CollegeBoard

Almost 31 Flavors of Multi-level Modeling in SAS

Presentation to the New York Area
SAS Users' Group

Brian F. Patterson
Assistant Research Scientist

December 9, 2008

connect to college success™
www.collegeboard.com

Overview

1. Introduction (10 min.)
2. Review of Statistics (10 min.)
3. SAS for Multi-level Models (20 min.)
4. Research Using MMs (5 min.)
5. Question / Answer Period (15 min.)

1.1 Introduction

- Why multi-level models?
  - Avoid aggregation bias and inflated std. err.
  - Disentangle group- and individual-level effects
  - Strengthen generalizability to group-level pop.
  - Model changes over time

1.2 Definition of Terms

- SAT
  - College admissions test consisting of Critical Reading, Math & Writing sections on 200-800 scale
- Advanced Placement
  - High school program created to give students the opportunity to take college-level courses
- Visit: http://www.collegeboard.com/

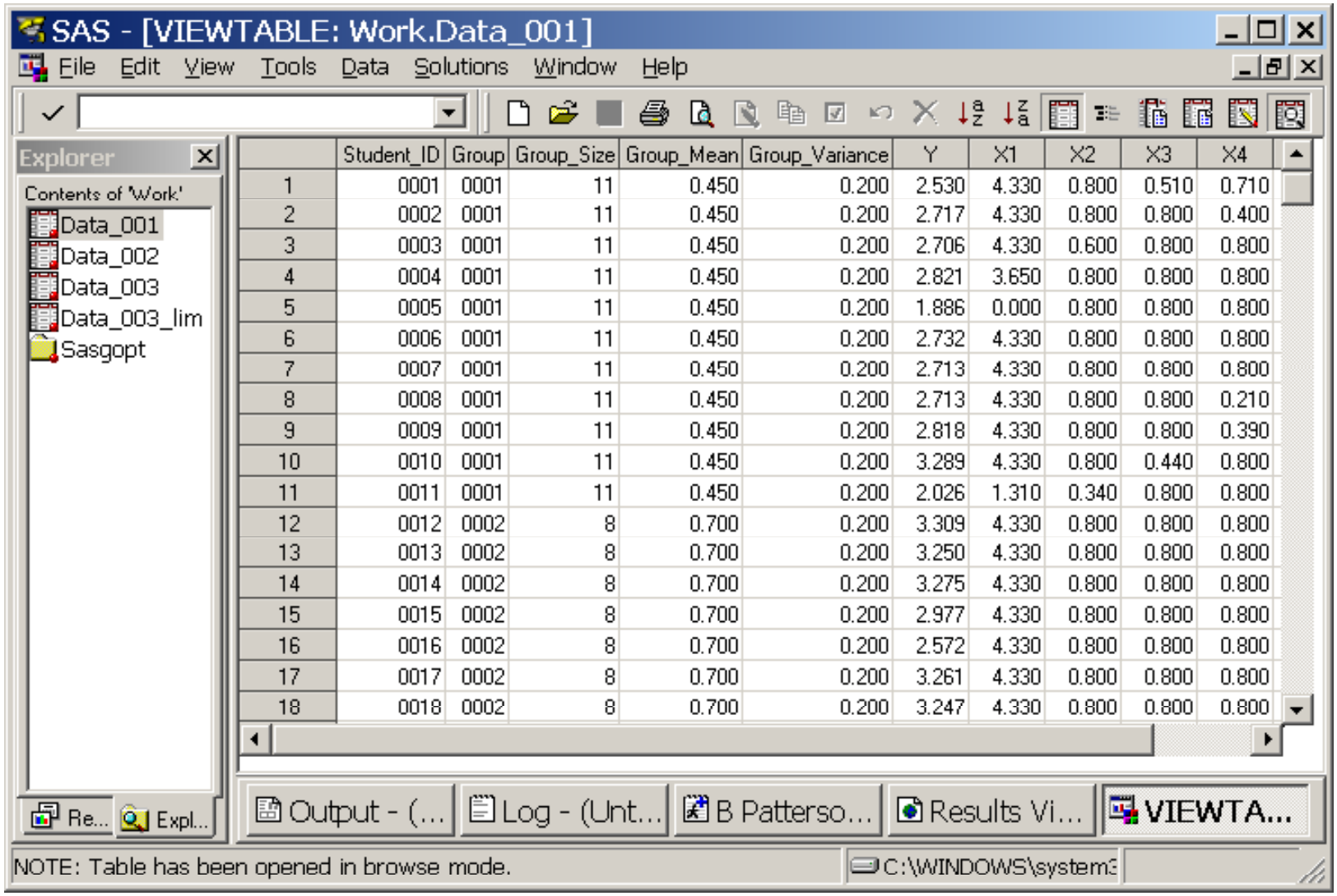
1.3 Multi-level Modeling Opportunities

- Model first-year grade point average (FYGPA) at 110 colleges and universities
- Model prob. of participation in Advanced Placement Program® at many high schools
- Model FYGPA with high-school- and college-variability with cross-classified data

1.4 SAS Requirements for Multi-level Modeling

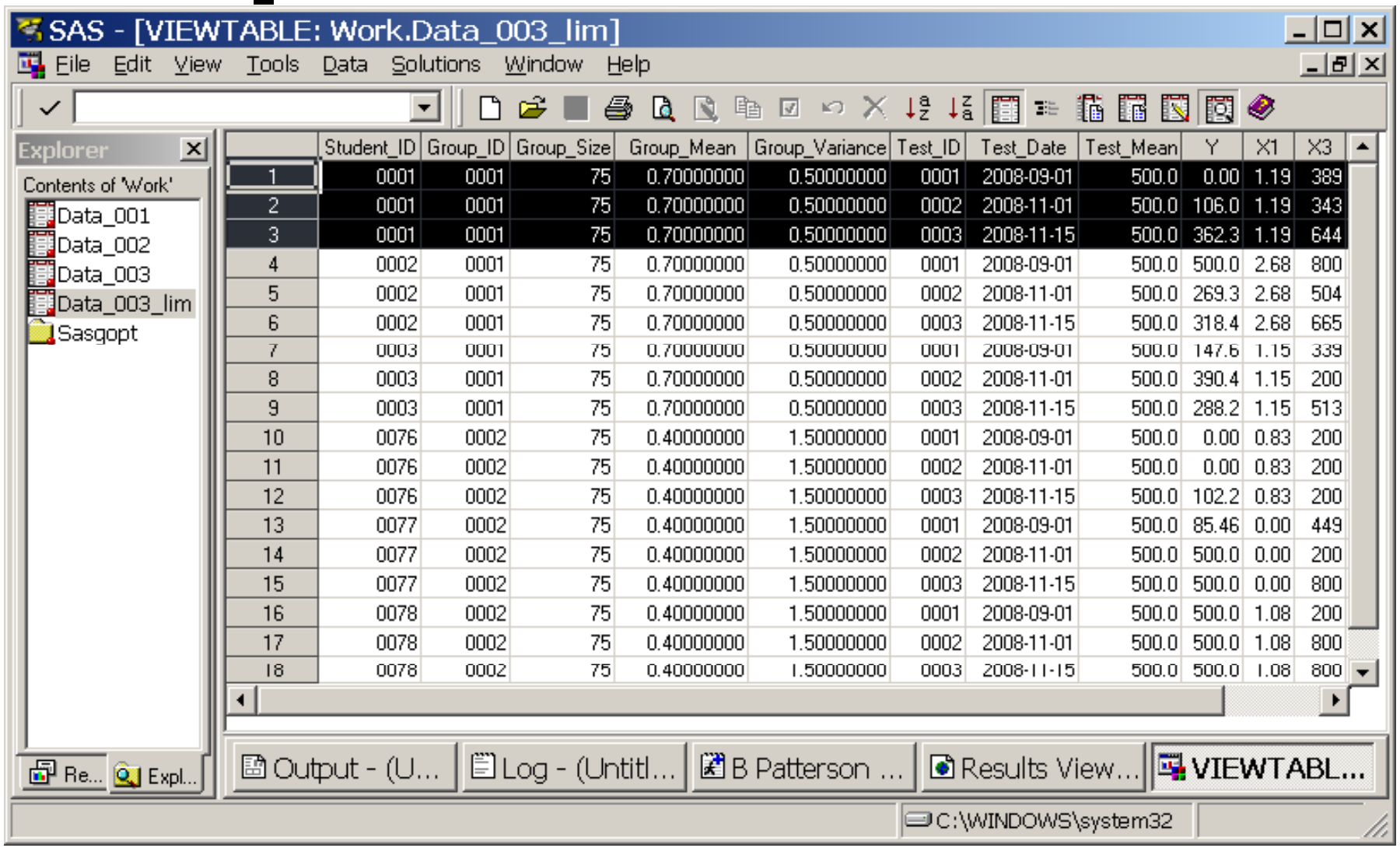
- Many observations
  - Large sample size within each group
  - Sufficient number of groups
- Computing power
  - Procedures (and optional statements within some procedures) are memory-intensive
- Data structure

