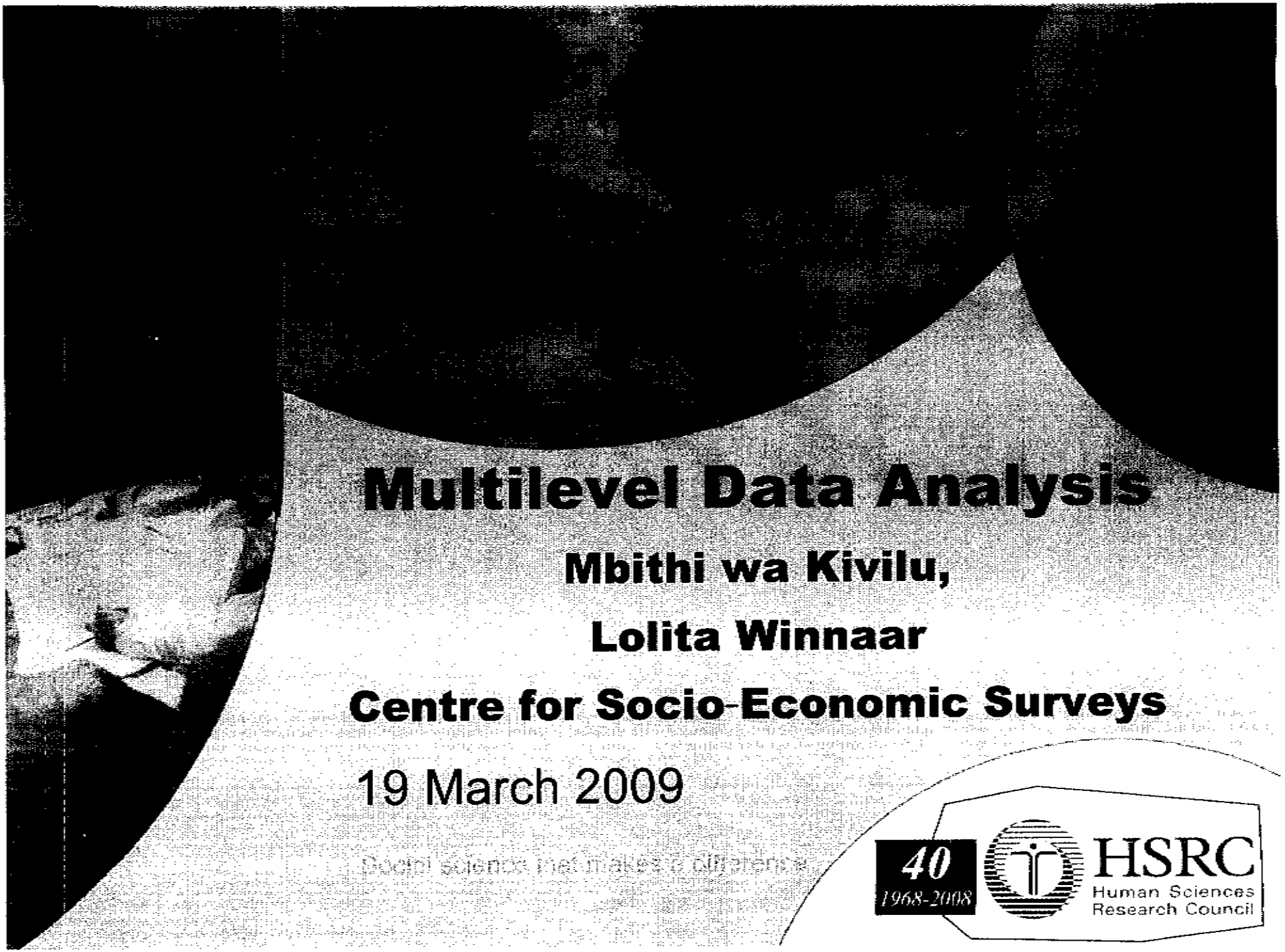


Multilevel Data Analysis

Mbithi wa Kivilu,

Lolita Winnaar

Centre for Socio-Economic Surveys

19 March 2009

Social science that makes a difference

40
1968-2008

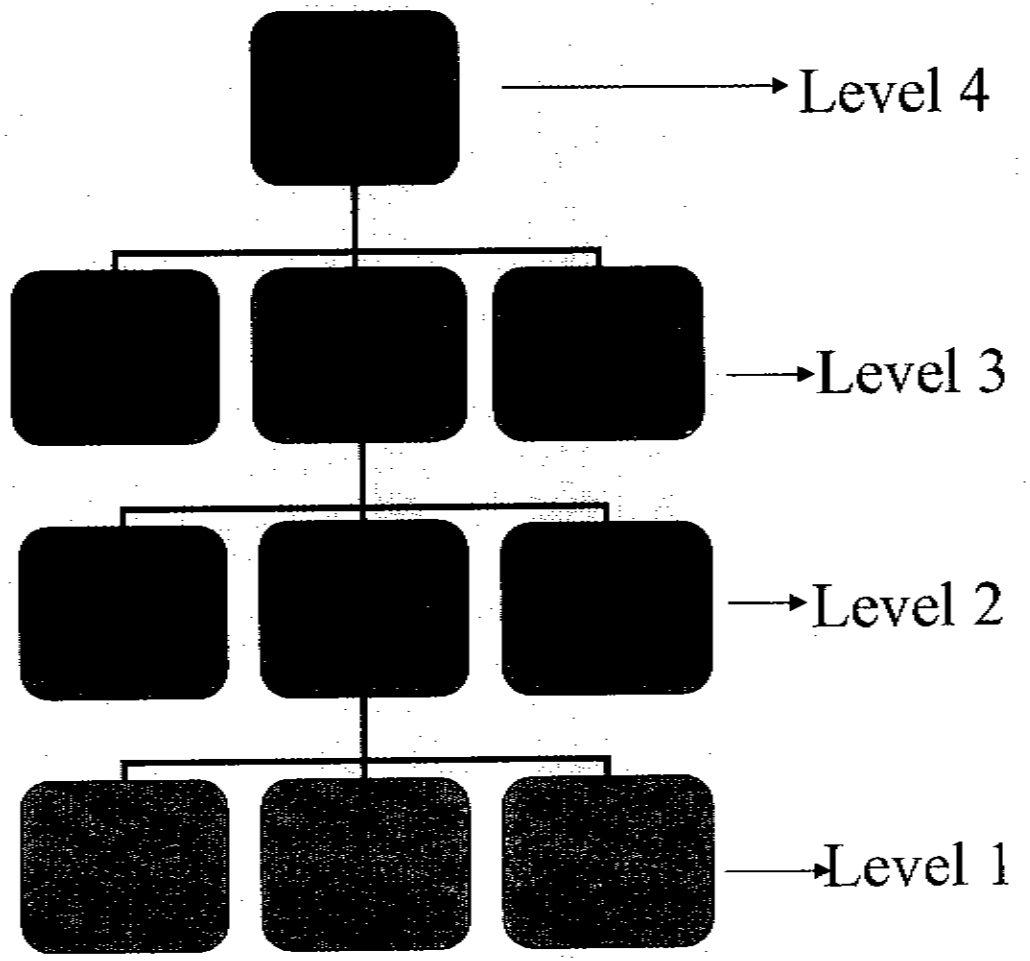
HSRC
Human Sciences
Research Council

Introduction

- The structure of data from social organizations, such schools, families, etc. are hierarchical or multi-level in nature.
- In education, for example, students are said to be nested within classrooms which are in turn nested within schools, schools are nested within school districts which are in turn nested within provinces.
- Despite the prevalence of hierarchical structures in behavioural and social research, researchers often fail to address them adequately in the data analysis.

Knowledge Systems

Hierarchical structure of educational data



Knowledge Systems

- Learners at level 1 are nested within schools at level 2 which are in turn nested within district at level 3, which are nested within province at level 4.
- Other data hierarchies include Repeated-measures data and meta-analytic data.
- Data repeatedly gathered on an individual is hierarchical as all the observations are nested within individuals

Why is a hierarchical data structure an issue

- Students within a particular classroom tend to come from community that is more homogeneous in terms of family background, socio-economic status, race or ethnicity or religion than the population as a whole.
- Furthermore , students in the same classroom share the experiences of being in the same teacher, physical environment and similar experiences, which may lead to increased homogeneity over time.

Assumptions of Regression models

- Hierarchical Linear Modelling (HLM) is commonly used when analyzing hierarchical or multi-level social science data.
- Like all other statistical applications (eg. Regression models) certain assumptions should be met before any HLM analysis is performed.
- With standard HLM modelling the critical assumptions are:
- The expected outcome must be expressed as a linear function of the regression coefficients, and
- The random effects (residuals) are normally distributed with constant variance (independence of observations/measurements).

