# Simple Linear Regression: Estimation

## EC 320: Introduction to Econometrics

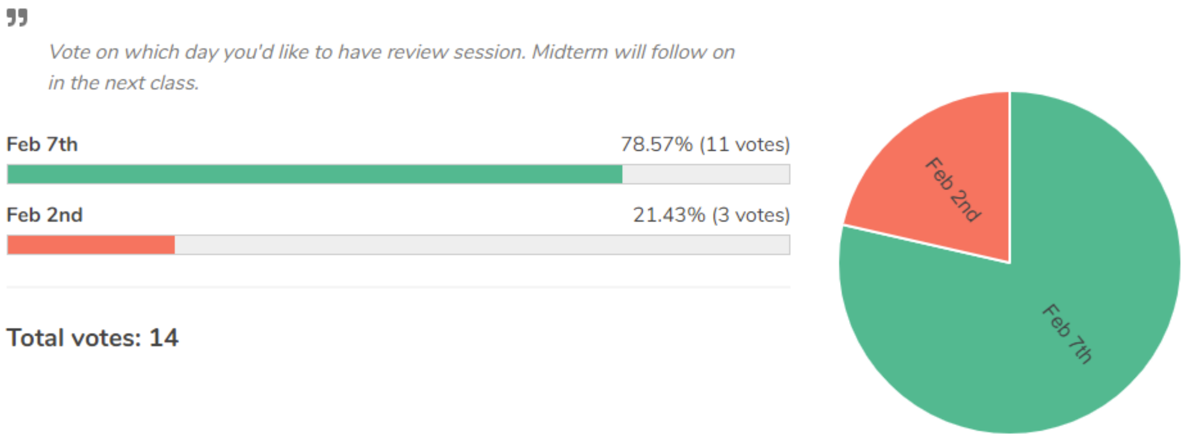
Philip Economides

Winter 2022

# Prologue

# Housekeeping

- **Midterm Review** votes are in! Syllabus updated




- Submissions for **PBS2**, great job overall
- Reminders: **PBS3** online (due 31st), **office hours** Thurs 10am, Friday 2pm, Midway Student Experience Survey

# PBS2 Q6

The fully manual version can be tedious.

```
df_1 = tibble(
  D   =c(rep(0,6), rep(1,6)),
  Y_1 =c(4,9,3,10,11,14,12,7,15,2,8,19),
  Y_0 =c(8,6,7, 8,12,19,15,9,16,1,6,12),
  Y   =c(8,6,7, 8,12,19,15,9,16,1,6,12),
  Y_c =c(8,9,7,10,12,19,15,9,16,2,8,19),
  C   =c(0,1,0,1,0,0,0,0,0,1,1,1)
)
```

Let's automate some of the code to avoid user error.

# PBS2 Q6

Consider the `if_else(arg1, arg2, arg3)` function. `arg1` is your condition. `arg2` is the given value if condition is **TRUE**. `arg3` is the given value if condition is **FALSE**.

```
df_2 <- tibble(
  D   =c(rep(0,6), rep(1,6)),
  C   =c(0,1,0,1,0,0,0,0,0,1,1,1),
  Y_1 =c(4,9,3,10,11,14,12,7,15,2,8,19),
  Y_0 =c(8,6,7, 8,12,19,15,9,16,1,6,12),
) %>%
  mutate(Y = if_else(D==1, Y_1, Y_0),
         Y_c = if_else(C==1, Y_1, Y_0))
```

# PBS2 Q6

To go one step further, consider the fact that people are choosing whether to get treated based on their respect outcomes $\{Y_{1i}, Y_{0i}\}$.

```
df_3 <- tibble(
  D   =c(rep(0,6), rep(1,6)),
  Y_1 =c(4,9,3,10,11,14,12,7,15,2,8,19),
  Y_0 =c(8,6,7, 8,12,19,15,9,16,1,6,12),
) %>%
  mutate( C = if_else(Y_1 >= Y_0, 1, 0),
          Y = if_else(D==1, Y_1, Y_0),
        Y_c = if_else(C==1, Y_1, Y_0))
```

With only three items of data: to whom random treatment is assigned and measures of our two possible outcomes for each individual, we are able to determine who selects into treatment and identify the selection bias present.

This would be the difference between answers to (iii) & (ii).

# PBS2 Q6

Resulting Dataframe

```
#> # A tibble: 12 x 6
#>         D   Y_1   Y_0     C     Y   Y_c
#>     <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
#>  1      0     4     8     0     8     8
#>  2      0     9     6     1     6     9
#>  3      0     3     7     0     7     7
#>  4      0    10     8     1     8    10
#>  5      0    11    12     0    12    12
#>  6      0    14    19     0    19    19
#>  7      1    12    15     0    12    15
#>  8      1     7     9     0     7     9
#>  9      1    15    16     0    15    16
#> 10      1     2     1     1     2     2
#> 11      1     8     6     1     8     8
#> 12      1    19    12     1    19    19
```

i) ATET (Random Assignment)

$$E[Y_i|D_i = 1] - E[Y_i|D_i = 0]$$

ii) ATET (Selection into Treatment)

$$E[Y_i^c|C_i = 1] - E[Y_i^c|C_i = 0]$$

iii) ATE (Unobservable in practice)

$$E[Y_{1i} - Y_{0i}] = \tau$$

# PBS2 Q6

i) Determine the average treatment effect, based on treatment assignment

$$E\left[Y_i|D_i=1\right]-E\left[Y_i|D_i=0\right]=E\left[Y_{1,i}|D_i=1\right]-E\left[Y_{0,i}|D_i=0\right]=0.5$$

ii) Determine the average treatment effect, based on choice.