1.5 Nested Data Structure in SAS



	Student_ID	Group	Group_Size	Group_Mean	Group_Variance	Y	X1	X2	X3	X4
1	0001	0001	11	0.450	0.200	2.530	4.330	0.800	0.510	0.710
2	0002	0001	11	0.450	0.200	2.717	4.330	0.800	0.800	0.400
3	0003	0001	11	0.450	0.200	2.706	4.330	0.600	0.800	0.800
4	0004	0001	11	0.450	0.200	2.821	3.650	0.800	0.800	0.800
5	0005	0001	11	0.450	0.200	1.886	0.000	0.800	0.800	0.800
6	0006	0001	11	0.450	0.200	2.732	4.330	0.800	0.800	0.800
7	0007	0001	11	0.450	0.200	2.713	4.330	0.800	0.800	0.800
8	0008	0001	11	0.450	0.200	2.713	4.330	0.800	0.800	0.210
9	0009	0001	11	0.450	0.200	2.818	4.330	0.800	0.800	0.390
10	0010	0001	11	0.450	0.200	3.289	4.330	0.800	0.440	0.800
11	0011	0001	11	0.450	0.200	2.026	1.310	0.340	0.800	0.800
12	0012	0002	8	0.700	0.200	3.309	4.330	0.800	0.800	0.800
13	0013	0002	8	0.700	0.200	3.250	4.330	0.800	0.800	0.800
14	0014	0002	8	0.700	0.200	3.275	4.330	0.800	0.800	0.800
15	0015	0002	8	0.700	0.200	2.977	4.330	0.800	0.800	0.800
16	0016	0002	8	0.700	0.200	2.572	4.330	0.800	0.800	0.800
17	0017	0002	8	0.700	0.200	3.261	4.330	0.800	0.800	0.800
18	0018	0002	8	0.700	0.200	3.247	4.330	0.800	0.800	0.800

1.6 Repeated Measures Data in SAS



	Student_ID	Group_ID	Group_Size	Group_Mean	Group_Variance	Test_ID	Test_Date	Test_Mean	Y	X1	X3
1	0001	0001	75	0.70000000	0.50000000	0001	2008-09-01	500.0	0.00	1.19	389
2	0001	0001	75	0.70000000	0.50000000	0002	2008-11-01	500.0	106.0	1.19	343
3	0001	0001	75	0.70000000	0.50000000	0003	2008-11-15	500.0	362.3	1.19	644
4	0002	0001	75	0.70000000	0.50000000	0001	2008-09-01	500.0	500.0	2.68	800
5	0002	0001	75	0.70000000	0.50000000	0002	2008-11-01	500.0	269.3	2.68	504
6	0002	0001	75	0.70000000	0.50000000	0003	2008-11-15	500.0	318.4	2.68	665
7	0003	0001	75	0.70000000	0.50000000	0001	2008-09-01	500.0	147.6	1.15	339
8	0003	0001	75	0.70000000	0.50000000	0002	2008-11-01	500.0	390.4	1.15	200
9	0003	0001	75	0.70000000	0.50000000	0003	2008-11-15	500.0	288.2	1.15	513
10	0076	0002	75	0.40000000	1.50000000	0001	2008-09-01	500.0	0.00	0.83	200
11	0076	0002	75	0.40000000	1.50000000	0002	2008-11-01	500.0	0.00	0.83	200
12	0076	0002	75	0.40000000	1.50000000	0003	2008-11-15	500.0	102.2	0.83	200
13	0077	0002	75	0.40000000	1.50000000	0001	2008-09-01	500.0	85.46	0.00	449
14	0077	0002	75	0.40000000	1.50000000	0002	2008-11-01	500.0	500.0	0.00	200
15	0077	0002	75	0.40000000	1.50000000	0003	2008-11-15	500.0	500.0	0.00	800
16	0078	0002	75	0.40000000	1.50000000	0001	2008-09-01	500.0	500.0	1.08	200
17	0078	0002	75	0.40000000	1.50000000	0002	2008-11-01	500.0	500.0	1.08	800
18	0078	0002	75	0.40000000	1.50000000	0003	2008-11-15	500.0	500.0	1.08	800

1.7 SAS Mixed Effects Procedures

- Main procedures discussed:
  - PROC MIXED
  - PROC GLIMMIX (experimental in v9.1.3; v9.2)
- Others with mixed model capabilities:
  - PROC GLM; PROC HPMIXED (v9.2); PROC LATTICE; PROC NESTED; PROC NLMIXED; PROC VARCOMP

2.1 From Linear to Multi-level Models

- Multi-level models (MMs) as generalization of linear and generalized linear models
- Assumption of independent error terms

$$Y_{ij} = \beta_0 + \beta_k X_{ijk} + r_{ij} \quad \text{with} \quad r_{ij} \sim N(0, \sigma^2)$$

```
PROC REG DATA= DATA_001;
  MODEL Y = X1-Xp;
RUN;
```

```
PROC MIXED DATA= DATA_001;
  MODEL Y = X1-Xp;
RUN;
```

2.2 Null Multi-level Model

- Two-level data; individuals within groups
- Intercepts vary by group; no other pred.
- Notation from Raudenbush & Bryk (2001).

