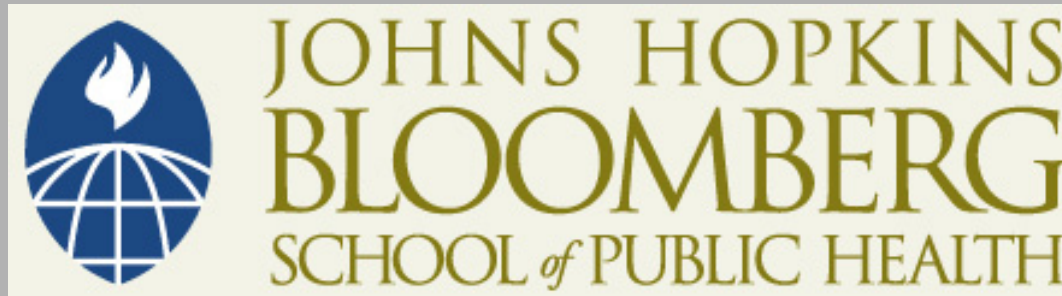


This work is licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike License](https://creativecommons.org/licenses/by-nc-sa/4.0/). Your use of this material constitutes acceptance of that license and the conditions of use of materials on this site.



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.

# Advanced Structural Equations Models II

Statistics for Psychosocial Research II:  
Structural Models

Qian-Li Xue

# Outline

- Multilevel Models
  - Special case: Latent Growth Curve Models

# Multilevel Models: Definition

- “Analysis models that contain variables measured at different levels of the hierarchy” (Kreft & De Leeuw, 2006)
- Examples:
  - SAT scores for students grouped by classes within schools (i.e. three-level hierarchy: school->class->student)
  - Evaluation of group therapy research in clinical psychology (i.e. two-level hierarchy: group->individual)
  - Growth curve analysis: repeated measurements on individuals (i.e. two-level hierarchy: individual->measurement at time t)

# Added Complexity Due to Multilevel Data Structure

- Recall: a fundamental assumption of a regression model is that the residuals are independent
- No longer true with multilevel data
- For example: if we believe that the education given by schools may influence outcome attainment, then students within a particular school will tend to be more similar than students from different schools
- The sharing of the same “context” contributes to within-level dependency among measurements!

# Key Concepts in Multilevel Modeling

- Contextual models

- “Any linear regression model that contains lower-level variables and higher-level characteristics that are aggregated or globally measured”

- For example:

Academic achievement of students nested within school classes

- ❖ Student-level (i.e. lower-level) characteristics (e.g. SES)
- ❖ School-level (i.e. higher-level) characteristics (e.g. student-to-teacher ratio)
- ❖ A model that contains both SES and student-to-teacher ratio as predictors of students’ academic achievement is referred to as a contextual model

# Key Concepts in Multilevel Modeling

- Intra-Class Correlation (ICC;  $\rho$ )
  - “A measure of the dependence of individuals”
  - “A measure of group homogeneity”
  - For a two-level hierarchical structure,  $\rho$  is defined as proportion of the variance in the outcome that is between the 2<sup>nd</sup>-level units (i.e. clusters)

$$\rho = \frac{\sigma^2_{between}}{\sigma^2_{between} + \sigma^2_{within}}$$

# Impact of ICC on Statistical Inference

- Inflate Type I error (i.e. the alpha level)

N	$\rho$		
	0.01	0.05	0.20
10	0.06	0.11	0.28
25	0.08	0.19	0.46
50	0.11	0.30	0.59
100	0.17	0.43	0.70

# Multilevel Model: Example

- Identify predictors of SAT score ( $Y_{ij}$ ) of students ( $i=1, \dots, n_j$ ) nested within schools ( $j=1, \dots, J$ )
- Student-level characteristics
  - SES of parents ( $X_1$ )
  - Parents' education level ( $X_2$ )
- School-level characteristics
  - Class size, measured by the student-teacher ratio ( $Z_1$ )
  - Education sector (private vs. public) ( $Z_2$ )

Level 1 (Student-Level):

$$Y_{ij} = \eta_{0j} + \eta_{1j}X_{1ij} + \eta_{2j}X_{2ij} + \varepsilon_{ij}$$

Random  
Coefficient

Level 2 (School-Level):

$$\eta_{0j} = \gamma_{00} + \gamma_{01}Z_{1j} + \gamma_{02}Z_{2j} + \zeta_{0j}$$

$$\eta_{1j} = \gamma_{10} + \gamma_{11}Z_{1j} + \gamma_{12}Z_{2j} + \zeta_{1j}$$

$$\eta_{2j} = \gamma_{20} + \gamma_{21}Z_{1j} + \gamma_{22}Z_{2j} + \zeta_{2j}$$

Fixed  
Coefficient

# Multilevel Model: Example

## Multilevel Expression

Level 1 (Student-Level):

$$Y_{ij} = \eta_{0j} + \eta_{1j}X_{1ij} + \eta_{2j}X_{2ij} + \varepsilon_{ij}$$

Level 2 (School-Level):

$$\eta_{0j} = \gamma_{00} + \gamma_{01}Z_{1j} + \gamma_{02}Z_{2j} + \zeta_{0j}$$

$$\eta_{1j} = \gamma_{10} + \gamma_{11}Z_{1j} + \gamma_{12}Z_{2j} + \zeta_{1j}$$

$$\eta_{2j} = \gamma_{20} + \gamma_{21}Z_{1j} + \gamma_{22}Z_{2j} + \zeta_{2j}$$

## Reduced-Form Expression

$$Y_{ij} = (\gamma_{00} + \gamma_{10}X_{1ij} + \gamma_{20}X_{2ij} + \gamma_{01}Z_{1j} + \gamma_{02}Z_{2j})$$

