# Quantitative Methods

## Weeks 9-11: Linear Regression

AY 2023-24

Department of Political Economy

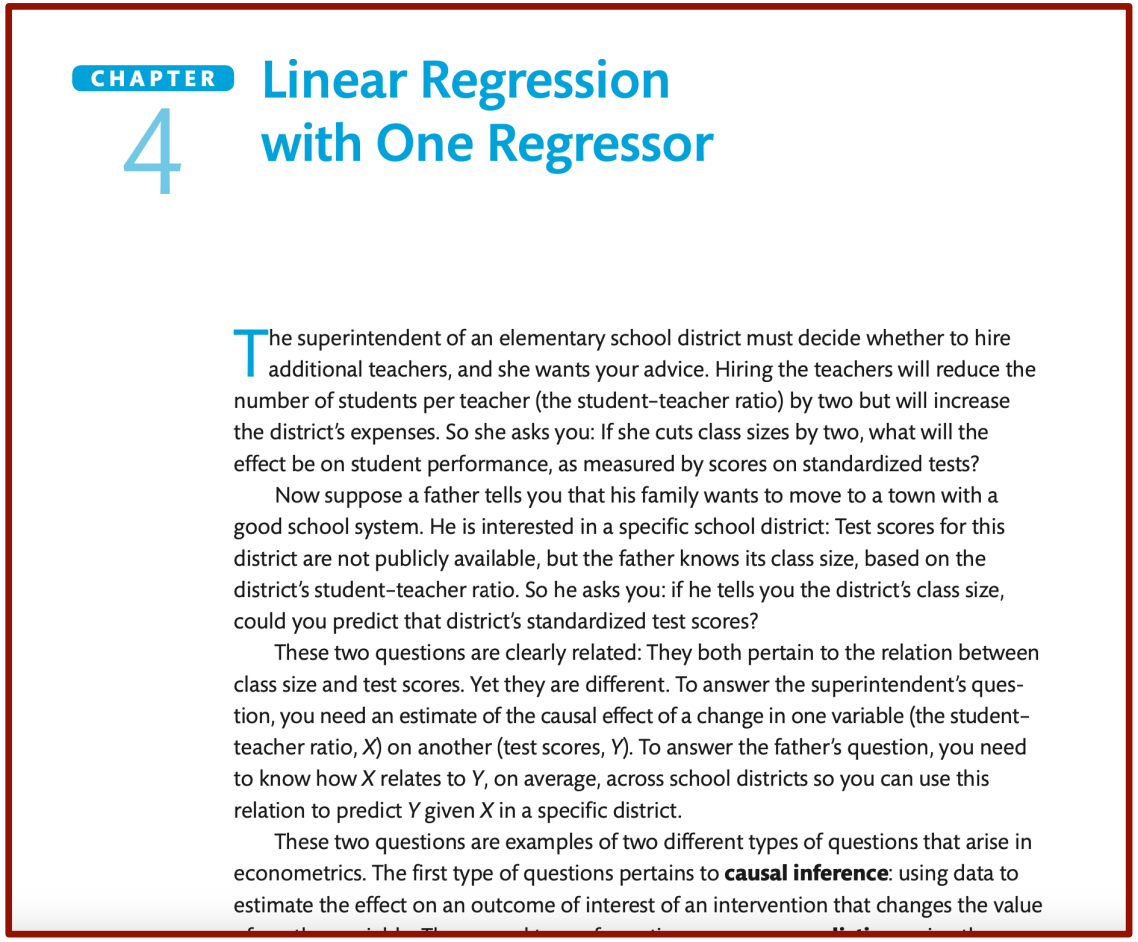
Instructor: Daniele Girardi

# Weeks 9 to 11
# Linear Regression


1. Linear regression with one regressor
   - *Stock & Watson Chapter 4*
2. Statistical inference about linear regression
   - *Stock & Watson Chapter 5*
3. Linear regression with multiple regressors
   - *Stock & Watson Chapters 6 & 7*

# **Write down** three things you learned from the reading

*(Stock & Watson Chapter 4)*



CHAPTER 4

## Linear Regression with One Regressor

The superintendent of an elementary school district must decide whether to hire additional teachers, and she wants your advice. Hiring the teachers will reduce the number of students per teacher (the student–teacher ratio) by two but will increase the district's expenses. So she asks you: If she cuts class sizes by two, what will the effect be on student performance, as measured by scores on standardized tests?

Now suppose a father tells you that his family wants to move to a town with a good school system. He is interested in a specific school district: Test scores for this district are not publicly available, but the father knows its class size, based on the district's student–teacher ratio. So he asks you: if he tells you the district's class size, could you predict that district's standardized test scores?

These two questions are clearly related: They both pertain to the relation between class size and test scores. Yet they are different. To answer the superintendent's question, you need an estimate of the causal effect of a change in one variable (the student–teacher ratio, *X*) on another (test scores, *Y*). To answer the father's question, you need to know how *X* relates to *Y*, on average, across school districts so you can use this relation to predict *Y* given *X* in a specific district.

