbol{I_{n_i}}) \circ \boldsymbol{\delta^{-1}_{1,i}\delta^{-1'}_{2,i}}, \\ \boldsymbol{C} &= \boldsymbol{Z_{1i}(\Sigma_{11}\Sigma_{12})} \circ \boldsymbol{J_{n_i\times q}\delta^{-1}_{1,i}}, \\ \boldsymbol{D} &= (\boldsymbol{Z_{2i}\Sigma_{22}Z'_{2i}} + \sigma_{22}\boldsymbol{I_{n_i}}) \circ \boldsymbol{\delta^{-1}_{2,i}\delta^{-1'}_{2,i}}, \\ \boldsymbol{E} &= \boldsymbol{Z_{2i}(\Sigma_{21}\Sigma_{22})} \circ \boldsymbol{J_{n_i\times q}\delta^{-1}_{2,i}}, \end{aligned}$$

$\boldsymbol{J_{n_i\times q}\delta^{-1}_{k,i}}$ is $n_i \times q$ matrix with columns $\boldsymbol{\delta^{-1}_{k,i}}$, $k = 1,\, 2$, and $\boldsymbol{I_{n_i}}$ is the $n_i \times n_i$ identity matrix.

Let us denote

$$\boldsymbol{M_i} = \begin{bmatrix} \boldsymbol{y_{1i_{new}}} - (\boldsymbol{X_{1i}\beta_1} - \boldsymbol{\alpha_{1,i}}) \circ \boldsymbol{\delta^{-1}_{1,i}} \\ \boldsymbol{y_{2i_{new}}} - (\boldsymbol{X_{2i}\beta_2} - \boldsymbol{\alpha_{2,i}}) \circ \boldsymbol{\delta^{-1}_{2,i}} \end{bmatrix} = \boldsymbol{y_{i_{new}}} - (\boldsymbol{X_i\beta} - \boldsymbol{\alpha_i}) \circ \boldsymbol{\delta^{-1}_i}$$

and

$$\boldsymbol{\Sigma_{B_i}} = \left(\begin{array}{cc} \boldsymbol{C'} & \boldsymbol{E'} \end{array} \right) \left(\begin{array}{cc} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{B'} & \boldsymbol{D} \end{array} \right)^{-1}.$$

Then the conditional distribution of $\boldsymbol{b_i}$ given $\boldsymbol{y_{1i_{new}}}$ and $\boldsymbol{y_{2i_{new}}}$ is again normal:

$$\boldsymbol{b_i}|\boldsymbol{y_{1i_{new}}}, \boldsymbol{y_{2i_{new}}} \sim N[\boldsymbol{\Sigma_{B_i} M_i}, \boldsymbol{\Sigma} - \boldsymbol{\Sigma_{B_i}} \left(\begin{array}{c} \boldsymbol{C} \\ \boldsymbol{E} \end{array} \right)] .$$

In the expressions for the estimators we have to calculate the following conditional expectations: $E(\boldsymbol{b_i}|\boldsymbol{y_i^*}), E(\boldsymbol{b_i b_i'}|\boldsymbol{y_i^*}), E(\boldsymbol{b_i y'_{i_{new}}}|\boldsymbol{y_i^*})$. We will show that they depend only on the first two moments of the distribution of the transformed latent variables given the observed variables, i.e. they depend on the first two moments of $\boldsymbol{y_{i_{new}}}|\boldsymbol{y_i^*}$, which distribution is truncated multivariate normal.

The expectation of the random effects given the observed variables is:

$$\begin{aligned} E(\boldsymbol{b_i}|\boldsymbol{y_i^*}) &= E[E(\boldsymbol{b_i}|\boldsymbol{y_{i_{new}}})|\boldsymbol{y_i^*}] \\ &= E[\boldsymbol{\Sigma_{B_i}}(\boldsymbol{y_{i_{new}}} - (\boldsymbol{X_i\beta} - \boldsymbol{\alpha_i}) \circ \boldsymbol{\delta_i^{-1}})|\boldsymbol{y_i^*}] \\ &= \boldsymbol{\Sigma_{B_i}}[E(\boldsymbol{y_{i_{new}}}|\boldsymbol{y_i^*}) - (\boldsymbol{X_i\beta} - \boldsymbol{\alpha_i}) \circ \boldsymbol{\delta_i^{-1}}]. \end{aligned}$$

