

# Structural Equation Modeling: Categorical Variables

Anders Skrondal<sup>1</sup> and Sophia Rabe-Hesketh<sup>2</sup>

<sup>1</sup>Department of Statistics  
London School of Economics and Political Science (LSE)

<sup>2</sup>Graduate School of Education and Graduate Group in Biostatistics  
University of California, Berkeley

## Abstract

In the behavioral sciences, response variables are often noncontinuous, common types being dichotomous, ordinal or nominal variables, counts and durations. Conventional structural equation models (SEMs) have thus been generalized to accommodate different kinds of responses.

Keywords: Structural equation model, categorical data, item response model, MIMIC model, generalized latent variable model

## Introduction

**Structural equation models** (SEMs) comprise two components, a measurement model and a structural model. The measurement model relates observed responses or ‘indicators’ to **latent variables** and sometimes to observed covariates. The structural model then specifies relations among latent variables and regressions of latent variables on observed variables. When the indicators are categorical, we need to modify the conventional measurement model for continuous indicators. However, the structural model can remain essentially the same as in the continuous case.

We first describe a class of structural equation models also accommodating dichotomous and ordinal responses [5]. Here, a conventional measurement model is specified for multivariate normal ‘latent responses’ or ‘underlying variables’. The latent responses are then linked to observed categorical responses via threshold models yielding probit measurement models.

We then extend the model to generalized latent variable models (e.g. Bartholomew and Knott [1]; Skrondal and Rabe-Hesketh [13]) where, conditional on the latent variables, the measurement models are **generalized linear models** which can be used to model a much wider range of response types.

Next, we briefly discuss different approaches to estimation of the models since estimation is considerably more complex for these models than for conventional structural equation models. Finally, we illustrate the application of structural equation models for categorical data in a simple example.

## SEMs for latent responses

### *Structural model*

The structural model can take the same form regardless of response type. Letting  $j$  index units or subjects, Muthén [5] specifies the structural model for latent variables  $\boldsymbol{\eta}_j$  as

$$\boldsymbol{\eta}_j = \boldsymbol{\alpha} + \mathbf{B}\boldsymbol{\eta}_j + \boldsymbol{\Gamma}\mathbf{x}_{1j} + \boldsymbol{\zeta}_j. \quad (1)$$

Here,  $\boldsymbol{\alpha}$  is an intercept vector,  $\mathbf{B}$  a matrix of structural parameters governing the relations among the latent variables,  $\boldsymbol{\Gamma}$  a regression parameter matrix for regressions of latent variables on observed explanatory variables  $\mathbf{x}_{1j}$  and  $\boldsymbol{\zeta}_j$  a vector of disturbances (typically multivariate normal with zero mean). Note that this model is defined conditional on the observed explanatory variables  $\mathbf{x}_{1j}$ . Unlike conventional SEMs where all observed variables are treated as responses, we need not make any distributional assumptions regarding  $\mathbf{x}_{1j}$ .

In the example considered later, there is a single latent variable  $\eta_j$  representing mathematical reasoning or ‘ability’. This latent variable is regressed on observed covariates (gender, race and their interaction),

$$\eta_j = \alpha + \boldsymbol{\gamma}\mathbf{x}_{1j} + \zeta_j, \quad \zeta_j \sim N(0, \psi), \quad (2)$$

where  $\boldsymbol{\gamma}$  is a row-vector of regression parameters.

### *Measurement model*

The distinguishing feature of the measurement model is that it is specified for *latent* continuous responses  $\mathbf{y}_j^*$  in contrast to observed continuous responses  $\mathbf{y}_j$  as in conventional SEMs,

$$\mathbf{y}_j^* = \boldsymbol{\nu} + \boldsymbol{\Lambda}\boldsymbol{\eta}_j + \mathbf{K}\mathbf{x}_{2j} + \boldsymbol{\epsilon}_j. \quad (3)$$

Here  $\boldsymbol{\nu}$  is a vector of intercepts,  $\boldsymbol{\Lambda}$  a factor loading matrix and  $\boldsymbol{\epsilon}_j$  a vector of unique factors or ‘measurement errors’. Muthén and Muthén [7] extend the measurement model in Muthén [5] by including the term  $\mathbf{K}\mathbf{x}_{2j}$  where  $\mathbf{K}$  is a regression parameter matrix for the regression of  $\mathbf{y}_j^*$  on observed explanatory variables  $\mathbf{x}_{2j}$ . As in the structural model, we condition on  $\mathbf{x}_{2j}$ .

