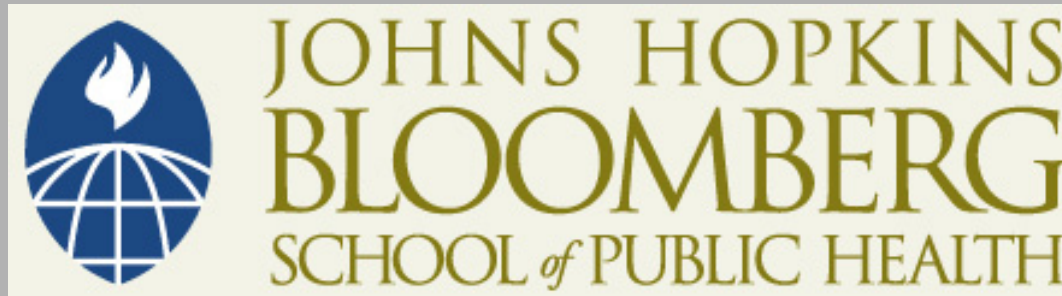


This work is licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike License](https://creativecommons.org/licenses/by-nc-sa/4.0/). Your use of this material constitutes acceptance of that license and the conditions of use of materials on this site.



Copyright 2007, The Johns Hopkins University and Qian-Li Xue. All rights reserved. Use of these materials permitted only in accordance with license rights granted. Materials provided "AS IS"; no representations or warranties provided. User assumes all responsibility for use, and all liability related thereto, and must independently review all materials for accuracy and efficacy. May contain materials owned by others. User is responsible for obtaining permissions for use from third parties as needed.

Latent Class Regression Part I

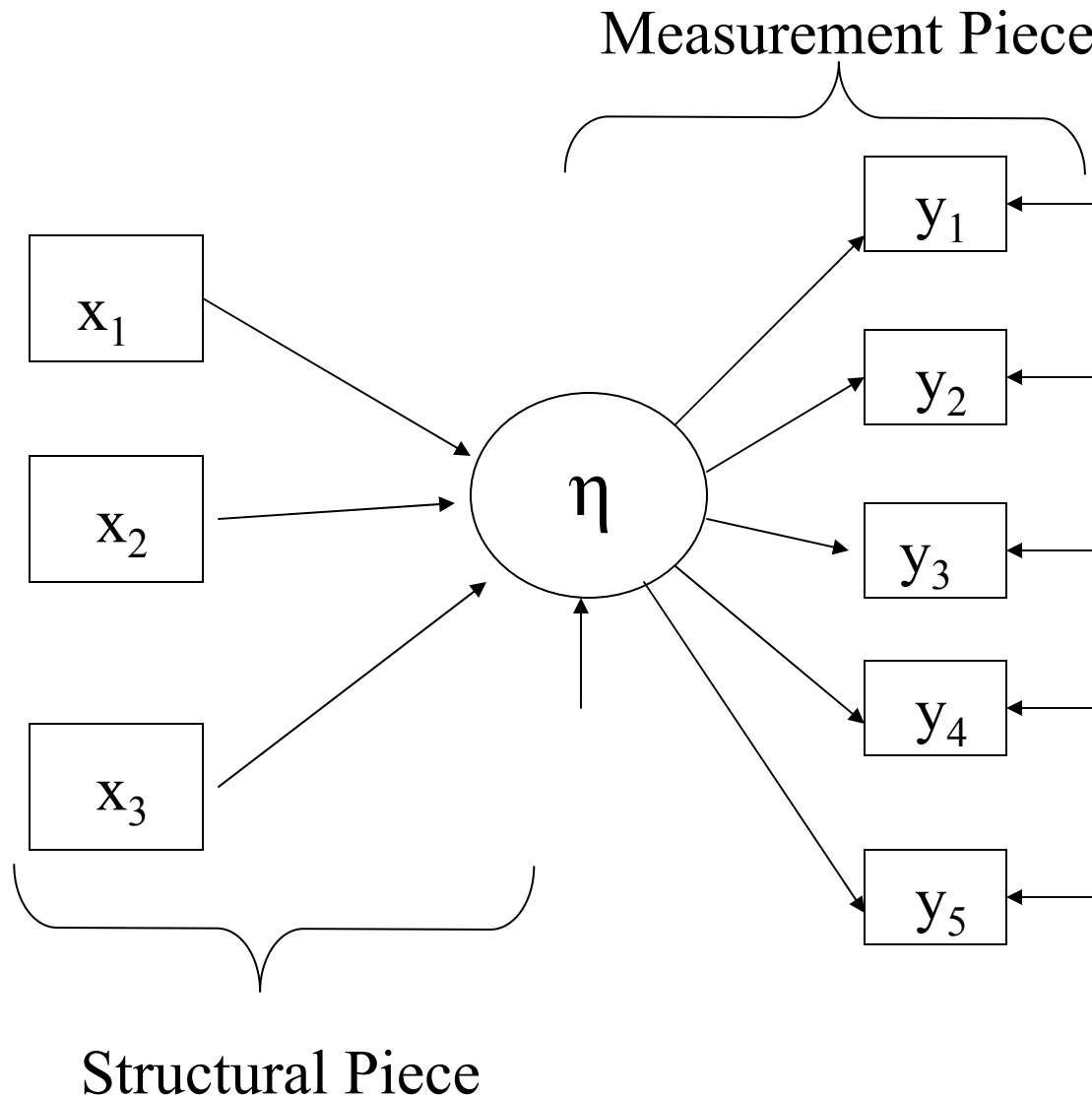
Statistics for Psychosocial Research II:
Structural Models

