

# POLS571 - Longitudinal Data Analysis

October 25, 2001

## Panel Data Models!

### 1 Introduction

#### 1.1 Terminology

We're going to be talking about data in which variables vary both over time and across cross-sectional units. We'll always refer to the units as  $i = 1, 2, \dots, N$ , and to the time points as  $t = 1, 2, \dots, T$ . The total number of *observations* (i.e., lines of data) is equal to  $NT$ . Some general conventions for naming these kind of data are:

- *Panel data* generally refers to data which are cross-sectionally dominated; that is, where  $N$  is significantly larger than  $T$ . Examples are the NES panel studies ( $N = 2000, T = 3$ ) or the Panel Study of Income Dynamics ( $N = \text{large}, T = 12$  or so). Such data usually have a fixed  $T$ , so that these data's asymptotics are in  $N$ , which is important (we'll come back to this).
- *Time-series cross-sectional* (TSCS) data usually means data in which either  $T$  is dominant, or  $N \approx T$ . These data are common in comparative politics. But, it can also refer to data where  $N$  is dominant, but  $T$  is larger than in panel data (e.g. all-dyads all-years IR data, with  $N = \text{several thousand}$  and  $T = 50$  or more). Here,  $N$  is usually fixed, and the asymptotics are in  $T$ ; moreover, if we have enough data, we can say something about the time-series properties of the data as well as the cross-sectional part.
- *Repeated measures data* is a term that gets used more in biostats. Its useful, because it can mean any of these things, but its also vague.

#### 1.2 Data Structure

In panel or TSCS data, we have multiple lines of data for each unit of observation. Such data are arranged as follows:

ID	T	Y	X <sub>1</sub>	...
1	1	250	3.4	...
1	2	290	3.3	...
⋮	⋮	⋮	⋮	...
2	1	160	4.7	...
2	2	150	4.9	...
⋮	⋮	⋮	⋮	...

When analyzing such data in Stata, its good practice to `-sort-` the data on the  $N$  and  $T$  identifier variables periodically.

The series of commands in Stata for analyzing such data all begin with the letters `-xt-` (for “**x**-sectional **t**ime-series”). We need to tell Stata that our data are of this format in order to use these commands (not unlike we did with `-tsset-`). We do this by specifying the  $i$  and  $t$  variables:

```
. iis ID
. tis T
```

Once we’ve done this, there are a number of commands that become available to us. Also, Stata has a few useful commands for managing TSCS/panel data...

### 1.2.1 The `-expand-` command

Stata’s `-expand-` creates multiple “copies” of observations already in the data. This is good as a first step, when you have data that don’t vary over time but are planing on adding/collecting some that does. Suppose we have a (very) small dataset on three countries – the U.S., the U.K., and Japan – which included data on variables that didn’t vary over time (e.g., government type, etc.):

ID	$X_1$	<i>YEARS</i>	...
US	250	7	...
UK	290	9	...
JP	150	5	...
⋮	⋮	⋮	⋮ ...

Suppose we wanted to collect data on ten years of data, 1991-2000, for each country (giving us  $NT = 30$ ). To create a dataset with 30 lines of data, and with the existing  $X$  variables retained for each country, we would simply type:

```
. expand 10
```

This would give us a dataset that had 10 exact copies of each existing observation. We could then assign each line of data a year by typing:

```
. sort ID
```

```
. gen year = 1991
```

```
. quietly by ID : replace year=year[_n-1]+1 if year[_n-1]! =.
```

This gives us a dataset ready for inputting or merging time-varying data. The `-expand-` command will also take variables as an argument; so if, for example, you wanted to create a number of years equal to the variable `YEARS` in the data, you would type:

```
. expand years
```

which would create 6 copies of the observation for the U.S., eight of the U.K., and four for Japan.

### 1.2.2 The `-reshape-` command

Stata's `-reshape-` command is useful for converting data from “wide” to “long” format and back. Think of the way we usually arrange (e.g.) TSCS data as “long”, in that we use rows rather than columns for storing information. So, if we had data on three years worth of GDP numbers for the three

countries mentioned (i.e.,  $NT = 9$ ), we'd typically have it arranged as:

NAME	YEAR	GDP
US	1980	280
US	1981	294
US	1982	303
UK	1980	121
UK	1981	124
UK	1982	131
JP	1980	176
JP	1981	192
JP	1982	212

Note that we could accomplish the same thing by storing the data with separate variables for each of the GDP-years:

NAME	GDP80	GDP81	GDP82
US	280	294	303
UK	121	124	131
JP	176	192	212

The `-reshape-` command converts data from one such format to the other. I won't go into the details of it right now (there are lots of options), but suffice it to say that, in many cases, you receive (e.g.) government data in "wide" format, and need to convert it to "long" format in order to analyze it. `-reshape-` makes this much easier.

