# Simple Linear Regression: Estimation

## EC 320: Introduction to Econometrics

Amna Javed
Winter 2020

# Prologue

# Housekeeping

**Midterm:**

- Midterm 1 is on Wednesday January 29th
- Will have extra office hours, announced through email

# Where Are We?

## Where we've been

### High Concepts

- Reviewed core ideas from statistics
- Developed a framework for thinking about causality

Also, **R**.

# Where Are We?

## Where we're going

**The Weeds!**

- Learn the mechanics of *how* OLS works
- Interpret regression results (mechanically and critically)
- Extend ideas about causality to a regression context
- Think more deeply about statistical inference
- Lay a foundation for more-sophisticated regression techniques.

Also, **moreR**.

# Simple Linear Regression

# Addressing Questions

## Example: Effect of police on crime

**Policy Question:** Do on-campus police reduce crime on campus?

- **Empirical Question:** Does the number of on-campus police officers affect campus crime rates? If so, by how much?

How can we answer these questions?

- Prior beliefs.
- Theory.
- **Data!**

## Example: Effect of police on crime

Search:

| | Police per 1000 Students | Crimes per 1000 students |
|---|---|---|
| 1 | 20.42 | 1.1 |
| 2 | 0.15 | 2 |
| 3 | 0.47 | 1.41 |
| 4 | 14.68 | 2.06 |
| 5 | 23.75 | 1.52 |
| 6 | 7.68 | 2.76 |

Showing 1 to 6 of 96 entries

Previous Next

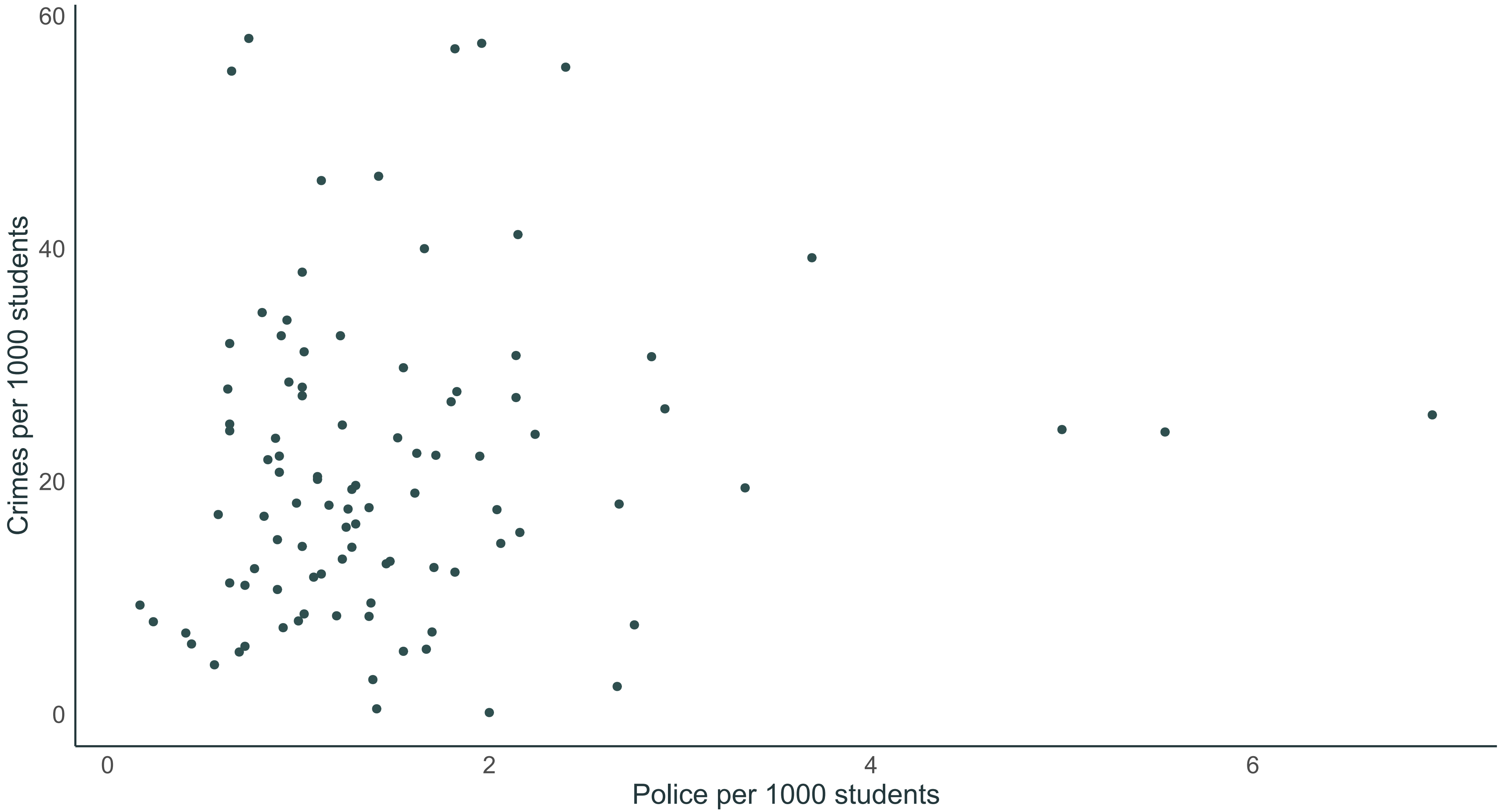
## Example: Effect of police on crime

*"Looking"* at data wasn't especially helpful.

