Latent Class Regression

Karen Bandeen-Roche

October 28, 2016
Objectives
For you to leave here knowing…

- What is the LCR model and its underlying assumptions?
- How are LCR parameters interpreted?
- How does one check the assumptions of an LCR model?
- Latent class regression analogs to LCA for fitting and identifiability
Motivating Example: Frailty

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

<table>
<thead>
<tr>
<th></th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Robust</td>
<td>80</td>
</tr>
<tr>
<td>Intermediate Frail</td>
<td>15</td>
</tr>
<tr>
<td>Frail</td>
<td>5</td>
</tr>
</tbody>
</table>
Motivating Example: Frailty

- Is frailty associated with age, education, and/or disease burden?

- Do age, education, and/or disease burden predict heightened risk for membership in some frailty classes as opposed to others?
Part I: Model
Recall the standard latent class model:

- Discrete latent variables & discrete indicator variables
- Indicators measure discrete “subpopulations” rather than underlying continuous scores
- Patterns of responses are thought to contain information above and beyond “aggregation” of responses
- The prediction goal is “clustering” individuals rather than continuous response variables

We add “structural” piece to model where covariates “explain” class membership
Latent Class Regression Model

**Structural Model**
- Test Results
  - Genotype
  - Biomarkers
  - Environment
  - Life Events
  - Demographics

**Measurement Model**
- Depression
  - Sadness
  - Appetite
  - Sleep
  - Guilt
  - Apathy

Follow-up Studies/Outcomes
Analysis of underlying subpopulations
Latent class regression

\[
P_1(X) \quad \ldots \quad P_J(X)
\]

\[
\Pi_{11} \quad \ldots \quad \Pi_{1M} \quad \Pi_{J1} \quad \ldots \quad \Pi_{JM}
\]

\[
Y_1 \quad \ldots \quad Y_M \quad Y_1 \quad \ldots \quad Y_M
\]

\[
U_i \quad \beta \quad X_i
\]

Dayton & Macready, 1988
Latent Variable Models

Latent Class Regression (LCR) Model

- **Model:** 
  \[ f_{Y|x}(y|x) = \sum_{j=1}^{J} P_j(x, \beta) \prod_{m=1}^{M} \pi_{mj}^{y_m} (1 - \pi_{mj})^{1-y_m} \]

- **Structural model:** 
  \[ [U_i|x_i] = \Pr[U_i = j|x_i] = P_j(x_i, \beta) = \frac{\exp(x_i \beta_j)}{1 + \sum_{k=1}^{J-1} \exp(x_i \beta_k)} \]

  - A latent polytomous logistic regression

- **Measurement model:** 
  \[ [Y_i|U_i] \]
  \[ \pi_{mj} = \Pr[Y_{im} = 1|U_i = j] = “conditional probabilities” \]

  **\( \pi \)** is MxJ
Structural Model

- With two classes, the latent variable (class membership) is dichotomous
Parameter Interpretation

- Consider simplest case: 2 classes (1 vs. 2)
  \[
  \log \left( \frac{\Pr(C_i = 2 \mid x_i)}{1 - \Pr(C_i = 2 \mid x_i)} \right) = \beta_0 + \beta_1 x_i
  \]
  or equivalently,
  \[
  = \log \left( \frac{\Pr(C_i = 2 \mid x_i)}{\Pr(C_i = 1 \mid x_i)} \right) = \beta_0 + \beta_1 x_i
  \]
  \( \beta_1 \) is a log odds ratio.

- This is a latent logistic regression.
Parameter Interpretation

- Consider simplest case: 2 classes (1 vs. 2)

\[
\log \left( \frac{\Pr(C_i = 2 \mid x_i)}{\Pr(C_i = 1 \mid x_i)} \right) = \beta_0 + \beta_1 x_i
\]

where \( \beta_1 \) is a log odds ratio.

