Practical 3: The random intercept model with covariates

Data

Recall that `High School & Beyond` is a nationally representative survey of U.S. public and Catholic high schools conducted by the National Center for Education Statistics (NCES). The data are a sub-sample of a survey conducted in 1982 and involving 7,185 students from 160 schools and are described in the book *Hierarchical Linear Models* by Raudenbush and Bryk.

1. The `High-School-and-Beyond data`.

   Data are held in `hsb_selected.dta`. The variables are:

   ```
   minority       Indicator of student ethnicity (1=minority, 0=other)
   female         Indicator of student being female
   ses            Standardized Socio-Economic Status score
   mathach        Measure of mathematics achievement
   size           School's total number of students
   sector         School's sector: 1=Catholic, 0=not Catholic
   schoolid       School identifier
   ```

Questions

1. Load and familiarize yourself again with the High-School-and-Beyond data. Create the indicator variable that picks up only one record per school:

   ```
   . egen pickone=tag(schoolid)
   ```

2. Fit a random intercept model for `mathach` on `ses` using both ML and REML. Are the results different in any way? Enter the results for $\hat{\sigma}_u$ and $\hat{\sigma}_e$ in the table on the next page.

3. Examine the distribution of the variable `sector`. Include it in the model as an explanatory variable: what happens to $\hat{\sigma}_u$ and $\hat{\sigma}_e$? Does it make a difference if you use ML or REML?

4. Now add `size` to the model. Consider centering it and dividing it by 100 to facilitate interpretation (i.e. 'per 100 extra students above the mean'). What happens to $\hat{\sigma}_u$ and $\hat{\sigma}_e$?

5. Assess whether the effect of `size` interacts with `sector`. What test will you use?

6. Summarize the results concerning $\hat{\sigma}_u$ and $\hat{\sigma}_e$ in the table below. Does $\hat{\sigma}_e$ change? Why? Is there a difference between the ML and REML results?

		REML		**ML**	
	Model with	$\hat{\sigma}_u$	$\hat{\sigma}_e$	$\hat{\sigma}_u$	$\hat{\sigma}_e$
1	`ses`				
2	`ses` & `sector`				
3	`ses`, `size` & `sector`				
4	`ses`, `size` & `sector` & interaction `size-sector`				

7. Now add `female` and `minority` to the model for the intercept and use REML to fit it. What happens now to $\hat{\sigma}_u$ and $\hat{\sigma}_e$? Why?

8. Examine whether `female` and `minority` are associated with the school level variable `sector` using:

```
. tab sector minority, chi col nokey
. tab sector female, chi col nokey
```

   Then assess whether the effect of `female` and `minority` on the school intercepts is modified by the school's sector. Make a note of the estimated σ_u and σ_e.

9. Predict the school level and the pupil level EB residuals corresponding to the last fitted model with:

```
. predict uhat_eb, reffects reses(uhat_eb_se)
. predict ehat, rstandard
```

   Then standardize the level 2 residuals using the marginal variance $R_j\hat{\sigma}_u^2$ and check the distribution of both residuals.

10. Identify the schools with the more extreme u.

The next two questions are optional

11. Generate the school level mean SES and also for each pupil his/her difference in SES relative to the school mean with:

```
. egen mean_ses=mean(ses),by(schoolid)
. gen dif_ses=ses-mean_ses
```

   Refit the model with only SES as the explanatory variable for the intercept and then the model with only `mean_ses` and `dif_ses`. Compare their results and interpret the regression coefficients.

12. Test whether the second specification of the model for the effect of SES gives a better fit to the data. Use the Wald test with the command:

```
. test mean_ses==dif_ses
```