Lecture 09

OLS // Regression

Ivan Rudik
AEM 4510

Roadmap

- Intro to regression and ordinary least squares

Regression and ordinary least squares

class: inverse, center, middle name: tidyverse

Why?

Let's start with a few **basic, general questions**

Why?

Let's start with a few **basic, general questions**

1. What is the goal of econometrics?

2. Why do economists (or other people) study or use econometrics?

Why?

Let's start with a few **basic, general questions**

1. What is the goal of econometrics?

2. Why do economists (or other people) study or use econometrics?

One simple answer: Learn about the world using data

Why? Example

GPA is an output from endowments (ability), and hours studied (inputs), and pollution exposure (externality)

Why? Example

GPA is an output from endowments (ability), and hours studied (inputs), and pollution exposure (externality)

One might hypothesize a model: $\mathrm{GPA} = f(I, P, \mathrm{SAT}, H)$

where H is hours studied, P is pollution exposure, SAT is SAT score and I is family income

Why? Example

GPA is an output from endowments (ability), and hours studied (inputs), and pollution exposure (externality)

One might hypothesize a model: $\mathrm{GPA} = f(I, P, \mathrm{SAT}, H)$

where H is hours studied, P is pollution exposure, SAT is SAT score and I is family income

We expect that GPA will rise with some variables, and decrease with others

Why? Example

GPA is an output from endowments (ability), and hours studied (inputs), and pollution exposure (externality)

One might hypothesize a model: $\mathrm{GPA} = f(I, P, \mathrm{SAT}, H)$

where H is hours studied, P is pollution exposure, SAT is SAT score and I is family income

We expect that GPA will rise with some variables, and decrease with others

But who needs to *expect*?

Why? Example

GPA is an output from endowments (ability), and hours studied (inputs), and pollution exposure (externality)

One might hypothesize a model: $\text{GPA} = f(I, P, \text{SAT}, H)$

where H is hours studied, P is pollution exposure, SAT is SAT score and I is family income

We expect that GPA will rise with some variables, and decrease with others

But who needs to *expect*?

We can test these hypotheses **using a regression model**

How?

We can write down a linear regression model of the relationship between GPA and (H, P, SAT, PCT):

$$\text{GPA}_i = \beta_0 + \beta_1 I_i + \beta_2 P_i + \beta_3 \text{SAT}_i + \beta_4 H_i + \varepsilon_i$$

How?

We can write down a linear regression model of the relationship between GPA and (H, P, SAT, PCT):

$$\text{GPA}_i = \beta_0 + \beta_1 I_i + \beta_2 P_i + \beta_3 \text{SAT}_i + \beta_4 H_i + \varepsilon_i$$

The left hand side of the equals sign is our **dependent variable** GPA

How?

We can write down a linear regression model of the relationship between GPA and (H, P, SAT, PCT):

$$\text{GPA}_i = \beta_0 + \beta_1 I_i + \beta_2 P_i + \beta_3 \text{SAT}_i + \beta_4 H_i + \varepsilon_i$$

The left hand side of the equals sign is our **dependent variable** GPA

The right hand side of the equals sign contains all of our **independent variables** (I, P, SAT, H), and an error term ε_i (described later)

How?

We can write down a linear regression model of the relationship between GPA and (H, P, SAT, PCT):

$$\text{GPA}_i = \beta_0 + \beta_1 I_i + \beta_2 P_i + \beta_3 \text{SAT}_i + \beta_4 H_i + \varepsilon_i$$

The left hand side of the equals sign is our **dependent variable** GPA

The right hand side of the equals sign contains all of our **independent variables** (I, P, SAT, H), and an error term ε_i (described later)

The subscript i means that the variable contains the value for some person i in our dataset where $i = 1, \ldots, N$

How?

$$\text{GPA}_i = \beta_0 + \beta_1 I_i + \beta_2 P_i + \beta_3 \text{SAT}_i + \beta_4 H_i + \varepsilon_i$$

We are interested in how pollution P affects GPA

How?

$$\text{GPA}_i = \beta_0 + \beta_1 I_i + \beta_2 P_i + \beta_3 \text{SAT}_i + \beta_4 H_i + \varepsilon_i$$

We are interested in how pollution P affects GPA

This is given by β_2

How?

$$\mathrm{GPA}_i = \beta_0 + \beta_1 I_i + \beta_2 P_i + \beta_3 \mathrm{SAT}_i + \beta_4 H_i + \varepsilon_i$$

We are interested in how pollution P affects GPA

This is given by β_2

Notice that $\beta_2 = \frac{\partial \mathrm{GPA}_i}{\partial \mathrm{P}_i}$

How?

$$\mathrm{GPA}_i = \beta_0 + \beta_1 I_i + \beta_2 P_i + \beta_3 \mathrm{SAT}_i + \beta_4 H_i + \varepsilon_i$$

We are interested in how pollution P affects GPA

This is given by β_2

Notice that $\beta_2 = \frac{\partial \mathrm{GPA}_i}{\partial \mathrm{P}_i}$

β_2 tells us how GPA changes, given a 1 unit increase in pollution!

How?

$$\text{GPA}_i = \beta_0 + \beta_1 I_i + \beta_2 P_i + \beta_3 \text{SAT}_i + \beta_4 H_i + \varepsilon_i$$

We are interested in how pollution P affects GPA

This is given by β_2

Notice that $\beta_2 = \frac{\partial \text{GPA}_i}{\partial \text{P}_i}$

β_2 tells us how GPA changes, given a 1 unit increase in pollution!

Our goal will be to estimate β_2, we denote estimates with hats: $\hat{\beta}_2$

How?

How do we estimate β_2?

How?

How do we estimate β_2?

First, suppose we have a set of estimates for all of our βs, then we can *estimate* the GPA ($\widehat{GPA}_i$) for any given person based on just (I, P, SAT, H):

$$\widehat{GPA}_i = \hat{\beta}_0 + \hat{\beta}_1 I_i + \hat{\beta}_2 P_i + \hat{\beta}_3 \text{SAT}_i + \hat{\beta}_4 H_i$$

How?

We estimate the βs with **linear regression**, specifically ordinary least squares

Ordinary least squares: choose all the βs so that the sum of squared errors between the *real* GPAs and model-estimated GPAs are minimized:

$$SSE = \sum_{i=1}^{N} (GPA_i - \widehat{GPA}_i)^2$$

How?

We estimate the βs with **linear regression**, specifically ordinary least squares

Ordinary least squares: choose all the βs so that the sum of squared errors between the *real* GPAs and model-estimated GPAs are minimized:

$$SSE = \sum_{i=1}^{N} (GPA_i - \widehat{GPA}_i)^2$$

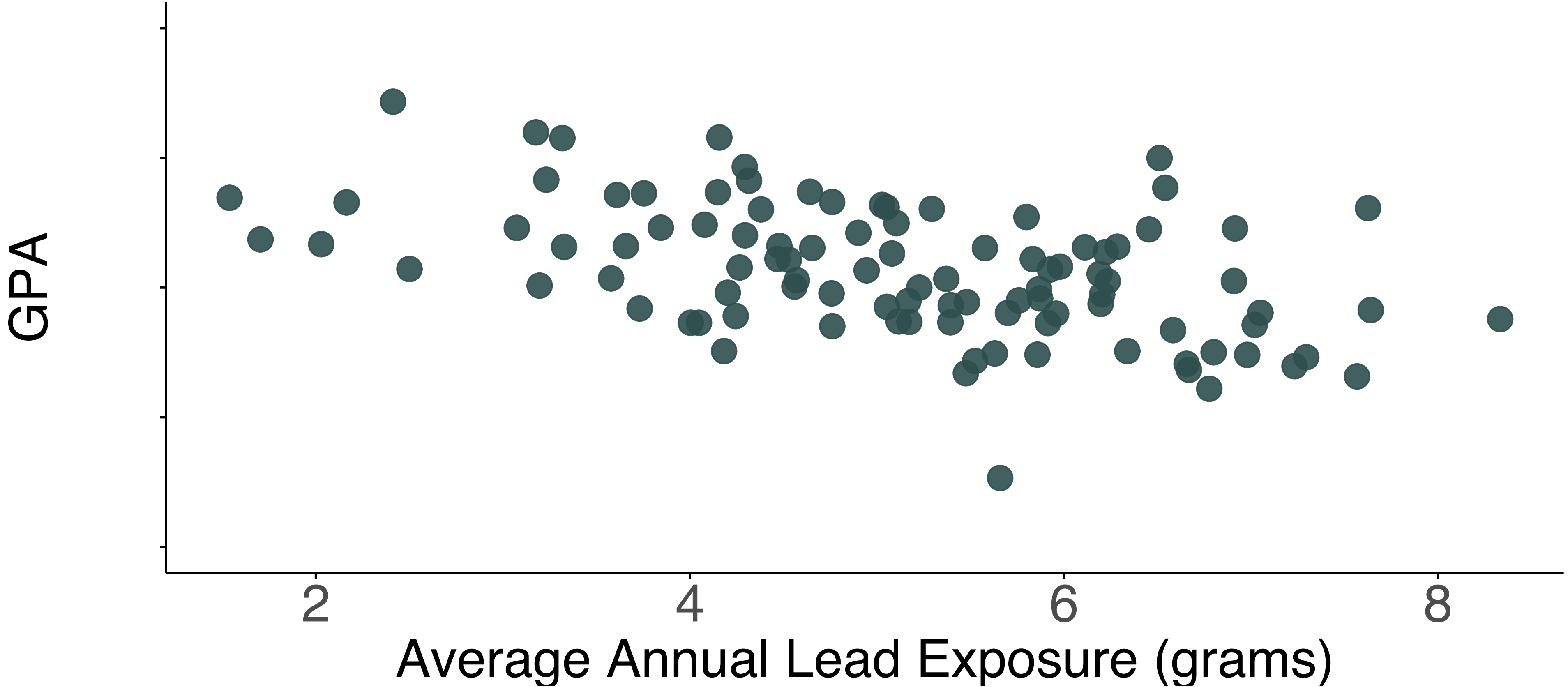
Choosing the βs in this fashion gives us the best-fit line through the data

How?

Simple example

Suppose we were only looking at GPA and pollution (lead/Pb):

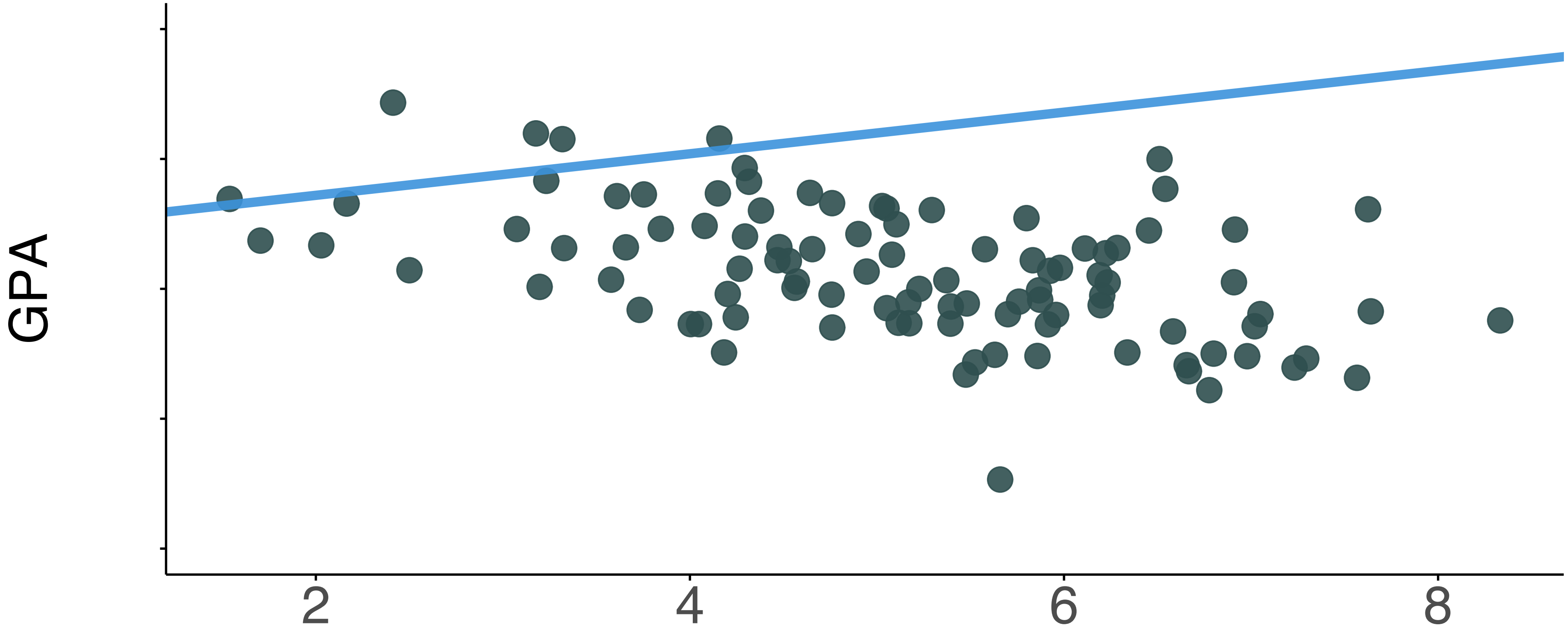
$$\mathrm{GPA}_i = \beta_0 + \beta_1 P_i + \varepsilon_i$$