Model

$$Y_{ij} = \beta_{0j} + r_{ij}$$
$$\beta_{0j} = \gamma_{00} + u_{0j}$$

Variance Components

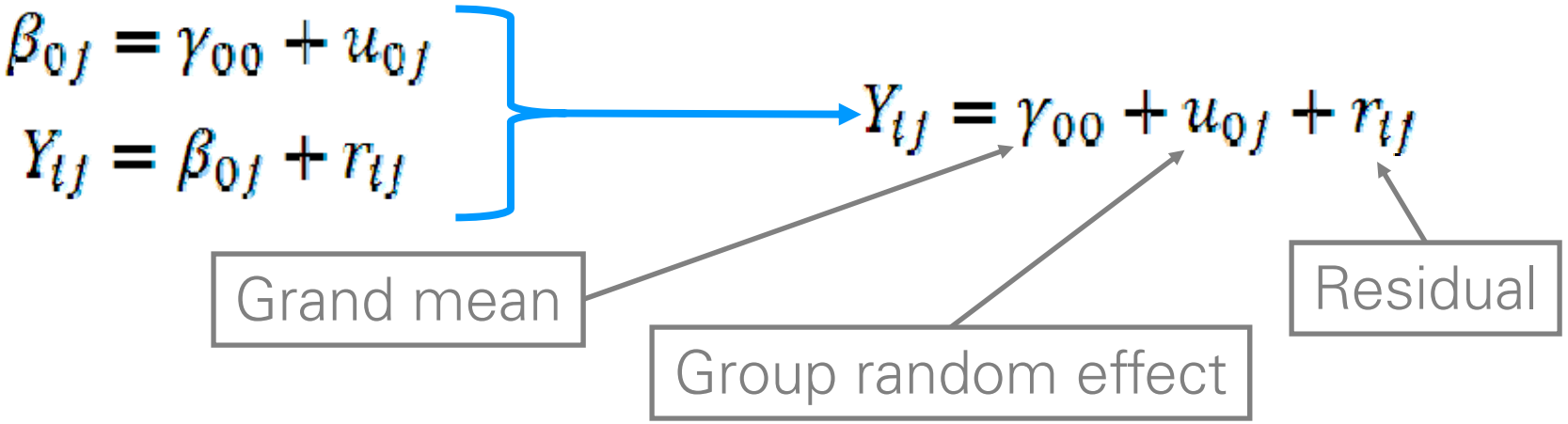
$$r_{ij} \sim N(0, \sigma^2) \quad T = [\tau_{00}]$$

SAS Code

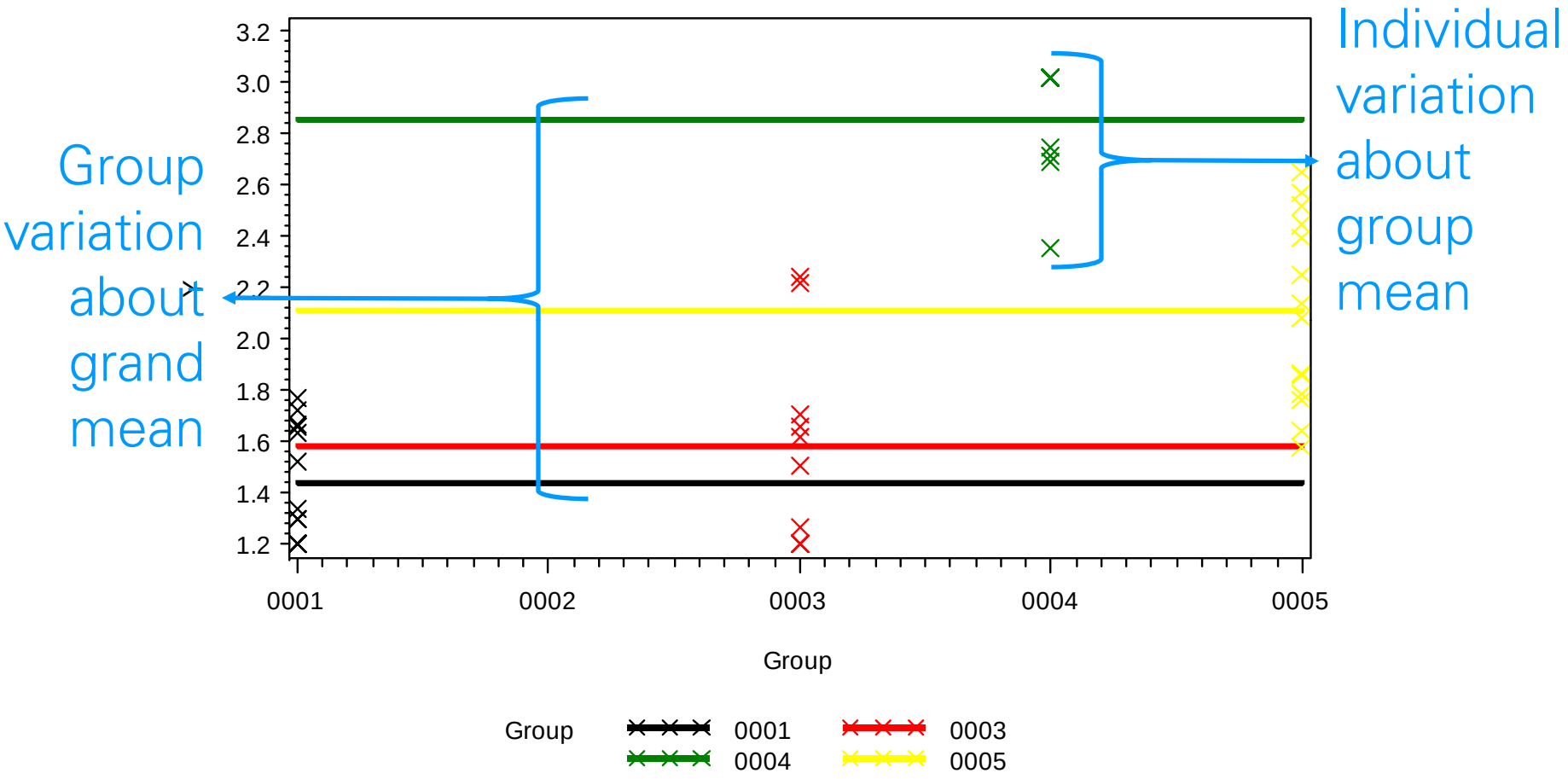
```
PROC MIXED DATA= DATA_001
   COVTEST;
 MODEL Y = / SOLUTION;
 RANDOM INTERCEPT /
  SUBJECT= Group TYPE= VC;
RUN;
```

2.3 Null MM as a Single Equation

- Substituting the expression for β_{0j} from the group-level into the individual-level equation:



2.4 Graph of Null MM



2.5 Do we Need a Multi-level Model?

- Based on the null model (or an ANOVA table) we compute the intra-class correlation coef. (ICC) for linear models
  - Indicates the proportion of error variance in the outcome relative to overall error variance.

