We will fail when we fail to try. - Rosa Parks

PANEL DATA

What is Panel Data?

- Panel data is a combination of cross-sectional data and time series data (time series does NOT need to be equally spaced)
- Panel data is also referred to as longitudinal data
- Examples:
  - Cost of tickets for 6 U.S. airlines across the years 1970-1984.
  - Gas prices across states between 1990 and 2010.
  - Profit of financial advisors measured across years.

Why Use Panel Data Methods?

- There are two advantages to using panel data methods.
  1. Increased sample size.
  2. Control for unobserved differences between individual subjects.

Why Use Panel Data Methods?

- There are two advantages to using panel data methods.
  1. Increased sample size.
  2. Control for unobserved differences between individual subjects.

- Cost of tickets for 6 U.S. airlines across the years 1970-1984.
  - Single cross-section only has 6 observations.
  - Single time-series only has 15 observations.
  - Panel data has $6 \times 15 = 90$ observations.

Why Use Panel Data Methods?

- There are two advantages to using panel data methods.
  1. Increased sample size.
  2. Control for unobserved differences between individual subjects.

- There exists some unobservable variables that we know influences our results.
- Could this cause a concern? OMITTED VARIABLE BIAS!

Why Use Panel Data Methods?

- There are two advantages to using panel data methods.
  1. Increased sample size.
  2. Control for unobserved differences between individual subjects.

- Profit of financial advisors measured across years.
  - Success could be related to motivation – how to measure?
  - Sales people with “it” quality.
- These unobserved variables influence our panel data model.

PANEL DATA MODEL

Data Structure

- Data can either be “Wide Data” or “Long Data”
- Most programs want their data to be Long Data

“Wide Data”

Person	T1	T2	T3	T4	T5
1	Y_{11}	Y_{12}	Y_{13}	Y_{14}	Y_{15}
2	Y_{21}	Y_{22}	Y_{23}	Y_{24}	Y_{25}
3	Y_{31}	Y_{32}	Y_{33}	Y_{34}	Y_{35}

“Wide Data”

Person	T1	T2	T3	T4	T5
1	Y_{11}	Y_{12}	Y_{13}	Y_{14}	Y_{15}
2	Y_{21}	Y_{22}	Y_{23}	Y_{24}	Y_{25}
3	Y_{31}	Y_{32}	Y_{33}	Y_{34}	Y_{35}

3 cross-sectional observations

“Wide Data”

Person	T1	T2	T3	T4	T5
1	Y_{11}	Y_{12}	Y_{13}	Y_{14}	Y_{15}
2	Y_{21}	Y_{22}	Y_{23}	Y_{24}	Y_{25}
3	Y_{31}	Y_{32}	Y_{33}	Y_{34}	Y_{35}

5 time series observations

"Long Data"

Person	Time	Y	X1	X2
1	1	Y_{11}		
1	2	Y_{12}		
1	3	Y_{13}		
1	4	Y_{14}		
1	5	Y_{15}		
2	1	Y_{21}		
2	2	Y_{22}		
2	3	Y_{23}		
2	4	Y_{24}		

Also easier to
show X variables

Balanced vs. Unbalanced

- You can have **balanced** or **unbalanced** panel data.
- Balanced panel data is defined as data with the number of time periods being equal across all of the different cross-sectional individuals.
- Unbalanced data is defined as data with an unequal number of time periods across different individual cross-sections.

Panel Data Model

- The following is a panel data model for $i = 1, \ldots, n$ cross-sections and $t = 1, \ldots, T$ periods in time:

$$y_{it} = \alpha_i + \beta_1 x_{1,it} + \beta_2 x_{2,it} + \cdots + \beta_k x_{k,it} + \varepsilon_{it}$$

Panel Data Model

- The following is a panel data model for $i = 1, \ldots, n$ cross-sections and $t = 1, \ldots, T$ periods in time:

$$y_{it} = \alpha_i + \beta_1 x_{1,it} + \beta_2 x_{2,it} + \cdots + \beta_k x_{k,it} + \varepsilon_{it}$$

Collection of all of the unobserved variables and their coefficients.

