Panel Regression: Fixed Effect and Demeaned Regressioin

Zahid Asghar

School of Economics, QAU, Islamabad

10/26/22

Types of Data I

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

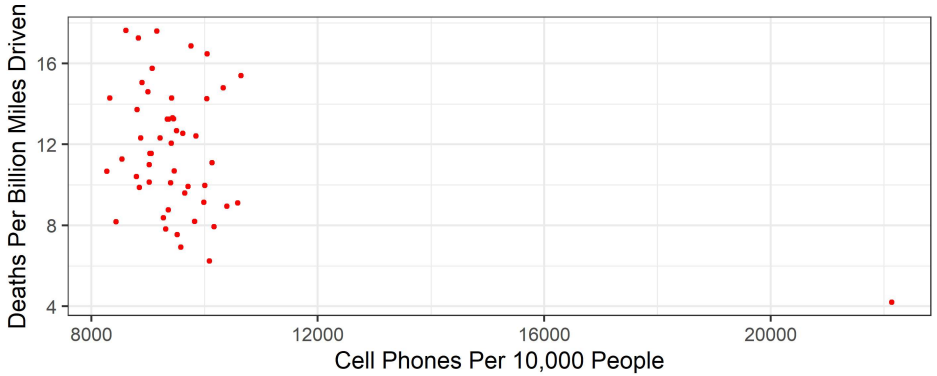
```
## # A tibble: 6 × 4
##   state      year  deaths cell_plans
##   <fct>      <fct>  <dbl>      <dbl>
## 1 Alabama    2012   13.3       9434.
## 2 Alaska     2012   12.3       8873.
## 3 Arizona    2012   13.7       8811.
## 4 Arkansas   2012   16.5      10047.
## 5 California 2012    8.76      9362.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

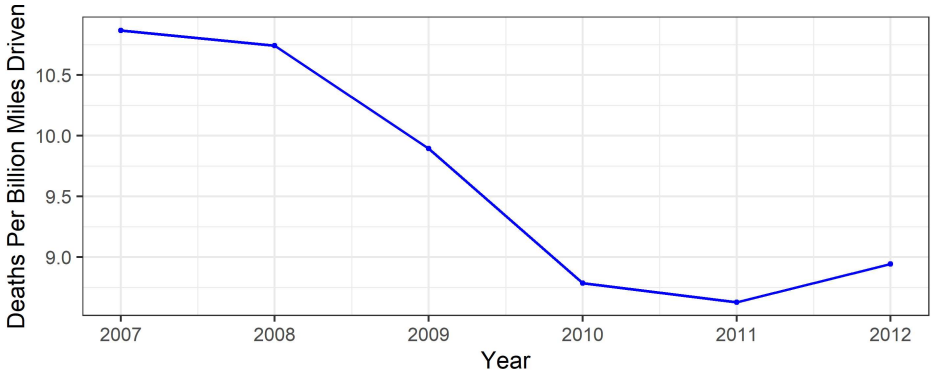
- **Time-series data**: track same individual $\bar{i}$ over different times t

```
## # 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 I

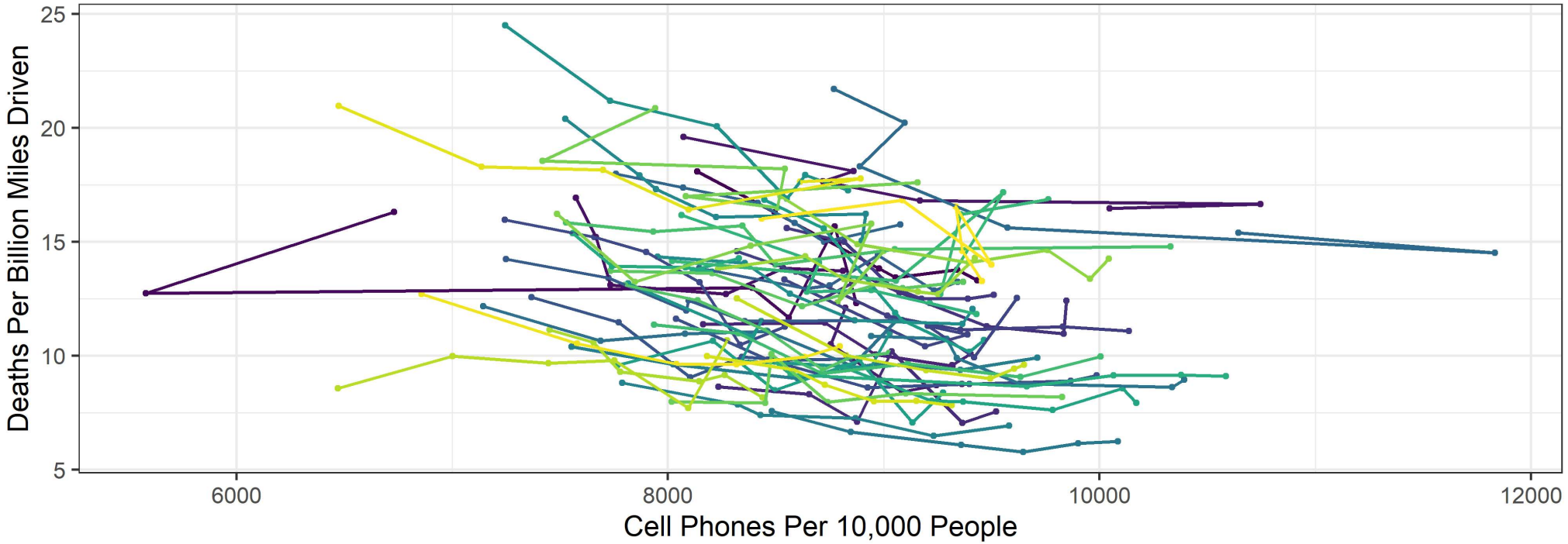
- **Cross-sectional data**: compare different individual i's at same time $\bar{t}$
- **Time-series data**: track same individual $\bar{i}$ over different times t





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

Panel Data I



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
## # ℹ Use `print(n = ...)` to see 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
## # ℹ Use `print(n = ...)` to see 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
## # ℹ Use `print(n = ...)` to see 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
- 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, 200…
## $ state         <fct> Alabama, Alaska, Arizona, Arkansas, California, Col…
## $ urban_percent <dbl> 30, 55, 45, 21, 54, 34, 84, 31, 100, 53, 39, 45, 11…
## $ cell_plans    <dbl> 8136, 6730, 7572, 8071, 8822, 8162, 8235, 8684, 159…
## $ cell_ban      <fct> 0, 0, 0, 0, 0, 0, 1, 0, 1, 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, …
## $ deaths        <dbl> 18.08, 16.30, 16.93, 19.60, 12.10, 11.37, 8.64, 12.…
## $ year_num      <dbl> 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 200…
```

The Data II