Simple example

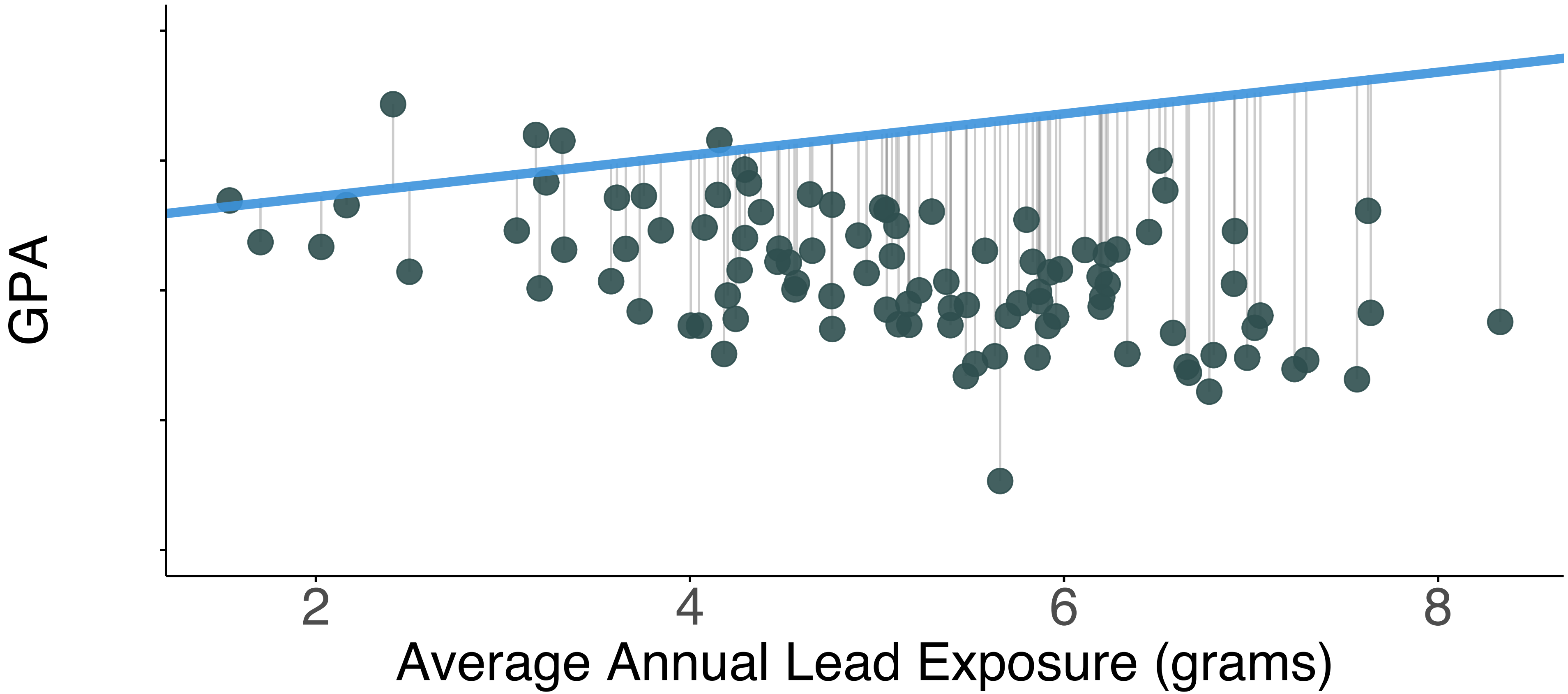
For any line $\left(\hat{GPA}_i = \hat{\beta}_0 + \hat{\beta}_1 P_i\right)$