Let's try using a scatter plot.

- Plot each data point in $(X, Y)$ -space.
- `Police` on the $X$ -axis.
- `Crime` on the $Y$ -axis.

# Example: Effect of police on crime



## Example: Effect of police on crime

The scatter plot tells us more than the spreadsheet.

- Somewhat weak *positive* relationship.
- Sample correlation coefficient of 0.14 confirms this.

But our question was

> Does the number of on-campus police officers affect campus crime rates? If so, by how much?

- The scatter plot and correlation coefficient provide only a partial answer.

## Example: Effect of police on crime

Our next step is to estimate a **statistical model.**

To keep it simple, we will relate an **explained variable**$Y$ to an **explanatory variable**$X$ in a linear model.

# Simple Linear Regression Model

We express the relationship between a **explained variable** and an **explanatory variable** as linear:

$$Y_i = \beta_1 + \beta_2 X_i + u_i.$$

- $\beta_1$ is the **intercept** or constant.
- $\beta_2$ is the **slope coefficient**.
- $u_i$ is an **error term** or disturbance term.

*Simple*= Only one explanatory variable.

# Simple Linear Regression Model

The **intercept** tells us the expected value of $Y_i$ when $X_i = 0$ .

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

Usually not the focus of an analysis.

# Simple Linear Regression Model

The **slope coefficient** tells us the expected change in $Y_i$ when $X_i$ increases by one.

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

"A one-unit increase in $X_i$ is associated with a $\beta_2$ -unit increase in $Y_i$ ."

Under certain (strong) assumptions about the error term, $\beta_2$ is the *effect of* $X_i$ *on* $Y_i$ .

- Otherwise, it's the *association of* $X_i$ *with* $Y_i$ .

# Simple Linear Regression Model

The **error term** reminds us that $X_i$ does not perfectly explain $Y_i$ .

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

Represents all other factors that explain $Y_i$ .

- Useful mnemonic: pretend that $u$ stands for *"unobserved"* or *"unexplained."*

## Example: Effect of police on crime

How might we apply the simple linear regression model to our question about the effect of on-campus police on campus crime?

- Which variable is $X$ ? Which is $Y$ ?

$$\text{Crime}_i = \beta_1 + \beta_2 \text{Police}_i + u_i.$$

- $\beta_1$ is the crime rate for colleges without police.
- $\beta_2$ is the increase in the crime rate for an additional police officer per 1000 students.

## Example: Effect of police on crime

How might we apply the simple linear regression model to our question?