Knowledge Systems

independence of the observations

- The average correlation between variables measured on students from the same schools will be higher than the average correlation between variables measured on students from different schools.
- Assumption of independence of the observations is violated. Estimates of standard errors are relatively very small leading to highly significant results.

Normality of the outcome variable

- The dependent variable should follow a normal distribution.
- Non-normally distributed variables which are skewed and have large kurtosis with substantial outliers can distort relationships and significance tests.
- Visually inspect data plots
- Skewness and kurtosis indices give information about normality

Knowledge Systems

Linear relationship between the dependent and independent variables

- Regression models can only accurately estimate the relationship between dependent and independent variables if the relationships are linear in nature.
- If the relationship is not linear, the results of the analysis will under-estimate the true relationship.
- Examine residual plots (plots of standardized residuals as a function of standardized predicted values)
- Routinely run regression analyses that incorporate curvilinear components (squared and cubic terms) utilize the non-linear regression option

Knowledge Systems

Unit of analysis

- When outcomes, for example academic achievement, are gathered at the individual level and other variables at classroom or school level e.g. school size, or student composition (race or gender) the question arises as to what the unit of analysis should be and how to deal with the cross-level nature of the data.
- We may convert variables from one level to another by aggregation or disaggregating.
- Aggregation means that variables at a lower level are transferred to a higher level, for instance, by computing the school mean score from individual student's test scores.
- Disaggregating means moving variables to a lower level, for instance by assigning to all students a variable that reflects the composition of the school they belong to such as gender or race.

HLM approach

- The HLM approach allows for explicit modeling of effects at the various levels of the hierarchy
- All estimated effects are adjusted for the individual-level and group-level influence on the dependent variable.
- A learner-level regression model is estimated for each school to predict learners' measure of performance using learner characteristics.
- Simultaneously, at the school level , a regression model is defined using school characteristics to estimate the parameters obtained at the learner-level
- Equations at each level are estimated at the same time and the variance at one level is taken into account in estimating the next level.
- The level 1 in our example will represent the relationships among the learner level variables, the level-2 model will capture the influence of school level factors and a level three may incorporate district level effects.

Basic HLM Illustrations

- Formally there are $i = 1,\ldots..n_j$ level-1 units, (e.g. learners) which are nested with each of $j = 1,\ldots J$ level-2 units (e.g. schools).

- LEVEL 1 model:

- $Y_{ij} = \beta_{0j} + \beta_{1j} X_{1j} + \ldots + \beta_{pj} X_{pj} + e_{ij}$
- Where:
- β_{pj} (p=0,1, …P) are level 1 coefficients
- X_{pj} is a level-1 predictor p for case i in level-2 unit J
- $e_{ij} \sim N(0,\sigma^2)$ normally distributed with mean zero and variance σ^2

- LEVEL 2 Model: Each of the β_{pj} coefficients in the level-1 model becomes an outcome variable in the Level-2 model

- $\beta_{0j} = \gamma_{00} + \mu_{0j} \quad \mu_{0j} \sim N(0, \tau_{00})$
- $\beta_{1j} = \gamma_{10} + \gamma_{11} Z_j + \mu_{1j}$

Where Z is the level 2 predictor

Knowledge Systems

Research Questions:

- How much do SA high schools vary if their mean mathematics achievement?
- Do schools with high MEAN SES also have high math achievement?
- Is the strength of association between student SES and math achievement similar across schools? Or is SES a more important predictor of achievement in some schools than in others?

Knowledge Systems

Variable Descriptions

LEVEL1		
id	Unique Identifier linking level 1 & level 2	
mathach	Math achievement score	continuous variable
female	Gender	1 = Female
		0 = Male
ses	Socio-economic status	-1 = Low
		0 = Average
		1 = High

LEVEL 2		
id	Unique Identifier linking level 1 & level 2	
sector	School type	1 = Private
		2 = Public
meanses	School mean socio-economic status	

Data for illustration (level 1 data)

id	minority	female	ses	mathach
1224	0	1	-1.528	5.876
1224	0	1	-0.588	19.708
1224	0	0	-0.528	20.349
1224	0	0	-0.668	8.781
1224	0	0	-0.158	17.898
1224	0	0	0.022	4.583
1288	0	1	-0.788	7.857
1288	1	0	-0.328	10.171
1288	0	0	0.472	15.699
1288	0	1	0.352	22.919
1288	1	1	-1.468	10.664
1288	0	1	0.202	13.543
1296	1	0	-0.608	8.773
1296	1	1	1.242	12.176
1296	0	0	-0.688	3.052

