Pooling Cross Sections Across Time and Simple Panel Data Methods

ECONOMETRICS (ECON 360)

BEN VAN KAMMEN, PHD

Introduction

So far this class has analyzed data that are *either* cross-sectional or time series.

Now it will examine data that have *both* dimensions. These come in two forms:

- multiple ("pooled") cross sections from different time periods and
- the *same* cross section ("panel") observed in multiple time periods.

The difference is that pooling cross sections means different elements are sampled in each period, whereas panel data follows the same elements through time.

The objective is to explore what problems can be solved with such "two dimensional" data that is difficult to do with a single cross section.

Outline

Pooling Independent Cross Sections Across Time.

Policy Analysis with Pooled Cross Sections.

Two-Period Panel Data Analysis.

Policy Analysis with Two-Period Panel Data.

Differencing with More than Two Time Periods.

Pooling independent cross sections across time

For many surveys, a cross-sectional sample is drawn periodically; the book uses the example of the Current Population Survey (CPS).

- Each CPS sample is quite large in its own right, but when they are pooled it becomes a very large sample—with all the attendant benefits in terms of precision.

As long as we're talking about cross sections drawn in periods that aren't *too* far removed from one another,

- i.e., in which the relationships among variables are unlikely to have changed notably,

pooling them doesn't introduce much of a problem statistically either.

Yes the distribution of the variables may change over time, but this can typically be accounted for in a regression model by estimating the coefficient on a time period indicator, e.g., year.

Pooling independent cross sections across time (continued)

A year indicator variable would be constructed:

$$year_t = \begin{cases} 1, & \&\text{observation is from year } t \\ 0, & \&\text{otherwise.} \end{cases}$$

When it is of interest, a year indicator can also be interacted with another “x” variable of interest to examine whether its effect changed in that year compared to the other period(s) in the sample.

The Chow test for structural change across time

As you have seen with Chapter 7.4 (in the context of differences across groups), interaction with an indicator can be taken to the extreme by estimating coefficients on interactions between the year indicator and all the variables in the model.

Their joint significance (*F* test) would be evidence to reject the null hypothesis that the model does not change between two periods.

- There are exercises in the Wooldridge book that apply this to more than two time periods as well.

Policy analysis with pooled cross sections

Empiricists are fortunate, on occasion, to observe natural experiments.

- These occur when some economic agents are exposed to an exogenous change in their incentives as a result of a locally enacted policy, for example, while others are not so exposed.

Natural experiments aren't quite as good as laboratory experiments because the treatment (exposure to the policy change) may not be applied to a group that is ex ante identical to the control group, as it is in a laboratory.

That's where the usefulness of multiple time periods comes in.

Policy analysis with pooled cross sections (continued)

To accurately measure the causal effect of some treatment,

- e.g., a state-wide ban on text messaging while driving,

a researcher would wish to compare a measure (*y*) of roadway safety in two hypothetical states:

$$causal\ effect = y_{1,post-ban} - y_{1,counterfactual},$$

i.e., the difference between what happened and what *would have* happened if the state(s) had not enacted the texting ban.

Policy analysis with pooled cross sections (continued)

Since this latter counterfactual is not observable, a researcher would be tempted to substitute observations of other states that did not enact texting bans, estimating:

$$observed\ diff = y_{1,post-ban} - y_{0,post-ban}.$$

The difference is between two different groups where the non-banning states are the control group (0) and the banning states are the treatment group (1).

This strategy may or may not be sound, depending on how comparable the two groups were prior to group (1) enacting their laws.

Policy analysis with pooled cross sections (continued)

If the term in the 2nd parentheses is zero (as it is in lab experiments), the non-ban states provide a good counterfactual for what would have happened in the absence of the ban in the treatment states.

- Then the cross sectional differences could be interpreted as causal effects.

If that is not the case (and it frequently is), two cross-sections can help solve the problem.

Label the above difference,

$$y_{1,post-ban} - y_{0,post-ban} \equiv observed\ diff_{post}, \text{and}$$

$$y_{1,pre-ban} - y_{0,pre-ban} \equiv observed\ diff_{pre}.$$

Difference in difference (DD) estimation

The difference between these two differences (you see the origin of the strategy's name) is:

$$DD \equiv \left(y_{1,post-ban} - y_{0,post-ban}\right) - \left(y_{1,pre-ban} - y_{0,pre-ban}\right).$$

This expression subtracts any pre-existing differences between the two groups from the observed post-treatment difference.

