JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

SEM for Categorical Outcomes

Statistics for Psychosocial Research II:
Structural Models

Qian-Li Xue

Outline

- Consequences of violating distributional assumptions with continuous observed variables
- SEM for categorical observed variables

Consequences of Violation of Multivariate Normality Assumption

	Properties of ML and GLS Estimators			
Observed Variable Dist.	Consistency	Asymptotic Efficiency	ACOV($\hat{\theta}$)	Chi-square Estimator
Multivariate Normal	Yes	Yes	Correct	Correct
No Kurtosis	Yes	Yes	Correct	Correct
"Arbitrary"	yes	no	Incorrect	incorrect

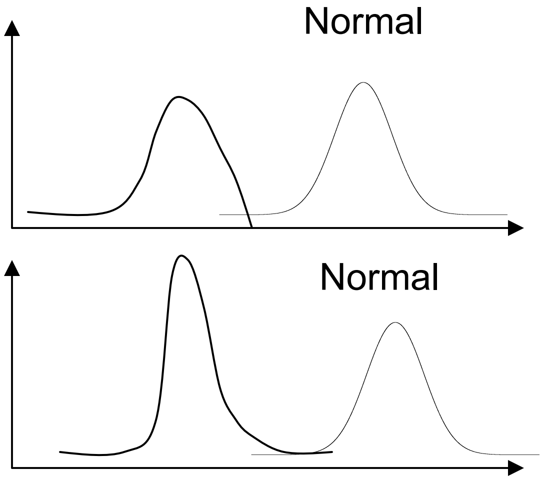
Adapted from Bollen's Structural Equations with Latent Variables, p. 416

Tests of Non-normality

- Definition
  - For a random variable X with a population mean of μ_1
  - The r[th] moment about the mean is

$$\mu_r = E(X - \mu_1)^r \quad \text{for } r > 1$$

	Population parameter	Sample Statistic	Normal Dist.
Skewness	$\dfrac{\mu_3}{(\mu_2)^{3/2}}$	$\dfrac{m_3}{(m_2)^{3/2}}$	0
Kurtosis	$\dfrac{\mu_4}{(\mu_2)^2}$	$\dfrac{m_4}{(m_2)^2}$	3



$$m_r = \Sigma\left(X - \bar{X}\right)^r / N$$

Tests of Non-normality

- Univariate Test
  - Calculate the first four sample moments of the observed variable
  - Calculate skewness and kurtosis based on these sample moments
  - Test H_0: skewness=0 and H_0: kurtosis=3 (D'Agostino, 1986, see Bollen, p.421)
  - Joint test of skewness and kurtosis equal to that of a normal distribution, i.e. H_0: skewness=0 and kurtosis=3 (if $N \geq 100$)
  - Using "sktest" command in STATA

- Multivariate Test for multivariate skewness and kurtosis
  - Using univariate tests with a Bonferroni adjustment based on the fact: Multivariate normality $\Rightarrow$ univariate normality
  - Mardia's multivariate test (see Bollen, pp.423-424) or
  - Use "multnorm" in STATA

Solutions for Non-normality

1. Transformation of the observed variables to achieve approximate normality
2. Post-estimation adjustments to the usual test statistics and standard errors (Browne, 1982, 1984)
3. Nonparametric tests via bootstrap resampling procedures

- However, neither 2 nor 3 corrects the lack of asymptotic efficiency of $\hat{\theta}$

A Better Solution for Non-normality

4. Weighted Least Squares (WLS) Estimators
   - To minimize the fitting function:

$$F_{WLS} = [s - \sigma(\theta)]' W^{-1} [s - \sigma(\theta)]$$

     where s is a vector of n(n+1)/2 non-redundant elements in S, $\sigma(\theta)$ si the vector of corresponding elements in $\Sigma(\theta)$, and W^{-1} is a $(n(n+1)/2)\times(n(n+1)/2)$ weight matrix

   - Optimal choice for W: asymptotic covariance matrix of the sample covariances (i.e. s)

   - With the optimal choice of W, the WLS fitting function is also termed "arbitrary distribution function (ADF)"

   - It can be shown that F_{GSL}, F_{MLS}, and F_{ULS} are special cases of F_{WLS}

Pros and Cons of the WLS Estimator

Pros

- Minimal assumptions about the distribution of the observed variables
- The WLS is a consistent and efficient estimator
- Provide valid estimates of asymptotic covariance matrix of $\hat{\theta}$ and a chi-square test statistic

