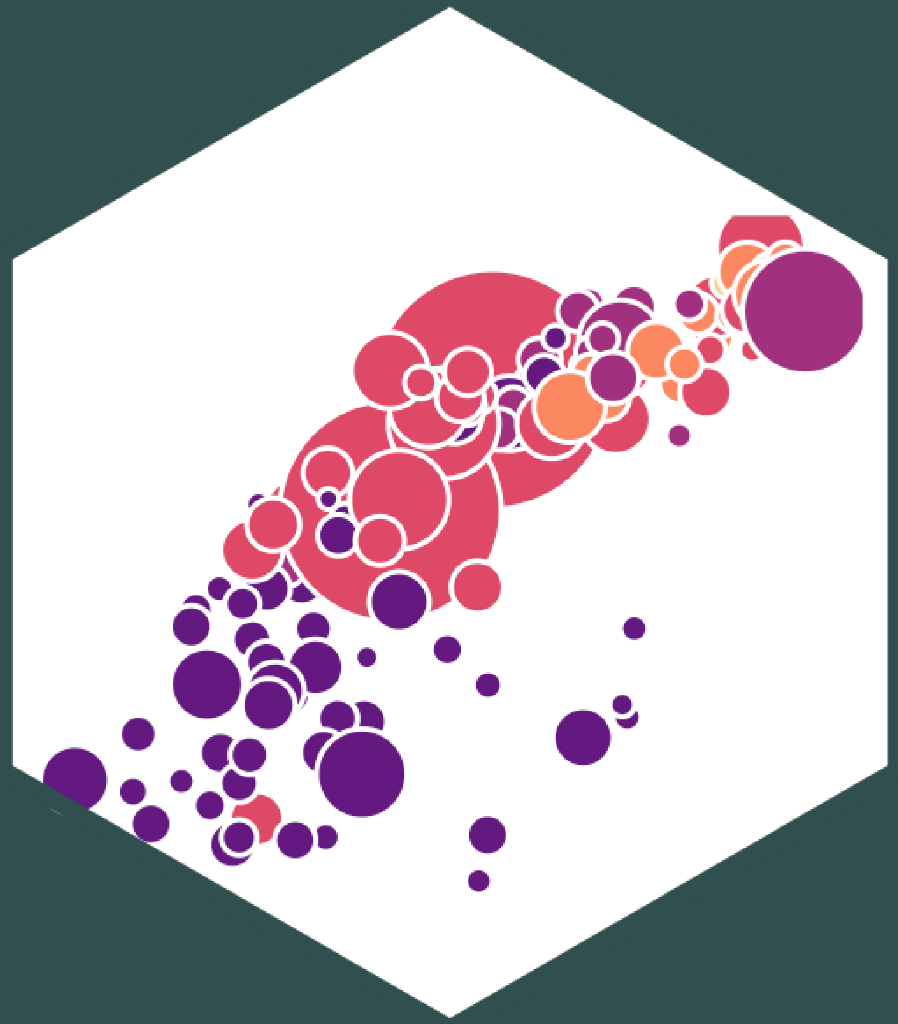
QM — Fixed Effects

Dr. Zahid Asghar
Professor of Economics
zasghar@qau.edu.pk
zahedasghar
zahidasghar.com



Contents

Panel Data

Types of Data I

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

```
# A tibble: 6 × 4
  state      year  deaths cell_plans
  <fct>      <fct>  <dbl>      <dbl>
1 Alabama    2012   13.3       9434.
2 Alaska     2012   12.3       8873.
3 Arizona    2012   13.7       8811.
4 Arkansas   2012   16.5      10047.
5 California 2012    8.76      9362.
6 Colorado   2012   10.1       9403.
```

Types of Data I

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

```
# A tibble: 6 × 4
  state       year  deaths cell_plans
  <fct>       <fct>  <dbl>      <dbl>
1 Alabama     2012   13.3       9434.
2 Alaska      2012   12.3       8873.
3 Arizona     2012   13.7       8811.
4 Arkansas    2012   16.5      10047.
5 California  2012    8.76      9362.
6 Colorado    2012   10.1       9403.
```

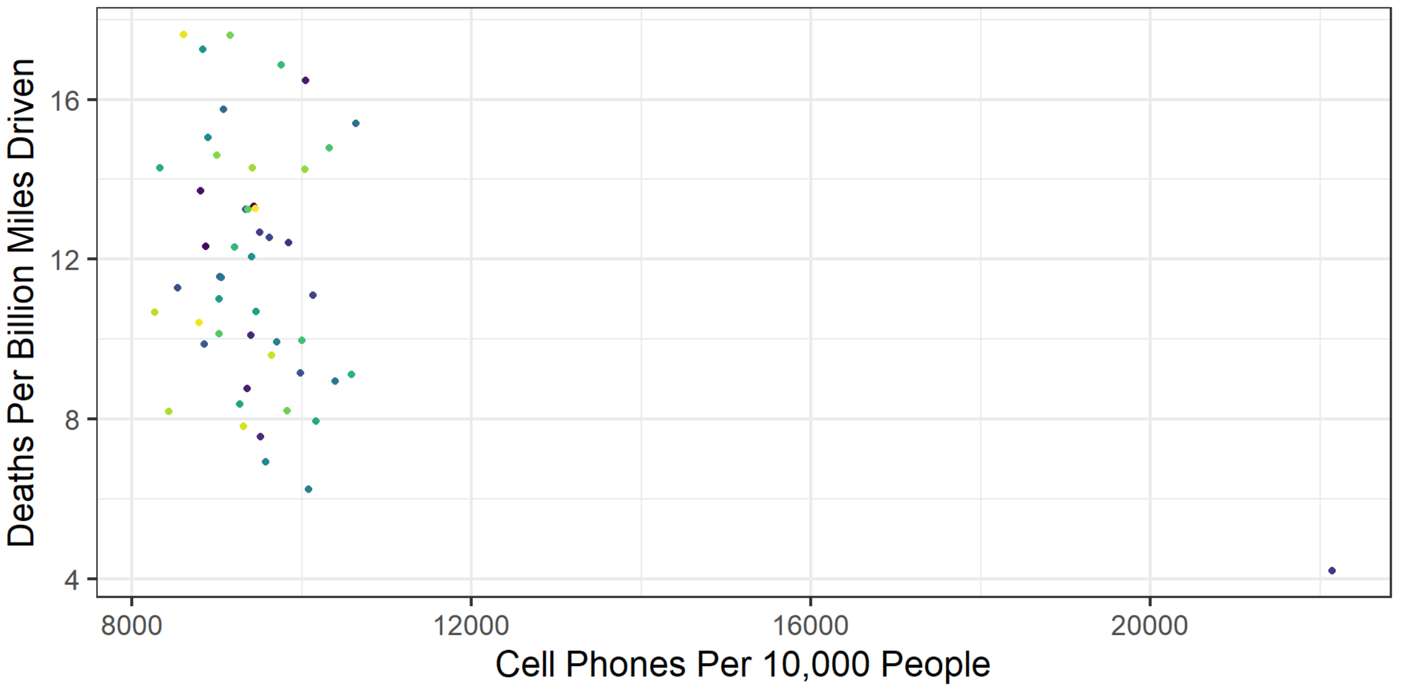
- **Time-series data**: track same individual $\bar{i}$ over different times t

```
# A tibble: 6 × 4
  state     year  deaths cell_plans
  <fct>     <fct>  <dbl>      <dbl>
1 Maryland 2007   10.9       8942.
2 Maryland 2008   10.7       9291.
3 Maryland 2009    9.89      9339.
4 Maryland 2010    8.78      9630.
5 Maryland 2011    8.63     10336.
6 Maryland 2012    8.94     10393.
```

Types of Data II

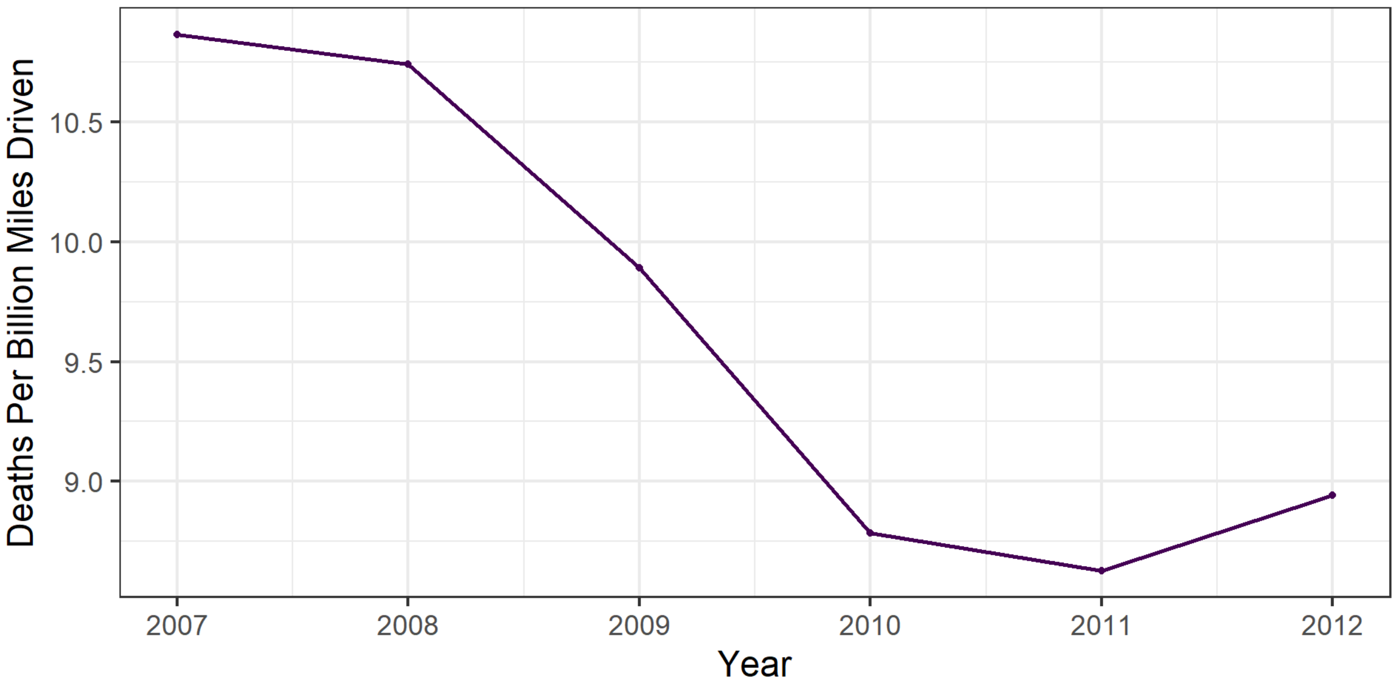
- **Cross-sectional data**: compare different individual i's at same time $\bar{t}$