- So it "controls for" how different the two groups are prior to the treatment.

DD estimation (continued)

The DD estimator makes it much more plausible that it estimates the object of interest:

$$DD = \left(y_{1,post-ban} - y_{1,counterfactual}\right) - \left(y_{0,post-ban} - y_{0,pre-ban}\right) + \left(y_{1,counterfactual} - y_{1,pre-ban}\right)$$

$$\Leftrightarrow DD = causal\ effect + (\Delta_1 - \Delta_0).$$

As long as there is no other confounding change happening in either group, it is plausible that the last two terms are both zero in expectation and the differences in differences estimator captures the treatment effect.

The key assumption is that, in the absence of the treatment, the treatment places *would have* changed at the same rate as the control places.

Regression DD

In a regression context, the estimator would appear in the model,

$$y = \beta_0 + \delta_0 d2 + \beta_1 dT + \delta_1 d2 * dT + error,$$

$$d2 = \begin{cases} 1, & \&\text{post treatment} \\ 0, & \&\text{otherwise,} \end{cases}$$

$$dT = \begin{cases} 1, & \&\text{treatment group} \\ 0, & \&\text{control group.} \end{cases}$$

Regression DD (continued)

In terms of parameters, the pre-treatment and post-treatment differences would be:

$$\Delta y_{d2=0} = \beta_1 \text{ and } \Delta y_{d2=1} = \beta_1 + \delta_1.$$

And DD would be:

$$DD = \Delta y_{d2=1} - \Delta y_{d2=0} = \beta_1 + \delta_1 - \beta_1 = \delta_1.$$

Policy analysis using pooled cross sections (concluded)

When expressed in its regression form, the DD estimator has more flexibility,

- i.e., one could include other control variables that vary through time but the power of multiple cross-sections is evident merely from comparing a set of 4 averages (pre and post, treatment and control).

To estimate the effect of a texting ban, one could simply average the rates of motor vehicle crashes in states that enacted bans and those that did not and compare the two differences (pre and post).

Two-period panel data analysis

Depending on the nature of the data used, the texting ban DD estimation could serve as an example of pooled cross sections *or* panel data analysis.

If the unit of observation was individual drivers, the data might consist of random samples from two different states in two consecutive years—and the drivers in the sample need not be the same in both periods.

- The outcome measure (y) may be the number of collisions each respondent was involved in, in a year.
- This data would be classified as pooled cross sections.

Two period panel data analysis (continued)

However automobile collisions are officially recorded by law enforcement agencies, and the statistics for each county and state are usually readily available.

If the unit of observation was states instead of drivers, the data would be classified as longitudinal or panel—in which the same elements (states) are observed over time.

DD analysis could be conducted in the same fashion, but now the empiricist would run into sample size issues, since there are only 50 states and only a small fraction (1? 2?) of them are likely to experiment with a law contemporaneously.

Consequently there isn't much variation in the treatment indicator, and the estimates are likely to be quite imprecise.

- Nonetheless it illustrates the difference between pooled cross sections and panel data.

Two period panel data analysis (continued)

In general one of the biggest advantages of using panel data, compared using one or more non-identical cross sections, is its negation of fixed effects.

Variables that are specific to the elements (individuals, cities, firms, schools, et al.) and are "fixed" (do not change or change very slowly) over time belong to this category.

In an individual-level data set these would include:

- Gender and race.
- Intrinsic ability, e.g., motivation, intelligence, other unobservable, but crystallized, skills.
- Characteristics of your ancestors or birth place.

Two period panel data analysis (continued)

These are distinguished from variables that vary across individuals *and* over time.

The two categories can be distinguished in the regression model by subscripts.

Fixed effects are only indexed by an *i* (specific to the individual) because they do not vary over time ($a_{i,t=0} = a_{i,t=1} = a_i$).

