Empirical Economics

Panel Data Regression

Andrew Proctor

andrew.proctor@phdstudent.hhs.se

October 5, 2018


STOCKHOLM SCHOOL OF ECONOMICS

Introduction

Stata Seminar Approach

The format of the Stata seminars for my portion of the class will be a little bit different.

The format of the Stata seminars will generally be:

1. A relatively short overview (<30 minutes) of any new Stata commands and recap / presentation of important econometric concepts.
2. A Stata exercise for the remainder of the seminar, in which you practice analyzing an empirical problem similar in nature to the upcoming assignment.

Today's Presentation

For today, I will be reviewing the following 4 things before you get started on the seminar exercise:

1. Recap panel regression methods and the variation they use.
2. Overview of panel methods in Stata.
3. Data preparation commands that might be useful for Assignment 3.
4. How to create graphs and output tables from Stata.

Econometric concepts

Panel estimators and the variation they use

- A key focus of panel data methods is understanding how different panel regression estimators are distinguished by their use of different types of methods.
- Today, we'll recap the following estimators and the variation they use:
  - Pooled OLS
  - Between estimator
  - Fixed effects estimator
  - Random effects estimator

Pooled OLS and the Between Estimator

- **Pooled OLS:** $y_{it} = \alpha + \beta x_{it} + v_{it}$
  - This is just normal OLS, using all variation across individuals over time and across individuals.
  - By default, doesn't treat observations of the same individual over time any differently than observations for a different individual.
- **"Between" Estimator**: $\bar{y}_i = \beta \bar{x}_i + \bar{v}_i$
  - OLS using only the time-averaged values for each individual as the explanatory variables.
  - Consequently, doesn't use variation in the explanatory variable for a given individual, only variation across individuals (only "between" variation).

Fixed and random effects estimators

Fixed and random effects estimators take the panel nature of the data into account by explicitly accounting for an individual effect in the error term:

$$y_{it} = \beta x_{it} + v_{it}, \quad v_{it} = a_i + u_{it}$$

Fixed and random effects take different approaches to addressing the unobserved individual effect in the error term.

Fixed effect estimator

With fixed effects, we explicitly control for the individual effect.

- Computationally, this is the same as differencing out the time-averages of the explanatory variables.

$$y_{it} - \bar{y}_i = \beta(x_{it} - \bar{x}_i) + a_i - a_i + u_{it} - \bar{u}_i$$

$$y_{it} - \bar{y}_i = \beta(x_{it} - \bar{x}_i) + u_{it} - \bar{u}_i$$

- We can also write the regression as:

$$\tilde{y}_{it} = \beta\tilde{x}_{it} + \tilde{u}_{it},$$

where $\sim$ indicates the demeaned variable, eg. $\tilde{x}_{it} = x_{it} - \bar{x}_i$.

Random effects estimator

If we suppose that $cov(\alpha_i, x_{it}) = 0$, then we don't need to worry about OVB by not controlling for α_i in the regression.

- But pooled OLS estimates are not fully efficient because they do not account for correlation in the error term implied by α_i.
  - Random effects transforms the variables of the regression using an estimate of this correlation, so that the resulting estimates will be efficient.

The resulting regression looks like:

$$y_{it} - \theta\bar{y}_i = \alpha(1-\theta) + \beta(x_{it} - \theta\bar{x}_i) + (v_{it} - \theta\bar{v}_i), \text{ where}$$

θ is a term measuring the relative strength of the correlation in the error term.

Within and Between Variation

To understand which of "between" and "within" variation each of these estimators uses, let's express overall variation as the sum of within and between variation.

Using the sample variance of a given variable w_{it}, we have:

$$\hat{\sigma}_w^2 = \frac{1}{N \cdot T} \sum_t \sum_i \overbrace{(w_{it} - \bar{w})^2}^{total\ \ variation} =$$

$$\frac{1}{N \cdot T} \sum_t \sum_i \underbrace{(w_{it} - \bar{w}_i)^2}_{within} + \frac{1}{N} \sum_i \underbrace{(\bar{w}_i - \bar{w})^2}_{between}$$

Estimators and Source of Variation

Pooled OLS: Since the explanatory variable is x_{it}:

- Within variation: $\sum\sum x_{it} - \bar{x}_i \neq 0$.
- Between variation: $\sum \bar{x}_i - \bar{x} \neq 0$

Between Estimator: The explanatory variable is $\bar{x}_i$.

- Within variation: $\sum\sum \bar{x}_i - \bar{x}_i = 0$.
- Between variation: $\sum \bar{x}_i - \bar{x} \neq 0$

Estimators and Source of Variation ctd

Fixed Effects: The explanatory variable is $\tilde{x}_{it} = x_{it} - \bar{x}_i$. Note time-demeaning the variable implies $\bar{\tilde{x}}_i = 0$ and $\bar{\tilde{x}} = 0$.