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## ℹ Please use `linewidth` instead.
```



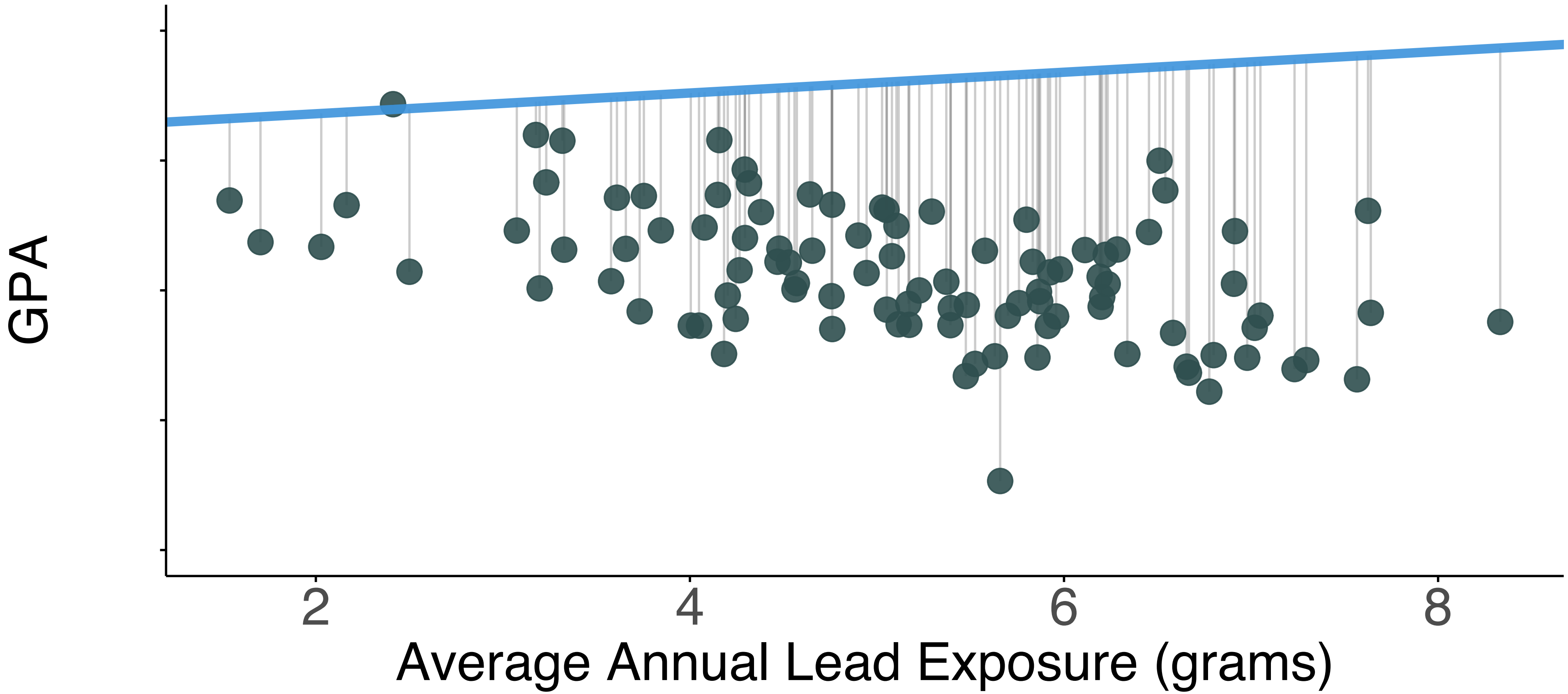
Simple example

For any line $\left(\hat{GPA}_i = \hat{\beta}_0 + \hat{\beta}_1 P_i\right)$, we calculate errors: $e_i = GPA_i - \hat{GPA}_i$



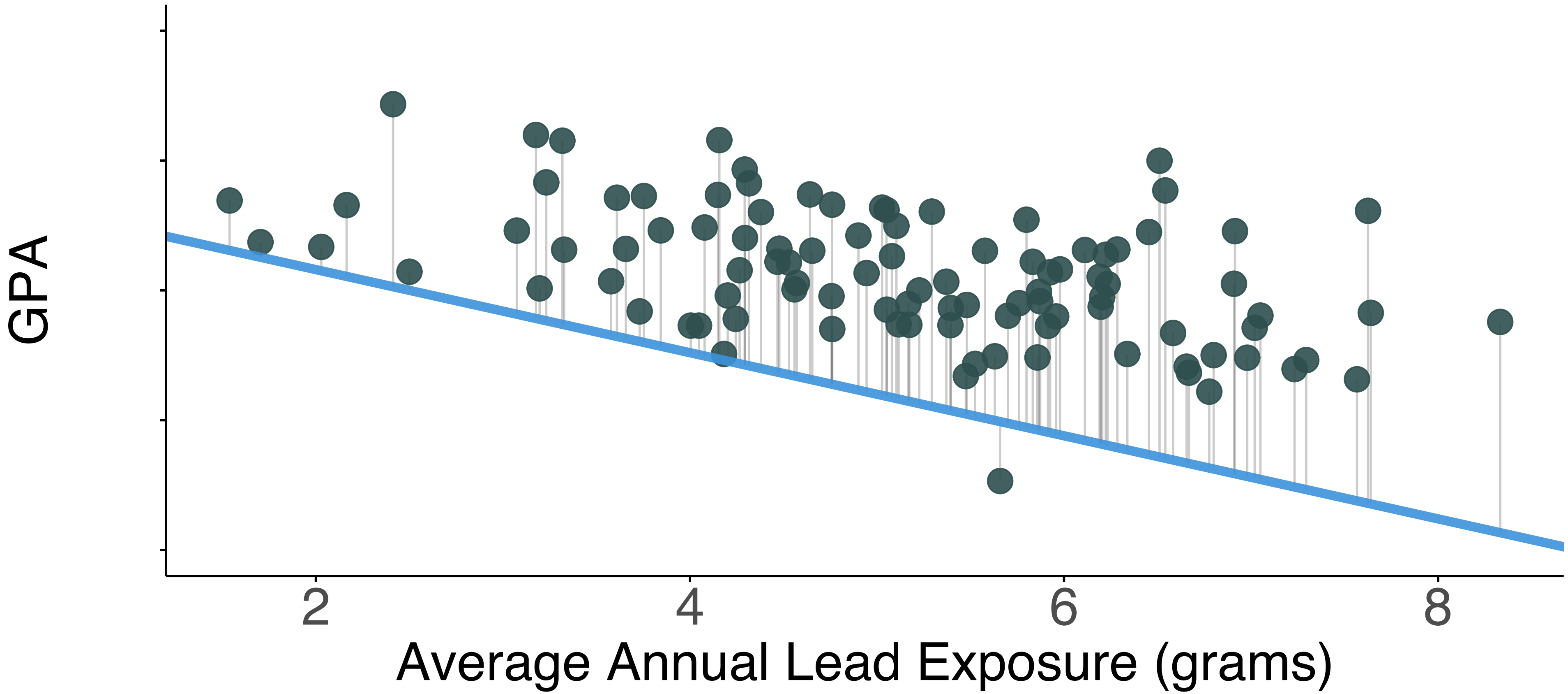
Simple example

For any line $\left(G\hat{P}A_i = \hat{\beta}_0 + \hat{\beta}_1 P_i\right)$, we calculate errors: $e_i = GPA_i - G\hat{P}A_i$



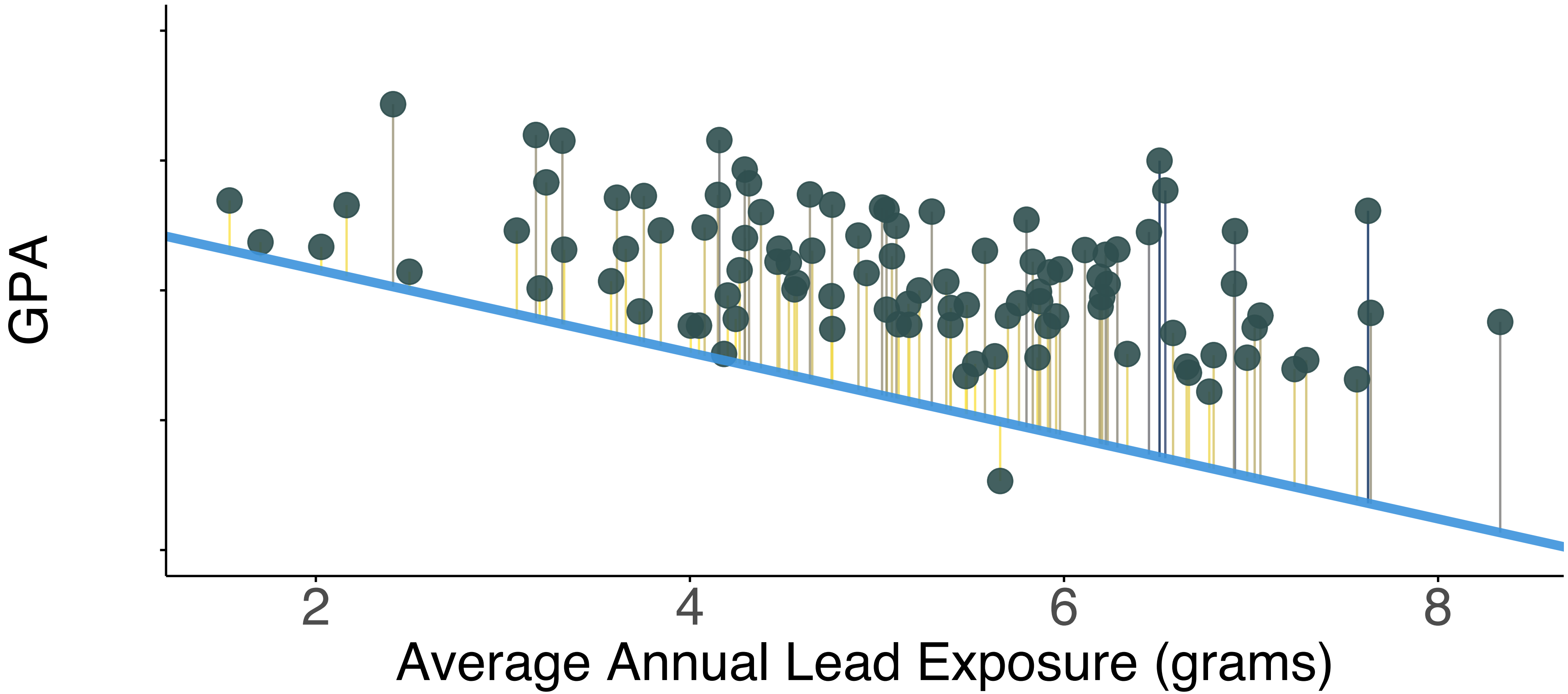
Simple example

For any line $\left(\hat{GPA}_i = \hat{\beta}_0 + \hat{\beta}_1 P_i\right)$, we calculate errors: $e_i = GPA_i - \hat{GPA}_i$



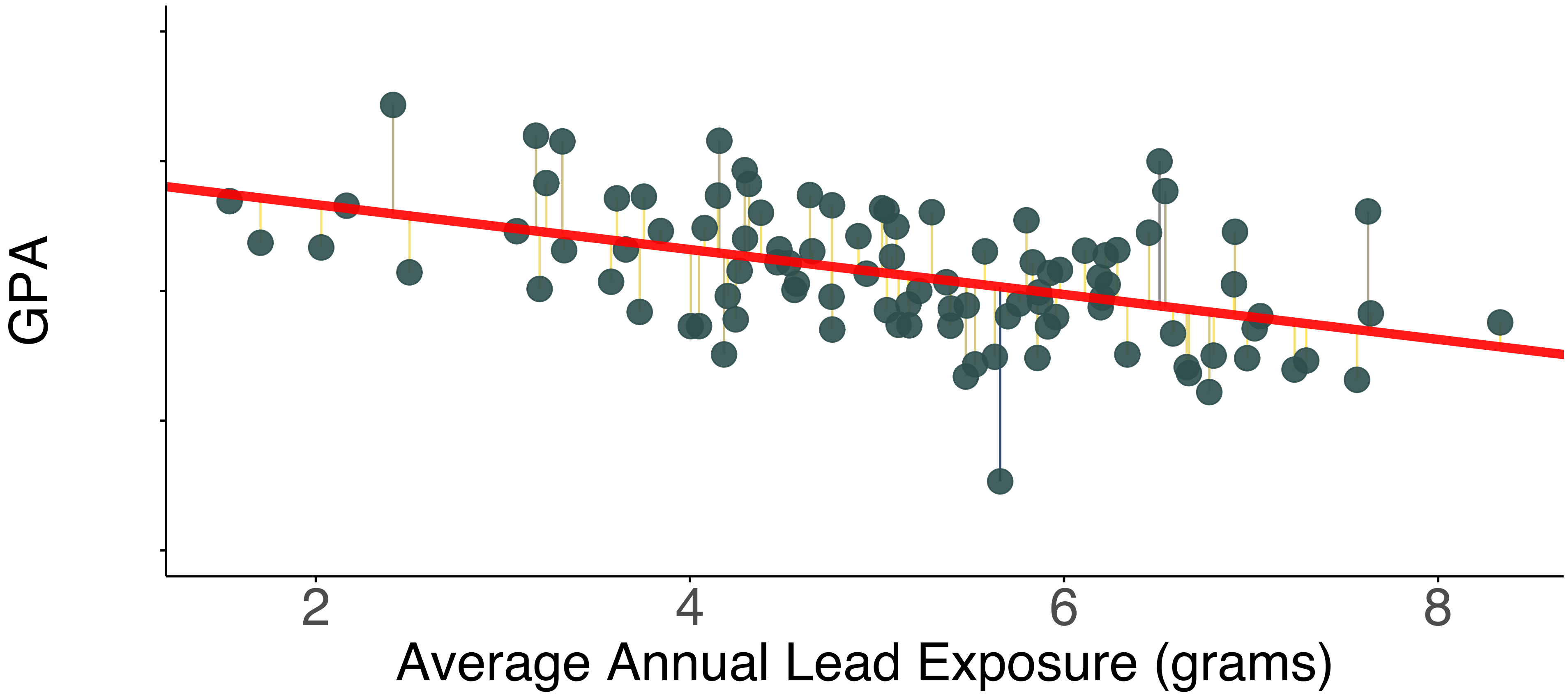
Simple example

SSE squares the errors $\left(\sum e_i^2\right)$: bigger errors get bigger penalties



Simple example

The OLS estimate is the combination of $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize SSE



OLS error term

So OLS is just the best-fit line through your data

OLS error term

So OLS is just the best-fit line through your data

OLS error term

So OLS is just the best-fit line through your data

Why?

OLS error term

So OLS is just the best-fit line through your data

Why?

Our model isn't perfect, the people in our dataset (i.e. our sample) may not perfectly match up to the entire population of people

OLS error term

There's **a lot** of other stuff that determines GPAs!

OLS error term

There's **a lot** of other stuff that determines GPAs!

We jam all that stuff into error term ε_i:

$$\text{GPA}_i = \beta_0 + \beta_1 I_i + \beta_2 P_i + \beta_3 \text{SAT}_i + \beta_4 H_i + \varepsilon_i$$

OLS error term

There's **a lot** of other stuff that determines GPAs!

We jam all that stuff into error term ε_i:

$$\text{GPA}_i = \beta_0 + \beta_1 I_i + \beta_2 P_i + \beta_3 \text{SAT}_i + \beta_4 H_i + \varepsilon_i$$

So ε_i contains all the determinants of GPA that we aren't explicitly addressing in our model like:

- Home environment
- Time studying

OLS error term

There's **a lot** of other stuff that determines GPAs!

We jam all that stuff into error term ε_i:

$$\text{GPA}_i = \beta_0 + \beta_1 I_i + \beta_2 P_i + \beta_3 \text{SAT}_i + \beta_4 H_i + \varepsilon_i$$

So ε_i contains all the determinants of GPA that we aren't explicitly addressing in our model like:

- Home environment
- Time studying

It is just a "catch-all", we don't actually know or see ε_i

OLS properties

OLS has one **very** nice property relevant for this class:

OLS properties

OLS has one **very** nice property relevant for this class:

Unbiasedness: `$E[\hat{\beta}] = \beta$`

OLS properties

Unbiasedness: $E[\hat{\beta}] = \beta$

On average, our estimate $\hat{\beta}$ exactly equals the **true** β

OLS properties

Unbiasedness: $E[\hat{\beta}] = \beta$

On average, our estimate $\hat{\beta}$ exactly equals the **true** β

The key is **on average:** we are estimating our model using only some sample of the data

OLS properties

Unbiasedness: $E[\hat{\beta}] = \beta$

On average, our estimate $\hat{\beta}$ exactly equals the **true** β

The key is **on average:** we are estimating our model using only some sample of the data

The estimated β won't exactly be right for the entire population, but on average, we expect it to match

OLS properties

Unbiasedness: $E[\hat{\beta}] = \beta$

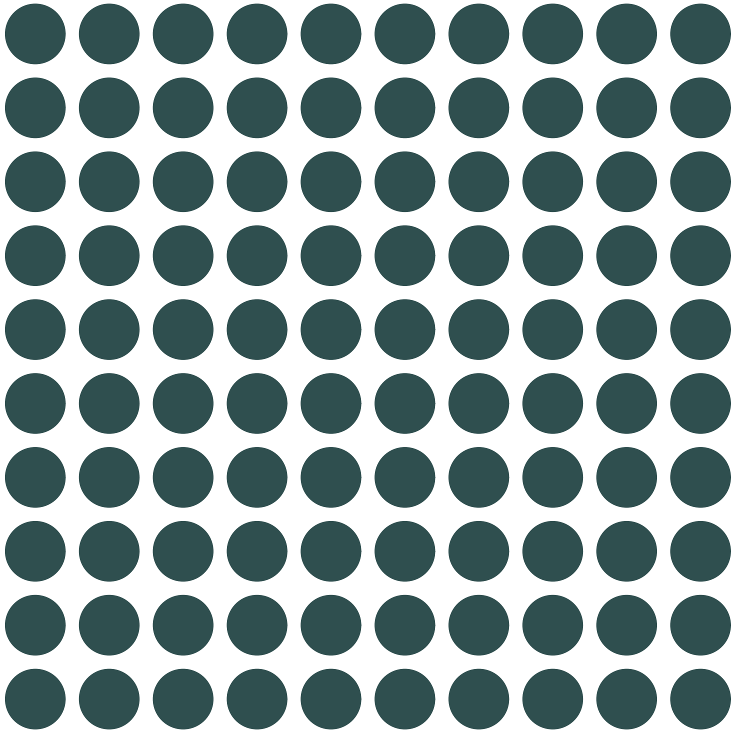
On average, our estimate $\hat{\beta}$ exactly equals the **true** β

The key is **on average:** we are estimating our model using only some sample of the data

The estimated β won't exactly be right for the entire population, but on average, we expect it to match

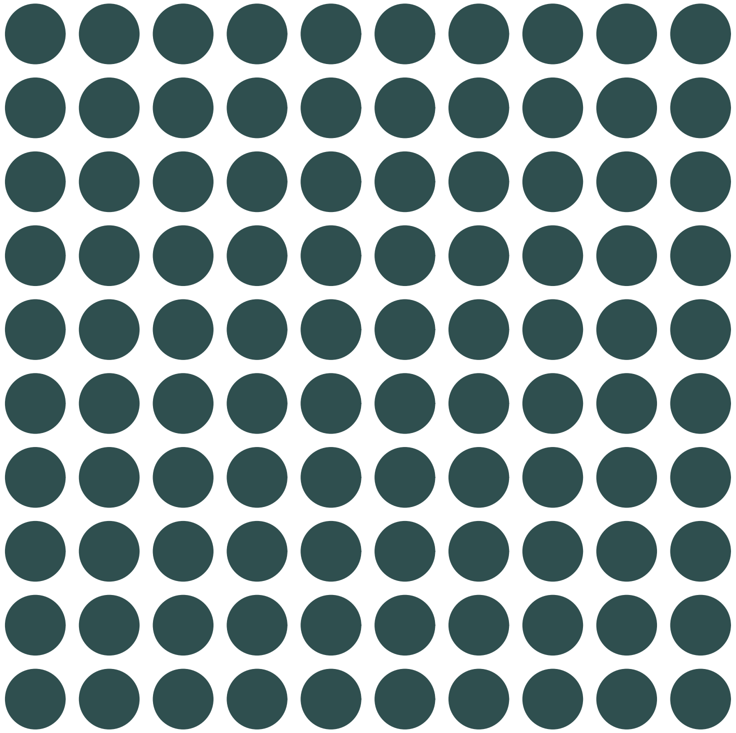
