Lecture 18: Panel Data and Fixed Effects

ECON 480 - Econometrics - Fall 2018

Ryan Safner

December 3, 2018

Panel Data

Fixed Effects Model

Two-Way Fixed Effects

Panel Data

- Panel or longitudinal data contains a time-series for *each* cross-sectional unit

Panel Data

- Panel or longitudinal data contains a time-series for *each* cross-sectional unit
    - i.e. tracking multiple entities (individuals, firms, cities, states, countries) over time

- Panel or longitudinal data contains a time-series for *each* cross-sectional unit
    - i.e. tracking multiple entities (individuals, firms, cities, states, countries) over time

Example
Do cell phones cause more traffic fatalities?

- Don't have a measure of cell phones *used* while driving

- Panel or longitudinal data contains a time-series for *each* cross-sectional unit
  - i.e. tracking multiple entities (individuals, firms, cities, states, countries) over time

Example
Do cell phones cause more traffic fatalities?

- Don't have a measure of cell phones *used* while driving
- Data on number of cell phone plans Per 10,000 people to **proxy** for cell phone use while driving (probably strongly positively correlated)

- Panel or longitudinal data contains a time-series for *each* cross-sectional unit
  - i.e. tracking multiple entities (individuals, firms, cities, states, countries) over time

Example
Do cell phones cause more traffic fatalities?

- Don't have a measure of cell phones *used* while driving
- Data on number of cell phone plans Per 10,000 people to **proxy** for cell phone use while driving (probably strongly positively correlated)
- **State-level data** over 6 years

```r
cell_deaths<-read.csv("../Data/cellphones.csv") # download data
str(cell_deaths) # look at structure of data
```

```
## 'data.frame':    306 obs. of  9 variables:
##  $ X                    : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ year                 : int  2007 2007 2007 2007 2007 2007 2007 2007 2007 2007 ...
##  $ state                : Factor w/ 51 levels "Alabama","Alaska",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ state_numeric        : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ urban_percent        : int  30 55 45 21 54 34 84 31 100 53 ...
##  $ cell_per10thous_pop  : num  8136 6730 7572 8071 8822 ...
##  $ cell_ban             : int  0 0 0 0 0 0 1 0 1 0 ...
##  $ text_ban             : int  0 0 0 0 0 0 1 0 1 0 ...
##  $ DeathsPerBillionMiles: num  18.1 16.3 16.9 19.6 12.1 ...
```

HOOD COLLEGE

Panel Data: Example (After Some Tidying)

- What the `data.frame` looks like (after some tidying - see code in `.Rmd`)

- What the `data.frame` looks like (after some tidying - see code in `.Rmd`)

year	state	cell_plans	deaths	urban_pct	cell_ban	text_ban
2007	Alabama	8135.525	18.07523	30	0	0
2008	Alabama	8494.391	16.28923	31	0	0
2009	Alabama	8979.108	13.83368	31	0	0
2010	Alabama	9054.894	13.43408	35	0	0
2011	Alabama	9340.501	13.77199	39	0	0
2012	Alabama	9433.800	13.31606	35	0	0
2007	Alaska	6730.282	16.30118	55	0	0
2008	Alaska	5580.707	12.74409	21	0	0
2009	Alaska	8389.730	12.97385	42	0	1
2010	Alaska	8560.595	11.67089	42	0	1
2011	Alaska	8772.439	15.67572	30	0	1
2012	Alaska	8872.799	12.31198	21	0	1

Panel Data: Example

```r
table(cell_deaths_tidy$year) # make table of counts of each year
```

```
##
## 2007 2008 2009 2010 2011 2012
##   51   51   51   51   51   51
```

```r
table(cell_deaths_tidy$state) # make table of counts of each state
```

```
##
##              Alabama               Alaska              Arizona
##                    6                    6                    6
##             Arkansas           California             Colorado
##                    6                    6                    6
##          Connecticut             Delaware District of Columbia
##                    6                    6                    6
##              Florida              Georgia               Hawaii
##                    6                    6                    6
##                Idaho             Illinois              Indiana
##                    6                    6                    6
##                 Iowa               Kansas             Kentucky
```

Panel Data: Example II

```
library("plm") # package for panel data regression models
pdim(cell_deaths_tidy, index=c("state","year")) # check N groups and T periods
```

```
## Balanced Panel: n = 51, T = 6, N = 306
```

Pooled Regression

$$\widehat{Y_{it}} = \beta_0 + \beta_1 X_{it} + \epsilon_{it}$$

- What if we just ran a standard regression model

$$\widehat{Y_{it}} = \beta_0 + \beta_1 X_{it} + \epsilon_{it}$$

- What if we just ran a standard regression model
    - N number of i groups (e.g. U.S. States)

Pooled Regression

$$\widehat{Y}_{it} = \beta_0 + \beta_1 X_{it} + \epsilon_{it}$$

- What if we just ran a standard regression model
  - N number of i groups (e.g. U.S. States)
  - T number of t periods (e.g. years)

$$\widehat{Y_{it}} = \beta_0 + \beta_1 X_{it} + \epsilon_{it}$$

- What if we just ran a standard regression model
    - N number of i groups (e.g. U.S. States)
    - T number of t periods (e.g. years)
- This is a pooled regression model: treats all observations as independent

