Panel Data

AECN 396/896-002

Before we start

Learning objectives

Understand the new econometric methods that can be used with panel datasets to address endogeneity problems.

Panel (longitudinal) Data

Definition

Data follow the same individuals, families, firms, cities, states or whatever, across time

Panel (longitudinal) Data

Definition

Data follow the same individuals, families, firms, cities, states or whatever, across time

Example

- Randomly select people from a population at a given point in time
- Then the same people are reinterviewed at several subsequent points in time, which would result in data on wages, hours, education, and so on, for the same group of people in different years.

Panel Data in data.frame

```
##   year  fcode employ    sales
## 1 1987 410032    100 47000000
## 2 1988 410032    131 43000000
## 3 1989 410032    123 49000000
## 4 1987 410440     12  1560000
## 5 1988 410440     13  1970000
## 6 1989 410440     14  2350000
## 7 1987 410495     20   750000
## 8 1988 410495     25   110000
## 9 1989 410495     24   950000
```

- `year`: year
- `fcode`: factory id
- `employ`: the number of employees
- `sales`: sales in USD

Central Question

Can we do anything to deal with endogeneity problem taking advantage of the panel data structure?

Panel Data Estimation Methods

Demand for massage (cross-sectional)

Location	Year	P	Q
Chicago	2003	75	2.0
Peoria	2003	50	1.0
Milwaukee	2003	60	1.5
Madison	2003	55	0.8

- `P`: the price of one massage
- `Q`: the number of massages received per capita

Demand for massage (cross-sectional)

Location	Year	P	Q
Chicago	2003	75	2.0
Peoria	2003	50	1.0
Milwaukee	2003	60	1.5
Madison	2003	55	0.8

Demand for massage (cross-sectional)

Location	Year	P	Q
Chicago	2003	75	2.0
Peoria	2003	50	1.0
Milwaukee	2003	60	1.5
Madison	2003	55	0.8

Question

Across the four cities, how are price and quantity are associated? Positive or negative?

Demand for massage (cross-sectional)

Location	Year	P	Q
Chicago	2003	75	2.0
Peoria	2003	50	1.0
Milwaukee	2003	60	1.5
Madison	2003	55	0.8

Question

Across the four cities, how are price and quantity are associated? Positive or negative?

Answer

They are positively correlated. So, does that mean people want more massages as their price increases? Probably not.

Demand for massage (cross-sectional)

Location	Year	P	Q
Chicago	2003	75	2.0
Peoria	2003	50	1.0
Milwaukee	2003	60	1.5
Madison	2003	55	0.8

Question

What could be causing the positive correlation?

Demand for massage (cross-sectional)

Location	Year	P	Q
Chicago	2003	75	2.0
Peoria	2003	50	1.0
Milwaukee	2003	60	1.5
Madison	2003	55	0.8

Question

What could be causing the positive correlation?

Answer

- Income (can be observed)
- Quality of massages (hard to observe)
- How physically taxing jobs are (?)

Demand for massage (cross-sectional)

Location	Year	P	Q	Ql
Chicago	2003	75	2.0	10
Peoria	2003	50	1.0	5
Milwaukee	2003	60	1.5	7
Madison	2003	55	0.8	6

Key

Massage quality was hidden (omitted) affecting both price and massages per capita.

Problem

Massage quality is not observable, and thus cannot be controlled for.

Mathematically

$Q = \beta_0 + \beta_1 P + v \;\; (= \beta_2 + Ql + u)$

- P: the price of one massage
- Q: the number of massages received per capita
- Ql: the quality of massages
- u: everything else that affect P

Endogeneity Problem

P is correlated with Ql.

Demand for massage (two-period panel)

Location	Year	P	Q	QI
Chicago	2003	75	2.0	10
Chicago	2004	85	1.8	10
Peoria	2003	50	1.0	5
Peoria	2004	48	1.1	5
Milwaukee	2003	60	1.5	7
Milwaukee	2004	65	1.4	7
Madison	2003	55	0.8	6
Madison	2004	60	0.7	6

Key

There are two kinds of variations:

- inte-rcity (across city) variation
- intra-city (within city) variation

The cross-sectional data offers only the inte-rcity (across city) variations.

Demand for massage (two-period panel)

Location	Year	P	Q	QI
Chicago	2003	75	2.0	10
Chicago	2004	85	1.8	10
Peoria	2003	50	1.0	5
Peoria	2004	48	1.1	5
Milwaukee	2003	60	1.5	7
Milwaukee	2004	65	1.4	7
Madison	2003	55	0.8	6
Madison	2004	60	0.7	6

Now, compare the massage price and massages per capita within each city (over time). What do you see?

Demand for massage (two-period panel)

Location	Year	P	Q	QI
Chicago	2003	75	2.0	10
Chicago	2004	85	1.8	10
Peoria	2003	50	1.0	5
Peoria	2004	48	1.1	5
Milwaukee	2003	60	1.5	7
Milwaukee	2004	65	1.4	7
Madison	2003	55	0.8	6
Madison	2004	60	0.7	6

Now, compare the massage price and massages per capita within each city (over time). What do you see?

Answer

Price and quantity are negatively correlated!

