4.1 — Panel Data and Fixed Effects

ECON 480 • Econometrics • Fall 2021

Ryan Safner

Assistant Professor of Economics

safner@hood.edu

ryansafner/metricsF21

metricsF21.classes.ryansafner.com

Outline

Pooled Regression Model

Types of Data I

- **Cross-sectional data**: compare different individual i's at same time $\bar{t}$

state	year	deaths	cell_plans
<fct>	<fct>	<dbl>	<dbl>
Alabama	2012	13.316056	9433.800
Alaska	2012	12.311976	8872.799
Arizona	2012	13.720419	8810.889
Arkansas	2012	16.466730	10047.027
California	2012	8.756507	9362.424
Colorado	2012	10.092204	9403.225

6 rows

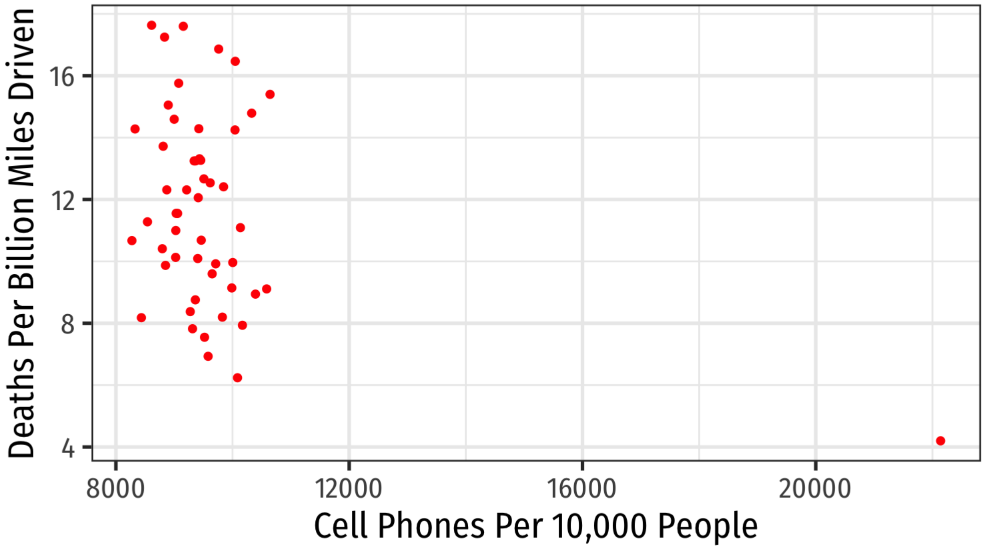
- **Time-series data**: track same individual $\bar{i}$ over different times t

state	year	deaths	cell_plans
<fct>	<fct>	<dbl>	<dbl>
Maryland	2007	10.866679	8942.137
Maryland	2008	10.740963	9290.689
Maryland	2009	9.892754	9339.452
Maryland	2010	8.783883	9630.120
Maryland	2011	8.626745	10335.795
Maryland	2012	8.941916	10393.295

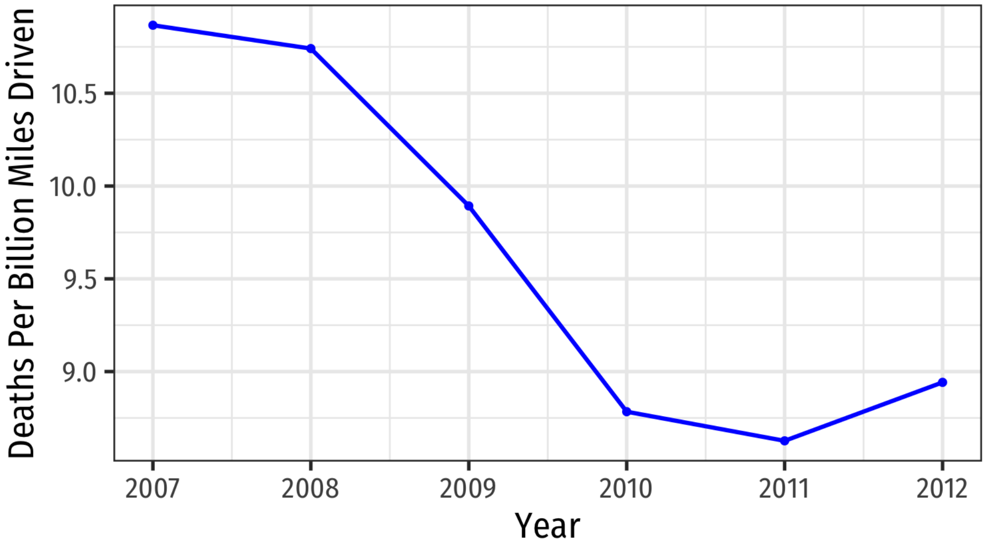
6 rows

Types of Data I

- **Cross-sectional data**: compare different individual i's at same time $\bar{t}$

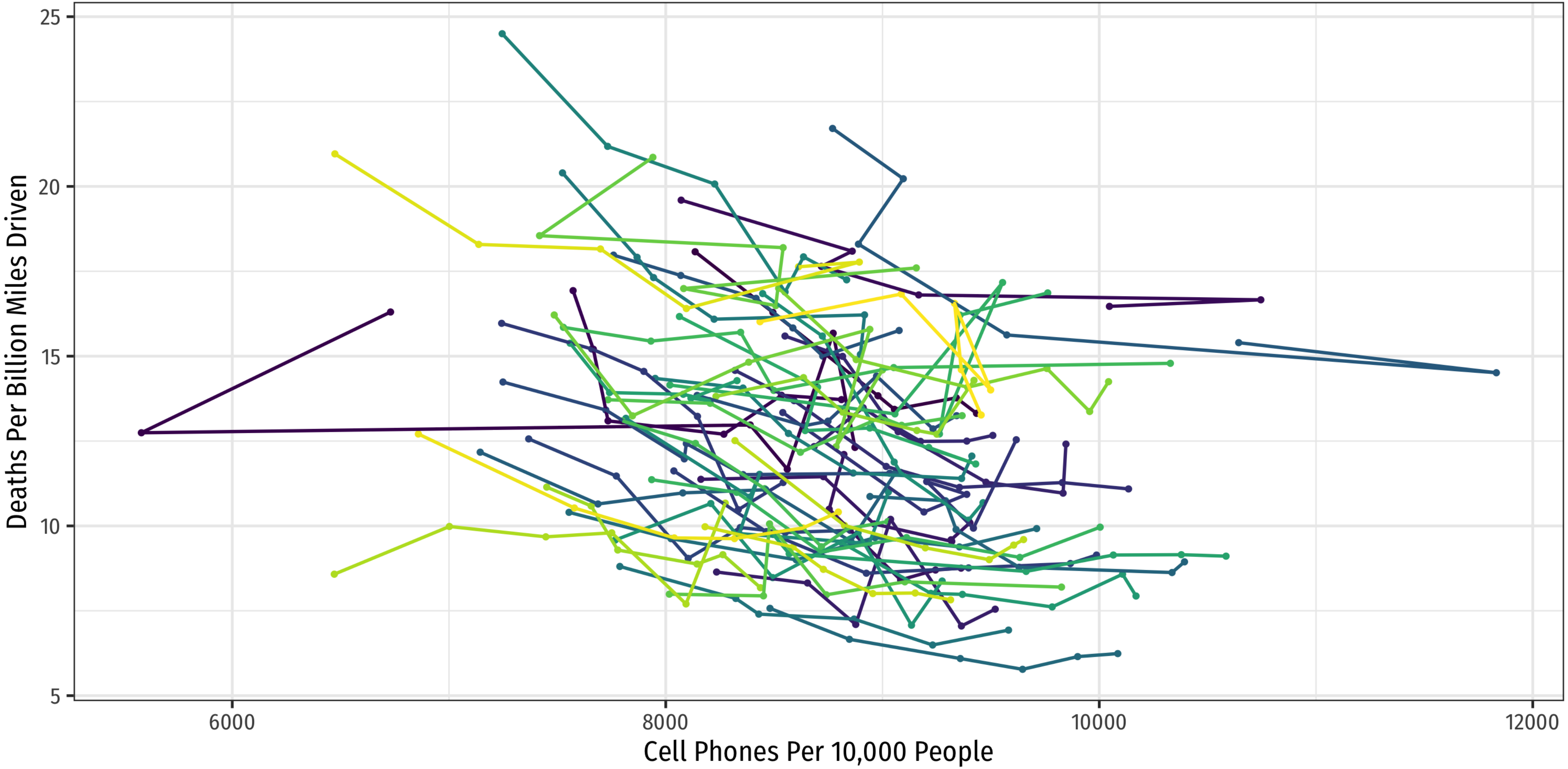


- **Time-series data**: track same individual $\bar{i}$ over different times t



- **Panel data**: combines these dimensions: compare all individual i's over all time t's

Panel Data I



Deaths Per Billion Miles Driven
Cell Phones Per 10,000 People
25
20
15
10
5
6000
8000
10000
12000

Panel Data II

state	ye...	deaths	cell_plans
<fct>	<fct>	<dbl>	<dbl>
Alabama	2007	18.075232	8135.525
Alabama	2008	16.289227	8494.391
Alabama	2009	13.833678	8979.108
Alabama	2010	13.434084	9054.894
Alabama	2011	13.771989	9340.501
Alabama	2012	13.316056	9433.800
Alaska	2007	16.301184	6730.282
Alaska	2008	12.744090	5580.707
Alaska	2009	12.973849	8389.730
Alaska	2010	11.670893	8560.595

1-10 of 306 rows Previous **1** 2 3 4 5 6 ... 31 Next

- **Panel** or **Longitudinal** data contains
  - repeated observations (t)
  - on multiple individuals (i)

Panel Data II

state	ye...	deaths	cell_plans
<fct>	<fct>	<dbl>	<dbl>
Alabama	2007	18.075232	8135.525
Alabama	2008	16.289227	8494.391
Alabama	2009	13.833678	8979.108
Alabama	2010	13.434084	9054.894
Alabama	2011	13.771989	9340.501
Alabama	2012	13.316056	9433.800
Alaska	2007	16.301184	6730.282
Alaska	2008	12.744090	5580.707
Alaska	2009	12.973849	8389.730
Alaska	2010	11.670893	8560.595

1-10 of 306 rows   Previous 1 2 3 4 5 6 ... 31 Next

- **Panel** or **Longitudinal** data contains
  - repeated observations (t)
  - on multiple individuals (i)
- Thus, our regression equation looks like:

```
\begin{align}
  \hat{Y}_{\color{red}{i}\color{blue}{t}}} = \beta_0 +
\end{align}
```

> for individual i in time t.