$$+ (\gamma_{11}Z_{1j}X_{1ij} + \gamma_{12}Z_{2j}X_{1ij} + \gamma_{21}Z_{1j}X_{2ij} + \gamma_{22}Z_{2j}X_{2ij})$$

$$+ (\zeta_{0j} + \zeta_{1j}X_{1ij} + \zeta_{2j}X_{2ij} + \varepsilon_{ij})$$

Cross-Level Interactions

Student-level deviation from school-level averages

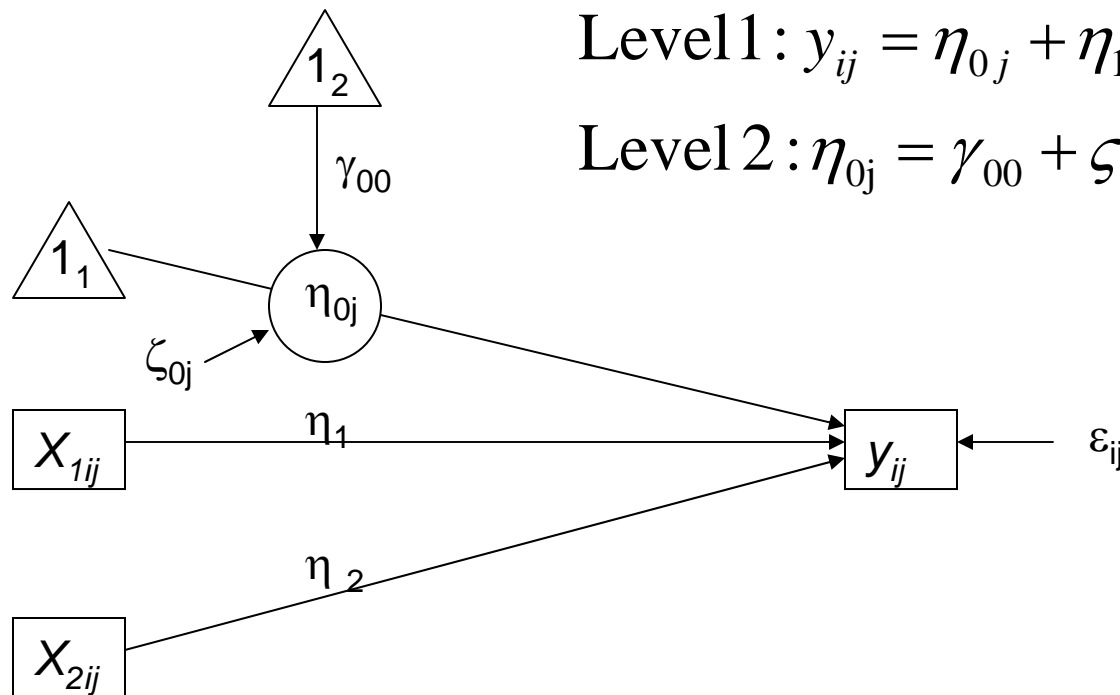
# Path Diagram for Multilevel Models

(Curran & Bauer, 2007)

- Boxes – observed variables (predictors or factor indicators)
- Triangles – intercept terms with subscripts to denote specific level
- Circles – random coefficients
- Straight Single-Headed Arrows – regression parameters
  - Triangle to cycle (or box) – intercept
  - Box to box – regression of one observed variable on another
  - Box to circle – regression of a lower level random coefficient on a higher level observed variable
  - Nothing to box or circle – residual or disturbance
- Multiheaded arrows – covariance (e.g. covariance between two random coefficients)

