

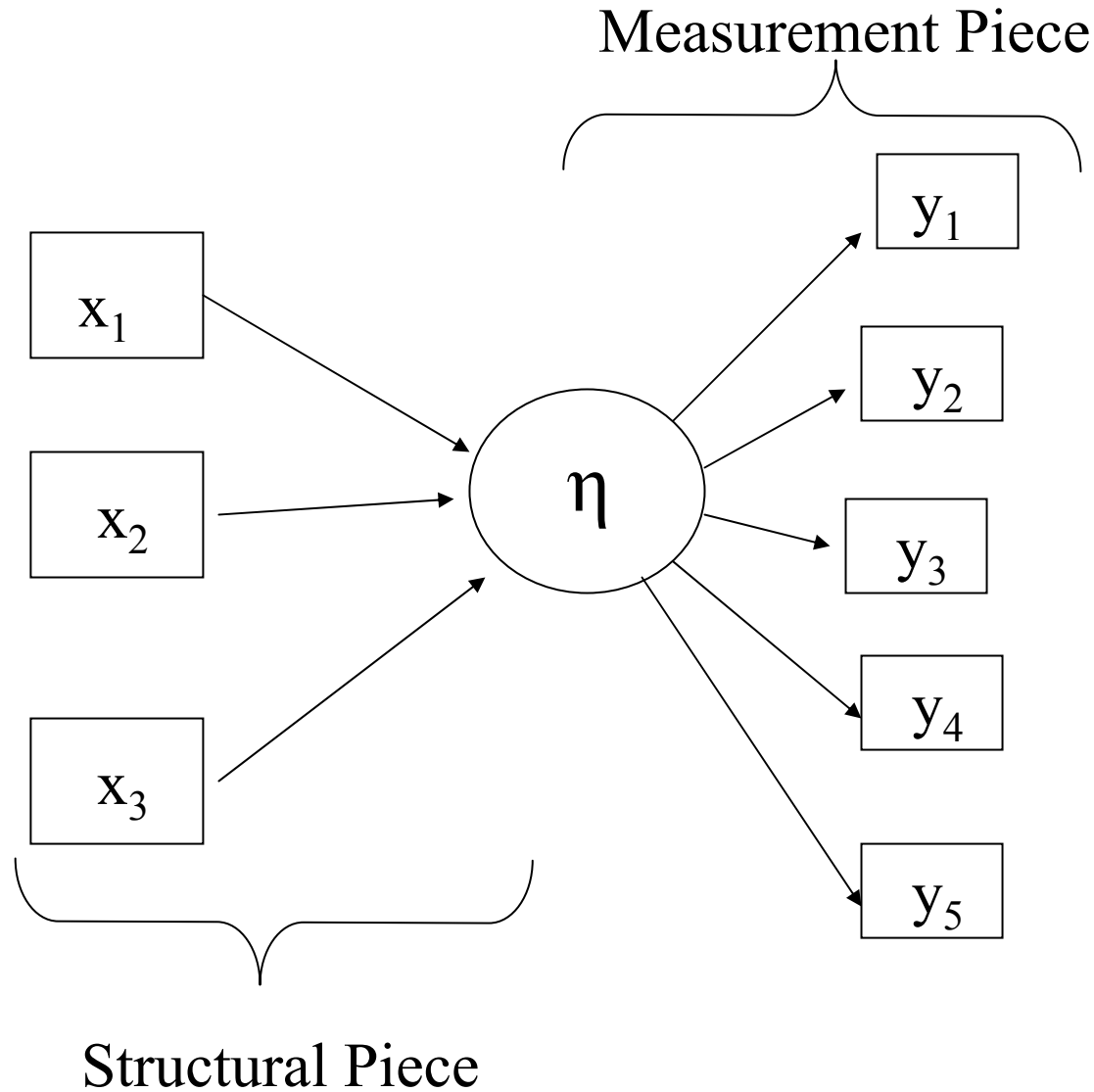
# Latent Class Regression

Statistics for Psychosocial  
Research II: Structural Models  
December 4 and 6, 2006

# 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



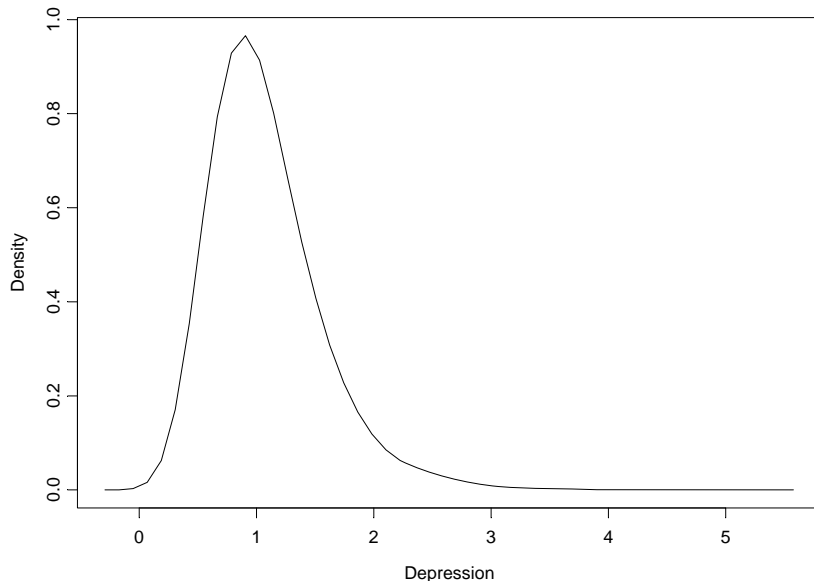
# When to use LCR

- Multiple discrete outcome variables
  - binary examples
    - yes/no questions
    - present/absent symptoms
  - all measuring same latent construct
  - We want to 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: Depression

Is depression continuous or categorical?

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



	<u>%</u>
class 1	80
class 2	15
class 3	5

# Recall LC model

- $M$ : number of latent classes
- $K$ : number of symptoms
- $p_{km}$ : probability of reporting symptom  $k$  given latent class  $m$
- $\pi_m$ : proportion of individuals in class  $m$
- $\eta_i$ : the true latent class of individual  $i$ ,  $i = 1, \dots, N$
- $m = 1, \dots, M$ ;  $k = 1, \dots, K$
- $y_{i1}, y_{i2}, \dots, y_{iK}$ : symptom presence/absence for individual  $i$ .

# ECA wave 3 data (1993)

- N=1126 in Baltimore
- Symptoms:
  - weight/appetite change
  - sleep problems
  - slow/increased movement
  - loss of interest/pleasure
  - fatigue
  - guilt
  - concentration problems
  - thoughts of death
  - dysphoria
- Covariates of interest
  - gender
  - age
  - marital status
  - education
  - income
- How are the above associated with depression?