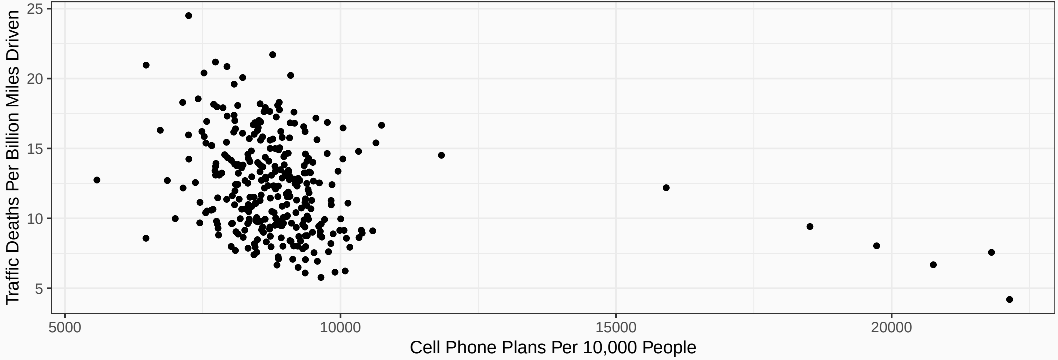
HOOD COLLEGE

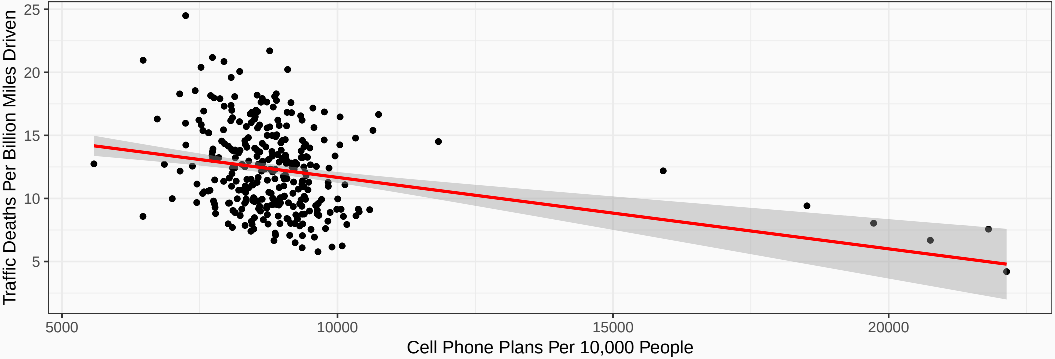
Pooled Regression

```
pooled<-lm(deaths~cell_plans, data=cell_deaths_tidy)
summary(pooled)
```

```
##
## Call:
## lm(formula = deaths ~ cell_plans, data = cell_deaths_tidy)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.0951 -2.6411 -0.2893  2.2755 11.2665
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 17.3371034  0.9753845  17.775  < 2e-16 ***
## cell_plans  -0.0005666  0.0001070  -5.297 2.26e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.279 on 304 degrees of freedom
## Multiple R-squared:  0.0845, Adjusted R-squared:  0.08148
```

Panel Regression II




Traffic Deaths Per Billion Miles Driven
25
20
15
10
5
5000
10000
15000
20000
Cell Phone Plans Per 10,000 People

$$\widehat{Y_{it}} = \beta_0 + \beta_1 X_{it} + \epsilon_{it}$$

- Pooled regression model is **biased** because it ignores:

$$\widehat{Y_{it}} = \beta_0 + \beta_1 X_{it} + \epsilon_{it}$$

- Pooled regression model is **biased** because it ignores:
    - Multiple observations come from the same group i (State)

Biases of Pooled Regression

$$\widehat{Y_{it}} = \beta_0 + \beta_1 X_{it} + \epsilon_{it}$$

- Pooled regression model is **biased** because it ignores:
    - Multiple observations come from the same group i (State)
    - Multiple observations come from the same time t (year)

Biases of Pooled Regression

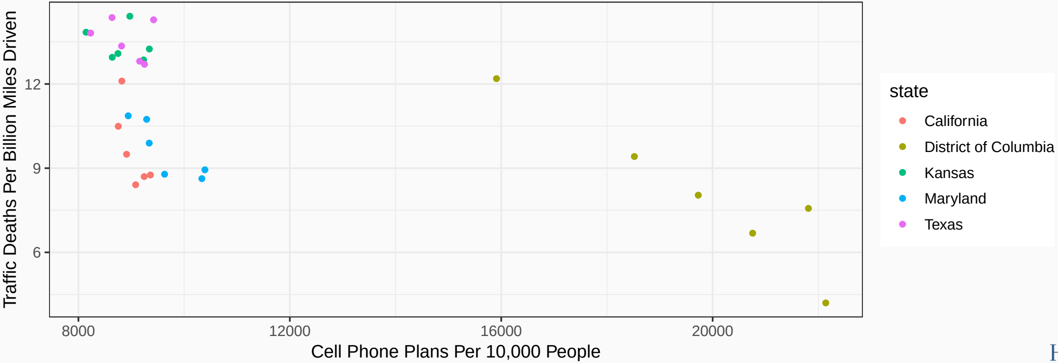
$$\widehat{Y_{it}} = \beta_0 + \beta_1 X_{it} + \epsilon_{it}$$

- Pooled regression model is **biased** because it ignores:
  - Multiple observations come from the same group i (State)
  - Multiple observations come from the same time t (year)
- Error terms ϵ_{it} of each observation will be **serially correlated** with error terms of other observations (e.g. from same group)

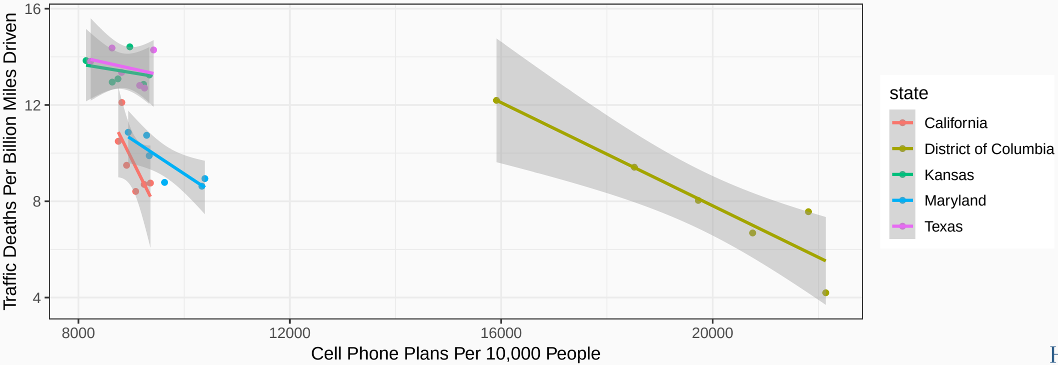
- Example: look only at 5 states

Panel Data: The Problem of Pooling

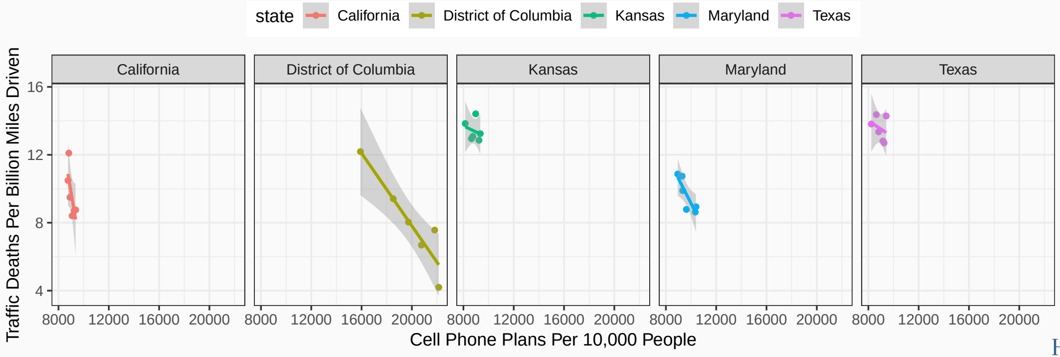
- Example: look only at 5 states



- Example: look only at 5 states

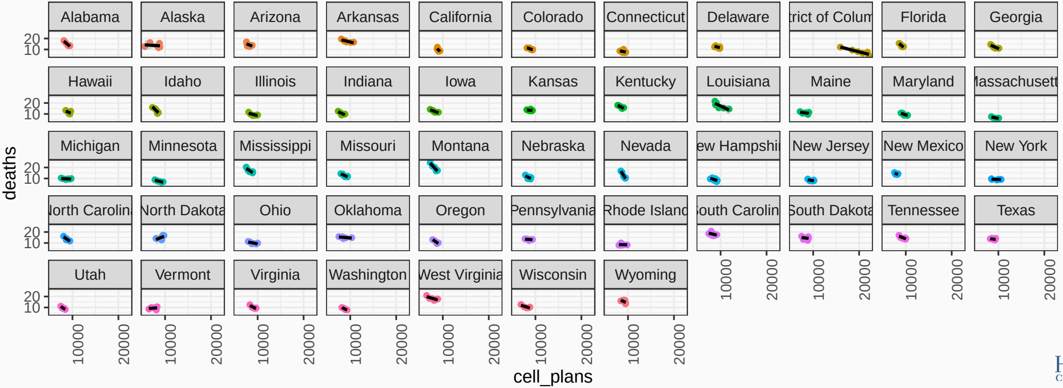


- Example: look only at 5 states



- Remember, we actually have 51 states (including D.C.)...