$$E\left[Y_i^c|C_i=1\right]-E\left[Y_i^c|C_i=0\right]=E\left[Y_{1,i}|C_i=1\right]-E\left[Y_{0,i}|C_i=0\right]=-2.69$$

iii) Determine the average treatment effect, based on individual differences in outcomes

$$E\left[Y_{1,i}-Y_{0,i}\right]=-0.42$$

# Last Time

We considered a simple linear regression of $Y_i$ on $X_i$:

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

- $\beta_1$ and $\beta_2$ are **population parameters** that describe the *"true"* relationship between $X_i$ and $Y_i$.

- **Problem:** We don't know the population parameters. The best we can do is to estimate them.

# Last Time

We derived the OLS estimator by picking estimates that minimize $\sum_{i=1}^{n} \hat{u}_i^2$.

- **Intercept:**

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

- **Slope:**

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

We used these formulas to obtain estimates of the parameters $\beta_1$ and $\beta_2$ in a regression of $Y_i$ on $X_i$.

# Last Time

With the OLS estimates of the population parameters, we constructed a regression line:

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i.$$

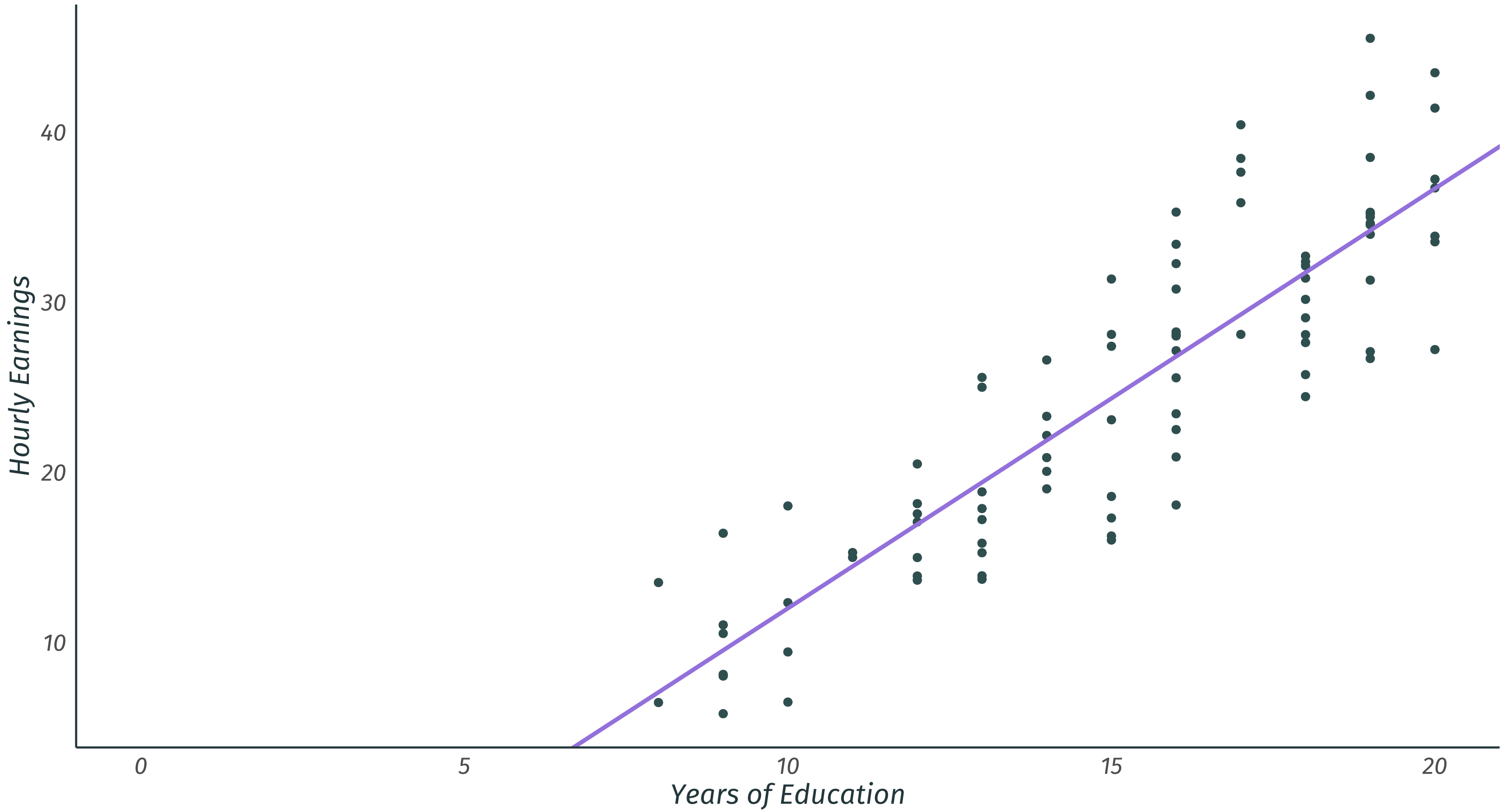
- $\hat{Y}_i$ are predicted or **fitted** values of $Y_i$.
- You can think of $\hat{Y}_i$ as an estimate of the average value of $Y_i$ given a particular of $X_i$.