Qian-Li Xue

Latent Class Regression (LCR)

- What is it and when do we use it?
- Recall the standard latent class model from last term:
 - Items measure “diagnoses” rather than underlying scores
 - Patterns of responses are thought to contain information above and beyond “aggregation” of responses
 - The goal is “clustering” individuals rather than response variables
- We add “structural” piece to model where covariates “predict” class membership

Structural Equation-Type Depiction



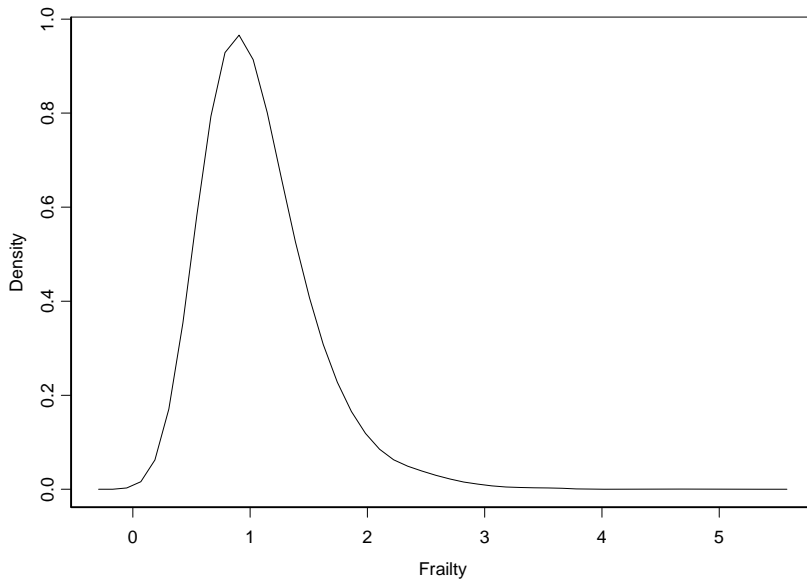
When to use LCR

- Multiple discrete outcome variables
 - Binary examples
 - ❖ yes/no questions
 - ❖ present/absent symptoms
 - All measuring same latent construct
 - We want to use construct as outcome variable
 - Responses to questions/items measure underlying states (i.e. classes) with error
- NOT appropriate for...
 - counts or other way of grouping response patterns
 - responses measure underlying score with error
- **Note: Latent Variable is DISCRETE**

Example: Frailty

Is frailty continuous or categorical?