Data for illustration (level 2 data)

id	sector	meanses
1224	0	0
1288	0	0
1296	0	0
1308	1	1
1317	1	0
1358	0	0
1374	0	0
1433	1	1
1436	1	1
1461	0	1
1462	1	-1
1477	1	0

HLM 2 example (base model)

- **LEVEL 1 (Learner Level):**

$$\text{MathAch} = \beta_{0j} + r_{ij} \qquad r_{ij} \sim N(0, \sigma^2)$$

- **LEVEL 2 (School Level):**

$$\beta_{0j} = \gamma_{00} + \gamma_{01} Z_j + \mu_{0j} \qquad \mu_{0j} \sim N(0, \tau_{00})$$

Knowledge Systems

Base Model output

Summary of the model specified (in equation format)

Level-1 Model

$$Y = B0 + R$$

Level-2 Model

$$B0 = G00 + U0$$

Sigma_squared = 39.14831

Tau
INTRCPT1,B0 8.61431

Tau (as correlations)
INTRCPT1,B0 1.000

Random level-1 coefficient	Reliability estimate
INTRCPT1, B0	0.901

Base Model output (cont.)

The outcome variable is MATHACH

Final estimation of fixed effects (with robust standard errors)

Fixed Effect	Coefficient	Standard Error	T-ratio	Approx. d.f.	P-value
For INTRCPT1, B0					
INTRCPT2, G00	12.636972	0.243628	51.870	159	0.000

Final estimation of variance components:

Random Effect		Standard Deviation	Variance Component	df	Chi-square	P-value
INTRCPT1,	U0	2.93501	8.61431	159	660.23259	0.000
level-1,	R	6.25686	39.14831			

Interpretation of results :Base Model

- The average school mean math achievement is 12.64 and standard error 0.24.
- The coefficient of 8.61 (with corresponding chi-square value of 660.23) indicates significant variability between schools in terms of their average math achievement.
- The reliability is 0.90 which tells us that the information we have for each school is reliable

Base Model :Intraclass Correlation Coefficient

The proportion of the total variance that is within schools.

$$\hat{\rho} = \hat{\tau}_{00} / (\hat{\tau}_{00} + \hat{\sigma}^2)$$

Where:

$\hat{\rho}$: proportion of variability explained

$\hat{\tau}$: between school variance

$\hat{\sigma}^2$: within school variance

Knowledge Systems

Variance explained by Base Model (cont.)

To determine the Intraclass Correlation Coefficient (ICC), look at the **"variance components"** section of the output.

The ICC = intercept (level 1) variance / total variance. In this case,

$$ICC = 8.61/(8.61+39.15) = 0.18$$

- which means that 18% of the total variability is attributable to individual differences between schools.
- Alternatively 82% of the total variance is within schools.

Level 1 Predictor variables added

Knowledge Systems

- **LEVEL 1:**

$$\text{MathAch} = \beta_{0j} + \beta_{1j}(\text{SES}) + r_{ij} \qquad r_{ij} \sim N(0, \sigma^2)$$

- **LEVEL 2:**

$$\beta_{0j} = \gamma_{00} + \mu_{0j} \qquad \mu_{0j} \sim N(0, \tau_{00})$$
$$\beta_{1j} = \gamma_{10} + \mu_{1j}$$

Level 1 output

The outcome variable is MATHACH
Final estimation of fixed effects (with robust standard errors)

Fixed Effect	Coefficient	Standard Error	T-ratio	Approx. d.f.	P-value
For INTRCPT1, B0					
INTRCPT2, G00	12.636137	0.243722	51.847	159	0.000
For SES slope, B1					
INTRCPT2, G10	2.191172	0.129367	16.938	7183	0.000

Final estimation of variance components:

Random Effect		Standard Deviation	Variance Component	df	Chi-square	P-value
INTRCPT1,	U0	2.94491	8.67252	159	1756.17196	0.000
level-1,	R	6.08362	37.01040			

Knowledge Systems

Social science that makes a difference

HSRC Human Sciences Research Council

Interpretation of results

- The average of school means is 12.64, standard error of 0.24
- Average SES-achievement slope is 2.19, std error 0.13 and t-ratio of 17.26.
- This indicates that on average that student SES is significantly positively related to math achievement within schools.

