PLISKA

STUDIA MATHEMATICA BULGARICA

ПЛИСКА

БЪЛГАРСКИ МАТЕМАТИЧЕСКИ СТУДИИ

Pliska Stud. Math. Bulgar. **22** (2013), 41–56

PLISKA
STUDIA MATHEMATICA
BULGARICA

# IMPLEMENTATION OF THE EM ALGORITHM FOR MAXIMUM LIKELIHOOD ESTIMATION OF A RANDOM EFFECTS MODEL FOR ONE LONGITUDINAL ORDINAL OUTCOME*

Denitsa Grigorova
Ralitza Gueorguieva

Abstract. Longitudinal data arise when we have repeated measures on subjects over time. The correlated probit model is frequently used for ordered longitudinal data since it allows to seamlessly incorporate different correlation structures. The estimation of the probit model parameters based on direct maximization of the limited information maximum likelihood is a numerically intensive procedure especially when we have repeated measures on subjects. We propose an extension of the EM algorithm for obtaining maximum likelihood estimates for one ordinal longitudinal outcome. The algorithm is implemented in the free software environment for statistical computing and graphics **R**. We use simulations to examine the performance of the developed algorithm and apply the model to data from the Health and Retirement Study (HRS). We apply a bootstrap approach for standard error approximation. Advantages of the presented algorithm include the potential of dealing with high-dimensional random effects and of extending the algorithm to combinations of ordinal and continuous longitudinal outcomes.

*The research was partially supported by appropriated state funds for research allocated to Sofia University (contract No 125/2012), Bulgaria

2010 *Mathematics Subject Classification:* 62J99.

*Key words:* correlated probit model, EM algorithm, free software environment for statistical computing and graphics **R**, ordinal longitudinal data.

**1. Introduction.** Longitudinal surveys follow up subjects over time. Modeling such data requires taking into account the correlation of measurements within subject. There are three main classes of models for longitudinal data: random effects models, transition models and marginal models. Marginal models are used when we are primarily interested in inferences for the population mean over time and the correlation structure is of secondary interest. Transition models are used when we are interested in modeling the response as a function of preceding outcomes and covariates. Random effects models are used when we are interested in inferences about individual change rather than average population change over time. Diggle et al. [10] describe the theoretical details of modeling longitudinal data while Weiss [29] provides a more applied overview.

Individuals often drop out of longitudinal studies. The mechanism of missingness needs to be taken into account when we model an incomplete data set. While marginal and transition models need to be extended when there are missing data, random effects models deal seamlessly with missing at random observations, that is, when the missingness depends only on observed outcomes and covariates. Another advantage of random effects models is that they simultaneously describe the mean structure and the correlation structure of the data.

In many longitudinal studies the variable of interest is ordinal. For example, in the Health and Retirement Study (HRS, http://hrsonline.isr.umich.edu/) self-rated health is categorical with five levels: excellent (coded as 1), very good (2), good (3), fair (4) and poor (5). The survey follows American citizens born between years 1931 and 1941 and their spouses over 14 years. There are seven waves of data collection at intervals of two years. At each interview the participants provided information about their self-rated health and multiple other variables. Our interest is in modeling how self-assessment of health varies over time. We use the correlated probit model.

Probit models were first proposed by Gaddum [14] and Bliss [3, 4] for binary data. Ashford and Sowden [2] introduced a multivariate extension of the probit model based on an underlying multivariate normal distribution. Aitchison and Silvey [1] proposed a probit model for ordinal data. Ochi and Prentice [27] first introduced a correlated probit model but only for exchangeable binary data. Extensions of this model were proposed by Hedeker and Gibbons [15], Catalano [7], Grilli and Rampichini [16], Gueorguieva and Sanacora [19] among others. The correlated probit model has been extensively used because it is easy to interpret and allows different correlation structures within subject. However, computational problems have always been a challenge and the issue of a general approach

to parameter estimation is still open. Gueorguieva [17] has a detailed overview on correlated probit models.

In the present article we consider a correlated probit model which is suitable for ordinal longitudinal data. A special feature of the model is the assumption of latent normal variables with thresholds that generate the observed ordinal responses. The latent variables can be interpreted as unobserved continuous measures that generate the observed responses. For example, in HRS we may consider health to be an underlying continuous measure. We can not directly observe it, rather we know only the level (categorical variable) at which the subjects rate their health at a particular time point. Conceptually, the unobserved continuous measure depends on measured covariates (both time-independent or time-dependent) via fixed effects and unmeasured covariates via random effects.

