Mile stone III

Modelling Panel Data

- The structure of panel data
- Natural experiments
- Simple panel data methods
- Advanced panel data methods

The structure of panel data

Panel data: both time series and cross sectional dimension

- Let Y a variable of interest (e.g. the return of a stock, the industrial production index, wages);

- Assume that this variable is observed for N units (e.g. for different firms, various countries, a group of individuals) over T periods of time (e.g. daily for T days, quarterly for T quarters):

  y_{it} . . . observation of Y in unit i at time t
  $i = 1, \ldots, N$ . . . unit number
  $t = 1, \ldots, T$ . . . observation number

Estimating regression models for panel data

Various types of regressors:

- same time-varying predictor x_t for each unit (exogenous economic conditions)
- unit-specific predictor x_i that does not change over time
- unit-specific, time-varying predictors x_{it}
- endogenous predictors through including lagged values of y_{it} (dynamic panels)

Estimating regression models for panel data

Cross-sectional data for a fixed time point t:

- Consider all N observations $(y_{it}, x_{it}, x_t, x_i)$ for $i = 1, \ldots, N$
- It is not possible to estimate the effect of x_t from the cross-sectional regression model

$$y_{it} = \beta_{0t} + \beta_{1t} x_{it} + \beta_{2t} x_t + \beta_{3t} x_i + u_{it}, \tag{13}$$

because x_t is constant!

Estimating regression models for panel data

Time series data for a fixed unit i:

- Consider all T observations $(y_{it}, x_{it}, x_t, x_i)$ for $t = 1, \ldots, T$
- It is not possible to estimate the effect of x_i from the ,,individual" regression model

$$y_{it} = \beta_{0i} + \beta_{1i} x_{it} + \beta_{2i} x_t + \beta_{3i} x_i + u_{it}, \tag{14}$$

because x_i is constant!

Estimating regression models for panel data

- Joint estimation for the whole panel allows to estimate the effect of x_i and x_t
- It is possible to deal with omitted variable bias
- Panel data have to be distinguished from independently pooled cross-section

Independently pooled cross-sections

Independently pooled cross-section are obtained by random sampling from populations at different points in time.

Advantages:

- increases the number of observations.
- allows us to investigate the effect of time (year dummies).
- allows us to investigate whether relationships have changed over time (interactions of year dummies with explanatory variables).
- is particularly suitable for policy analysis, if we have data collected before and after an event.

Natural Experiments

A natural experiment has a

- control group C not affected by an event (e.g. policy change)
- treatment group T (assumed to be affected by policy change)

In true experiments (like in a medical experiment): random assignment to treatment and control groups. One can then simply compare the change in outcomes across the treatment and control groups to estimate the treatment effect.

Natural Experiments

In natural experiments, systematic differences between control and treatment group must be accounted for.

Therefore, we need at least two periods of data (before/after the event), which breaks our sample down into four groups:

- Control group before/after change
- Treatment group before/after change

Difference-in-difference estimator

The regression model of interest is:

$$y_{it} = \beta_0 + \beta_1 D_i^T + \beta_2 D_t^P + \beta_3 D_t^P D_i^T + (\text{other factors}) + u_{it},$$

where

- D_i^T is a dummy variable, taking the value 1 for the treatment group,
- D_t^P is a dummy variable, taking the value 1 for period 2.

The average treatment effect is equal to β_3.

Difference-in-difference estimator: the simple case

Simple case without control variates:

$$y_{it} = \beta_0 + \beta_1 D_i^T + \beta_2 D_t^P + \beta_3 D_t^P D_i^T + u_{it}.$$

	before	after	after-before
Control	β_0	$\beta_0 + \beta_2$	β_2
Treatment	$\beta_0 + \beta_1$	$\beta_0 + \beta_1 + \beta_2 + \beta_3$	$\beta_2 + \beta_3$
Treatment-Control	β_1	$\beta_1 + \beta_3$	β_3

Difference-in-difference estimator: the simple case

For the simple case, the estimate of β_3 is the difference-in-differences in the group means:

$$\begin{aligned}\hat{\beta}_3 &= (\overline{y}_{2,T} - \overline{y}_{2,C}) - (\overline{y}_{1,T} - \overline{y}_{1,C}) \\ &= (\overline{y}_{2,T} - \overline{y}_{1,T}) - (\overline{y}_{2,C} - \overline{y}_{1,C})\end{aligned}$$

The usual regression framework can be used to estimate the regression parameters, and hence the treatment effect under the presence of control variables.

Panel Data - Pooled OLS estimation

Estimate a pooled regression model using all data $t = 1, \ldots, T$ from all units $i = 1, \ldots, N$, e.g.:

$$y_{it} = \beta_0 + \beta_1 x_{it} + \beta_2 x_t + \beta_3 x_i + u_{it},$$

- All covariates may be vectors instead of scalars!
- To include a yearly dummy, define for $t > 1$ the dummy variable $D_t^j = 1$, iff $t = j$. For T periods, $T - 1$ dummies are included.

