Multivariate ordinal regression models using the R package MultOrd

Rainer Hirk & Kurt Hornik & Laura Vana

WU Vienna University of Economics and Business

Introduction

Multivariate ordinal regression models build up on *cumulative link models* which are amongst the most popular models for univariate ordinal data analysis. In cumulative link models the observed ordinal outcome Y is assumed to be a coarser (categorized) version of a latent continuous variable $\widetilde{Y}$. If multiple observations on the same subject are observed, univariate cumulative link models can be extended to a multivariate framework. These repeated measurements for each subject may take place either at the same time yielding a cross-sectional multivariate ordinal regression model or at different points in time yielding a panel multivariate ordinal regression model (Bhat et al., 2010).

Model Class

Model Formulation

Let us suppose to have J repeated measurements on n different subjects i, where each repeated ordinal observation (indexed by $j \in J$) is denoted by Y_{ij}. Each observable categorical outcome Y_{ij} and the unobservable latent variable $\widetilde{Y}_{ij}$ are connected by:

$$Y_{ij} = r_{ij} \quad \Leftrightarrow \quad \theta_{j,r_{ij}-1} < \widetilde{Y}_{ij} \leq \theta_{j,r_{ij}}, \qquad r_{ij} \in \{1, \dots, K_j\},$$

where r_{ij} is a category out of K_j ordered categories and $\boldsymbol{\theta}_j$ is a vector of suitable threshold parameters for outcome j with the following restriction: $-\infty \equiv \theta_{j,0} < \theta_{j,1} < \cdots < \theta_{j,K_j} \equiv \infty$. The number of threshold categories K_j as well as the threshold parameters themselves are allowed to vary across outcome dimensions $j \in J$ in order to account for differences in the repeated measurements. Given an $n \times p$ matrix X_j of covariates for each $j \in J$, where each $\boldsymbol{x}_{ij}$ is a p-dimensional vector (i-th row of X_j) for subject i and repeated measurement j, the following linear model for the relationship between $\widetilde{Y}_{ij}$ and the vector of covariates $\boldsymbol{x}_{ij}$ is assumed:

$$\widetilde{Y}_{ij} = \beta_{j0} + \boldsymbol{x}_{ij}^\top \boldsymbol{\beta}_j + \epsilon_{ij}, \qquad \boldsymbol{\epsilon}_i = (\epsilon_{i1}, \epsilon_{i2}, \dots, \epsilon_{iJ})^\top \sim F_J(\boldsymbol{0}, \boldsymbol{\Sigma}), \tag{1}$$

where

- β_{j0} is an intercept term corresponding to outcome j,
- $\boldsymbol{\beta}_j = (\beta_{j1}, \dots, \beta_{jp})^\top$ is a vector of regression coefficients corresponding to outcome j,
- ϵ_{ij} is an error term with mean zero and distributed according to a J-dimensional distribution function F_J.

Identifiability: As the absolute scale and the absolute location are not identifiable in ordinal models further restrictions on the parameter set need to be imposed. Assuming a full covariance matrix $\boldsymbol{\Sigma}$ with diagonal elements σ_j^2, only the quantities $\boldsymbol{\beta}_j/\sigma_j$ and $(\theta_{j,r_{ij}} - \beta_{j0})/\sigma_j$ are identifiable in model (1). The scale can be fixed either by restricting the full variance-covariance matrix $\boldsymbol{\Sigma}$ to be a correlation matrix $\boldsymbol{R}$ or by fixing two threshold parameters. For the location either the intercept β_{j0} or one threshold parameter has to be fixed to some value. Therefore, in order to obtain an identifiable model the following typical constraints on the parameters can be imposed for all $j \in J$:

- Fixing the intercept β_{j0} (e.g., to zero), using flexible thresholds $\boldsymbol{\theta}_j$ and fixing σ_j (e.g., to unity).
- Leaving the intercept β_{j0} unrestricted, fixing one threshold parameter (e.g., $\theta_{j,1} = 0$) and fixing σ_j (e.g., to unity).
- Leaving the intercept β_{j0} unrestricted, fixing two threshold parameters (e.g., $\theta_{j,1} = 0$ and $\theta_{j,2} = 1$) and leaving σ_j unrestricted.
- Fixing the intercept β_{j0} (e.g., to zero), fixing one threshold parameter (e.g., $\theta_{j,1} = 0$) and leaving σ_j unrestricted for all $j \in J$.