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



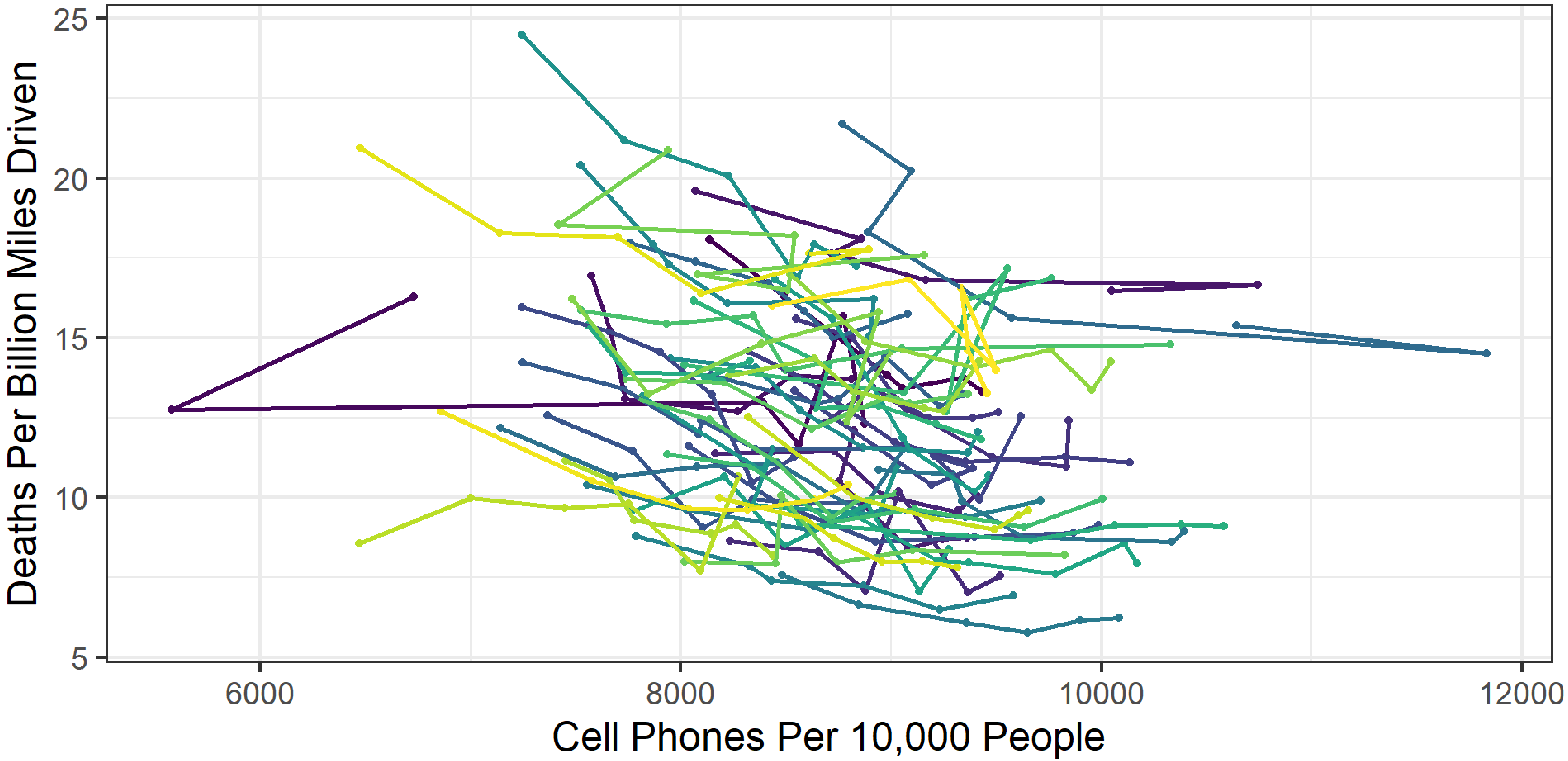
- **Time-series data**: track same individual $\bar{i}$ over different times t

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



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

Panel Data I



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

Panel Data II

```
# A tibble: 306 × 4
   state    year  deaths cell_plans
   <fct>    <fct>  <dbl>      <dbl>
 1 Alabama 2007     18.1      8136.
 2 Alabama 2008     16.3      8494.
 3 Alabama 2009     13.8      8979.
 4 Alabama 2010     13.4      9055.
 5 Alabama 2011     13.8      9341.
 6 Alabama 2012     13.3      9434.
 7 Alaska  2007     16.3      6730.
 8 Alaska  2008     12.7      5581.
 9 Alaska  2009     13.0      8390.
10 Alaska  2010     11.7      8561.
# … with 296 more rows
```

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

Panel Data II

```
# A tibble: 306 × 4
   state    year  deaths cell_plans
   <fct>    <fct>  <dbl>      <dbl>
 1 Alabama 2007    18.1       8136.
 2 Alabama 2008    16.3       8494.
 3 Alabama 2009    13.8       8979.
 4 Alabama 2010    13.4       9055.
 5 Alabama 2011    13.8       9341.
 6 Alabama 2012    13.3       9434.
 7 Alaska  2007    16.3       6730.
 8 Alaska  2008    12.7       5581.
 9 Alaska  2009    13.0       8390.
10 Alaska  2010    11.7       8561.
# … with 296 more rows
```

- **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

```
# A tibble: 306 × 4
   state    year  deaths cell_plans
   <fct>    <fct>  <dbl>      <dbl>
 1 Alabama 2007    18.1       8136.
 2 Alabama 2008    16.3       8494.
 3 Alabama 2009    13.8       8979.
 4 Alabama 2010    13.4       9055.
 5 Alabama 2011    13.8       9341.
 6 Alabama 2012    13.3       9434.
 7 Alaska  2007    16.3       6730.
 8 Alaska  2008    12.7       5581.
 9 Alaska  2009    13.0       8390.
10 Alaska  2010    11.7       8561.
# … with 296 more rows
```

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

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

```
# A tibble: 51 × 2
   state                    n
   <fct>                <int>
 1 Alabama                  6
 2 Alaska                   6
 3 Arizona                  6
 4 Arkansas                 6
 5 California               6
 6 Colorado                 6
 7 Connecticut              6
 8 Delaware                 6
 9 District of Columbia     6
10 Florida                  6
# … with 41 more rows
```

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

```
# A tibble: 6 × 2
  year      n
  <fct> <int>
1 2007     51
2 2008     51
3 2009     51
4 2010     51
5 2011     51
6 2012     51
```

The Data III

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

```
# A tibble: 51 × 1
   state
   <fct>
 1 Alabama
 2 Alaska
 3 Arizona
 4 Arkansas
 5 California
 6 Colorado
 7 Connecticut
 8 Delaware
 9 District of Columbia
10 Florida
# … with 41 more rows
```

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

```
# A tibble: 6 × 1
  year
  <fct>
1 2007
2 2008
3 2009
4 2010
5 2011
6 2012
```

The Data IV

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

```
# A tibble: 1 × 2
  States Years
   <int> <int>
1     51     6
```

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

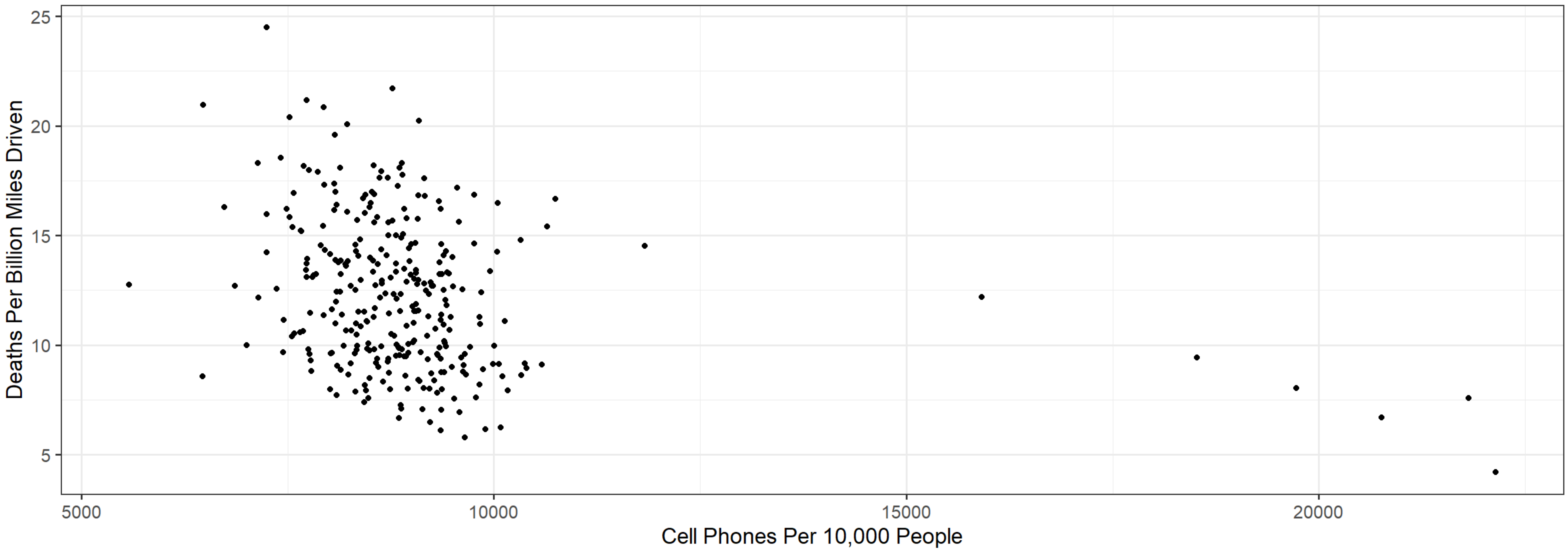
Pooled Regression II

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

```
# A tibble: 2 × 5
  term         estimate std.error statistic  p.value
  <chr>           <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept) 17.3       0.975         17.8  5.82e-49
2 cell_plans  -0.000567  0.000107      -5.30 2.26e- 7
```

