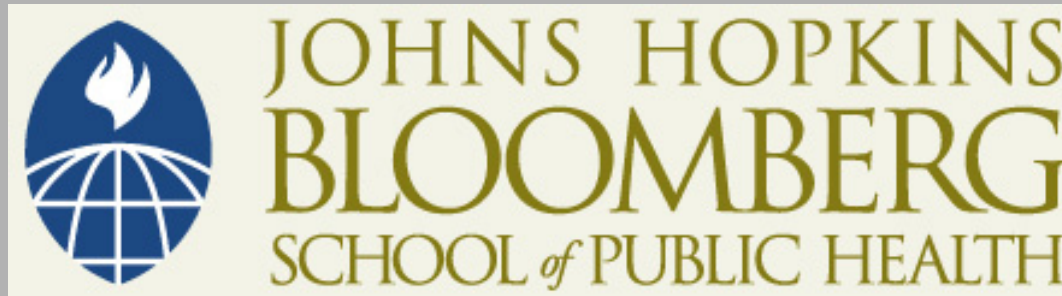


This work is licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike License](https://creativecommons.org/licenses/by-nc-sa/4.0/). Your use of this material constitutes acceptance of that license and the conditions of use of materials on this site.



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 II

Statistics for Psychosocial Research II:
Structural Models

Qian-Li Xue

Outline

- Model Checking
 - Check model fit
 - Check model assumption
 - ❖ Conditional independence
 - ❖ Non-differential measurement
- Identifiability

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? [Conditional Independence]
 - 2) Do Y's relate to the X's as expected given the model [Non-Differential Measurement]

Checking Whether the Model Fits

- 1) Do Y's aggregate as expected given the model? [Conditional Independence]
 - ❖ Compare observed pattern frequencies to predicted pattern frequencies

Frailty Example: Observed versus Expected Response Patterns: Ignoring Covariates

| | 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 | 24.4 | 13.1 |
| | (p<.0001) | (p=.22) | (p=.52) |

| | | | |
|-----|------|------|------|
| AIC | 3560 | 3389 | 3390 |
|-----|------|------|------|

| | | | |
|-----|------|------|------|
| BIC | 3583 | 3440 | 3467 |
|-----|------|------|------|

(Bandeem-Roche et al. 2006)

MPLUS Input

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 sweight; MISSING ARE ALL (999999); CATEGORICAL ARE
shrink-kcal; CLASSES = frailty(2);
WEIGHT IS sweight; 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: TECH1 TECH10

Contains observed vs. expected
frequencies of response patterns

SAVEDATA:

FILE IS "C:\teaching\140.658.2007\lcasave.out";

SAVE=CPROB;

Save estimated posterior
probabilities of class membership

MPLUS Output

TECHNICAL 10 OUTPUT

MODEL FIT INFORMATION FOR THE LATENT CLASS INDICATOR MODEL PART

RESPONSE PATTERNS

| No. | Pattern | No. | Pattern | No. | Pattern | No. | Pattern |
|-----|---------|-----|---------|-----|---------|-----|---------|
| 1 | 00000 | 2 | 10000 | 3 | 01000 | 4 | 11000 |
| 5 | 00100 | 6 | 10100 | 7 | 01100 | 8 | 11100 |
| 9 | 00010 | 10 | 10010 | 11 | 01010 | 12 | 11010 |
| 13 | 00110 | 14 | 10110 | 15 | 01110 | 16 | 11110 |

RESPONSE PATTERN FREQUENCIES AND CHI-SQUARE CONTRIBUTIONS

| Response Pattern | Frequency | | Standardized Residual (z-score) | Chi-square Contribution | | |
|---------------------|-----------|-----------|---------------------------------------|-------------------------|---------------|---------|
| | Observed | Estimated | | Pearson | Loglikelihood | Deleted |
| 1 | 344.81 | 331.36 | 0.99 | 0.11 | 3.01 | |
| 2 | 30.14 | 28.91 | 0.23 | 0.05 | 3.31 | |
| 3 | 37.13 | 38.02 | -0.15 | -0.06 | -5.15 | |
| 4 | 7.95 | 5.36 | 1.12 | 1.25 | 5.99 | |
| 5 | 73.46 | 76.67 | -0.39 | 0.12 | -3.94 | |
| 6 | 8.27 | 11.27 | -0.90 | 0.73 | -3.63 | |

Recall: standardized residual = $O - E / (E)^{1/2}$

Do Y patterns behave as model predicts?