# Path Diagram for Multilevel Models: Example (Curran & Bauer, 2007)

- Two-predictor regression with a random intercept and fixed slopes

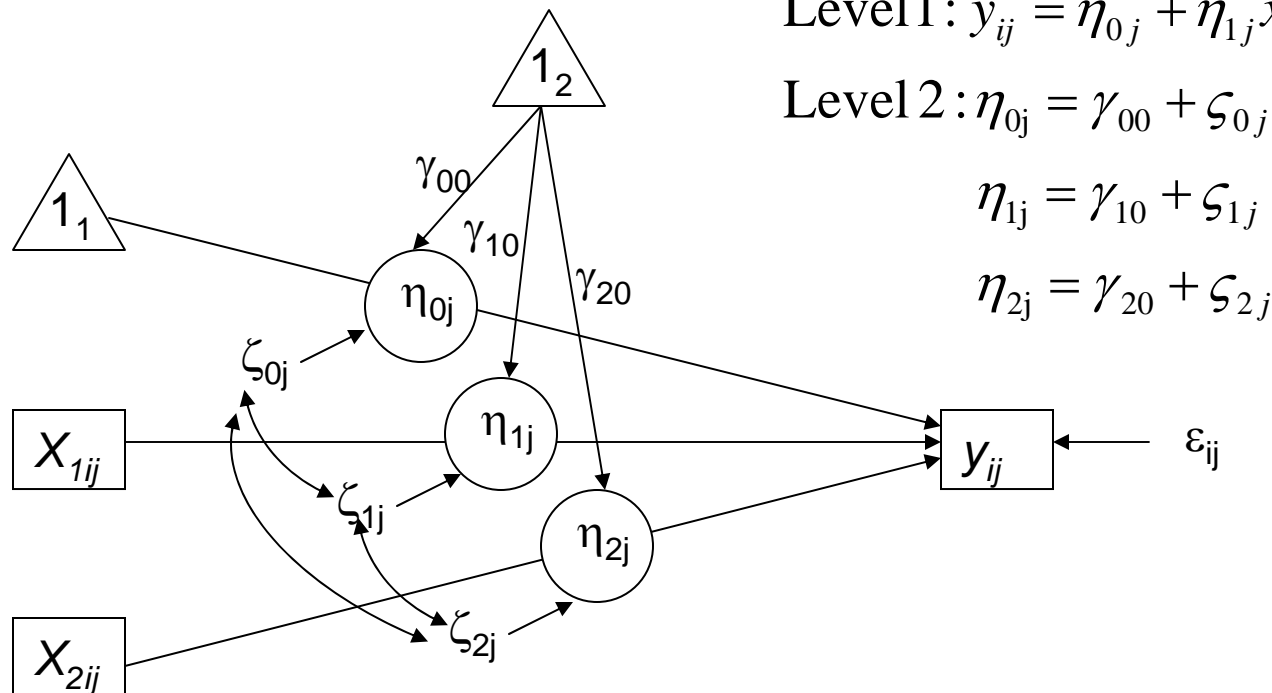


$$\text{Level 1: } y_{ij} = \eta_{0j} + \eta_1 x_{1ij} + \eta_2 x_{2ij} + \varepsilon_{ij}$$

$$\text{Level 2: } \eta_{0j} = \gamma_{00} + \zeta_{0j}$$

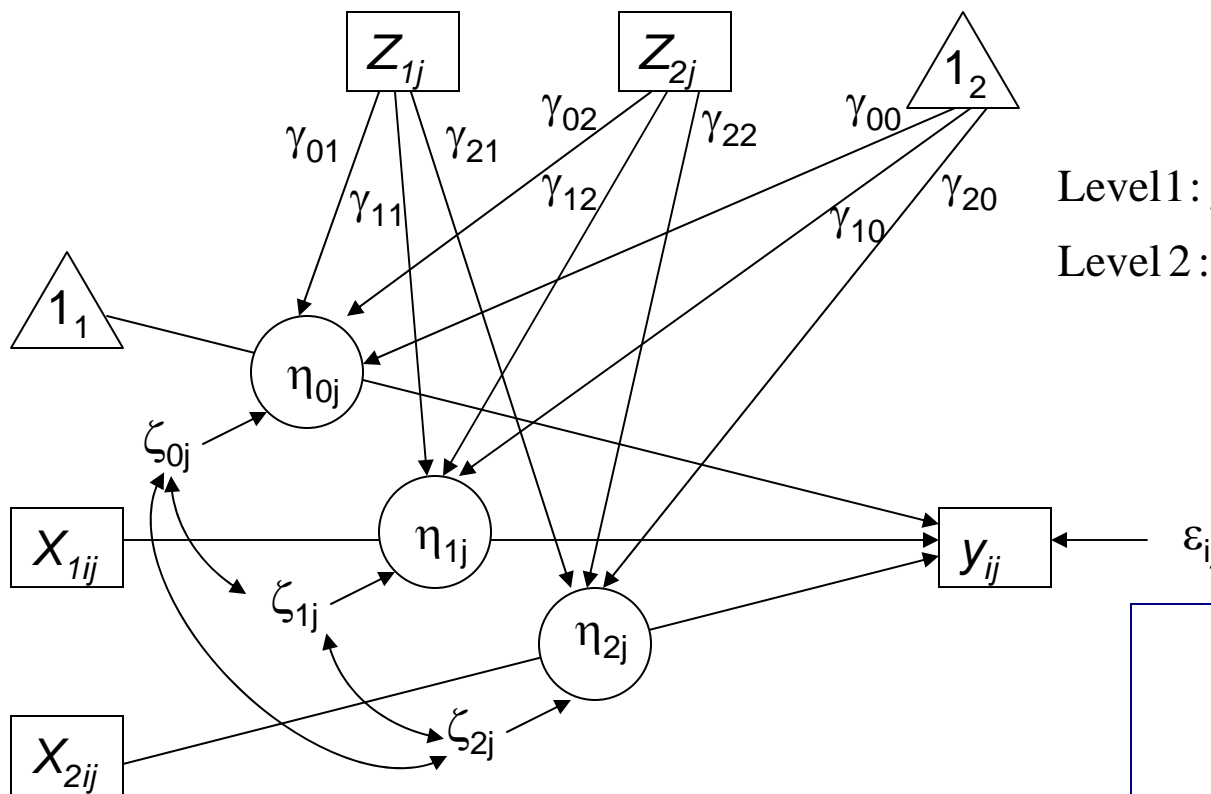
# Path Diagram for Multilevel Models: Example (Curran & Bauer, 2007)

- Two-predictor regression with a random intercept and random slopes



# Path Diagram for Multilevel Models: the SAT Example (continued)

- Multilevel Model with both Level-1 and Level-2 predictors



$$\text{Level 1: } y_{ij} = \eta_{0j} + \eta_{1j}x_{1ij} + \eta_{2j}x_{2ij} + \varepsilon_{ij}$$

$$\text{Level 2: } \eta_{0j} = \gamma_{00} + \gamma_{01}z_{1j} + \gamma_{02}z_{2j} + \zeta_{0j}$$

$$\eta_{1j} = \gamma_{10} + \gamma_{11}z_{1j} + \gamma_{12}z_{2j} + \zeta_{1j}$$

$$\eta_{2j} = \gamma_{20} + \gamma_{21}z_{1j} + \gamma_{22}z_{2j} + \zeta_{2j}$$

Student-Level  
 $X_1$ : SES of parents  
 $X_2$ : Parents' education level  
 School-Level  
 $Z_1$ : Student-teacher ratio ( $Z_1$ )