OLS still produces prediction errors: $\hat{u}_i = Y_i - \hat{Y}_i$.

- Put differently, there is a part of $Y_i$ we can explain and a part we cannot: $Y_i = \hat{Y}_i + \hat{u}_i$.

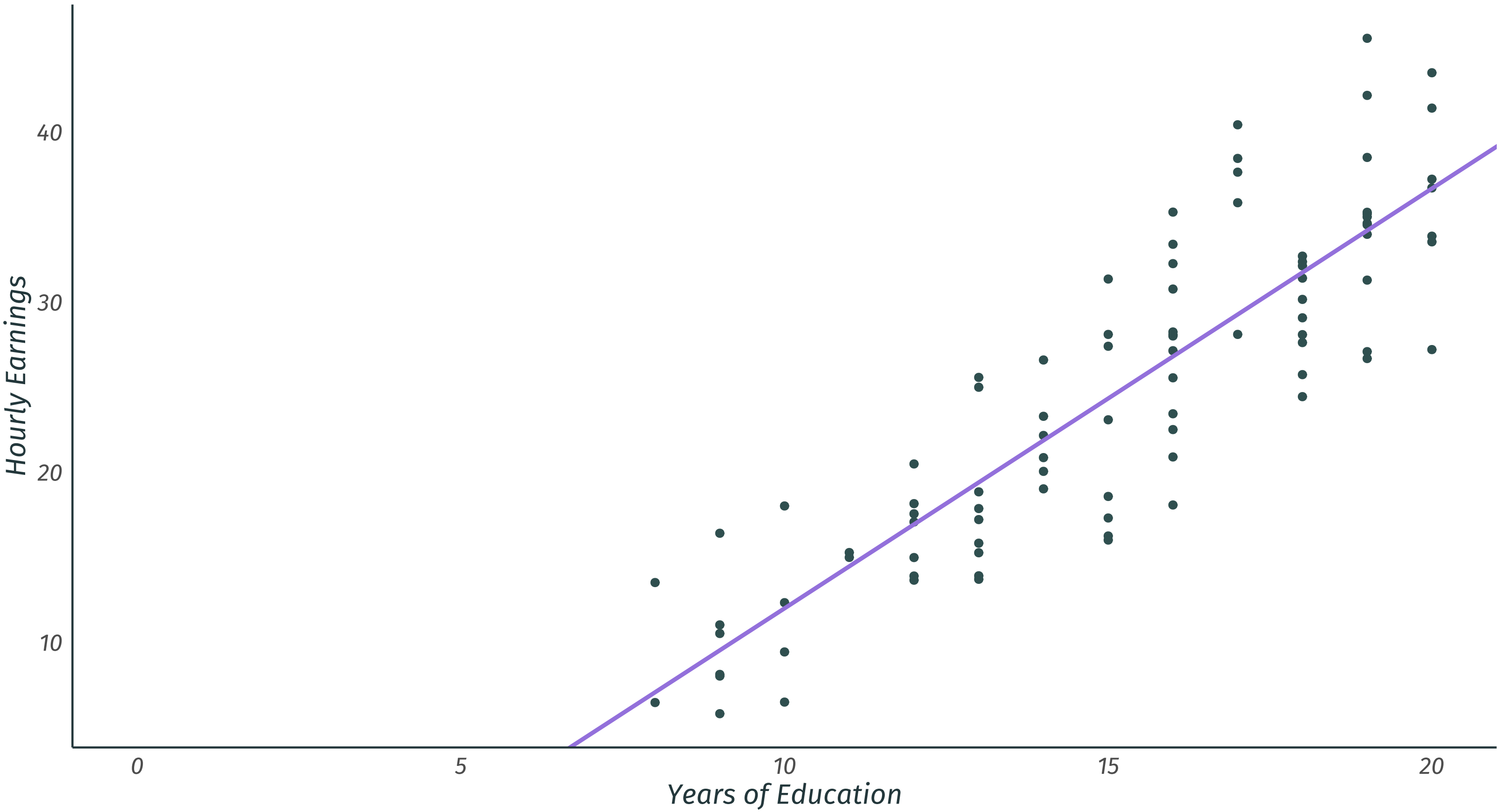
# Review

What is the equation for the regression model estimated below?