When  $\boldsymbol{\epsilon}_j$  is assumed to be multivariate normal, this model, combined with the threshold model described below, is a *probit* model (see **Probits**). The variances of the latent responses are not separately identified and some constraints are therefore imposed. Muthén sets the total variance of the latent responses (given the covariates) to 1.

### *Threshold model*

Each observed categorical response  $y_{ij}$  is related to a latent continuous response  $y_{ij}^*$  via a threshold model.

For ordinal observed responses it is assumed that

$$y_{ij} = \begin{cases} 0 & \text{if } -\infty < y_{ij}^* \leq \kappa_{1i} \\ 1 & \text{if } \kappa_{1i} < y_{ij}^* \leq \kappa_{2i} \\ \vdots & \vdots \\ S & \text{if } \kappa_{Si} < y_{ij}^* \leq \infty. \end{cases} \quad (4)$$

This is illustrated for three categories ( $S = 2$ ) in Figure 1 for normally distributed  $\epsilon_i$ , where the areas under the curve are the probabilities of the observed responses.

Either the constants  $\boldsymbol{\nu}$  or the thresholds  $\kappa_{1i}$  are typically set to 0 for identification. Dichotomous observed responses simply arise as the special case where  $S = 1$ .

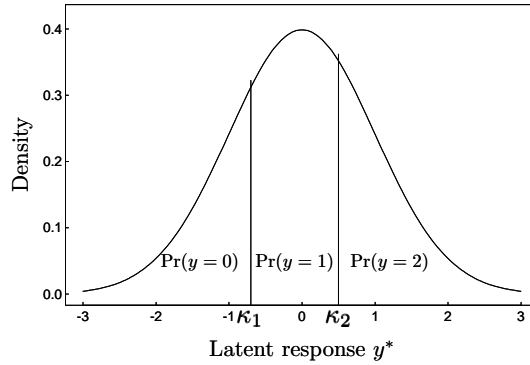


Figure 1: Threshold model for ordinal responses with three categories (from [13])

## Generalized latent variable models

In generalized latent variable models, the measurement model is a generalized linear model of the form

$$\mathbf{g}(\boldsymbol{\mu}_j) = \boldsymbol{\nu} + \boldsymbol{\Lambda}\boldsymbol{\eta}_j + \mathbf{K}\mathbf{x}_{2j}, \quad (5)$$

where  $\mathbf{g}(\cdot)$  is a vector of link functions which may be of different kinds handling mixed response types (for instance both continuous and dichotomous observed responses or ‘indicators’).  $\boldsymbol{\mu}_j$  is a vector of conditional means of the responses given  $\boldsymbol{\eta}_j$  and  $\mathbf{x}_{2j}$  and the other quantities are defined as in (3). The conditional models for the observed responses given  $\boldsymbol{\mu}_j$  are then distributions from the exponential family (see **Generalized Linear Models (GLM)**). Note that there are no explicit unique factors in the model because the variability of the responses for given values of  $\boldsymbol{\eta}_j$  and  $\mathbf{x}_{2j}$  is accommodated by the conditional response distributions. Also note that the responses are implicitly specified as conditionally independent given the latent variables  $\boldsymbol{\eta}_j$  (see **Conditional Independence**).

In the example, we will consider a single latent variable measured by four dichotomous indicators or ‘items’  $y_{ij}$ ,  $i=1, \dots, 4$ , and use models of the form

$$\text{logit}(\mu_{ij}) \equiv \ln \left( \frac{\Pr(\mu_{ij})}{1 - \Pr(\mu_{ij})} \right) = \nu_i + \lambda_i \eta_j, \quad \nu_1 = 0, \lambda_1 = 1. \quad (6)$$

These models are known as two-parameter logistic item response models because two parameters ( $\nu_i$  and  $\lambda_i$ ) are used for each item  $i$  and the logit link is used (see **Item Response Theory Models for Dichotomous Responses**). Conditional on the latent variable, the responses are Bernoulli distributed (see **Catalogue of Probability Density Functions**) with expectations  $\mu_{ij} = \Pr(y_{ij} = 1 | \eta_j)$ . Note that we have set  $\nu_1 = 0$  and  $\lambda_1 = 1$  for identification because the mean and variance of  $\eta_j$  are free parameters in (2). Using a probit link in the above model instead of the more commonly used logit would yield a model accommodated by the Muthén framework discussed in the previous section.