- Latent trait (IRT) assumes it is continuous.
- Latent class model assumes it is discrete



	<u>%</u>
Robust	80
Intermediate Frail	15
Frail	5

Latent Class Regression: Notation

- $Y_i = (Y_{i1}, Y_{i2}, \dots, Y_{iM})$: a vector of M binary outcome indicators (e.g. presence (=1)/absence (=0) of symptom m , $m=1, \dots, M$) for individual i , $i = 1, \dots, N$.
- η_i : the true latent class of individual i ,
- $P_{ij}(x_i) = \Pr(\eta_i = j | x_i)$: probability of individual i with covariate vector x_i in class j , $j=1, \dots, J$
- $\sum_j P_{ij}(x_i) = 1$
- $\pi_{m|j} = \Pr(Y_{im} = 1 | \eta_i = j)$: probability of reporting symptom m given latent class j ($j = 1, \dots, J$)

Assumptions

- **Conditional Independence:**
 - given an individual's depression class, his symptoms are independent
 - $\Pr(y_{im}, y_{in} | \eta_i) = \Pr(y_{im} | \eta_i) \Pr(y_{in} | \eta_i)$
- **Non-differential Measurement:**
 - given an individual's depression class, covariates are not associated with symptoms
 - $\Pr(y_{im} | x_i, \eta_i) = \Pr(y_{im} | \eta_i)$

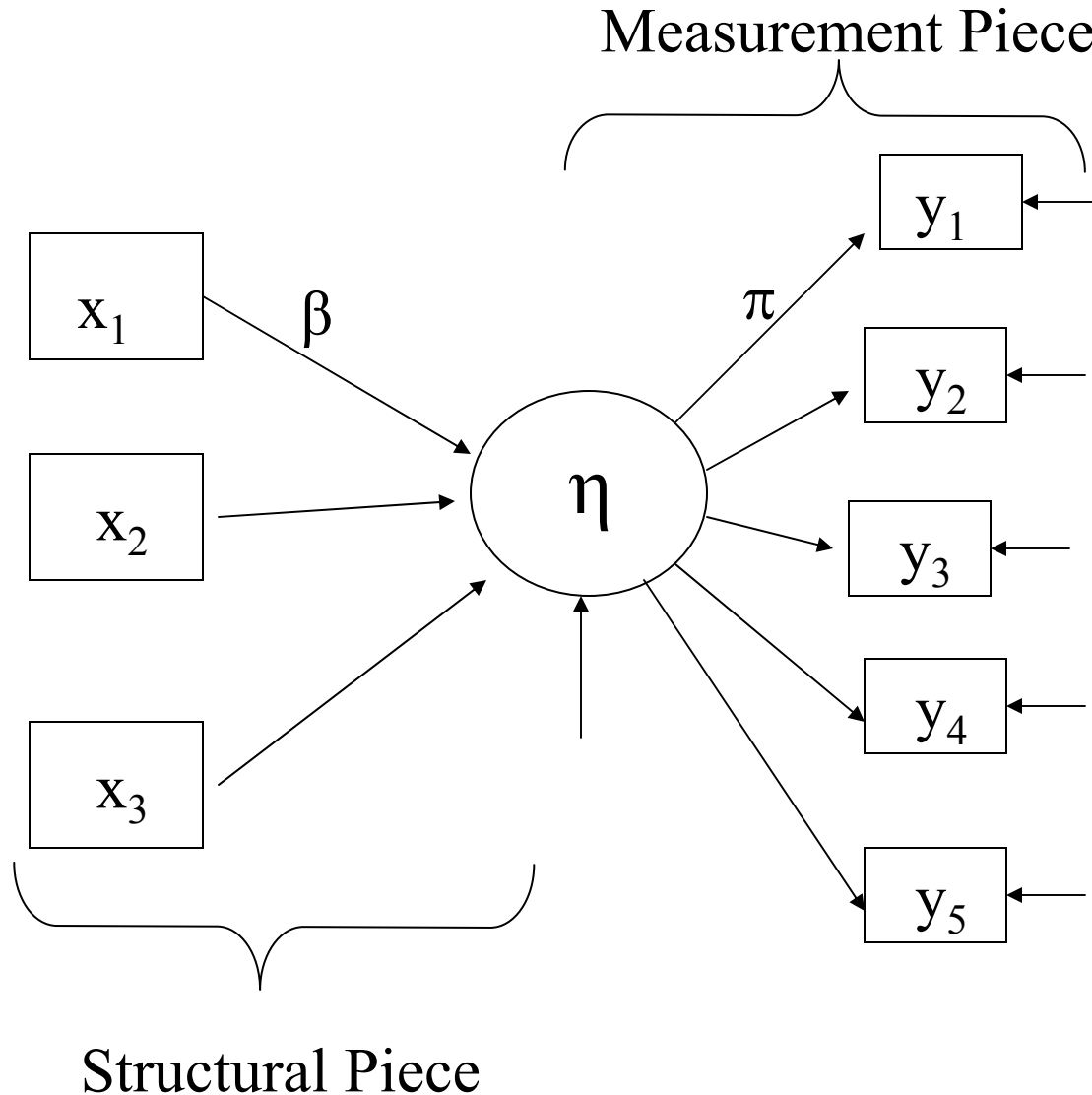
Why LCR may be better than another analytic method