Different error structures

error.structure	Cov. structure ($\boldsymbol{\Sigma}$)	Corr. structure ($\boldsymbol{R}$)	Factor dependent	Covariate dependent
corGeneral(~1)		✓		
corGeneral(~f)		✓	✓	
covGeneral(~1)	✓			
covGeneral(~f)	✓		✓	
corEqui(~1)		✓		
corEqui(~X)		✓		✓
corAR1(~1)		✓		
corAR1(~X)		✓		✓

Table 1: Overview of the different error structures.

Estimation

For a given parameter vector $\boldsymbol{\Gamma}$ which contains the threshold parameters $\boldsymbol{\Theta}$, the regression coefficients $\boldsymbol{B}$ and the variance-covariance (correlation) parameters $\boldsymbol{\Sigma}$ that have to be estimated, the likelihood has the following form:

$$L(\boldsymbol{\Gamma}) = \prod_{i=1}^{n} \mathbb{P}(Y_{i1} = r_{i1}, Y_{i2} = r_{i2}, \dots, Y_{iJ} = r_{iJ})^{w_i}$$
$$= \prod_{i=1}^{n} \left(\int_{\theta_{1,r_{i1}-1} - \beta_{j0} - \boldsymbol{x}_{ij}^\top \boldsymbol{\beta}_j}^{\theta_{1,r_{i1}} - \beta_{j0} - \boldsymbol{x}_{ij}^\top \boldsymbol{\beta}_j} \cdots \int_{\theta_{J_i,r_{iJ_i}-1} - \beta_{j0} - \boldsymbol{x}_{ij}^\top \boldsymbol{\beta}_j}^{\theta_{J_i,r_{iJ_i}} - \beta_{j0} - \boldsymbol{x}_{ij}^\top \boldsymbol{\beta}_j} f_J(v_{i1}, \dots, v_{iJ}; \boldsymbol{R}) dv_{i1} \dots dv_{iJ} \right)^{w_i}$$

where f_q denotes the density of q-dimensional distribution F_q. In order to estimate the model parameters we approximate full likelihood is by a composite likelihood, where a pseudolikelihood is constructed from bivariate marginal distributions F_2 (Pagui et al., 2015). Using transformed upper $U_{ij} = \theta_{j,r_{ij}} - \beta_{j0} - \boldsymbol{x}_i^\top \boldsymbol{\beta}_j$ and $L_{ij} = \theta_{j,r_{ij}-1} - \beta_{j0} - \boldsymbol{x}_i^\top \boldsymbol{\beta}_j$ the lower integration bounds, the pairwise log-likelihood function is obtained by

$$c\ell^{PL}(\boldsymbol{\Gamma}) = \sum_{i=1}^{n} \sum_{k=1}^{J-1} \sum_{l=k+1}^{J} w_i \log \left(\mathbb{P}(Y_{ik} = r_{ik}, Y_{il} = r_{il}) \right)$$
$$= \sum_{i=1}^{n} \sum_{k=1}^{J-1} \sum_{l=k+1}^{J} w_i \log \left(\int_{L_{ik}}^{U_{ik}} \int_{L_{il}}^{U_{il}} f_2(v_{ik}, v_{il} | \rho_{kl}) dv_{ik} dv_{il} \right) \tag{2}$$

The maximum composite likelihood estimates $\hat{\boldsymbol{\Gamma}}_{c\ell}$ are obtained by direct maximization of the composite likelihood given in using general purpose optimizers of the R package `optimx`. Standard errors are computed by means of the Godambe information matrix in order the standard errors to quantify the uncertainty of the maximum composite likelihood estimates (Varin, 2008).

Implementation

Multivariate ordinal regression models in the R package **MultOrd** are fitted using the function `multord`. The usage of the function `multord` is explained by means of a short credit rating example based on the following dataset `data`:

Credit ratings data

```
        firmID  raterID  rating     X1         X2         X3              f
   1     254    Moody's   Aaa    0.453214   2.394723   0.862093   manufacturing
   2     259      S&P     BBB    0.645985   1.928982   1.229113    retail trade
   ⋮      ⋮        ⋮        ⋮        ⋮          ⋮          ⋮              ⋮
2999     537      S&P      AA    0.583231   2.598759   0.882301       mining
3000     537     Fitch     AA    0.583231   2.598759   0.882301       mining
```

