Northwestern University STAT 356: Hierarchical Linear Modeling

Final Project

Elliott Evans

The following is an analysis of the Early Childhood Longitudinal Study with respect to mathematics achievement, using hierarchical linear models.

Elliott Evans
STAT 356
6/10/2015

Final Project

We proceed to use longitudinal growth models to examine the level 2 variables that may be risk factors for low rate mathematics achievement growth. We will also attempt to discover the risk factors most important to predicting mathematics achievement growth. The data analyzed is part of the Early Childhood Longitudinal Study (ECLS), a longitudinal study of a representative sample of children from Kindergarten who were followed through grade 8. Math achievement is measured by *MATH*, a mathematics achievement score. Firstly, we notice that a cubic two-level longitudinal growth model with varying intercept appears to be appropriate:

$$\underline{\text{Level 1}} : MATH_{ti} = \pi_{0i} + \pi_{1i}(TIME\_C_{ti}) + \pi_{2i}(TIME\_C\_2_{ti}) + \pi_{3i}(TIME\_C\_3_{ti}) + e_{ti}$$

$$\underline{\text{Level 2}} : \pi_{0i} = \beta_{00} + r_{0i}$$

$$\pi_{1i} = \beta_{10}$$

$$\pi_{2i} = \beta_{20}$$

$$\pi_{3i} = \beta_{30}$$

Here, level 1 represents the within-student level (i.e. measures of the same children at different grades) and level 2 represents the between-student level. A summary of the fixed and random effects is given in Table 1.

Because the P-values for all of the fixed effects are signifiant at the $\alpha = .05$ level, we assert that the quadratic and cubic terms are necessary to explain the growth trajectories in conjunction with the linear term. We notice several other things from this model, including the expected math achievement score ($\hat{\beta}_{00} \approx 137$) for a child at the mean grade (3.4). We also notice the linear, quadratic (acceleration), and cubic growth rates of 14.58, -1.684, and

Parameter	Estimate	Std. Error	t-value	P-Value
Fixed Effects				
β_{00}	109.55648	0.97279	112.62	$< 2 \times 10^{-16}$
β_{10}	14.57563	0.29887	48.77	$< 2 \times 10^{-16}$
β_{20}	-1.68430	0.05477	-30.75	$< 2 \times 10^{-16}$
β_{30}	0.05299	0.02189	2.42	0.0156
Parameter	Estimate	d.f.	χ^2	P-Value
Variance Components (Level 1, 2)				
$\sigma^2 = \text{Var}\{e_{ti}\}$	104.4	-	-	-
$\tau_{00} = \text{Var}(r_{0i})$	271.4	-	-	-

Table 1: Fixed and Random Effects for our first model

0.053, respectively, of math achievement. The average growth trajectory is given in Fig. 1.

The model also estimates that the residual within-student variance of reading scores after controlling for growth rates is approximately $\hat{\sigma}^2 \approx 104.4$, or a standard deviation of about 10.22 points. The model also suggests that the variance of math scores between students at the mean grade is approximately 271.4, or a standard deviation of 16.47 points. We see clearly and unsurprisingly that achievement in math increases with grade level.

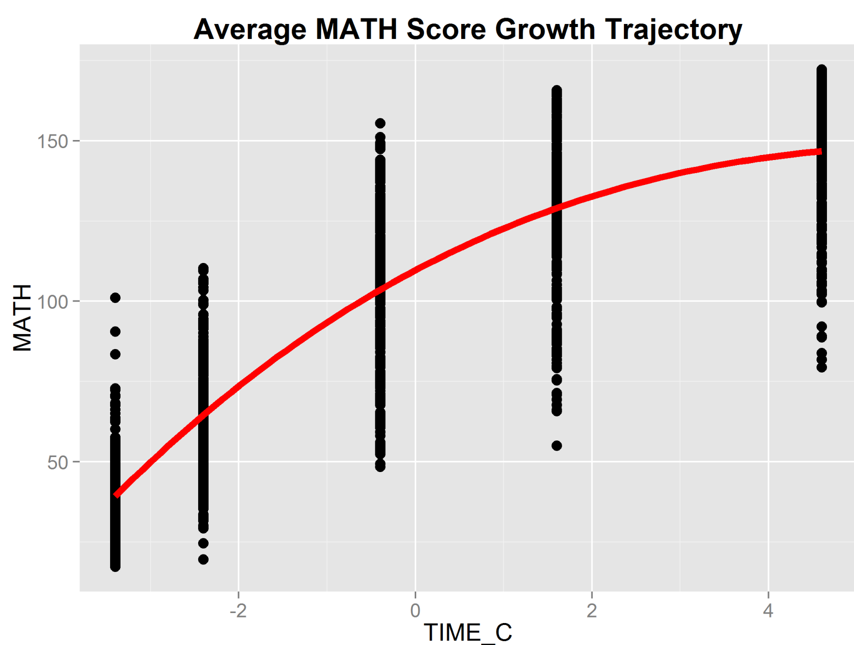

Figure 1: Includes linear, quadratic, and cubic terms.

