Quantitative Data Analysis

Master in Management - Academic Year 2023/2024

Ricardo Moura

ISEG

March 19, 2024

Intro to panel models

Now we have n units, $i = 1, \ldots, n$ but also more than one observation for every unit, T observations $t = 1, \ldots, T$.

Advantages:
Effect over time can be analysed;
Gained efficiency, as the sample size increases.

Cross-sectional versus panel data
Cross-sectional: independent units implies independent observations
Panel: same units observed through time implies that units are still independent, but for each unit observations are time dependent

Intro to panel models

Short panel:
- Very large sample n but with short time horizon (small T)
- time dependence for the observation of each unit is allowed and individuals are independent

Balanced panel
All units have observations for all t ($\forall i = 1, ..., n : T_i = T$)

Unbalanced panel
- There is missing information at some moments for some units ($T_i \neq T$), maybe because after a period there are no more observations (or decided not to provide it)
- Most of the estimators can be used

Decomposition of the variation

The variability of y_{it} is decomposed into:

$$\sum_{i=1}^{n}\sum_{t=1}^{T}(y_{it}-\bar{y})^2 = \sum_{i=1}^{n}\sum_{t=1}^{T}(y_{it}-\bar{y}_i+\bar{y}_i-\bar{y})^2$$

$$= \sum_{i=1}^{n}\sum_{t=1}^{T}(y_{it}-\bar{y}_i)^2 + \sum_{i=1}^{n}(\bar{y}_i-\bar{y})^2$$

First parcel: "Within variation" – variability of unit i through time

Second parcel: "between variation" – variability across units

Models for panel data

$$y_{it} = \alpha_i + \mathbf{x}'_{it}\boldsymbol{\beta} + u_{it}, \qquad i = 1, \ldots, n; t = 1, \ldots, T$$

- α_i: individual effects, time invariant and not observed
- x_{it} – explanatory variables:
  - x_{it}: characteristics that are different across individuals and change trough time
  - x_i: characteristics that are different across individuals and **do not** change trough time
  - d_t: time *dummy* at t
  - $d_t x_{it}$: interaction variables
- u_{it}: idiosyncratic error - differs across i and t

Models for panel data - Time-dummies

The objective of these *dummies* is to analyse time effects

For T years, the first year is the reference and suppressed and $T-1$ dummies, one for each of the remaining year, are created.

Example: panel data for 2016, 2017 and 2018 will only need two dummies: D2017 and D2018.

- $\hat{\beta}_{D2017}$ estimates the variation on the mean of Y in 2017 relative to 2016, caused by external factors to the considered regressors.
- $\hat{\beta}_{D2018}$ estimates the variation on the mean of Y in 2018 relative to 2016, caused by external factors to the considered regressors.

Random vs. Fixed effects

The model may be written as

$$y_{it} = \alpha_i + \mathbf{x}'_{it}\boldsymbol{\beta} + u_{it} = \mathbf{x}'_{it}\boldsymbol{\beta} + (\alpha_i + u_{it})$$

where the error term has two components, with α_i correlated or not with the explanatory variables:

Random effects:
– Assumption: α_i and x_{it} are not correlated
– Estimators addressed: Pooled and Random effects.

Fixed effects:
– Assumption: α_i and x_{it} may be correlated
– Estimators addressed: Fixed effects or "Within" and First differences.

Decision: Hausman test.

Random effects - Pooled Estimator

The model may be written as

$$y_{it} = \alpha + \mathbf{x}'_{it}\boldsymbol{\beta} + v_{it}$$

for $v_{it} = \alpha_i - \alpha + u_{it}$, when we assume that the individual-specific effects are assumed constant (homogeneity across units, treating data as one large cross-sectional dataset). (Use Breusch-Pagan test for homoskedasticity. If variance equal zero is not rejected use pooled regression)

The estimation is made by OLS with cluster or similar option for the variance
Stata: `regress Y` $X_1 \ldots X_p$`, vce(cluster` *clustvar*`)`