$$\text{Crime}_i = \beta_1 + \beta_2 \text{Police}_i + u_i$$

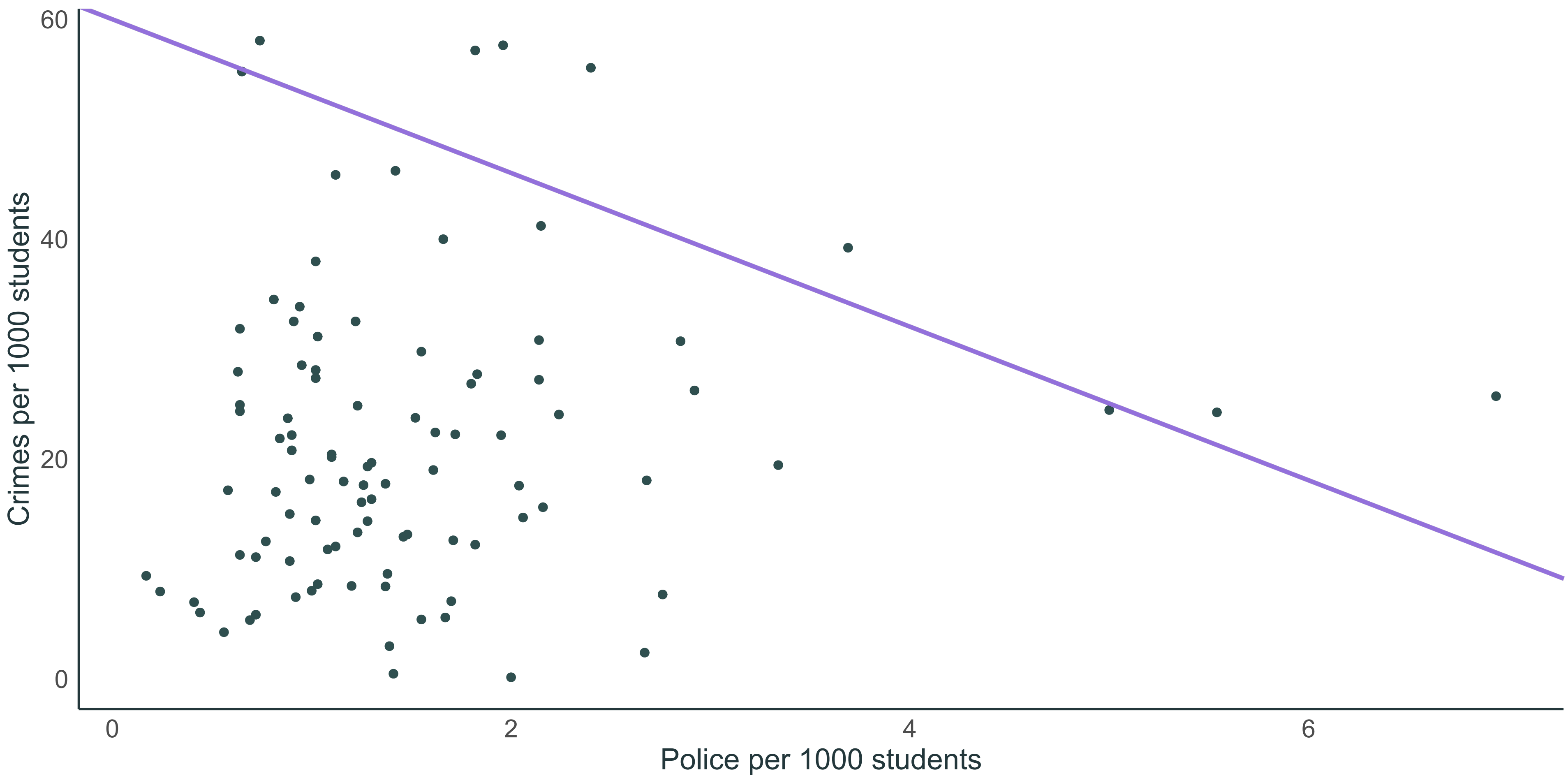
$\beta_1$ and $\beta_2$ are the population parameters we want, but we cannot observe them.

Instead, we must estimate the population parameters.

- $\hat{\beta}_1$ and $\hat{\beta}_2$ generate predictions of $\text{Crime}_i$ called $\widehat{\text{Crime}_i}$ .
- We call the predictions of the dependent variable **fitted values.**
- Together, these trace a line: $\widehat{\text{Crime}_i} = \hat{\beta}_1 + \hat{\beta}_2 \text{Police}_i$ .