# Assumptions

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



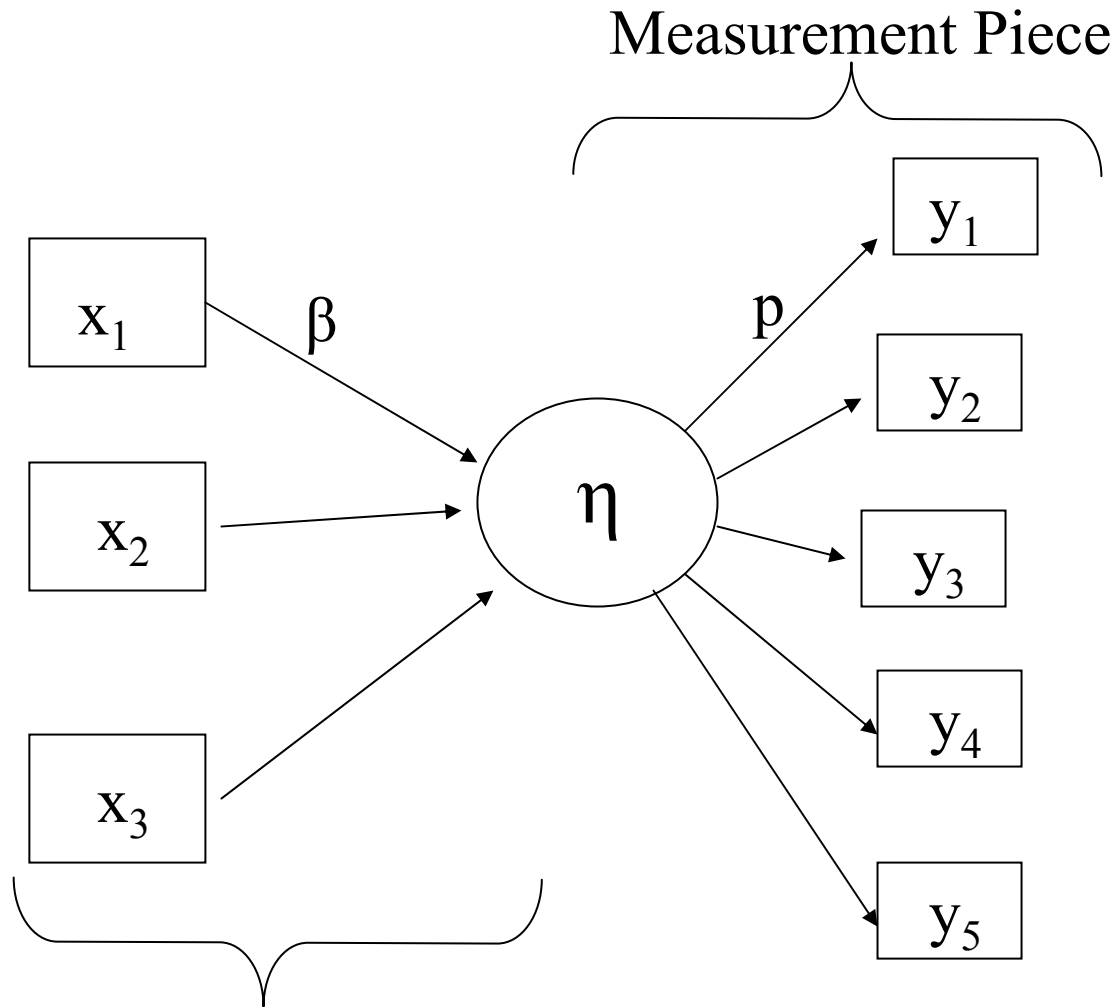
# Why LCR may be better than another analytic method

- 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
    - precision and estimates of regression coefficients reflect 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 dimensional (i.e. continuous)
    - identifiability harder to achieve (need big sample)
- Practically
  - Pro:
    - Allows for disease/disorder classification which is useful in a treatment vs. no treatment setting

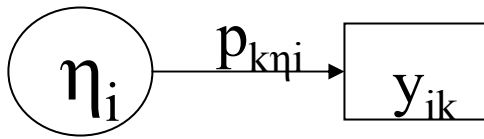
# Structural Equation-type Depiction



What are the parameters that the arrows represent? In other words, what are  $\beta$  and  $p$  in the LCR model?

# Parameter Interpretation

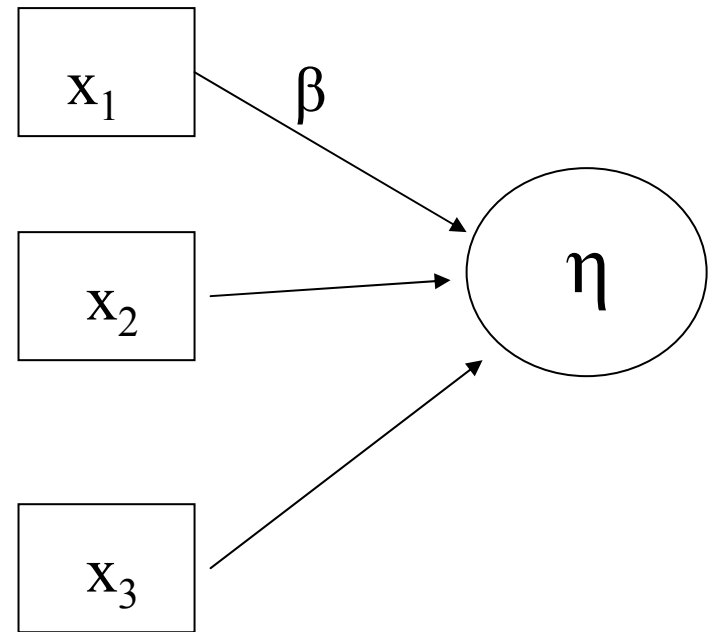
- Measurement Piece (p's)
  - $p_{km}$ : probability that an individual from class  $m$  reports symptom  $k$ .



- Same as standard latent class model from last term

# Parameter Interpretation

- How do we relate  $\eta$ 's and  $\beta$ 's?
- In “classic” SEM, we have linear model.
- What about when  $\eta$  is categorical?
- What if  $\eta$  is binary?



# Parameter Interpretation

- How do we relate  $\eta_i$  to  $x_i$ 's ?
- Consider simplest case: 2 classes

$$\log\left(\frac{P(\eta_i = 2)}{1 - P(\eta_i = 2)}\right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$$

or equivalently,

$$\log\left(\frac{\pi_{2i}}{\pi_{1i}}\right) = \log\left(\frac{P(\eta_i = 2)}{P(\eta_i = 1)}\right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$$

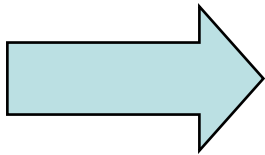
- $\beta_1$  and  $\beta_2$  are log odds ratios

# Model Results

- $\rho$ 
  - same as last term
  - $K \times M$   $\rho$ 's
- $\pi_{ji} = P(\eta_i = j)$ 
  - 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  $\pi_{ij}$  for all  $i$
- $\beta$ 
  - $(M-1) \times (H+1)$   $\beta$ 's where  $H$  = number of covariates
  - $M-1$ : one class is reference class so all of its  $\beta$  coefficients are technically zero
  - $H+1$ : for each class, there is one  $\beta$  for each covariate plus another for the intercept.