The correlated probit model does not have closed form expression for the likelihood function and hence approximations need to be used. 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 [13] pp. $306-307$ or addaptive Gaussian Quadrature [22]. However, these approaches become too computationally intensive and are not feasible for models with many random effects. Another approach is based on analytical approximations (Breslow and Clayton [5], Wolfinger and O'Connell [31]) but it has been shown to produce bias in the parameter estimates especially for binary data or ordinal data with few categories. Stochastic approximations appear most suitable for correlated probit models for longitudinal data since the computational complexity does not increase exponentially with the increase of number of random effects and they provide unbiased results as the number of generated samples increases.

We consider a stochastic extension of the Expectation Maximization (EM) algorithm [9] or more precisely of the Expectation Conditional Maximization (ECM) algorithm [26]. Ruud [28] is the first to apply the EM algorithm for the estimation of the parameters of probit models. Kawakatsu and Largey [20] extend Ruud's work to a joint model of a single ordinal and multivariate normal outcomes. Chan and Kuk [8] consider a correlated model for a clustered binary variable and propose an ECM algorithm for parameter estimation. Gueorguieva and Agresti [18] extend their approach to correlated binary and continuous outcomes. Our algorithm extends the approach of Chan and Kuk [8] and Gueorguieva and Agresti [18] to ordinal data by using the parameter transformation proposed by Kawakatsu and Largey [20] for estimation of the threshold parameters.

Since the EM algorithm does not provide direct estimates for the standard errors of the parameters, we use bootstrap for standard error estimation. Bootstrap methods were first introduced by Efron [11] and are resampling methods. They are very useful when the theoretical distribution of a statistic of interest is complicated or unknown. Bootstrap methods can be applied to a broad class of problems (e.g. standard error estimation, hypothesis testing, confidence intervals construction).

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 algorithm. An application of the model to the HRS data is included in Section 4. Section 5 contains concluding remarks and discussion about possible extensions of the algorithm.

**2. Correlated probit model for ordinal longitudinal data.** Let $y^*_{ij}$ denote the observed ordinal variable with $m$ levels on the $i$th subject at time $j$. We assume that there is a latent normal variable $y_{ij}$ that generated the observed variable. We consider the following random effects model:

$$y_{ij} = \boldsymbol{x'_{ij}}\boldsymbol{\beta} + \boldsymbol{z'_{ij}}\boldsymbol{b_i} + \epsilon_{ij}. \tag{1}$$

The rule that relates the latent variable to the observed ordinal variable is:

$$y^*_{ij} = \begin{cases} 1, \ y_{ij} \leq \alpha_1; \\ j, \ \alpha_{j-1} < y_{ij} \leq \alpha_j, \ j = 2, \ldots, m-1; \\ m, \ y_{ij} > \alpha_{m-1}; \end{cases} \tag{2}$$

for some unknown thresholds $\alpha_1, \ldots, \alpha_{m-1}$.

The vector of random effects is assumed to be normally distributed $q$-dimensional and is denoted by $\boldsymbol{b_i} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$. The error term is normally distributed $\epsilon_{ij} \sim N(0, \sigma^2)$ and is independent of the random effects.

The regression parameters for the fixed effects in model (1) are denoted by the $p$-dimensional vector $\boldsymbol{\beta}$. The vector of predictors for the fixed effects is $\boldsymbol{x_{ij}}$ and the vector of predictors for the random effects is $\boldsymbol{z_{ij}}$. The covariance matrix $\boldsymbol{\Sigma}$ is a quadratic $q \times q$ positive semi-definite matrix.

From the observed data we can not estimate all of the unknown parameters, so we impose the following identifiability restrictions: the first threshold $\alpha_1$ is set to zero and the variance of the normal error term $\sigma^2$ is set to 1. Other restrictions and re-parameterizations are possible.

The correlated model formulation is very appealing because the underlying normal distribution allows for very rich correlation structure. Also, the model has intuitive interpretation and can be easily extended to multiple ordinal and continuous outcomes. Furthermore, estimating the parameters using maximum likelihood allows the use of all results concerning maximum likelihood estimates. Hence Wald, score and likelihood ratio tests can be used for hypothesis testing and confidence interval construction. Likelihood ratio tests can also be used to compare nested models while information criteria such as Akaike Information Criterion or Schwartz-Bayesian Criterion can be used for model selection.

