Handling item non-response in Structural Equation Modelling with ordinal variables

Myrsini Katsikatsou, Irini Moustaki

Department of Statistics, London School of Economics, UK

1. Framework

- Ordinal variables (items), cross-sectional data, N independent observations.
- Structural equation modelling (SEM)

Let $\mathbf{y}$ be the vector of ordinal items of dimension p, $\boldsymbol{\eta}$ the vector of factors, and $y_i^\star$ the underlying continuous variable of the ordinal variable y_i, where $y_i = a \Leftrightarrow \tau_{i,a-1} < y_i^\star < \tau_{i,a}$, $i = 1, \ldots, p$, a is the a-th response category of variable y_i, $a = 1, \ldots, c_i$, $\tau_{i,a}$ is the a-th threshold, $\tau_{i,0} = -\infty$, and $\tau_{i,c_i} = +\infty$.

$$\mathbf{y}^\star = \boldsymbol{\nu} + \Lambda\boldsymbol{\eta} + \boldsymbol{\varepsilon}$$

$$\boldsymbol{\eta} = \boldsymbol{\alpha} + B\boldsymbol{\eta} + \boldsymbol{\zeta},$$

where $\boldsymbol{\nu}$ and $\boldsymbol{\alpha}$ are intercept vectors, $\boldsymbol{\varepsilon} \sim \mathcal{N}_p(\mathbf{0}, \Theta_\varepsilon)$, $\boldsymbol{\zeta} \sim \mathcal{N}_q(\mathbf{0}, \Psi)$, $Cov(\boldsymbol{\eta}, \boldsymbol{\varepsilon}) = Cov(\boldsymbol{\eta}, \boldsymbol{\zeta}) = Cov(\boldsymbol{\varepsilon}, \boldsymbol{\zeta}) = \mathbf{0}$, $I - \mathrm{B}$ is non-singular, and I is the identity matrix.
Let $\boldsymbol{\theta}$ be the model parameter vector; it includes $\boldsymbol{\nu}$, $\boldsymbol{\alpha}$, Λ, B, Θ_ε, Ψ, $\boldsymbol{\tau}$.

- Item non-response (at least one variable observed in each sample unit)

Any type of missing pattern (monotone/ non-monotone) is allowed.

2. Background information on estimation

- Maximum likelihood is computationally unfeasible for SEM with ordinal items.
- Conventional estimation approach: three-stage diagonally weighted least squares (DWLS).
- When data are missing at random (MAR) (Rubin, 1976) multiple imputation followed by DWLS (MI-DWLS) is recommended.
- Alternative estimation approach: **pairwise likelihood** (**PL**),

$$pl(\boldsymbol{\theta}; \mathbf{y}) = \sum_{n=1}^{N} \sum_{i<j} \log f(y_{ni}, y_{nj}; \boldsymbol{\theta}),\ i, j = 1, \ldots, p.$$

In SEM with ordinal variables a bivariate probability is modeled as

$$\pi(y_{ni} = a, y_{nj} = b; \boldsymbol{\theta}) = \int_{\tau_{i,a-1}}^{\tau_{i,a}} \int_{\tau_{j,b-1}}^{\tau_{j,b}} f(y_i^\star, y_j^\star)\, dy_i^\star dy_j^\star.$$

2.1 Treatment of missing values under PL

- The **complete-pairs PL** (**CP-PL**) defined as

$$pl^{CP}(\boldsymbol{\theta}; \mathbf{y}) = \sum_{n=1}^{N} \sum_{i<j} \tilde{r}_{n,ij} \log f(y_{ni}, y_{nj}; \boldsymbol{\theta}),$$

where $\tilde{r}_{n,ij}$ takes the value 1 if both y_{ni}, y_{nj} are observed and 0 otherwise.
- The **available- case PL** (**AC-PL**) defined as

$$pl^{AC}(\boldsymbol{\theta}; \mathbf{y}) = pl^{CP}(\boldsymbol{\theta}; \mathbf{y}) + \sum_{n=1}^{N} m_n \sum_{i=1}^{p} r_{ni} \log f(y_{ni}; \boldsymbol{\theta}),$$

where m_n is the number of items with missing value for the nth sample unit, and r_{ni} takes the value 1 if y_{ni} is observed and 0 otherwise.
- Molenberghs et al. (2011) argue that, in general, CP-PL and AC-PL yield biased estimators under MAR.