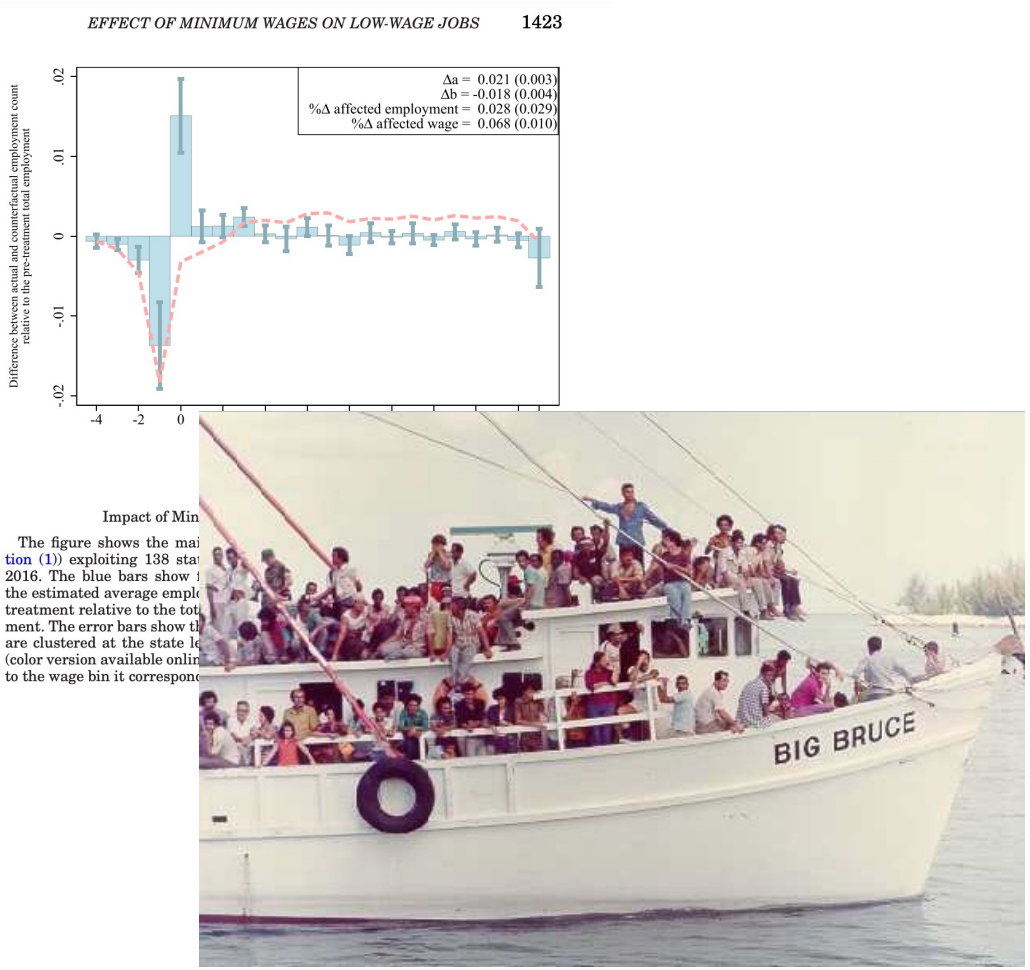
These two questions are examples of two different types of questions that arise in econometrics. The first type of questions pertains to **causal inference**: using data to estimate the effect on an outcome of interest of an intervention that changes the value

*If you couldn't do the reading this week:*

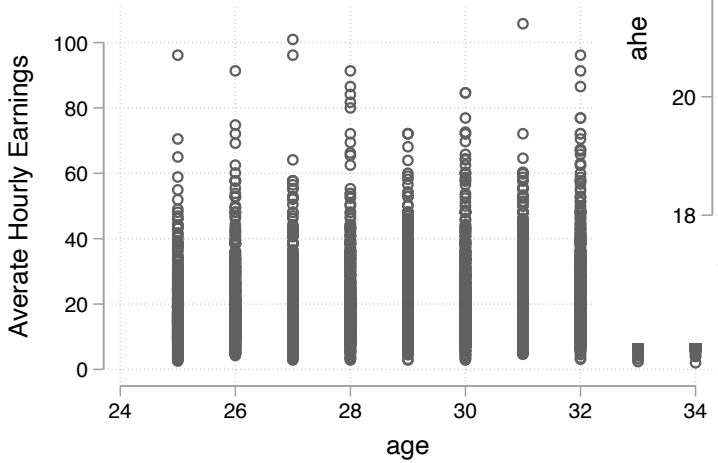
Write three things you remember from the last class.

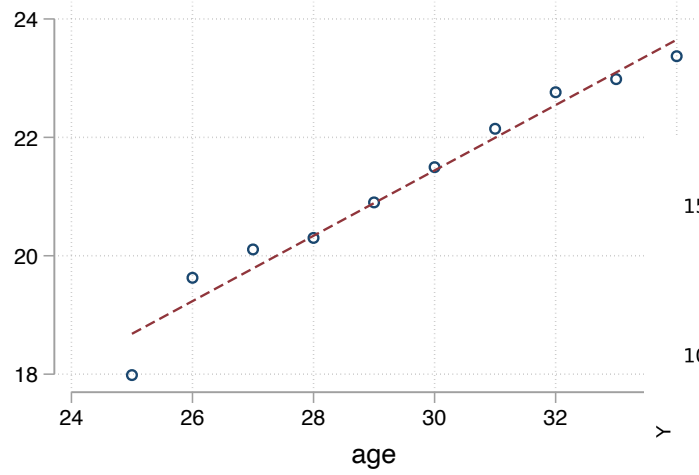
# Linear Regression: Overview

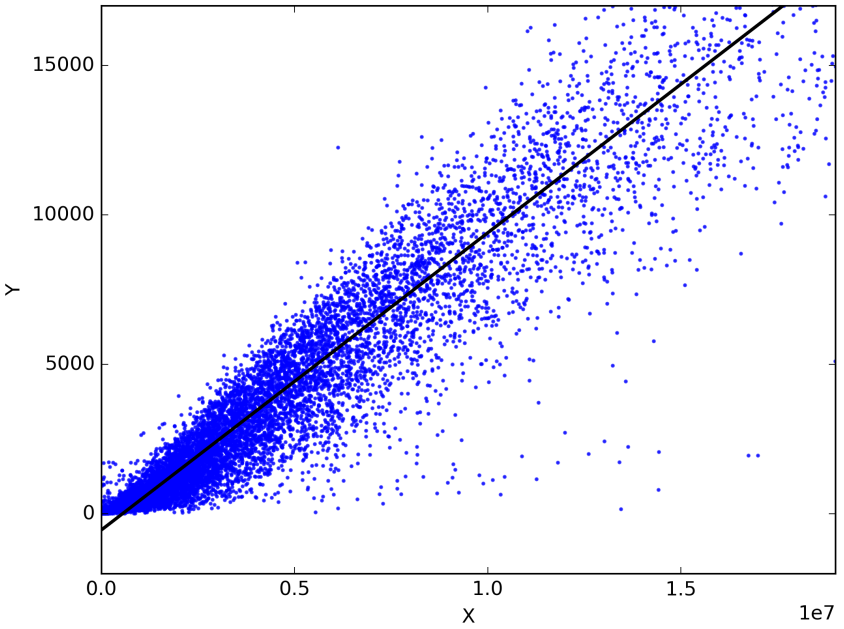
- Does a minimum wage decrease employment?
- Does immigration lower wages for native workers?
- Is a recession likely in the next year?
- ….