Let's see in an example where we only have a subsample of the full population of data

OLS properties

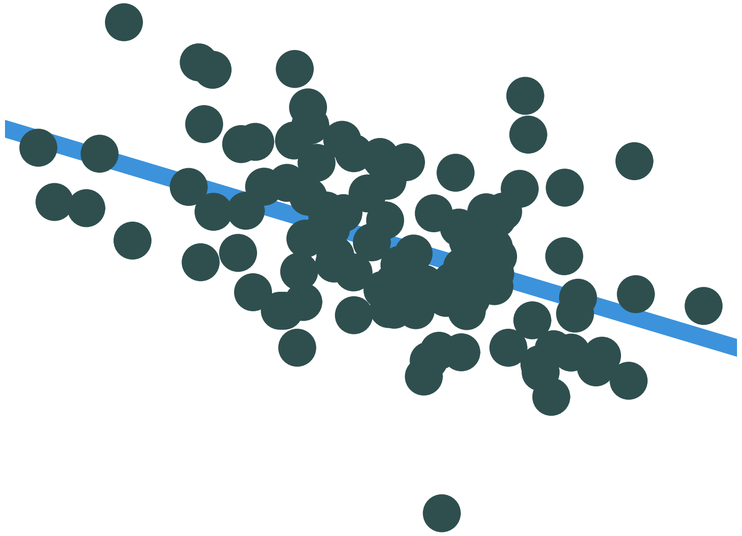


Population

OLS properties



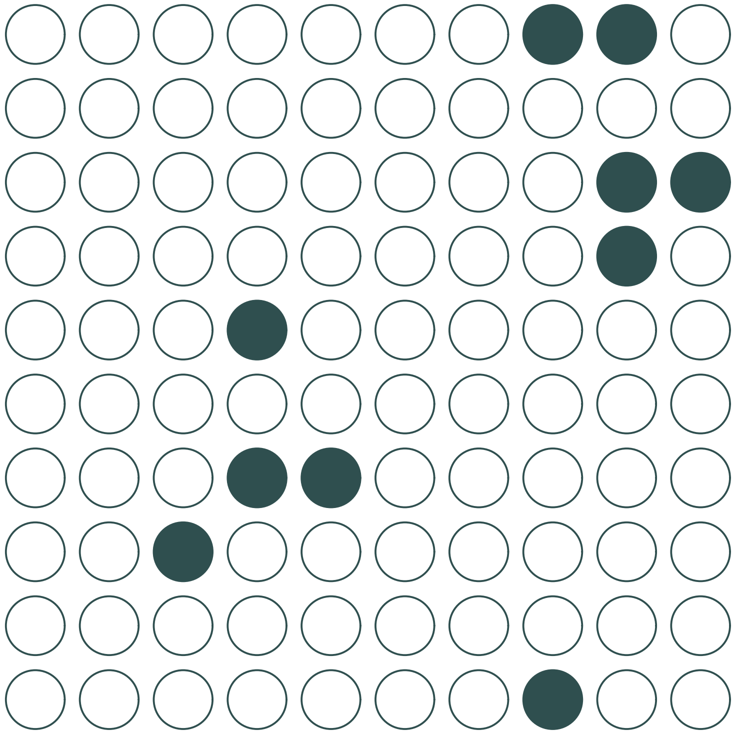
Population



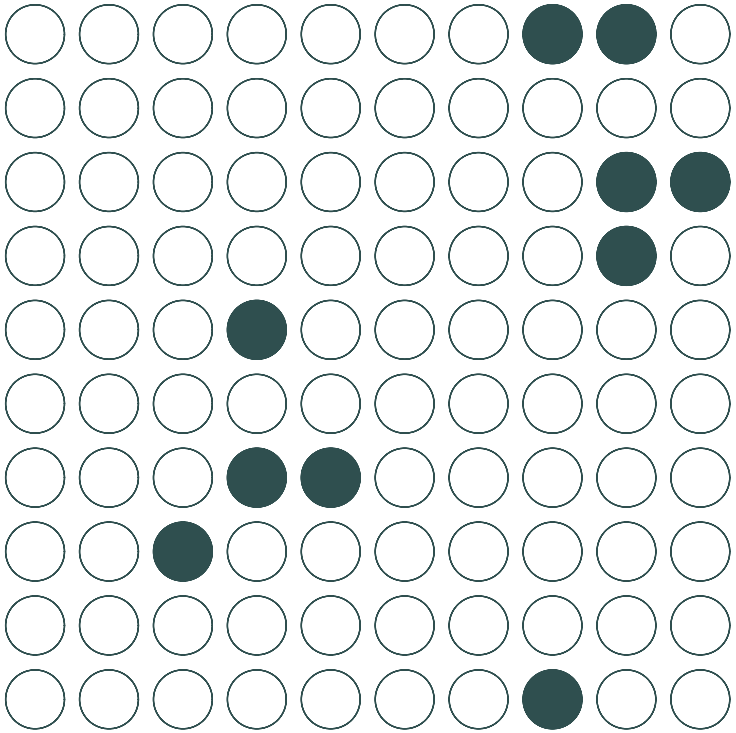
Population relationship

$$y_i = 2.53 + -0.43 x_i + u_i$$

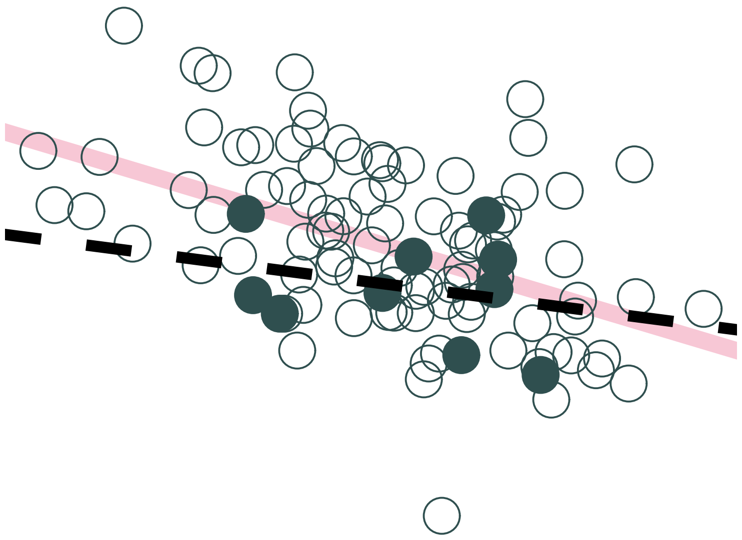
$$y_i = \beta_0 + \beta_1 x_i + u_i$$



Sample 1: 10 random individuals



Sample 1: 10 random individuals

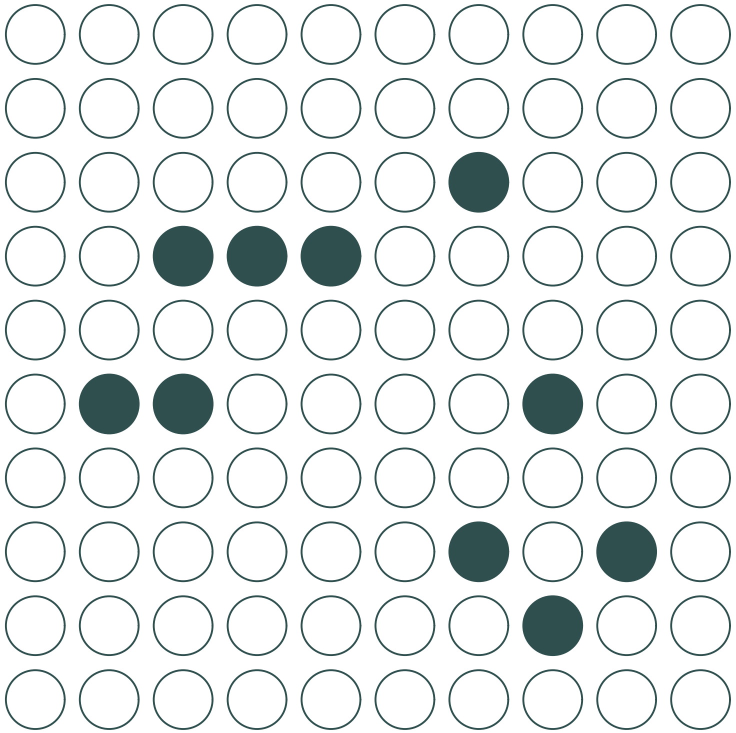


Population relationship

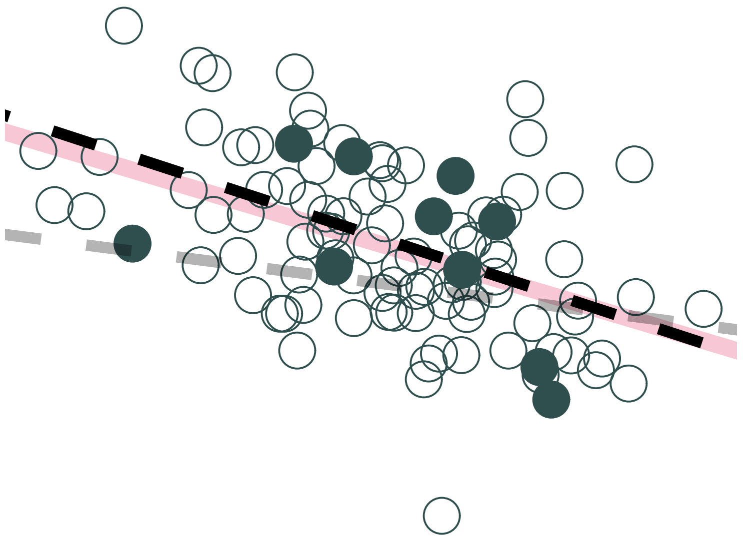
$$y_i = 2.53 + -0.43 x_i + u_i$$

Sample relationship

$$\hat{y}_i = 0.72 + -0.19 x_i$$



Sample 2: 10 random individuals

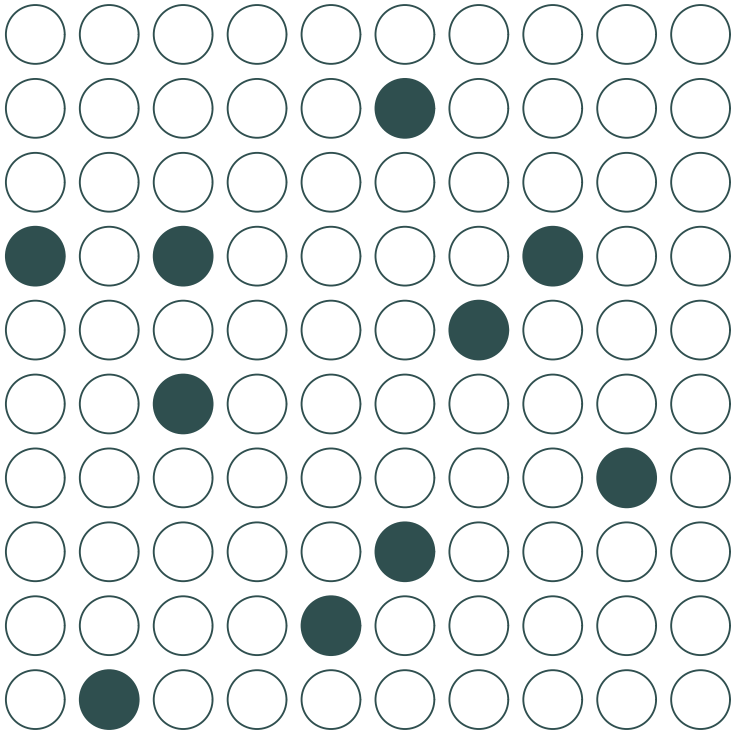


Population relationship

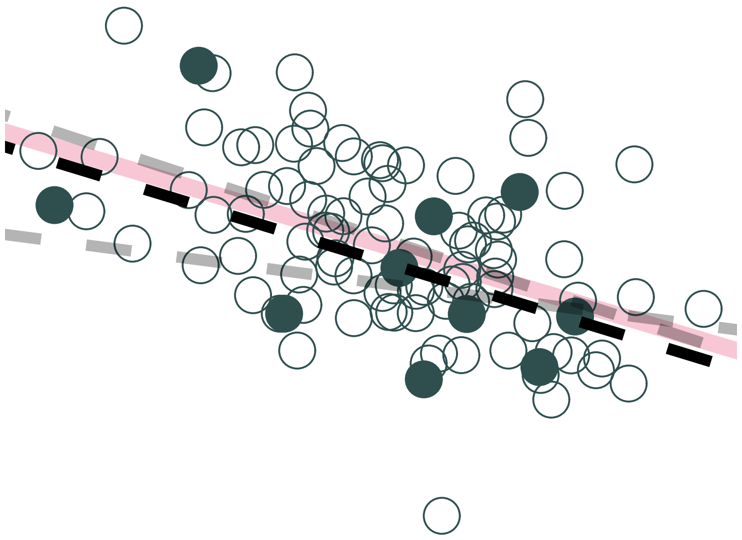
$$y_i = 2.53 + -0.43x_i + u_i$$

Sample relationship

$$\hat{y}_i = 2.82 + -0.47x_i$$



Sample 3: 10 random individuals



Population relationship

$$y_i = 2.53 + -0.43x_i + u_i$$

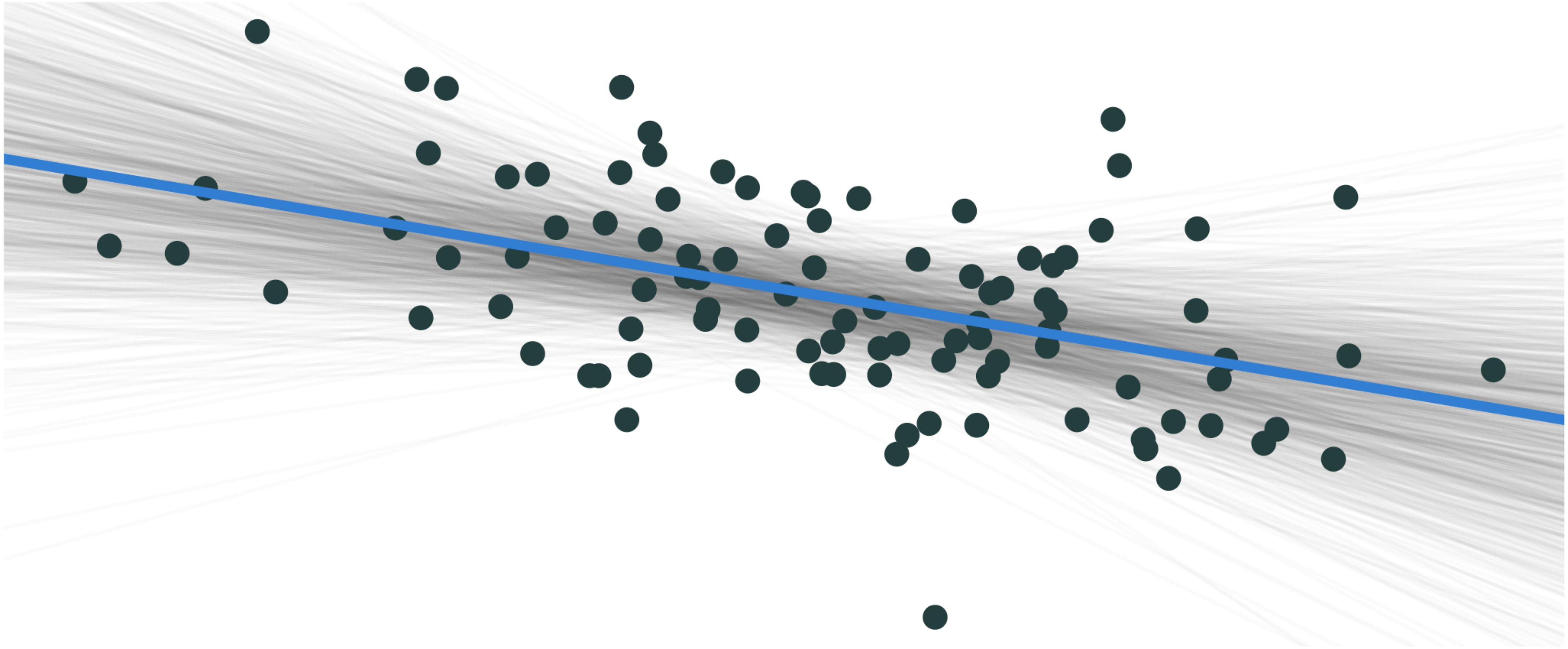
Sample relationship

$$\hat{y}_i = 2.32 + -0.44x_i$$

Let's repeat this **1,000 times**.

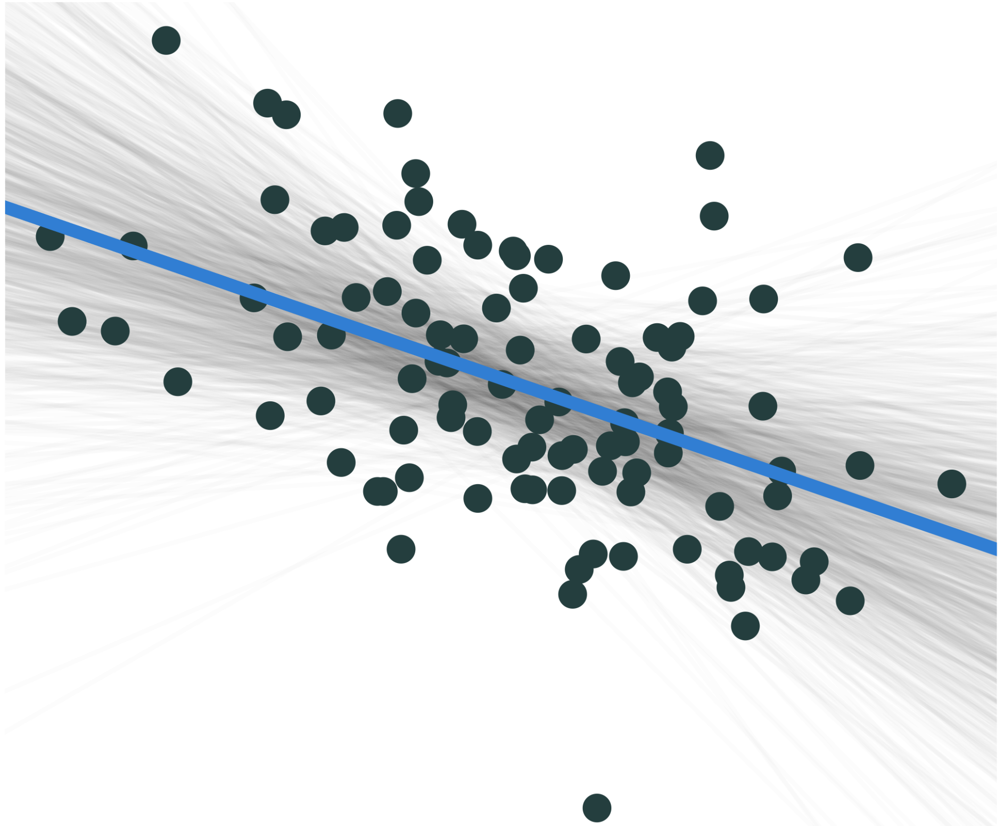
(This exercise is called a (Monte Carlo) simulation.)

Population *vs.* sample



Population *vs.* sample

Question: Why do we care about *population vs. sample*?



On **average**, our regression lines match the population line very nicely

However, **individual lines** (samples) can really miss the mark

Population *vs.* sample

Answer: Uncertainty/randomness matters!

Population *vs.* sample

Answer: Uncertainty/randomness matters!

$\hat{\beta}$ itself is will depend on the sample of data we have

Population *vs.* sample

Answer: Uncertainty/randomness matters!

$\hat{\beta}$ itself is will depend on the sample of data we have

When we take a sample and run a regression, we don't know if it's a 'good' sample ($\hat{\beta}$ is close to β) or a 'bad sample' (our sample differs greatly from the population)

Unbiasedness

For OLS to be unbiased and give us, on average, the causal effect of some X on some Y we need a few assumptions to hold

Unbiasedness

For OLS to be unbiased and give us, on average, the causal effect of some X on some Y we need a few assumptions to hold

Whether or not these assumptions are true is why you often hear *correlation is not causation*

Unbiasedness

For OLS to be unbiased and give us, on average, the causal effect of some X on some Y we need a few assumptions to hold

Whether or not these assumptions are true is why you often hear *correlation is not causation*

If we want some $\hat{\beta}_1$ on a variable x to be unbiased we x to be **uncorrelated** with the error term:

$$E[x\varepsilon] = 0 \quad \leftrightarrow \quad \text{correlation}(x, \varepsilon) = 0$$

Unbiasedness

The variable you are interested in **cannot** be correlated with the error term

Unbiasedness

The variable you are interested in **cannot** be correlated with the error term

What does this mean in words?

Unbiasedness

The variable you are interested in **cannot** be correlated with the error term

What does this mean in words?

The error term contains all variables that determine y, but we *omitted* from our model

Unbiasedness

The variable you are interested in **cannot** be correlated with the error term

What does this mean in words?

The error term contains all variables that determine y, but we *omitted* from our model

We are assuming that our variable of interest, x, is not correlated with any of these omitted variable

Unbiasedness

The variable you are interested in **cannot** be correlated with the error term

What does this mean in words?

The error term contains all variables that determine y, but we *omitted* from our model

