Lecture notes X

Regression with panel data

SW ch. 10

Panel data means that an observation unit (e.g. individual) is observed several times

In an ordinary regression this may lead to correlated error terms, i.e. for the same individual: Repeated observations on the same unit tend to be correlated.

Data structure:

(X_{it}, Y_{it}) where $i=1,\ldots,n$ and $t=1,\ldots,T$
subscript it refers to unit i at time t.
X_{it} is either a single regressor or a vector of regressors:

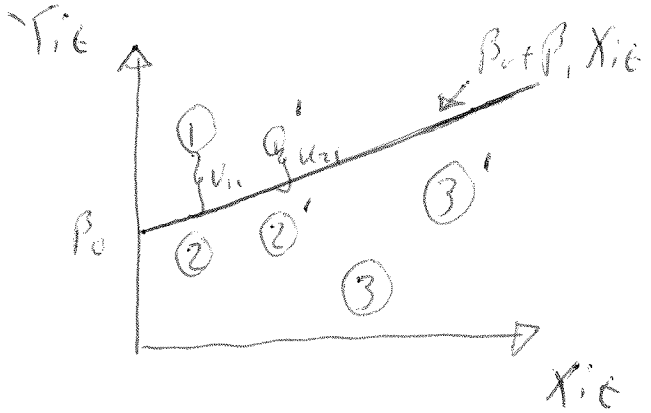
$$X_{it} = (X_{1,it}, \ldots, X_{k,it})$$

Balanced panels: All units observed on all time points
Unbalanced panels: Some units have "missing" data on some time points.

Example :

$$Y_{it} = \beta_0 + \beta_1 X_{it} + U_{it}, \quad \begin{matrix} i=1,\dots,n \\ t=1,\dots,T \end{matrix}$$

Here $Cov(U_{it}, U_{is}) \neq 0$ when $s \neq t$



① - unit i at t=1

①' - unit i at t=2

Assume Cobb-Douglas production function

$$Q_{it} = A_{it} L_{it}^{\beta} K_{it}^{1-\beta}$$

$$\Downarrow$$

$$\ln\left(\frac{Q_{it}}{L_{it}}\right) = (1-\beta)\ln\left(\frac{K_{it}}{L_{it}}\right) + \ln A_{it}$$

$$\rightarrow \quad Y_{it} \equiv \ln \frac{Q_{it}}{L_{it}}$$

$$X_{it} = \ln \frac{K_{it}}{L_{it}}$$

$$\ln A_{it} = \underbrace{\beta_0}_{\sim} + U_{it}$$

U_{it} is the productivity of firm i at t relative to the population mean $E(\ln A_{it})$.

$$\beta_0 = E(\ln A_{it})$$

$$U_{it} = \ln A_{it} - E(\ln A_{it})$$

One can think of U_{it} as an error term that captures unobserved variables that: 1) varies across time for given unit (firm) but also 2) are constant over time for given unit i →

modified regression:

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \varepsilon_{it},$$

where $\varepsilon_{it} = \alpha_i + U_{it}$

captures both the fixed effect (α_i) and the idiosyncratic components (U_{it})

The fixed effect, α_i, captures 2), while the "new" error term, U_{it}, captures 1). It is assumed $Cov(U_{it}, U_{js}) = 0$ if $i \neq j$ or $t \neq s$.

Example 2: R&D subsidies in Written paper-II in appended data set (2003 and 2004)

The fixed effect regression equation:

(1) $Y_{it} = \beta_0 + \beta_1 X_{it} + \alpha_i + u_{it}$

(1) can be formulated within OLS-framework as follows:

Define dummy variables D_{ki} which is 1 if $i=k$ and 0 else:

$$D_{1i} = \begin{cases} 1 & \text{if } i=1 \\ 0 & \text{else} \end{cases}$$

$$D_{2i} = \begin{cases} 1 & \text{if } i=2 \\ 0 & \text{else} \end{cases}$$

$$\vdots$$

$$D_{Ni} = \begin{cases} 1 & \text{if } i=N \\ 0 & \text{else} \end{cases}$$

Then (1) can be written as follows:

(2) $Y_{it} = \beta_0 + \beta_1 X_{it} + \alpha_1 D_{1i} + \alpha_2 D_{2i} + \dots + \alpha_N D_{Ni} + u_{it}$

Problem multicollinearity: $D_{1i} + D_{2i} + \dots + D_{Ni} = 1$

To obtain identification we can either drop an arbitrary ~~individual (e.g. $i=1$)~~ individual-dummy, e.g. D_{1i}, or drop the constant term (β_0).

