Panel Data Models

(GB: Chapter 6.5, 7.8, 10)

Isabel Casas

Summary: Censored Regression Models

- Two types of censored models
  - Data censoring
  - Corner solution outcomes
- Tobit model
- Expected values and marginal effects
- Estimation (MLE)
- Reporting results
- Specification issues
  - Heteroskedasticity
  - Endogeneity

Summary today's class

1. Motivation
   - What is panel data?
   - Data sets and notation
   - Review of the linear model
2. The linear panel data model, and four estimation methods:
   - Pooled OLS
   - Fixed effects
   - First differences
   - Random effects
   - Models comparison
3. Next week: Probit, Logit and Tobit with panel data

Motivation

- The regression model is an essential statistical model in econometrics
- However, regression lines from economic data often cannot give a causal interpretation
- Although explanatory variables might be correlated with unobservables (endogenous)
- The regression model assumes that such correlation does not exist

Motivation

Examples of expected endogeneity:

1. supply-and-demand simultaneous problem

$$\text{Demand: } q_t = \alpha_1 + \alpha_2\ p_t + \epsilon_{1t}$$
$$\text{Supply: } q_t = \beta_1 + \beta_2\ p_t + \epsilon_{2t}$$

2. Measurement error. The explanatory variable is wrongly recorded and the error is correlated to the original variable.
3. Unobserved heterogeneity. If variables that have an effect on Y and X are omitted then the explanatory variables are correlated with the errors

Solutions:

- Instrumental variables
- Multiple regressions (2SLS)
- However we often lack the data or instruments needed

Motivation

> *'The major motivation for using panel data has been the ability to control for possibly correlated, time-invariant heterogeneity without observing it'.*

– Manuel Arellano (2003). "Panel Data Econometrics"

Therefore we are going to study the use of the extra information in a panel to estimate models with time-invariant omitted variables.

Motivation

Suppose we have a cross-sectional regression of the form:

$$y_{j1} = \beta_0 + \mathbf{x}_{j1}\beta + c_j + \epsilon_{j1} \quad \text{with} \quad E(\epsilon_{j1}|\mathbf{x}_{j1}, c_j) = 0$$

- If c_j is observed $\Rightarrow$ OLS
- If c_j is omitted and $Cov(\mathbf{x}_{j1}, c_j) = 0$

$$\beta = \frac{Cov(\mathbf{x}_{j1}, y_{j1})}{Var(\mathbf{x}_{j1})}$$

- If c_j is omitted and $Cov(\mathbf{x}_{j1}, c_j) \neq 0$ then we need an instrument z_j such that $Cov(z_j, c_j) = 0$

$$\beta = \frac{Cov(z_j, y_{j1})}{Cov(z_j, \mathbf{x}_{j1})}$$

Suppose that either of these cases is available. However, we have more data: $(\mathbf{x}_{j2}, y_{j2})$.

Motivation

$$
\begin{aligned}
y_{j1} &= \beta_0 + \mathbf{x}_{j1}\beta + c_j + \epsilon_{j1} \\
y_{j2} &= \beta_0 + \mathbf{x}_{j2}\beta + c_j + \epsilon_{j2}
\end{aligned}
$$

where $E(\epsilon_{jt}|\mathbf{x}_{j1}, \mathbf{x}_{j2}, c_j) = 0$.

Q: What do you get if you do $y_{j2} - y_{j1}$?

Motivation

$$
\begin{aligned}
y_{j1} &= \beta_0 + \mathbf{x}_{j1}\beta + c_j + \epsilon_{j1} \\
y_{j2} &= \beta_0 + \mathbf{x}_{j2}\beta + c_j + \epsilon_{j2}
\end{aligned}
$$

where $E(\epsilon_{jt}|\mathbf{x}_{j1}, \mathbf{x}_{j2}, c_j) = 0$.

Q: What do you get if you do $y_{j2} - y_{j1}$?

$$
\begin{aligned}
y_{j2} - y_{j1} &= \beta(\mathbf{x}_{j2} - \mathbf{x}_{j1}) + \epsilon_{j2} - \epsilon_{j1} \\
\Delta_t y_j &= \beta \Delta_t \mathbf{x}_j + \Delta\epsilon_j
\end{aligned}
$$

Then β is identified in the regression of the first difference.