# 1. The linear regression model



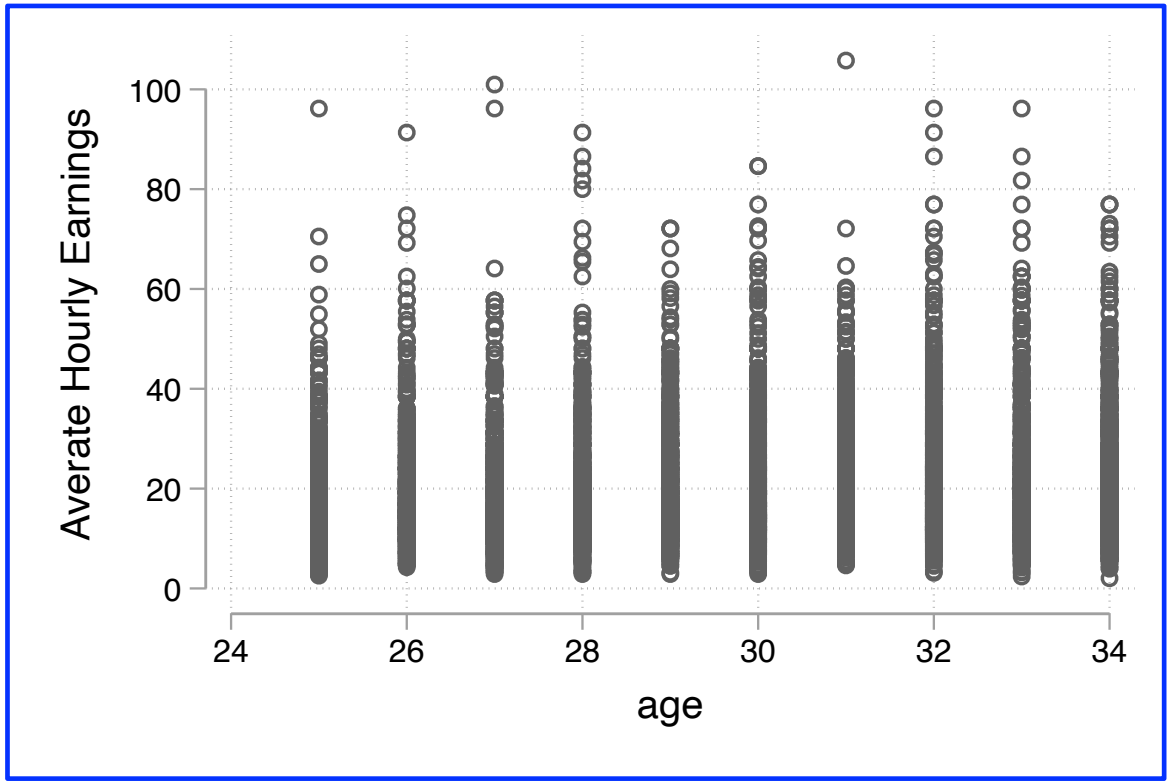



# Conditional expectations

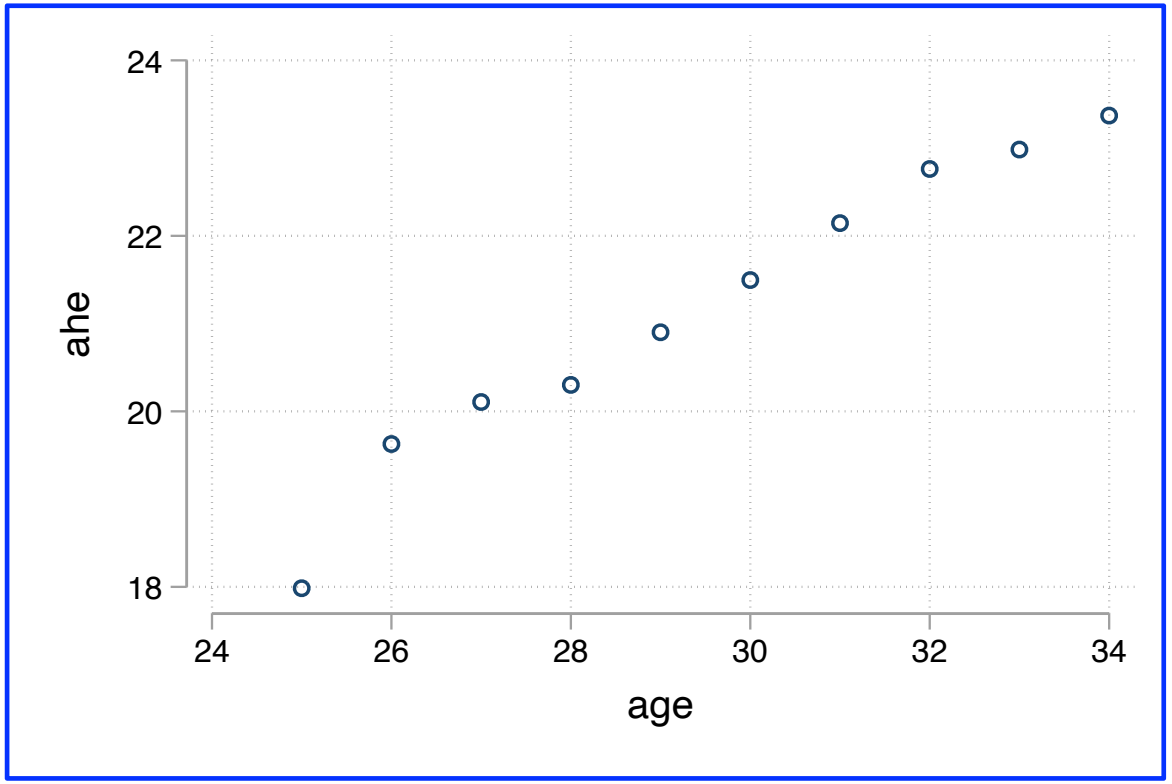
- How do average earnings vary with age?
- Conditional Expectation: $E(AHE|AGE = x)$
- Imagine we observe the population.

*Scatterplot of the (hypothetical) population data*



# Conditional expectations

- Compute conditional mean $E(AHE|AGE = x)$ for each possible value *x* that AGE can take.
- Can be used to *predict* earnings based on age.
- Not possible when X is continuous.
- Does not give us a single "average effect" of age.