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”
Solving for $P_j(x_i) = \Pr(C_i=j|x_i)$

$$\log\left(\frac{p_2(x_i)}{p_1(x_i)}\right) = \log\left(\frac{\Pr(C_i = 2 \mid x_{i1}, x_{i2})}{\Pr(C_i = 1 \mid x_{i1}, x_{i2})}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$$

Using the fact: $p_1(x_i) + p_2(x_i) = 1$, we obtain

$$p_2(x_i) = \Pr(C_i = 2 \mid x_{i1}, x_{i2}) = \frac{e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}}}{1 + e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}}}$$

$$p_1(x_i) = \Pr(C_i = 1 \mid x_{i1}, x_{i2}) = \frac{1}{1 + e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}}}$$
Parameter Interpretation
General Case

- \((J-1)(p+1)\) \(\beta\)'s where \(p\) = number of covariates

- \(J-1\): one class is reference class so all of its \(\beta\) coefficients are technically zero

- \(p+1\): for each class (except the reference), there is one \(\beta\) for each covariate plus another for the intercept.
Parameter Interpretation

General Case

Need more than one equation
Choose a reference class (e.g. class 1)

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

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

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

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

$e^{\beta_{13}} / e^{\beta_{12}} = e^{\beta_{13} - \beta_{12}} = \text{OR for class 3 versus class 2 for females versus males}$
Solving for \( P_j(x_i) = Pr(C_i|x_i) \)

\[
\log \left( \frac{P_2(x_i)}{P_1(x_i)} \right) = \log \left( \frac{Pr(C_i = 2|x_{i1}, x_{i2})}{Pr(C_i = 1|x_{i1}, x_{i2})} \right) = \beta_{02} + \beta_{12}x_{i1} + \beta_{22}x_{i2}
\]

\[
P_2(x_i) = Pr(C_i = 2|x_{i1}, x_{i2}) = \frac{e^{\beta_{02} + \beta_{12}x_{i1} + \beta_{22}x_{i2}}}{1 + e^{\beta_{02} + \beta_{12}x_{i1} + \beta_{22}x_{i2}} + e^{\beta_{03} + \beta_{13}x_{i1} + \beta_{23}x_{i2}}}
\]

\[
= \frac{3}{\sum_{j=1}^{3} e^{\beta_{0j} + \beta_{1j}x_{i1} + \beta_{2j}x_{i2}}}
\]

\[
\beta_{01} = \beta_{11} = \beta_{21} = 0 \quad \text{(Since class 1 is the reference class)}
\]
Assumptions

- Conditional Independence:
  - given an individual’s class, his symptoms are independent
  - \( \Pr(y_{im}, y_{ir} | C_i) = \Pr(y_{im} | C_i) \times \Pr(y_{ir} | C_i) \)

- Non-differential Measurement:
  - given an individual’s class, covariates are not associated with symptoms
  - \( \Pr(y_{im} | x_i, C_i) = \Pr(y_{im} | C_i) \)
Conditional Dependence
Differential Measurement

![Diagram showing structural model and measurement model related to depression.]

- **Structural Model**
  - Test Results
    - Genotype
    - Biomarkers
    - Environment
    - Life Events
    - Demographics

- **Measurement Model**
  - Depression
    - Sadness
    - Appetite
    - Sleep
    - Guilt
    - Apathy
Part II: Fitting
Model Building

- **Step 1:**
  - Get the measurement part right
  - Fit standard latent class model first.
  - Works: “marginalization” property

\[
 f_y(y) = \sum_{j=1}^{J} \left\{ \int P_j(x) dG(x) \right\} \prod_{m=1}^{M} \pi_{mj}^y (1 - \pi_{mj})^{1-y_m} = \sum_{j=1}^{J} P_j \prod_{m=1}^{M} \pi_{mj}^y (1 - \pi_{mj})^{1-y_m}
\]

Bandeen-Roche et al., *J Am Statist Assoc.*, 1997

- **Step 2:**
  - Model building as in multiple logistic regression
Model Estimation

- Maximum likelihood estimator
- Latent Class Regression Likelihood

\[
\Pr(Y_i = y_i \mid x_i) = \Pr(Y_{i1} = y_{i1}, Y_{i2} = y_{i2}, \ldots, Y_{iM} = y_{iM} \mid x_i)
\]

\[
= \sum_{j=1}^{J} p_j(x_i) \prod_{m=1}^{M} \pi_{mj}^{y_{im}} (1 - \pi_{mj})^{(1-y_{im})}
\]

where

\[
p_j(x_i) = \frac{e^{\beta_j x_i}}{\sum_{j=1}^{J} e^{\beta_j x_i}}
\]

- EM Algorithm
  - E-step – as before
  - M-step – polytomous logistic regression with posterior probabilities as “outcomes”

