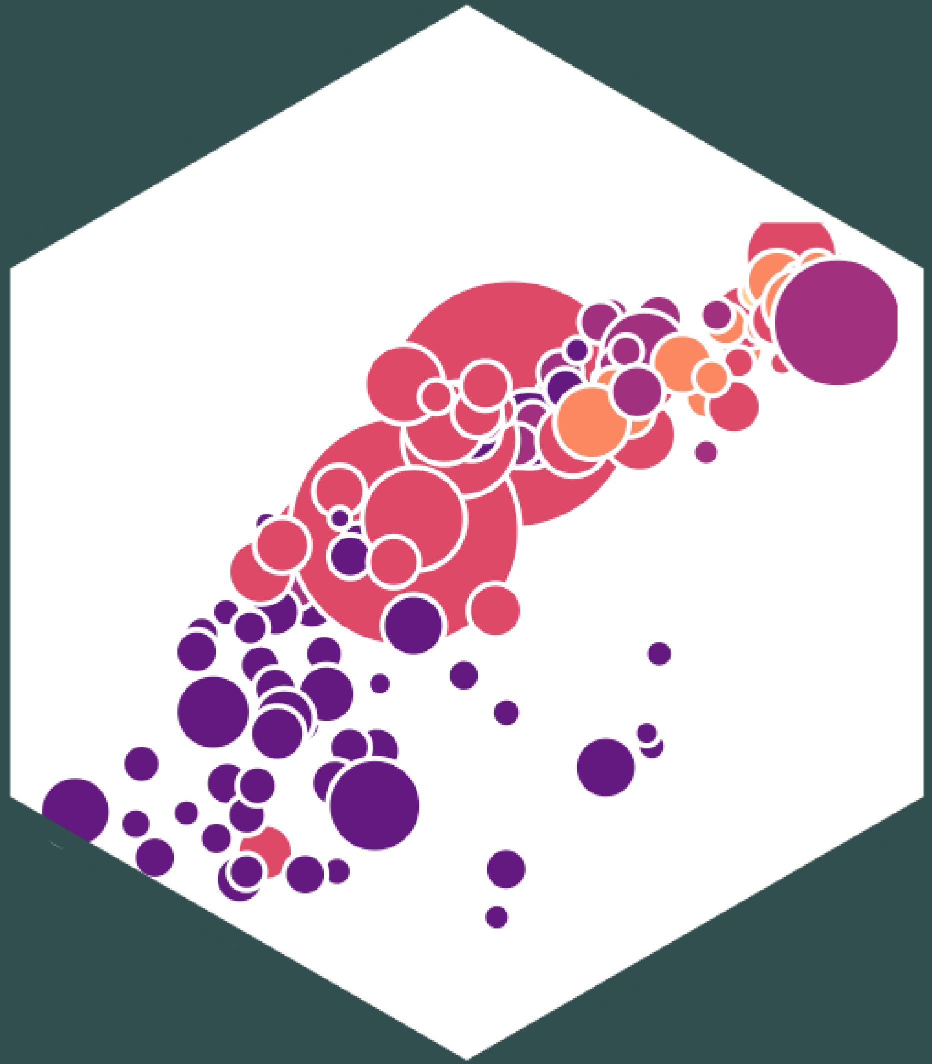
5.1 — Fixed Effects

ECON 480 • Econometrics • Fall 2022

Dr. Ryan Safner
Associate Professor of Economics

safner@hood.edu
ryansafner/metricsF22
metricsF22.classes.ryansafner.com



Contents

Panel Data

Types of Data I

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

state <fct>
Alabama
Alaska
Arizona
Arkansas
California
Colorado

6 rows | 1-1 of 4 columns

Types of Data I

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

state <fct>
Alabama
Alaska
Arizona
Arkansas
California
Colorado

6 rows | 1-1 of 4 columns

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

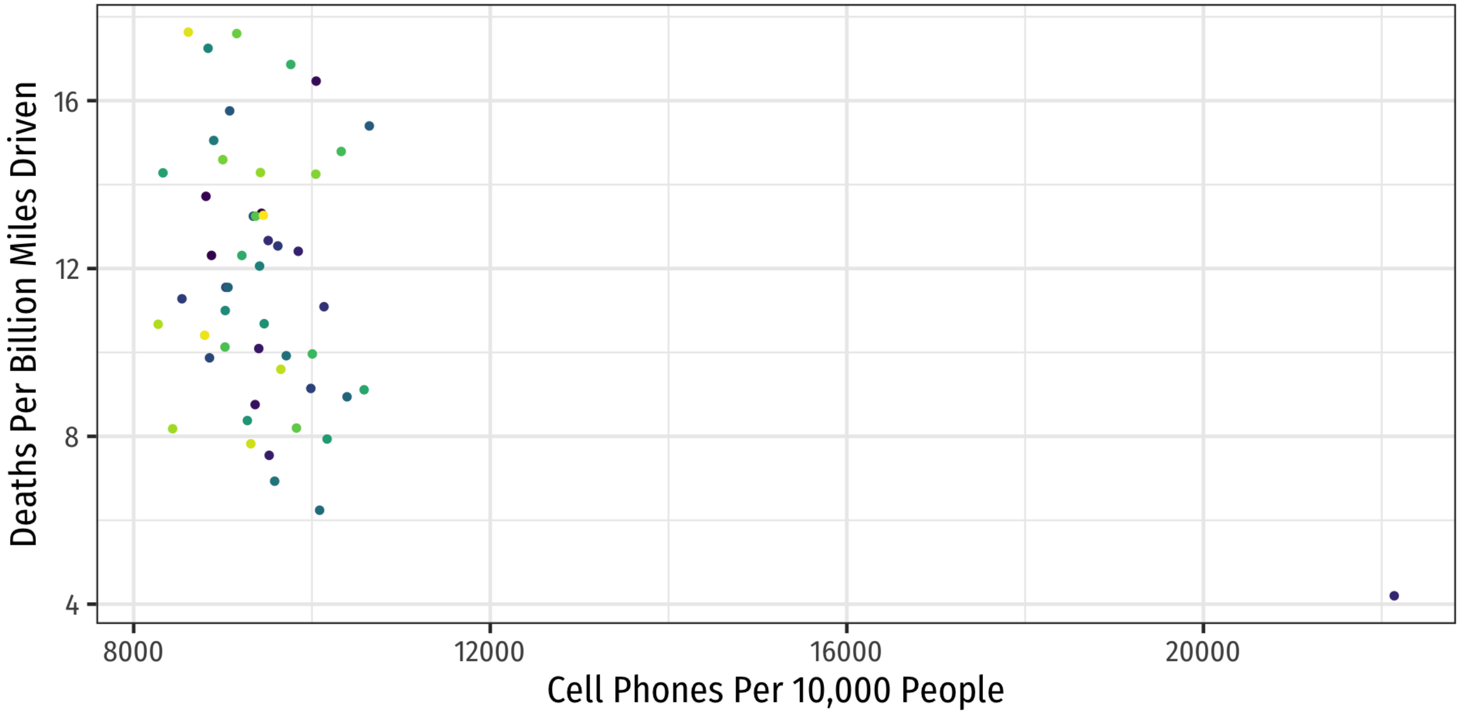
state <fct>
Maryland
Maryland
Maryland
Maryland
Maryland
Maryland

6 rows | 1-1 of 4 columns

Types of Data II

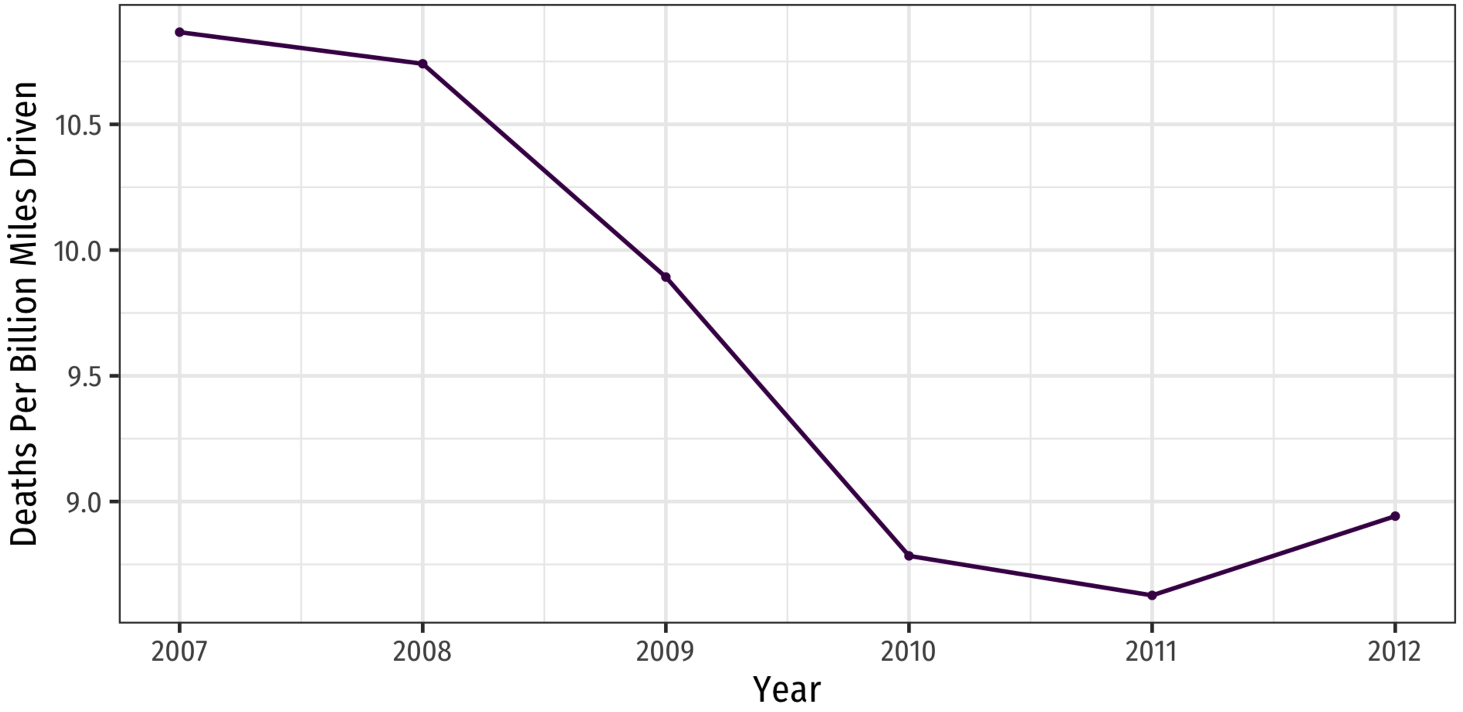
- **Cross-sectional data**: compare different individual i's at same time $\bar{t}$