- LCR versus using counts (e.g. number of symptoms)
 - Pros:
 - ❖ distinguishes meaningful patterns from trivially different ones which may be hard to discern empirically
 - ❖ acknowledges measurement error
 - Cons:
 - ❖ may overdistinguish prevalent patterns and mask differences in rare ones
 - ❖ violation of assumptions make inferences invalid

Why LCR may be better than another analytic method (continued)

- Versus factor-type methods
 - Pros:
 - ❖ less severe assumptions (statistically)
 - ❖ easier to check assumptions
 - Cons:
 - ❖ lose statistical power if construct is actually continuous
 - ❖ estimability harder to achieve (need big sample)
- Practically
 - Pro:
 - ❖ Allows for disease/disorder classification which is useful in a treatment vs. no treatment setting

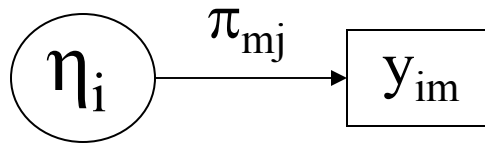
Structural Equation-Type Depiction



What are the parameters that the arrows represent? In other words, what are β and π in the LCR model?

Parameter Interpretation

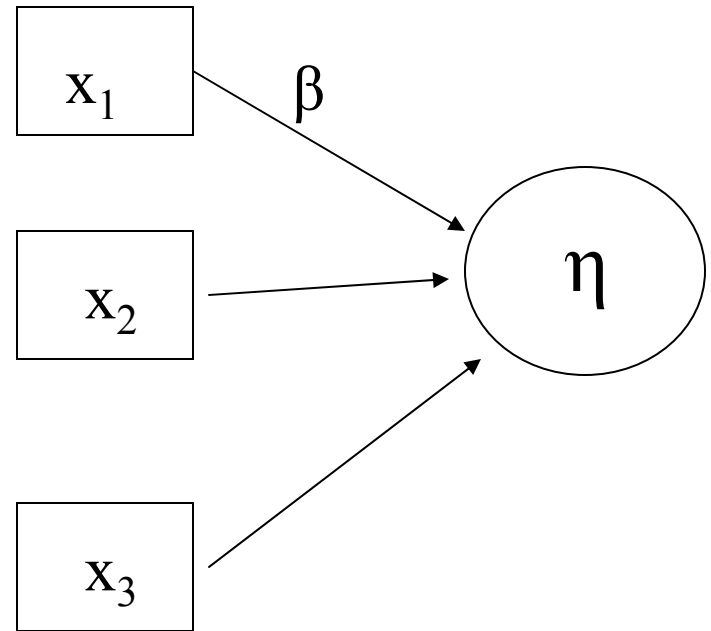
- Measurement Piece (p's)
 - π_{mj} : probability that an individual from class j reports symptom m .



- Same as standard latent class model from last term

Parameter Interpretation

- How do we relate η 's and β 's?
- In “classic” SEM, we have linear model
- What about when η is categorical?
- What if η is binary?