### 1.2.3 The `-stack-` command

`-stack-` does exactly that; it takes existing variables and "stacks" them into a single column. This is useful when you have data that are in a variation of "wide" format, in that it can act as a combination of `-reshape-` and `-expand-`. Suppose your data look like this:

NAME	US	UK	JP
1980	280	121	176
1981	294	124	192
1982	303	131	212

The `-stack-` command will convert this onto normal (“long”-format) data in one fell swoop:

```
. stack US UK JP, into(gdp)
```

Your new data are now in “long” format, albeit minus labels (so keep good track of what goes where). `-stack-` is generally more useful for smaller datasets with few variables; otherwise, `-reshape-` is more flexible.

## 2 Variation in the TSCS context

Variables in TSCS data can, obviously, vary across units, or over time, or both. Consider data on Supreme Court justices, by year. A variable for whether or not the justice is from the South will vary only between justices, never “within” any particular justice. Conversely, a variable for the party of the sitting president will not vary across justices in any given year (“between”), but will vary over time (“within”).

We can think of these two kinds of variation as reflecting variations around some mean. Consider, for example, the variable for the number of majority and dissenting opinions written by a justice in a given year (we call this variable *WRITING*). Simply examining the mean and standard deviation yields:

$$\mu = 17.94 \quad \sigma = 14.14 \quad \text{Minimum} = 0 \quad \text{Maximum} = 103 \quad NT = 1765$$

This variation occurs both “within” and “between” justices, however. One way to separate these two concepts is to consider the justice-specific mean  $\bar{X}_i = \frac{1}{T_i} \sum_{t=1}^{T_i} X_{it}$ . This value represents the average within-justice level of writing; comparing these differences tells us what the between-justice differences in writing levels are. The difference between this average and the

observed value in any given year is  $X_{it} - \bar{X}_i$ ; we think of this deviation as the within-justice variation around the mean.

If we examine the “between” versus the “within” variation in *WRITING* separately (using Stata’s `-xtsum-` command), we find:

$$\begin{aligned}\sigma_{BETWEEN} &= 11.24 \text{ Min} = 0 \text{ Max} = 65.53 \\ \sigma_{WITHIN} &= 8.46 \text{ Min} = -26.59 \text{ Max} = 85.24\end{aligned}$$

This suggests that there is generally greater variation in levels of writing “between” justices than there is within any given justice’s career. (This ought not be too surprising). We’ll come back to these ideas again numerous times in the next few weeks.

### 3 General TSCS Regression Issues

Think of a general regression model for cross-sectional data:

$$Y_i = \alpha + \beta X_i + u_i \tag{1}$$

This model assumes several things:

- All the usual OLS assumptions, plus
- that the constant term is constant across different *is*, and
- that the effect of any given variable *X* on *Y* is constant across observations (at least, to the extent that non-constancy isn’t specified in the model, e.g. through interaction terms).

We can write a similar model in the TSCS context as follows:

$$Y_{it} = \alpha + \beta X_{it} + u_{it} \tag{2}$$

Note that this model assumes the same things as the earlier ones, especially about the effects of constants and covariates.

In *any* regression context, the two assumptions mentioned are critical; violating them leads to a form of specification bias. In the TSCS context,

these two assumptions are often going to be problematic. This is because, since we're observing multiple units over time, there's usually some reason to believe that there may be differences in either  $\alpha$  or  $\beta$  over either  $i$  or  $t$ . Consider each of these possibilities.

### 3.1 Variable intercepts

One possible violation of the above assumptions is that the intercepts vary. The most common way this occurs is for different units to have varying intercepts:

$$Y_{it} = \alpha_i + \beta X_{it} + u_{it} \quad (3)$$

The slopes for each unit are the same, but the intercepts are different. Its also possible that the intercepts vary over time, rather than over units:

$$Y_{it} = \alpha_t + \beta X_{it} + u_{it} \quad (4)$$

or even over both  $i$  and  $t$ :

$$Y_{it} = \alpha_{it} + \beta X_{it} + u_{it} \quad (5)$$

Most of the time, however, it is unit differences that concern us most. If we have data that correspond to (3), but estimate a model like (2), we can get biased coefficients. To see how this is true, consider the data in Figure 1.

The actual slope (the effect of  $X$  on  $Y$ ) is equal to 1.0; however we overestimate it significantly (here, we get  $\hat{\beta} = 3.18$ , with a standard error of 0.14) because of the different intercepts. Its just as likely that the bias will be the other way, however; in most instances, we just don't know.

