Fundamentals of Econometrics

Lecture 8: Pooling Cross Sections across Time: Simple Panel Data Methods

Section 1

Simple Panel Data Methods

Panel Data

- Panel data is a combination of time series and cross-sectional data.
- It consists of observations on multiple entities (individuals, firms, countries, etc.) over multiple time periods.
- To gather a panel data set, we can either:
  - Collect data on the same individuals over time (e.g., a survey of the same individuals at different points in time).
  - Collect data on different individuals at different points in time (e.g., a survey of different individuals at different points in time).

Independently Pooled Cross Sections

- Collection of independent, random samples from the same population at multiple periods of time.

Some Data Sources

- The Panel Study of Income Dynamics (PSID): Collected by the University of Michigan
- National Longitudinal Surveys (NLS): Collected by the Bureau of Labor Statistics
- Medical Expenditure Panel Survey (MEPS): Collected by the Agency for Healthcare Research and Quality
- National Health and Nutrition Examination Survey (NHANES): Collected by the National Center for Health Statistics at the CDC
- American Community Survey (ACS): Collected by the U.S. Census Bureau
- Current Population Survey (CPS): Collected by the Bureau of Labor Statistics
- American Time Use Survey (ATUS): Collected by the Bureau of Labor Statistics

Using Pooled Cross Sections Data

- We want to pool all cross-sections over time into one single dataset.

Advantages:

1. Increased sample size: By pooling data from multiple cross-sections, we can increase the sample size, which can lead to more precise estimates.
2. Improve generalizability: By pooling data from different time periods, we can improve the generalizability of our results to the population.
3. Can be used to estimate the effect of a policy change or event on a population (natural experiment).

Section 2

Policy Analysis with Pooled Cross Sections

Natural or Quasi-Experiments

- A natural experiment is when an exogenous shock occurs to a system and affects individual behavior.
- We have a group of individuals affected by the shock and a group of individuals that are not. So this is similar in principle to a laboratory experiment where there is a treatment group (affected by shock) and control group (not affected by shock).
- A quasi-experiment is when a researcher uses a natural experiment to estimate the effect of a treatment on an outcome.

Two requirements:

- Two time periods (one before and one after the policy change)
- Two groups (treatment and control)

Natural Experiment Framework

Goal: To determine differences between treatment and control groups due to an exogenous shock.

1. Pool the data from the two time periods.
2. Include a dummy variable for time and group.

- d2 = 1 if obs occurs after event
- dT = 1 if obs occurs in treatment group

3. Include additional variables and an interaction term between the two dummy variables

$$y = \beta_0 + \underbrace{\delta_0 d2}_{\substack{\text{Controls for}\\ \text{unobserved}\\ \text{changes affecting}\\ \text{both groups}}} + \underbrace{\delta_1 dT}_{\substack{\text{Controls for}\\ \text{initial difference}\\ \text{between groups}}} + \delta_2 d2 \cdot dT + \underbrace{\text{other factors}}_{\substack{\text{Controls for}\\ \text{observable differences}\\ \text{between treatment}\\ \text{and control group}}} + u$$

Group	Before	After	After - Before
Treatment	$\hat{\beta}_0 + \hat{\delta}_1$	$\hat{\beta}_0 + \hat{\delta}_0 + \hat{\delta}_1 + \hat{\delta}_2$	$\hat{\delta}_0 + \hat{\delta}_2$
Control	$\hat{\beta}_0$	$\hat{\beta}_0 + \hat{\delta}_0$	$\hat{\delta}_0$
Treatment - Control	$\hat{\delta}_1$	$\hat{\delta}_1 + \hat{\delta}_2$	$\hat{\delta}_2$

Effect of Garbage Incinerator Locations on Housing Prices

Example - Blacksburg is considering the location of a new garbage incinerator. We want to conduct an analysis of the impact of the incinerator on property values.

- We could consider a town similar to Blacksburg where an incinerator was built. We will use Boston housing data from Kiel and McCain (1995). We have data from 1971 and in 1981 (when the incinerator was built).
- Similar town with:
  - 30,000 residents
  - Small college of 2,000 students
  - 25 miles from the nearest city

```r
l.before <- lm(rprice ~ nearinc, data = kielmc, subset = year == 1978)
l.after <- lm(rprice ~ nearinc, data = kielmc, subset = year == 1981)

stargazer(l.before, l.after,
  font.size = "scriptsize",
  title = "Pooled Cross Sections", column.labels = c("Before", "After"),
  omit.stat = c("f", "ser", "adj.rsq"), header = FALSE
)
```

Table 2: Pooled Cross Sections

	Dependent variable:	
	rprice	
	Before	After
	(1)	(2)
nearinc	−18,824.000***	−30,688.000***
	(4,745.000)	(5,828.000)
Constant	82,517.000***	101,308.000***
	(2,654.000)	(3,093.000)
Observations	179	142
R^2	0.082	0.165
Note:	*p<0.1; **p<0.05; ***p<0.01	

