Panel Data

Jerome
Dumortier

Simple Panel
Data Methods

Fixed Effects
Model

Random
Effects Model

Panel Data

Jerome Dumortier

17 August 2023

Panel Data

Jerome
Dumortier

Simple Panel
Data Methods

Fixed Effects
Model

Random
Effects Model

Overview

Required R Packages

- plm
- lmtest

Documentation

- Panel data econometrics in R

Note regarding Notation

- Use of the package stargazer due to the large number of variables.

Panel Data

Jerome
Dumortier

Simple Panel
Data Methods

Fixed Effects
Model

Random
Effects Model

Introduction I

Pooled data: Combination of multiple cross-sectional data over time

- Two or more different observational units over time
- Grades in an economics class based on students' concentration combined from multiple semesters
- American Community Survey (ACS)

Panel data: Repeated measurement on the same individuals i over time t.

- Individual units can be people, states, firms, counties, countries, etc.
- National Longitudinal Survey (NLSY79)
- Necessary adjustments of standard error due to correlation across time.

For NLSY79: Accessing data $\Rightarrow$ investigator $\Rightarrow$ Begin searching as guest $\Rightarrow$ Pick income as an example

Panel Data

Jerome
Dumortier

Simple Panel
Data Methods

Fixed Effects
Model

Random
Effects Model

Introduction II

Some assumptions about linear panel models:

- Regular time intervals
- Errors are correlated
- Parameters may vary across individuals or time
- Intercept: Individual specific effects model (fixed or random)

Note that the General Social Survey (GSS) is not a panel data set because different respondents are questioned every year.

Panel Data

Jerome
Dumortier

Simple Panel
Data Methods

Fixed Effects
Model

Random
Effects Model

Examples and Advantages

Panel Study of Income Dynamics (PSID)

- Data on approximately 5,000 families on various socioeconomic and demographic variables

Survey of Income and Program Participation (SIPP)

- Interviews about economic condition of respondents

Advantages

- Takes into account heterogeneity among observational units, e.g., firms, states, counties, etc.
- Better understanding on the dynamics of change for observational units over time.
- Combines cross-sectional data with time series data leading to more complete behavioral models

Panel Data

Jerome
Dumortier

Simple Panel
Data Methods
Fixed Effects
Model
Random
Effects Model

Terminology and Types

Balanced versus unbalance panel:

- A balanced panel has the same number of time-series observations for each subject or observational unit, whereas an unbalanced panel does not.

Short versus long panel:

- A short panel has a larger number of subjects or observational units than there are time periods. A long panel has a greater number of time periods than observational units.

Types of regression models:

- Pooled Ordinary Least Square model
- Fixed effects model
- Random effects model

Panel Data

Jerome
Dumortier

Simple Panel
Data Methods

Fixed Effects
Model

Random
Effects Model

Simple Panel Data Methods

Panel Data

Jerome
Dumortier

Simple Panel
Data Methods

Fixed Effects
Model

Random
Effects Model

fertil1: Variables

Data from the General Social Survey for the years 1974 to 1984

- *year*: 72 to 84, even
- *educ*: years of schooling
- *meduc* and *feduc*: mother's and father's education
- *kids*: number children ever born
- *east*, *northcentral*, and *west*: 1 if lived in at 16
- *farm*: 1 if on farm at 16
- *otherrural*: 1 if other rural at 16
- *town*: 1 if lived in town at 16
- *smallcity*: 1 if in small city at 16

Source: Jeffrey Wooldridge, Introductory Econometrics: A Modern Approach

Panel Data

Jerome
Dumortier

Simple Panel
Data Methods

Fixed Effects
Model

Random
Effects Model

fertil1: Estimation

```
## 
## ================================================
##                        Dependent variable:
##                    -----------------------------
##                                kids
## ------------------------------------------------
## educ                     -0.130*** (0.019)
## age                       0.499*** (0.141)
## I(age2)                  -0.005*** (0.002)
## east                      0.061 (0.133)
## northcentral              0.220* (0.121)
## west                      0.051 (0.168)
## y82                      -0.414** (0.174)
## y84                      -0.565*** (0.177)
## Constant                 -6.785** (3.099)
## ------------------------------------------------
## Observations              1,129
## R2                        0.099
## Adjusted R2               0.086
## Residual Std. Error       1.581 (df = 1112)
## F Statistic           7.671*** (df = 16; 1112)
## ================================================
## Note:            *p<0.1; **p<0.05; ***p<0.01
```

Panel Data

Jerome
Dumortier

Simple Panel
Data Methods

Fixed Effects
Model

Random
Effects Model

fertil1: Interpretation

Evolution of fertility rates over time after controlling of other observable factors:

- Base year: 1972
- Negative coefficients indicate a drop in fertility in the early 1980's
- Coefficient of $y82$ (-0.41) indicates that women had on average 0.41 less children, i.e., 100 women had 41 kids less than 1972
- This drop is independent from education since we are controlling for education.
- More educated women have fewer children
- Assumes that the effect of each explanatory variable remains constant.