Parameter Interpretation

- How do we relate η_i to x_i 's ?
- Consider simplest case: 2 classes (1 vs. 2)

$$\log\left(\frac{\Pr(\eta_i = 2 | x_i)}{1 - \Pr(\eta_i = 2 | x_i)}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$$

or equivalently,

$$\log\left(\frac{p_{i2}(x_i)}{p_{i1}(x_i)}\right) = \log\left(\frac{\Pr(\eta_i = 2 | x_i)}{\Pr(\eta_i = 1 | x_i)}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$$

β_1 and β_2 are log odds ratios

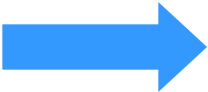
Model Results

- π_{mj}
 - Probability of response given in class j
 - $M \times J$ π 's
- $p_{ij}(x_i) = Pr(\eta_i = j | x_i)$
 - Conditional on x 's
 - No longer 'proportion of individuals in class'
 - Now, only can interpret to mean 'probability of class membership given covariates for individual i '
 - To get size of class j , can sum of p_{ij} for all i
- β
 - $(J-1) \times (H+1)$ β 's where H = number of covariates
 - $J-1$: one class is reference class so all of its β coefficients are technically zero
 - $H+1$: for each class, there is one β for each covariate plus another for the intercept.

Solving for $p_{ij}(x_i) = \Pr(\eta_i=j)$

$$\log\left(\frac{p_{i2}(x_i)}{p_{i1}(x_i)}\right) = \log\left(\frac{\Pr(\eta_i = 2 | x_{1i}, x_{2i})}{\Pr(\eta_i = 1 | x_{1i}, x_{2i})}\right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$$

Using the factor: $p_{i1} + p_{i2} = 1$, we obtain


$$p_{i2} = \Pr(\eta_i = 2 | x_{i1}, x_{i2}) = \frac{e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}}}{1 + e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}}}$$

$$p_{i1} = \Pr(\eta_i = 1 | x_{i1}, x_{i2}) = \frac{1}{1 + e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}}}$$

Parameter Interpretation

Example: $\exp(\beta_1) = 2$ and $x_{i1} = 1$ if female, 0 if male

“Women have twice the odds of being in class 2 versus class 1 than men, holding all else constant”

$$e^{\beta_1} = \frac{\Pr(\eta_i = 2 \mid x_{i1} = 1, x_{i2} = c)}{\Pr(\eta_i = 1 \mid x_{i1} = 1, x_{i2} = c)} \bigg/ \frac{\Pr(\eta_i = 2 \mid x_{i1} = 0, x_{i2} = c)}{\Pr(\eta_i = 1 \mid x_{i1} = 0, x_{i2} = c)}$$

More than two classes?

Need more than one equation

Need to choose a reference class

$$\log\left(\frac{\Pr(\eta_i = 2 \mid x_{i1}, x_{i2})}{\Pr(\eta_i = 1 \mid x_{i1}, x_{i2})}\right) = \beta_{02} + \beta_{12}x_{i1} + \beta_{22}x_{i2}$$

$$\log\left(\frac{\Pr(\eta_i = 3 \mid x_{i1}, x_{i2})}{\Pr(\eta_i = 1 \mid x_{i1}, x_{i2})}\right) = \beta_{03} + \beta_{13}x_{i1} + \beta_{23}x_{i2}$$

$e^{\beta_{12}}$ = OR for class 2 versus class 1 for females versus males

$e^{\beta_{13}}$ = OR for class 3 versus class 1 for females versus males

$e^{\beta_{13}} / e^{\beta_{12}} = e^{\beta_{13} - \beta_{12}}$ = OR for class 3 versus class 2 for

females versus males

Solving for $p_{ij}(x_i) = p(\eta_i|x_i)$

$$\log\left(\frac{p_{i2}(x_i)}{p_{i1}(x_i)}\right) = \log\left(\frac{\Pr(\eta_i = 2)}{\Pr(\eta_i = 1)}\right) = \beta_{02} + \beta_{12}x_{i1} + \beta_{22}x_{i2}$$



$$p_{i2} = \Pr(\eta_i = 2 | x_{i1}, x_{i2}) = \frac{e^{\beta_{02} + \beta_{12}x_{i1} + \beta_{22}x_{i2}}}{1 + e^{\beta_{02} + \beta_{12}x_{i1} + \beta_{22}x_{i2}} + e^{\beta_{03} + \beta_{13}x_{i1} + \beta_{23}x_{i2}}}$$
$$= \frac{e^{\beta_{02} + \beta_{12}x_{i1} + \beta_{22}x_{i2}}}{\sum_{j=1}^3 e^{\beta_{0j} + \beta_{1j}x_{i1} + \beta_{2j}x_{i2}}}$$