```
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
## # ℹ Use `print(n = ...)` to see more rows
```

```
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

```
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
## # ℹ Use `print(n = ...)` to see more rows
```

```
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
```

The Data: With plm

```
# install.packages("plm")
library(plm)

pdim(phones, index=c("state","year"))
## Balanced Panel: n = 51, T = 6, N = 306
```

- `plm` **package** for panel data in R
- `pdim()` checks dimensions of panel dataset
  - `index=` vector of “group” & “year” variables
- Returns with a summary of:
  - `n` groups
  - `T` periods
  - `N` total observaiton

Pooled Regression I

- What if we just ran a standard regression:

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

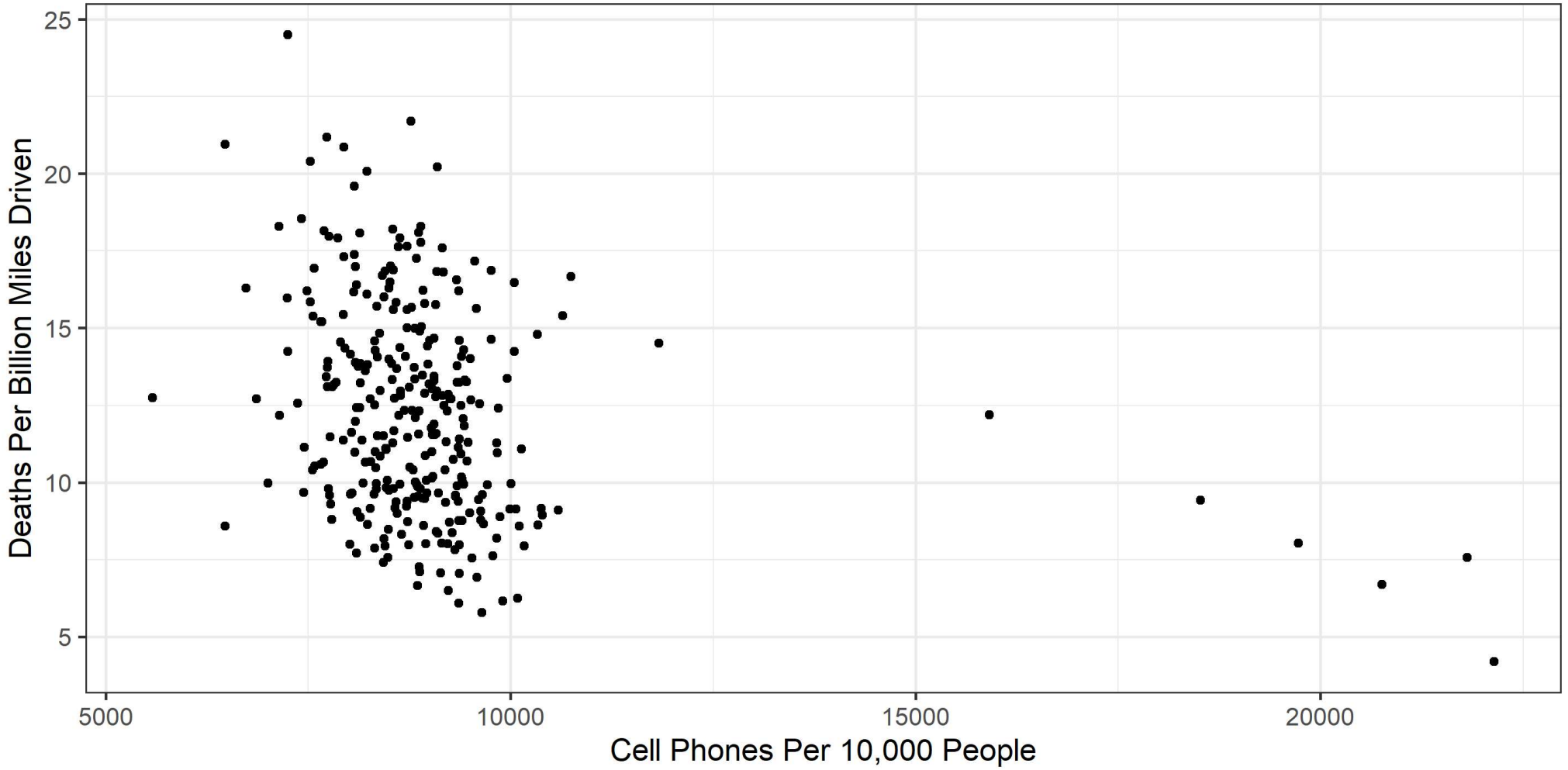
- N number of i groups (e.g. U.S. States)
- T number of t periods (e.g. years)
- This is a pooled regression model: treats all observations as independent

Pooled Regression II

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

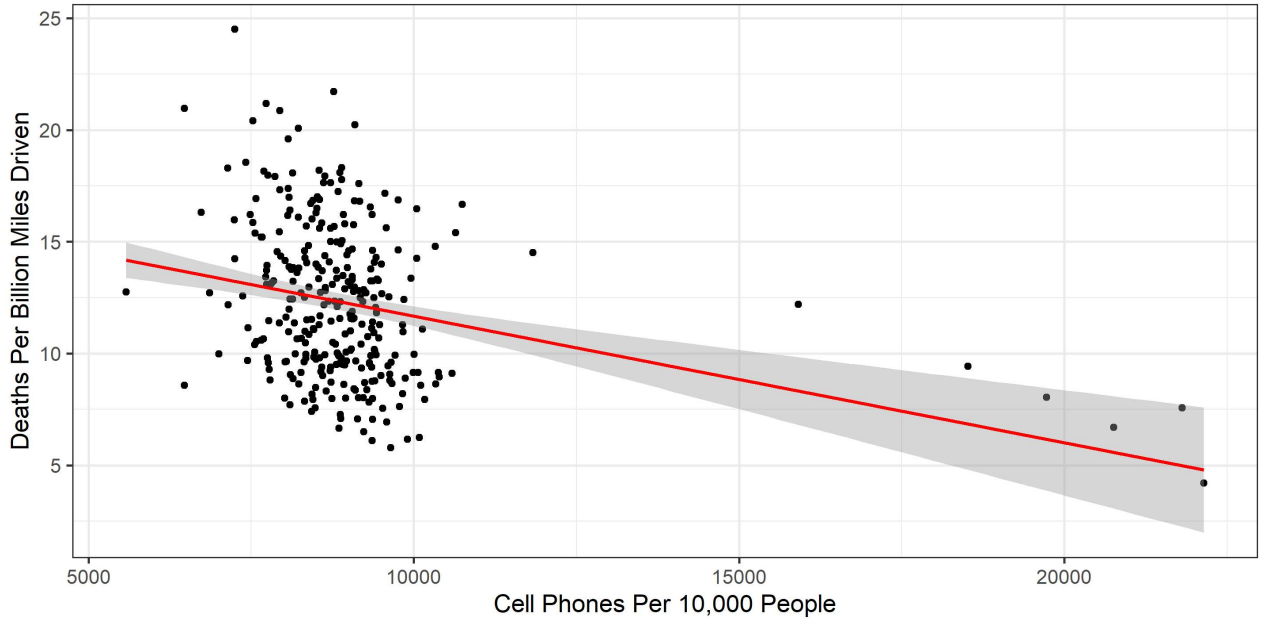
Pooled Regression III

```
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)
```