The expectation of the second moment of the random effects given the observed variables is:

$$\begin{aligned} E(\boldsymbol{b_i b_i'}|\boldsymbol{y_i^*}) &= E[E(\boldsymbol{b_i b_i'}|\boldsymbol{y_{i_{new}}})|\boldsymbol{y_i^*}] \\ &= E[Var(\boldsymbol{b_i}|\boldsymbol{y_{i_{new}}}) + E(\boldsymbol{b_i}|\boldsymbol{y_{i_{new}}})E(\boldsymbol{b_i'}|\boldsymbol{y_{i_{new}}})|\boldsymbol{y_i^*}]. \end{aligned}$$

The last expectation that we need is:

$$\begin{aligned} E(\boldsymbol{b_i y'_{i_{new}}}|\boldsymbol{y_i^*}) &= E[E(\boldsymbol{b_i y'_{i_{new}}}|\boldsymbol{y_{i_{new}}})|\boldsymbol{y_i^*}] \\ &= E\{\boldsymbol{\Sigma_{B_i}}[\boldsymbol{y_{i_{new}}} - (\boldsymbol{X_i\beta} - \boldsymbol{\alpha_i}) \circ \boldsymbol{\delta_i^{-1}}]\boldsymbol{y'_{i_{new}}}|\boldsymbol{y_i^*}\} \\ &= \boldsymbol{\Sigma_{B_i}}[E(\boldsymbol{y_{i_{new}} y'_{i_{new}}}|\boldsymbol{y_i^*}) - (\boldsymbol{X_i\beta} - \boldsymbol{\alpha_i}) \circ \boldsymbol{\delta_i^{-1}} E(\boldsymbol{y'_{i_{new}}}|\boldsymbol{y_i^*})]. \end{aligned}$$

2.1.3. $(K+1)$-ST ITERATION OF THE EM ALGORITHM

We use an extension of the EM algorithm called Expectation/Conditional Maximization (ECM) algorithm [23]. The E-step at the $(k+1)$-st iteration of the proposed algorithm consists of finding of the following expectations: $E(\boldsymbol{b_i}|\boldsymbol{y_i^*}; \boldsymbol{\Gamma^k})$, $E(\boldsymbol{b_i b_i'}|\boldsymbol{y_i^*}; \boldsymbol{\Gamma^k})$, $E(\boldsymbol{b_i y'_{i_{new}}}|\boldsymbol{y_i^*}; \boldsymbol{\Gamma^k})$, where $\boldsymbol{\Gamma^k}$ are the k-th estimates of the unknown parameters $\boldsymbol{\Gamma}$. The M-step consists of several computationally simpler CM-steps. In each CM-step we maximise the expectation of the complete data log-likelihood function in respect to some parameters while the other parameters are held fixed. The algorithm starts with initial values for the uknown parameters $\boldsymbol{\Gamma^0}$, iterates between the E-step and M-step and stops when $|\boldsymbol{\Gamma^{k+1}} - \boldsymbol{\Gamma^k}| < \epsilon$ for each element of the vector, where ϵ is a preselected small number (for example $\epsilon = 0.0001$).

We use the bootstrap method for standard errors approximation described in [22, pp. 130–131]. The steps are as follows:

1. We fit model (2.1) to the observed data set consisting of n individuals using the proposed ECM algorithm and obtain the estimates of the unknown parameters $\hat{\mathbf{\Gamma}} = (\hat{\boldsymbol{\beta}}_1', \hat{\boldsymbol{\beta}}_2', \hat{\boldsymbol{\delta}}_1', \hat{\boldsymbol{\delta}}_2', \boldsymbol{vect}(\hat{\mathbf{\Sigma}})', \hat{\lambda})'$. To generate a bootstrap sample first we generate n random effects $\boldsymbol{b}_i^{\boldsymbol{b}}$ from $N(\mathbf{0}, \hat{\mathbf{\Sigma}})$, $i = 1, \ldots, n$. Next we simulate normal values $\boldsymbol{y}_{\mathbf{1}i}^{\boldsymbol{b}}$ and $\boldsymbol{y}_{\mathbf{2}i}^{\boldsymbol{b}}$ of dimension n_i according to the fitted model for every random effect $\boldsymbol{b}_i^{\boldsymbol{b}}$. We use the estimated via $\hat{\boldsymbol{\delta}}_{\mathbf{1}}$ and $\hat{\boldsymbol{\delta}}_{\mathbf{2}}$ thresholds to determine in which interval the normal data $\boldsymbol{y}_{\mathbf{1}i}^{\boldsymbol{b}}$ and $\boldsymbol{y}_{\mathbf{2}i}^{\boldsymbol{b}}$ fall and determine the levels of the bootstrap categorical variables $\boldsymbol{y}_{\mathbf{1}i}^{\boldsymbol{b}*}$ and $\boldsymbol{y}_{\mathbf{2}i}^{\boldsymbol{b}*}$. The bootstrap sample consists of the categorical variables $\boldsymbol{y}_{\mathbf{1}i}^{\boldsymbol{b}*}$ and $\boldsymbol{y}_{\mathbf{2}i}^{\boldsymbol{b}*}$, $i = 1, \ldots, n$.

2. We apply the ECM algorithm to the bootstrap data $\boldsymbol{y}_{\mathbf{1}i}^{\boldsymbol{b}*}$ and $\boldsymbol{y}_{\mathbf{2}i}^{\boldsymbol{b}*}$, $i = 1, \ldots, n$ to obtain estimates for the generated bootstrap data set $\mathbf{\Gamma}^{\boldsymbol{b}}$.

3. We use Monte Carlo method to approximate the bootstrap covariance matrix. That means that we repeat step 1 and step 2 B times and calculate the covariance matrix of the B estimated parameters $\mathbf{\Gamma}^{\boldsymbol{b}}$, $\ b = 1, \ldots, B$:

$$Cov(\hat{\mathbf{\Gamma}}) \approx \sum_{b=1}^{B} \frac{(\mathbf{\Gamma}^{\boldsymbol{b}} - \bar{\mathbf{\Gamma}})(\mathbf{\Gamma}^{\boldsymbol{b}} - \bar{\mathbf{\Gamma}})'}{B-1},$$

where $\bar{\mathbf{\Gamma}} = \sum_{b=1}^{B} \mathbf{\Gamma}^{\boldsymbol{b}}/B$.

3. SIMULATIONS

We simulated values from the following random intercept model:

$$\begin{aligned} y_{1ij} &= \beta_{10} + \beta_{11} t_{ij} + b_{1i} + \epsilon_{1ij}, \ j = 1, \ldots, 6, \\ y_{2ij} &= \beta_{20} + \beta_{21} t_{ij} + b_{2i} + \epsilon_{2ij}, \ j = 1, \ldots, 6, \end{aligned} \tag{3.1}$$

where $\beta_{10} = -0.5$, $\beta_{11} = 1$, $\beta_{20} = 1$, $\beta_{21} = -0.5$, $\alpha_{1,1} = 0$, $\alpha_{1,2} = 1.2$, $\alpha_{1,3} = 3$, $\alpha_{2,1} = 0$, $\alpha_{2,2} = 2$, $\lambda = 0.8$. The covariance matrices of errors $\mathbf{\Sigma}_{\boldsymbol{\epsilon}}$ and of the random effects $\mathbf{\Sigma}$ are:

$$\mathbf{\Sigma}_{\boldsymbol{\epsilon}} = \begin{pmatrix} \boldsymbol{\sigma}_{\mathbf{11}} & \boldsymbol{\sigma}_{\mathbf{12}} \\ \boldsymbol{\sigma}_{\mathbf{12}} & \boldsymbol{\sigma}_{\mathbf{22}} \end{pmatrix} = \begin{pmatrix} \mathbf{1} & \mathbf{0.8} \\ \mathbf{0.8} & \mathbf{1.64} \end{pmatrix},$$