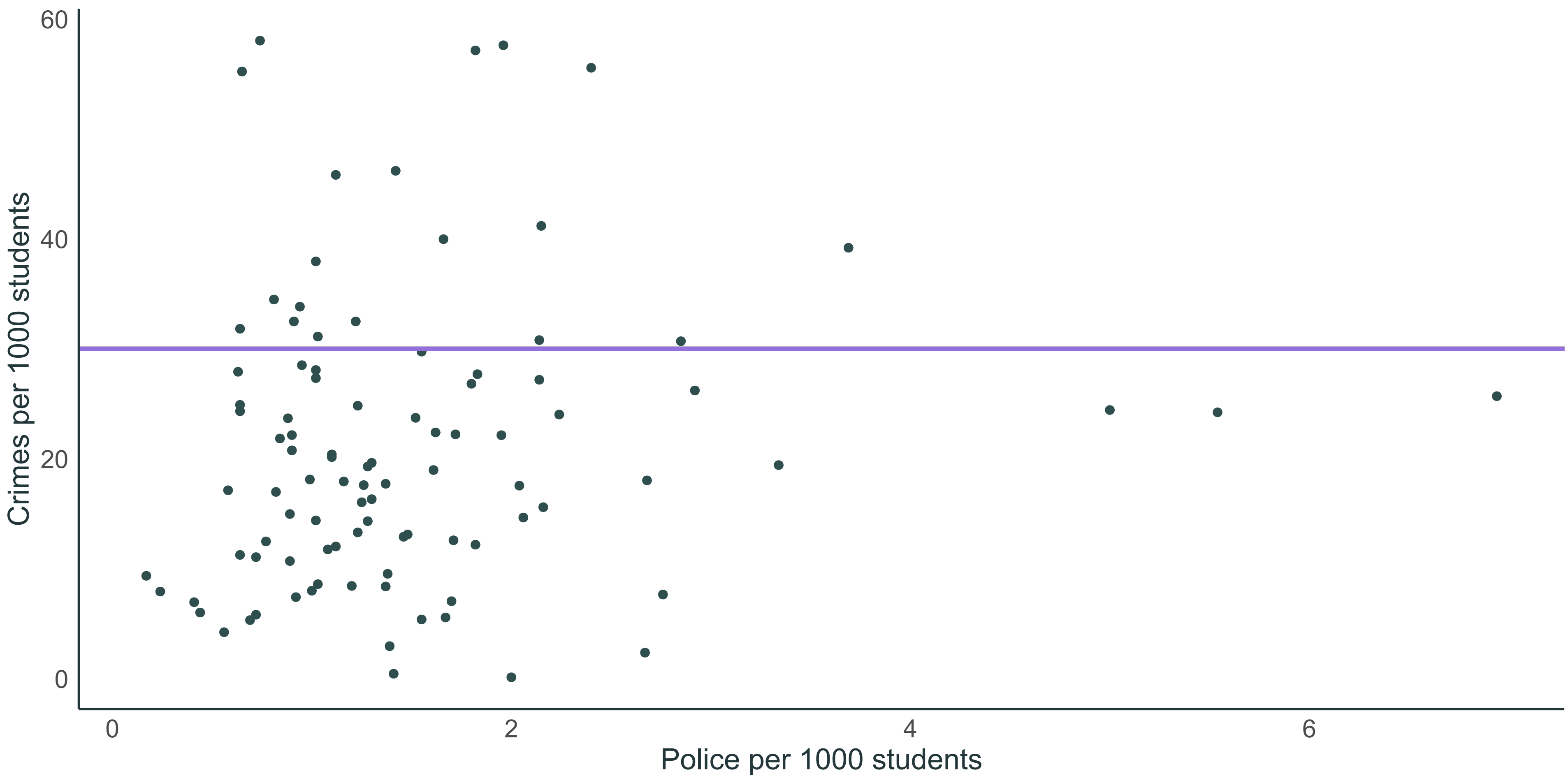
## Example: Effect of police on crime

Guess: $\hat{\beta}_1 = \mathbf{60}$ and $\hat{\beta}_2 = -\mathbf{7}$ .



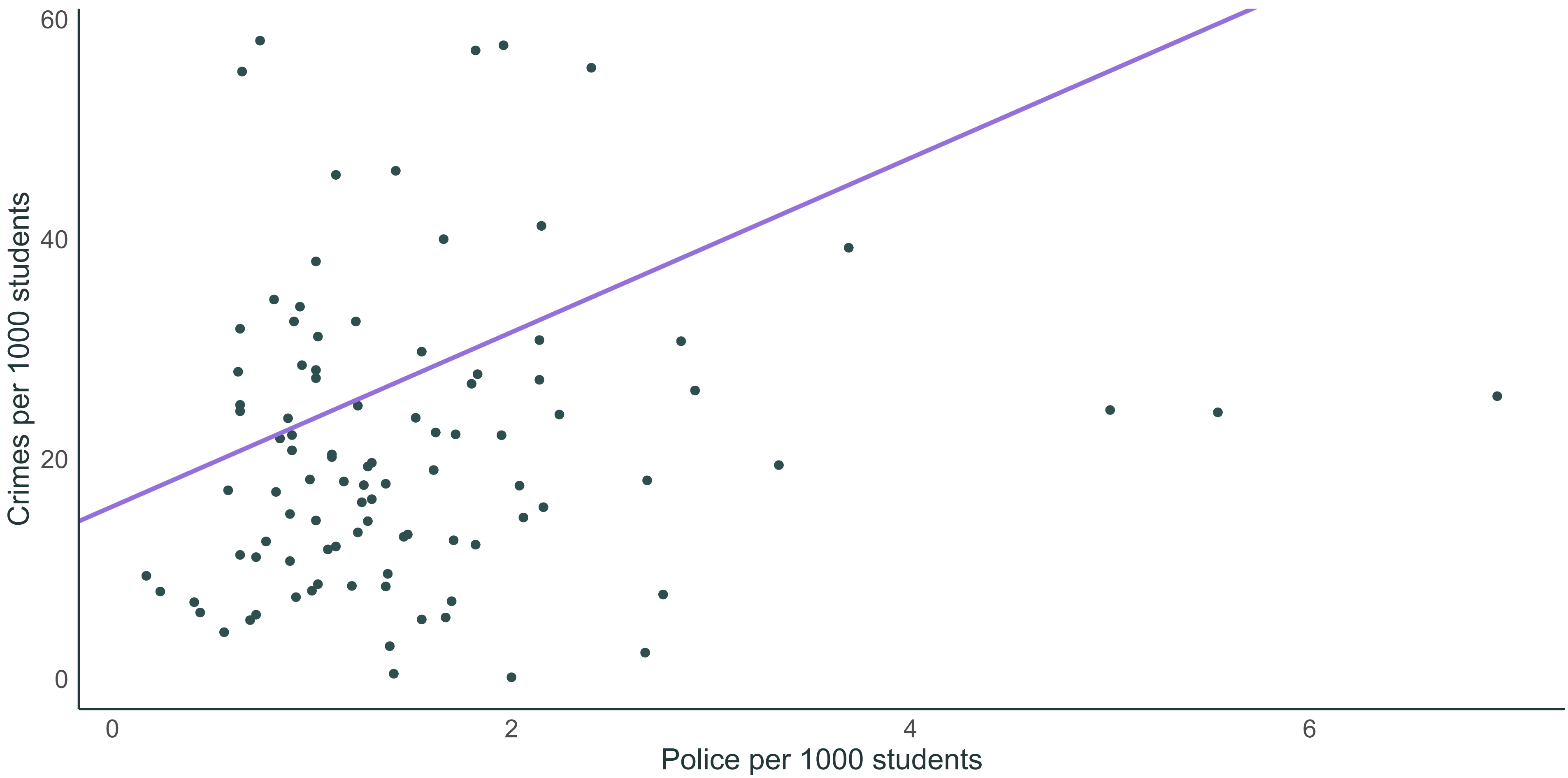
## Example: Effect of police on crime

Guess: $\hat{\beta}_1 = \mathbf{30}$ and $\hat{\beta}_2 = \mathbf{0}$ .