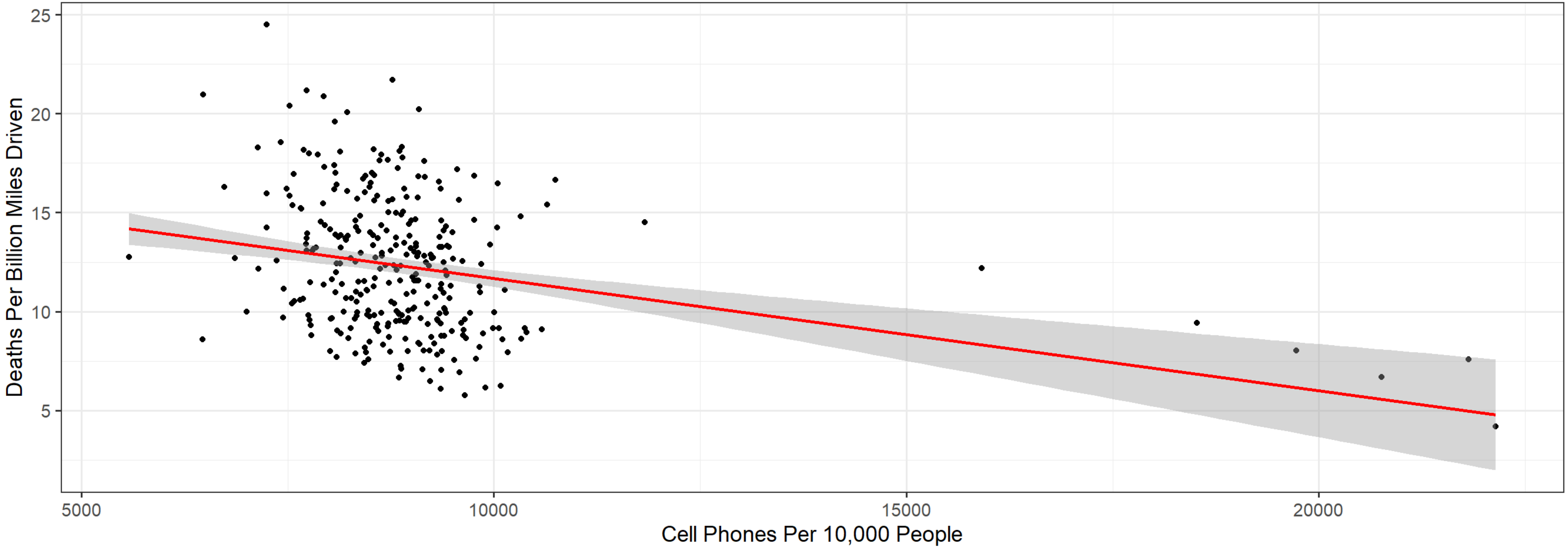
Pooled Regression III

▶ 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

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

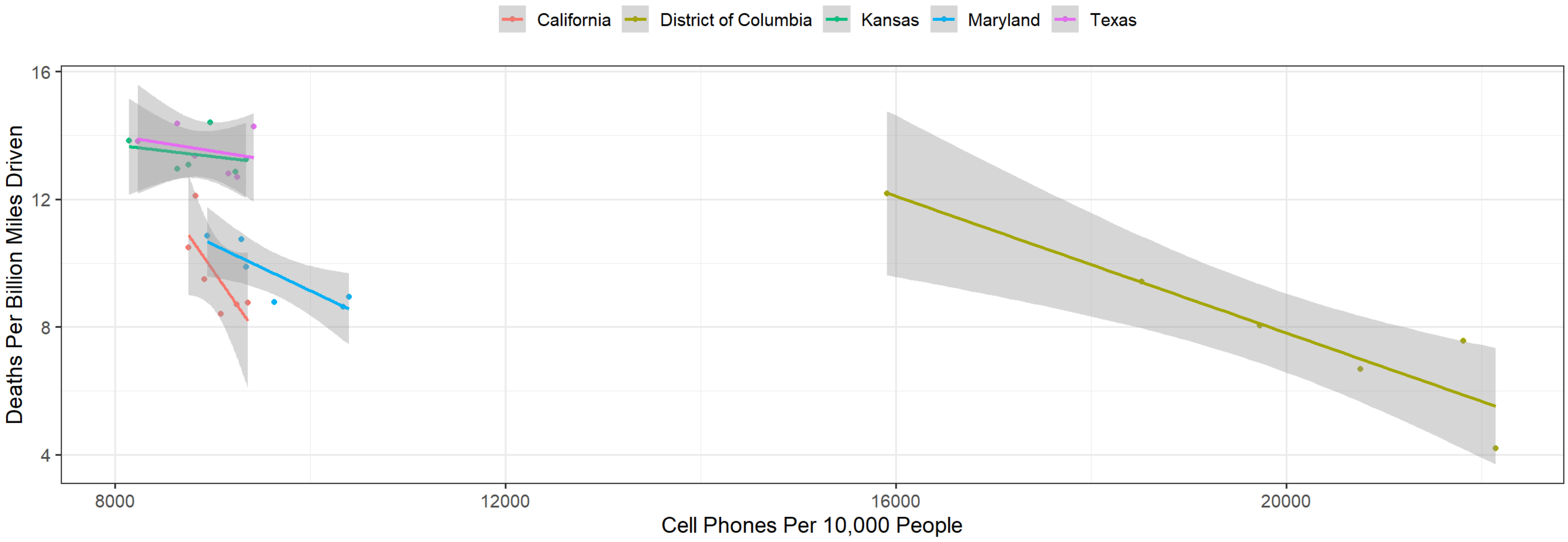
...