# Solving for $\pi_{ji} = P(\eta_i = j)$

$$\log\left(\frac{\pi_{2i}}{\pi_{1i}}\right) = \log\left(\frac{P(\eta_i = 2 | x_{1i}, x_{2i})}{P(\eta_i = 1 | x_{1i}, x_{2i})}\right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$$



$$\pi_{2i} = P(\eta_i = 2 | x_{1i}, x_{2i}) = \frac{e^{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}}}{1 + e^{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}}}$$

$$\pi_{1i} = P(\eta_i = 1 | x_{1i}, x_{2i}) = \frac{1}{1 + e^{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}}}$$



# Parameter Interpretation

Example:  $e^{\beta_1} = 2$  and  $x_{1i} = 1$  if female, 0 if male

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

$$e^{\beta_1} = \frac{P(\eta_i = 2 | x_{1i} = 1, x_{2i} = c)}{P(\eta_i = 1 | x_{1i} = 1, x_{2i} = c)} \bigg/ \frac{P(\eta_i = 2 | x_{1i} = 0, x_{2i} = c)}{P(\eta_i = 1 | x_{1i} = 0, x_{2i} = c)}$$

# More than two classes?

Need more than one equation

Need to choose a reference class

$$\log\left(\frac{P(\eta_i = 2 | x_{1i}, x_{2i})}{P(\eta_i = 1 | x_{1i}, x_{2i})}\right) = \beta_{02} + \beta_{12}x_{1i} + \beta_{22}x_{2i}$$

$$\log\left(\frac{P(\eta_i = 3 | x_{1i}, x_{2i})}{P(\eta_i = 1 | x_{1i}, x_{2i})}\right) = \beta_{03} + \beta_{13}x_{1i} + \beta_{23}x_{2i}$$

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

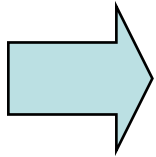
$e^{\beta_{13}}$  = OR for class 3 versus class 1 for females versus males

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

females versus males

# Solving for $\pi_{ji} = P(\eta_i)$

$$\log\left(\frac{\pi_{2i}}{\pi_{1i}}\right) = \log\left(\frac{P(\eta_i = 2)}{P(\eta_i = 1)}\right) = \beta_{02} + \beta_{12}x_{1i} + \beta_{22}x_{2i}$$



$$\begin{aligned}\pi_{2i} = P(\eta_i = 2) &= \frac{e^{\beta_{02} + \beta_{12}x_{1i} + \beta_{22}x_{2i}}}{1 + e^{\beta_{02} + \beta_{12}x_{1i} + \beta_{22}x_{2i}} + e^{\beta_{03} + \beta_{13}x_{1i} + \beta_{23}x_{2i}}} \\ &= \frac{e^{\beta_{02} + \beta_{12}x_{1i} + \beta_{22}x_{2i}}}{\sum_{r=1}^3 e^{\beta_{0r} + \beta_{1r}x_{1i} + \beta_{2r}x_{2i}}}\end{aligned}$$

Where we assume that  $\beta_{01} = \beta_{11} = \beta_{21} = 0$

## Depression Example:

LCR coefficients (**log ORs**) in 3 class model

	Class 3 vs 1	Class 2 vs 1	Class 3 vs 2
Log(age)	-1.2*	-1.5*	0.23
Female	0.85*	0.76*	0.09
Single	0.44	0.38	-0.05
Sep/wid/div	0.86*	0.83*	-0.01
HS diploma	-0.01	-0.56*	0.51

\* indicates significant at the 0.10 level

Note: class 1 is non-depressed, class 2 is mild, class 3 is severe

Depression Example:  
**ODDS RATIOS** in 3 class model

	Class 3 vs 1	Class 2 vs 1	Class 3 vs 2
Log(age)	0.3*	0.22*	1.26
Female	2.34*	2.13*	1.09
Single	1.55	1.46	0.95
Sep/wid/div	2.36*	2.29*	0.99
HS diploma	0.99	0.57*	1.67

\* indicates significant at the 0.10 level

Note: class 1 is non-depressed, class 2 is mild, class 3 is severe

# 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
- 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)

# Estimation

- Same *caveats* as last term
- Maximum likelihood:
  - Iterative fitting procedure.
  - Packages
    - Mplus
    - Splus, R
    - SAS
- Bayesian approach
  - Computationally intensive
    - WinBugs
    - Splus, R
    - SAS

# Properties of Estimates ( $\beta$ , $p$ )

- 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).



# Specifics Statistically

- Standard LCM Likelihood

$$\begin{aligned} P(Y_i = y_i) &= P(Y_{i1} = y_{i1}, Y_{i2} = y_{i2}, Y_{i3} = y_{i3}, Y_{i4} = y_{i4}, Y_{i5} = y_{i5}) \\ &= \sum_{m=1}^M \pi_{mi} \prod_{k=1}^K p_{km}^{y_{ik}} (1 - p_{km})^{(1-y_{ik})} \end{aligned}$$

- Latent Class Regression Likelihood

$$\begin{aligned} P(Y_i = y_i) &= P(Y_{i1} = y_{i1}, Y_{i2} = y_{i2}, Y_{i3} = y_{i3}, Y_{i4} = y_{i4}, Y_{i5} = y_{i5} \mid x) \\ &= \sum_{m=1}^M \pi_{mi}(x) \prod_{k=1}^K p_{km}^{y_{ik}} (1 - p_{km})^{(1-y_{ik})} \end{aligned}$$

where  $\pi_{mi}(x_i) = \frac{e^{\beta_m x_i}}{\sum_{m=1}^M e^{\beta_m x_i}}$

# Example: 3 class model

<b>coefficient</b>	<b>estimate</b>	<b>se</b>	<b>95% confidence interval</b>	
b02	-3.11	0.21	-3.52	-2.71
b01	-1.80	0.15	-2.08	-1.52
b2age	-1.21	0.74	-2.65	0.27
b3age	-1.44	0.53	-2.48	-0.38
b2sex	0.86	0.38	0.15	1.64
b3sex	0.77	0.25	0.32	1.34
<hr/>				
p[1,1]	0.83	0.06	0.69	0.93
p[1,2]	0.40	0.05	0.31	0.50
p[1,3]	0.02	0.01	0.01	0.03
p[2,1]	0.84	0.061	0.72	0.94
p[2,2]	0.41	0.05	0.31	0.52
p[2,3]	0.02	0.01	0.01	0.04

.....etc.

.

# Some Additional Concepts

- (1)  $\eta$  is a NOMINAL variable
- (2) Data Setup: Centering covariates can help.
  - Due to need to “initialize” algorithm in ML.
  - Due to priors on  $\beta$ 's in Bayesian setting
  - Will be meaningful in model checking, too.
  - Need to choose starting values for model estimation for regression coefficients in some ML packages. This is easier if they are centered.
  - Not an issue for Mplus: only need starting values for measurement part.