$$\mathbf{\Sigma} = \boldsymbol{Var} \begin{pmatrix} \boldsymbol{b}_{\mathbf{1}i} \\ \boldsymbol{b}_{\mathbf{2}i} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\sigma}_{\mathbf{11}}^{\boldsymbol{b}} & \boldsymbol{\sigma}_{\mathbf{12}}^{\boldsymbol{b}} \\ \boldsymbol{\sigma}_{\mathbf{21}}^{\boldsymbol{b}} & \boldsymbol{\sigma}_{\mathbf{22}}^{\boldsymbol{b}} \end{pmatrix} = \begin{pmatrix} 1 & -0.8 \\ -0.8 & 1 \end{pmatrix}.$$

Table 1: Estimates and standard errors in the simulation model 3.1

parameter	β_{10}	β_{11}	β_{20}	β_{21}	$\delta_{1,2}$	$\delta_{1,3}$	$\delta_{2,2}$	λ	σ^b_{11}	σ^b_{12}	σ^b_{22}
	Sample size **1500**										
true value	-0.5	1	1	-0.5	1.2	1.8	2	0.8	1	-0.8	1
mean est.	-0.51	1	0.99	-0.50	1.20	1.80	1.99	0.79	1.00	-0.79	0.99
stand.dev. of estim.	0.041	0.010	0.047	0.012	0.018	0.024	0.039	0.026	0.047	0.039	0.055
mean of boot.st.er.	0.038	0.010	0.046	0.012	0.020	0.023	0.040	0.026	0.049	0.042	0.058
	Sample size **3000**										
mean est.	-0.51	1.00	1.01	-0.50	1.20	1.80	2.00	0.79	1.00	-0.79	0.98
stand.dev. of estim.	0.025	0.006	0.036	0.009	0.011	0.019	0.029	0.023	0.033	0.027	0.041
mean of boot.st.er.	0.027	0.007	0.032	0.008	0.013	0.016	0.028	0.019	0.034	0.029	0.040

We simulated 100 samples with two different sample sizes ($\boldsymbol{n} = \mathbf{1500}$ and $\boldsymbol{n} = \mathbf{3000}$). For each approximation of the standard errors we used 50 bootstrap samples which is within the recommended range of 50 to 100 bootstrap replications (Efron and Tibshirani [9]). The results are presented in Table 1.

Note that due to the re-parametrization we estimate the differences in the thresholds rather than the thresholds themselves, but they coincide in the case of only three levels of the categorical variables. In both simulation studies most of the averages of the estimated parameters are equal to the parameter values from which the samples were generated and where they differ the difference is smaller than $\mathbf{0.02}$.

As expected the standard errors get smaller when we increase the sample size. All of the estimates are statistically significantly different from zero.

The approximate equality of the standard deviations of the estimates and the means of the bootstrap standard errors confirms that the algorithm is converging as expected. However, larger simulation study that varies the parameter settings is necessary to confirm the above observations.

3.1. IMPLEMENTATION OF THE ALGORITHM

For the implementation of the algorithm we used the free software environment for statistical computing and graphics R [25]. The R code for fitting the presented models is available from the author.

We want to point out several things regarding the implementation of the proposed ECM algorithm. In the package **tmvtnorm** [28] there are functions for analytical finding of the first two moments of multivariate truncated normal distribution based on the work by Manjunath and Wilhelm [21]. There are also functions for generating random numbers using Gibbs sampling [27] which allows stochastic approximation of the first two moments of the truncated normal distribution. But for these models we recommend stochastic approximation because the analytical calculation could be very slow when we have many observations per subject.

A good choice for starting points for the regression parameters in model (2.1) and thresholds in (2.2) for the proposed ECM algorithm are estimates from model without random effects. Selecting large values as starting points for the variances of the random effects should be avoided. Problems with performance of the algorithm may occur with starting points corresponding to a multivariate truncated normal distribution for which the truncation area is close to 0. In such cases finding analytical solutions for the moments of the truncated normal distribution may fail. Generating random numbers via Gibbs sampling may also fail.