Models for counts can be specified using a log link and Poisson distribution (see **Catalogue of Probability Density Functions**). Importantly, many other response types can be handled including ordered and unordered categorical responses, rankings, durations, and mixed responses;

see e.g., [1], [2], [4], [9], [11], [12] and [13] for theory and applications. A recent book on generalized latent variable modeling [13] extends the models described here to ‘generalized linear latent and mixed models’ (GLLAMMs) [9] which can handle multilevel settings and discrete latent variables.

## Estimation and Software

In contrast to the case of multinormally distributed continuous responses, maximum likelihood estimation cannot be based on sufficient statistics such as the empirical covariance matrix (and possibly mean vector) of the observed responses. Instead, the likelihood must be obtained by somehow ‘integrating out’ the latent variables  $\boldsymbol{\eta}_j$ . Approaches which work well but are computationally demanding include adaptive Gaussian quadrature [10] implemented in `gllamm` [8] and Markov Chain Monte Carlo methods (typically with noninformative priors) implemented in BUGS [14] (see **Markov Chain Monte Carlo and Bayesian Statistics**).

For the special case of models with multinormal latent responses (principally probit models), Muthén suggested a computationally efficient limited information estimation approach [6] implemented in Mplus [7]. For instance, consider a structural equation model with dichotomous responses and no observed explanatory variables. Estimation then proceeds by first estimating ‘tetrachoric correlations’ (pairwise correlations between the latent responses). Secondly, the asymptotic covariance matrix of the tetrachoric correlations is estimated. Finally, the parameters of the SEM are estimated using weighted least squares (see **Least Squares Estimation**), fitting model-implied to estimated **tetrachoric correlations**. Here, the inverse of the asymptotic covariance matrix of the tetrachoric correlations serves as weight matrix.

Skrondal and Rabe-Hesketh [13] provide an extensive overview of estimation methods for SEMs with noncontinuous responses and related models.

## Example

### Data

We will analyze data from the Profile of American Youth (U.S. Department of Defense [15]), a survey of the aptitudes of a national probability sample of Americans aged 16 through 23. The responses (1: correct, 0: incorrect) for four items of the arithmetic reasoning test of the Armed Services Vocational Aptitude Battery (Form 8A) are shown in Table 1 for samples of white males and females and black males and females. These data were previously analyzed by Mislevy [3].

### Model specification

The most commonly used measurement model for ability is the two-parameter logistic model in (6) and (2) without covariates.

Item characteristic curves, plots of the probability of a correct response as a function of ability, are given by

$$\Pr(y_{ij} = 1 | \eta_j) = \frac{\exp(\nu_i + \lambda_i \eta_j)}{1 + \exp(\nu_i + \lambda_i \eta_j)}.$$

and shown for this model (using estimates under  $\mathcal{M}_1$  in Table 2) in Figure 2.

We then specify a structural model for ability  $\eta_j$ . Considering the covariates

- [Female]  $F_j$ , a dummy variable for subject  $j$  being female
- [Black]  $B_j$ , a dummy variable for subject  $j$  being black

Table 1: Arithmetic reasoning data

Item				White Males	White Females	Black Males	Black Females
Response	1	2	3				
0	0	0	0	23	20	27	29
0	0	0	1	5	8	5	8
0	0	1	0	12	14	15	7
0	0	1	1	2	2	3	3
0	1	0	0	16	20	16	14
0	1	0	1	3	5	5	5
0	1	1	0	6	11	4	6
0	1	1	1	1	7	3	0
1	0	0	0	22	23	15	14
1	0	0	1	6	8	10	10
1	0	1	0	7	9	8	11
1	0	1	1	19	6	1	2
1	1	0	0	21	18	7	19
1	1	0	1	11	15	9	5
1	1	1	0	23	20	10	8
1	1	1	1	86	42	2	4
Total:				263	228	140	145

Source: Mislevy [3]

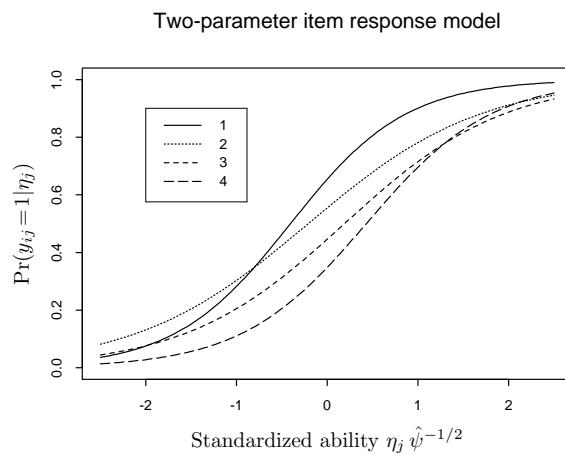


Figure 2: Item characteristic curves for items 1 to 4 (from [13])

we allow the mean abilities to differ between the four groups,

$$\eta_j = \alpha + \gamma_1 F_j + \gamma_2 B_j + \gamma_3 F_j B_j + \zeta_j.$$

This is a MIMIC model where the covariates affect the response via a latent variable only.

