Econ 582
Introduction to Pooled Cross Section and Panel Data

Eric Zivot

May 22nd, 2012

Outline

- Pooled Cross Section and Panel Data
- Analysis of Pooled Cross Section Data
- Two Period Panel Data
- Multi-period Panel Data

Pooled Cross Section and Panel Data

Definition 1 *(Pooled cross-section data) Randomly sampled cross sections of individuals at different points in time*

Example: Current population survey (CPS) in 1978 and 1988

Definition 2 *(Panel Data) Observe cross sections of the same individuals at different points in time*

Example: National Longitudinal Survey of Youth (NLSY)

Pooled Cross Section Data

- Pooling makes sense if cross sections are randomly sampled (like one big sample)
- Time dummy variables can be used to capture structural change over time
- Observations across different time periods allows for policy analysis

Example: Women's fertility over time (Wooldridge)

National Opinion Research Center's General Social Survey for even years from 1972-1984

$$\begin{aligned} kids_i &= \delta_0 + \delta_1 d74_i + \cdots + \delta_6 d84_i + \boldsymbol{\beta}'\mathbf{x}_i + \varepsilon_i \\ d74_i &= 1 \text{ if year} = 74, 0 \text{ otherwise (year dummy)} \\ \mathbf{x}_i &= (educ_i, age_i, age_i^2, black_i, east_i, \ldots, smcity_i) \end{aligned}$$

Q: After controlling for observable factors (educ etc), what has happened to fertility over time?

A: Time effects of fertility are captured by dummy variables

$$\begin{aligned} E[kids_i|\mathbf{x}_i, year &= 72] = \delta_0 + \boldsymbol{\beta}'\mathbf{x}_i \\ E[kids_i|\mathbf{x}_i, year &= 74] = \delta_0 + \delta_1 + \boldsymbol{\beta}'\mathbf{x}_i \end{aligned}$$

$$E[kids_i|\mathbf{x}_i, year = 74] - E[kids_i|\mathbf{x}_i, year = 72] = \delta_1$$

Hence, δ_1 = change in fertility between 1972 and 1974 controlling for $\mathbf{x}_i$.

Some complications:

- $var(\varepsilon_i)$ may change over time. Best to use HC standard errors
- Other coefficients may not be constant over time

Example cont'd

To allow coefficients on $\mathbf{x}_i$ to vary over time, add interaction terms with the dummy variable:

$$\begin{aligned} kids_i &= \delta_0 + \delta_1 d74_i + \cdots + \delta_6 d84_i + \boldsymbol{\beta}'\mathbf{x}_i \\ &\quad \boldsymbol{\gamma}_1' (d74_i \times \mathbf{x}_i) + \cdots + \boldsymbol{\gamma}_6 (d84_i \times \mathbf{x}_i) + \varepsilon_i \end{aligned}$$

Then

$$\begin{aligned} E[kids_i|\mathbf{x}_i, year &= 72] = \delta_0 + \boldsymbol{\beta}'\mathbf{x}_i \\ E[kids_i|\mathbf{x}_i, year &= 74] = \delta_0 + \delta_1 + (\boldsymbol{\beta} + \boldsymbol{\gamma}_1)'\mathbf{x}_i \end{aligned}$$

and

$$E[kids_i|\mathbf{x}_i, year = 74] - E[kids_i|\mathbf{x}_i, year = 72] = \delta_1 + \boldsymbol{\gamma}_1'\mathbf{x}_i$$

Testing for Structural Change (Chow Test)

H_0 : (no structural change) $\delta_1 = \cdots = \delta_6 = 0$ and $\boldsymbol{\gamma}_1 = \cdots = \boldsymbol{\gamma}_6 = \mathbf{0}$

H_1 : (structural change) some $\delta_i \neq 0$ and/or $\boldsymbol{\gamma}_i \neq 0$

- Use F-test or Wald test
- Advisable to correct for possible heteroskedasticity

Policy Analysis with Pooled Cross Section Data

- Pooled cross-sections can be useful for evaluating the impact of certain events or policy interventions
- Event or policy intervention must be a "natural experiment" - i.e., must be exogenously imposed on data
- Control variable must be exogenous (no endogenous regressors)

Example: Effect of Garbage Incinerator Location on House Values in North Andover MA

- 2 year pooled cross section of data for 1978 and 1981
- New incinerator built in 1981 and online in 1985
- Knowledge of incinerator project not known in 1978
- Q: Did house values near the incinerator decline in value?