- Remember, we actually have 51 states (including D.C.)...



Example

$$\widehat{\text{Deaths}_{it}} = \beta_0 + \beta_1 \text{Cell Phones}_{it} + \epsilon_{it}$$

- Cell Phones_{it} is *endogenous*:

Example

$$\widehat{\text{Deaths}_{it}} = \beta_0 + \beta_1 \text{Cell Phones}_{it} + \epsilon_{it}$$

- Cell Phones_{it} is *endogenous*:

$$corr(\epsilon_{it}, X_{it}) \neq 0 \qquad\qquad E[\epsilon_{it}|X_{it}] \neq 0$$

Example

$$\widehat{\text{Deaths}_{it}} = \beta_0 + \beta_1 \text{Cell Phones}_{it} + \epsilon_{it}$$

- Cell Phones_{it} is *endogenous*:

$$corr(\epsilon_{it}, X_{it}) \neq 0 \qquad\qquad E[\epsilon_{it}|X_{it}] \neq 0$$

- What's in ϵ_{it} that is correlated with Cell Phones?
    - Infrastructure spending, population, urban vs. rural, more/less cautious citizens, cultural attitudes towards driving, texting, etc

Example

$$\widehat{\text{Deaths}_{it}} = \beta_0 + \beta_1 \text{Cell Phones}_{it} + \epsilon_{it}$$

- Cell Phones_{it} is *endogenous*:

$$corr(\epsilon_{it}, X_{it}) \neq 0 \qquad\qquad E[\epsilon_{it}|X_{it}] \neq 0$$

- What's in ϵ_{it} that is correlated with Cell Phones?
  - Infrastructure spending, population, urban vs. rural, more/less cautious citizens, cultural attitudes towards driving, texting, etc
  - A lot of things vary systematically by state!

Example

$$\widehat{\text{Deaths}_{it}} = \beta_0 + \beta_1 \text{Cell Phones}_{it} + \epsilon_{it}$$

- Cell Phones_{it} is *endogenous*:

$$corr(\epsilon_{it}, X_{it}) \neq 0 \qquad E[\epsilon_{it} | X_{it}] \neq 0$$

- What's in ϵ_{it} that is correlated with Cell Phones?
  - Infrastructure spending, population, urban vs. rural, more/less cautious citizens, cultural attitudes towards driving, texting, etc
  - A lot of things vary systematically by state!
  - ϵ_{it_1} for state i in year t_1 correlates with ϵ_{it_2} for state i in different year t_2

Example

$$\widehat{\text{Test Scores}}_{it} = \beta_0 + \beta_1 \text{Private School}_{it} + \epsilon_{it}$$

- What's in ϵ_{it}?

Example

$$\widehat{\text{Test Scores}}_{it} = \beta_0 + \beta_1 \text{Private School}_{it} + \epsilon_{it}$$

- What's in ϵ_{it}?
- **Unit-specific** (i.e. individual person) factors stable over time correlated with Private school?

Example

$$\widehat{\text{Test Scores}}_{it} = \beta_0 + \beta_1 \text{Private School}_{it} + \epsilon_{it}$$

- What's in ϵ_{it}?
- **Unit-specific** (i.e. individual person) factors stable over time correlated with Private school?
    - Gender?

Example

$$\widehat{\text{Test Scores}_{it}} = \beta_0 + \beta_1 \text{Private School}_{it} + \epsilon_{it}$$

- What's in ϵ_{it}?
- **Unit-specific** (i.e. individual person) factors stable over time correlated with Private school?
  - Gender?
  - Whether a student was sick the day of the test?

Example

$$\widehat{\text{Test Scores}_{it}} = \beta_0 + \beta_1 \text{Private School}_{it} + \epsilon_{it}$$

- What's in ϵ_{it}?
- **Unit-specific** (i.e. individual person) factors stable over time correlated with Private school?
    - Gender?
    - Whether a student was sick the day of the test?
    - Parent's education level?

Example

$$\widehat{\text{Test Scores}_{it}} = \beta_0 + \beta_1 \text{Private School}_{it} + \epsilon_{it}$$

- What's in ϵ_{it}?
- **Unit-specific** (i.e. individual person) factors stable over time correlated with Private school?
    - Gender?
    - Whether a student was sick the day of the test?
    - Parent's education level?
- Unobservables:

Example

$$\widehat{\text{Test Scores}_{it}} = \beta_0 + \beta_1 \text{Private School}_{it} + \epsilon_{it}$$

- What's in ϵ_{it}?
- **Unit-specific** (i.e. individual person) factors stable over time correlated with Private school?
  - Gender?
  - Whether a student was sick the day of the test?
  - Parent's education level?
- Unobservables:
  - Ability, culture, etc.

Example

$$\widehat{\text{Test Scores}}_{it} = \beta_0 + \beta_1 \text{Private School}_{it} + \epsilon_{it}$$

- What's in ϵ_{it}?
- **Unit-specific** (i.e. individual person) factors stable over time correlated with Private school?
    - Gender?
    - Whether a student was sick the day of the test?
    - Parent's education level?
- Unobservables:
    - Ability, culture, etc.
- ϵ_{it} of observations on same person i (in different years) are correlated

Fixed Effects Model

Fixed Effects Model: Decomposing ϵ_{it}

- Much of the endogeneity in X_{it} can be explained by systematic differences between i-groups

- Much of the endogeneity in X_{it} can be explained by systematic differences between i-groups
- Exploit the systematic variation across groups with a fixed effects model

Fixed Effects Model: Decomposing ϵ_{it}

- Much of the endogeneity in X_{it} can be explained by systematic differences between i-groups
- Exploit the systematic variation across groups with a fixed effects model
- Decompose the error term:

$$\epsilon_{it} = \alpha_i + \nu_{it}$$

Fixed Effects Model: α_i

$$\epsilon_{it} = \alpha_i + \nu_{it}$$

- α_i are group-specific fixed effects

$$\epsilon_{it} = \alpha_i + \nu_{it}$$

- α_i are group-specific fixed effects
  - Group i (e.g. Maryland) tends to have higher or lower Y than other groups (e.g. Texas) given regressor(s) X_{it}

