Hierarchical Linear Modeling

Dan Byrd

UC Office of the President

Refresher on OLS models

1. OLS regression assumes that residuals (observed value- predicted value) are normally distributed and that each observation is independent from others and that the relationship between the IV and DV is linear.
2. While small violations of OLS assumptions are not generally a problem, a large violation can affect the interpretation

Source: Dewey (2018)

What is Hierarchal linear modeling?

- HLM is a type of regression model that takes into account the nested nature of data
  - Data is structured in groups “Nested” and coefficients can very by group (Gelman & Hill, 2007)

What is nested data?

1. When examining changes in responses over time, peoples' responses are likely to be correlated with each other. Because of the correlated nature of their responses, it isn't advisable to use a regular OLS model as the observations are not independent.
2. Students from a particular department and or campus may be more similar in their behavior when compared to a random sample from the UC population

Source: University of Bristol

Application to Panel Data

1. Panel data (aka longitudinal data) is a type of data where individuals are measured at multiple time points
   - Think back to meeting one where we discussed within subjects designs
2. Panel data allows researchers to control for measures you can't observe
   - When using cross sectional data one cannot determine if changes over time are due to differences in the population, verses actual changes over time
3. Has advantages over a repeated measures ANOVA (or a mixed ANOVA) as you don't have to drop a subject if you have missing data at one time point
   - Also due to the way sum of squares (measures how far an individual measurement is from the mean) is calculated in repeated measures ANOVAS, one cannot conduct post hoc-tests (e.g., simple main effects)
   - You can also treat time as a continuous variable (you still need at least three time points to do this)

Sources: Torres-Reyna (2007) Grace-Martin (2018)

Why Use Hierarchal Linear Modeling?

Correct inferences: *Traditional multiple regression techniques treat the units of analysis as independent observations. One consequence of failing to recognize hierarchical structures is that standard errors of regression coefficients will be underestimated*

Source: University of Bristol

Random and fixed effects

1. Researchers may want to acknowledge the affect of a value where specific values are not of interest
   - "in a study looking for the effect of a new drug on blood pressure, different doctors may prescribe the pill to different patients. The effect of a specific physician is not of theoretical interest, yet the investigator may suspect that different health care providers can contribute to a patient's outcome"
     - Because the doctors draw on a random sample, their impact is considered to be a random effect
     - Fixed effects refer to the levels of the factor that the researcher is interested in. In this example, fixed effects would be prescribing the drug or not prescribing the drug

Source: Albright (2016)

HLM Software (We will focus on SAS)

1. Stata
2. SAS
3. R
4. HLM software (hard to work with, requires two different data sets)

Walkthrough of an HLM output

Datasets for Exercise

- UCUES Panel Data (N=530)
  - 2007 Entry Cohort
  - Took UCUES three times mostly 2008, 2010 and 2012
  - Fake IDs and Fake Campus Names were used to protect student privacy

HLM Example Research Question

- Have beliefs about how free students feel to express their political beliefs changed over time?
  - Have they changed by campus

Three level model:

Panel data= nested within individual

Panel data=nested within campuses

SAS Code for HLM

- This code utilizes the UCUES panel dataset
- The DV is “I feel free to express my political beliefs on campus”
- Random effects are specified for fake ID and campus. In this case the data is nested both within the individual and the campus.
- SATTERTHWAITE= the pooled degrees of freedom(the number of independent values that can vary in an analysis, just think back to the F statement)

```
proc mixed data=Final covtest;
  class Fake_ID Fake_Campus
(ref="Atlanta")  App_Status Discipline
(ref="Engineering/Computer Sciences")  ;
  model RUCAGRXPRSPOLI = Fake_Campus
Svy_Yr|App_Status svy_YR|Fake_Campus  /
solution ddfm = SATTERTHWAITE ;
  random intercept / type=vc sub=Fake_ID;
 random intercept / type=vc
sub=Fake_Campus;
run;
```

SAS Output 1

- The key things to take away from this section of the output is
  - 1) 527 people were used in this analysis

Life Sciences Physical Sciences/Math Professional Fields Social Sciences/Psychology

Dimensions	
Covariance Parameters	3
Columns in X	24
Columns in Z per Subject	527
Subjects	1
Max Obs per Subject	1443

Number of Observations	
Number of Observations Read	1531
Number of Observations Used	1443
Number of Observations Not Used	88

SAS output 2

- The intercept is the estimate individual between subjects variance
- The residual is the estimated within subjects variance
  - Effects are significant which means its okay to include as random effects in the model
- “a variance component estimate can equal zero; in this case, you might want to drop the corresponding random effect from the model. However, be aware that changing the model in this fashion can affect degrees-of-freedom calculations.”

