Panel models

Hugo Harari-Kermadec

EPOG
Econometrics

Oct. 22, 2020

Panel data

indexed be both individual and time : $Y_{i,t}$.

```
panel<- read.csv("./panel_large.csv", header=TRUE,sep=",", dec="
panel_toy<-panel %>% select(Company_Name,year,marketcap)
head(panel_toy)
                 Company_Name year  marketcap IQ_TOTAL_ASSETS
1                 Pfizer Inc. 1994 24316.5977       11099.000
2 Verizon Communications Inc. 1994 21701.4941       24271.800
3     Exxon Mobil Corporation 1994 75418.4297       87862.000
4         Chevron Corporation 1994 29080.7754       34407.000
5            Time Warner Inc. 1994   659.8445         148.288
6  Motors Liquidation Company 1994 52889.3242      198598.700
```

Panel data

Many data on the same individual : $Y_{i,t},\ Y_{i,t+1},\ Y_{i,t+2}\ldots.$

```
panel_toy<-panel_toy %>% arrange(Company_Name)
head(panel_toy)
```

```
  Company_Name year marketcap IQ_TOTAL_ASSETS
1   AT&T Corp. 1994  78543.06           79262
2   AT&T Corp. 1995 103073.28           62395
3   AT&T Corp. 1996  70279.82           55382
4   AT&T Corp. 1997  99587.03           61095
5   AT&T Corp. 1998 132833.48           59550
6   AT&T Corp. 1999 162363.52          169406
```

Individuals are different $\Rightarrow$ $Y_{i,t}$ is likely to change with i.

Panel data

$Y_{i,t}$ is very dependent of $Y_{i,t+1}$. We should not use a standard linear model.

$$Y_{i,t} = a + bX_{i,t} + \varepsilon_{i,t} \;\leftrightarrow\; Y_{i,t} - a - bX_{i,t} = \varepsilon_{i,t}$$

The residuals are not likely to be iid.

Pooled model

It's the simple linear model : as if there was no panel. It assumes that each individual behave exactly as the others.

$$Y_{i,t} = a + bX_{i,t} + \varepsilon_{i,t}$$

Or more precisely, that every significative specificity is captured by the Xs. Not very likely.

Pooled model

It's the simple linear model : as if there was no panel. It assumes that each individual behave exactly as the others.

$$Y_{i,t} = a + bX_{i,t} + \varepsilon_{i,t}$$

Or more precisely, that every significative specificity is captured by the Xs. Not very likely.

Theorem : under the following hypothesis

H_1 : Columns of X are linearly independent,

H_2 : $\varepsilon_{i,t}$ residuals have 0 expectation and are uncorrelated with $X_{i,t}$.

H_3 : $\varepsilon_{i,t}$ residuals are uncorrelated with common variance σ^2.

The OLS estimator is unbiased and as minimum variance among linear estimators.

Breaking the panel ?

An alternative approach is to build a specific model for each individual

$$Y_{i,t} = a_i + b_i X_{i,t} + \varepsilon_{i,t}$$

or for each year (but not both !)

$$Y_{i,t} = a_t + b_t X_{i,t} + \varepsilon_{i,t}$$

Taking into account the panel

We can controlled for unobserved constant characteristics of individuals (and/or years : crisis, boom) with dummies.

$$Y_{i,t} = a_i + bX_{i,t} + \varepsilon_{i,t}$$

Fixed effects vs. random effects

- If a_i and $X_{i,t}$ are correlated, it's a fixed effect model
- If a_i and $X_{i,t}$ are not correlated, it's a random effect model

Panel package

```
install.packages("plm")
library("plm")
```

Fixed effects $Y_{i,t} = a_i + bX_{i,t} + \varepsilon_{i,t}$

Intercepts

We introduce an intercept for each individual.
This forbids to have an global intercept :
$Y_{i,t} = a_0 + a_i + bX_{i,t} + \varepsilon_{i,t}$
$\Leftrightarrow \quad Y_{i,t} = a_0 \times 1 + \sum_j a_j \times \mathbb{1}_{i=j} + bX_{i,t} + \varepsilon_{i,t}$
and $1 = \sum_j \mathbb{1}_{j=i}$ because every data refers to some individual.

Fixed effects $Y_{i,t} = a_i + bX_{i,t} + \varepsilon_{i,t}$

Intercepts

We introduce an intercept for each individual.
This forbids to have an global intercept :
$Y_{i,t} = a_0 + a_i + bX_{i,t} + \varepsilon_{i,t}$
$\Leftrightarrow \quad Y_{i,t} = a_0 \times 1 + \sum_j a_j \times \mathbb{1}_{i=j} + bX_{i,t} + \varepsilon_{i,t}$
and $1 = \sum_j \mathbb{1}_{j=i}$ because every data refers to some individual.

