DUKE FUQUA SCHOOL OF BUSINESS

Empirical Analysis for Strategy

Professor McDevitt

Winter 2021

Class 2

Announcements

OA1 feedback

OA2 graded next week

OA3 due Feb 27 8:59am

Status check on class so far

Agenda

Case Nike Vaporfly 4% Better?

Lecture Fixed Effects

Case Low Birth Weights

Roadmap

Last Experimental Design

Next Matching Models

Are Nike Vaporfly 4% Really 4% Faster?

The New York Times

Nike Says Its $250 Running Shoes Will Make You Run Much Faster. What if That's Actually True?

Key Facts

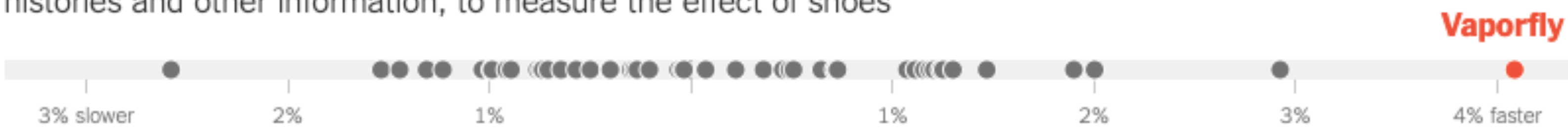
- The distinctive and controversial Nike Vaporfly 4% running shoe is supposed to improve running ease and speed by as much as 4%
- Using public race reports and shoe records from Strava, *The Times* found that runners in Vaporflys ran 3 to 4% faster than similar runners wearing other shoes, and more than 1% faster than the next-fastest racing shoe
- Runners choose to wear Vaporflys — they are not randomly assigned them

Conceptual Questions

- Would finding that runners who wore Vaporflys ran faster than those who wore other shoes be enough to conclude that the Vaporfly causes faster times?
- What are the other explanations for faster times ruled out in the article?
- Consider the Colorado study where runners wore 3 shoes in terms of fixed effects — is this a credible research design?
- What would be the ideal experiment to test the shoe's effectiveness?
- Compare the merits of each of the four methods for measuring the shoe's effectiveness that are described in the article

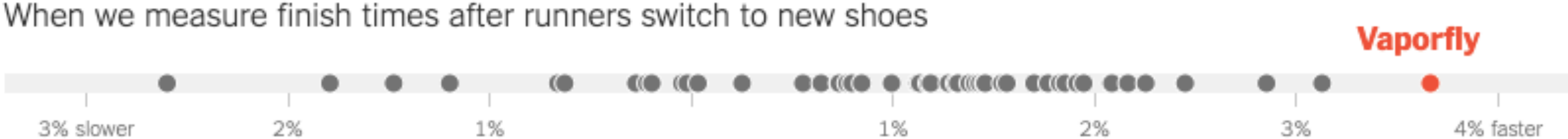
How the Nike Vaporflys compare with other popular running shoes ...

When we use a statistical model, based on runners' ages, genders, race histories and other information, to measure the effect of shoes

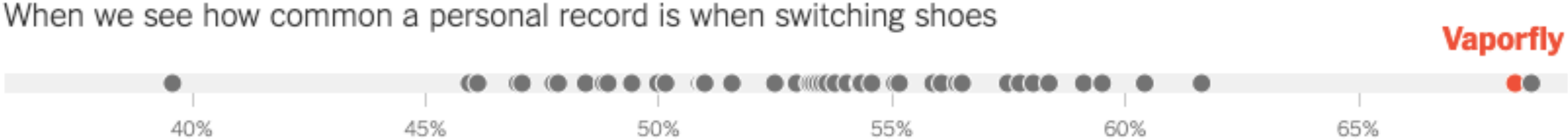


When we compare changes in race times among groups of runners who ran the same pairs of races

When we measure finish times after runners switch to new shoes



When we see how common a personal record is when switching shoes



Approach #1: Regression May Have Selection Bias

Measuring shoe effects using statistical models

Pros of this approach: Tries to control for race conditions, weather, gender, age, pre-race training and a runner's previous race times.

Cons of this approach: Still not a randomized controlled trial.

Approach #2: Matching Runners Reduces Some Bias, Not All (Next Class)

Comparing groups of runners who completed the same two races

Pros of this approach: Follows athletes of similar ability who ran in identical conditions.

Cons of this approach: Runners could save their special shoes for when they expect to have a fast race.

Approach #3: Runner Fixed Effect Better, But Still Some Selection Bias

Following runners as they switch to a new kind of racing shoe

Pros of this approach: Accounts for runners of varying skills over several races.

Cons of this approach: Runners could save Vaporflys for when they expect to be faster than normal, or Vaporfly wearers could be different in some way from other kinds of runners.

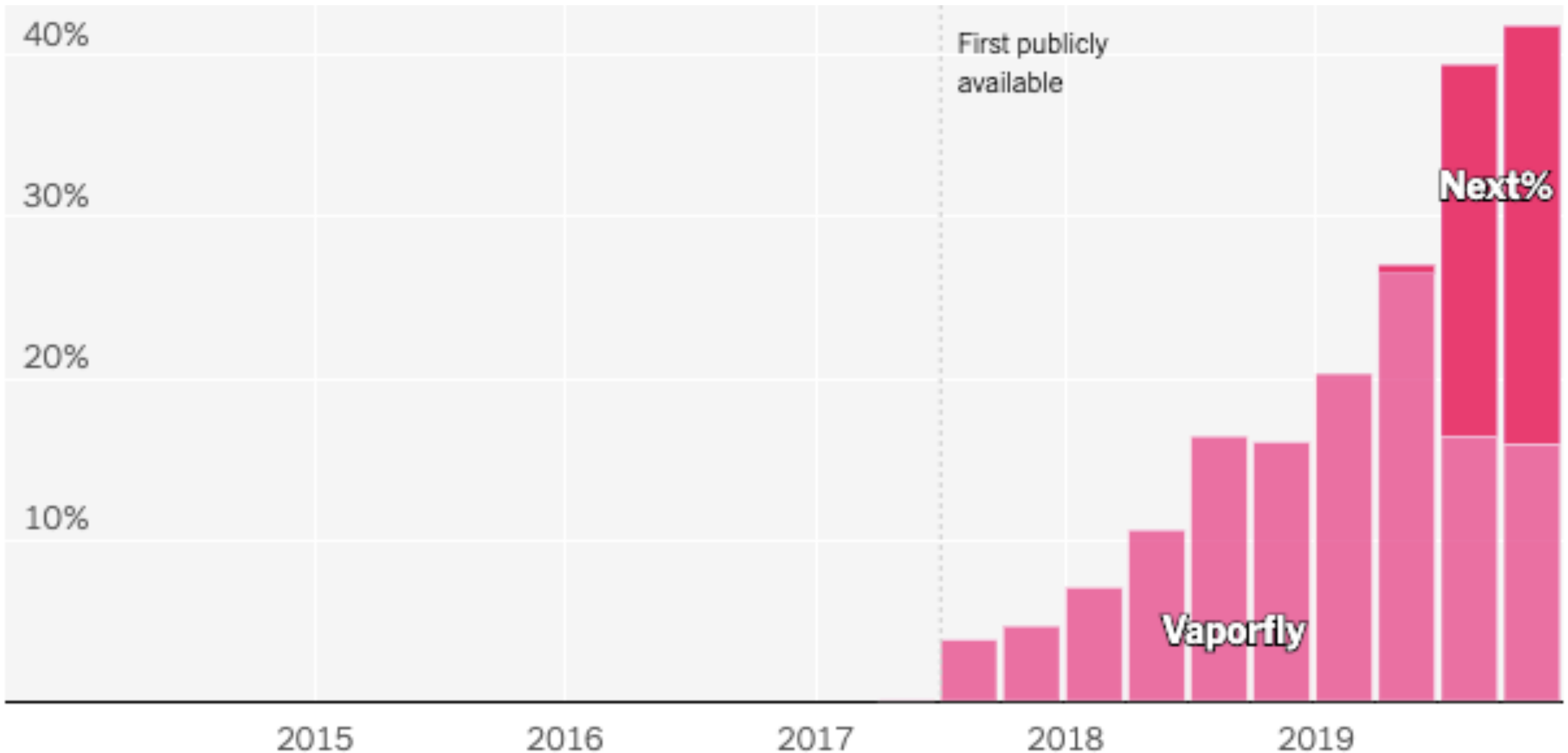
Approach #4: PR Differences-in-Differences an Intriguing Idea (Class 6)

Measuring the likelihood of a personal best

Pros of this approach: A measure of race performance most runners know by heart.

Cons of this approach: Doesn't account for race conditions, increased training miles or aging. Runners who switch to Vaporflys could be different from other runners.

Share of sub-3 marathons in which a runner reported wearing Vaporflys or Next%



Going from Nikes to ICDs → Outcomes Improve

EXHIBIT 1
Major Trials Of Implantable Cardioverter Defibrillators (ICDs), 1996–2004

Trial	Year published	Number of patients randomized	Hazard ratio (confidence limits)
Secondary prevention			
AVID	1997	1,016	0.62 (0.47–0.81)
CIDS	2000	659	0.82 (0.61–1.1)
CASH	2000	288	0.82 (0.6–1.1)
Primary prevention			
MADIT-I	1996	196	0.46 (0.26–0.82)
CABG-Patch	1997	900	1.07 (0.81–1.42)
MADIT-II	2002	1,242	0.69 (0.51–0.93)
DEFINITE	2004	458	0.65 (0.40–1.06)
COMPANION	2004	903	0.64 (0.48–0.86)
DINAMIT	–	674	1.08 (0.76–1.55)
SCD-HeFT	–	1,676	0.77 (0.62–0.96)

SOURCE: See Note 8 in text for an article summarizing the major trials. Individual citations for the trials are available from the authors.

NOTE: For more details on these trials, see descriptions in text.

Going from Nikes to ICDs → Is It Cost Effective?

EXHIBIT 2
Cost-Effectiveness Of Implantable Cardioverter Defibrillators (ICDs)

Indication	Life years added by ICD	Cost added by ICD ($)	Cost-effectiveness ratio ($)
Secondary prevention	0.69	37,400	54,000
Primary prevention			
EF <30	1.01	53,600	53,000
EF 31–40	0.51	53,100	104,000
EF >40	0.26	59,800	230,000

SOURCE: See Notes 12 and 13 in text.
NOTE: EF is ejection fraction.

Fixed Effects Regressions

Regression Overview

Regression is a statistical technique used to compare treatment and control groups while accounting for observed characteristics

- Causal inference requires that when key observed variables have been made equal across treatment and control groups, selection bias from the things we can't see has been mostly eliminated

- For instance, since the decision to have health insurance isn't made randomly, we must control for all factors that determine both health insurance status and health
  - For example, income, education, gender, etc.