- **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

$$cor(u \quad u \quad) \neq 0$$

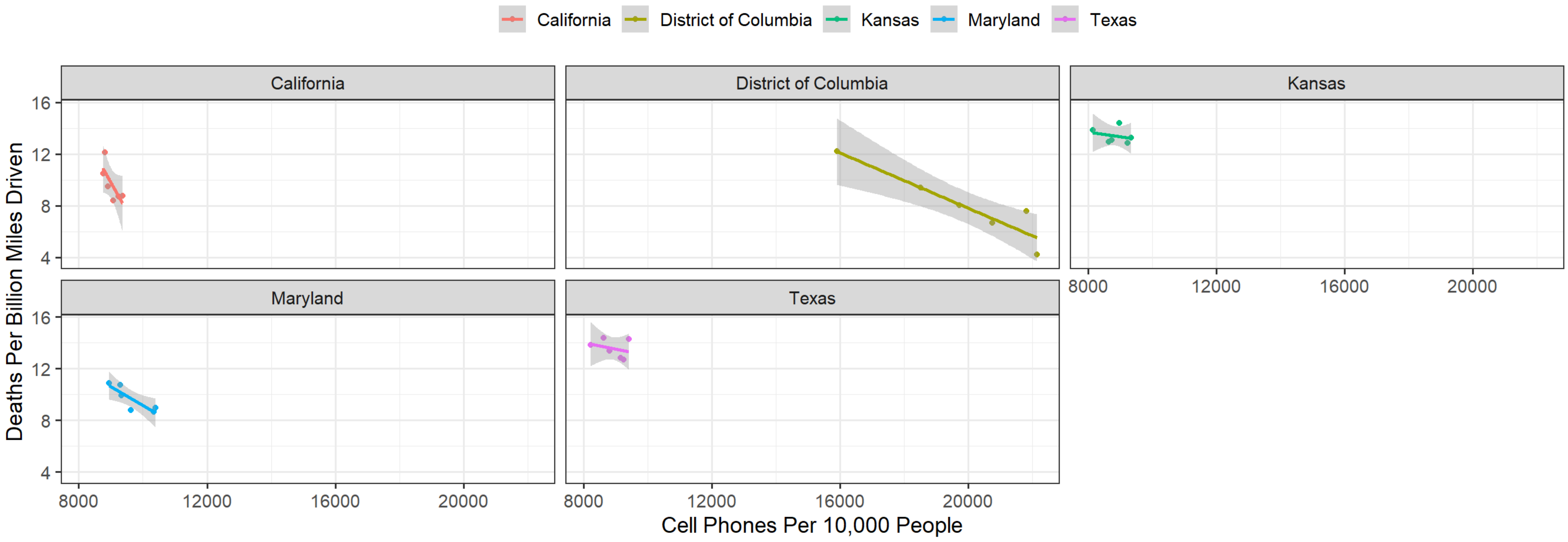
Example: Consider Just 5 States

- Code



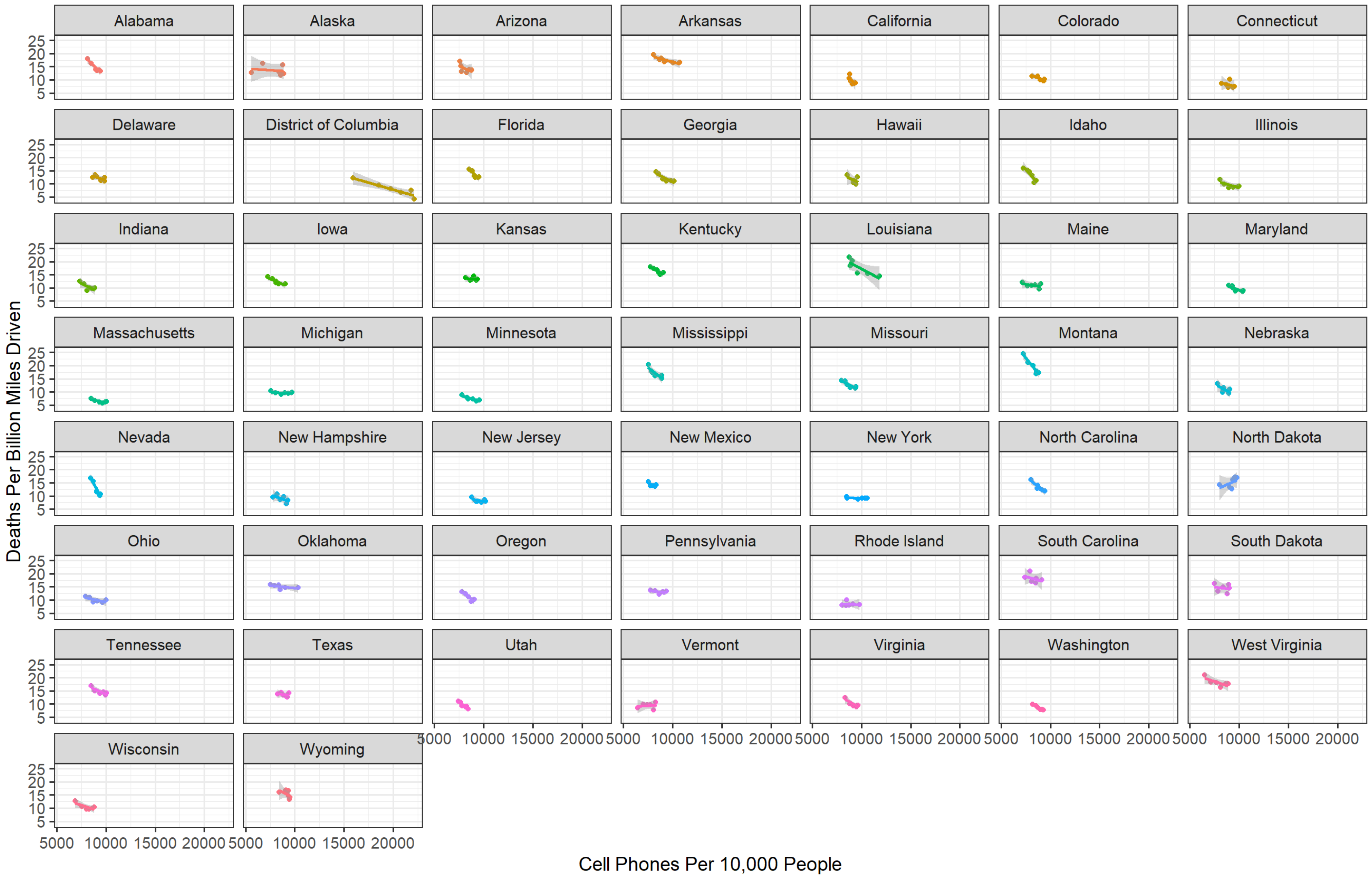
Example: Consider Just 5 States

- 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**:

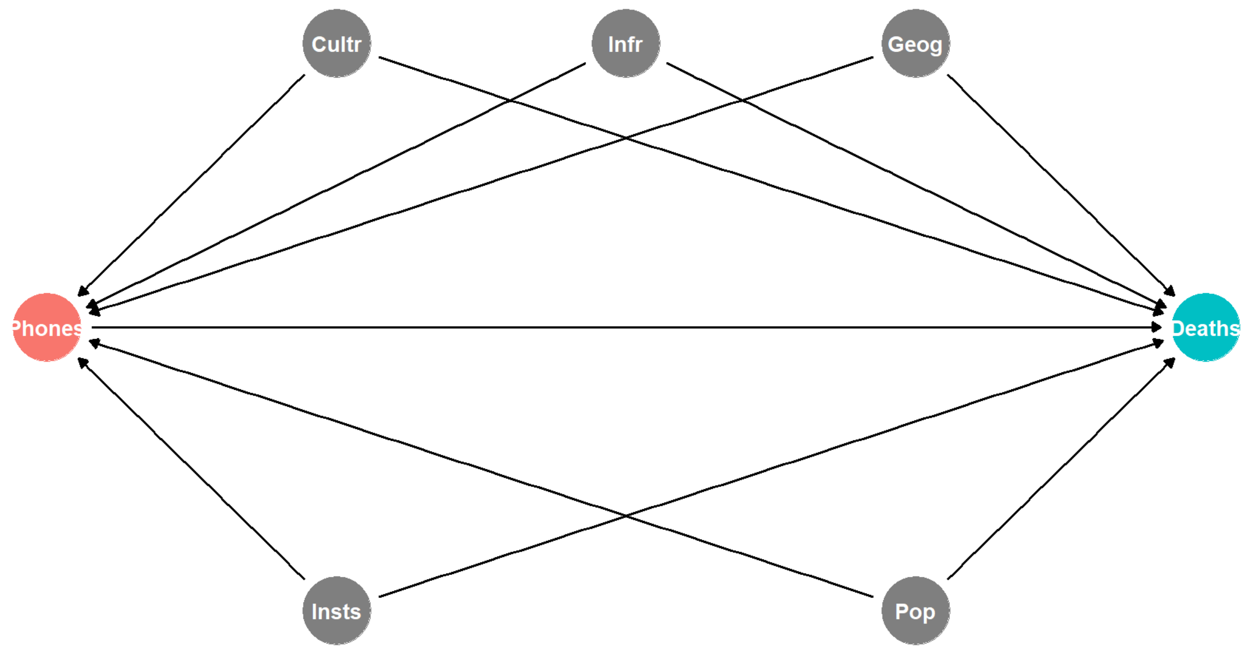
$$cor(\text{u}_{it}, \text{Cell Phones}_{it}) \neq 0 \qquad E[\text{u}_{it} | \text{Cell Phones}_{it}] \neq 0$$

- 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

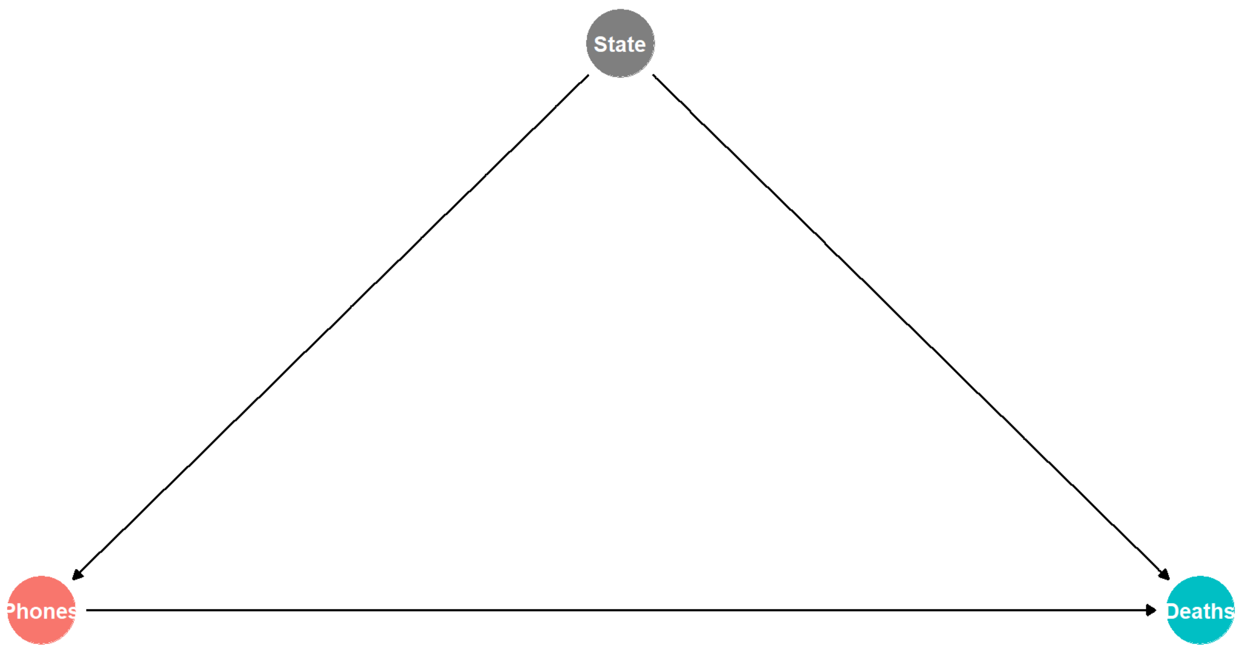
Fixed Effects: DAG I

- 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:

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

Fixed Effects: α_i

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

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

- 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:

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

- 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()
```

```
# A tibble: 52 × 5
   term                       estimate std.error statistic  p.value
   <chr>                         <dbl>     <dbl>     <dbl>    <dbl>
 1 (Intercept)                25.5      1.02          25.1  1.24e-70
 2 cell_plans                 -0.00120  0.000101     -11.9  3.48e-26
 3 stateAlaska                -2.48     0.675         -3.68 2.82e- 4
 4 stateArizona               -1.51     0.670         -2.25 2.51e- 2
 5 stateArkansas               3.19     0.666          4.79 2.83e- 6
 6 stateCalifornia            -4.98     0.666         -7.48 1.21e-12
 7 stateColorado              -4.34     0.665         -6.53 3.59e-10
 8 stateConnecticut           -6.60     0.665         -9.91 8.70e-20
 9 stateDelaware              -2.10     0.667         -3.15 1.84e- 3
10 stateDistrict of Columbia  6.36     1.29           4.93 1.50e- 6
# … with 42 more rows
```

Our Example in R: Regression II

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

```
# A tibble: 1 × 12
  r.squared adj.r.squ…¹ sigma stati…²   p.value    df logLik   AIC   BIC devia…³
      <dbl>       <dbl> <dbl>   <dbl>     <dbl> <dbl>  <dbl> <dbl> <dbl>   <dbl>
1     0.905       0.887  1.15    47.7 7.78e-104    51  -449. 1004. 1202.    337.
# … with 2 more variables: df.residual <int>, nobs <int>, and abbreviated
#   variable names ¹adj.r.squared, ²statistic, ³deviance
```

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

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

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

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

- Within each group i, the de-meaned variables $\tilde{Y}_{it}$ and $\tilde{X}_{it}$'s all have a mean of 0[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

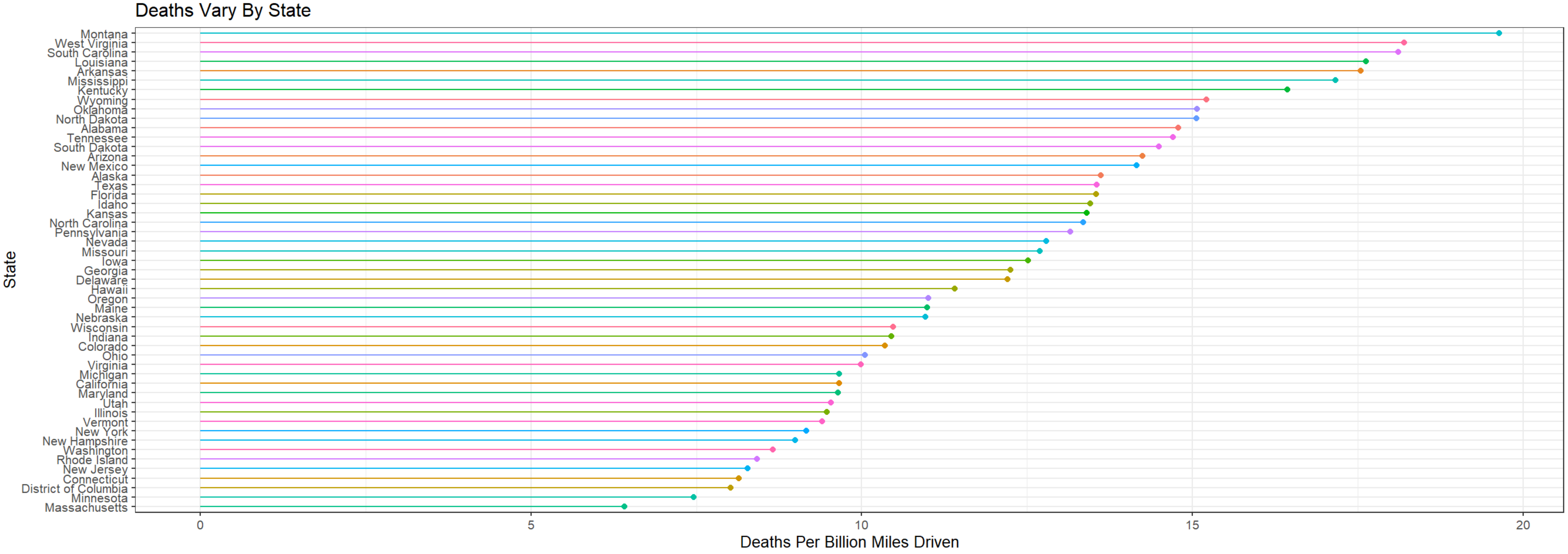
```r
# get means of Y and X by state
means_state <- phones %>%
  group_by(state) %>%
  summarize(avg_deaths = mean(deaths),
            avg_phones = mean(cell_plans))

# look at it
means_state
```

```
# A tibble: 51 × 3
   state                avg_deaths avg_phones
   <fct>                     <dbl>      <dbl>
 1 Alabama                   14.8       8906.
 2 Alaska                    13.6       7818.
 3 Arizona                   14.2       8097.
 4 Arkansas                  17.5       9268.
 5 California                 9.66      9030.
 6 Colorado                  10.4       8982.
 7 Connecticut                8.14      8948.
 8 Delaware                  12.2       9304.
 9 District of Columbia       8.02     19811.
10 Florida                   13.5       9079.
# … with 41 more rows
```

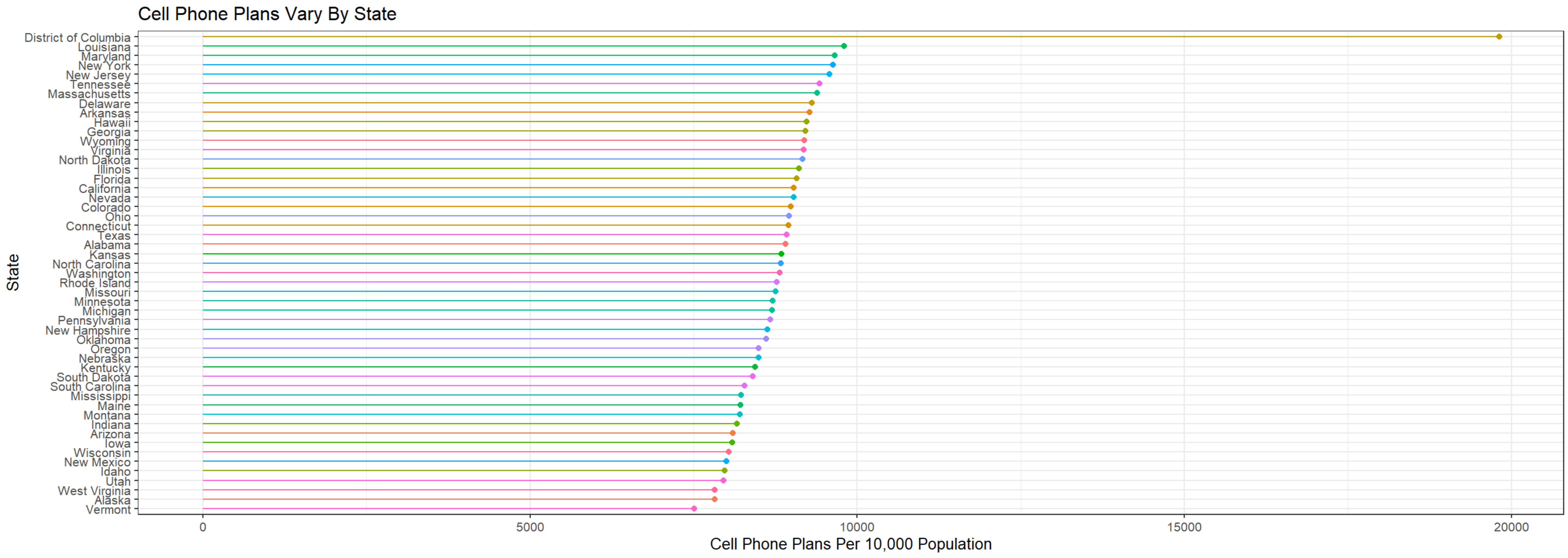
Looking at the Data in R II

- Code



Looking at the Data in R III

- Code



De-Meaning the Data in 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
```

```
# A tibble: 306 × 6
# Groups:   state [51]
   state                year  cell_plans deaths phones_dm deaths_dm
   <fct>                <fct>      <dbl>  <dbl>     <dbl>     <dbl>
 1 Alabama              2007       8136.  18.1      -771.     3.29
 2 Alaska               2007       6730.  16.3     -1087.     2.69
 3 Arizona              2007       7572.  16.9      -525.     2.68
 4 Arkansas             2007       8071.  19.6     -1197.     2.05
 5 California           2007       8822.  12.1      -208.     2.44
 6 Colorado             2007       8162.  11.4      -820.     1.02
 7 Connecticut          2007       8235.   8.64     -713.     0.500
 8 Delaware             2007       8684.  12.3      -620.     0.128
 9 District of Columbia 2007      15910.  12.2     -3901.     4.18
10 Florida              2007       8550.  15.6      -528.     2.05
# … with 296 more rows
```