Garbage Incinerator Example

- What is the effect of the garbage incinerator on housing prices?
- Being close to an incinerator depresses prices, but location was one with lower prices to begin with.
- The effect of the garbage incinerator on housing prices is given by the difference in the coefficients of the `nearinc` variable in the two models.

$$\hat{\delta} = -30688.274 - (-18824.370) = \boxed{-11863.903}$$

$\hat{\delta}$ is known as the **difference-in-difference (DiD)** estimator. It can be expressed as the difference over time in the average difference in housing prices between the two groups (near and far from the incinerator):

$$\hat{\delta} = \left(\overline{rprice}_{81,near} - \overline{rprice}_{81,far}\right) - \left(\overline{rprice}_{78,near} - \overline{rprice}_{78,far}\right)$$

Final Words

Basic Setup of DiD:

- Two groups (treatment ($D_i = 1$) and control ($D_i = 0$))
- Two time periods (before ($T = 0$) and after the treatment (T=1))

Group (D_i)	After ($T_i = 1$)	Before ($T_i = 0$)
Treated ($D_i = 1$)	$E[Y_{1i}(1)\|D_i = 1]$	$E[Y_{0i}(0)\|D_i = 1]$
Control ($D_i = 0$)	$E[Y_{0i}(1)\|D_i = 0]$	$E[Y_{0i}(0)\|D_i = 0]$

The **fundamental challenge**: We cannot observe $E[Y_{0i}(1)|D_i = 1]$– i.e., the **counterfactual outcome** for the treated group had they not received treatment.

- DiD estimates the **Average Treatment Effect on the Treated** (ATT) as:

$$E[Y_{1i}(1) - Y_{0i}(1)|D_i = 1] = E[Y_{1i}(1)|D_i = 1] - E[Y_{0i}(0)|D_i = 1]$$

DiD Estimator in Regressions

$$rprice_{it} = \beta_0 + \delta_0 after + \beta_1 nearinc + \delta_1 after \cdot nearinc + u_{it}$$

- The differential effect of being in the location **and** after the incinerator was built is given by δ_1.
- The DiD estimator is the difference in the coefficients of `nearinc` in the two models.
- If houses sold before and after the incinerator was built were systematically different, further explanatory variables should be included.
  - **This will also reduce the error variance and thus standard errors.**

DiD Estimator in Regressions

Table 4: Pooled Cross Sections

	Dependent variable:		
	rprice		
	(1)	(2)	(3)
y81	18,790.000***	21,321.000***	13,928.000***
	(4,050.000)	(3,444.000)	(2,799.000)
nearinc	−18,824.000***	9,398.000*	3,780.000
	(4,875.000)	(4,812.000)	(4,453.000)
y81:nearinc	−11,864.000	−21,920.000***	−14,178.000***
	(7,457.000)	(6,360.000)	(4,987.000)
Constant	82,517.000***	89,117.000***	13,808.000
	(2,727.000)	(2,406.000)	(11,167.000)
Other Controls	No	Age, AgeSq	Full Set
Observations	321	321	321
R^2	0.174	0.414	0.660
Note:			*p<0.1; **p<0.05; ***p<0.01

```r
l.difference <- lm(rprice ~ y81 + nearinc + y81:nearinc,
  data = kielmc
)

l.difference2 <- lm(
  rprice ~ y81 + nearinc + y81:nearinc +
    age + agesq,
  data = kielmc
)

l.difference3 <- lm(
  rprice ~ y81 + nearinc + y81:nearinc +
    age + agesq + intst + land + area +
    rooms + baths,
  data = kielmc
)

stargazer(l.difference, l.difference2, l.difference3,
  font.size = "scriptsize",
  title = "Pooled Cross Sections",
  # Keep only y81, nearinc, y81:nearinc, "Constant")
  keep = c("Constant", "y81", "nearinc", "y81:nearinc"),
  add.lines = list(c("Other Controls", "No", "Age, AgeSq", "Full Set")),
  omit.stat = c("f", "ser", "adj.rsq"), header = FALSE
)
```

Table 5: Pooled Cross Sections

	Dependent variable:		
	log(rprice)		
	(1)	(2)	(3)
y81	0.193***	0.220***	0.162***
	(0.045)	(0.037)	(0.028)
nearinc	−0.340***	0.007	0.032
	(0.055)	(0.052)	(0.047)
y81:nearinc	−0.063	−0.185***	−0.132**
	(0.083)	(0.068)	(0.052)
Constant	11.300***	11.400***	7.650***
	(0.031)	(0.026)	(0.416)
Other Controls	No	Age, AgeSq	Full Set
Observations	321	321	321
R^2	0.246	0.509	0.733
Note:	*p<0.1; **p<0.05; ***p<0.01		