- Within variation: $\sum\sum \tilde{x}_{it} - \bar{\tilde{x}}_i = \sum\sum \tilde{x}_{it} \neq 0$.
- Between variation: $\sum \bar{\tilde{x}}_i - \bar{\tilde{x}} = 0 - 0 = 0$

Random Effects: The explanatory variable is $\check{x}_{it} = x_{it} - \theta\bar{x}_i$.

- Clearly, the behavior is an intermediate case of pooled OLS and fixed effects, depending on the strength of θ.
- If $\theta = 0$, random effects is the same as pooled OLS, while if $\theta = 1$, it is identical to the fixed effects estimator.

Panel Methods in Stata

Declaring the panel variables

Before you can perform a panel data regression in Stata, you must first declare the panel structure of the data.

To do so, use the **xtset** command. The syntax is:

```
xtset PANELVAR TIMEVAR
```

- Where PANELVAR is the name of the variable that corresponds to the identifier for individual / group variable for the dataset.
- And TIMEVAR is the name of the variable indicating time period.

For example, for a cross-country growth regression, you might write:

```
xtset country year
```

Panel regression in Stata

Once you've declared the panel structure, standard panel data regression estimators can be done using the **xtreg** command (except for pooled OLS, which uses **regress**).

The basic syntax of **xtreg** is the same as **regress**:

```
xtreg yvar xvar, OPTIONS
```

xtreg options

Key options include:

- What type of panel regression to run:
  - "be" - between
  - "re" - random effects
  - "fe" - fixed effects
- As always, I'd suggest using the "robust" option. In fixed and random effects estimators, "robust" computes cluster-robust errors by default.

```
xtreg testgrade stdyhours, fe robust
```

Data prep commands

Using gen to create a variable

The basic way of creating a variable in Stata is to use the **gen** command. You can then *modify* values of a variable with the **replace** command.

```
gen weeklypay = hourlypay * wklyhours
```

Modifying a variable with replace

To change the values of a variable, use the **replace** command.

- Often, you will want to conditionally replace values. Conditions can be attached to most commands in stata by stating ***if*** after the main command.

```
generate has_child = .
replace has_child = 0 if (children == "No children")
replace has_child = 1 if children > 0 ///
  & !missing(children)
```

Note When assigning values of 1, the command is restricted to run only for obs where the *children* value is *not missing*. This is because missing values are coded as the largest possible values.

Creating a variable with egen

While **gen** is sufficient for creating variables that are simple transformations of another variable value for the same observation, often you are interested in creating variables that summarize information ***across*** observations in the dataset.

- The way to do that in Stata is by using the Stata command **egen**.
  - **egen** uses helper functions to summarize the datas, including **min, max, mean, sd, count, total** (ie sum), and **rank**.

```
* Create aggregate variable
egen firmemployees = count(employindic), by(firm)
egen countymeaneduc = mean(educ), by(county)
egen totalchildren = total(household_children)
```

Renaming a variable

- When doing your analysis, it is typically convenient to rename variables so that they are more intuitive (and concise). The syntax for this is:

```
rename OLDVARNAME NEWVARNAME
```

For example:

```
* Rename Hispanic origin indicator variable
rename horigin hispanic
```

Labelling a variable

Another major formatting option is labelling a variable.

- When you create regression tables, graphs, etc, the variable label is then shown instead of the variable name.
  - This is very useful for creating reports (like your assignment!)
- To create a variable lable, use the syntax:

```
label variable varname "Your label here"
```

For example:

```
* Label hourly wage variable
label variable wage "Hourly wage (in 2017 USD)"
```

Changing the dataset with keep and drop

If you want to restrict the dataset (to certain variables or observations that meet certain conditions), use **keep** or **drop**.

- With **keep**, only observations that meet a given condition are kept, while **drop** removes observations meeting a given condition.
- If variables are specified after keep or drop, only those variables will be kept or drop respectively from the dataset.

```
* Keep only the following variables
keep wage educ experience gender ethnicity county
* Drop observations later than 1996
drop if year < 1996
```

Formatting graphs and tables

Formatting options for graphs

Hopefully, you should be aware of how to make common types of graphs and plots in Stata now. I want to briefly touch on a few graphics options.

- As an option for graphics commands, you can add different titles:
  - Main Title — **title("Title here")**
  - Y/X Axis Titles: **ytitle("Title here") / xtitle("Title here")**
- To change the scale of an axis, use: **xscale(range(a b))**
- Finally, after creating a graph for a report, save it into your working directory. You can do this with **graph export**.
  - The syntax of this command is: **graph export "filename.type", replace**.
  - Good graph types include **.emf**, **.pdf**, **.png**, and **.tif**.

