

Categorical and Zero Inflated Growth Models



Alan C. Acock*

July, 2008

*Alan C. Acock, Department of Human Development and Family Sciences, Oregon State University, Corvallis OR 97331 (alan.acock@oregonstate.edu) . This was supported in part by 1R01DA13474, The Positive Action Program: Outcomes and Mediators, A Randomized Trial in Hawaii and R305L030072 CFDA U.S. Department of Education; Positive Action for Social and Character Development. Randomized trial in Chicago, Brian Flay, PI. ; and R215S020218 CFDA, Uintah Character Education Randomized Trial, U.S. Department of Education.

Topics to Be Covered



- Predicting Rare Events
- Binary Growth Curves
- Count Growth curves
- Zero-Inflated Poisson Growth Curves
- Latent Class Zero-Inflated Poisson Models
- A detailed presentation of the ideas is available at
www.oregonstate.edu/~acock/growth

Predicting Rare Events

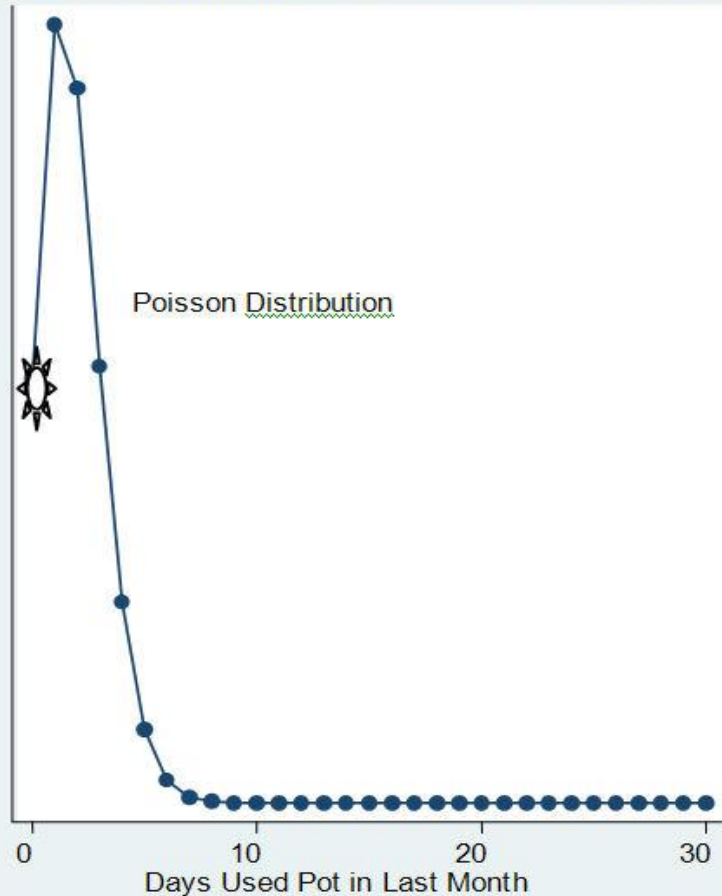


- Physical conflict in romantic relationships
- Frequency of depressive symptoms
- Frequency of Parent-Child Conflict
- Frequency of risky sex last month

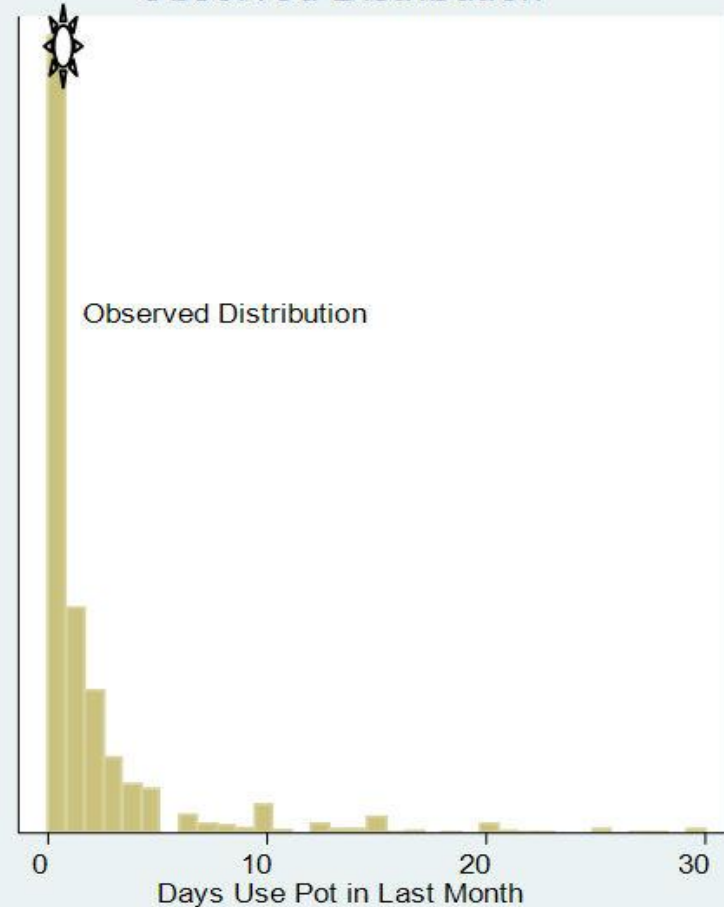
When a Poisson Distribution Fails: An Excess of Zeros