**2.1. Maximum likelihood estimation via the EM algorithm.** We extend the stochastic ECM algorithm of Chan and Kuk [8] to estimate the unknown parameters in model (1). The first step is to re-parameterize the thresholds so that they can be explicitly included in the complete data log-likelihood. For this we use the approach of Kawakatsu and Largey [20]. We define the differences between consecutive thresholds with $\delta_i = \alpha_i - \alpha_{i-1},\ i = 2, \ldots, m-1$. We also define $\delta_1 = \delta_m = 1$ for completeness and future use. Then we consider a new variable which is a linear transformation of the latent variable: $y_{ij_{new}} = (y_{ij} - \alpha_{y^*_{ij}-1})/\delta_{y^*_{ij}}$. For completeness we denote $\alpha_0 = 0$. For example, if $y^*_{ij} = u,\ u = 1, \ldots, m$ then $y_{ij_{new}} = (y_{ij} - \alpha_{u-1})/\delta_u$. Because the new variable is a linear transformation of the latent variable then it is also normally distributed. However, conditional on the observed categorical variable, it is a truncated normal variable. If we observe the first level of $y^*$ the new variable is truncated at $(-\infty, 0]$, if $y^*$ is between the first and the last level the new variable is truncated at $(0, 1]$, and if we observe the last level of $y^*$ the new variable is truncated at $(0, \infty)$.

***2.1.1. Complete data log-likelihood.*** Complete data log-likelihood $\ln L$ is:

$$\ln L = \ln f(\boldsymbol{b}, \boldsymbol{y_{new}}) = \sum_{i=1}^{n} \ln f(\boldsymbol{b_i}) f(\boldsymbol{y_{i_{new}}} | \boldsymbol{b_i}).$$

Apart from the constants the log likelihood has the following closed form:

$$\begin{aligned} \ln L \quad = \quad & -0.5 \sum_{i=1}^{n} \ln |\boldsymbol{\Sigma}| - 0.5 \sum_{i=1}^{n} \boldsymbol{b_i'} \boldsymbol{\Sigma^{-1}} \boldsymbol{b_i} + \\ & + \sum_{i=1}^{n} \sum_{j=1}^{n_i} \ln \delta_{y^*_{ij}} - 0.5 \sum_{i=1}^{n} \sum_{j=1}^{n_i} [\delta_{y^*_{ij}} y_{ij_{new}} - (\boldsymbol{x'_{ij}} \boldsymbol{\beta} + \boldsymbol{z'_{ij}} \boldsymbol{b_i} - \alpha_{y^*_{ij}-1})]^2. \end{aligned}$$

Thus we obtain closed form expressions for the estimators of the unknown parameters $\mathbf{\Gamma} = (\boldsymbol{\beta}, \mathbf{\Sigma}, \delta_2, \ldots, \delta_{m-1})$ by setting the first derivatives of the complete data log-likelihood to zero.

***2.1.2. Closed form expressions for the estimators.*** The estimator for the covariance matrix $\mathbf{\Sigma}$ of the random effects is:

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

Regression parameters for the fixed effects satisfy the following equation:

$$\sum_{i=1}^{n}\sum_{j=1}^{n_i} \boldsymbol{x_{ij}}\boldsymbol{x_{ij}'}\boldsymbol{\beta} = \sum_{i=1}^{n}\sum_{j=1}^{n_i} [\delta_{y_{ij}^*} y_{ij_{new}} - \boldsymbol{z_{ij}'}\boldsymbol{b_i} + \alpha_{y_{ij}^*-1})]\boldsymbol{x_{ij}}.$$

It follows that the regression parameters $\boldsymbol{\beta}$ are a least squares solution of the regression of $\tilde{y}_{ij}$ on $\boldsymbol{x_{ij}}$, where $\tilde{y}_{ij} = \delta_{y_{ij}^*} y_{ij_{new}} - \boldsymbol{z_{ij}'}\boldsymbol{b_i} + \alpha_{y_{ij}^*-1}$.

The equations for $\delta_k,\ k = 2, \ldots, m-1$ are quadratic equations of the form $a\delta_k^2 + b\delta_k + c = 0$ which always have real roots and the bigger root is always positive. The constants $a, b, c$ are as follows:

$$\begin{aligned}
a &= \sum_{i,j}\sum_{y_{ij}^*=k} (y_{ij_{new}}^2) + n_{k+1} + \ldots + n_m,\\
b &= -\sum_{i,j}\sum_{y_{ij}^*=k} y_{ij_{new}}(\boldsymbol{x_{ij}'}\boldsymbol{\beta} + \boldsymbol{z_{ij}'}\boldsymbol{b_i} - \alpha_{k-1}) +\\
&\quad \sum_{i,j}\sum_{y_{ij}^*>k} (\delta_{y_{ij}^*} y_{ij_{new}} - \boldsymbol{x_{ij}'}\boldsymbol{\beta} - \boldsymbol{z_{ij}'}\boldsymbol{b_i} + \delta_1 + \ldots + \delta_{k-1} + \delta_{k+1} + \ldots + \delta_{y_{ij}^*-1}),\\
c &= -n_k,
\end{aligned}$$

where $n_k$ is the number of observations categorical variable at $k$-th level.