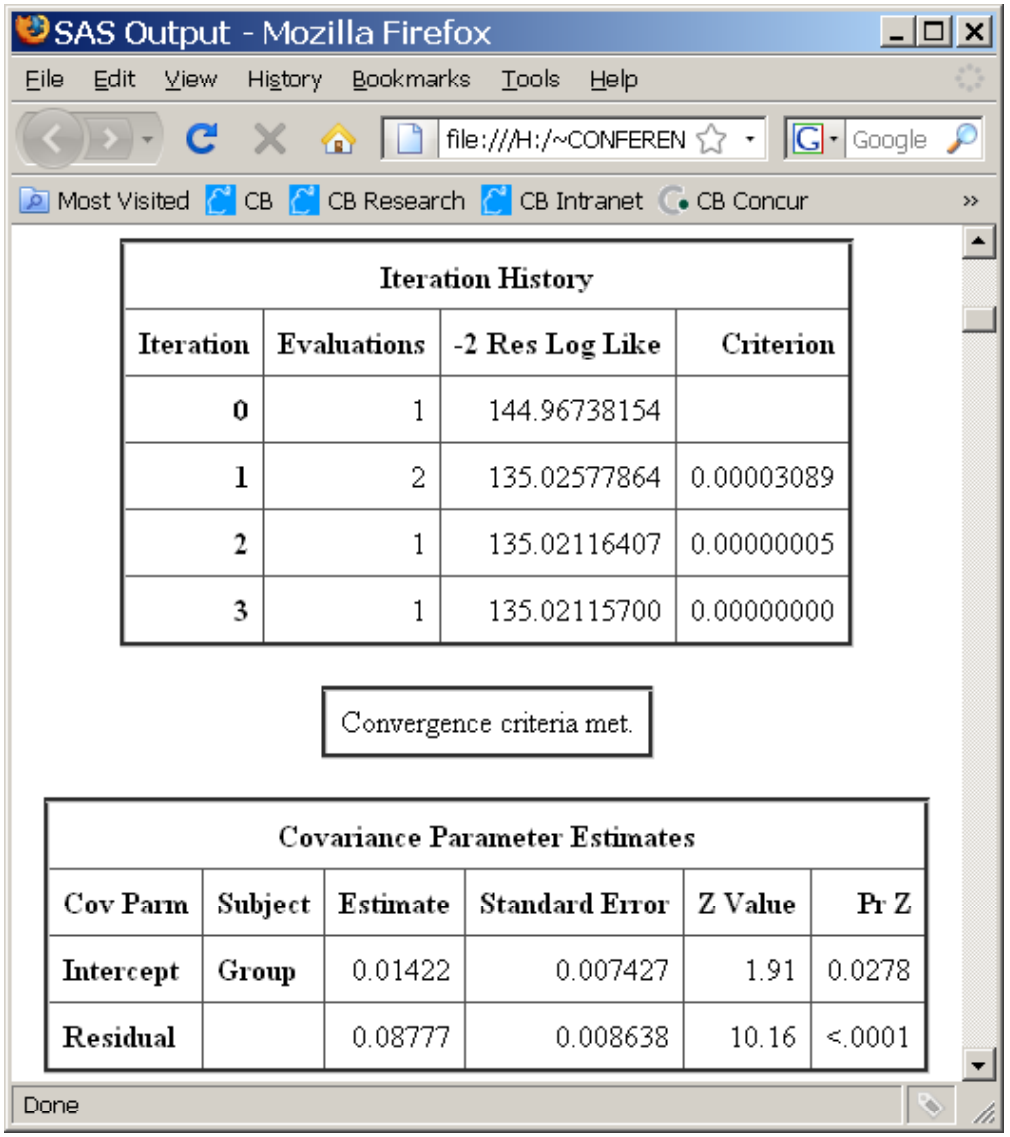
$$\rho = \frac{\tau_{00}}{\tau_{00} + \sigma^2} = \frac{\text{Group-level Error Var.}}{\text{Total Error Var.}}$$

2.6 SAS Output: Null Model

- ICC = 0.014 / 0.10

  = 0.14

- ICC not meaningful for GLMM
- COVTEST option in PROC MIXED



Iteration History

Iteration	Evaluations	-2 Res Log Like	Criterion
0	1	144.96738154	
1	2	135.02577864	0.00003089
2	1	135.02116407	0.00000005
3	1	135.02115700	0.00000000

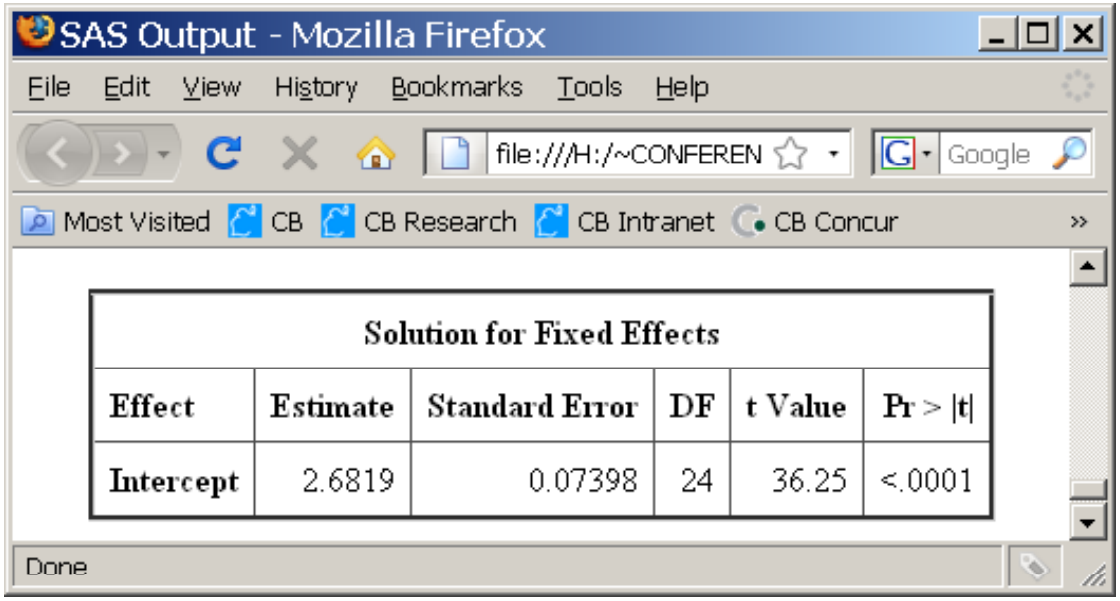
Convergence criteria met.

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error	Z Value	Pr Z
Intercept	Group	0.01422	0.007427	1.91	0.0278
Residual		0.08777	0.008638	10.16	<.0001

2.7 SAS Output: Null Model

- Random intercept specified
- No other predictors included in the model



Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	2.6819	0.07398	24	36.25	<.0001