Model in Logs

```r
l.difference.log <- lm(log(rprice) ~ y81 + nearinc + y81:nearinc,
  data = kielmc
)
l.difference2.log <- lm(
  log(rprice) ~ y81 + nearinc + y81:nearinc +
    age + agesq,
  data = kielmc
)
l.difference3.log <- lm(
  log(rprice) ~ y81 + nearinc + y81:nearinc +
    age + agesq + lintst + lland + larea +
    rooms + baths,
  data = kielmc
)
stargazer(l.difference.log, l.difference2.log, l.difference3.log,
  font.size = "scriptsize",
  title = "Pooled Cross Sections",
  # Keep only y81, nearinc, y81:nearinc
  keep = c("Constant", "y81", "nearinc", "y81:nearinc"),
  add.lines = list(c("Other Controls", "No", "Age, AgeSq", "Full Set")),
  omit.stat = c("f", "ser", "adj.rsq"), header = FALSE
)
```

Adding Additional Control Groups

- An implicit assumption of the DiD estimator is that the treatment and control groups are similar in all respects except for the treatment.
- We assume that potential trends in the outcome, y, would trend at the same rate in the absence of the treatment.
- This is known as the **parallel trends assumption**.
- If this assumption is violated, the DiD estimator will be biased.
- We can add an additional control group (say a state that has not received the treatment but has a similar trend pre-treatment) to the model.
- Leads us to a **triple difference** estimator (**Difference-in-Difference-in-Difference**).

Adding Additional Control Groups

Example: The effects of Health care expansion to Low-Income families in a state

- Let L denote low-income families (eligible for the policy) and M be middle-income families (not eligible).
- Let B denote states that implemented the policy and A be states that did not implement the policy.
- The policy is implemented in period 1 $(T = 2)$, but no policy exists in period 0 $(T = 1)$.
- The additional control group (income level) allows for more flexibility if we assume that potential differences in trends in health outcomes between low and middle income families is similar across states.

Adding Additional Control Groups

$$y = \beta_0 + \beta_1 dL + \beta_2 dB + \beta_3 dL \cdot dB + \delta_0 dT + \delta_1 dL \cdot dT + \delta_2 dB \cdot dT + \delta_3 dL \cdot dB \cdot dT +$$

$$\hat{\delta}_3 = [(\bar{y}_{2,L,B} - \bar{y}_{1,L,B})] - [(\bar{y}_{2,L,A} - \bar{y}_{1,L,A})] - [(\bar{y}_{2,M,B} - \bar{y}_{1,M,B})] - [(\bar{y}_{2,M,A} - \bar{y}_{1,M,A})]$$

$$\hat{\delta}_3 = \delta_{DD,B} - \delta_{DD,A} = \delta_{DDD}$$

Key Notes

- The difference-in-difference-in differences estimator has two components
  - A DD estimator looking only at states that implemented the policy.
  - A DD estimator looking only at states that did not implement the policy.
- If health trends between the L and M groups do not differ in non-implementation states, then the second component vanishes and we are back to the standard DiD setup.
  - However, we include this second term to account for possibly different trends in the L and M groups that are common across both states A and B.

Generalizing the Framework for Policy Analysis

$$y_{i,g,t} = \lambda_t + \gamma_g + \beta x_{g,t} + \mathbf{z}_{i,g,t}\boldsymbol{\gamma} + u_{i,g,t}$$

where:

- $y_{i,g,t}$ is the outcome variable for individual i in group g at time t.
- λ_t is an aggregate time effect common across all groups.
- γ_g is a group-specific effect common across all individuals in group g.
- $x_{g,t}$ is a group-specific treatment variable (captures the **policy effect**).
- $\mathbf{z}_{i,g,t}$ is a vector of individual- and group-specific control variables.

Generalizing the Framework for Policy Analysis

What happens if we believe the policy, $p_{g,t}$, takes time to have an effect?

- We can include a lagged policy effect in the model.

$$y_{i,g,t} = \lambda_t + \gamma_g + + \beta_1 p_{g,t} + \beta_2 p_{g,t-1} + \beta_3 p_{g,t-2} + \gamma_1 z_{i,g,t1} + \ldots + \gamma_J z_{i,g,tJ} + u_{i,g,t}$$

What happens if we believe the policy has different effects between groups?

- We can include groups specific policy effects in the model.

Section 3

Advanced Panel Data Methods

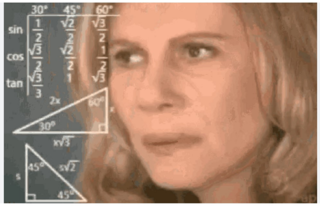
Panel Data (Motivation)

```
lm(crmrte ~ unem, data = crime2,
   subset = c(year == 87)) |> summary() |> coef()
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   128.38      20.76    6.18  1.8e-07
## unem           -4.16       3.42   -1.22  2.3e-01
```

Takeaways

- Interpreting this simple regression model causally, we get that a 1% increase in the unemployment rate will *lower* the crime rate by 4.16%.



Panel Data (Motivation)

- Many factors in the real world can be hard to control for.
- In the panel framework we view the unobserved factors affecting the dependent variable as either:
  1. fixed (constant) or
  2. random (vary) over time.

