Week 2: Fixed-effects and random-effects models

POLI803

Howard Liu

August, 2024

University of South Carolina

Outline for today

1. Panel data
2. Fixed-effects model
3. Random-effects (mixed-effects) model

Dataset types

1. Cross-sectional data
2. Time-series data
3. Time-series cross-sectional data, also called **panel data**

Panel data

	country	year	spend	left	trade	fdi	gdppc
1.	Australia	1981	34.3	0	32.8	1.8	12689
2.	Australia	1982	36.9	0	33.5	1.8	12132
3.	Australia	1983	37.1	75	30.5	2.1	12784
4.	Australia	1984	38.4	100	32.3	1	13274
5.	Australia	1985	38.8	100	36	2.3	13583
6.	Austria	1981	50.3	100	77.9	.8	10407
7.	Austria	1982	50.9	100	74.4	.5	10484
8.	Austria	1983	51.2	88	73.5	.6	10728
9.	Austria	1984	50.8	80	77.8	.3	10877
10.	Austria	1985	51.7	80	81.3	.4	11131
11.	Belgium	1981	63.9	47	137.9	1.4	10829
12.	Belgium	1982	63.9	0	144.6	1.5	10986
13.	Belgium	1983	63.9	0	147.4	1.9	10972
14.	Belgium	1984	62.6	0	156.3	.9	11236
15.	Belgium	1985	62.3	0	151.1	1.5	11285
16.	Canada	1981	41.5	0	53.7	.7	14555
17.	Canada	1982	46.6	0	48.2	.1	13740
18.	Canada	1983	47.2	0	48	.9	14105
19.	Canada	1984	46.8	0	53.7	1.1	14954
20.	Canada	1985	47.1	0	54.4	.2	15589
21.	Denmark	1981	59.8	100	72.3	.4	11153
22.	Denmark	1982	61.2	75	72.3	.3	11526
23.	Denmark	1983	61.6	0	70.8	.4	11828
24.	...	...	...	...	...	...	...

We can't (shouldn't) apply simple OLS

We can't (shouldn't) apply simple OLS

Regular regression models assume the data set is cross-sectional.

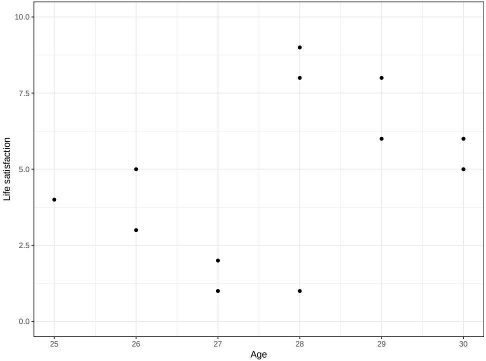
- = observations are **independent** across unit and across time (i.i.d. independent and identically distributed random variables);
- = we can meaningfully compare any pairs observations in the data set (but can we really compare United States 2001 with Switzerland 1990, for example?);
- = unit-level idiosyncrasies and time-level idiosyncrasies are ignorable.

Running standard regression models with panel data may lead to **biased inferences**.

What would happen if we did?

Age and life satisfaction (lsat)

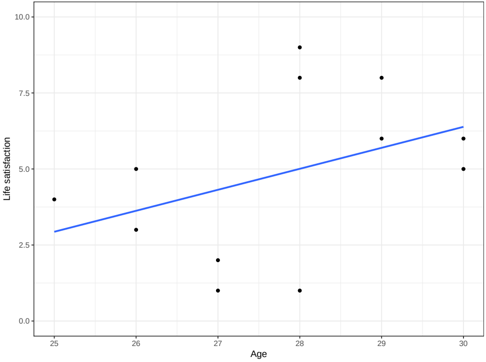
	name	year	age	lsat
1.	John	1968	28	8
2.	John	1969	29	6
3.	John	1970	30	5
4.	Paul	1968	26	5
5.	Paul	1969	27	2
6.	Paul	1970	28	1
7.	George	1968	25	4
8.	George	1969	26	3
9.	George	1970	27	1
10.	Ringo	1968	28	9
11.	Ringo	1969	29	8
12.	Ringo	1970	30	6



What would happen if we did?