Cons

- Computational burden
- Larger sample size requirement for convergence compared to other estimators
- Not clear about the degree to which WLS outperforms F_{GSL}, F_{MLS}, and F_{ULS} in the case of minor violation of normality

SEM with Categorical Observed Variables

- So far, we have assumed that the observed and latent variables are continuous
- What happens if we have observed variables taking ordinal or binary values?
- Are the estimators and significance tests for continuous variables still valid for categorical variables?
- We will deal with categorical latent variables in next lecture

Consequences of Using Ordinal Indicators as if They were Continuous

1. $y \neq \Lambda_y \eta + \varepsilon$

2. $x \neq \Lambda_x \xi + \delta$

3. $\Sigma \neq \Sigma(\theta)$

4. $ACOV(s_{ij}, s_{gh}) \neq ACOV(s_{ij}{}^{*}, s_{gh}{}^{*})$

Corrective Procedures for 1 and 2

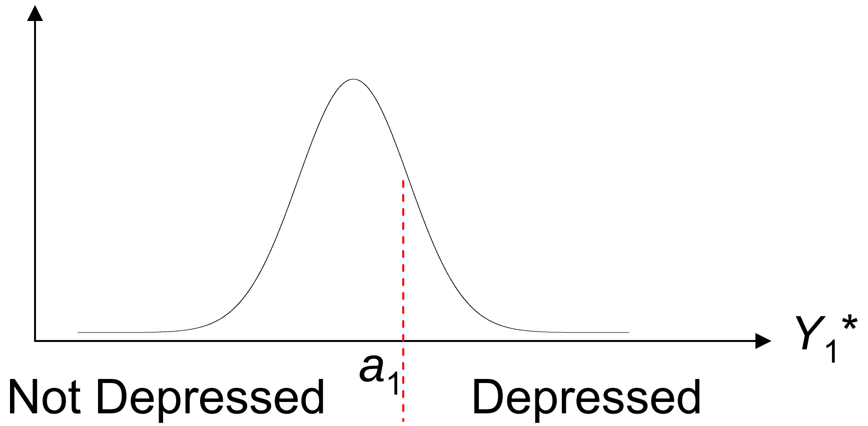
- Define a nonlinear function relating the observed categorical variables (y and/or x) to the latent continuous variables (y* and/or x*)

- Assume $y^{*} = \Lambda_{y}\eta + \varepsilon$ and $x^{*} = \Lambda_{x}\xi + \delta$

- For example,

$$y_1 = \begin{cases} 0 & \text{if } \mathrm{y}_1^{*} \le a_1 \\ 1 & \text{if } \mathrm{y}_1^{*} > a_1 \end{cases}$$

Where a_1 is the category threshold.



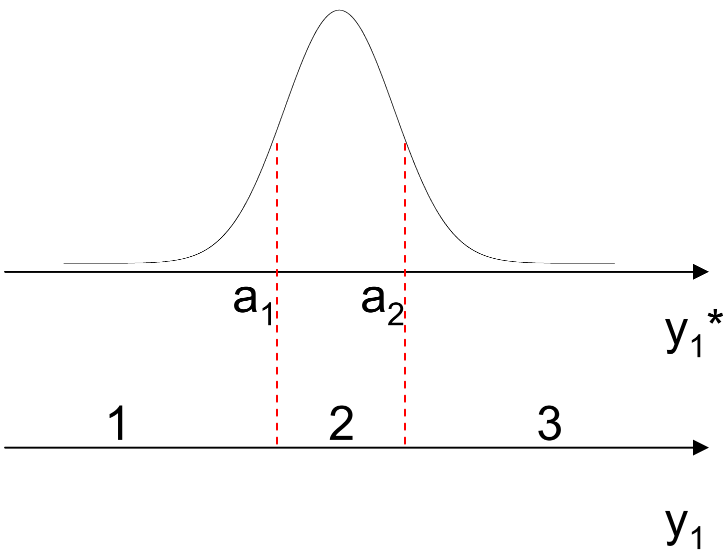
Corrective Procedures for 1 and 2

- In general, define

$$y_1 = \begin{cases} 1 & \text{if } y_1^* \le a_1 \\ 2 & \text{if } a_1 < y_1^* \le a_2 \\ \vdots & \\ c-1 & \text{if } a_{c-2} < y_1^* \le a_{c-1} \\ c & \text{if } y_1^* > a_{c-1} \end{cases}$$