# Choosing Values for Initialization

## A: Measurement model

1. Use results from standard latent class model

## B: Structural piece

1. choose all  $\beta$ 's equal to 0 (will work if there is a LOT of data and no ID problems)
2. a. Make a “surrogate” latent class (e.g. choose cutoffs based on number of symptoms)
  - b. Perform “mlogit” on surrogate with covariates
  - c. Use log ORs as starting values

# Choosing Values for Initialization

3. Use ML “pseudo-class” approach
  - a. Using pseudo-classes from standard LC model, treat class assignment as fixed
  - b. Regress class membership on covariates (polytomous logistic regression)
  - c. Model building strategy -- gives initial idea of which covariates are associated.
  - d. Also, can use this as a model checking strategy post hoc
4. Use MCMC class assignment approach: same as 3, but with classes assigned using MCMC model

# Important Identifiability Issue

Must run model more than once using different starting values to check identifiability!

# Model Checking

- Very important step in LCR
- LCR can give misleading findings if measurement model assumptions are violated
- Two types of model checks:
  - (1) model fit
    - “do y patterns behave as model would predict?”
  - (2) violation of assumptions
    - “do y’s relate to x’s as expected?”

# ECA wave 3 data (1993)

- N=1126 in Baltimore
- Symptoms:
  - weight/appetite change
  - sleep problems
  - slow/increased movement
  - loss of interest/pleasure
  - fatigue
  - guilt
  - concentration problems
  - thoughts of death
  - dysphoria
- Covariates of interest
  - gender
  - age
  - marital status
  - education
  - income
- How are the above associated with depression?



# Models

- Model A:  $\log(\text{age})$ , gender, race
- Model B:  $\log(\text{age})$ , gender, race, diploma

# Do y patterns behave as model predicts?

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

o = 2 class  
 x = 3 class  
 ▲ = 4 class

Figure 5: PFC for 2, 3, and 4 class models

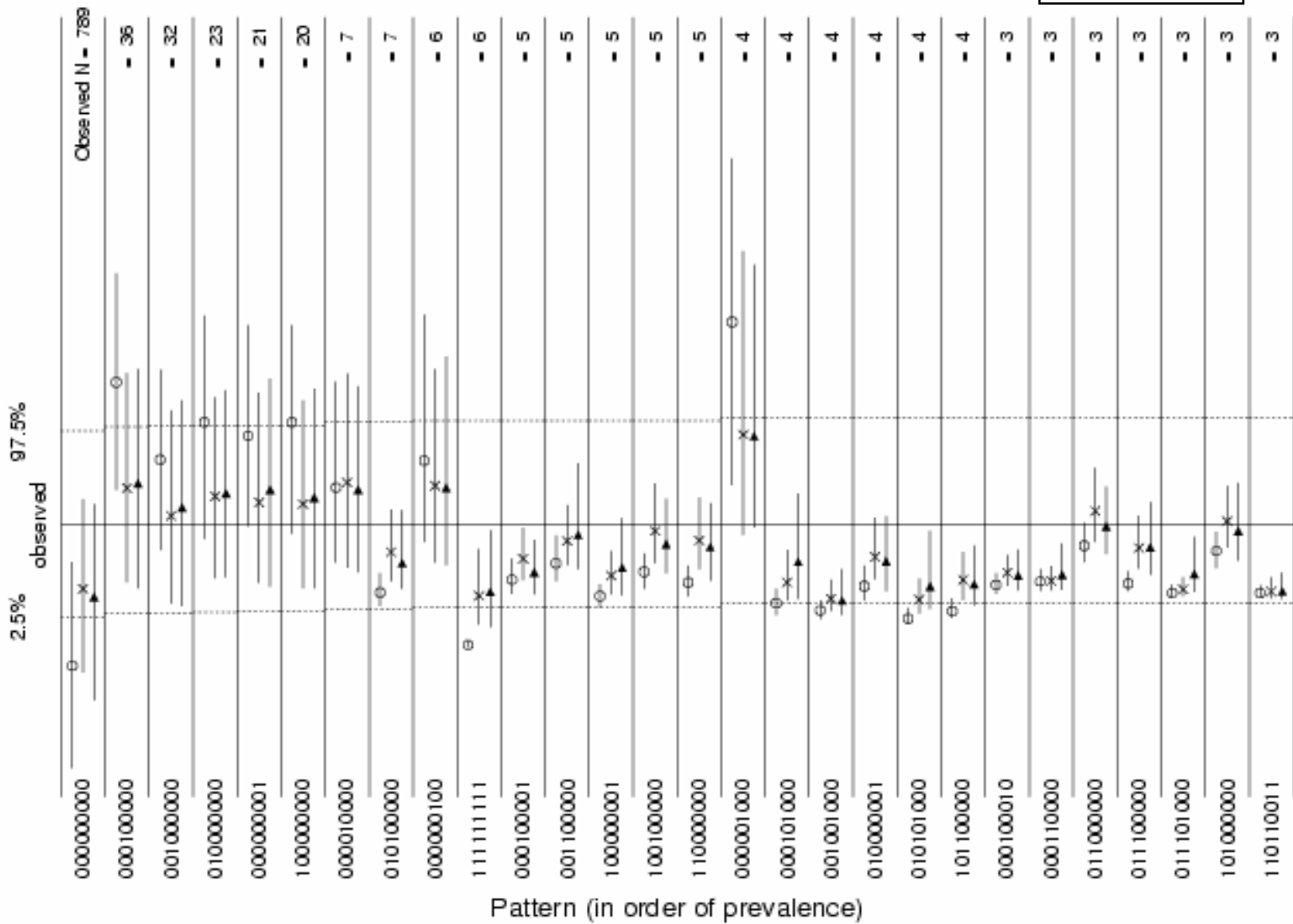
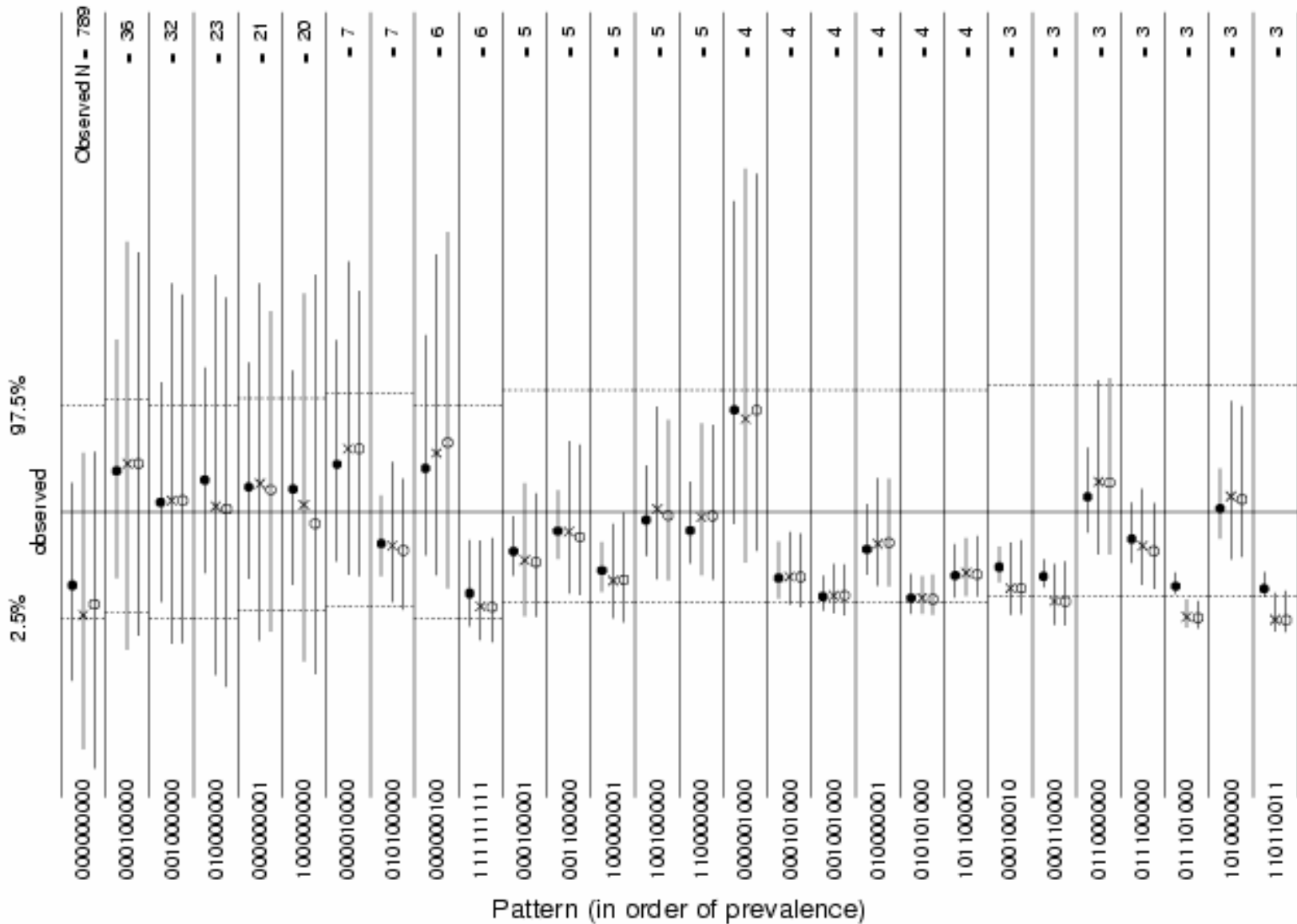