$$\hat{Y}_i = \beta_0 + \beta_1 X_i + u_i$$



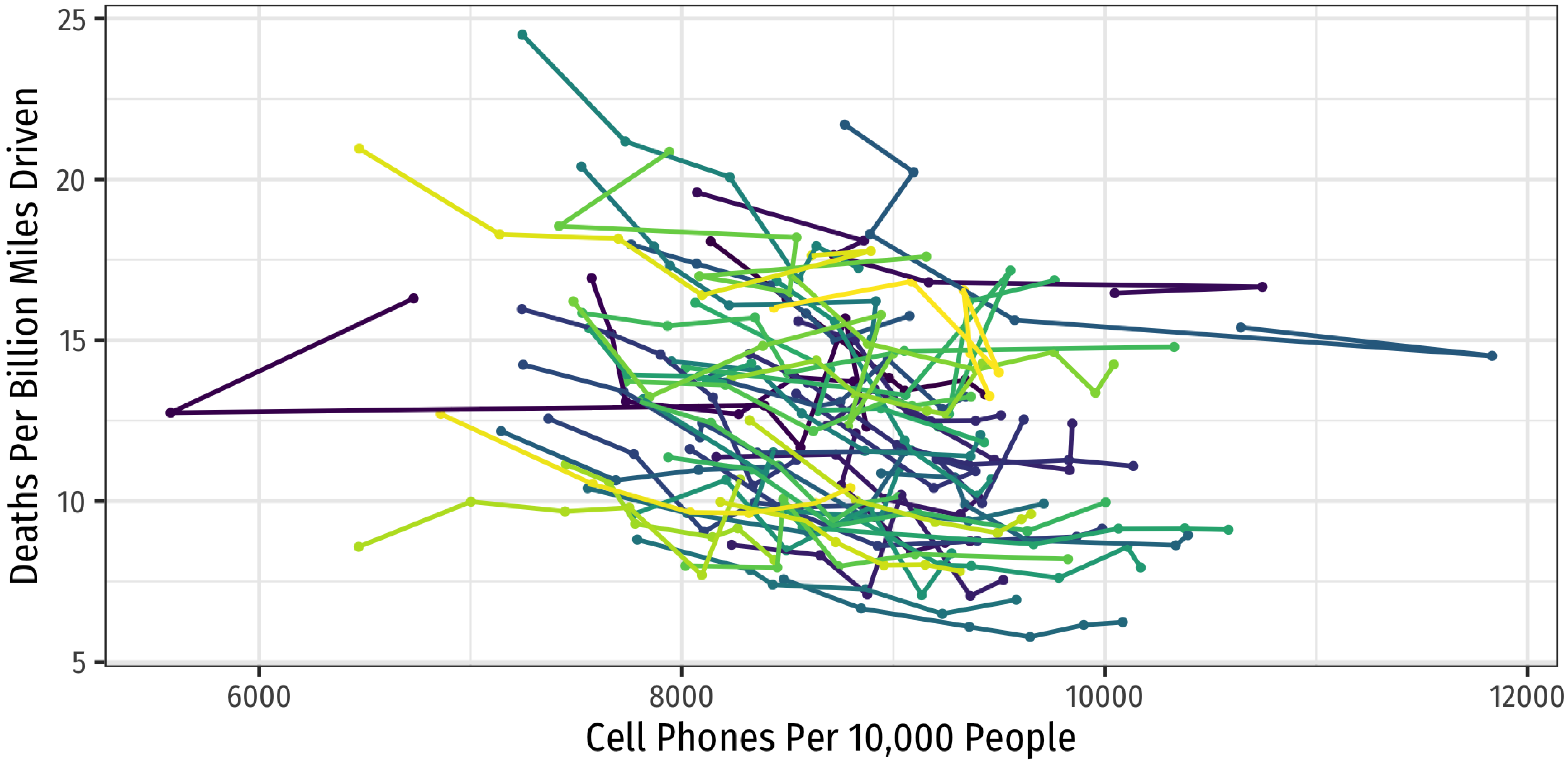
- **Time-series data**: track same individual $\bar{i}$ over different times t

$$\hat{Y}_t = \beta_0 + \beta_1 X_t + u_t$$



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

Panel Data I


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

Panel Data II

state <fct>
Alabama
Alabama
Alabama
Alabama
Alabama
Alabama
Alaska
Alaska
Alaska
Alaska

1-10 of 306 rows | 1-1 of 4 columns Previous **1** 2 3 ... 31 Next

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

Panel Data II

state <fct>
Alabama
Alabama
Alabama
Alabama
Alabama
Alabama
Alaska
Alaska
Alaska
Alaska

1-10 of 306 rows | 1-1 of 4 columns Previous **1** 2 3 ... 31 Next

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

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

for individual i in time t.

Panel Data: Our Motivating Example

state <fct>
Alabama
Alabama
Alabama
Alabama
Alabama
Alabama
Alaska
Alaska
Alaska
Alaska

1-10 of 306 rows | 1-1 of 4 columns Previous **1** 2 3 ... 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
- U.S. State-level data over 6 years

The Data I

```r
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

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

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

1-10 of 51 rows | 1-1 of 2 columns   Previous **1** 2 3 … 6 Next

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

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

6 rows

The Data III

```r
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 ... 6 Next

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

year <fct>
2007
2008
2009
2010
2011
2012

6 rows

The Data IV

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

States	Years
<int>	<int>
51	6

1 row

Pooled Regression

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

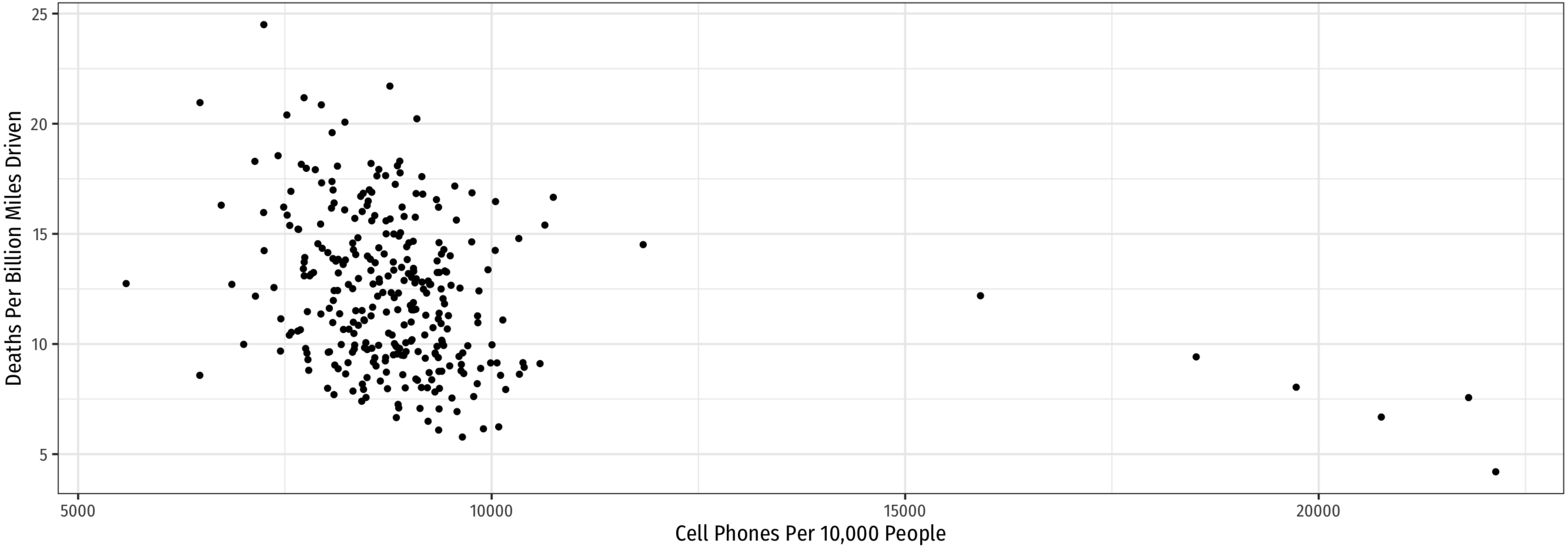
```r
pooled <- lm(deaths ~ cell_plans, data = phones)
pooled %>% tidy()
```