Two-Period Fixed Effects Model

- Panel data contains observations on the same individuals in every time period.
- The main advantage is that we can control for all of the unobserved features of individuals that do not change over time. We call these individual-specific characteristics either unobserved or fixed effects.

Two Period Fixed Effects Model:

$$y_{it} = \beta_0 + \gamma_0 d2 + \beta_1 x_{i1t} + \ldots + \beta_k x_{ikt} + \theta_i + \varepsilon_{it}, \quad i = 1, \ldots, N, t = 1, 2$$

where:

- $d2 = 1$ if the observation occurs after the event (t=2), and zero otherwise.
- θ_i is the **unobserved** or **fixed effect** for individual i.

Two-Period Fixed Effects Model

The composite error term is given by:

$$u_{it} = \theta_i + \varepsilon_{it}$$

In general, we need $E[u_{it}|x_{it}, \theta_i] = 0$.

But, OLS is unbiased if $\mathrm{E}[\varepsilon_{it}|x_{it}, \theta_i] = 0$.

- $Corr[X, \theta_i] \neq 0$ is okay
- Can relax usual conditional mean assumptions for error term.
- All comes out of measurement on the same individuals over multiple time periods.

First Difference Model

$$
\begin{aligned}
\text{Time 1:} \quad & y_{i1} = \beta_0 + \beta_1 X_{i11} + \ldots + \beta_k X_{ik1} + \theta_i + \varepsilon_{i1} \\
\text{Time 2:} \quad & y_{i2} = (\beta_0 + \gamma_0) + \beta_1 X_{i12} + \ldots + \beta_k X_{ik2} + \theta_i + \varepsilon_{i2}
\end{aligned}
$$

Subtracting the first equation from the second:

$$y_{i2} - y_{i1} = \gamma_0 + \beta_1(X_{i12} - X_{i11}) + \ldots + \beta_k(X_{ik2} - X_{ik1}) + (\varepsilon_{i2} - \varepsilon_{i1})$$

The FD Estimator is then:

$$\Delta y_i = \gamma_0 + \beta_1 \Delta X_{i1} + \ldots + \beta_k \Delta X_{ik} + \Delta u_i$$

We will then:

- Use data to calculate the first difference.
- Estimate the model using OLS.

First Difference Model

Disadvantages of the First Difference Model

- All time invariant variables are dropped from the model (e.g. binary variables like `gender`, `race`, etc.)
  - This is a problem if we want to include these variables in the model and to determine their effect on the dependent variable.
- If there is little variation in the independent over time the variance of the first difference will be small and the standard errors will be large.
  - This is a problem if we want to include these variables in the model and to determine their effect on the dependent variable.

Do open container law decrease auto fatalities?

Dataset: `traffic1`

- 50 states plus DC
- dthrte = deaths per 100 million miles driven
- open = 1 if state has open container law, 0 otherwise
- admn = 1 if state can suspend license before DD conviction, 0 otherwise

Table 6: States with Open Container Law

	Deaths	Admn	Open
1985	137.699999332428	21	19
1990	109.899999141693	29	22

```r
traffic1 |>
  summarize(
    across(
      c(dthrte85, dthrte90, admn85, admn90, open85, open90),
      ~sum(.x) |> format(3)
    )
  ) |>
  matrix(nrow = 2, dimnames = list(
    c("1985", "1990"), c("Deaths", "Admn","Open"))) |>
  knitr::kable(caption = "States with Open Container Law")
```

Model & DiD Estimation

Model: $dthrte_{it} = \beta_0 + \gamma_t + \beta_1 open_{it} + \beta_2 admn_{it} + \theta_i + \varepsilon_{it}$

DiD Estimator:

$dthrte_{it} = \beta_0 + \gamma_1 d2 + \beta_1 open_{it} + \delta_1 d2 \cdot open_{it} + \beta_2 admn_{it} + \delta_2 d2 \cdot admn_{it} + \varepsilon_{it}$

```r
lm(dthrte ~ d2 * (open + admn), data = traff) |> summary()
```

```r
lm(dthrte ~ d2 * (open + admn), data = traff) |> summary()
```

```
##
## Call:
## lm(formula = dthrte ~ d2 * (open + admn), data = traff)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -0.795 -0.356 -0.115  0.330  1.715
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.685      0.118   22.67   <2e-16 ***
## d2            -0.590      0.176   -3.35   0.0011 **
## open          -0.213      0.162   -1.31   0.1926
## admn           0.230      0.160    1.44   0.1532
## d2:open        0.272      0.229    1.19   0.2378
## d2:admn       -0.169      0.227   -0.74   0.4590
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.561 on 96 degrees of freedom
## Multiple R-squared:  0.227,  Adjusted R-squared:  0.187
## F-statistic: 5.64 on 5 and 96 DF,  p-value: 0.000134
```