Figure 7: PFC for Standard 3 Class Model, Model A, and Model B

● = LCA  
 x = LCR-A  
 o = LCR-B



# Does pattern frequency behave as predicted by covariates?

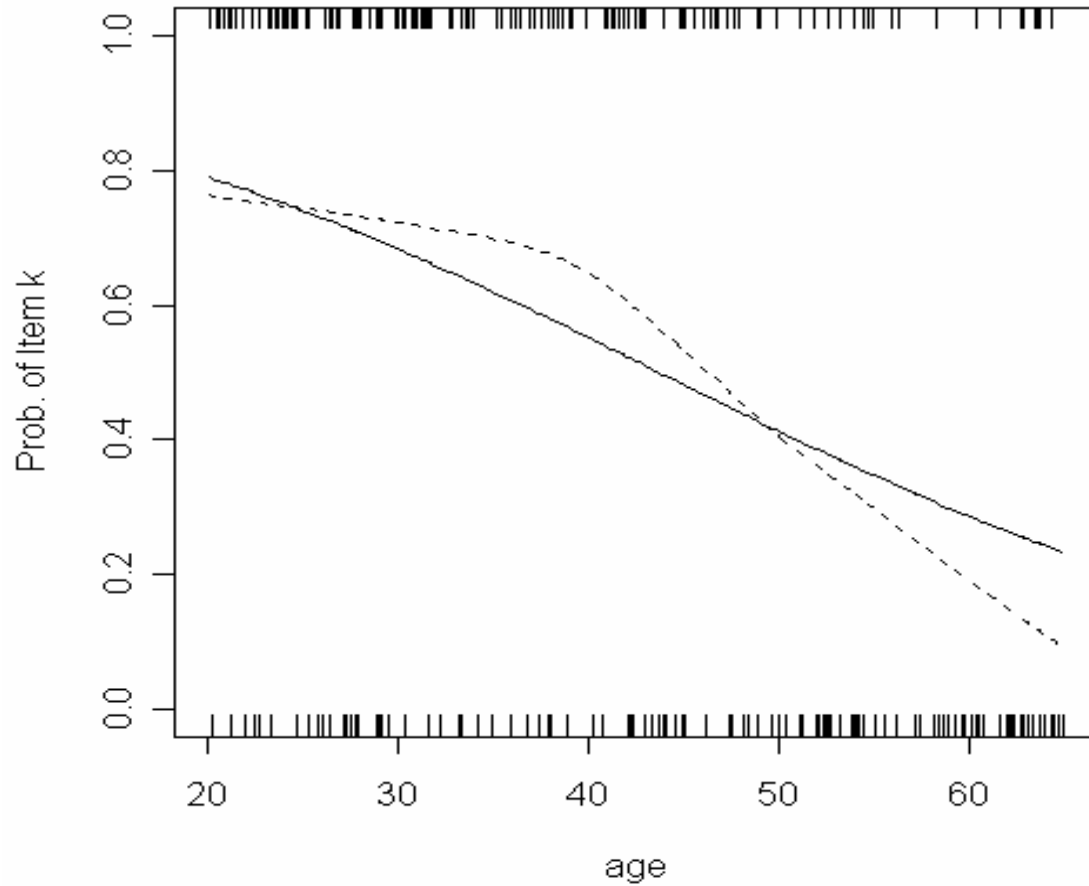
- Idea: focus on one item at a time
- Recall:

$$P(Y_{1i} = y_{1i}, \dots, Y_{Ki} = y_{ki} \mid x_i) = \sum_{m=1}^M \pi_i(x_i) \prod_{k=1}^K p_{km}^{y_{ki}} (1 - p_{km})^{(1-y_{ki})}$$

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

$$P(Y_{ri} = y_{ri} \mid x_i) = \sum_{m=1}^M \pi_i(x_i) p_{rm}^{y_{ri}} (1 - p_{rm})^{(1-y_{ri})}$$

# Comparing Fitted to Observed




# Categorical Covariates

- Easier than continuous (computationally)
- Example
  - Calculate:
    - Predicted males with guilt
    - Observed males with guilt
    - Predicted females with guilt
    - Observed females with guilt

- Assume LC regression model with only gender
- Gender = 0 if male, 1 if female
- Item of interest if guilt.
- Want find how many class 2 men we would expect to report guilt based on the model

$$\begin{aligned}
 P(\text{guilt and male and class} = 2) &= P(\text{guilt and class} = 2 | \text{male}) P(\text{male}) \\
 &= P(\text{guilt} | \text{male and class} = 2) P(\text{class} = 2 | \text{male}) P(\text{male}) \\
 &= P(\text{guilt} | \text{class} = 2) P(\text{class} = 2 | \text{male}) P(\text{male}) \\
 &= p_{km} \times \frac{e^{\beta_0}}{1 + e^{\beta_0}} \times P(\text{male})
 \end{aligned}$$

  $Expected(\text{guilt and male and class}=2) = N \times p_{km} \times \frac{e^{\beta_0}}{1 + e^{\beta_0}} \times P(\text{male})$

Calculate this for each of the classes and sum up:  
 Will tell us the expected number of males reporting guilt.



