General structural model – Part 2: Categorical variables and beyond

Psychology 588: Covariance structure and factor models

Categorical variables

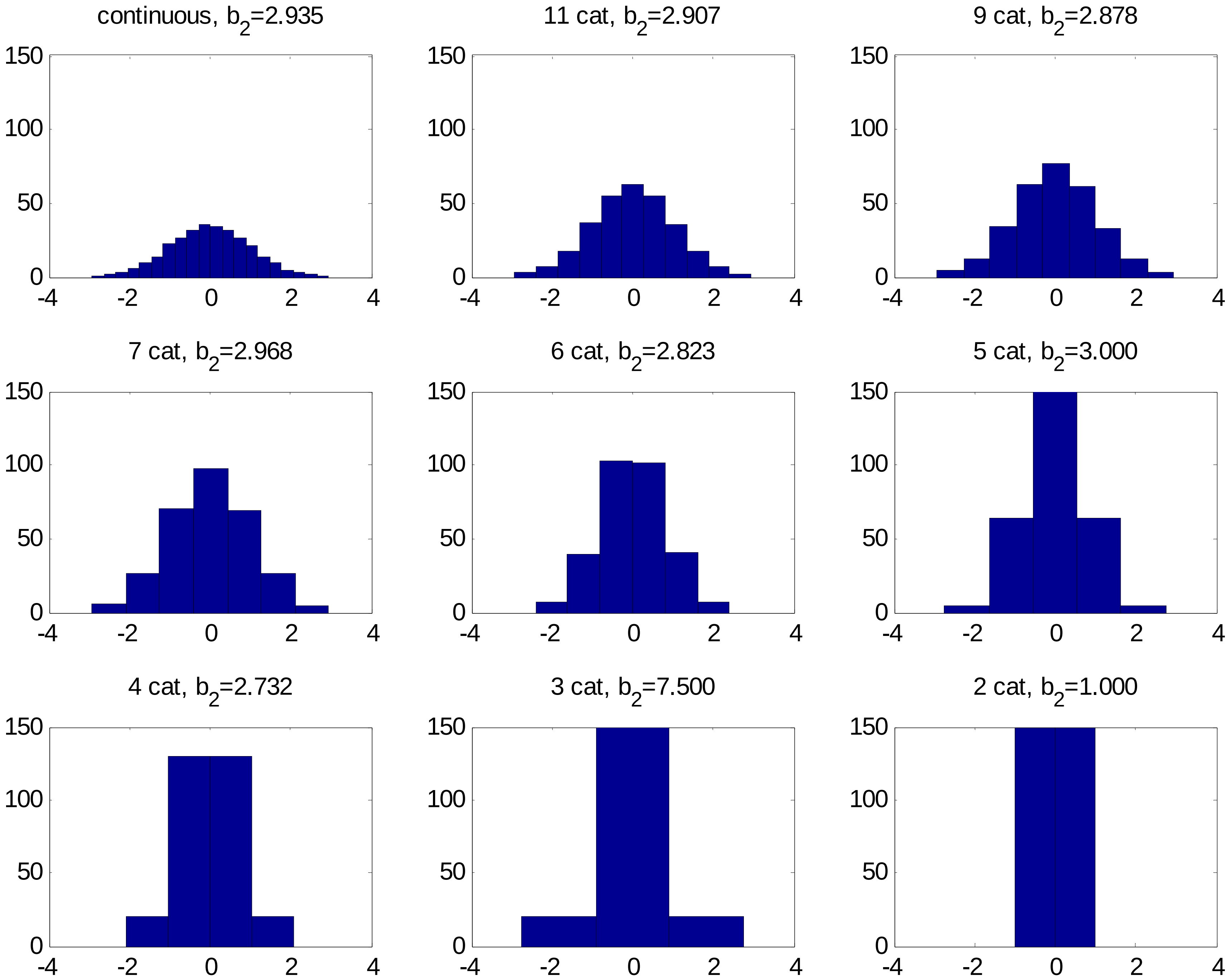
- Conventional (linear) SEM assumes <u>continuous</u> observed variables (except for exogenous $\mathbf{x}$) --- thus, SE modeling of categorical variables not fully justifiable

- Empirical (discretized) vs. conceptual categories:
  - Length measured in quarter-inch intervals
  - # of deaths for heart failure
  - Political affiliation, ethnicity
  - Color

- Dichotomies as quantitative variables --- dichotomous (and polytomous) variables used for "quantification" of nominal variables and any quantitative analysis/interpretation with them meaningful up to distinction of the categories

Why problem?