Age and life satisfaction (lsat)

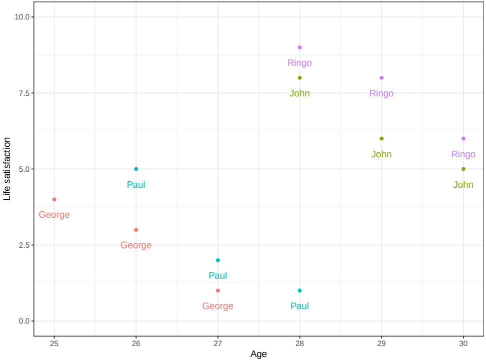
	name	year	age	lsat
1.	John	1968	28	8
2.	John	1969	29	6
3.	John	1970	30	5
4.	Paul	1968	26	5
5.	Paul	1969	27	2
6.	Paul	1970	28	1
7.	George	1968	25	4
8.	George	1969	26	3
9.	George	1970	27	1
10.	Ringo	1968	28	9
11.	Ringo	1969	29	8
12.	Ringo	1970	30	6



What would happen if we did?

Age and life satisfaction (lsat)

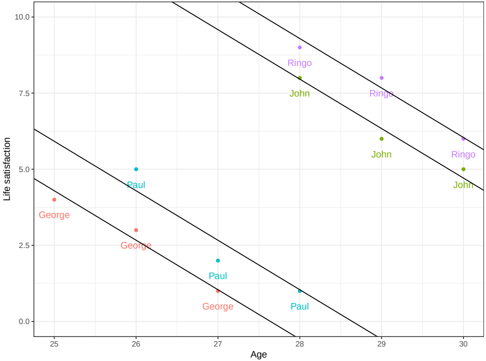
```
     +---------------------------+
     |   name   year   age   lsat |
     |---------------------------|
  1. |   John   1968    28      8 |
  2. |   John   1969    29      6 |
  3. |   John   1970    30      5 |
     |---------------------------|
  4. |   Paul   1968    26      5 |
  5. |   Paul   1969    27      2 |
  6. |   Paul   1970    28      1 |
     |---------------------------|
  7. | George   1968    25      4 |
  8. | George   1969    26      3 |
  9. | George   1970    27      1 |
     |---------------------------|
 10. |  Ringo   1968    28      9 |
 11. |  Ringo   1969    29      8 |
 12. |  Ringo   1970    30      6 |
     +---------------------------+
```



What would happen if we did?

Age and life satisfaction (lsat)

	name	year	age	lsat
1.	John	1968	28	8
2.	John	1969	29	6
3.	John	1970	30	5
4.	Paul	1968	26	5
5.	Paul	1969	27	2
6.	Paul	1970	28	1
7.	George	1968	25	4
8.	George	1969	26	3
9.	George	1970	27	1
10.	Ringo	1968	28	9
11.	Ringo	1969	29	8
12.	Ringo	1970	30	6



What we should do instead

We need to fit four regression lines, rather than one

How do we do this?

What we should do instead

We need to fit four regression lines, rather than one

How do we do this?

- Create a series of dummy variables, one for each person

What we should do instead

We need to fit four regression lines, rather than one

How do we do this?

- Create a series of dummy variables, one for each person
- Include these four dummy variables, while dropping the intercept

age	0.690	−1.625***
	(0.491)	(0.166)
John		53.458***
		(4.819)
Paul		46.542***
		(4.488)
Ringo		54.792***
		(4.819)
George		44.917***
		(4.322)
Constant	−14.322	
	(13.656)	
Observations	12	12
R^2	0.165	0.996
Adjusted R^2	0.081	0.993
Residual Std. Error	2.612 (df = 10)	0.469 (df = 7)
Note:		*p<0.1; **p<0.05; ***p<0.01

Fixed-effects models

When we run a regression model that gives each unit (e.g., country, individual, etc.) a different intercept, we say we run a **fixed-effects** (FE) model

- Unit-specific intercepts are called unit-specific fixed-effects
- Such a model allows us to control for any unit-specific confounders
- We are essentially making a **within-unit comparison**
  - Compare Ringo's lsat when he was 28 with Ringo's lsat when he was 29 (within)
  - We never compare Ringo's lsat when he was 28 with Paul's lsat when he was 28 (between)