We are assuming that our variable of interest, x, is not correlated with any of these omitted variable

If x is correlated with any of them, then we will have something called **omitted variable bias**

Omitted variable bias

Here's an intuitive example

Omitted variable bias

Here's an intuitive example

Suppose we wanted to understand the effect of lead exposure P on GPAs

Omitted variable bias

Here's an intuitive example

Suppose we wanted to understand the effect of lead exposure P on GPAs

lead harm's children's brain development, especially before age 6

Omitted variable bias

Here's an intuitive example

Suppose we wanted to understand the effect of lead exposure P on GPAs

lead harm's children's brain development, especially before age 6

We should expect early-life lead exposure to reduce future GPAs

Omitted variable bias

Our model might look like:

$$\mathrm{GPA}_i = \beta_0 + \beta_1 \mathrm{P}_i + \varepsilon_i$$

Omitted variable bias

Our model might look like:

$$\text{GPA}_i = \beta_0 + \beta_1 \text{P}_i + \varepsilon_i$$

We want to know β_1

Omitted variable bias

Our model might look like:

$$\mathrm{GPA}_i = \beta_0 + \beta_1 \mathrm{P}_i + \varepsilon_i$$

We want to know β_1

What would happen if we took a sample of *real world data* and used OLS to estimate $\hat{\beta}_1$?

Omitted variable bias

We would have omitted variable bias

Omitted variable bias

We would have omitted variable bias

Why? What are some examples?

Omitted variable bias

We would have omitted variable bias

Why? What are some examples?

Who is more likely to be exposed to lead?

Omitted variable bias

We would have omitted variable bias

Why? What are some examples?

Who is more likely to be exposed to lead?

Poorer families likely have more lead exposure, why?

Omitted variable bias

We would have omitted variable bias

Why? What are some examples?

Who is more likely to be exposed to lead?

Poorer families likely have more lead exposure, why?

Richer families can move away, pay to replace lead paint, lead pipes, etc

Omitted variable bias

We would have omitted variable bias

Why? What are some examples?

Who is more likely to be exposed to lead?

Poorer families likely have more lead exposure, why?

Richer families can move away, pay to replace lead paint, lead pipes, etc

This means lead exposure is correlated with lower income

Omitted variable bias

Why does this correlation cause us problems?

Omitted variable bias

Why does this correlation cause us problems?

Family income *also* matters for GPA, it is in ε_i, so our assumption that $\text{correlation}(x, \varepsilon) = 0$ is violated

Omitted variable bias

Why does this correlation cause us problems?

Family income *also* matters for GPA, it is in ε_i, so our assumption that $\text{correlation}(x, \varepsilon) = 0$ is violated

Children from richer families tend to have higher GPAs

Omitted variable bias

Why does this correlation cause us problems?

Family income *also* matters for GPA, it is in ε_i, so our assumption that $\text{correlation}(x, \varepsilon) = 0$ is violated

Children from richer families tend to have higher GPAs

Why?

Omitted variable bias

Why does this correlation cause us problems?

Family income *also* matters for GPA, it is in ε_i, so our assumption that $\text{correlation}(x, \varepsilon) = 0$ is violated

Children from richer families tend to have higher GPAs

Why?

Access to tutoring, better schools, parental pressure, etc, etc

Omitted variable bias

If we just look at the effect of lead exposure on GPAs without addressing its correlation with income, lead exposure will look worse than it actually is

Omitted variable bias

