**Regression with Clustered Data Using Proc Genmod**

OLS regression assumes that the residuals are independent. The **elemapi2** dataset contains data on 400 schools that come from 37 school districts. It is very possible that the scores within each school district may not be independent, and this could lead to residuals that are not independent within districts. SAS **proc genmod** is used to model correlated data.  We can use the **class** statement and the **repeated** statement to indicate that the observations are clustered into districts (based on **dnum**) and that the observations may be correlated within districts, but would be independent between districts.

First of all, the data are in a funny format. It is “sasb7dat” format. In order to import these data, you need to do the following:

* Create a folder to hold the data. We may have more of this type data in the future
* Download the data file into that folder
* Create a libname in SAS using the following commands – you will need the correct pathname to your file

libname DataLib '\\Client\H$\MSDS 6372\Data';

**proc** **print** data=DataLib.elemapi2; **run**;

Once the data are in SAS, you will need to call it with any procedures you use. In other words, you will always need the “data” option.

First, let’s run this in OLS just to see what happens.

**proc** **reg** data=DataLib.elemapi2;

model api00 = acs\_k3 acs\_46 full enroll /clb stb ;

**run**; **quit**;

SAS didn’t blow up even though you ran correlated regression data. This underscores the point that the analyst needs to be smarter than the computer. Here we go with PROC GENMOD.

**proc** **genmod** data=DataLib.elemapi2;

class dnum;

model api00 = acs\_k3 acs\_46 full enroll ;

repeated subject=dnum / type=ind ;

**run**; **quit**;

The estimates of the coefficients are the same as the OLS estimates, but the standard errors take into account that the observations within districts are non-independent.  Even though the standard errors are larger in this analysis, the three variables that were significant in the OLS analysis are significant in this analysis as well.

If you examined the residuals from Proc Reg, you probably saw some evidence of nonconstant variance. Do we still see that effect when we account for the correlation structure? To find out, use the output statement to output some residual statistics. The syntax is

output out=resids STDRESDEV=res DCLS=DCLS;

The output statement belongs after the repeated statement in proc genmod. This statement outputs the standardized residuals and the clustered cook’s distance to a file called “resids”. You need to run proc univariate or proc gplot on the resids data file to examine the residuals.

After plotting the residuals from proc genod, make any transformation you think are necessary and rerun proc genmod. If none are necessary, then you’re done!