Econometrics 1
Lecture 8: Repeated cross sections, panel data

ENSAE 2014/2015

Michael Visser (CREST-ENSAE)

Outline of lecture

In today's lecture we have a closer look at data with a time dimension

- Repeated cross sections across time (difference in differences estimation)
- Simple panel data (first difference estimation)

Repeated cross-sections

Many data collections are repeated on a regular basis. For example:

- The Current Population Survey (CPS), which randomly samples households each year
- The "Budget de Famille" survey (BdF), which randomly samples households every five years

Each sample constitutes a random sample from the population. Pooling the different random samples therefore gives an independently pooled cross section. This implies in particular that the standard regression techniques studied so far can be used.

Advantages of pooled cross sections:

- increases the sample size (allows to get more precise estimates)
- but, more importantly, they allow us to analyze changes over time

Repeated cross-sections (cntd)

We illustrate this with two examples.

The first example example uses pooled data from 1972 to 1984 (from surveys organized every two years) to study how the fertility of women in the U.S. has changed.

The second example uses pooled data from 1978 and 1985 on wages, education, and gender (and some other variables). The purpose is to examine how the return to education has changed during this period. Another objective is to study whether the gender wage gap has changed over the seven-year period.

Example: Changes in fertility

```
. tab year

       year |      Freq.     Percent        Cum.
------------+-----------------------------------
         72 |         156       13.82       13.82
         74 |         173       15.32       29.14
         76 |         152       13.46       42.60
         78 |         143       12.67       55.27
         80 |         142       12.58       67.85
         82 |         186       16.47       84.32
         84 |         177       15.68      100.00
------------+-----------------------------------
      Total |       1,129      100.00

. sum kids educ age age2 black east northcen west farm othrural town smcity

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        kids |      1129    2.743136    1.653899         0          7
        educ |      1129    12.69088    2.640236         0         20
         age |      1129     43.4845    5.836421        35         54
        age2 |      1129    1924.935    515.8564      1225       2916
       black |      1129     .085031    .2790514         0          1
-------------+--------------------------------------------------------
        east |      1129    .2488928    .4325632         0          1
    northcen |      1129     .319752    .4665871         0          1
        west |      1129    .1080602     .310594         0          1
        farm |      1129    .1984057     .398976         0          1
    othrural |      1129    .1018601    .3025982         0          1
-------------+--------------------------------------------------------
        town |      1129    .3170948    .4655509         0          1
      smcity |      1129     .125775    .3317426         0          1
```

Example: Changes in fertility

```
. reg kids y??
[output omitted]
. estimates store basic

. reg kids y?? educ age* black east northcen west farm othrural town smcity
[output omitted]
. estimates store controls

. estout basic controls, cells("b(fmt(3)) se(par)")

------------------------------------------------------------
                    basic                   controls
                        b            se             b            se
------------------------------------------------------------
y74                 0.182       (0.180)         0.268       (0.173)
y76                -0.223       (0.185)        -0.097       (0.179)
y78                -0.221       (0.188)        -0.069       (0.182)
y80                -0.209       (0.189)        -0.071       (0.183)
y82                -0.622       (0.177)        -0.522       (0.172)
y84                -0.788       (0.179)        -0.545       (0.175)
educ                                           -0.128       (0.018)
age                                             0.532       (0.138)
age2                                           -0.006       (0.002)
black                                           1.076       (0.174)
east                                            0.217       (0.133)
northcen                                        0.363       (0.121)
west                                            0.198       (0.167)
farm                                           -0.053       (0.147)
othrural                                       -0.163       (0.175)
town                                            0.084       (0.125)
smcity                                          0.212       (0.160)
_cons               3.026       (0.130)        -7.742       (3.052)
------------------------------------------------------------
```