If we just look at the effect of lead exposure on GPAs without addressing its correlation with income, lead exposure will look worse than it actually is

This is because our data on lead exposure is also proxying for income (since $\text{correlation}(x, \varepsilon) = 0$)

Omitted variable bias

If we just look at the effect of lead exposure on GPAs without addressing its correlation with income, lead exposure will look worse than it actually is

This is because our data on lead exposure is also proxying for income (since $\text{correlation}(x, \varepsilon) = 0$)

So $\hat{\beta}_1$ will pick up the effect of both!

Omitted variable bias

If we just look at the effect of lead exposure on GPAs without addressing its correlation with income, lead exposure will look worse than it actually is

This is because our data on lead exposure is also proxying for income (since $\text{correlation}(x, \varepsilon) = 0$)

So $\hat{\beta}_1$ will pick up the effect of both!

Our estimate $\hat{\beta}_1$ is **biased** and overstates the negative effects of lead

Omitted variable bias

How do we fix this bias?

Omitted variable bias

How do we fix this bias?

Make income not omitted: control for it in our model

Omitted variable bias

How do we fix this bias?

Make income not omitted: control for it in our model

If we have data on family income I we can instead write our model as:

$$\text{GPA}_i = \beta_0 + \beta_1 \text{P}_i + \beta_2 \text{I}_i + \varepsilon_i$$

I is no longer omitted

Omitted variable bias

How do we fix this bias?

Make income not omitted: control for it in our model

If we have data on family income I we can instead write our model as:

$$\text{GPA}_i = \beta_0 + \beta_1 \text{P}_i + \beta_2 \text{I}_i + \varepsilon_i$$

I is no longer omitted

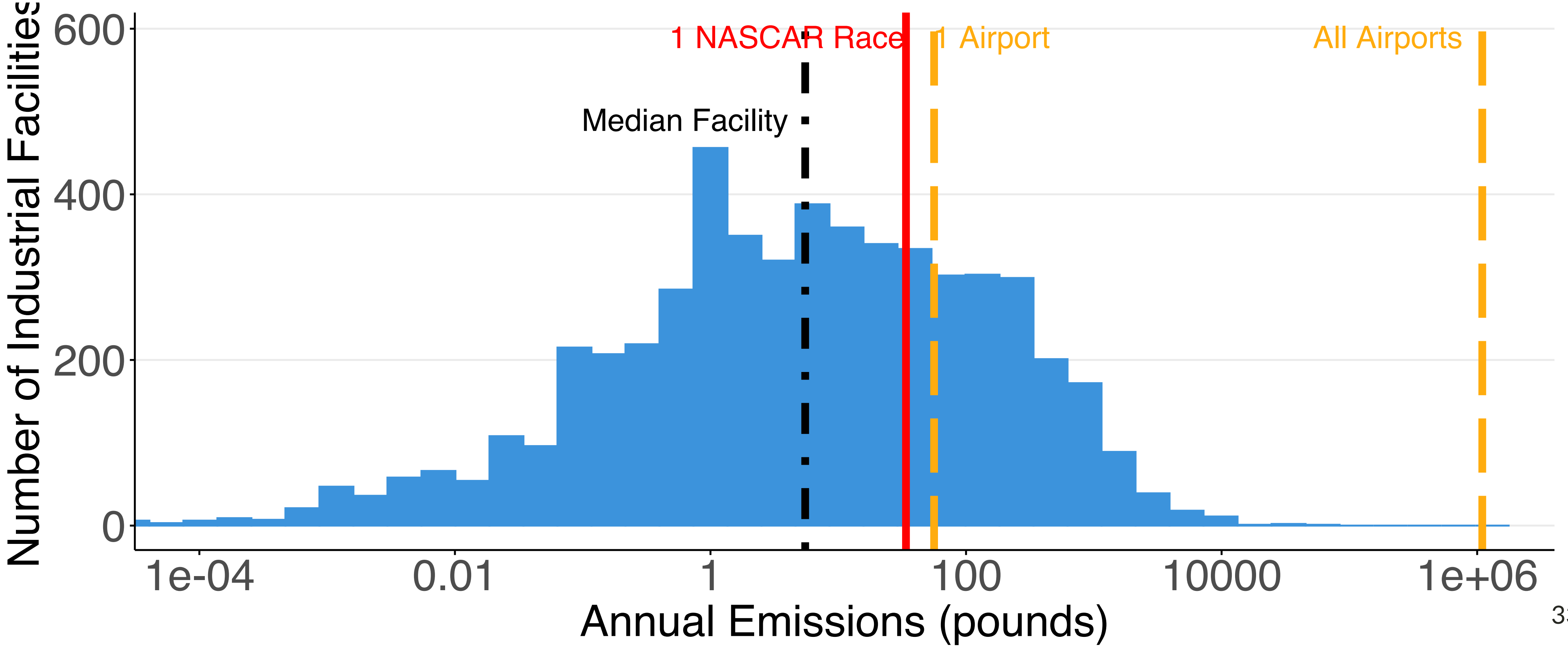
Independent variables in our model that we include to address bias are called **controls**

Hands-on pollution education example

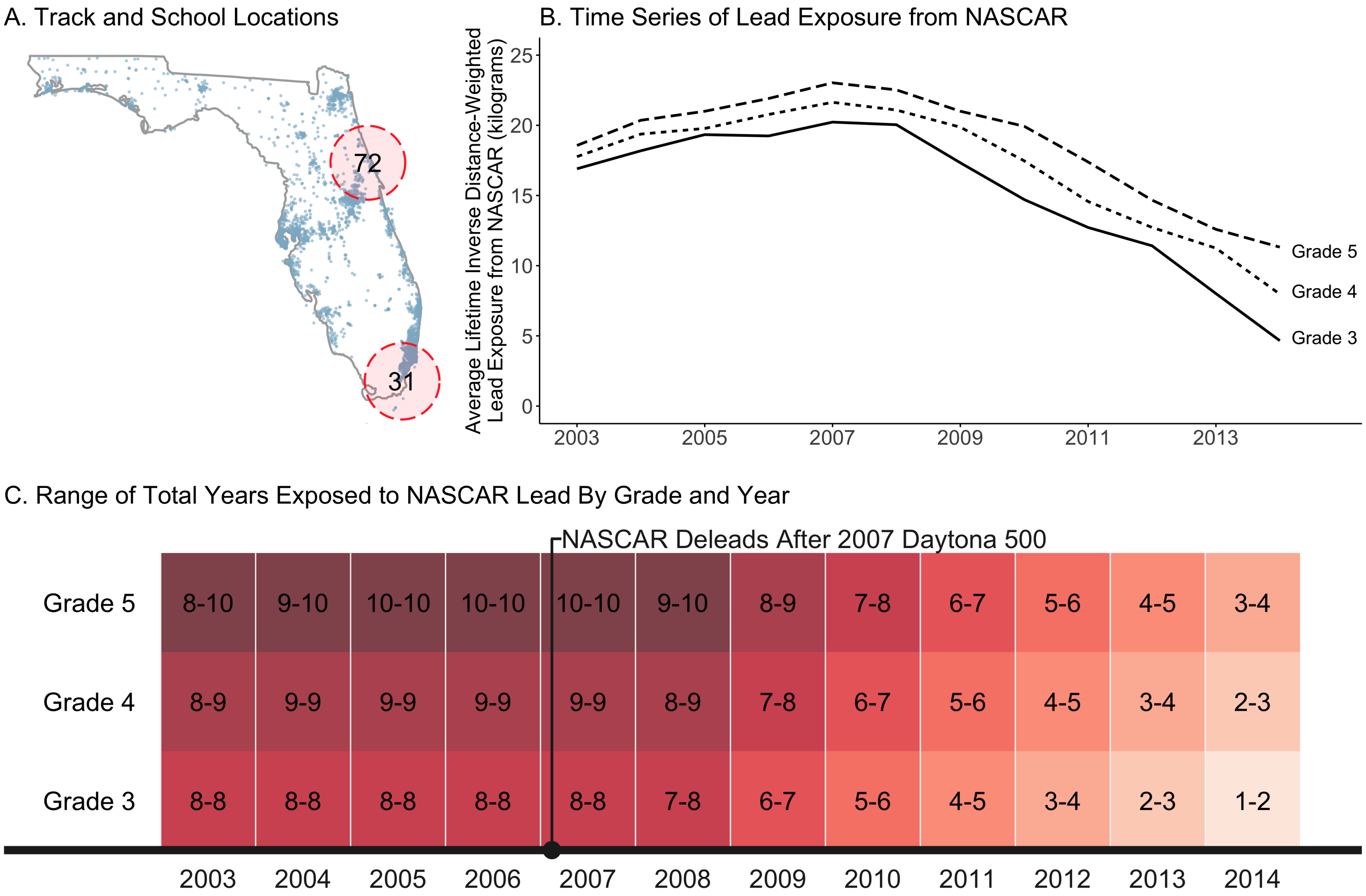
Real pollution education example

Real pollution education example

In **3 hours**, one NASCAR race emits more lead than a majority of industrial facilities do in an **entire year**



We will look at Florida



All the data are public, you can look at scores yourself!

FLORIDA DEPARTMENT OF EDUCATION
fldoe.org

About Us | Newsroom | How Do I? | Contact Us | Public Records

AAA What are you looking for?

FEATURED TOPICS | ACADEMICS | SCHOOLS | TEACHING | ACCOUNTABILITY | POLICY | FINANCE

Home | Accountability | Assessments | K-12 Student Assessment | Assessments & Publications Archive | FCAT Historical | Scores & Reports

SCORES & REPORTS

- 2003 Score Reports
- 2004 Score Reports
- 2005 Score Reports
- 2006 Score Reports
- 2007 Score Reports
- 2008 Score Reports
- 2009 Score Reports
- 2010 Score Reports

Scores & Reports

For results by reporting year, use the link for the appropriate year in the left-hand navigation panel. For results by subject area (e.g., reading, science), use one of the links below. FCAT 2.0 results are also available, and the interactive reporting resources provide access to databases that allow users to generate reports for the state, districts or schools for certain educational areas.

- FCAT Reading & Mathematics SSS Scores (1998-2011)
- FCAT Science SSS Scores (2003-2011)
- FCAT Writing Scores (1997-2012)
- FCAT Norm-Referenced Test Scores (1995-2008)
- Longitudinal Data: FWAP / FCAT / HSCT 1995-2000

Additional Resources for Understanding Results

- Understanding FCAT Reports
- FCAT Achievement Level Definitions/Tables (PDF)
- Developmental Score Scale Memo (04/14/02) (PDF)
- Guidance on Content Area Scores
- Content Focus Reports

Let's look at the data