A path diagram of the structural equation model is shown in Figure 3. Here, observed variables are represented by rectangles whereas the latent variable is represented by a circle. Arrows represent regressions (not necessary linear) and short arrows residual variability (not necessarily an additive error term). All variables vary between subjects  $j$  and therefore the  $j$  subscripts are not shown.

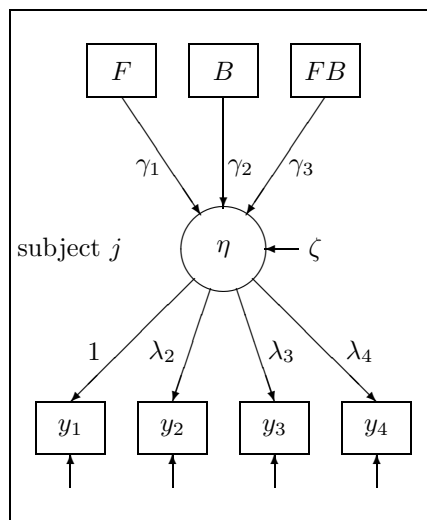


Figure 3: Path diagram of MIMIC model

We can also investigate if there are direct effects of the covariates on the responses, in addition to the indirect effects via the latent variable. This could be interpreted as ‘item bias’ or ‘differential item functioning’ (DIF), i.e., where the probability of responding correctly to an item differs for instance between black women and others with the same ability (see **Differential Item Functioning**). Such item bias would be a problem since it suggests that candidates cannot be fairly assessed by the test. For instance, if black women perform worse on the first item ( $i = 1$ ) we can specify the following model for this item:

$$\text{logit}[\text{Pr}(y_{1j} = 1|\eta_j)] = \beta_1 + \beta_5 F_j B_j + \lambda_1 \eta_j.$$

### Results

Table 2 gives maximum likelihood estimates based on 20-point adaptive quadrature estimated using `gllamm` [8]. (Note that the specified models are not accommodated in the Muthén framework because we are using a logit link.) Estimates for the two-parameter logistic IRT model (without covariates) are given under  $\mathcal{M}_1$ , for the MIMIC model under  $\mathcal{M}_2$  and for the MIMIC model with item bias for black women on the first item under  $\mathcal{M}_3$ . Deviance and Pearson  $X^2$  statistics are also reported in the table, from which we see that  $\mathcal{M}_2$  fits better than  $\mathcal{M}_1$ . The variance estimate of the disturbance decreases from 2.47 for  $\mathcal{M}_1$  to 1.88 for  $\mathcal{M}_2$  because some of the variability in ability is ‘explained’ by the covariates. There is some evidence for a [Female] by [Black] interaction. While being female is associated with lower ability among white people, this is not the case among

Table 2: Estimates for ability models

Parameter	$\mathcal{M}_1$		$\mathcal{M}_2$		$\mathcal{M}_3$	
	Est	(SE)	Est	(SE)	Est	(SE)
Intercepts						
$\nu_1$ [Item1]	0	–	0	–	0	–
$\nu_2$ [Item2]	–0.21	(0.12)	–0.22	(0.12)	–0.13	(0.13)
$\nu_3$ [Item3]	–0.68	(0.14)	–0.73	(0.14)	–0.57	(0.15)
$\nu_4$ [Item4]	–1.22	(0.19)	–1.16	(0.16)	–1.10	(0.18)
$\nu_5$ [Item1] $\times$ [Black] $\times$ [Female]	0	–	0	–	–1.07	(0.69)
Factor loadings						
$\lambda_1$ [Item1]	1	–	1	–	1	–
$\lambda_2$ [Item2]	0.67	(0.16)	0.69	(0.15)	0.64	(0.17)
$\lambda_3$ [Item3]	0.73	(0.18)	0.80	(0.18)	0.65	(0.14)
$\lambda_4$ [Item4]	0.93	(0.23)	0.88	(0.18)	0.81	(0.17)
Structural model						
$\alpha$ [Cons]	0.64	(0.12)	1.41	(0.21)	1.46	(0.23)
$\gamma_1$ [Female]	0	–	–0.61	(0.20)	–0.67	(0.22)
$\gamma_2$ [Black]	0	–	–1.65	(0.31)	–1.80	(0.34)
$\gamma_3$ [Black] $\times$ [Female]	0	–	0.66	(0.32)	2.09	(0.86)
$\psi$	2.47	(0.84)	1.88	(0.59)	2.27	(0.74)
Log-likelihood	–2002.76		–1956.25		–1954.89	
Deviance	204.69		111.68		108.96	
Pearson $X^2$	190.15		102.69		100.00	