- Estimated variance amongst school means $\tau_{00} = 8.64$ with chi-square statistic of 1770.5
- We infer that highly significant differences exist among school means.

Interpretation of results (cont.)

- Variance of slopes $\tau_{11} = 0.65$ with chi-square of 213.4 and d.f of 159 at p-value < 0.05.
- We reject the $H_0 : \tau_{11} = 0$ and we infer that the relationship between SES and math achievement within schools vary significantly across population of schools.
- The reliability is 0.901 which tells us that the information we have for each school is reliable.

Variance explained after adding predictor to level 1

$$\hat{\rho} = (\sigma_2\,(random\ ANOVA) - \sigma 2\ (SES))/\sigma 2\ (random\ ANOVA)$$

$$= 39.15 - 37.01\ /39.15$$

$$= 0.06$$

Level 1 residual variance (σ_2) has been reduced to 37.01, compared to 39.15 in our base model.

Here we see that adding SES as a predictor of math achievement reduced the within-school variance by only 6%.

Level 2 predictor variables added

LEVEL 1:

$$\text{MathAch} = \beta_0 + \beta_1(\text{SES}) + r_{ij} \qquad r_{ij} \sim N(0,\sigma^2)$$

LEVEL 2:

$$\beta_0 = \gamma_{00} + \gamma_{01} \text{*(SECTOR)} + \gamma_{02} \text{*(MEANSES)} + \mu_0$$
$$\beta_1 = \gamma_{10} + \gamma_{11} \text{*(SECTOR)} + \gamma_{12} \text{*(MEANSES)} + \mu_1$$

Level 2 output

Sigma_squared = 36.70313

Tau

INTRCPT1,B0	**2.37996**	**0.19058**
SES,B1	**0.19058**	**0.14892**

Tau (as correlations)

INTRCPT1,B0	**1.000**	**0.320**
SES,B1	**0.320**	**1.000**

Random level-1 coefficient	**Reliability estimate**
INTRCPT1, B0	**0.733**
SES, B1	**0.073**

Knowledge Systems

Level 2 output (cont.)

Knowledge Systems

Final estimation of fixed effects
(with robust standard errors)

Fixed Effect	Coefficient	Standard Error	T-ratio	d.f.	Approx. P-value
For INTRCPT1, B0					
INTRCPT2, G00	12.631549	0.140082	90.173	157	0.000
SECTOR, G01	1.226384	0.308484	3.976	157	0.000
MEANSES, G02	5.333056	0.334600	15.939	157	0.000
For SES slope, B1					
INTRCPT2, G10	2.219870	0.108224	20.512	157	0.000
SECTOR, G11	-1.640954	0.237401	-6.912	157	0.000
MEANSES, G12	1.034427	0.332785	3.108	157	0.003

Final estimation of variance components:

Random Effect		Standard Deviation	Variance Component	df	Chi-square	P-value
INTRCPT1,	U0	1.54271	2.37996	157	605.29503	0.000
SES slope,	U1	0.38590	0.14892	157	162.30867	0.369
level-1,	R	6.05831	36.70313			

Results:

- The coefficient of **2.37996** (with corresponding chi-square value of **605**) indicates significant variability among schools in terms of their average math achievement.
- The largest variance component is at level-1 of the model (**36.70313**), indicating that quite a lot of the variation in the outcome remains unexplained by this model.
- There is no significant variability in terms of the SES slopes for the level-2 units, as indicated by the estimate of **0.14892** (with p-value of **0.369**) for the level-2 component.

Variance explained after adding level 2 predictor variables

$$\hat{\rho} = (\tau_{00}\ (\textit{random coeff.}) - \tau_{00}\ (\textit{current model}))/\ \tau_{00}\ (\textit{random coeff.})$$

$$= (8.67 - 2.38) / 8.67$$

$$= 0.73$$

Sector (private, public), and MeanSES explain 73 percent of the variance in average math achievement.

Knowledge Systems

Conclusion

- We need to understand data in terms of structure, type of variables, and relationships being investigated
- Problem of unit of analysis is avoided and data is no longer aggregated or disaggregated.
- We obtain accurate and reliable estimation of each level effects
- All estimated effects are adjusted for individual-level and group-level influence on the outcome variable.
- The only draw back of applying HLM is that it requires advanced level of sophistication in statistics.

Knowledge Systems

THANK YOU!!

HSRC
Human Sciences
Research Council