Panel Data Structure and Useful Stata Commands

Panel Data Structure

- Panel data = longitudinal data.
- There are multiple/repeated observations for the same units (= individuals, states, occupation, industry, family...) over time.
- If we say the unit of observation is "individual-year", it means the same individuals are observed repeatedly over years. In this case, we have one record per year per individual. If the unit of observation is "country-month", it means there are multiple observations for countries over time. The interval of the observation is a month.
- Yet for another example, "occupation-year" can be a unit of observation. Occupational mean wage can be a dependent variable and the % of the highly educated, % female, % mid-west... in each occupation can be independent variables. Thus we can measure as % female changes over time, how much income changes over time across occupations on average.
- "individual", "country", or "occupation" in the previous examples is called "group" in textbooks and statistical programs. It can be confusing, but here 'group' does not mean a group of individuals, countries or occupations. It is called 'group' because there are multiple observations over time for the same individuals, countries or occupations.
- i subscript is usually used to identify the 'group', and j or t subscript is usually used to identify the time dimension. Thus, Y_{ij} or Y_{it} refers to the dependent variable Y for group i at time j or t.
- The main goal of the panel methods is the same as OLS. One of the main advantages of panel models over OLS is that we can estimate the expected change in y over time when x's change over time with panel models.
- Typical data structure looks like this:

id	year	y	x1	x2	x3
1	2010	10.0	4	25	0
1	2011	10.2	4	26	0
1	2012	10.5	4	27	1
2	2010	9.1	2	30	1
2	2011	9.2	2	31	1
2	2012	9.0	2	32	0
2	2013	9.3	2	33	1
3	2010	10.5	5	28	0
3	2011	10.6	5	29	1
3	2012	10.7	5	30	1

- Here individual i (=id) is observed over years.

Transpose from "wide" format to "long" format

- To apply panel models, your data should be in the format of the previous table. Each row should be one observation for each group each time. This format of data is called "long" format.
- Sometimes, your data are not arranged in that format. Instead it looks somethng like the below. This is called "wide" format. Here `famid` is i. There are three observations for each `famid`.
- Example 1: wide format

```
        famid   faminc96   faminc97   faminc98
1.          3      75000      76000      77000
2.          1      40000      40500      41000
3.          2      45000      45400      45800
```

- You need to reshape the data to "long" if your data format is "wide." Use the Stata command, `reshape`.
- Example 1: long format

```
. reshape long faminc, i(famid) j(year)

        famid       year     faminc
1.          1         96      40000
2.          1         97      40500
3.          1         98      41000
4.          2         96      45000
5.          2         97      45400
6.          2         98      45800
7.          3         96      75000
8.          3         97      76000
9.          3         98      77000
```

- In the 1st example, the unit of observation for the long format is "family-year". The unit of observation for the wide format is family.
- Example 2: wide format

```
         famid      birth        ht1        ht2
1.          1          1        2.8        3.4
2.          1          2        2.9        3.8
3.          1          3        2.2        2.9
4.          2          1          2        3.2
5.          2          2        1.8        2.8
6.          2          3        1.9        2.4
7.          3          1        2.2        3.3
8.          3          2        2.3        3.4
9.          3          3        2.1        2.9
```

- Example 2: long format

```
. reshape long ht, i(famid birth) j(age)

        famid      birth        age         ht
1.          1          1          1        2.8
```

```
  2.        1          1          2        3.4
  3.        1          2          1        2.9
  4.        1          2          2        3.8
  5.        1          3          1        2.2
  6.        1          3          2        2.9
  7.        2          1          1          2
  8.        2          1          2        3.2
  9.        2          2          1        1.8
 10.        2          2          2        2.8
 11.        2          3          1        1.9
 12.        2          3          2        2.4
 13.        3          1          1        2.2
 14.        3          1          2        3.3
 15.        3          2          1        2.3
 16.        3          2          2        3.4
 17.        3          3          1        2.1
 18.        3          3          2        2.9
```

- In the 2nd example, the unit of observation for the long format is "each birth within family (individual)-age". The unit of observation for the wide format is individual (within family).
- Using the `reshape` command, you can convert the long format into the wide format.

Types of Panel Data

1. You can use the panel data. Two most popular panel datasets in the US are NLSY and PSID. Check https://www.bls.gov/nls/home.htm
2. You can convert pooled (individual-levle) cross-sectional data into group-level panel data. For example, you can compute state-level statistics for each year (e.g., mean personal income by state, % female, % highly educated, % married...) using CPS/ACS, and append multiple years' data into one data-set. Then, it becomes state-year panel data. Stata's `collapse` command is very useful for this.
3. You can collect summary statistics by country (or by other units) over time from web or an organization's (e.g., OECD) portal and then combine them to make your own panel data. Political scientists and macro-economists are doing this often.

`collapse` command

- `collapse (mean) meanwage=wage age female year (sd) sdwage=wage (percent) immig, by(state)`
- The above command will create new data which have 6 variables: meanwage, mean age, % female, year, standard deviation of wage, and % immigrant. The unit of observation is state.
- You do this multiple times for each year.
- And then append all data.
- For example,

```
use ACS2000.dta, clear
collapse (mean) meanwage=wage age female year (sd) sdwage=wage (percent) immig, by(state)
save newdata2000.dta

use ACS2001.dta, clear
collapse (mean) meanwage=wage age female year (sd) sdwage=wage (percent) immig, by(state)
save newdata2001.dta

use newdata2000.dta, clear
append using newdata2001.dta
save newdataall.dta
```