Model & DiD Estimation

Key Takeaways

- We would conclude that open container laws have no effect on fatalities.
- Issues?
  - Causality: it may be that states enact open container laws because they have higher death rates.
    - So death rates may be causing a change in laws, not the other way around
  - Neither model accounts for much variation in fatalities. There are many unobserved variables that might be correlated with the open container laws. For example,
    - states with open container laws may also have stricter DUI laws, which may also reduce fatalities.
    - $ amount of fines for speeding or other traffic infractions,
    - maintenance and safety of roads,
    - use of seat belts, etc.

```r
traff <- traffic1 |>
  select(state:speed85) |>
  pivot_longer(
    cols = -state,
    names_to = c("var", "year"),
    # chars before last 2 digits as var, last 2 as year
    names_pattern = "(.*)(\\d{2})",
    values_to = "value"
  ) |>
  mutate(d2 = ifelse(year == "90", 1, 0)) |>
  pivot_wider(
    names_from = var,
    values_from = value
  )
```

FD Estimator: $\Delta dthrte_{it} = \beta_0 + \beta_1 \Delta open_{it} + \beta_2 \Delta admn_{it} + \varepsilon_{it}$

```r
lm(dthrte90 - dthrte85 ~ I(open90 - open85) + I(admn90 - admn85), data = traffic1)
```

```
##
## Call:
## lm(formula = dthrte90 - dthrte85 ~ I(open90 - open85) + I(admn90 -
##     admn85), data = traffic1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.2526 -0.1434 -0.0032  0.1968  0.7968
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -0.4968     0.0524   -9.48  1.4e-12 ***
## I(open90 - open85)  -0.4197     0.2056   -2.04    0.047 *
## I(admn90 - admn85)  -0.1506     0.1168   -1.29    0.204
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.344 on 48 degrees of freedom
## Multiple R-squared:  0.119,  Adjusted R-squared:  0.0819
## F-statistic: 3.23 on 2 and 48 DF,  p-value: 0.0482
```

What was the impact of changes in the open container law on traffic fatalities?

- Reduction in fatality for states that did not change the law?
- Reduction in fatality for states that changed the law?
- Reliability of the results?

Takeaways: Impacts of open contain law

- In 1985, there were an average of 2.7 deaths per 100 million miles driven
- For states that did not change their open container law:
  - 18.4% reduction in fatalities between 1985 and 1990 $^{(2.7-0.497)}/_{2.7}$
- For states that enacted open container law: 15.5% additional reduction in fatalities between 1985 and 1990 $^{(2.7-0.42)}/_{2.7}$

First Difference Model via plm

```r
require(plm)

plm(dthrte ~ d2+ open  + admn, index = c("state", "year"),
    method = "fd", data = traff) |> summary()
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = dthrte ~ d2 + open + admn, data = traff, inde
##     "year"), method = "fd")
##
## Balanced Panel: n = 51, T = 2, N = 102
##
## Residuals:
##    Min. 1st Qu.  Median 3rd Qu.    Max.
## -0.6263 -0.0984  0.0000  0.0984  0.6263
##
```

FD Model Assumption

- Estimate true model
- Random sample from each cross-section
- Variation in each independent variable and no perfect collinearity
- $E[\varepsilon_{it}|x_{i1},\ldots,x_{iK},a_i]=0$ and $E[\triangle\varepsilon_{it}|\triangle x_{i1},\ldots,\triangle x_{iK},a_i]=0$ for all t
- Homoskedasticity: $var(\triangle\varepsilon_{it}|\triangle x_{i1},\ldots,\triangle x_{iK})=\sigma^2 \quad \forall t$
- No autocorrelation: $cov(\triangle\varepsilon_{it},\varepsilon_{is}|\triangle x_{i1},\ldots,\triangle x_{iK})=0$ for all $t\neq s$
- Normality of differenced errors

Unbiased with 1-4

Section 4

Fixed Effects (FE) Estimator

Fixed Effect (FE) Estimator

$$y_{it} = \beta_0 + \beta_1 x_{i1t} + \cdots + \beta_K x_{iKt} + a_i + \varepsilon_{it}$$

Take the average over time:

$$\bar{y}_i = \beta_0 + \beta_1 \bar{x}_{i1} + \cdots + \beta_K \bar{x}_{iK} + a_i + \bar{\varepsilon}_i$$

Subtract out the average:

$$y_{it} - \bar{y}_i = (\beta_0 - \beta_0) + \beta_1(x_{i1t} - \bar{x}_{i1}) + \cdots + \beta_K(x_{ikt} - \bar{x}_{iK}) + (a_i - a_i) + (\varepsilon_{it} - \bar{\varepsilon}_i)$$

The fixed effects (FE) estimator is then:

$$\widetilde{y}_{it} = \beta_1 \widetilde{x}_{i1t} + \cdots + \beta_K \widetilde{x}_{ikt} + \widetilde{\varepsilon}_{it}$$

Fixed Effects (FE) Estimator

- By subtracting the average over time, we are "time demeaning" the data.
    - Averages are computed for **each individual** over time, not across all individuals (not across the full sample).
    - Like the FD, the unobserved effect, a_i, is eliminated.
    - This transformation is also called the **within transformation**.
- The FE is similar to the FD except that the constant term drops out.
- Estimate the FE by pooling data and using OLS.
- dof: $N(T-1)-k$ where N is the number of individuals and T is the number of time periods.
    - We lose one observation in **EACH** cross-section due to time demeaning.