Fixed Effects Model: α_i

$$\epsilon_{it} = \alpha_i + \nu_{it}$$

- α_i are group-specific fixed effects
    - Group i (e.g. Maryland) tends to have higher or lower Y than other groups (e.g. Texas) given regressor(s) X_{it}
    - A separate α for every group i

Fixed Effects Model: α_i

$$\epsilon_{it} = \alpha_i + \nu_{it}$$

- α_i are group-specific fixed effects
    - Group i (e.g. Maryland) tends to have higher or lower Y than other groups (e.g. Texas) given regressor(s) X_{it}
    - A separate α for every group i
    - Includes **all** factors that do not change *within* group i over time

Fixed Effects Model: α_i

$$\epsilon_{it} = \alpha_i + \nu_{it}$$

- α_i are group-specific fixed effects
    - Group i (e.g. Maryland) tends to have higher or lower Y than other groups (e.g. Texas) given regressor(s) X_{it}
    - A separate α for every group i
    - Includes **all** factors that do not change *within* group i over time
    - i.e. **all** group-wide factors that vary *across* groups

Fixed Effects Model: ν_{it}

$$\epsilon_{it} = \alpha_i + \nu_{it}$$

- ν_{it} is the remaining random error[1]

[1] Rewritten as ν instead of ϵ only because it's different!

$$\epsilon_{it} = \alpha_i + \nu_{it}$$

- ν_{it} is the remaining random error[1]
    - Like usual ϵ_{it} in OLS, we assume it is random with mean $E[\nu_{it}] = 0$ and constant variance σ^2_{ν}

[1] Rewritten as ν instead of ϵ only because it's different!

Fixed Effects Model: ν_{it}

$$\epsilon_{it} = \alpha_i + \nu_{it}$$

- ν_{it} is the remaining random error[1]
    - Like usual ϵ_{it} in OLS, we assume it is random with mean $E[\nu_{it}] = 0$ and constant variance σ^2_ν
    - Includes all other factors affecting Y_{it} **not** specific to group fixed effects (e.g. differences *within* each group that *change* over time)

[1] Rewritten as ν instead of ϵ only because it's different!

$$\epsilon_{it} = \alpha_i + \nu_{it}$$

- ν_{it} is the remaining random error[1]
    - Like usual ϵ_{it} in OLS, we assume it is random with mean $E[\nu_{it}] = 0$ and constant variance σ^2_ν
    - Includes all other factors affecting Y_{it} **not** specific to group fixed effects (e.g. differences *within* each group that *change* over time)
    - Can still have endogeneity if non-group-specific factors correlated with X_{it}!

[1] Rewritten as ν instead of ϵ only because it's different!

Fixed Effects Model: New Regression Equation

$$\widehat{Y_{it}} = \beta_0 + \beta_1 X_{it} + \alpha_i + \nu_{it}$$

- Pull α_i out of the error term and include it in the regression

Fixed Effects Model: New Regression Equation

$$\widehat{Y_{it}} = \beta_0 + \beta_1 X_{it} + \alpha_i + \nu_{it}$$

- Pull α_i out of the error term and include it in the regression
- Essentially we will estimate a different intercept for each group

Fixed Effects Model: New Regression Equation

$$\widehat{Y_{it}} = \beta_0 + \beta_1 X_{it} + \alpha_i + \nu_{it}$$

- Pull α_i out of the error term and include it in the regression
- Essentially we will estimate a different intercept for each group
- Must have multiple observations (over time) for each group (i.e. panel data)

Example

$$\widehat{\text{Deaths}}_{it} = \beta_0 + \beta_1 \text{Cell Phones}_{it} + \alpha_i + \nu_{it}$$

- α_i is the **state fixed effect**

Example

$$\widehat{\text{Deaths}}_{it} = \beta_0 + \beta_1 \text{Cell Phones}_{it} + \alpha_i + \nu_{it}$$

- α_i is the **state fixed effect**
  - Captures everything unique about state *i* that does not vary in time

Example

$$\widehat{\text{Deaths}}_{it} = \beta_0 + \beta_1 \text{Cell Phones}_{it} + \alpha_i + \nu_{it}$$

- α_i is the **state fixed effect**
  - Captures everything unique about state *i* that does not vary in time
  - Includes **all stable factors** we could never measure or think of: culture, institutions, history, geography, climate, etc!

Example

$$\widehat{\text{Deaths}}_{it} = \beta_0 + \beta_1 \text{Cell Phones}_{it} + \alpha_i + \nu_{it}$$

- α_i is the **state fixed effect**
  - Captures everything unique about state *i* that does not vary in time
  - Includes **all stable factors** we could never measure or think of: culture, institutions, history, geography, climate, etc!
- There could *still* be factors in ν_{it} that are correlated with Cell Phones!

Example

$$\widehat{\text{Deaths}_{it}} = \beta_0 + \beta_1 \text{Cell Phones}_{it} + \alpha_i + \nu_{it}$$

- α_i is the **state fixed effect**
  - Captures everything unique about state *i* that does not vary in time
  - Includes **all stable factors** we could never measure or think of: culture, institutions, history, geography, climate, etc!
- There could *still* be factors in ν_{it} that are correlated with Cell Phones!
  - Maybe some states passed a cell phone ban while driving, but only during *some* years in our data (i.e. *not* identical across every observation we have for that state)

$$\widehat{Y_{it}} = \beta_0 + \beta_1 X_{it} + \alpha_i + \nu_{it}$$

- There are two ways we can estimate fixed effects models

Strategies to Implement Fixed Effects

$$\widehat{Y_{it}} = \beta_0 + \beta_1 X_{it} + \alpha_i + \nu_{it}$$

- There are two ways we can estimate fixed effects models
    1. Least Squares Dummy Variable (LSDV) Approach

$$\widehat{Y_{it}} = \beta_0 + \beta_1 X_{it} + \alpha_i + \nu_{it}$$

- There are two ways we can estimate fixed effects models
    1. Least Squares Dummy Variable (LSDV) Approach
    2. De-Meaned Data Approach

$$\widehat{Y_{it}} = \beta_0 + \beta_1 X_{it} + \beta_2 D_{1i} + \beta_3 D_{2i} + \cdots \beta_N D_{(N-1)i} + \nu_{it}$$

- A dummy variable D_i for every possible group (e.g. state) $= 1$ if observation *it* is from group *i*, else $= 0$

Least Squares Dummy Variable (LSDV) Approach

$$\widehat{Y_{it}} = \beta_0 + \beta_1 X_{it} + \beta_2 D_{1i} + \beta_3 D_{2i} + \cdots \beta_N D_{(N-1)i} + \nu_{it}$$

- A dummy variable D_i for every possible group (e.g. state) $= 1$ if observation *it* is from group *i*, else $= 0$
- If there are N groups:

Least Squares Dummy Variable (LSDV) Approach

$$\widehat{Y_{it}} = \beta_0 + \beta_1 X_{it} + \beta_2 D_{1i} + \beta_3 D_{2i} + \cdots \beta_N D_{(N-1)i} + \nu_{it}$$