Panel data set

j	t	y_{jt}	x_{jt}^1	x_{jt}^2	$\dots$	x_{jt}^k
1	1	y_{11}	x_{11}^1	x_{11}^2	$\dots$	x_{11}^k
1	2	y_{12}	x_{12}^1	x_{12}^2	$\dots$	x_{12}^k
1	3	y_{13}	x_{13}^1	x_{13}^2	$\dots$	x_{13}^k
2	1	y_{21}	x_{21}^1	x_{21}^2	$\dots$	x_{21}^k
2	2	y_{22}	x_{22}^1	x_{22}^2	$\dots$	x_{22}^k
2	3	y_{23}	x_{23}^1	x_{23}^2	$\dots$	x_{23}^k
$\vdots$					$\vdots$	
n	1	y_{n1}	x_{n1}^1	x_{n1}^2	$\dots$	x_{31}^k
n	2	y_{n2}	x_{n2}^1	x_{n2}^2	$\dots$	x_{32}^k
n	3	y_{n3}	x_{n3}^1	x_{n3}^2	$\dots$	x_{33}^k

Motivation

Panel data:

- Each individual is observed several times
- I.e. time series data for each individual
- $j = 1, \ldots, n$ individuals
- $t = 1, \ldots, T$ periods

Note:

- A balanced panel has $n \times T$ observations.
- All individuals need not be observed in all periods: an unbalanced panel has less than $n \times T$.
- We assume throughout that n is large and T is fixed
- The asymptotic properties are for $n \to \infty$ (not for $T \to \infty$).

Motivation

Advantages of panel data:

- More observations compared to cross-sectional data (where $T = 1$).
- Additional possibilities for dealing with endogenous variables on the RHS when endogeneity is due to an unobserved individual effect (e.g. ability):

$$y_{jt} = \beta_0 + \beta_1 \mathbf{x}_{jt}^1 + \ldots + \beta_k \mathbf{x}_{jt}^k + c_j + \epsilon_{jt}$$

- c_j is unobserved and correlated with one or more of the $\mathbf{X}$'s

Review: Multivariate linear model

Remember the multivariate linear model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

with assumptions:

- OLS.1 $E(\boldsymbol{\epsilon}|\mathbf{X}) = 0$
- OLS.2 rank $E(\mathbf{X}'\mathbf{X}) = k + 1$ ($\mathbf{X}$ includes the intercept)
- OLS.3 $E(\boldsymbol{\epsilon}^2|\mathbf{X}) = \sigma^2$

and assume that we have a random sample.

Review: Multivariate linear model

- OLS.1 + OLS.2 $\Rightarrow$
  1. The OLS estimator $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ is consistent
  2. unbiased
- OLS.1 + OLS.2 +OLS.3 $\Rightarrow$
  1. The OLS estimator is asymptotically normal.

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \rightarrow^{d} N(\mathbf{0}, \sigma^2\mathbf{A}^{-1}) \quad \text{for } \mathbf{A} = E(\mathbf{X}'\mathbf{X})$$

  2. Therefore, the estimator is approx. normal in large samples with variance

$$\hat{\mathbf{V}} = AVar(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$$

Review: Multivariate linear model

If the model looks like:

$$y_j = \mathbf{x}_j \beta + c_j + \epsilon_j$$

where c_j is an unobserved individual effect that is correlated with $\mathbf{x}_j$.

$$\Downarrow$$

Then OLS.1 is violated and the OLS estimator becomes inconsistent

Solutions:

- OLS with proxies for c_j or
- Instrumental variables for $\mathbf{x}_j$

However panel data provides another solution to this...

Linear panel data model (LPDM)

- The basic unobserved effects model.
- For observation i (individual i), the model is written as:

$$y_{jt} = \beta_0 + \beta_1 \mathbf{x}_{jt}^1 + \ldots + \beta_k \mathbf{x}_{jt}^k + c_j + \epsilon_{jt}, \quad t = 1, 2, \ldots, T$$

- Where c_j is an unobserved **time-invariant** individual effect which may or may **not** be correlated with $\mathbf{x}_{jt}$
- $\mathbf{x}_{jt}$ may vary across i or t or both
- Depending on the assumptions we put on c_j, the model above can be consistently estimated with panel data using different techniques

Linear panel data model (LPDM)

The basic unobserved effects model:

$$y_{jt} = \mathbf{x}_{jt}\beta + c_j + \epsilon_{jt}, \quad t = 1, 2, \ldots, T$$

Definition (Strict exogeneity of X)

$$E(y_{jt}|\mathbf{x}_{jt}) = \mathbf{x}_{jt}\beta \quad \Rightarrow \quad E(\mathbf{x}'_{jt}\epsilon_{jt}) = 0 \text{ and } E(\mathbf{x}'_{jt}c_j) = 0$$

$$\Downarrow$$

$$E(c_j|\mathbf{x}_{jt}) = E(c_j) \text{ (independence of } \mathbf{x}_j \text{ and } c_j)$$

Linear panel data model (LPDM)

The basic unobserved effects model:

$$y_{jt} = \mathbf{x}_{jt}\beta + c_j + \epsilon_{jt}, \quad t = 1, 2, \ldots, T$$

Definition (Strict exogeneity of X conditional on the unobserved effect)

$$E(y_{jt}|\mathbf{x}_{jt}, c_j) = \mathbf{x}_{jt}\beta + c_j \quad \Leftrightarrow \quad E(\epsilon_{jt}|\mathbf{x}_{jt}, c_j) = 0$$

$$\Downarrow$$

$E(\mathbf{x}'_{js}\epsilon_{jt}) = 0$ (it is more than contemporaneous uncorrelation)

Linear panel data model (LPDM)

- The strict exogeneity assumption is more restrictive than the conditional one.
- We always used the strict exogeneity for the previous models with cross-sectional data
- However, when using panel data, this assumption can be relaxed for the pooled OLS, RE, FE and FD models
- Two questions to be answered:
  1. Is the unobserved effect c_j uncorrelated with $\mathbf{x}_{jt}$ for all t?
  2. Is the conditional strict exogeneity of X conditioned on c_j reasonable?