Panel Data Model

- The following is a panel data model for $i = 1, \dots, n$ cross-sections and $t = 1, \dots, T$ periods in time:

$$y_{it} = \alpha_i + \beta_1 x_{1,it} + \beta_2 x_{2,it} + \cdots + \beta_k x_{k,it} + \varepsilon_{it}$$

- How we treat this α_i directly influences the panel data model.

Fixed Effects versus Random Effects

In panel data, fixed effects means that we assume the coefficient (for example α_i) is a fixed unknown constant. This assumes there is some correlation between the unobserved (omitted) variables and α_i.

A random effects model assumes that the coefficient, α_i, varies randomly around some unknown mean (for example μ). In this case, each coefficient $\alpha_i = \mu + v_i$, where $v_i \sim N(0, \sigma_v)$. This assumes the omitted variables and α_i are uncorrelated.

FIXED EFFECTS MODEL

Fixed Effects Model (Cross-section)

$$y_{it} = \alpha_i + \beta_1 x_{1,it} + \beta_2 x_{2,it} + \cdots + \beta_k x_{k,it} + \varepsilon_{it}$$

- In the fixed effects model we are assuming that the α_i's are some fixed unknown quantity and there is some correlation to the omitted variables.

Fixed Effects Model

$$y_{it} = \alpha_i + \beta_1 x_{1,it} + \beta_2 x_{2,it} + \cdots + \beta_k x_{k,it} + \varepsilon_{it}$$

- In the fixed effects model we are assuming that the α_i's are some fixed unknown quantity and there is some correlation to the omitted variables.

Subject specific constant terms

Fixed *Individual* Effects Model

$$y_{it} = \alpha_i + \beta_1 x_{1,it} + \beta_2 x_{2,it} + \cdots + \beta_k x_{k,it} + \varepsilon_{it}$$

- In the fixed effects model we are assuming that the α_i's are some fixed unknown quantity and there is some correlation to the omitted variables.

Different intercepts across subjects with the slopes remaining the same.

Assumptions

- Since the fixed effects model is slightly different than typical OLS, the assumptions change slightly:
  1. For each subject *i,* the following model holds:

$$y_{it} = \alpha_i + \beta_1 x_{1,it} + \beta_2 x_{2,it} + \cdots + \beta_k x_{k,it} + \varepsilon_{it}$$

Assumptions

- Since the fixed effects model is slightly different than typical OLS, the assumptions change slightly:
  1. For each subject *i*, the following model holds:

     $$y_{it} = \alpha_i + \beta_1 x_{1,it} + \beta_2 x_{2,it} + \cdots + \beta_k x_{k,it} + \varepsilon_{it}$$

  2. No perfect collinearity between predictor variables **and** each predictor variable changes across time for some subject.

Assumptions

- Since the fixed effects model is slightly different than typical OLS, the assumptions change slightly:

1. For each subject *i*, the following model holds:

$$y_{it} = \alpha_i + \beta_1 x_{1,it} + \beta_2 x_{2,it} + \cdots + \beta_k x_{k,it} + \varepsilon_{it}$$

2. No perfect collinearity between predictor variables **and** each predictor variable changes across time for some subject.

3. $\varepsilon_{it} \sim N(0, \sigma^2)$

Fixed *Time* Effects

- The one-way fixed effects model for time is :

$$y_{it} = \alpha_{\mathrm{t}} + \beta_1 x_{1,it} + \beta_2 x_{2,it} + \cdots + \beta_k x_{k,it} + \varepsilon_{it}$$

Notice that the subscript on α is now *t* (for time), this effect is now being estimated for each time point

Two-way Fixed Effects Model

- Combine both cross-sectional and time components into a **two-way fixed effects model**:

$$y_{it} = \alpha_i + \beta_1 x_{1,it} + \beta_2 x_{2,it} + \cdots + \beta_k x_{k,it} + \gamma_t + \varepsilon_{it}$$

Airline Example

Data

Christenson Associates airline data (Greene, 2000) measures costs, prices of inputs and utilization rates for 6 airlines from 1970-1984).

