Introduction

In this lab we’re going to cover panel data and estimating the appropriate models in Stata. Just like we did with count and duration data, we have to designate our data as panel data; this gives us a series of built-in functions related to panel data. We will also cover the appropriate times to implement fixed and random effects, and very quickly demonstrate how to do this in Stata.

Once we’re through the obligatory Stata, we’ll jump into a demo in R that will illustrate why we are making the corrections we do when we think there are groupings and serial-correlation.

R Section

Random Effects

What if we had the following data. Student test scores in a classroom over the course of a semester. Imagine that they’re kindergartners and you, as the administrator of a magnet school are interested in placing people in the best colleges. Clearly, you’ve got to start testing at an appropriately early age, right?

Suppose the data comes from the following set of equations and random-variables:

\[
\text{score}_{it} = 0 + 0.75 \times \text{hours}_{it} + \alpha_i + \epsilon_{it}
\]

\[
\alpha_i = N(0, 0.7^2); \quad i \in \{1 \cdots 100\}
\]

\[
\epsilon_{it} = N(0, 0.2^2); \quad t \in \{1, 2, 3, 4, 5\}
\]

Notice that by construction, the relationship should hold at the individual-level, and that individuals are given a random-intercept (\(\alpha_i\)) that is their underlying ability.

You want to know the relationship between hours spent studying latin\(^4\) and score in juried performances of Virgil’s \textit{Aeneid} – but we’ll just call it assignment grade. You thought \textit{Dance Mom’s} was catty? These performances are fierce!

\(^1\)Is responsible for the successful parts of this lab
\(^2\)Who is responsible for the unsuccessful parts of this lab
\(^3\)Most of this is from Chris Adolph’s great course on Panel Data.
\(^4\)\textit{Iuppiter te perdat!}
But, we have more data that we weren’t using in the last set of plots – repeated measurements of each student in the classroom. Let’s re-draw the same plot as before, but this time, we will color the nodes by the student. The colors will repeat, but it will be clear which points are repeated observations of the same student.

With this kind of information, we can make a better estimate of the effect of studying on a student’s performance – rather than estimating a pooled regression where we do not differentiate between individuals, instead we can estimate the relationship for each student. This will give each student her own regression line (with estimated slope and intercept), making this a random effects model.

Upon inspection we see a few things:

1. The relationship holds for each of the students – the fitted line is roughly parallel to the pooled estimate.
2. The vertical distance between the individual regression line and the pooled regression line is just the random-effect.

3. What would happen if we were to subtract out each individual’s random effect? What would the plot of our data look like?

Recall that we constructed our random effects to be normally distributed? Are they?

Figure 3: The line on the side of the plot-frame is the kernel density of the random effects – but this is mostly just Chris Adolph showing off his plot tools!

Let’s think about the hierarchical nature of this because random effects are part of a large class of models known as hierarchical models. **Level 1**: the student level is shaped by \( \alpha_i \) and sits “above” **Level 2**: the student & assignment level. It is important to note that in setting up a random effects model, we haven’t controlled for any possible omitted variables. So, if unmeasured ability is correlated with study effort our estimate of the regression coefficient will still be biased.

**Fixed Effects**

Rather than believing that individual level differences are caused at random, frequently we think that there might be some set of variables that have caused differences.

To expand on the last example, imagine we are comparing performance scores across schools and some of the schools are located in the relatively affluent La Jolla school district and others are in the relatively less affluent Pacific Beach school district. While we think that affluence has something to do with any difference we might observe, we’re also aware that there are long histories within each of these communities that might still be having some effect.
This is frequently just the goal of using fixed-effects – including fixed effects controls for all time-invariant variables, which captures unobserved variance that is potentially correlated with the RHS regressors that you’re interested in studying.

This sounds good, right? But it comes at a cost, because we are effectively removing all of the cross-sectional variance that we had in the data. We can still interpret the within group effects, and even the between group effects, but we have no ability to interpret the across time effects in a Fixed Effect setup.

Example Imagine the following example related to the work that Thad does – we have time-series and cross-sectional data on republican vote share (as a percent of total votes cast) in a number of counties (that are located in a number of states). This is just the setup of Gelmen et al’s “Red State, Blue State, Rich State, Poor State”, but we will make up the data to work with.

The general form of the model is

\[ y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 z_i + \epsilon_{ij} \]
\[ \epsilon_{ij} \sim N(0, \sigma^2) \]
\[ i \in \{1, \cdots, N\} \]
\[ j = \text{in}\{1, \cdots, M_i\} \]

where \( j \) indexes a set of \( M_i \) counties drawn from state \( i \). Following Adolph’s example, imagine there are \( N = 15 \) states and we drew \( M = M_i = 15 \) counties from each state.

Lets be more explicit about what we’re measuring in our model:

\[ \text{RVS}_{ij} = \beta_0 + \beta_1 \text{Income}_{ij} + \beta_2 \text{ConservativeCulture}_{i} + \epsilon_{ij} \]
\[ \epsilon_{ij} \sim N(0, \sigma^2) \]

But imagine that we have the problem that actually measuring conservative culture is difficult. Sure, we could make up an IR-style variable, have one guy code it, and then have everybody in quantitative-IR use it... I digress. If we include a mis-measured ConservativeCulture we will generate considerable bias in our \( \hat{\beta}_1 \). If we omit ConservativeCulture we will do the same. Are we snookered?