Other variables are indexed with the individual as well as which time period in which they are observed. Examples of these include:

- age, years of labor market experience.
- marital status, number of children, current place of residence.
- years of schooling (except maybe among older individuals),
- occupation and industry in which one works.

Two period panel data analysis (continued)

So a regression with fixed effects (fixed effects model) would look like this:

$$y_{it} = \beta_0 + \delta_0 t + \beta_1 x_{it} + a_i + u_{it}; \text{for simplicity } t \in \{0,1\}, \text{where}$$

β_0 is intercept for period 0, and the intercept for period 1 is $\beta_0 + \delta_0$, x_{it} are variables that vary in the cross section and over time, and u_{it} is the idiosyncratic error.

The usefulness is that *x* variables of interest are correlated with the fixed effects and also that the fixed effects are not observed.

So in an OLS regression, the fixed effects are relegated to the composite error term:

$$v_{it} \equiv a_i + u_{it}.$$

Two period panel data analysis (continued)

Here is where the really cool part comes in.

- Since a_i does not change over time, it is negated when you take the first difference the model, resulting in the first-differenced equation,

$$\Delta y_{it} \equiv y_{i1} - y_{i0} = \delta_0 + \beta_1 \Delta x_{it} + a_i - a_i + \Delta u_{it} = \delta_0 + \beta_1 \Delta x_{it} + \Delta u_{it},$$

where $\Delta x_{it} \equiv x_{i1} - x_{i0}$.

Differencing turns the sample from two cross sections into a single cross section, i.e., two observations are necessary to form one difference (technically we could drop the subscript t).

But now the fixed effects have been negated and no longer appear in the model.

Two period panel data analysis (continued)

The differenced model can be estimated by OLS, with the resulting estimator of β_1 known as the first-differenced estimator ($\widehat{\beta}_{FD}$).

- The properties of unbiasedness and consistency will prevail as long as both values of *x* are uncorrelated with both idiosyncratic errors, i.e.,

$$E[(x_{i1} - x_{i0})(u_{i1} - u_{i0})] = 0.$$

This may or may not be a good assumption.

- After all the value of *x* in the latter period could respond to a particularly severe shock (large or small u_{i0}) in the former period, inducing such correlation.

But there are plenty of instances in which this assumption is: plausible,

- much better than the one required to use a single cross-section, and
- useful for resolving omitted variables bias.

Two period panel data analysis (continued)

An underappreciated fact about panel data analysis is that there has to be temporal variation in *x* for β_1 to be identified!

- For instance if the unit of observation is states and a *national* law is passed between periods 0 and 1, $\Delta x_{i1} = 1$ for all states; there is no variation from which to estimate the coefficient.
- What the research design needs is *local* laws enacted only by a subset of places within the country.

Inference based on estimating a first-differenced model by OLS depends on homoskedasticity, which is nothing new.

- The issues raised by, and solutions suggested to remedy violation of homoskedasticity, however, have already been covered in Chapter 8.

Two period panel data analysis (concluded)

First differencing can also accommodate more than 2 time periods, as well as a whole vector of *x* variables, as in multiple OLS.

To generalize the model to *k* regressors, it would look like:

$$y_{it} = \beta_0 + \delta_0 t + \beta_1 x_{it1} + \ldots \beta_k x_{itk} + a_i + u_{it}; T \geq 2.$$

- As in Example 13.6 in Wooldridge, more than 2 periods can be useful for estimating finite distributed lag (FDL) models (covered in Chapter 10), in which lagged values of the regressors enter the model as well as contemporary ones.

Organizing panel data

Earlier I alluded to data as a “spreadsheet” with variables as columns and observations as rows.

A unique question posed by panel data is whether the data should be organized as “long” or “wide”.

Data stored in the “long” format feature *time* as a variable, differentiating each observation of an individual from the others.

“Long” data storage

Person (“i”)	Year (“t”)	Gender (1=“Male”)	Age (Years)
1	2010	1	20
1	2011	1	21
. . .			
n	2010	0	36
n	2011	0	37

It is called “long” because the number of observations equals (assuming the panel is “balanced”) *n*T*: the cross-sectional sample size times the length of the time series.

There is one observation per combination of *i* and *t*,

- i.e., a long list of observations.

