PAINEL DATA MODELS

PROF. DR. VASCONCELOS REIS WAKIM

Tipos de Dados

- Concepts rescuing

There are 3 kind of data: Cross-Section, Panel Data and Time Series;

In the specific case of the Panel data→ we have the same unit of analysis or different unit of analysis in the long time

The panel data model can be called: *pooled data*

Why we should use panel data?

- What the advantage of the used of panel data?

1) Control of the effects do not unobserved;
2) Control the heterogeneity problem;
3) As we have cross-section + time series → more informative data; more variability, less collinearity between variables, more liberty degree, and more efficiency (Gujarati; Porter, 2011);
4) Panel data model are more adequate to examine the change dynamic;
5) The panel data model can be reflet better the facts than a cross-section model

Characteristic of panel data model

- We have: balanced panel; unbalanced panel; short panel or long panel
- **Balanced Panel:** the de panel is balanced if exist in the data all observation (time and units);
- **Unbalanced Panel:** the de panel is unbalanced if not exist in the data some observation (time and units)
- **Short Panel:** the number of units of analysis (N) is higher than the number of periods (T) → $N < T$
- **Long Panel:** the number of units of analysis (N) is minor than the number of periods (T) → $N > T$

Options to estimate the panel data model

- Estimating by Pooled Data;
- OLS with dummy variable;
- Fixed Effect Model; and
- Randon Effect Model.

Estimating by Pooled Data

- In this case, we just group the data as was a big database, and estimating it as an OLS - we ignore if exist a cross-section and/or times series;

$$Y_{it} = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \beta_3 X_{3t} + e_{it}$$

Estimating by Pooled Data

- In this case, we just group the data as was a big database, and estimating it as an OLS - we ignore if exist a cross-section and/or times series;

$$Y_{it} = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \beta_3 X_{3t} + e_{it}$$

- i = it's the unit of analysis that we are analyzed
- t = it's the period of analysis that we are analyzed

Estimating by Pooled Data

$$Y_{it} = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \beta_3 X_{3t} + e_{it}$$

- We presuppose that each coefficient are the same all sample;
- All units of analysis are equal to other units - there aren't differences between they;
- It’s hard to keep this presuppose;
- Assume that the explanatory variable is totally exogenous → current value does not depend on the past value

Estimating by Pooled Data

$$Y_{it} = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \beta_3 X_{3t} + e_{it}$$

- The results of the pooled model are totally significant, and the R2 statistic is very high, but the Durbin-Watson (DW) statistic é very low → DW test it's for autocorrelation of the residuals

- This way, suggest no existing spatial correlation between the resids. Or can be error of the model specification

Estimating by Pooled Data

$$Y_{it} = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \beta_3 X_{3t} + e_{it}$$

- In this kind of model (Pooled) we are camouflaging heterogeneity that existing between the units of analysis.

- If X_{3t} it's invariant in the time, we cannot observe directly your effect over the Y_{it}, but we can obtain your effect if rewrite the equation as:

$$Y_{it} = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \alpha_i + u_{it}$$

WAKIM
KNOWLEDGE TO LIBERTY

Estimating by Pooled Data

$$Y_{it} = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \alpha_i + u_{it}$$

- Where α_i it's the unobserved effect or heterogeneity;

- If estimating the Pooled model, we are ignoring this effect and all the units of analysis will be considered equal → what it's not true

- How to say Gujarati and Porter (2011) the heterogeneity is a nuisance parameter and must be treated

OLS WITH DUMMY VARIABLE

OLS WITH DUMMY VARIABLE

- This kind of model considers the heterogeneity that exists between the variables. Look at the equation:

$$Y_{it} = \beta_{0i} + \beta_1 X_{1t} + \beta_2 X_{2t} + \beta_3 X_{3t} + e_{it}$$

- β_{0i} show that each intercept of each unit of analysis can be different
- This difference can reflect the aspects do not unobserved of each unit

OLS WITH DUMMY VARIABLE

$$Y_{it} = \beta_{0i} + \beta_1 X_{1t} + \beta_2 X_{2t} + \beta_3 X_{3t} + e_{it}$$

- We can call this kind of model a Fixed Effect, because exist one or more variable which is invariable in the time
- Each intercept of each unit of analysis do not variant in the time
- But if we write β_{0i} as β_{0it} this suggests that constant is variant in the time
- But how to capture the heterogeneity among the units of analysis? We can use dummy variables

OLS WITH DUMMY VARIABLE

$$Y_{it} = \beta_{0i} + \beta_1 X_{1t} + \beta_2 X_{2t} + \beta_3 X_{3t} + e_{it}$$

- We rewrite equation above as:

$$Y_{it} = \alpha_0 + \alpha_1 D_{1t} + \alpha_2 D_{2t} + \alpha_3 D_{3t} + \beta_1 X_{1t} + \beta_2 X_{2t} + \beta_3 X_{3t} + e_{it}$$