Example: Agricultural Cobb–Douglas

$$y_{jt} = \mathbf{x}_{jt}\beta + c_j + \epsilon_{jt} \quad \text{Production function}$$

- Agricultural Cobb–Douglas Production Function
- i: farm
- t: time period
- y_{jt}: log production
- $\mathbf{x}_{jt}$: log of a variable input (labour)
- c_j: soil quality (constant over time)
- ϵ_{jt}: rainfall (outside farmer's control)

Example: Agricultural Cobb–Douglas

- Farmer knows how to work different soils but the econometrician doesn't know variable c_j
- Therefore the econometrician will suggest wrong ways of working the soil to the farmer if we only use period $(\mathbf{x}_{j1}, y_{j1})$
- However, if we know variables for period 2
- Moreover, the rainfall on the second period is unpredictable from the first period (uncorrelation in the errors)
- We can then estimate β using this information

Example: Returns to education

- Cross–sectional estimates of returns are not trusted because of the omitted $ability$
- i: individual and t: time period
- y_{jt}: log wage
- $\mathbf{x}_{jt}$: Years of full-time education
- c_j: ability
- β: returns to education

Taking the first difference does not work in this example, why?

Example: Returns to education

- $\mathbf{x}_{jt}$: Years of full-time education, is not time–variant
- If education is the same in two periods for all individuals then $Var(\Delta X) = 0$
- If it is not, the changes are too small so the new information is not enough to get a good estimate of β

Fill in this table

	Assumptions	Estimation procedure	Pros and cons
Pooled OLS			
FE			
FD			
RE			

Types of Models

$$y_{jt} = \beta_0 + \beta_1 \mathbf{x}_{jt}^1 + \ldots + \beta_k \mathbf{x}_{jt}^k + c_j + \epsilon_{jt}, \quad t = 1, 2, \ldots, T$$

If we assume that c_j is part of the error $\nu_{jt} = c_j + \epsilon_{jt}$ and uncorrelated to $\mathbf{x}_{jt}$

- Pooled OLS
- Random effects

If we assume that c_j may be correlated with $\mathbf{x}_{jt}$,

- Fixed effects,
- First difference difference

LPDM: Pooled OLS

A panel data model without unknown effects:

$$y_{jt} = \beta_0 + \beta_1 \mathbf{x}_{jt}^1 + \ldots + \beta_k \mathbf{x}_{jt}^k + \epsilon_{jt}, \quad t = 1, 2, \ldots, T$$

- $(y_j, \mathbf{x}_j)$ has T rows that should be ordered chronologically

LPDM: Pooled OLS

A panel data model without unknown effects:

$$y_{jt} = \beta_0 + \beta_1 \mathbf{x}_{jt}^1 + \ldots + \beta_k \mathbf{x}_{jt}^k + \epsilon_{jt}, \quad t = 1, 2, \ldots, T$$

- $(y_j, \mathbf{x}_j)$ has T rows that should be ordered chronologically

If

POLS.1 $E(\epsilon_{jt}\mathbf{x}_{jt}) = 0$
- ϵ_{jt} and $\mathbf{x}_{jt}$ are uncorrelated

POLS.2 rank $E(\sum_t \mathbf{x}'_{jt}\mathbf{x}_{jt}) = k + 1$

POLS.3 Strong homoskedasticity assumption
- $E(\epsilon_t^2 \mathbf{x}'_t X) = \sigma^2 E(\mathbf{x}'_t \mathbf{x}_t)$
- $E\epsilon_t\epsilon_s \mathbf{x}'_t \mathbf{x}_t) = 0, t \neq s,\ t, s = 1, \ldots T$

LPDM: Pooled OLS

A panel data model without unknown effects:

$$y_{jt} = \beta_0 + \beta_1 \mathbf{x}_{jt}^1 + \ldots + \beta_k \mathbf{x}_{jt}^k + \epsilon_{jt}, \quad t = 1, 2, \ldots, T$$

- $(y_j, \mathbf{x}_j)$ has T rows that should be ordered chronologically

If

POLS.1 $E(\epsilon_{jt}\mathbf{x}_{jt}) = 0$
- ϵ_{jt} and $\mathbf{x}_{jt}$ are uncorrelated

POLS.2 rank $E(\sum_t \mathbf{x}'_{jt}\mathbf{x}_{jt}) = k + 1$

POLS.3 Strong homoskedasticity assumption
- $E(\epsilon_t^2 \mathbf{x}'_t X) = \sigma^2 E(\mathbf{x}'_t \mathbf{x}_t)$
- $E\epsilon_t \epsilon_s \mathbf{x}'_t \mathbf{x}_t) = 0, t \neq s,\ t, s = 1, \ldots T$