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

Pooled Regression III

```
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

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

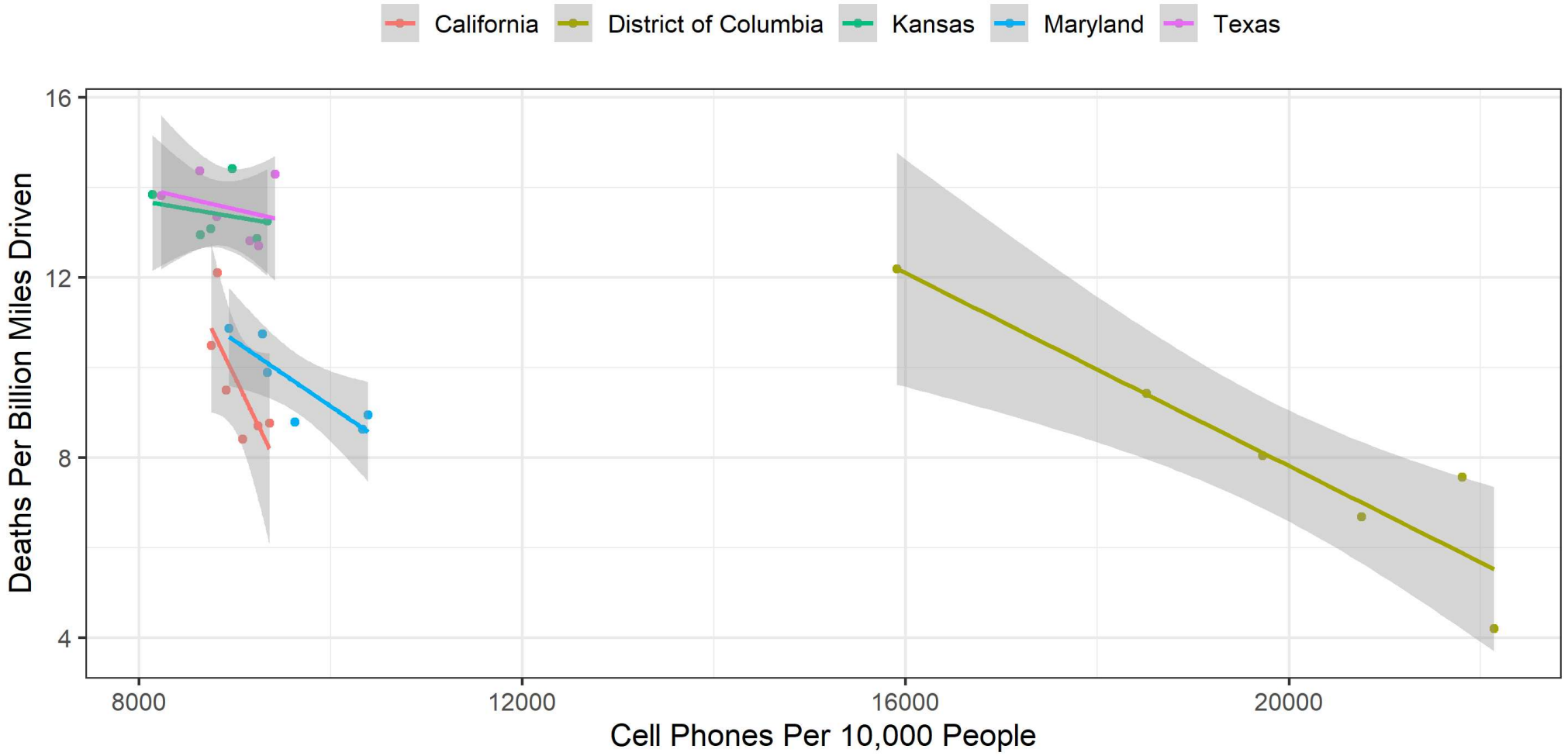
Biases of Pooled Regression: Our Example

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

- 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

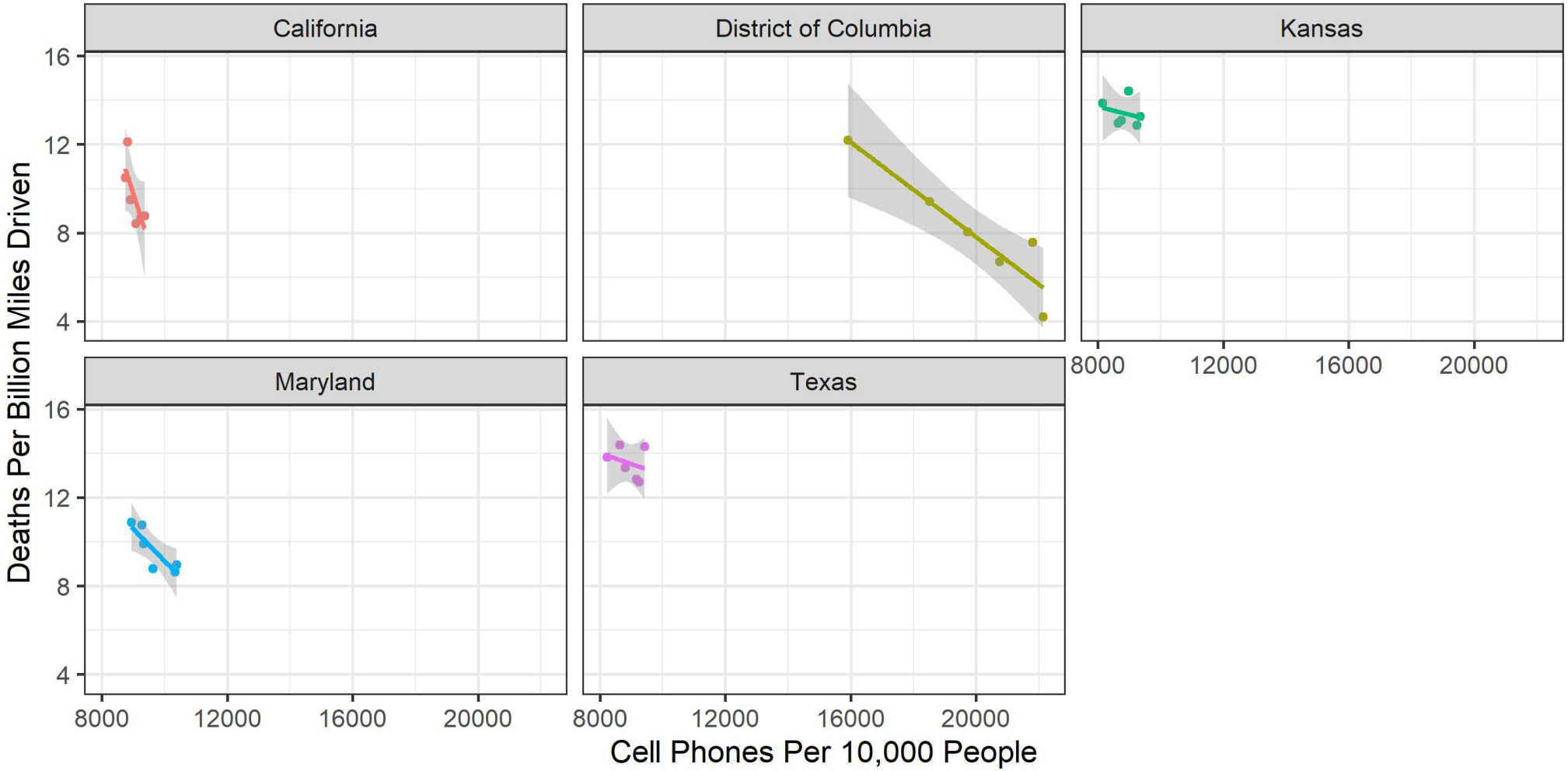
Example: Consider Just 5 States

