Econometrics

Chapter 13: Simple Panel Data Methods

In Choi

Sogang University

A True Panel vs. A Pooled Cross Section

- Often loosely use the term panel data to refer to any data set that has both a cross-sectional dimension and a time-series dimension.
- More precisely it's only data following the same cross-section units over time. Otherwise it's a pooled cross-section. (e.g., random sample on hourly wages, education, experieince, etc. collected in every year)

Pooled Cross Sections

- We may want to pool cross sections just to get bigger sample sizes.
- We may want to pool cross sections to investigate the effect of time.
- We may want to pool cross sections to investigate whether relationships have changed over time.

Pooled Cross Sections

Example

(Effects of time on fertility; FERTIL1.RAW used; Wooldrige p.429)
Model

$$\begin{aligned} y_{it} &= \beta_0 + \beta_1 x_{1it} + \cdots + \beta_k x_{kit} + \delta_{74} d_{74} + \delta_{76} d_{76} \\ &\quad + \delta_{78} d_{78} + \delta_{80} d_{80} + \delta_{82} d_{82} + \delta_{84} d_{84} + u_{it}. \end{aligned}$$

The data starts from 1972, but the year dummy for 1972 is not included to avoid the problem of multicollinearity. The year dummies for 1982 and 1984 are significant.

Pooled Cross Sections

Dep var: # of Kinds		
Indep var.	*Coef.*	*s.e.*
edu	*-0.128*	*0.018*
age	*0.532*	*0.138*
age^2	*-0.0058*	*0.0016*
black	*1.076*	*0.174*
east	*0.217*	*0.133*
northern	*0.363*	*0.121*
west	*0.198*	*0.167*
farm	*-0.053*	*0.147*
othrural	*-0.163*	*0.175*
town	*0.084*	*0.124*
smcity	*0.212*	*0.160*
y74	*0.268*	*0.173*
y76	*-0.097*	*0.179*
y78	*-0.069*	*0.182*
y80	*-0.071*	*0.183*
y82	*-0.522*	*0.172*
y84	*-0.545*	*0.175*
constant	*-7.742*	*3.052*
n=1,129;	$R^2 = 0.1295$	$\bar{R}^2 = 0.1162$

Pooled Cross Sections

Example

(Changes in the return to education and the gender wage gap)
Data: pooled cross section data from years 1978 and 1985, CPS78_85.RAW
Model

$$\begin{aligned}\log(wage) &= \beta_0 + \delta_0 y85 + \beta_1 edu + \delta_1 y85 \cdot edu + \beta_2 \exp er \\ &\quad + \beta_3 \exp er^2 + \beta_4 union + \beta_5 female + \delta_5 y85 \cdot female + u\end{aligned}$$

(i) Parameter δ_1 measures how the return to another year of education has changed over the seven-year period. If $\delta_1 > 0$, higher return in 1985.
(ii) Parameter δ_5 measure how the gender gap has changes over the years. If $\delta_5 > 0$, it means decrease in the gender gap over the years.

Pooled Cross Sections

Estimation results

Indep var	Coef	s.e.
y85	0.118	0.124
educ	0.0747	0.0067
y85·educ	0.0185	0.0094
exper	0.0296	0.0036
$exper^2$	-0.00040	0.00008
union	0.202	0.030
female	-0.317	0.037
y85·female	0.085	0.051
constant	0.459	0.093
n=1,084	$R^2 = 0.427$	$\bar{R}^2 = 0.422$

1. Return to education in 1978 is about 7.5%, while that in 1985 is about 9.35%.
2. The gender gap has fallen over the years.

Advantages of using panel data

- Large number of data points – better efficiency
- Panel data allow a researcher to study a number of important economic questions that cannot be addressed using cross-sectional or time series data sets.

Difference-in-Differences

- Suppose that a new policy was initiated at the end of period 1 that affects only one group (treatment group), while it does not affect another group (control group). Data for these groups in two periods (1 and 2) are known. How can we analyze the effect of the policy over time?

Difference-in-Differences

- If $y_{i,t}$ denotes the effect of the policy for group i in period t, the effect of the policy over time can be measured by

$$(y_{B,2} - y_{A,2}) - (y_{B,1} - y_{A,1}),$$

(changes over time in the group differences)

or equivalently

$$(y_{B,2} - y_{B,1}) - (y_{A,2} - y_{A,1}).$$

(changes over group in the time differences)

This is called the difference (in time) -in-differences (between the groups).

- A regression framework using time and treatment dummy variables can calculate this difference-in-difference as well.