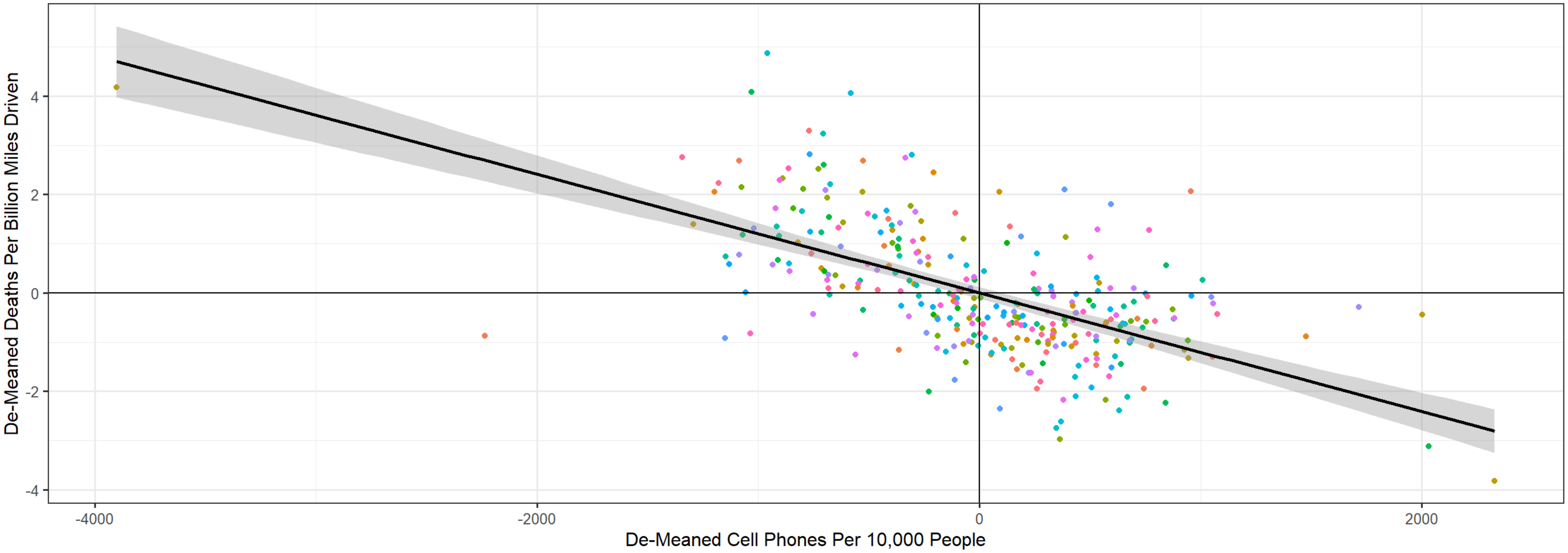
De-Meaning the Data in R II

```r
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
```

```
# A tibble: 51 × 5
   state                mean_deaths sd_deaths mean_phones sd_phones
   <fct>                      <dbl>     <dbl>       <dbl>     <dbl>
 1 Alabama                        0      1.95           0      502.
 2 Alaska                         0      1.9            0     1348.
 3 Arizona                        0      1.57           0      514.
 4 Arkansas                       0      1.18           0      970.
 5 California                     0      1.41           0      242 
 6 Colorado                       0      0.85           0      478.
 7 Connecticut                    0      1.19           0      471.
 8 Delaware                       0      0.94           0      489.
 9 District of Columbia           0      2.68           0     2333.
10 Florida                        0      1.38           0      358.
# … with 41 more rows
```

De-Meaning the Data in R: Visualizing

▶ Code



De-Meaning the Data in R: Regression I

```
# A tibble: 2 × 5
  term         estimate std.error statistic  p.value
  <chr>           <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept) -8.62e-16 0.0602    -1.43e-14 1.00e+ 0
2 phones_dm   -1.20e- 3 0.0000926 -1.30e+ 1 5.07e-31
```

De-Meaning the Data in R: Regression II

```
# A tibble: 1 × 12
  r.squared adj.r.squa…¹ sigma stati…²  p.value    df logLik   AIC   BIC devia…³
      <dbl>        <dbl> <dbl>   <dbl>    <dbl> <dbl>  <dbl> <dbl> <dbl>   <dbl>
1     0.357        0.355  1.05    169. 5.07e-31     1  -449.  904.  915.    337.
# … with 2 more variables: df.residual <int>, nobs <int>, and abbreviated
#   variable names ¹adj.r.squared, ²statistic, ³deviance
```