# Model Fitting: MPLUS

TITLE: This is an example of a two-level regression for a continuous dependent variable (SAT) on both student-level and school-level characteristics

DATA: FILE IS c:\teaching\140.658.2007\SAT.dat;

VARIABLE: NAMES ARE sat x1 x2 z1 z2 clus;  
WITHIN = x1 x2;  
BETWEEN = z1 z2;  
CLUSTER = clus;  
CENTERING = GRANDMEAN (x1 x2 z1);

ANALYSIS: TYPE = TWOLEVEL RANDOM;

MODEL:

```
%WITHIN%  
s1 | sat ON x1;  
s2 | sat ON x2;  
%BETWEEN%  
sat s1 s2 ON z1 z2;  
s1 WITH s2 SAT;  
s2 WITH SAT;
```

Estimate random slopes for regression of SAT on X1 and SAT on X2

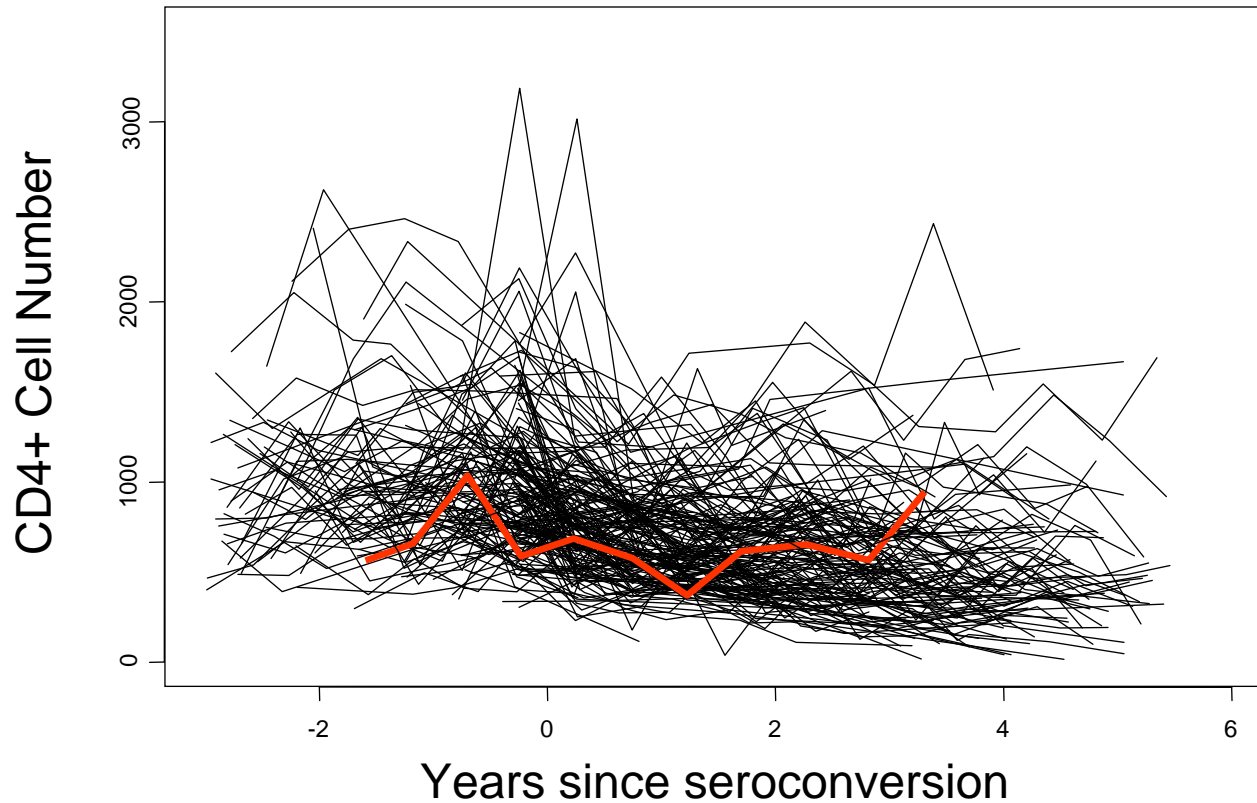
Cross-level Interactions: Regression of random intercept and slopes on Z1 and Z2

# Special Case of Multilevel Models: Growth Curve Models

# Growth Curve Models (GCM)

- Focus on 'growth' trajectories
  - Outcome of interest can be anything that varies systematically over time
  - Estimation involves the estimation of either population average trajectory or
  - Individual specific trajectory, with a slope and intercept (and possibly a quadratic term) for every individual in the analysis pool (i.e. Latent GC models)