- A dummy variable D_i for every possible group (e.g. state) $= 1$ if observation *it* is from group *i*, else $= 0$
- If there are N groups:
  - If we want to keep β_0: include $N-1$ dummies (to avoid **dummy variable trap**) and β_0 is the reference category

Least Squares Dummy Variable (LSDV) Approach

$$\widehat{Y_{it}} = \beta_0 + \beta_1 X_{it} + \beta_2 D_{1i} + \beta_3 D_{2i} + \cdots \beta_N D_{(N-1)i} + \nu_{it}$$

- A dummy variable D_i for every possible group (e.g. state) $= 1$ if observation *it* is from group i, else $= 0$
- If there are N groups:
  - If we want to keep β_0: include $N-1$ dummies (to avoid **dummy variable trap**) and β_0 is the reference category
  - If we drop β_0: include all N dummies

Least Squares Dummy Variable (LSDV) Approach

$$\widehat{Y}_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 D_{1i} + \beta_3 D_{2i} + \cdots \beta_N D_{(N-1)i} + \nu_{it}$$

- A dummy variable D_i for every possible group (e.g. state) $= 1$ if observation *it* is from group *i*, else $= 0$
- If there are N groups:
  - If we want to keep β_0: include $N - 1$ dummies (to avoid **dummy variable trap**) and β_0 is the reference category
  - If we drop β_0: include all N dummies
  - In either case, β_0 takes the place of one category-dummy

Least Squares Dummy Variable (LSDV) Approach

$$\widehat{Y_{it}} = \beta_0 + \beta_1 X_{it} + \beta_2 D_{1i} + \beta_3 D_{2i} + \cdots \beta_N D_{(N-1)i} + \nu_{it}$$

- A dummy variable D_i for every possible group (e.g. state) $= 1$ if observation *it* is from group *i*, else $= 0$
- If there are N groups:
    - If we want to keep β_0: include $N-1$ dummies (to avoid **dummy variable trap**) and β_0 is the reference category
    - If we drop β_0: include all N dummies
    - In either case, β_0 takes the place of one category-dummy
    - So we are estimating different intercepts for each group

Least Squares Dummy Variable (LSDV) Approach

$$\widehat{Y_{it}} = \beta_0 + \beta_1 X_{it} + \beta_2 D_{1i} + \beta_3 D_{2i} + \cdots \beta_N D_{(N-1)i} + \nu_{it}$$

- A dummy variable D_i for every possible group (e.g. state) $= 1$ if observation *it* is from group *i*, else $= 0$
- If there are N groups:
    - If we want to keep β_0: include $N-1$ dummies (to avoid **dummy variable trap**) and β_0 is the reference category
    - If we drop β_0: include all N dummies
    - In either case, β_0 takes the place of one category-dummy
    - So we are estimating different intercepts for each group
- Sounds like a lot of work, automatic in R

Least Squares Dummy Variable (LSDV) Approach

$$\widehat{Y_{it}} = \beta_0 + \beta_1 X_{it} + \beta_2 D_{1i} + \beta_3 D_{2i} + \cdots \beta_N D_{(N-1)i} + \nu_{it}$$

- A dummy variable D_i for every possible group (e.g. state) $= 1$ if observation *it* is from group i, else $= 0$
- If there are N groups:
  - If we want to keep β_0: include $N-1$ dummies (to avoid **dummy variable trap**) and β_0 is the reference category
  - If we drop β_0: include all N dummies
  - In either case, β_0 takes the place of one category-dummy
  - So we are estimating different intercepts for each group
- Sounds like a lot of work, automatic in R
- This soaks up *anything* in the error term fixed within groups over time!

Example

$$\widehat{\text{Deaths}}_{it} = \beta_1 \text{Cell Phones}_{it} + Alabama_i + Alaska_i + \cdots + Wyoming_i + \nu_{it}$$

- R is generous, if `cell_deaths_tidy$state` is a `factor` variable, just put it in regression:

- R is generous, if `cell_deaths_tidy$state` is a `factor` variable, just put it in regression:
  - R automatically makes all N state-dummies and adds them in a regression (and drops β_0)

Example in R

- R is generous, if `cell_deaths_tidy$state` is a `factor` variable, just put it in regression:
  - R automatically makes all N state-dummies and adds them in a regression (and drops β_0)

```r
lsdv.1<-lm(deaths~cell_plans+state, data = cell_deaths_tidy)
summary(lsdv.1)
```

```
##
## Call:
## lm(formula = deaths ~ cell_plans + state, data = cell_deaths_tidy)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5617 -0.6577 -0.1353  0.5997  3.7087
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)               25.5076799  1.0176400  25.066  < 2e-16 ***
## cell_plans                -0.0012037  0.0001013 -11.881  < 2e-16 ***
## stateAlaska               -2.4841648  0.6745076  -3.683 0.000282 ***
## stateArizona              -1.5105774  0.6704570  -2.253 0.025109 *
## stateArkansas              3.1926629  0.6664384   4.791 2.83e-06 ***
## stateCalifornia           -4.9786687  0.6655468  -7.481 1.21e-12 ***
## stateColorado             -4.3445535  0.6654735  -6.529 3.59e-10 ***
## stateConnecticut          -6.5951855  0.6654429  -9.911  < 2e-16 ***
## stateDelaware             -2.0983936  0.6666483  -3.148 0.001842 **
```

Aside: Doing it Manually

- To better understand what R is doing in a regression with fixed effects, we *could* make our own dummies and include all of them in the regression

- To better understand what R is doing in a regression with fixed effects, we *could* make our own dummies and include all of them in the regression
    - Easy when we have just a few categories

Aside: Doing it Manually

- To better understand what R is doing in a regression with fixed effects, we *could* make our own dummies and include all of them in the regression
  - Easy when we have just a few categories
  - Think back to how we use the `ifelse()` function to create `male` or `female`

Aside: Doing it Manually

- To better understand what R is doing in a regression with fixed effects, we *could* make our own dummies and include all of them in the regression
    - Easy when we have just a few categories
    - Think back to how we use the `ifelse()` function to create `male` or `female`

```
set.seed(1) # make a random data.frame of Sex variable with values "Male" or "Female"
ex<-data.frame(sex=sample(c("Male","Female"),5,replace=TRUE))

ex <- ex %>% # using dplyr, generate Male dummy and Female dummy
  mutate(male=ifelse(sex=="Male",1,0),
         female=ifelse(sex=="Female",1,0))

ex # look at data.frame
```

```
##      sex male female
## 1   Male    1      0
## 2   Male    1      0
## 3 Female    0      1
## 4 Female    0      1
## 5   Male    1      0
```

Aside: Doing it Manually

- Much more tedious if we have 51 different categories

Aside: Doing it Manually

- Much more tedious if we have 51 different categories
- Rule of computer science: DRY: Don't Repeat Yourself

Aside: Doing it Manually

- Much more tedious if we have 51 different categories
- Rule of computer science: **DRY**: **D**on't **R**epeat **Y**ourself
    - Want to avoid making mistakes, computers $>$ you at repetitive monotonous tasks

