

## **Dataset 1 - Strength of HLM over OLS.**

### **For use with either S-PLUS or R**

In this dataset, users are shown that sometimes the result of an HLM analysis can be exactly the opposite from the results of Ordinary Least Squares (OLS). The dataset that we will be using in this example is called “example.data”. In this example, the dependent variable is a science achievement test score measured at the individual level. The independent variable called “urban” is a composite score from scores on student SES and other markers of student “urbanicity”. Thus, high student scores on “urban”, represent a student who has low SES and also has many of the markers associated with students in a highly urban setting. The data represent 160 students nested in 16 schools (10 students in each school). The level 2 grouping structure is identified by the variable “group”.

First, let’s look at the results from the OLS procedure. By using the commands:

```
lm.out<-lm(data=example.data, SCIENCE~URBAN)
summary(lm.out)
```

we can produce the following estimates:

```
Call: lm(formula = SCIENCE ~ URBAN, data = example.data)
```

```
Residuals:
```

```
   Min       1Q   Median       3Q      Max
-5.336 -2.129  0.4919  2.043  5.009
```

```
Coefficients:
```

```
              Value Std. Error  t value Pr(>|t|)
(Intercept) -1.2511   0.5937   -2.1072  0.0367
          URBAN  0.8276   0.0386   21.4248  0.0000
```

```
Residual standard error: 2.592 on 158 degrees of freedom
```

```
Multiple R-Squared:  0.7439
```

```
F-statistic: 459 on 1 and 158 degrees of freedom, the p-value is 0
```

```
Correlation of Coefficients:
```

```
      (Intercept)
URBAN -0.9386
```

If we were to trust these results without looking at further HLM estimates, we might assume that as students become more “urban”, their scores on the science achievement variable tend to increase (c.f., slope coefficient of 0.8276). As we will see from the HLM analysis, this would be an egregious error.

We first fit the null model, also called the unconditional model, in which just the intercept is estimated in a random coefficients model. The full model can be represented as:

$$y_{ij} = \gamma_{00} + u_{0j} + e_{ij}$$

with a level-1 model of:

$$y_{ij} = \beta_{0j} + e_{ij}$$

where  $y_{ij}$  represents the scores for student “ $i$ ” in school “ $j$ ” on the dependent variable,  $\beta_{0j}$  is the mean science achievement for school “ $j$ ”, and  $e_{ij}$  are the student-level random deviates around school “ $j$ ’s” mean. Further extrapolated, we can see that for each student at level-1, their score is represented by the following equations:

$$\begin{cases} y_{11} = \beta_1 + e_{11} \\ y_{21} = \beta_1 + e_{21} \\ \dots \\ y_{ij} = \beta_j + e_{ij} \end{cases} .$$

The level-2 model would then be:

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

where  $\beta_{0j}$  is the mean science achievement for school “ $j$ ”,  $\gamma_{00}$  is the overall grand intercept, and  $u_{0j}$  is school “ $j$ ’s” random deviate around the grand mean.

We can estimate this model with the command lines:

```
lme.null<-lme(data=example.data, SCIENCE~1, random = ~ 1 | GROUP)
summary(lme.null)
```

Notice that we are “predicting” the “SCIENCE” variable with a command of “SCIENCE~1”. In S-PLUS and R, this has the effect of having the “SCIENCE” variable predicted by a

Doing this will yield the following results:

```

*** Linear Mixed Effects Model ***

Linear mixed-effects model fit by REML
Data: example.data
   AIC      BIC    logLik
643.8561 653.0628 -318.9281

Random effects:
Formula: ~ 1 | GROUP
(Intercept) Residual
StdDev:    5.052846 1.406829

Fixed effects: SCIENCE ~ 1
              Value Std.Error DF t-value p-value
(Intercept) 10.6875  1.268098 144  8.427975 <.0001

Standardized Within-Group Residuals:
      Min       Q1       Med       Q3      Max
-1.463671 -0.7214126 -0.006493055  0.7084265  1.450685

Number of Observations: 160
Number of Groups: 16

```

Multiply by -2 and this is a Chi-square

Square this value and it is the  $\sigma_{r_{ij}}^2$  or residual variance

Square this value and it is the  $\sigma_{\mu_j}^2$  or variance for the intercept

Maximum likelihood estimation of the grand mean of the dependent variable

We can then add the fixed effect for our “urban” variable by running the commands:

```

lme.one<-lme(data=example.data, SCIENCE~URBAN, random = ~ 1 | GROUP)
summary(lme.one)

```

Which will produce the following results:

```

*** Linear Mixed Effects Model ***

Linear mixed-effects model fit by REML
Data: example.data
   AIC      BIC    logLik
508.094 520.3444 -250.047

Random effects:
Formula: ~ 1 | GROUP
(Intercept) Residual
StdDev:    9.298169 0.8094491

Fixed effects: SCIENCE ~ URBAN
              Value Std.Error DF t-value p-value
(Intercept) 22.30291  2.426310 143  9.19211 <.0001
URBAN      -0.80523  0.047999 143 -16.77609 <.0001

Standardized Within-Group Residuals:
      Min       Q1       Med       Q3      Max
-2.893221 -0.6874918  0.01311717  0.6303738  2.323769

Number of Observations: 160
Number of Groups: 16