# Example 1. CD4+ cells decrease in number in time from infection with HIV

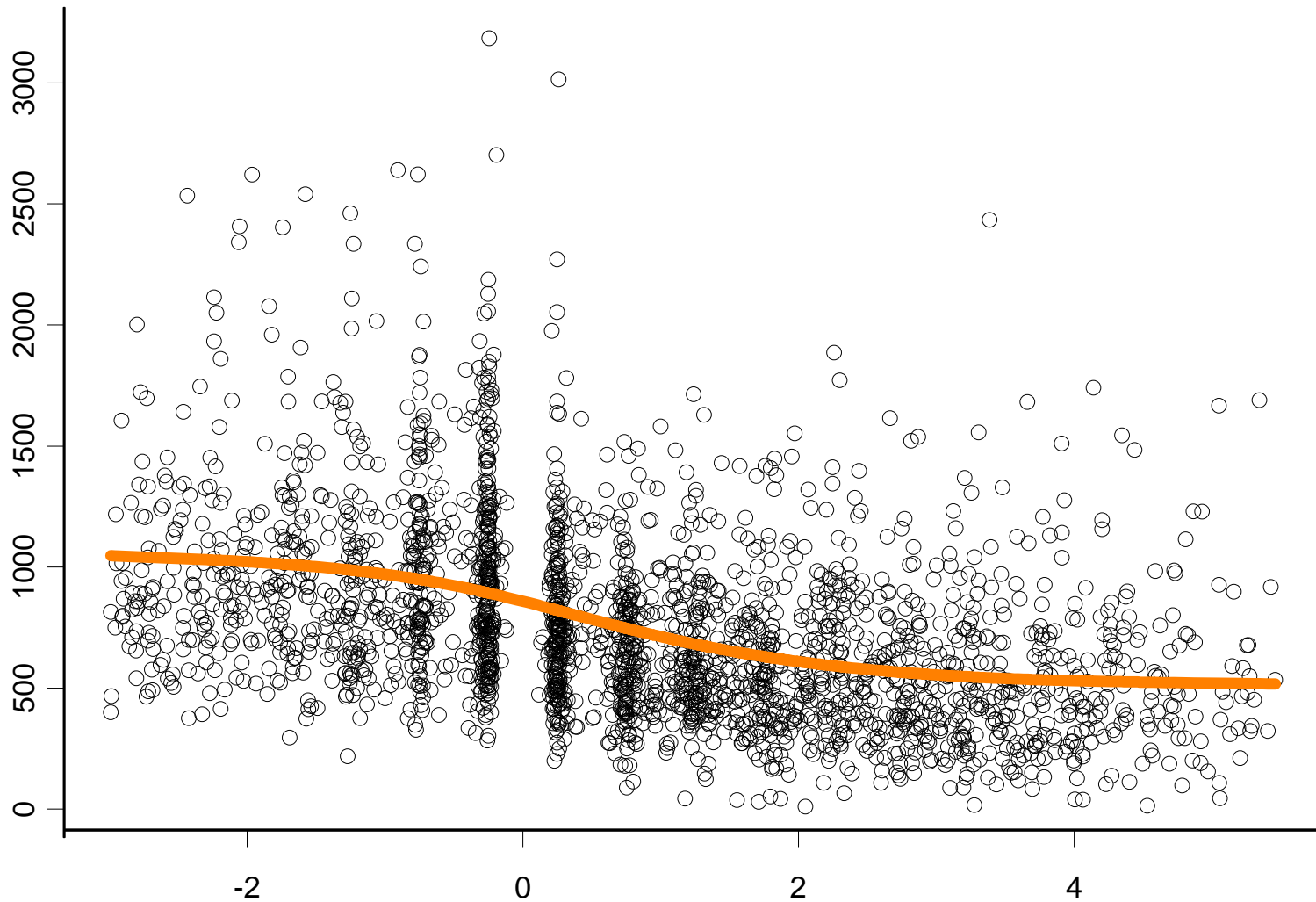


Diggle PJ, Heagerty P, Liang KY and Zeger SL (2001)

# Growth Curve Models: Marginal Models

- Marginal (i.e. population average) models
  - Well-defined homogeneous target population
  - Consistency of a pattern across subjects
  - Model “marginal mean”  $E(Y_{ij}|X_{ij})$
  - Coefficients describe/compare sub-populations
  - Same interpretation as in ordinary regression
  - Growth curve models when  $X$  is time

# Population Mean Trajectory



# Multilevel Formulation of Latent Growth Curve Models

- GC models can be formulated as multilevel models
- Repeated measurements nested within individuals
- Let  $Y_{ij}$  – outcome for subject  $j$  ( $=1, \dots, N$ ) at time  $i$  ( $=1, \dots, n_j$ )

Level 1 (within-subject):

$$Y_{ij} = \eta_{0j} + \eta_{1j}T_{ij} + \varepsilon_{ij}$$

Level 2 (between-subject):