Example: Changes in the return to education and the gender wage gap

```
. tab year

       year |      Freq.     Percent        Cum.
------------+-----------------------------------
         78 |         550       50.74       50.74
         85 |         534       49.26      100.00
------------+-----------------------------------
      Total |       1,084      100.00

. reg lwage y85 educ y85educ exper expersq union female y85fem

      Source |       SS       df       MS              Number of obs =    1084
-------------+------------------------------           F(  8,  1075) =   99.80
       Model |  135.992074     8  16.9990092           Prob > F      =  0.0000
    Residual |  183.099094  1075  .170324738           R-squared     =  0.4262
-------------+------------------------------           Adj R-squared =  0.4219
       Total |  319.091167  1083   .29463635           Root MSE      =   .4127

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         y85 |   .1178062   .1237817     0.95   0.341     -.125075    .3606874
        educ |   .0747209   .0066764    11.19   0.000     .0616206    .0878212
     y85educ |   .0184605   .0093542     1.97   0.049      .000106     .036815
       exper |   .0295843   .0035673     8.29   0.000     .0225846     .036584
     expersq |  -.0003994   .0000775    -5.15   0.000    -.0005516   -.0002473
       union |   .2021319   .0302945     6.67   0.000     .1426888    .2615749
      female |  -.3167086   .0366215    -8.65   0.000    -.3885663    -.244851
      y85fem |    .085052    .051309     1.66   0.098    -.0156251     .185729
       _cons |   .4589329   .0934485     4.91   0.000     .2755707     .642295
------------------------------------------------------------------------------
```

Repeated cross-sections (cntd)

Probably the most common use of repeated cross-sections is to analyze the impact of a certain event or policy.

To illustrate how pooled cross-sections can be used to analyze the occurrence of an event, consider the following example.

In 1979 rumors started to spread out about a new garbage incinerator to be constructed in North Andover, a town in Massachusetts.

Construction of the incinerator began in 1981, and the incinerator began operating in 1985.

Kiel and McClain (1995) have data on prices of houses that were sold in 1978 (before rumors began) and in 1981. All prices are in 1978 U.S. dollars.

A house is defined as being near the incinerator if it is located within three miles of the incinerator. Let *nearinc* be the dummy variable indicating whether a house is near the incinerator.

Repeated cross-sections (cntd)

Let us first estimate the following model using data of 1981 only:

$$price = \gamma_0 + \gamma_1 nearinc + u \qquad (1)$$

This gives (standard errors in parentheses):

$$\widehat{(price)} = \underset{(3{,}093.0)}{101{,}307.5} \underset{(5{,}827.71)}{-30{,}688.27}\, nearinc$$

$$n = 142, R^2 = 0.165.$$

We can reject the null hypothesis that $\gamma_1 = 0$ at all usual significance levels (since the t-statistic is larger than 5 in absolute value). In 1981 a house near the incinerator sold for more than 30 thousand less than a house located far away.

Estimating (1) using data of 1978 only gives:

$$\widehat{(price)} = \underset{(2{,}653.79)}{82{,}517.23} \underset{(4{,}744.594)}{-18{,}824.37}\, nearinc$$

$$n = 179, R^2 = 0.082.$$

Repeated cross-sections (cntd)

The indicator *nearinc* is again statistically different from zero (t-statistic almost 4 in absolute value), and the estimate implies that, in 1978, the value of a home near the incinerator is almost 19 thousand less than one located not near the incinerator (almost 83 thousand).

So even before rumors began the houses near the site were valued less! This suggests that the incinerator was built in an area where house prices were lower.

A better way to evaluate the impact of the incinerator on house prices is to look at how the coefficient associated with *nearinc* changed between 1978 and 1981:

$$\hat{\delta}_1 \equiv -30,688.27 - (-18,824.37) = -11,863.9$$

This estimator of the effect of the incinerator is often called the *difference-in-differences* estimator, since it can be written as

$$\hat{\delta}_1 = (\overline{price}_{81,near} - (\overline{price}_{81,far}) - (\overline{price}_{78,near} - (\overline{price}_{78,far})$$

Repeated cross-sections (cntd)

So $\hat{\delta}_1$ is the difference across time in the average difference of house prices in the two areas (far and near the garbage incinerator).

The estimator $\hat{\delta}_1$ can also be obtained by estimating the following model

$$price = \beta_0 + \delta_0 y81 + \beta_1 nearinc + \delta_1 y81 \cdot nearinc + u \qquad (2)$$

using the pooled data set (combining data from 1978 and 1981), and where *y81* is a dummy equal to one if an observation is from 1981 and zero otherwise. This allows us in particular to get the standard error of $\hat{\delta}_1$ and perform tests of the form $H_0 : \delta_1 = 0$.

In model (2) β_0 (resp. $\beta_0 + \beta_1$) is the average price in 1978 of a house located far from (resp. close to) the incinerator.

δ_0 is the increase of average price over time for houses far from the incinerator, and $\delta_0 + \delta_1$ the increase for houses near the incinerator.

β_1 measures the average price effect of the location (the fact that, even in 1978, houses near the future incinerator had different prices compared to houses far from the site).

Repeated cross-sections (cntd)

δ_1 measures the change in average prices due to introduction of the new incinerator.