3.1 MM with Level-1 Random Intercepts & Fixed Slopes

- Intercepts vary by group; other predictors are fixed

Model

$$Y_{ij} = \beta_{0j} + \sum_{k=1}^{p} \beta_{kj} X_{kj} + r_{ij}$$

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

$$\beta_{kj} = \gamma_{k0}$$

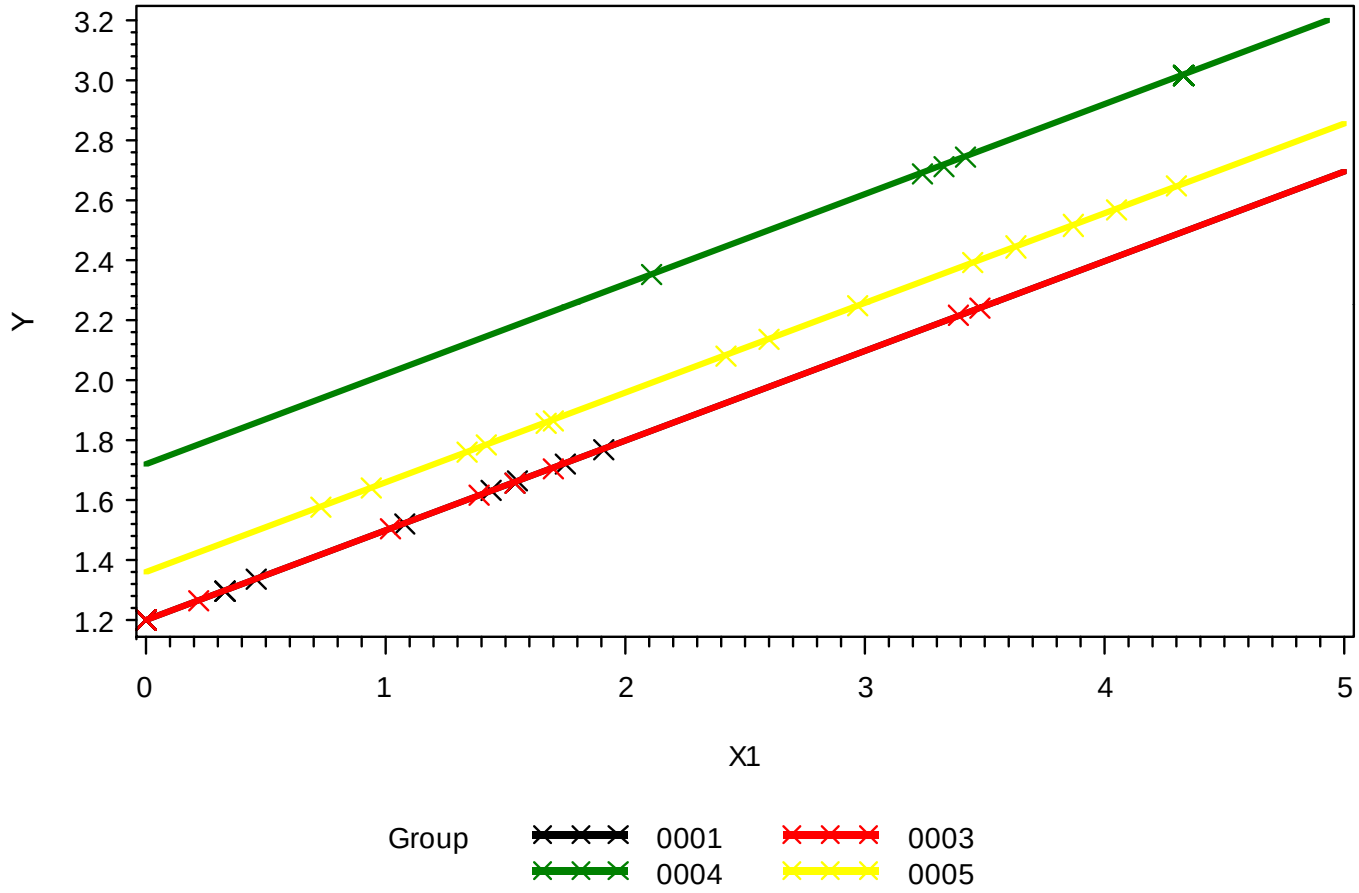
Variance Components

$$r_{ij} \sim N(0, \sigma^2) \quad \mathrm{T} = [\tau_{00}]$$

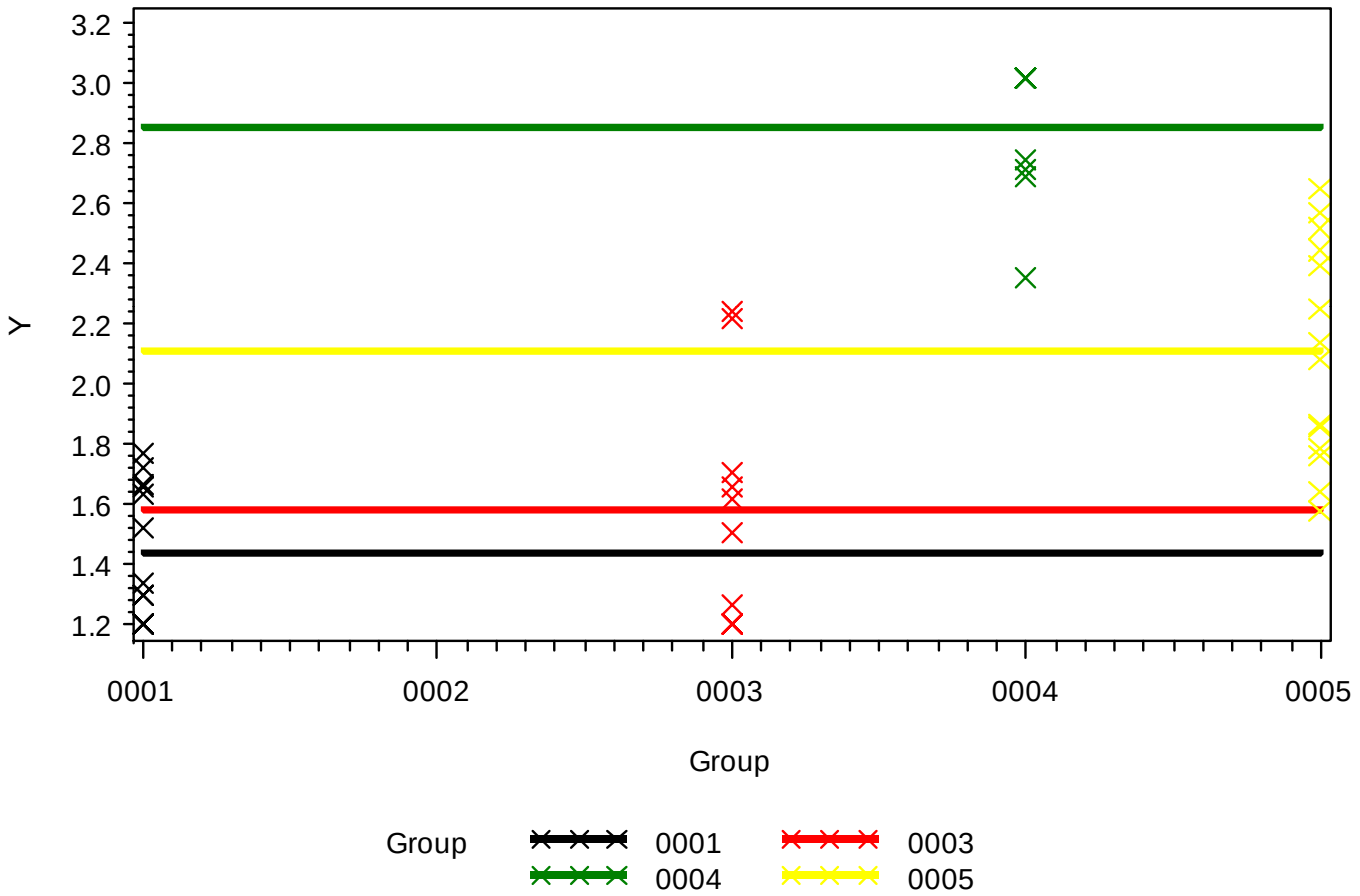
SAS Code

```
PROC MIXED DATA= DATA_001
   COVTEST;
MODEL Y= X1-Xp / SOLUTION;
RANDOM INTERCEPT /
 SUBJECT= Group TYPE= VC;
RUN;
```