- Compare observed pattern frequencies to expected pattern frequencies
- How does addition of regression change interpretation?
- Evaluating fit of measurement piece
 - Will be “same” as in standard LCA model unless.....

Do Y patterns behave as model predicts? With Categorical Covariates

- Easier than continuous (computationally)
- Example
 - Calculate:
 - ❖ Predicted whites with weight loss
 - ❖ Observed whites with weight loss
 - ❖ Predicted blacks with weight loss
 - ❖ Observed blacks with weight loss

Do Y patterns behave as model predicts?

Categorical Covariates

- Assume LC regression model with only race as covariate
- Race (x) = 0 if white, 1 if black
- Item of interest: weight loss ($y_m=1$ – yes; 0 – no)
- Want find how many class 2 whites we would expect to report weight loss based on the model

$$\begin{aligned}\Pr(y_{im} = 1, \eta_i = j, x_i = 0) \\ &= \Pr(y_{im} = 1 \mid \eta_i = j, x_i = 0) \Pr(\eta_i = j \mid x_i = 0) \Pr(x_i = 0) \\ &= \Pr(y_{im} = 1 \mid \eta_i = j) \Pr(\eta_i = j \mid x_i = 0) \Pr(x_i = 0)\end{aligned}$$



$$\text{Expected}(\text{weight loss, whites, and class} = 2) = N \times \pi_{mj} \times \frac{e^{\beta_0}}{1 + e^{\beta_0}} \times \text{Prop.}(\text{whites})$$

- Calculate this for each of the classes and sum up:
- Will tell us the expected number of whites reporting weight loss

Checking Whether the Model Fits

2) Do Y's relate to the X's as expected given the model [Non-Differential Measurement]

- ❖ Idea: focus on one item at a time
- ❖ Recall:

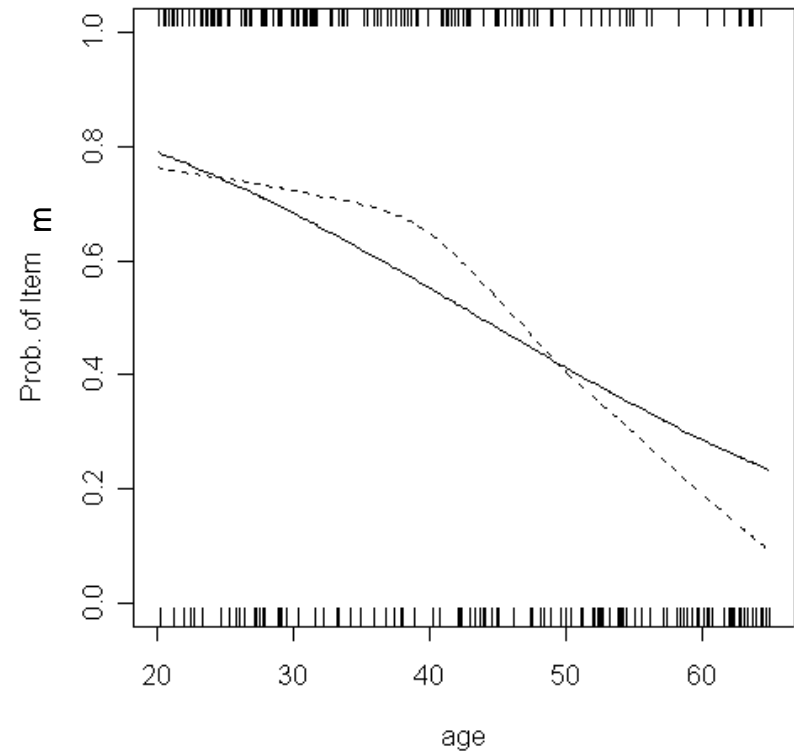
$$P(Y_{i1} = y_{i1}, \dots, Y_{iM} = y_{iM} \mid x_i) = \sum_{j=1}^J p_{ij}(x_i) \prod_{m=1}^M \pi_{mj}^{y_{im}} (1 - \pi_{mj})^{(1-y_{im})}$$

