An introduction to econometrics of panel data

F. Legendre

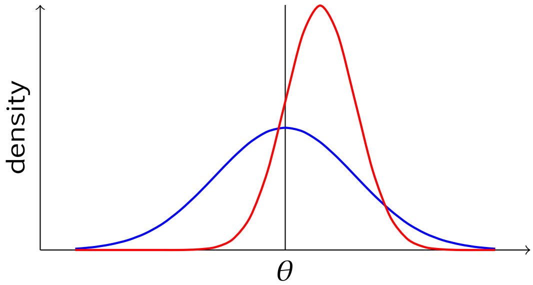
Université Paris-Est

Seminars given at UKMA departement of finance

How to find an estimator? - I

What is a good estimator?

- An good estimator is unbiased: $\mathrm{E}(\hat{\theta}) = \theta$ – it is a weak property



- The variance of a good (unbiased) estimator is lower than the variance of the other (unbiased) estimators
- The variance of a good estimator decreases when the number of observations increases

How to find an estimator? - II

First way: guess and enjoy

- Let us suppose that I want to estimate the mean of a distribution and that I got a sample of size 5 $(x_1, x_2, \ldots, x_5) = (1.80, 1.70, 1.85, 1.78, 1.87)$ independently identically distributed $(\mu, 1)$
- guess $\hat{\theta} = \overline{x} = 1.80$
- $\hat{\theta}$ is unbiased
- $V(\hat{\theta}) = V\left(\frac{1}{5}x_1 + \frac{1}{5}x_2 + \cdots + \frac{1}{5}x_5\right) = \left(\frac{1}{5}\right)^2 (1 + 1 + \cdots + 1) = \frac{1}{5}$
- another guess $\tilde{\theta}$ / throw the lower x_i, throw the higher x_i, then compute the average of the rest $\frac{1.80+1.85+1.78}{3} = 1.81$
- $\tilde{\theta}$ is unbiased
- $V(\tilde{\theta})$ Oops, to hard to evaluate

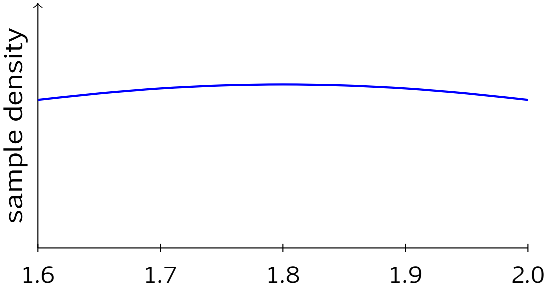
How to find an estimator? - III

Second way: use the estimator of the maximum of the likelihood (but make an assumption on the distribution) Fischer 1922

In the gaussian case, one can get

$$f(x_i) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}(x_i - \mu)^2\right] \quad \forall\, i$$

$$\text{sample density} = f(x_1) \times f(x_2) \times \cdots \times f(x_5)$$



How to find an estimator? - IV

Likelihood of the sample

$$\mathcal{L}(\mu) = \prod_{i=1}^{5} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}(x_i - \mu)^2\right]$$

$$\mathcal{L}(\mu) = \left(\frac{1}{\sqrt{2\pi}}\right)^5 \exp\left[-\frac{1}{2}\sum_{i=1}^{5}(x_i - \mu)^2\right]$$

First order condition (of the log-likelihood)

$$-\sum_{i=1}^{5}(x_i - \mu) = 0$$

So

$$\hat{\theta} = \frac{\sum_{i=1}^{5} x_i}{5} = \overline{x}$$

How to find an estimator? - V

Properties of the ML estimator (be careful for the support)

- $\hat{\theta}$ is asymptotically unbiased
- $\hat{\theta}$ is asymptotically the best

How to find an estimator? - V

ML of the linear model

$$\underline{y} = X\underline{a} + \underline{u} \quad \text{and} \quad \begin{cases} \mathrm{E}(u_i) = 0 \ \forall\, i \\ \mathrm{V}(u_i) = \sigma^2 \ \forall\, i \\ \mathrm{Cov}(u_i, u_j) = 0 \ \forall\, i \neq j \end{cases} \qquad \begin{cases} \mathrm{E}(\underline{u}) = \underline{0} \\ \mathrm{V}(\underline{u}) = \sigma^2 I_N \end{cases}$$

$$\mathcal{L}(\underline{a}, \sigma^2) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma} \exp\left[-\frac{1}{2} \left(\frac{y_i - \underline{X}_i' \underline{a}}{\sigma} \right)^2 \right]$$

$$\mathcal{L}(\underline{a}, \sigma^2) = \left(\frac{1}{\sqrt{2\pi}} \right)^N \left(\frac{1}{\sigma} \right)^N \exp\left[-\frac{1}{2\sigma^2} (\underline{y} - X\underline{a})' (\underline{y} - X\underline{a}) \right]$$

Then you get the OLS estimator

Random versus fixed effect - I

The need to capture individual specificities

$$u_{it} = \mu_i + v_{it}$$

But the random effect needs to be uncorrelated with the explanatory variables
Then, an alternative specification is the fixed effect hypothesis

$$y_{it} = ax_{it} + b_i + u_{it}$$

aka LSDV (least squares dummy variables) as the regression includes N dummy variables