```
nascar_df
```

```
## # A tibble: 68,858 × 12
##    school_id school_name   grade  year zscore nascar_lead nascar_lead_weighted years_leaded indust…¹
##        <dbl> <chr>         <dbl> <dbl>  <dbl>       <dbl>                <dbl>        <dbl>    <dbl>
##  1        56 HAMILTON ELEM     3  2003 -0.186        72.2                 2.53            8  822328.
##  2        56 HAMILTON ELEM     4  2003  0.101        80.4                 2.81            8  822639.
##  3        56 HAMILTON ELEM     5  2003 -0.206        88.0                 3.08            8  822909.
##  4        56 HAMILTON ELEM     3  2004 -0.686        74.0                 2.59            8  967077.
##  5        56 HAMILTON ELEM     4  2004 -0.633        82.4                 2.88            8  967352.
##  6        56 HAMILTON ELEM     5  2004  0.352        90.5                 3.17            8  967663.
##  7        56 HAMILTON ELEM     3  2005 -1.14         77.0                 2.69            8 1061570.
##  8        56 HAMILTON ELEM     4  2005 -0.649        84.7                 2.97            8 1062071.
##  9        56 HAMILTON ELEM     5  2005 -0.336        92.0                 3.26            8 1062346.
## 10        56 HAMILTON ELEM     3  2006 -0.333        79.9                 2.80            8 1164072.
## # … with 68,848 more rows, and abbreviated variable names ¹industrial_lead, ²median_income, ³unemp_r
```

My sister is in these observations!

```
nascar_df |> # only keep Saturn Elementary School
  filter(school_name == "SATURN ELEM")
```

```
## # A tibble: 21 × 12
##    school_id school_name grade  year  zscore nascar_lead nascar_lead_weighted years_leaded industr…¹
##        <dbl> <chr>       <dbl> <dbl>   <dbl>       <dbl>                <dbl>        <dbl>     <dbl>
##  1      2067 SATURN ELEM     3  2003  0.105            0                    0            0   823844.
##  2      2067 SATURN ELEM     4  2003 -0.0633           0                    0            0   824155.
##  3      2067 SATURN ELEM     5  2003  0.163            0                    0            0   824425.
##  4      2067 SATURN ELEM     3  2004  0.655            0                    0            0   967646.
##  5      2067 SATURN ELEM     4  2004  0.586            0                    0            0   967921.
##  6      2067 SATURN ELEM     5  2004  0.679            0                    0            0   968232.
##  7      2067 SATURN ELEM     3  2005  1.03             0                    0            0  1059953.
##  8      2067 SATURN ELEM     4  2005  0.131            0                    0            0  1060454.
##  9      2067 SATURN ELEM     5  2005  0.696            0                    0            0  1060729.
## 10      2067 SATURN ELEM     3  2006  0.599            0                    0            0  1161336.
## # … with 11 more rows, and abbreviated variable names ¹industrial_lead, ²median_income, ³unemp_rate,
```

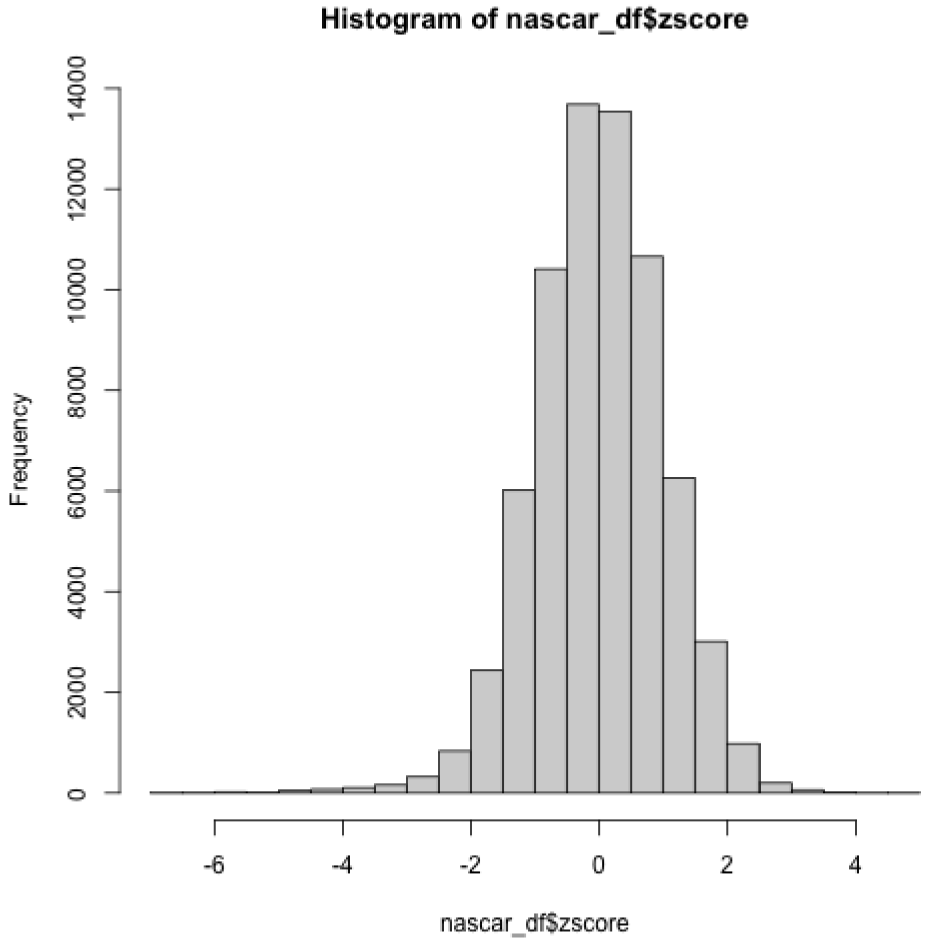
Let's look at the data