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error	Z Value	Pr > Z
Intercept	Fake_ID	0.4389	0.04732	9.28	<.0001
Intercept	Fake_Campus	0	.	.	.
Residual		0.7990	0.03714	21.51	<.0001

Source: SAS and the University of Michigan

SAS Output 3

- Intercept: This is the value at which the fitted line crosses the y-axis
  - Generally this isn't meaningful
- Main effects
  - Seattle campus relative to Atlanta Campus
  - Washington D.C. campus relative to Atlanta campus
  - New Orleans campus trend
  - Survey year
  - Freshman v transfers
- Interaction Effects
  - Survey year and app status
  - Survey year and Seattle campus
  - Survey year and Washington D.C. Campus

Source: Minilab

Solution for Fixed Effects

Effect	Fake_Campus	App_Status	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept			539.69	234.61	970	2.30	0.0216
Fake_Campus	Chicago		112.29	94.5910	953	1.19	0.2355
Fake_Campus	Detroit		18.4735	184.18	952	0.10	0.9201
Fake_Campus	Miami		-42.1655	103.22	966	-0.41	0.6830
Fake_Campus	Milwaukee		17.8877	96.7136	953	0.18	0.8533
Fake_Campus	New Orleans		247.66	140.80	979	1.76	0.0789
Fake_Campus	Seattle		275.98	120.25	967	2.30	0.0219
Fake_Campus	St. Louis		128.57	122.48	943	1.05	0.2941
Fake_Campus	Washington D.C		323.49	141.50	976	2.29	0.0225
Fake_Campus	Atlanta		0	.	.	.	.
svy_yr			-0.2663	0.1167	970	-2.28	0.0228
App_Status		Freshman	-473.16	227.70	971	-2.08	0.0380
App_Status		Transfer	0	.	.	.	.
svy_yr*App_Status		Freshman	0.2354	0.1133	971	2.08	0.0380
svy_yr*App_Status		Transfer	0	.	.	.	.
svy_yr*Fake_Campus	Chicago		-0.05585	0.04706	953	-1.19	0.2356
svy_yr*Fake_Campus	Detroit		-0.00902	0.09164	952	-0.10	0.9216
svy_yr*Fake_Campus	Miami		0.02103	0.05136	966	0.41	0.6822
svy_yr*Fake_Campus	Milwaukee		-0.00886	0.04812	954	-0.18	0.8539
svy_yr*Fake_Campus	New Orleans		-0.1233	0.07006	980	-1.76	0.0788
svy_yr*Fake_Campus	Seattle		-0.1372	0.05983	967	-2.29	0.0221
svy_yr*Fake_Campus	St. Louis		-0.06397	0.06094	943	-1.05	0.2941
svy_yr*Fake_Campus	Washington D.C		-0.1610	0.07040	976	-2.29	0.0224
svy_yr*Fake_Campus	Atlanta		0	.	.	.	.

SAS Output 4

- These are the overall tests the effects
- Survey year
- App status
- App status and year are all significant effects
- It's best to plot interaction effects

Type 3 Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
Fake_Campus	8	960	1.82	0.0693
svy_yr	1	969	13.55	0.0002
App_Status	1	971	4.32	0.0380
svy_yr*App_Status	1	971	4.32	0.0380
svy_yr*Fake_Campus	8	960	1.82	0.0693

Interaction effects

- Jeremy Dawson's excel files
- Interaction effects in R
- Interaction effects in SAS

SAS and HLM Continued

- To get R-Squared you have to do more work
  - *"Run the full model and get the variance component estimates. Then run the model again (with no fixed effects, which is the intercept only model), but specify the variance component(s) in a PARMS statement, and use the hold= option to fix them at the same values"*
  - *You can also get the R-squared by dividing the intercept from the random effects by the total sum of the random intercept and the residuals*
- You can't get standardized coefficients