Regression using 1981 data

$$\begin{aligned} rprice_i &= \gamma_0 + \gamma_1 nearinc_i + u_i \\ &= \underset{(3{,}093)}{101{,}307} - \underset{(5{,}827)}{30{,}688} \cdot nearinc_i \\ nearinc_i &= 1 \text{ if near incinerator, 0 otherwise} \\ n &= 142,\ R^2 = 0.665 \end{aligned}$$

Note

$$\begin{aligned} E[rprice_i|nearinc_i &= 1 \text{ in } 1981] - E[rprice_i|nearinc_i = 0 \text{ in } 1981] \\ &= \gamma_1 = -30{,}688 \end{aligned}$$

Regression using 1978 data

$$\begin{aligned} \widehat{rprice_i} &= \underset{(2{,}653)}{82{,}517} - \underset{(5{,}287)}{18{,}824} \cdot nearinc_i \\ n &= 142,\ R^2 = 0.665 \end{aligned}$$

Note

$$\begin{aligned} E[rprice_i|nearinc_i &= 1 \text{ in } 1978] - E[rprice_i|nearinc_i = 0 \text{ in } 1978] \\ &= -18{,}824 \end{aligned}$$

so that it appears that the incinerator was build in a low income/house value area.

Difference in Differences (Diff-in-Diff) Estimate

To determine the impact of the incinerator on house values, we need to compare the differences between the treatment and control groups across the two time periods (compute the difference in the difference)

$$\begin{aligned} E[rprice_i|nearinc_i &= 1 \text{ in } 1981] - E[rprice_i|nearinc_i = 0 \text{ in } 1981] \\ -E[rprice_i|nearinc_i &= 1 \text{ in } 1978] - E[rprice_i|nearinc_i = 0 \text{ in } 1978] \\ &= -30,688 - (-18,824) \\ &= -11,863 \end{aligned}$$

Dummy Variable Formulation of Diff-in-Diff Estimation

$$rprice_i = \beta_0 + \delta_0 d81_i + \beta_1 nearinc_i + \delta_1 (d81_i \times nearinc_i) + \varepsilon_i$$

Then

$$\begin{aligned} E[rprice_i|nearinc_i &= 1,\ d81_i = 1] = \beta_0 + \delta_0 + \beta_1 + \delta_1 \\ E[rprice_i|nearinc_i &= 0,\ d81_i = 1] = \beta_0 + \delta_0 \\ Diff_{81} &= \beta_1 + \delta_1 \\ E[rprice_i|nearinc_i &= 1,\ d81_i = 0] = \beta_0 + \beta_1 \\ E[rprice_i|nearinc_i &= 0,\ d81_i = 0] = \beta_0 \\ Diff_{78} &= \beta_1 \\ Diff_{81} - Diff_{78} &= \delta_1 \end{aligned}$$

Dummy variable regression results

$$\begin{aligned}
\widehat{rprice}_i &= \underset{(2{,}726)}{82{,}517} + \underset{(4{,}050)}{18{,}790} \times d81_i - \underset{(4{,}875)}{18{,}824} \times nearinc_i \\
&\quad \underset{(7{,}456)}{-11{,}863} \times d81_i \times nearinc_i \\
\hat{\delta}_1 &= -11{,}863 = Diff_{81} - Diff_{78} \\
t_{\delta_1=0} &= \frac{-11{,}863}{7{,}456} = 1.59
\end{aligned}$$

Note: Dummy variable formulation allows the standard error on $\hat{\delta}_1$ to be computed.