If we drop constant term, we get

$$Y_{it} = \beta_1 X_{it} + \sum_{k=1}^{N} \alpha_k D_{ki} + u_{it}$$

If we include constant term, but drop D_{1i} we get

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \sum_{k=2}^{N} \tilde{\alpha}_k D_{ki} + u_{it}$$

Thus:

intercept of $i=1$: $\beta_0 = \alpha_1$

intercept of $i=2$: $\beta_0 + \tilde{\alpha}_2 = \alpha_2$

intercept of $i=3$: $\beta_0 + \tilde{\alpha}_3 = \alpha_3$

$\vdots$

$\Rightarrow$ $\tilde{\alpha}_2 = \alpha_2 - \alpha_1$

$\tilde{\alpha}_3 = \alpha_3 - \alpha_1$

$\vdots$

fixed effect $\tilde{\alpha}_i$ denotes change in intercept relative to reference unit $i=1$

Problem with fixed effects regression:
if the regressor X_{it} is constant over time for the same unit i:

$$X_{it} = X_i \quad \text{for } t = 1, \ldots, T$$

i.e. no variation in X_{it} for the same unit i.

Then we will have a multicollinearity problem: $Y_{it} = \beta_1 X_{it} + \alpha_i + u_{it}$

$$\Leftrightarrow \quad Y_{it} = \beta_1 \underbrace{X_{it}}_{= X_i} + \sum_{k=1}^{N} \alpha_k D_{ki} + u_{it}$$

But $\sum_{k=1}^{N} X_k D_{ki} = X_i = X_{it} \Rightarrow$

perfect multicollinearity.

Conclusion: Cannot include regressors that are constant over time in fixed effect regression. Any such effect is captured by the fixed effect α_i

Example: Earnings equation

Y_{it} = ln Earnings for individual i at time t

X_{it} = Years of Schooling

Then if $X_{it} = S_i$ for each unit — i.e. years of schooling does not change during observation period — we cannot estimate the model

$$\underbrace{Y_{it}}_{\text{ln Earnings}} = \beta_1 \underbrace{X_{it}}_{=S_i} + \alpha_i + u_{it}$$

due to collinearity.

~~Near collinearity~~

Example: School-fixed effects

Y_{it} = grades obtained on average in school i at t

X_{it} = teachers per student in that school at t

$$Y_{it} = \beta_1 X_{it} + \alpha_i + u_{it}$$

If $X_{it} = X_i \Rightarrow$ cannot identify β_1 (the effect of class size)

How to estimate a fixed effects model

(1) $Y_{it} = \beta_1 X_{it} + \alpha_i + u_{it}$

Alt. 1: Run OLS on the following equation:

$$Y_{it} = \beta_1 X_{it} + \sum_{k=1}^{N} \alpha_k D_{ki} + u_{it}$$

assuming u_{it} satisfies the usual assumptions

Alt. 2: Subtract the mean from all observations.

(1) $\Rightarrow$ $\bar{Y}_{i\cdot} = \beta_1 \bar{X}_{i\cdot} + \alpha_i + \bar{u}_{i\cdot}$

where $\bar{Y}_{i\cdot} = \frac{1}{T}\sum_{t=1}^{T} Y_{it}$, $\bar{X}_{i\cdot} = \frac{1}{T}\sum_{t=1}^{T} X_{it}$, etc.

$\Downarrow$

$$\underbrace{Y_{it} - \bar{Y}_{i\cdot}}_{=\tilde{Y}_{it}} = \beta_1 \underbrace{(X_{it} - \bar{X}_{i\cdot})}_{\tilde{X}_{it}} + \underbrace{u_{it} - \bar{u}_{i\cdot}}_{\tilde{u}_{it}}$$

Then run OLS on: $\tilde{Y}_{it} = \beta_1 \tilde{X}_{it} + \tilde{u}_{it}$

Note, however, that $\text{Cov}(\tilde{u}_{it}, \tilde{u}_{is}) \neq 0$

Alt. 1 and Alt 2. give same estimates $\hat{\beta}_1^{OLS}$

Regression with fixed time effects

(1) $Y_{it} = \beta_1 X_{it} + \lambda_t + u_{it}$

Different intercept, λ_t, for each period
Then λ_t are called fixed time effects.