term <chr>	estimate <dbl>
(Intercept)	17.3371034167
cell_plans	-0.0005666385

2 rows | 1-2 of 5 columns

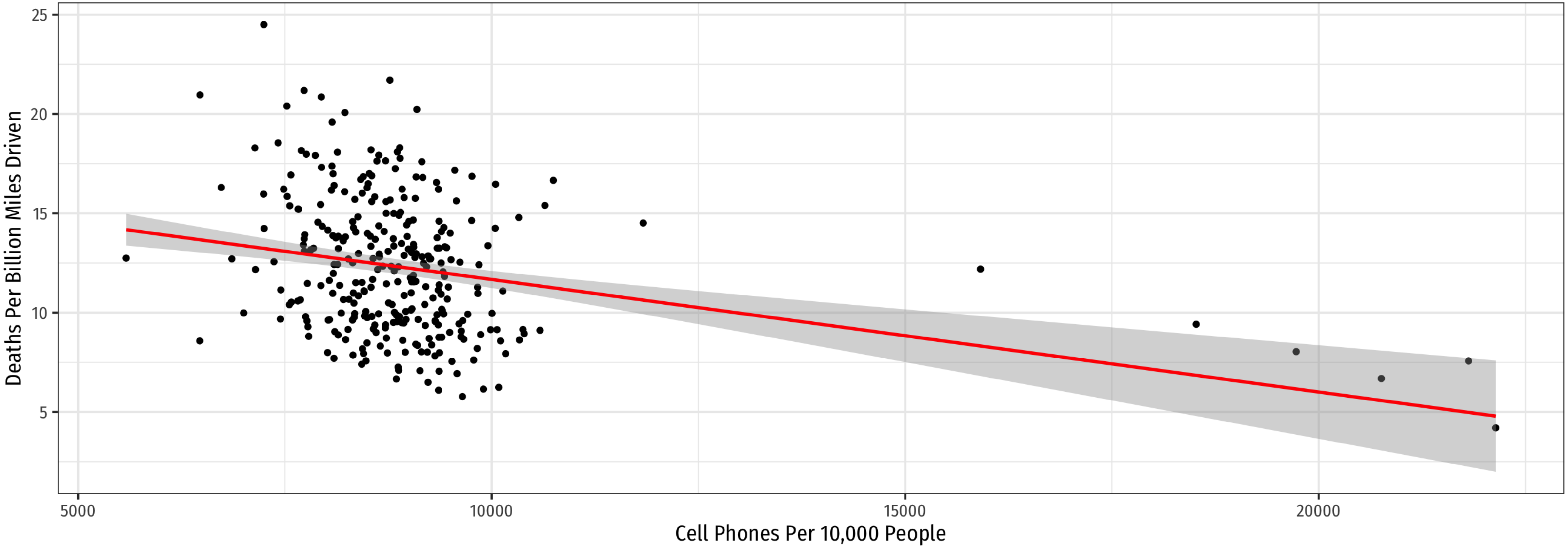
Pooled Regression III

- Code



Pooled Regression III

- Code



Recall: Assumptions about Errors

- We make **4 critical assumptions about u**:

1. The expected value of the errors is 0

$$\mathbb{E}[u] = 0$$

2. The variance of the errors 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

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

- **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 come from same state i**
  - Probably similarities among u_t for obs in same state i
  - Residuals on observations from same state are likely correlated

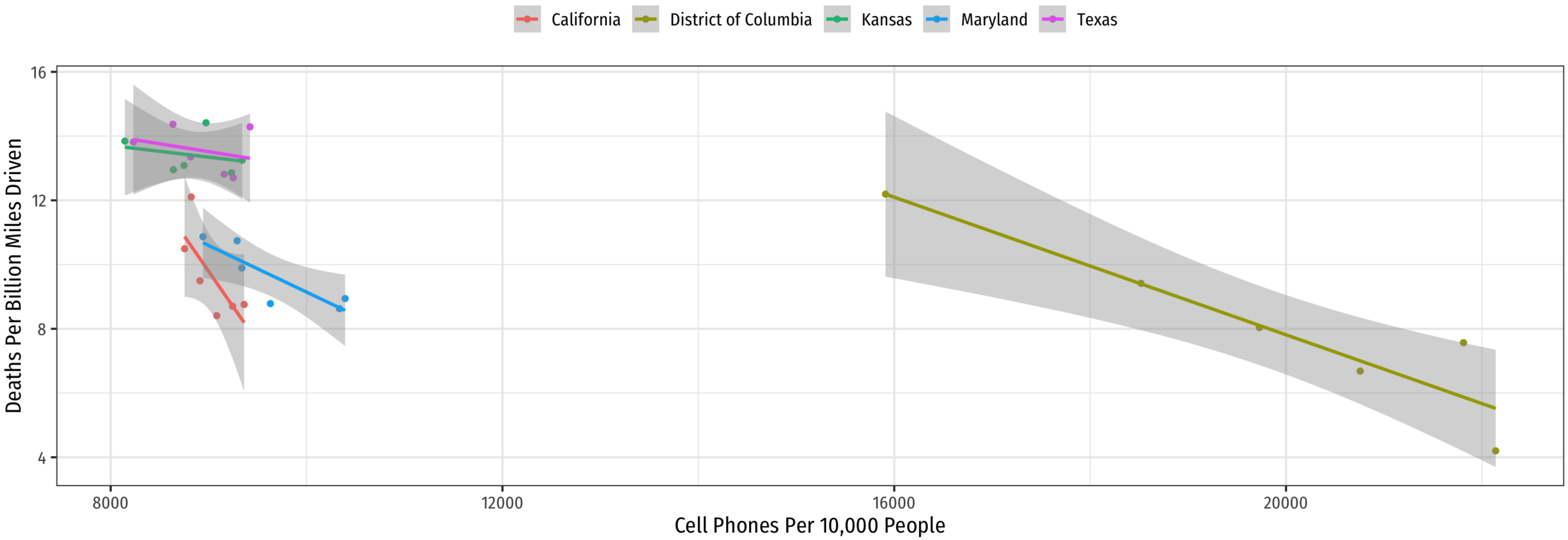
$$cor(u_{\text{MD, 2008}}, u_{\text{MD, 2009}}) \neq 0$$

- **Multiple observations come from same year t**
  - Probably similarities among u_i for obs in same year t
  - Residuals on observations from same year are likely correlated

$$cor(u_{\text{MD, 2008}}, u_{\text{VA, 2008}}) \neq 0$$

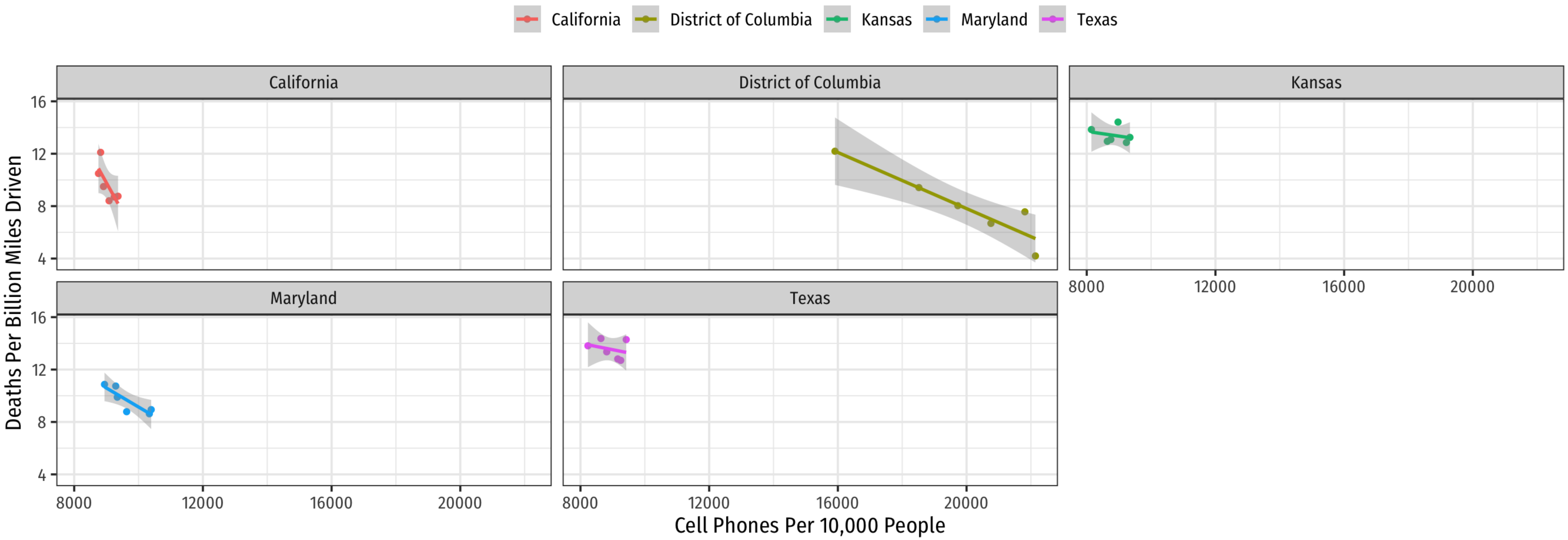
Example: Consider Just 5 States

- Code



Example: Consider Just 5 States

- Code



Example: Consider All 51 States

- Code

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**:

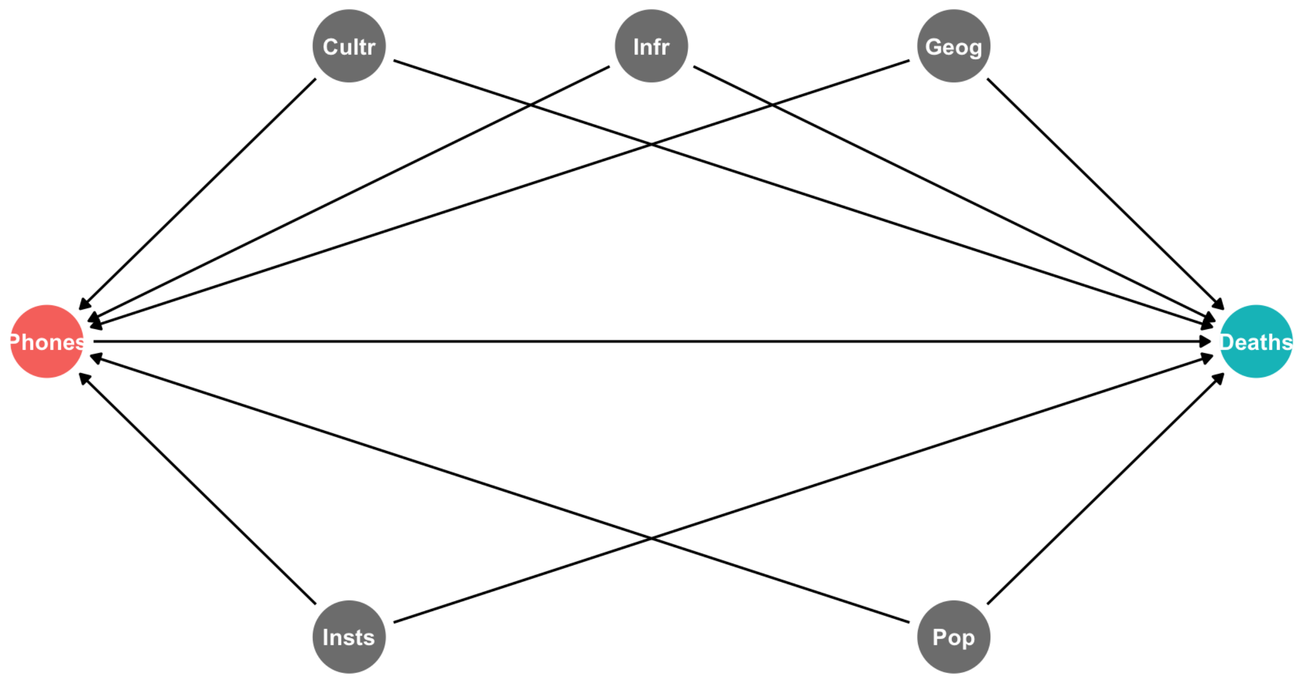
$$cor(\text{u}_{it}, \text{Cell Phones}_{it}) \neq 0 \qquad E[\text{u}_{it}|\text{Cell Phones}_{it}] \neq 0$$

- 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

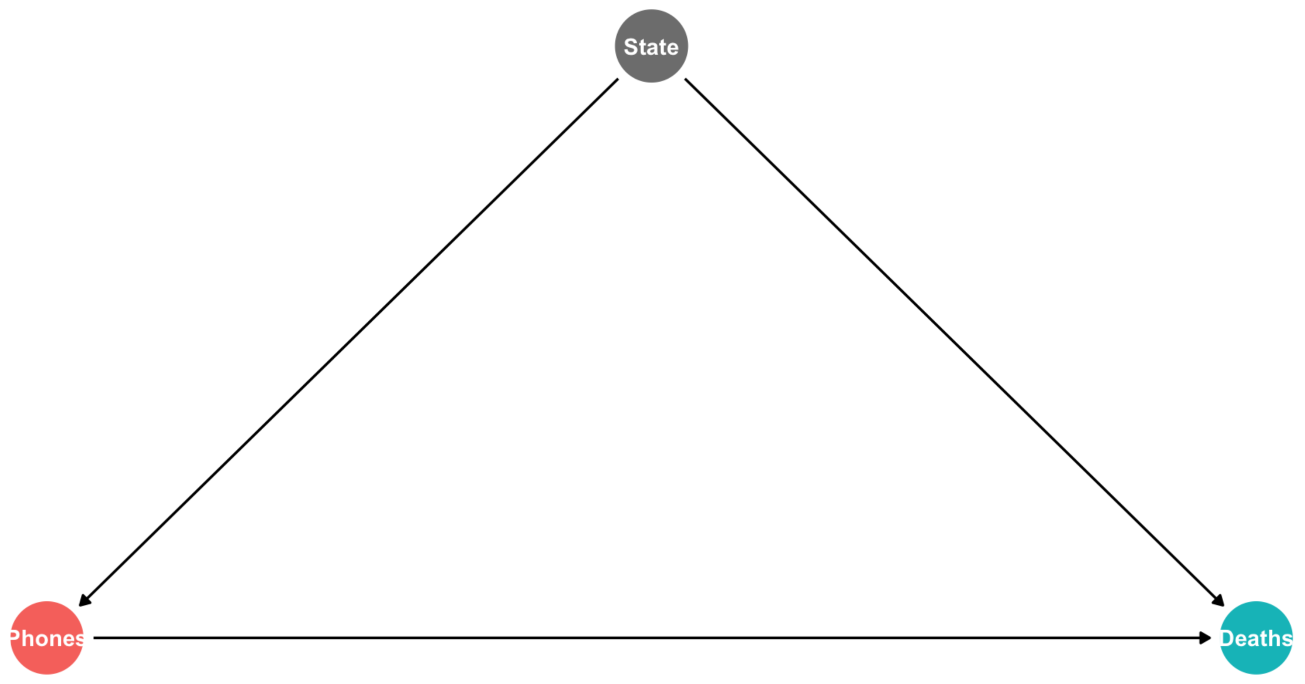
Fixed Effects: DAG I

- 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 II

- 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 ("intercept") 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:

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

- 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

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

- 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

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

- 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

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

- Create a dummy variable $D_i = \{0, 1\}$ for each possible group,
$$\begin{cases} = 1 & \text{if observation } it \text{ is from group } i \\ = 0 & \text{otherwise} \end{cases}$$

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

- Sounds like a lot of work, automatic in R

1. 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 dummy for each of the other U.S. States

Our Example in R

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

- If `state` variable is a `factor`, can 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 (Alabama)

Our Example in R: Regression I

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

term <chr>	estimate <dbl>
(Intercept)	25.507679925
cell_plans	-0.001203742
stateAlaska	-2.484164783
stateArizona	-1.510577383
stateArkansas	3.192662931
stateCalifornia	-4.978668651
stateColorado	-4.344553493
stateConnecticut	-6.595185530
stateDelaware	-2.098393628
stateDistrict of Columbia	6.355790010

1-10 of 52 rows | 1-2 of 5 columns

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

Our Example in R: Regression II

```
fe_reg_1 %>% glance()
```

r.squared <dbl>	adj.r.squared <dbl>	sigma <dbl>	statistic <dbl>
0.9054987	0.886524	1.152558	47.72144

1 row | 1-4 of 12 columns

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 mean of each variable (over time, t):

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

- $\bar{Y}_i$: 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

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

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

$$Y_{it} - \bar{Y}_i = \beta_1 (X_{it} - \bar{X}_i) + \alpha_i + \epsilon_{it} - \bar{\alpha}_i - \bar{\epsilon}_{it}$$
$$\tilde{Y}_{it} = \beta_1 \tilde{X}_{it} + \tilde{\epsilon}_{it}$$

- Within each group i, the de-meaned variables $\tilde{Y}_{it}$ and $\tilde{X}_{it}$'s all have a mean of 0[1]
- Variables that don't change over time will drop out of analysis altogether
- **Removes any source of variation <u>across</u> groups (all now have mean of 0) to only work with variation <u>within</u> each group**