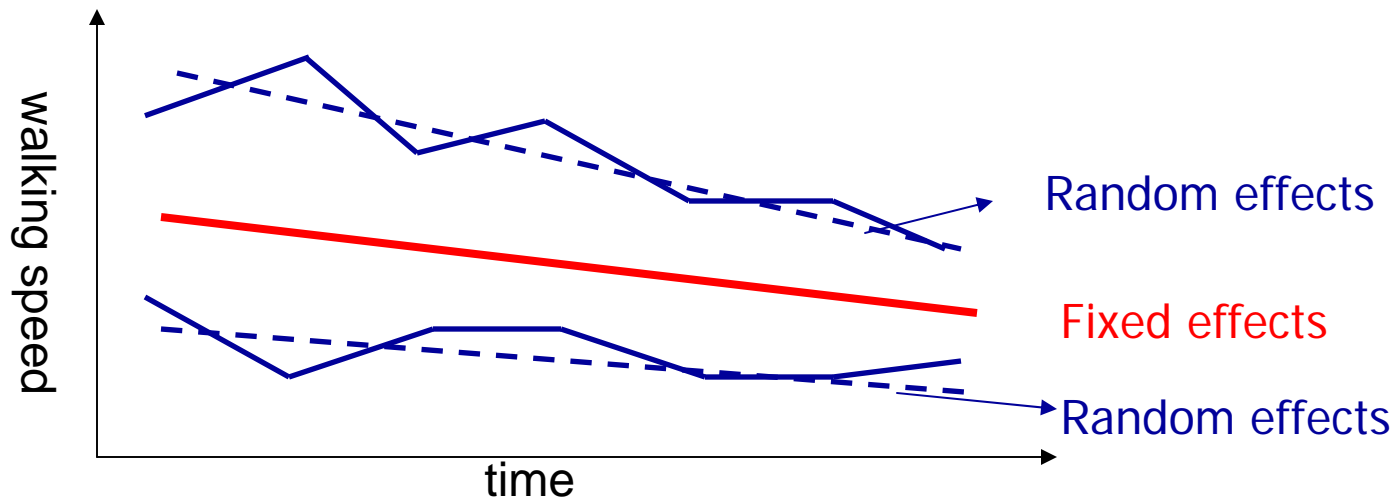
$$\eta_{0j} = \gamma_0 + \zeta_{0j}$$

$$\eta_{1j} = \gamma_1 + \zeta_{1j}$$

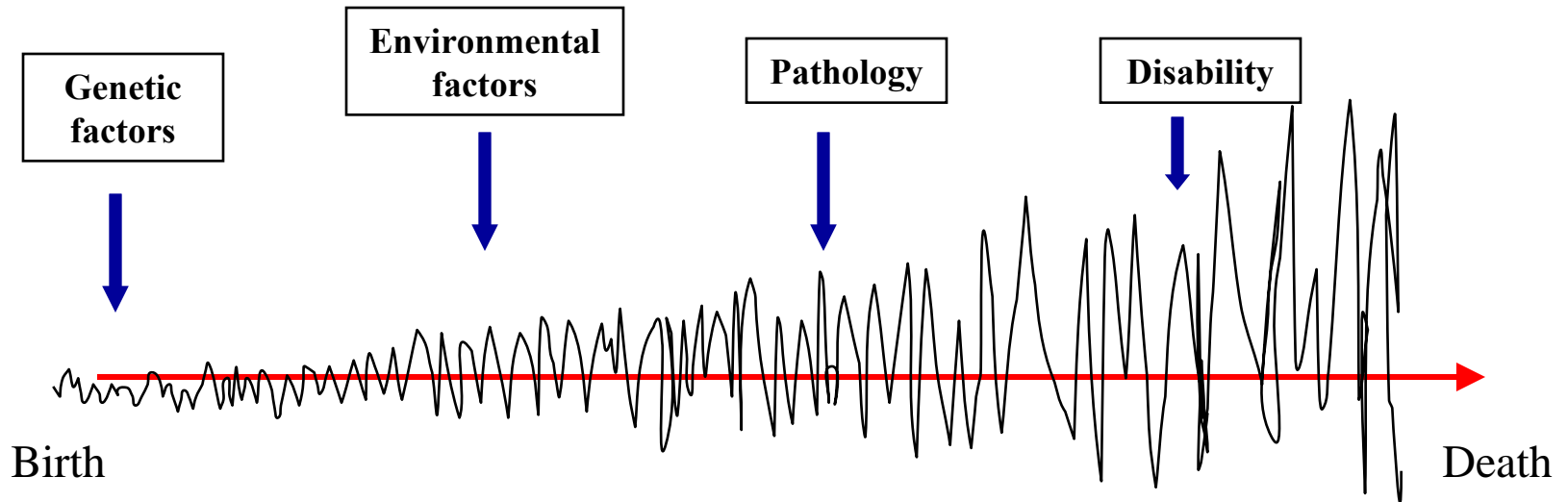
# Latent Growth Curve Models

- Latent GC model (also termed “Random Effects” Model)
- Reduced-form expression:

$$Y_{ij} = \underbrace{\gamma_0 + \gamma_1 T_{ij}}_{\text{Fixed Effects}} + \underbrace{\zeta_{0j} + \zeta_{1j} T_{ij}}_{\text{Random Effects}} + \varepsilon_{ij}$$