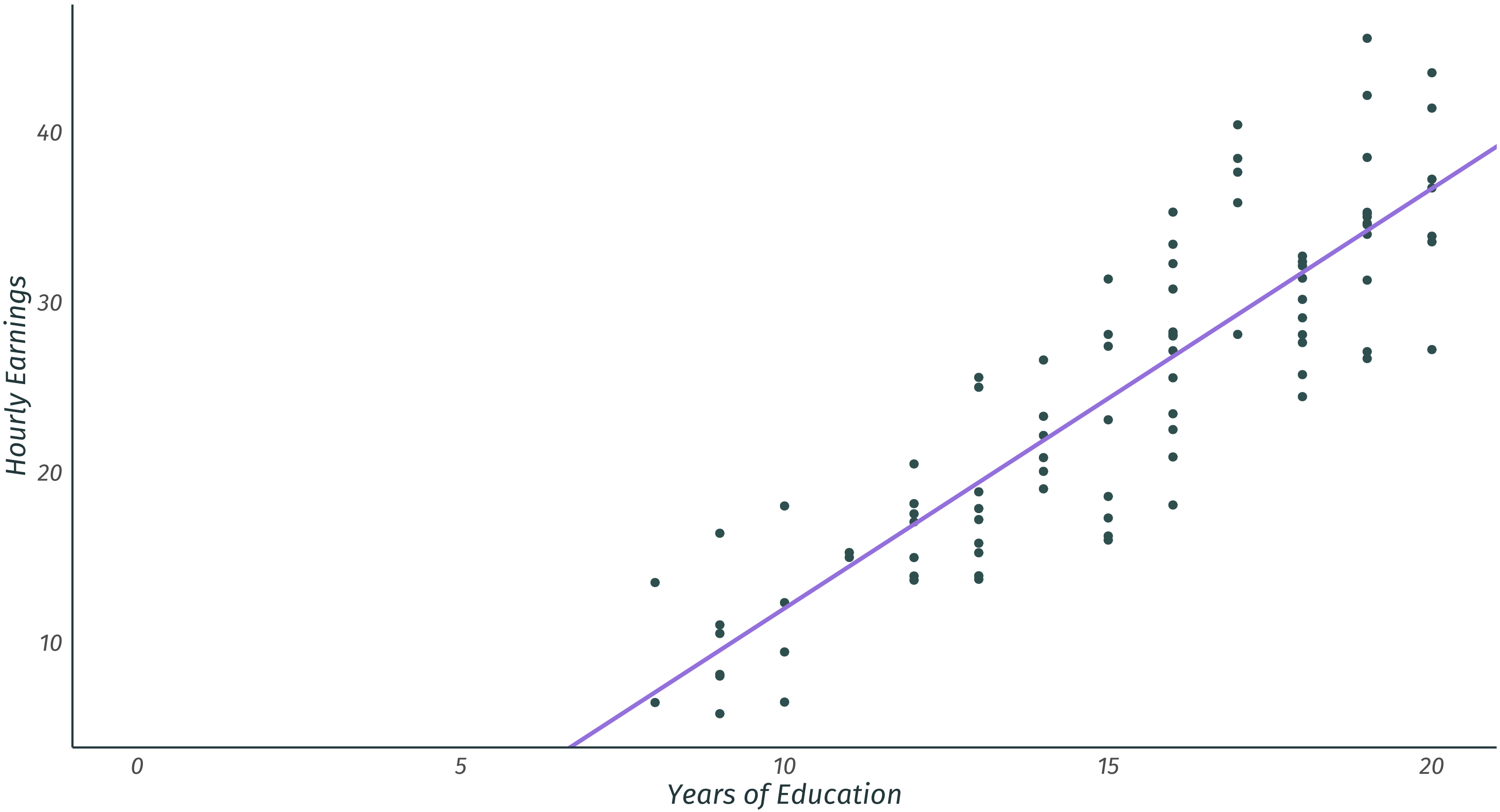
# Review

The estimated **intercept** is -12.67. What does this tell us?



# Review

The estimated **slope** is 2.47. How do we interpret it?



# Today

## Agenda

1. Highlight important properties of OLS.
2. Discuss goodness of fit: how well does one variable explain another?
3. Units of measurement.

# OLS Properties

# OLS Properties

The way we selected OLS estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ gives us three important properties:

1. Residuals sum to zero: $\sum_{i=1}^{n} \hat{u}_i = 0$.
2. The sample covariance between the independent variable and the residuals is zero: $\sum_{i=1}^{n} X_i \hat{u}_i = 0$.
3. The point $(\bar{X}, \bar{Y})$ is always on the regression line.

You will **prove** (i) and (ii) in the upcoming problem set.

# OLS Regression Line

The point $(\bar{X}, \bar{Y})$ is always on the regression line.

- Start with the regression line: $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$.
- $\hat{Y}_i = \bar{Y} - \hat{\beta}_2 \bar{X} + \hat{\beta}_2 X_i$.
- Plug $\bar{X}$ into $X_i$:

$$\begin{aligned} \hat{Y}_i &= \bar{Y} - \hat{\beta}_2 \bar{X} + \hat{\beta}_2 \bar{X} \\ &= \bar{Y}. \end{aligned}$$