Panel Data: Our Motivating Example

state	ye...	deaths	cell_plans
<fct>	<fct>	<dbl>	<dbl>
Alabama	2007	18.075232	8135.525
Alabama	2008	16.289227	8494.391
Alabama	2009	13.833678	8979.108
Alabama	2010	13.434084	9054.894
Alabama	2011	13.771989	9340.501
Alabama	2012	13.316056	9433.800
Alaska	2007	16.301184	6730.282
Alaska	2008	12.744090	5580.707
Alaska	2009	12.973849	8389.730
Alaska	2010	11.670893	8560.595

1-10 of 306 rows Previous **1** 2 3 4 5 6 ... 31 Next

Example: Do cell phones cause more traffic fatalities?

- No measure of cell phones *used* while driving
  - `cell_plans` as a **proxy** for cell phone usage
- State-level data over 6 years

The Data I

```
glimpse(phones)
```

```
## Rows: 306
## Columns: 8
## $ year          <fct> 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 20…
## $ state         <fct> Alabama, Alaska, Arizona, Arkansas, California, Colorado…
## $ urban_percent <dbl> 30, 55, 45, 21, 54, 34, 84, 31, 100, 53, 39, 45, 11, 56,…
## $ cell_plans    <dbl> 8135.525, 6730.282, 7572.465, 8071.125, 8821.933, 8162.0…
## $ cell_ban      <fct> 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,…
## $ text_ban      <fct> 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,…
## $ deaths        <dbl> 18.075232, 16.301184, 16.930578, 19.595430, 12.104340, 1…
## $ year_num      <dbl> 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 20…
```

The Data II

```r
phones %>%
  count(state)
```

state	n
<fct>	<int>
Alabama	6
Alaska	6
Arizona	6
Arkansas	6
California	6
Colorado	6
Connecticut	6
Delaware	6
District of Columbia	6
Florida	6

1-10 of 51 rows Previous **1** 2 3 4 5 6 Next

```r
phones %>%
  count(year)
```

year	n
<fct>	<int>
2007	51
2008	51
2009	51
2010	51
2011	51
2012	51

6 rows

The Data III

```
phones %>%
  distinct(state)
```

state <fct>
Alabama
Alaska
Arizona
Arkansas
California
Colorado
Connecticut
Delaware
District of Columbia
Florida

1-10 of 51 rows Previous 1 2 3 4 5 6 Next

```
phones %>%
  distinct(year)
```

year <fct>
2007
2008
2009
2010
2011
2012

6 rows

The Data IV

```
phones %>%
  summarize(States = n_distinct(state),
            Years = n_distinct(year))
```

	States	Years
	<int>	<int>
	51	6

1 row

Pooled Regression I

- What if we just ran a standard regression:

$$\hat{Y}_{it} = \beta_0 + \beta_1 X_{it} + u_{it}$$

- 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

Pooled Regression II

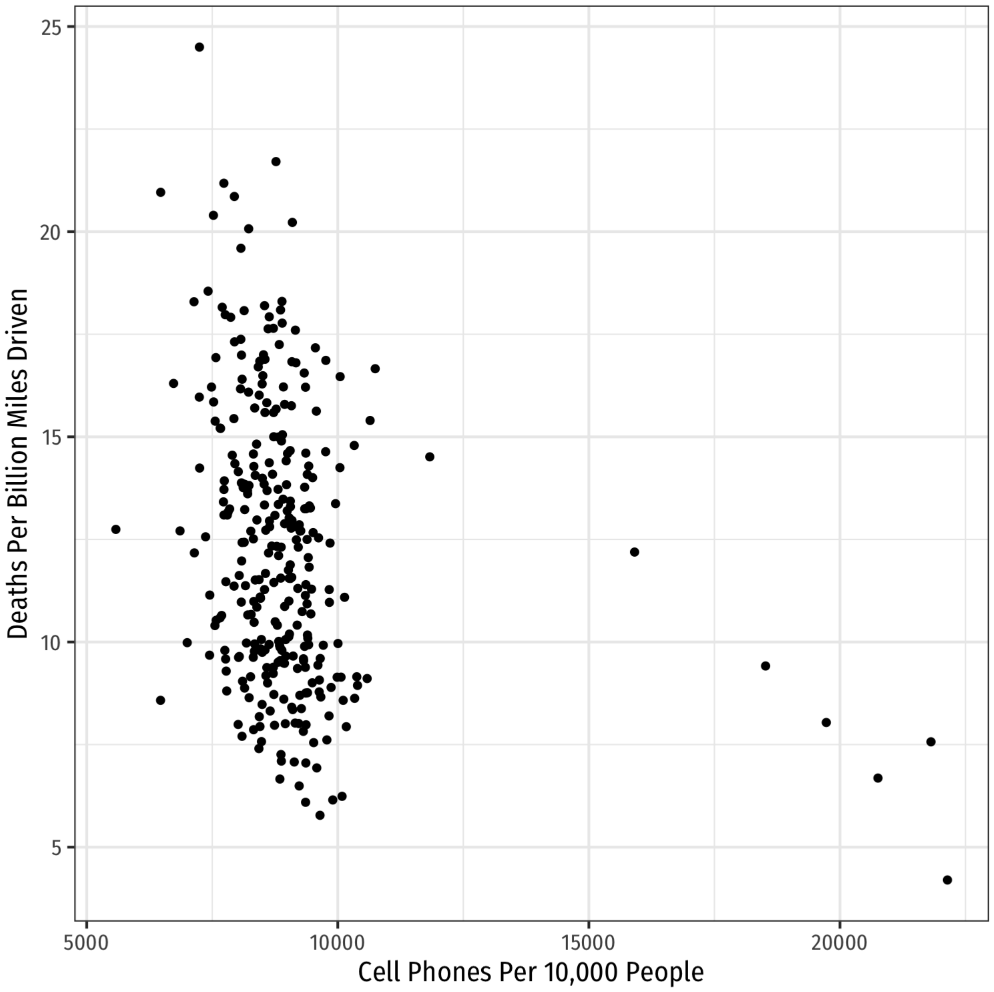
```
pooled <- lm(deaths ~ cell_plans, data = phones)
pooled %>% tidy()
```

term	estimate	std.error	statistic	p.value
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
(Intercept)	17.3371034167	0.975384504	17.774635	5.821724e-49
cell_plans	-0.0005666385	0.000106975	-5.296926	2.264086e-07