Where c is the number of categories for y_1, a_i (i=1,2, …,c-1) is the category threshold, and y_1^* is the latent continuous indicator

Determine the Thresholds



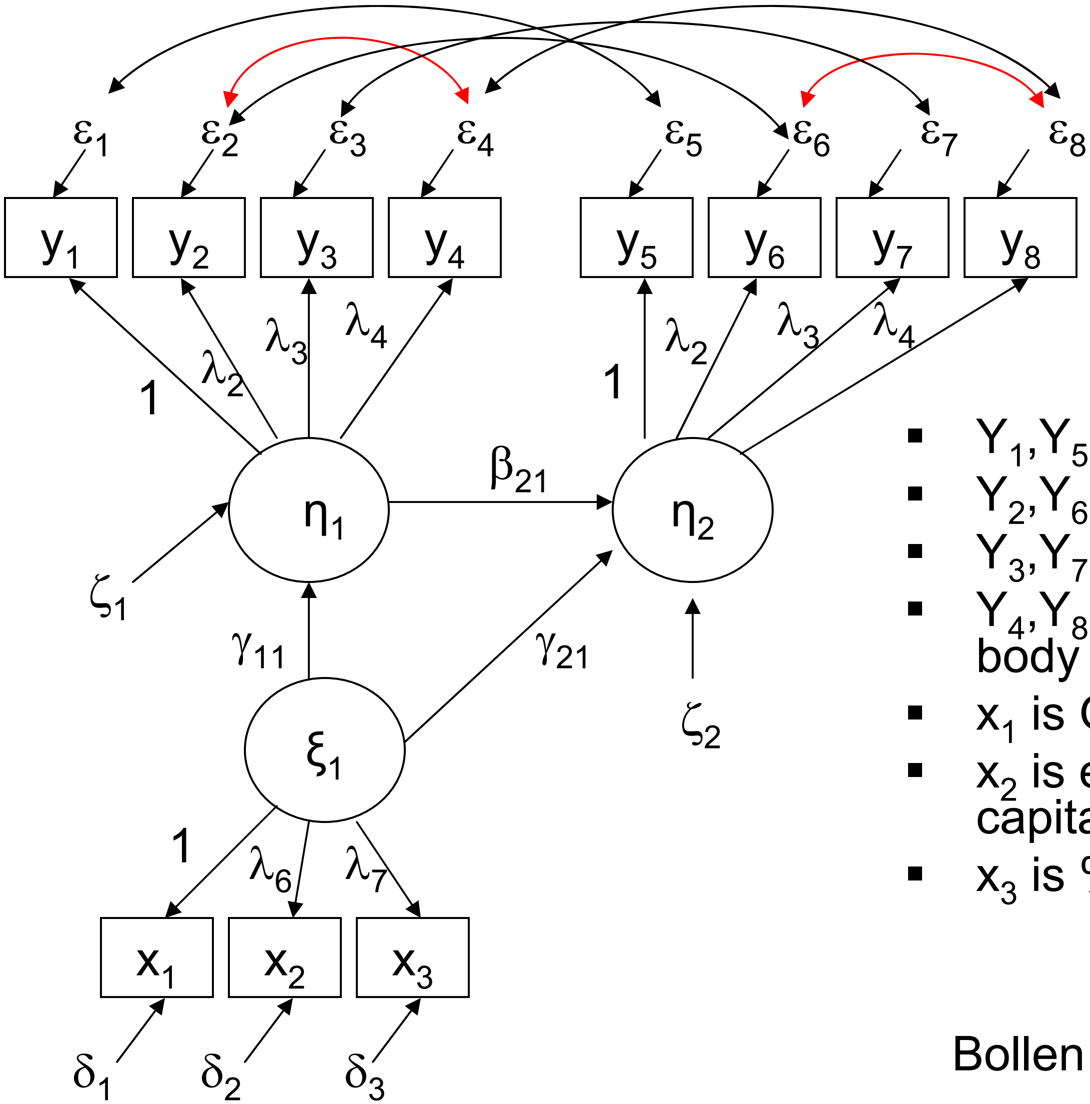
- y* and x* ~ multivariate normal
- Such that each variable of y* and x* ~ univariate normal
- Standardize each variable to a mean of 0 and a variance of 1
- An estimate of the threshold is:

$$a_i = \Phi^{-1}\left(\sum_{k=1}^{i} \frac{N_k}{N} \right)$$

- Where Φ is the standardized normal distribution function

Adapted from Bollen's Structural Equations with Latent Variables, p.440

Example: Industrialization and Political Democracy



- Y_1, Y_5: freedom of press
- Y_2, Y_6: freedom of group opposition
- Y_3, Y_7: fairness of election
- Y_4, Y_8: effectiveness of legislative body
- x_1 is GNP per capita
- x_2 is energy consumption per capita
- x_3 is % labor force