Regression Setup

Consider the standard regression model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

- Observations are indexed by i = 1,…,n
- Y is the dependent variable, or outcome of interest (e.g., health status)
- X_1 and X_2 are the independent variables (e.g., health insurance, income)
- β_0 is the unknown intercept
- β_1 is the effect on Y of a change in X_1, holding X_2 constant
- β_2 is the effect on Y of a change in X_2, holding X_1 constant
- ε_i is the regression error, which reflects all omitted factors
  - That is, anything that affects Y other than X_1 and X_2
  - ε_i = "εverything εlse"

Interpreting Coefficients

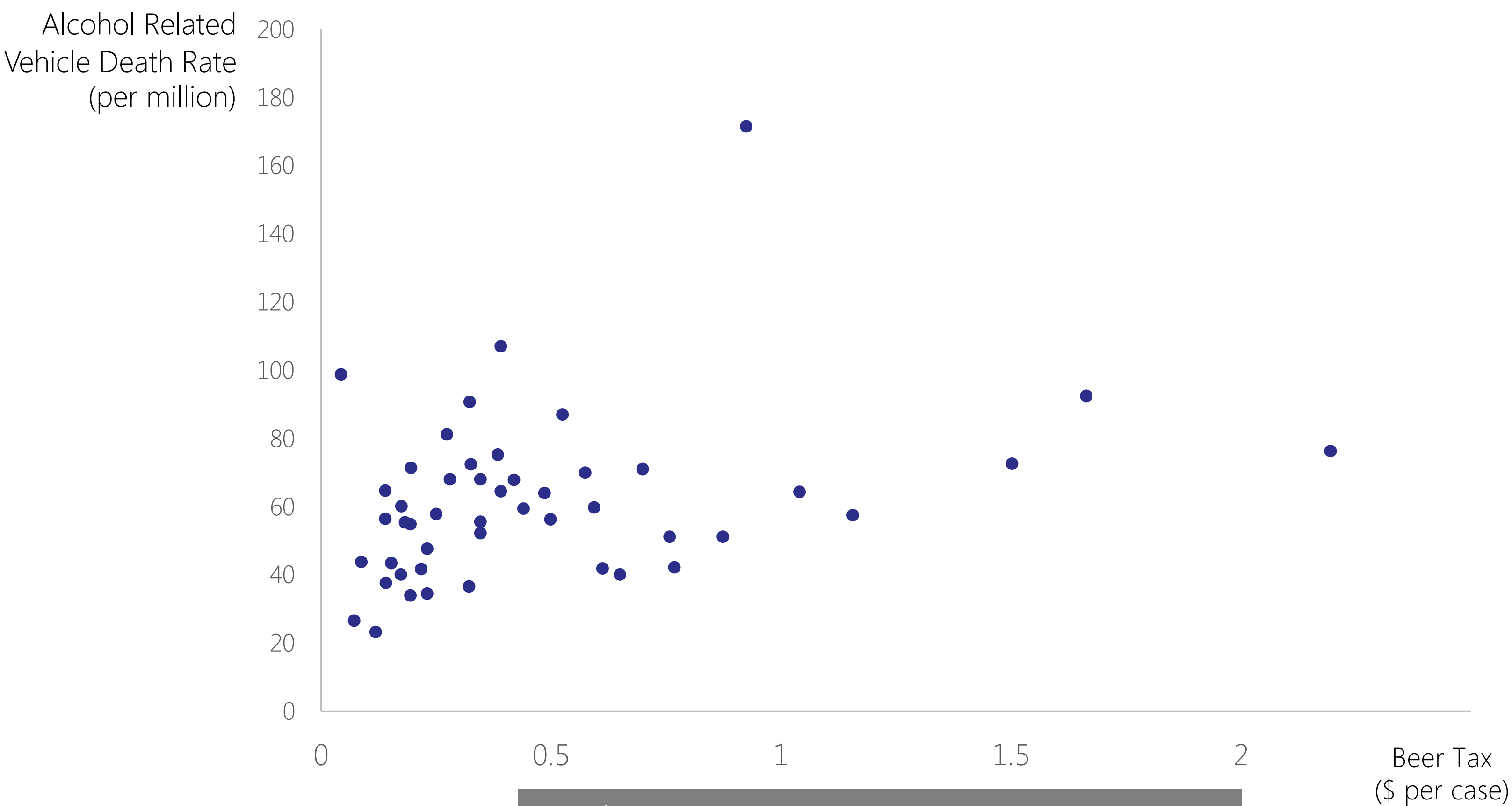
Consider changing X_1 by $\triangle X_1$ while holding X_2 constant

- Estimated line before the change is $Y = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}$
- Estimated line after the change is $Y + \triangle Y = \beta_0 + \beta_1 (X_1 + \triangle X_1) + \beta_2 X_2$

So the difference is $\triangle Y = \beta_1 \triangle X_1$

- $\beta_1 = \triangle Y / \triangle X_1$, holding X_2 fixed
- β_0 = predicted value of Y when $X_1 = X_2 = 0$

Example: Alcohol Related Traffic Deaths in 1988

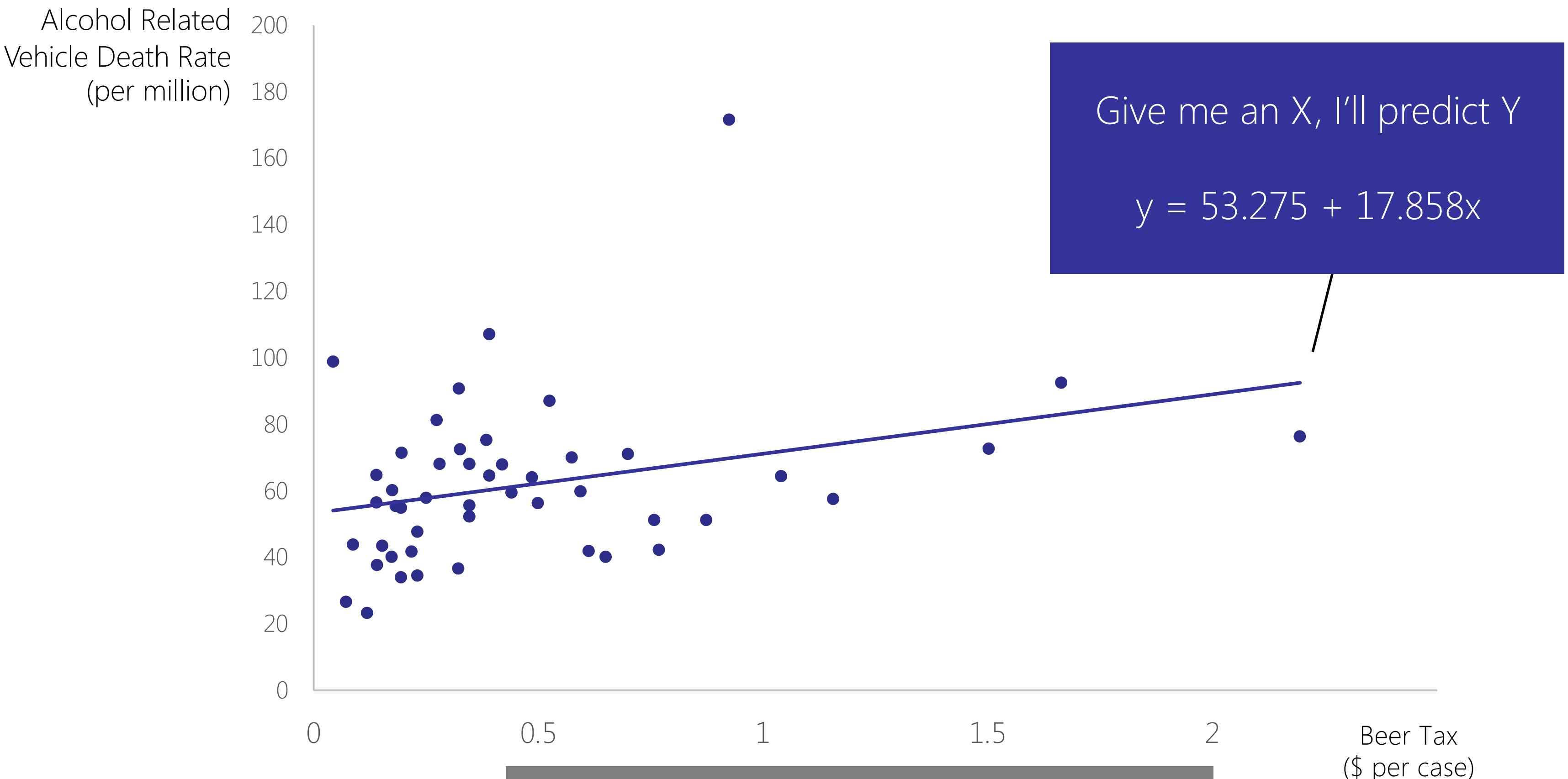


X_i = beer tax in state i
Y_i = alcohol related vehicle death rate in state i

Regression Output

```
. reg deaths_per_mil beertax if year==1988

      Source |       SS           df       MS      Number of obs   =        48
-------------+----------------------------------   F(1, 46)        =      5.13
       Model |  2834.25114         1  2834.25114   Prob > F        =    0.0283
    Residual |  25429.2467        46  552.809711   R-squared       =    0.1003
-------------+----------------------------------   Adj R-squared   =    0.0807
       Total |  28263.4978        47  601.351018   Root MSE        =    23.512

------------------------------------------------------------------------------
deaths_per~l |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 β1 beertax |   17.85849   7.887029     2.26   0.028     1.982725    33.73426
    β0 _cons |   53.27521   5.083102    10.48   0.000     43.04345    63.50696
------------------------------------------------------------------------------
```

Regression Best-Fit Line



X_i = beer tax in state i
Y_i = alcohol related vehicle death rate in state i

Omitted Variable Bias

Omitted variable bias occurs when we omit a variable from the regression that affects both X and Y

- Omitting that variable — denoted W — means the error term is correlated with the regressors (a technicality → violates assumptions for OLS)
  - We often refer to W as a confound
  - We wish we had data for W but we don't :(
- Example: regression of health insurance on health outcomes omits income, finds positive effect
  - higher income → more likely to have insurance (↑X)
  - higher income → better health (↑Y) irrespective of health insurance

By omitting W, we will mistakenly conclude that all of the impact on Y comes from X, even though part of it actually came from W

Omitted Variable Bias

CORONAVIRUS | 130,196 views | Jun 6, 2020, 11:26am EDT

Bald Men At Higher Risk Of Severe Coronavirus Symptoms

Marla Milling Contributor ⓘ

Healthcare

I am a Forbes.com Contributor specializing in geriatric health and women's health articles.

Omitted Variable Bias

Updated (6/8/20) This piece has been clarified to note that the study did not control for age, which is a risk factor for hair loss and severe Covid-19.