Random effects - Random effects Estimator

The model may be written as

$$y_{it} = \alpha + \mathbf{x}_{it}'\boldsymbol{\beta} + v_{it}$$

for $v_{it} = \alpha_i - \alpha + u_{it}$, but we assume $Var(\alpha_i) = \sigma_\alpha^2$ and $Var(u_{it}) = \sigma_u^2$

The estimation is made by generalized LS with cluster or similar option for the variance (efficient estimator) by using model:

$$y_{it} - \hat{\theta}_i \bar{y}_i = (1 - \hat{\theta}_i)\alpha + (\mathbf{x}_{it} - \hat{\theta}_i \bar{\mathbf{x}}_i)'\boldsymbol{\beta} + v_{it}$$

where $\hat{\theta}_i = 1 - \sqrt{\hat{\sigma}_u^2 / (T_i \hat{\sigma}_\alpha^2 + \hat{\sigma}_\alpha^2)}$ and $v_{it} = (1 - \hat{\theta}_i)\alpha_i + (u_{it} - \hat{\theta}_i \bar{u}_i)$, exploiting the correlation between u_{it} and u_{is}. (the pooled doesn't exploit the panel nature apart from variance calculation in cluster robust form)

Stata: `xtreg Y` $X_1 \ldots X_p$`, vce(cluster` *clustvar*`)`

xt stands for "cross-sectional time-series"

Fixed effects - Fixed effects Estimator

The model may be written as

$$y_{it} - \bar{y}_i = (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)'\boldsymbol{\beta} + (u_{it} - \bar{u}_{it})$$

Estimation is made applying OLS to the transformed variables with cluster version for the variance

Stata: `xtreg Y` $X_1 \ldots X_p$`, fe vce(cluster` *clustvar*`)`

It is robust, but has the following disadvantages, given that random effects are not required:

- Eliminates all time invariant explanatory variables
- Eliminates all time variant explanatory variables that change in time by a constant (Example: age, experience...)

Fixed effects - First difference Estimator

The model may be written as

$$y_{it} - y_{i(t-1)} = (\mathbf{x}_{it} - \mathbf{x}_{i(t-1)})'\boldsymbol{\beta} + (u_{it} - u_{i(t-1)}) \Leftrightarrow \Delta y_{it} = \Delta \mathbf{x}'_{it}\boldsymbol{\beta} + \Delta u_{it}$$

Estimation is made applying OLS to the transformed variables with cluster version for the variance

Stata: `regress D.Y D.`X_1`...D.`X_p`, vce(cluster` *clustvar*`)`

Displays the same disadvantages of the FE estimator and in fact is numerically equal to the FE estimator for T=2.

Hausman Test

Test if effects are fixed or random

$$H_0 : E(\alpha_i|\mathbf{x}_{it}) = 0 \quad versus \quad E(\alpha_i|\mathbf{x}_{it}) \neq 0$$

If not rejected, RE and FE are consistent but only RE is efficient. If rejected, FE is consistent but RE is inconsistent.

$$H = (\hat{\beta}_{FE} - \hat{\beta}_{RA})' \left[Var(\hat{\beta}_{FE}) - Var(\hat{\beta}_{RE}) \right]^{-} 1(\hat{\beta}_{FE} - \hat{\beta}_{RA}) \sim \chi^2_p$$

Stata: Just applies to models estimated without robust or cluster options

`xtreg Y` $X_1 \ldots X_p$`, fe`

`estimates store ModelFE`

`xtreg Y` $X_1 \ldots X_p$

`estimates store ModelRE`

`hausman ModelFE ModelRE`

Policy analysis based on panel data, T=2

Consider a sample where individuals are observed twice (observed before and after the programme implementation) and we have individuals of two types: affected (cases / treated) and not affected (controls) Model:

$$y_{it} = \alpha + \delta d_2 + \beta prog_{it} + \alpha_i + u_{it}$$

where $prog = 1$ if treated/affected and $d_2 = 1$ after the programme implementation.
Model based on differences:

$$\Delta y_{it} = \delta + \beta prog_{it} + \Delta u_{it}$$

Effect of the programme: β

Policy analysis based on panel data, T=2

Example: Wooldridge

Aim: investigate whether the scrap rate (% products that are not in conditions to be sold), scrap, changes as a consequence of the participation in a training programme, (Grant=1 if training was received), in 1988. Panel data for 1987 and 1988 are available and include sampling units with Grant=1 and Grant=0.