Here’s where it gets excellent. We know there is a grouping structure to this data. Groups of counties are drawn from the same state and it is possible that there is some unobserved conservative culture that is present in different quantities in different states. It would seem that in contrast to the first pooled model that we estimated, there is in fact a positive relationship between income and republican vote share (a trend we might have expected).
Figure 4: On the left are the data plotted without any grouping structure, and on the right is the best-fit line through those points. This doesn’t look quite right, does it?

But how should we estimate this model? In any appropriate model to fit this data, we will have to control for the state-level variation – we have to model the differences. We could do this either by collecting all the variables that we think might lead to state level variation – percentage minority, VRA, civil war factions, ... – clearly the list of variables here is long, and the task would be, um, daunting to get right.

Or, we could brute force the solution by including an indicator variable for each state.

How are fixed and random effects different?

1. Fixed effects control for omitted variables (at the grouping level) – random effects don’t

2. Random effects allow the effects to follow a specific distribution – fixed effects don’t
Gulp...Stata

Setting Data as Time Series

To designate your data as panel data in Stata, it needs to be in long form, that is, organized by country-year. Wide-form would have a column for each variable in each year. If you’ve got data in one or another of the forms, remember that the reshape command will make that transformation for you.

We’ve got data that is already in long-form, so we can jump immediately to using it. This is data put together by BDM. Its data about the world so that you can forecast the future!

To designate the data as time-series data, use

tset ccode year

Which sets the panel on ccode and the time on year. Once we have set the data into a time-series dataframe Stata will provide us a whole bunch of commands with the xt command appended to the front of them – xtdescribe, xtsum, xtline are a few that come immediately to mind. It is kind of like this where you cross some threshold and then can do all kinds of wacky stuff you couldn’t before.

Fixed and Random Effects

We could go ahead and estimate a pooled OLS model, but this would be pretty wrong, for all the reasons that we discussed in class: last year’s gdp is probably highly predictive of this years gdp. And you won’t meet the exogeneity conditions we need to estimate OLS (think about the path diagrams from last week).

\begin{verbatim}
reg WB_gdppc_con PR CL open auton ELF60
\end{verbatim}

Instead, we could estimate this in a couple of ways. First, the xi command will allow us to run the dummy variable version of this. In this version, we would create a dummy vector for each country (widening our dataset) that takes a 1 when the country is present and 0 otherwise. This should give us a consistent estimate of the coefficients, but it won’t get the standard errors “right” for a number of reasons (chief among them serial-correlation) unless we mess with the variance-covariance matrix in the appropriate ways.

\begin{verbatim}
xi: reg WB_gdppc_con PR CL open auton ELF60 i.ccode;
drop _I*;
areg WB_gdppc_con PR CL open auton ELF60, absorb(ccode) vce(cluster ccode);
xtreg WB_gdppc_con PR CL open auton ELF60, fe;
\end{verbatim}

What is reported is the within-group effects – because they are usually what we’re most interested in. Stata will report you between-group effects if you like, but you’ve got to ask nicely.
xtreg WB_gdppc_con PR CL open auton ELF60, be

Maybe we’re interested in what the coefficients look like for each of our panel units? How different is the effect of an independent judiciary in Angola compared to the United States?

statsby, by(ccode) clear: reg WB_gdppc_con PR CL open auton ELF60
format _b* %9.2f
list, clean

It is possible, but unlikely most of the time in political science that we think that the group level intercept is truly random across our cases – that conditioning on our data, the expectation of the intercept is zero – \( E(\alpha_i|x_i) = 0 \). Like we said, it is very unlikely that you can meet these assumptions with your data – certainly global data will not meet it. Putting aside that this we can’t meet these assumptions, we could estimate this model using:

xtreg WB_dgppc_con PR CL open auton ELF60, re

### Checking Assumptions

To check (a-theoretically) if our data meets this assumption about \( \alpha_i \) we can conduct a Hausman test for the two models:

xtreg WB_gdppc_con PR CL open auton ELF60, fe;
est sto fixed;
xtreg WB_gdppc_con PR CL open auton ELF60, re theta;
est sto random;
hausman fixed random;

We can still estimate (and get confused by...) instrumental variables in this model – just append the `xt` before the instrumental variables regression call. We will use the built in `xtivreg` call here, but there are additional packages that would allow you to use the `xtivreg2` call that is consistent with the syntax we presented last week.

xtivreg WB_gdppc_con (PR = polity) CL open auton ELF60, fe

There are a bunch of other panel models that you can use for panel and dynamic panel data – `xtabond`, `xtgls`, `xtpcse`, `xtlogitx` and a bunch of others. But remember two things:

1. Always make sure that you know the model you are estimating; and,
2. Don’t assume it is a good idea just because it is available in Stata.
Wrapping up Stata

All the teasing of Stata aside, it can be an adequate package. However, frequently researchers get into trouble when they do not know the assumptions, requirements, or routines that are going into the processes they are estimating. You do NOT want to get caught with your pants down in a talk or in a journal for estimating an incorrect model. And when you come to a talk with Stata plots or regressions, the me’s (Chris Fariss’s) in the audience are going to be primed to poke at your knowledge.

So, if there is one take away message from this: If you have any questions about what a model requires, what the assumptions are, how the routine is working look it up. Do your stata coding with an appropriate graduate-level econometrics book nearby. Write your self notes and convince yourself that the math works in your case. The same goes for researchers who are using built-in function calls and add-on packages in R.