Then (See Wooldridge, Chapter 7.8.1),

- The POLS estimator is consistent, asymptotically normal and we can find the estimator of the asymptotic variance
- The usual t-test and F-test are valid

LPDM: Pooled OLS

The estimator looks like:

$$\hat{\beta}^{POLS} = \left(\sum_{j=1}^{n} \sum_{t=1}^{T} \mathbf{x}_{jt}' \mathbf{x}_{jt} \right)^{-1} \left(\sum_{j=1}^{n} \sum_{t=1}^{T} \mathbf{x}_{jt}' y_{jt} \right)$$

LPDM: Pooled OLS

However, if we have individual unknown effects:

$$y_{jt} = \beta_0 + \beta_1 \mathbf{x}_{jt}^1 + \ldots + \beta_k \mathbf{x}_{jt}^k + \underbrace{c_j + \epsilon_{jt}}_{\nu_{jt}}, \quad t = 1, 2, \ldots, T$$

If:

POLS.1 $E(\nu_{jt}\mathbf{x}_{jt}) = 0$

- c_j and $\mathbf{x}_{jt}$ are uncorrelated, $E(\mathbf{x}'_{jt} c_j) = 0$ and
- $E(\mathbf{x}'_{jt}\epsilon_{jt}) = 0$

POLS.2 rank $E(\sum_t \mathbf{x}'_{jt}\mathbf{x}_{jt}) = k + 1$

POLS.3 Strong homoskedasticity assumption is not satisfied

- Problem: the error terms $\nu_{jt} = c_j + \epsilon_{jt}$, $\nu_{j,t-1} = c_j + \epsilon_{j,t-1}$: serially correlated $cov(\nu_{jt}, \nu_{j,t-1}) = Var(c_j) \Rightarrow$ we need robust variance matrix estimators and robust t-test (Wooldridge, Chapter 7.8.4) to do the inference

Pooled OLS in R

- Package plm
- Tell the program which indexes from your data set are for individuals and which for time

$$data.plm = pdata.frame(mydata, c(\text{"}id\text{"}, \text{"}time\text{"}))$$

- Estimate the model using the formula as usual

$$model.pols < -plm(formula, data = data.plm, \\ model = \text{"}pooling\text{"}, effect = \text{"}individual\text{"})$$

- It returns: $coefficients, residuals, fitted.values, vcov,$ $df.residual$ and $call$

Example: wagepan.dat

variable name	variable label
Individual definition	
nr	person identifier
year	1980 to 1987
educ	years of schooling
exper	labor mkt experience
expersq	exper^2
union	=1 if in union
married	=1 if married
black	=1 if black
hisp	=1 if Hispanic
poorhlth	=1 if in poor health
hours	annual hours worked
lwage	log(wage)
nrthcen	=1 if north central
nrtheast	=1 if north east
rur	=1 if live in rural area
south	=1 if south
Industry dummies	
agric	=1 if in agriculture
manuf	=1 if in manufacturing
min	=1 if mining
fin	=1 if finance
tra	=1 if transportation
trad	=1 if trade
per	=1 if personal service
pro	=1 if professional & related
pub	=1 if public administration
bus	=1 if business & repair serv.
construc	=1 if in construction
ent	=1 if entertainment

Example: wagepan.dat

```
variable name        variable label
-----------------------------------------
Ocuppational dummies
occ1                 =1 if professional, technical
occ2                 =1 if mgr, official, proprietor
occ3                 =1 if sales
occ4                 =1 if clerical
occ5                 =1 if craftsman, foreman
occ6                 =1 if operative
occ7                 =1 if laborer, farmer
occ8                 =1 if farm laborer, foreman
occ9                 =1 if service

Years dummies
d81                  =1 if year == 1981
d82                  =1 if year == 1982
d83                  =1 if year == 1983
d84                  =1 if year == 1984
d85                  =1 if year == 1985
d86                  =1 if year == 1986
d87                  =1 if year == 1987
```

Example: wagepan.dat

- Is there a causal relationship between years of education and log(wages)?
- We know $educ$ is endogenous but now we have data for several years of each individual
- Run a Pooled OLS of $lwage$ on
  $educ, exper, expersq, union, married, black, hisp, pub$

Example: wagepan.dat

```
> library(plm)
> library(lmtest)
> data<-read.table("./Exercises/wagepan.dat", h=T)
> #Describe the data by the individual and time indexes
> data2<-pdata.frame(data, index=c("nr", "year"))
> # Pooled OLS
> wage.pool<-plm(lwage ~ educ+ exper+ expersq+union+ marrie
+ black+ hisp+ pub, data = data2, model = "pooling")
```