1. Recall **Rule 4** from the 2.3 class appendix 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

Looking at the Data in R I

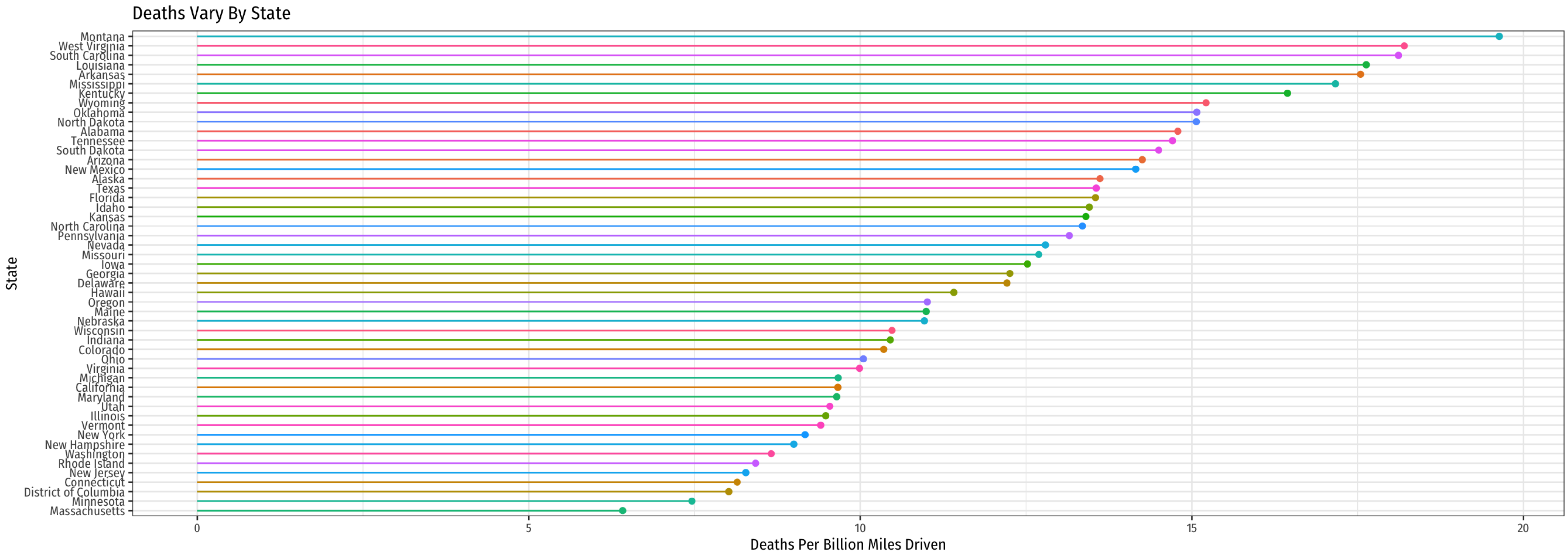
```r
# get means of Y and X by state
means_state <- phones %>%
  group_by(state) %>%
  summarize(avg_deaths = mean(deaths),
            avg_phones = mean(cell_plans))

# 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

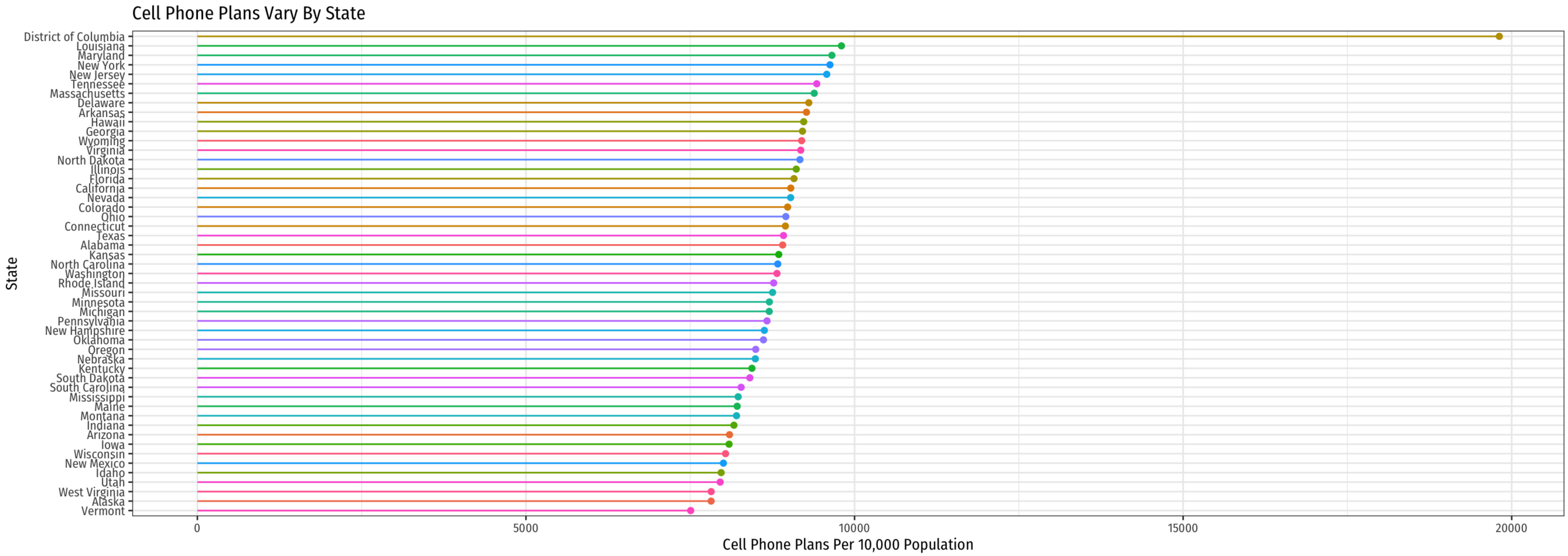
Looking at the Data in R II

- Code



Looking at the Data in R III

- Code



De-Meaning the Data in R

```r
phones_dm <- phones %>%
  select(state, year, cell_plans, deaths) %>%
  group_by(state) %>% # for each state...
  mutate(phones_dm = cell_plans - mean(cell_plans), # de-mean X
         deaths_dm = deaths - mean(deaths)) # de-mean Y
phones_dm
```

state <fct>	year <fct>	cell_plans <dbl>
Alabama	2007	8135.525
Alaska	2007	6730.282
Arizona	2007	7572.465
Arkansas	2007	8071.125
California	2007	8821.933
Colorado	2007	8162.065
Connecticut	2007	8234.567
Delaware	2007	8684.450
District of Columbia	2007	15910.466
Florida	2007	8550.103

1-10 of 306 rows | 1-3 of 6 columns

Previous **1** 2 3 4 5 6 ... 31 Next

De-Meaning the Data in R II

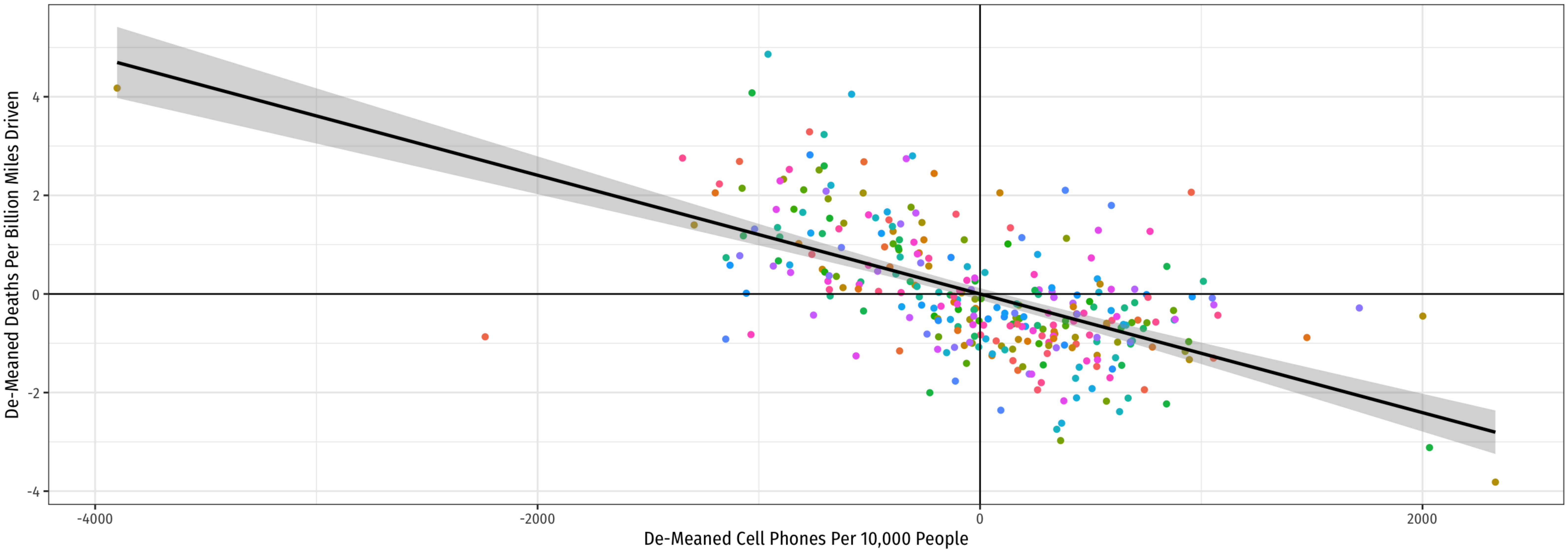
```
phones_dm %>%
 #ungroup() %>% # it's still grouped by state
  summarize(mean_deaths = round(mean(deaths_dm),2), sd_deaths = round(sd(deaths_dm),2), mean_phones = round(mean(phones_dm),2), sd_phones = round(sd(phone
```

state <fct>	mean_deaths <dbl>	sd_deaths <dbl>
Alabama	0	1.95
Alaska	0	1.90
Arizona	0	1.57
Arkansas	0	1.18
California	0	1.41
Colorado	0	0.85
Connecticut	0	1.19
Delaware	0	0.94
District of Columbia	0	2.68
Florida	0	1.38

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

De-Meaning the Data in R: Visualizing

- Code



De-Meaning the Data in R: Regression I

term <chr>	estimate <dbl>
(Intercept)	-8.618515e-16
phones_dm	-1.203742e-03

2 rows | 1-2 of 5 columns

De-Meaning the Data in R: Regression II

r.squared	adj.r.squared	sigma	statistic
<dbl>	<dbl>	<dbl>	<dbl>
0.3572378	0.3551234	1.05352	168.9587

1 row | 1-4 of 12 columns

Using `fixest` I

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