Poisson Distribution Fails: An Excess of Zeros

Expected Distribution for Poisson Variate



Observed Distribution



Binary: Does Behavior Occur



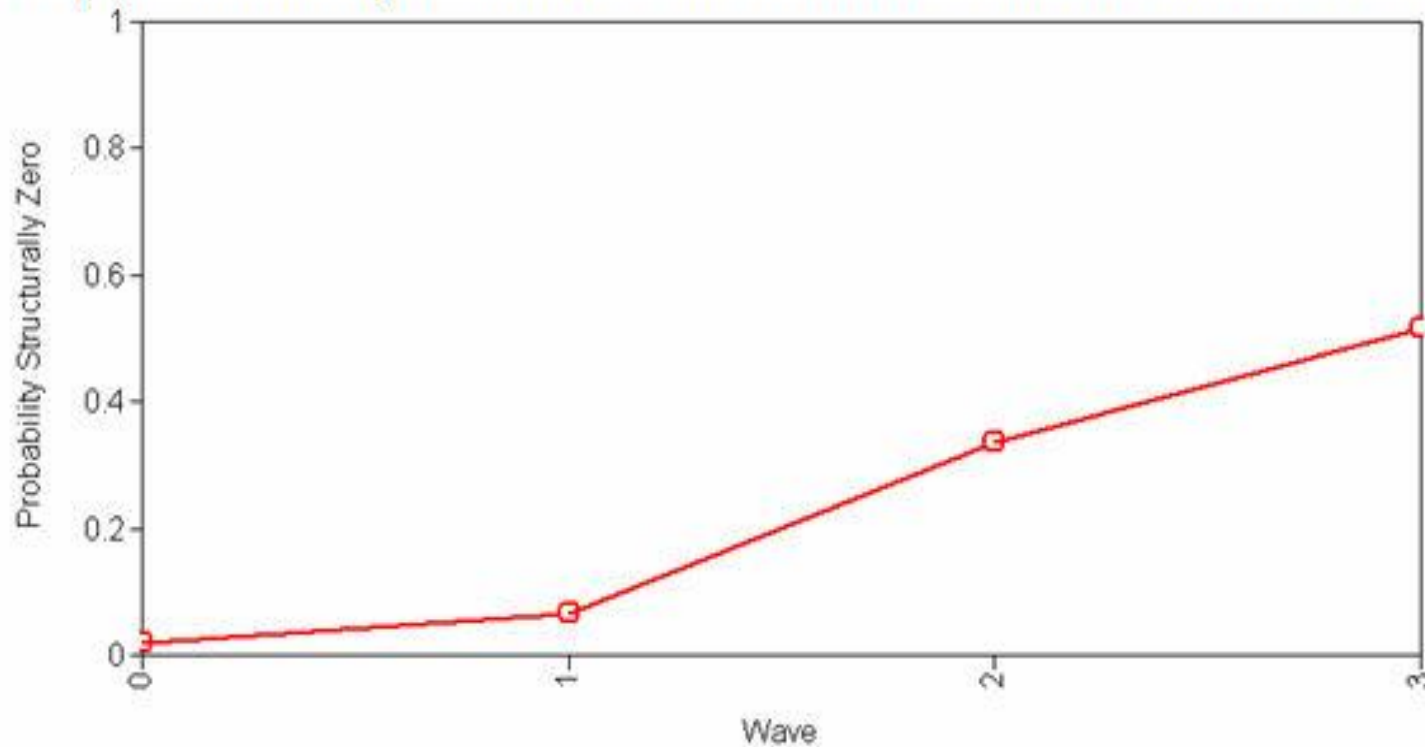
- **Structural zeros**—behavior is not in behavioral repertoire
 - Do not smoke marijuana \therefore Didn't smoke last month
- **Chance zeros**—Behavior part of repertoire, just not last month
 - No fight with spouse last week, but you should have seen the week before that!

Count Component

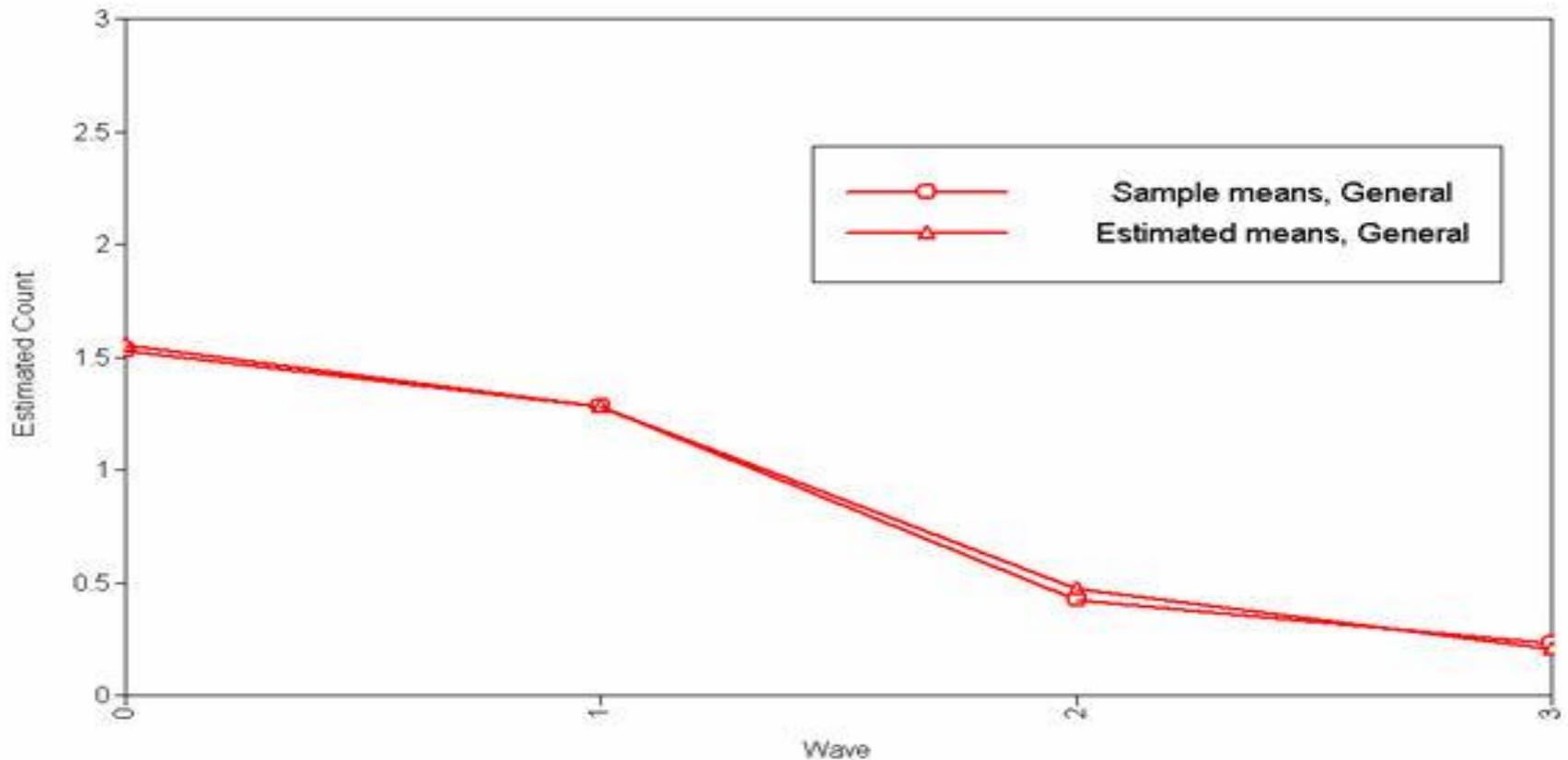
- **Two-part model**
 - Equation for zero vs. not zero
 - Equation for those not zero. Zeros are missing values
- **Zero-inflated model**—includes both those who are structural zeros and chance zeros