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

$$P(Y_{im} = y_{im} \mid x_i) = \sum_{j=1}^J p_{ij}(x_i) \pi_{mj}^{y_{im}} (1 - \pi_{mj})^{(1-y_{im})}$$

Comparing Fitted to Observed

- ❖ Construct the predicted curve by plotting this conditional probability versus any given x
 - Add a smooth spline to reveal systematic trend
- ❖ Superimpose it with an “observed” item response curve by
 - Plot item response (0 or 1) by x
 - Add smooth spline to reveal systematic trend



Checking How the Model Fails to Fit

- Check Assumptions
 - non-differential measurement
 - conditional independence
- Non-differential Measurement:
 - $P(y_{im} | x_i, \eta_i) = P(y_{im} | \eta_i)$
 - In words, within a class, there is no association between y's and x's.
 - Check this using logistic regression approach

Checking How the Model Fails to Fit

- Basic ideas:
 - Suppose the model is true
 - If we know 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 or observed patterns 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 Conditional Independence

- In words, within a class, there is no association between y_m and y_n , $m \neq n$.
- Define:

$$OR_{mn|j} = \frac{P(y_m = 1, y_n = 1 | \eta = j) / P(y_m = 0, y_n = 1 | \eta = j)}{P(y_m = 1, y_n = 0 | \eta = j) / P(y_m = 0, y_n = 0 | \eta = j)}$$

- $OR \sim 1$

Checking Non-differential Measurement

- For binary covariates and for each class j and item m consider

$$OR_{m|jx} = \frac{P(y_m = 1 | x = 1, \eta = j) / P(y_m = 0 | x = 1, \eta = j)}{P(y_m = 1 | x = 0, \eta = j) / P(y_m = 0 | x = 0, \eta = j)}$$

- If assumption holds, this OR will be approximately equal to 1
- What about continuous covariates?
 - Use same general idea, but estimate the logOR within classes by logistic regressionExample: age

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
 - Recall: posterior probability is defined as
$$\Pr(\eta_i = j | x_i, y_i) = \frac{\Pr(y_i | \eta_i = j) \Pr(\eta_i = j | x_i)}{\sum_{j=1}^J \Pr(y_i | \eta_i = j) \Pr(\eta_i = j | x_i)}$$
 - Different from latent class probabilities $\Pr(\eta_i = j | x_i)$!!
- Analyze as described before, except using “pseudo” class membership rather than true ones

MPLUS Output: Estimated Posterior Probabilities

SAVEDATA:

FILE IS "C:\teaching\140.658.2007\lcasave.out";

SAVE=CPROB;

| | Response Data | | | | | Sampling weights | Posterior Class Prob. | | |
|---|---------------|-------|-------|--------|-------|------------------|-----------------------|-------|-------|
| | 0.000 | 1.000 | 1.000 | 0.000 | 0.000 | 0.845 | 0.211 | 0.789 | 2.000 |
| | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 0.737 | 0.001 | 0.999 | 2.000 |
| | 0.000 | 1.000 | 1.000 | 0.000 | 1.000 | 0.737 | 0.022 | 0.978 | 2.000 |
| | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 0.737 | 0.001 | 0.999 | 2.000 |
| | 0.000 | 0.000 | 1.000 | 0.000 | 1.000 | 0.766 | 0.215 | 0.785 | 2.000 |
| * | | 1.000 | 1.000 | 0.000* | | 0.718 | 0.102 | 0.898 | 2.000 |
| | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 0.742 | 0.001 | 0.999 | 2.000 |
| | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 | 0.791 | 0.786 | 0.214 | 1.000 |
| | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 | 0.791 | 0.786 | 0.214 | 1.000 |