2 rows

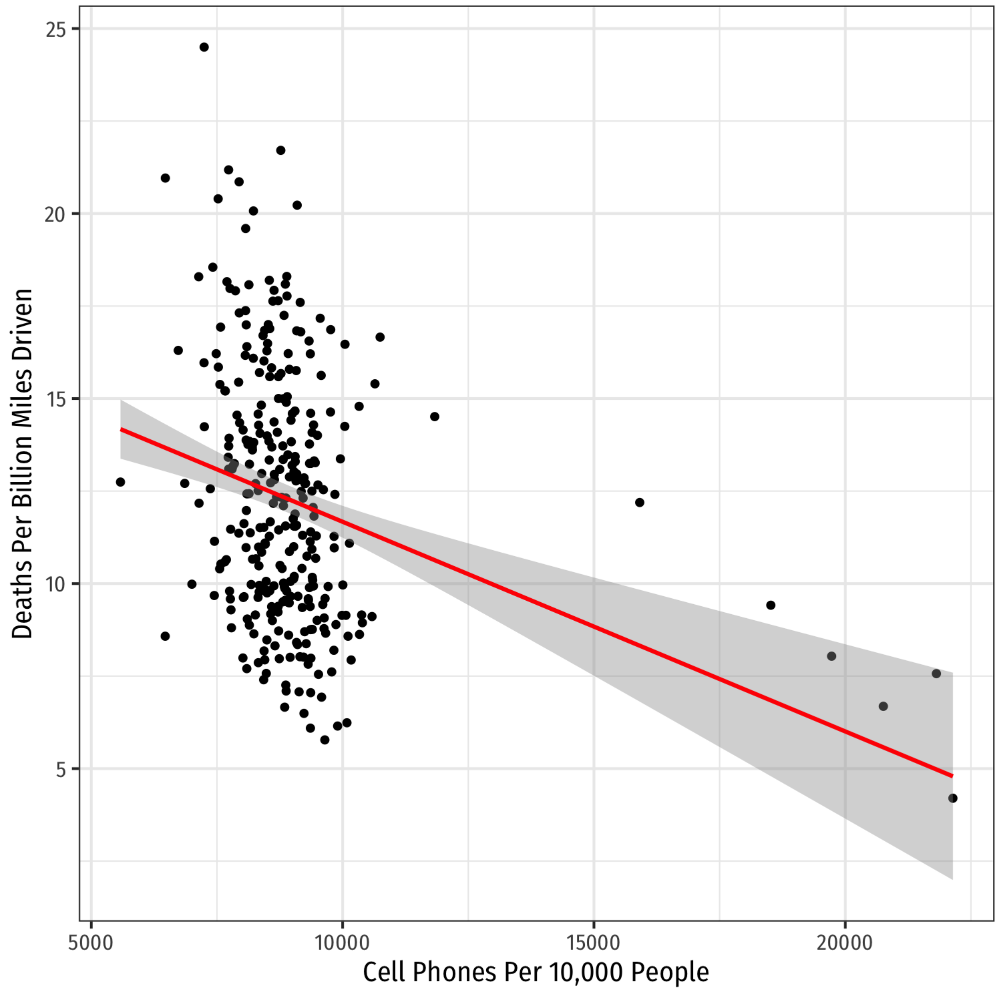
Pooled Regression III

```r
ggplot(data = phones)+
  aes(x = cell_plans,
      y = deaths)+
  geom_point()+
  labs(x = "Cell Phones Per 10,000 People",
       y = "Deaths Per Billion Miles Driven")+
  theme_bw(base_family = "Fira Sans Condensed",
           base_size=14)
```



Pooled Regression III

```r
ggplot(data = phones)+
  aes(x = cell_plans,
      y = deaths)+
  geom_point()+
  geom_smooth(method = "lm", color = "red")+
  labs(x = "Cell Phones Per 10,000 People",
       y = "Deaths Per Billion Miles Driven")+
  theme_bw(base_family = "Fira Sans Condensed",
           base_size=14)
```



Recap: Assumptions about Errors

- Recall the **4 critical assumptions about u**:

1. The expected value of the residuals is 0

$$E[u] = 0$$

2. The variance of the residuals over X is constant:

$$var(u|X) = \sigma_u^2$$

3. **Errors are not correlated across observations:**

$$cor(u_i, u_j) = 0 \quad \forall i \neq j$$

4. There is no correlation between X and the error term:

$$cor(X, u) = 0 \text{ or } E[u|X] = 0$$

Biases of Pooled Regression

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

- **Assumption 3**: $cor(u_i, u_j) = 0 \quad \forall\, i \neq j$

- Pooled regression model is **biased** because it ignores:
  - Multiple observations from same group i
  - Multiple observations from same time t

- Thus, errors are **serially** or **auto-correlated**; $cor(u_i, u_j) \neq 0$ within same i and within same t

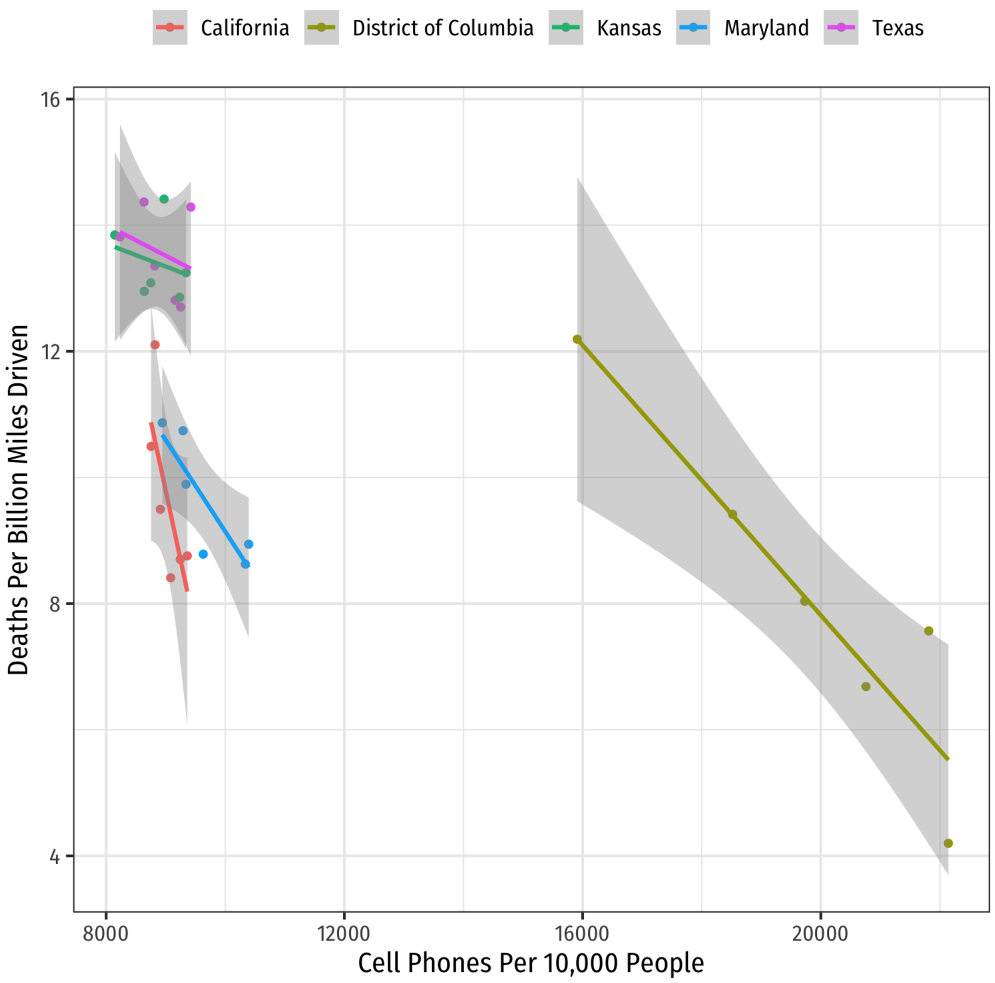
Biases of Pooled Regression: Our Example

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

- **Multiple observations from same state i**
    - Probably similarities among u for obs in same state
    - Residuals on observations from same state are likely correlated
- **Multiple observations from same year t**
    - Probably similarities among u for obs in same year
    - Residuals on observations from same year are likely correlated

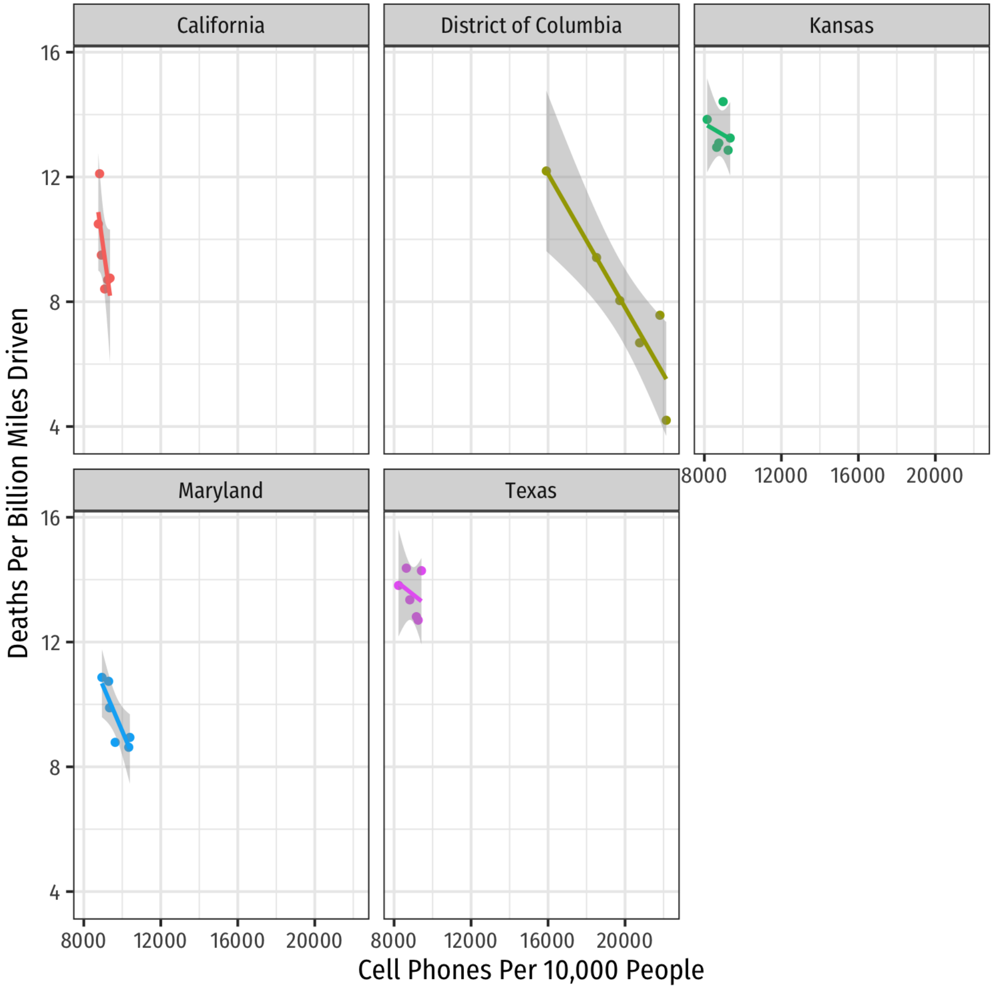
Example: Consider Just 5 States

```
phones %>%
  filter(state %in% c("District of Columbia",
                      "Maryland", "Texas",
                      "California", "Kansas")) %>%
ggplot(data = .)+
  aes(x = cell_plans,
      y = deaths,
      color = state)+
  geom_point()+
  geom_smooth(method = "lm")+
  labs(x = "Cell Phones Per 10,000 People",
       y = "Deaths Per Billion Miles Driven",
       color = NULL)+
  theme_bw(base_family = "Fira Sans Condensed",
           base_size=14)+
  theme(legend.position = "top")
```