$$\beta_{01} = \beta_{11} = \beta_{21} = 0$$

Model Building

- Step 1:
 - Get the measurement part right!
 - Fit standard latent class model first.
 - Use methods we discussed last term to choose appropriate model (e.g. number of classes)
- Step 2:
 - add covariates one at a time
 - It is useful to perform “simple” regressions to see how each covariate is associated with latent variable before adjusting for others.
 - Many of same issues in linear and logistic regression (e.g. multicollinearity)

Model Estimation

- Maximum likelihood estimator
- Standard LCM Likelihood

$$\begin{aligned}\Pr(Y_i = y_i) &= \Pr(Y_{i1} = y_{i1}, Y_{i2} = y_{i2}, Y_{i3} = y_{i3}, Y_{i4} = y_{i4}, Y_{i5} = y_{i5}) \\ &= \sum_{j=1}^J p_{ij} \prod_{m=1}^M \pi_{m|j}^{y_{im}} (1 - \pi_{m|j})^{(1-y_{im})}\end{aligned}$$

- Latent Class Regression Likelihood

$$\begin{aligned}\Pr(Y_i = y_i) &= \Pr(Y_{i1} = y_{i1}, Y_{i2} = y_{i2}, Y_{i3} = y_{i3}, Y_{i4} = y_{i4}, Y_{i5} = y_{i5} \mid x) \\ &= \sum_{j=1}^J p_{ij}(x) \prod_{m=1}^M \pi_{m|j}^{y_{im}} (1 - \pi_{m|j})^{(1-y_{im})}\end{aligned}$$

$$\text{where } p_{ij}(x_i) = \frac{e^{\beta_j x_i}}{\sum_{j=1}^J e^{\beta_j x_i}}$$

Properties of Estimates (β , π)

- If N is large, coefficients are approximately normal \Rightarrow confidence intervals and Z-tests are appropriate.
- Nested models can be compared by using chi-square test.
- But, recall problems of chi-square test when sample size is large!
- And problems when the sample size is small!
- Also can use AIC, BIC, etc. to compare nested AND non-nested models (e.g. is age as continuous better than 3 age categories).

Example: Frailty

- Study Population: Women's Health and Aging Studies I and II; N = 829
- Community-dwelling women 70-79 yrs
- Represent whole spectrum of functional status of older women
- Outcome:
 - Frailty by 5 binary indicators
- Predictor:
 - Age, education, disease burden

Frailty Phenotype

Frailty

Binary Criteria:

Shrinking (weight loss)

Weakness

Poor endurance

Slowed walking speed

Low physical activity

Clinical Classification:

Non-frail: 0/5 criteria

Pre-frail: 1 or 2/5 criteria

Frail: 3,4, or 5/5 criteria

Step 1. Measurement Model

- To evaluate our measure's convergent validity
 - Run LCA to determine
 - ❖ the number of such subpopulations ("classes")
 - ❖ each subpopulation's prevalence in the overall population
 - ❖ per subpopulation and criterion, the proportion having the criterion (π - "conditional probabilities")
- To validate the hypothesis that our frailty criteria are syndromic in their occurrence by
 - Examining the number of latent classes and patterns of conditional probabilities across classes

MPLUS fitting of LCA

TITLE: Weighted Latent Class Analysis of Frailty Components Using Combined WHAS I and II Data Age 70-79

DATA:

FILE IS "c:\whas\frail\paper\lcaw.txt";

VARIABLE:

NAMES ARE baseid shrink weak slow exhaust kcal sweight;
USEVARIABLES ARE shrink weak slow exhaust kcal sweight;
MISSING ARE ALL (999999);
CATEGORICAL ARE shrink-kcal;
CLASSES = frailty(2);

WEIGHT IS sweight;

ANALYSIS:

TYPE IS MIXTURE MISSING;

MODEL:

%OVERALL%

%frailty#1%

[shrink\$1*-1 weak\$1*-1 slow\$1*0 exhaust\$1*-1 kcal\$1*-1];