Example: wagepan.dat

```
> summary(wage.pool)
Oneway (individual) effect Pooling Model
Call:
plm(formula = lwage ~ educ + exper + expersq + union + married +
    black + hisp + pub, data = data2, model = "pooling")

Balanced Panel: n=545, T=8, N=4360
Residuals :
   Min. 1st Qu.  Median 3rd Qu.    Max.
-5.2700 -0.2490  0.0332  0.2960  2.5600

Coefficients :
               Estimate  Std. Error t-value  Pr(>|t|)
(Intercept) -0.03437236  0.06467230 -0.5315    0.5951
educ         0.09936782  0.00468289 21.2194 < 2.2e-16 ***
exper        0.08913804  0.01012149  8.8068 < 2.2e-16 ***
expersq     -0.00284682  0.00070771 -4.0226 5.854e-05 ***
union        0.17990425  0.01721460 10.4507 < 2.2e-16 ***
married      0.10762117  0.01570528  6.8525 8.271e-12 ***
black       -0.14382268  0.02356305 -6.1037 1.126e-09 ***
hisp         0.01565030  0.02081966  0.7517    0.4523
pub          0.00354610  0.03747396  0.0946    0.9246
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    1236.5
Residual Sum of Squares: 1005.8
R-Squared      :  0.18659
      Adj. R-Squared :  0.1862
F-statistic: 124.759 on 8 and 4351 DF, p-value: < 2.22e-16
```

LPDM: Fixed effects (FE)

- If c_j is correlated with $\mathbf{x}_{jt}$, pooled OLS fails (inconsistent estimates)
- FE estimation allows c_j to be correlated with $\mathbf{x}_j$.

Basic idea:

- We treat c_j as a parameter to be estimated. This is possible with several observations for each individual

In practice:

1. We transform data by subtracting individual averages: *demeaned* variables
2. This removes c_j from the equation.
3. Run POLS on demeaned variables

LPDM: Fixed effects (FE)

Formally, averaging:

$$y_{jt} = \mathbf{1}\beta_0 + \mathbf{x}_{jt}\beta + c_j + \epsilon_{jt}$$

the average for individual i over time:

$$\bar{y}_j = 1 \cdot \beta_0 + \bar{\mathbf{x}}_j\beta + c_j + \bar{\epsilon}_j$$

where $\bar{y}_j = T^{-1}\sum_t y_{jt}$.

LPDM: Fixed effects (FE)

The demeaned equation:

$$y_{jt} - \bar{y}_j = (\mathbf{x}_{jt} - \bar{\mathbf{x}}_j)\beta + (\epsilon_{jt} - \bar{\epsilon}_j)$$

Which we just write as:

$$\ddot{y}_{jt} = \ddot{\mathbf{x}}_{jt}\beta + \ddot{\epsilon}_{jt}$$

- This "within" transformation has eliminated the time-invariant unobserved effect c_j.
- The $\hat{\beta}^{FE}$ is the pooled OLS estimator from the above regression
- Problem: we also eliminate other time-invariant variables among the X and the intercept (example: $educ$)

LPDM: Fixed effects (FE)

- Pooled OLS on the transformed equation: $\ddot{y}_{jt} = \ddot{\mathbf{x}}_{jt}\beta + \ddot{\epsilon}_{jt}$
- yields the consistent estimator

$$\hat{\beta}^{FE} = \left(\sum_{j=1}^{n} \sum_{t=1}^{T} \ddot{\mathbf{x}}'_{jt} \ddot{\mathbf{x}}_{jt} \right)^{-1} \left(\sum_{j=1}^{n} \sum_{t=1}^{T} \ddot{\mathbf{x}}'_{jt} \ddot{y}_{jt} \right)$$

- If

  FE.1 Strict exogeneity conditional on the unobserved effect
  - $E(\epsilon_{jt}|\mathbf{x}_{jt}, c_j) = 0$
  - $E(\mathbf{x}_{jt}, c)$ may be $\neq 0$

  FE.2 rank $\sum_{t=1}^{T} E(\ddot{\mathbf{x}}'_{jt}\ddot{\mathbf{x}}_{jt}) = k$
  - Therefore, the model does not include intercept or time-invariant explanatory variables

LPDM: Fixed effects (FE)

To ensure the efficiency of the estimator:

FE.3 $Var(\epsilon_{jt}|\mathbf{x}_j, c_j) = \sigma^2 I_T$

- Homokedaticity: $Var(\epsilon_{jt}|\mathbf{x}_j, c_j) = \sigma^2$
- Serially uncorrelated errors:
  $cov(\epsilon_{jt}\epsilon_{js}|\mathbf{x}_j, c_j) = 0, s \neq t$

LPDM: Fixed effects (FE)

Pros:

- We can consistently estimate partial effects in present of time-constant omitted variables

Cons:

- We cannot include time-invariant factors such as $gender$, $race$, $industry$ of a firm, etc. in $\mathbf{x}_{jt}$
- This can be a drawback in certain applications
- However, if our application uses only time-varying explanatory variables then this is the model to use
- Variables such as $educ$ can be constant for some periods but must vary in some part of the sample.