```r
library(fixest)
feols(y ~ x | g, # after |, g is the group variable
      data = df)
```

Using `fixest` II

```r
fe_reg_1_alt <- feols(deaths ~ cell_plans | state,
                      data = phones)

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
fe_reg_1_alt %>% tidy()
```

term <chr>	estimate <dbl>	std.error <dbl>
cell_plans	-0.001203742	0.0001430118

1 row | 1-3 of 5 columns

Comparing FE Approaches

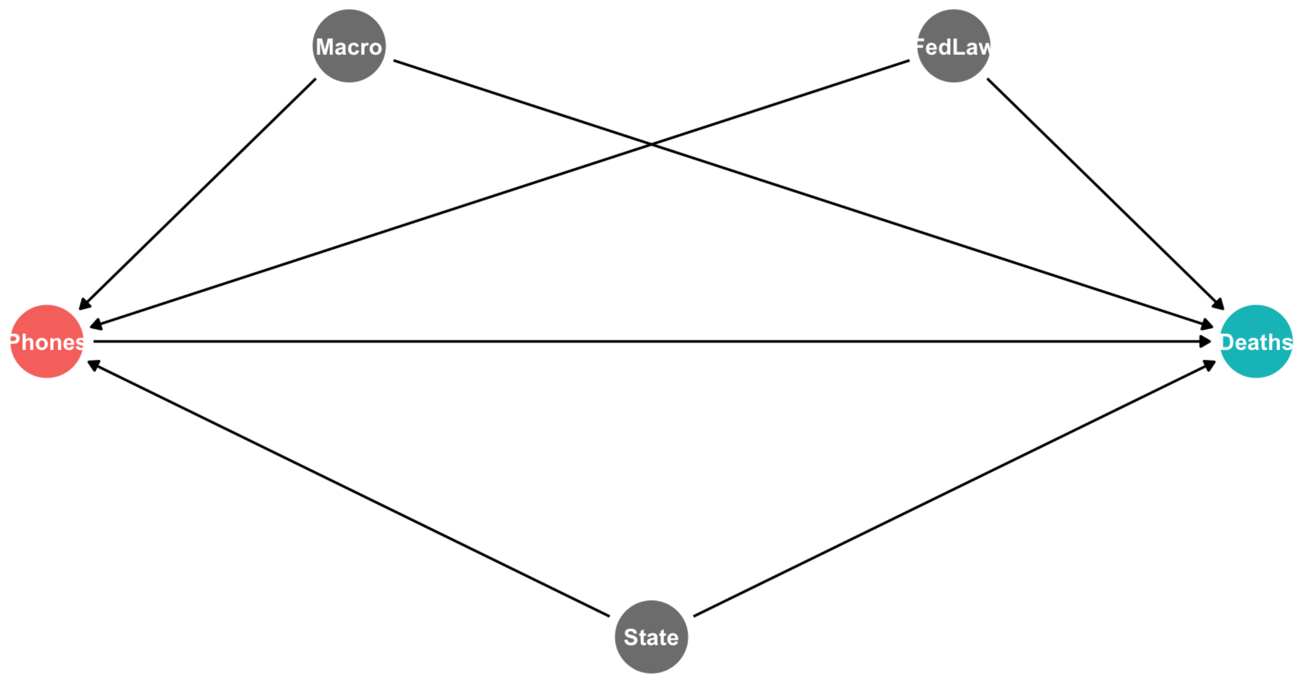
	Pooled Regression	FE: LSDV Method	FE: De-Meaned	FE: fixest
Constant	17.33710***	25.50768***	0.00000	
	(0.97538)	(1.01764)	(0.06023)	
Cell Phone Plans	−0.00057***	−0.00120***	−0.00120***	−0.00120***
	(0.00011)	(0.00010)	(0.00009)	(0.00014)
n	306	306	306	306
Adj. R^2	0.08	0.89	0.36	
SER	3.27	1.05	1.05	1.05

* p < 0.1, ** p < 0.05, *** p < 0.01

Two-Way Fixed Effects

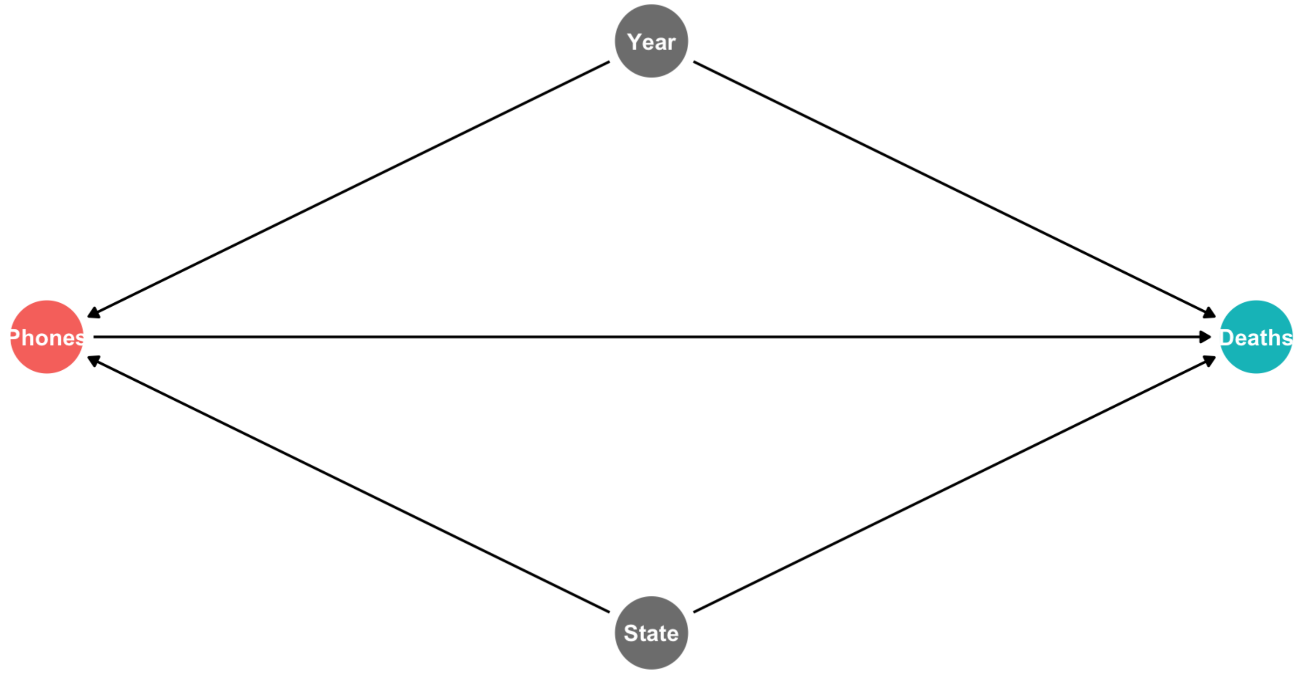
Two-Way Fixed Effects

- 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 (TWFE)** 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
  - all remaining factors that affect Y_{it} that vary by state *and* change over time

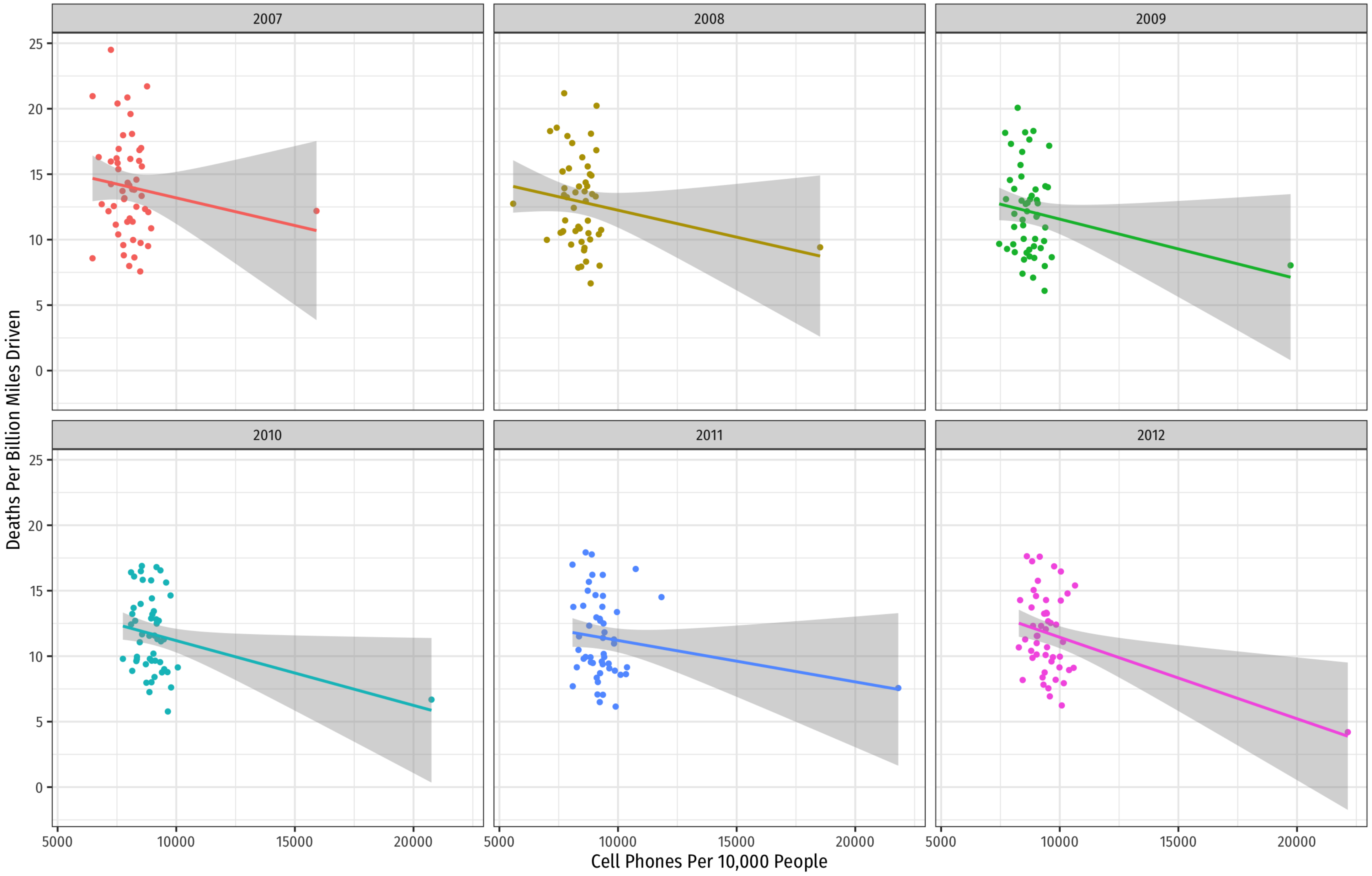
Two-Way Fixed Effects: Our Example

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