## Example: Effect of police on crime

Guess: $\hat{\beta}_1 = \mathbf{15.6}$ and $\hat{\beta}_2 = \mathbf{7.94}$ .



# Residuals

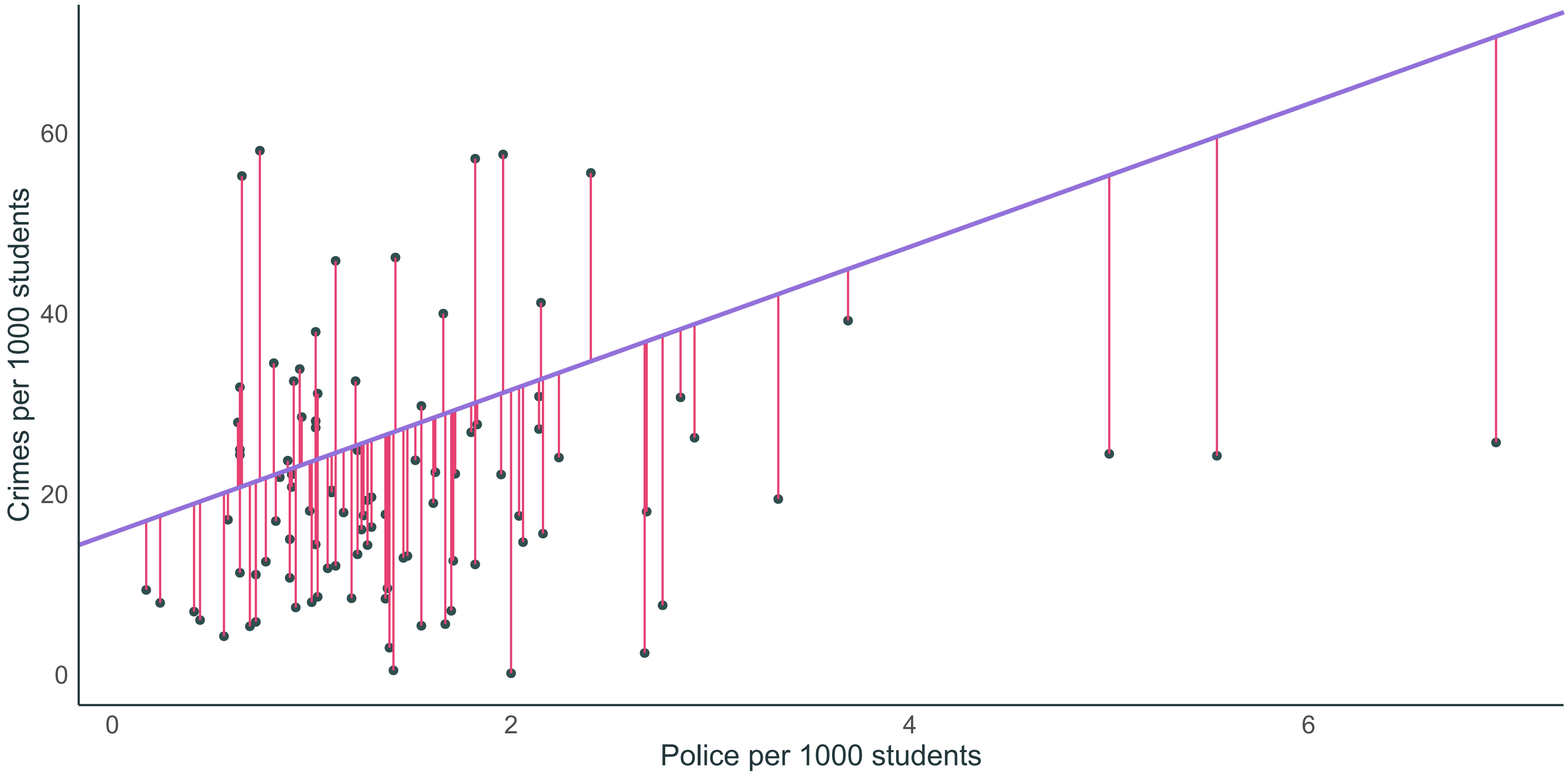
Using $\hat{\beta}_1$ and $\hat{\beta}_2$ to make $\hat{Y}_i$ generates misses called **residuals**:

$$\hat{u}_i = Y_i - \hat{Y}_i.$$

- Sometimes called $e_i$ .