FE Assumptions

1. Same as FD assumption 1
2. Same as FD assumption 2
3. Same as FD assumption 3
4. $E[\varepsilon_{it} \mid x_{i1}, \ldots, x_{ik}, a_i] = 0 \quad \forall t$

Assumptions 1 – 4: Estimators are unbiased.

5. Homoskedasticity: $var(\varepsilon_{it} \mid x_{i1}, \ldots, x_{iK}) = \sigma^2 \quad \forall t$
6. No autocorrelation: $cov(\varepsilon_{it}, \varepsilon_{is} \mid x_{i1}, \ldots, x_{iK}) = 0 \quad \forall t \neq s$

Assumptions 1 – 6: Estimators are BLUE

7. Normality of errors, $\varepsilon_{it} \overset{iid}{\sim} \mathcal{N}$

- Both FD and FE are unbiased under the same assumptions.
- The difference is in assumption 6, where FE is always more efficient than FD.

FE with Dummies

- Instead of "demeaning" the data, we could use intercept-shifting dummies for each "individual" in the panel.

Recall in the regression model, the intercept is:

$$\hat{\beta}_0 = \bar{y} - \sum_k \hat{\beta}_k \bar{x}_k$$

Then:

$$\widehat{\beta}_{0i} = \bar{y}_i - \sum_k \hat{\beta}_k \bar{x}_{ik} - \bar{a}_i - \bar{\varepsilon}_i$$

Then:

$$y_{it} = \beta_0 + \sum_k \beta_k x_{ikt} + \nu_{it} + \beta_{0i}$$

Is equivalent to:

$$y_{it} - \bar{y}_i = (\beta_0 + \beta_{0i}) + \sum_k \beta_k (x_{ikt} - \bar{x}_{ik}) + (\varepsilon_{it} - \bar{\varepsilon}_i)$$

FE with Dummies

- Thus, we could rewrite the FE estimator as:

$$y_{it} = \delta_i d_i + \sum_k \beta_k x_{ikt} + (\varepsilon_{it} - \bar{\varepsilon}_i)$$

where $d_i = 1$ for individual i and zero otherwise. This adds N dummy variables to the model (or have a general model intercept with $N - 1$ dummies).

- May prefer this approach for functional form reasons or because we have endogenous sample selection.

Elimination of Time Constant Variables

- In both FD and FE approaches, the time-constant independent variables are eliminated.
- What if we are interested in those time-constant variables?

Solution

- Interact with time dummy variables.

Example:

$$y_{it} = \beta_0 + \beta_1 x_{i1t} + \beta_2 x_{i1t} \cdot D_{2t} + \beta_3 x_{i1t} \cdot D_{3t} + \theta_i + \varepsilon_{it}$$

- FE Model:

$$(y_{it} - \bar{y}_i) = \beta_2 \Big[x_{i1t} D_{2t} - \bar{x}_{i1} D_{2t} \Big] + \beta_3 \Big[x_{i1t} D_{3t} - \bar{x}_{i1} D_{3t} \Big] + (\varepsilon_{it} - \bar{\varepsilon}_i)$$

Worker Training Example

We want to evaluate the effect of a Michigan job training program on worker productivity.

Variables:

- `scrap` = number of items, per 100, that must be scrapped due to defect
- `grant` = 1 if firm received job training grant
- `d88` and `d89` are year dummy variables

Original model is:

$$y_{it} = \beta_0 + \beta_1 d88 + \beta_2 d89 + \beta_3 grant_{it} + \beta_4 grant_{it-1} + \theta_i + \epsilon_{it}$$

```r
# Compute fcode specific means

m.jtrain <- jtrain |>
  group_by(fcode) |>
  mutate(
    dm.d88 = d88 - mean(d88),
    dm.d89 = d89 - mean(d89),
    dm.grant = grant - mean(grant),
    dm.grant_1 = grant_1 - mean(grant_1),
    dm.scrap = lscrap - mean(lscrap)
  )

(grant <- lm(dm.scrap ~ -1 + dm.d88 + dm.d89 + dm.grant + dm.grant_1,
             data = m.jtrain) )|> summary() |> coef()
```

```
##             Estimate Std. Error t value Pr(>|t|)
## dm.d88       -0.0802     0.0888  -0.903   0.3678
## dm.d89       -0.2472     0.1081  -2.287   0.0235
## dm.grant     -0.2523     0.1222  -2.065   0.0406
## dm.grant_1  -0.4216     0.1705  -2.472   0.0145
```

```r
# via `plm`
(grant.fe <- plm(lscrap ~ d88 + d89 + grant + grant_1,
  data = jtrain, index = c("fcode", "year"),
  model = "within"))|> summary() |> coef()
```

```
##          Estimate Std. Error t-value Pr(>|t|)
## d88       -0.0802      0.109  -0.733   0.4654
## d89       -0.2472      0.133  -1.856   0.0663
## grant     -0.2523      0.151  -1.675   0.0969
## grant_1   -0.4216      0.210  -2.006   0.0475
```

Key Takeaways

1. What is the impact of the grant in the first year?
   - There is a 22.3% reduction in the number of defective products.
2. What is the impact of the grant in the second year?
   - There is a 34.4% reduction in the number of defective products.
   - Does it make sense that the second year effect is larger than the first year effect?
3. How do we interpret the 1989 dummy coefficient?
   - The average scrap rate in 1989 was 21.902% lower across all firms than it was in the base year of 1987.
   - If we had not controlled for the 1988 dummy, we could incorrectly attribute some of this overall reduction to the job training program.