- 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

Looking at the Data: Change Over Time

- Code



Looking at the Data: Change Over Time II

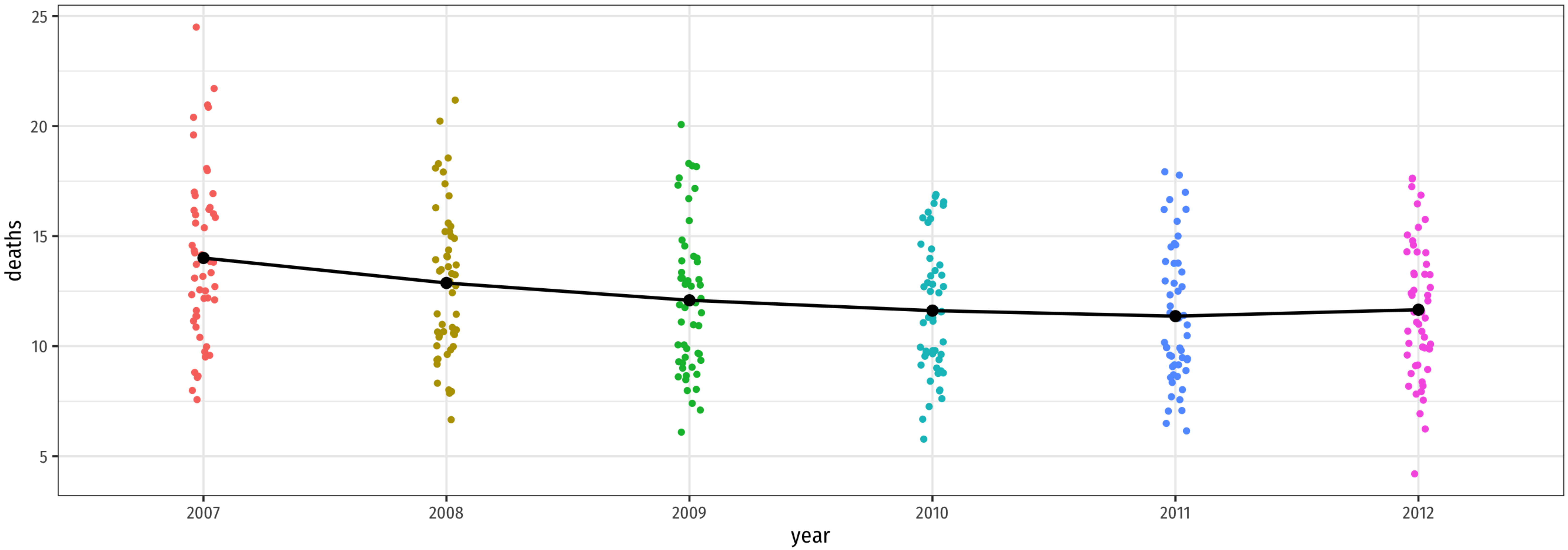
```r
means_year <- phones %>%
  group_by(year) %>%
  summarize(avg_deaths = mean(deaths),
            avg_phones = mean(cell_plans))
means_year
```

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

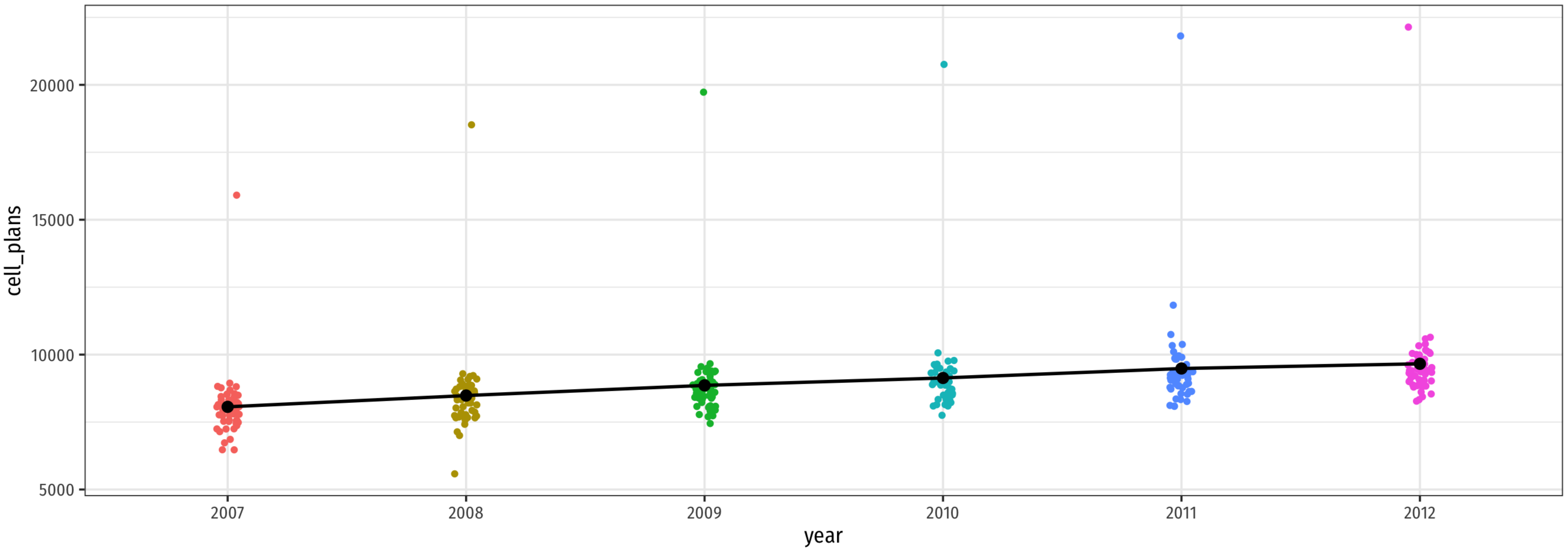
Looking at the Data: Change In *Deaths* Over Time

▶ Code



Looking at the Data: Change in *Cell Phones* Over Time

- Code



Estimating Two-Way Fixed Effects

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

- 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: $\widetilde{var}_{it} = var_{it} - \overline{var}_t - \overline{var}_i$

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

LSDV Method

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

fe2_reg_1 %>% tidy()
```

term <chr>	estimate <dbl>
(Intercept)	18.9304707399
cell_plans	-0.0002995294
stateAlaska	-1.4998292482
stateArizona	-0.7791714713
stateArkansas	2.8655344756
stateCalifornia	-5.0900897113
stateColorado	-4.4127241692
stateConnecticut	-6.6325834801
stateDelaware	-2.4579829953
stateDistrict of Columbia	-3.5044963616