Because it has been shown that SES is an important variable in predicting achievement in school, we will add SES as a time varying level 1 covariate to the model above. The new model is as follows:

$$\underline{\text{Level 1}} : MATH_{ti} = \pi_{0i} + \pi_{1i}(SES\_CONT_{ti}) +$$
$$\pi_{2i}(TIME\_C_{ti}) + \pi_{3i}(TIME\_C\_2_{ti}) + \pi_{4i}(TIME\_C\_3_{ti}) + e_{ti}$$
$$\underline{\text{Level 2}} : \pi_{0i} = \beta_{00} + r_{0i}$$
$$\pi_{1i} = \beta_{10}$$
$$\pi_{2i} = \beta_{20}$$
$$\pi_{3i} = \beta_{30}$$
$$\pi_{4i} = \beta_{40}$$

The new fixed and random effects from this model are given in Table 2.

Parameter	Estimate	Std. Error	t-Value	P-Value
Fixed Effects				
β_{00}	109.55648	0.97279	112.62	$< 2 \times 10^{-16}$
β_{10}	7.25485	0.85632	8.472	$< 2 \times 10^{-16}$
β_{20}	14.59491	0.30073	48.532	$< 2 \times 10^{-16}$
β_{30}	-1.69976	0.05515	-30.823	$< 2 \times 10^{-16}$
β_{40}	0.05995	0.02204	2.719	0.00663
Parameter	Estimate	d.f.	χ^2	P-Value
Variance Components (Level 1, 2)				
$\sigma^2 = \text{Var}\{e_{ti}\}$	105.7	-	-	-
$\tau_{00} = \text{Var}(r_{0i})$	212.3	-	-	-

Table 2: Fixed and random effects for the second model

We find that all of the fixed effects, including the coefficient of SES_CONT, are significant at the $\alpha = .05$ level, giving us reasonable evidence that time-varying SES is needed to partially explain math achievement growth trajectory. We see that as time-varying SES increases, math achievement also tends to increase. The fixed effects from the first model do not appear to have changed drastically, however.

We now begin to determine what level 2 variables may be risk factors for low rate mathematics achievement growth. We proceed to add level 2 covariates separately into the model. For instance, first we add *FEMALE* to the model above. Then we remove it. Then we add *RACE* to the model above. We proceed to do this with every level 2 covariate. Centering the level 2 covariates does not drastically change the coefficients and P-values observed, so we will add the uncentered variables to our model. Thus, each model we build will have the form

$$\begin{aligned}
&\underline{\text{Level 1}} : MATH_{ti} = \pi_{0i} + \pi_{1i}(SES_CONT_{ti}) + \\
&\qquad \pi_{2i}(TIME_C_{ti}) + \pi_{3i}(TIME_C_2_{ti}) + \pi_{4i}(TIME_C_3_{ti}) + e_{ti} \\
&\underline{\text{Level 2}} : \pi_{0i} = \beta_{00} + \beta_{01}(NEW_COVARIATE_{i}) + r_{0i} \\
&\qquad \pi_{1i} = \beta_{10} \\
&\qquad \pi_{2i} = \beta_{20} \\
&\qquad \pi_{3i} = \beta_{30} \\
&\qquad \pi_{4i} = \beta_{40}
\end{aligned}$$

Since we have so many categorical variables, however, we will only consider our new covariates that are significant at the $\alpha = .01$ level of significance. Note that dummy coding has been used for many of the variables, including *RACE*, *WKMOMED*, *WKDADED*, *C1SCREEN*, *P1HSEVER*, *P1FIRKDG*, and *WKHMOMAR*. The reference values for these dummy codings are “White, Non-Hispanic”, “High-School Diploma/Equiv.”, “High-School Diploma/Equiv.”, “Speak English at home”, “Child never in head start”, “Not a first-time kindergartener”, and “Not married at time of birth”, respectively. Because there were no instances of “never” in *WKHMOMAR*, it is treated as a dichotomous variable.