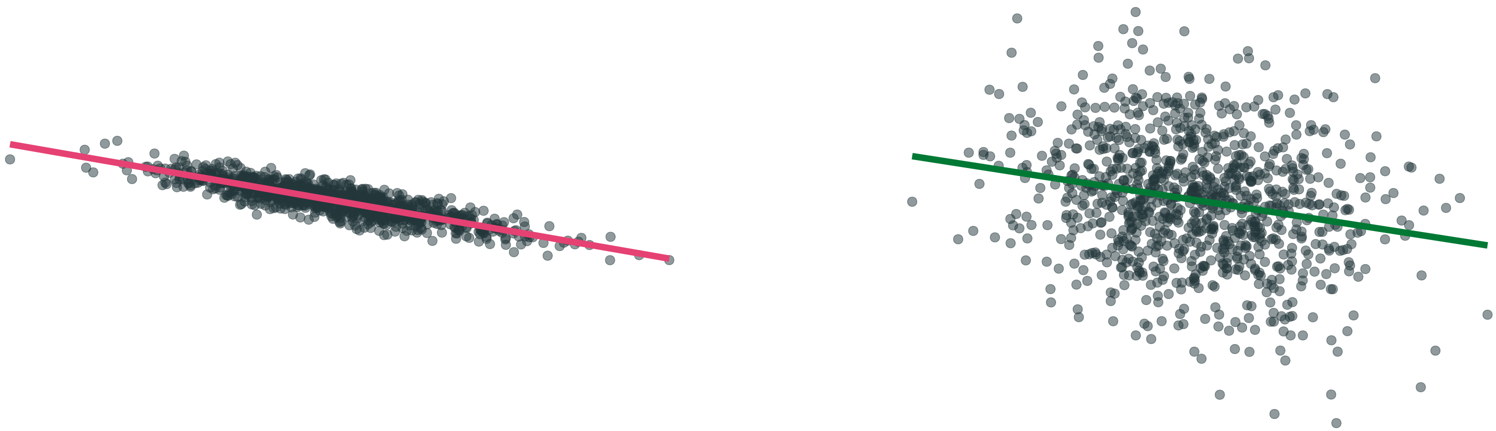
# Goodness of Fit

# Goodness of Fit

## Regression 1 *vs.* Regression 2

- Same slope.
- Same intercept.

**Q:** Which fitted regression line *"explains"*[*] the data better?



[*] *Explains = fits.*

# Goodness of Fit

## Regression 1 *vs.* Regression 2

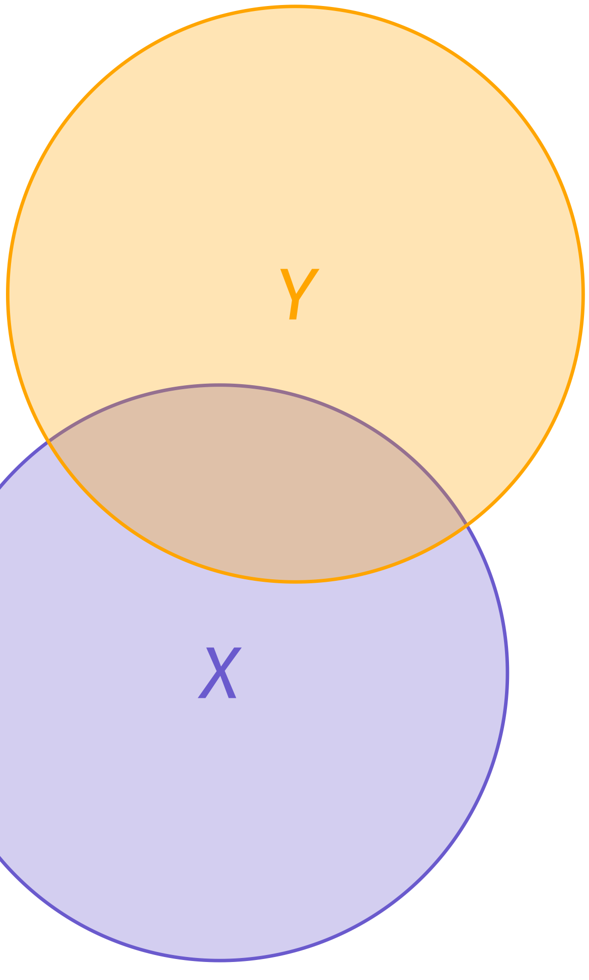
The **coefficient of determination** $R^2$ is the fraction of the variation in $Y_i$ *"explained"* by $X_i$ in a linear regression.

- $R^2 = 1 \implies X_i$ explains *all* of the variation in $Y_i$.
- $R^2 = 0 \implies X_i$ explains *none* of the variation in $Y_i$.