# Sources of Heterogeneity



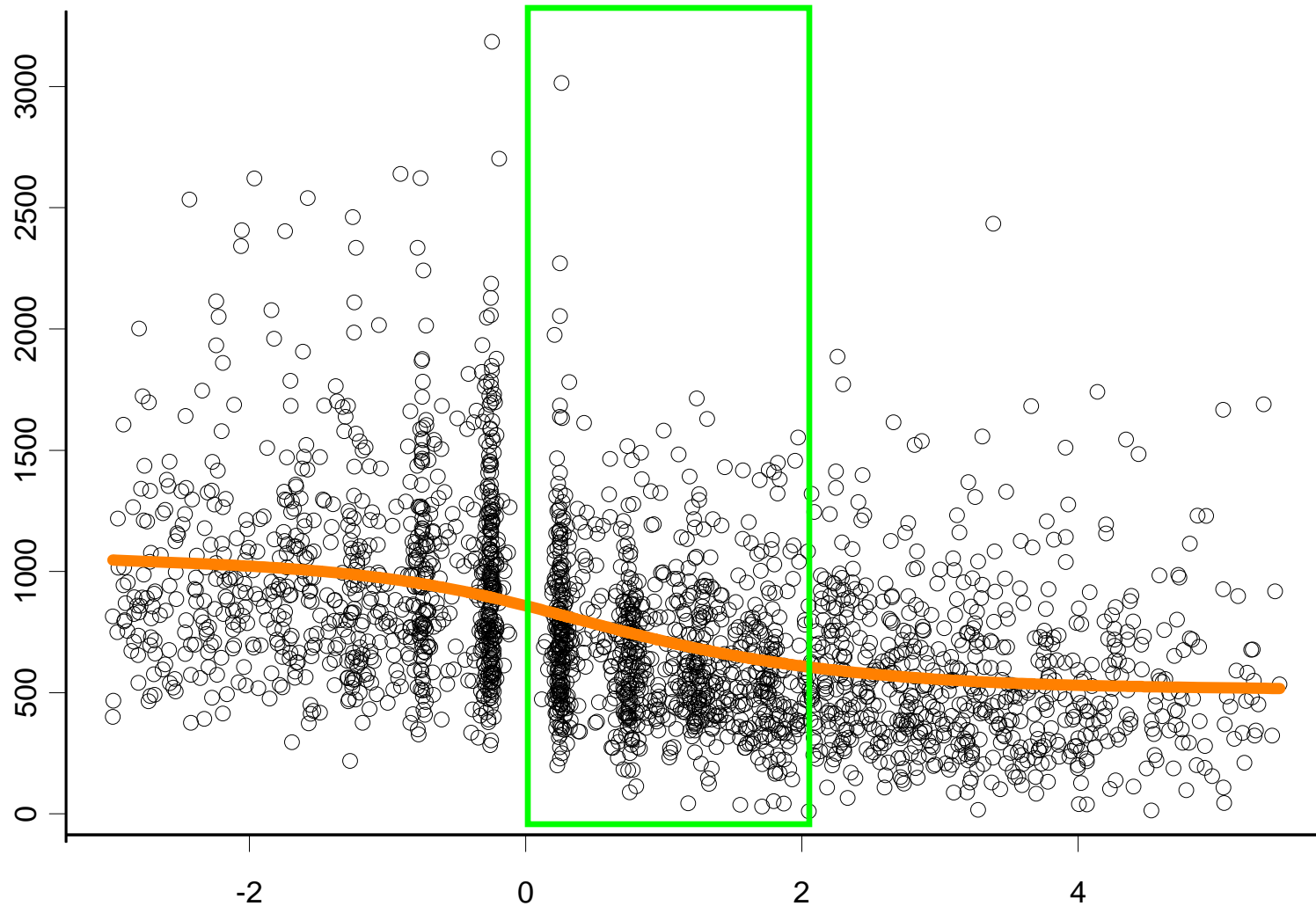
# Analysis of Growth Curve Data: Model Comparison

	<b>Marginal via GEE</b>	<b>Random Effects</b>
<b>Between-Subject Heterogeneity</b>	—	+
<b>Model assumptions</b>	+	—
<b>Handling Missing Data</b>	—	+
<b>Irregular Time Intervals</b>	—	+
<b>Cluster Size</b>	+	—
<b>Computation</b>	+	—

# Example: CD4+ Level

- HIV attacks CD4+ cell which regulates the body's immunoresponse to infectious agent
- 2376 values of CD4+ cell number plotted against time since seroconversion for 369 infected men enrolled in the MACS
- Specific aims:
  - Aim 1: Estimate population-average time course of CD4+ cell depletion
  - Aim 2: Identify factors that predict subject-specific baseline and rate of change in CD4+ cell count over time

# Population Mean Trajectory



# Data Format For GCM Fitting

- Data in “wide” format (one record per person)

Individually-varying times  
of observation

	id	time1	time2	time3	time4	cd41	cd42	cd43	cd44	packs1	packs2	packs3	packs4	drugs1	drugs2	drugs3	drugs4
13	10029	0.25	0.77	1.41	1.81	6.71	6.76	6.56	6.63	0	0	0	0	1	1	1	1
23	10048	0.31	0.81	1.10	1.59	6.19	6.41	6.78	5.86	0	0	0	0	1	1	1	1
39	10088	0.28	0.82	1.18	2.22	6.75	6.56	6.25	6.86	0	0	0	0	1	0	0	0
54	10131	0.25	1.23	1.76	2.26	7.73	6.96	7.26	7.54	3	3	3	3	1	1	1	1
66	10132	0.24	0.73	1.26	1.83	6.53	5.79	6.17	5.41	3	3	2	3	1	1	1	1

	sex1	sex2	sex3	sex4	cesd1	cdsd2	cdsd3	cdsd4	age
13	5	5	5	5	7	15	21	25	2.64
23	5	5	5	0	-7	-5	7	-7	17.99
39	1	-2	0	-1	9	1	6	1	1.64
54	-2	-2	-3	-3	-2	1	1	4	5.93
66	5	5	0	-4	-7	-7	-7	-6	9.33

# MPLUS Input for Aim 1: GCM Fitting

TITLE: Growth Curve Modeling of CD4 Cell Depletion Trajectory

DATA: FILE IS c:/teaching/140.658.2007/cd4w.dat;

VARIABLE:

NAMES ARE id time1-time4 cd41-cd44 pack1-pack4  
drug1-drug4 sex1-sex4 cesd1 cesd4 age;

USEVARIABLES ARE cd41-cd44 time1-time4;

TSCORES=time1-time4;

ANALYSIS: TYPE=RANDOM;