3. Research Objective

- Study the performance of CP-PL and AC-PL under MAR via a simulation study.
  Performance criteria:
  - **Standardised Bias** of parameter estimates, $\frac{\bar{\hat{\theta}} - \theta}{SE(\hat{\theta})}$,
    where $\bar{\hat{\theta}}$ and $SE(\hat{\theta})$ are, respectively, the average and standard deviation of the replicated estimates.
  - Root Mean Square Error (**RMSE**) of parameter estimates ,$\frac{\sum_{r=1}^{R}\left(\hat{\theta}_r - \theta\right)^2}{R}$,
    where $\hat{\theta}_r$ is the rth replicated estimate, and R is the number of replications.
  - **Coverage rate of the 95% confidence interval** (CI),
    where $\hat{\theta}_r \pm 1.96 * \hat{se}(\hat{\theta}_r)$ is the rth replicated 95% CI, and
    $\hat{se}(\hat{\theta}_r)$ is the estimated standard error at the rth replication.
  - **Bias** of standard errors, $\frac{\sum_{r=1}^{R} \hat{se}(\hat{\theta}_r)}{R} - SE(\hat{\theta})$.
  - **Type I error rate** of pairwise likelihood ratio test (**PLRT**) for overall goodness-of-fit at 5% and 1% significance level.
- CP-PL and AC-PL performances are benchmarked against that of PL with complete data set.
- Compare the performances of CP-PL and AC-PL to that of MI-DWLS.

4. Simulation study design

10 experimental conditions
Sample size 300 & 1000 in each of the 5 experimental conditions below.

	Experimental Conditions				
	1	2	3	4	5
Factors		1		2	
Items		6		12 (6 items per factor)	
Item loadings	0.8 for all y's	0.4 for y_1 0.8 for $y_2, \ldots, y_6$	0.6 for all y's	$0.4, 0.5, 0.6, 0.7, 0.8, 0.9$ (for each 6-item set)	
Factor correlation				0.3	0.6

Missing rate	approximately 30% in each item except for the first item in each factor which is always observed.
Missing mechanism	MAR, where $\Pr(y_i \text{ missing}) = \frac{1}{1+\exp(-0.63+0.1*y_j)}$, $i = 2, \ldots, 6$ when $j = 1$, and $i = 8, \ldots, 12$ when $j = 7$.
Replications	1000 in each of the total 10 experimental conditions.
Software	Our R code for PL (with complete data), CP-PL, AC-PL, which has been incorporated into package `lavaan`. **Mplus** for MI-DWLS; the covariance model used to produce 10 imputed data sets.

References

Molenberghs, G., Kenward, M.G., Verbeke, G., Birhanu, T. (2011). Pseudo-likelihood estimation for incomplete data, *Statistica Sinica* (21), 187-206.

Rubin, D. B. (1976). Inference and missing data (with discussion). *Biometrika 63*, 581-592.

Acknowledgments: Many thanks to Prof. Yves Rosseel who incorporated our R code into his R package `lavaan`. The research is supported by ES/L009838/1 ESRC grant.

5. Simulation results

Loadings and factor correlation

- **All three** methods, CP-PL, AC-PL, and MI-DWLS, show **acceptable performance** regarding **all performance criteria in all conditions**.
  However, **CP-PL and AC-PL exhibit lower standardised bias and a better coverage rate than MI-DWLS**.
- **CP-PL and AC-PL** exhibit standardised bias and coverage rate fairly **close to** those of **PL with complete data**, especially for sample size 1000.
- A **sample size** increase seems to be associated with better performance in all criteria.
- A smaller factor **loading for** y_1 which determines the level of MAR, seems to be associated with slightly larger bias of both estimates and standard errors but this is the case in PL with complete data as well.
- Smaller **loadings** in all items seem to be associated with larger RMSE.
- A higher **factor correlation** seems to be associated with improved coverage rate.

Thresholds

- **MI-DWLS clearly outperforms CP-PL and AC-PL in all criteria**.
- **No clear preference between AC-PL and CP-PL as:**
  - AC-PL exhibits acceptable standardised bias (average per condition up to 11%), while standardised bias in CP-PL may exceed 40%.
  - **But**, AC-PL systematically under-estimates the standard errors leading to unacceptable coverage rate. CP-PL shows similar levels of standard errors bias as MI-DWLS and acceptable coverage in most occasions.
- A **hybrid PL**, which uses the AC-PL threshold estimates and the corresponding CP-PL standard errors, exhibits acceptable coverage rate in all conditions.

PLRT for overall goodness-of-fit

- Both CP-PL and AC-PL show rates of Type I error very close to the nominal levels 5% and 1% with two exceptions:
  - in Ex. Con. 6 with sample size 300, where the rates are smaller than the nominal levels, and
  - in Ex. Con. 3, where the rates are a bit larger than the nominal ones, but actually this occurs in PL with complete data as well.

6. Discussion

- The general result that CP-PL and AC-PL yield biased estimates do not seem to hold in SEM, especially for loadings and factor correlations.
- **Potential advantages of CP-PL and AC-PL over MI-DWLS**: a) MI-DWLS requires a model for imputing, b) in MI-DWLS, it is no clear how to use the fit indices to judge overall fit.
- Worthy to develop a **doubly-robust PL** (Molenberghs et al., 2011) and compared it to CP-PL and AC-PL.