4. APPLICATION OF THE MODEL

We analyzed the first seven waves of HRS data with 12,543 individuals. We fitted the following correlated probit model to the data:

$$\begin{aligned}
\boldsymbol{y_{1ij}} =& \boldsymbol{\beta_{10} + \beta_{11}t_{ij} + \beta_{12}I(smoker) + \beta_{13}I(female)} \\
&\boldsymbol{+ \beta_{14}t_{ij}I(smoker) + \beta_{15}t_{ij}I(female)} \\
&\boldsymbol{+ \beta_{16}I(smoker)I(female) + \beta_{17}t_{ij}I(smoker)I(female)} \\
&\boldsymbol{+ b_{1i1} + b_{1i2}t_{ij} + \epsilon_{1ij}}, \\
\boldsymbol{y_{2ij}} =& \boldsymbol{\beta_{20} + \beta_{21}t_{ij} + \beta_{22}I(smoker) + \beta_{23}I(female)} \\
&\boldsymbol{+ \beta_{24}t_{ij}I(smoker) + \beta_{25}t_{ij}I(female)} \\
&\boldsymbol{+ \beta_{26}I(smoker)I(female) + \beta_{27}t_{ij}I(smoker)I(female)} \\
&\boldsymbol{+ b_{2i1} + b_{2i2}t_{ij} + \epsilon_{2ij}}, \\
\boldsymbol{y^*_{kij}} =& \begin{cases} \boldsymbol{1,\ y_{kij} \le \alpha_{k,1} = 0}, \\ \boldsymbol{l,\ \alpha_{k,l-1} < y_{kij} \le \alpha_{k,l},\ l = 2, \dots, m_k - 1} \\ \boldsymbol{m_k,\ y_{kij} > \alpha_{k,m_k-1}}, \end{cases}
\end{aligned} \tag{4.1}$$

where $\boldsymbol{k = 1, 2, m_1 = 5, m_2 = 4}$ and where the covariance matrix of the errors is: $\boldsymbol{\Sigma_\epsilon} = \begin{pmatrix} \boldsymbol{\sigma_{11}} & \boldsymbol{\sigma_{12}} \\ \boldsymbol{\sigma_{21}} & \boldsymbol{\sigma_{22}} \end{pmatrix}$ and the covariance matrix of the random effects is:

$$\boldsymbol{\Sigma = Var} \begin{pmatrix} \boldsymbol{b_{1i1}} \\ \boldsymbol{b_{1i2}} \\ \boldsymbol{b_{2i1}} \\ \boldsymbol{b_{2i2}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\sigma^b_{11}} & \boldsymbol{\sigma^b_{12}} & \boldsymbol{\sigma^b_{13}} & \boldsymbol{\sigma^b_{14}} \\ \boldsymbol{\sigma^b_{21}} & \boldsymbol{\sigma^b_{22}} & \boldsymbol{\sigma^b_{23}} & \boldsymbol{\sigma^b_{24}} \\ \boldsymbol{\sigma^b_{31}} & \boldsymbol{\sigma^b_{32}} & \boldsymbol{\sigma^b_{33}} & \boldsymbol{\sigma^b_{34}} \\ \boldsymbol{\sigma^b_{41}} & \boldsymbol{\sigma^b_{42}} & \boldsymbol{\sigma^b_{43}} & \boldsymbol{\sigma^b_{44}} \end{pmatrix}.$$

The estimates of the parameters, their standard errors and z-scores are presented in Table 2 and Table 3. Z-scores are computed before rounding off the estimates and their standard errors, and then rounded to the second decimal point in Table 2 and third decimal point in Table 3.

