Correlated probit analysis of two longitudinal ordinal outcomes[a]

Denitsa Grigorova

dgrigorova@fmi.uni-sofia.bg

Sofia University "St. Kliment Ohridski"

16 September 2017
Sofia

[a] 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.

Contents

[a] 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.

1 Introduction

- Probit models [Bliss, 1934] are suitable for ordered categorical variables. We make an assumption for a latent continuous variable, which is categorized Fig. 1.
- Correlated probit models (CPM) are used when we have multiple outcomes per subject/group, e.g. longitudinal or clustered data. They are easy for interpretation and allow rich correlation structure within subject/group.
- CPMs have several advantages: testing several parameters while controlling type I error, improving efficiency of the estimates, assessing correlation between clustered measurements.
- Methods for finding of MLE for correlated binary data are well developed but not for correlated ordinal data.

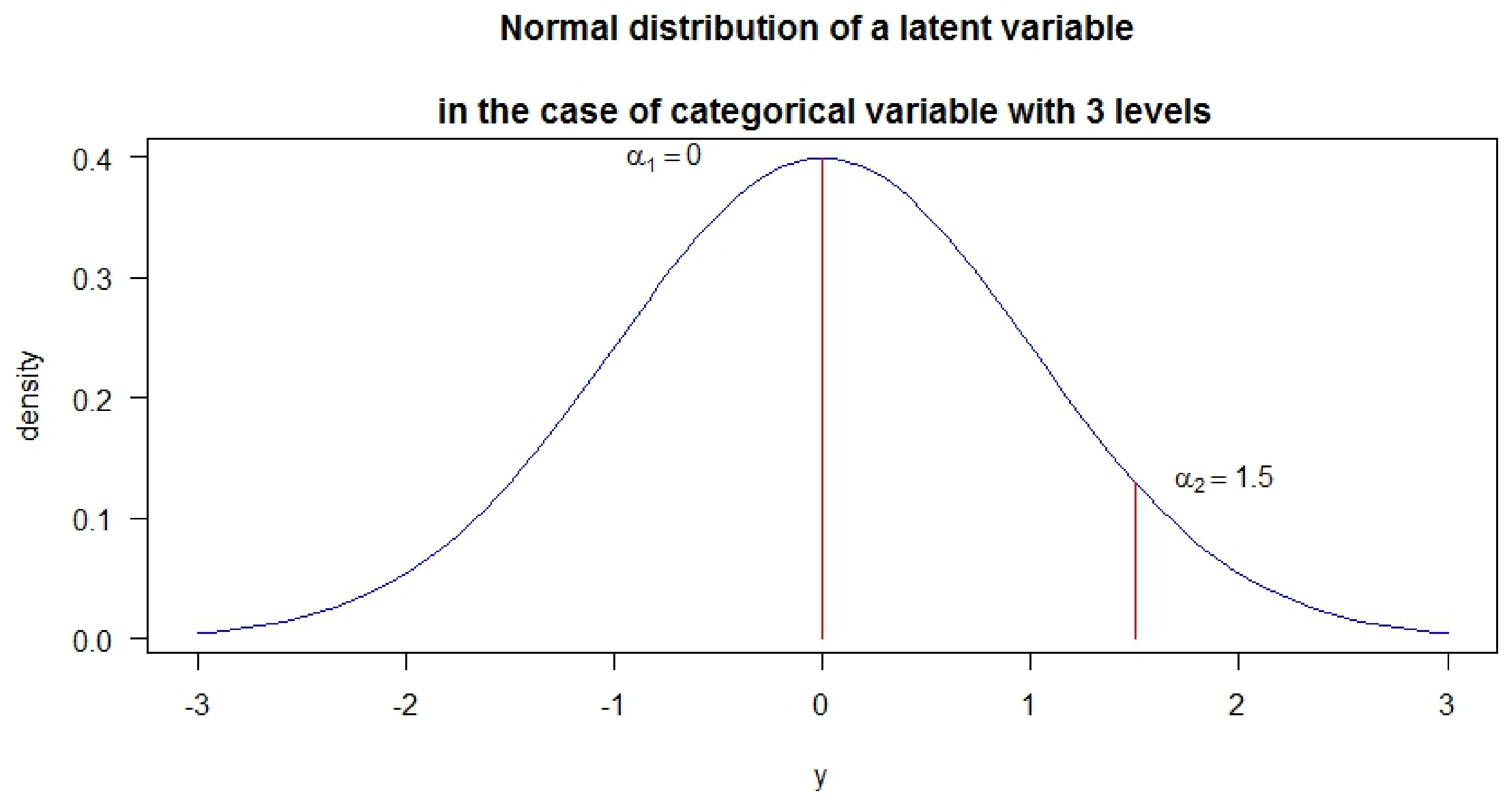

Figure 1: Density of a latent normal variable. Examples: logit model (logistic distribution) and probit model (normal distribution) [Kutner *et al.*, 2005]

2 Literature review

- The EM algorithm is an option for finding of MLE for probit model [Ruud, 1991].
- [Kawakatsu & Largey, 2009] extend his approach to finding MLE for model for a single ordinal outcome and multivariate normal outcome.
- [Grigorova *et al.*, 2013] propose EM algorithm for MLE of a model for multiple ordinal outcomes.
- Alternative method for finding of estimates: Numerical methods (Gaussian quadrature) have been used by [Todem *et al.*, 2007] and [Liu & Hedeker, 2006].
- We extend the EM algorithm approach to finding MLE for CPM for two repeatedly measured ordinal outcomes.
- **The computations using EM algorithm do not grow exponentially with the increase of the dimension of the random effects.**