Example: Consider Just 5 States

```
phones %>%
  filter(state %in% c("District of Columbia",
                      "Maryland", "Texas",
                      "California", "Kansas")) %>%
ggplot(data = .)+
  aes(x = cell_plans,
      y = deaths,
      color = state)+
  geom_point()+
  geom_smooth(method = "lm")+
  labs(x = "Cell Phones Per 10,000 People",
       y = "Deaths Per Billion Miles Driven",
       color = NULL)+
  theme_bw(base_family = "Fira Sans Condensed",
           base_size=14)+
  theme(legend.position = "none")+
  facet_wrap(~state, ncol=3)
```



Look at All States

```r
ggplot(data = phones)+
  aes(x = cell_plans,
      y = deaths,
      color = state)+
  geom_point()+
  geom_smooth(method = "lm")+
  labs(x = "Cell Phones Per 10,000 People",
       y = "Deaths Per Billion Miles Driven",
       color = NULL)+
  theme_bw(base_family = "Fira Sans Condensed")+
  theme(legend.position = "none")+
  facet_wrap(~state, ncol=7)
```

The Bias in our Pooled Regression

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

- Cell Phones_{it} is **endogenous**:

$$cor(\text{u}_{it}, \text{Cell Phones}_{it}) \neq 0 \qquad E[\text{u}_{it} | \text{Cell Phones}_{it}] \neq 0$$

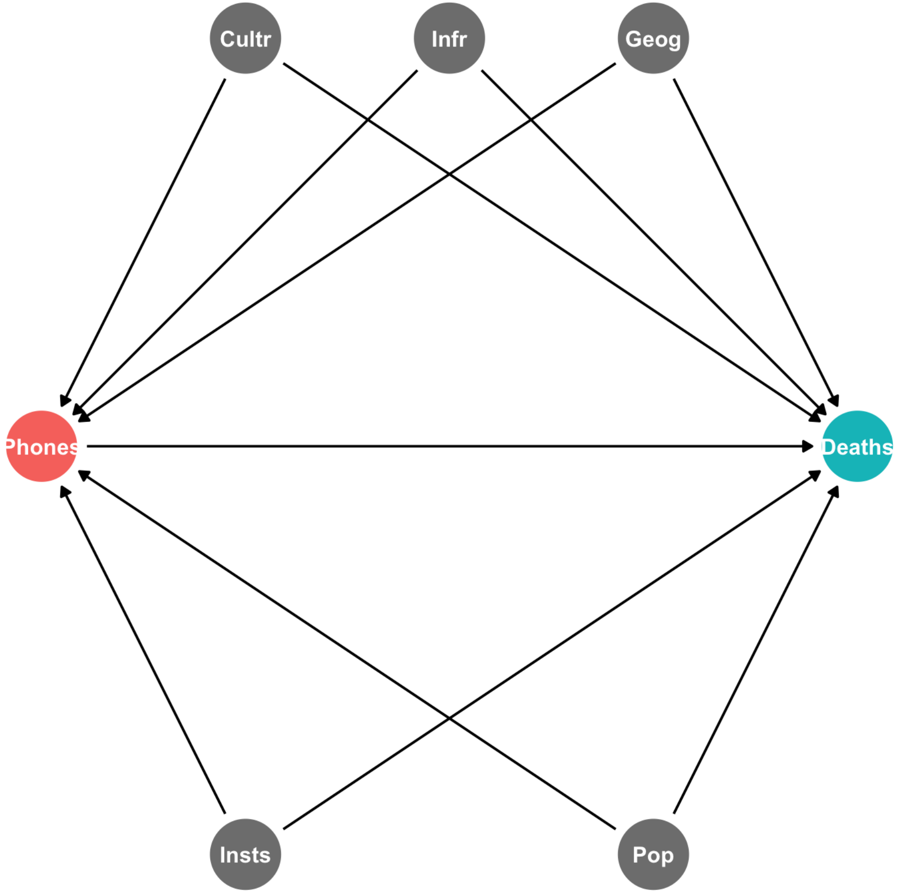
- Things in u_{it} correlated with Cell phones_{it}:
  - infrastructure spending, population, urban vs. rural, more/less cautious citizens, cultural attitudes towards driving, texting, etc

- A lot of these things vary systematically **by State**!
  - $cor(\text{u}_{it_1}, \text{u}_{it_2}) \neq 0$
    - Error in State i during t_1 correlates with error in State i during t_2
    - things in State i that don't change over time

Fixed Effects Model

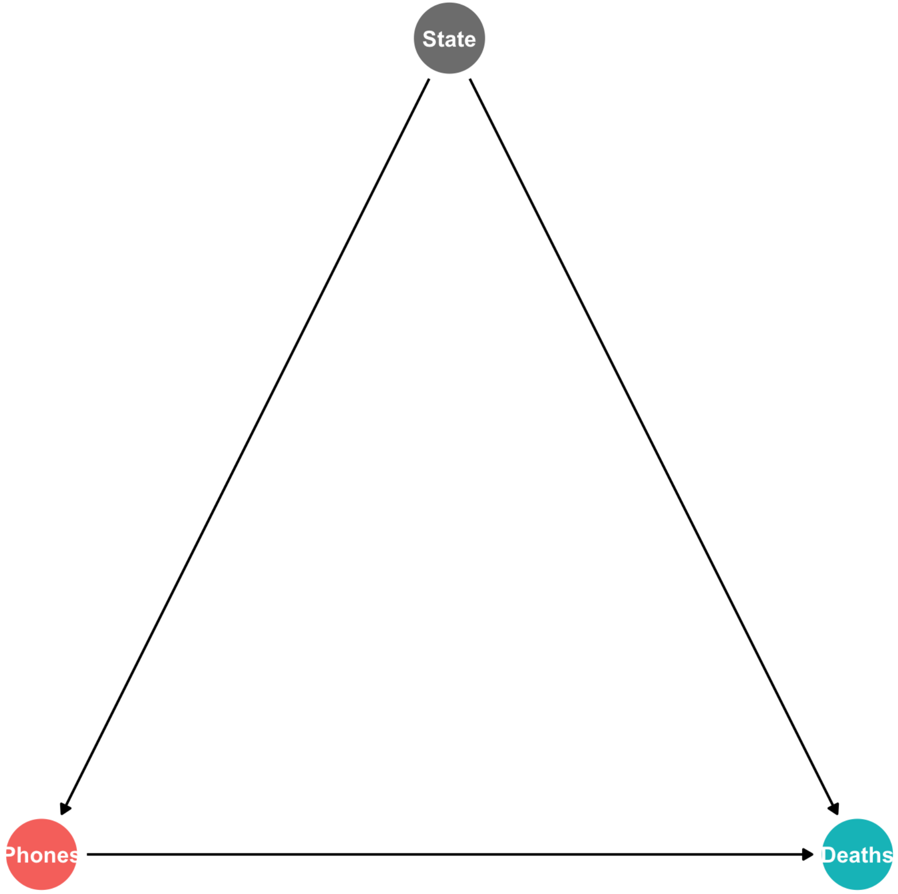
Fixed Effects: DAG

- A simple pooled model likely contains lots of omitted variable bias
- Many (often unobservable) factors that determine both Phones & Deaths
  - Culture, infrastructure, population, geography, institutions, etc



Fixed Effects: DAG

- A simple pooled model likely contains lots of omitted variable bias
- Many (often unobservable) factors that determine both Phones & Deaths
  - Culture, infrastructure, population, geography, institutions, etc
- But the beauty of this is that **most of these factors systematically vary by U.S. State and are stable over time!**
- We can simply **“control for State”** to safely remove the influence of all of these factors!