Clustered Standard Errors

- The FE estimator violates Assumption 5 of the MLR because error terms are correlated over time.
  - $Cov(\varepsilon_{it}, \varepsilon_{is}) \neq 0$ for $t \neq s$.
  - Intuitively, ε_{it} contains time-varying unobserved factors that impact y and are correlated over time.
  - Not correcting for this correlation leads to biased (and artificially lower) standard errors.
- Instead, compute and report the HAC-standard errors (clustered s.e.) that are robust to arbitrary correlation within clusters (entities), robust to heteroskedasticity, and assume no correlation across entities.

Clustered Standard Errors

- Some common methods:
  - "white1" - for general heteroskedasticity but no serial correlation. *Recommended for random effects.*
  - "white2" - is "white1" restricted to a common variance within groups. *Recommended for random effects.*
  - "arellano" - both heteroskedasticity and serial correlation. Recommended for fixed effects.

```r
coeftest(grant.fe, vcovHC(grant.fe, type = "HC3"))
```

```
##
## t test of coefficients:
##
##         Estimate Std. Error t value Pr(>|t|)
## d88      -0.0802     0.0978   -0.82    0.414
## d89      -0.2472     0.1968   -1.26    0.212
## grant    -0.2523     0.1438   -1.75    0.082 .
## grant_1  -0.4216     0.2837   -1.49    0.140
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table 7: LSDV Model

	lscrap	
	panel linear	*OLS*
	FE	LSDV
	(1)	(2)
d88	−0.080	−0.080
	(0.109)	(0.109)
d89	−0.247*	−0.247*
	(0.133)	(0.133)
grant	−0.252*	−0.252*
	(0.151)	(0.151)
grant_1	−0.422**	−0.422**
	(0.210)	(0.210)
factor(fcode)410538		3.900***
		(0.406)
factor(fcode)410563		4.720***
		(0.406)
factor(fcode)410565		4.440***
		(0.406)
factor(fcode)410566		4.620***
		(0.406)
factor(fcode)410567		2.280***
		(0.406)
factor(fcode)410577		3.420***

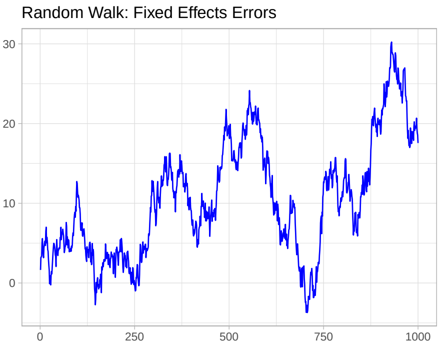
```r
grant.dv <- lm(lscrap ~ d88 + d89 +  grant +  grant_1 + as.fac
   data = jtrain)

stargazer(grant.fe, grant.dv, font.size = "scriptsize",
   title = "LSDV Model",
   dep.var.caption = "",
   column.labels = c("FE", "LSDV"),
   omit.stat = c("f", "ser"),
   no.space = TRUE,
   digits.extra = 3
   )
```

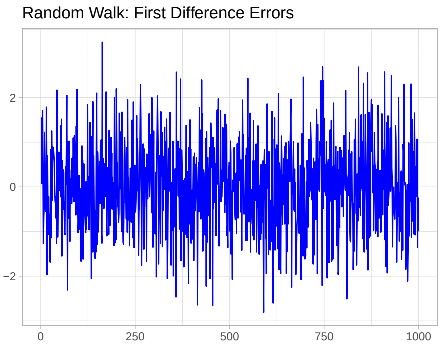
Section 5

Random Effects (RE) Model

Impact of Positive Autocorrelation

- Consider a random walk process: $\varepsilon_{it} = \varepsilon_{it-1} + r_t, \quad r_t \sim (0, \sigma_r)$. Here r_t has a mean of zero and constant variance.





Takeaways

- With random walk process:
  - Demeaning the data does NOT remain the autocorrelation
  - However, first-differencing DOES remove the autocorrelation

- Recall the FE model:

$$y_{it} = \beta_0 + \beta_1 x_{i1t} + \cdots + \beta_K x_{iKt} + \underbrace{a_i + \varepsilon_{it}}_{u_{it}}$$

- In the FD and FE models, we eliminate the unobserved effect, a_i, by differencing or demeaning the data.
- Suppose that a_i is actually uncorrelated with the independent variables, x_{it}.

$$cov(x_{it}, a_i) = 0 \implies E[u_{it} \mid x_{iKt}] = 0; \quad \forall i, t, k$$

Random Effects (RE) Model

If a_i is uncorrelated with $\mathbf{X}$, then why not just use pooled OLS? - The residuals are correlated over time:

$$corr(u_{it}, u_{is}) = \frac{cov(u_{it}, u_{is})}{\sqrt{var(u_{it})}\sqrt{var(u_{is})}} \neq 0$$

for $t \neq s$.

- This renders the pooled OLS inefficient.
- We can remove this autocorrelation. Derivation of this functional form is beyond the scope of this course.

Random Effects (RE) Estimator

$$\lambda = 1 - \sqrt{\frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + T a_i^2}}$$

The transformed model become:

$$y_{it} - \lambda \bar{y}_i = \beta_0 + \beta_1(x_{i1t} - \lambda \bar{x}_{i1}) + \cdots + \beta_K(x_{ikt} - \lambda \bar{x}_{ik}) + (u_{it} - \lambda \bar{u}_i)$$

- This looks like the FE estimator except that we only subtract a fraction (λ) of the averages over time.
- In the RE model, therefore, we are **"quasi-demeaning"** the data.
- Unlike the FE model, the time invariant variables **are not eliminated** since $0 \leq \lambda \leq 1$.
  - If $\lambda = 0$, then the RE model is equivalent to the pooled OLS model.
  - If $\lambda = 1$, then the RE model is equivalent to the FE model.