Panel Data

Jerome
Dumortier

Simple Panel
Data Methods

Fixed Effects
Model

Random
Effects Model

cps7885: Setup

Interact year dummy with key explanatory variables to see if the effect of that variable has changed over time:

$$\ln(wage) = \beta_0 + \gamma_0 \cdot y85 + \beta_1 \cdot educ + \gamma_1 \cdot y85 \cdot educ + \beta_2 \cdot exper + \beta_3 \cdot exper^2$$
$$+ \beta_4 \cdot union + \beta_5 \cdot female + \gamma_5 \cdot y85 \cdot female$$

Interpretation:

- β_0 is the 1978 intercept
- $\beta_0 + \gamma_0$ is the 1985 intercept
- β_1 is the return to education in 1978
- $\beta_1 + \gamma_1$ is the return to education in 1985
- γ_1 measures how the return to education has changed over the seven year period

Panel Data

Jerome
Dumortier

Simple Panel
Data Methods

Fixed Effects
Model

Random
Effects Model

cps7885: Estimation

```
cps7885$y85    = ifelse(cps7885$year==85,1,0)
bhat           = lm(log(wage)~y85+educ+y85:educ+exper+
                   I(exper^2)+union+female+y85:female,
                   data=cps7885)
```

Panel Data

Jerome
Dumortier

Simple Panel
Data Methods

Fixed Effects
Model

Random
Effects Model

cps7885: Results

```
stargazer(bhat,type="text",no.space=TRUE,single.row=TRUE)
```

```
##
## ===============================================
##                     Dependent variable:
##                  ---------------------------
##                          log(wage)
## -----------------------------------------------
## y85                     0.118 (0.124)
## educ                   0.075*** (0.007)
## exper                  0.030*** (0.004)
## I(exper2)             -0.0004*** (0.0001)
## union                  0.202*** (0.030)
## female                -0.317*** (0.037)
## y85:educ               0.018** (0.009)
## y85:female             0.085* (0.051)
## Constant               0.459*** (0.093)
## -----------------------------------------------
## Observations               1,084
## R2                         0.426
## Adjusted R2                0.422
## Residual Std. Error   0.413 (df = 1075)
## F Statistic       99.804*** (df = 8; 1075)
## ===============================================
## Note:             *p<0.1; **p<0.05; ***p<0.01
```

Panel Data

Jerome
Dumortier

Simple Panel
Data Methods

Fixed Effects
Model

Random
Effects Model

cps7885: Interpretation

Interpretation

- 1978 return to education: 7.47%
- 1985 return to education: 7.47%+1.85% = 9.32%
- 1978 gender gap: 31.67%
- 1985 gender gap: 31.67% - 8.51% = 23.16%

Panel Data

Jerome
Dumortier

Simple Panel
Data Methods

Fixed Effects
Model

Random
Effects Model

Data kiel: Setup I

Data set about home values near the location of an garbage incinerator

- Run 1981 data
- Run 1978 data

Difference-in-difference estimator: -\$30,688-(-\$18,824)=-\$11,864

$$\hat{\delta}_1 = (price_{81,near} - price_{81,far}) - (price_{78,near} - price_{78,far})$$

where $\hat{\delta}_1$ represents the difference over time in average differences in housing prices in the two locations.

Panel Data

Jerome
Dumortier

Simple Panel
Data Methods

Fixed Effects
Model

Random
Effects Model

Data kiel: Setup II

To determine statistical significance:

$$price = \beta_0 + \gamma_0 \cdot y81 + \beta_1 \cdot nearinc + \gamma_1 \cdot y81 \cdot nearinc$$

Interpretation

- β_0: Average home value which is not near the garbage incinerator
- $\gamma_0 \cdot y81$: Average change in housing values for all homes
- $\beta_1 \cdot nearinc$: Location effect that is not due to the incinerator
- γ_1: Decline in housing values due to incinerator

Homes have lost 9.3% in values when including additional independent variables and using the natural logarithm of price.