Fixed-effects models

- How does FE work specifically?

$$y_{it} = \beta_1 x_{ij} + a_i + u_{ij} \qquad (1)$$

Fixed-effects models

- How does FE work specifically?

$$y_{it} = \beta_1 x_{ij} + a_i + u_{ij} \qquad (1)$$

- Now, for each i, average this equation over time. We get

$$\bar{y}_i = \beta_1 \bar{x}_i + a_i + \bar{u}_i \qquad (2)$$

Fixed-effects models

- How does FE work specifically?

$$y_{it} = \beta_1 x_{ij} + a_i + u_{ij} \qquad (1)$$

- Now, for each i, average this equation over time. We get

$$\bar{y}_i = \beta_1 \bar{x}_i + a_i + \bar{u}_i \qquad (2)$$

- Because a_i is fixed over time, it appears in both equations. If we subtract (2) from (1) for each t, we wind up with

$$y_{it} - \bar{y}_i = \beta_1(x_{it} - \bar{x}_i) + u_{it} - \bar{u}_i \qquad (3)$$

Fixed-effects models

- How does FE work specifically?

$$y_{it} = \beta_1 x_{ij} + a_i + u_{ij} \qquad (1)$$

- Now, for each i, average this equation over time. We get

$$\bar{y}_i = \beta_1 \bar{x}_i + a_i + \bar{u}_i \qquad (2)$$

- Because a_i is fixed over time, it appears in both equations. If we subtract (2) from (1) for each t, we wind up with

$$y_{it} - \bar{y}_i = \beta_1 (x_{it} - \bar{x}_i) + u_{it} - \bar{u}_i \qquad (3)$$

- This is called **time-demeaning** or **within transformation** because we only estimate time-demeaned variables and the unobserved effect (like country-specific effects) a_i disappeared

The `plm` package

We use the `plm` (panel linear model) package to make this easier

- Install the package: `install.packages("plm", dependencies = TRUE)`
- Load the package: `library(plm)`
- Declare the data to be a panel data:

```
pdata.frame(data, index = c("name", "year"))
```

The plm package

To run a simple model (i.e., a model that ignores the panel structure),

```
plm(y ~ x, data, model = "pooling")
```

To run a fixed-effects model (i.e., a model that fits a different line to a different unit),

```
plm(y ~ x, data, model = "within")
```

Example: Effect of globalization on welfare state

Garrett and Mitchell (2001): “Globalization, government spending and taxation in the OECD”

- IDV: globalization (total trade, imports from low wage economies, FDI, market integration)
- DV: welfare effort (government spending and taxation)
- Data: OECD countries (18 advanced economies for 1961–1994)

Numerical and graphical summaries

`summary(data)` would be hardly enough

```
> summary(gm)
   country              cnum          year         govcons         govconsl
 Length:612         Min.   : 1.0   Min.   :1961   Min.   : 7.30   Min.   : 7.30
 Class :character   1st Qu.: 5.0   1st Qu.:1969   1st Qu.:14.30   1st Qu.:14.10
 Mode  :character   Median : 9.5   Median :1978   Median :17.00   Median :16.80
                    Mean   : 9.5   Mean   :1978   Mean   :16.91   Mean   :16.74
                    3rd Qu.:14.0   3rd Qu.:1986   3rd Qu.:18.93   3rd Qu.:18.90
                    Max.   :18.0   Max.   :1994   Max.   :29.60   Max.   :29.60

     sstran          sstranl           trade           lowwage           fdi
 Min.   : 3.70   Min.   : 3.700   Min.   :  9.40   Min.   : 6.20   Min.   : 0.000
 1st Qu.: 9.50   1st Qu.: 9.175   1st Qu.: 39.27   1st Qu.:12.80   1st Qu.: 0.600
 Median :13.35   Median :13.050   Median : 52.90   Median :16.65   Median : 1.000
 Mean   :13.69   Mean   :13.378   Mean   : 57.10   Mean   :19.05   Mean   : 1.449
 3rd Qu.:17.02   3rd Qu.:16.700   3rd Qu.: 71.72   3rd Qu.:23.43   3rd Qu.: 1.800
 Max.   :28.90   Max.   :28.900   Max.   :156.30   Max.   :46.00   Max.   :10.300
 NA's   :48      NA's   :44                                        NA's   :68
```