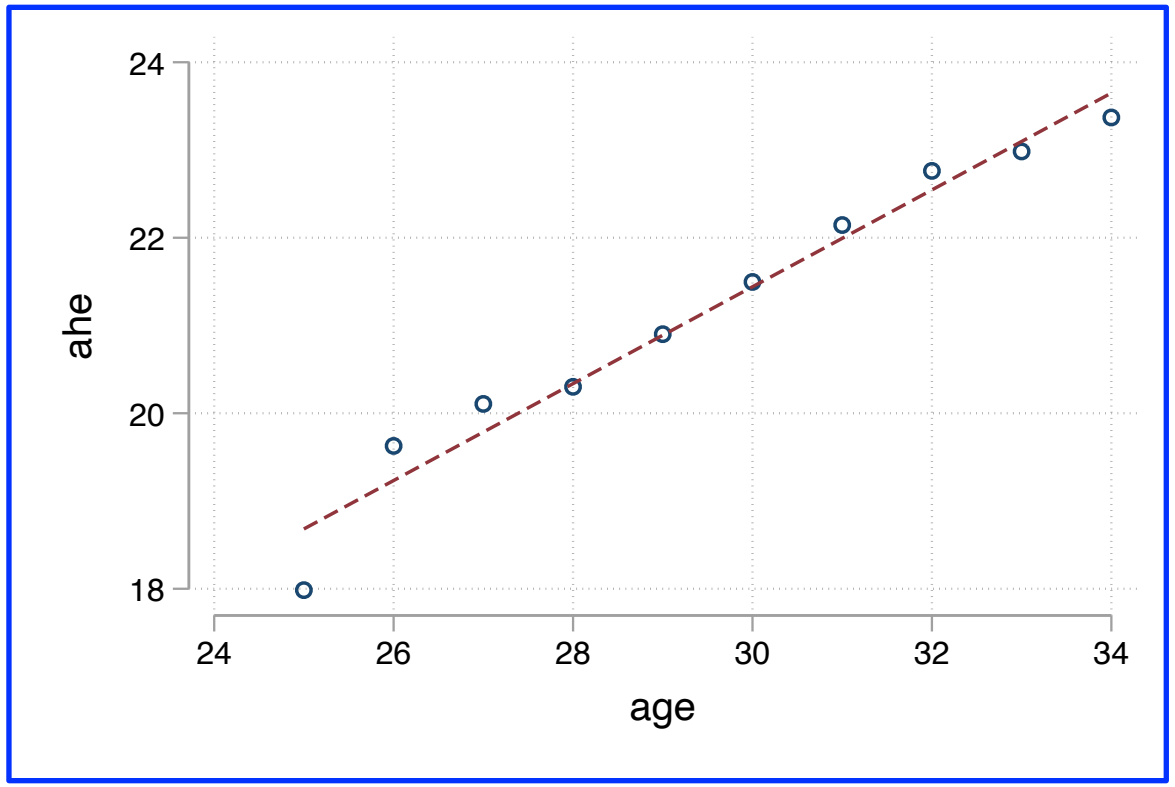


# Conditional expectations

- What if we assume the relation is linear?
- Linear CE:

$$E(AHE|AGE) = \beta_0 + \beta_1 AGE$$

- Works with continuous X.
- $\beta_1$ gives an "average effect"
- Can be used to *predict* earnings based on age.

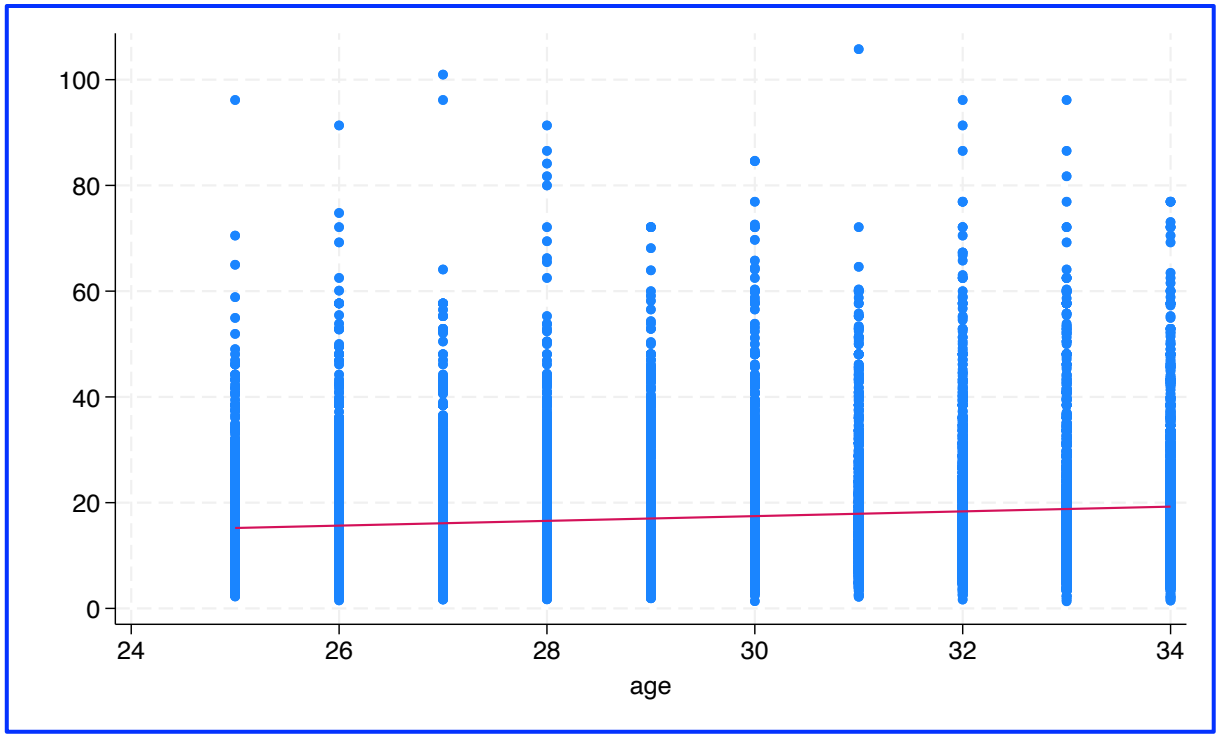


# Conditional expectations

- For an individual worker *i* in the population:

$$AHE_i = E(AHE|AGE_i) + u_i$$

$$= \underbrace{\beta_0 + \beta_1 AGE_i}_{\text{Predicted value based on the worker's age}} + \underbrace{u_i}_{\text{Error term (individual's deviation from the predicted value)}}$$