Bollen pp322-323

Determine the Threshold

- Consider a categorized version of the 1960 free press measure Y_1

	1	2	3	4	5	6	7	8
Frequency	8	13	5	13	5	22	4	5
Proportion	0.11	0.17	0.07	0.17	0.07	0.29	0.05	0.07
Cum.Prop.	0.11	0.28	0.35	0.52	0.59	0.88	0.93	1.00

Threshold	a_1	a_2	a_3	a_4	a_5	a_6	a_7
Estimate	-1.24	-0.58	-0.39	0.05	0.22	1.17	1.50

Adapted from Bollen’s Structural Equations with Latent Variables, p. 440-441

Corrective Procedures for 3 (i.e. $\Sigma \neq \Sigma(\theta)$)

- Assume:
  - $\Sigma^* = \Sigma(\theta)$, where Σ^* is the covariance matrix of y^* and x^*
  - y^* and x^* ~ multivariate normal
- Idea: estimate correlation between each pair of latent variables y_i^* and x_j^*
- If are both y_i and x_j are continuous, calculate Pearson correlation
- If are both y_i and x_j are ordinal, calculate polychoric correlation between y_i^* and x_j^*
  - If are both y_i and x_j are binary, calculate tetrachoric correlation between y_i^* and x_j^*
- If one is ordinal and the other is continuous, calculate polyserial correlation between y_i^* and x_j^*

Pros and Cons of Polychoric and Tetrachoric Correlation (Pearson, 1901)

Pros

- In a familiar form of a correlation coefficient
- Separately quantify association and similarity of category definitions
- Independent of number of categories
- Assumptions underlying the polychoric and tetrachoric correlation can be easily tested
- Estimation software is routinely available

Cons

- Model assumptions are not always appropriate
- With only two variables, the assumptions of the tetrachoric correlation can not be tested

(Uebersax JS)

Maximum Likelihood Estimation of the Polychoric Correlation

- For example, the log likelihood for estimation of the polychoric correlation based on a I×J table of two ordinal variables x and y is

$$\ln L = \sum_{i=1}^{I} \sum_{j=1}^{J} N_{ij} \ln(\pi_{ij}) + C$$

$$\pi_{ij} = \Phi_2(a_i, b_j) - \Phi_2(a_{i-1}, b_j) -$$

$$\Phi_2(a_i, b_{j-1}) + \Phi_2(a_{i-1}, b_{j-1})$$

		y		
		1	2	3
x	1			
	2			
	3			

where N_{ij} is the frequency of observations in the *i*th and *j*th categories, C is a constant, a_i and b_j are thresholds for x and y, respectively, and Φ_2 is the bivariate normal distribution function with correlation ρ

- An iterative search algorithm tries different combinations for ai, bj and ρ to find a “optimal” combination for minimizing the difference between the expected counts to the observed counts

(Olsson, 1979)

A Few Important Facts

- The polychoric correlation matrix Σ_p based on y and x is a consistent estimator of Σ^*
- Analysis of Σ_p via F_{ML}, F_{GLS}, or F_{ULS} yields consistent estimators of θ
- However, standard errors, significant tests (e.g. chi-square tests) are incorrect!!
- A better choice is F_{WLS}:

$$F_{WLS} = [\hat{\rho} - \sigma(\theta)]' W^{-1} [\hat{\rho} - \sigma(\theta)]$$

where $\hat{\rho}$ is $[n(n+1)/2] \times 1$ vector of the polychoric correlations, $\sigma(\theta)$ is the implied covariance matrix, and W is the asymptotic covariance matrix of $\hat{\rho}$ (Muthen, 1984).