Graph example code

```
histogram wage,  ///
            title("Distribution of Wages") ///
            subtitle("(all years)") ///
            xtitle("Wages")
graph export "histogram_wage.tif", replace
```

Graph example

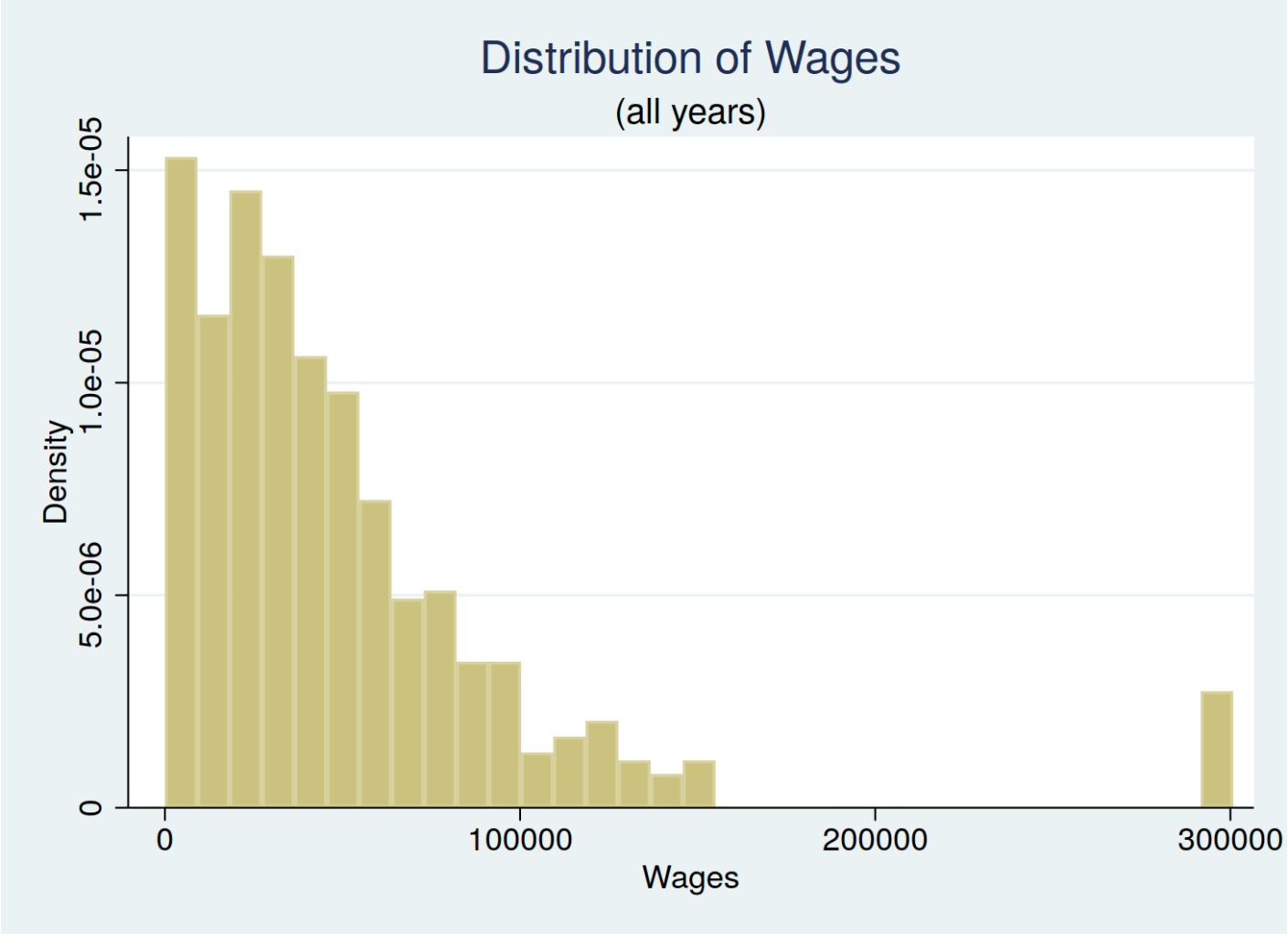

Figure 1:

Creating table output with ESTOUT

Finally, a very useful package in Stata to generate reports is **estout**.

- **Estout** generates publication-quality tables for both summary statistics and regression output, in a variety of formats including word documents (.rtf) and Latex.
- To install this package, in the Stata console type: **ssc install estout**
- Once you've installed the package, the first step is to write **eststo :** immediately before (on the same line) you type a normal regression command.
    - Do this for each regression you want to store on the same table

Using esttab to output regression tables

- When you have saved all the regression you want in an individual table, output the table to a file with the command **esttab using "FILENAME"**.
  - You can find options for **esttab** at http://repec.org/bocode/e/estout/esttab.html
  - But I will give you suggested options for the assignment since these can be a little complicated.
  - The main thing to know is how to change the titles (and model titles using the **mtitles()** option).
- After using **esttab**, always use **eststo clear** to clear the saved regression in memory before saving further regressions for output.

ESTOUT Example

```
eststo clear
eststo:  xtreg lwage exper educ union black hisp ///
  nrthcen nrtheast south i.year

esttab using "Wooldridge_earnings", ///
    title("FE Regression of earnings") ///
    se label wrap noabbrev rtf drop(19*) ///
    star(* 0.10 ** 0.05 *** 0.01) b(%8.2g) ///
  compress one replace
eststo clear
```