Aside: Doing it Manually

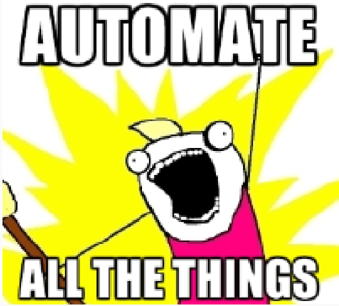
- Much more tedious if we have 51 different categories
- Rule of computer science: **DRY**: **D**on't **R**epeat **Y**ourself
  - Want to avoid making mistakes, computers $>$ you at repetitive monotonous tasks
  - In other words:

Aside: Doing it Manually

- Much more tedious if we have 51 different categories
- Rule of computer science: **DRY**: **D**on't **R**epeat **Y**ourself
    - Want to avoid making mistakes, computers $>$ you at repetitive monotonous tasks
    - In other words:

Aside: Doing it Manually

- Much more tedious if we have 51 different categories
- Rule of computer science: **DRY**: **D**on't **R**epeat **Y**ourself
  - Want to avoid making mistakes, computers $>$ you at repetitive monotonous tasks
  - In other words:



- A `for` loop in R can tackle a repetitive task, like making 51 separate dummy variables and assigning them a value based on the value of `state`.

- A `for` loop in R can tackle a repetitive task, like making 51 separate dummy variables and assigning them a value based on the value of `state`.

```
for (something in some.object){
  Do.this.thing
}
```

- A `for` loop in R can tackle a repetitive task, like making 51 separate dummy variables and assigning them a value based on the value of `state`.

```
for (something in some.object){
  Do.this.thing
}
```

- R will run through each individual `something` in `some.object`, and for each `something`, it will `Do.this.thing` to it

```r
# take the first 7 integers and square each one, and then print the result:
for (x in 1:7){
  print(x^2)
}
```

```
## [1] 1
## [1] 4
## [1] 9
## [1] 16
## [1] 25
## [1] 36
## [1] 49
```

```r
# First Line: take each unique value for state, for each value i:
# Second Line: create variable that is the name of the state, then populate it with 1 if obs is in state 1, 0
for(i in unique(cell_deaths_tidy$state)){
  cell_deaths_tidy[i] <- ifelse(cell_deaths_tidy$state == i, 1, 0)
}
```

Advanced R: A for Loop III

```r
# check and confirm it worked

head(cell_deaths_tidy[,1:10]) # look only at first 10 columns (to fit on slide)
```

```
##   year   state cell_plans   deaths urban_pct cell_ban text_ban Alabama
## 1 2007 Alabama   8135.525 18.07523        30        0        0       1
## 2 2008 Alabama   8494.391 16.28923        31        0        0       1
## 3 2009 Alabama   8979.108 13.83368        31        0        0       1
## 4 2010 Alabama   9054.894 13.43408        35        0        0       1
## 5 2011 Alabama   9340.501 13.77199        39        0        0       1
## 6 2012 Alabama   9433.800 13.31606        35        0        0       1
##   Alaska Arizona
## 1      0       0
## 2      0       0
## 3      0       0
## 4      0       0
## 5      0       0
## 6      0       0
```

- Alternatively, we can hold group fixed effects constant without directly estimating them

- Alternatively, we can hold group fixed effects constant without directly estimating them
- We simply **de-mean** the data for each group

- Alternatively, we can hold group fixed effects constant without directly estimating them
- We simply **de-mean** the data for each group
- For each group (state), find the mean

$$\bar{Y}_i = \beta_0 + \beta_1 \bar{X}_i + \bar{\alpha}_i + \bar{\nu}_i$$

- Alternatively, we can hold group fixed effects constant without directly estimating them
- We simply **de-mean** the data for each group
- For each group (state), find the mean

$$\bar{Y}_i = \beta_0 + \beta_1 \bar{X}_i + \bar{\alpha}_i + \bar{\nu}_i$$

  - $\bar{Y}_i$: average value of Y_{it} for group i

- Alternatively, we can hold group fixed effects constant without directly estimating them
- We simply **de-mean** the data for each group
- For each group (state), find the mean

$$\bar{Y}_i = \beta_0 + \beta_1 \bar{X}_i + \bar{\alpha}_i + \bar{\nu}_i$$

  - $\bar{Y}_i$: average value of Y_{it} for group i
  - $\bar{X}_i$: average value of X_{it} for group i

- Alternatively, we can hold group fixed effects constant without directly estimating them
- We simply **de-mean** the data for each group
- For each group (state), find the mean

$$\bar{Y}_i = \beta_0 + \beta_1 \bar{X}_i + \bar{\alpha}_i + \bar{\nu}_i$$

  - $\bar{Y}_i$: average value of Y_{it} for group i
  - $\bar{X}_i$: average value of X_{it} for group i
  - $\bar{\alpha}_i$: average value of α_i for group i $(= \alpha_i)$

- Alternatively, we can hold group fixed effects constant without directly estimating them
- We simply **de-mean** the data for each group
- For each group (state), find the mean

$$\bar{Y}_i = \beta_0 + \beta_1 \bar{X}_i + \bar{\alpha}_i + \bar{\nu}_i$$

  - $\bar{Y}_i$: average value of Y_{it} for group i
  - $\bar{X}_i$: average value of X_{it} for group i
  - $\bar{\alpha}_i$: average value of α_i for group i $(= \alpha_i)$
  - $\bar{\nu}_i = 0$, by assumption

De-meaned Approach II

- Subtracting the means equation from the pooled equation:

[2] Recall **Rule 4** from the **Handout** on the Summation Operator: $\sum(X_i - \bar{X}) = 0$

- Subtracting the means equation from the pooled equation:

$$Y_i - [\bar{Y}_i] = \beta_0 + \beta_1 X_{it} + \alpha_i + \nu_{it} - [\beta_0 + \beta_1 \bar{X}_i + \bar{\alpha}_i + \bar{\nu}_i]$$

[2] Recall **Rule 4** from the **Handout** on the Summation Operator: $\sum(X_i - \bar{X}) = 0$

- Subtracting the means equation from the pooled equation:

$$
\begin{aligned}
Y_i - [\bar{Y}_i] &= \beta_0 + \beta_1 X_{it} + \alpha_i + \nu_{it} - [\beta_0 + \beta_1 \bar{X}_i + \bar{\alpha}_i + \bar{\nu}_i] \\
Y_i - \bar{Y}_i &= \beta_1 (X_{it} - \bar{X}_i) + \tilde{\nu}_{it}
\end{aligned}
$$

[2] Recall **Rule 4** from the **Handout** on the Summation Operator: $\sum (X_i - \bar{X}) = 0$

De-meaned Approach II

- Subtracting the means equation from the pooled equation:

$$
\begin{aligned}
Y_i - [\bar{Y}_i] &= \beta_0 + \beta_1 X_{it} + \alpha_i + \nu_{it} - [\beta_0 + \beta_1 \bar{X}_i + \bar{\alpha}_i + \bar{\nu}_i] \\
Y_i - \bar{Y}_i &= \beta_1 (X_{it} - \bar{X}_i) + \tilde{\nu}_{it} \\
\tilde{Y}_{it} &= \beta_1 \tilde{X}_{it} + \tilde{\nu}_{it}
\end{aligned}
$$