When we complete this process, we find that eight variables have significant P-values:

- *DMOMED2*: Value of 1 if the mother's education level is 9th-12th grade, and 0 otherwise (P-value = 0.00125, $\hat{\beta}_{01} = -13.02773$).

- *DMOMED5*: Value of 1 if the mother's education is Bachelor's Degree level, and 0 otherwise (P-value = 0.00312, $\hat{\beta}_{01} = 6.07317$).
- *DDADED7*: Value of 1 if the father's education is Master's Degree level, and 0 otherwise (P-value = 0.00369, $\hat{\beta}_{01} = 10.12380$).
- *DRACE1*: Value of 1 if the student is Black and Non-Hispanic, 0 otherwise (P-value = 0.00726, $\hat{\beta}_{01} = -9.55675$).
- *DRACE3*: Value of 1 if the student is Hispanic, race not specified (P-value = 0.000954, $\hat{\beta}_{01} = -9.93003$).
- *WKINCOME*: 0 to 1,000,000 (P-Value = 0.00732, $\hat{\beta}_{01} = 4.465 \times 10^{-5}$).
- *C1SCREEN*: Value of 1 if a non-English language is spoken at home, 0 otherwise (P-Value = 0.00161, $\hat{\beta}_{01} = -8.04401$).
- *P1HSEVER*: Value of 1 if child was ever in the Head Start Program, 0 otherwise (P-value = 0.000133, $\hat{\beta}_{01} = 10.73666$).

We notice several interesting things from this process. A mother's education level being only 9th-12th grade gives her child a disadvantage at the mean grade relative to students whose mothers had a high school diploma. These students start behind other students with more educated mothers by about 13 points in math achievement. Students whose mothers had a BA found themselves with an advantage of about 6 points at the mean grade and students whose fathers had a Master's level education had a 10 point advantage relative to students of high school-educated fathers.

In terms of race, Black students started behind White students at the mean grade at an estimated 10 points. Students whose race was Hispanic and not specified also had a disadvantage of about 10 points relative to white students.

Students with higher family incomes tended to have higher math achievement scores, as the positive coefficient of *WKINCOME* reveals. Students living in a home where a non-

English language was spoken started behind students living in a home where English was spoken at the mean grade by about 8 points. Finally, children who were in the Head Start Program had an advantage over students never in the program by almost 11 points.

A two-level cubic longitudinal growth model including only some of these variables may be both simpler and more accurate for prediction. We illustrate this when we create a model using all of these variables:

$$\underline{\text{Level 1}} : MATH_{ti} = \pi_{0i} + \pi_{1i}(SES\_CONT_{ti}) +$$
$$\pi_{2i}(TIME\_C_{ti}) + \pi_{3i}(TIME\_C\_2_{ti}) + \pi_{4i}(TIME\_C\_3_{ti}) + e_{ti}$$
$$\underline{\text{Level 2}} : \pi_{0i} = \beta_{00} + \beta_{01}(DMOMED2_i) + \beta_{02}(DMOMED5_i) + \beta_{03}(DDADED7_i) +$$
$$\beta_{04}(DRACE1_i) + \beta_{05}(DRACE3_i) + \beta_{06}(WKINCOME_i) + \beta_{07}(C1SCREEN_i) +$$
$$\beta_{08}(P1HSEVER_i) + r_{0i}$$
$$\pi_{1i} = \beta_{10}$$
$$\pi_{2i} = \beta_{20}$$
$$\pi_{3i} = \beta_{30}$$
$$\pi_{4i} = \beta_{40}$$

The fixed and random effect estimates are given in Table 3. From the estimates in Table 3, we notice that several of the coefficients of our level 2 covariates are no longer significant, even at the $\alpha = .05$ level. If we abandon these covariates from our model, we remain with *DMOMED2*, *DDADED7*, *DRACE1*, *DRACE3*, and *P1HSEVER*, along with our time variables and the level 1 *SES_CONT*. This will be our final model, made up of only the most important level 2 covariates.