The results show that all of the parameters in the model are statistically significantly different from zero, except the regression parameters for the three-way interactions between time, smoking and gender in both sub-models ($\boldsymbol{\beta_{17}}$ and $\boldsymbol{\beta_{27}}$),

Table 2: Table of estimates, standard errors and z-scores of the regression parameters and threshold differences in model 4.1 fitted to the first seven waves of HRS data

Regression parameters for latent self-rated health								
parameter	$\boldsymbol{\beta_{10}}$	$\boldsymbol{\beta_{11}}$	$\boldsymbol{\beta_{12}}$	$\boldsymbol{\beta_{13}}$	$\boldsymbol{\beta_{14}}$	$\boldsymbol{\beta_{15}}$	$\boldsymbol{\beta_{16}}$	$\boldsymbol{\beta_{17}}$
estimate	1.37	0.13	0.64	0.06	0.03	-0.03	-0.26	0.01
stand. error	0.019	0.004	0.039	0.025	0.009	0.006	0.053	0.012
z-score	71.11	30.61	16.46	2.44	2.81	-4.73	-4.91	1.09
Regression parameters for latent categorized body mass index								
parameter	$\boldsymbol{\beta_{20}}$	$\boldsymbol{\beta_{21}}$	$\boldsymbol{\beta_{22}}$	$\boldsymbol{\beta_{23}}$	$\boldsymbol{\beta_{24}}$	$\boldsymbol{\beta_{25}}$	$\boldsymbol{\beta_{26}}$	$\boldsymbol{\beta_{27}}$
estimate	6.00	0.05	-0.88	-0.28	-0.01	0.05	-0.09	0.01
stand. error	0.053	0.006	0.038	0.030	0.010	0.007	0.061	0.015
z-score	113.4	8.94	-23.17	-9.36	-0.77	6.89	-1.53	0.53
Threshold parameters for both latent variables								
parameter	$\boldsymbol{\delta_{1,2}}$	$\boldsymbol{\delta_{1,3}}$	$\boldsymbol{\delta_{1,4}}$		$\boldsymbol{\delta_{2,2}}$	$\boldsymbol{\delta_{2,3}}$		
estimate	1.64	1.56	1.45		4.97	3.31		
stand. error	0.010	0.010	0.012		0.145	0.074		
z-score	172.18	164.15	120.11		34.28	44.91		

Table 3: Table of estimates, standard errors and z-scores of the covariance parameters in model 4.1 fitted to the first seven waves of HRS data

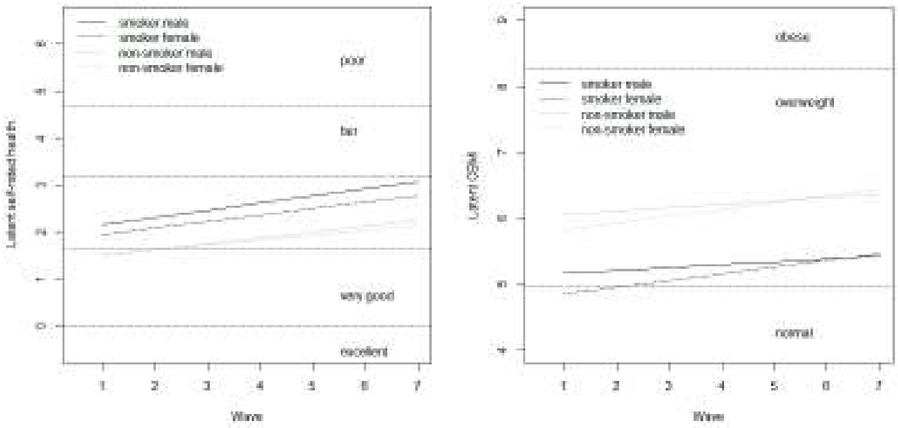
parameter	$\boldsymbol{\sigma^b_{11}}$	$\boldsymbol{\sigma^b_{22}}$	$\boldsymbol{\sigma^b_{33}}$	$\boldsymbol{\sigma^b_{44}}$	$\boldsymbol{\sigma^b_{12}}$	$\boldsymbol{\sigma^b_{13}}$	$\boldsymbol{\sigma^b_{14}}$
estimate	3.541	0.038	8.185	0.082	-0.194	1.308	-0.074
stand. error	0.065	0.001	0.311	0.005	0.006	0.071	0.007
z-score	54.872	47.948	26.294	17.507	-30.277	18.345	-10.151
parameter	$\boldsymbol{\sigma^b_{23}}$	$\boldsymbol{\sigma^b_{24}}$	$\boldsymbol{\sigma^b_{34}}$		$\boldsymbol{\lambda}$		
estimate	-0.020	0.002	-0.160		-0.002		
stand. error	0.008	0.001	0.020		0.007		
z-score	-2.619	2.638	-8.122		-0.267		

the regression parameter $\boldsymbol{\beta_{24}}$ for the two-way interaction between time and smoking and the regression parameter $\boldsymbol{\beta_{26}}$ for the two-way interaction between smoking and gender in the sub-model for the latent CBMI.