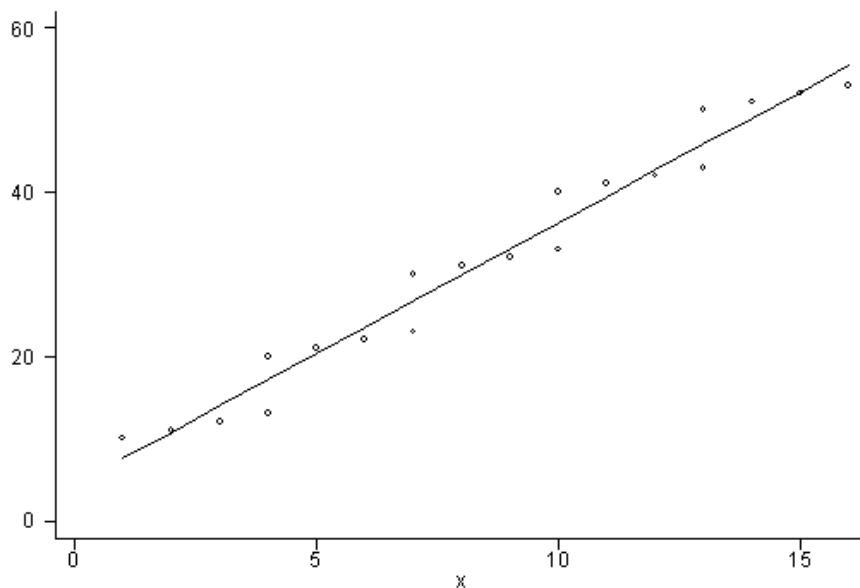
### 3.2 Variable Slopes

The other obvious possibility is that we have a constant intercept, but the effects of  $X$  on  $Y$  differs across either units or (less likely) time; e.g.:

$$Y_{it} = \alpha + \beta_i X_{it} + u_{it} \quad (6)$$

We could also have variation in  $\beta$  over time, or even over both units and time.

Figure 1: Regression Results with Varying Intercepts



A model like in (6) assumes that the regression lines all pass through the same point on the Y-axis, but that their slopes differ (perhaps tremendously). As a result, the estimate of  $\hat{\beta}$  we'll get will be an “average” of those for the individual  $i$ s.

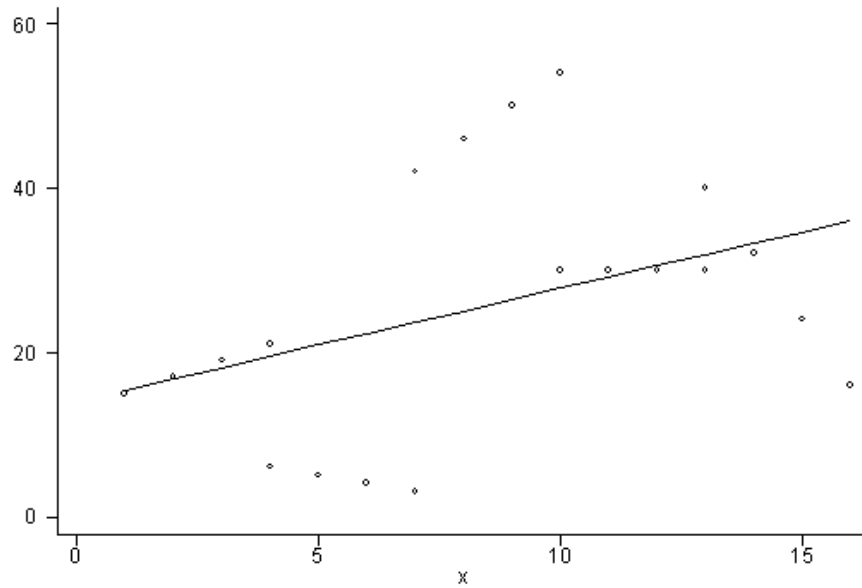
### 3.3 Variable Slopes and Intercepts

This is when things really start to get difficult. We might, for example, have variable slopes and intercepts for each unit  $i$ :

$$Y_{it} = \alpha_i + \beta_i X_{it} + u_{it} \quad (7)$$

and could, of course, also have different  $\alpha$ s and  $\beta$ s for every time point, or for both different units and time points. The example in (7) is illustrated in Figure 2, which shows what you get if you estimate a model like (2) when the data correspond to (7).

Figure 2: Regression Results with Varying Slopes and Intercepts



Not surprisingly, the results you get are nonsensical. This points up how important accurately modeling slope- and intercept-variation in TSCS models can be. On Tuesday, we'll start addressing these issues.