```
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")
```


California
District of Columbia
Kansas
Maryland
Texas
16
12
8
4
Deaths Per Billion Miles Driven
8000
12000
16000
20000
Cell Phones Per 10,000 People

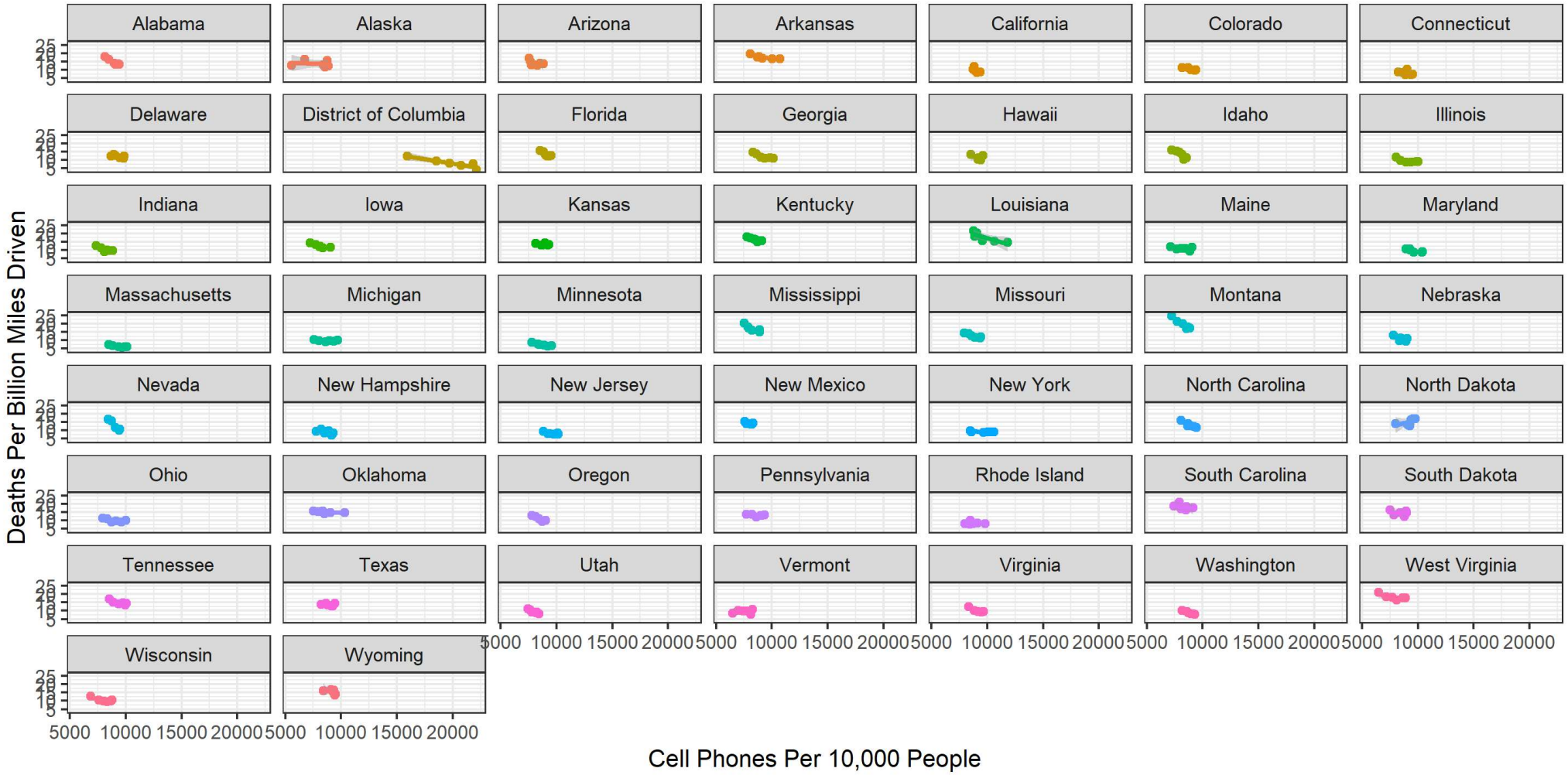
Example: Consider Just 5 States

```
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) #<<
```


California
District of Columbia
Kansas
Maryland
Texas
Deaths Per Billion Miles Driven
16
12
8
4
8000
12000
16000
20000
Cell Phones Per 10,000 People

Look at All States