Demand for massage (two-period panel)

Location	Year	P	Q	Ql
Chicago	2003	75	2.0	10
Chicago	2004	85	1.8	10
Peoria	2003	50	1.0	5
Peoria	2004	48	1.1	5
Milwaukee	2003	60	1.5	7
Milwaukee	2004	65	1.4	7
Madison	2003	55	0.8	6
Madison	2004	60	0.7	6

Question

Why looking at the intra-city (within city) variation seemed to help us estimate the impact of massage price on demand more credibly?

Demand for massage (two-period panel)

Location	Year	P	Q	Ql
Chicago	2003	75	2.0	10
Chicago	2004	85	1.8	10
Peoria	2003	50	1.0	5
Peoria	2004	48	1.1	5
Milwaukee	2003	60	1.5	7
Milwaukee	2004	65	1.4	7
Madison	2003	55	0.8	6
Madison	2004	60	0.7	6

Question

Why looking at the intra-city (within city) variation seemed to help us estimate the impact of massage price on demand more credibly?

Answer

The omitted variable, massage quality, did not change over time within city, which means it is controlled for as long as you look only at the intra-city variations (you do not compare across cities).

Demand for massage (two-period panel)

Location	Year	P	Q	QI
Chicago	2003	75	2.0	10
Chicago	2004	85	1.8	10
Peoria	2003	50	1.0	5
Peoria	2004	48	1.1	5
Milwaukee	2003	60	1.5	7
Milwaukee	2004	65	1.4	7
Madison	2003	55	0.8	6
Madison	2004	60	0.7	6

Question

But, what if massage quality changed from 2003 to 2004?

Answer

Looking at the intra-city variations is problematic just like looking at the inter-city variations.

Question

So, how do we use only the intra-city variations in a regression framework?

Question

So, how do we use only the intra-city variations in a regression framework?

One way

One way to do this is to compute the changes in prices and th changes in quantities in each city $(\Delta P$ and $\Delta Q)$ and then regress ΔQ and ΔP.

Question

So, how do we use only the intra-city variations in a regression framework?

One way

One way to do this is to compute the changes in prices and th changes in quantities in each city (ΔP and ΔQ) and then regress ΔQ and ΔP.

First-differenced Data

Location	Year	P	Q	Ql	P_dif	Q_dif	Ql_dif
Chicago	2003	75	2.0	10	NA	NA	NA
Chicago	2004	85	1.8	10	10	-0.2	0
Peoria	2003	50	1.0	5	NA	NA	NA
Peoria	2004	48	1.1	5	-2	0.1	0
Milwaukee	2003	60	1.5	7	NA	NA	NA
Milwaukee	2004	65	1.4	7	5	-0.1	0
Madison	2003	55	0.8	6	NA	NA	NA
Madison	2004	60	0.7	6	5	-0.1	0

Question

So, how do we use only the intra-city variations in a regression framework?

One way

One way to do this is to compute the changes in prices and th changes in quantities in each city $(\Delta P$ and $\Delta Q)$ and then regress ΔQ and ΔP.

First-differenced Data

Location	Year	P	Q	Ql	P_dif	Q_dif	Ql_dif
Chicago	2003	75	2.0	10	NA	NA	NA
Chicago	2004	85	1.8	10	10	-0.2	0
Peoria	2003	50	1.0	5	NA	NA	NA
Peoria	2004	48	1.1	5	-2	0.1	0
Milwaukee	2003	60	1.5	7	NA	NA	NA
Milwaukee	2004	65	1.4	7	5	-0.1	0
Madison	2003	55	0.8	6	NA	NA	NA
Madison	2004	60	0.7	6	5	-0.1	0

Key

Variations in quality is eliminated after first differentiation!! (quality is controlled for)

A new way of writing a model

$$Q_{i,t} = \beta_0 + \beta_1 P_{i,t} + v_{i,t} \ \ (= \beta_2 Ql_{i,t} + u_{i,t})$$

- `i`: indicates city
- `t`: indicates time

A new way of writing a model

$$Q_{i,t} = \beta_0 + \beta_1 P_{i,t} + v_{i,t} \;\; (= \beta_2 Ql_{i,t} + u_{i,t})$$

- `i`: indicates city
- `t`: indicates time

First differencing

$$Q_{i,1} = \beta_0 + \beta_1 P_{i,1} + v_{i,1} \;\; (= \beta_2 Ql_{i,1} + u_{i,1})$$

$$Q_{i,2} = \beta_0 + \beta_1 P_{i,2} + v_{i,2} \;\; (= \beta_2 Ql_{i,2} + u_{i,2})$$

A new way of writing a model

$$Q_{i,t} = \beta_0 + \beta_1 P_{i,t} + v_{i,t} \quad (= \beta_2 Ql_{i,t} + u_{i,t})$$

- `i`: indicates city
- `t`: indicates time

First differencing

$$Q_{i,1} = \beta_0 + \beta_1 P_{i,1} + v_{i,1} \quad (= \beta_2 Ql_{i,1} + u_{i,1})$$

$$Q_{i,2} = \beta_0 + \beta_1 P_{i,2} + v_{i,2} \quad (= \beta_2 Ql_{i,2} + u_{i,2})$$

$$\Rightarrow$$

$$\Delta Q = \beta_1 \Delta P + \Delta v (= \beta_2 \Delta Ql + \Delta u)$$

Endogeneity Problem?