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()
```

```
# A tibble: 1 × 5
  term       estimate std.error statistic  p.value
  <chr>         <dbl>     <dbl>     <dbl>    <dbl>
1 cell_plans -0.00120  0.000143     -8.42 3.79e-11
```

Comparing FE Approaches

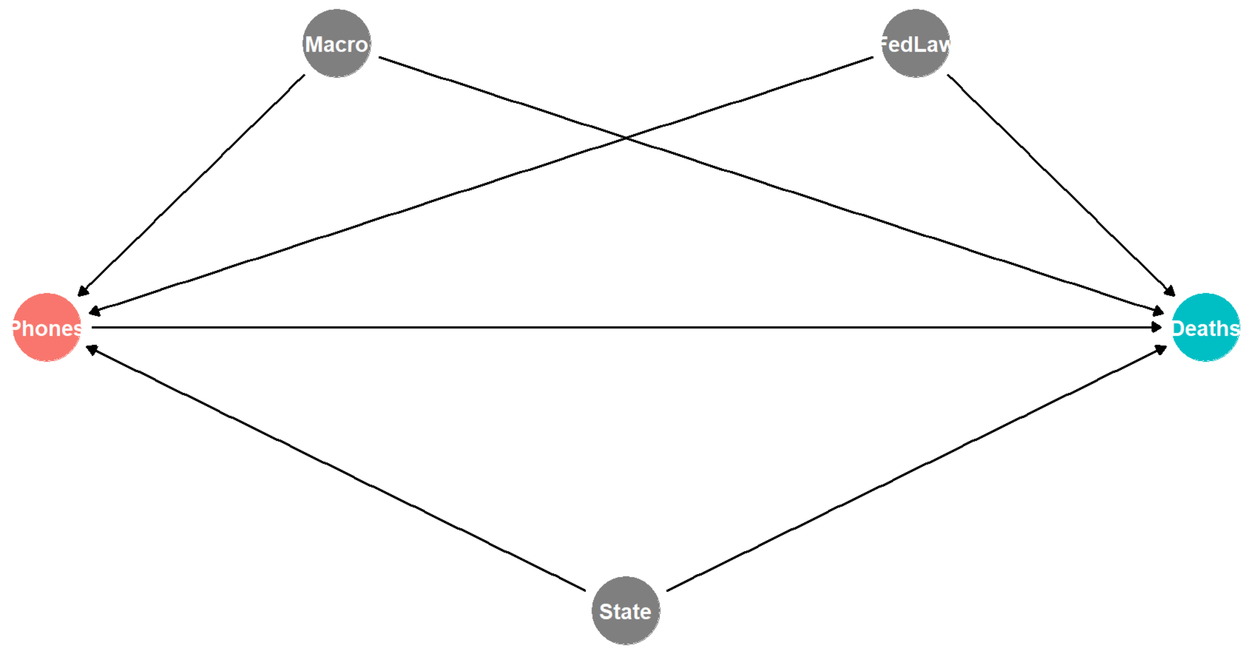
	Pooled Regression	FE: LSDV Method	FE: De-Meaned	FE: fixest
Constant	17.33710***	25.50768***	0.00000	
	(0.97538)	(1.01764)	(0.06023)	
Cell Phone Plans	−0.00057***	−0.00120***	−0.00120***	−0.00120***
	(0.00011)	(0.00010)	(0.00009)	(0.00014)
n	306	306	306	306
Adj. R^2	0.08	0.89	0.36	
SER	3.27	1.05	1.05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

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

Two-Way Fixed Effects

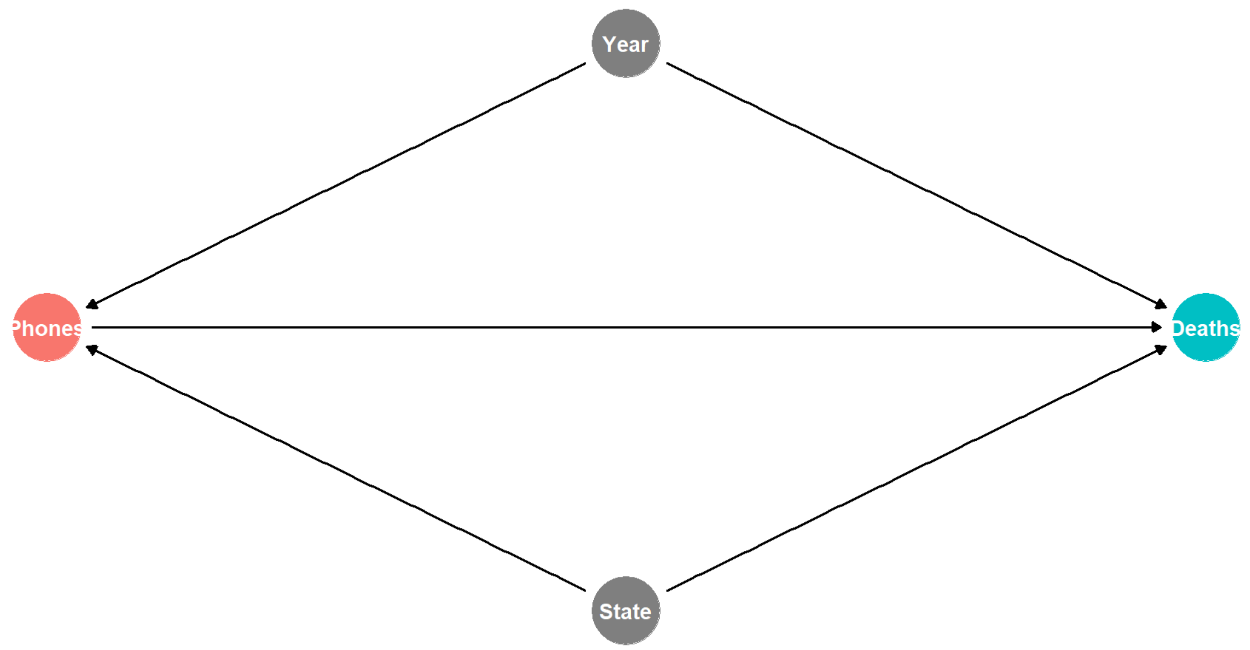
Two-Way Fixed Effects

- 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

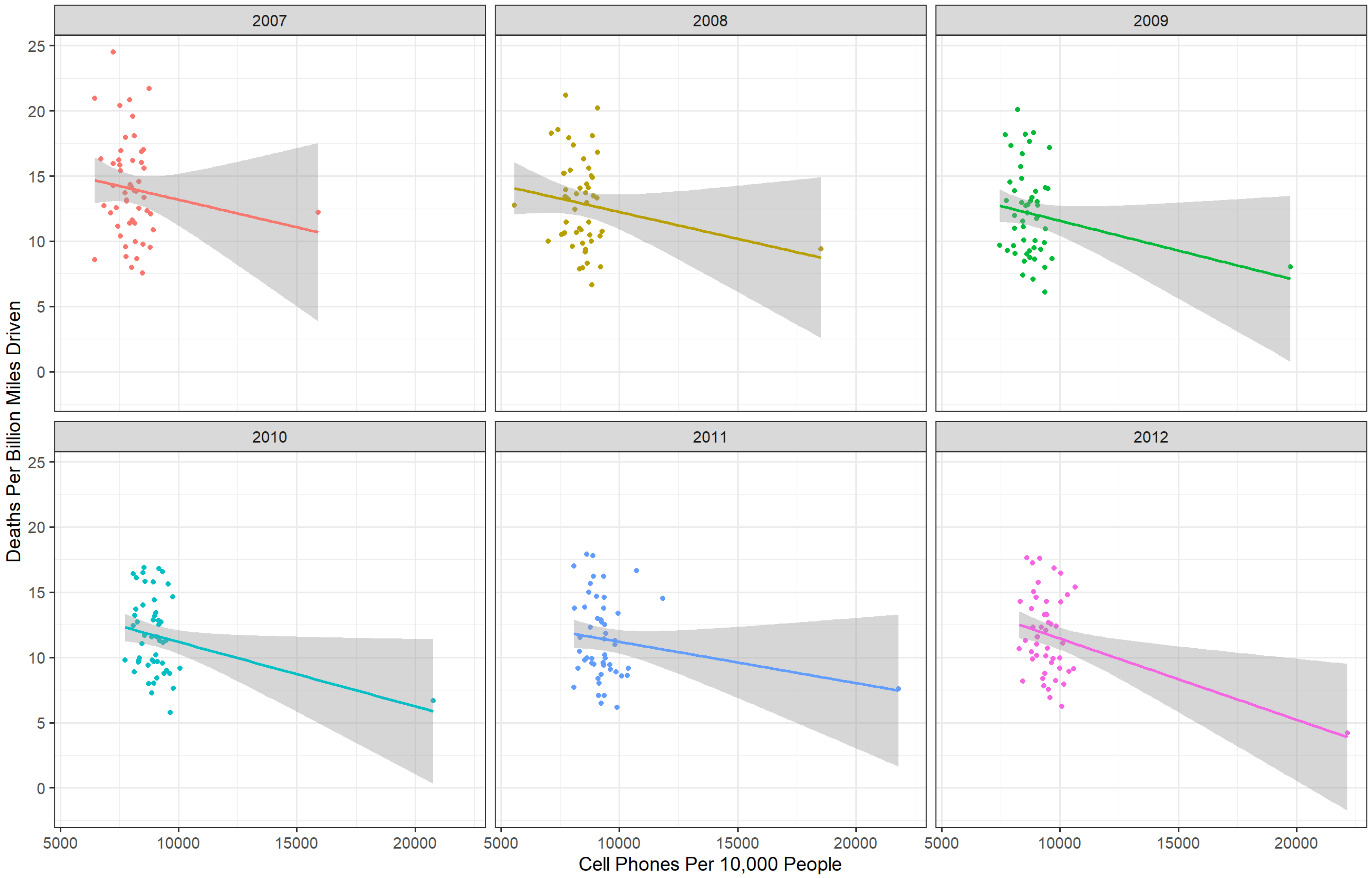
Two-Way Fixed Effects: Our Example

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

- 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



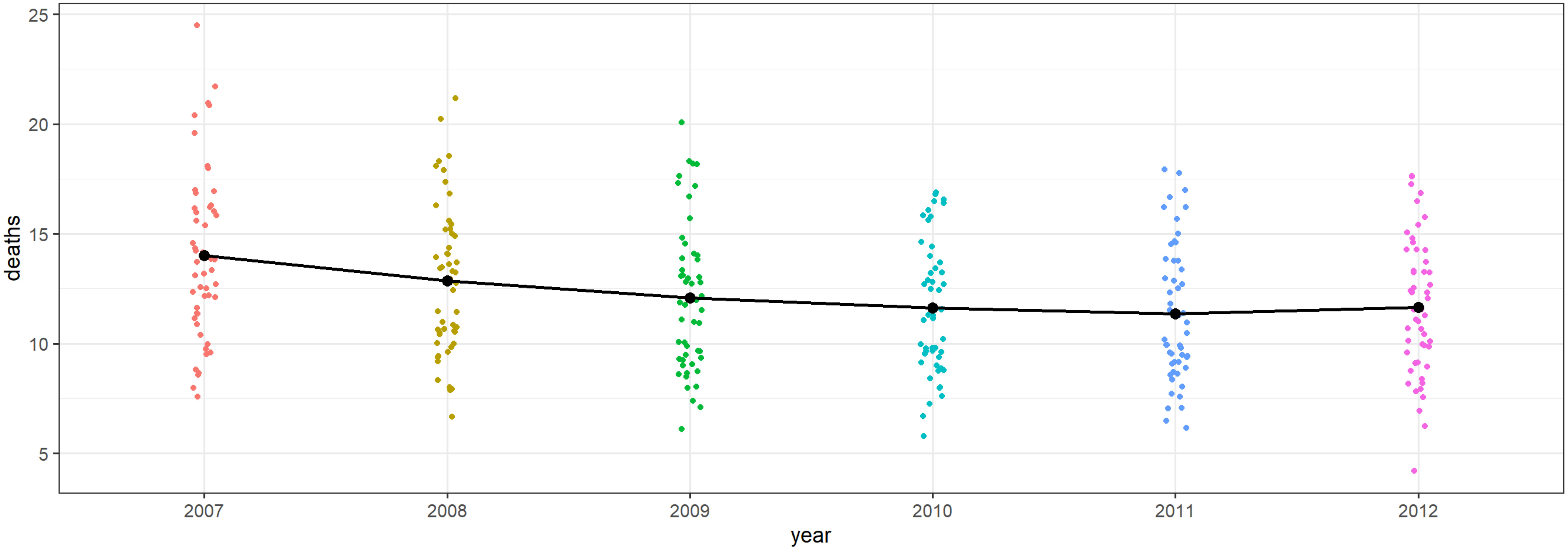
Looking at the Data: Change Over Time II

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

```
# A tibble: 6 × 3
  year  avg_deaths avg_phones
  <fct>      <dbl>      <dbl>
1 2007        14.0      8065.
2 2008        12.9      8483.
3 2009        12.1      8860.
4 2010        11.6      9135.
5 2011        11.4      9485.
6 2012        11.7      9660.
```

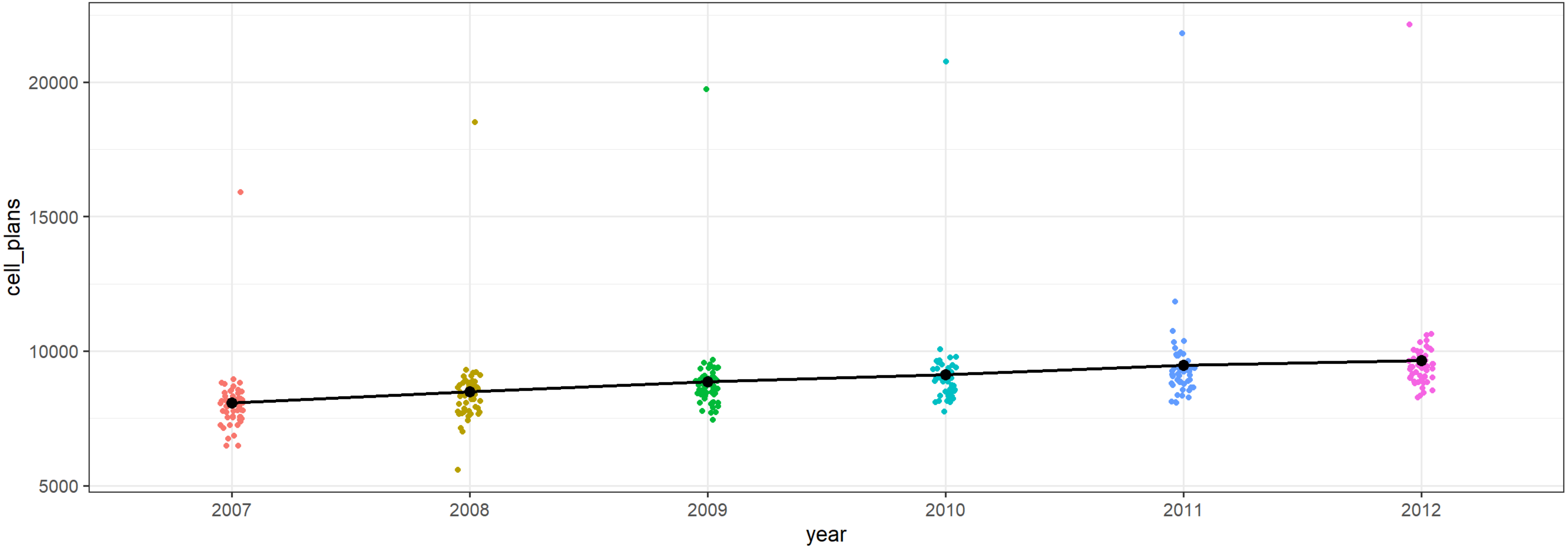
Looking at the Data: Change In *Deaths* Over Time

▶ 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()
```

```
# A tibble: 57 × 5
   term                        estimate std.error statistic  p.value
   <chr>                          <dbl>     <dbl>     <dbl>    <dbl>
 1 (Intercept)                 18.9      1.45          13.0  5.43e-30
 2 cell_plans                  -0.000300 0.000172      -1.74 8.34e- 2
 3 stateAlaska                 -1.50     0.624         -2.40 1.70e- 2
 4 stateArizona                -0.779    0.611         -1.27 2.04e- 1
 5 stateArkansas                2.87     0.599          4.79 2.90e- 6
 6 stateCalifornia             -5.09     0.596         -8.55 1.30e-15
 7 stateColorado               -4.41     0.595         -7.41 1.95e-12
 8 stateConnecticut            -6.63     0.595        -11.1  1.17e-23
 9 stateDelaware               -2.46     0.599         -4.10 5.55e- 5
10 stateDistrict of Columbia -3.50     1.97          -1.78 7.66e- 2
# … with 47 more rows
```