```r
(grant.re <- plm(lscrap ~ d88 + d89 + grant + grant_1,
  data = jtrain, index = c("fcode", "year"), model = "random")) |> summary()
```

```
## Oneway (individual) effect Random Effect Model
##    (Swamy-Arora's transformation)
##
## Call:
## plm(formula = lscrap ~ d88 + d89 + grant + grant_1, data = jtrain,
##     model = "random", index = c("fcode", "year"))
##
## Balanced Panel: n = 54, T = 3, N = 162
##
## Effects:
##                 var std.dev share
## idiosyncratic 0.248   0.498  0.11
## individual    1.983   1.408  0.89
## theta: 0.8
##
## Residuals:
##    Min. 1st Qu.  Median 3rd Qu.    Max.
## -2.5760 -0.2026  0.0164  0.2403  1.6207
##
## Coefficients:
##             Estimate Std. Error z-value Pr(>|z|)
## (Intercept)   0.5974     0.2033    2.94   0.0033 **
## d88          -0.0935     0.1090   -0.86   0.3907
## d89          -0.2714     0.1315   -2.06   0.0390 *
## grant        -0.2144     0.1476   -1.45   0.1463
## grant_1      -0.3729     0.2051   -1.82   0.0690 .
```

RE vs FE Model

```
stargazer(grant.fe, grant.re, font.size = "scriptsize",
  title = "RE vs FE Model", dep.var.caption = "",
  column.labels = c("FE", "RE"), omit.stat = c("f", "ser"),
  no.space = TRUE, digits.extra = 3, header = FALSE
)
```

Table 8: RE vs FE Model

	lscrap	
	FE	RE
	(1)	(2)
d88	−0.080	−0.094
	(0.109)	(0.109)
d89	−0.247*	−0.271**
	(0.133)	(0.131)
grant	−0.252*	−0.214
	(0.151)	(0.148)
grant_1	−0.422**	−0.373*
	(0.210)	(0.205)
Constant		0.597***
		(0.203)
Observations	162	162
R^2	0.201	0.139

Choosing between RE and FE Model

- We can test whether the RE or FE model is appropriate using the **Hausman test**.

Logics

- The RE model assumes that the unobserved effect is uncorrelated with the independent variables.
- If $\boxed{cov(\mathbf{X}, a_i) \neq 0}$, then the FE is preferred.
- If $\boxed{cov(\mathbf{X}, a_i) = 0}$, then the RE is preferred as it is more efficient.

Estimator	Appropriate when
RE	$cov(\mathbf{X}, a_i) = 0$
FE	$cov(\mathbf{X}, a_i) \neq 0$ $cov(\varepsilon_{it}, \varepsilon_{is}) = 0$
FD	$cov(\mathbf{X}, a_i) \neq 0$ $cov(\varepsilon_{it}, \varepsilon_{is}) > 0$

Hausman Test Procedure

1. Estimate and Store model using FE (`fe.model`)
2. Estimate and Store model using RE (`re.model`)
3. Perform the Hausman test:
   - `phtest(fe.model, re.model)`
   - Null hypothesis: $cov(\mathbf{X}, a_i) = 0$ (RE is appropriate)
   - Alternative hypothesis: FE is appropriate

Worker Training Example

- Conclusion??

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

```
##
##  Hausman Test
##
## data:  lscrap ~ d88 + d89 + grant + grant_1
## chisq = 2, df = 4, p-value = 0.7
## alternative hypothesis: one model is inconsistent
```