Since $Ql_{i,1} = Ql_{i,2}$, $\Delta Ql = 0$.

A new way of writing a model

$$Q_{i,t} = \beta_0 + \beta_1 P_{i,t} + v_{i,t} \;\; (= \beta_2 Ql_{i,t} + u_{i,t})$$

- `i`: indicates city
- `t`: indicates time

First differencing

$$Q_{i,1} = \beta_0 + \beta_1 P_{i,1} + v_{i,1} \;\; (= \beta_2 Ql_{i,1} + u_{i,1})$$

$$Q_{i,2} = \beta_0 + \beta_1 P_{i,2} + v_{i,2} \;\; (= \beta_2 Ql_{i,2} + u_{i,2})$$

$$\Rightarrow$$

$$\Delta Q = \beta_1 \Delta P + \Delta v (= \beta_2 \Delta Ql + \Delta u)$$

Endogeneity Problem?

Since $Ql_{i,1} = Ql_{i,2}$, $\Delta Ql = 0$.

$$\Rightarrow$$

$$\Delta Q = \beta_0 + \beta_1 \Delta P + \Delta u$$

No endogeneity problem after first differentiation!

Data

```
massage_data_fd %>% head(5)
```

```
## # A tibble: 5 × 8
##   Location   Year     P     Q    Ql P_dif  Q_dif Ql_dif
##   <chr>     <dbl> <dbl> <dbl> <dbl> <dbl>  <dbl>  <dbl>
## 1 Chicago    2003    75   2      10    NA NA         NA
## 2 Chicago    2004    85   1.8    10    10 -0.2        0
## 3 Peoria     2003    50   1       5    NA NA         NA
## 4 Peoria     2004    48   1.1     5    -2  0.100      0
## 5 Milwaukee  2003    60   1.5     7    NA NA         NA
```

```
## # A tibble: 5 × 8
##   Location   Year     P     Q    Ql P_dif  Q_dif Ql_dif
##   <chr>     <dbl> <dbl> <dbl> <dbl> <dbl>  <dbl>  <dbl>
## 1 Chicago    2003    75   2      10    NA NA         NA
## 2 Chicago    2004    85   1.8    10    10 -0.2        0
## 3 Peoria     2003    50   1       5    NA NA         NA
## 4 Peoria     2004    48   1.1     5    -2  0.100      0
## 5 Milwaukee  2003    60   1.5     7    NA NA         NA
```

OLS on the original data:

```
feols(Q ~ P, data = massage_data_fd)
```

```
## OLS estimation, Dep. Var.: Q
## Observations: 8
## Standard-errors: IID
##              Estimate Std. Error   t val
## (Intercept) -0.496511   0.614204 -0.8083
## P            0.028659   0.009696  2.9558
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01
## RMSE: 0.27893   Adj. R2: 0.525004
```

OLS on the first-differenced data:

```
feols(Q_dif ~ P_dif, data = massage_data_f
```

```
## OLS estimation, Dep. Var.: Q_dif
## Observations: 4
## Standard-errors: IID
##              Estimate Std. Error   t va
## (Intercept)  0.039041   0.012750   3.06
## P_dif       -0.025342   0.002055 -12.33
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01
## RMSE: 0.012414   Adj. R2: 0.980534
```

Summary

- As long as the omitted variable that affects both the dependent and independent variables are constant over time (time-invariant), then using only the variations over time (ignoring variations across cross-sectional units) can eliminate the omitted variable bias
- First-differencing the data and then regressing changes on changes does the trick of ignoring variations across cross-sectional units
- Of course, first-differencing is possible only because the same cross-sectional units are observed multiple times over time.

Multi-year panel datasets

- If we have lots of years of data, we could, in principle, compute all of the first differences (i.e., 2004 versus 2003, 2005 versus 2004, etc.) and then run a single regression. But there is an easier way.

Multi-year panel datasets

- If we have lots of years of data, we could, in principle, compute all of the first differences (i.e., 2004 versus 2003, 2005 versus 2004, etc.) and then run a single regression. But there is an easier way.
- Instead of thinking of each year's observation in terms of how much it differs from the prior year for the same city, let's think about how much each observation differs from the average for that city.

How much each observation differs from the average for that city?

Location	Year	P	P_mean	P_dev	Q	Q_mean	Q_dev	QI	QI_mean	QI_dev
Chicago	2003	75	80.0	-5.0	2.0	1.90	0.10	10	10	0
Chicago	2004	85	80.0	5.0	1.8	1.90	-0.10	10	10	0
Peoria	2003	50	49.0	1.0	1.0	1.05	-0.05	5	5	0
Peoria	2004	48	49.0	-1.0	1.1	1.05	0.05	5	5	0
Milwaukee	2003	60	62.5	-2.5	1.5	1.45	0.05	7	7	0
Milwaukee	2004	65	62.5	2.5	1.4	1.45	-0.05	7	7	0
Madison	2003	55	57.5	-2.5	0.8	0.75	0.05	6	6	0
Madison	2004	60	57.5	2.5	0.7	0.75	-0.05	6	6	0

How much each observation differs from the average for that city?