MPLUS Fitting of CFA with Categorical Indicators

```
TITLE: this is an example of a CFA with
categorical factor indicators
DATA: FILE IS ex5.2.dat;
VARIABLE: NAMES ARE u1-u6;
CATEGORICAL ARE u1-u6;
MODEL: f1 BY u1-u3;
f2 BY u4-u6;
```

U1-u6 are binary indicators

Declare U1-u6 to be categorical indicators

The default estimator is robust weighted least squares estimator

MPLUS Fitting of CFA with Continuous and Categorical Indicators

TITLE: this is an example of a CFA with continuous and categorical factor indicators
DATA: FILE IS ex5.3.dat;
VARIABLE: NAMES ARE u1-u3 y4-y6;
CATEGORICAL ARE u1 u2 u3;
MODEL: f1 BY u1-u3;
f2 BY y4-y6;

By default, MPLUS treats y4-u6 as continuous indicators

Declare only u1-u3 to be categorical indicators

Example: Frailty and Disability

- Study Population: Women's Health and Aging Studies I; N = 1002
- Community-dwelling women 65-101 yrs;
- Represent one-third most disabled women
- Outcome:
  - Frailty by 5 binary indicators
  - Disability by 5 4-level ordinal indicators
- Predictor:
  - Age, education, disease burden

Outcome Definitions

Frailty	**Mobility Disability**
Binary Criteria:	Ordinal Criteria:
Shrinking (weight loss)	Walk ¼ mile
Weakness	Climb up 10 steps
Poor endurance	Lift 10 lbs
Slowed walking speed	Transfer from bed to chair
Low physical activity	Heavy housework
Classification:	Each rated on a four-point scale:
Non-frail: 0/5 criteria	0 – no difficulty
Pre-frail: 1 or 2/5 criteria	1 – a little difficulty
Frail: 3,4, or 5/5 criteria	2 – some difficulty
	3 – a lot of difficulty/unable