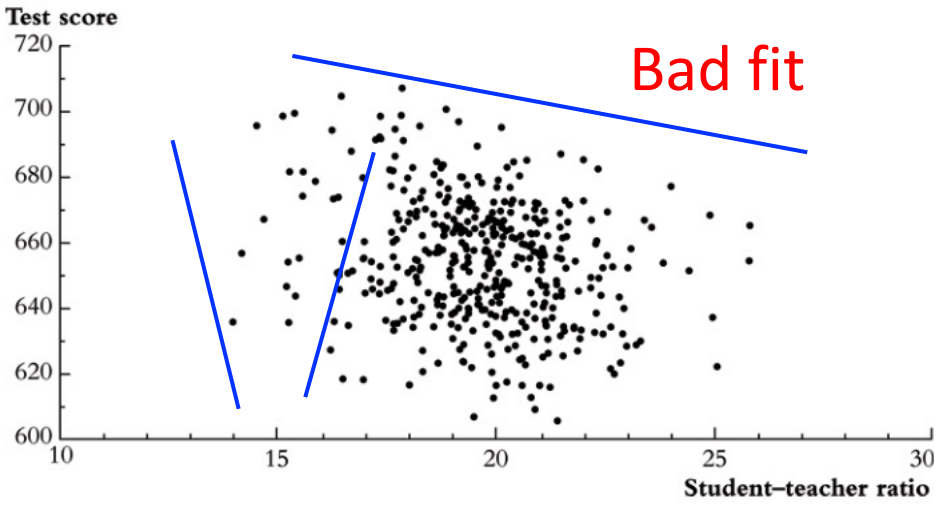
# The linear regression model

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

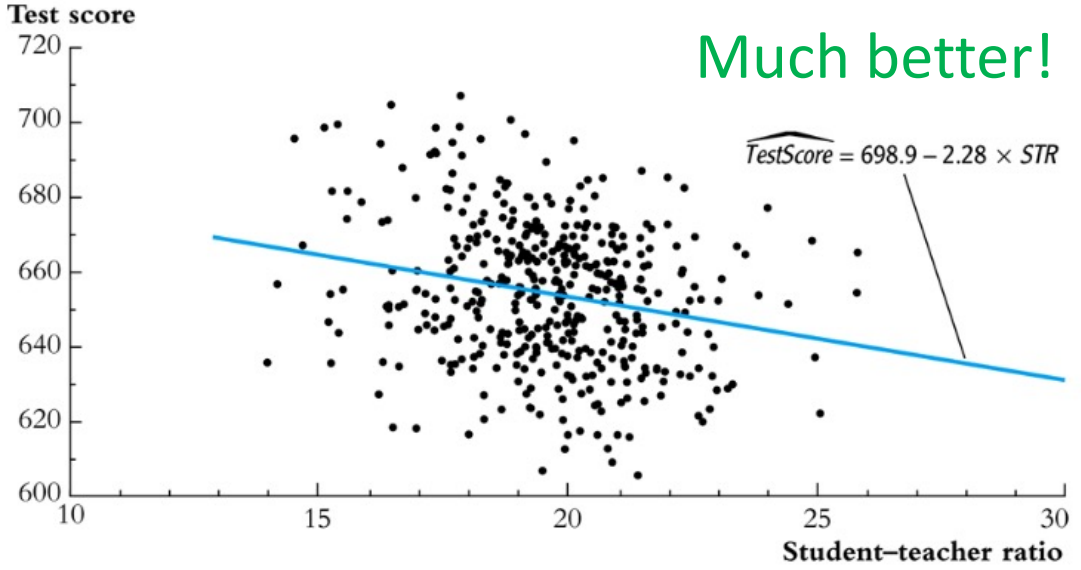
- Y = dependent variable
- X = independent variable (or regressor)
- $\beta_0 + \beta_1 X_i$ = population regression function
- $\beta_0$ = population intercept
- $\beta_1$ = population slope (= change in E(Y) for a unit increase in X)
- $u_i$ = population error term

# Estimating the linear regression model

- We can *estimate* $\beta_0$ and $\beta_1$ from a sample.
- Choose $\widehat{\beta_0}$ & $\widehat{\beta_1}$ to *best fit* the data.




Note: Data from 420 CA school districts



# The OLS estimator

- "Best fit" = minimize (squared) prediction mistakes:

$$\min_{b_0,\, b_1} \sum_{i=1}^{n} (Y_i - [b_0 + b_1 X_i])^2$$

- The $\hat{\beta}_0$ & $\hat{\beta}_1$ that minimize this are called Ordinary Least Squares (OLS) estimators:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{s_{XY}}{s_X^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

Unbiased estimators of the population regression coefficients $\beta_0$ & $\beta_1$