```
##    school_id        zscore             nascar_lead    industrial_lead   median_income    num_students
##  Min.   :   3   Min.   :-6.765987   Min.   : 0.00   Min.   :      0   Min.   :25201   Min.   : 10.0
##  1st Qu.: 961   1st Qu.:-0.630857   1st Qu.: 0.00   1st Qu.: 300489   1st Qu.:41184   1st Qu.: 72.0
##  Median :1811   Median : 0.012807   Median : 0.00   Median : 562856   Median :44635   Median :100.0
##  Mean   :1832   Mean   : 0.000358   Mean   :12.88   Mean   :1197073   Mean   :44712   Mean   :102.5
##  3rd Qu.:2702   3rd Qu.: 0.661761   3rd Qu.:16.38   3rd Qu.:2040709   3rd Qu.:48772   3rd Qu.:130.0
##  Max.   :4110   Max.   : 4.884255   Max.   :92.02   Max.   :6454837   Max.   :67238   Max.   :447.0
```

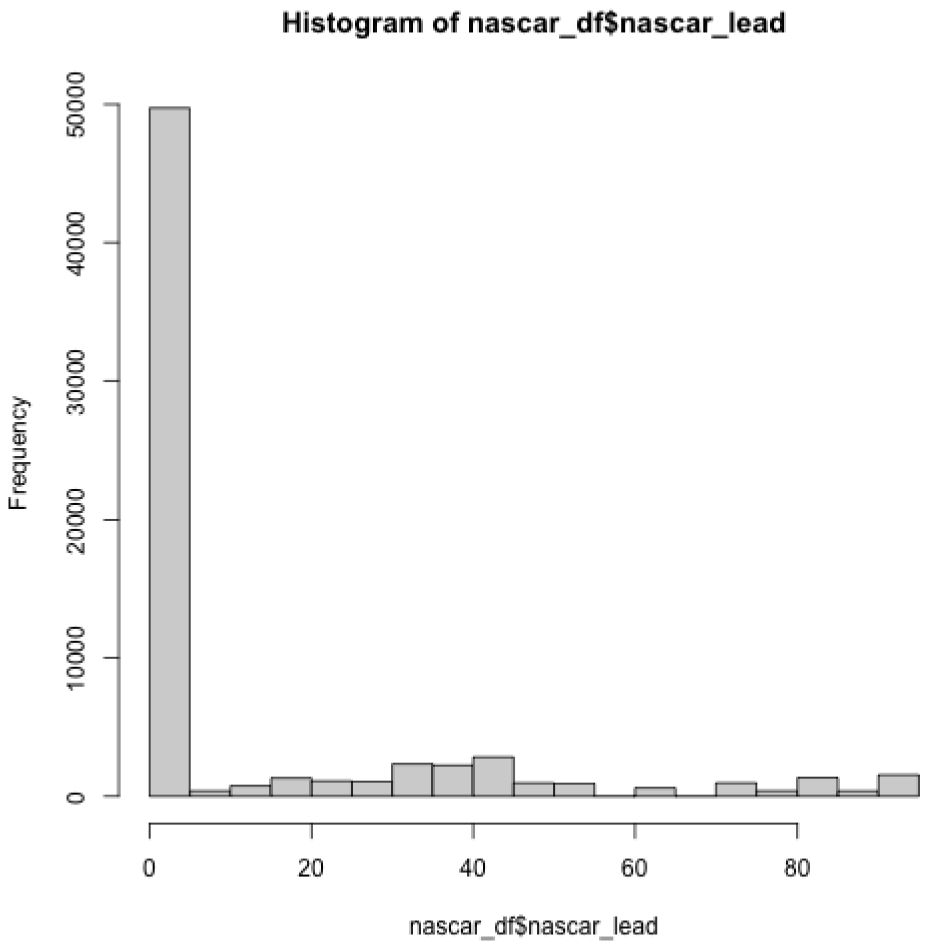
The variables

- **zscore**: the school's score for the average student in terms of standard deviations above or below the state-wide average
- **nascar lead**: lifetime exposure to lead emissions from NASCAR tracks within 50 miles
- **industrial lead**: lead emissions from industrial sources (e.g. factories) within 50 miles
- **median income**: the school district's median incoe
- **num students**: the number of students at the school
- **school id, school name, grade, and year**: self-explanatory

What does the distribution of scores look like?



What about exposure to NASCAR lead

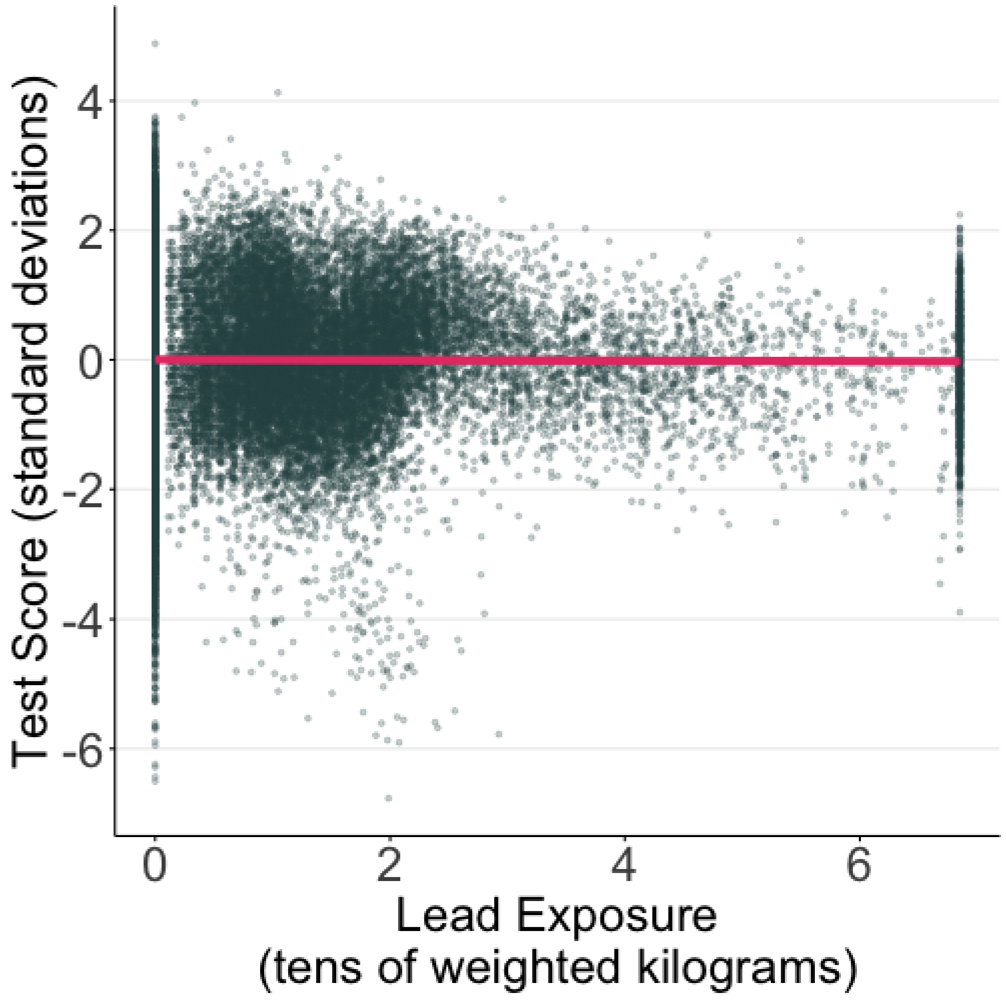


Most schools have zero exposure

Some have a lot

Units are 10s of kilograms

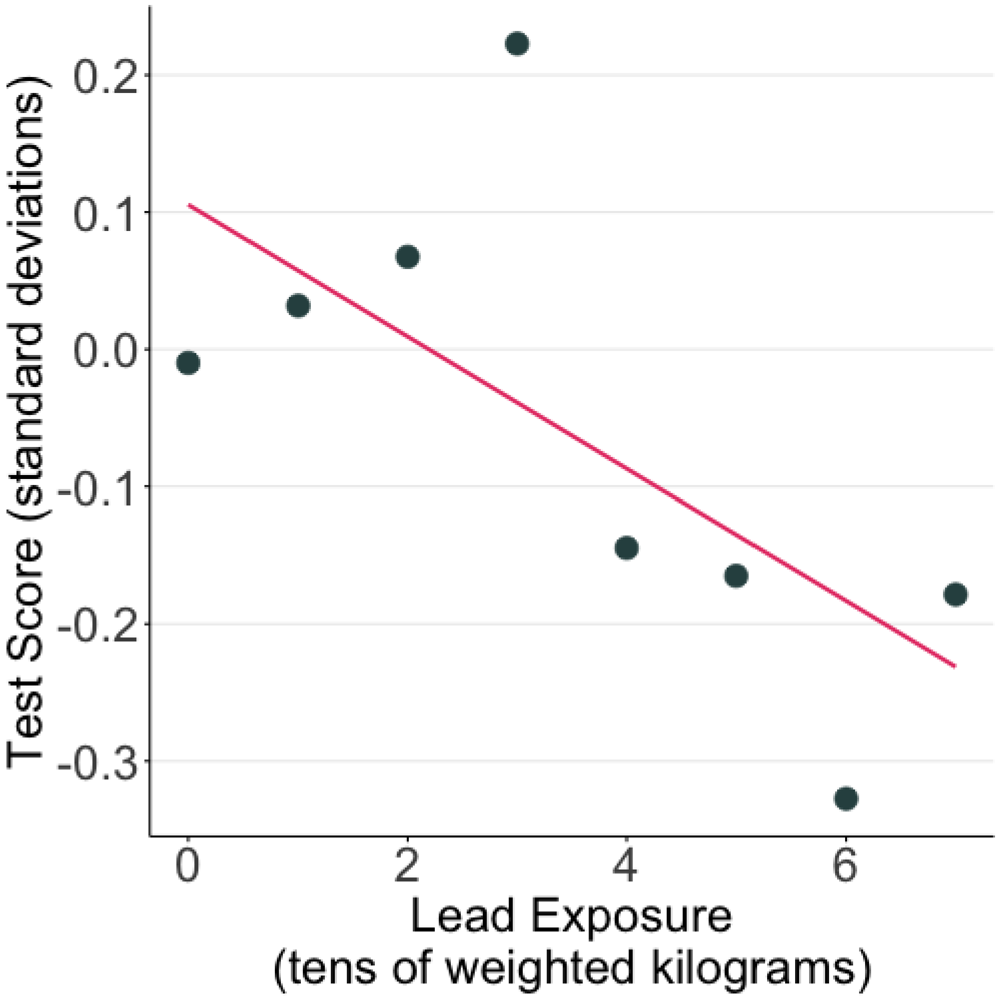
What is the association between lead and scores?



Let's look at the pure correlation between test scores and lead

There's a lot of data so it's kind of hard to see but it appears there's a **negative** association: lead is bad for test scores

What is the association between lead and scores?



Lets **bin** the data to see the pattern more clearly

All I'm doing is:

- Rounding lead to the nearest integer
- Taking the average of test scores for that bin
- Plot the average scores versus rounded lead

What is the association between lead and scores?

We can get a better sense by running a regression:

$$zscore_{sgy} = \beta_0 + \beta_1 nascar\_lead\_weighted_{sgy}$$

(s is school, g is grade, y is year)

What is the association between lead and scores?

```
## Estimation Results
##   parameter                               estimate
## 1 beta_0 (Intercept)                      0.002
## 2 beta_1 nascar_lead_weighted            -0.004
```

What does this mean?

An additional 10 kg of lead exposure is associated with a school having an average test score 0.004 standard deviations lower

Do we believe this number?

What's a potential issue with just looking at the raw association?

Do we believe this number?

What's a potential issue with just looking at the raw association?

Schools near NASCAR tracks are probably a lot different than schools further away

Do we believe this number?

What's a potential issue with just looking at the raw association?

Schools near NASCAR tracks are probably a lot different than schools further away

We want to control for things that are potentially correlated with both test scores and being close to NASCAR

Do we believe this number?

What's a potential issue with just looking at the raw association?

Schools near NASCAR tracks are probably a lot different than schools further away

We want to control for things that are potentially correlated with both test scores and being close to NASCAR

Two broad important things: lead emissions from other sources, socioeconomic status

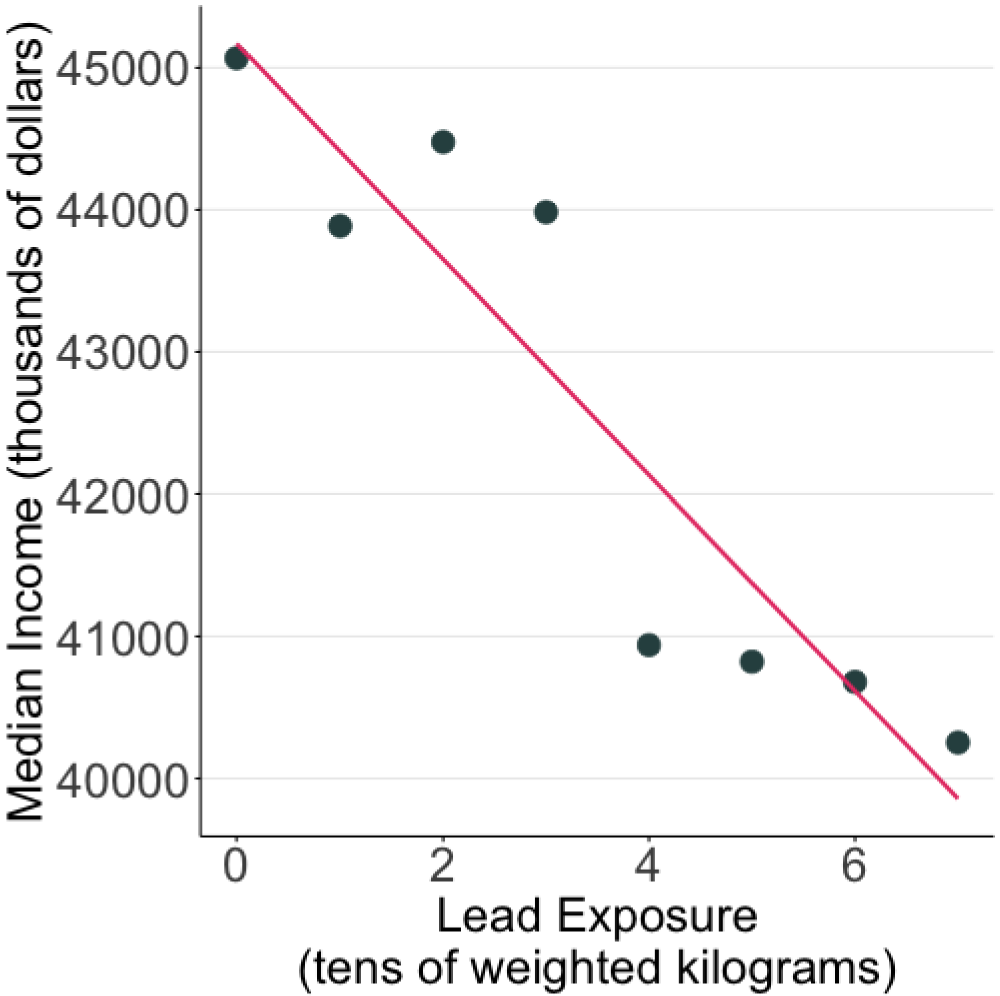
Do we believe this number?

$$zscore_{sgy} = \beta_0 + \beta_1 nascar\_lead\_weighted_{sgy} + \beta_2 other\_lead_{sgy} + \beta_3 income_{sgy}$$

```
## Estimation results
##   parameter                          estimate
## 1 beta_0 (Intercept)                 -0.846
## 2 beta_1 nascar_lead_weighted       -0.0008    (versus -0.004 above)
## 3 beta_2 other_lead                 -0.00000006 (other lead = bad!)
## 4 beta_3 income                      0.00002 (rich family = good!)
```

Controlling for other things matters: new estimate is 1/4 the size

Why did this matter?



Mainly because places with NASCAR tracks tend to be poorer