%frailty#2%

[shrink\$1*1 weak\$1*1 slow\$1*1 exhaust\$1*1 kcal\$1*2];

Assign label "frailty" to the latent class variable and specify number of classes = 2

Run weight analysis using probability sampling weights in variable "sweight"

Assign starting values for thresholds (optional)

Frailty Criteria Patterns and Latent Class Analysis Fit:

Women's Health and Aging Studies I and II

	Criterion					Pattern Frequencies			
	Weight loss	Weak	Slow	Exhaustion	Low Activity	Observed	Expected		
							1-class	2-class	3-class
5 most frequently observed patterns among the non-frail	N	N	N	N	N	310	247.8	339.5	341.2
	N	N	Y	N	N	76	107.2	74.9	72.4
	N	N	N	N	Y	40	61.9	36.0	34.1
	N	Y	N	N	N	36	63.9	39.7	38.5
	N	N	Y	N	Y	32	26.8	22.8	28.7
5 most frequently observed patterns among the frail	N	Y	Y	N	Y	16	6.9	18.8	14.5
	N	Y	Y	Y	Y	12	1.1	9.5	8.2
	N	N	Y	Y	Y	10	4.3	9.5	7.3
	Y	Y	Y	Y	Y	10	0.2	3.3	7.6
	Y	Y	Y	N	Y	9	1	6.4	7

Latent Class Model Fit statistics

Pearson Chi-Square 568 (p<.0001) 24.4 (p=.22) 13.1 (p=.52)

AIC 3560 3389 3390

BIC 3583 3440 3467

(Bandeem-Roche et al. 2006)

Frailty Group Profiles: Conditional Probabilities of Meeting Criteria within Latent Classes Women's Health and Aging Studies

Criterion	2-Class Model		3-Class Model		
	CLASS 1 NON-FRAIL	CLASS 2 FRAIL	CLASS 1 ROBUST	CLASS 2 INTERMEDIATE	CLASS 3 FRAIL
Weight Loss	.073	.26	.072	.11	.54
Weakness	.088	.51	.029	.26	.77
Slowness	.15	.70	.004	.45	.85
Low Physical Activity	.078	.51	.000	.28	.70
Exhaustion	.061	.34	.027	.16	.56
Class Prevalence (%)	73.3	26.7	39.2	53.6	7.2

(Bandeem-Roche et al. 2006)

Step 2. Structural Model

- To evaluate the associations between frailty and age, education, and disease burden
 - Run latent class regression while fixing the number of latent classes derived from the LCA
 - Assuming “conditional independence” and “non-differential measurement”

MPLUS fitting of LCR

TITLE: Weighted Latent Class Analysis of Frailty Components Using Combined WHAS I and II Data Age 70-79

DATA:

FILE IS "c:\whas\frail\paper\lcaw.txt";

VARIABLE:

NAMES ARE baseid shrink weak slow exhaust kcal sweight age educ disease;
USEVARIABLES ARE shrink weak slow exhaust kcal sweight age educ disease;
CENTERING GRANDMEAN(age educ);
MISSING ARE ALL (999999);
CATEGORICAL ARE shrink-kcal;
CLASSES = frailty(2);

WEIGHT IS sweight;

ANALYSIS:

TYPE IS MIXTURE MISSING;

MODEL:

%OVERALL%

frailty#1 ON age educ disease;

Centering predictors age and education for meaningful interpretation of intercept

Structural regression model using Class 2 as the reference group

MPLUS Output

Categorical Latent Variables

	Estimates	S.E.	Est./S.E.
FRAILITY# ON			
AGE	-0.125	0.053	-2.355
EDUC	0.350	0.063	5.567
DISEASE	-0.939	0.151	-6.208
Intercepts			
FRAILITY#1	2.337	0.337	6.925

Using Class 2 as the
reference group

ALTERNATIVE PARAMETERIZATIONS FOR THE CATEGORICAL LATENT VARIABLE REGRESSION

Parameterization using Reference Class 1

FRAILITY# ON			
AGE	0.125	0.053	2.355
EDUC	-0.350	0.063	-5.567
DISEASE	0.939	0.151	6.208
Intercepts			
FRAILITY#2	-2.337	0.337	-6.925