First two columns identify airlines ("individuals") and time

I and T

Q=Revenue passenger miles (LnQ…Log of Q)

C=Total cost, in thousands (LnC…Log of Cost)

PF=Fuel price (LnPF…Log of Fuel Price)

LF=Load Factor

One-Way Fixed Individual Effects Model

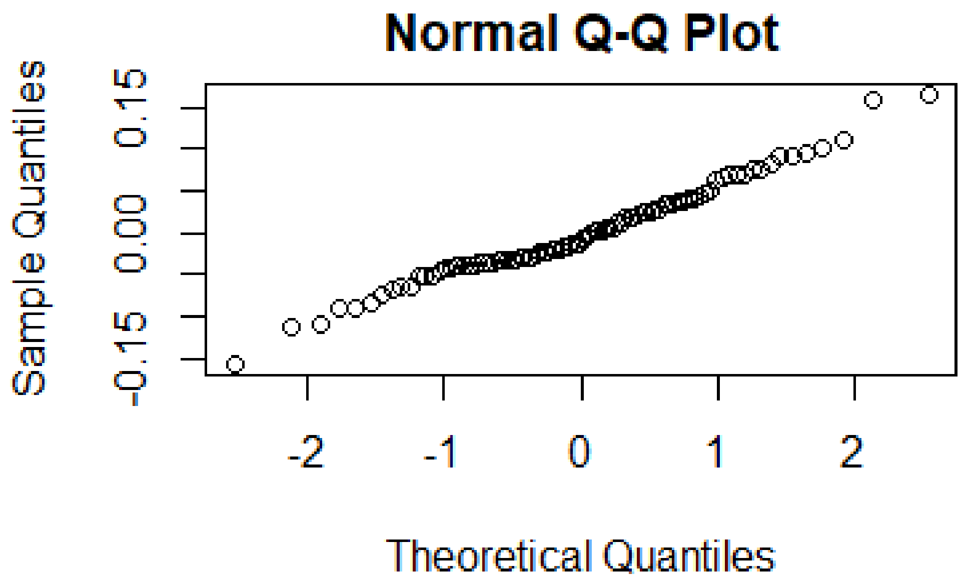
```
model1=plm(LnC~LnQ+LnPF+LF,data=airlines,model="within")
summary(model1)
fixef(model1,type="dmean")
fixef(model1,type="level")
qqnorm(model1$residuals)
model1.pred=predict(model1)
plot(as.numeric(model1.pred),as.numeric(model1$residuals),
xlab="Predict",ylab="Residuals")
```

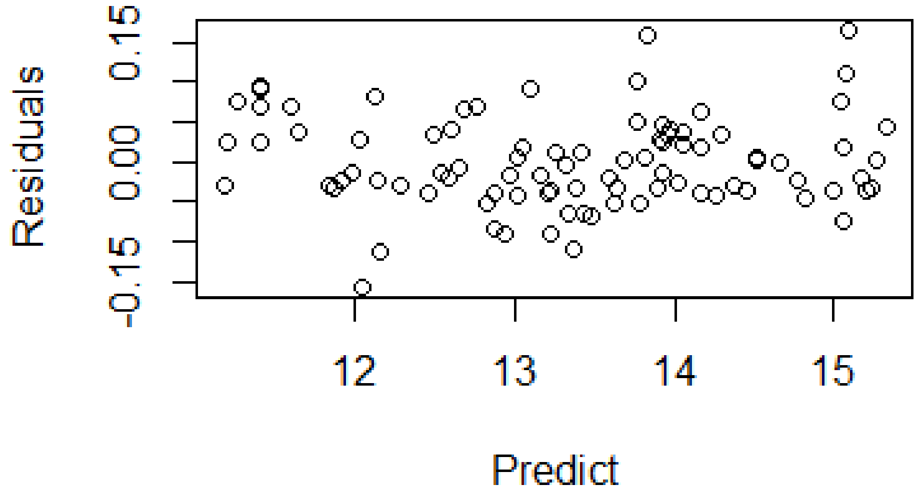
Coefficients:

	Estimate	Std. Error	t-value	Pr(>|t|)
LnQ	0.919285	0.029890	30.7555	< 2e-16
LnPF	0.417492	0.015199	27.4682	< 2e-16
LF	-1.070396	0.201690	-5.3071	9.5e-07

Total Sum of Squares: 39.361
Residual Sum of Squares: 0.29262
R-Squared: 0.99257
Adj. R-Squared: 0.99183

```
> fixef(model1,type="dmean")
        1         2         3        4        5        6
-0.007586 -0.048822 -0.216507 0.176970 0.016469 0.079476
> fixef(model1,type="level")
     1      2      3      4      5      6
9.7059 9.6647 9.4970 9.8905 9.7300 9.7930
```


Normal Q-Q Plot
Sample Quantiles
-0.15
0.00
0.15
-2
-1
0
1
2
Theoretical Quantiles


Residuals
-0.15
0.00
0.15
12
13
14
15
Predict

Time Fixed Effect Model

```
model1=plm(LnC~LnQ+LnPF+LF,data=airlines,model="within",
effect="time")
summary(model1)
fixef(model1,type="dmean")
fixef(model1,type="level")
```

Coefficients:

	Estimate	Std. Error	t-value	Pr(>|t|)
LnQ	0.867727	0.015408	56.3159	< 2.2e-16
LnPF	-0.484485	0.364109	-1.3306	0.1875
LF	-1.954403	0.442378	-4.4179	3.447e-05

R-Squared: 0.98582
Adj. R-Squared: 0.98247

> fixef(model1,type="level")

1	2	3	4	5	6
20.496	20.578	20.656	20.741	21.200	21.412
7	8	9	10	11	12
21.503	21.654	21.830	22.114	22.465	22.651
13	14	15			
22.617	22.552	22.537			

Two-Way Fixed Effects Model

```
model2=plm(LnC~LnQ+LnPF+LF,data=airlines,model="
within",effect="twoways")
summary(model2)
fixef(model2,effect="individual")
fixef(model2,effect="time")
```

Coefficients:

	Estimate	Std. Error	t-value	Pr(>|t|)
LnQ	0.817249	0.031851	25.6586	< 2.2e-16
LnPF	0.168611	0.163478	1.0314	0.306064
LF	-0.882812	0.261737	-3.3729	0.001239

R-Squared: 0.91391

Adj. R-Squared: 0.88564

1	2	3	4	5	6
12.421	12.358	12.103	12.427	12.200	12.247

1	2	3	4	5	6
12.421	12.476	12.519	12.572	12.641	12.687
7	8	9	10	11	12
12.718	12.774	12.842	12.887	13.003	13.081
13	14	15			
13.097	13.096	13.114			

To pool or not to pool?

In each of these models, we allowed different intercepts (or levels) across individuals, time or both. Do we need to have different levels, or can we pool this information into one common level? Since best model (and the one that makes the most sense) was assuming individual levels, we will demonstrate this test:

$$H_0: Pooled\ effect$$

$$H_A: Effects\ are\ significant$$

```
ind.test <- plm(LnC~LnQ+LnPF+LF, data=airlines, model="pooling")
plmtest(ind.test, effect="individual", type="kw")
```

```
Lagrange Multiplier Test - (King and
        Wu) for balanced panels

data:  LnC ~ LnQ + LnPF + LF
normal = 18.299, p-value < 2.2e-16
alternative hypothesis: significant effects
```

RANDOM EFFECTS MODEL

Random Effects Model

$$y_{it} = \alpha_i + \beta_1 x_{1,it} + \beta_2 x_{2,it} + \cdots + \beta_k x_{k,it} + \varepsilon_{it}$$

- In the random effects model we are assuming that $\alpha_i = \mu + v_i$

$$\alpha_i = \mu + v_i$$

Random Effects Model

$$y_{it} = \alpha_i + \beta_1 x_{1,it} + \beta_2 x_{2,it} + \cdots + \beta_k x_{k,it} + \varepsilon_{it}$$

- In the random effects model we are assuming that $\alpha_i = \mu + v_i$

$$\alpha_i = \mu + v_i$$

Common (fixed) effect across all subjects.