Panel Data

Jerome
Dumortier

Simple Panel
Data Methods

Fixed Effects
Model

Random
Effects Model

Data `kiel`: Naive Implementation in R

```r
kiel81         = subset(kiel,year==1981)
bhat81         = lm(rprice~nearinc,data=kiel81)
kiel78         = subset(kiel,year==1978)
bhat78         = lm(rprice~nearinc,data=kiel78)
```

Panel Data

Jerome
Dumortier

Simple Panel
Data Methods

Fixed Effects
Model

Random
Effects Model

Data `kiel`: Naive Results

```
## 
## =====================================================================
##                              Dependent variable:
##                    -------------------------------------------------
##                                       rprice
##                          (1)                       (2)
## ---------------------------------------------------------------------
## nearinc           -30,688.270*** (5,827.709)  -18,824.370*** (4,744.594)
## Constant          101,307.500*** (3,093.027)   82,517.230*** (2,653.790)
## ---------------------------------------------------------------------
## Observations              142                       179
## R2                       0.165                     0.082
## Adjusted R2              0.159                     0.076
## Residual Std. Error  31,238.040 (df = 140)     29,431.960 (df = 177)
## F Statistic       27.730*** (df = 1; 140)     15.741*** (df = 1; 177)
## =====================================================================
## Note:                                    *p<0.1; **p<0.05; ***p<0.01
```

Panel Data

Jerome
Dumortier

Simple Panel
Data Methods

Fixed Effects
Model

Random
Effects Model

Data `kiel`: Implementation in R

```
bhat1    = lm(rprice~y81+nearinc+y81:nearinc,data=kiel)
bhat2    = lm(rprice~y81+nearinc+y81:nearinc+age+agesq,data=kiel)
bhat3    = lm(rprice~y81+nearinc+y81:nearinc+age+agesq+cbd+rooms+
            area+land+baths,data=kiel)
```

Panel Data

Jerome
Dumortier

Simple Panel
Data Methods

Fixed Effects
Model

Random
Effects Model

Data `kiel`: Results

```
## 
## =====================================================================================
##                                      Dependent variable:
##                    -----------------------------------------------------------------
##                                             rprice
##                         (1)                  (2)                   (3)
## -------------------------------------------------------------------------------------
## y81                18,790.290***        21,321.040***         14,115.710***
##                     (4,050.065)          (3,443.631)           (2,802.303)
## nearinc            -18,824.370***         9,397.936*            3,618.020
##                     (4,875.322)          (4,812.222)           (4,644.530)
## rooms                                                          3,310.163**
##                                                                (1,665.357)
## area                                                            17.920***
##                                                                  (2.310)
## land                                                            0.136***
##                                                                  (0.031)
## y81:nearinc         -11,863.900         -21,920.270***        -14,269.820***
##                     (7,456.646)          (6,359.745)           (4,999.499)
## Constant           82,517.230***        89,116.540***          13,055.760
##                     (2,726.910)          (2,406.051)          (11,272.270)
## -------------------------------------------------------------------------------------
## Observations           321                  321                   321
## R2                    0.174                0.414                 0.658
## Adjusted R2           0.166                0.405                 0.647
## Residual Std. Error 30,242.900 (df = 317) 25,543.290 (df = 315) 19,667.290 (df = 310)
## F Statistic     22.251*** (df = 3; 317) 44.591*** (df = 5; 315) 59.742*** (df = 10; 310)
## =====================================================================================
## Note:                                                *p<0.1; **p<0.05; ***p<0.01
```

Panel Data

Jerome
Dumortier

Simple Panel
Data Methods

Fixed Effects
Model

Random
Effects Model

Grunfeld Data

The data set is used in many textbooks and comes also with the package plm. Data on 10 companies over the period 1935 to 1954:

- *inv*: Investment
- *value*: Value of the firm
- *capital*: Capital stock

Companies of interest for this class: GM (firm 1), U.S. Steel (firm 2), GE (firm 3), Westinghouse (firm 8)

Panel Data

Jerome
Dumortier

Simple Panel
Data Methods

Fixed Effects
Model

Random
Effects Model

Pooled OLS Model

Pooling all cross-sectional and time series observations into a single data set and running an OLS regression.

$$inv_i = \beta_0 + \beta_1 \cdot value_i + \beta_2 \cdot capital_i$$

General formulation of the pooled model

$$y_{it} = \beta_0 + \beta_1 \cdot x_i + \epsilon_i$$

Issues with pooled OLS model:

- Ignores heterogeneity among the observations and time.

With heterogeneity: Biased and inconsistent estimates due to correlation between independent variables and error term.