New research is showing why a larger percentage of men—particularly bald men—are

Omitted Variable Bias Simulation

Simulate data to resemble typical regression

- W, our confounding variable, can be equal to 0 or 1
- X, our explanatory variable, depends on W in the following way:

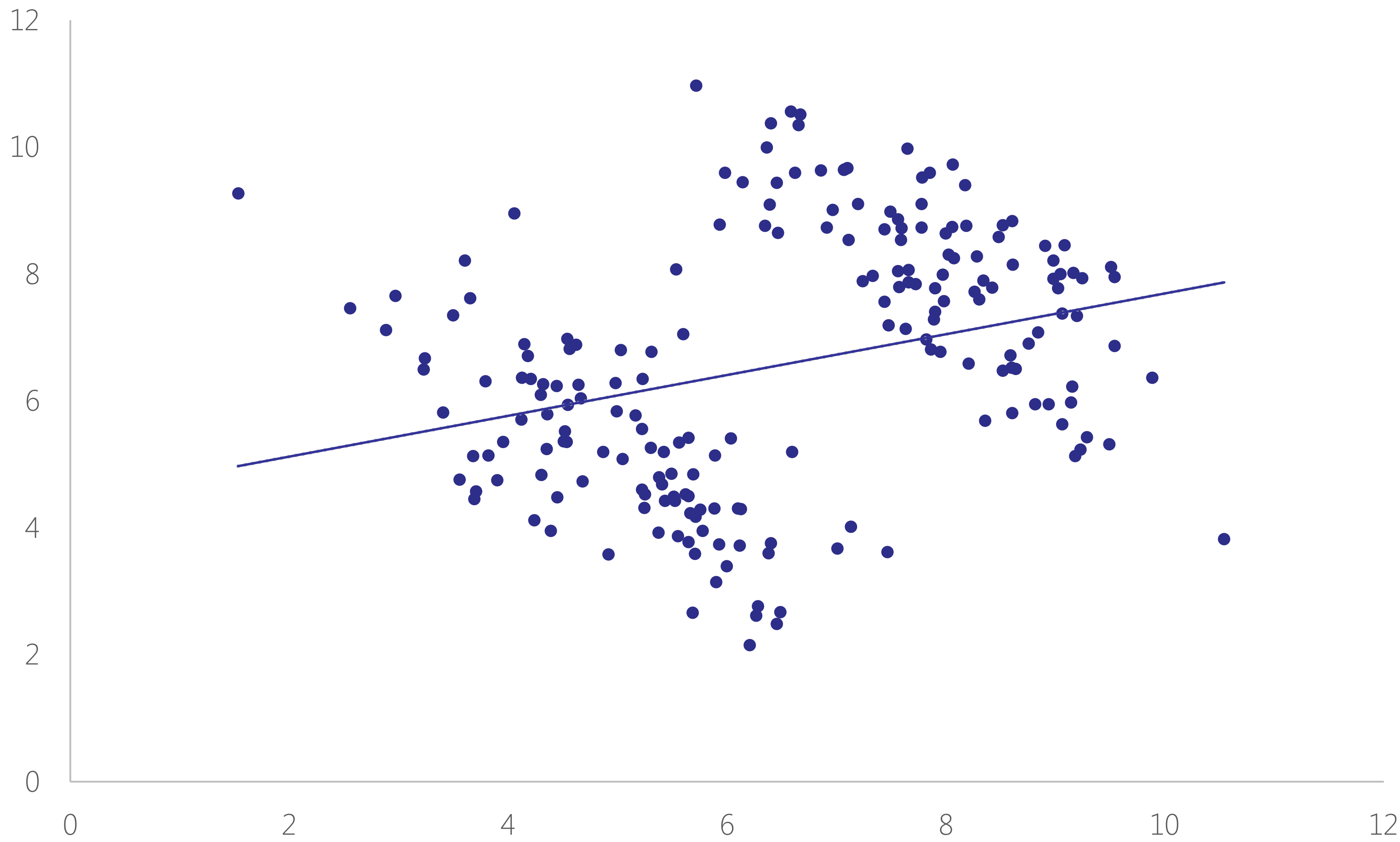
$$X = 5 + 3 \cdot W + \text{NORM}(0,1)$$

- Y, our outcome variable, depends on X & W in the following way:

$$Y = -1 \cdot X + 6 \cdot W + 10 + \text{NORM}(0,1)$$

- Notice that W affects both X and Y, so omitting it from a regression biases our results
- The causal effect of X on Y is -1
  - What we want to recover from our regression after controlling for W

Simulated Data Not Accounting for W

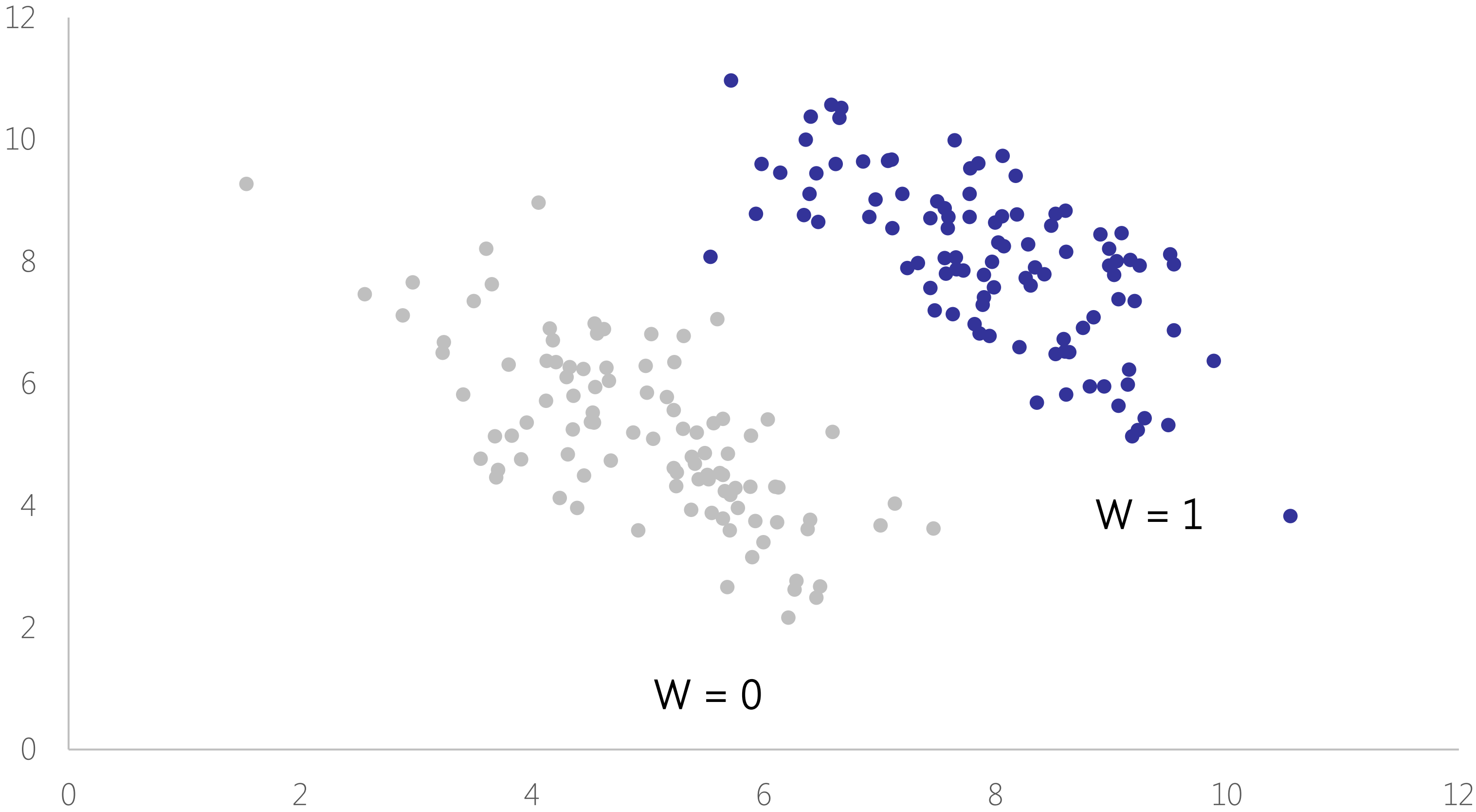


Omitted Variable W Biases Our Regression

```
. reg y x

      Source |       SS           df       MS      Number of obs   =       200
-------------+----------------------------------   F(1, 198)       =     17.64
       Model |  68.7438135         1  68.7438135   Prob > F        =    0.0000
    Residual |  771.826289       198  3.89811257   R-squared       =    0.0818
-------------+----------------------------------   Adj R-squared   =    0.0771
       Total |  840.570103       199  4.22397037   Root MSE        =    1.9744

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   .3214736   .0765518     4.20   0.000      .170512    .4724351
       _cons |   4.480085   .5155686     8.69   0.000     3.463374    5.496795
------------------------------------------------------------------------------

. reg y x w

      Source |       SS           df       MS      Number of obs   =       200
-------------+----------------------------------   F(2, 197)       =    317.21
       Model |  641.402577         2  320.701289   Prob > F        =    0.0000
    Residual |  199.167526       197  1.01100267   R-squared       =    0.7631
-------------+----------------------------------   Adj R-squared   =    0.7607
       Total |  840.570103       199  4.22397037   Root MSE        =    1.0055

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |  -.9861896   .0673704   -14.64   0.000    -1.119049   -.8533298
           w |   5.848274   .2457287    23.80   0.000     5.363678     6.33287
       _cons |   10.03389   .3512761    28.56   0.000     9.341145    10.72663
------------------------------------------------------------------------------
```