Fixed Effects: Decomposing u_{it}

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

- Exploit the systematic variation across groups with a **fixed effects model**

- *Decompose* the model error term into two parts:

$$\text{u}_{it} = \alpha_i + \epsilon_{it}$$

Fixed Effects: α_i

- *Decompose* the model error term into two parts:

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

- α_i are **group-specific fixed effects**
    - group i tends to have higher or lower $\hat{Y}$ than other groups given regressor(s) X_{it}
    - estimate a separate α_i for each group i
    - essentially, estimate a separate constant (intercept) *for each group*
    - notice this is stable over time within each group (subscript only i, no t)

- **This includes all factors that do not change *within* group *i* over time**

Fixed Effects: ϵ_{it}

- *Decompose* the model error term into two parts:

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

- ϵ_{it} is the **remaining random error**
  - As usual in OLS, assume the 4 typical assumptions about this error:
  - $E[\epsilon_{it}] = 0, var[\epsilon_{it}] = \sigma^2_{\epsilon}, cor(\epsilon_{it}, \epsilon_{jt}) = 0, cor(\epsilon_{it}, X_{it}) = 0$
- ϵ_{it} includes all other factors affecting Y_{it} *not* contained in group effect α_i
  - i.e. differences *within* each group that *change* over time
  - Be careful: X_{it} **can still be endogenous due to other factors!**
    - $cor(X_{it}, \epsilon_{it}) \neq 0$

Fixed Effects: New Regression Equation

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

- We've pulled α_i out of the original error term into the regression
- Essentially we'll estimate an intercept **for each group** (minus one, which is β_0)
  - avoiding the dummy variable trap
- Must have multiple observations (over time) for each group (i.e. panel data)

Fixed Effects: Our Example

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

- α_i is the **State fixed effect**
  - Captures everything unique about each state i that *does not change over time*
  - culture, institutions, history, geography, climate, etc!
- There could *still* be factors in ϵ_{it} that are correlated with Cell phones_{it}!
  - things that do change over time within States
  - perhaps individual States have cell phone bans for *some* years in our data

Estimating Fixed Effects Models

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

- Two methods to estimate fixed effects models:

1. Least Squares Dummy Variable (LSDV) approach

2. De-meaned data approach

Least Squares Dummy Variable Approach

Least Squares Dummy Variable Approach

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

- A dummy variable $D_i = \{0, 1\}$ for each possible group
  - $= 1$ if observation it is from group i, otherwise $= 0$

- If there are N groups:
  - Include $N - 1$ dummies (to avoid **dummy variable trap**) and β_0 is the reference category[†]
  - So we are estimating a different intercept for each group

- Sounds like a lot of work, automatic in `R`

[†] If we do not estimate β_0, we could include all N dummies. In either case, β_0 takes the place of one category-dummy.

Least Squares Dummy Variable Approach: Our Example

Example:

$$\widehat{\text{Deaths}}_{it} = \beta_0 + \beta_1 \text{Cell Phones}_{it} + \text{Alaska}_i + \cdots + \text{Wyoming}_i$$

- Let Alabama be the reference category (β_0), include all other States

Our Example in R I

$$\widehat{\text{Deaths}}_{it} = \beta_0 + \beta_1 \text{Cell Phones}_{it} + \text{Alaska}_i + \cdots + \text{Wyoming}_i$$

- If `state` is a `factor` variable, just include it in the regression
- `R` automatically creates $N-1$ dummy variables and includes them in the regression
    - Keeps intercept and leaves out first group dummy

Our Example in R II

```r
fe_reg_1 <- lm(deaths ~ cell_plans + state, data = phones)
fe_reg_1 %>% tidy()
```

term	estimate	std.error	statistic	p.value
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
(Intercept)	25.507679925	1.0176400289	25.06552337	1.241581e-70
cell_plans	-0.001203742	0.0001013125	-11.88147584	3.483442e-26
stateAlaska	-2.484164783	0.6745076282	-3.68293060	2.816972e-04
stateArizona	-1.510577383	0.6704569688	-2.25305643	2.510925e-02
stateArkansas	3.192662931	0.6664383936	4.79063476	2.829319e-06
stateCalifornia	-4.978668651	0.6655467951	-7.48056889	1.206933e-12
stateColorado	-4.344553493	0.6654735335	-6.52851432	3.588784e-10
stateConnecticut	-6.595185530	0.6654428902	-9.91097152	8.698802e-20
stateDelaware	-2.098393628	0.6666483193	-3.14767707	1.842218e-03
stateDistrict of Columbia	6.355790010	1.2897172620	4.92804911	1.499627e-06

1-10 of 52 rows   Previous **1** 2 3 4 5 6 Next

De-meaned Approach

De-meaned Approach I

- Alternatively, we can control our regression for group fixed effects without directly estimating them

- We simply **de-mean the data for each group** to remove the group fixed-effect

- For each group i, find the means (over time, t):

$$\bar{Y}_i = \beta_0 + \beta_1 \bar{X}_i + \bar{\alpha}_i + \bar{\epsilon}_{it}$$

- Where:
  - $\bar{Y}_i$: 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 $(= \alpha_i)$
  - $\bar{\epsilon}_{it} = 0$, by assumption 1 about errors

De-meaned Approach II

$$\begin{aligned}\widehat{Y_{it}} &= \beta_0 + \beta_1 X_{it} + u_{it} \\ \bar{Y}_i &= \beta_0 + \beta_1 \bar{X}_i + \bar{\alpha}_i + \bar{\epsilon}_i\end{aligned}$$

- Subtract the means equation from the pooled equation to get:

$$\begin{aligned}Y_{it} - \bar{Y}_i &= \beta_1 (X_{it} - \bar{X}_i) + \tilde{\epsilon}_{it} \\ \tilde{Y}_{it} &= \beta_1 \tilde{X}_{it} + \tilde{\epsilon}_{it}\end{aligned}$$

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

[†] Recall **Rule 4** from the 2.3 class notes on the Summation Operator: $\sum (X_i - \bar{X}) = 0$

De-meaned Approach III

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

- Yields identical results to dummy variable approach
- More useful when we have many groups (would be many dummies)
- Demonstrates **intuition** behind fixed effects:
    - Converts all data to deviations from the mean of each group
    - All groups are “centered” at 0, no variation across groups
    - Fixed effects are often called the **“within” estimators**, they exploit variation *within* groups, not *across* groups

De-meaned Approach IV

- We are basically comparing groups *to themselves* over time
    - apples to apples comparison
    - e.g. Maryland in 2000 vs. Maryland in 2005
- Ignore all differences *between* groups, only look at differences *within* groups over time

De-Meaning the Data in R I

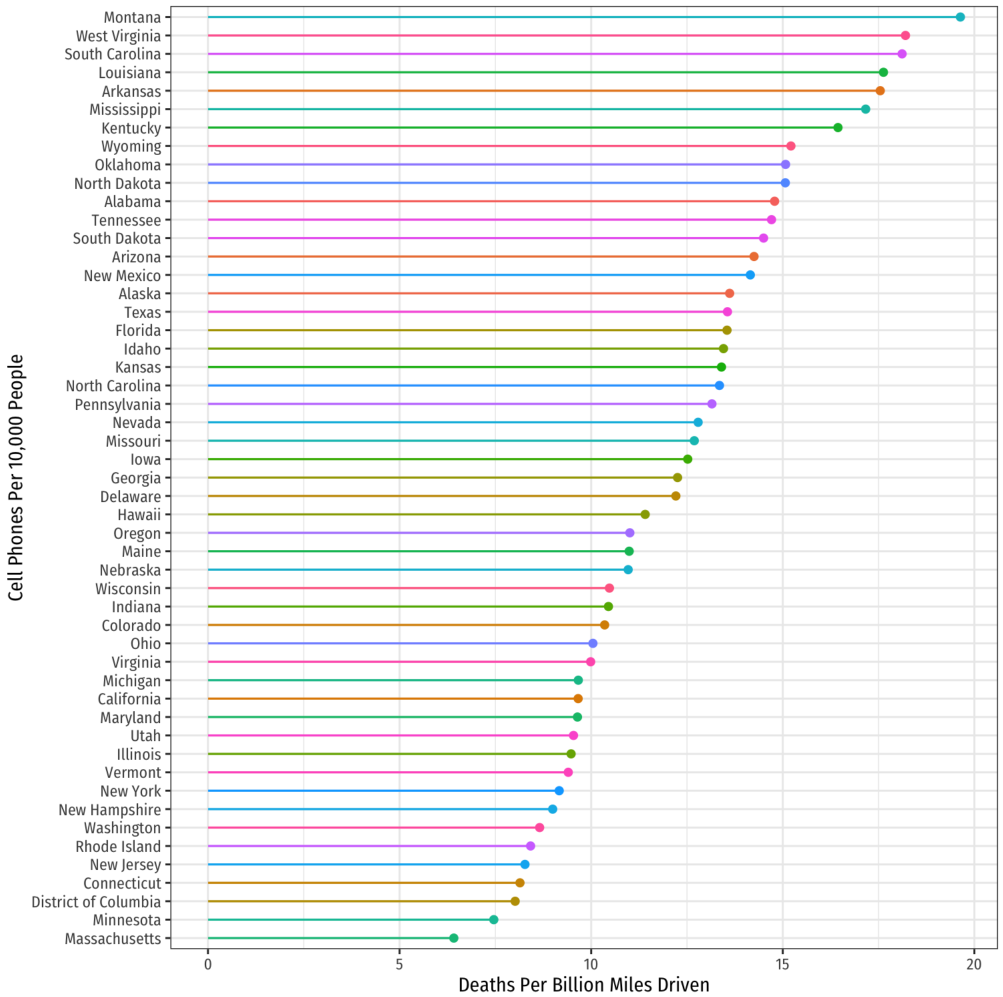
```r
# get means of Y and X by state
means_state <- phones %>%
  group_by(state) %>%
  summarize(avg_deaths = mean(deaths),
            avg_phones = mean(cell_plans))

# look at it
means_state
```

state <fct>	avg_deaths <dbl>	avg_phones <dbl>
Alabama	14.786711	8906.370
Alaska	13.612953	7817.759
Arizona	14.249825	8097.482
Arkansas	17.543881	9268.153
California	9.659712	9029.594
Colorado	10.351405	8981.762
Connecticut	8.141739	8947.729
Delaware	12.209610	9304.052
District of Columbia	8.015895	19811.205
Florida	13.544635	9078.592