1-10 of 57 rows | 1-2 of 5 columns 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
```

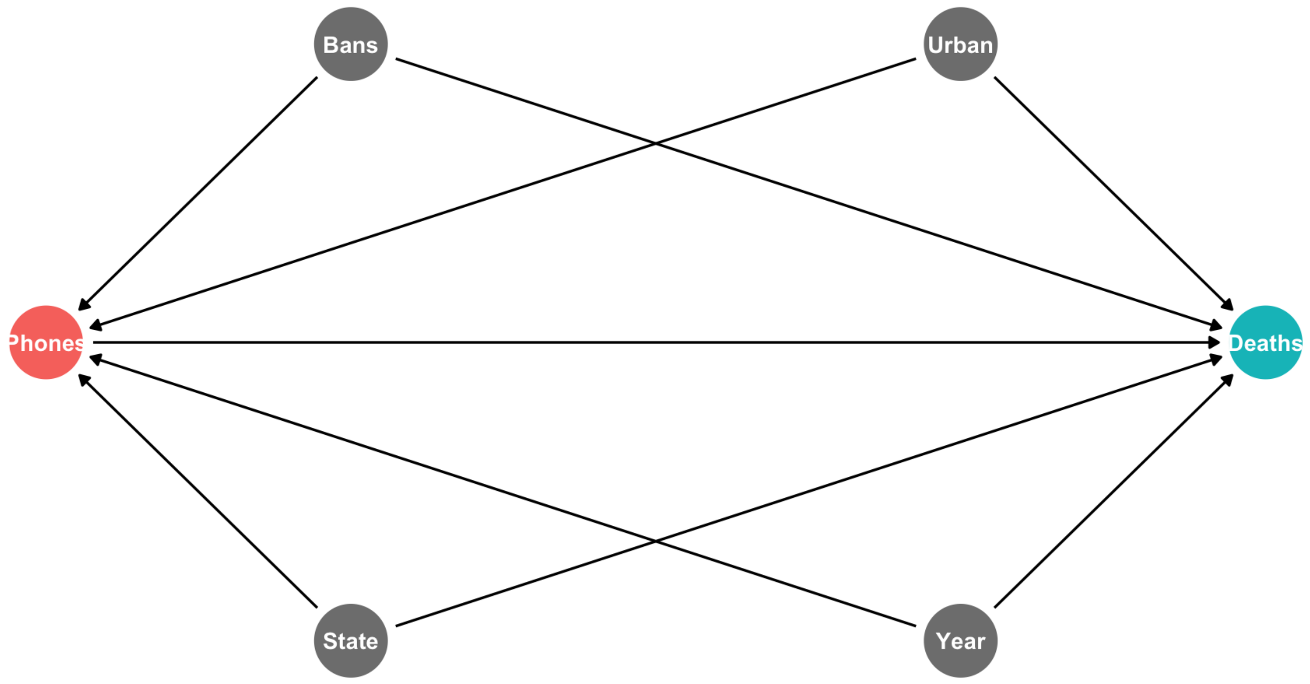
```r
fe2_reg_2 %>% tidy()
```

term <chr>	estimate <dbl>
cell_plans	-0.0002995294

1 row | 1-2 of 5 columns

Adding Covariates I

- 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 — Necessary?

```
phones %>%
  group_by(year) %>%
  count(cell_ban) %>%
  pivot_wider(names_from = cell_ban, values_from = n) %>%
  rename(`States Without a Ban` = `0`,
         `States With Cell Phone Ban` = `1`)
```

year <fct>	States Without a Ban <int>
2007	46
2008	46
2009	44
2010	43
2011	41
2012	40

6 rows | 1-2 of 3 columns

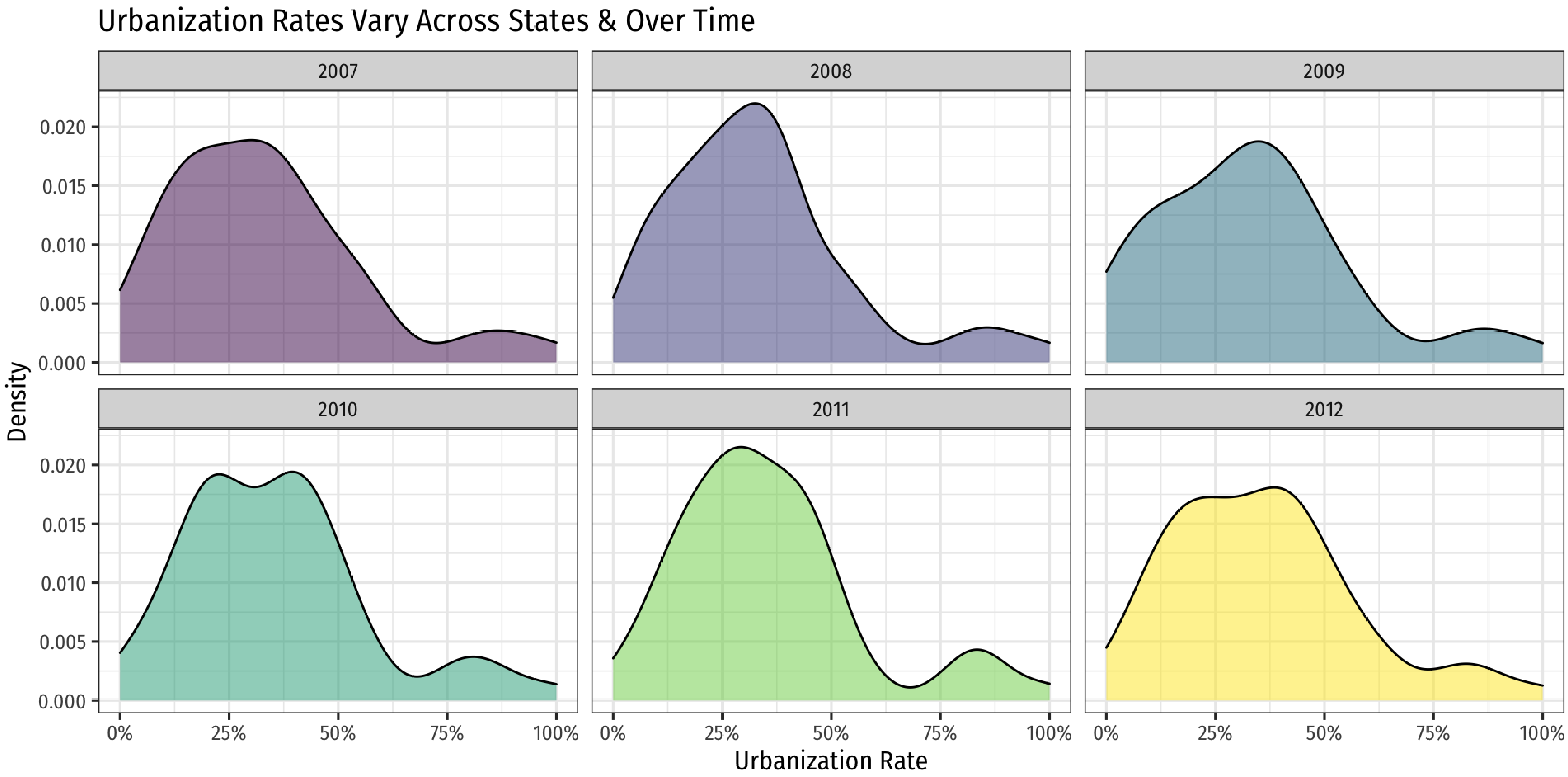
Adding Covariates — Necessary?

```
phones %>%
  group_by(year) %>%
  count(text_ban) %>%
  pivot_wider(names_from = text_ban, values_from = n) %>%
  rename(`States Without a Ban` = `0`,
         `States With a Texting Ban` = `1`)
```

year <fct>	States Without a Ban <int>
2007	49
2008	47
2009	42
2010	30
2011	20
2012	16

6 rows | 1-2 of 3 columns

Adding Covariates — Necessary?



Adding Covariates II

$$\widehat{\text{Deaths}}_{it} = \beta_1 \text{ Cell Phones}_{it} + \alpha_i + \theta_t + \beta_2 \text{ urban pct}_{it} + \beta_3 \text{ cell ban}_{it} + \beta_4 \text{ text ban}_{it}$$

- 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 III (`fixest`)

```r
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
```

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

term <chr>	estimate <dbl>	std.error <dbl>
cell_plans	-0.0003403735	0.0002772212
text_ban1	0.2559261569	0.2434442111
urban_percent	0.0131347657	0.0098150705
cell_ban1	-0.6797956522	0.3356553662

4 rows | 1-3 of 5 columns

Comparing Models

	Pooled Regression	State FE	State & Year FE	TWFE with Controls
Constant	17.33710***			
	(0.97538)			
Cell Phone Plans	-0.00057***	-0.00120***	-3e-04	-0.00034
	(0.00011)	(0.00014)	(0.00031)	(0.00028)
text_ban1				0.25593
				(0.24344)
urban_percent				0.01313
				(0.00982)
cell_ban1				-0.67980**
				(0.33566)
n	306	306	306	306
Adj. R^2	0.08			
SER	3.27	1.05	0.93	0.92

* p < 0.1, ** p < 0.05, *** p < 0.01