Panel Data

Jerome
Dumortier

Simple Panel
Data Methods

Fixed Effects
Model

Random
Effects Model

Data Preparation and Pooled OLS Model

Use of firms 1, 2, 3, and 8:

```
grunfeld = subset(grunfeld,grunfeld$firm %in% c(1,2,3,8))
```

To use the functions from plm, define data as a panel data set:

```
grunfeld = pdata.frame(grunfeld,index=c("firm","year"))
```

Running a simple OLS model on the data:

- Using the regular lm() function
- Using the plm() function and specifying the model as **pooling**
- Name the outputs grunfeld.ols and grunfeld.pooling

Panel Data

Jerome
Dumortier

Simple Panel
Data Methods

Fixed Effects
Model

Random
Effects Model

Pooled OLS Model

```
grunfeld.ols      = lm(inv~value+capital,data=grunfeld)
grunfeld.pooling  = plm(inv~value+capital,data=grunfeld,
                        model="pooling")
```

Panel Data

Jerome
Dumortier

Simple Panel
Data Methods

Fixed Effects
Model

Random
Effects Model

Pooled OLS Model

```
stargazer(grunfeld.ols,grunfeld.pooling,no.space=TRUE,
          single.row=TRUE,type="text")
```

```
##
## ============================================================
##                            Dependent variable:
##                    ----------------------------------------
##                                      inv
##                          OLS                   panel
##                                                linear
##                          (1)                    (2)
## ------------------------------------------------------------
## value            0.111*** (0.014)      0.111*** (0.014)
## capital          0.300*** (0.049)      0.300*** (0.049)
## Constant         -62.832** (29.725)    -62.832** (29.725)
## ------------------------------------------------------------
## Observations            80                    80
## R2                     0.755                 0.755
## Adjusted R2            0.748                 0.748
## Residual Std. Error  142.916 (df = 77)
## F Statistic (df = 2; 77)  118.424***          118.424***
## ============================================================
## Note:                      *p<0.1; **p<0.05; ***p<0.01
```

Panel Data

Jerome
Dumortier

Simple Panel
Data Methods

Fixed Effects
Model

Random
Effects Model

Fixed Effects Model

Panel Data

Jerome
Dumortier

Simple Panel
Data Methods

Fixed Effects
Model

Random
Effects Model

Theoretical Concepts I

Fixed effects model or Least-Squares Dummy Variable (LSDV) regression

• Constant slope coefficients but varying intercept over i

Regression equation:

$$inv_{it} = \beta_{0i} + \beta_1 \cdot value_{it} + \beta_2 \cdot capital_{it}$$

with $i = 1, 2, 3, 4$ and $t = 1, 2, \ldots, 20$. This model can also be written as

$$inv_{it} = \alpha_0 + \alpha_1 \cdot D_{1i} + \alpha_2 \cdot D_{2i} + \alpha_3 \cdot D_{3i} + \beta_1 \cdot value_{it} + \beta_2 \cdot capital_{it}$$

Individual specific effects:

$$y_{it} = \alpha_i + \beta_i \cdot x_{it} + \epsilon_{it}$$

α_i can be fixed or random

Panel Data

Jerome
Dumortier

Simple Panel
Data Methods

Fixed Effects
Model

Random
Effects Model

Theoretical Concepts II

Fixed effects model

- Intercept β_{0i} is firm specific.
- For an individual, this could be education and/or ability, possibly correlated with independent variables
- Intercept is time-invariant.
- Slope coefficients do not vary across individuals (firms) or time

Implementation in R

```
grunfeld.fixed = plm(inv~value+capital,data=grunfeld,
                     model="within")
```

Panel Data

Jerome
Dumortier

Simple Panel
Data Methods

Fixed Effects
Model

Random
Effects Model

Results

```
stargazer(grunfeld.fixed,no.space=TRUE,
          single.row=TRUE,type="text")
```

```
##
## ========================================
##                 Dependent variable:
##              ---------------------------
##                           inv
## ----------------------------------------
## value          0.108*** (0.018)
## capital        0.345*** (0.027)
## ----------------------------------------
## Observations           80
## R2                    0.806
## Adjusted R2           0.792
## F Statistic   153.291*** (df = 2; 74)
## ========================================
## Note:        *p<0.1; **p<0.05; ***p<0.01
```

Panel Data

Jerome
Dumortier

Simple Panel
Data Methods

Fixed Effects
Model

Random
Effects Model

Firm-Specific Intercepts and Hypothesis Test

In order to get the firm specific intercepts:

```
fixef(grunfeld.fixed)
```

```
##       1       2       3       8
## -85.515  94.988 -246.228 -59.386
```

Testing whether a fixed effects or OLS is appropriate (H_0: OLS better):

```
pFtest(grunfeld.fixed,grunfeld.ols)
```

```
##
##  F test for individual effects
##
## data:  inv ~ value + capital
## F = 67.215, df1 = 3, df2 = 74, p-value < 2.2e-16
## alternative hypothesis: significant effects
```

If the p-value is below 0.05 then the fixed effects model is a better choice.