# Failure in Fit

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

# Checking Non-differential Measurement Assumption

- For binary covariates and for each class  $m$  and item  $k$  consider

$$OR_{kmx} = \frac{P(y_k = 1 | x = 1, \eta = m) / P(y_k = 0 | x = 1, \eta = m)}{P(y_k = 1 | x = 0, \eta = m) / P(y_k = 0 | x = 0, \eta = m)}$$

- If assumption holds, this OR will be approximately equal to 1.
- Why may this get tricky?
  - We don't KNOW class assignments.
  - Need a strategy for assigning individuals to classes.

# Checking NDM: Maximum Likelihood Approach

- (a) assign individuals to “pseudo-classes” based on posterior probability of class membership
  - recall posterior probability based on observed pattern
  - e.g. individual with 0.20, 0.05, 0.75
    - better chance of being in class 3
    - not necessarily in class 3
- (b) calculate OR's within classes.
- (c) repeat (a) and (b) at least a few times
- (d) compare OR's to 1.

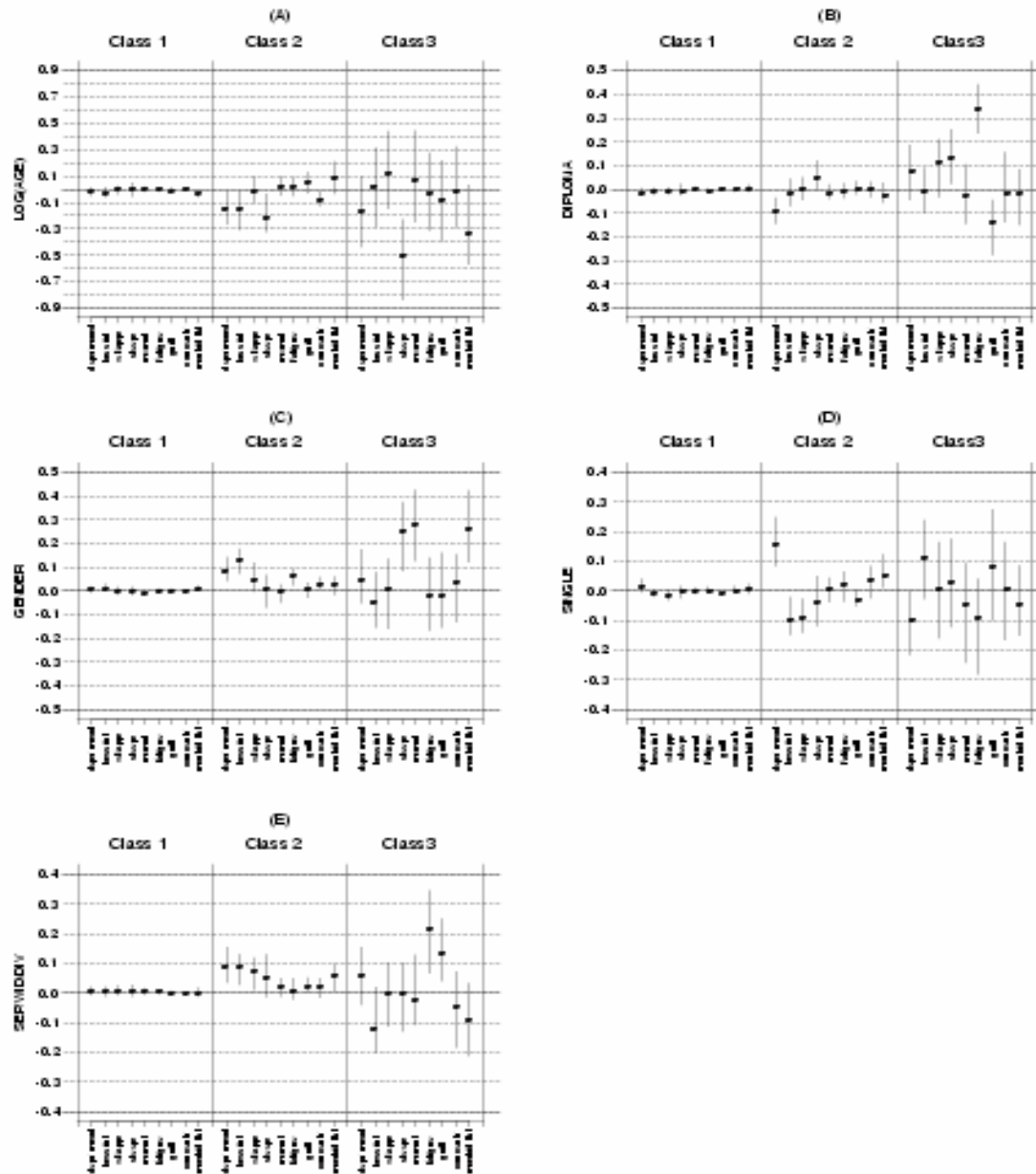
# Checking NDM: Maximum Likelihood Approach

- What about continuous covariates?
- Use same general idea, but estimate the logOR within classes by logistic regression
- Example: age