Trajectory of the Likelihood of the Behavior Occurring?

Graph of Binary Part of Zero-Inflated Model with No Covariates



Trajectory of the Count of the Behavior Occurring?



Why Aren't Both Lines Straight?

- We use a linear model of the growth curve
- We predict the **log of the expected count**
- We predict **log odds for the binary component**

Why Aren't Both Lines Straight?

- For the count we are predicting
 - Expected $\ln(\lambda) = \alpha + \beta T_i$
 - T_i (0, 1, 2, . . .) is the time period
 - α is the intercept or initial value
 - β is the slope or rate of growth
- Expected probability or expected count, are no longer linear

Time Invariant Covariates



- Time invariant covariates** are constants over the duration of study
- May influence growth in the binary and count components
 - May influence initial level of binary and count components
 - Different effects a major focus

Time Invariant Covariates



- Mother's education might influence likelihood of being structurally zero
- Mother's education might be negatively related to the rate of growth

Time Varying Covariates



Time Varying Covariates—variables that can change across waves

- Peer pressure may increase each year between 12 and 18
- The peer pressure each wave can directly influence drug usage that year

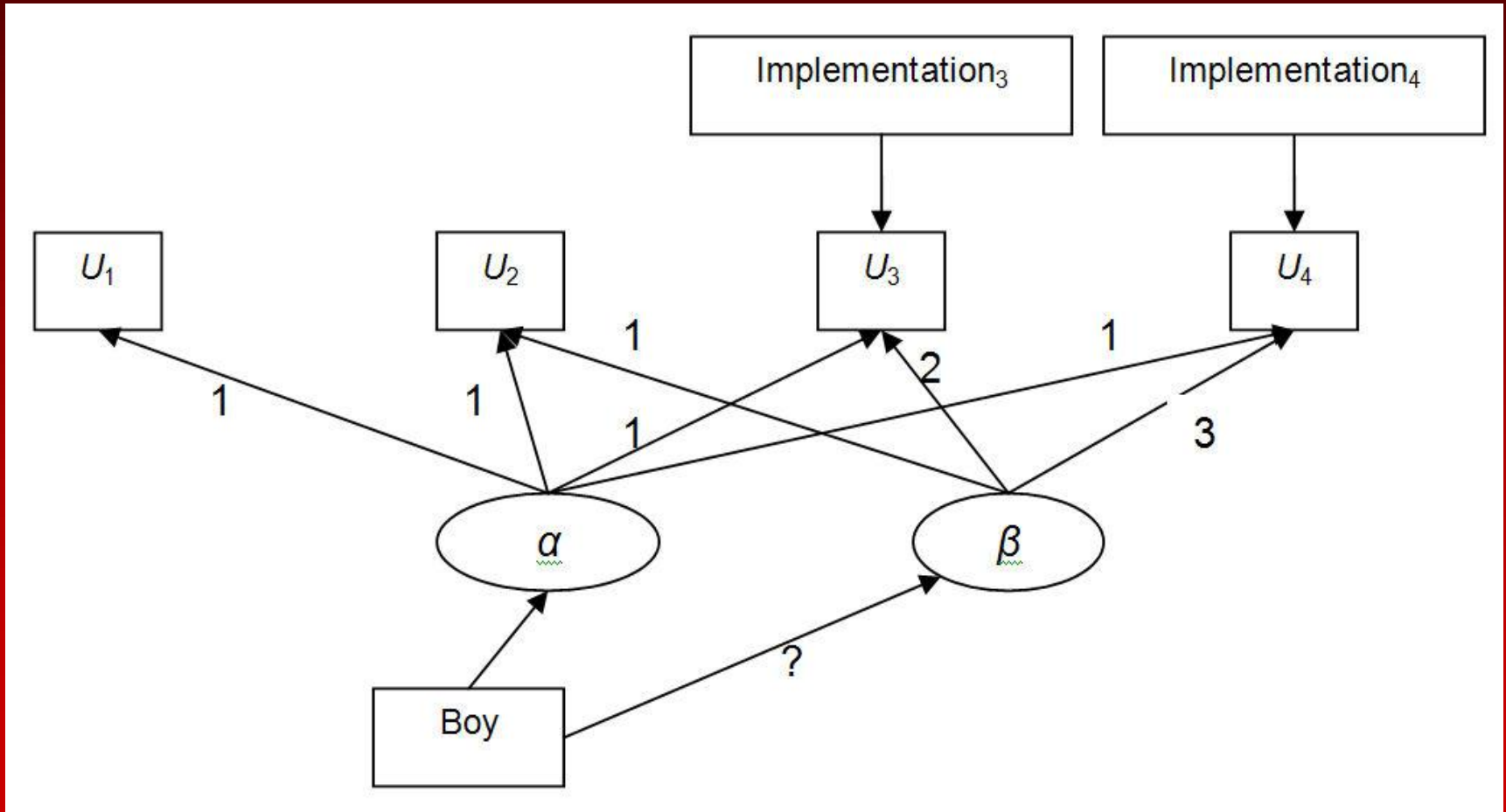
Estimating a Binary Growth Curve



Example of Binary Component

- Brian Flay has a study in Hawaii evaluating the **Positive Action Program** in Grades 1 - 4
 - Key outcome—reducing negative responses to behaviors that Positive Action promotes
 - Gender is a time invariant covariate—boys higher initially but to have just as strong a negative slope
 - Level of implementation is a time varying covariate—the nearly 200 classrooms vary. A Latent Profile Analysis produced two classes on implementation