The correlation between the random intercept and random slope for the latent self-rated health is estimated to $\mathbf{-0.53}$ and for the BMI is estimated to $\mathbf{-0.20}$. The estimate for the correlation between the random intercepts is 0.24.

The trajectories of the latent variables for SRH and CBMI over time for four individuals with zero random effects are presented in Figure 1. They reflect the progress of the variables on average over time for each group of people: smoker male, smoker female, non-smoker male and non-smoker female. As expected, for the four groups of people on average the self-evaluation of health is worsening over time and people are gaining weight with time. The group with the most gentle slope for the self-rated health is the group of non-smoker female and with the steepest slope - smoker male, which means that, according to their own opinion on their own health, smoker males are worsening most quickly, while non-smoker females are worsening most slowly on average. For the BMI, smoker males are gaining weight most slowly and non-smoker females are gaining weight most quickly on average over time.

Figure 1: Latent SRH and latent CBMI over time for four individuals with zero random effects



5. DISCUSSION

In this paper we considered a correlated probit model for a joint analysis of two longitudinal ordinal outcomes. We proposed an extension of the EM algorithm of Chan and Kuk [7] and the ECM algorithm of Grigorova and Gueorguieva [13] for obtaining maximum likelihood estimates. The algorithm is implemented in the free software environment for statistical computing and graphics R [25]. We studied its performance via simulations. We illustrated the approach on the data from the Health and Retirement Study. Our approach has advantages over alternative estimation methods in that it can handle a large dimension of the multivariate outcome, it can be easily extended to any combination of binary, ordinal and continuous outcomes and it provides asymptotically unbiased estimates. It is also easily implemented in the free open-source software environment R.

We used bootstrap method for standard error estimation which is computationally very intensive. While the bootstrap algorithm can always be applied, it is not efficient. Other approaches may be possible. For example, one might consider the Louis's approximation method [20].

Further research is needed to extend the algorithm to the estimation of a joint model for time to drop-out and combinations of ordinal and continuous longitudinal outcomes. Model selection and model diagnostics are also open areas of research.

ACKNOWLEDGEMENT. This work is partially supported by the financial funds allocated to the Sofia University "St. Kl. Ohridski", Grants No. 013/2016 and No. 80-10-146/21.04.2017.

6. REFERENCES

[1] Aitchison, J., Silvey, S. D.: The generalization of probit analysis to the case of multiple responses. *Biometrika*, **44**, no. 1-2, 1957, 131–140.

[2] Ashford, J. R., Sowden, R. R.: Multi-variate probit analysis. *Biometrics*, **26**, no. 3, 1970, 535–546.

[3] Bliss, C. I.: The method of probits. *Science*, **79**, 2037, 1934, 38–39.

[4] Bliss, C. I.: The method of probits - a correction. *Science*, **79**, 2053, (1934), 409–410.

[5] Breslow, N. E., Clayton, D. G.: Approximate inference in generalized linear mixed models. *J. Amer. Stat. Assoc.*, **88**, 1993, 9–25.

[6] Catalano, P. J.: Bivariate modelling of clustered continuous and ordered categorical outcomes. *Statistics in Medicine*, **16**, no. 8, 1997, 883–900.

[7] Chan, J. S. K., Kuk, A. Y. C.: Maximum likelihood estimation for probit-linear mixed models with correlated random effects. *Biometrics*, **53**, 1997, 86–97.

[8] Dempster, A. P., Laird, N. M., Rubin, D. B.: Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Stat. Soc.*, Series B (Methodological), **39**, no. 2, 1977, 1–22.