Figure out

- Cross-sectional unit
- Time-series unit

Numerical and graphical summaries

- Cross-sectional unit

```
> table(gm $ country)

  Australia     Austria     Belgium      Canada     Denmark     Finland      France 
         34          34          34          34          34          34          34 
    Germany     Ireland       Italy       Japan Netherlands New Zealand      Norway 
         34          34          34          34          34          34          34 
     Sweden Switzerland          UK          US 
         34          34          34          34 
>
```

- Time-series unit

```
> table(gm $ year)

1961 1962 1963 1964 1965 1966 1967 1968 1969 1970 1971 1972 1973 1974 1975 1976 1977 
  18   18   18   18   18   18   18   18   18   18   18   18   18   18   18   18   18 
1978 1979 1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994 
  18   18   18   18   18   18   18   18   18   18   18   18   18   18   18   18   18 
```

Once you figure these two things out, then provide numerical and graphical summaries of X and Y for **each unit** and/or **over time**

Numerical and graphical summaries

To obtain numerical summaries by unit, we use the `by` function

Numerical and graphical summaries

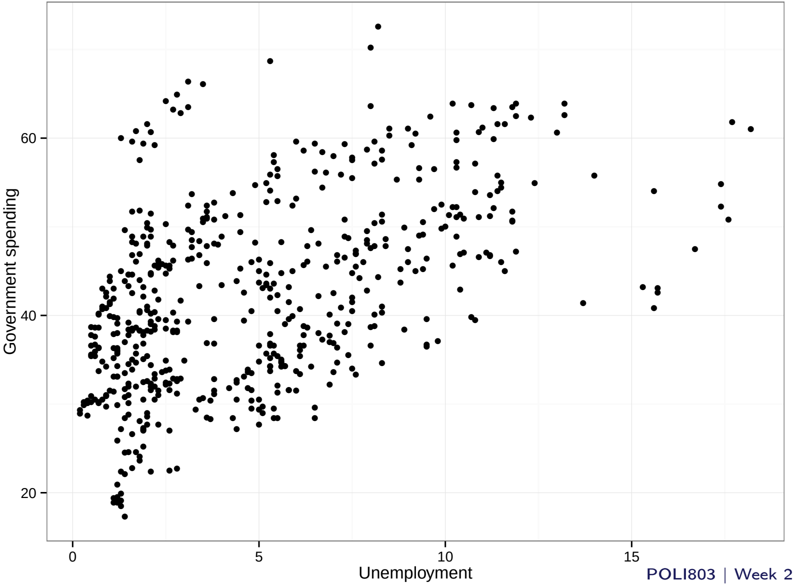
To obtain numerical summaries by unit, we use the `by` function

```
by(X, ID, FUNCTION)
```

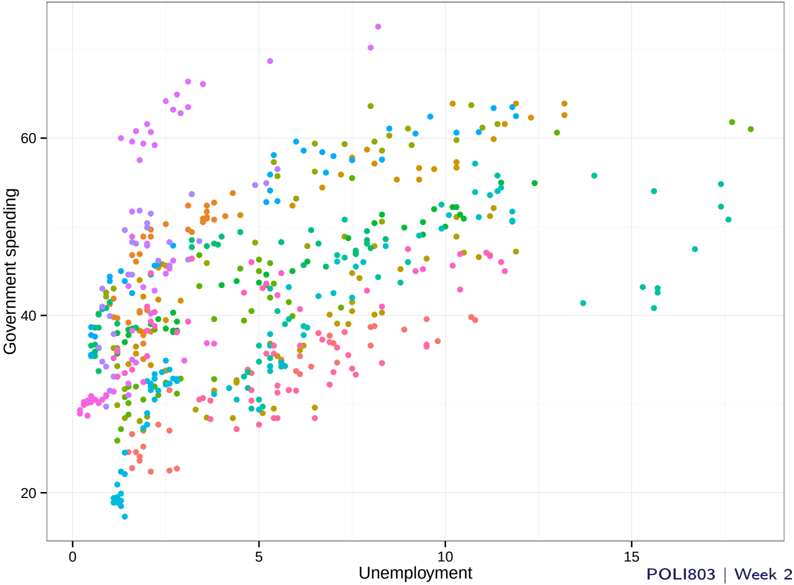
```
> by(gm $ spend, gm $ country, summary)
gm$country: Australia
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  22.40   25.20   33.70   31.62   36.90   39.80       1
-------------------------------------------------------------
gm$country: Austria
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  35.10   39.92   47.90   45.67   50.88   53.80
-------------------------------------------------------------
gm$country: Belgium
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  33.80   41.87   54.85   50.66   57.67   63.90
-------------------------------------------------------------
gm$country: Canada
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  28.40   33.23   40.10   39.73   46.55   52.10
-------------------------------------------------------------
```