Binary Model for Reducing Negative Responses to Good Behavior

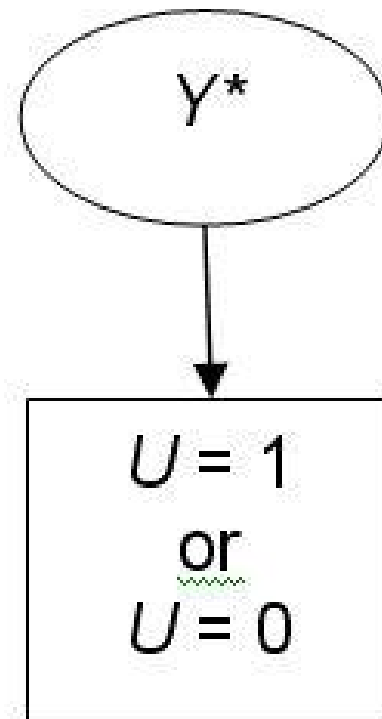


Intercept and Slope are Continuous: Threshold

A Continuous Latent Factor and a Binary Response Variable

Thresholds

Rule: $U = 1$ if $Y^* \geq \tau$,
 $U = 0$ if $Y^* \leq \tau$



Binary Growth Curve Program— Part 1

Title: workshop binary growth.inp

Data: File is workshop_growth.dat ;

Variables:

Names are

idnum s1flbadc s2flbadc s3flbadc s4flbadc male s1flbadd s2flbadd
s3flbadd s4flbadd s1flbadm s2flbadm s3flbadm s4flbadm c3 c4
s3teacher room ;

Usevariables are male s1flbadd s2flbadd s3flbadd s4flbadd c3 c4 ;

Categorical are s1flbadd s2flbadd s3flbadd s4flbadd ;

Missing are all (-9999) ;

Analysis:

Type = Missing ;

Estimator = ML ;

Binary Growth Curve Program— Part 2



Model:

```
alpha beta | s1flbadd@0 s2flbadd@1 s3flbadd@2 s4flbadd@3 ;  
alpha on male ;  
beta on male ;  
s3flbadd on c3 ;  
s4flbadd on c4 ;
```

Output:

```
Patterns sampstat standardized tech8;
```

Sample Proportions and Model Fit

S1FLBADD

Category 1 0.331

Category 2 0.669

S2FLBADD

Category 1 0.372

Category 2 0.628

S3FLBADD

Category 1 0.547

Category 2 0.453

S4FLBADD

Category 1 0.744

Category 2 0.256

Loglikelihood

H0 Value -2049.169

Information Criteria

Number of Free Parameters 9

Akaike (AIC) 4116.338

Bayesian (BIC) 4160.390

Sample-Size

Adjusted BIC 4131.806

Proportion of Negative Responses drop each year

Model Estimates-

		Estimates	S.E.	Est./S.E.	Std	StdYX
ALPHA	ON					
MALE		0.548	0.184	2.980	0.464	0.232
BETA	ON					
MALE		0.033	0.088	0.371	0.077	0.038
S3FLBADD	ON					
C3		-0.231	0.085	-2.714	-0.231	-0.055
S4FLBADD	ON					
C4		-0.642	0.144	-4.476	-0.642	-0.147
BETA	WITH					
ALPHA		-0.344	0.217	-1.581	-0.686	-0.686
Intercepts						
ALPHA		0.000	0.000	0.000	0.000	0.000
BETA		-0.475	0.094	-5.078	-1.120	-1.120

*Current version of Mplus print Std (STDY) and StdYX seaparately.

Gender Effects



- Unstandardized effect of male on the intercept, α , is .548, $z = 2.98$, $p < .01$
- Standardized Beta weight is .232
- **Partially standardized (standardized on latent variable only) is .464**
- Path to slope is not significant, $B = .03$, partially standardized path is .08
- However effective the program is at reducing negative feelings, it is about **as effective for boys as for girls**

Implementation Effects



- Wave 3—Unstandardized effect of implementation for the Binary Component has a $B = -.23$, $z = -2.71$, $p < .05$ --
Exponentiated odds ratio is $e^{-.23} = .79$
- Wave 4—the unstandardized effect of implementation for the Binary Component has a $B = -.64$, $z = -4.476$, $p < .001$ --
Exponentiated odds ratio is $e^{-.64} = .53$

MODEL RESULTS (cont.)

Thresholds	Estimates	S.E.	Est./S.E.	Std	StdYX
S1FLBADD\$1	-0.714	0.139	-5.137	-0.714	-0.330
S2FLBADD\$1	-0.714	0.139	-5.137	-0.714	-0.349
S3FLBADD\$1	-0.714	0.139	-5.137	-0.714	-0.354
S4FLBADD\$1	-0.714	0.139	-5.137	-0.714	-0.342
Residual Variances					
ALPHA	1.321	0.529	2.497	0.946	0.946
BETA	0.180	0.113	1.589	0.999	0.999
LOGISTIC REGRESSION ODDS RATIO RESULTS					
S3FLBADD	ON				
C3		0.794			
S4FLBADD	ON				
C4		0.526			

Thresholds & Graphs



- Mplus does not graph estimated probabilities when there are covariates because variances depend on the covariate level
- \therefore We cannot estimate initial probability using threshold value. Could if no covariates
- If you want a series of graphs (e.g., boy/low intervention both wave 3 and wave 4), you need to treat each combination as a separate group
- Each group would have no covariates; just be a subset of children.
- Results might not be consistent with the model using all of the data

Estimating a Count Growth Curve



Count Component

- Mplus uses a Poisson Distribution for estimating counts
 1. The Poisson distribution is a single parameter distribution with $\lambda = M = \sigma^2$
 2. Without adjusting for the excess of zeros, the $\sigma^2 > M$