Location	Year	P	P_mean	P_dev	Q	Q_mean	Q_dev	Ql	Ql_mean	Ql_dev
Chicago	2003	75	80.0	-5.0	2.0	1.90	0.10	10	10	0
Chicago	2004	85	80.0	5.0	1.8	1.90	-0.10	10	10	0
Peoria	2003	50	49.0	1.0	1.0	1.05	-0.05	5	5	0
Peoria	2004	48	49.0	-1.0	1.1	1.05	0.05	5	5	0
Milwaukee	2003	60	62.5	-2.5	1.5	1.45	0.05	7	7	0
Milwaukee	2004	65	62.5	2.5	1.4	1.45	-0.05	7	7	0
Madison	2003	55	57.5	-2.5	0.8	0.75	0.05	6	6	0
Madison	2004	60	57.5	2.5	0.7	0.75	-0.05	6	6	0

Note

We call this data transformation within-transformation or demeaning.

How much each observation differs from the average for that city?

Location	Year	P	P_mean	P_dev	Q	Q_mean	Q_dev	Ql	Ql_mean	Ql_dev
Chicago	2003	75	80.0	-5.0	2.0	1.90	0.10	10	10	0
Chicago	2004	85	80.0	5.0	1.8	1.90	-0.10	10	10	0
Peoria	2003	50	49.0	1.0	1.0	1.05	-0.05	5	5	0
Peoria	2004	48	49.0	-1.0	1.1	1.05	0.05	5	5	0
Milwaukee	2003	60	62.5	-2.5	1.5	1.45	0.05	7	7	0
Milwaukee	2004	65	62.5	2.5	1.4	1.45	-0.05	7	7	0
Madison	2003	55	57.5	-2.5	0.8	0.75	0.05	6	6	0
Madison	2004	60	57.5	2.5	0.7	0.75	-0.05	6	6	0

Note

We call this data transformation within-transformation or demeaning.

Fixed Effects Regression

- Dependent variable: `Q_dev`
- Independent variable: `P_dev`

How much each observation differs from the average for that city?

Location	Year	P	P_mean	P_dev	Q	Q_mean	Q_dev	Ql	Ql_mean	Ql_dev
Chicago	2003	75	80.0	-5.0	2.0	1.90	0.10	10	10	0
Chicago	2004	85	80.0	5.0	1.8	1.90	-0.10	10	10	0
Peoria	2003	50	49.0	1.0	1.0	1.05	-0.05	5	5	0
Peoria	2004	48	49.0	-1.0	1.1	1.05	0.05	5	5	0
Milwaukee	2003	60	62.5	-2.5	1.5	1.45	0.05	7	7	0
Milwaukee	2004	65	62.5	2.5	1.4	1.45	-0.05	7	7	0
Madison	2003	55	57.5	-2.5	0.8	0.75	0.05	6	6	0
Madison	2004	60	57.5	2.5	0.7	0.75	-0.05	6	6	0

Note

We call this data transformation within-transformation or demeaning.

Fixed Effects Regression

- Dependent variable: `Q_dev`
- Independent variable: `P_dev`

Key

In calculating `P_dev` (deviation from the mean by city), `Ql_dev` is eliminated.

Within-transformation

$$Q_{i,1} = \beta_0 + \beta_1 P_{i,1} + v_{i,1} \quad (= \beta_2 Ql_{i,1} + u_{i,1})$$

$$Q_{i,2} = \beta_0 + \beta_1 P_{i,2} + v_{i,2} \quad (= \beta_2 Ql_{i,2} + u_{i,2})$$

$$\vdots$$

$$Q_{i,T} = \beta_0 + \beta_1 P_{i,T} + v_{i,T} \quad (= \beta_2 Ql_{i,T} + u_{i,T})$$

Within-transformation

$$Q_{i,1} = \beta_0 + \beta_1 P_{i,1} + v_{i,1} \;\; (= \beta_2 Ql_{i,1} + u_{i,1})$$

$$Q_{i,2} = \beta_0 + \beta_1 P_{i,2} + v_{i,2} \;\; (= \beta_2 Ql_{i,2} + u_{i,2})$$

$$\vdots$$

$$Q_{i,T} = \beta_0 + \beta_1 P_{i,T} + v_{i,T} \;\; (= \beta_2 Ql_{i,T} + u_{i,T})$$

$$\Rightarrow$$

$$Q_{i,t} - \bar{Q}_i = \beta_1 [P_{i,t} - \bar{P}_i] + [v_{i,t} - \bar{v}_i] (= \beta_2 [Ql_{i,t} - \bar{Ql}_i] + [u_{i,t} - \bar{u}_i])$$

Endogeneity Problem?

$$Ql_{i,1} = Ql_{i,2} = \cdots = Ql_{i,T} = \bar{Ql}_i$$

Within-transformation

$$Q_{i,1} = \beta_0 + \beta_1 P_{i,1} + v_{i,1} \ \ (= \beta_2 Ql_{i,1} + u_{i,1})$$

$$Q_{i,2} = \beta_0 + \beta_1 P_{i,2} + v_{i,2} \ \ (= \beta_2 Ql_{i,2} + u_{i,2})$$

$$\vdots$$

$$Q_{i,T} = \beta_0 + \beta_1 P_{i,T} + v_{i,T} \ \ (= \beta_2 Ql_{i,T} + u_{i,T})$$

$$\Rightarrow$$

$$Q_{i,t} - \bar{Q}_i = \beta_1 [P_{i,t} - \bar{P}_i] + [v_{i,t} - \bar{v}_i] (= \beta_2 [Ql_{i,t} - \bar{Ql}_i] + [u_{i,t} - \bar{u}_i])$$

Endogeneity Problem?

$$Ql_{i,1} = Ql_{i,2} = \cdots = Ql_{i,T} = \bar{Ql}_i$$

$$\Rightarrow$$

$$Q_{i,t} - \bar{Q}_i = \beta_1 [P_{i,t} - \bar{P}_i] + [u_{i,t} - \bar{u}_i]$$

No endogeneity problem after the within-transformation!