- Discretized variables are necessarily censored at the tails and center becomes taller with fewer categories --- deviation from normality gets severe with 2 or 3 categories
  - If continuous variable discretized, is it polytomous or ordinal?
- Crude measurement (too much rounding) --- increased measurement error
- Individual differences in where to put thresholds --- may create some systematic tendency (bias) or add more measurement error at best
- Following histograms show effects on kurtosis by even-interval categorization (N = 300)


continuous, b_2=2.935
11 cat, b_2=2.907
9 cat, b_2=2.878
7 cat, b_2=2.968
6 cat, b_2=2.823
5 cat, b_2=3.000
4 cat, b_2=2.732
3 cat, b_2=7.500
2 cat, b_2=1.000
150
100
50
0
-4
-2
0
2
4

Consequences

- Suppose a linear structure holds for true, unobserved continuous indicators $\mathbf{y}^*$ as:

$$\mathbf{y}^* = \boldsymbol{\Lambda}_y \boldsymbol{\eta} + \boldsymbol{\varepsilon}$$

then the categorized indicators $\mathbf{y}$ don't agree with the model:

$$\mathbf{y} \neq \boldsymbol{\Lambda}_y \boldsymbol{\eta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\Sigma} \neq \boldsymbol{\Sigma}(\boldsymbol{\theta}) \rightarrow \text{ biased } \hat{\boldsymbol{\theta}}$$

$$\text{acov}\left(s_{ij}, s_{gh}\right) \neq \text{acov}\left(s_{ij}^*, s_{gh}^*\right) \rightarrow \text{ invalid stat testing}$$

Simulation results

- Excessive kurtosis and skewness created by categorization result in too large chi-square (more rejection of correct parsimonious models than it should) and too large SE (more rejection of correct non-zero θ)

- Chi-square estimates tend to be more influenced by excessive kurtosis and skewness than by # of categories

- Generally coefficients (β and γ) and loadings are attenuated toward 0 --- in that categorization adds measurement errors

- When unobserved continuous indicators are highly correlated, categorization into few categories may artificially increase factorial complexity (resulting in correlated errors) --- since mis-classifying has a bigger consequence (than less correlated cases) and the consequence is likely to vary by variables

Correction of $\boldsymbol{\Sigma}$

- Assuming the unobserved, continuous $\mathbf{y}^*$ takes certain distributional form (most often normal), $\boldsymbol{\Sigma}^*$ (i.e., tetrachoric or polychoric correlations) may be estimated based on observed proportions at bivariate combinations of categories, by maximizing the likelihood:

$$\ln L = A + \sum_{i=1}^{c}\sum_{j=1}^{d} N_{ij} \ln\left(\pi_{ij}\right)$$

$$\pi_{ij} = \Phi_2\left(a_i, b_j\right) - \Phi_2\left(a_{i-1}, b_j\right) - \Phi_2\left(a_i, b_{j-1}\right) + \Phi_2\left(a_{i-1}, b_{j-1}\right)$$

where N_{ij} and π_{ij} are, respectively, frequency and probability at the ij-th category of y_1 and y_2; Φ_2 is CDF of bivariate normal distribution; and a_i and b_j are thresholds for the ij-th category

- Any continuous y is used as observed so that the entries of $\boldsymbol{\Sigma}^*$ are Pearson, polyserial (biserial) or polychoric (tetrachoric) correlations

- ML estimation of these correlations requires intensive computation --- thus, unstable with small samples

- Given $\boldsymbol{\Sigma}^*$, the usual SEM estimators will provide consistent estimates of $\boldsymbol{\theta}$, but WLS is recommended for correct statistical testing --- available in PRELIS (included in LISREL)

- See the examples, Tables 9.6 & 9.8

Nonlinear measurement models

- Relationship between observed and latent variables is defined as, e.g., the logistic or ogive function:

  - If $\mathbf{y}^*$ is normal, $\Pr(y < c)$ follows the normal CDF (ogive function) with varying central locations

  - Assuming only one latent variable, it becomes “graded item response” or “2 parameter logistic” model

  - The generalized latent variable modeling approach allows for such nonlinear relationships, along with other relationships for counts and duration (survival), by adopting the generalized linear modeling (GLM) approach --- offered e.g., by Mplus

- Comprehensive treatment of the generalized modeling approach --- Skrondal A. & Rabe-Hesketh S. (2004). *Generalized latent variable modeling*, CRC

- Short introduction --- Muthen B.O. (2002). Beyond SEM: General latent variable modeling. *Behaviormetrika, 29*, 81-117. (available in the course website)

Further developments of SEM

- Latent growth curve modeling
- Multilevel SEM for hierarchically designed data
- Categorical latent variables
  - When one latent categorical variable assumed with multiple categorical indicators, it becomes latent class model
  - More general modeling framework is what's known as "finite mixture" modeling --- possible with continuous indicators
  - It yields probabilistic membership as "latent variable scores"
  - Such idea of "latent clusters" can be applied to any SEM modeling approaches