Count Component

$$\Pr(Y = k) = \frac{e^{-\lambda} \lambda^k}{k!}$$

where $k = \text{count}$

$$\lambda = E(Y) = \text{Var}(Y)$$

Example--What is Probability of exactly 0, 1, or 2 successes, if $\lambda = 1.8345$

$$\Pr(Y = 0) = \frac{e^{-1.8345} \times 1.8345^0}{0!} = e^{-1.8345} = .16$$

$$\Pr(Y = 1) = \frac{e^{-1.8345} \times 1.8345^1}{1!} = .29$$

$$\Pr(Y = 2) = \frac{e^{-1.8345} \times 1.8345^2}{2!} = .27$$

Count Program—Part 1

Title: workshop count growth fixed effects.inp

Data:

File is workshop_growth.dat ;

Variable:

Names are

idnum s1flbadc s2flbadc s3flbadc s4flbadc male
s1flbadd s2flbadd s3flbadd s4flbadd s1flbadm

s2flbadm

s3flbadm s4flbadm c3 c4 s3techer room ;

Usevariables are s1flbadc s2flbadc s3flbadc
s4flbadc ;

Missing are all (-9999) ;

Count are s1flbadc s2flbadc s3flbadc s4flbadc ;

Count Program—Part 2

Model:

```
alpha beta | s1flbadc@0 s2flbadc@1 s3flbadc@2 s4flbadc@3 ;  
alpha@0 ; !fixes variance of intercept at zero-optional  
beta@0 ; !fixes variance of slope at zero-optional
```

Output:

```
residual tech1 tech4 tech8;
```

Plot:

```
Type = Plot3 ;  
Series = s1flbadc s2flbadc s3flbadc s4flbadc(*) ;
```

Count Model Output-

MODEL RESULTS	Estimates	S.E.	S.E./Est.
Means			
ALPHA	0.559	0.018	31.199
BETA	-0.644	0.012	-53.641
Variances			
ALPHA	0.000	0.000	0.000
BETA	0.000	0.000	0.000

Interpreting the Est. Intercept



- We fixed the residual variances at zero
- The mean intercept is .56, $z = 31.199$, $p < .001$
- We can exponentiate this when there are no covariates to get the expected count at the intercept, $e^{.56} = 1.75$

Interpreting the Est. Slope

- The mean slope is $-.64$, $z = -53.64$, $p < .001$. With no covariates we use exponentiation to obtain the expected count for each wave

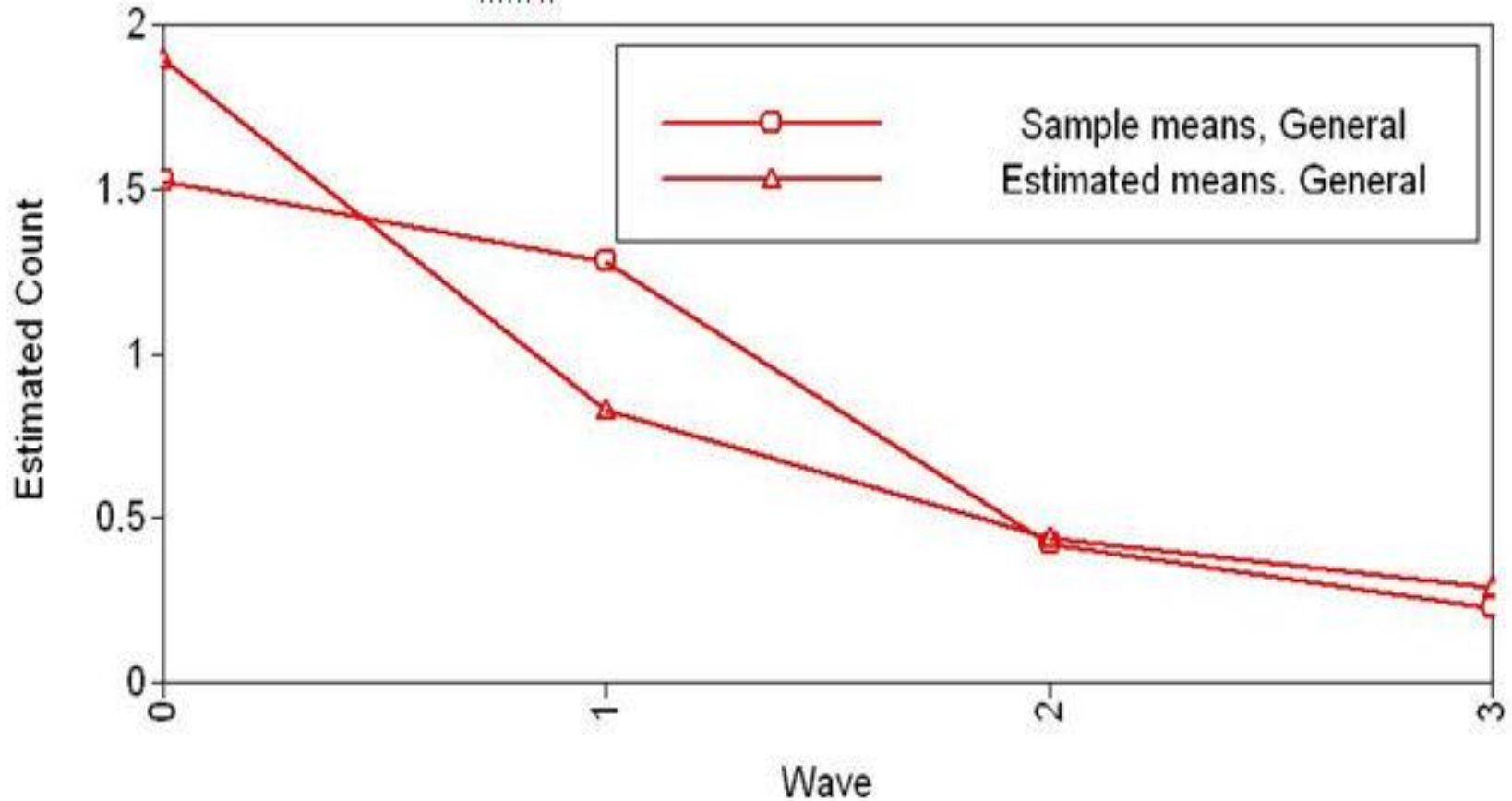
$$\text{Expected count (wave1)} = e^{\alpha} \times e^{\beta \times 0} = 1.75$$

$$\text{Expected count (wave2)} = e^{\alpha} \times e^{\beta \times 1} = .92$$

$$\text{Expected count (wave3)} = e^{\alpha} \times e^{\beta \times 2} = .48$$

$$\text{Expected count (wave4)} = e^{\alpha} \times e^{\beta \times 3} = .25$$

Sample and Estimated Count



Interpreting Model Results: W/O fixed variances of α & β

MODEL RESULTS

	Estimates	S.E.	Est./S.E.	Std	StdYX
Beta WITH					
Alpha	-0.086	0.021	-4.056	-0.276	-0.276
Means					
Alpha	0.396	0.028	14.194		
Beta	-0.842	0.021	-40.115		
Variances					
Alpha	0.483	0.040	12.064		
Beta	0.199	0.016	12.077		