Bandeen-Roche et al., *J Am Statist Assoc.*, 1997
Example: Frailty of older adults
Step 1. Measurement model

<table>
<thead>
<tr>
<th>Criterion</th>
<th>2-Class Model</th>
<th>3-Class Model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CLASS 1 NON-FRAIL</td>
<td>CLASS 2 FRAIL</td>
</tr>
<tr>
<td>Weight Loss</td>
<td>.073</td>
<td>.26</td>
</tr>
<tr>
<td>Weakness</td>
<td>.088</td>
<td>.51</td>
</tr>
<tr>
<td>Slowness</td>
<td>.15</td>
<td>.70</td>
</tr>
<tr>
<td>Low Physical Activity</td>
<td>.078</td>
<td>.51</td>
</tr>
<tr>
<td>Exhaustion</td>
<td>.061</td>
<td>.34</td>
</tr>
<tr>
<td>Class Prevalence (%)</td>
<td>73.3</td>
<td>26.7</td>
</tr>
</tbody>
</table>

(Bandeen-Roche et al. 2006)
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”
Frailty: Structural Model
Reference group = Non-frail

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coeff</th>
<th>OR</th>
<th>Coeff SE</th>
<th>Coeff Z</th>
<th>Coeff CI</th>
<th>OR CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age(^{1,2})</td>
<td>0.137</td>
<td>1.147</td>
<td>0.049</td>
<td>2.787</td>
<td>(0.041,0.233)</td>
<td>(1.04,1.26)</td>
</tr>
<tr>
<td>Education(^{1,2})</td>
<td>-0.362</td>
<td>0.696</td>
<td>0.053</td>
<td>-6.842</td>
<td>(-0.466,-0.258)</td>
<td>(0.63,0.77)</td>
</tr>
<tr>
<td>Diseases(^3)</td>
<td>0.890</td>
<td>2.435</td>
<td>0.137</td>
<td>6.510</td>
<td>(0.621,1.158)</td>
<td>(1.86,3.19)</td>
</tr>
<tr>
<td>Intercept</td>
<td>-2.139</td>
<td>0.118(^4)</td>
<td>0.338</td>
<td>-6.335</td>
<td>(-2.801,-1.478)</td>
<td>(0.06,0.23)(^4)</td>
</tr>
</tbody>
</table>

1 Centered at means
2 Years
3 Number (count)
4 Odds (among the non-frail; rather than odds ratios)
### Frailty: Structural Model

Reference group = Non-frail

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coeff</th>
<th>OR</th>
<th>Coeff SE</th>
<th>OR CI</th>
<th>Coeff Z</th>
<th>Coeff CI</th>
<th>OR CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age(^1,2)</td>
<td>0.137</td>
<td>1.147</td>
<td>0.049</td>
<td>(0.041,0.233)</td>
<td>2.787</td>
<td>(0.041,0.233)</td>
<td>(1.04,1.26)</td>
</tr>
<tr>
<td>Education(^1,2)</td>
<td>-0.362</td>
<td>0.696</td>
<td>0.053</td>
<td>(-0.466,-0.258)</td>
<td>-6.842</td>
<td>(-0.466,-0.258)</td>
<td>(0.63,0.77)</td>
</tr>
<tr>
<td>Diseases(^3)</td>
<td>0.890</td>
<td>2.435</td>
<td>0.137</td>
<td>(0.621,1.158)</td>
<td>6.510</td>
<td>(0.621,1.158)</td>
<td>(1.86,3.19)</td>
</tr>
<tr>
<td>Intercept</td>
<td>-2.139</td>
<td>0.118</td>
<td>0.338</td>
<td>(-2.801,-1.478)</td>
<td>-6.335</td>
<td>(-2.801,-1.478)</td>
<td>(0.06,0.23)</td>
</tr>
</tbody>
</table>

\(^1\) Prob = odds/(1+odds)  
\(^2\) = estimated frail prevalence among those with mean age and education & no diseases

\(^3\) Prob = 0.118/1.118  
\(^4\) = 0.106
Frailty: Structural Model  
Reference group = Non-frail