The difference-in-differences estimator $\hat{\delta}_1$ has this causal interpretation if in particular we make the assumption that, had the incinerator *not* been constructed, prices of both types of houses (far and near the site) would have evolved in the same way between 1978 and 1981.

Estimation by OLS of (2) leads to

$$\widehat{price} = \underset{(2{,}653.79)}{82{,}517.23} + \underset{(4{,}050.07)}{18{,}790.29}\, y81 - \underset{(4{,}744.594)}{18{,}824.37}\ nearinc$$
$$- \underset{(7{,}456.65)}{11{,}863.90}\, y81 \cdot nearinc, \qquad n = 321, R^2 = 0.174.$$

The t-statistic on $\hat{\delta}_1$ is -1.59, so we cannot reject $H_0 : \delta_1 = 0$.

Repeated cross-sections (cntd)

The difference-in-difference (DID) strategy applied in the previous example was probably first pioneered at the end of the 19th century, by a physician called John Snow.

Snow studied cholera epidemics in London in the mid-nineteenth century. He wanted to know whether cholera was transmitted by contaminated drinking water (and not by "bad" air, the prevailing explanation at that time).

Snow compared death rates in districts serviced by two water utilities, and exploited the fact that, in 1852, one utility changed its water supply (it no longer used water from the dirty Thames in central London, but more upriver, in a part were the Thames was cleaner).

In the districts served by the water company that changed its supply the death rates fell more over time than in the districts served by the company that did not change.

Repeated cross-sections (cntd)

In economics and econometrics the DID strategy is nowadays often applied when the data arise from a so-called *natural experiment.*

A natural experiment occurs when some exogenous event modifies the environment of economic agents (consumers, firms, etc.)

The exogenous event is frequently a change in government policy (new traffic laws, new divorce laws, changes in minimum wages, changes in the tax system, etc.).

A natural experiment always has a *control group*, made up of agents that are not affected by the policy change, and a *treatment group* made up of agents that are affected.

Unlike true experiments where control and treatment groups are constituted by random assignment, the groups in natural experiments are determined by the particular policy change.

Repeated cross-sections (cntd)

To account for possible unobserved differences between members of both groups, at least two sets of data are necessary: data before the policy change (say year 1) and data after the policy change (year 2).

The pooled sample can be broken up in four sub-samples: the control group before the change, the control group after the change, the treatment group before the change, and the treatment group after the change.

Let A denote the control group, B the treatment group, dB a dummy equal to one if an observation belongs to group B, and $d2$ a dummy equal to one if an observation is from year 2.

Consider then the regression model

$$y = \beta_0 + \delta_0 d2 + \beta_1 dB + \delta_1 d2 \cdot dB + u \qquad (3)$$

Repeated cross-sections (cntd)

The parameter δ_1 measures the effect of the policy change. It can be shown that the OLS estimator of this parameter, $\hat{\delta}_1$, can be written as:

$$\hat{\delta}_1 = (\overline{y}_{2B} - \overline{y}_{2A}) - (\overline{y}_{1B} - \overline{y}_{1A}) \tag{4}$$

where $(\overline{y}_{2B})$ is the average of the outcome variable for those in group B in year 2, etc. Because of this form, $\hat{\delta}_1$ is called the DID estimator.

Remarks:

- A key identifying assumption is the *common trend assumption:* had there not been a treatment (i.e., no policy change), the average of the outcome variable would have evolved (over time) in the same way for groups A and B.
- The common trend assumption can be tested if multiple cross sections before the policy change are available. The idea is to investigate whether before the implementation of the change there is a common trend over time in the outcome variable of groups A and B.

Repeated cross-sections (cntd)

- When other explanatory variables are added to model (3) (to control for the fact that the sampled populations differ in the two periods), the OLS estimator $\hat{\delta}_1$ can no longer be expressed as (4), but can be interpreted in a similar way.