```
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) #<<
```


Alabama
Alaska
Arizona
Arkansas
California
Colorado
Connecticut
Delaware
District of Columbia
Florida
Georgia
Hawaii
Idaho
Illinois
Indiana
Iowa
Kansas
Kentucky
Louisiana
Maine
Maryland
Massachusetts
Michigan
Minnesota
Mississippi
Missouri
Montana
Nebraska
Nevada
New Hampshire
New Jersey
New Mexico
New York
North Carolina
North Dakota
Ohio
Oklahoma
Oregon
Pennsylvania
Rhode Island
South Carolina
South Dakota
Tennessee
Texas
Utah
Vermont
Virginia
Washington
West Virginia
Wisconsin
Wyoming
Deaths Per Billion Miles Driven
Cell Phones Per 10,000 People

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**:

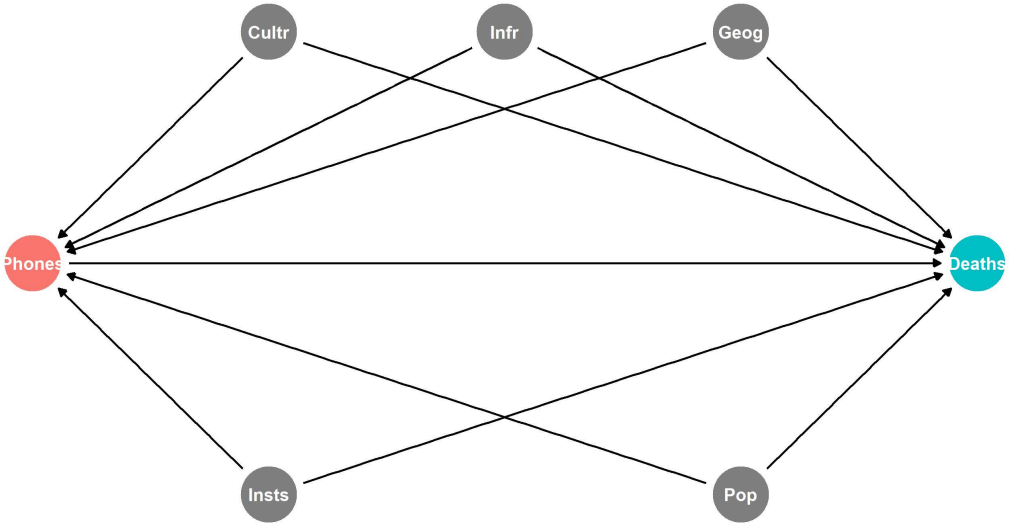
$$cor(\text{u}_{it}, \text{cell phones}_{it}) \neq 0 \qquad E[\text{u}_{it} | \text{cell phones}_{it}] \neq 0$$

- 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**!
  - $cor(\mathbf{u}_{it_1}, \mathbf{u}_{it_2}) \neq 0$
    - Error in State i during t_1 correlates with error in State i during t_2
    - things in State that don't change over time

Fixed Effects Model

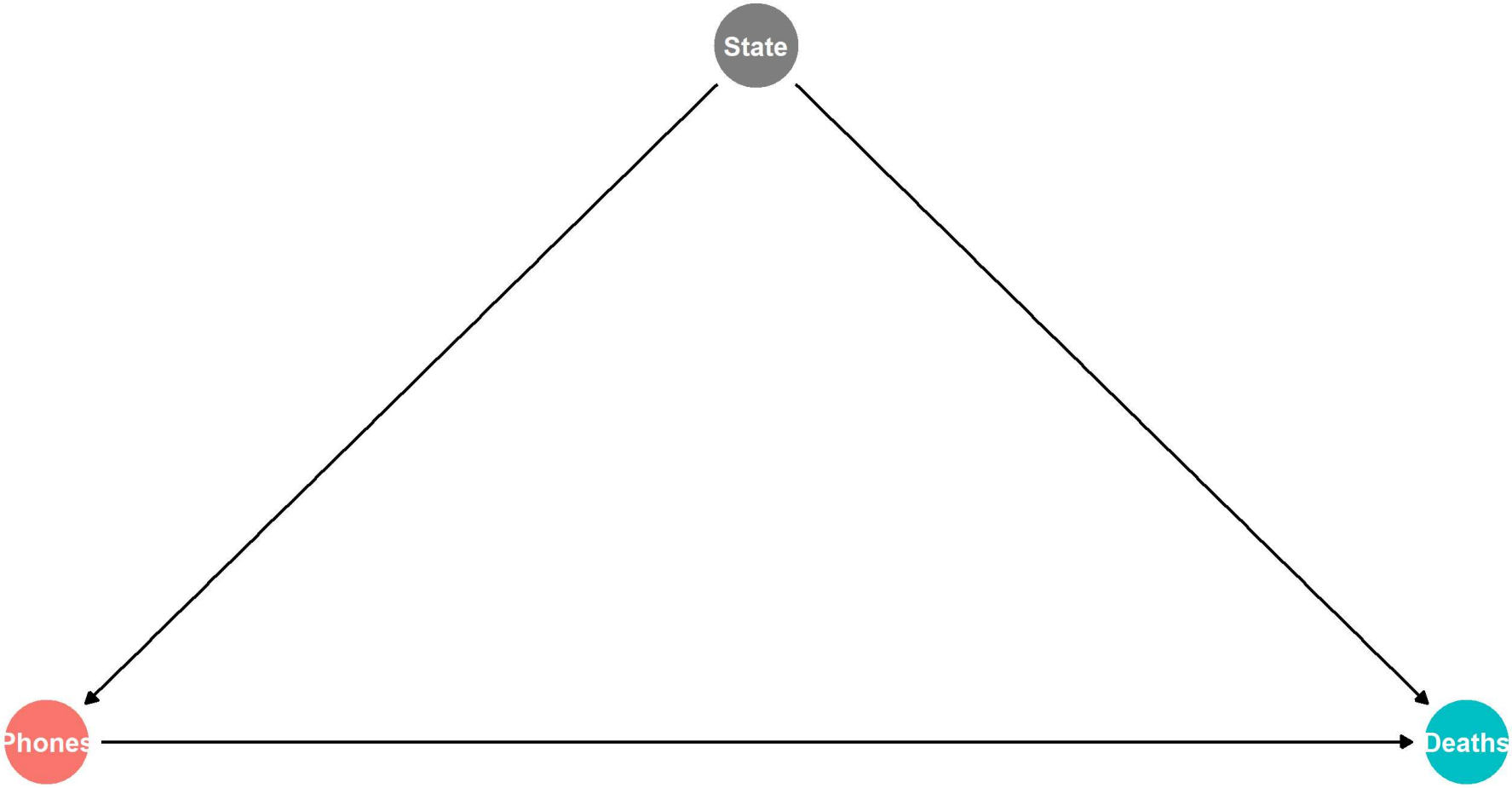
Fixed Effects: DAG

- 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:

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

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 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}

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

- ϵ_{it} is the remaining random error
  - As usual in OLS, assume the 4 typical assumptions about this error:
    - $E[\epsilon_{it}] = 0, var[\epsilon_{it}] = \sigma_\epsilon^2, cor(\epsilon_{it}, \epsilon_{jt}) = 0, cor(\epsilon_{it}, X_{it}) = 0$
- 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 from other factors!

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 .pink[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 .hi-pink[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

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

##

- 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.magenta[†]
  - 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

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

- 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