Natural Experiment

- Some exogenous event (e.g., change in government policy) changes the environment in which individuals, families, firms, cities, etc., operate
- Control group is not affected by the policy change
- Treatment group is thought to be affected by the policy change
- No random assignment to control and treatment groups

Group comparison

Group	Period 1	Period 2	
Control	before	after	Diff
Treatment	before	after	Diff
			Diff in Diff

Two Period Panel Data

- Observe cross section on the same individuals, cities, countries etc., in two time periods $t = t_1$ and $t = t_2$
- Panel data structure makes it possible to deal with certain types of endogeneity without the use of exogenous instruments
- Extends the natural experiment framework to situations in which there may be endogeneity

Example: Determine the effect of the unemployment rate on crime rates (Wooldridge)

Data on crime rates and unemployment for 46 cities for 1982 and 1987

Regression for 1982

$$\begin{aligned} \widehat{crmrte}_i &= \underset{(20.76)}{128.38} - \underset{(3.42)}{4.16} \times \textit{umemp}_i \\ n &= 46,\ R^2 = 0.033 \end{aligned}$$

- It appears that increases in unemp lowers crime rate (but not significant) !

- Bias likely due to omitted variables (unemp is endogenous)

Error Components Framework for Two Period Panel Data

$$\begin{aligned} y_{it} &= \beta_0 + \delta_0 d2_t + \boldsymbol{\beta}'\mathbf{x}_{it} + \varepsilon_{it},\ t = t_1, t_2 \\ &= \beta_0 + \delta_0 d2_t + \boldsymbol{\beta}'\mathbf{x}_{it} + (a_i + u_{it}) \\ d2_t &= 1 \text{ if } t = t_2; 0 \text{ otherwise} \\ a_i &= \text{unobserved heterogeneity (fixed effect)} \\ u_{it} &= \text{idiosyncratic error} \end{aligned}$$

- a_i represents unobserved omitted variables that vary across individuals but stay fixed over time (e.g., race, gender, ability)

- $\mathbf{x}_{it}$ is endogenous if it is correlated with a_i and pooled OLS is biased and inconsistent

Example: Pooled OLS estimates in crime rate regression

$$\begin{aligned} \widehat{crmrte}_{it} &= \underset{(12.74)}{93.42} + \underset{(7.98)}{7.94} \times d87_t + \underset{(1.188)}{.427} \times unemp_{it} \\ n &= 92\ (46 \times 2),\ R^2 = 0.012 \end{aligned}$$

- unemp is not significant in pooled regression
- It is likely that unemp is endogenous; e.g., correlated with omitted time invariant city specific demographic variables like age, race, education levels, attitudes towards crime etc.

Eliminating Endogeneity in Two Period Panel Data

$$y_{it} = \beta_0 + \delta_0 d2_t + \boldsymbol{\beta}'\mathbf{x}_{it} + a_i + u_{it},\ t = t_1, t_2$$

Then

$$\begin{aligned} t &= t_1 : y_{it_1} = \beta_0 + \boldsymbol{\beta}'\mathbf{x}_{it_1} + a_i + u_{it_1} \\ t &= t_2 : y_{it_2} = \beta_0 + \delta_0 + \boldsymbol{\beta}'\mathbf{x}_{it_2} + a_i + u_{it_2} \\ diff &: \Delta y_{it} = \delta_0 + \boldsymbol{\beta}'\boldsymbol{\Delta}x_{i2t} + \Delta u_{it_2} \end{aligned}$$

- First differencing eliminates the unobserved fixed effect a_i!
- OLS on first differenced data gives consistent estimates of $\boldsymbol{\beta}$ (provided $\boldsymbol{\Delta}x_{i2t}$ is uncorrelated with Δu_{it_2})

Example: First Difference Estimates in crime rate regression

$$\begin{aligned} \Delta\widehat{crmrte}_{it} &= \underset{(4.70)}{15.40} + \underset{(0.88)}{2.22}\,\Delta unemp_{it} \\ n &= 46,\ R^2 = .127 \\ t_{\beta=0} &= \frac{2.22}{0.88} = 2.52 \end{aligned}$$

- coef on $\Delta unemp$ is of expected sign and is significant

Potential Problems with First Difference Regression

- First differencing removes variables that don't vary with time (e.g. gender, race, etc.)