3.2 Graph of MM with Level-1 Fixed X1 Slope

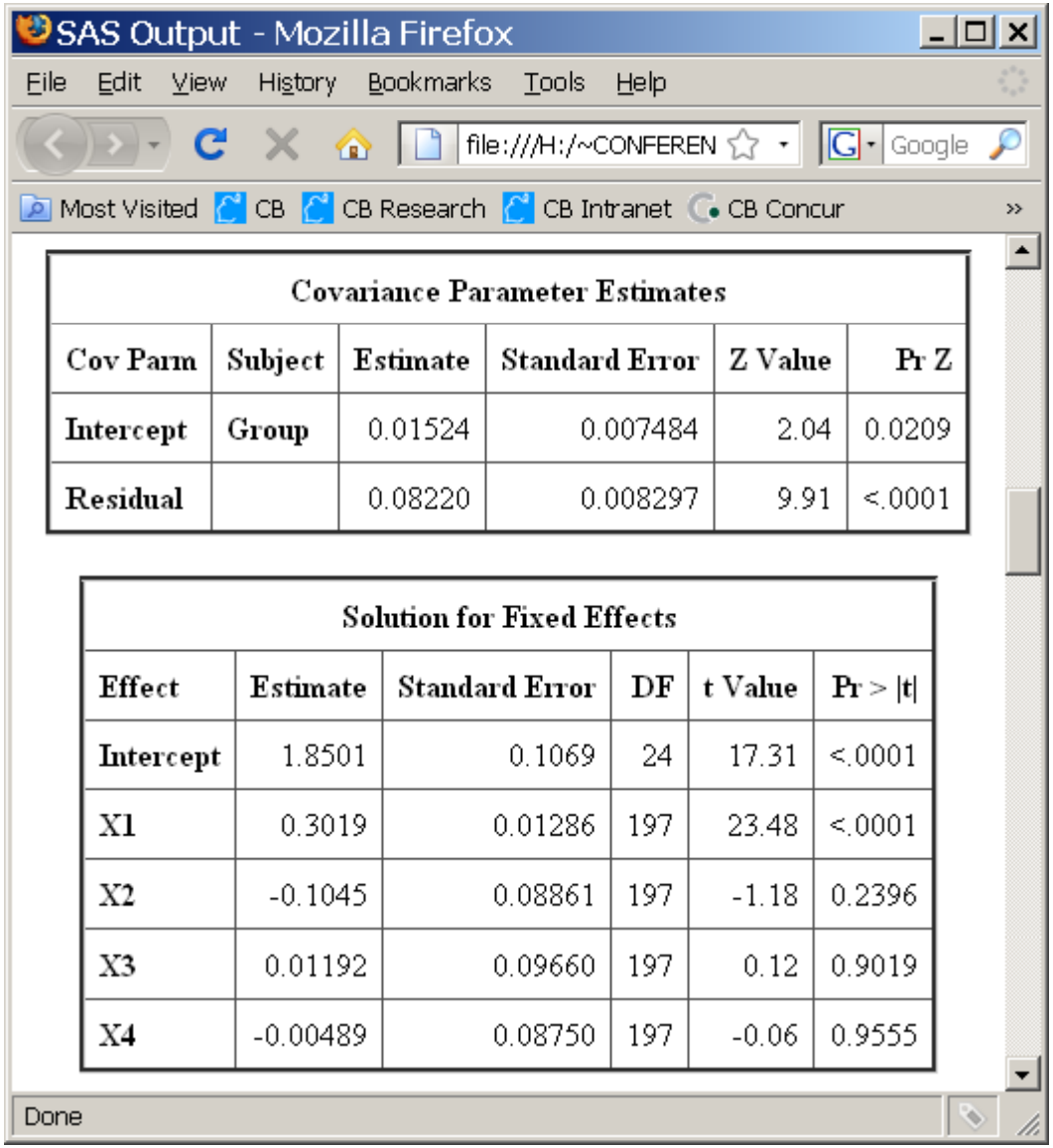


3.3 Graph of Null MM



3.4 SAS Output: Fixed Slopes

- Same random effects as null
- Slopes are constrained to be equal across groups



Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error	Z Value	Pr Z
Intercept	Group	0.01524	0.007484	2.04	0.0209
Residual		0.08220	0.008297	9.91	<.0001

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	1.8501	0.1069	24	17.31	<.0001
X1	0.3019	0.01286	197	23.48	<.0001
X2	-0.1045	0.08861	197	-1.18	0.2396
X3	0.01192	0.09660	197	0.12	0.9019
X4	-0.00489	0.08750	197	-0.06	0.9555

3.5 MM with Fixed- & Random-Intercept and Slope Effects

- Intercepts and other parameters vary

Model

$$Y_{ij} = \beta_{0j} + \sum_{k=1}^{p} \beta_{kj} X_{kj} + r_{ij}$$

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

$$\beta_{kj} = \gamma_{k0} + u_{kj}$$

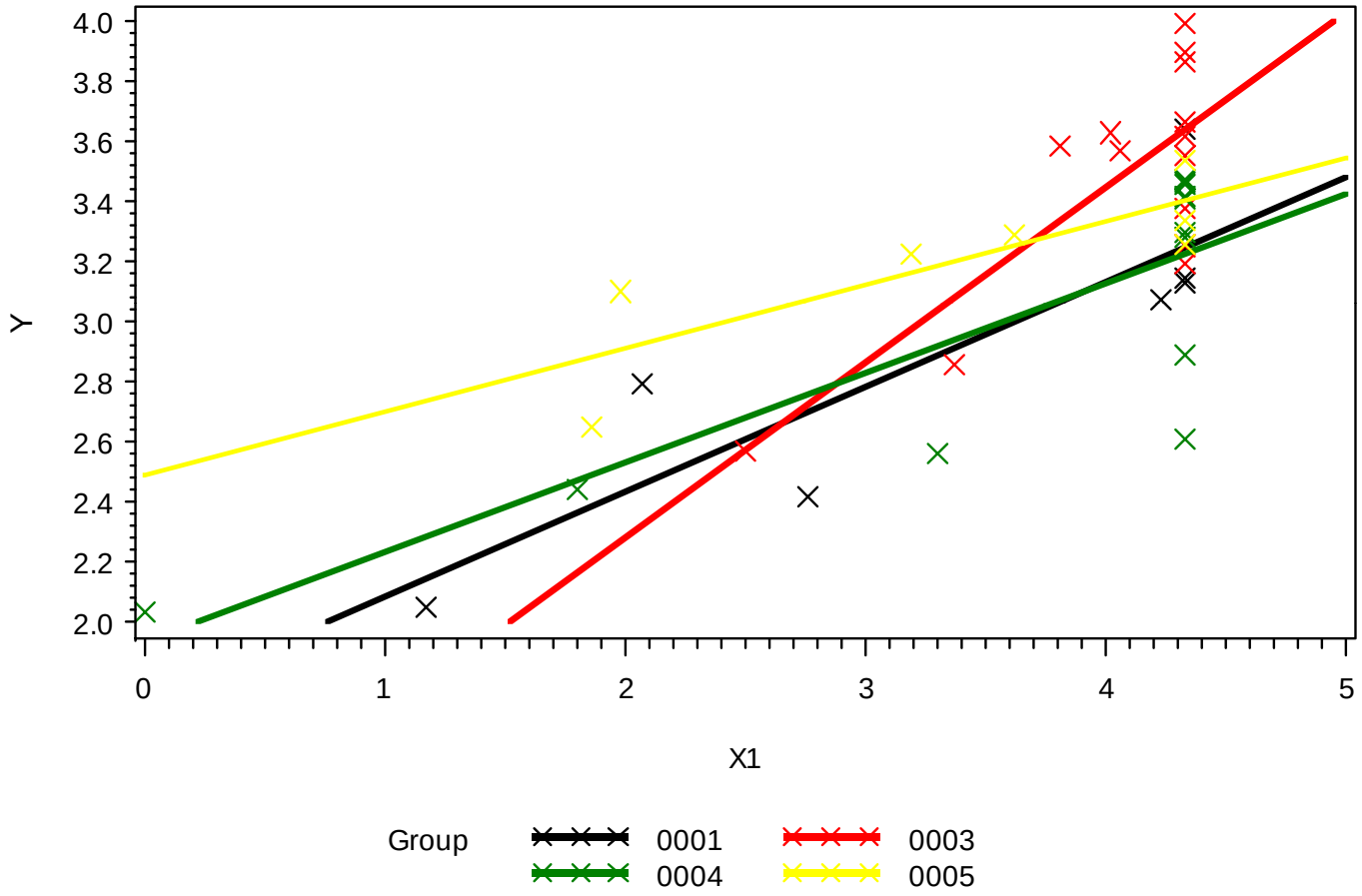
Variance Components

$$r_{ij} \sim N(0, \sigma^2) \quad \mathrm{T} = \begin{bmatrix} \tau_{00} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \tau_{pp} \end{bmatrix}$$