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coeff</th>
<th>OR</th>
<th>Coeff SE</th>
<th>Coeff Z</th>
<th>Coeff CI</th>
<th>OR CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age(^1,2)</td>
<td>0.137</td>
<td>1.147</td>
<td>0.049</td>
<td>2.787</td>
<td>(0.041,0.233)</td>
<td>(1.04,1.26)</td>
</tr>
<tr>
<td>Education(^1,2)</td>
<td>-0.362</td>
<td>0.696</td>
<td>0.053</td>
<td>-6.842</td>
<td>(-0.466,-0.258)</td>
<td>(0.63,0.77)</td>
</tr>
<tr>
<td>Diseases(^3)</td>
<td>0.890</td>
<td><strong>2.435</strong></td>
<td>0.137</td>
<td>6.510</td>
<td>(0.621,1.158)</td>
<td>(1.86,3.19)</td>
</tr>
<tr>
<td>Intercept</td>
<td>-2.139</td>
<td>0.118(^4)</td>
<td>0.338</td>
<td>-6.335</td>
<td>(-2.801,-1.478)</td>
<td>(0.06,0.23)(^4)</td>
</tr>
</tbody>
</table>

we estimate that the odds of being frail (vs. nonfrail) increases 2.4-fold with each added disease
Part III: Evaluating Fit
Model Checking

- Analog = residual checking in linear regression

- IT’S CRITICALLY IMPORTANT!
  - Can give misleading findings if measurement model assumptions are unwarranted
  - Philosophical opinion: we learn primarily by specifying how simple models fail to fit, not by observing that complex models happen to fit

- Two types of checking
  - Whether the model fits (e.g. observed vs. expected)
  - How a model may fail to fit (ASSUMPTIONS)
Checking Whether the Model Fits

- Means

1) Do Y’s aggregate as expected given the model?
   - Check the measurement model (LCA)
   - Check whether the measurement model is comparable without (LCA) and with (LCR) covariates
Checking Whether the Model Fits

2) Do Y’s relate to the X’s as expected given the model

- **Idea**: focus on one item at a time

\[
P(Y_{i1} = y_{i1}, \ldots, Y_{iM} = y_{iM} \mid x_i) = \sum_{j=1}^{J} P_j(x_i) \prod_{m=1}^{M} \pi_{mj}^{y_{im}} (1 - \pi_{mj})^{(1-y_{im})}
\]

- If interested in item m, ignore (“marginalize over”) other items:

\[
P(Y_{im} = y_{im} \mid x_i) = \sum_{j=1}^{J} P_j(x_i) \pi_{mj}^{y_{im}} (1 - \pi_{mj})^{(1-y_{im})}
\]
Comparing Fitted to Observed

- Construct the predicted curve by plotting this probability versus any given x
  - Add a smooth spline (e.g. use “lowess” in STATA) to reveal systematic trend (solid line)

- Superimpose it with an “observed” item response curve by
  - Plot item response (0 or 1) by x
  - Add smooth spline to reveal systematic trend (dashed line)
Checking *How the Model Fails to Fit*

- Check Assumptions
  - conditional independence
  - non-differential measurement
Checking *How the Model Fails to Fit*

### Basic ideas:
- Suppose the model is true
- If we knew persons’ latent class memberships, we would check directly:
  - Stratify into classes, then, **within classes**:
    - Check correlations or pairwise odds ratios among the item responses (*Conditional Independence*)
    - Regress item responses on covariates (*non-differential measurement*)
  - Regress class memberships on covariates, hope for
    - Similar findings re regression coefficients
    - No strong effects of outliers
    - Identify strongly nonlinear covariates effects
Checking How the Model Fails to Fit

- But in reality, we don’t know the true latent class membership!
- Latent class memberships must be estimated
  - Randomize people into “pseudo” classes using their posterior probabilities or assign to “most likely class” corresponding to the highest posterior probability
  - Posterior probability is defined as

\[
\Pr(C_i = j | x_i, y_i) = \frac{\Pr(y_i | C_i = j) \Pr(C_i = j | x_i)}{\sum_{j=1}^{J} \Pr(y_i | C_i = j) \Pr(C_i = j | x_i)}
\]

- Analyze as described before, except using “pseudo” class membership rather than true ones

Bandeen-Roche et al., *J Am Statist Assoc.*, 1997
Utility of Model Checking

- May modify interpretation to incorporate lack of fit/violation of assumption
- May help elucidate a transformation that would be more appropriate (e.g. \( \log(\text{age}) \) versus age)
- May suggest how to improve measurement (e.g. better survey instrument)
- May lead to believe that LCR is not appropriate
Part IV: Identifiability / Estimability
Identifiability
Latent class (binary Y) regression