- We must pay attention --> Only with the rule of the dummy variables → we have that exclude 1 dummy to do not to fall in the dummies variable trap → **perfect collinearity**

OLS WITH DUMMY VARIABLE

$$Y_{it} = \alpha_0 + \alpha_1 D_{1t} + \alpha_2 D_{2t} + \alpha_3 D_{3t} + \beta_1 X_{1t} + \beta_2 X_{2t} + \beta_3 X_{3t} + e_{it}$$

- The effect of unit 1, for example, it's the sum between $(\alpha_0 + \alpha_1)$, so α_2 and α_3 show the differences that exist among the unit of analysis

- $\alpha_0 + \alpha_1$ → gives the intercept of unit 1
- $\alpha_0 + \alpha_2$ → gives the intercept of unit 2
- $\alpha_0 + \alpha_3$ → gives the intercept of unit 3

REMEMBER, if you put all dummies variables, you must exclude the intercept of the equation!
Not to fall in the dummies variable trap

PERFECT COLLINEARITY!

OLS WITH DUMMY VARIABLE

$$Y_{it} = \alpha_0 + \alpha_1 D_{1t} + \alpha_2 D_{2t} + \alpha_3 D_{3t} + \beta_1 X_{1t} + \beta_2 X_{2t} + \beta_3 X_{3t} + e_{it}$$

- **ATTENTION:** If we estimating the model above, we loss much liberty degrees
- This is do not recommended
- We can estimate the model using the **fixed-effect** model
- The Stata command is:
- **Xtreg Y_{it} X_{1t} X_{2t} X_{3t}, fe (Where fe = fixed-effect)**

RANDON EFFECT MODEL

RANDON EFFECT MODEL

$$Y_{it} = \beta_{0i} + \beta_1 X_{it} + \beta_2 X_{it} + \beta_3 X_{it} + e_{it} \quad (1)$$

- In the equation above, β_{0i} it's not fixed, we should, now, presuppose that he is aleatory → β_0 without the superscript I
- $\beta_{0i} = \beta_0 + \varepsilon_i$ (2)
- With this, we say, that our sample was collected of the one universe bigger and that they have a mean common value of the intercept (β_0), and each difference between units to be in the error term.

RANDON EFFECT MODEL

$Y_{it} = \beta_{0i} + \beta_1 X_{it} + \beta_2 X_{it} + \beta_3 X_{it} + e_{it}$ (1) $\beta_{0i} = \beta_0 + \varepsilon_i$ (2)

- If we replace (2) in (1), equation 1 can be rewritten as:

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 X_{it} + \beta_3 X_{it} + e_{it} + \varepsilon_i$$

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 X_{it} + \beta_3 X_{it} + w_{it}$$

- Where:

$$w_{it} = e_{it} + \varepsilon_i$$

RANDON EFFECT MODEL

$$w_{it} = e_{it} + \varepsilon_i$$

- e_{it} have the combination of the time series with cross-section; and ε_i its the error term of each individual

- The difference between fixed-effect and Random effect model is: the fixed-effect model has an intercept common to all variables, in your turn, the random effect, the intercept represents the value means of all intercepts

- We'll see that exist so much resemblance between the results of the fixed-effect and random effect models. However, what model to estimates?

HAUSMAN TEST

HAUSMAN TEST

- If not exist differences significant between fixed-effect and random-effects models, what model we should estimates?

- To answer this question, we should use the Hausman Test (1974)

- The Null Hypothesis (H_0) of the Hausman Test is that the estimators of the fixed-effect and random effect model do not differ much

HAUSMAN TEST

- The Null Hypothesis (H_0) of the Hausman Test is that the estimators of the fixed-effect and random effect model do not differ much

- If the p-value of the Hausman test is significant, we are accepting the Null Hypothesis (H_0) what signified which the random effect is better than fixed-effect;

- However, if p-value of the Hausman test do not significant, we are rejecting the Null Hypothesis (H_0) what signified which the fixed-effect is better than the random effect;

HAUSMAN TEST

- **ATTENTION:** If the p-value of the Hausman test is significant, we are accepting the Null Hypothesis (H_0) what signified which the random effect is better than fixed-effect;
- So we must compare the random effect with the pooled data model, for this, we should use again the Hausman test. In this case:
- If the p-value of the Hausman test is significant, we are accepting the Null Hypothesis (H_0) what signified which the random effect is better than pooled model

HAUSMAN TEST

- However, if the p-value of the Hausman test do not significant, we are rejecting the Null Hypothesis (H_0) what signified which the pooled model is better than random effect

- In this case, when pooled model is better than the random effect, we must estimate the traditional OLS

HAUSMAN TEST

- The commands in the Stata is:
- xtreg Y X1 X2 X3, fe
- estimates store fe (stored in the memory of the program)
- xtreg Y X1 X2 X3, re
- estimates store re (stored in the memory of the program)
- hausman fe re, sgimamore

H_0 The random effect is better

H_1 The fixed-effect is better

P-value significative → accepts H_0

P-value not significative → rejects H_0