Estimated model (standard deviations above coefficients)

$$\Delta \widehat{ln(scrap)} = -\underset{}{\overset{(0.097)}{0.057}} - \overset{(0.164)}{0.317}\, grant, n = 57, R^2 = 0.067$$

Policy analysis based on panel data, T=2

Example: Wooldridge

Aim: investigate whether the scrap rate (% products that are not in conditions to be sold), scrap, changes as a consequence of the participation in a training programme, (Grant=1 if training was received), in 1988. Panel data for 1987 and 1988 are available and include sampling units with Grant=1 and Grant=0.

Estimated model (standard deviations above coefficients)

$$\Delta \widehat{ln(scrap)} = -\underset{}{\overset{(0.097)}{0.057}} - \overset{(0.164)}{0.317}\, grant, n = 57, R^2 = 0.067$$

- Training reduce the scrap rate in $(e^{0.317} - 1)100\% = 27.2\%$ - observe the std

Policy analysis based on panel data, T=2

Example: Wooldridge

Aim: investigate whether the scrap rate (% products that are not in conditions to be sold), scrap, changes as a consequence of the participation in a training programme, (Grant=1 if training was received), in 1988. Panel data for 1987 and 1988 are available and include sampling units with Grant=1 and Grant=0.

Estimated model (standard deviations above coefficients)

$$\Delta \widehat{ln(scrap)} = -\overset{(0.097)}{0.057} - \overset{(0.164)}{0.317}\, grant, n = 57, R^2 = 0.067$$

- Training reduce the scrap rate in $(e^{0.317} - 1)100\% = 27.2\%$ - observe the std
- The scrap rate reduced in $(e^{0.057} - 1)100\% = 5.9\%$ due to factors which are not the training programme participation - observe the std

Policy analysis based on panel data, T=2

Example: Wooldridge

Aim: investigate whether the scrap rate (% products that are not in conditions to be sold), scrap, changes as a consequence of the participation in a training programme, (Grant=1 if training was received), in 1988. Panel data for 1987 and 1988 are available and include sampling units with Grant=1 and Grant=0.

Estimated model (standard deviations above coefficients)

$$\Delta \widehat{ln(scrap)} = -\underset{}{\overset{(0.097)}{0.057}} - \overset{(0.164)}{0.317}\, grant, n = 57, R^2 = 0.067$$

- Training reduce the scrap rate in $(e^{0.317} - 1)100\% = 27.2\%$ - observe the std
- The scrap rate reduced in $(e^{0.057} - 1)100\% = 5.9\%$ due to factors which are not the training programme participation - observe the std
- The R-squared value of 0.067 indicates that around 6.7% of the variation in the change in log scrap rates can be explained by the model. (even if it is low in policy analysis can make an impact)

Policy analysis based on panel data, T=2

Example: Wooldridge

Aim: investigate whether the scrap rate (% products that are not in conditions to be sold), scrap, changes as a consequence of the participation in a training programme, (Grant=1 if training was received), in 1988. Panel data for 1987 and 1988 are available and include sampling units with Grant=1 and Grant=0.

Estimated model (standard deviations above coefficients)

$$\Delta \widehat{ln(scrap)} = -\underset{}{\overset{(0.097)}{0.057}} - \overset{(0.164)}{0.317}\, grant, n = 57, R^2 = 0.067$$

- $-0.057 \pm 1.96 \times 0.097 = (-0.247, 0.133)$
- $-0.317 \pm 1.96 \times 0.164 = (-0.638, 0.004)$