Let us assume that we have a dataset of corporate credit ratings of different firms from different raters at the same point in time or from the same rater at different points in time. Each row of `data` corresponds to a single credit rating observations from one rater of a firm together with its covariates `X1, X2, X3` and `f`. A character vector of length two `index` specifies the subject index i (`firmID`) and the repeated measurement index j (`raterID`). Let us further assume that we have three different raters (`response.names = c("S\&P", "Moody's", "Fitch")`) with different categories[1]. If the categories differ across repeated measurements one needs to specify the `response.levels` explicitly by:

```
response.levels <- list(c("AAA", "AA", "A", "BBB", "BB","B", "CCC\C"),
                        c("Aaa", "Aa", "A", "Baa", "Ba", "B", "Caa\C", "D"),
                        c("AAA", "AA", "A", "BBB", "BB", "B", "CCC\C"))
```

For a given repeated measurement index `raterID` and covariates `X1`, `X2` and `X3` the `formula` in a model **without intercept** has the following form:

```
formula1 <- rating ~ 0 + X1 + X2 + X3
```

In analogy, in a model **with intercept** we have:

```
formula2 <- rating ~ X1 + X2 + X3, or  formula3 <- rating ~ 1 + X1 + X2 + X3
```

Two different link functions can be used, either the probit link (`link = "probit"`), or the logit link (`link = "logit"`). Furthermore, constraints on the coefficients can be imposed in the following way:

Constraints on regression coefficients

```
coef.constraints <- cbind(c(1, 2, 3),
                          c(1, 2, 1),
                          c(1, NA, 1))

coef.values <- cbind(c(NA, NA, NA),
                     c(NA, NA, NA),
                     c(2, 0, 2))
```

gives the following model:

$$\widetilde{Y}_{i1} = \beta_{10} + \beta_{11} x_{i1} + \beta_{12} x_{i2} + 2 x_{i3},$$
$$\widetilde{Y}_{i2} = \beta_{20} + \beta_{21} x_{i1} + \beta_{22} x_{i2},$$
$$\widetilde{Y}_{i3} = \beta_{30} + \beta_{31} x_{i1} + \beta_{12} x_{i2} + 2 x_{i3}.$$

In addition, constraints on the threshold coefficients can be imposed by:

Constraints on threshold coefficients

```
threshold.constraints <- c(1, 2, 1)

threshold.values <- list(c(-3, NA, NA, NA, NA, NA),
                         c(-3.5, NA, NA, NA, NA, NA, NA),
                         c(-3, NA, NA, NA, NA, NA))
```

gives

$$\boldsymbol{\theta}_1 = \boldsymbol{\theta}_3,$$
$$\theta_{11} = -3 < \theta_{12} < \theta_{13} < \theta_{14} < \theta_{15} < \theta_{16},$$
$$\theta_{21} = -3.5 < \theta_{22} < \theta_{23} < \theta_{24} < \theta_{25} < \theta_{26} < \theta_{27}.$$

A multivariate ordinal regression model for the credit rating example is then fitted by the call:

Function call

```
multord(formula = formula2, data = data, index = c("firmID", "rater"),
        response.names = c("S&P", "Moody's", "Fitch"),
        response.levels = response.levels, link = "probit",
        error.structure = corGeneral(~f), coef.constraints = coef.constraints,
        coef.values = coef.values, threshold.constraints = threshold.constraints,
        threshold.values = threshold.values, se = TRUE, start.values = NULL,
        solver = "newuoa", PL.lag = NULL)
```

In addition, several methods like `summary`, `print`, `coef`, `threshold`, `sigma` and `predict` are implemented for the class `'multord'`.

[1] S&P and Fitch: AAA, AA, A, BBB, BB, B, CCC\C. Moody's: Aaa, Aa, A, Baa, Ba, B, Caa\C, D

Conclusion

- R-package **MultOrd** offers a flexible framework for multivariate ordinal regression models.
- Different error structures allow for cross-sectional and panel models.
- Constraints on regression coefficients as well as threshold parameters can be imposed.

References

Bhat, C. R., Varin, C., and Ferdous, N. (2010). A comparison of the maximum simulated likelihood and composite marginal likelihood estimation approaches in the context of the multivariate ordered-response model. *Advances in Econometrics*, 26:65.

Pagui, K., Clovis, E., and Canale, A. (2015). Pairwise likelihood inference for multivariate ordinal responses with applications to customer satisfaction. *Applied Stochastic Models in Business and Industry.*

Varin, C. (2008). On composite marginal likelihoods. *AStA Advances in Statistical Analysis*, 92(1):1–28.