- Effective sample size is reduced

Policy Analysis with Two-Period Panel Data

- Two period panel data is often used for program evaluation studies in which there is likely to be endogeneity

Example: Evaluation of Michigan Job Training Program

- Data for two years (1987 and 1988) on the same manufacturing firms in Michigan

- Some firms received job training grants in 1988 and some did not (training was available on first come first serve basis)

Panel data regression

$$
\begin{aligned}
scrap_{it} &= \beta_0 + \delta_0 \times d88_t + \beta_1 grant_{it} + a_i + u_{it} \\
scrap_{it} &= \text{scrap rate (\% of items scrapped due to defects)} \\
grant_{it} &= \text{1 if firm } i \text{ received a training grant in 1988} \\
a_i &= \text{unobserved firm fixed effects (e.g. worker productivity)} \\
cov(grant_{it}, a_i) &\neq \text{0 (why?)}
\end{aligned}
$$

First Difference transformation

$$
\begin{aligned}
\Delta scrap_{it} &= \delta_0 + \beta_1 \Delta grant_{it} + \Delta u_{it} \\
&= \delta_0 + \beta_1 grant_{i88} + \Delta u_{it}
\end{aligned}
$$

Here, $\beta_1 =$ "average treatment effect"

$$\begin{aligned} Diff_t &= E[scrap_{i,88}|grant_{i,88} = 1] - E[scrap_{87}|grant_{i,88} = 1] = \delta_0 \\ Diff_c &= E[scrap_{i,88}|grant_{i,88} = 0] - E[scrap_{87}|grant_{i,88} = 0] = \delta_0 \\ Diff_t - Diff_c &= \beta_1 \end{aligned}$$

Example: First Differences Regression

$$\begin{aligned} \Delta\widehat{scrap}_{it} &= \underset{(.405)}{-.564} - \underset{(.683)}{.739}\,\Delta grant_{it} \\ n &= 54,\ R^2 = .022 \\ t_{\beta_1=0} &= \frac{-.739}{.683} = 1.08 \end{aligned}$$

Panel Data with More than 2 Time Periods

Suppose $t = t_1, t_2$ and t_3

$$\begin{aligned} y_{it} &= \delta_1 + \delta_2 d2_t + \delta_3 d3_t + \boldsymbol{\beta}'\mathbf{x}_{it} + a_i + u_{it} \\ d2_t &= 1 \text{ if } t = t_2; 0 \text{ otherwise} \\ d3_t &= 1 \text{ if } t = t_3; 0 \text{ otherwise} \end{aligned}$$

Then

$$\begin{aligned} t &= t_1 : y_{it_1} = \delta_1 + \boldsymbol{\beta}'\mathbf{x}_{it_1} + a_i + u_{it_1} \\ t &= t_2 : y_{it_2} = \delta_1 + \delta_2 + \boldsymbol{\beta}'\mathbf{x}_{it_2} + a_i + u_{it_2} \\ t &= t_3 : y_{it_3} = \delta_1 + \delta_3 + \boldsymbol{\beta}'\mathbf{x}_{it_2} + a_i + u_{it_3} \end{aligned}$$

First differencing gives

$$\Delta y_{it} = \delta_1 + \delta_2 \Delta d2_t + \delta_3 \Delta d3_t + \boldsymbol{\beta}'\boldsymbol{\Delta} x_{i2t} + \Delta u_{it_2},\ t = t_2, t_3$$

That is,

$$\begin{aligned} t &= t_2 : \Delta y_{i2} = \delta_2 + \boldsymbol{\beta}'\boldsymbol{\Delta} x_{i2t} + \Delta u_{it_2} \\ t &= t_3 : \Delta y_{i3} = -\delta_2 + \delta_3 + \boldsymbol{\beta}'\boldsymbol{\Delta} x_{i2t} + \Delta u_{it_3} \end{aligned}$$

because

$$\Delta d2_3 = d2_3 - d2_2 = -1$$

- Estimation is by pooled OLS on first differenced data

- Error terms for a given i are correlated across time

$$\begin{aligned} cov(\Delta u_{it_3}, \Delta u_{it_2}) &= cov(u_{it_3} - u_{it_2}, u_{it_2} - u_{it_1}) \\ &= -var(u_{it_2}) \end{aligned}$$

  Hence, Gauss-Markov assumptions are violated and OLS is not efficient.