HLM in R

- You will need to download the lme4 package
- library(readxl)
  - HLM_UCUES <- read_excel("Q:/dbyrd/Research Methods Users Group/HLM/HLM_UCUES.xlsx")
  - View(HLM_UCUES)
- UCLA Stats has a great example of how to do this analysis in R
- https://stats.idre.ucla.edu/r/examples/mlm-imm/r-kreft-chp-3/

```
lmer(math ~ homework + public + homework:public + (homework|schnum), REML=FALSE)

Linear mixed model fit by maximum likelihood
Formula: math ~ homework + public + homework:public + (homework | schnum)
  AIC  BIC logLik deviance REMLdev
 1767 1795 -875.4     1751    1739
Random effects:
 Groups   Name        Variance Std.Dev. Corr
 schnum   (Intercept) 40.503   6.3642
          homework    21.577   4.6451   -0.982
 Residual             42.954   6.5540
Number of obs: 260, groups: schnum, 10

Fixed effects:
              Estimate Std. Error t value
(Intercept)    59.2102     6.5976   8.975
homework        1.0946     4.6688   0.234
public        -15.9419     6.9775  -2.285
homework:public   0.9472     4.9385   0.192

Correlation of Fixed Effects:
            (Intr) homwrk public
homework    -0.966
public      -0.946  0.913
homwrk:pblc  0.913 -0.945 -0.965
```

lmer(y ~ time * tx + (time | therapist:subjects) + (time | therapist), data=df)
Three level code

Sources: R Psychologist and STATS UCLA

HLM in Stata

- Here is stata code for a basic two level multilevel model

```
lmer(math ~ homework + public + homework:public + (homework|schnum), REML=FALSE)

Linear mixed model fit by maximum likelihood
Formula: math ~ homework + public + homework:public + (homework | schnum)
  AIC  BIC logLik deviance REMLdev
 1767 1795 -875.4     1751    1739
Random effects:
 Groups   Name        Variance Std.Dev. Corr
 schnum   (Intercept) 40.503   6.3642
          homework    21.577   4.6451   -0.982
 Residual             42.954   6.5540
Number of obs: 260, groups: schnum, 10

Fixed effects:
                Estimate Std. Error t value
(Intercept)      59.2102     6.5976   8.975
homework          1.0946     4.6688   0.234
public          -15.9419     6.9775  -2.285
homework:public   0.9472     4.9385   0.192

Correlation of Fixed Effects:
            (Intr) homwrk public
homework    -0.966
public      -0.946  0.913
homwrk:pblc  0.913 -0.945 -0.965
```

Source: Stats UCLA

Group work

- Break into small groups
- formulate a research question
- Import the data and run a HLM model
- Interpret the results
- Report back to group on findings

Additional Information Slides

SAS HLM Output Interpretation

Element	Purpose
Dimensions	Size of relevant matrices
Number of observations	The number of cases used in the analysis from the dataset
Iteration history	Residual log likelihood – the residual of the natural logarithm of the likelihood function This table tells you how many times the objective function is evaluated during each iteration
Convergence Parameter Estimates	“procedure computes one-sided p-values for the residual variance and for covariance parameters with a lower bound of 0. The procedure computes two-sided p-values otherwise. These statistics constitute Wald tests of the covariance parameters, and they are valid only asymptotically.”
Type 3 Tests of Fixed Effects	The "Type 3 Tests of Fixed Effects" table contains hypothesis tests for the significance of each of the fixed effects—that is, those effects you specify in the MODEL statement.

Source: SAS

Additional information

- Type refers to the covariance(just an unstandardized version of a correlation) structure
  - You can use the sphericity test to tell you which is most appropriate
  - VC=standard variance-default
  - AR=autoregressive – as measures get further apart they get less correlated
  - CS=Compound symmetry-the correlation is the same regardless of how far apart the measurements are
  - UN=unstructured- this model allows every term to be different. It fits most models
  - AR=Toeplitz-measures that are next to each other have the same correlation

Source: SAS