==

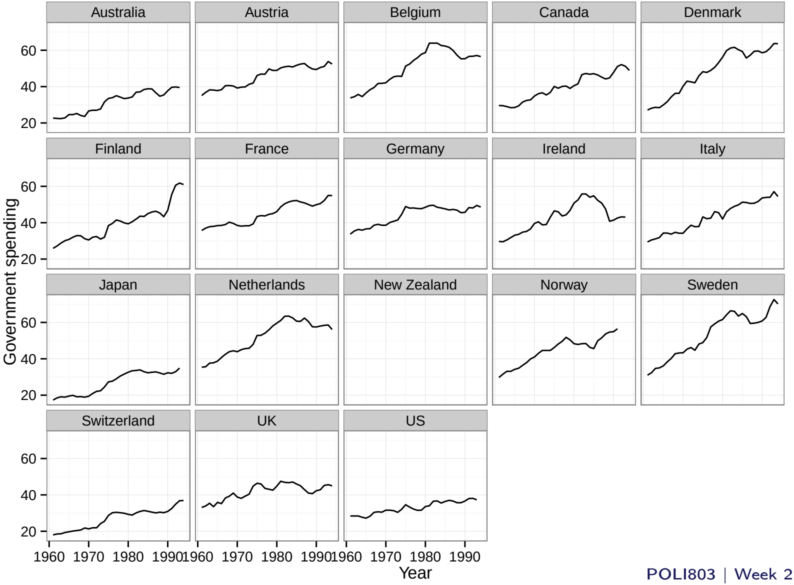
Government spending and unemployment



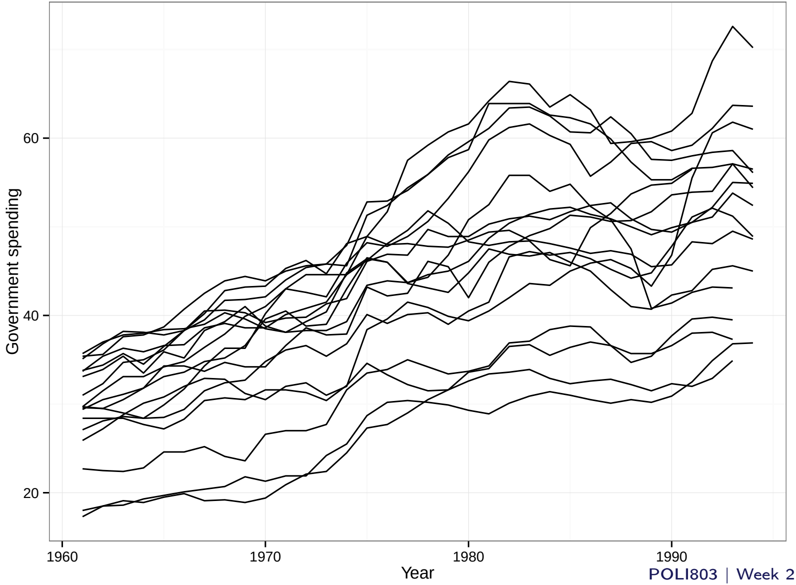
Government spending and unemployment



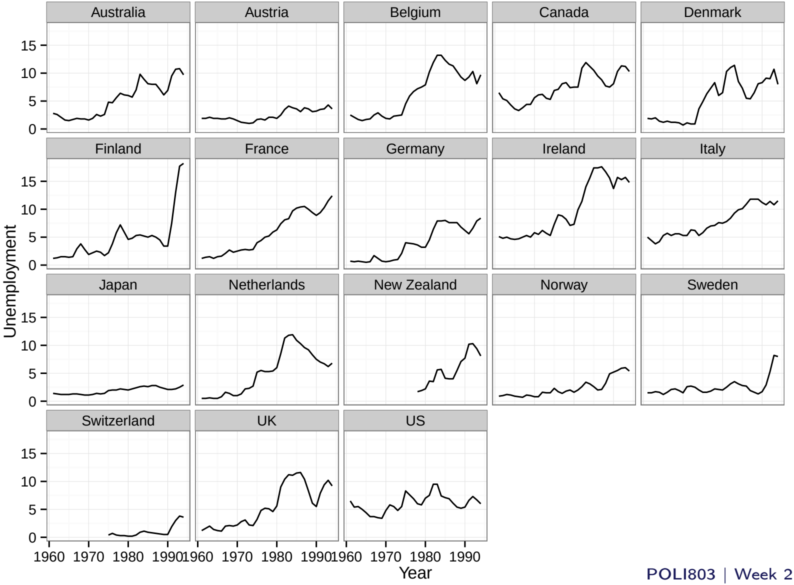
Government spending



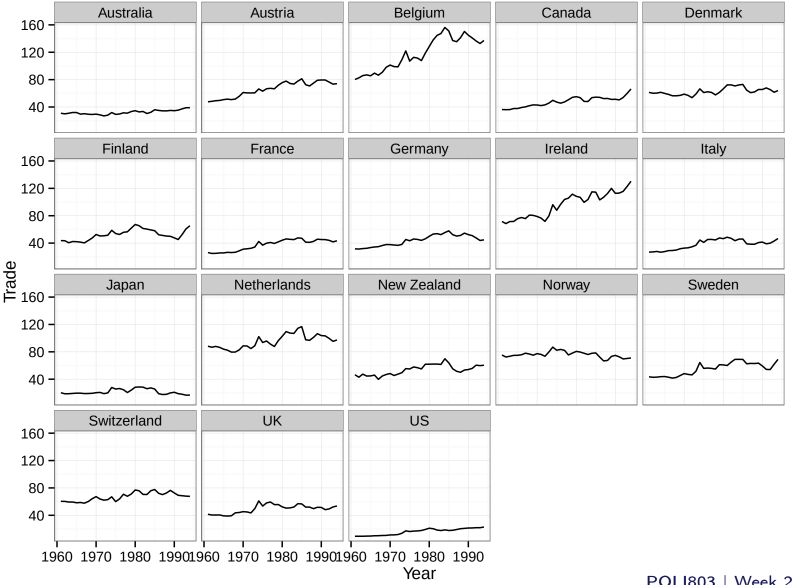
Government spending



Unemployment



Trade


Australia
Austria
Belgium
Canada
Denmark
Finland
France
Germany
Ireland
Italy
Japan
Netherlands
New Zealand
Norway
Sweden
Switzerland
UK
US
Trade
160
120
80
40
1960 1970 1980 1990
Year

Estimate regression models

1. Tell R that this is a panel data set `bm.p <- pdata.frame(bm, index = c("country", "year"))`
2. Estimate
   2. Pooled model
      `plm(Y ∼ X, data, model = "pooling")`
   2. FE (within-effect) model
      `plm(Y ∼ X, data, model = "within")`
3. Compare the results

	Pooling	FE
Unemployment	1.120***	1.366***
	(0.089)	(0.087)
Trade	0.143***	0.202***
	(0.012)	(0.026)
Leftist	0.066***	−0.012*
	(0.009)	(0.007)
Growth	−1.014***	−0.830***
	(0.126)	(0.085)
Christian Democrat	0.044***	−0.051***
	(0.012)	(0.015)
Constant	28.396***	
	(0.886)	
Observations	557	557
R^2	0.569	0.700
Adjusted R^2	0.563	0.672
Note:	*p<0.1; **p<0.05; ***p<0.01	