The fixed effect model is equivalent to a Least Squares Dummy Variables (LSDV) model :

$$Y_{i,t} = \sum_j a_j \times \mathbb{1}_{i=j} + bX_{i,t} + \varepsilon_{i,t}$$

Fixed effects $Y_{i,t} = a_i + bX_{i,t} + \varepsilon_{i,t}$

Multicolinearity between a_i and $X_{i,t}$

- a_i and $X_{i,t}$ can be correlated, but not multicolinear
- therefore $X_{i,t}$ can't be constant with time.
- For example, you can't include the sector in the Xs.

Fixed effects $Y_{i,t} = a_i + bX_{i,t} + \varepsilon_{i,t}$

Multicolinearity between a_i and $X_{i,t}$

- a_i and $X_{i,t}$ can be correlated, but not multicolinear
- therefore $X_{i,t}$ can't be constant with time.
- For example, you can't include the sector in the Xs.

The fixed effect a_i captures all of the time invariant individual specificity. That's why you can't include a time invariant $X_{i,t}$ (actually a X_i). And then if $Y_{i,t}$ changes over time, it can only be because of the variation of $X_{i,t}$.
The model $Y_{i,t} = a_i + bX_{i,t} + \varepsilon_{i,t}$ estimates the effect of X on Y within each country (assuming it's the same for every country).
To work well, we need Xs that vary strongly with times.

Fixed effect panel model estimation

$$Y_{i,t} = a_i + bX_{i,t} + \varepsilon_{i,t}$$

```
model.fe<-plm(marketcap~IQ_TOTAL_ASSETS,data=panel_toy,
index=c("Company_Name","year"),model="within")
summary(model.fe)
fixef(model.fe)
```

Random effects $Y_{i,t} = a + bX_{i,t} + A_i + \varepsilon_{i,t}$

For the random effect model, we suppose that A_i is random, therefore uncorrelated with everything (but constant across time).

Random effect panel model estimation

$$Y_{i,t} = A_i + bX_{i,t} + \varepsilon_{i,t}$$

```
model.re<-plm(marketcap~IQ_TOTAL_ASSETS,data=panel_toy,
index=c("Company_Name","year"),model="random")
summary(model.re)
```

Hausman test : RE vs. FE

Random effect are the null assumption (0 correlation between a_i and $X_{i,t}$) If the p-value is small, we reject the null and choose the Fixed effect model. But then we can't use any time invariant X.

```
phtest(model.fe,model.re)
```

Fisher test

Suppose I have a decent model :

```
model.fe<-plm(marketcap~IQ_TOTAL_ASSETS)
```

Should I add variables ? For example

```
model.fe.time<-plm(marketcap~IQ_TOTAL_ASSETS+factor(year))
```

A Fisher test compares two models (not only panel models)

```
pFtest(model.fe.time,model.fe)
```

If p-value <0.05, we reject the null explanatory power of the new variables $\Rightarrow$ keep them.
If p-value is large, come back to the smaller model.

Fisher test

```
F test for individual effects

data:  marketcap ~ IQ_TOTAL_ASSETS + factor(year)
F = 2.51, df1 = 17, df2 = 158, p-value = 0.001516
alternative hypothesis: significant effects
```

p-value <0.05 so we keep the year dummies.

Breusch-Godfrey/Wooldridge test for autocorrelation in residuals

The model is valid if the residuals are independent and identically distributed
$\Rightarrow$ no correlation among them (and homoskedasticity)

$$Y_{i,t} = a_i + bX_{i,t} + \varepsilon_{i,t}$$

For panel data, there is a strong risk of dependence between $\varepsilon_{i,t}$ and $\varepsilon_{i,t+1}$

```
pbgtest(model.fe)
```

If p-value <0.05, we reject the null correlation

Breusch-Godfrey/Wooldridge test for autocorrelation in residuals

```
data:  marketcap ~ IQ_TOTAL_ASSETS
chisq = 90.503, df = 11, p-value = 1.329e-14
alternative hypothesis: serial correlation in idiosyncratic
```

p-value $<<<0.05 \Rightarrow$ the model is not valid, there is more to explain.
Standard-errors are underestimated $\Rightarrow$ significativity of explanatory variables is overestimated

Intermediate report

For November 12th, prepare a 5 to 8 pages report

- a short explanation of the paper results
- a short presentation of the databases (Z1, compustat and WIOD)
- choose a variable of one of these databases and study it from a quantification perspective
- replicate figures 1, 2 and 5