• Latent class analysis (measurement only)
  • \( Y \sim \text{multinomial} \left( \underbrace{p}_{\text{dim}} \right), \dim(\underbrace{p}_{\text{dim}}) = \)

• Unconstrained J-class model:

• Need \( 2^M - 1 \geq M(J - 1) \)

• Latent class regression
  • As above + full-rank X

Bandeen-Roche et al, JASA, 1997; Huang & Bandeen-Roche, Psychometrika, 2004
Identifiability

- To best assure identification
  - Incorporate a priori theory as much as possible
    - Set $\pi$s to 0 or 1 where it makes sense to do so
    - Set $\pi$s equal to each other
  - If program fails to converge
    - Run the program longer
    - Re-initialize in very different places
    - Add constraints (e.g. set $\pi$s to 0 or 1 where sensible)
    - Stop (attempting to do too much with one’s data)
Objectives
For you to leave here knowing…

- What is the LCR model and its underlying assumptions?
- How are LCR parameters interpreted?
- How does one check the assumptions of an LCR model?
- Latent class regression analogs to LCA for fitting and identifiability
Appendix

- Mplus codes
TITLE: Latent Class Analysis of Frailty Components
Using Combined WHAS I and II Data Age 70-79
DATA: FILE IS "h:\teaching\140.658\2007\lcr.dat";
VARIABLE: NAMES ARE baseid shrink weak slow exhaust kcal sweight age educ disease;
USEVARIABLES ARE shrink weak slow exhaust kcal;
MISSING ARE ALL (999999);
CATEGORICAL ARE shrink-kcal;
CLASSES = frailty(2);
ANALYSIS: TYPE IS MIXTURE;
MODEL:
%OVERALL%

%frailty#1%
[shrink$1*2 weak$1*2 slow$1*2 exhaust$1*2 kcal$1*2];

%frailty#2%
[shrink$1*-1 weak$1*-1 slow$1*-1 exhaust$1*-1 kcal$1*-2];

OUTPUT: TECH10 TECH11;
SAVEDATA:
FILE IS "h:\teaching\140.658\2010\lcasave.out";
SAVE=CPROB;

Declare missing value code
Assign label “frailty” to the latent class variable and specify number of classes = 2
Assign starting values for thresholds (optional)
TECH10: output of observed vs. estimated frequencies of response patterns; TECH11, TECH14: results of Lo-Mendell-Rubin test and bootstrapped likelihood ratio test for comparing models with k vs. k-1 classes
Save posterior class probabilities
Step 2. Structural Model

<table>
<thead>
<tr>
<th>TECHNICAL 11 OUTPUT</th>
<th>VUONG-LO-MENDELL-RUBIN LIKELIHOOD RATIO TEST FOR 2 (H0) VERSUS 3 CLASSES</th>
</tr>
</thead>
<tbody>
<tr>
<td>H0 Loglikelihood Value</td>
<td>-1890.373</td>
</tr>
<tr>
<td>2 Times the Loglikelihood Difference</td>
<td>12.482</td>
</tr>
<tr>
<td>Difference in the Number of Parameters</td>
<td>6</td>
</tr>
<tr>
<td>Mean</td>
<td>6.334</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>6.562</td>
</tr>
<tr>
<td>P-Value</td>
<td>0.1441</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>LO-MENDELL-RUBIN ADJUSTED LRT TEST</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
</tr>
<tr>
<td>P-Value</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TECHNICAL 14 OUTPUT</th>
<th>PARAMETRIC BOOTSTRAPPED LIKELIHOOD RATIO TEST FOR 2 (H0) VERSUS 3 CLASSES</th>
</tr>
</thead>
<tbody>
<tr>
<td>H0 Loglikelihood Value</td>
<td>-1890.373</td>
</tr>
<tr>
<td>2 Times the Loglikelihood Difference</td>
<td>12.482</td>
</tr>
<tr>
<td>Difference in the Number of Parameters</td>
<td>6</td>
</tr>
<tr>
<td>Approximate P-Value</td>
<td>0.0896</td>
</tr>
<tr>
<td>Successful Bootstrap Draws</td>
<td>67</td>
</tr>
</tbody>
</table>
MPLUS fitting of LCR