Was .559

Was -.644

Putting the Binary and Count Growth Curves Together



Putting the Binary and Count Models Together

Two-Part Solution

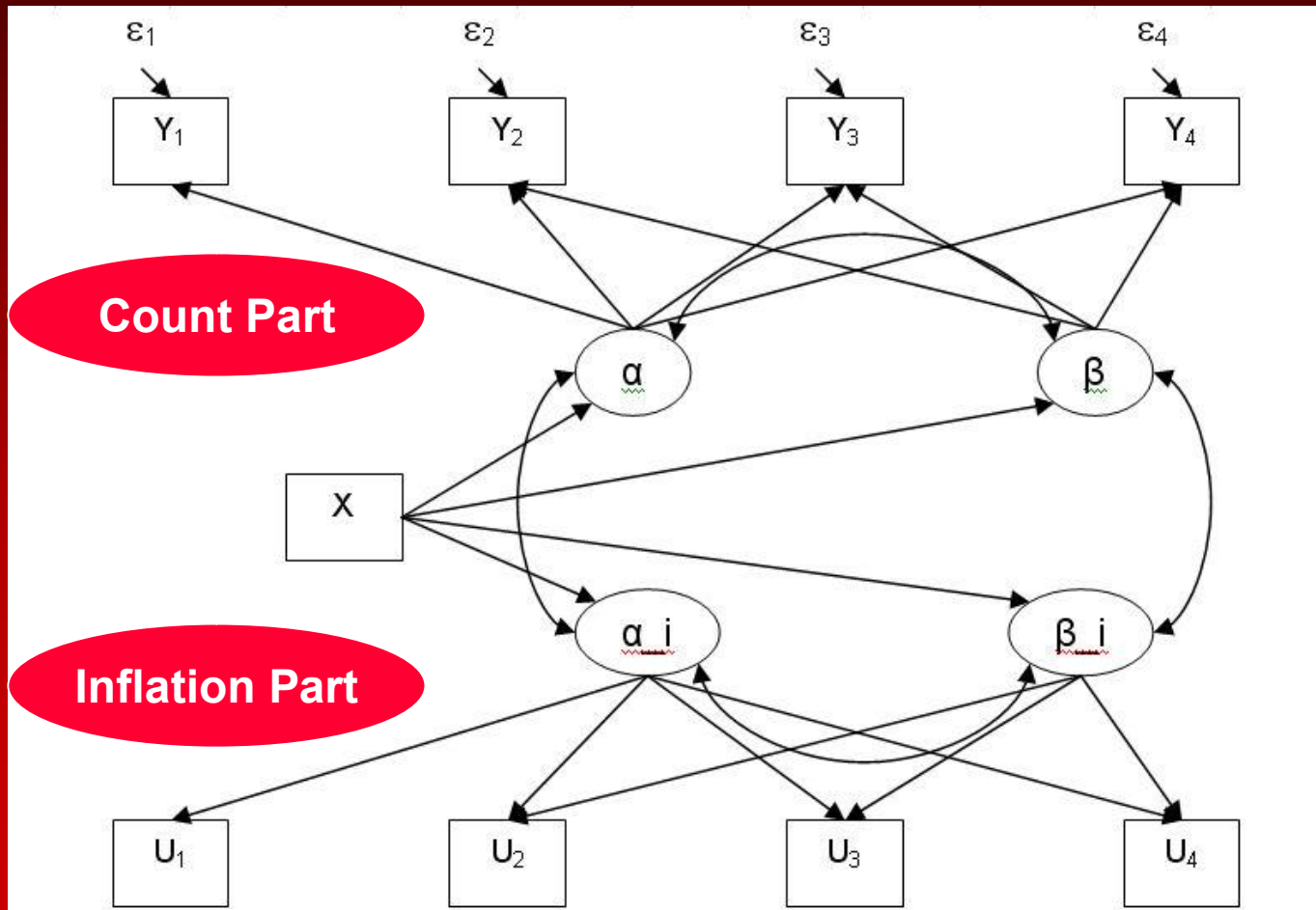
- First part models binary outcome as we did here with binary data
- Second part deletes all people who have a count of zero at any wave. This leaves only children who have a count of at least 1 for every wave
- Second part estimated using a Poisson Model

Putting the Binary and Count Models Together

Zero-Inflated Growth Curve

- Model estimates growth curve for structural zeros and for the count simultaneously
- Binary component includes all observations
- Count component includes all observations but is modeling only those zeros that are explainable by a random Poisson process

Zero-Inflated Poisson Regression



Latent Class Growth Analysis Using Zero-Inflated Poisson Model (LCGA Poisson)




LCGA Poisson Models



We use **mixture** models

- A single population may have two subpopulations, i.e., our Implementation variable is a class variable
- Usually assumes class membership explains differences in trajectory, thus a fixed effects model

Latent Profile Analysis for Implementation



<i>Variable</i>	<i>Overall Item Means</i>	<i>Two Classes</i>	
		<i>First Class</i>	<i>Second Class</i>
Stickers for PA	1.74	2.18	1.52
Word of the week	1.14	1.80	.81
You put notes in icu box	1.20	2.30	.50
Teacher read ICU notes about you	1.00	2.24	.36
Teacher read your ICU notes	1.00	2.46	.24
Tokens for meeting goals	1.55	2.16	1.24
PA Assembly activities	1.53	1.96	1.30
Assembly Balloon for PA	.61	.93	.45
Whole school PA	1.21	1.55	1.03
Days/wk taught PA	2.42	2.78	2.24
<i>N</i>	1,550	1,021	529