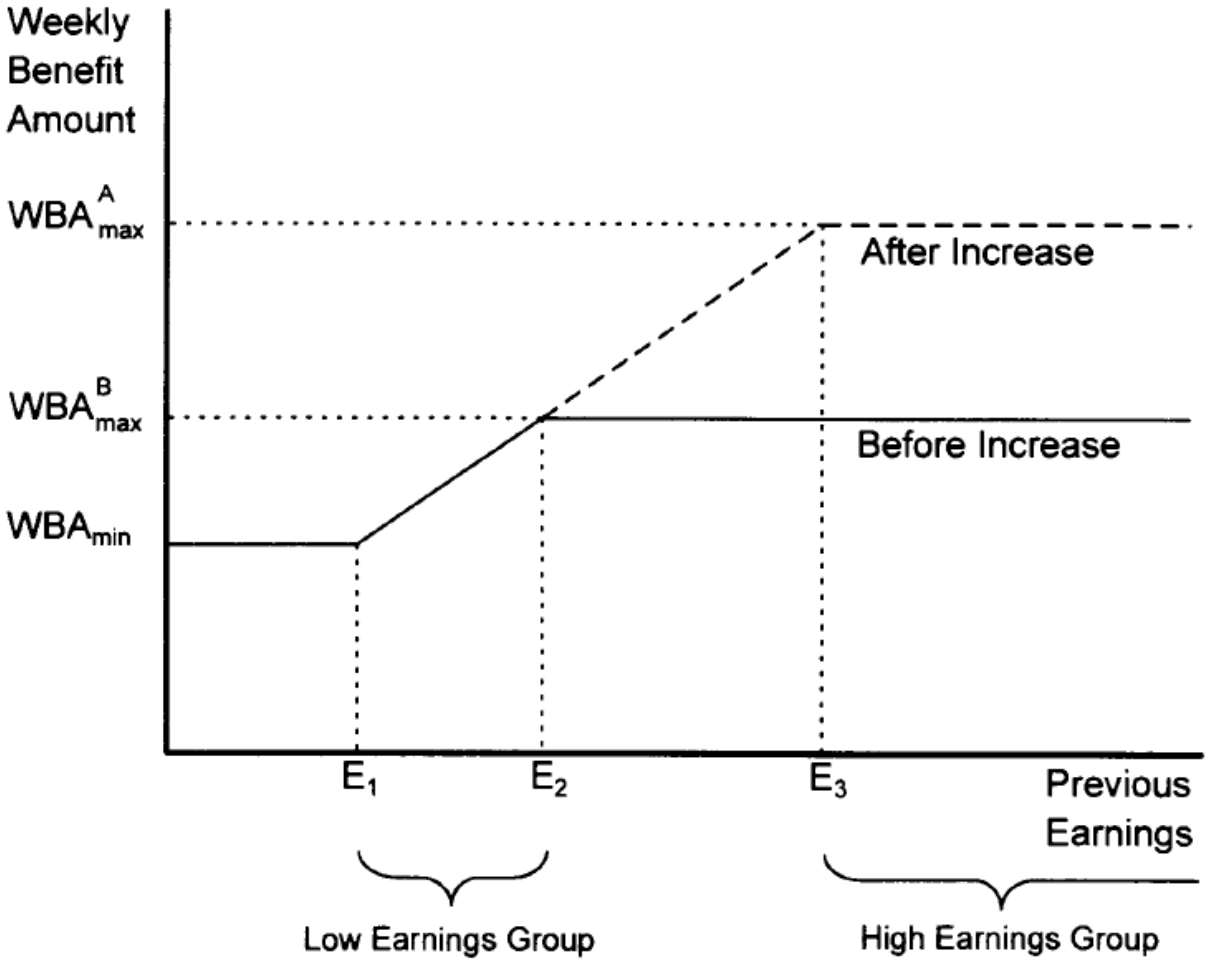
Let us illustrate the DID estimator using data of Meyer, Viscusi, and Durbin (1995).

These researchers studied the length of time in weeks that an injured worker receives workers' compensation (disability payments).

On July 15, 1980, Kentucky raised the cap on weekly earnings that were covered by workers' compensation.

The increase in the cap had no effect on the benefit for low-income workers (control group), but it made it less costly for high-income workers (treatment group) to stay on worker's compensation (see diagram next slide).

Example: Disability Payments



Example: Disability Payments

Let *durat* be the duration the worker remains out of work, *afchnge* the dummy indicating that an observation is after the policy change, and *highearn* the dummy indicating that the individual is a high-income worker.

Let us first consider the regression $log(durat)$ on a constant and *afchnge,* on the high-earning workers only. The results are

$$\widehat{log(durat)} = \underset{(0.035)}{1.414} + \underset{(0.050)}{0.212}\, afchnge$$

$$n = 2,852,\ R^2 = 0.006.$$

The estimate of the coefficient associated with *afchnge* is 0.212, suggesting that the introduction of more generous workers' compensation has increased the duration out of work by 21.2%. The t-statistic is 4.27, we strongly reject the null hypothesis that the effect is zero.