```
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
## # ℹ Use `print(n = ...)` to see more rows
```

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**
- 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

De-meaned Approach II

$$

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

$$

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

$$Y_i - \bar{Y}_i = \beta_1 (X_{it} - \bar{X}_i) + \tilde{\epsilon}_{it}$$
$$\tilde{Y}_{it} = \beta_1 \tilde{X}_{it} + \tilde{\epsilon}_{it}$$

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

Recall : 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
  - Fixed effects are often called the .hi-pink[“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

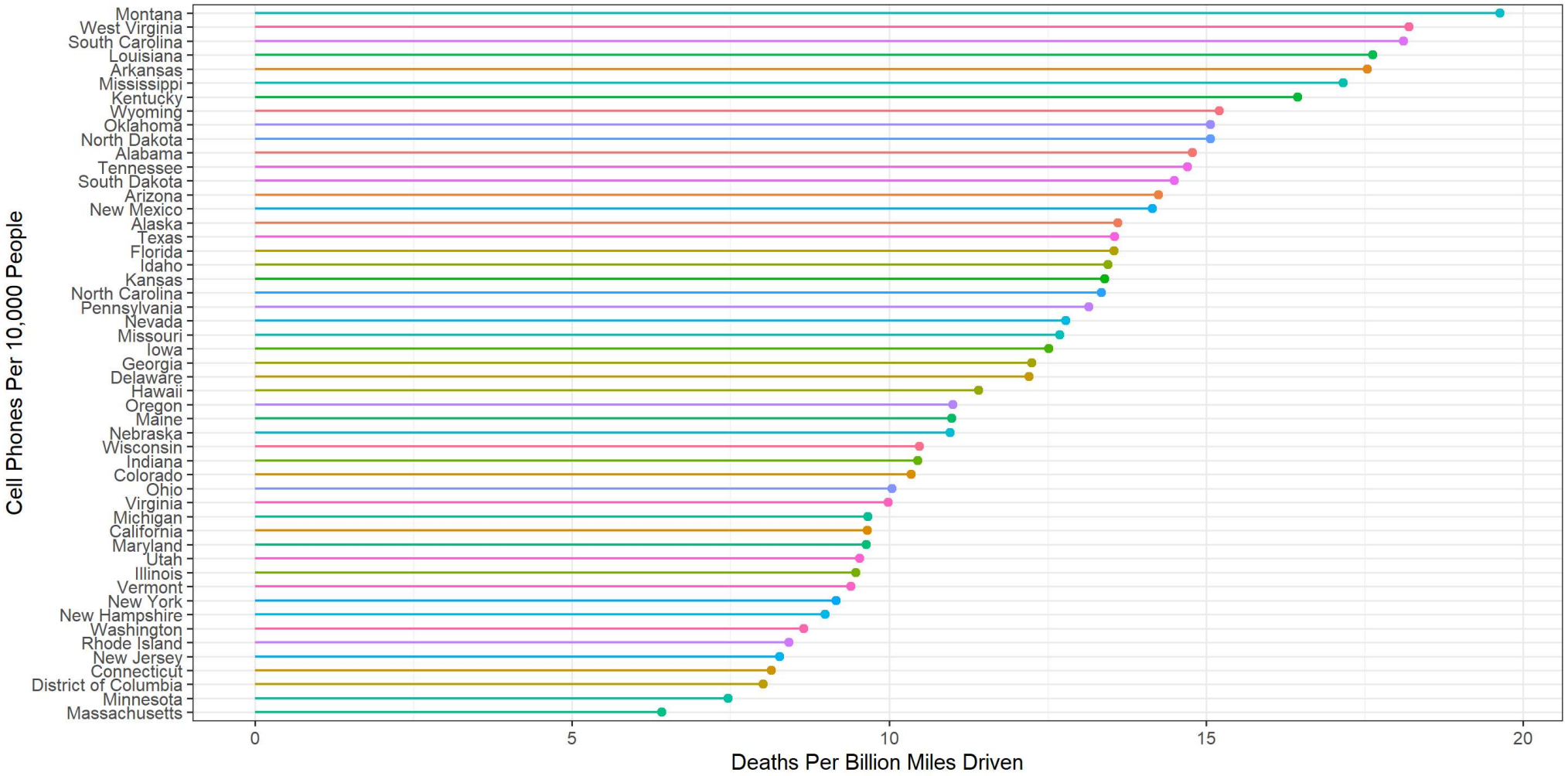
```
# get means of Y and X by state
means_state<-phones %>%
  group_by(state) %>%
  summarize(avg_deaths = mean(deaths),
            avg_phones = mean(cell_plans))

# 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
## # ℹ Use `print(n = ...)` to see more rows
```

De-Meaning the Data in R II