Applied To Zero-Inflated Growth Models

- We can use a Latent Class Analysis in combination with a zero-inflated growth model to
 - See if there are several classes
 - Classes are distinct from each other
 - Members of a class share a homogeneous growth trajectory

Applied To Zero-Inflated Growth Models

- Sometimes referred to as **Case or Person Centered** rather than Variable Centered
- Subgroups of children rather than of variables
- Has advantages in ease of interpretation of results
- No w/n group variance of intercepts or slope—assumes each subgroup is homogeneous

LCGA Using ZIP Model, No Covariates—One Class Solution



- Serves as a baseline for multi-class solutions
- Add **Mixture** to **Analysis:** section because we are doing a mixture model
- Add **%Overall%** to **Model:** section
- Later, we will add commands so each class can have differences

LCGA ZIP Model Program—Part 1

Title: workshop LCA zip poisson model NO covariates c1.inp
Latent Class Growth Analysis for a count outcome
using a ZIP Model with no covariates and with just one
class

Data:

File is workshop_growth.dat ;

Variable:

Names are

idnum s1flbadc s2flbadc s3flbadc s4flbadc male s1flbadd
s2flbadd s3flbadd s4flbadd s1flbadm s2flbadm s3flbadm
s4flbadm c3 c4 s3techer room ;

Usevariables are s1flbadc s2flbadc s3flbadc s4flbadc ;

Missing are all (-9999) ;

Count are s1flbadc s2flbadc s3flbadc s4flbadc (i) ;

Classes = c(1) ;

! this says there is a single class

LCGA ZIP Model Program—Part 2

Analysis:

```
Type = Mixture missing ;
```

Model:

```
%Overall%
```

```
Alpha Beta | s1flbadc@0 s2flbadc@1 s3flbadc@2 s4flbadc@3 ;
```

```
Alpha_i Beta_i | s1flbadc#1@0 s2flbadc#1@1 s3flbadc#1@2 s4flbadc#1@3;
```

Output:

```
sampstat residual tech1 tech8 ;
```

Plot:

```
Type = Plot3 ;
```

```
! Series = s1flbadc s2flbadc s3flbadc s4flbadc(*) ;
```

```
Series = s1flbadc#1 s2flbadc#1 s3flbadc#1 s4flbadc#1(*) ;
```

```
! We estimate the model twice, once with each series commented out
```

LCGA Using ZIP Model, No Covariates—Two Classes

- We change `classes = c(1)` to `= c(2)` The following table compares 1 to 4 classes
- We will focus on the 2 class solution
 - A normative class (902 children) and a deviant class with just 85
 - The biggest improvement in fit is from 1 to 2 classes

Comparison of 1 to 4 Classes

	<i>1 Class</i>	<i>2 Classes</i>	<i>3 Classes</i>	<i>4 Classes</i>
Free Parameters	8	11	14	17
AIC	8647.18	8232.32	8126.24	8118.08
BIC	8686.33	8286.16	8194.71	8201.29
Sample Adjusted BIC	8660.34	8251.23	8150.30	8147.30
Entropy		.83	.82	.52
Lo, Mendell, Rubin	<i>na</i>	2 v 1 Value =401.45 <i>p</i> < .001	3 v 2 Value = 106.31 <i>p</i> < .01	4 v 3 Value = 13.50 <i>p</i> = .179
<i>N</i> for each class	<i>C1</i> = 987	<i>C1</i> =85 <i>C2</i> =902	<i>C1</i> =55 <i>C2</i> =41 <i>C3</i> =891	<i>C1</i> =36 <i>C2</i> =44 <i>C3</i> =386 <i>C4</i> =521

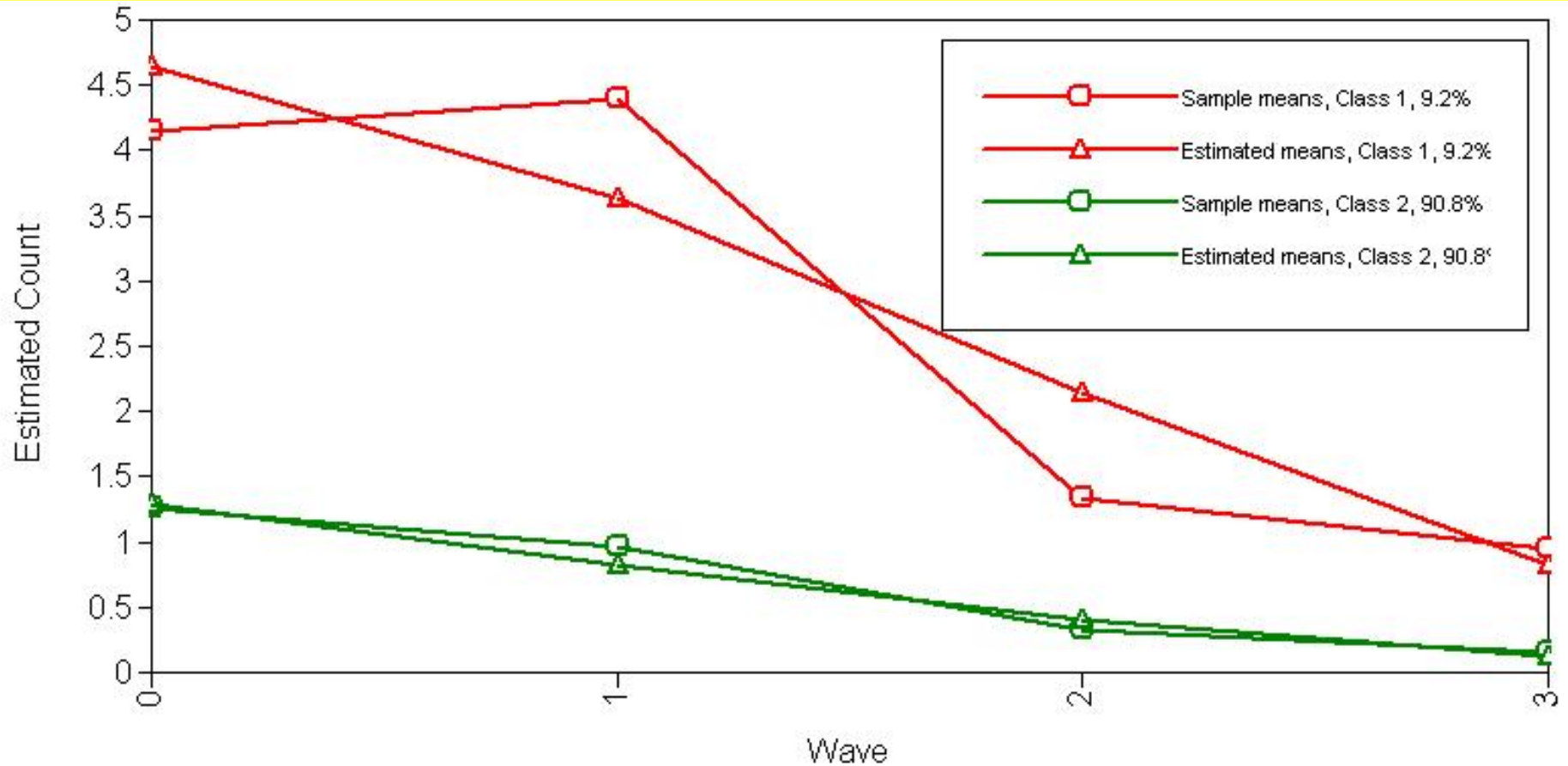
Two Class Solution

```
Entropy                                0.832
CLASSIFICATION OF INDIVIDUALS BASED ON THEIR MOST LIKELY
LATENT CLASS MEMBERSHIP
Class Counts and Proportions
  Latent
  Classes
      1          85          0.08612
      2         902          0.91388
Average Latent Class Probabilities for Most Likely Latent
Class Membership (Row)
by Latent Class (Column)
      1          2
  1   0.893     0.107
  2   0.034     0.966
```