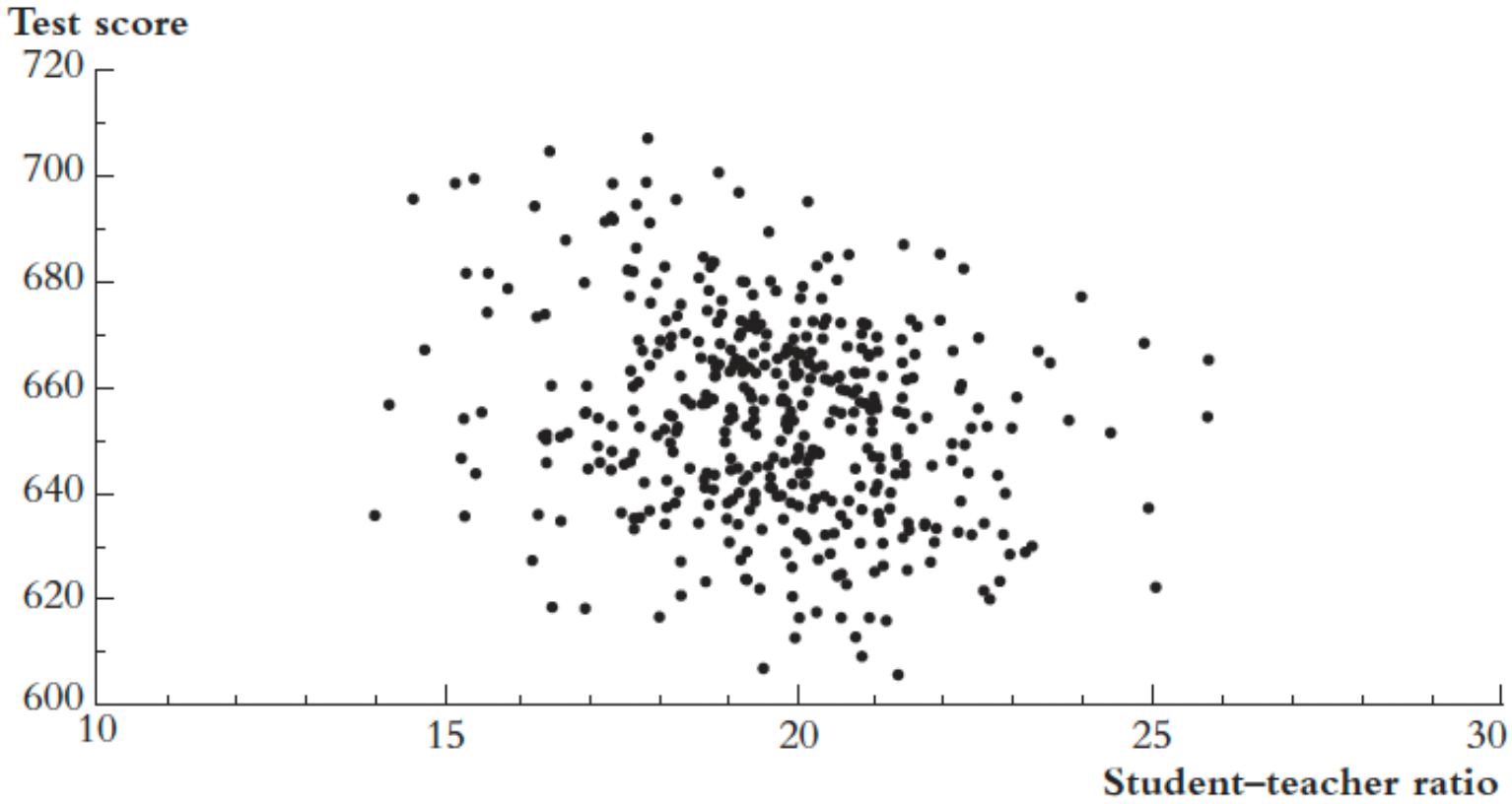
# The OLS estimator

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{u}_i$$

- Linear regression model…
- …but with sample OLS coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ as estimators of population coefficients $\beta_0$ and $\beta_1$.
- $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ = predicted value of $Y_i$ based on $X_i$
- $\hat{u}_i$ = regression residual (estimator of population error term $u_i$)

# Application: class size & test scores

- Dataset from 420 California school districts.
- *Test scores* = average score from a standardized test.
- *Class size* = Average students/teacher ratio.
- *Is larger class size associated with lower test scores?*



# OLS regression IN STATA

```
regress testscr str, robust

Regression with robust standard errors                Number of obs =     420
                                                      F(  1,   418) =   19.26
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.0512
                                                      Root MSE      =  18.581
------------------------------------------------------------------------------
             |               Robust
    testscr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         str |  -2.279808   .5194892    -4.39   0.000    -3.300945   -1.258671
       _cons |   698.933    10.36436    67.44   0.000     678.5602    719.3057
------------------------------------------------------------------------------
```

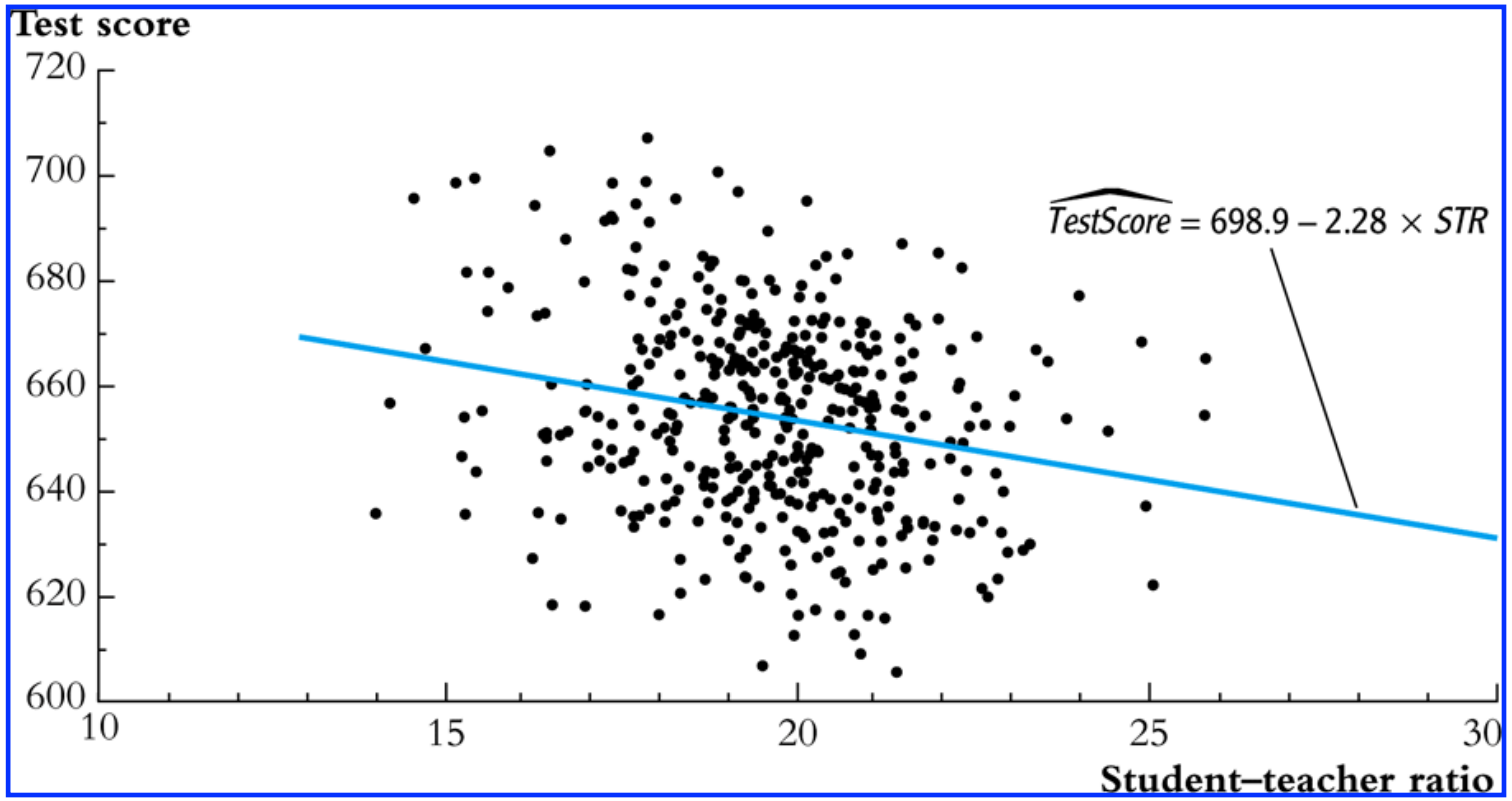
Estimated slope ($\hat{\beta}_1$)