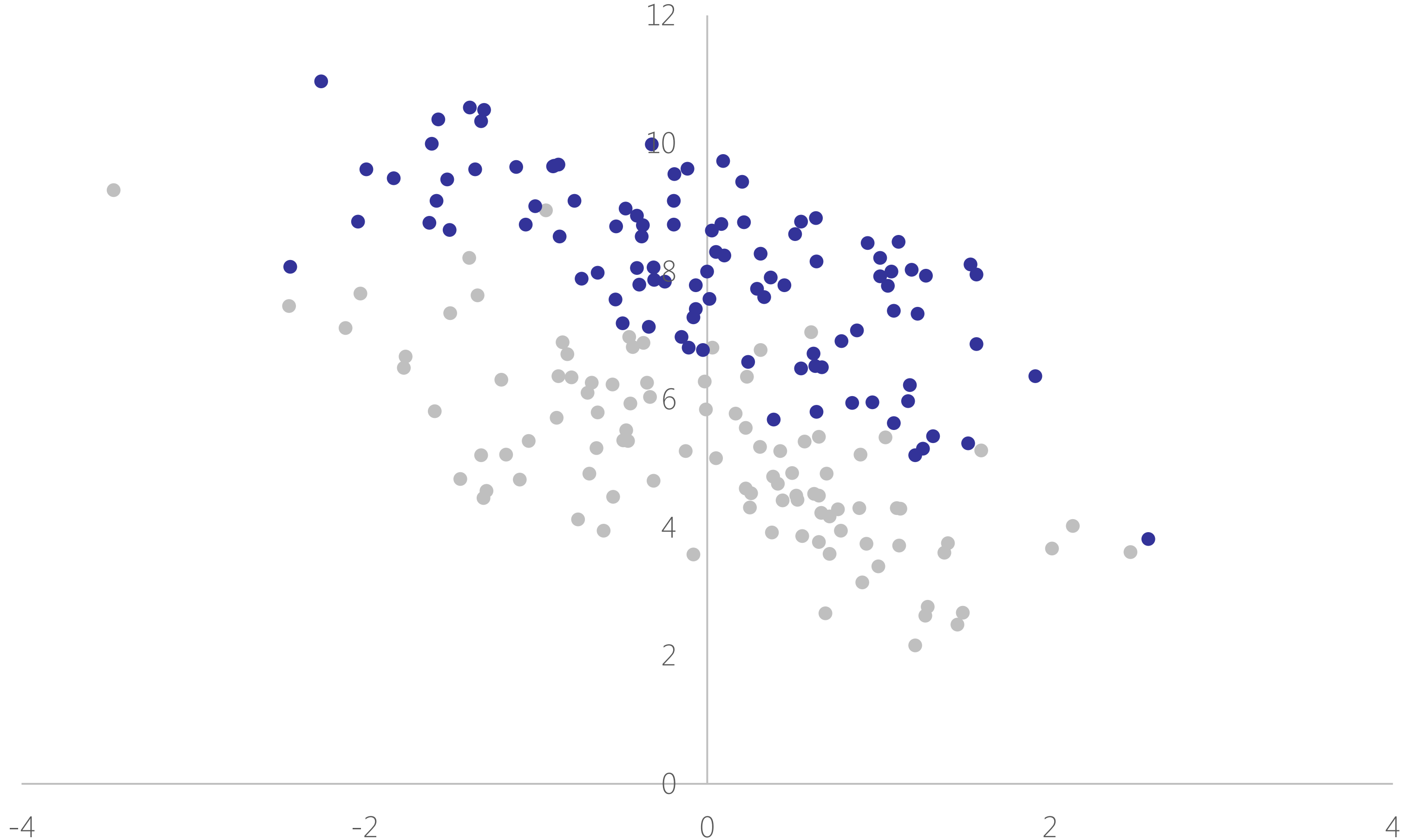
OVB =
0.321 – (-0.986) =
1.308

Need to Correct for Effect of W



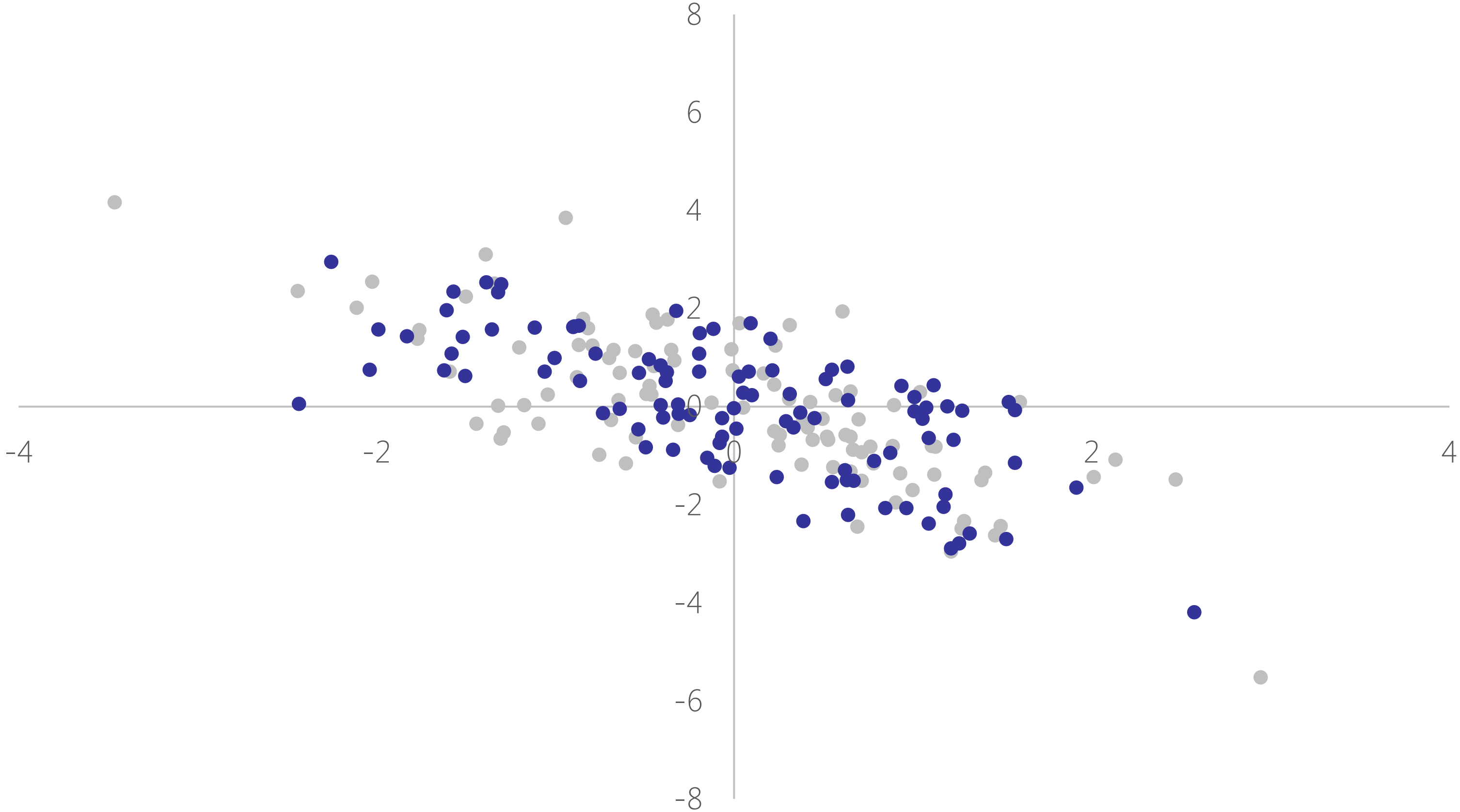
W = 1 makes both X & Y larger, but it looks like ↑X leads to ↑Y

Regression with W Removes Differences in X Explained by W



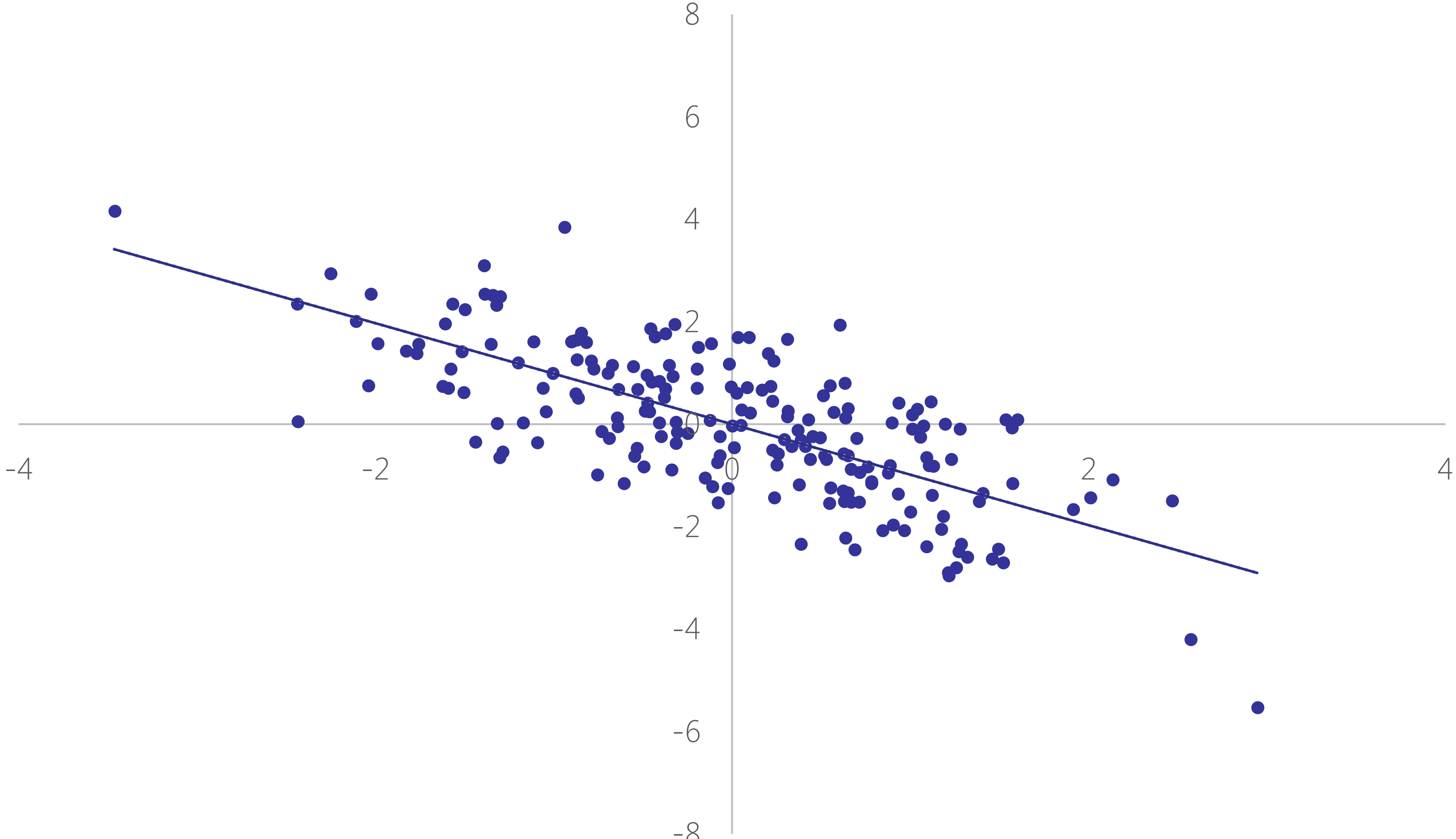
Controlling for W takes out the part of X driven by W

…As Well As the Differences in Y Explained by W



Controlling for W takes out the part of Y driven by W

…Which Allows Us to Estimate the Causal Impact of X on Y



Holding W fixed, this line tells us how much Y changes for any given X

Final Regression Output: Include W or De-Mean X & Y Gives Same Result

```
. reg y x w

      Source |       SS           df       MS      Number of obs   =       200
-------------+----------------------------------   F(2, 197)       =    317.21
       Model |  641.402577         2  320.701289   Prob > F        =    0.0000
    Residual |  199.167526       197  1.01100267   R-squared       =    0.7631
-------------+----------------------------------   Adj R-squared   =    0.7607
       Total |  840.570103       199  4.22397037   Root MSE        =    1.0055

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |  -.9861896   .0673704   -14.64   0.000    -1.119049   -.8533298
           w |   5.848274   .2457287    23.80   0.000     5.363678     6.33287
       _cons |   10.03389   .3512761    28.56   0.000     9.341145    10.72663
------------------------------------------------------------------------------

. reg demeanY demeanX

      Source |       SS           df       MS      Number of obs   =       200
-------------+----------------------------------   F(1, 198)       =    215.37
       Model |  216.637881         1  216.637881   Prob > F        =    0.0000
    Residual |  199.167525       198  1.00589659   R-squared       =    0.5210
-------------+----------------------------------   Adj R-squared   =    0.5186
       Total |  415.805406       199   2.0894744   Root MSE        =    1.0029

------------------------------------------------------------------------------
     demeanY |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     demeanX |  -.9861896   .0672001   -14.68   0.000    -1.118709   -.8536698
       _cons |   1.25e-07   .0709188     0.00   1.000    -.1398531    .1398533
------------------------------------------------------------------------------
```

Panel Data

A panel dataset contains observations on multiple entities (e.g., individuals, states, hospitals), where each entity is observed at two or more points in time

Person	Year	Age	Sex	Income	Education	Cancer
1	2010	45	F	$40,000	College	No
1	2011	46	F	$42,000	College	Yes
1	2012	47	F	$44,000	College	No
2	2010	53	M	$30,000	High School	No
2	2011	54	M	$30,000	High School	No
2	2012	55	M	$31,000	High School	No

Benefits of Panel Data

With panel data we can control for factors that

- Vary across entities but do not vary over time within entity (e.g., eye color)
- Could cause omitted variable bias if they are left out
- Are unobserved and therefore cannot be included in the regression directly

If an omitted variable does not change over time, then any changes in Y that occur over time could not have been caused by the omitted variable

→ this is the key idea of a fixed effects regression

Unobserved Heterogeneity

Unobserved heterogeneity refers to omitted factors that vary across individuals, like race, gender, family background, or innate ability