Fixed effects regression

$N = 3$ and $T = 2$

$$\begin{pmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{pmatrix} = \begin{pmatrix} x_{11} & 1 & 0 & 0 \\ x_{12} & 1 & 0 & 0 \\ x_{21} & 0 & 1 & 0 \\ x_{22} & 0 & 1 & 0 \\ x_{31} & 0 & 0 & 1 \\ x_{32} & 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} a \\ b_1 \\ b_2 \\ b_3 \end{pmatrix} + \begin{pmatrix} u_{11} \\ u_{12} \\ u_{21} \\ u_{22} \\ u_{31} \\ u_{32} \end{pmatrix}$$

More generally

$$\underline{y} = X\underline{a} + Z\underline{b} + \underline{u}$$

X variables of interest and Z control variables

Fixed effects regression and Frisch-Waugh theorem

$$\underline{y} = X\underline{\hat{a}} + Z\underline{\hat{b}} + \underline{\hat{u}} \quad \text{(a)}$$

$$p_Z(\underline{y}) = p_Z(X\underline{\hat{a}} + Z\underline{\hat{b}} + \underline{\hat{u}}) = p_Z(X)\underline{\hat{a}} + p_Z(Z)\underline{\hat{b}} + p_Z(\underline{\hat{u}}) \quad \text{(b)}$$

But $p_Z(Z) = Z$ and $p_Z(\underline{\hat{u}}) = \underline{0}$

$$\left[\underline{y} - p_Z(\underline{y})\right] = [X - p_Z(X)]\,\underline{\hat{a}} + \underline{\hat{u}} \quad \text{(a) - (b)}$$

The matrix of $P_Z(\cdot)$ is the B matrix

$$\left[y_{it} - \overline{y}_i\right] = [x_{it} - \overline{x}_i]\,\hat{a} + \hat{u}_{it} \quad \text{(a) - (b)}$$

A menagerie of estimators - I

1. OLS on the raw data

$$y_{it} = a\,x_{it} + b + u_{it}$$

2. RE hypothesis and FGLS estimator

$$y_{it} = a\,x_{it} + b + u_{it} \quad \text{and} \quad u_{it} = \mu_i + v_{it}$$

3. FE hypothesis and LSDV estimator

$$y_{it} = a\,x_{it} + b_i + u_{it}$$

   thanks to Frisch-Waugh theorem

$$\left[y_{it} - \overline{y}_i\right] = a\left[x_{it} - \overline{x}_i\right] + u_{it}$$

   aka within estimator

4. OLS on individual averages

$$\overline{y}_i = a\,\overline{x}_i + b + u_i$$

   aka between estimator

A menagerie of estimators - II

1. OLS on the raw data

```
MyPool.LS lq? lk? ll? t
```

2. RE hypothesis and FGLS estimator

```
MyPool.LS(R) lq? lk? ll? t
```

3. FE hypothesis and LSDV estimator aka within estimator

```
MyPool.LS(F) lq? lk? ll? t
```

   drop time-constant variables (e.g. sex)

4. OLS on individual averages aka between estimator

```
MyPool.LS @mean(lq?) @mean(lk?) @mean(ll?)
```

   wrong degrees of freedom, drop individual-constant variables (e.g. macro event)

A menagerie of results

Now only ***industry***: the first 19 sectors
$N = 19$, $T = 30$, and $NT = 570$

$$\ln(q_{it}) = \alpha \ln(k_{it}) + \beta \ln(\ell_{it}) + g\, t + a + u_{it}$$

Estimator	α	β	$\alpha+\beta$	g %
OLS on the raw data	0.29	0.34	0.63	1.7
RE hypothesis	0.26	0.76	1.02	2.5
Within estimator	0.31	0.85	1.16	2.6
Between estimator	0.29	0.31	0.60	–

Random versus fixed effect - II

$$y_{it} = a\, x_{it} + b + u_{it}$$

	RE	FE
Specification	$u_{it} = \mu_i + v_{it}$	b_i
Main hypothesis	$\mu \perp x$	irrelevance of between variability (time-constant variables)
Parsimonious?	yes: σ^2_μ	no: N dummy variables
Fragility	as usual	*i)* omitted variables with high within variability *ii)* measurement errors
Capture of heterogeneity	not enough	too much

Mundlack 1978 model

Production function in agricultural
As usual

$$y_{it} = a\,x_{it} + b + u_{it}$$

As econometricians do, RE hypothesis

$$u_{it} = \mu_i + v_{it}$$

But correlation between the random effect and the explanatory variable

$$\mu_i = \pi\,\overline{x}_i + \varepsilon_i$$

Mundlack "theorem"
FIML estimator of a is the within estimator of a

Measurement errors

Asymetry between the dependant variable and the explanatory variable

$$y_{it} = a x_{it} + b + u_{it}$$

u_i is the measurement error on y
Suppose that x is measured with error: $x^{\star}_{it} = x_{it} + v_{it}$

$$y_{it} = a x^{\star}_{it} + b + u_{it}$$

$\hat{a}$ is biased toward 0 but within estimator is much more biased than between estimator
Hint: the bias is linked to the "share of noise with respect to the signal"

Another way to rule out the RE

As usual (linear model plus RE hypothesis)

$$y_{it} = a\,x_{it} + b + \mu_i + v_{it}$$

Take the difference over time

$$[y_{it} - y_{i\,t-1}] = a\,[x_{it} - x_{i\,t-1}] + [v_{it} - v_{i\,t-1}]$$

Good news: b and μ_i are eliminated
Bad news: some time dependencies (moving average pattern) are introduced

Thanks for your attention

Feel free to send me e-mail: F.Legendre@u-pec.fr