Estimated constant ($\hat{\beta}_0$)

$$\widehat{TestScore} = 698.9 - 2.28 \times STR$$

# OLS APPLICATION:
# CLASS SIZE & TEST SCORES

$$\widehat{TestScore} = 698.9 - 2.28 \times STR$$




**In STATA:**

graph twoway (scatter testscr str) (lfit testscr str)

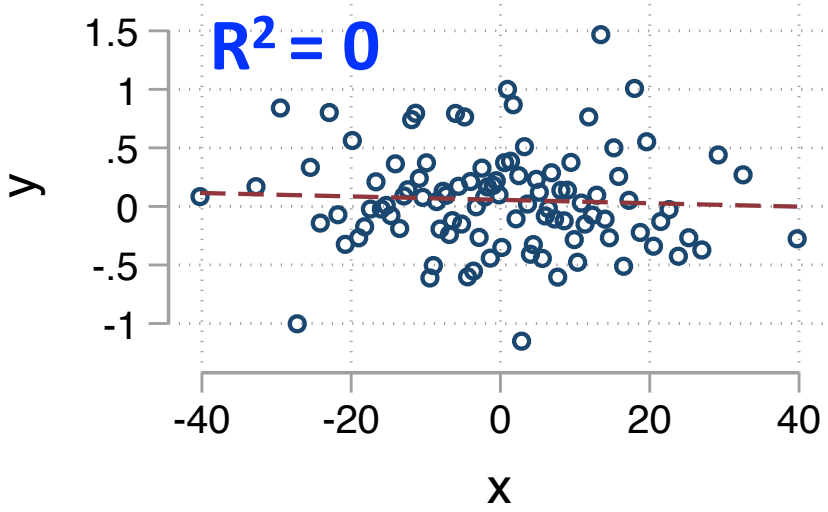
# OLS application: class size & test scores

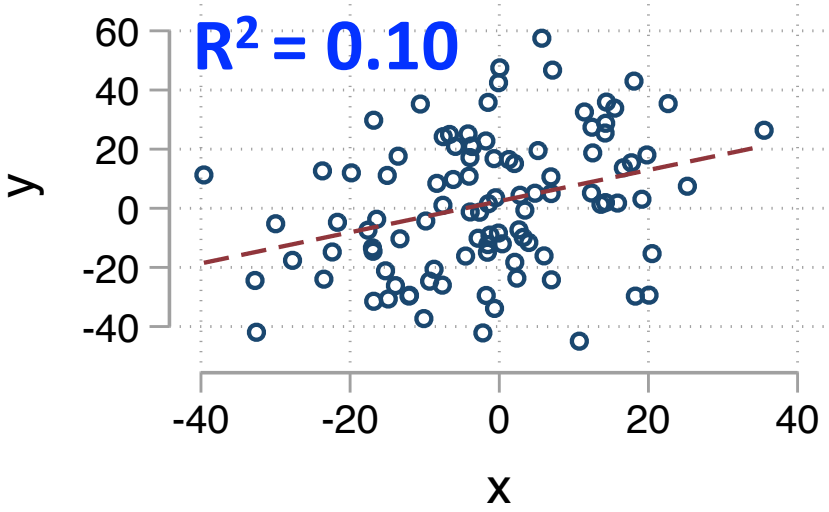
$$\widehat{TestScore} = 698.9 - 2.28 \times STR$$

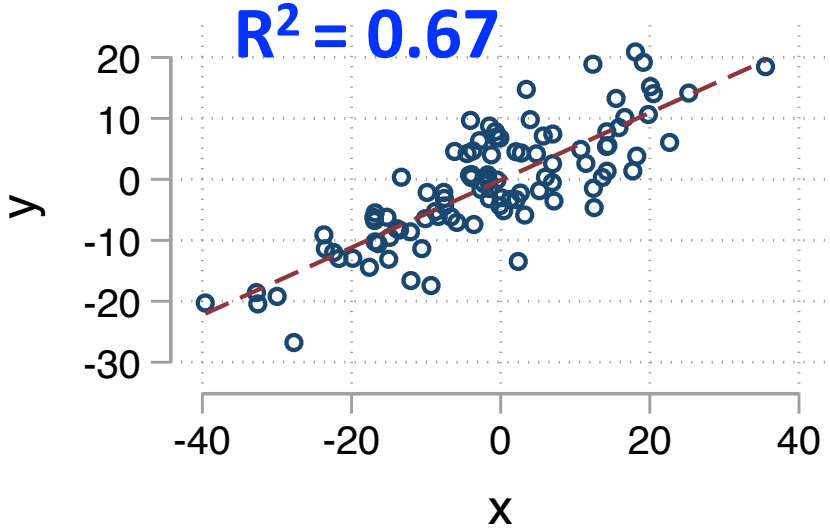
- Is the estimated slope -2.28 large or small?
  - *1 unit increase in STR decreases expected test scores by 2.28 points.*
- To answer, we need descriptive statistics on class size and test scores.
  - *STR* has mean of 19.6 and SD 1.9
  - *TestScore* has a mean of 654.2 and SD of 19.1
- So the slope is pretty small (flat regression line)