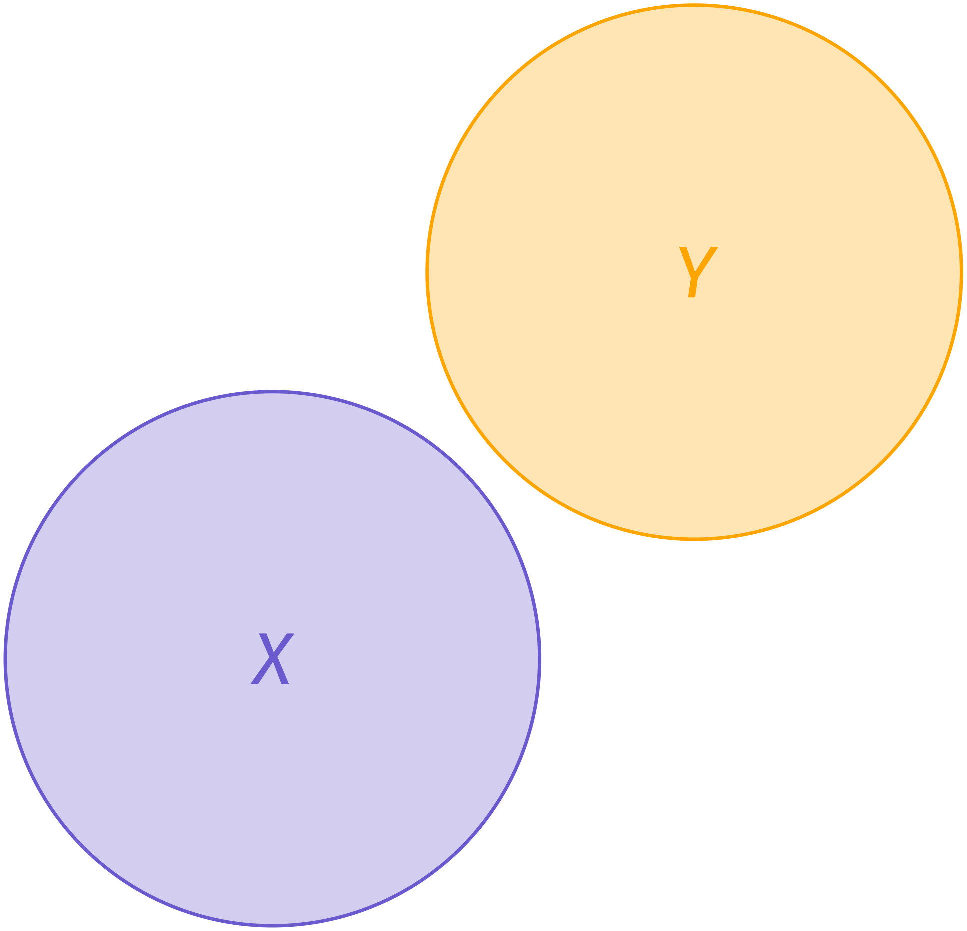
$R^2$ = 0.74

$R^2$ = 0.05

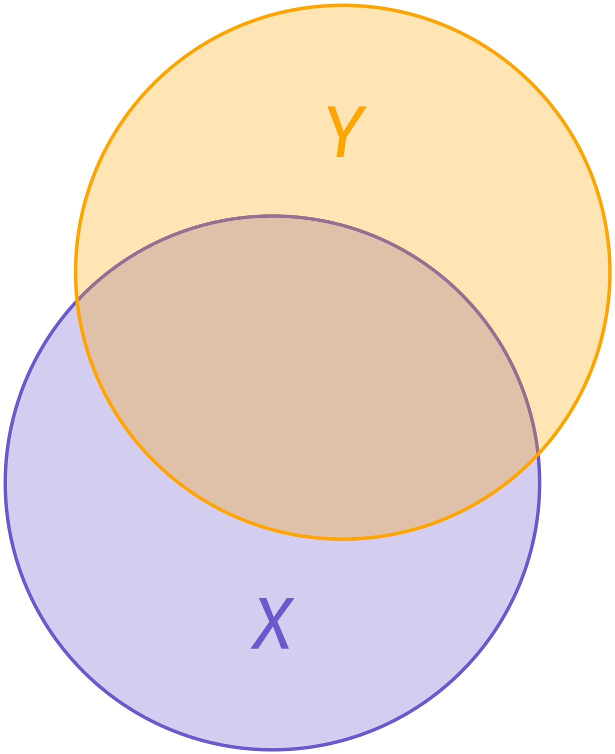
# Goodness of Fit



# Goodness of Fit



# Goodness of Fit



# Explained and Unexplained Variation

Residuals remind us that there are parts of $Y_i$ we can't explain.

$$Y_i = \hat{Y}_i + \hat{u}_i$$

- Sum the above, divide by $n$, and use the fact that OLS residuals sum to zero to get $\bar{\hat{u}} = 0 \implies \bar{Y} = \bar{\hat{Y}}$.

**Total Sum of Squares (TSS)** measures variation in $Y_i$:

$$\text{TSS} \equiv \sum_{i=1}^{n} (Y_i - \bar{Y})^2.$$

- We will decompose this variation into explained and unexplained parts.

# Explained and Unexplained Variation

**Explained Sum of Squares (ESS)** measures the variation in $\hat{Y}_i$:

$$\text{ESS} \equiv \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2.$$

**Residual Sum of Squares (RSS)** measures the variation in $\hat{u}_i$:

$$\text{RSS} \equiv \sum_{i=1}^{n} \hat{u}_i^2.$$

**Goal:** Show that $\text{TSS} = \text{ESS} + \text{RSS}$.

**Step 1:** Plug $Y_i = \hat{Y}_i + \hat{u}_i$ into TSS.

$$
\begin{aligned}
\text{TSS} & \\
&= \textstyle\sum_{i=1}^n (Y_i - \bar{Y})^2 \\
&= \textstyle\sum_{i=1}^n ([\hat{Y}_i + \hat{u}_i] - [\bar{\hat{Y}} + \bar{\hat{u}}])^2
\end{aligned}
$$

**Step 2:** Recall that $\bar{\hat{u}} = 0$ and $\bar{Y} = \bar{\hat{Y}}$.

$$
\begin{aligned}
\text{TSS} & \\
&= \textstyle\sum_{i=1}^n \left([\hat{Y}_i - \bar{Y}] + \hat{u}_i\right)^2 \\
&= \textstyle\sum_{i=1}^n \left([\hat{Y}_i - \bar{Y}] + \hat{u}_i\right) \left([\hat{Y}_i - \bar{Y}] + \hat{u}_i\right) \\
&= \textstyle\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^n \hat{u}_i^2 + 2 \sum_{i=1}^n \left((\hat{Y}_i - \bar{Y})\hat{u}_i\right)
\end{aligned}
$$

**Step 3:** Notice **ESS** and **RSS**.

$$
\begin{aligned}
&\text{TSS} \\
&\quad = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n} \hat{u}_i^2 + 2\sum_{i=1}^{n}\left((\hat{Y}_i - \bar{Y})\hat{u}_i\right) \\
&\quad = \text{ESS} + \text{RSS} + 2\sum_{i=1}^{n}\left((\hat{Y}_i - \bar{Y})\hat{u}_i\right)
\end{aligned}
$$