- If these factors affect both treatment and outcome, they will cause OVB

- Can see this in regression form

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \varepsilon_{it}$$

α_i
individual constant

u_{it}
time varying error term

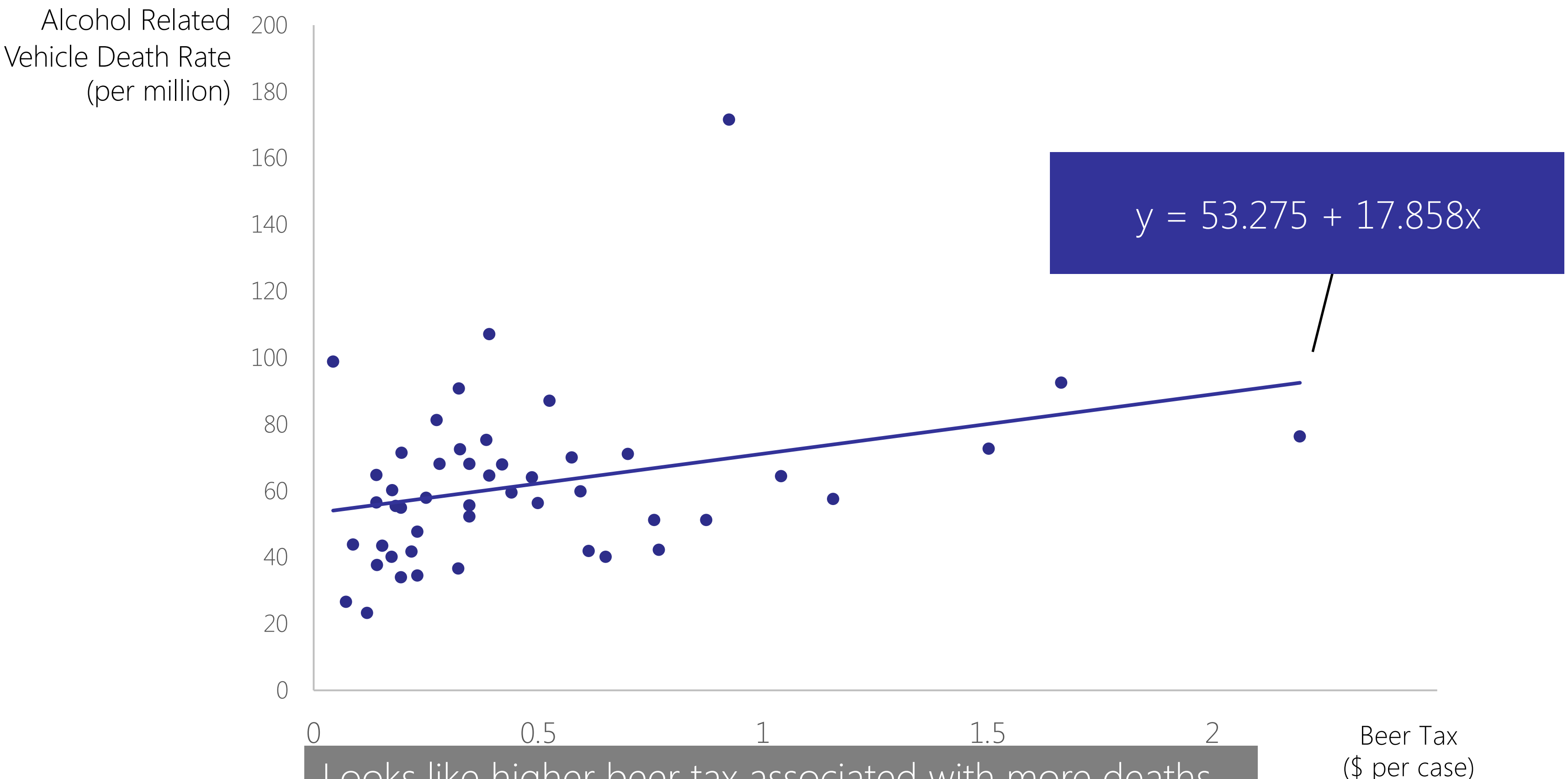
Fixed effects account for the **time-invariant** portion of unobserved heterogeneity, α_i, to reduce omitted variable bias

Example: Alcohol Related Traffic Deaths from 1982-1988

We have panel data for 48 states across 7 years

- i = state, n = 48
- t = 1982, … ,1988
- Balanced panel, so total # observations = 7•48 = 336
- Variables include
  - Traffic fatality rate (# traffic deaths per capita in that state in that year)
  - Tax on a case of beer
  - Other factors that might be important, like legal driving age, income, etc.

Recall Regression Using Cross-Section from 1988



Looks like higher beer tax associated with more deaths
→ demand slopes upward?!? → no, this is due to OVB

Many Omitted Variables Affect Both Taxes & Drunk Driving

Many other factors affect drunk driving death rate that may also be correlated with tax rate

- Quality of roads
  - States with high beer taxes have bad roads, bad roads kill people?
- Traffic density
  - States with high beer taxes have bad traffic, bad traffic kills people?
- Culture around drinking and driving
  - States with high beer taxes have bad culture, bad culture kills people?

Any of these unobserved confounds that remain constant within a state over time can be addressed using state fixed effects

Fixed Effects Intuition Using Two Years

Consider the regression

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

α_i
individual constant

u_{it}
time varying error term

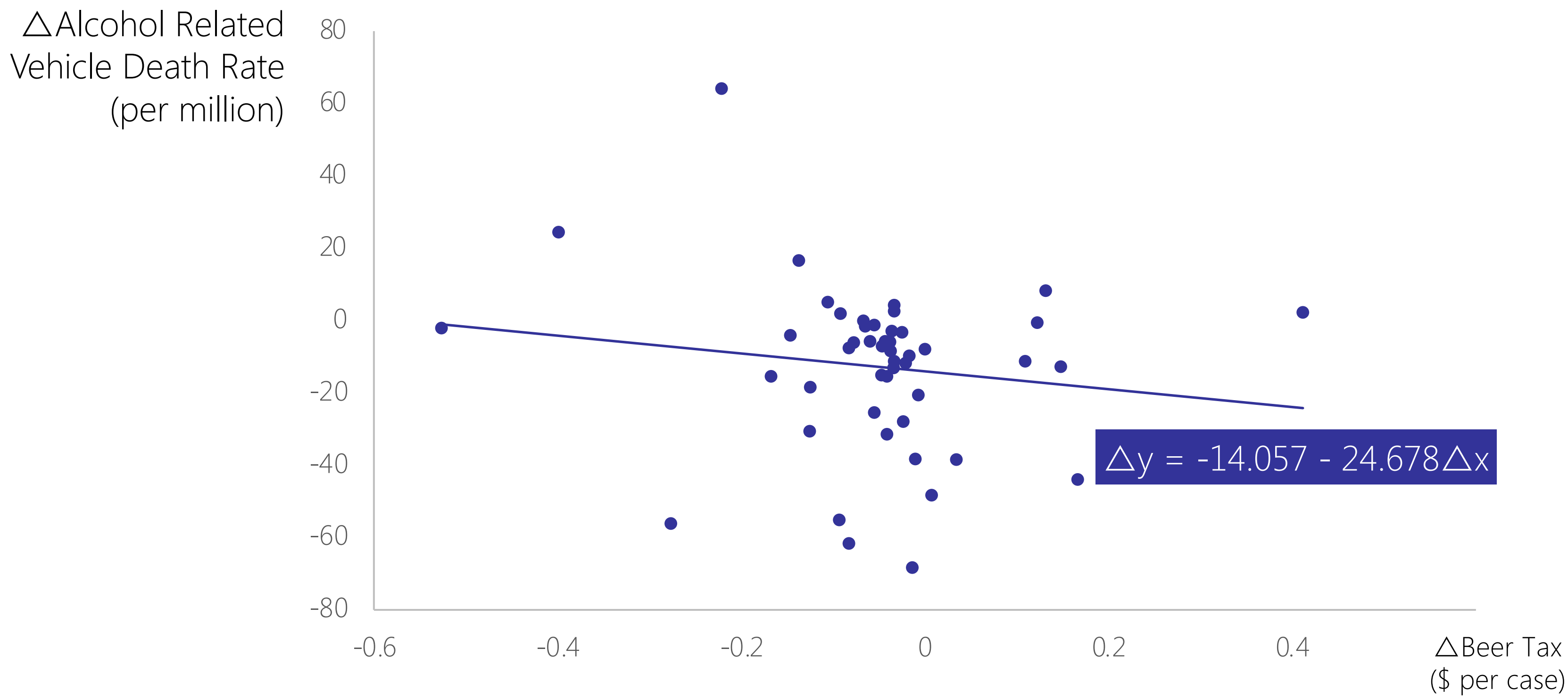
And the difference between 1988 and 1982

$$Y_{i,1988} - Y_{i,1982} = \beta_1(X_{i,1988} - X_{i,1982}) + (u_{i,1988} - u_{i,1982})$$

$\alpha_i Z_i$
cancels out

Any change in Y_i from 1982 to 1988 cannot be caused by Z_i because Z_i does not change between 1982 and 1988

First-Differences Regression Using 1982 & 1988 Data



State fixed effects capture constant unobserved heterogeneity → corrects OVB → now have negative relationship btwn death & taxes

Back to Regression Model: Consider OLS without Controls

```
. reg mraidall beertax

      Source |       SS           df       MS      Number of obs   =       336
-------------+----------------------------------   F(1, 334)       =     30.89
       Model |  1.9121e-08         1  1.9121e-08   Prob > F        =    0.0000
    Residual |  2.0678e-07       334  6.1910e-10   R-squared       =    0.0846
-------------+----------------------------------   Adj R-squared   =    0.0819
       Total |  2.2590e-07       335  6.7432e-10   Root MSE        =    2.5e-05

------------------------------------------------------------------------------
    mraidall |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     beertax |   .0000158   2.84e-06     5.56   0.000     .0000102    .0000214
       _cons |   .0000578   1.99e-06    29.00   0.000     .0000539    .0000617
------------------------------------------------------------------------------
```

Even with 7 years of panel data it still looks like beer taxes have a statistically significant, positive effect on drunk driving deaths

Adding Controls Brings Down Effect of Beer Tax

```
. reg mraidall beertax unrate perinc mlda vmiles

      Source |       SS           df       MS      Number of obs   =       336
-------------+----------------------------------   F(5, 330)       =     33.11
       Model |  7.5466e-08         5  1.5093e-08   Prob > F        =    0.0000
    Residual |  1.5043e-07       330  4.5586e-10   R-squared       =    0.3341
-------------+----------------------------------   Adj R-squared   =    0.3240
       Total |  2.2590e-07       335  6.7432e-10   Root MSE        =    2.1e-05

------------------------------------------------------------------------------
    mraidall |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     beertax |   4.26e-06   2.73e-06     1.56   0.119    -1.10e-06    9.63e-06
      unrate |   6.57e-07   6.19e-07     1.06   0.289    -5.60e-07    1.87e-06
      perinc |  -5.20e-09   7.16e-10    -7.26   0.000    -6.61e-09   -3.79e-09
        mlda |  -1.66e-06   1.35e-06    -1.23   0.219    -4.31e-06    9.90e-07
      vmiles |   3.23e-09   8.61e-10     3.75   0.000     1.54e-09    4.93e-09
       _cons |   .0001395   .0000325     4.29   0.000     .0000756    .0002035
------------------------------------------------------------------------------
```

Adding proper controls makes beer tax statistically insignificant because they reduce OVB → income and miles driven matter more