Fixed Effects (FE) Estimation (in general)

Consider the following general model

$$y_{i,t} = \beta_1 x_{i,t} + \alpha_i + u_{i,t}$$

- α_i: the impact of time-invariant unobserved factor that is specific to i (also termed individual fixed effect)
- α_i is thought to be correlated with $x_{i,t}$

Fixed Effects (FE) Estimation (in general)

Consider the following general model

$$y_{i,t} = \beta_1 x_{i,t} + \alpha_i + u_{i,t}$$

- α_i: the impact of time-invariant unobserved factor that is specific to i (also termed individual fixed effect)
- α_i is thought to be correlated with $x_{i,t}$

For each i, average this equation over time, we get

$$\frac{\sum_{t=1}^{T} y_{i,t}}{T} = \frac{\sum_{t=1}^{T} x_{i,t}}{T} + \alpha_i + \frac{\sum_{t=1}^{T} u_{i,t}}{T}$$

Note, $\frac{\sum_{t=1}^{T} \alpha_i}{T} = \alpha_i$

Fixed Effects (FE) Estimation (in general)

Consider the following general model

$$y_{i,t} = \beta_1 x_{i,t} + \alpha_i + u_{i,t}$$

- α_i: the impact of time-invariant unobserved factor that is specific to i (also termed individual fixed effect)
- α_i is thought to be correlated with $x_{i,t}$

For each i, average this equation over time, we get

$$\frac{\sum_{t=1}^{T} y_{i,t}}{T} = \frac{\sum_{t=1}^{T} x_{i,t}}{T} + \alpha_i + \frac{\sum_{t=1}^{T} u_{i,t}}{T}$$

Note, $\frac{\sum_{t=1}^{T} \alpha_i}{T} = \alpha_i$

Subtracting the second equation from the first one,

$$(y_{i,t} - \frac{\sum_{t=1}^{T} y_{i,t}}{T}) = \beta_1 (x_{i,t} - \frac{\sum_{t=1}^{T} x_{i,t}}{T}) + (u_{i,t} - \frac{\sum_{t=1}^{T} u_{i,t}}{T})$$

Fixed Effects (FE) Estimation (in general)

Consider the following general model

$$y_{i,t} = \beta_1 x_{i,t} + \alpha_i + u_{i,t}$$

- α_i: the impact of time-invariant unobserved factor that is specific to i (also termed individual fixed effect)
- α_i is thought to be correlated with $x_{i,t}$

For each i, average this equation over time, we get

$$\frac{\sum_{t=1}^{T} y_{i,t}}{T} = \frac{\sum_{t=1}^{T} x_{i,t}}{T} + \alpha_i + \frac{\sum_{t=1}^{T} u_{i,t}}{T}$$

Note, $\frac{\sum_{t=1}^{T} \alpha_i}{T} = \alpha_i$

Subtracting the second equation from the first one,

$$(y_{i,t} - \frac{\sum_{t=1}^{T} y_{i,t}}{T}) = \beta_1(x_{i,t} - \frac{\sum_{t=1}^{T} x_{i,t}}{T}) + (u_{i,t} - \frac{\sum_{t=1}^{T} u_{i,t}}{T})$$

Important

α_i is gone!

Fixed Effects (FE) Estimation (in general)

Consider the following general model

$$y_{i,t} = \beta_1 x_{i,t} + \alpha_i + u_{i,t}$$

- α_i: the impact of time-invariant unobserved factor that is specific to i (also termed individual fixed effect)
- α_i is thought to be correlated with $x_{i,t}$

For each i, average this equation over time, we get

$$\frac{\sum_{t=1}^{T} y_{i,t}}{T} = \frac{\sum_{t=1}^{T} x_{i,t}}{T} + \alpha_i + \frac{\sum_{t=1}^{T} u_{i,t}}{T}$$

Note, $\frac{\sum_{t=1}^{T} \alpha_i}{T} = \alpha_i$

Subtracting the second equation from the first one,

$$(y_{i,t} - \frac{\sum_{t=1}^{T} y_{i,t}}{T}) = \beta_1 (x_{i,t} - \frac{\sum_{t=1}^{T} x_{i,t}}{T}) + (u_{i,t} - \frac{\sum_{t=1}^{T} u_{i,t}}{T})$$

Important

α_i is gone!

We then regress $(y_{i,t} - \frac{\sum_{t=1}^{T} y_{i,t}}{T})$ on $(x_{i,t} - \frac{\sum_{t=1}^{T} x_{i,t}}{T})$ to estimate β_1.

When is FE estimation unbiased?

Here is the data after within-transformation:

$$(y_{i,t} - \frac{\sum_{t=1}^{T} y_{i,t}}{T}) = \beta_1(x_{i,t} - \frac{\sum_{t=1}^{T} x_{i,t}}{T}) + (u_{i,t} - \frac{\sum_{t=1}^{T} u_{i,t}}{T})$$

When is FE estimation unbiased?