LPDM: Least squares dummy variables (LSDV)

A way around the constraints from FE.2:

$$y_{jt} = \theta_1 \; d1_{jt} + \ldots + \theta_N \; dN_{jt} + w_{jt}\delta + \epsilon_{jt}$$

- $d1_{jt}, \ldots, dN_{jt}$ are individual dummies so that $di_{jt} = 1$ if $i = j$ and zero otherwise
- z_j is the vector of time constant variables (gender, race, ...) not in the model because they are multicollineal with the dummies
- w_{jt} is the vector of time-varying variables

FE in R

The whole process of demeaning the variables and running the OLS on the transformed equation is done by the command:

$$plm(formula, data = data.plm, model = "within")$$

```
> wage.fe <- plm(lwage ~ educ + exper + expersq +
+      union + married + black + hisp + pub, data = data2,
+      model = "within")
```

FE in R

```
> summary(wage.fe)

Oneway (individual) effect Within Model

Call:
plm(formula = lwage ~ educ + exper + expersq + union + married +
    black + hisp + pub, data = data2, model = "within")

Balanced Panel: n=545, T=8, N=4360

Residuals :
    Min.  1st Qu.   Median  3rd Qu.     Max.
-4.17000 -0.12600  0.00992  0.15900  1.47000

Coefficients :
           Estimate  Std. Error t-value  Pr(>|t|)
exper    0.11645698  0.00843090 13.8131 < 2.2e-16 ***
expersq -0.00428857  0.00060544 -7.0834 1.668e-12 ***
union    0.08120303  0.01931592  4.2039 2.683e-05 ***
married  0.04510615  0.01831141  2.4633   0.01381 *
pub      0.03492668  0.03860819  0.9046   0.36571
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    572.05
Residual Sum of Squares: 470.1
R-Squared      :  0.17822
      Adj. R-Squared :  0.15574
F-statistic: 165.256 on 5 and 3810 DF, p-value: < 2.22e-16
```

LSDV in R

It is using pooling but you have to create hte dummy variables. Check the program.

LPDM: First difference (FD)

$$y_{jt} = \delta\ t + \mathbf{x}_{jt}\beta + c_j + \epsilon_{jt} \quad t = 1, \ldots, T$$

- An alternative to fixed effects is to use first differences.
- This also eliminates c_j from the above equation by taking first differences
- Transformed equation

$$\Delta y_{jt} = \delta + \Delta \mathbf{x}_{jt}\beta + \Delta \epsilon_{jt} \quad t = 2, \ldots, T$$

  where $\Delta y_{jt} = y_{jt} - y_{j(t-1)}$
- The pooled OLS estimator on the transformed equation is consistent

LPDM: First difference (FD)

- OLS on the transformed equation: $\Delta y_{jt} = y_{jt} - y_{j(t-1)}$ yields a consistent estimator

$$\hat{\beta}^{FD} = (\Delta X' \Delta X)^{-1} \Delta X' \Delta Y$$

- If
  - FD.1 Strict exogeneity conditional on the unobserved effect
  - FD.2 rank $\sum_{t=2}^{T} E(\Delta \mathbf{x}'_{jt} \Delta \mathbf{x}_{jt}) = k$
    - The model does not include intercept or time-invariant explanatory variables

LPDM: First difference (FD)

Efficiency if

FD.3 $Var(\Delta\epsilon_j|\mathbf{x}_{j1},\ldots,\mathbf{x}_{jT},c_j) = \sigma^2_{\Delta\epsilon} I_{T-1}$

- Homokedaticity: $Var(\epsilon_{jt}|\mathbf{x}_j, c_j) = \sigma^2$
- The first difference of the errors is serially uncorrelated

If this condition is violated then, we can compute a robust variance matrix for inference.

LPDM: First difference (FD)

$$\Delta y_{jt} = \delta + \Delta \mathbf{x}_{jt}\beta + \Delta \epsilon_{jt}$$

- FD exploits changes between two periods to estimate parameters (one period of observations is lost!)
- FE exploits deviations from average over time
- With $T = 2$, FE and FD are identical
- If time variation is small then we get imprecise estimates (educ)
- ϵ_{jt} uncorrelated with $\mathbf{x}_{j(t-1)}$, $\mathbf{x}_{jt}$ and $\mathbf{x}_{j(t+1)}$ is the necessary condition for consistency
- If $\Delta\epsilon_{jt} = \epsilon_{jt} - \epsilon_{j(t-1)}$ is serially correlated $\Rightarrow$ we need robust variance estimator; see equation (10.70) in Wooldridge

FD in R

$$plm(Y\ X, data = data.plm, model = "fd")$$

```
>wage.fd<-plm(lwage ~ educ+ exper+ expersq+union+ married
+ black+ hisp+ pub, data = data2, model = "fd")
```