## Example: Effect of police on crime

Using $\hat{\beta}_1 = \mathbf{15.6}$ and $\hat{\beta}_2 = \mathbf{7.94}$ to make $\hat{\text{Crime}}_i$ generates **residuals**.



# Residuals

We want an estimator that makes fewer big misses.

Why not minimize $\sum_{i=1}^{n} \hat{u}_i$ ?

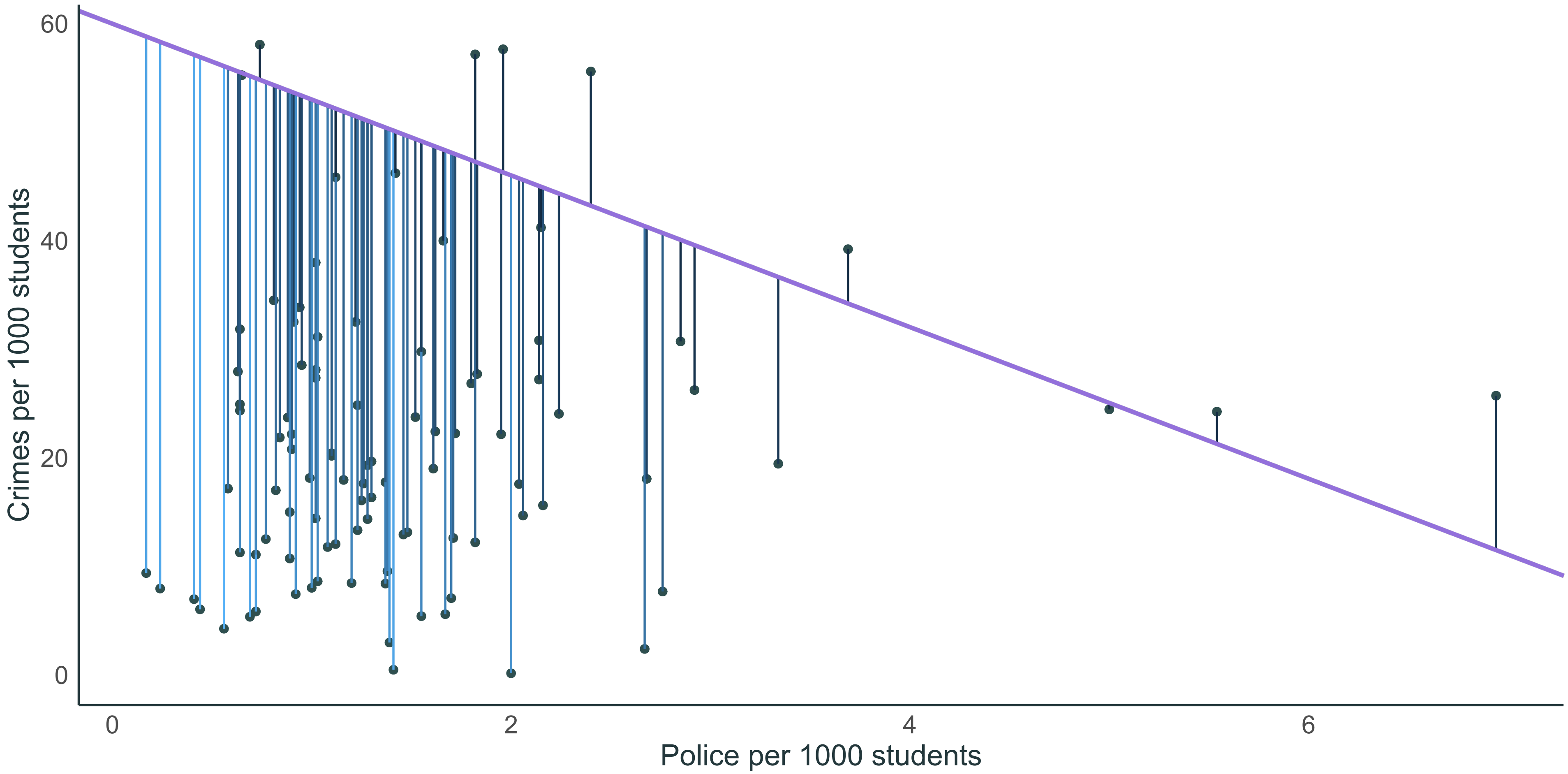
- There are positive *and* negative residuals $\implies$ no solution (can always find a line with more negative residuals).

**Alternative:** Minimize the sum of squared residuals a.k.a. the **residual sum of squares (RSS)**.

- Squared numbers are never negative.

## Example: Effect of police on crime

**RSS** gives bigger penalties to bigger residuals.



## Minimizing RSS

We could test thousands of guesses of $\hat{\beta}_1$ and $\hat{\beta}_2$ and pick the pair that minimizes RSS.

- Or we just do a little math and derive some useful formulas that give us RSS-minimizing coefficients without the guesswork.