Difference-in-Differences

- Consider the model for $t = 1, 2$ and $i = A, B$:

$$\begin{aligned} y_{it} &= \beta_0 + \delta_0 d2_t + \beta_1 dB_i + \delta_1 d2_t * dB_i \\ &\quad + u_{it}; \\ d2_t &= \begin{cases} 0 \text{ for the first period} \\ 1 \text{ for the second period} \end{cases} \\ dB_i &= \begin{cases} 0 \text{ for the control group} \\ 1 \text{ for the treatment group} \end{cases} \end{aligned}$$

  The estimated δ_1 will be the difference-in-differences in the group means, and it measures the effect of policy change. In fact,

$$\hat{\delta}_1 = (\bar{y}_{B,2} - \bar{y}_{A,2}) - (\bar{y}_{B,1} - \bar{y}_{A,1}).$$

- Additional regressors can be added to the regression to control for differences across the treatment and control groups
- Sometimes referred to as a "natural experiment" especially when a policy change is being analyzed.

Difference-in-Differences

Example

(Effects of nuclear power plant on housing prices)
Model

$$\ln(rprice_{it}) = \beta_0 + \delta_0 d2_t + \beta_1 dB_i + \delta_1 d2_t * dB_i + other\ variables_{it} + u_{it};$$

$$d2_t = \begin{cases} 0 \text{ before the power plant} \\ 1 \text{ after the power plant} \end{cases}$$

$$dB_i = \begin{cases} 0 \text{ houses far away from the power plant} \\ 1 \text{ houses close to the power plant} \end{cases}$$

(i) Parameter δ_1 measures the effect of the nuclear power plant on the real housing prices in percentage terms (when multiplied by 100).
(ii) Other variables need to be included such as house size, lot size, age of house, school district, distance to public transportation, distance to public facilities, etc.

Two-Period Panel Data

- It's possible to use a panel just like pooled cross-sections, but can do more than that.
- Panel data can be used to address some kinds of omitted variable bias.

Unobserved Fixed Effects

- Suppose the population model is

$$y_{it} = \beta_0 + \beta_1 x_{it1} + \ldots + \beta_k x_{itk} + \mu_i + u_{it}.$$

- Here we have added a time-constant component to the error.
- Assume x and u are not correlated.
- If μ_i is correlated with the x's, OLS will be biased, since μ_i is part of the error term.
- With panel data, we can difference-out the unobserved fixed effect.

First-differences

- We can subtract one period from the other, to obtain

$$\Delta y_{it} = \beta_1 \Delta x_{i1t} + \ldots + \beta_k \Delta x_{ikt} + \Delta u_{it}$$

- This model has no correlation between the x's and the error terms, so no bias.
- This model is the one cross section data because there are only two periods.

First-differences

Example

(Earnings equation for two periods)
Model

$$\log(wage_{it}) = \beta_0 + \delta_0 d2_t + \beta_1 educ_{it} + \mu_i + u_{it}$$

Here μ_i is unobserved ability that is correlated with $educ_{it}$. The differenced model is

$$\Delta \log(wage_{it}) = \delta_0 + \beta_1 \Delta educ_{it} + \Delta u_{it}.$$

The problem of this model is $\Delta educ_{it}$ is zero for most adults.

First-differences

Example

(Sleep and working)
Data: SLP75_81.RAW
Model

$$\begin{aligned} slpnap_{it} &= \beta_0 + \delta_0 d81_t + \beta_1 totwrk_{it} + \beta_2 educ_{it} \\ &\quad + \beta_3 marr_{it} + \beta_4 yngkid_{it} + \beta_5 gdhlth_{it} + \mu_i + u_{it}, \end{aligned}$$

where
$slpnap_{it}$: minutes slept including nap
$totwrk_{it}$: minutes worked per week
$educ_{it}$: years of education
$marr_{it}$ =1 if married in year t
$yngkid_{it}$ =1 if child < 3 in year t
$gdhlth_{it}$ =1 if good health in t

First-differences

Differencing across two years gives

$$\begin{aligned}\Delta slpnap_i &= \delta_0 + \beta_1 \Delta totwrk_i + \beta_2 \Delta educ_i \\ &\quad + \beta_3 \Delta marr_i + \beta_4 \Delta yngkid_i + \beta_5 \Delta gdhlth_i + \Delta u_i.\end{aligned}$$

Estimation results

Indep var	Coef	s.e.
$\Delta totwrk$	-0.227	0.036
$\Delta educ_i$	-0.024	48.759
$\Delta marr_i$	104.21	92.86
$\Delta yngkid_i$	94.67	87.65
$\Delta gdhlth_i$	87.58	76.60
constant	-92.63	45.87
n=239	$R^2 = 0.150$	

1. One more hour of work is associated with 0.227×60=13.62 less minutes of sleeping.
2. All other variables are insignificant.

Differencing with Multiple Periods

- Can extend this method to more periods.
- Simply difference adjacent periods.
- So if 3 periods, then subtract period 1 from period 2, period 2 from period 3 and have 2 observations per individual.
- Simply estimate by OLS, assuming the Δu_{it} are uncorrelated over time.