Random Effects Model

$$y_{it} = \alpha_i + \beta_1 x_{1,it} + \beta_2 x_{2,it} + \cdots + \beta_k x_{k,it} + \varepsilon_{it}$$

- In the random effects model we are assuming that $\alpha_i = \mu + v_i$

$$\alpha_i = \mu + v_i$$

Random variable with a mean of zero and constant variance that accounts for subject specific disturbances.

Assumptions

- Since the random effects model is slightly different than typical OLS, the assumptions change slightly:
  1. For each subject *i,* the following model holds:

$$y_{it} = \alpha_i + \beta_1 x_{1,it} + \beta_2 x_{2,it} + \cdots + \beta_k x_{k,it} + \varepsilon_{it}$$

Assumptions

- Since the random effects model is slightly different than typical OLS, the assumptions change slightly:
  1. For each subject *i*, the following model holds:

     $$y_{it} = \alpha_i + \beta_1 x_{1,it} + \beta_2 x_{2,it} + \cdots + \beta_k x_{k,it} + \varepsilon_{it}$$

  2. No perfect collinearity between predictor variables.

Assumptions

- Since the random effects model is slightly different than typical OLS, the assumptions change slightly:
  1. For each subject *i*, the following model holds:
  $$y_{it} = \alpha_i + \beta_1 x_{1,it} + \beta_2 x_{2,it} + \cdots + \beta_k x_{k,it} + \varepsilon_{it}$$
  2. No perfect collinearity between predictor variables.
  3. There is no relationship between unobserved differences in “individuals” and the responses **and** $\boldsymbol{\nu}_i$ has a mean of 0 with constant variance $\sigma^2{}_{\nu}$

Assumptions

- Since the random effects model is slightly different than typical OLS, the assumptions change slightly:

1. For each subject *i*, the following model holds:

   $$y_{it} = \alpha_i + \beta_1 x_{1,it} + \beta_2 x_{2,it} + \cdots + \beta_k x_{k,it} + \varepsilon_{it}$$

2. No perfect collinearity between predictor variables.
3. There is no relationship between unobserved differences in "individuals" and the responses **and** ν_i has a mean of 0 with constant variance $\sigma^2{}_\nu$
4. $\varepsilon_{it} \sim N(0, \sigma^2)$

One-Way Random Effects Model

```
model3=plm(LnC~LnQ+LnPF+LF,data=airlines,model="random")
summary(model3)
```

Effects:

	var	std.dev	share
idiosyncratic	0.003613	0.060105	0.188
individual	0.015597	0.124889	0.812

Coefficients:

	Estimate	Std. Error	z-value	Pr(>|z|)
(Intercept)	9.627909	0.210164	45.8114	< 2.2e-16
LnQ	0.906681	0.025625	35.3827	< 2.2e-16
LnPF	0.422778	0.014025	30.1451	< 2.2e-16
LF	-1.064498	0.200070	-5.3206	1.034e-07

Two-Way Random Effects Model

```
model4=plm(LnC~LnQ+LnPF+LF,data=airlines,model="random",
effect="twoways")
summary(model4)
```

Effects:

	var	std.dev	share
idiosyncratic	2.640e-03	5.138e-02	0.144
individual	1.566e-02	1.251e-01	0.853
time	6.831e-05	8.265e-03	0.004

Testing for Random Effects

- There is a test to examine if we should be fitting a random effects or fixed effects model:

Hausman test:

H_0: Random effects good

H_a: Fixed effects might be better

```
f.model <-
plm(LnC~LnQ+LnPF+LF,data=airlines,model="within")
r.model <-
plm(LnC~LnQ+LnPF+LF,data=airlines,model="random")
phtest(f.model, r.model)
```

Hausman Test

data: LnC ~ LnQ + LnPF + LF
chisq = 2.1247, df = 3, p-value = 0.5469
alternative hypothesis: one model is inconsistent

Would choose one-way random effects for model.

References

https://cran.r-project.org/web/packages/plm/vignettes/A_plmPackage.html

https://cran.r-project.org/web/packages/plm/plm.pdf