“Wide” data storage

Person (“i”)	Gender (1=“Male”)	Age2010	Age2011
1	1	20	21
. . .			
n	0	36	37

By contrast the “wide” format for storing panel data stores each period for time-variant variables as a separate variable and does not have a separate variable for time, itself.

The same data set from the last slide would look like this in wide format.

It is called “wide” because the length of the list is now merely *n*.

But there are columns for each value of *t* for each variable that varies over time, e.g.,

- *Age2010* and *Age2011* are separate variables in the wide format.

Organizing panel data (continued)

Remarkably Stata has commands for converting a data set from long to wide (and back). To demonstrate this using the **lowbirth.dta** file that accompanies the text, the following code would convert the data from long to wide.

use "[Location of your data followed by \]LOWBRTH.DTA", clear

This is state-level data with *T=2* and $t \in \{1987, 1990\}$.

egen id=group(stateabb)

This generates a numerical *id* variable that takes values unique to each state.

drop cafdcprc clpcinc clphysic clowbrth cinfmort clafdcpy cafdcinc clbedspc cpovrate cafdcpsq clphypc clpopul

Organizing panel data (continued)

This gets rid of all the differenced variables that are only observed for the 2nd period

- and which we could easily re-generate with the difference operator (beyond the scope of this tutorial).

reshape wide lowbrth- lpcinc lphysic afdcpay- lafdcpay beds- lbedspc povrate afdcpsq physicpc lphypc lpopul, i(id) j(year)

The command is called “reshape”.

- The next input is what kind of data you want to turn it into, i.e., “wide” because the data is already “long”.
- Then you input a list of all the *x* variables, i.e., time variant ones.
- Finally the options include designations of the cross sectional index (“i”) and time index (“j”) variables.

Output using Stata's "reshape" command

```
. use "F:\ECON 360 Econometrics\Wooldridge Data 5E\LOWBRTH.DTA", clear

. egen id=group(stateabb)

. drop cafdcprc clpcinc clphysic clowbrth cinfmort clafdcpy cafdcinc clbedspc cpovrate cafdcpsq clphypc clpopul

. reshape wide lowbrth- lpcinc lphysic afdcpay- lafdcpay beds- lbedspc povrate afdcpsq physicpc lphypc lpopul, i(id) j(year)
(note: j = 1987 1990)

Data                               long   ->   wide
-----------------------------------------------------------------------------
Number of obs.                      100   ->      50
Number of variables                  25   ->      45
j variable (2 values)              year   ->   (dropped)
xij variables:
                                lowbrth   ->   lowbrth1987 lowbrth1990
                                infmort   ->   infmort1987 infmort1990
                                afdcprt   ->   afdcprt1987 afdcprt1990
                                  popul   ->   popul1987 popul1990
                                  pcinc   ->   pcinc1987 pcinc1990
                                 physic   ->   physic1987 physic1990
                                afdcprc   ->   afdcprc1987 afdcprc1990
                                    d90   ->   d901987 d901990
                                 lpcinc   ->   lpcinc1987 lpcinc1990
                                lphysic   ->   lphysic1987 lphysic1990
                                afdcpay   ->   afdcpay1987 afdcpay1990
                                afdcinc   ->   afdcinc1987 afdcinc1990
                               lafdcpay   ->   lafdcpay1987 lafdcpay1990
                                   beds   ->   beds1987 beds1990
                                 bedspc   ->   bedspc1987 bedspc1990
                                lbedspc   ->   lbedspc1987 lbedspc1990
                                povrate   ->   povrate1987 povrate1990
                                afdcpsq   ->   afdcpsq1987 afdcpsq1990
                               physicpc   ->   physicpc1987 physicpc1990
                                 lphypc   ->   lphypc1987 lphypc1990
                                 lpopul   ->   lpopul1987 lpopul1990
-----------------------------------------------------------------------------
```

Organizing panel data (concluded)

All the *x* variables are expanded to *T* (=2) and given suffixes specific to the years to which they correspond.

Were you to encounter the data set in the wide format and wish to convert it to long, the command for doing so would be:

reshape long lowbrth infmort afdcprt popul pcinc physic afdcprc d90 lpcinc lphysic afdcpay afdcinc lafdcpay beds bedspc lbedspc povrate afdcpsq physicpc lphypc lpopul, i(id) j(year),

with the only notable differences being that “long” has replaced “wide” as the desired format and all of the variables must be listed individually.