With `fixest`

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

fe2_reg_2 %>% summary()
```

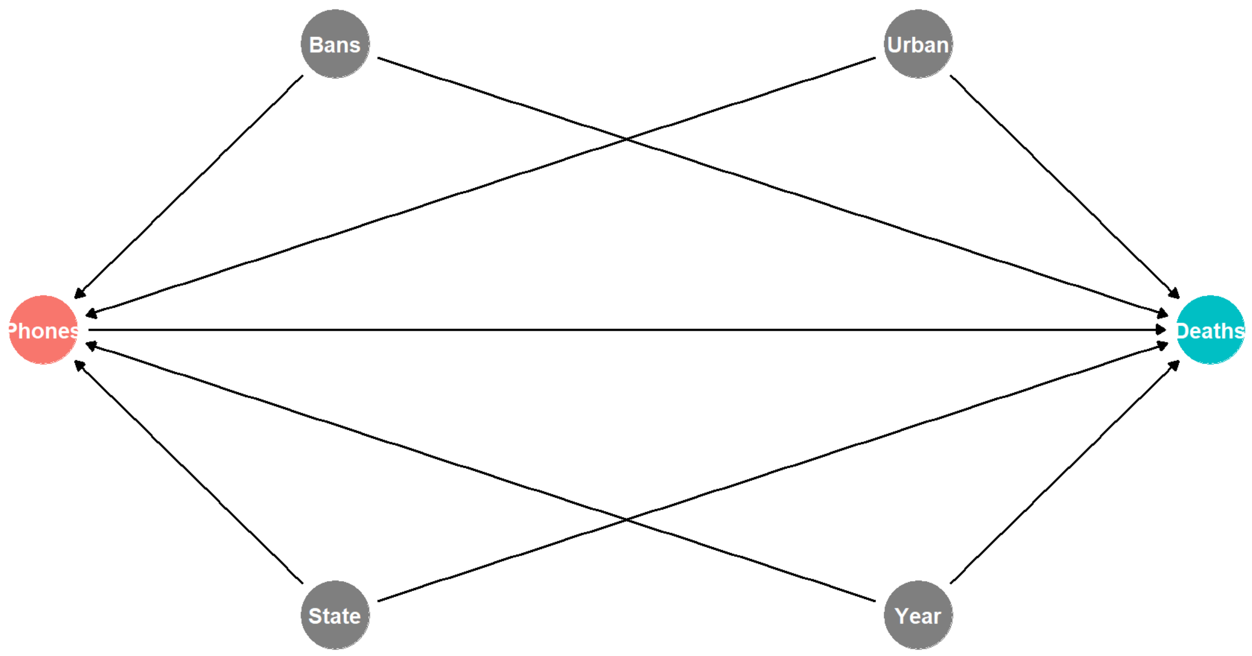
```
OLS estimation, Dep. Var.: deaths
Observations: 306
Fixed-effects: state: 51,  year: 6
Standard-errors: Clustered (state)
           Estimate Std. Error   t value Pr(>|t|)
cell_plans   -3e-04   0.000305 -0.980739  0.33144
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 0.930036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

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

```
# A tibble: 1 × 5
  term        estimate std.error statistic p.value
  <chr>          <dbl>     <dbl>     <dbl>   <dbl>
1 cell_plans -0.000300  0.000305    -0.981   0.331
```

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?

```r
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`)
```

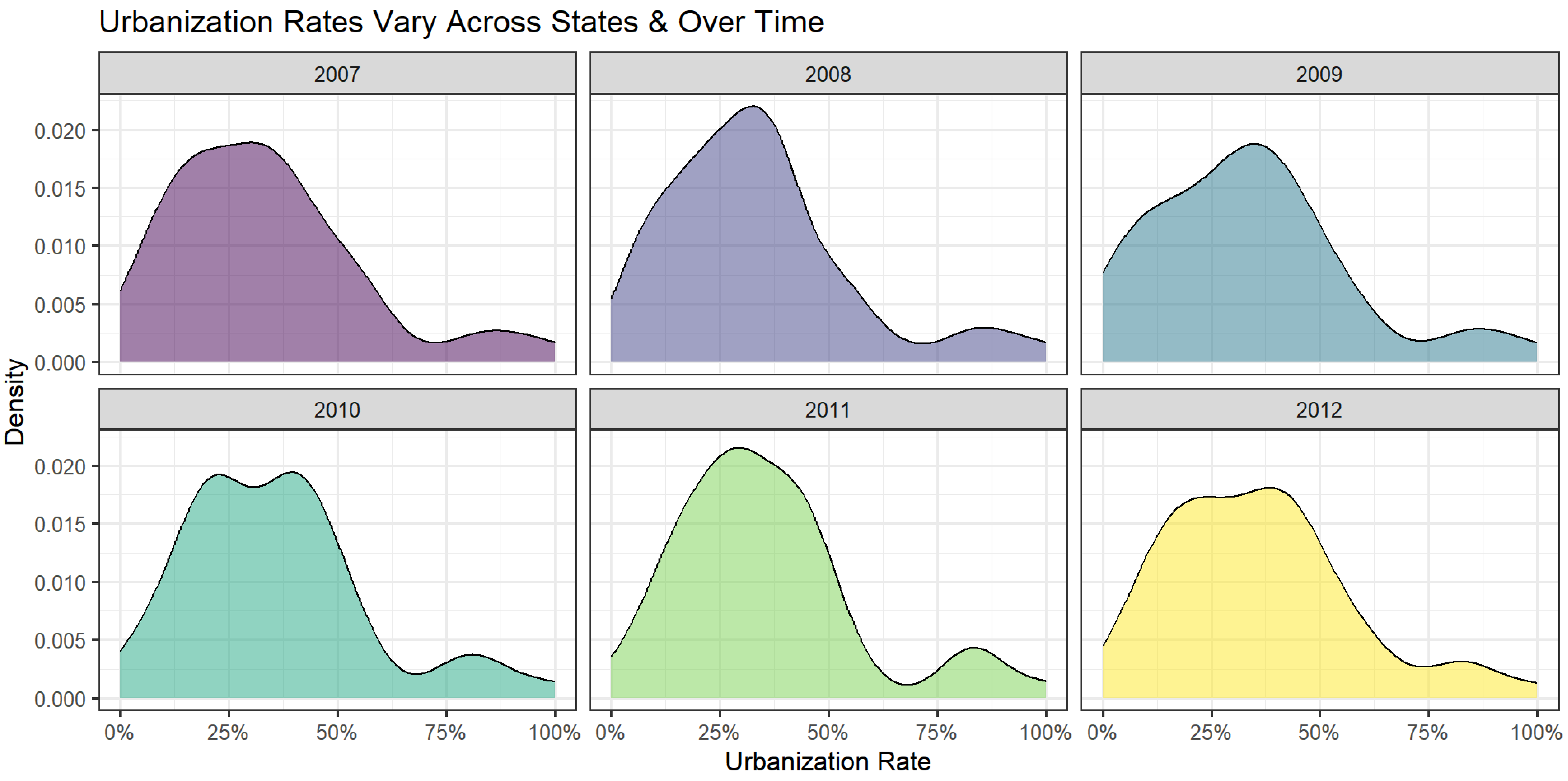
```
# A tibble: 6 × 3
# Groups:   year [6]
  year  `States Without a Ban` `States With Cell Phone Ban`
  <fct>                  <int>                        <int>
1 2007                      46                            5
2 2008                      46                            5
3 2009                      44                            7
4 2010                      43                            8
5 2011                      41                           10
6 2012                      40                           11
```

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`)
```

```
# A tibble: 6 × 3
# Groups:   year [6]
  year  `States Without a Ban` `States With a Texting Ban`
  <fct>                  <int>                       <int>
1 2007                      49                           2
2 2008                      47                           4
3 2009                      42                           9
4 2010                      30                          21
5 2011                      20                          31
6 2012                      16                          35
```

Adding Covariates — Necessary?



Adding Covariates II

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

- 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()
```

```
# A tibble: 4 × 5
  term           estimate std.error statistic p.value
  <chr>             <dbl>     <dbl>     <dbl>   <dbl>
1 cell_plans    -0.000340  0.000277     -1.23  0.225
2 text_ban1      0.256     0.243         1.05  0.298
3 urban_percent  0.0131    0.00982       1.34  0.187
4 cell_ban1     -0.680     0.336        -2.03  0.0482
```

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

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