Panel Data - Pooled OLS estimation

Pooled OLS estimation assumes that

- the regression coefficients β_{jt} in the cross-sectional regression model (13) are identical for all $t = 1, \ldots, T$, i.e.: $\beta_{jt} \equiv \beta_j$;
- the regression coefficients β_{jt} in the individual regression model (14) are identical for all $i = 1, \ldots, N$, i.e.: $\beta_{ji} \equiv \beta_j$.

Some of these regression parameters may be different across units or change over time!

Unobserved Fixed Effects

Unobserved heterogeneity: an important time-invariant variable $x_i^\star$ is not observed (e.g. ability):

$$y_{it} = \beta_0 + \beta_1 x_{it} + \beta_2 x_t + \beta_3 x_i + \beta_4 x_i^\star + u_{it}.$$

Since $x_i^\star$ is not observed, we cannot estimate β_4, however, we may define a so-called fixed effect a_i for each unit i by $a_i = \beta_4 x_i^\star$. This leads to the fixed-effects model

$$y_{it} = \beta_0 + \beta_1 x_{it} + \beta_2 x_t + \beta_3 x_i + a_i + u_{it}. \tag{15}$$

Unobserved Fixed Effects

If repeated measurements are available for each unit, it is possible to estimate all parameters of interest.

Model (15) may be regarded as following regression model,

$$y_{it} = \beta_0 + \beta_1 x_{it} + \beta_2 x_t + \beta_3 x_i + \tilde{u_{it}}, \tag{16}$$

where the errors $\tilde{u_{it}} = a_i + u_{it}$ have following properties:

Unobserved Fixed Effects

- If the missing covariate $x_i^\star$ is correlated with the other regressors, then $\tilde{u_{it}}$ is correlated with the other regressors and the basic assumption for the unbiasedness of OLS estimation $\mathrm{E}(\tilde{u_{it}}|\mathbf{X}_i)$ is violated $\Rightarrow$ OLS estimation of $\beta_0, \beta_1, \ldots$ from regression model (16) is biased.

- If the missing covariate $x_i^\star$ is uncorrelated with the remaining regressors, then OLS estimation is unbiased, but inefficient, because the residuals $\tilde{u_{it}}$ are correlated across time.

First differencing

First differencing (FD) eliminates the fixed effect, by subtracting observations in subsequent periods:

$$\Delta y_{it} = \beta_1 \Delta x_{it} + \beta_2 \Delta x_t + \Delta \tilde{u_{it}}. \tag{17}$$

Advantages:

- Allows for correlation between the missing covariate and the remaining covariates, i.e. $\mathrm{Cov}(a_i, x_{it}) = \mathrm{Cov}(x_i^\star, x_{it}) \neq 0$.
- FD eliminates the fixed effect and leads to a regression model, where the error term $\Delta \tilde{u_{it}} = \Delta u_{it}$ is not correlated with the remaining regressors.

First differencing

Disadvantages:

- It is not possible to estimate the effect of individual regressors x_i.
- It is not possible to estimate the fixed effects a_i.
- Estimation for certain regression coefficients might be inefficient, if the corresponding time-varying variable shows little, i.e. Δx_{it} is equal or close to 0 for most of the time.

Fixed-effects estimation

Model (15) may be regarded as a large regression model, where a unit specific intercept is estimated for each unit:

$$y_{it} = \beta_0 + \beta_1 x_{it} + \beta_2 x_t + \beta_3 x_i + a_1 D_i^1 + \ldots + a_N D_i^N + u_{it},$$

where $D_i^j = 1$, iff $i = j$, i.e. $D_i^1 = 1$ only for observations from unit 1, etc. Note that either the intercept β_0 has to be removed from the model or the constraint $\sum_i a_i = 0$ has to be imposed.

Advantages:

- It is possible to estimate the fixed effects a_i, i.e. to identify units which are above or below the expected average value.

Fixed-effects estimation

Disadvantages:

- The dimension of the regression parameter $(\beta_0, \ldots, \beta_K, a_1, a_2, \ldots, a_N)$ may be large, if N is large, i.e. the panel contains many units.
- If T is small compared to N, then the standard errors for a_i might be large.
- The fixed effects a_i may not be estimated consistently for fixed T, even if N increases, because with each additional unit a new regression parameter a_i is introduced.

Comparing FE estimation versus FD estimation

- Both estimation methods yield the same estimators, if $T = 2$.
- The two methods are different for $T > 2$.
- Under certain assumptions, both methods yield unbiased estimators and are consistent.
- The relative efficiency depends on assumptions concerning the correlation in the idiosyncratic errors u_{it} in the fixed-effects model (15).

Comparing FE estimation versus FD estimation

- If the original errors u_{it} are uncorrelated across time, then the errors $\Delta u_{it} = u_{it} - u_{i,t-1}$ in the FD regression model (17) are not independent, but correlated. In this case, first differencing (FD) may lead to wrong standard errors and FE estimation is preferable.

- However, if the original errors u_{it} have positive correlation across time, then the correlation of the errors $\Delta u_{it} = u_{it} - u_{i,t-1}$ in the FD regression model (17) is reduced. In this case, first differencing (FD) may be more efficient than FE estimation.