Source: Skrandal and Rabe-Hesketh [13]

black people where males and females have similar abilities. Black people have lower mean abilities than both white men and white women. There is little evidence suggesting that item 1 functions differently for black females.

Note that none of the models appear to fit well according to absolute fit criteria (see **Model Fit: Assessment of**). For example, for  $\mathcal{M}_2$ , the deviance is 111.68 with 53 degrees of freedom, although the Table 1 is perhaps too sparse to rely on the  $\chi^2$  distribution.

## Conclusion

We have discussed generalized structural equation models for noncontinuous responses. Muthén suggested models for continuous, dichotomous, ordinal and censored (tobit) responses based on multivariate normal latent responses and introduced a limited information estimation approach for his model class.

Recently, considerably more general models have been introduced. These models handle (possibly mixes of) responses such as continuous, dichotomous, ordinal, counts, unordered categorical (polytomous), and rankings. The models can be estimated using maximum likelihood or Bayesian analysis.

## References

- [1] D. J. Bartholomew and M. Knott (1999). *Latent Variable Models and Factor Analysis*. Arnold, London.
- [2] J. P. Fox and C. A. W. Glas (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika*, 66:271–288.

- [3] R. J. Mislevy (1985). Estimation of latent group effects. *Journal of the American Statistical Association*, 80:993–997.
- [4] I. Moustaki (2003). A general class of latent variable models for ordinal manifest variables with covariate effects on the manifest and latent variables. *British Journal of Mathematical and Statistical Psychology*, 56:337–357.
- [5] B. O. Muthén (1984). A general structural equation model with dichotomous, ordered categorical and continuous latent indicators. *Psychometrika*, 49:115–132.
- [6] B. O. Muthén and A. Satorra (1996). Technical aspects of Muthén’s LISCOMP approach to estimation of latent variable relations with a comprehensive measurement model. *Psychometrika*, 60:489–503.
- [7] L. K. Muthén and B. O. Muthén (2004). *Mplus User’s Guide*. Muthén & Muthén, Los Angeles, CA.
- [8] S. Rabe-Hesketh, A. Skrondal and A. Pickles (2004a). GLLAMM Manual. *U.C. Berkeley Division of Biostatistics Working Paper Series*. Working Paper 160. Downloadable from <http://www.bepress.com/ucbbiostat/paper160/>
- [9] S. Rabe-Hesketh, A. Skrondal, and A. Pickles (2004b). Generalized multilevel structural equation modeling. *Psychometrika*, 69:167–190.
- [10] S. Rabe-Hesketh, A. Skrondal, and A. Pickles (2005). Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. *Journal of Econometrics*, in press.
- [11] M. D. Sammel, L. M. Ryan, and J. M. Legler (1997). Latent variable models for mixed discrete and continuous outcomes. *Journal of the Royal Statistical Society, Series B*, 59:667–678.
- [12] A. Skrondal and S. Rabe-Hesketh (2003). Multilevel logistic regression for polytomous data and rankings. *Psychometrika*, 68:267–287.
- [13] A. Skrondal and S. Rabe-Hesketh (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. Chapman & Hall/CRC, Boca Raton, FL.
- [14] D. J. Spiegelhalter, A. Thomas, N. G. Best, and W. R. Gilks (1996). *BUGS 0.5 Bayesian Analysis using Gibbs Sampling. Manual (version ii)*. MRC-Biostatistics Unit, Cambridge. Downloadable from <http://www.mrc-bsu.cam.ac.uk/bugs/documentation/contents.shtml>.
- [15] U. S. Department of Defense (1982). *Profile of American Youth*. Office of the Assistant Secretary of Defense for Manpower, Reserve Affairs, and Logistics, Washington DC.