All we need to do in order to update the parameter estimates at each step of the algorithm is to find the conditional expectations in the closed form expressions of the estimators. We will show that all of the conditional expectations depend only on the first two moments of the truncated multivariate normal distribution.

***2.1.3. Conditional expectations.*** Let us introduce the following notation:

$$\boldsymbol{X_i} = \begin{pmatrix} \boldsymbol{x'_{i1}} \\ \boldsymbol{x'_{i2}} \\ \vdots \\ \boldsymbol{x'_{in_i}} \end{pmatrix}, \boldsymbol{Z_i} = \begin{pmatrix} \boldsymbol{z'_{i1}} \\ \boldsymbol{z'_{i2}} \\ \vdots \\ \boldsymbol{z'_{in_i}} \end{pmatrix}, \boldsymbol{\alpha_i} = \begin{pmatrix} \alpha_{y^*_{i1}-1} \\ \alpha_{y^*_{i2}-1} \\ \vdots \\ \alpha_{y^*_{in_i}-1} \end{pmatrix}, \boldsymbol{\delta_i^{-1}} = \begin{pmatrix} 1/\delta_{y_{i1}} \\ 1/\delta_{y_{i2}} \\ \vdots \\ 1/\delta_{y_{in_i}} \end{pmatrix}.$$

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

$$\begin{pmatrix} \boldsymbol{y_{i_{new}}} \\ \boldsymbol{b_i} \end{pmatrix} \sim N\left[\begin{pmatrix} \boldsymbol{(X_i\beta - \alpha_i) \circ \delta_i^{-1}} \\ \boldsymbol{0} \end{pmatrix}, \boldsymbol{V}\right],$$

where $\boldsymbol{J_{n_i \times q}\delta_i^{-1}}$ is a $n_i \times q$ matrix with columns $\boldsymbol{\delta_i^{-1}}$ and $\circ$ is the Hadamard (elementwise) product and

$$\boldsymbol{V} = \begin{pmatrix} \boldsymbol{(Z_i\Sigma Z'_i + I_{n_i}) \circ \delta_i^{-1}\delta_i^{-1'}} & \boldsymbol{Z_i\Sigma \circ (J_{n_i \times q}\delta_i^{-1})} \\ \boldsymbol{\Sigma Z'_i \circ (J_{n_i \times q}\delta_i^{-1})'} & \boldsymbol{\Sigma} \end{pmatrix}.$$

Let us denote $\boldsymbol{\Sigma_{B_i} = (\Sigma Z'_i \circ (J_{n_i \times q}\delta_i^{-1})')[(Z_i\Sigma Z'_i + I_{n_i}) \circ \delta_i^{-1}\delta_i^{-1'}]^{-1}}$.
Then the conditional distribution of $\boldsymbol{b_i}$ given $\boldsymbol{y_{i_{new}}}$ is again normal:

$$\boldsymbol{b_i|y_{i_{new}}} \sim N[\boldsymbol{\Sigma_{B_i}(y_{i_{new}} - (X_i\beta - \alpha_i) \circ \delta_i^{-1}), \Sigma - \Sigma_{B_i}(Z_i\Sigma \circ (J_{n_i \times q}\delta_i^{-1}))}].$$

In the expressions for the estimators we have to find the following conditional expectations: $E(\boldsymbol{b_i|y_i^*}), E(\boldsymbol{b_ib'_i|y_i^*}), E(y_{ij_{new}}\boldsymbol{b_i|y_i^*})$. We will show that they depend only on the first two moments of $\boldsymbol{y_{i_{new}}|y_i^*}$.

The expectation of the random effects conditional on the observed variable is:

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

Let us denote $\boldsymbol{M_i = y_{i_{new}} - (X_i\beta - \alpha_i) \circ \delta_i^{-1}}$. For the estimator of the covariance matrix of the random effects we need:

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

$$
\begin{aligned}
&= \boldsymbol{\Sigma} - \boldsymbol{\Sigma}_{\boldsymbol{B_i}}(\boldsymbol{Z_i}\boldsymbol{\Sigma} \circ (\boldsymbol{J}_{\boldsymbol{n_i}\times\boldsymbol{q}}\boldsymbol{\delta}_{\boldsymbol{i}}^{-\mathbf{1}})) + \boldsymbol{\Sigma}_{\boldsymbol{B_i}} E[\boldsymbol{M_i}\boldsymbol{M_i'}|\boldsymbol{y_i^*}]\boldsymbol{\Sigma}'_{\boldsymbol{B_i}} \\
&= \boldsymbol{\Sigma} - \boldsymbol{\Sigma}_{\boldsymbol{B_i}}(\boldsymbol{Z_i}\boldsymbol{\Sigma} \circ (\boldsymbol{J}_{\boldsymbol{n_i}\times\boldsymbol{q}}\boldsymbol{\delta}_{\boldsymbol{i}}^{-\mathbf{1}})) + \\
&\quad \boldsymbol{\Sigma}_{\boldsymbol{B_i}}[Var(\boldsymbol{y}_{\boldsymbol{i_{new}}}|\boldsymbol{y_i^*}) + E(\boldsymbol{y}_{\boldsymbol{i_{new}}}|\boldsymbol{y_i^*})E(\boldsymbol{y'}_{\boldsymbol{i_{new}}}|\boldsymbol{y_i^*}) \\
&\qquad -E(\boldsymbol{y}_{\boldsymbol{i_{new}}}|\boldsymbol{y_i^*})[(\boldsymbol{X_i}\boldsymbol{\beta} - \boldsymbol{\alpha_i}) \circ \boldsymbol{\delta}_{\boldsymbol{i}}^{-\mathbf{1}}]' \\
&\qquad -[(\boldsymbol{X_i}\boldsymbol{\beta} - \boldsymbol{\alpha_i}) \circ \boldsymbol{\delta}_{\boldsymbol{i}}^{-\mathbf{1}}]E(\boldsymbol{y'}_{\boldsymbol{i_{new}}}|\boldsymbol{y_i^*}) \\
&\qquad +[(\boldsymbol{X_i}\boldsymbol{\beta} - \boldsymbol{\alpha_i}) \circ \boldsymbol{\delta}_{\boldsymbol{i}}^{-\mathbf{1}}][(\boldsymbol{X_i}\boldsymbol{\beta} - \boldsymbol{\alpha_i}) \circ \boldsymbol{\delta}_{\boldsymbol{i}}^{-\mathbf{1}}]']\boldsymbol{\Sigma}'_{\boldsymbol{B_i}}.
\end{aligned}
$$

In the expression for the conditional expectation of $b$ we need:

$$
\begin{aligned}
E(y_{ij_{new}}\boldsymbol{b_i}|\boldsymbol{y_i^*}) &= E[E(y_{ij_{new}}\boldsymbol{b_i}|\boldsymbol{y}_{\boldsymbol{i_{new}}})|\boldsymbol{y_i^*}] \\
&= E[y_{ij_{new}}\boldsymbol{\Sigma}_{\boldsymbol{B_i}}(\boldsymbol{y}_{\boldsymbol{i_{new}}} - (\boldsymbol{X_i}\boldsymbol{\beta} - \boldsymbol{\alpha_i}) \circ \boldsymbol{\delta}_{\boldsymbol{i}}^{-\mathbf{1}})|\boldsymbol{y_i^*}] \\
&= \boldsymbol{\Sigma}_{\boldsymbol{B_i}} E[y_{ij_{new}}\boldsymbol{y}_{\boldsymbol{i_{new}}} - y_{ij_{new}}(\boldsymbol{X_i}\boldsymbol{\beta} - \boldsymbol{\alpha_i}) \circ \boldsymbol{\delta}_{\boldsymbol{i}}^{-\mathbf{1}}|\boldsymbol{y_i^*}] \\
&= \boldsymbol{\Sigma}_{\boldsymbol{B_i}}[Cov(y_{ij_{new}}\boldsymbol{y}_{\boldsymbol{i_{new}}}|\boldsymbol{y_i^*}) + E(y_{ij_{new}}|\boldsymbol{y_i^*})E(\boldsymbol{y}_{\boldsymbol{i_{new}}}|\boldsymbol{y_i^*}) \\
&\qquad -E(y_{ij_{new}}|\boldsymbol{y_i^*})(\boldsymbol{X_i}\boldsymbol{\beta} - \boldsymbol{\alpha_i}) \circ \boldsymbol{\delta}_{\boldsymbol{i}}^{-\mathbf{1}}].
\end{aligned}
$$

The conditional expectations above are available in closed forms [24] but their calculation is computationally very intensive and proved to be inefficient especially when the dimension of the truncated multivariate distribution $d$ is bigger than 2. We describe a stochastic approximation of the conditional expectations which has good practical properties. In order to find the above conditional expectations which we showed to depend only on the first two moments of $\boldsymbol{y}_{\boldsymbol{i_{new}}}|\boldsymbol{y_i^*}$ we use a Monte Carlo method. We generate values from the truncated normal distribution given observed data using Gibbs sampling with the help of the **rtmvnorm** function in the **R** package **tmvtnorm**. The algorithm for the generation of random numbers is described in detail in [30]. We use the sample mean and the sample variance of simulated values to approximate $E(\boldsymbol{y}_{\boldsymbol{i_{new}}}|\boldsymbol{y_i^*})$ and $Var(\boldsymbol{y}_{\boldsymbol{i_{new}}}|\boldsymbol{y_i^*})$. In mathematical notation this is expressed as:

$$
\hat{E}(\boldsymbol{y}_{\boldsymbol{i_{new}}}|\boldsymbol{y_i^*}) = \frac{1}{m}\sum_{k=1}^{m}\boldsymbol{y}_{\boldsymbol{i_{new}}}^{(\boldsymbol{k})},
$$

$$\hat{Var}(\boldsymbol{y_{i_{new}}}|\boldsymbol{y_i^*}) = \frac{1}{m-1}\sum_{k=1}^{m}(\boldsymbol{y_{i_{new}}^{(k)}} - \hat{\boldsymbol{E}}(\boldsymbol{y_{i_{new}}}|\boldsymbol{y_i^*}))(\boldsymbol{y_{i_{new}}^{(k)}} - \hat{\boldsymbol{E}}(\boldsymbol{y_{i_{new}}}|\boldsymbol{y_i^*}))',$$

where $\boldsymbol{y_{i_{new}}^{(k)}}$ is the $k-$th realisation of $\boldsymbol{y_{i_{new}}}|\boldsymbol{y_i^*}$ in the generated sample of $m$ random numbers. Our experience shows that even for small $m$ (e.g. 150 or 200) we get adequate results.

***2.1.4. $(k+1)$-st iteration of the ECM algorithm.*** The estimates of the unknown parameters at the $k+1-$st step of the proposed ECM algorithm are updated as follows:

- The $(k+1)$-st estimates of the regression parameters $\boldsymbol{\beta^{k+1}}$ are the least square solution of the regression of $E(\tilde{y}_{ij}|\boldsymbol{y_i^*;\Gamma^k})$ on $\boldsymbol{x_{ij}}$.
- The $(k+1)$-st estimate of $\delta_u,\ u=2,\ldots,m-1$ is: $\delta_u^{k+1} = (-E[b|\boldsymbol{y^*;\Gamma^k}] + \sqrt{(E[b|\boldsymbol{y^*;\Gamma^k}]^2 - 4E[a|\boldsymbol{y^*;\Gamma^k}]E[c|\boldsymbol{y^*;\Gamma^k}])})/2E[a|\boldsymbol{y^*;\Gamma^k}]$. In the expression for the expectations of $a,b,c$ we use the already updated estimates $\boldsymbol{\beta^{k+1}}$, $\delta_i^{k+1}, i=2,\ldots,u-1$.
- The $(k+1)$-st estimate of the covariance matrix of random effects is $\hat{\boldsymbol{\Sigma}}^{k+1} = \frac{1}{n}\sum_{i=0}^{n}E(\boldsymbol{b_i b_i'}|\boldsymbol{y_i^*;\Gamma^k})$.

In order to update the estimates we use the approximatations of the expectations and the variances described in the previous section.

**2.2. Standard error estimation.** We use the bootstrap method for standard errors approximation described in [25] pp. $130-131$. The steps are as follows:

1. We fit model (1) to the observed data set consisting of $n$ individuals using the proposed EM algorithm and obtain the estimates of the unknown parameters denoted by $\hat{\boldsymbol{\Gamma}} = (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Sigma}}, \hat{\boldsymbol{\delta}})$. To generate a bootstrap sample first we generate $n$ random effects $\boldsymbol{b_k^b}$ from $N(\boldsymbol{0}, \hat{\boldsymbol{\Sigma}})$, $k=1,\ldots,n$. Next we simulate normal values $\boldsymbol{y_k^b}$ of dimension $n_k$ according to model (1) for every random effect $\boldsymbol{b_k^b}$. We use the estimated thresholds via $\hat{\boldsymbol{\delta}}$ to determine in which interval the normal data $\boldsymbol{y_k^b}$, $k=1,\ldots,n$ fall and thus using (2) determine the level of the bootstrap categorical variable $\boldsymbol{y_k^{b*}}$. The bootstrap sample consists of the categorical variables $\boldsymbol{y_k^{b*}}$, $k=1,\ldots,n$.

2. We apply the EM algorithm to the bootstrap data $\boldsymbol{y_k^{b*}}$, $k = 1, \ldots, n$ to get estimates for the generated data set $\boldsymbol{\Gamma^b}$.

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

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

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

**3. Simulations.** For the implementation of the algorithm we used the free software environment for statistical computing and graphics **R**. The **R** code for fitting the presented models is available from the authors.

We simulated values from the following random intercept model:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + b_i + \epsilon_{ij},$$

where $\beta_0 = -0.5, \beta_1 = 1, Var(b_i) = \sigma^2 = 0.01, Var(\epsilon_{ij}) = 1$ with thresholds $\alpha_1 = 0, \alpha_2 = 1.5, \alpha_3 = 3, \alpha_4 = 4$.