Two Class solution

<i>Parameter Estimate</i>	<i>Class 1 (N = 85)</i>	<i>Class 2 (N=902)</i>
Mean α	1.26***	.022 ^{ns}
Mean β	.050 ^{ns}	-.095 [†]
Mean α_i (fixed)	.000	.000
Mean β_i	.989***	.989***
Threshold	-3.160***	-3.160***
Gender $\rightarrow \alpha$.199 [†]	.199 [†]
Gender $\rightarrow \beta$.057 ^{ns}	.057 ^{ns}
Implement (3) \rightarrow s3count	-.266***	-.266***
Implement (4) \rightarrow s4count	-.385***	-.385***

Two Class solution—Count Part



Freeing Additional Parameters



- We probably are not happy with the default equality constraints
- We can free the effects of the covariates to see if these vary across classes
- We can free the inflation part of the model to see if it varies. Doing this may lead to estimation problems

Program for Freeing Constraint-

Model:

```
%Overall%  
Alpha Beta | s1flbadc@0 s2flbadc@1 s3flbadc@2 s4flbadc@3 ;  
Alpha_i Beta_i | s1flbadc#1@0 s2flbadc#1@1 s3flbadc#1@2  
s4flbadc#1@3 ;  
Alpha on male ;  
Beta on male ;  
S3flbadc on c3 ;  
S4flbadc on c4 ;  
%c#2%  
[s1flbadc#1 s2flbadc#1 s3flbadc#1 s4flbadc#1](1) ;  
[Beta_i] ;
```

Explaining New Commands



- After the `%Overall%` subsection we add a new subsection
- `%c#2%`
 - `[s1flbadc#1 s2flbadc#1 s3flbadc#1 s4flbadc#1](1) ;`
 - `[Beta_i] ;`
- **C#2** means we are removing constraints on class 2 parameter estimates
- We allow a different set of thresholds. The (1) at end requires that all of these are equal in class 2
- The `[Beta_i]` indicates that the slope for the inflation portion can be different for class 2

Key Results

<i>Parameter Estimate</i>	<i>Class 1 (N = 101)</i>	<i>Class 2 (N=886)</i>
Mean α	1.271***	-.028ns
Mean β	.030 ^{ns}	-.159*
Mean α_i (fixed)	.000	.000
Mean β_i	.552***	4.853***
Threshold	-1.505***	-15.000 ^{fixed}
Gender α	.192 [†]	.192 [†]
Gender β	.065 ^{ns}	.065 ^{ns}
Implement (3) s3count	-.292***	-.292***
Implement (4) s4count	-.335**	-.335**

Interpretation



- This solution has the deviant group start with a higher initial count (α) and showing no significant improvement (β)
- The Normative group starts with a lower likelihood of being always zero (threshold) but this increases dramatically (β_i) across waves. The deviant group also increases, but not nearly as rapidly

Next Steps



- If you find distinct classes of participants who have different growth trajectories you can save the class of each participant. This is shown in the detailed document
- You can then compare the classes on whatever variable you think might be important in explaining the differentiation, e.g., parental support for program
- This will generate a new set of important covariates for subsequent research

Next Steps



- An introduction to growth curves and a detailed presentation of the ideas we've discussed is available at www.oregonstate.edu/~acock/growth

Summary—Three Models Available from Mplus



Traditional growth modeling where

- There is a common expectation for the trajectory for a sample
- Parameter estimates will have variances across individuals around the common expected trajectory (random effects)
- Covariates may explain some of this variance

Summary—Three Models Available from Mplus



- Latent Class Growth Models where
 - We expect distinct classes that have different trajectories
 - Class membership explains all of the variance in the parameters. Classes are homogeneous with respect to their growth curves (fixed effects)

Summary—Three Models Available from Mplus



Mixture Models extending Latent Class Growth Models where

- We expect distinguishable classes that each have a different common trajectory
- Residual variance not explained by class membership are allowed (random effects)
- Covariates may explain some of this residual variance

Summary



- Mplus offers many features that are especially useful for longitudinal studies of individuals and families
- Many outcomes for family members are best studied using longitudinal data to identify growth trajectories
- Studies of growth trajectories can utilize time invariant, time variant, and distal outcomes

Summary



- Some outcomes for family members are successes or failures and the binary growth curves are useful for modeling these processes
- Some outcomes for family members are counts of how often some behavior or outcome occurs
- Many outcome involve both the binary and the count components