[9] Efron, B., Tibshirani, R. J.: *An Introduction to the Bootstrap*, First edition, Monographs on Statistics & Applied Probability, **57**, Chapman & Hall, New York, 1994.

[10] Fahrmeir, L., Tutz, G.: *Multivariate Statistical Modelling Based on Generalized Linear Models*, Second edition, Springer-Verlag, New York, 2001.

[11] Gaddum, J. H.: Methods of biological assay depending on a quantal response. *Reports on biological standards*, III, 1933.

[12] Gibbons, R. D., Hedeker, D.: Application of random-effects probit regression models. *J. Consulting and Clinical Psychology*, **62**, no. 2, 1994, 285–296.

[13] Grigorova, D., Gueorguieva, R.: Implementation of the EM algorithm for maximum likelihood estimation of a random effects model for one longitudinal ordinal outcome. *Pliska Stud. Math. Bulgar.*, **22**, 2013, 41–56.

[14] Grigorova, D., Gueorguieva, R.: Correlated probit analysis of repeatedly measured ordinal and continuous outcomes with application to the Health and Retirement Study. *Statistics in Medicine*, **35**, no. 23, 2016, 4202–4225, sim. 6982.

[15] Grilli, L., Rampichini, C.: Alternative specifications of multivariate multilevel probit ordinal response models. *J. Educational and Behavioral Statistics*, **28**, 2003, 31–44.

[16] Gueorguieva, R. V.: Correlated probit model. In: *Encyclopedia of Biopharmaceutical Statistics*, 2006, Chapt. 59, pp. 355–362.

[17] Gueorguieva, R. V., Sanacora, G.: Joint analysis of repeatedly observed continuous and ordinal measures of disease severity. *Statistics in Medicine*, **25**, 2006, 1307–1322.

[18] Kawakatsu, H., Largey, A. G.: EM algorithms for ordered probit models with endogenous regressors. *Econometrics Journal*, **12**, 2009, 164–186.

[19] Liu, Q., Pierce, D. A.: A note on Gauss-Hermite quadrature. *Biometrika*, **81**, no. 3, 1994, 624–629.

[20] Louis, T. A.: Finding the observed information matrix when using the EM algorithm. *J. Royal Stat. Soc.*, Series B, **44**, no. 2, 1982, 226–233.

[21] Manjunath, B. G.,Wilhelm, S.: Moments calculation for the double truncated multivariate normal density. `http://ssrn.com/abstract=1472153`, September 11, 2009.

[22] McLachlan, G. J., Krishnan, T.: *The EM Algorithm and Extensions*, Wiley Series in Probability and Statistics, Second edition, Wiley-Interscience, 2008.

[23] Meng, X.-L., Rubin, D. B.: Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, **80**, no. 2, 1993, 267–278.

[24] Ochi, Y., Prentice, R. L.: Likelihood inference in a correlated probit regression model. *Biometrika*, **71**, no. 3, 1984, 531–543.

[25] R Core Team: R: *A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2013, ISBN 3-900051-07-0.

[26] Ruud, P. A.: Extensions of estimation methods using the EM algorithm. *Journal of Econometrics*, **49**, no. 3, 305–341.

[27] Wilhelm, S.: Gibbs sampler for the truncated multivariate normal distribution. `http://cran.r-project.org/web/packages/tmvtnorm/vignettes/GibbsSampler.pdf`, April 6, 2012.

[28] Wilhelm, S., Manjunath, B. G.: *tmvtnorm: Truncated Multivariate Normal and Student t Distribution*, 2012, R package version 1.4-7.

[29] Wolfinger, R., O'Connell, M.: Generalized linear mixed models: A pseudo-likelihood approach. *J. Stat. Comp. Sim.*, *48*, 1993, 233–243.

Received on October 17, 2017

Denitsa Grigorova

Faculty of Mathematics and Informatics
"St. Kl. Ohridski" University of Sofia
Department of Probability, Operations Research and Statistics
5, J. Bourchier blvd., BG-1164 Sofia
BULGARIA

E-mail: dgrigorova@fmi.uni-sofia.bg