We simulated 100 samples for two different sample sizes ($n = 100$ and $n = 500$) with 5 repeated measures on each individual. For each approximation of the standard errors we used 75 bootstrap samples which is within the recommended range of 50 to 100 bootstrap replications (Efron and Tibshirani [12]). The results are presented in Table 1.

Note that due to the re-parametrization we estimate the differences in thresholds rather than the thresholds themselves. In both simulations the averages of the estimated parameters are equal within two significant digits after the decimal point to the parameter values from which the samples were generated. The only exception is the estimate of the last threshold difference at the smaller sample size setting but even this estimate is close to the true value. Thus we empirically confirm the unbiasedness of the algorithm.

As expected the estimates get closer to the real values and the standard errors get smaller when we increase the sample size. All of the estimates are statistically significantly different from zero except the variance of the random intercept for

Table 1. Table of estimates and standard errors of both simulation studies

| real values | $\beta_0 = -0.5$ | $\beta_1 = 1$ | $\sigma^2 = 0.01$ | $\delta_2 = 1.5$ | $\delta_3 = 1.5$ | $\delta_4 = 1$ |
|---|---|---|---|---|---|---|
| Simulation 1: number of subjects = 500, $n_i = 5$ | | | | | | |
| mean of estimates | −0.503 | 1.001 | 0.010 | 1.50 | 1.50 | 0.997 |
| stand. dev. of estimates | 0.06 | 0.023 | 0.0006 | 0.053 | 0.052 | 0.036 |
| mean of bootstrap stand. errors | 0.059 | 0.023 | 0.0006 | 0.052 | 0.049 | 0.040 |
| Simulation 2: number of subjects = 100, $n_i = 5$ | | | | | | |
| mean of estimates | −0.498 | 1.007 | 0.011 | 1.512 | 1.498 | 1.009 |
| stand. dev. of estimates | 0.114 | 0.051 | 0.010 | 0.107 | 0.108 | 0.090 |
| mean of bootstrap stand. errors | 0.138 | 0.052 | 0.012 | 0.122 | 0.112 | 0.094 |

the smaller sample size. This is not surprising since this variance is small and the estimates at the smaller sample size are not as efficient as at the larger sample size.

Finally, the approximate equality of the standard deviations of the estimates and 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.

**4. Application of the model.** We apply the proposed model to the HRS data. The variable of main interest in the study (self-rated health) takes values from excellent (1) to poor (5). Categories (2), (3) and (4) mean very good, good and fair self-rated health respectively. We examine how self-rated health changes over time. We fit the following probit model to the data:

$$\text{Self rated health }_{ij} = \beta_0 + \beta_1 \text{Wave }_{ij} + b_i + \epsilon_{ij}$$

In the analysis we include 7550 individuals in the study who have complete set of observations. The results are presented in Table 2:

Table 2. Table of estimates and standard errors of the model fitted to HRS data

| | $\beta_0$ | $\beta_1$ | $\sigma^2$ | $\delta_2$ | $\delta_3$ | $\delta_4$ |
|---|---|---|---|---|---|---|
| estimates | 1.23 | 0.12 | 2.15 | 1.58 | 1.50 | 1.42 |
| standard errors | 0.015 | 0.0026 | 0.049 | 0.011 | 0.011 | 0.017 |

Table 2 shows that all of the parameters in the model are statistically significantly different from zero. The parameter of most interest is the regression coefficient $\beta_1$. It is positive and the z-test statistic for this parameter is $z = 0.12/0.0026 = 46.15$, $p-value < 0.0001$ and thus we conclude that self-rated health deteriorates significantly over time. Further study including additional covariates may reveal whether this change is associated with particular subject characteristics.

We also note that the variance of the random intercept is significantly different from 0. This implies that the between-subject variability of the self-reported health measurements is large and that there is strong correlation between the repeated measurements on a particular individual.

**5. Conclusions.** In this paper we considered a correlated probit model for the analysis of repeatedly measured ordinal outcomes. We proposed an extension of the EM algorithm of Chan and Kuk [8] for obtaining maximum likelihood estimates, implemented it in the free software environment for statistical computing and graphics **R** and studied its performance using simulations. We also illustrated the approach on self-reported health data from the Health and Retirement Study (HRS). Our approach has advantages over alternative estimation methods in that it can handle a large number of random effects, it can be easily extended to any combination of binary, ordinal and continuous outcomes and it provides unbiased estimates. It is also easily implemented in the open-source software environment **R**. Using free software is a premise for wider usage and quicker improvement of the code.