Define $D_{st} = \begin{cases} 1 & \text{if } s = t \\ 0 & \text{else} \end{cases}$

(1) can be written as

(2) $Y_{it} = \beta_1 X_{it} + \sum_{s=1}^{T} \lambda_s D_{st} + u_{it}$

(2) can be estimated by OLS.

Note: typically $T \ll N \Rightarrow$ fewer fixed effects than in model with fixed individual effects (α_i)

Note: $\sum_{s=1}^{T} D_{st} = D_{1t} + D_{2t} + \cdots + D_{Tt} = 1$

$\Rightarrow$ cannot have a constant term in the model. Alt.: Use $t=1$ as reference year $\Leftrightarrow$ drop λ_1 from regression

Problem: If X_{it} does not vary over individuals at given t.

~~Assume that~~

That is: $X_{it} = X_t$ for all i

Then:

(1) $$Y_{it} = \beta_1 X_t + \sum_{s=1}^{T} \lambda_s D_{st} + u_{it}$$

and, for all observations (i,t)

$\sum_{s=1}^{T} \alpha_s D_{st} = X_t \Leftrightarrow$ perfect multicollinearity. Conclusion: cannot estimate (1)

Problem: The fixed time effects capture the effect of <u>all</u> variables (observed and unobserved) that are common to <u>all</u> units (but may vary over time)

Models with both fixed time and individual effects

$$Y_{it} = \beta_1 X_{it} + \alpha_i + \lambda_t + u_{it}$$

or equivalently:

$$Y_{it} = \beta_1 X_{it} + \sum_{k=1}^{N} \alpha_k D_{ki} + \sum_{s=1}^{T} \lambda_s D_{st} + u_{it}$$

Not possible that either

(1) $X_{it} = X_i$ for all $t = 1, \ldots, T$ and each i

or that (2) $X_{it} = X_t$ for all $i = 1, \ldots, N$ and each t

Moreover:

$$\sum_{k=1}^{N} D_{ki} = \sum_{s=1}^{T} D_{st} = 1 \quad \text{for all } (i,t)$$

Thus we still have a problem with perfect multicollinearity.

Solution: Drop D_{1t} — i.e. use $t=1$ as reference year $\Leftrightarrow \lambda_1 = 0$

$$Y_{it} = \beta_1 X_{it} + \alpha_i + \lambda_t + u_{it}, \text{ with } \lambda_1 = 0$$

$$\Leftrightarrow Y_{it} = \beta_1 X_{it} + \sum_{k=1}^{N} \alpha_k D_{ki} + \sum_{s=2}^{T} \lambda_s D_{st} + u_{it}$$

Example

$\ln Earnings_{it} = \beta_1 Age_{it} + \beta_2 S_{it} + d_i + l_t + u_{it}$

S_{it} = years of schooling unit i at t

Age_{it} = age of " "

If $S_{it} = S_i$ [does not vary with t] ⇒ cannot identify β_2, i.e. the effect of S_i is captured by d_i

If all units are born in the same year, say τ, then $Age_{it} = t - \tau$ for all i and t

- i.e. all units have same age in the same year (t) ⇒ cannot identify β_1 [effect of Age is captured by l_t]

The fixed effects regression assumptions

Let X_{it} denote a single regressor or a vector of regressors: $X_{it} = (X_{1,it}, X_{2,it}, \ldots, X_{k,it})$

A.1 $E(u_{it} | X_{i1}, \ldots, X_{iT}) = 0$ for all i

A.2 $(X_{i1}, \ldots, X_{iT}, u_{i1}, \ldots, u_{iT})$ for $i = 1, \ldots, n$ are i.i.d vectors

A.3 Large outliers are unlikely: $E X_{it}^4 < \infty$ and $E u_{it}^4 < \infty$

A.4 No perfect multicollinearity

A.5 $Cov(u_{it}, u_{is} | X_{i1}, \ldots, X_{iT}) = 0$ for all i and $t \neq s$ — i.e. no auto-correlation in u_{it}

If A.5 does not hold, ordinary standard errors produced by STATA are not valid. Instead use heteroscedasticity and auto-correlation consistent (HAC) standard errors.