Example: Disability Payments

This is sometimes called the *before-after* estimate of the effect of the increased benefits. If there is a trend in the duration over time (if the duration out of work has changed because of changes in macro-economic factors for instance), the before-after estimate is biased.

The DID estimator takes possible trend effects into account, and can be obtained by regressing $log(durat)$ on a constant, *afchnge*, *highearn*, and $afchnge \cdot highearn$. The results are

$$\widehat{log(durat)} = \underset{(0.031)}{1.126} + \underset{(0.045)}{0.007}\, afchnge + \underset{(0.045)}{0.256}\, highearn + \underset{(0.069)}{0.191}\, afchnge \cdot highearn, \quad n = 5,626,\ R^2 = 0.021.$$

The estimate of the coefficient associated with $afchnge \cdot highearn$ is 0.191, suggesting that the introduction of the new policy has increased the duration out of work by 19.1%. The t-statistic is 2.77, we again strongly reject the null hypothesis that the effect is zero.

Panel data

Panel data are data where we have repeated observations for each unit.

Such data are relatively easy to collect for cities, regions, or countries. For individuals and households, however, it is more complicated and costly to obtain repeated observations.

Example: Suppose we are interested in the relationship between the unemployment rate and crime level in cities and consider the following regression model

$$crime_{it} = \beta_{0t} + \beta_1 unem_{it} + u_{it}$$

We have data on 46 US cities for 1982 and 1987. Let us first perform the cross-sectional regression by OLS, only for the year 1987.

Example: Unemployment and crime

Cross-sectional regression

```
. reg crmrte unem if year==87

      Source |       SS       df       MS              Number of obs =      46
-------------+------------------------------           F(  1,    44) =    1.48
       Model |  1775.90928     1  1775.90928            Prob > F      =  0.2297
    Residual |  52674.6428    44  1197.15097            R-squared     =  0.0326
-------------+------------------------------           Adj R-squared =  0.0106
       Total |  54450.5521    45  1210.01227            Root MSE      =    34.6

------------------------------------------------------------------------------
      crmrte |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        unem |  -4.161134   3.416456    -1.22   0.230    -11.04655     2.72428
       _cons |   128.3781   20.75663     6.18   0.000     86.54589     170.2104
------------------------------------------------------------------------------
```

The estimator of the slope coefficient β_1 is unbiased (and hence the estimate has a causal interpretation) only if

$$E[u_{it}|unem_{it}] = 0$$

Decomposing $u_{it} = a_i + e_{it}$, the necessary condition such that the cross-sectional estimator is unbiased thus becomes

$$E[a_i + e_{it}|unem_{it}] = 0.$$

Example: Unemployment and crime

First differences

With panel data we can estimate a first-differenced equation:

$$\Delta crime_{it} = \Delta\beta_{0t} + \beta_1 \Delta unem_{it} + \Delta e_{it}$$

```
. reg Dcrmrte Dunem

      Source |       SS       df       MS              Number of obs =      46
-------------+------------------------------           F(  1,    44) =    6.38
       Model |  2566.43056     1  2566.43056            Prob > F      =  0.0152
    Residual |  17689.5426    44  402.035059            R-squared     =  0.1267
-------------+------------------------------           Adj R-squared =  0.1069
       Total |  20255.9732    45  450.132737            Root MSE      =  20.051

------------------------------------------------------------------------------
     Dcrmrte |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       Dunem |   2.217996   .8778657     2.53   0.015     .4487745    3.987218
       _cons |   15.40219   4.702116     3.28   0.002     5.925702    24.87869
------------------------------------------------------------------------------
```

This estimator is the first-difference estimator of β_1. It is unbiased under the weaker condition $E[\Delta e_{it}|\Delta unem_{it}] = 0$.

First differencing

The first-difference estimator with two periods is a special case of the so-called fixed effects estimator (which we will study next week).

The errors a_i are called the fixed effects since they represent the unobserved variables of unit i that are fixed over time.

The errors a_i are random variables, and the advantage of fixed effects estimation is that they are allowed to be correlated with the explanatory variables: $E[a_i|unem_{it}] \neq 0$.

The only condition that is required for unbiasedness of the first-difference estimator is: $E[\triangle e_{it}|\triangle unem_{it}] = 0$. Note that the slope parameters can only be estimated if the associated regressors are time-varying.

I understand/can apply...

- Repeated cross-sections
- Before-after comparisons
- Common trend assumption
- DID estimator
- Two-period panel data
- Fixed-effect estimator