# Ordinary Least Squares (OLS)

# OLS

The **OLS estimator** chooses the parameters $\hat{\beta}_1$ and $\hat{\beta}_2$ that minimize the **residual sum of squares (RSS)**:

$$\min_{\hat{\beta}_1, \hat{\beta}_2} \quad \sum_{i=1}^{n} \hat{u}_i^2$$

This is why we call the estimator ordinary **least squares.**

## Outline

Step 1. Replace $\sum_{i=1}^{n} \hat{u}_i^2$ with an equivalent expression involving $\hat{\beta}_1$ and $\hat{\beta}_2$ .

Step 2. Take partial derivatives of our RSS expression with respect to $\hat{\beta}_1$ and $\hat{\beta}_2$ and set each one equal to zero (first-order conditions).

Step 3. Use the first-order conditions to solve for $\hat{\beta}_1$ and $\hat{\beta}_2$ in terms of data on $\boldsymbol{Y}_i$ and $\boldsymbol{X}_i$
.

Step 4. Check second-order conditions to make sure we found the $\hat{\beta}_1$ and $\hat{\beta}_2$ that minimize RSS.

# OLS Formulas

For details, see the handout posted on Canvas.

**Slope coefficient**

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

**Intercept**

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2\bar{X}$$

# Slope coefficient

The slope estimator is equal to the sample covariance divided by the sample variance of $X$ :

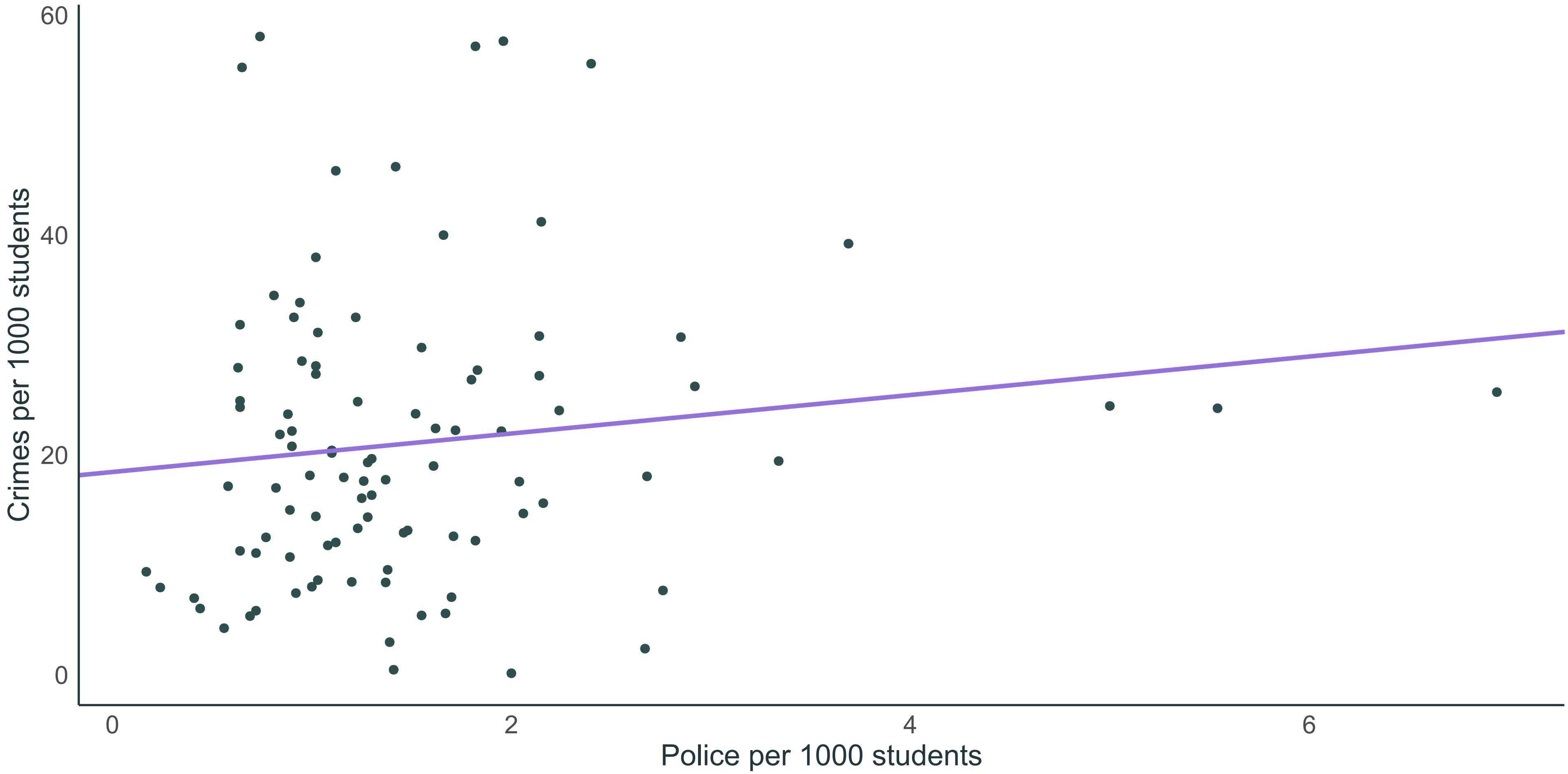
$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$= \frac{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X})}{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$= \frac{S_{XY}}{S_X^2}.$$

## Example: Effect of police on crime

Using the OLS formulas, we get $\hat{\beta}_1$ = 18.41 and $\hat{\beta}_2$ = 1.76.