FD in R

```
> summary(wage.fd)
Oneway (individual) effect First-Difference Model

Call:
plm(formula = lwage ~ educ + exper + expersq + union + married +
    black + hisp + pub, data = data2, model = "fd")

Balanced Panel: n=545, T=8, N=4360

Residuals :
   Min. 1st Qu.  Median 3rd Qu.    Max.
-4.5800 -0.1460 -0.0124  0.1340  4.8400

Coefficients :
              Estimate Std. Error t-value  Pr(>|t|)
(intercept)  0.1154086  0.0195893  5.8914 4.161e-09 ***
expersq     -0.0038755  0.0013863 -2.7956  0.005207 **
union        0.0425429  0.0196588  2.1641  0.030521 *
married      0.0377588  0.0229311  1.6466  0.099719 .
pub          0.0421258  0.0409964  1.0275  0.304228
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    751.19
Residual Sum of Squares: 747.83
R-Squared      :  0.0044794
      Adj. R-Squared :  0.0044735
F-statistic: 4.28584 on 4 and 3810 DF, p-value: 0.0018406
```

LPDM: Random effects (RE)

- One problem with FE and FD is that they eliminate all time-invariant explanatory variables, e.g. race, gender, etc.
- Often, however, we are interested in the effects of these variables
- If c_j is uncorrelated with $\mathbf{x}_{jt}$, we can used pooled OLS on the original model.
- However, if we are willing to impose more structure (more assumptions on the model). Then, we can get more precise estimates with RE estimation

LPDM: Random effects (RE)

$$y_{jt} = \mathbf{x}_{jt}\beta + c_j + \epsilon_{jt}$$

RE.1 A) $E(\epsilon_{jt}|\mathbf{x}_{jt}, c_j) = 0$ for $t = 1, \dots T$

B) $E(c_j|\mathbf{x}_j) = E(c_j) = 0$ (independence)

RE.2 rank $\sum_t E(\mathbf{x}'_{jt}\Omega^{-1}\mathbf{x}_{jt}) = k + 1$

- Assumption RE.1 A) is the strict exogeneity of $\mathbf{x}_{jt}$ conditional on the unobserved effect
- Assumption RE.1 B) means that $\mathbf{x}_j$ and c_j are independent (this implies uncorrelated)
- Therefore $\nu_{jt} = \epsilon_{jt} + c_j$ is uncorrelated with $\mathbf{x}_{jt}$
- However there is a serial correlation in ν_{jt} and therefore pooled OLS will not be efficient.

LPDM: Random effects (RE)

Idea: Unless the OLS, the Generalised Least Squares (GLS) can exploit the serial correlation of ν

- I.e. we impose some "structure" (assumptions) on the errors for each person, and
- use repeated observations for each individual to get more efficient estimates

LPDM: Random effects (RE)

Structure assumptions (homoskedasticity)

RE.3 A) $E(\epsilon_{jt}\epsilon'_{jt}|\mathbf{x}_j, c_j) = \sigma^2_\epsilon I_T$
B) $E(c_j^2|\mathbf{x}_j) = \sigma^2_c$

- Assumption A) means:
  - errors are homoskedasticity
  - errors are not autocorrelated. I.e. for each individual i the error term at time t is not correlated with previous or future errors
- Assumption B) means that c_j is homoskedastic (its variance does not change for each individual)

LPDM: Random effects (RE)

$$y_{jt} = x_{jt}\beta + c_j + \epsilon_{jt}$$

- Define: $\nu_{jt} = c_j + \epsilon_{jt}$
- Under RE.1 + RE.3:

$$Var(\nu_{jt}) = Var(c_j + \epsilon_{jt}) = \sigma_c^2 + \sigma_\epsilon^2$$
$$Cov(\nu_{jt}, \nu_{js}) = Cov(c_j + \epsilon_{jt}, c_j + \epsilon_{js}) = \sigma_c^2$$

LPDM: Random effects (RE)

Hence, the co-variance matrix of ν_{jt} for person i is:

$$\Omega = E(\nu_j \nu_j') = \begin{pmatrix} \sigma_c^2 + \sigma_\epsilon^2 & \sigma_c^2 & \dots & \sigma_c^2 \\ \sigma_c^2 & \sigma_c^2 + \sigma_\epsilon^2 & \dots & \vdots \\ \vdots & & \ddots & \sigma_c^2 \\ \sigma_c^2 & & & \sigma_c^2 + \sigma_\epsilon^2 \end{pmatrix}$$

- This matrix is the same for all individuals
- There is no correlation across error terms for different individuals

LPDM: Random effects (RE)

For the standard linear model, without the c_j:

$$\Omega = E(\nu_j \nu_j') = \begin{pmatrix} \sigma_\epsilon^2 & 0 & \dots & 0 \\ 0 & \sigma_\epsilon^2 & \dots & \vdots \\ \vdots & & \ddots & 0 \\ 0 & & & \sigma_\epsilon^2 \end{pmatrix} = \sigma_\epsilon^2 I_T$$

LPDM: Random effects (RE)