The long format is usually preferable because it enables you to use the “xt” settings in Stata, which make using lags, leads and differences easier, as well as performing fixed effects regressions generally.

Policy analysis with two-period panel data

Performing a program evaluation, i.e., measuring the effect of a policy, with panel data can be performed like Differences in Differences (DD).

The major advantage of using panel data is the non-necessity of aggregating observations to make temporal differencing relevant.

- The agents are already observed before and after the program implementation, so differencing of the data can be done at the “micro” level.

Furthermore participation being involuntary, e.g., because of a law applying to everyone within a county or state, or because participation is assigned by a lottery, is no longer a requisite.

Policy analysis with two-period panel data (continued)

Participation in the program is allowed to be correlated with individuals' fixed effects because the fixed effects will be negated by differencing.

Examples:

- More (less) productive firms can be more likely to participate in a job training program without biasing the *differenced* model's estimates of training's effect on productivity.
- States with more (less) prevalence of drunk driving can be more likely to enact drunk driving laws without biasing the *differenced* model's effect of laws on traffic fatalities.
- In a study measuring the effect of a voluntary personal finance class on saving behavior, the participation can be positively (negatively) correlated with individuals' pre-class frugality without biasing the *differenced* model's effect on saving.

Policy analysis with two-period panel data (concluded)

A model that enables a researcher to overcome unobserved fixed effects by differencing is:

$$y_{it} = \beta_0 + \delta_0 1[t = 2] + \beta_1 prog_{it} + a_i + u_{it}, \text{where}$$

$$1[t = 2] \text{ is an indicator function for the 2nd period, and}$$
$$prog_{it} \text{ is the indicator for participation } (prog_{it} = 1).$$

Assuming that participation occurs for a subset of the sample and only in period 2, estimating the differenced model,

$$\Delta y_{it} = \delta_0 1[t = 2] + \beta_1 \Delta prog_{it} + \Delta u_{it},$$

would yield an estimate of β_1 that is unbiased by program participation's correlation with a_i.

Differencing with more than two time periods

The method of differencing can be generalized to $T \geq 2$ periods, primarily by accounting for intercepts specific to each time period, i.e.,

$$y_{it} = \delta_1 + \sum_{t=2}^{T} \delta_t 1[time = t] + \sum_{j=1}^{k} \beta_j x_{itj} + a_i + u_{it}.$$

The strict exogeneity assumption ("FD.4" in Wooldridge appendix) must also be generalized to include independence between each combination of time period:

$$FD.4 \rightarrow E\left(u_{it} \middle| x_{isj}, a_i\right) = 0, \forall\, j, t, s.$$

Differencing with more than two time periods (continued)

Along with FD 1-3 (model specification, random sampling, and the rank condition), this is what is necessary for the first difference estimator to be unbiased and consistent.

- It is consistent under an even weaker version of FD.4.

Inference about first differenced estimates can be complicated by the possibility that the errors (Δu_{it}) in the transformed model can be serially correlated, e.g.,

$$E(\Delta u_{i2} * \Delta u_{i1}) = E[(u_{i2} - u_{i1})(u_{i1} - u_{i0})] = -\sigma^2; u_{it} \sim i.i.d(0, \sigma^2).$$

Differencing with more than two time periods (concluded)

A thorough discussion of how serial correlation in the errors may be detected and remedied is inappropriate for this class, however, software such as Stata has options to make the standard errors robust to serial correlation within a cluster ("i").

In Stata the language should go at the end of a regression command.

- The syntax looks like this:

reg *yvar listofregressors*, vce(cluster *id*),

where *id* is a variable that uniquely identifies each element in the cross section with a different value.

Conclusion

Panel data can be enormously helpful in empirical applications in which biased estimators arise because of unobserved effects.

Unfortunately panel data is still somewhat rare, partly because it is expensive to track many individuals accurately over a period of time.

Since cross sectional data is more common, there is still plenty of need for methods that can be used to counteract biases in single cross section samples.

- The next method we will study (instrumental variables) is an example.