There are several possible directions in which the algorithm implementation can be improved. There is a possible extention of the algorithm, called parameter expanded ECM algorithm [21] that can accelerate the speed of convergence of the algorithm. Rather than restrict some parameters (e.g. the variance of the error term) for parameter identifiability up front, this extension allows estimation of all parameters free of restrictions and at the last iteration calculates fully iden-

tifiable functions of the parameter (e.g. the ratios of the regression parameters and the squared root of the variance of the errors estimate). An example of implementation of this algorithm can be found in Gueorguieva and Agresti [18].

It is also possible to improve the implementation of the algorithm by choosing different **R** functions or improving the efficiency of the code. We already applied one such optimization. Although there are functions for finding the first two moments of the multivariate truncated normal distribution in the package **mvtnorm** based on the work by Manjunath and Wilhelm [24] they are rather slow. Generating random numbers using Gibbs sampling [6] and finding the first two moments based on that sample proved to be quicker.

Standard error estimation 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 [23].

Further research is also needed to extend the algorithm to combinations of ordinal and continuous longitudinal outcomes. Model selection and model diagnostics are also open areas of research.

## REFERENCES

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

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

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

[4] C. I. Bliss. The method of probits. *Science*, 1934, 409–410.

[5] N. E. Breslow, D. G. Clayton. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, **88** (1993), 925.

[6] G. Casella, E. I. George. Explaining the Gibbs sampler. *The American Statistician* **46** (1992), No 3, 167–174, .

[7] P. J. Catalano. Bivariate modelling of clustered continuous and ordered categorical outcomes. *Statistics in Medicine* **16** (1997), No 8, 883–900.

[8] J. Chan, A. Kuk. Maximum likelihood estimation for probit-linear mixed models with correlated random effects. *Biometrics* 53 (1997), 86–97.

[9] A. P. Dempster, N. M. Laird, D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, **39** (1977) No, 1–22

[10] P. J. Diggle, K.-L. Liang, S. L. Zeger. Analysis of Longitudinal Data New York, Oxford University Press, 1996.

[11] B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, **7**(1979), No 1, 1–26.

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

[13] L. Fahrmeir, G. Tutz. Multivariate Statistical Modelling Based on Generalized Linear Models, 2nd edition. New York, Springer-Verlag, 2001.

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

[15] R. D. Gibbons, D. Hedeker. Application of random-effects probit regression models. *Journal of Consulting and Clinical Psychology* **62** (1994) No 2, 285–296.

[16] L. Grilli, C. Rampichini. Alternative specifications of multivariate multilevel probit ordinal response models. *Journal of Educational and Behavioral Statistics* **28** (2003), 3144.

[17] R. V. Gueorguieva. Correlated probit model. In: Encyclopedia of Biopharmaceutical Statistics, chapter 59, 2006, 355–362.

[18] R. V. Gueorguieva, A. Agresti. A correlated probit model for joint modeling of clustered binary and continuous responses. *Journal of the American Statistical Association* **96** (2001), 1102–1112.

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

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

[21] C. Liu, D. Rubin, Y. Wu. Parameter expansion to accelerate EM: The PX-EM algorithm. *Biometrika* **85**(1998), No 4, 755–770.

[22] Q. Liu, D. A. Pierce. A note on Gauss-Hermite quadrature. *Biometrika* **81** (1994), No 3, 624–629.

[23] T. A. Louis. Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society. Series B* **44** (1982), No 2, 226–233.

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

[25] G. J. McLachlan, T. Krishnan. The EM Algorithm and Extensions, 2 edition. New York, Wiley, Mar. 2008.

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

[27] Y. Ochi, R. L. Prentice. Likelihood inference in a correlated probit regression model. *Biometrika* **71** (1984), No 3, 531–543.

[28] P. A. Ruud. Extensions of estimation methods using the EM algorithm. *Journal of Econometrics* **49** (September 1991), No 3, 305–341.

[29] R. E. Weiss. *Modeling Longitudinal Data*. New York, Springer Science+Business Media, 2005.

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

[31] R. Wolfinger, M. O'Connell. Generalized linear mixed models: A pseudo-likelihood approach. *Journal of Statistical Computation and Simulation* **48** (1993), 233–243.

*Denitsa Grigorova*
*Sofia University "St. Kliment Ohridski"*
*Faculty of Mathematics and Informatics*
*James Bourchier Blvd*
*Sofia 1164*
*Bulgaria*
*e-mail:* dgrigorova@fmi.uni-sofia.bg

*Ralitza Gueorguieva*
*Department of Biostatistics*
*Yale School of Public Health*
*60 College St*
*New Haven, CT 06520*
*USA*
*e-mail:* ralitza.gueorguieva@yale.edu