3 Correlated probit model for two longitudinal ordinal outcomes

3.1 Motivating example: Health and Retirement Study (HRS)

- Health and Retirement Study (http://hrsonline.isr.umich.edu/)
- The variables of main interest are: self-rated health (SRH) and categorized body mass index (CBMI).
  - SRH has five levels: excellent (coded as 1), very good (2), good (3), fair (4) and poor (5).
  - 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).
- Objective of the study: modeling how self-assessment of health and categorized body mass index depend on gender and smoking status and vary over time.

3.2 Model definition

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 second ordinal outcome with m_2 levels on the same subject at the same time. We assume that there are latent normal variables y_{1ij} and y_{2ij} that generated the observed ordinal variables. We consider the following random effects model:

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

$$y^*_{kij} = \begin{cases} 1, \ y_{kij} \leq \alpha_{k,1}; \\ l, \ \alpha_{k,l-1} < y_{kij} \leq \alpha_{k,l}, \ l = 2, \ldots, m_k - 1; \\ m_k, \ y_{kij} > \alpha_{m_k-1}; \end{cases}$$

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

The vector of random effects: $\boldsymbol{b_i} = (\boldsymbol{b'_{1i}}, \boldsymbol{b'_{2i}})' \sim N(\boldsymbol{0_q}, \boldsymbol{\Sigma})$, where

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

The error terms: $(\epsilon_{1ij}, \epsilon_{2ij})' \sim N(\boldsymbol{0_2}, \boldsymbol{\Sigma_\epsilon})$, where

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

Regression parameters for the fixed effects: $\boldsymbol{\beta_1}, \boldsymbol{\beta_2}$.

The vectors of predictors for the fixed effects: $\boldsymbol{x_{1ij}}, \boldsymbol{x_{2ij}}, j = 1, \ldots, n_i$.

The vectors of predictors for the random effects: $\boldsymbol{z_{1ij}}, \boldsymbol{z_{2ij}}, j = 1, \ldots, n_i$.