Here is the data after within-transformation:

$$(y_{i,t} - \frac{\sum_{t=1}^{T} y_{i,t}}{T}) = \beta_1(x_{i,t} - \frac{\sum_{t=1}^{T} x_{i,t}}{T}) + (u_{i,t} - \frac{\sum_{t=1}^{T} u_{i,t}}{T})$$

So,

$(x_{i,t} - \frac{\sum_{t=1}^{T} x_{i,t}}{T})$ needs to be uncorrelated with $(u_{i,t} - \frac{\sum_{t=1}^{T} u_{i,t}}{T})$.

When is FE estimation unbiased?

Here is the data after within-transformation:

$$(y_{i,t} - \frac{\sum_{t=1}^{T} y_{i,t}}{T}) = \beta_1(x_{i,t} - \frac{\sum_{t=1}^{T} x_{i,t}}{T}) + (u_{i,t} - \frac{\sum_{t=1}^{T} u_{i,t}}{T})$$

So,

$(x_{i,t} - \frac{\sum_{t=1}^{T} x_{i,t}}{T})$ needs to be uncorrelated with $(u_{i,t} - \frac{\sum_{t=1}^{T} u_{i,t}}{T})$.

The above condition is satisfied if

$$E[u_{i,s}|x_{i,t}] = 0 \ \ ^{\forall} s, \ \ t, \ \ \text{and} \ \ j$$

When is FE estimation unbiased?

Here is the data after within-transformation:

$$(y_{i,t} - \frac{\sum_{t=1}^{T} y_{i,t}}{T}) = \beta_1 (x_{i,t} - \frac{\sum_{t=1}^{T} x_{i,t}}{T}) + (u_{i,t} - \frac{\sum_{t=1}^{T} u_{i,t}}{T})$$

So,

$(x_{i,t} - \frac{\sum_{t=1}^{T} x_{i,t}}{T})$ needs to be uncorrelated with $(u_{i,t} - \frac{\sum_{t=1}^{T} u_{i,t}}{T})$.

The above condition is satisfied if

$$E[u_{i,s}|x_{i,t}] = 0 \;\; {}^{\forall} s, \;\; t, \;\; \text{and} \;\; j$$

e.g., $E[u_{i,1}|x_{i,4}] = 0$

Fixed effects estimation

Regress within-transformed `Q` on within-transformed `P`:

```
feols(Q_dev ~ P_dev, data = massage_data_wth)
```

```
## OLS estimation, Dep. Var.: Q_dev
## Observations: 8
## Standard-errors: IID
##                  Estimate Std. Error       t value   Pr(>|t|)
## (Intercept)  2.780000e-17   0.006044  4.590000e-15 1.0000e+00
## P_dev       -2.077922e-02   0.001948 -1.066667e+01 4.0041e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 0.014804   Adj. R2: 0.941558
```

An alternative way to view the Fixed Effects estimation methods

Important

The two approached below will result in the same coefficient estimates (mathematically identical).

- Running OLS on the within-tranformed (demeaned) data
- Running OLS on the untransformed data but including the dummy variables for the individuals

You can use the original data (no within-transformation) and include dummy variables for all the cities except one.

```
(
  massage_data_d <- massage_data_2p %>%
    mutate(
      Peoria_D = ifelse(Location == "Peoria", 1, 0),
      Milwaukee_D = ifelse(Location == "Milwaukee", 1, 0),
      Madison_D = ifelse(Location == "Madison", 1, 0)
    )
)
```

```
##    Location Year  P   Q Ql Peoria_D Milwaukee_D Madison_D
## 1   Chicago 2003 75 2.0 10        0           0         0
## 2   Chicago 2004 85 1.8 10        0           0         0
## 3    Peoria 2003 50 1.0  5        1           0         0
## 4    Peoria 2004 48 1.1  5        1           0         0
## 5 Milwaukee 2003 60 1.5  7        0           1         0
## 6 Milwaukee 2004 65 1.4  7        0           1         0
## 7   Madison 2003 55 0.8  6        0           0         1
## 8   Madison 2004 60 0.7  6        0           0         1
```

Fixed effects estimation (alternative way)

```
feols(Q ~ P + Peoria_D + Milwaukee_D + Madison_D, data = massage_data_d)
```

```
## OLS estimation, Dep. Var.: Q
## Observations: 8
## Standard-errors: IID
##              Estimate Std. Error   t value   Pr(>|t|)
## (Intercept)  3.562338   0.221059  16.11488 0.00051976 ***
## P           -0.020779   0.002755  -7.54247 0.00483179 **
## Peoria_D    -1.494156   0.088759 -16.83378 0.00045649 ***
## Milwaukee_D -0.813636   0.053933 -15.08599 0.00063230 ***
## Madison_D   -1.617532   0.066534 -24.31141 0.00015255 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 0.014804   Adj. R2: 0.997324
```

Note that the coefficient estimate on `P` is exactly the same as the one we saw earlier when we regressed `Q_dev` on `P_dev`.

What does this tell us?

By including individual dummies (individual fixed effects), you are effectively eliminating the between (inter-city) variations and using only the clean within (within-city) variations for estimation.