```
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")
```


Montana
West Virginia
South Carolina
Louisiana
Arkansas
Mississippi
Kentucky
Wyoming
Oklahoma
North Dakota
Alabama
Tennessee
South Dakota
Arizona
New Mexico
Alaska
Texas
Florida
Idaho
Kansas
North Carolina
Pennsylvania
Nevada
Missouri
Iowa
Georgia
Delaware
Hawaii
Oregon
Maine
Nebraska
Wisconsin
Indiana
Colorado
Ohio
Virginia
Michigan
California
Maryland
Utah
Illinois
Vermont
New York
New Hampshire
Washington
Rhode Island
New Jersey
Connecticut
District of Columbia
Minnesota
Massachusetts
Cell Phones Per 10,000 People
0
5
10
15
20
Deaths Per Billion Miles Driven

Visualizing "Within Estimates" for the 5 States

Visualizing "Within Estimates" for All 51 States

De-meaned Approach in R I

- The `plm` package is designed for panel data
- `plm()` function is just like `lm()`, with some additional arguments:
  - `index="group_variable_name"` set equal to the name of your `factor` variable for the groups
  - `model=` set equal to `"within"` to use fixed-effects (within-estimator)

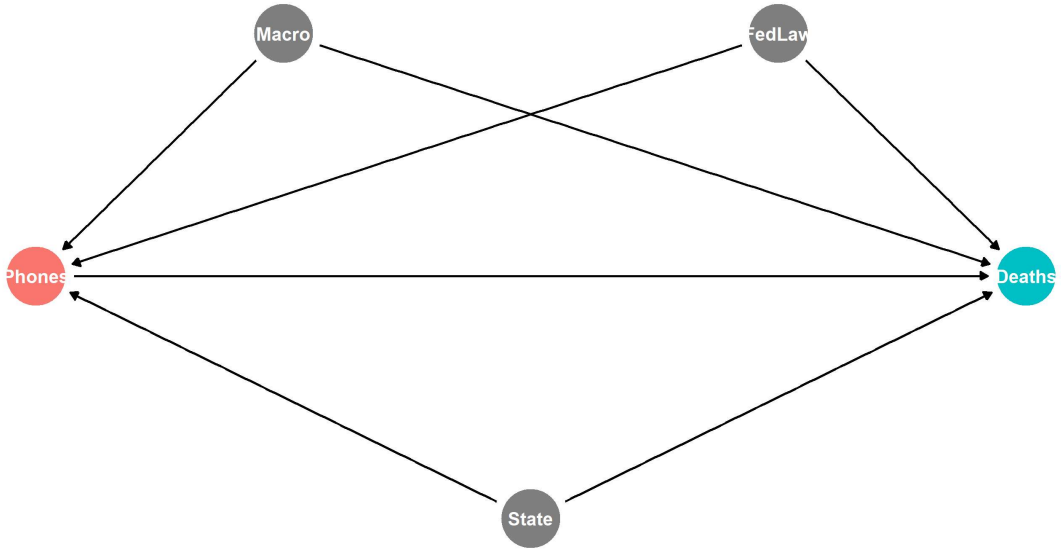
```
#install.packages("plm")
library(plm)
fe_reg_1_alt<-plm(deaths ~ cell_plans,
                  data = phones,
                  index = "state",
                  model = "within")
```

De-meaned Approach in R II

```
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  0.000101     -11.9 3.48e-26
```

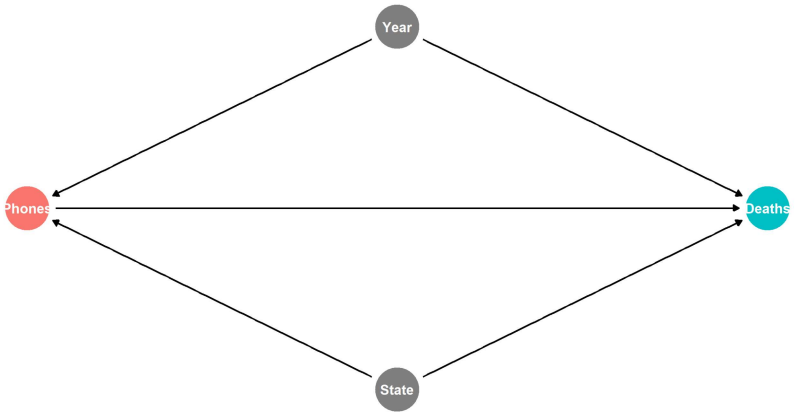
Two-Way Fixed Effects

- 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 .hi[one-way fixed effects model] estimates a fixed effect for **groups**
- .hi[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
  - 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

Visualizing Year Effects I

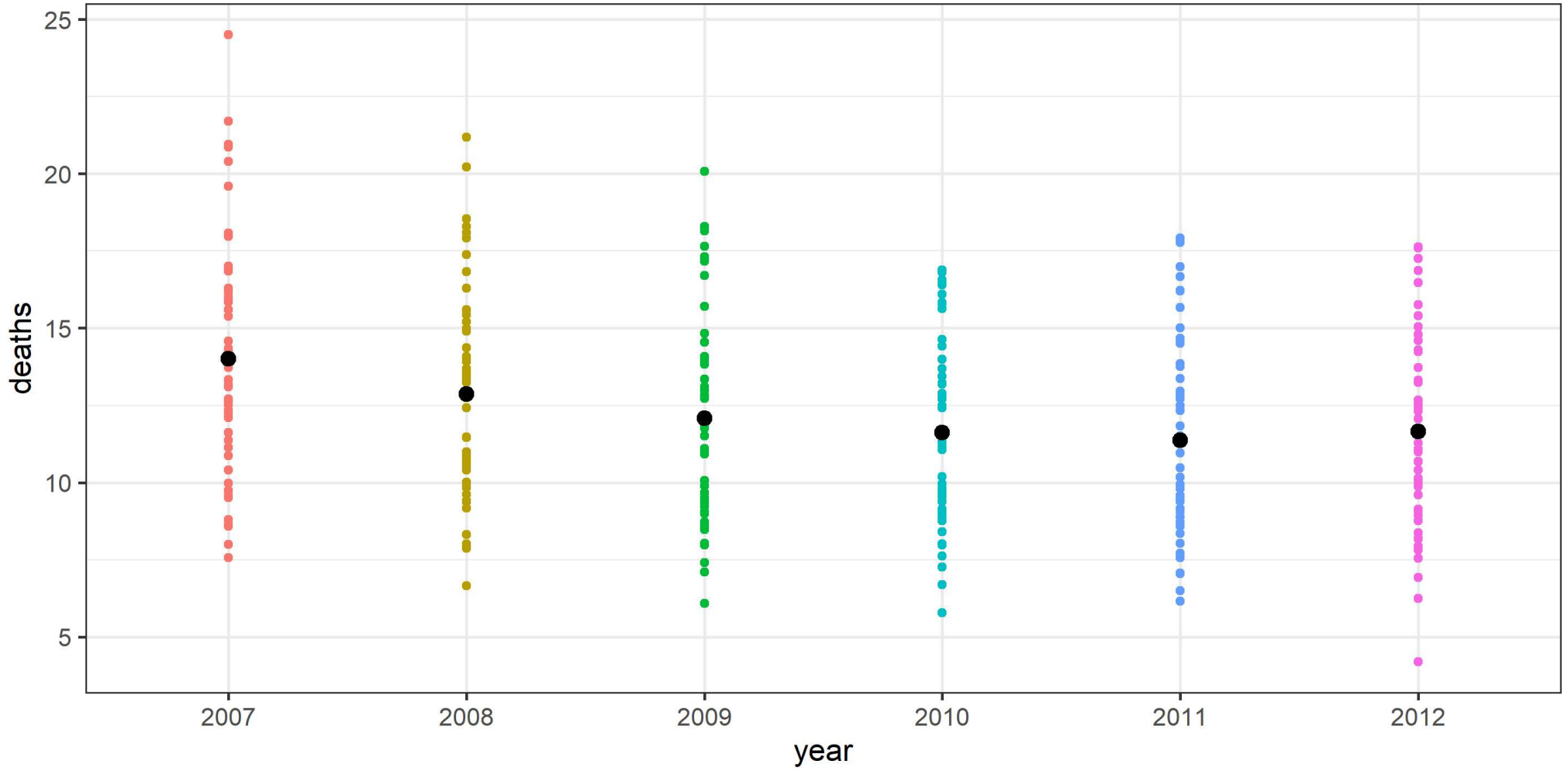
```
# find averages for years
means_year<-phones %>%
  group_by(year) %>%
  summarize(avg_deaths = mean(deaths),
            avg_phones = mean(cell_plans))
means_year
## # A tibble: 6 × 3
##   year  avg_deaths avg_phones
##   <fct>      <dbl>      <dbl>
## 1 2007        14.0      8065.
## 2 2008        12.9      8483.
## 3 2009        12.1      8860.
## 4 2010        11.6      9135.
## 5 2011        11.4      9485.
## 6 2012        11.7      9660.
```

Visualizing Year Effects II