MODEL:

i s| cd41-cd44 AT time1-time4;

cd41-cd44 (1);

OUTPUT: TECH1;

Allow individually-varying  
times of observation

Allow random intercept  
and random slope for  
regression of cd4 on time

# Data Format for Multilevel Fitting

- Data in “long” format (multiple records per person)

	Time	CD4	Age	Packs	Drugs	Sex	Cesd	ID	n	seqnum	
13	0.251882	824	2.64	0	1	5	7	10029	5	1	Same Individual ↗
14	0.769336	866	2.64	0	1	5	15	10029	5	2	
15	1.412731	704	2.64	0	1	5	21	10029	5	3	
16	1.806982	757	2.64	0	1	5	25	10029	5	4	
23	0.306639	486	17.99	0	1	5	-7	10048	7	1	
24	0.813142	605	17.99	0	1	5	-5	10048	7	2	
25	1.095140	880	17.99	0	1	5	7	10048	7	3	
26	1.593429	352	17.99	0	1	0	-7	10048	7	4	
39	0.279261	858	1.64	0	1	1	9	10088	10	1	
40	0.815880	709	1.64	0	0	-2	1	10088	10	2	

# MPLUS Input for Aim 1: Multilevel Model Fitting

TITLE: Growth Curve Modeling of CD4 Cell Depletion Trajectory

DATA: FILE IS c:/teaching/140.658.2007/cd4.dat;

VARIABLE:

NAMES ARE time cd4 age packs drugs sex cesd id n seqnum;

USEVARIABLES ARE time cd4;

WITHIN = time;

CLUSTER = id;

ANALYSIS: TYPE = TWOLEVEL RANDOM;

MODEL:

%WITHIN%

s | cd4 ON time;

%BETWEEN%

s WITH cd4;

OUTPUT: TECH1;

Estimate random slope for regression of cd4 on time

Estimate covariance between random intercept and random slope

# MPLUS Output for Aim 1

## MODEL RESULTS

	Estimates	S.E.	Est./S.E.	
Within Level				
Residual Variances				
CD4	0.074	0.010	7.412	
Between Level				
S	WITH			
CD4	-0.039	0.028	-1.422	Covariance between random coefficients
Means				
CD4	6.655	0.033	203.570	Population-average logCD4+ cell count at baseline
S	-0.193	0.028	-6.984	Population-average rate of CD4+ cell depletion over time
Variances				
CD4	0.115	0.026	4.489	Variances of random coefficients
S	0.085	0.043	1.974	

# GCM with Covariates

- Goal: Take into account all the covariates that might have an effect on the response
  - Does AZT (time-varying) have an effect on CD4+ cell count?
  - Does age at seroconversion (time-invariant) affect baseline and rate of CD4+ cell depletion?

# MPLUS Input for Aim 2: GCM Fitting

TITLE: Growth Curve Modeling of CD4 Cell Depletion Trajectory

DATA: FILE IS c:/teaching/140.658.2007/cd4w.dat;

VARIABLE:

NAMES ARE id time1-time4 cd41-cd44 pack1-pack4  
drug1-drug4 sex1-sex4 cesd1 cesd4 age;

USEVARIABLES ARE cd41-cd44 time1-time4  
drug1-drug4 age;

TSCORES=time1-time4;

ANALYSIS: TYPE=RANDOM;

MODEL:

i s| cd41-cd44 AT time1-time4;

i s ON age;

cd41 ON drug1 (1);

cd42 ON drug2 (1);

cd43 ON drug3 (1);

cd44 ON drug4 (1);

cd41-cd44 (2);

OUTPUT: TECH1;

Allow time-invariant age effect on subject specific baseline and rate of CD4+ cell depletion over time

Allow time-varying covariate (AZT) on population-average CD4+ cell count

# MPLUS Input for Aim 2: Multilevel Model Fitting

TITLE: Growth Curve Modeling of CD4 Cell Depletion Trajectory

DATA: FILE IS c:/teaching/140.658.2007/cd4.dat;

VARIABLE:

NAMES ARE time cd4 age packs drugs sex cesd id n seqnum;

USEVARIABLES ARE time cd4 drugs age;

WITHIN = time drugs;

BETWEEN=age;

CLUSTER = id;

ANALYSIS: TYPE = TWOLEVEL RANDOM;

MODEL:

```
%WITHIN%  
s | cd4 ON time;  
cd4 ON drugs;  
%BETWEEN%  
cd4 s ON age;  
s WITH cd4;
```

Estimate random slope for  
regression of cd4 on time

Estimate covariance  
between random intercept  
and random slope

OUTPUT: TECH1;

# MPLUS Output for Aim 2

## MODEL RESULTS

		Estimates	S.E.	Est./S.E.	
Within Level					
CD4	ON				AZT was associated with significant higher mean logCD4+ cell count
DRUGS		0.120	0.039	3.066	
Residual Variances					
CD4		0.073	0.010	7.356	
Between Level					
S	ON				No sig. effect of age on subject-specific baseline or rate of CD4+ cell depletion
AGE		0.000	0.005	-0.038	
CD4	ON				
AGE		0.003	0.005	0.664	
S	WITH				
CD4		-0.039	0.028	-1.417	
Intercepts					
CD4		6.544	0.049	132.611	Population-average rate of CD4+ cell depletion over time
S		-0.182	0.035	-5.264	
Residual Variances					
CD4		0.110	0.025	4.425	Population-average logCD4+ cell count at baseline
S		0.087	0.044	1.986	