Most likely Class

Checking Conditional Independence

- 1) Assign individuals to “pseudo-classes” based on posterior probability of class membership
 - e.g. individual with 0.20, 0.05, 0.75
 - ❖ better chance of being in class 3
 - ❖ not necessarily in class 3
- 2) Calculate OR's within classes.
- 3) Repeat 1) and 2) at least a few times
- 4) Compare OR's to 1

Utility of Model Checking

- May modify interpretation to incorporate lack of fit/violation of assumption
- May help elucidate a transformation that 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

Select Number of Classes

- Method 1: analyze for a theorized number of classes, check for violation of conditional dependence, refit with more classes if clear patterns of dependence emerge
 - Labor-intensive!
- Method 2: Conduct a LCA ignoring covariates; select classes as discussed last term
 - This works assuming non-differential measurement
 - Y's are latent class distributed ignoring covariates
 - With same number of classes and same π s as in regression model

Identifiability

- General Idea: different parameters can lead to the same model fit
- 2-step rule: If
 - (a) polytomous logistic regression is identified
 - (b) standard LCM is identifiedThen model is identified
- T-rule: need more data cells than parameters
 - Recall:
 - ❖ for binary indicators: # of parameters (unknowns) = $[M \times J] \pi_s$ + $[(J-1) \times (H+1)] \beta$'s, where H is # of covariates
 - Data cell count is $2^n - 1$ per covariate combination
 - Each possible covariate combination creates a “stratum”
 - For continuous covariates, t-rule likely satisfied

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

MPLUS Output: Good Stability

RANDOM STARTS RESULTS RANKED FROM THE BEST TO THE WORST LOGLIKELIHOOD
VALUES

Final stage loglikelihood values at local maxima, seeds, and initial stage
start numbers:

| | | |
|-----------|--------|-----|
| -2750.059 | 373505 | 88 |
| -2750.059 | 642909 | 251 |
| -2750.059 | 569131 | 26 |
| -2750.059 | 467339 | 66 |
| -2750.059 | 903369 | 134 |
| -2750.059 | 652266 | 490 |
| -2750.059 | 765392 | 382 |
| -2750.059 | 315029 | 471 |
| -2750.059 | 22089 | 143 |
| -2750.059 | 127215 | 9 |
| -2750.059 | 252949 | 487 |

MPLUS Output: Bad Stability

RANDOM STARTS RESULTS RANKED FROM THE BEST TO THE WORST LOGLIKELIHOOD
VALUES

Final stage loglikelihood values at local maxima, seeds, and initial stage
start numbers:

| | | |
|-----------|--------|-----|
| -2750.059 | 373505 | 192 |
| -2759.012 | 642909 | 425 |
| -2755.426 | 852154 | 19 |
| -2750.059 | 467339 | 66 |
| -2743.785 | 158965 | 111 |
| -2750.059 | 965321 | 336 |
| -2743.785 | 765392 | 382 |
| -2743.785 | 178526 | 189 |
| -2750.059 | 56325 | 258 |
| -2750.059 | 128963 | 41 |
| -2759.012 | 56833 | 401 |

Identifiability is in
question!!

Identifiability

- To best assure identification
 - Incorporate a priori theory as much as possible
 - ❖ Set π s to 0 or 1 where it makes sense to do so
 - ❖ Set π s equal to each other
 - If program fails to converge
 - ❖ Run the program longer
 - ❖ Re-initialize in very different place
 - ❖ Add constraints (e.g. set π s to 0 or 1 where sensible)

Identifiability

- Cautions
 - Even if no warning message, some other solution than the one you've identified may fit as well or better than yours
 - ❖ Try several initializations
 - Models that appeared identified at the latent class analysis stage may generate an error message at the regression stage
 - ❖ Estimability vs. global identifiability
 - ❖ May need to add further measurement model constraints
 - If fit very unstable: should reconsider using LCR at all

Parameter Constraints: Mplus Example

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 π in class 1

u1 has π equal to 1 in class 2

u2 and u3 have same π 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)