The random effect estimator:

$$\hat{\beta}^{RE} = \left(\sum_{j=1}^{n} \mathbf{x}_j' \hat{\Omega}^{-1} \mathbf{x}_j \right)^{-1} \sum_{j=1}^{n} \mathbf{x}_j' \hat{\Omega}^{-1} y_j$$

- NB! We need to estimate Ω to obtain this estimator (Wooldridge, Chapter 10.4.2)
- If $\Omega = \sigma_\epsilon^2 I_T$ then the RE estimator is the OLS estimator

LPDM: Random effects (RE)

In summary:

- OLS and RE require c_j to be uncorrelated with $\mathbf{x}_{jt}$
- This is a strong (even unrealistic) assumption if we believe that unobservable effects are indeed present.
- RE can get more efficient estimates than OLS ? but do so by adding more assumptions (RE.3)

RE in R

$$plm(formula, data = data.plm, model = "random")$$

You can choose amongst 4 methods to estimate Ω, by default $random.method = "swar"$. The other options are $"walhus", "amemiya"$ and $"nerlove"$.

```
> wage.re<-plm(lwage ~ educ+ exper+ expersq+union+ married
+ black+ hisp+ pub, data = data2, model = "random",
random.method="swar")
```

RE in R

```
> summary(wage.re)
Oneway (individual) effect Random Effect Model
   (Swamy-Arora's transformation)

Call:
plm(formula = lwage ~ educ + exper + expersq + union + married +
    black + hisp + pub, data = data2, model = "random", random.method = "swar")
Balanced Panel: n=545, T=8, N=4360
Effects:
                 var std.dev share
idiosyncratic 0.1234  0.3513 0.539
individual    0.1055  0.3248 0.461
theta:  0.6429

Residuals :
   Min. 1st Qu.  Median 3rd Qu.    Max.
-4.5800 -0.1450  0.0234  0.1860  1.5400

Coefficients :
               Estimate  Std. Error t-value  Pr(>|t|)
(Intercept) -0.10431124  0.11083404 -0.9411 0.3466813
educ         0.10102372  0.00892187 11.3232 < 2.2e-16 ***
exper        0.11178514  0.00827093 13.5154 < 2.2e-16 ***
expersq     -0.00405745  0.00059198 -6.8540 8.189e-12 ***
union        0.10641338  0.01786690  5.9559 2.791e-09 ***
married      0.06254648  0.01677617  3.7283 0.0001952 ***
black       -0.14400263  0.04764392 -3.0225 0.0025218 **
hisp         0.01972690  0.04263026  0.4627 0.6435709
pub          0.03015542  0.03646707  0.8269 0.4083267
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    656.8
```

Compare estimators of $educ$

	education	significant?	expersq	significant?
Pooled OLS				
FE				
FD				
RE				

FE vs FD

With $T = 2$:

- FD and FE estimates are the same
- FD is easier to implement
- We can apply heteroskedastic-robust inference directly in the FD

With $T > 2$:

- FE is more efficient under assumption FE.3 (ϵ_{jt} are serially uncorrelated)
- FD is more efficient when ϵ_{jt} follows a random walk
- Correlations between ϵ_{jt} and $\mathbf{x}_{jt}$ tend to appear when there are measurement errors, omitted variables or simultaneity. This causes FD and FE to be inconsistent

RE vs FE

- When $\mathbf{x}_t$ does not vary much over time, FE and FD can lead to imprecise estimates
- If $E(c_j|\mathbf{x}_{jt}) = E(c_j)$ then the RE has smaller variances than the FE or FD estimators
- Comparing between the RE or the FE model can be done with the Hausman test
  - FE is consistent when c_j is correlated with $\mathbf{x}_{jt}$, but RE is inconsistent
  - The models should not have time dummies or time effects.
  - Hausman null hypothesis $H_0 : E(c_j|\mathbf{x}_{jt}) = 0$
  - In R: $phtest(model.fe, model.re)$ or $phtest(formula, data)$

Hausman test in R

```
> phtest(wage.re, wage.fd)

Hausman Test

data:  lwage ~ educ + exper + expersq + union + married
+ black + hisp + pub
chisq = 66.9911, df = 4, p-value = 9.791e-14
alternative hypothesis: one model is inconsistent
```

The null hypothesis is rejected, then the omitted variables c_j are correlated with X and FE is preferible to RE.

Summary

If there is no individual unknown effects $\Rightarrow$ Pooled OLS is the most efficient estimator

If there are individual unknown effects c_j:

- If c_j is uncorrelated with $\mathbf{x}_j$
  - POLS is consistent but inefficient (POLS.3 is not satisfied)
  - Random effects is the most efficient estimator if $RE.3$ is satisfied
- If c_j is correlated with $\mathbf{x}_j$ $\Rightarrow$
  - The POLS is inconsistent
  - Then FE or FD are the best methods
- If $\mathbf{x}_j$ contains time-invariant variables $\Rightarrow$ RE or LSDV