State Fixed Effects Capture Constant Unobserved Factors for Each State

```
. reg mraidall beertax unrate perinc mlda vmiles i.state

      Source |       SS           df       MS      Number of obs   =       336
-------------+----------------------------------   F(52, 283)      =     18.48
       Model |  1.7450e-07        52  3.3557e-09   Prob > F        =    0.0000
    Residual |  5.1401e-08       283  1.8163e-10   R-squared       =    0.7725
-------------+----------------------------------   Adj R-squared   =    0.7306
       Total |  2.2590e-07       335  6.7432e-10   Root MSE        =    1.3e-05

------------------------------------------------------------------------------
   mraidall |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     beertax |  -.0000187    .000014    -1.33   0.185    -.0000463    9.00e-06
      unrate |  -1.22e-06   7.84e-07    -1.56   0.120    -2.77e-06    3.21e-07
      perinc |  -2.99e-09   1.65e-09    -1.81   0.072    -6.24e-09    2.65e-10
        mlda |  -2.50e-06   1.45e-06    -1.72   0.086    -5.36e-06    3.55e-07
      vmiles |  -1.10e-09   7.45e-10    -1.48   0.141    -2.57e-09    3.66e-10
             |
       state |
        AZ   |  -.0000342   .0000196    -1.75   0.082    -.0000728    4.32e-06
        AR   |  -5.85e-06   .0000164    -0.36   0.722    -.0000382    .0000265
        CA   |  -.0000438    .000023    -1.91   0.058     -.000089    1.43e-06
                     OTHER STATES ESTIMATED BUT RESULTS OMITTED
        WI   |  -.0000517   .0000218    -2.37   0.018    -.0000946   -8.77e-06
        WY   |   2.41e-06    .000023     0.10   0.917    -.0000428    .0000476
             |
       _cons |   .0002167   .0000437     4.96   0.000     .0001308    .0003026
```

State + Time Fixed Effects More Robust → Now Estimate Negative Tax Effect

```
. reg mraidall beertax unrate perinc mlda vmiles i.state i.year

      Source |       SS           df       MS      Number of obs   =       336
-------------+----------------------------------   F(58, 277)      =     20.17
       Model |  1.8266e-07        58  3.1493e-09   Prob > F        =    0.0000
    Residual |  4.3242e-08       277  1.5611e-10   R-squared       =    0.8086
-------------+----------------------------------   Adj R-squared   =    0.7685
       Total |  2.2590e-07       335  6.7432e-10   Root MSE        =    1.2e-05

------------------------------------------------------------------------------
    mraidall |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     beertax |  -.0000249   .0000132    -1.88   0.061    -.0000509    1.14e-06
      unrate |  -3.66e-06   8.47e-07    -4.32   0.000    -5.33e-06   -2.00e-06
      perinc |   3.85e-10   1.71e-09     0.22   0.822    -2.99e-09    3.76e-09
        mlda |  -3.69e-07   1.43e-06    -0.26   0.796    -3.18e-06    2.44e-06
      vmiles |  -4.03e-10   7.04e-10    -0.57   0.568    -1.79e-09    9.82e-10
             |
       state |
         AZ  |  -.0000575   .0000188    -3.06   0.002    -.0000945   -.0000205
              OTHER STATES ESTIMATED BUT RESULTS OMITTED
             WY  |  -.0000214   .0000219    -0.98   0.330    -.0000645    .0000217
             |
        year |
        1983 |  -7.11e-06   2.56e-06    -2.78   0.006    -.0000121   -2.07e-06
        1984 |   -.000015   2.98e-06    -5.01   0.000    -.0000208   -9.08e-06
        1985 |  -.0000202   3.08e-06    -6.54   0.000    -.0000262   -.0000141
        1986 |  -.0000181   3.33e-06    -5.44   0.000    -.0000246   -.0000115
        1987 |  -.0000245   3.82e-06    -6.42   0.000    -.0000321    -.000017
        1988 |  -.0000279   4.30e-06    -6.47   0.000    -.0000363   -.0000194
             |
       _cons |   .0001813   .0000428     4.24   0.000     .0000971    .0002656
------------------------------------------------------------------------------
```

↓ trend in deaths

Fixed Effects Regression Simulation

Simulate data to resemble typical fixed effects regression

- W can be equal to 0, 1, 2, or 3 (we have four individuals in the data)
- X, our explanatory variable, varies across W in the following way

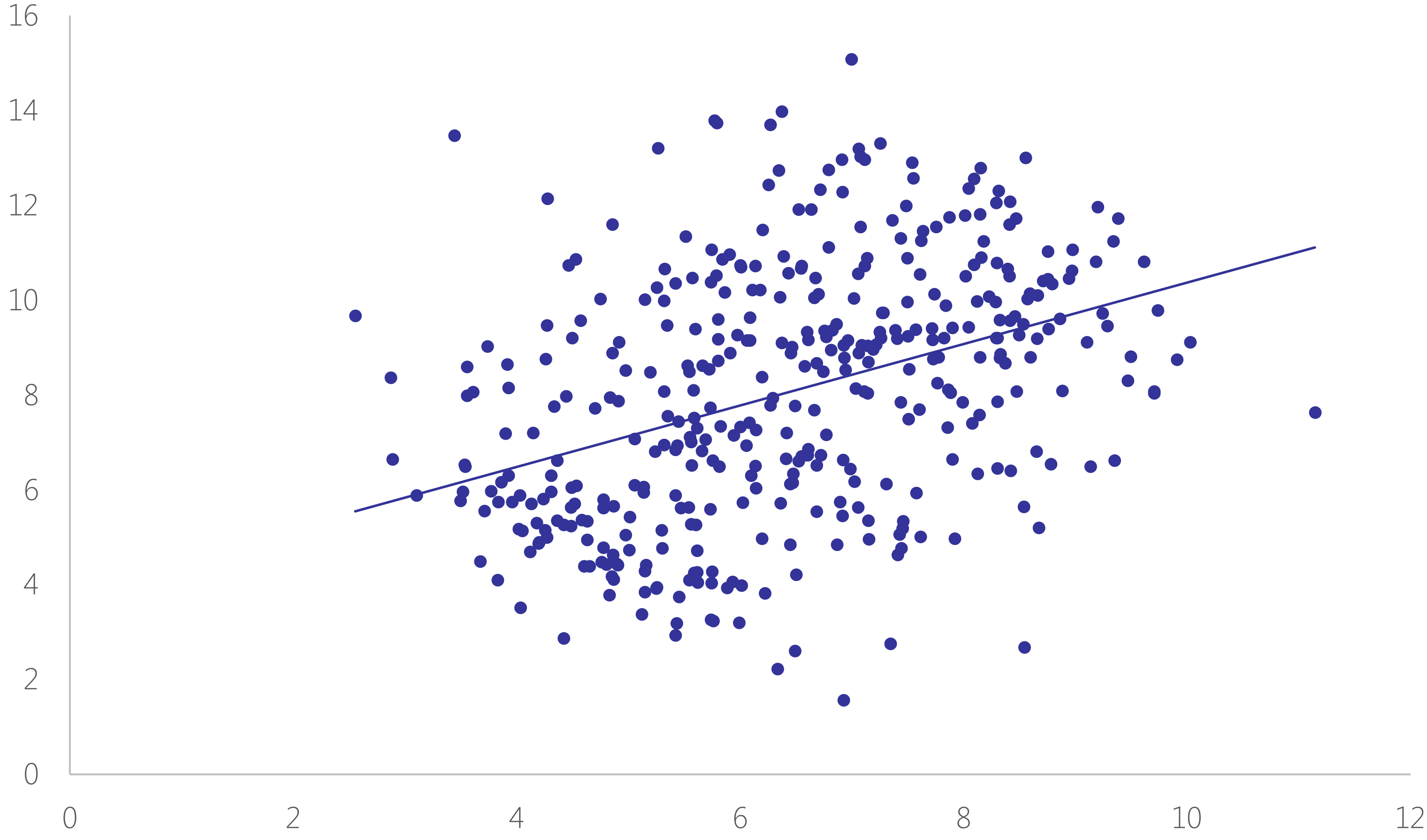
$$X = 5 + 1 \cdot W + \text{NORM}(0,1)$$

- Y, our outcome variable, depends on X & W in the following way

$$Y = -1 \cdot X + 3 \cdot W + 10 + \text{NORM}(0,1)$$

- W affects both X and Y, so omitting it from a regression will bias our results
  - We would mistakenly conclude that all of the impact on Y comes from X, even though part of it actually came from differences across individuals
- The causal effect of X on Y is -1
  - This is what we want to recover from our regression after controlling for the underlying differences across the four individuals

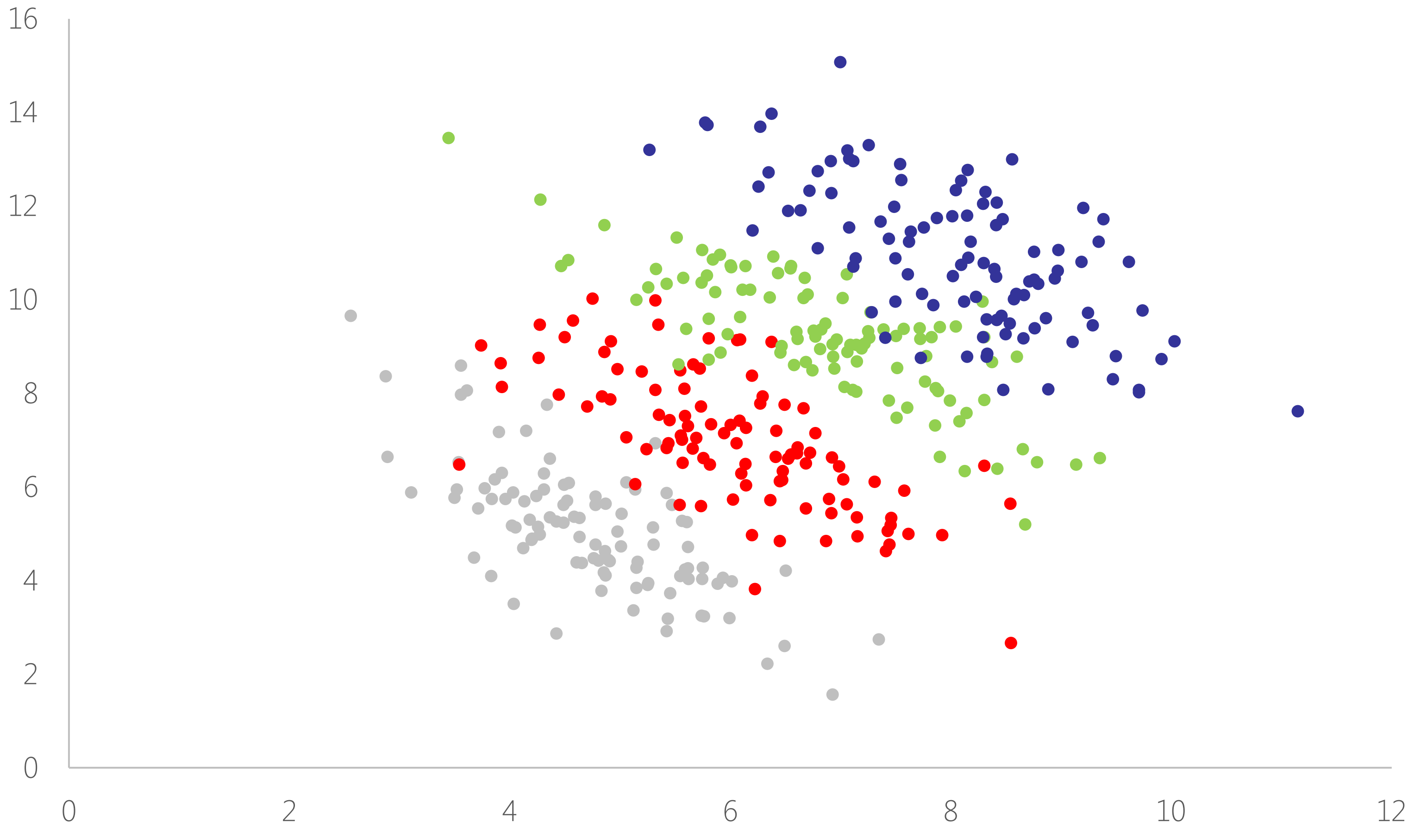
Simulated Data: Looks Like Positive Effect of X on Y without FE