SAS Code

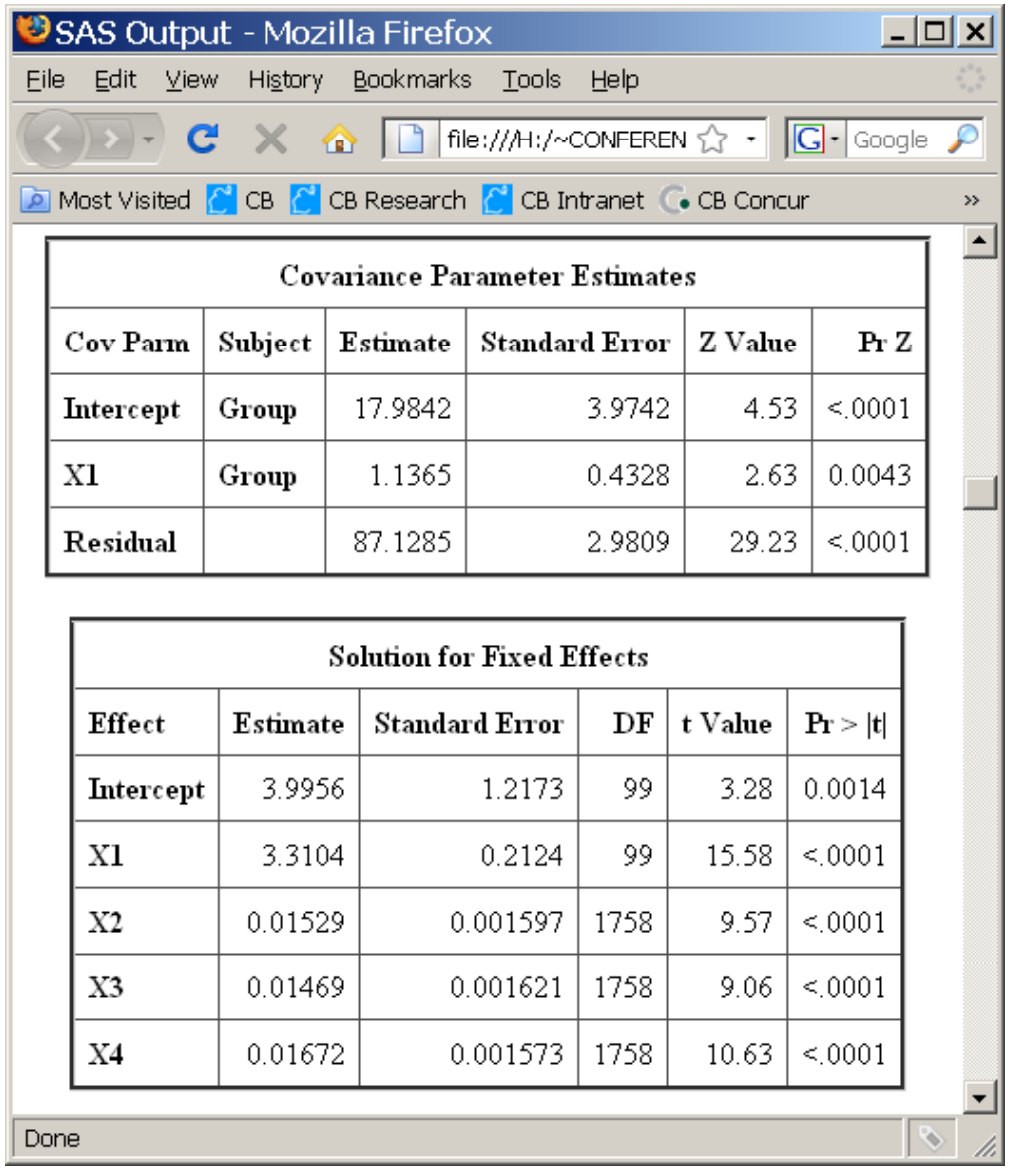
```
PROC MIXED DATA= DATA_001
   COVTEST;
 MODEL Y= X1-Xp / SOLUTION;
 RANDOM INTERCEPT X1-Xp /
  SUBJECT= Group TYPE= VC;
RUN;
```

3.6 Graph of MM with Level-1 Fixed- and Random-Effects



3.7 SAS Output: Random Slopes

- Same random effects as null
- Slope for X1 allowed to be vary across groups
- Notice τ_{11} in CovParms table



Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error	Z Value	Pr Z
Intercept	Group	17.9842	3.9742	4.53	<.0001
X1	Group	1.1365	0.4328	2.63	0.0043
Residual		87.1285	2.9809	29.23	<.0001

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	3.9956	1.2173	99	3.28	0.0014
X1	3.3104	0.2124	99	15.58	<.0001
X2	0.01529	0.001597	1758	9.57	<.0001
X3	0.01469	0.001621	1758	9.06	<.0001
X4	0.01672	0.001573	1758	10.63	<.0001

3.8 Generalized Linear Multi-level Models

- With a non-normal outcome, it may be appropriate to use a generalized linear multi-level model (GLMM)
  - PROC GLIMMIX will meet most needs; PROC NLMIXED may be appropriate otherwise
  - Recall ICC not meaningful; rely on size and standard error of cov. parameter estimates
  - Prediction sluggish in exp. GLIMMIX

3.9 SAS Specification of GLMM

```
PROC GLIMMIX DATA = Data_004
   METHOD= RSPL INITGLM IC= PQ
   NAMELEN= 65 NOCLPRINT;
 CLASS Group_ID;
 NLOPTIONS TECHNIQUE= NRRIDG;
 MODEL Event= X1-Xp / DIST= BINARY
   LINK= LOGIT SOLUTION CL;
 RANDOM INTERCEPT / SUBJECT= Group_ID;
RUN;
```

3.10 PROC GLIMMIX for Propensity Score Matching

- Add the following to our PROC GLIMMIX:

```
OUTPUT OUT= Data_004_Pred
   PREDICTED(BLUP ILINK)= Treat_Prob_Pred
   STDERR(BLUP ILINK)= Treat_Prob_SE;
```

- Use the predicted probabilities to perform propensity score matching.

3.11 Notes on Covariance Structure

- Theory and subject matter expertise should dictate the covariance structure
- For hierarchically structured data:
  - Generally most restrictive (*J* parm.): **`TYPE= VC`**
  - Generally least restrictive (J^2 parm.): **`TYPE= UN`**
- *Many* other options; check OnlineDoc

3.12 What Can Go Wrong?

- Convergence problems
  - PROC HPMIXED? Need SAS 9.2.
  - Another estimation method?
  - Fewer fixed- and random-effects?
- Variance components not estimated
  - Consider modifying RANDOM statement.

3.13 What *Else* Can Go Wrong?

- "Out of memory" Error
  - Some mixed procedures are prone to this error
  - "PROC MIXED is looking for a contiguous memory space and cannot find one large enough due to the relocation of the operating system modules." -SAS Problem Note 15060.
  - **Solution**: Try to defragment the target drive.

3.14 Diagnosing and Solving Problems

- Diagnose what's going wrong and when
  - Use the LOGNOTE option on PROC MIXED for estimation progress step-by-step
  - Remove statements / options to find the issue
  - If convergence is a problem, consider re-parameterizing the model
    - Especially in the case of generalized linear MMs

3.15 Example of LOGNOTE Output

NOTE: Levelizing effects.

NOTE: Processing subject and group effects.

NOTE: Setting up data.

NOTE: Loading data.

NOTE: Computing likelihood in iteration 0.

NOTE: Computing G derivatives in iteration 1.

NOTE: Computing likelihood in iteration 1.

NOTE: Computing G derivatives in iteration 1.

NOTE: Convergence criteria met but final hessian is not positive definite.

NOTE: Computing likelihood in iteration 2.

NOTE: A linear combination of covariance parameters is confounded with the residual variance.

NOTE: Computing Cholesky root of cross-products matrix.