Restrictions for identifiability of the model:
$\alpha_{k,1} = 0, k = 1, 2,\ \sigma_{11} = 1,\ \sigma_{2|1} = \sigma_{22} - \sigma_{12}^2/\sigma_{11} = 1.$

3.3 EM algorithm for MLE

- Linear transformation of the latent variables (approach [Kawakatsu & Largey, 2009]): $y_{kij_{new}} = (y_{kij} - \alpha_{k,y^*_{kij}-1})/\delta_{k,y^*_{kij}}$, where $\delta_{k,i} = \alpha_{k,i} - \alpha_{k,i-1}$, $i = 2, \ldots, m_k - 1$ $(\delta_{k,1} = \delta_{k,m_k} = 1)$, $k = 1, 2$.
  Unknown parameters: $\mathbf{\Gamma} = (\boldsymbol{\beta_1}, \boldsymbol{\beta_2}, \mathbf{\Sigma}, \boldsymbol{\delta_1}, \boldsymbol{\delta_2}, \lambda)$, where $\boldsymbol{\delta_k} = (\delta_{k,2}, \ldots, \delta_{k,m_k-1}), k = 1, 2$ and $\lambda = \sigma_{12}$;
- Complete data log-likelihood $\ln L = \ln f(\boldsymbol{b}, \boldsymbol{y_{1_{new}}}, \boldsymbol{y_{2_{new}}})$;
- Steps of the ECM algorithm ([Meng & Rubin, 1993]):
  - E-step: finding of the expectations of the closed form expressions of estimators (we show that they depend only on the first two moments of multivariate truncated normal distribution);
  - M-step: several simpler conditional maximization steps.

  We start with initial values for the parameters, iterate between the E-step and the M-step until convergence.
- Standard error estimation: Monte Carlo approach to the bootstrap method for standard errors approximation ([McLachlan & Krishnan, 2008] pp. 130-131).

3.4 Simulation study

We simulated values from the following random intercept model with sample size 1500 and 3000:

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

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 matrix of errors is: $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix} = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1.64 \end{pmatrix}.$

The covariance matrix of the random effecs is:

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

Table 1: Estimates and standard errors in the simulation model 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	1.8	0 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

4 Application of the correlated probit model to the HRS data

$$
\begin{aligned}
y_{1ij} = {} & \beta_{10} + \beta_{11}t_{ij} + \beta_{12}I(smoker) + \beta_{13}I(female) + \\
& \beta_{14}t_{ij}I(smoker) + \beta_{15}t_{ij}I(female) + \beta_{16}I(smoker)I(female) \\
& +\beta_{17}t_{ij}I(smoker)I(female) + b_{1i1} + b_{1i2}t_{ij} + \epsilon_{1ij}, \qquad (2) \\
y_{2ij} = {} & \beta_{20} + \beta_{21}t_{ij} + \beta_{22}I(smoker) + \beta_{23}I(female) + \\
& \beta_{24}t_{ij}I(smoker) + \beta_{25}t_{ij}I(female) + \beta_{26}I(smoker)I(female) \\
& +\beta_{27}t_{ij}I(smoker)I(female) + b_{2i1} + b_{2i2}t_{ij} + \epsilon_{2ij},
\end{aligned}
$$

$$
y^*_{kij} = \begin{cases} 1, \; y_{kij} \le \alpha_{k,1} = 0, \\ l, \; \alpha_{k,l-1} < y_{kij} \le \alpha_{k,l}, \; l = 2, \ldots, m_k - 1 \\ m_k, \; y_{kij} > \alpha_{k,m_k-1}, \end{cases}
$$

where $m_1 = 5, m_2 = 4$.

where the covariance matrix of the errors is: $\mathbf{\Sigma}_{\boldsymbol{\epsilon}} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}$ and

the covariance matrix of the random effects is:

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

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

Regression parameters for latent self-rated health								
parameter	β_{10}	β_{11}	β_{12}	β_{13}	β_{14}	β_{15}	β_{16}	β_{17}
estimate	1.37	0.13	0.64	0.06	0.03	-0.03	-0.26	0.01
z-score	71.11	30.61	16.46	2.44	2.81	-4.73	-4.91	1.09
Regression parameters for body mass index								
parameter	β_{20}	β_{21}	β_{22}	β_{23}	β_{24}	β_{25}	β_{26}	β_{27}
estimate	6.00	0.05	-0.88	-0.28	-0.01	0.05	-0.09	0.01
z-score	113.4	8.94	-23.17	-9.36	-0.77	6.89	-1.53	0.53
Threshold parameters for both variables								
parameter	$\delta_{1,2}$	$\delta_{1,3}$	$\delta_{1,4}$		$\delta_{2,2}$	$\delta_{2,3}$		
estimate	1.64	1.56	1.45		4.97	3.31		
z-score	172.18	164.15	120.11		34.28	44.91		

