Regression II
Vanderbilt University
Prof. Sean P. Corcoran

October 17, 2022

Lecture 4 In-Class Examples: Panel Commands and Fixed Effects

---

Example 1. This example provides a simple illustration of `reshape`-ing panel data. The dataset used here is called *Census_states_1970_2000.dta* and is found on Github.

```
use https://github.com/spcorcor18/LPO-8852/raw/main/data/Census_states_1970_2000.dta
```

(a) This dataset includes unemployment rates and median household income by state for the Census years 1970, 1980, and 2000. What is the cross-sectional (panel) unit? What is the time variable? Is the panel balanced? Which variables are time invariant and which change over time?

(b) `reshape` this data to wide format. Confirm that the reshape worked correctly using the `browse` command.

(c) Now, return the data to long format using the `reshape` command. Confirm that the reshape worked correctly using the `browse` command.

Example 2. This example provides an illustration of panel data descriptives. The dataset used here is called *State_school_finance_panel_1990_2010.dta* and is found on Github. These data include annual observations on K-12 school revenues and expenditures by state for the years 1990 through 2010.

```
use https://github.com/spcorcor18/LPO-8852/raw/main/data/State_school_finance_panel_1990_2010.dta
```

(a) Use `xtset` to declare the data as a panel. What is the cross-sectional (panel) unit? What is the time variable? Is the panel balanced? Remember that cross-sectional unit identifiers must be numeric for `xtset`. Use `encode` to create a new numeric cross-sectional identifier. (Note that `encode` will automatically assign value labels for you).

(b) Use `xtdescribe` to see the patterns of data availability in the panel. (This command will be more useful in a later example). Use `summarize` and `xtsum` to get some descriptive statistics on real per-pupil current school expenditure. (Current expenditure on schools is all expenditure other than "capital", such as construction. The expenditures are in real terms because they have been adjusted for inflation, with 2009 as the base year).

(c) Create a dummy variable that is equal to one for observations in the Mid-Atlantic region (states DC, DE, MD, NJ, NY, PA, VA, and WV), and zero otherwise. Do a two-way cross-tabulation of *year* and this new variable to ensure you have 8 Mid-Atlantic states in each year.

(d) Use `xtline` to see the trend in per-pupil current expenditure between 1990 and 2010 for states in the Mid-Atlantic region. Do this plot again but using the `overlay` option.

(e) Use `egen` to create a new variable that contains the mean per-pupil expenditure by state. (Recall that there are 21 observations per state).

(f) Get a visual picture of the within-state over-time variation in per-pupil expenditure, using a scatter plot of real per-pupil expenditure against the state number (the numeric state number you created in (a)). Do this a second time, but add the following options to make the graph more useful: `xlabel(1(1)51, valuelabel alternate) xsize(10) ysize(4)`

```
twoway scatter expp_co2 state2
twoway scatter expp_co2 state2, xlabel(1(1)51, valuelabel alternate) xsize(10) ysize(4)
```

(g) Create a bar graph of that shows the mean per-pupil expenditure (that you created above) by state (you can use the string variable *state*), and display the bars in descending order by mean per-pupil expenditure. Hint: for the bar graph, graph the actual values (`asis`) for only one year that you designate. The specific year does not matter, since the variable created is the same in all years within a state.

```
graph bar (asis) meanexp if year==1990, over(state, sort(meanexp) descending) xsize(10) ysize(4)
```

Example 3. This example illustrates fixed effects regression models using several different approaches. The dataset here is called *Texas_elementary_panel_2004_2007.dta* and is found on Github. This dataset consists of academic performance and other characteristics of all elementary schools in Texas between 2004 and 2007.

```
use https://github.com/spcorcor18/LPO-8852/raw/main/data/Texas_elementary_panel_2004_2007.dta
```

(a) Use `xtset` to declare the data as a panel. The *campus* (school) is the cross-sectional unit while *year* is the time variable. Use `xtdescribe` to see the patterns of participation in the panel. Is the panel balanced? (It often pays to use `duplicates report` to check the panel and ensure there is only one observation per cross-sectional unit and time period.

(b) The variable *ca311tar* is the average passing rate in a school-year across all state tests. Rename this variable *avgpassing*. Use `xtsum` to get some descriptive statistics on this average passing rate variable. Does there appear to be more variation in passing rates between schools, or within schools over time?

(c) The variables *cpctg01a* – *cpctgmea* are average class sizes for various grades in the school. Use `egen` with the `rowmean` function to create an *avgclass* variable that contains the average of these for each school and year.

(d) Estimate an OLS regression of *avgpassing* on *avgclass* for 2007 only (a cross-sectional regression). How is class size related to achievement on the Texas state tests in 2007? Does this relationship seem sensible to you?

(e) Now estimate an OLS regression of the first difference of *avgpassing* on the first difference of *avgclass*, again for 2007 only (the first difference will use 2006 as the lag year). How does this change your estimate of the effect of class size on achievement?

(f) How much within-school variation is there in *avgpassing* and *avgclass*? Use the `histogram` and `summarize` command to look at variation in the first difference of these variables.

(g) Next estimate an OLS regression using first differences, as in (e), but without restricting the sample to 2007. How do the results compare? Create a table of *year* using the observations used in the regression (`if e(sample)`) to see how many schools from each year were used in the estimation.

(h) Estimate an OLS least squares dummy variable (LSDV) regression of *avgpassing* on *avgclass*, but only for schools in the Houston Independent School District (***houston*==1**). We make this restriction here because there are a very large number of schools in Texas, and this limits the number of unique school dummy variables. Do this first by including dummy variables directly in the regression model (using Stata's factor variable notation), and then again using `areg`. How does the coefficient on class size compare to the ones in earlier steps? How does the coefficient compare between `reg` and `areg` here?

(i) To see how the coefficients on school effects should be interpreted, re-estimate the model using only the 8 largest schools in Houston (see sample do file for one approach). Get the predicted values after the LSDV regression and see if you can reconstruct them using the OLS regression output and observed values of *avgclass*.

(j) Estimate a "within" regression using `xtreg, fe`. Do this for both the Houston 2006+ subsample and for all schools. How do the coefficients on class size compare to the ones in earlier steps?

(k) Obtain the estimated fixed effects using `predict` following `xtreg`. Inspect the results estimates and interpret.

(l) Finally, estimate a two-way fixed effects version of the model in (j) using all schools and all years by including dummy variables for the years. (Use the Stata factor variable notation for the year effects). Alternatively, use the `reghdfe` command to do the same thing.