```

Maximum likelihood estimation of the slope of the independent variable

## Testing the different models.

Generally, we can just look at the AIC, BIC, and change in Chi-Square to determine if we have a better model (smaller numbers are better). We can also perform an ANOVA on the two separate models with the command:

```
> anova(lme.null, lme.one)
Warning messages:
  Fitted objects with different fixed effects. REML
  comparisons are not meaningful. in: anova(lme.null,
  lme.one)
  Model df      AIC      BIC    logLik  Test  L.Ratio
lme.null  1  3 643.8561 653.0628 -318.9281
lme.one   2  4 508.0940 520.3444 -250.0470 1 vs 2 137.7621

      p-value
lme.null
lme.one <.0001
```

This shows us that the “lme.one” model is a better model than the “lme.null”.

Next, we can add the random effect with the command line:

```
lme.two<-lme(data=example.data, SCIENCE~URBAN, random = ~ URBAN | GROUP)
summary(lme.two)
```

Doing so will produce the following results:

```
*** Linear Mixed Effects Model ***

Linear mixed-effects model fit by REML
Data: example.data
      AIC      BIC    logLik
424.1713 442.5469 -206.0857

Random effects:
Formula: ~ URBAN | GROUP
Structure: General positive-definite
              StdDev  Corr
(Intercept) 10.6584840 (Inter
  URBAN      0.5019958 -0.625
  Residual   0.5202538

Fixed effects: SCIENCE ~ URBAN
              Value Std.Error DF  t-value p-value
(Intercept) 22.39124  2.717018 143  8.241111 <.0001
  URBAN     -0.86701  0.129811 143 -6.678989 <.0001

Standardized Within-Group Residuals:
      Min      Q1      Med      Q3      Max
-2.376494 -0.7914869 0.007393824 0.6221058 2.169879

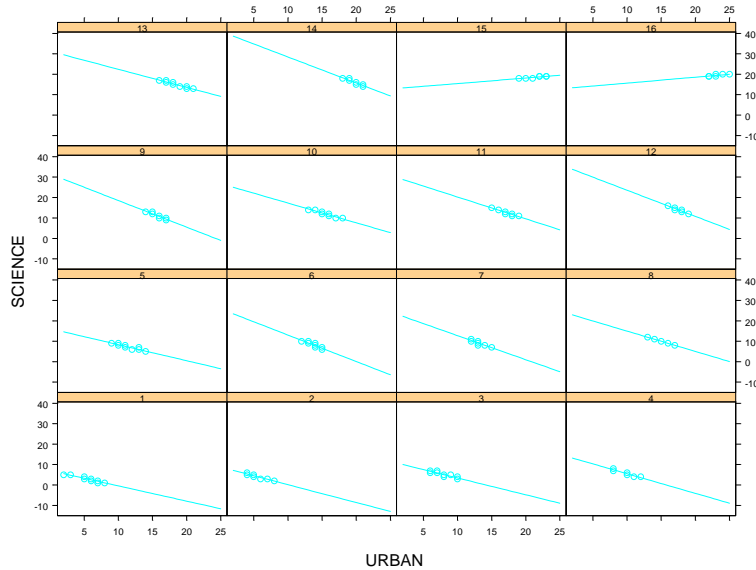
Number of Observations: 160
Number of Groups: 16
```

Notice that we now have a variance component for “URBAN” which is  $\sigma_{\mu_{1j}}^2$ . This is the random effect that we added.

We can see from the AIC and the BIC that this model has better fit to the data than the previous model.

If we wish to see our results graphically, we can run the command

```
plot(augPred(lme.two))
```



Finally, we can see from the last model that the relationship between “science” and “urban” in this dataset is actually negative, meaning that as a student has higher classification of the urban composite variable, scores on the science achievement variable tend to decrease. Remember that this is exactly the opposite from the conclusion that we drew from the OLS analysis. Also, we can see from either the fitted estimates or from the above graph that schools “15” and “16” have slopes that are dramatically different from the other schools. We can produce the Empirical Bayesian estimates for these schools with:

```
> coef(lme.two, level=1)
  (Intercept)      URBAN
1    7.038482 -0.7468700
2    8.901293 -0.8742814
3   11.668626 -0.8227974
4   15.130284 -0.9607321
5   16.185686 -0.7849303
6   26.028960 -1.2969891
7   24.550955 -1.1780442
8   24.894022 -0.9929565
9   31.570300 -1.3020020
10  26.967205 -0.9657086
11  30.982599 -1.0705300
12  36.360105 -1.2779106
13  31.267604 -0.8842928
14  41.201522 -1.2730693
15  12.724031  0.2710722
16  12.788238  0.2879622
```

This document was created with Win2PDF available at <http://www.daneprairie.com>.  
The unregistered version of Win2PDF is for evaluation or non-commercial use only.