Panel Data

Jerome
Dumortier

Simple Panel
Data Methods

Fixed Effects
Model

Random
Effects Model

Implementation in R using `lm()`

Implementation of a fixed effects model with the command `lm()`

```
bhat = lm(inv~value+capital+factor(firm),data=grunfeld)
```

Panel Data

Jerome
Dumortier

Simple Panel
Data Methods

Fixed Effects
Model

Random
Effects Model

Random Effects Model

Panel Data

Jerome
Dumortier

Simple Panel
Data Methods

Fixed Effects
Model

Random
Effects Model

Theoretical Concepts

The general fixed effects model can be expressed as

$$y_{it} = \beta_{0i} + \beta_1 \cdot x_{1,it} + \beta_2 \cdot x_{2,it} + \epsilon_{it}$$

Instead of treating β_{0i} as fixed, the random model assumes

$$\beta_{0i} = \beta_0 + \upsilon_i$$

where υ_i is random error term with a mean of zero and variance σ_υ^2. According to Gujarati: *What we are essentially saying is that the four firms included in our sample are a drawing from a much larger universe of such companies and that they have a common mean value for the intercept β_0 and the individual differences in the intercept values of each company are reflected in the error term υ_i.*

Panel Data

Jerome
Dumortier

Simple Panel
Data Methods

Fixed Effects
Model

Random
Effects Model

Implementation in R

```
grunfeld.random = plm(inv~value+capital,data=grunfeld,model="random")
stargazer(grunfeld.random,no.space=TRUE,single.row=TRUE,type="text")
```

```
##
## ========================================
##                   Dependent variable:
##                 ---------------------------
##                           inv
## ----------------------------------------
## value           0.108*** (0.017)
## capital         0.345*** (0.027)
## Constant        -73.085 (81.172)
## ----------------------------------------
## Observations            80
## R2                     0.803
## Adjusted R2            0.798
## F Statistic        314.851***
## ========================================
## Note:         *p<0.1; **p<0.05; ***p<0.01
```

Panel Data

Jerome
Dumortier

Simple Panel
Data Methods

Fixed Effects
Model

Random
Effects Model

Breusch-Pagan Lagrange Multiplier (LM) Test

For random effects models: Null hypothesis of no panel effect, i.e., OLS is better. If p-value is below 0.05, we reject the null hypothesis and thus, a random effects model is more appropriate than the OLS.

```
plmtest(grunfeld.pooling,type=c("bp"))
```

```
##
##  Lagrange Multiplier Test - (Breusch-Pagan)
##
## data:  inv ~ value + capital
## chisq = 378.44, df = 1, p-value < 2.2e-16
## alternative hypothesis: significant effects
```

Panel Data

Jerome
Dumortier

Simple Panel
Data Methods

Fixed Effects
Model

Random
Effects Model

Hausman Test: Fixed or Random Model

The Hausman Test tests the null hypothesis that the preferred model is a random effects model. It basically tests whether the unique errors are correlated with the regressors.

```
phtest(grunfeld.random,grunfeld.fixed)
```

```
##
##  Hausman Test
##
## data:  inv ~ value + capital
## chisq = 0.14882, df = 2, p-value = 0.9283
## alternative hypothesis: one model is inconsistent
```

Panel Data

Jerome
Dumortier

Simple Panel
Data Methods

Fixed Effects
Model

Random
Effects Model

Testing for Heteroscedasticity

```
bptest(inv~value+capital+factor(firm),data=grunfeld)
```

```
##
##   studentized Breusch-Pagan test
##
## data:  inv ~ value + capital + factor(firm)
## BP = 25.375, df = 5, p-value = 0.0001179
```

If the p-value is below 0.05, then we have heteroscedasticity.

Panel Data

Jerome
Dumortier

Simple Panel
Data Methods

Fixed Effects
Model

Random
Effects Model

Heteroscedasticity Consistent Coefficients and Standard Errors

```
coeftest(grunfeld.fixed,vcovHC)
```

```
##
## t test of coefficients:
##
##           Estimate Std. Error t value  Pr(>|t|)
## value    0.108400   0.014293  7.5839 7.902e-11 ***
## capital  0.345058   0.031152 11.0765 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```