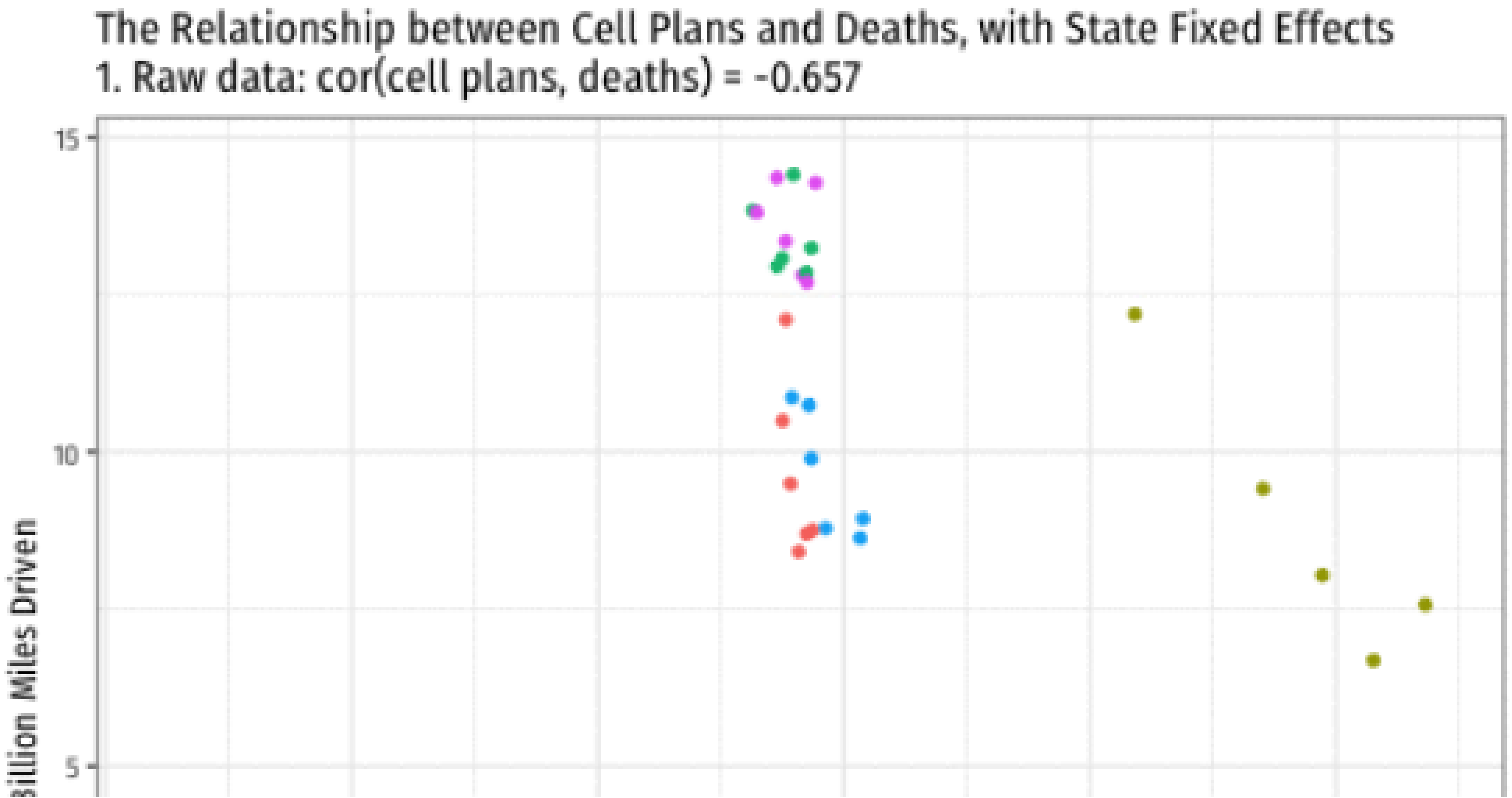
1-10 of 51 rows Previous 1 2 3 4 5 6 Next

De-Meaning the Data in R II

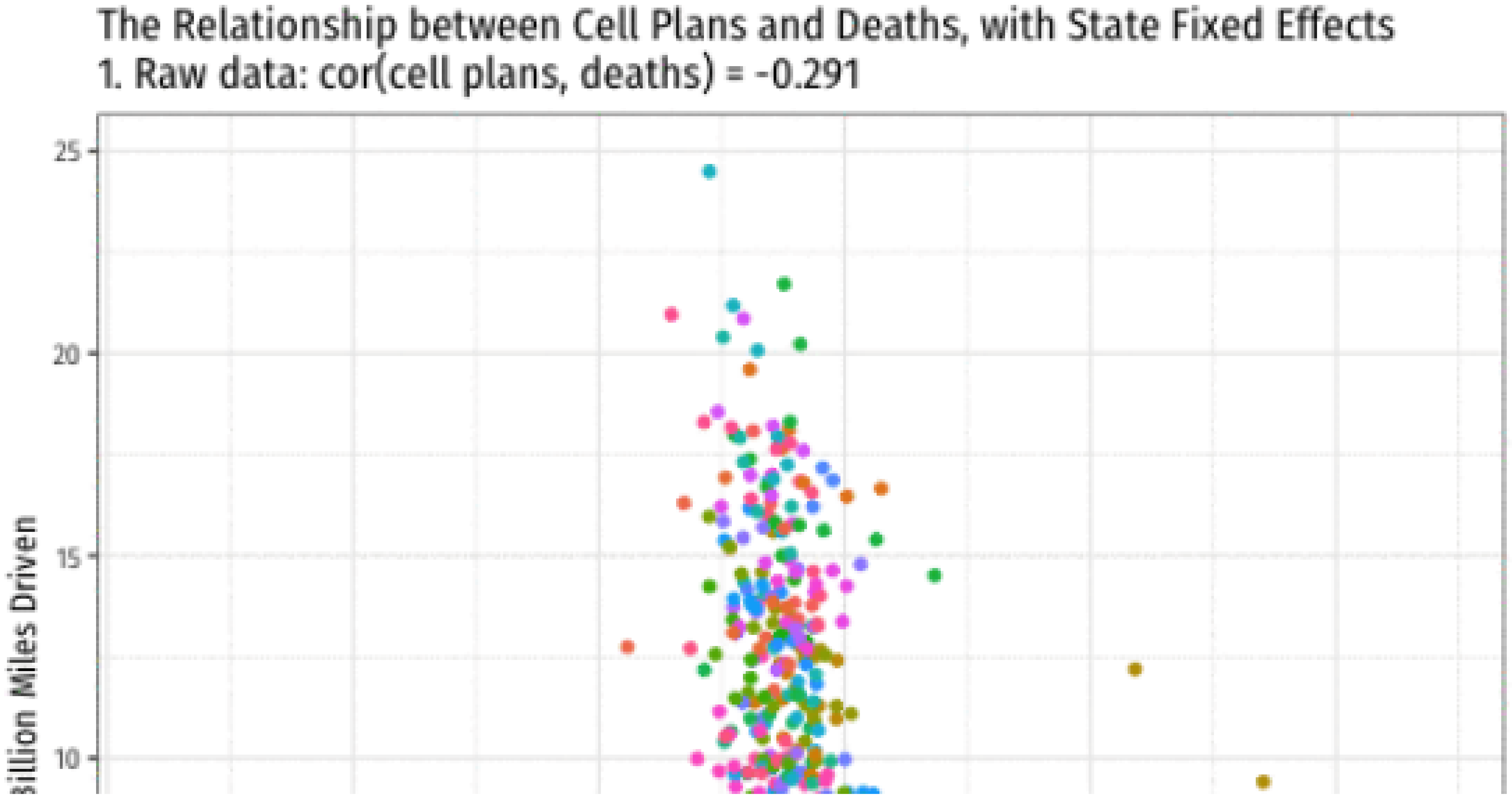
```r
ggplot(data = means_state)+
  aes(x = fct_reorder(state, avg_deaths),
      y = avg_deaths,
      color = state)+
  geom_point()+
  geom_segment(aes(y = 0,
                   yend = avg_deaths,
                   x = state,
                   xend = state))+
  coord_flip()+
  labs(x = "Cell Phones Per 10,000 People",
       y = "Deaths Per Billion Miles Driven",
       color = NULL)+
  theme_bw(base_family = "Fira Sans Condensed",
           base_size=10)+
  theme(legend.position = "none")
```



Visualizing "Within Group" Estimates for the 5 States



Visualizing "Within Group" Estimates for All 51 States



De-meaned Approach in R I

- The `fixest` package is designed for running regressions with fixed effects
- `feols()` function is just like `lm()`, with some additional arguments:

```
#install.packages("fixest")
library(fixest)
fe_reg_1_alt <- feols(deaths ~ cell_plans | state,
                      data = phones)
```

De-meaned Approach in R II

```r
fe_reg_1_alt %>% summary()
```

```
## OLS estimation, Dep. Var.: deaths
## Observations: 306
## Fixed-effects: state: 51
## Standard-errors: Clustered (state)
##             Estimate Std. Error  t value  Pr(>|t|)
## cell_plans -0.001204   0.000143 -8.41708 3.792e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 1.05007     Adj. R2: 0.886524
##                 Within R2: 0.357238
```

```r
# or using broom's tidy()
fe_reg_1_alt %>% tidy()
```

term	estimate	std.error	statistic	p.value
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
cell_plans	-0.001203742	0.0001430118	-8.417077	3.791955e-11