What does this tell us?

By including individual dummies (individual fixed effects), you are effectively eliminating the between (inter-city) variations and using only the clean within (within-city) variations for estimation.

Very Important

More generally, including dummy variables of a categorical variable (like city in the example above), eliminates the variations between the elements of the category (e.g., different cities), and use only the variations within each of the element of the category.

Let's go back to the avovado example

This was how the avocado makert worked:

At the beginning of each month, avocado suppliers make a plan for what avocado prices will be each week in that month, and never change their plans until the next month.

We agreed that we should use only the within-month varaitions instead of between-month variations.

We have three months of avocado purchase and price observed weekly.

Let's go back to the avovado example

This was how the avocado makert worked:

At the beginning of each month, avocado suppliers make a plan for what avocado prices will be each week in that month, and never change their plans until the next month.

We agreed that we should use only the within-month varaitions instead of between-month variations.

We have three months of avocado purchase and price observed weekly.

Question

What should we do?

Let's go back to the avovado example

This was how the avocado makert worked:

At the beginning of each month, avocado suppliers make a plan for what avocado prices will be each week in that month, and never change their plans until the next month.

We agreed that we should use only the within-month varaitions instead of between-month variations.

We have three months of avocado purchase and price observed weekly.

Question

What should we do?

Answer

Include month dummy variables

We have two years of avocado purchase and price observed weekly.

We have two years of avocado purchase and price observed weekly.

Question

What should we do?

We have two years of avocado purchase and price observed weekly.

Question

What should we do?

Answer

- Include month-year dummy variables.
- Including month dummy variables will not do it. Because the observations in the same month in two different years are considered to be belong to the same group. That is, variations between two different years of the same month will be used for estimation. (e.g., January in 2014 and January in 2015)

Fixed Effects Estimation in Practice Using R

Advice

- Do not within-transform the data yourself and run a regression
- Do not create dummy variables yourself and run a regression with the dummies

Fixed Effects Estimation in Practice Using R

Advice

- Do not within-transform the data yourself and run a regression
- Do not create dummy variables yourself and run a regression with the dummies

In practice

We will use the `fixest` package.

Syntax

```
feols(dep_var ~ indep_vars | FE, data)
```

- `FE`: the name of the variable that identifies the cross-sectional units that are observed over time (`Location` in our example)
- `dep_var`: (non-transformed) dependet variable
- `indep_vars`: list of (non-transformed) independent variables

Data

```
massage_data_2p
```

```
##    Location Year  P   Q Ql
## 1   Chicago 2003 75 2.0 10
## 2   Chicago 2004 85 1.8 10
## 3    Peoria 2003 50 1.0  5
## 4    Peoria 2004 48 1.1  5
## 5 Milwaukee 2003 60 1.5  7
## 6 Milwaukee 2004 65 1.4  7
## 7   Madison 2003 55 0.8  6
## 8   Madison 2004 60 0.7  6
```

Example

```
feols(Q ~ P | Location, data = massage_data_2p)
```

```
## OLS estimation, Dep. Var.: Q
## Observations: 8
## Fixed-effects: Location: 4
## Standard-errors: Clustered (Location)
##    Estimate Std. Error t value   Pr(>|t|)
## P -0.020779   0.001159 -17.923 0.00037879 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 0.014804     Adj. R2: 0.997324
##                  Within R2: 0.949907
```

Random Effects (RE) Model

- Can be more efficient than FE
- If α_i and independent variables are correlated, then RE estimators are biased
- Unless α_i and independent variables are not correlated (which does not hold most of the time unless you got data from controlled experiments), RE is not an attractive option
- You almost never see this estimation method used in papers that use non-experimental data

Random Effects (RE) Model

- Can be more efficient than FE
- If α_i and independent variables are correlated, then RE estimators are biased
- Unless α_i and independent variables are not correlated (which does not hold most of the time unless you got data from controlled experiments), RE is not an attractive option
- You almost never see this estimation method used in papers that use non-experimental data

Note

We do not cover this estimation method as you almost certainly would not use this estimation method.

Year Fixed Effects

Year Fixed Effects

Definition

Just a collection of year dummies, which takes 1 if in a specific year, 0 otherwise.

```
##    id year income educ FE_2015 FE_2016 FE_2017
## 1   1 2015     77   12       1       0       0
## 2   1 2016     82   13       0       1       0
## 3   1 2017     84   14       0       0       1
## 4   1 2015    110   18       1       0       0
## 5   2 2016    120   19       0       1       0
## 6   2 2017    131   20       0       0       1
## 7   2 2015     56   10       1       0       0
## 8   2 2016     60   11       0       1       0
## 9   3 2017     61   12       0       0       1
## 10  3 2015     70   13       1       0       0
## 11  3 2016     71   14       0       1       0
## 12  3 2017     74   15       0       0       1
```

What do year FEs do?

They capture anything that happened to all the individuals for a specific year relative to the base year

Example

Education and wage data from 2012 to 2014,

$$log(income) = \beta_0 + \beta_1 educ + \beta_2 exper + \sigma_1 FE_{2012} + \sigma_2 FE_{2013}$$

- σ_1: captures the difference in $log(income)$ between 2012 and 2014 (base year)
- σ_2: captures the difference in $log(income)$ between 2013 and 2014 (base year)

What do year FEs do?