**Step 4:** Simplify.

$$
\begin{aligned}
&\mathrm{TSS} \\
&\quad = \mathrm{ESS} + \mathrm{RSS} + 2\sum_{i=1}^{n}\left((\hat{Y}_i - \bar{Y})\hat{u}_i\right) \\
&\quad = \mathrm{ESS} + \mathrm{RSS} + 2\sum_{i=1}^{n}\hat{Y}_i\hat{u}_i - 2\bar{Y}\sum_{i=1}^{n}\hat{u}_i
\end{aligned}
$$

**Step 5:** Shut down the last two terms. Notice that

$$
\begin{aligned}
&\sum_{i=1}^{n}\hat{Y}_i\hat{u}_i \\
&\quad = \sum_{i=1}^{n}(\hat{\beta}_1 + \hat{\beta}_2 X_i)\hat{u}_i \\
&\quad = \hat{\beta}_1\sum_{i=1}^{n}\hat{u}_i + \hat{\beta}_2\sum_{i=1}^{n}X_i\hat{u}_i \\
&\quad = 0
\end{aligned}
$$

As previously highlighted, these two terms will be equal to zero, as you will all prove in the upcoming assignment.

# Goodness of Fit

What percentage of the variation in our $Y_i$ is *apparently* explained by our model? The $R^2$ term represents this percentage.

Total variation is represented by **TSS** and our model is capturing the 'explained' sum of squares, **ESS**.

Taking a simple ratio reveals how much variation our model explains.

- $R^2 = \frac{\text{ESS}}{\text{TSS}}$ varies between 0 and 1
- $R^2 = 1 - \frac{\text{RSS}}{\text{TSS}}$, 100% less the unexplained variation

$R^2$ is related to the correlation between the actual values of $Y$ and the fitted values of $Y$. Can show that $R^2 = (r_{Y,\hat{Y}})^2$.

# Goodness of Fit

## So what?

In the social sciences, low $R^2$ values are common.

Low $R^2$ doesn't mean that an estimated regression is useless.

- In a randomized control trial, $R^2$ is usually less than 0.1

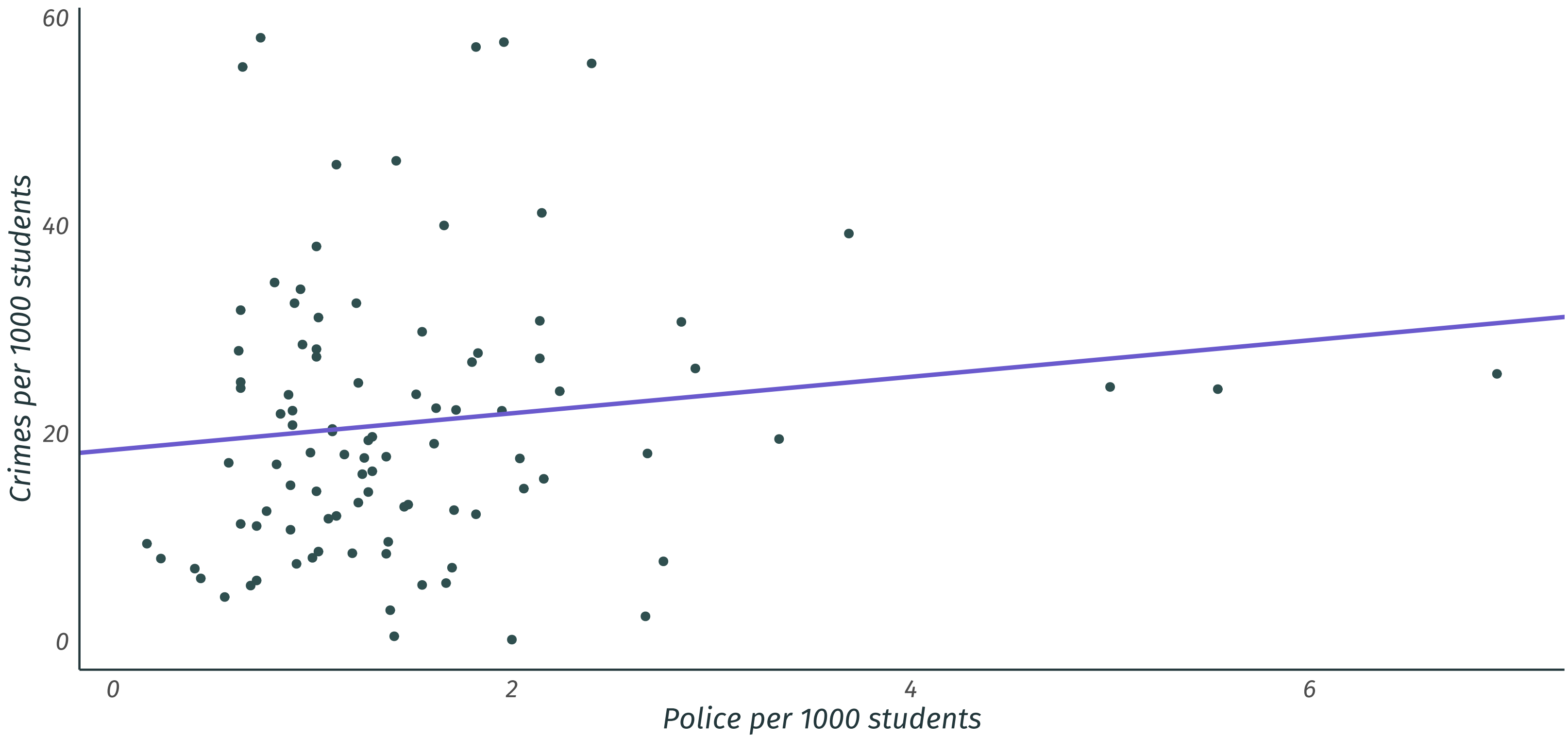
High $R^2$ doesn't necessarily mean you have a *"good"* regression.