1 row

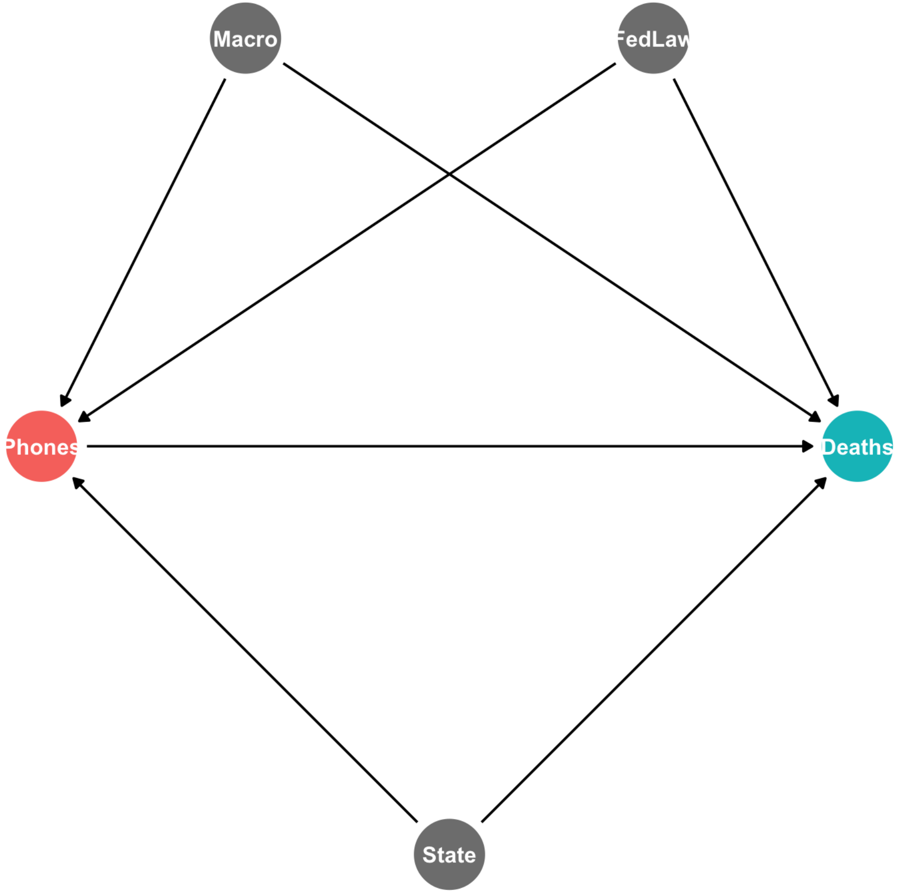
Two-Way Fixed Effects

Two-Way Fixed Effects

- State fixed effect controls for all factors that vary by state but are stable over time

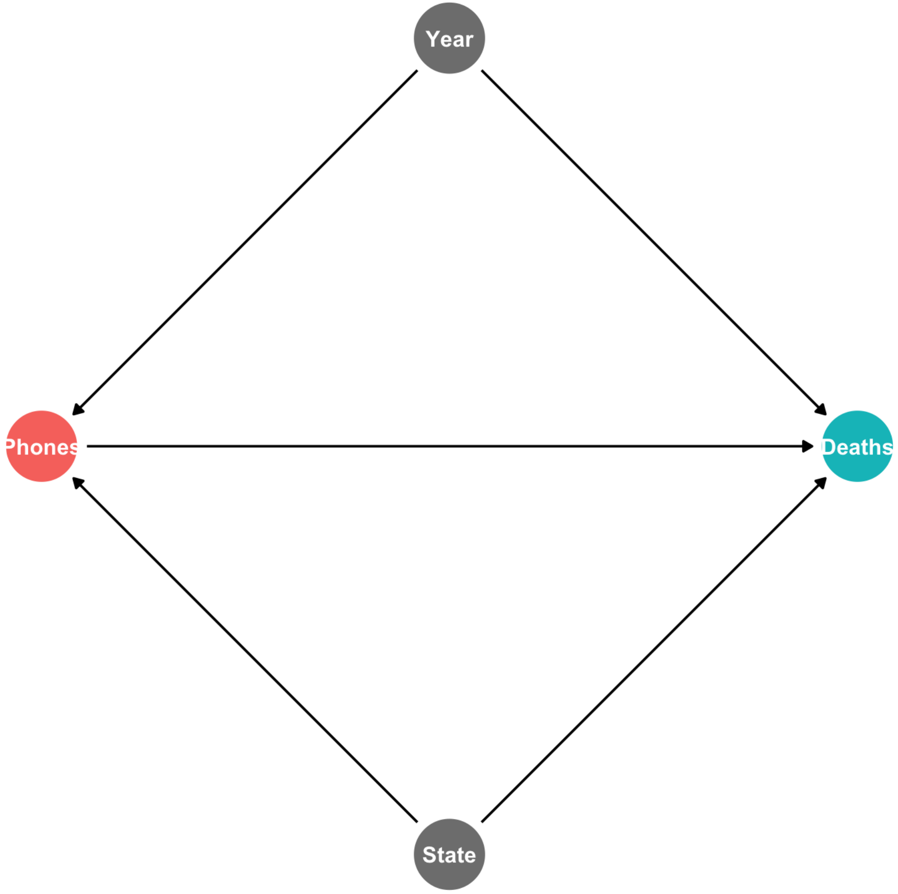
- But there are still other (often unobservable) factors that affect both Phones and Deaths, that *don't* vary by State

  - The country's macroeconomic performance, federal laws, etc



Two-Way Fixed Effects

- State fixed effect controls for all factors that vary by state but are stable over time
- But there are still other (often unobservable) factors that affect both Phones and Deaths, that *don't* vary by State
  - The country's macroeconomic performance, federal laws, etc
- If these factors systematically vary over time, but are the same by State, then we can **"control for Year"** to safely remove the influence of all of these factors!



Two-Way Fixed Effects

- A **one-way fixed effects model** estimates a fixed effect for **groups**
- **Two-way fixed effects model** estimates fixed effects for *both* **groups** *and* **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
  - accounts for **group-invariant differences over time**
- ν_{it} remaining random error

Two-Way Fixed Effects: Our 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** (note subscript i only)
  - e.g. geography, culture, (unchanging) state laws
- θ_t: Year fixed effects
  - differences **over time** that are **stable across states** (note subscript t only)
  - e.g. economy-wide macroeconomic changes, *federal* laws passed

Visualizing Year Effects I

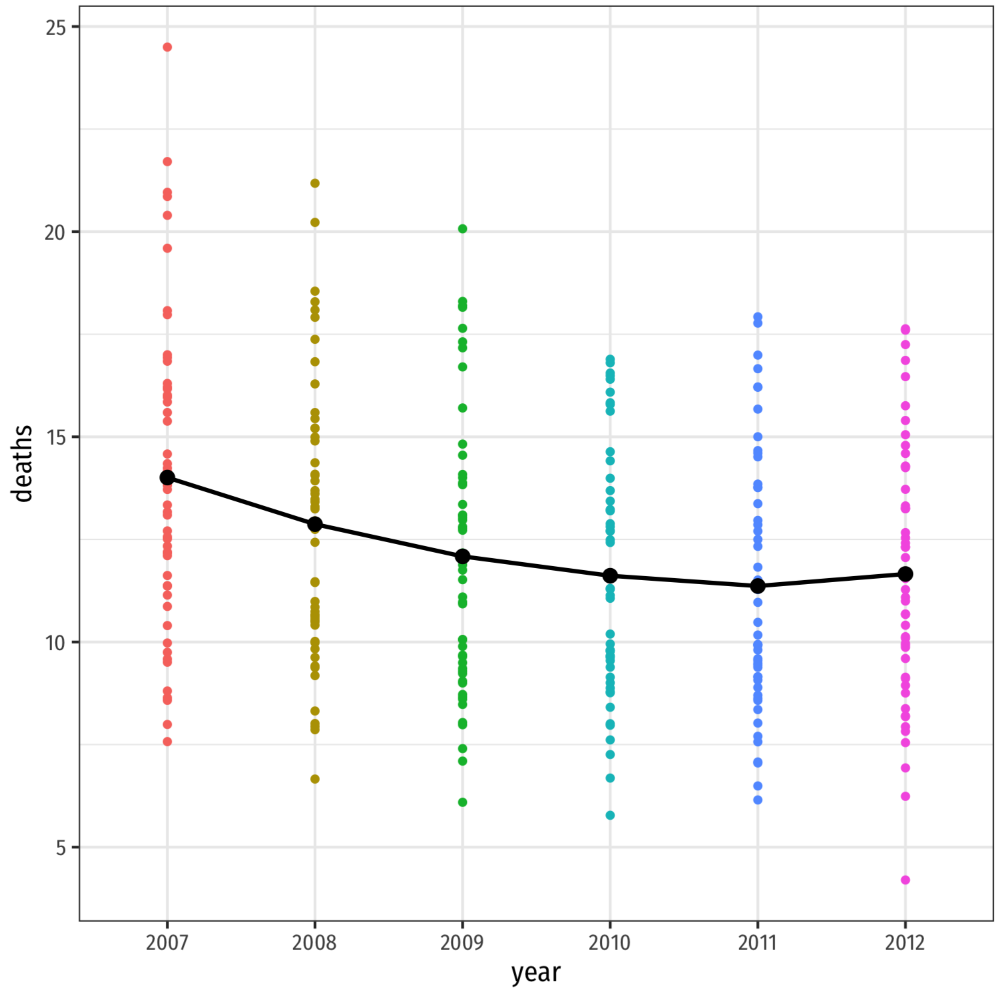
```r
# find averages for years
means_year <- phones %>%
  group_by(year) %>%
  summarize(avg_deaths = mean(deaths),
            avg_phones = mean(cell_plans))
means_year
```

year <fct>	avg_deaths <dbl>	avg_phones <dbl>
2007	14.00751	8064.531
2008	12.87156	8482.903
2009	12.08632	8859.706
2010	11.61487	9134.592
2011	11.36431	9485.238
2012	11.65666	9660.474