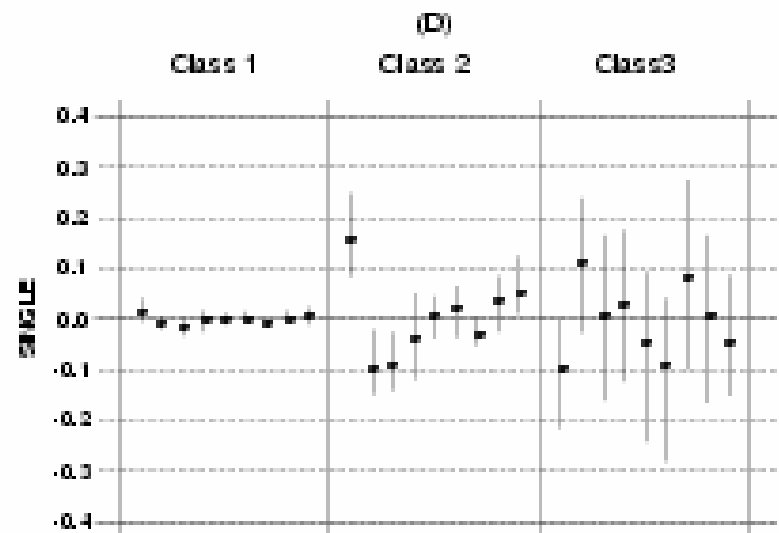
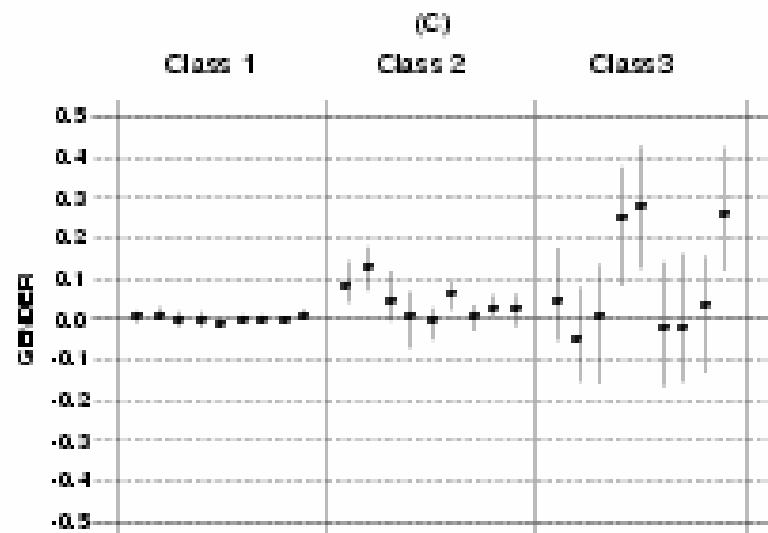
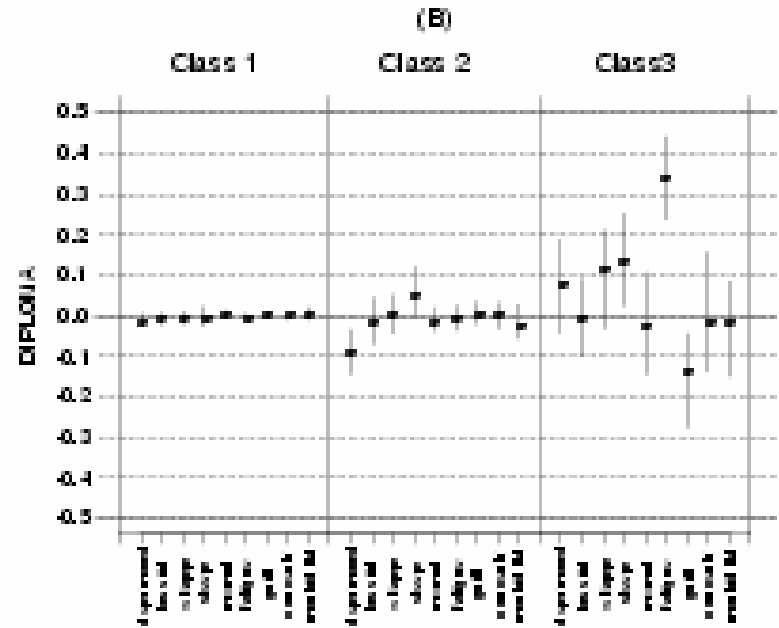
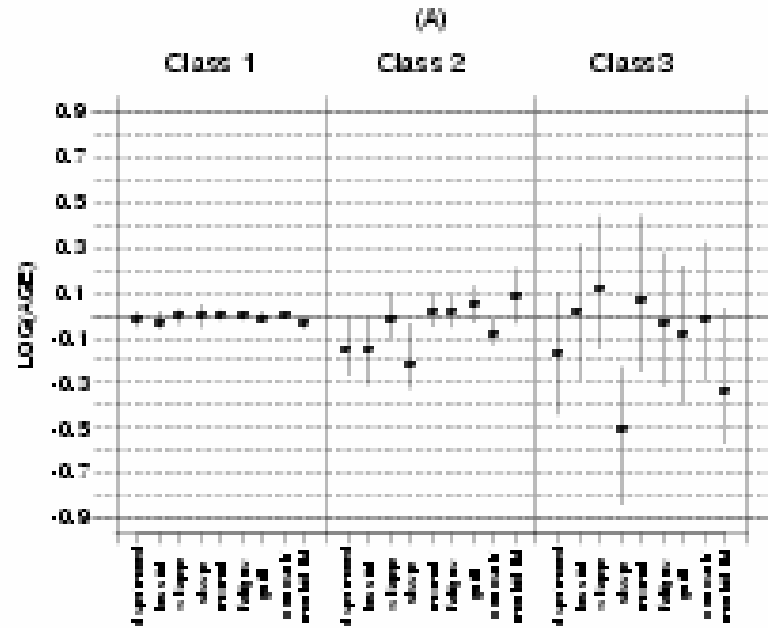
# Checking NDM: MCMC (Bayesian) approach

- At each iteration in Gibbs sampler, individuals are automatically assigned to classes  $\longrightarrow$  no need to “manually” assign.
- At each iteration, simply calculate the OR’s of interest.
- Then, “marginalize” or average over all iterations.
- Results is posterior distribution of OR

Figure 2 displays the estimated values for the log odds ratios. Vertical lines range from the 2.5th percentile to the 97.5th percentile of the posterior distribution. Posterior median estimates are plotted with  $\bullet$ . Vertical lines which do not overlap 0 indicate evidence of violation of differential measurement assumption.



estimates are plotted with "o". Vertical lines which do not overlap 0 indicate evidence of violation of differential measurement assumption.