Parameter	Estimate	Std. Error	t-Value	P-Value
Fixed Effects				
β_{00}	96.11	5.684	16.909	$< 2 \times 10^{-16}$
β_{01}	-8.749	4.073	-2.148	0.032407
β_{02}	3.913	2.016	1.941	0.053069
β_{03}	10.27	3.335	3.080	0.002235
β_{04}	-8.654	3.613	-2.395	0.017144
β_{05}	-6.527	3.240	-2.014	0.044768
β_{06}	2.379×10^{-5}	1.638×10^{-5}	1.453	0.147006
β_{07}	-4.699	2.703	-1.738	0.083056
β_{08}	6.138	2.883	2.129	0.033945
β_{10}	3.551	1.029	3.452	0.000576
β_{20}	14.59	0.2991	48.770	$< 2 \times 10^{-16}$
β_{30}	-1.694	.05485	-30.886	$< 2 \times 10^{-16}$
β_{40}	.05705	.02193	2.602	0.009378
Parameter	Estimate	d.f.	χ^2	P-Value
Variance Components (Level 1, 2)				
$\sigma^2 = \text{Var}\{e_{ti}\}$	104.6	-	-	-
$\tau_{00} = \text{Var}(r_{0i})$	191.5	-	-	-

Table 3: Fixed and random effects for the third model

Our final model is as follows:

$$\underline{\text{Level 1}} : MATH_{ti} = \pi_{0i} + \pi_{1i}(SES\_CONT_{ti}) +$$

$$\pi_{2i}(TIME\_C_{ti}) + \pi_{3i}(TIME\_C\_2_{ti}) + \pi_{4i}(TIME\_C\_3_{ti}) + e_{ti}$$

$$\underline{\text{Level 2}} : \pi_{0i} = \beta_{00} + \beta_{01}(DMOMED2_i) + \beta_{02}(DDADED7_i) +$$

$$\beta_{03}(DRACE1_i) + \beta_{04}(DRACE3_i) + \beta_{05}(P1HSEVER_i) + r_{0i}$$

$$\pi_{1i} = \beta_{10}$$

$$\pi_{2i} = \beta_{20}$$

$$\pi_{3i} = \beta_{30}$$

$$\pi_{4i} = \beta_{40}$$

A summary of the fixed and random effects is given in Table 4. All of the fixed effects are significant at the $\alpha = .05$ level and most of them at the $\alpha = .01$ level. Using this model, we may target the risk factors of low rate mathematics achievement growth. Firstly, two risk factors appear to be the educations associated with both the mother and the father. From this model, a student whose mother has just a 9th-12th grade education finds himself at an approximate 9 point disadvantage at the mean grade, relative to students whose mothers have just a high school diploma. Alternatively, students whose fathers have a Master's level education start with 10 point advantage at the mean grade.

Parameter	Estimate	Std. Error	t-Value	P-Value
Fixed Effects				
β_{00}	94.76584	5.53014	17.136	$< 2 \times 10^{-16}$
β_{01}	-9.04709	4.09605	-2.209	0.02784
β_{02}	10.21619	3.35527	3.045	0.00250
β_{03}	-8.47931	3.58980	-2.362	0.01873
β_{04}	-9.14332	3.03463	-3.013	0.00277
β_{05}	7.82715	2.83683	2.759	0.00610
β_{10}	4.81119	0.91921	5.234	2.07×10^{-7}
β_{20}	14.59065	0.29945	48.724	$< 2 \times 10^{-16}$
β_{30}	-1.69567	0.05492	-30.878	$< 2 \times 10^{-16}$
β_{40}	0.05793	0.02195	2.639	0.00841
Parameter	Estimate	d.f.	χ^2	P-Value
Variance Components (Level 1, 2)				
$\sigma^2 = \text{Var}\{e_{ti}\}$	104.8	-	-	-
$\tau_{00} = \text{Var}(r_{0i})$	194.2	-	-	-

Table 4: Fixed and random effects for the final model

Black students faced an approximate 8 point disadvantage at the mean grade relative to White students. Similarly, Hispanic students of a race not specified faced a 9 point disadvantage. Finally, students who were at some point in the Head Start program had an estimated 8 point advantage at the mean grade relative to students who had never been in the Head Start Program.

These risk factors are notable, considering the model's estimated standard deviation of math achievement scores between students at the mean grade is approximately 13.93. Thus, a child who has a father with a Master's Degree is already a more than half a standard deviation ahead of a student whose father has just a high school diploma and a black student more than half a standard deviation below a white student. The analysis of this data and the construction of these models reveal that risk factors for low rate of mathematics achievement growth are mainly race (in particular Black, Hispanic, or White), education of the parents, and the child's participation in the Head Start Program.

$*$ Note that the d.f., t-Value, and P-Value are not included for $\hat{\sigma}^2$ and $\hat{\tau}_{00}$ for any of the tables because they are not included in the output for `R`, the statistical software used for this project.