Table 3: Table of estimates and z-scores of the covariance parameters in model 2 fitted to the first seven waves of HRS data

parameter	σ^b_{11}	σ^b_{22}	σ^b_{33}	σ^b_{44}	σ^b_{12}	σ^b_{13}	σ^b_{14}	σ^b_{23}	σ^b_{24}	σ^b_{34}	λ
estimate	3.541	0.038	8.185	0.082	-0.194	1.308	-0.074	-0.020	0.002	-0.160	-0.002
z-score	54.872	47.948	26.294	17.507	-30.277	18.345	-10.151	-2.619	2.638	-8.122	-0.267

Standard deviations of the random effects are:

1.882	0.195	2.861	0.286

The correlation between the random intercept and random slope for the latent self-rated health: -0.53 and for the BMI is: -0.20. The correlation between the random intercepts is 0.24.

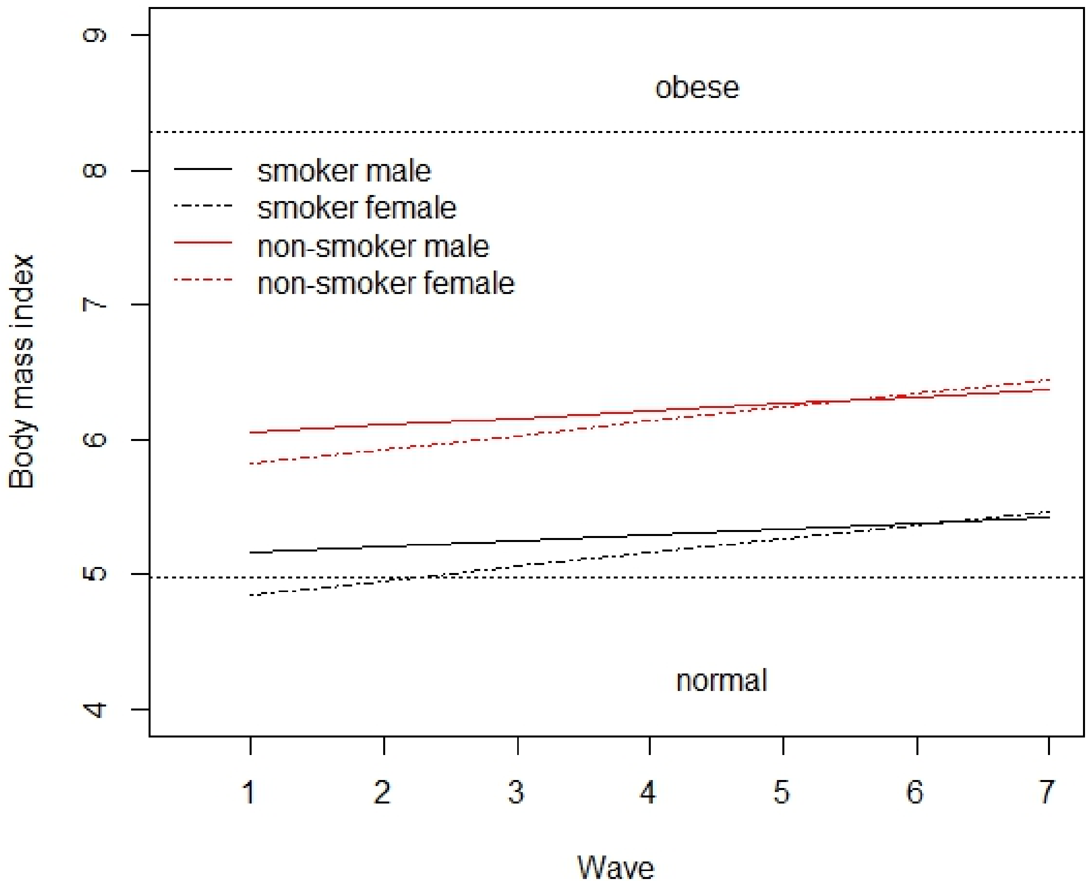

Figure 2: Latent CBMI over time for four individuals with zero random effects

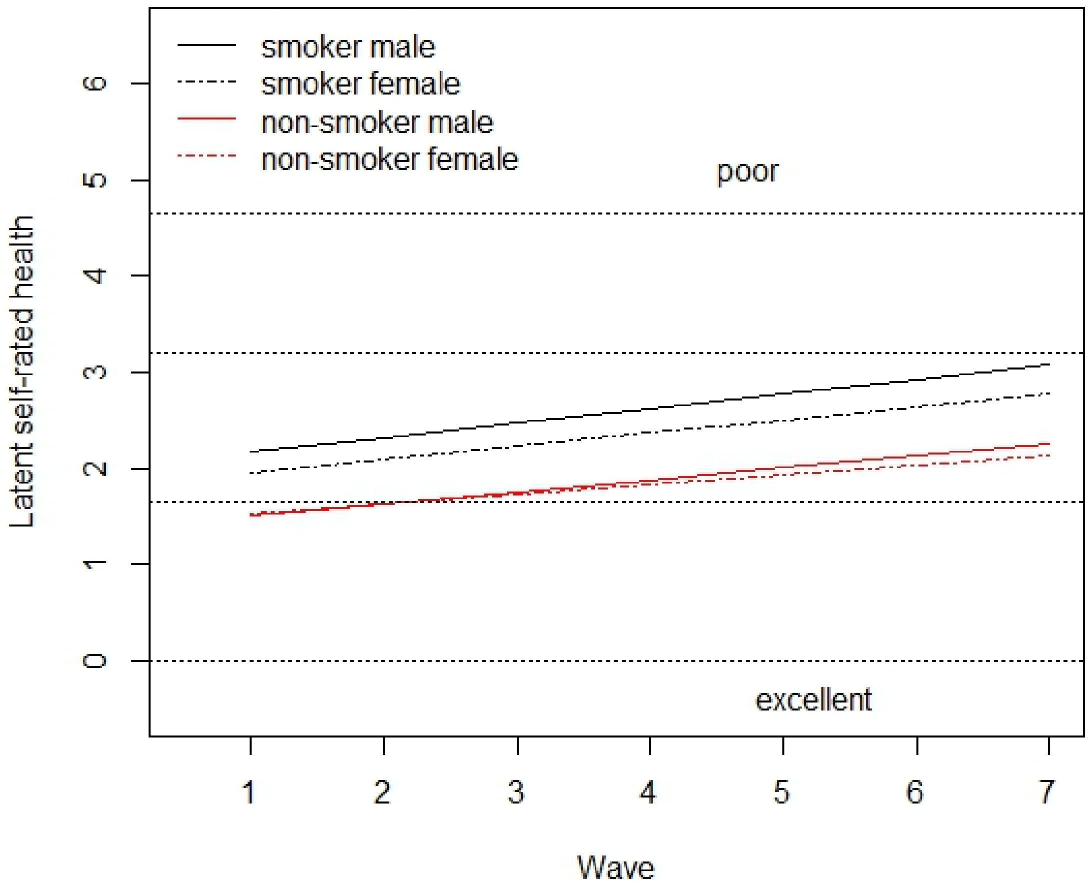

Figure 3: Latent SRH over time for four individuals with zero random effects

Conclusions from the analysis:

- The three-way interactions between time, gender and smoking is not statistically significant in both sub-models;
- The two-way interactions between time, gender and smoking are statistically significant in the sub-model for latent self-rated health while only the two-way interaction between time and gender is statistically significant in the sub-model for BMI;
- The estimates of the standard deviations of the random intercepts are bigger compared to the estimates of the standard deviations of the random slopes.
- There is no strong correlation between any of the random effects. The random intercept and random slope are negatively correlated for both variables. The correlation is stronger for the latent self-rated health.