```
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")+

  geom_path(data = means_year,
            aes(x = year,
                y = avg_deaths),
            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: $\tilde{var}_{it} = var_{it} - \overline{var}_t - \overline{var}_i$

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

LSDV Method

```
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
## # ℹ Use `print(n = ...)` to see more rows
```

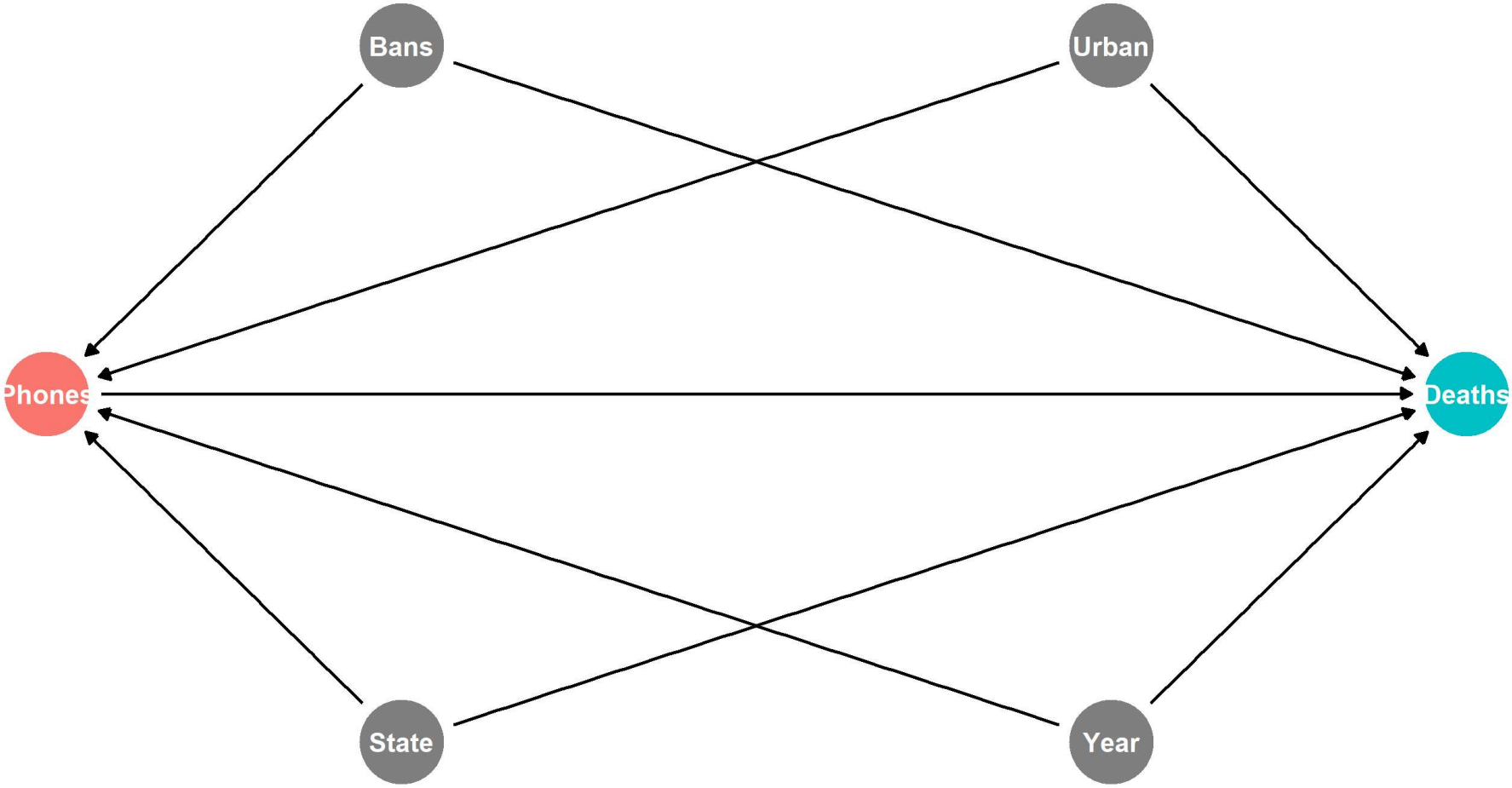
With plm

```
fe2_reg_2 <- plm(deaths ~ cell_plans,
                 index = c("state", "year"),
                 model = "within",
                 data = phones)
fe2_reg_2 %>% tidy()
## # A tibble: 1 × 5
##   term       estimate std.error statistic  p.value
##   <chr>         <dbl>     <dbl>     <dbl>    <dbl>
## 1 cell_plans -0.00120  0.000101     -11.9 3.48e-26
```

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

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 .hi-pink[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 <- plm(deaths ~ cell_plans + text_ban + urban_percent + cell_ban,
                        data = phones,
                        index = c("state","year"),
                        model = "within",
                        effect = "twoways")

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  0.000173     -1.97  0.0502
## 2 text_ban1      0.256     0.222         1.15  0.251
## 3 urban_percent  0.0131    0.0112        1.17  0.242
## 4 cell_ban1     -0.680     0.403        -1.69  0.0929
```

Comparing Models

```
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)	(0.0002)
Cell Ban				-0.6798
				(0.4029)
Texting Ban				0.2559
				(0.2222)
Urbanization Rate				0.0131
				(0.0112)
N	306	306	306	306

R-Squared	0.0845	0.9055	0.9259	0.0329
SER	3.2791	1.1526	1.0310	

*** p < 0.001; ** p < 0.01; * p < 0.05.