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

DATA:
  FILE IS "h:\whas\frail\paper\lcr.dat";

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

ANALYSIS:
  TYPE IS MIXTURE;

MODEL:
  %OVERALL%
    frailty#1 ON age educ disease;
  
    %frailty#1%
    [shrink$1*3 weak$1*3 slow$1*3 exhaust$1*3 kcal$1*3];
  
    %frailty#2%
    [shrink$1*-2 weak$1*-1 slow$1*-1 exhaust$1*-1 kcal$1*-2];

Centering predictors age and education for meaningful interpretation of intercept

Structural regression model using Class 2 as the reference group

Why have starting values?
Categorical Latent Variables

<table>
<thead>
<tr>
<th></th>
<th>Estimates</th>
<th>S.E.</th>
<th>Est./S.E.</th>
</tr>
</thead>
<tbody>
<tr>
<td>FRAILTY#1 ON</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AGE</td>
<td>-0.137</td>
<td>0.049</td>
<td>-2.787</td>
</tr>
<tr>
<td>EDUC</td>
<td>0.362</td>
<td>0.053</td>
<td>6.842</td>
</tr>
<tr>
<td>DISEASE</td>
<td>-0.890</td>
<td>0.137</td>
<td>-6.510</td>
</tr>
<tr>
<td>Intercepts</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FRAILTY#1</td>
<td>2.139</td>
<td>0.338</td>
<td>6.335</td>
</tr>
</tbody>
</table>

ALTERNATIVE PARAMETERIZATIONS FOR THE CATEGORICAL LATENT VARIABLE REGRESSION

Parameterization using Reference Class 1

<table>
<thead>
<tr>
<th></th>
<th>Estimates</th>
<th>S.E.</th>
<th>Est./S.E.</th>
</tr>
</thead>
<tbody>
<tr>
<td>FRAILTY# ON</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AGE</td>
<td>0.137</td>
<td>0.049</td>
<td>2.787</td>
</tr>
<tr>
<td>EDUC</td>
<td>-0.362</td>
<td>0.053</td>
<td>-6.842</td>
</tr>
<tr>
<td>DISEASE</td>
<td>0.890</td>
<td>0.137</td>
<td>6.510</td>
</tr>
<tr>
<td>Intercepts</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FRAILTY#2</td>
<td>-2.139</td>
<td>0.338</td>
<td>-6.335</td>
</tr>
</tbody>
</table>
Checking Whether the Model Fits

1) Do Y’s aggregate as expected given the model?
   - Compare observed pattern frequencies to predicted pattern frequencies
## Frailty Example: Observed versus Expected Response Patterns: Ignoring Covariates

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Weight loss</th>
<th>Weak</th>
<th>Slow</th>
<th>Exhaustion</th>
<th>Low Activity</th>
<th>Pattern Frequencies</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Observed</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1-class</td>
</tr>
<tr>
<td>5 most 5 frequently observed patterns among the non-frail</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>5 most 5 frequently observed patterns among the frail</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>5 most 5 frequently observed patterns among the frail</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>5 most 5 frequently observed patterns among the frail</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>5 most 5 frequently observed patterns among the frail</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>5 most 5 frequently observed patterns among the frail</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>5 most 5 frequently observed patterns among the frail</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>5 most 5 frequently observed patterns among the frail</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>5 most 5 frequently observed patterns among the frail</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>5 most 5 frequently observed patterns among the frail</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>Y</td>
</tr>
</tbody>
</table>

### Latent Class Model Fit statistics

<table>
<thead>
<tr>
<th>Fit statistic</th>
<th>Observed</th>
<th>Expected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson Chi-Square</td>
<td>568 (p&lt;.0001)</td>
<td>24.4 (p=.22)</td>
</tr>
<tr>
<td>AIC</td>
<td>3560</td>
<td>3389</td>
</tr>
<tr>
<td>BIC</td>
<td>3583</td>
<td>3440</td>
</tr>
</tbody>
</table>