NOTE: Computing H matrix.

NOTE: Computing Type 3 sums of squares.

NOTE: PROCEDURE MIXED used (Total process time):

real time	0.35 seconds
cpu time	0.14 seconds

4.1 Research & SAS Applications of Multi-level Models

- Model first-year grade point average at 110 colleges and universities
- Model probability participation in Advanced Placement Program® at many high schools
- Model high-school- and college-variability in the effect of AP® participation

4.2 Model FYGPA Across 110 Colleges

- Outcome: FYGPA, as normal continuous
- Levels: student and college (nested)
- Possible SAS Procedures:
  - PROC MIXED
  - PROC HPMIXED (new in 9.2)
    - For large number of fixed- or random-effects

4.3 Model AP Participation Across Many High Schools

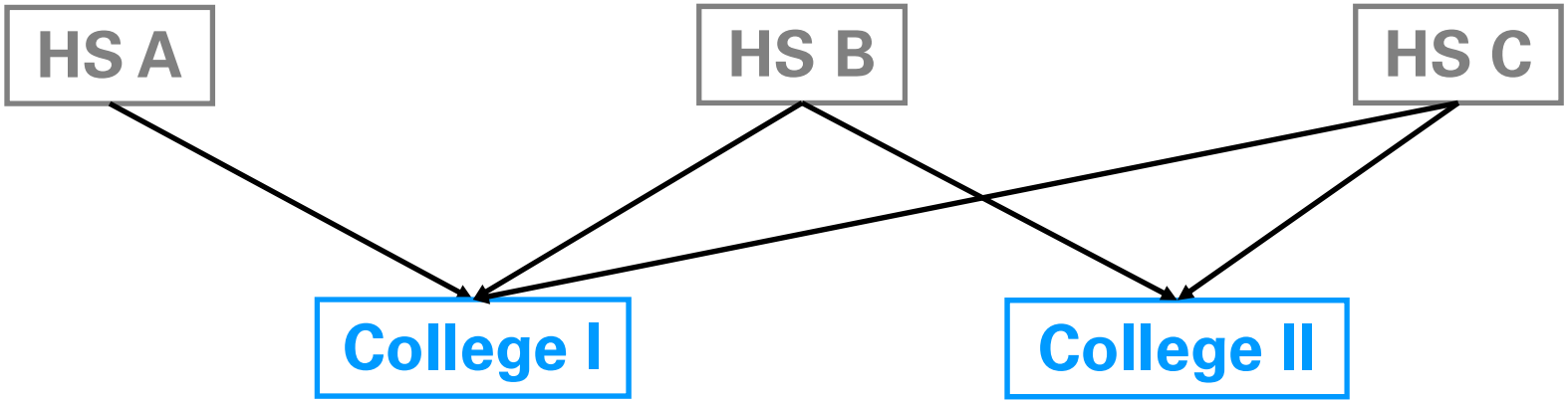
- Outcome: Indicator of AP Participation
- Levels: student and high school (nested)
  - High schools may be nested within districts
- Possible SAS Procedures:
  - PROC GLIMMIX (9.2; exp. in 9.1.3)
  - PROC NLMIXED

4.4 Model Effect of AP® Participation across HS & Colleges

- Outcome: FYGPA, as normal continuous
- Levels: student, high school and college
  - Data are cross-classified by HS and college
- Possible SAS Procedures:
  - PROC MIXED
  - PROC HPMIXED (new in 9.2)

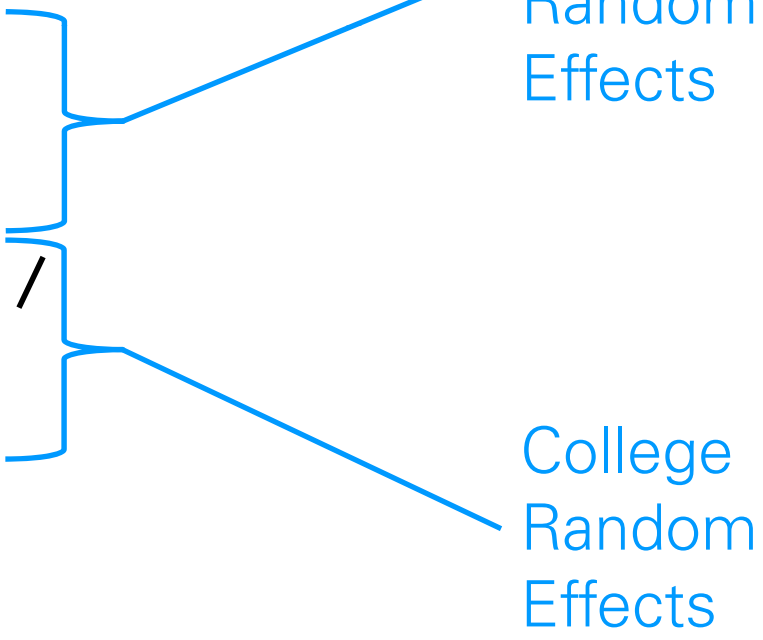
4.5 HS & College Cross-classified Data

- Data are not strictly nested
  - Students from high schools A, B and C attend colleges I and II
  - Random effects: high school and college levels



4.6 SAS Specification of CCMM

```
PROC MIXED DATA = Data_002 COVTEST;
 CLASS HS_ID College_ID;
 MODEL Y= X1-Xp / SOLUTION CL;
 RANDOM INTERCEPT X1 /
   SUBJECT= HS_ID;
 RANDOM INTERCEPT X1-Xp /
   SUBJECT= College_ID;
RUN;
```



4.7 Non-Educational Applications

- Multi-site clinical trials
  - Patients strictly nested within sites or cross-classified across sites
- Retail applications
  - PROC GLIMMIX to estimate same-store sales over quarterly observations

4.8 General References

- Multi-level Modeling Texts:
  - Raudenbush, S. W. and Bryk, A. S. 2002. Hierarchical Linear Models: Applications and Data Analysis Methods. Second Edition.
  - Snijders, T. A. B. and Bosker, R. J. 1999. Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling.

4.9 SAS-Oriented Reference

- Multi-level Modeling Articles for SAS:
  - Singer, J. D. 1998. "Using SAS PROC MIXED to Fit Multilevel Models, Hierarchical Models, and Individual Growth Models." Journal of Educational and Behavioral Statistics, Vol. 24, 323-355.

5.1 Question / Answer Period

- Any questions? Statistical, SAS-oriented or otherwise?

Thank You

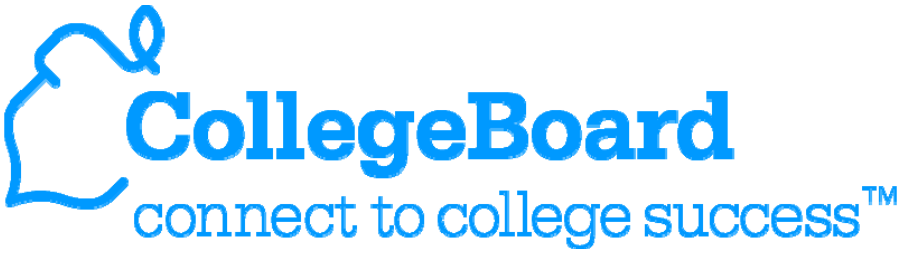
- Thank you all very much for having me.
- Special thanks to Martin Feuerman of NYASUG and Mary-Margaret Kerns of The College Board.

Contact Information

Brian F. Patterson
The College Board, Research & Development
Assistant Research Scientist

212-713-7714
bpatterson@collegeboard.org

http://www.collegeboard.com/research


CollegeBoard
connect to college success™