- Worries about selection bias and omitted variables still apply
- Some 'powerfully high' $R^2$ values are the result of simple accounting exercises, and tell us nothing about causality
(e.g. $Y = C + I + G + X - M$)

# Units of Measurement

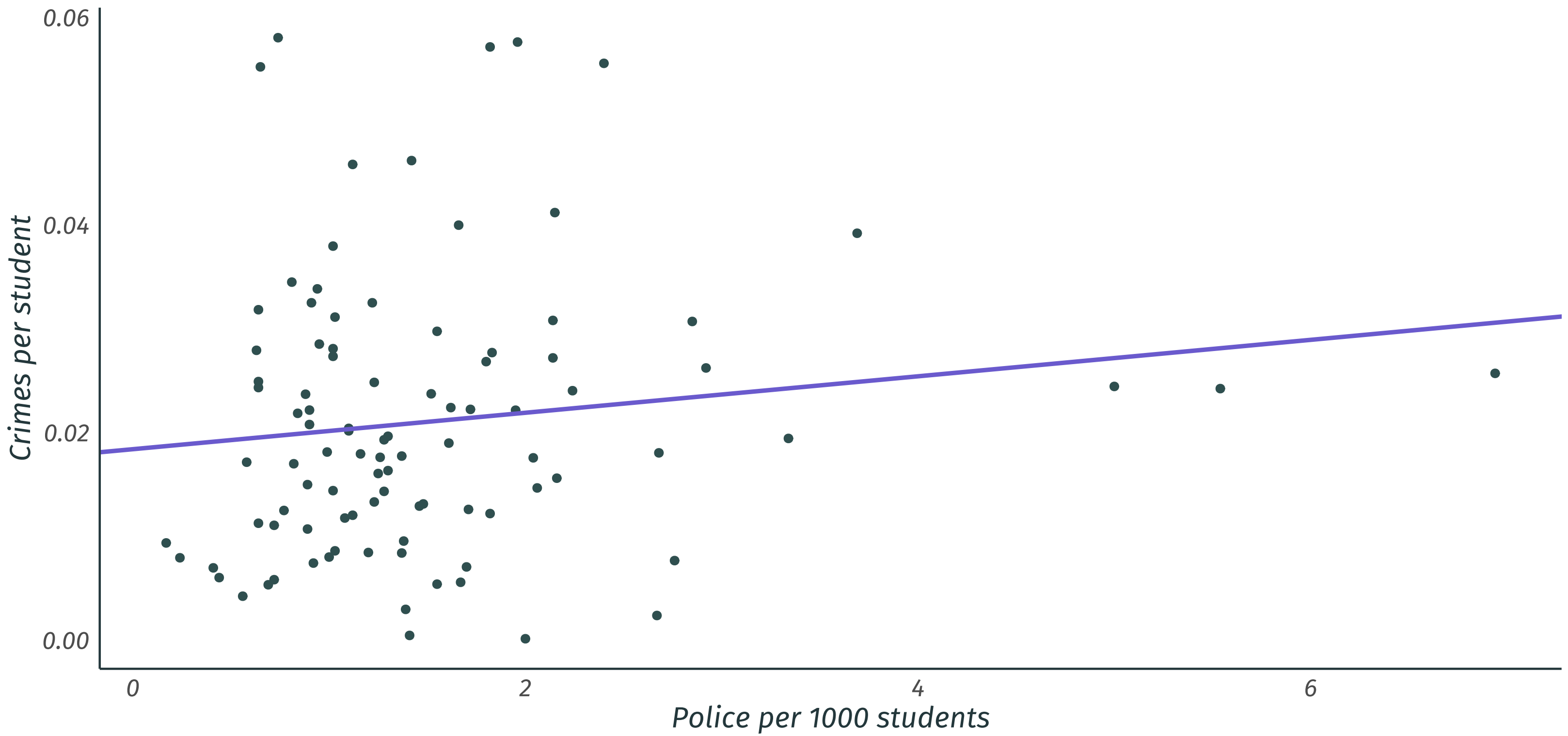
# Last Time

We ran a regression of crimes per 1000 students on police per 1000 students. We found that $\hat{\beta}_1 = 18.41$ and $\hat{\beta}_2 = 1.76$.



# Last Time

What if we had run a regression of crimes per student on police per 1000 students? What would happen to the slope?




$\hat{\beta}_2$ = 0.001756.

# Demeaning

## Practice problem

Suppose that, before running a regression of $Y_i$ on $X_i$, you decided to *demean* each variable by subtracting off the mean from each observation. This gave you $\tilde{Y}_i = Y_i - \bar{Y}$ and $\tilde{X}_i = X_i - \bar{X}$.

Then you decide to estimate

$$\tilde{Y}_i = \beta_1 + \beta_2 \tilde{X}_i + u_i.$$

What will you get for your intercept estimate $\hat{\beta}_1$?