Practical 1: Simple treatment of dependent data

Data

If you have not already done so, please copy **all the data and programs needed for the course**. These are in **U:\Download\Teach\MedSt_Hier**. **Please copy them all onto a dedicated folder in your own workspace. Remember to make sure that you set the working directory of Stata to be that folder so that you can easily access data and ado programs.**

In this session, we are going to use two datasets already introduced in the lectures:

1. The `PEFR data` includes measures of peak-expiratory-flow rate (PEFR) taken using two instruments on 17 people in an experiment reported by Bland and Altman (*Lancet* I, 1986, 307-310). The two instruments were the Standard Wright and the Mini Wright peak flow meter. Each method was used twice for each person so we have clustered data where clusters are individual people. The variables are:

`id`	`Participant identifier`
`wp1`	`Standard Wright measure at 1st occasion`
`wp2`	`Standard Wright measure at 2nd occasion`
`wm1`	`Mini Wright measure at 1st occasion`
`wm2`	`Mini Wright measure at 2nd occasion`

2. The `High-School-and-Beyond data` is from 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, involving 7,185 students from 160 schools and are described in the book *Hierarchical Linear Models* by Raudenbush and Bryk.

   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. Read in and familiarise yourself with the PEFR dataset introduced in the lecture. This time we are going to analyse the Standard Wright measures. Load and examine the data. In which format are they (wide or long)?

2. Plot the two Standard Wright measures against the subject identifier with:

   ```
   . twoway (scatter wp1 id,ms(circle))(scatter wp2 id,ms(circle_hollow)), ///
   xtitle(Subject id) xlabel(1/17) ytitle(W Measurements) ///
   legend( order(1 "Occasion 1 " 2 "Occasion 2")) yline( 447.8824)
   ```

   Compare this plot with that for the Mini Wright measures that was reported in the lecture.

3. Generate the subject specific means and examine their distribution. Then reshape the data in long format (remember to check that the reshaping is not distorting the data):

   ```
   . drop wm*
   . reshape long  wp, i(id) j(occasion)
   ```

4. Carry out an ANOVA for the individual variations in `wp` measures using `loneway`. What does the value of 'SD within effect' mean? Check that the value of 'SD of id effect' reported by Stata is equal to $\sqrt{\frac{\text{(MMS-MSE)}}{n}}$ (this will be discussed more fully in the next lecture), where n is the number of repeats for each subject. (Note that the observed variance of the estimated fixed effects is 117.47).

5. Fit the equivalent regression model, including a constant with:

   ```
   . regress wp i.id
   ```

   Write down the model algebraically. Describe what the coefficients for the individual dummies represent.

6. Refit the model but now replace the dependent variable `wp` with its difference from the overall mean and remove the constant:

   ```
   . summ wp
   . gen c_wp=wp-r(mean)
   . regress c_wp ibn.id,nocon
   ```

   Write down this last model algebraically. Compare these results with those obtained in the previous two questions.

7. Now load and familiarize yourself with the High-School-and-Beyond data. In particular examine the school level characteristics after creating an indicator variable that picks up only one record per school. For example try:

   ```
   . sort schoolid
   . egen pickone=tag(schoolid)
   . tab sector if pickone
   . summ size if pickone
   ```

8. For simplicity in this session we are going to work on just the first 5 schools and their 188 pupils. Select them as shown below and also generate the mean maths and mean SES score for each school:

```
. keep if schoolid<1320
. egen mean_ses=mean(ses),by(schoolid)
. egen mean_math=mean(mathach),by(schoolid)
```

9. We are interested in the relationship between Maths scores and SES. Start by examining the scatter of these two variables. Then fit an overall (ie Total) regression model and save the predicted values:

```
. scatter mathach ses
. regress  mathach ses
. predict pred_T, xb
```

10. Now fit the between regression model and save the predicted values with:

```
. regress mean_math mean_ses if pickone
. predict pred_B if pickone
```

11. Finally fit the within regression models, save the results in a file called `ols.dta` and then merge it to the main dataset with:

```
. statsby inter=_b[_cons] slope=_b[ses] seslope=_se[ses], by(schoolid) ///
       saving(ols,replace):  regress mathach ses
. sort schoolid
. merge m:1 schoolid using ols
```

    Check that the merge has been successful. Generate the Maths scores predicted by these separate regression models with:

```
. generate pred_W = inter + slope*ses
```

12. Compare the results from these three models. Are they consistent? For what reasons do they differ?

13. Plot all the predictions on the same plot:

```
. sort school ses
. twoway (line pred_W ses, connect(ascending) lcol(green) lpat(dash)) ///
   (line pred_T ses, connect(ascending) lcol(blue) lw(medium) ) ///
   (line pred_B mean_ses, connect(ascending) sort lcol(red) lw(thick)) ///
   (scatter mean_math mean_ses, msym(T) mcol(red)), ///
   xtitle(SES) ///
   ytitle(Fitted regression lines) ///
   legend(order(1 "Within" 2 "Total" 3 "Between" 4 "School Mean"))
```

14. Fit a fixed effect model on these data:

```
. regress mathach ses i.schoolid
```

    What do you conclude?