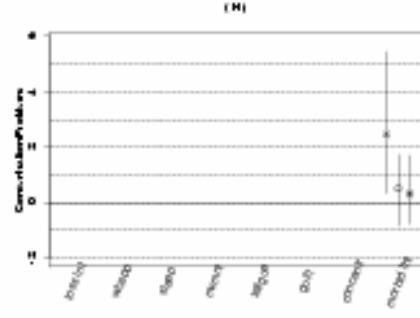
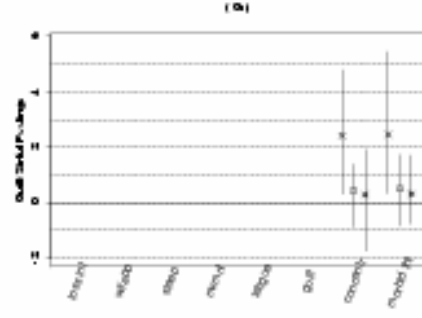
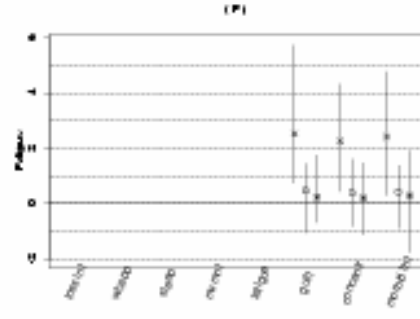
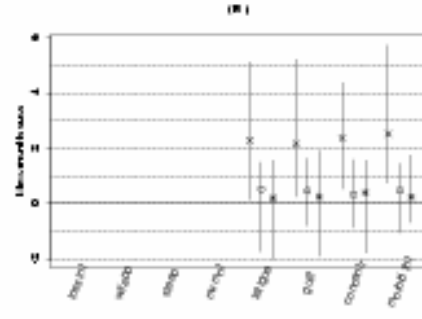
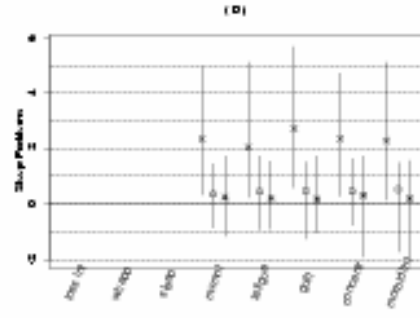
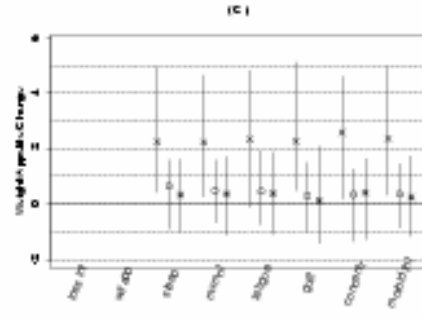
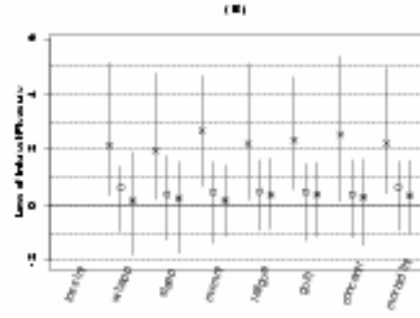
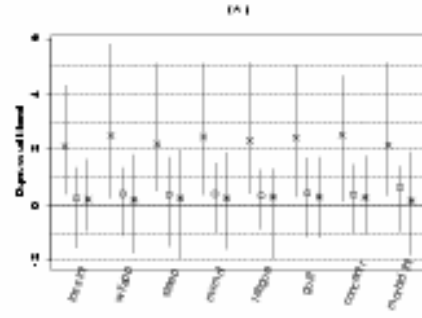
# Checking Conditional Independence Assumption

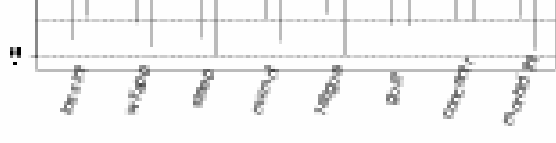
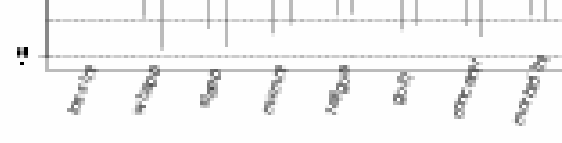
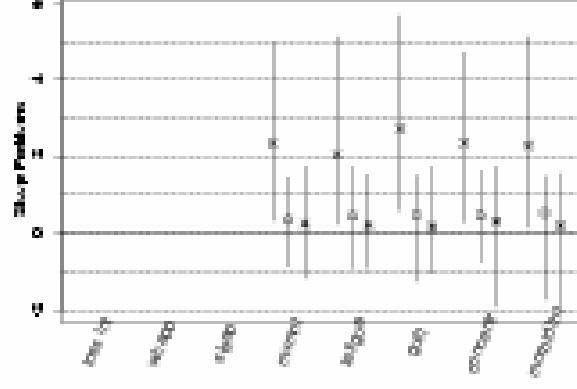
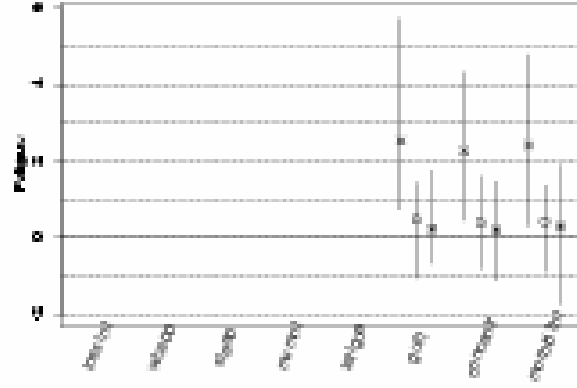
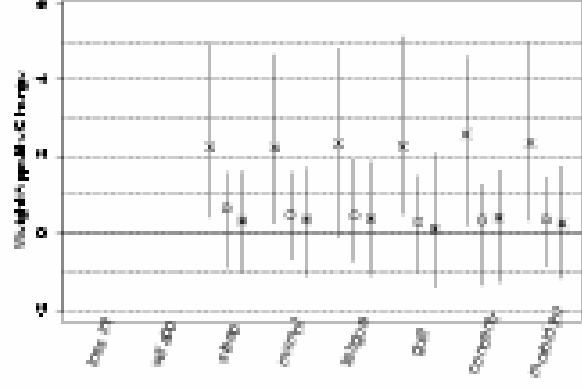
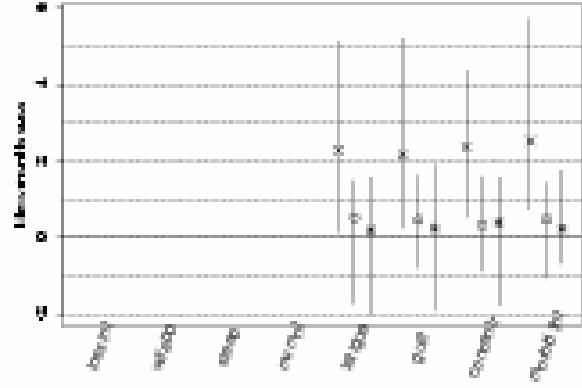
- In words, within a class, there is no association between  $y_k$  and  $y_j$ ,  $j \neq k$ .
- Same approach
- Only difference:

$$OR_{jkm} = \frac{P(y_j = 1, y_k = 1 | \eta = m) / P(y_j = 0, y_k = 1 | \eta = m)}{P(y_j = 1, y_k = 0 | \eta = m) / P(y_j = 0, y_k = 0 | \eta = m)}$$

- Still use “pseudo-class” assignment (ML) or class assignment at each iteration (MCMC)







# Identifiability (briefly)

- General Idea: different parameters can lead to the same model fit
- 2 step rule: If
  - (a) polytomous logistic regression is ID'ed
  - (b) standard LCM is ID'edThen model is ID'ed
- t-rule: need more data cells than parameters
  - complication: continuous covariates, but they usually don't make unID'ed.

# 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 lead to believe that LCR is not appropriate.