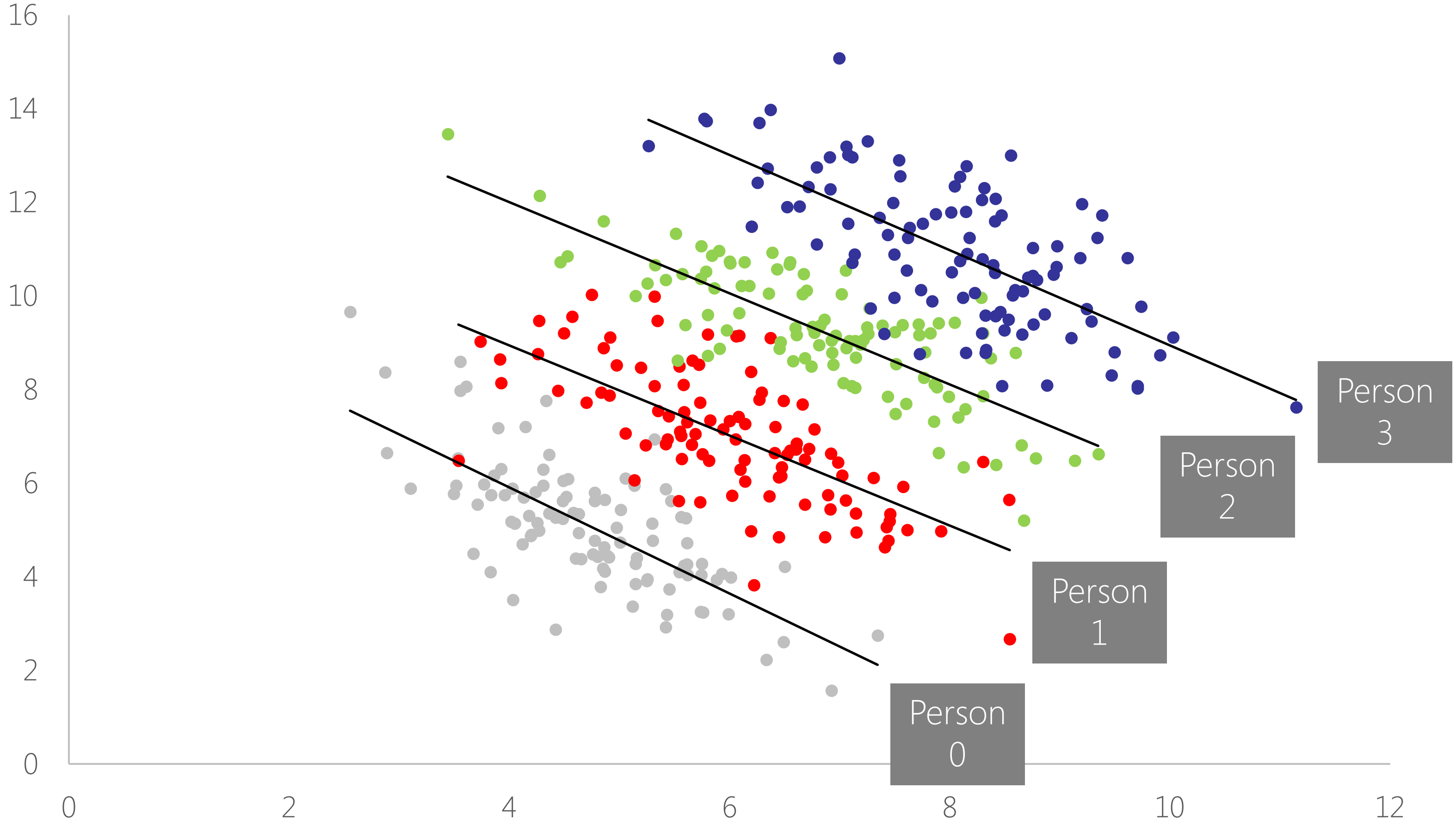
Omitted Differences Across Individuals Bias Regression

```
. reg y x

      Source |       SS           df       MS      Number of obs   =       400
-------------+----------------------------------   F(1, 398)       =     70.37
       Model |  418.922132         1  418.922132   Prob > F        =    0.0000
    Residual |  2369.49599       398  5.95350751   R-squared       =    0.1502
-------------+----------------------------------   Adj R-squared   =    0.1481
       Total |  2788.41812       399   6.9885166   Root MSE        =      2.44

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   .6477044   .0772141     8.39   0.000     .4959059    .7995028
       _cons |   3.892196   .5105287     7.62   0.000     2.888526    4.895866
------------------------------------------------------------------------------

. reg y x i.w

      Source |       SS           df       MS      Number of obs   =       400
-------------+----------------------------------   F(4, 395)       =    574.76
       Model |  2379.58005         4  594.895012   Prob > F        =    0.0000
    Residual |  408.838075       395   1.0350331   R-squared       =    0.8534
-------------+----------------------------------   Adj R-squared   =    0.8519
       Total |  2788.41812       399   6.9885166   Root MSE        =    1.0174

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |  -1.012382   .0501237   -20.20   0.000    -1.110925   -.9138395
           w |
          1  |   3.224634   .1577151    20.45   0.000     2.914568      3.5347
          2  |   6.294948   .1788431    35.20   0.000     5.943345    6.646551
          3  |   9.207425   .2200774    41.84   0.000     8.774756    9.640094
       _cons |   9.868695   .2583439    38.20   0.000     9.360794     10.3766
------------------------------------------------------------------------------
```

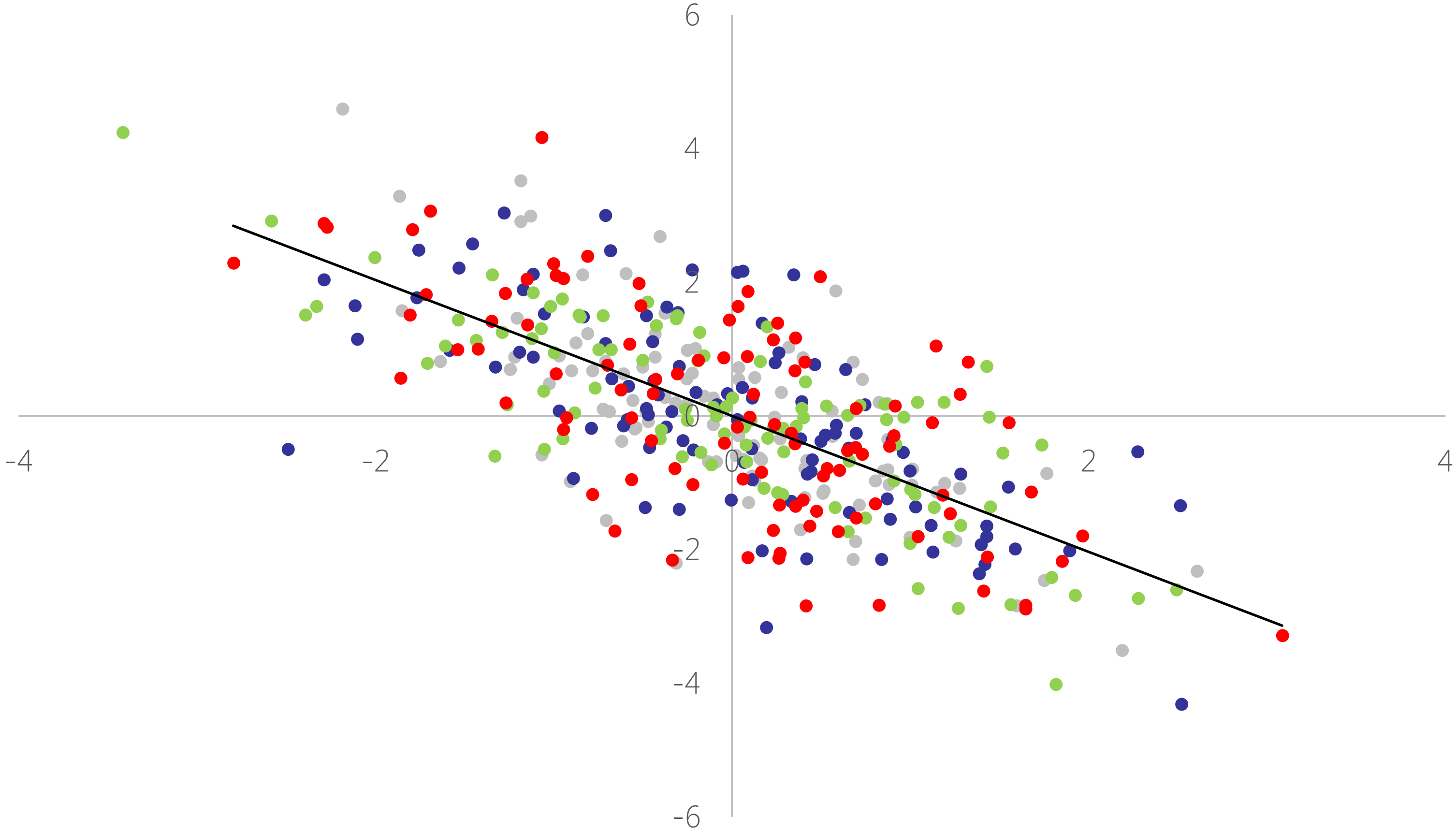
Biased Effect of X on Y Due to Differences Across Individuals



True Effect of X on Y Is Negative



Individual Fixed Effects De-Mean X and Y, Recover True Effect



Individual Fixed Effects De-Mean X and Y, Recover True Effect

```
. reg y x i.w

      Source |       SS           df       MS      Number of obs   =       400
-------------+----------------------------------   F(4, 395)       =    574.76
       Model |  2379.58005         4  594.895012   Prob > F        =    0.0000
    Residual |  408.838075       395   1.0350331   R-squared       =    0.8534
-------------+----------------------------------   Adj R-squared   =    0.8519
       Total |  2788.41812       399   6.9885166   Root MSE        =    1.0174

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |  -1.012382   .0501237   -20.20   0.000    -1.110925   -.9138395
             |
           w |
          1  |   3.224634   .1577151    20.45   0.000     2.914568      3.5347
          2  |   6.294948   .1788431    35.20   0.000     5.943345    6.646551
          3  |   9.207425   .2200774    41.84   0.000     8.774756    9.640094
             |
       _cons |   9.868695   .2583439    38.20   0.000     9.360794     10.3766
------------------------------------------------------------------------------
```

Person FE

Fixed effects reflect that each individual has Y about 3 times her index

→ person 1 is about 3 more than person 0, person 2 is 6 more

Summary of Fixed Effects Regressions

- With panel data and fixed effects we can control for factors that
  - Vary across entities but do not vary over time (e.g., state fixed effects)
  - Vary across time but do not vary across entities (i.e., time fixed effects)
- One caveat: you need variation in other key variables within entity
  - If beer tax is constant within states over time, you couldn't estimate its effect while also using state fixed effects
- One limitation: you can't tell the effect of that omitted variable
  - If eye color matters for health outcomes, a person fixed effect will account for that factor but won't tell you how much it matters
- This is very easy to implement using software
  - Just create a dummy variable for each entity and/or time period

The Cost of Low Birth Weight

The Cost of Low Birth Weight

Almond, Chay, & Lee (2005)

Key Facts

- Low-birth-weight infants experience severe health and developmental difficulties that can impose enormous costs on society
- It's not clear that efforts to prevent low-birth-weight infants would lead to commensurate cost savings and health improvements — some causes of low birth weight may be invariant to policy changes

Conceptual Questions

- What observable characteristics would be important to include in a regression of healthcare expenses on low birth weight?
- Would this regression likely provide a credible causal estimate of how low birth weights affect healthcare expenses?
- Why might estimates of the returns from preventing low birth weight using cross-sectional data be potentially biased? Why is this important for health policies?
- How does this study account for omitted variable bias to estimate a causal link between low birth weight and health care expenses?