(Bandeen-Roche et al. 2006)
TITLE: Weighted Latent Class Analysis of Frailty Components Using Combined WHAS I and II Data Age 70-79
DATA:  FILE IS "C:\teaching\140.658.2007\lcr.dat";
VARIABLE: NAMES ARE baseid shrink weak slow exhaust kcal sweight age educ disease;
USEVARIABLES ARE shrink weak slow exhaust kcal;
MISSING ARE ALL (999999);
CATEGORICAL ARE shrink-kcal;
CLASSES = frailty(2);
ANALYSIS: TYPE IS MIXTURE;
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];

OUTPUT: TECH10
SAVEDATA:
FILE IS "C:\teaching\140.658.2007\lcasave.out";
SAVE=CPROB;

Contains observed vs. expected frequencies of response patterns
Save estimated posterior probabilities of class membership
TECHNICAL 10 OUTPUT

MODEL FIT INFORMATION FOR THE LATENT CLASS INDICATOR MODEL PART

RESPONSE PATTERNS

<table>
<thead>
<tr>
<th>No.</th>
<th>Pattern</th>
<th>No.</th>
<th>Pattern</th>
<th>No.</th>
<th>Pattern</th>
<th>No.</th>
<th>Pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>00000</td>
<td>2</td>
<td>10000</td>
<td>3</td>
<td>01000</td>
<td>4</td>
<td>11000</td>
</tr>
<tr>
<td>5</td>
<td>00100</td>
<td>6</td>
<td>10100</td>
<td>7</td>
<td>01100</td>
<td>8</td>
<td>11100</td>
</tr>
<tr>
<td>9</td>
<td>00010</td>
<td>10</td>
<td>10010</td>
<td>11</td>
<td>01010</td>
<td>12</td>
<td>11010</td>
</tr>
<tr>
<td>13</td>
<td>00110</td>
<td>14</td>
<td>10110</td>
<td>15</td>
<td>01110</td>
<td>16</td>
<td>11110</td>
</tr>
</tbody>
</table>

RESPONSE PATTERN FREQUENCIES AND CHI-SQUARE CONTRIBUTIONS

<table>
<thead>
<tr>
<th>Response Pattern</th>
<th>Frequency</th>
<th>Standardized Residual (z-score)</th>
<th>Chi-square Contribution</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Observed</td>
<td>Estimated</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>310.00</td>
<td>293.64</td>
<td>1.23</td>
</tr>
<tr>
<td>2</td>
<td>25.00</td>
<td>23.97</td>
<td>0.21</td>
</tr>
<tr>
<td>3</td>
<td>36.00</td>
<td>36.89</td>
<td>-0.15</td>
</tr>
<tr>
<td>4</td>
<td>6.00</td>
<td>4.82</td>
<td>0.54</td>
</tr>
<tr>
<td>5</td>
<td>76.00</td>
<td>81.85</td>
<td>-0.69</td>
</tr>
<tr>
<td>6</td>
<td>10.00</td>
<td>11.15</td>
<td>-0.35</td>
</tr>
</tbody>
</table>

Recall: for a bivariate table, standardized residual = (O – E) / [(E)¹/²*(1-E/N)¹/²]
Empirical Testing of Identifiability

- Must run model more than once using different starting values to check identifiability!
- Mplus input:

```
ANALYSIS: TYPE IS MIXTURE;
STARTS = 500 50;
STITERATIONS=20;
```

Number of initial stage random sets of starting values and the number of final stage optimizations to use

Maximum number of iteration allowing in the initial stage
TITLE: this is an example of a LCA with binary latent class indicators and parameter constraints
DATA:   FILE IS ex7.13.dat;
VARIABLE:   NAMES ARE u1-u4;
            CLASSES = c (2);
            CATEGORICAL = u1-u4;
ANALYSIS:   TYPE = MIXTURE;
MODEL: 
            %OVERALL%
            %c#1%
              [u1$1*-1];
              [u2$1-u3$1*-1] (1);
              [u4$1*-1] (p1);
            %c#2%
              [u1$1@-15];
              [u2$1-u3$1*1] (2);
              [u4$1*1] (p2);
MODEL CONSTRAINT:
            p2 = - p1;
OUTPUT: TECH1 TECH8;

u2 and u3 have same \( \pi \) in class 1
u1 has \( \pi \) equal to 1 in class 2
u2 and u3 have same \( \pi \) in class 2
The threshold of u4 in class 1 is equal to 
-1*threshold of u4 in class 2 (i.e. same error rate)

MPLUS User’s Guide p. 153