## Example: Effect of police on crime

Using OLS gives us the fitted line

$$\widehat{\text{Crime}}_i = \hat{\beta}_1 + \hat{\beta}_2 \text{Police}_i.$$

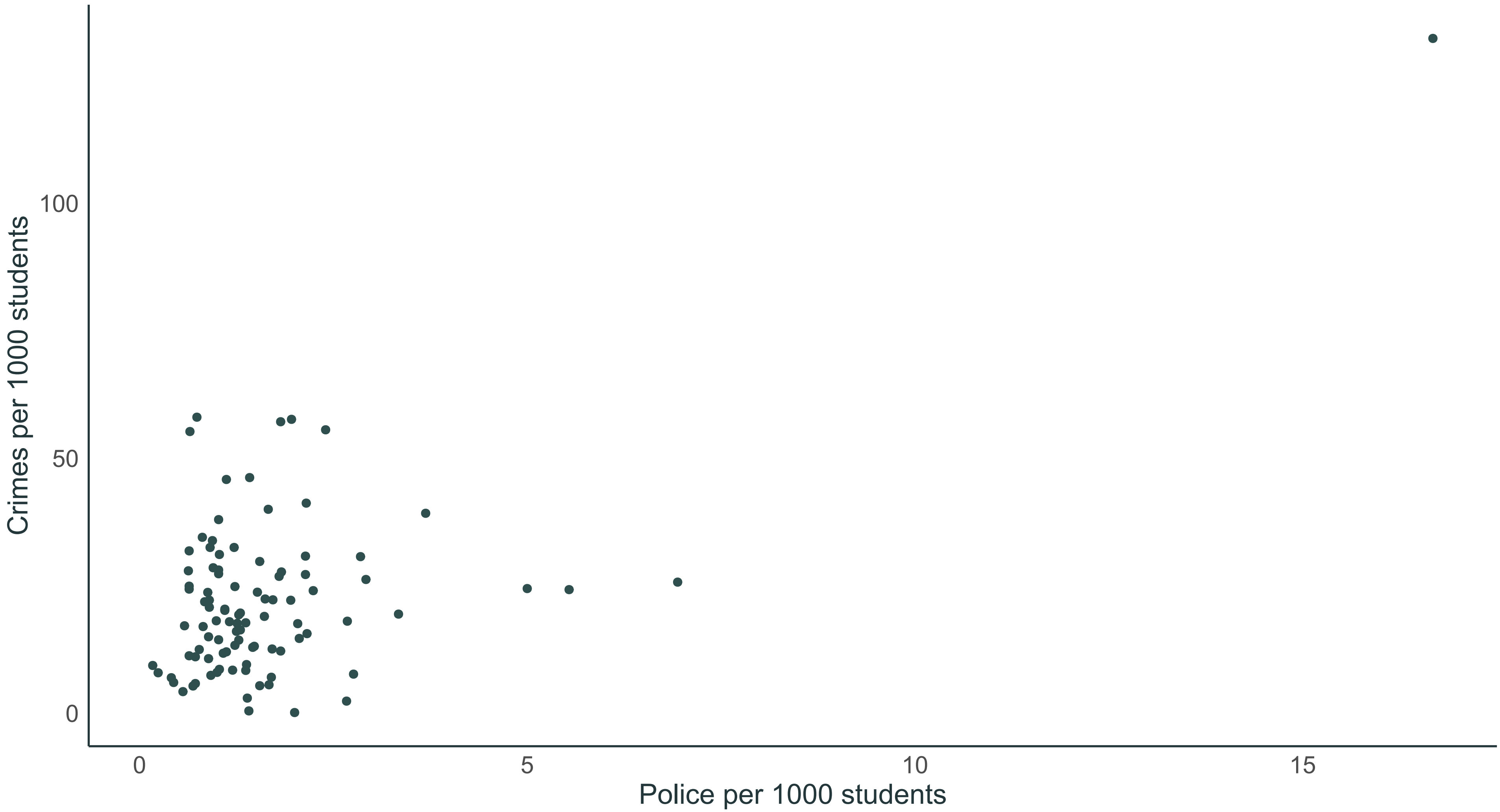
What does $\hat{\beta}_1$ = 18.41 tell us?

What does $\hat{\beta}_2$ = 1.76 tell us?

**Gut check:** Does this mean that police *cause* crime?

- Probably not. **Why?**

## Example: Association of police with crime



## Example: Association of police with crime

**Fitted line** without outlier. **Fitted line** with outlier.