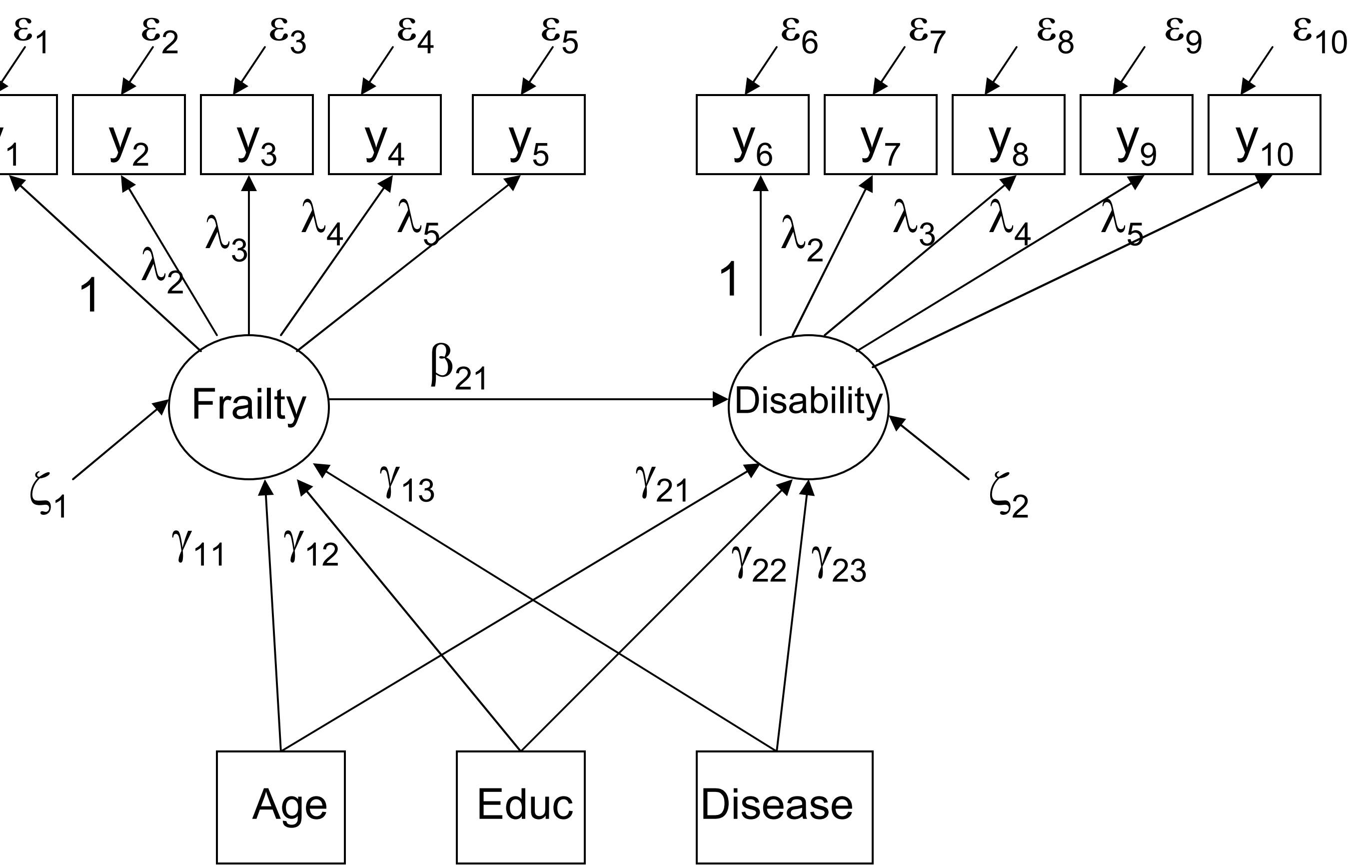
Example: Frailty and Disability

- Study Aims
  1) Evaluate the association between frailty and mobility disability

  2) Study potential risk factors of frailty and mobility disability
     - Age, education, number of chronic diseases

  3) Assess racial differences in 1) and 2)

Example: Frailty and Disability



Example: Measurement Model for Mobility

By default, MPLUS sets loadings and thresholds to be the same across groups (i.e. a more restricted model)

```
TITLE:     this is an example of a multiple group CFA
        with categorical factor indicators for mobility disability
        and a threshold structure
DATA:FILE IS c:\teaching\140.658.2007\catna.dat;
        VARIABLE: NAMES ARE baseid age race educ disease
             shrink strength speed exhaust physical
             lift walk stairs transfer hhw;
        USEVARIABLES ARE race lift-hhw;
        CATEGORICAL ARE lift-hhw;
        GROUPING IS race (0=white 1=black);
ANALYSIS: TYPE = MEANSTRUCTURE;
        DIFFTEST IS c:\teaching\140.658.2007\deriv.dat;
MODEL:
        mobility BY lift* walk@1 stairs-hhw;
OUTPUT: SAMPSTAT;
```

See output file: catcfad1.out

Example: Measurement Model for Mobility

Set loadings and thresholds for lift, stairs, and hhw to be different across groups (i.e. a less restricted model)

```
… (SAME AS BEFORE)
ANALYSIS: TYPE = MEANSTRUCTURE;
        DIFFTEST IS c:\teaching\140.658.2007\deriv.dat;
MODEL:
        mobility BY lift* walk@1 stairs-hhw;
MODEL black:
        mobility BY lift;
        [lift$1 lift$2 lift$3];
        {lift@1};
        mobility BY stairs;
        [stairs$1 stairs$2 stairs$3];
        {stairs@1}
        mobility BY transfer;
        [transfer$1 transfer$2 transfer$3];
        {transfer@1};
        mobility BY hhw;
        [hhw$1 hhw$2 hhw$3];
        {hhw@1};
SAVEDATA: DIFFTEST is c:\teaching\140.658.2007\deriv.dat;
OUTPUT: SAMPSTAT;
```

See output file: catcfad.out

Example: Structural Models for Mobility and Frailty

```
TITLE: this is an example of a multiple group CFA with covariates and
categorical factor indicators for mobility and frailty and a threshold structure
DATA:           FILE IS c:\teaching\140.658.2007\catna.dat;
VARIABLE: NAMES ARE baseid age race educ disease
          shrink strength speed exhaust physical
          lift walk stairs transfer hhw;
   USEVARIABLES ARE race age educ disease
          shrink-hhw;
        CATEGORICAL ARE shrink-hhw;
        GROUPING IS race (0=white 1=black);
ANALYSIS: TYPE = MEANSTRUCTURE;
MODEL:
   frailty BY shrink-physical;
   mobility BY lift* walk@1 stairs-hhw;
   mobility ON frailty;
   mobility frailty ON age educ disease;

MODEL black:
   mobility BY lift;
   [lift$1 lift$2 lift$3];
   {lift@1};
   mobility BY stairs;
   [stairs$1 stairs$2 stairs$3];
   {stairs@1};
   mobility BY transfer;
   [transfer$1 transfer$2 transfer$3];
   {transfer@1};
   mobility BY hhw;
   [hhw$1 hhw$2 hhw$3];
   {hhw@1};

   frailty BY strength;
   [strength$1];
   {strength@1};
```

See output file: catreg.out