They capture anything that happened to all the individuals for a specific year relative to the base year

Example

Education and wage data from 2012 to 2014,

$$log(income) = \beta_0 + \beta_1 educ + \beta_2 exper + \sigma_1 FE_{2012} + \sigma_2 FE_{2013}$$

- σ_1: captures the difference in $log(income)$ between 2012 and 2014 (base year)
- σ_2: captures the difference in $log(income)$ between 2013 and 2014 (base year)

Interpretation

$\sigma_1 = 0.05$ would mean that $log(income)$ is greater in 2012 than 2014 by 5% on average for whatever reasons with everything else fixed.

Recommendation

It is almost always a good practice to include year FEs if you are using a panel dataset with annual observations.

Recommendation

It is almost always a good practice to include year FEs if you are using a panel dataset with annual observations.

Why?

- Remember year FEs capture anything that happened to all the individuals for a specific year relative to the base year
- In other words, all the unobserved factors that are common to all the individuals in a specific year is controlled for (taken out of the error term)

Recommendation

It is almost always a good practice to include year FEs if you are using a panel dataset with annual observations.

Why?

- Remember year FEs capture anything that happened to all the individuals for a specific year relative to the base year
- In other words, all the unobserved factors that are common to all the individuals in a specific year is controlled for (taken out of the error term)

Example

Economic trend in:

$$log(income) = \beta_0 + \beta_1 educ + \sigma_1 FE_{2012} + \sigma_2 FE_{2013}$$

- Education is non-decreasing through time
- Economy might have either been going down or up during the observed period

Recommendation

It is almost always a good practice to include year FEs if you are using a panel dataset with annual observations.

Why?

- Remember year FEs capture anything that happened to all the individuals for a specific year relative to the base year
- In other words, all the unobserved factors that are common to all the individuals in a specific year is controlled for (taken out of the error term)

Example

Economic trend in:

$$log(income) = \beta_0 + \beta_1 educ + \sigma_1 FE_{2012} + \sigma_2 FE_{2013}$$

- Education is non-decreasing through time
- Economy might have either been going down or up during the observed period
- Without year FE, β_1 may capture the impact of overall economic trend.

R implementation

In order to include year FEs to individual FEs, you can simply add the variable that indicates year like below:

```
feols(I(log(income)) ~ educ | id + year, data = year_fe) %>%
  tidy()
```

```
## # A tibble: 1 × 5
##   term  estimate std.error statistic p.value
##   <chr>    <dbl>     <dbl>     <dbl>   <dbl>
## 1 educ    0.0801   0.00931      8.60  0.0132
```

Caveats

- Year FEs would be perfectly collinear with variables that change only across time, but not across individuals.
- If your variable of interest is such a variable, you cannot include year FEs, which would then make your estimation subject to omitted variable bias due to other unobserved yearly-changing factors.

Standard Error Estimation for Panel Data Methods

Heteroskedasticity

Just like we saw for OLS using cross-sectional data, heteroskedasticity leads to biased estimation of the standard error of the coefficient estimators if not taken into account

Serial Correlation

Correlation of errors over time, which we call serial correlation

Serial Correlation

Correlation of errors over time, which we call serial correlation

Consequences of serial correlation

- just like heteroskedasticity, serial correlation could lead to biased estimation of the standard error of the coefficient estimators if not taken into account
- do not affect the unbiasedness and consistency property of your estimators

Serial Correlation

Correlation of errors over time, which we call serial correlation

Consequences of serial correlation

- just like heteroskedasticity, serial correlation could lead to biased estimation of the standard error of the coefficient estimators if not taken into account
- do not affect the unbiasedness and consistency property of your estimators

Important

- Taking into account the potential of serial correlation when estimating the standard error of the coefficient estimators can dramatically change your conclusions about the statistical significance of some independent variables!!
- When serial correlation is ignored, you tend to underestimate the standard error (why?), inflating t-statistic, which in turn leads to over-rejection that you should.

Bertrand, Duflo, and Mullainathan (2004)

- Examined how problematic serial correlation is in terms of inference via Monte Carlo simulation
  - generate a fake treatment dummy variable in a way that it has no impact on the outcome (dependent variable) in the dataset of women's wages from the Current Population Survey (CPS)
  - run regression of the oucome on the treatment variable
  - test if the treatment variable has statistically significant effect via t-test
- They rejected the null 67.5% at the 5% significance level!!

SE robust to heteroskedasticity and serial correlation

- You can take into account both heteroskedasticity and serial correlation by clustering by individual (whatever the unit of individual is: state, county, farmer)
- Cluster by individual allows correlation within individuals (over time)

SE robust to heteroskedasticity and serial correlation

- You can take into account both heteroskedasticity and serial correlation by clustering by individual (whatever the unit of individual is: state, county, farmer)
- Cluster by individual allows correlation within individuals (over time)

R implementation

The last partition is used for clustering standard error estimation by variable like below.

```
feols(I(log(income)) ~ educ | id + year, cluster = ~id, data = year_fe) %>% tidy()
```

```
## # A tibble: 1 × 5
##   term  estimate std.error statistic p.value
##   <chr>    <dbl>     <dbl>     <dbl>   <dbl>
## 1 educ    0.0801   0.00931      8.60  0.0132
```