Testing if FE is better than pooled

Whenever you run a FE model, perform a test (Lagrange Multiplier Test) that compares it with the pooled model

```
> pFtest(mod.fe, mod.pool)

        F test for individual effects

data:  spend ~ unem + trade + left + growthpc + cdem
F = 55.6187, df1 = 16, df2 = 535, p-value < 2.2e-16
alternative hypothesis: significant effects
```

The null hypothesis: **FE = pooled** (FE doesn't improve)

- A small *p*-value $\rightsquigarrow$ FE needed
- A *p*-value $> 0.10 \rightsquigarrow$ FE not necessary

Random-effects model

FE models have several drawbacks:

- Efficiency problem: The number of intercepts may get very large. But, the degree of freedom $= n - k$ must be positive (where k is the number of αs and βs) for us to be able to identify unique values of αs and βs
- Time-invariant variable cannot be included on the RHS!

Random-effects model

A **random-effects** (RE) model can be an alternative:

- Statistian called "mixed effect model": Including both within and across unit variation together (Z)
- Instead of estimating unit-specific intercepts directly, RE models estimate the standard deviation of the intercepts
- You can estimate random intercepts (with same slopes) or random intercepts and slopes → more flexibility

$$y_{it} = \beta_1 x_{ij} + a_i + u_{ij} \qquad (1)$$

$$a_i = \beta_0 + \beta_2 Z_t + e_t$$

- where the latent vairable, Z_t, contains both within and between variation to be explained. So RE is a hierarchical/muti-level model
- Based on a set of assumptions
  - REs follow a normal distribution
  - REs are not correlated with Xs (covariates not correlated with unit-specific structure)

Which model to use – FE or RE?

- Theoretical answer
  - If you can be absolutely certain that unit-specific intercepts are uncorrelated with the Xs, use the RE model (it's more **efficient**)
  - If you are sure that unit-specific intercepts are correlated with the Xs, use the FE approach (it's more **flexible**)
- Reality:
  - If you have a time-invariant variable as your main treatment variable, go for RE
  - If your theory cares not only within unit comparison, but also cross unit comparison, go for RE (e.g. econ ineqaulity vs civil war)
  - Causal inference folks care about eliminating heterogenous treatment effects, so use FE more.
- Hausman test tests this empirically:

```
phtest(mod.re, mod.fe)
```

  - The null hypothesis is that RE and FE are equivalent

Which model to use – FE or RE?

- Theoretical answer
  - If you can be absolutely certain that unit-specific intercepts are uncorrelated with the Xs, use the RE model (it's more **efficient**)
  - If you are sure that unit-specific intercepts are correlated with the Xs, use the FE approach (it's more **flexible**)

- Reality:
  - If you have a time-invariant variable as your main treatment variable, go for RE
  - If your theory cares not only within unit comparison, but also cross unit comparison, go for RE (e.g. econ ineqaulity vs civil war)
  - Causal inference folks care about eliminating heterogenous treatment effects, so use FE more.

- Hausman test tests this empirically:

```
phtest(mod.re, mod.fe)
```

  - The null hypothesis is that RE and FE are equivalent
  - When p-value is small enough, you have to use FE

	Pooled	FE	RE
Unemployment	1.120***	1.366***	1.359***
	(0.089)	(0.087)	(0.083)
Trade	0.143***	0.202***	0.199***
	(0.012)	(0.026)	(0.023)
Leftist	0.066***	−0.012*	−0.009
	(0.009)	(0.007)	(0.007)
Growth	−1.014***	−0.830***	−0.838***
	(0.126)	(0.085)	(0.086)
Christian Democrat	0.044***	−0.051***	−0.044***
	(0.012)	(0.015)	(0.015)
Constant	28.396***		27.065***
	(0.886)		(1.768)
Observations	557	557	557
R^2	0.569	0.700	0.689
Adjusted R^2	0.563	0.672	0.682

Note: $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Testing if FE is better than RE

```
> phtest(mod.fe, mod.re)

        Hausman Test

data:  spend ~ unem + trade + left + growthpc + cdem
chisq = 43.4071, df = 5, p-value = 3.056e-08
alternative hypothesis: one model is inconsistent
```

The null hypothesis: FE = RE

- A small p-value $\rightsquigarrow$ FE needed
- A p-value > 0.10 $\rightsquigarrow$ FE not necessary (RE is OK)