[2] Recall **Rule 4** from the **Handout** on the Summation Operator: $\sum (X_i - \bar{X}) = 0$

- Subtracting the means equation from the pooled equation:

$$
\begin{aligned}
Y_i - [\bar{Y}_i] &= \beta_0 + \beta_1 X_{it} + \alpha_i + \nu_{it} - [\beta_0 + \beta_1 \bar{X}_i + \bar{\alpha}_i + \bar{\nu}_i] \\
Y_i - \bar{Y}_i &= \beta_1 (X_{it} - \bar{X}_i) + \tilde{\nu}_{it} \\
\tilde{Y}_{it} &= \beta_1 \tilde{X}_{it} + \tilde{\nu}_{it}
\end{aligned}
$$

- Within each group, the de-meaned variables $\tilde{Y}_{it}$ and $\tilde{X}_{it}$'s all have a mean of 0[2]
  - Variables that don't change over time will drop out of analysis altogether
- Removes any source of variation **across** groups to only work with variation **within** each group

[2] Recall **Rule 4** from the **Handout** on the Summation Operator: $\sum (X_i - \bar{X}) = 0$

De-meaned Approach II

$$\tilde{Y}_{it} = \beta_1 \tilde{X}_{it} + \tilde{\nu}_{it}$$

- Will yield identical results to LSDV approach

De-meaned Approach II

$$\tilde{Y}_{it} = \beta_1 \tilde{X}_{it} + \tilde{\nu}_{it}$$

- Will yield identical results to LSDV approach
- More useful when we have many groups (would be many dummies)

De-meaned Approach II

$$\tilde{Y}_{it} = \beta_1 \tilde{X}_{it} + \tilde{\nu}_{it}$$

- Will yield identical results to LSDV approach
- More useful when we have many groups (would be many dummies)
- Demonstrates the **intution** behind fixed effects:

De-meaned Approach II

$$\tilde{Y}_{it} = \beta_1 \tilde{X}_{it} + \tilde{\nu}_{it}$$

- Will yield identical results to LSDV approach
- More useful when we have many groups (would be many dummies)
- Demonstrates the **intution** behind fixed effects:
    - Converts data to deviations from the mean levels of each group's variables

De-meaned Approach II

$$\tilde{Y}_{it} = \beta_1 \tilde{X}_{it} + \tilde{\nu}_{it}$$

- Will yield identical results to LSDV approach
- More useful when we have many groups (would be many dummies)
- Demonstrates the **intution** behind fixed effects:
    - Converts data to deviations from the mean levels of each group's variables
    - Fixed effects also called **"within"** estimators, exploit variation *within* groups, not *across* groups

HOOD COLLEGE

Example

Obs.	State	Year	X_{it}	$\bar{X}_i$	$(X_{it} - \bar{X}_i)$ $\tilde{X}_{it}$	Y_{it}	$\bar{Y}_i$	$(Y_{it} - \bar{Y}_i)$ $\tilde{Y}_{it}$
1	California	2015	4	5	-1	12	10	2
2	California	2016	5	5	0	10	10	0
3	California	2017	6	5	1	8	10	-2
4	Maryland	2015	10	15	-5	8	8	0
5	Maryland	2016	20	15	5	4	8	-4
6	Maryland	2017	15	15	0	12	8	4

- Maryland average X_i is higher than California average X_i
- Maryland average Y_i is lower than California average Y_i

- Use `plm()` function from `plm` package (loaded above)

De-meaned Approach III

- Use **plm()** function from **plm** package (loaded above)

```
fe.state<-plm(deaths~cell_plans, data=cell_deaths_tidy,index="state", model="within")
summary(fe.state)
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = deaths ~ cell_plans, data = cell_deaths_tidy, model = "within",
##     index = "state")
##
## Balanced Panel: n = 51, T = 6, N = 306
##
## Residuals:
##     Min.  1st Qu.   Median  3rd Qu.     Max.
## -3.56170 -0.65772 -0.13533  0.59971  3.70868
##
## Coefficients:
##               Estimate  Std. Error t-value  Pr(>|t|)
## cell_plans -0.00120374  0.00010131 -11.882 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    524.94
## Residual Sum of Squares: 337.41
## R-Squared:      0.35724
```

Two-Way Fixed Effects

- So far, we've looked at a one-way fixed effects model that estimates a fixed effect for **groups**

- So far, we've looked at a one-way fixed effects model that estimates a fixed effect for **groups**
- Two-way fixed effects model estimates a fixed effect for *both* the **groups** *and* the **time periods**

$$\hat{Y}_{it} = \beta_0 + \beta_1 X_{it} + \alpha_i + \theta_t + \nu_{it}$$

- So far, we've looked at a one-way fixed effects model that estimates a fixed effect for **groups**
- Two-way fixed effects model estimates a fixed effect for *both* the **groups** *and* the **time periods**

$$\hat{Y}_{it} = \beta_0 + \beta_1 X_{it} + \alpha_i + \theta_t + \nu_{it}$$

  - α_i: group fixed effects

- So far, we've looked at a one-way fixed effects model that estimates a fixed effect for **groups**
- Two-way fixed effects model estimates a fixed effect for *both* the **groups** *and* the **time periods**

$$\hat{Y_{it}} = \beta_0 + \beta_1 X_{it} + \alpha_i + \theta_t + \nu_{it}$$

  - α_i: group fixed effects
    - accounts for **time-invariant differences across groups**

Two-Way Fixed Effects

- So far, we've looked at a one-way fixed effects model that estimates a fixed effect for **groups**
- Two-way fixed effects model estimates a fixed effect for *both* the **groups** *and* the **time periods**

$$\hat{Y}_{it} = \beta_0 + \beta_1 X_{it} + \alpha_i + \theta_t + \nu_{it}$$

  - α_i: group fixed effects
    - accounts for **time-invariant differences across groups**
  - θ_t: time fixed effects

Two-Way Fixed Effects

- So far, we've looked at a one-way fixed effects model that estimates a fixed effect for **groups**
- Two-way fixed effects model estimates a fixed effect for *both* the **groups** *and* the **time periods**

$$\hat{Y}_{it} = \beta_0 + \beta_1 X_{it} + \alpha_i + \theta_t + \nu_{it}$$

- α_i: group fixed effects
    - accounts for **time-invariant differences across groups**
- θ_t: time fixed effects
    - acounts for **group-invariant differences over time**

Example

$$\widehat{\text{Deaths}}_{it} = \beta_0 + \beta_1 \text{Cell Phones}_{it} + \alpha_i + \theta_t + \nu_{it}$$

- α_i: state fixed effects

Example

$$\widehat{\text{Deaths}}_{it} = \beta_0 + \beta_1 \text{Cell Phones}_{it} + \alpha_i + \theta_t + \nu_{it}$$