Empirical Strategy

Use twin fixed effect to control for stable unobservable characteristics because twins have same mother, same environment, etc.

- $Y_{i1} = \beta\, bw_{i1} + \gamma\, X_{i1} + \alpha_i + \varepsilon_{i1}$
- $Y_{i2} = \beta\, bw_{i2} + \gamma\, X_{i2} + \alpha_i + \varepsilon_{i2}$

$$\triangle Y = Y_{i2} - Y_{i1} = \beta\, (bw_{i2} - bw_{i1}) + \gamma\, (X_{i2} - X_{i1}) + (\varepsilon_{i2} - \varepsilon_{i1})$$

Fixed Effects Allow Us to Distinguish Birth Weight from Other Factors

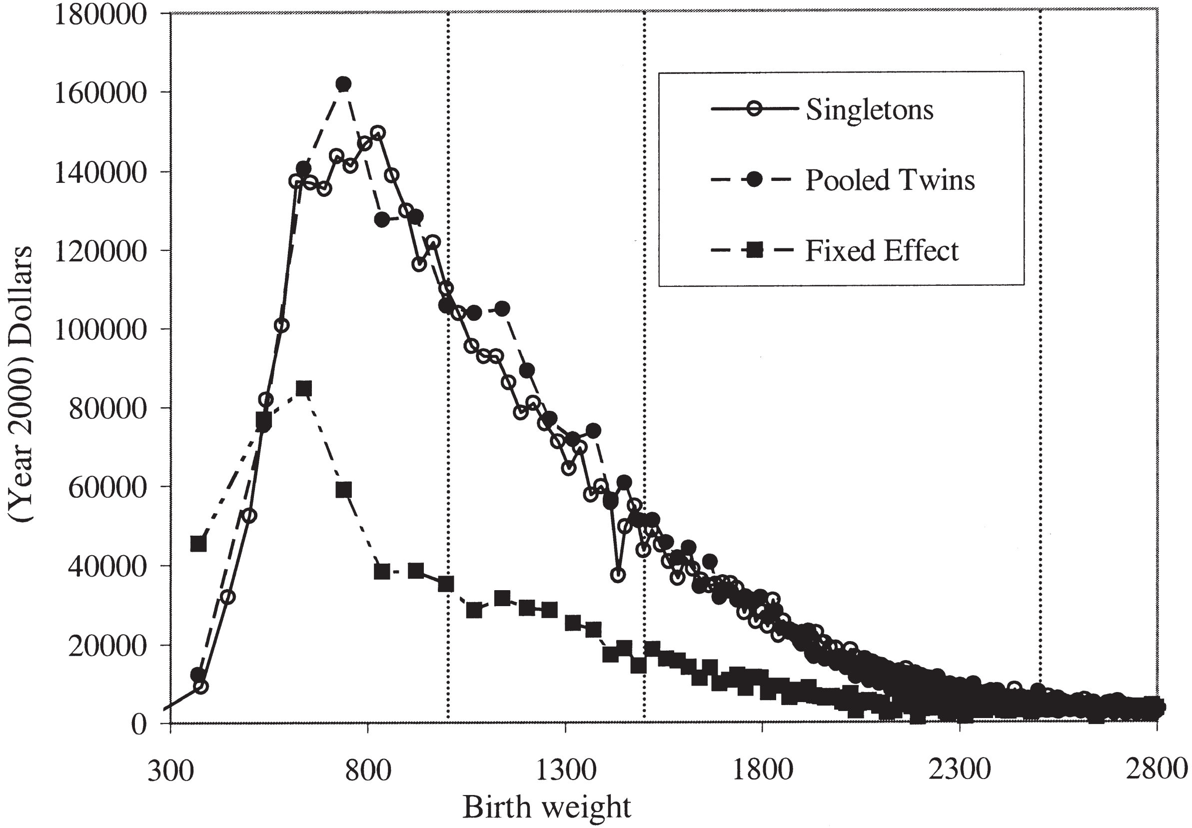


FIGURE Ia
Hospital Costs and Birth Weight
Note: 1995–2000 NY/NJ Hospital Discharge Microdata.

Unobserved Factors Lead Us to Overstate Effect of Birth Weight

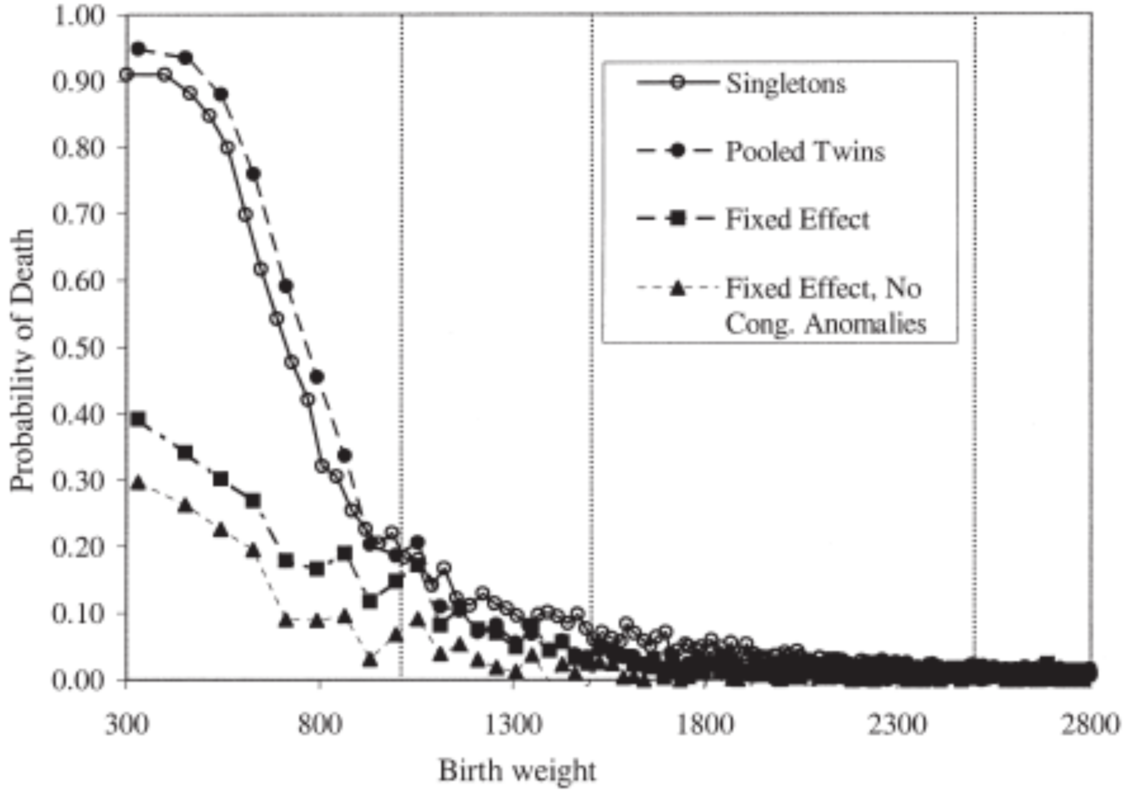


FIGURE Ib
Infant Mortality (1-year) and Birth Weight
Note: Linked Birth-Death certificate data, 1989.

All Effects Fall Considerably with a Fixed Effects Specification → OVB Matters

TABLE III
POOLED OLS AND TWINS FIXED EFFECTS ESTIMATES OF THE EFFECT OF BIRTH WEIGHT

Birth weight coefficient	Including congenital anomalies		Excluding congenital anomalies	
	Pooled OLS	Fixed effects	Pooled OLS	Fixed effects
Hospital costs	−29.95	−4.93	—	—
(in 2000 dollars)	(0.84)	(0.44)	—	—
	[−0.506]	[−0.083]	—	—
Adj. R^2	0.256	0.796	—	—
Sample size	44,410	44,410	—	—
Mortality, 1-year	−0.1168	−0.0222	−0.1069	−0.0082
(per 1000 births)	(0.0016)	(0.0016)	(0.0017)	(0.0012)
	[−0.412]	[−0.078]	[−0.377]	[−0.029]
Adj. R^2	0.169	0.585	0.164	0.629
Sample size	189,036	189,036	183,727	183,727
Mortality, 1-day	−0.0739	−0.0071	−0.0675	−0.0003
(per 1000 births)	(0.0015)	(0.0010)	(0.0015)	(0.0006)
	[−0.357]	[−0.034]	[−0.326]	[−0.001]
Adj. R^2	0.132	0.752	0.127	0.809
Sample size	189,036	189,036	183,727	183,727
Mortality, neonatal	−0.105	−0.0154	−0.0962	−0.0041
(per 1000 births)	(0.0016)	(0.0013)	(0.0016)	(0.0008)
	[−0.415]	[−0.061]	[−0.38]	[−0.016]
Adj. R^2	0.173	0.683	0.169	0.745
Sample size	189,036	189,036	183,727	183,727
5-min. APGAR score	0.1053	0.0117	0.1009	0.0069
(0–10 scale,	(0.0011)	(0.0012)	(0.0011)	(0.0011)
divided by 100)	[0.506]	[0.056]	[0.485]	[0.033]
Adj. R^2	0.255	0.663	0.248	0.673
Sample size	159,070	159,070	154,449	154,449

Estimated Effect of Birth Weight on Mortality Falls by a Factor of 10!

Birth weight coefficient	Including congenital anomalies		Excluding congenital anomalies	
	Pooled OLS	Fixed effects	Pooled OLS	Fixed effects
Mortality, neonatal	−0.105	−0.0154	−0.0962	−0.0041
(per 1000 births)	(0.0016)	(0.0013)	(0.0016)	(0.0008)
	[−0.415]	[−0.061]	[−0.38]	[−0.016]
Adj. R^2	0.173	0.683	0.169	0.745
Sample size	189,036	189,036	183,727	183,727

And Fixed Effects Explain Large Portion of Variation (the Unobserved Factors)

Birth weight coefficient	Including congenital anomalies		Excluding congenital anomalies	
	Pooled OLS	Fixed effects	Pooled OLS	Fixed effects
Mortality, neonatal (per 1000 births)	−0.105	−0.0154	−0.0962	−0.0041
	(0.0016)	(0.0013)	(0.0016)	(0.0008)
	[−0.415]	[−0.061]	[−0.38]	[−0.016]
Adj. R^2	0.173	0.683	0.169	0.745
Sample size	189,036	189,036	183,727	183,727

Q + A