6 rows

Visualizing Year Effects II

```r
ggplot(data = phones)+
  aes(x = year,
      y = deaths)+
  geom_point(aes(color = year))+
  # Add the yearly means as black points
  geom_point(data = means_year,
             aes(x = year,
                 y = avg_deaths),
             size = 3,
             color = "black")+
  # connect the means with a line
  geom_line(data = means_year,
            aes(x = as.numeric(year),
                y = avg_deaths),
            color = "black",
            size = 1)+
  theme_bw(base_family = "Fira Sans Condensed",
           base_size = 14)+
  theme(legend.position = "none")
```



Estimating Two-Way Fixed Effects

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

- 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**:

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

where for each variable: $v\tilde{a}r_{it} = var_{it} - \overline{var}_t - \overline{var}_i$

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

LSDV Method

```
fe2_reg_1 <- lm(deaths ~ cell_plans + state + year,
                data = phones)

fe2_reg_1 %>% tidy()
```

term	estimate	std.error	statistic	p.value
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
(Intercept)	18.9304707399	1.4511323962	13.0453092	5.427406e-30
cell_plans	-0.0002995294	0.0001723149	-1.7382677	8.339982e-02
stateAlaska	-1.4998292482	0.6241082951	-2.4031554	1.698648e-02
stateArizona	-0.7791714713	0.6113519094	-1.2745057	2.036724e-01
stateArkansas	2.8655344756	0.5985062952	4.7878101	2.895040e-06
stateCalifornia	-5.0900897113	0.5956293282	-8.5457338	1.299236e-15
stateColorado	-4.4127241692	0.5953924847	-7.4114543	1.945083e-12
stateConnecticut	-6.6325834801	0.5952933996	-11.1417051	1.169797e-23
stateDelaware	-2.4579829953	0.5991822226	-4.1022295	5.546475e-05
stateDistrict of Columbia	-3.5044963616	1.9710939218	-1.7779449	7.663326e-02

1-10 of 57 rows

Previous **1** 2 3 4 5 6 Next

With fixest

```r
fe2_reg_2 <- feols(deaths ~ cell_plans | state + year,
                   data = phones)

fe2_reg_2 %>% summary()
```

```
## OLS estimation, Dep. Var.: deaths
## Observations: 306
## Fixed-effects: state: 51,  year: 6
## Standard-errors: Clustered (state)
##             Estimate Std. Error   t value Pr(>|t|)
## cell_plans   -3e-04   0.000305 -0.980739  0.33144
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 0.930036     Adj. R2: 0.909197
##                  Within R2: 0.011989
```

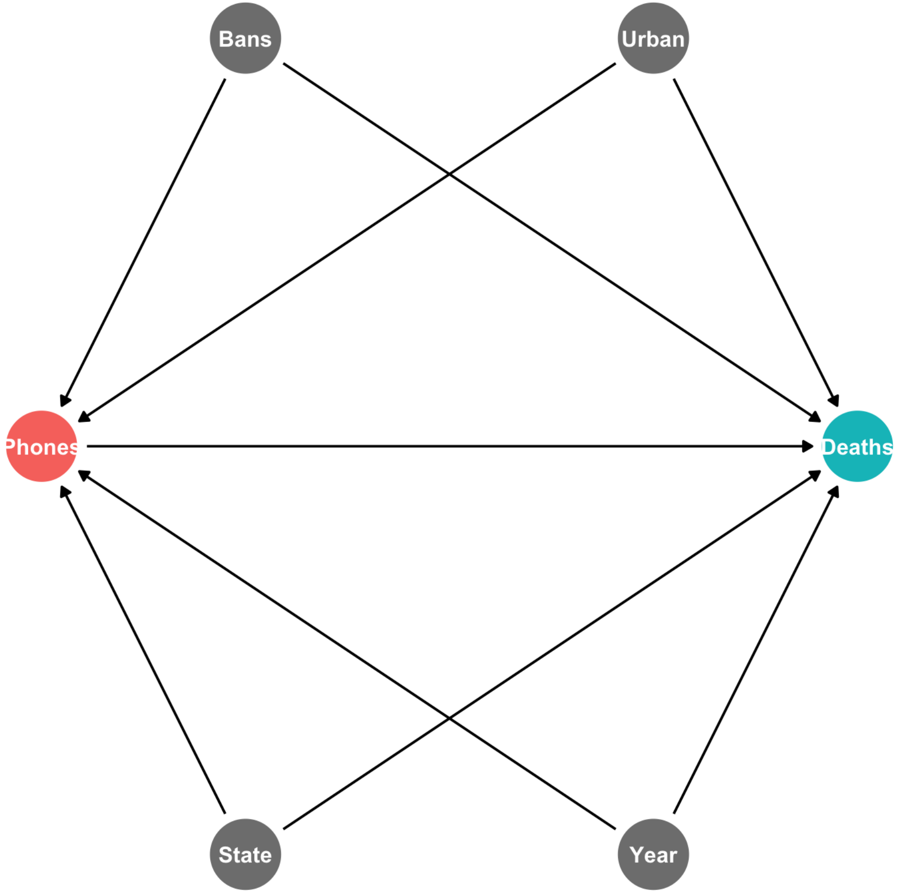
```r
fe2_reg_2 %>% tidy()
```

term	estimate	std.error	statistic	p.value
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
cell_plans	-0.0002995294	0.0003054118	-0.9807394	0.3314431

1 row

Adding Covariates

- State fixed effect absorbs all unobserved factors that vary by state, but are constant over time
- Year fixed effect absorbs all unobserved factors that vary by year, but are constant over States
- But there are still other (often unobservable) factors that affect both Phones and Deaths, that *vary* by State *and* change over time!
    - *Some* States *change* their laws during the time period
    - State *urbanization* rates *change* over the time period
- We will also need to **control for these variables** (*not* picked up by fixed effects!)
    - Add them to the regression



Adding Covariates I

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

- 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 II

```
fe2_controls_reg <- feols(deaths ~ cell_plans + text_ban + urban_percent + cell_ban | state + year,
                          data = phones)

fe2_controls_reg %>% summary()
```

```
## OLS estimation, Dep. Var.: deaths
## Observations: 306
## Fixed-effects: state: 51,  year: 6
## Standard-errors: Clustered (state)
##                Estimate Std. Error  t value Pr(>|t|)
## cell_plans    -0.000340   0.000277 -1.22780 0.225269
## text_ban1      0.255926   0.243444  1.05127 0.298188
## urban_percent  0.013135   0.009815  1.33822 0.186878
## cell_ban1     -0.679796   0.335655 -2.02528 0.048194 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 0.920123     Adj. R2: 0.910039
##                  Within R2: 0.032939
```

```
fe2_controls_reg %>% tidy()
```

term <chr>	estimate <dbl>	std.error <dbl>	statistic <dbl>	p.value <dbl>
cell_plans	-0.0003403735	0.0002772212	-1.227805	0.22526919
text_ban1	0.2559261569	0.2434442111	1.051272	0.29818803
urban_percent	0.0131347657	0.0098150705	1.338224	0.18687751
cell_ban1	-0.6797956522	0.3356553662	-2.025279	0.04819377

Comparing Models

```r
library(huxtable)
huxreg("Pooled" = pooled,
       "State Effects" = fe_reg_1,
       "State & Year Effects" = fe2_reg_1,
       "With Controls" = fe2_controls_reg,
       coefs = c("Intercept" = "(Intercept)",
                 "Cell phones" = "cell_plans",
                 "Cell Ban" = "cell_ban1",
                 "Texting Ban" = "text_ban1",
                 "Urbanization Rate" = "urban_percent"),
       statistics = c("N" = "nobs",
                      "R-Squared" = "r.squared",
                      "SER" = "sigma"),
       number_format = 4)
```

	Pooled	State Effects	State & Year Effects	With Controls
Intercept	17.3371 ***	25.5077 ***	18.9305 ***	
	(0.9754)	(1.0176)	(1.4511)	
Cell phones	-0.0006 ***	-0.0012 ***	-0.0003	-0.0003
	(0.0001)	(0.0001)	(0.0002)	(0.0003)
Cell Ban				-0.6798 *
				(0.3357)
Texting Ban				0.2559
				(0.2434)
Urbanization Rate				0.0131
				(0.0098)
N	306	306	306	306
R-Squared	0.0845	0.9055	0.9259	0.9274
SER	3.2791	1.1526	1.0310	1.0262