- α_i: state fixed effects
  - differences across states that are stable over time

Example

$$\widehat{\text{Deaths}}_{it} = \beta_0 + \beta_1 \text{Cell Phones}_{it} + \alpha_i + \theta_t + \nu_{it}$$

- α_i: state fixed effects
  - differences across states that are stable over time
  - e.g. geography, culture, (unchanging) state laws

Example

$$\widehat{\text{Deaths}}_{it} = \beta_0 + \beta_1 \text{Cell Phones}_{it} + \alpha_i + \theta_t + \nu_{it}$$

- α_i: state fixed effects
  - differences across states that are stable over time
  - e.g. geography, culture, (unchanging) state laws
- θ_t: time fixed effects

Example

$$\widehat{\text{Deaths}_{it}} = \beta_0 + \beta_1 \text{Cell Phones}_{it} + \alpha_i + \theta_t + \nu_{it}$$

- α_i: state fixed effects
  - differences across states that are stable over time
  - e.g. geography, culture, (unchanging) state laws
- θ_t: time fixed effects
  - differences over time that are stable across states

Two-Way Fixed Effects: Example

Example

$$\widehat{\text{Deaths}_{it}} = \beta_0 + \beta_1 \text{Cell Phones}_{it} + \alpha_i + \theta_t + \nu_{it}$$

- α_i: state fixed effects
  - differences across states that are stable over time
  - e.g. geography, culture, (unchanging) state laws
- θ_t: time fixed effects
  - differences over time that are stable across states
  - e.g. U.S. population growth, epidemics, macroeconomic conditions, *federal* laws passed

- As before, several equivalent ways to estimate two-way fixed effects models:

[3] Where each $\widetilde{variable}_{it} = variable_{it} - \overline{variable}_t - \overline{variable}_i$

- As before, several equivalent ways to estimate two-way fixed effects models:

  1. **Least Squares Dummy Variable (LSDV) Approach**: add dummies for both groups and time periods (separate intercepts for groups and times)

[3] Where each $\widetilde{variable}_{it} = variable_{it} - \overline{variable}_t - \overline{variable}_i$

- As before, several equivalent ways to estimate two-way fixed effects models:

  1. **Least Squares Dummy Variable (LSDV) Approach**: add dummies for both groups and time periods (separate intercepts for groups and times)
  2. **Fully De-meaned data**:[3]

$$\tilde{Y}_{it} = \beta_1 \tilde{X}_{it} + \tilde{\nu}_{it}$$

[3] Where each $\widetilde{variable}_{it} = variable_{it} - \overline{variable}_{t} - \overline{variable}_{i}$

- As before, several equivalent ways to estimate two-way fixed effects models:

1. **Least Squares Dummy Variable (LSDV) Approach**: add dummies for both groups and time periods (separate intercepts for groups and times)
2. **Fully De-meaned data**:[3]

$$\tilde{Y}_{it} = \beta_1 \tilde{X}_{it} + \tilde{\nu}_{it}$$

3. **Hybrid**: de-mean for one effect (groups or times) and add dummies for the other effect (times or groups)

[3] Where each $\widetilde{variable}_{it} = variable_{it} - \overline{variable}_t - \overline{variable}_i$

- `plm()` command allows for multiple effects to be fit inside `index=c("group", "time")`

```
fe.2way<-plm(deaths~cell_plans, data=cell_deaths_tidy,index=c("state","year"), model="within")
summary(fe.2way)
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = deaths ~ cell_plans, data = cell_deaths_tidy, model = "within",
##     index = c("state", "year"))
##
## Balanced Panel: n = 51, T = 6, N = 306
##
## Residuals:
##     Min.  1st Qu.   Median  3rd Qu.     Max.
## -3.56170 -0.65772 -0.13533  0.59971  3.70868
##
## Coefficients:
##               Estimate  Std. Error t-value  Pr(>|t|)
## cell_plans -0.00120374  0.00010131 -11.882 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    524.94
## Residual Sum of Squares: 337.41
## R-Squared:      0.35724
```

- Can still add covariates to remove endogeneity not soaked up by fixed effects

Adding Covariates

- Can still add covariates to remove endogeneity not soaked up by fixed effects
    - Factors that change within groups over time

- Can still add covariates to remove endogeneity not soaked up by fixed effects
    - Factors that change within groups over time
    - e.g. some states pass bans over the time period in data (some years before, some years after)

Adding Covariates

- Can still add covariates to remove endogeneity not soaked up by fixed effects
  - Factors that change within groups over time
  - e.g. some states pass bans over the time period in data (some years before, some years after)

$$\widehat{\text{Deaths}_{it}} = \beta_1 \text{Cell Phones}_{it} + \alpha_i + \theta_t + \text{urban pct}_{it} + \text{cell ban}_{it} + \text{text ban}_{it}$$

Adding Covariates

- Can still add covariates to remove endogeneity not soaked up by fixed effects
  - Factors that change within groups over time
  - e.g. some states pass bans over the time period in data (some years before, some years after)

$$\widehat{\text{Deaths}_{it}} = \beta_1 \text{Cell Phones}_{it} + \alpha_i + \theta_t + \text{urban pct}_{it} + \text{cell ban}_{it} + \text{text ban}_{it}$$

```r
controls<-plm(deaths~cell_plans+text_ban+urban_pct+cell_ban, data=cell_deaths_tid
summary(controls)
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = deaths ~ cell_plans + text_ban + urban_pct + cell_ban,
##     data = cell_deaths_tidy, model = "within", index = c("state",
##         "year"))
##
## Balanced Panel: n = 51, T = 6, N = 306
##
## Residuals:
##     Min.  1st Qu.   Median  3rd Qu.     Max.
## -3.34700 -0.61022 -0.12759  0.53430  3.76050
##
## Coefficients:
##                  Estimate  Std. Error t-value Pr(>|t|)
## cell_plans -0.00116937  0.00012047 -9.7064  < 2e-16 ***
## text_ban   -0.03388558  0.22564932 -0.1502  0.88075
## urban_pct   0.01132216  0.01246072  0.9086  0.36442
## cell_ban   -0.96915929  0.44538291 -2.1760  0.03049 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    524.94
```

	Dependent variable:			
	deaths			
	Pooled	State Effects	State and Time Effects	State and Time Effects
	(1)	(2)	(3)	(4)
cell_plans	−0.001***	−0.001***	−0.001***	−0.001***
	(0.0001)	(0.0001)	(0.0001)	(0.0001)
text_ban				−0.034
				(0.226)
urban_pct				0.011
				(0.012)
cell_ban				−0.969**
				(0.445)
Constant	17.337***			
	(0.975)			
Observations	306	306	306	306
R^2	0.084	0.357	0.357	0.373
Adjusted R^2	0.081	0.228	0.228	0.238
Residual Std. Error	3.279 (df = 304)			
F Statistic	28.057*** (df = 1; 304)	141.169*** (df = 1; 254)	141.169*** (df = 1; 254)	37.327*** (df = 4; 251)