5 The R package EMcorrProbit[a]

Main functionality of the package:

- function `emcorrprobit` computes:
  - estimates of the parameters of correlated probit model for one longitudinal ordinal outcome via EM algorithm
  - log-likelihood;
  - random effects;
  - AIC and BIC.
- function `summary.emcorrprobit` - approximates the covariance matrix of the estimates
- flexible design allowing for expansion for:
  - joint model for one normal and one ordinal longitudinal outcomes [Grigorova & Gueorguieva, 2016];
  - model for bivariate ordinal outcome;
  - joint model for two longitudinal ordinal outcomes.

[a] 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.

Instalation of the package

- The EMcorrProbit package is available on GitHub (https://github.com/ninard/EMcorrProbit) and in the near future will be available on CRAN.
- Any R user can install the package using the following two command lines:

```
install.packages("devtools")
devtools::install_github("ninard/EMcorrProbit")
```

Example of how to use the two main functions:

```
example1=emcorrprobit(model = "oneord", y=your.ordinal.data,
                      xfixed=predictors.fixed,
                      xrand=predictors.random,
                      start.values.beta=beta,
                      start.values.delta=delta,
                      start.values.sigma.rand=sigma.rand,
                      exact=TRUE,montecarlo=100,epsilon=.001)



ex1.se=summary(example1, doParallel=TRUE,
                bootstrap.samples=50)
ex1.se$vcov
```

6 Discussion

- CPMs are useful when the correlation within clustered observations can not be ignored. Advantage of CPMs is the possibility of testing several parameters, while controlling type I error. They are easy for interpretation and allow different correlation structures using random effects and/or correlated errors.
- The EM algorithm is a suitable option for finding of MLE of the parameters of CPM because the computational complexity does not increase exponentially with the dimension of the random effects.
- The main aim of the R package EMcorrProbit is to provide the whole functionality of such models in a compact and user-friendly way.
- Future work related to the complication ‘dropout’ from a longitudinal study is the definition and estimation of a joint model for two ordinal longitudinal outcomes and time to dropout.

MANY THANKS FOR YOUR ATTENTION!

References

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

[Grigorova & Gueorguieva, 2016] Grigorova, D., & Gueorguieva, R. 2016. Correlated probit analysis of repeatedly measured ordinal and continuous outcomes with application to the Health and Retirement Study. *Statistics in Medicine*, **35**(23), 4202–4225. sim.6982.

[Grigorova *et al.*, 2013] Grigorova, D., Encheva, E., & Gueorguieva, R. 2013. Implementation of the EM algorithm for maximum likelihood estimation of a random effects model for multiple ordinal outcome. *Serdica Journal of Computing*, **7**(3), 227–244.

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

[Kutner *et al.*, 2005] Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. 2005. *Applied Linear Statistical Models*. McGraw-Hill/Irwin.

[Liu & Hedeker, 2006] Liu, L. C., & Hedeker, D. 2006. A Mixed-Effects Regression Model for Longitudinal Multivariate Ordinal Data. *Biometrics*, **62**(1), 261–268.

[McLachlan & Krishnan, 2008] McLachlan, G. J., & Krishnan, T. 2008. *The EM Algorithm and Extensions (Wiley Series in Probability and Statistics)*. 2 edn. Wiley-Interscience.

[Meng & Rubin, 1993] Meng, X.-L., & Rubin, D. B. 1993. Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, **80**(2), 267–278.

[Ruud, 1991] Ruud, P. A. 1991. Extensions of estimation methods using the EM algorithm. *Journal of Econometrics*, **49**(3), 305–341.

[Todem *et al.*, 2007] Todem, D., Kim, K., & Lesaffre, E. 2007. Latent-variable models for longitudinal data with bivariate ordinal outcomes. *Statistics in Medicine*, **26**, 1034–1054.