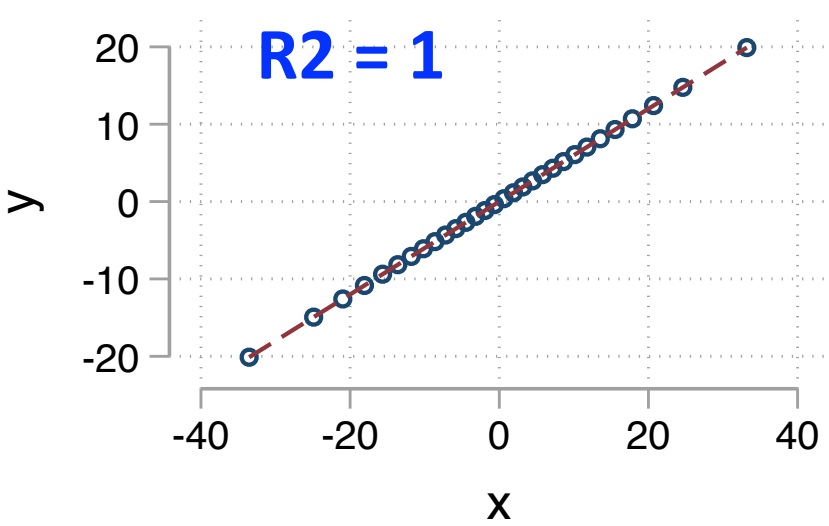
# The $R^2$

- How well does our estimated regression line fit the data?
- $R^2$: the fraction of variation in $Y_i$ explained by the regression.
- $Y_i = \hat{Y}_i + \hat{u}_i = \text{OLS prediction} + \text{OLS residual}$
- $\text{var}(Y_i) = \text{var}(\hat{Y}_i) + \text{var}(\hat{u}_i)$
- $R^2 = \frac{var(\hat{Y}_i)}{var(\hat{Y}_i)+var(\hat{u}_i)} = \frac{var(\hat{Y}_i)}{var(Y_i)} = \frac{ESS}{TSS}$

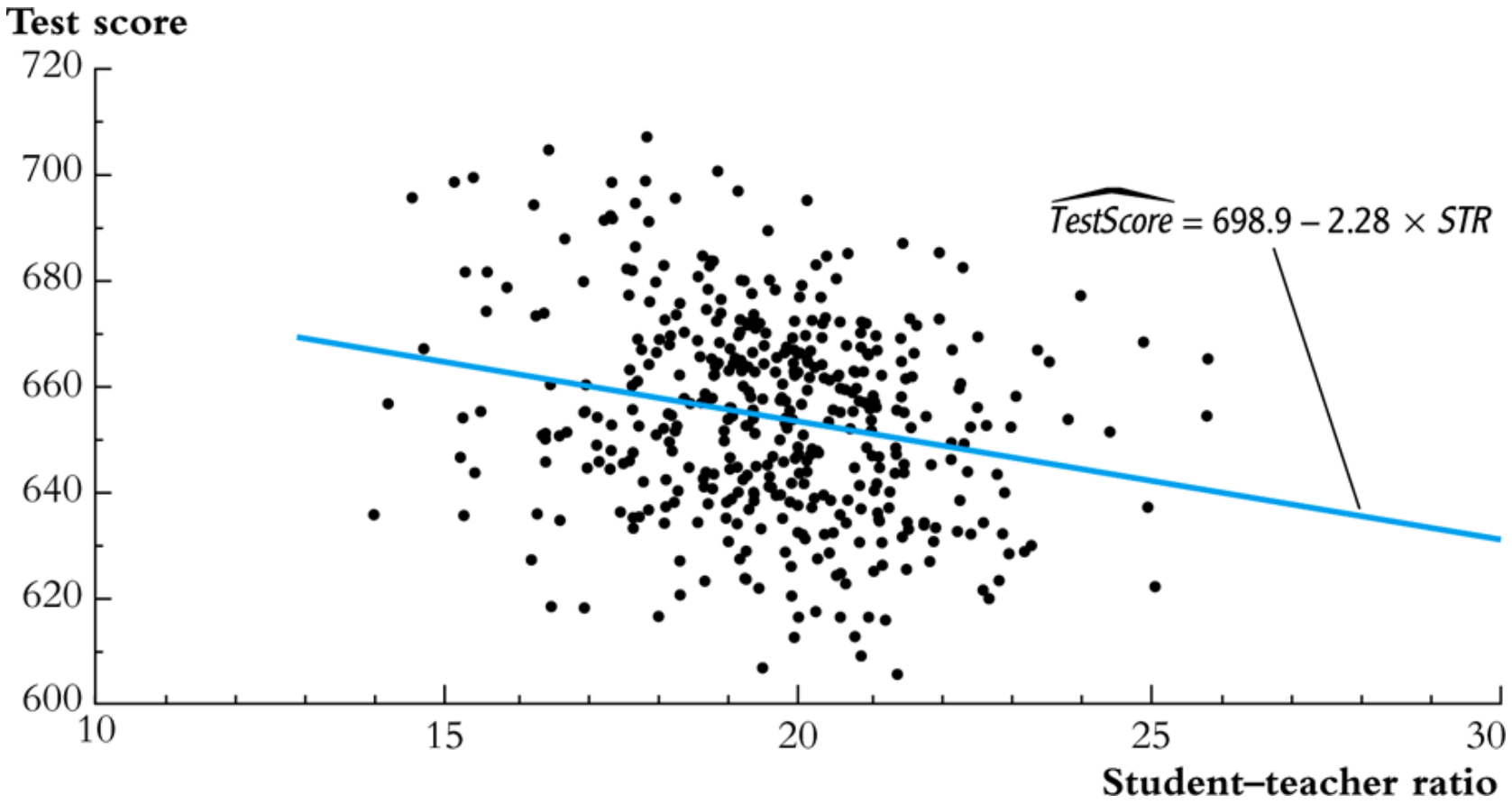

$R^2 = 0$
1.5
1
.5
0
-.5
-1
y
-40
-20
0
20
40
x


$R^2 = 0.10$
60
40
20
0
-20
-40
y
-40
-20
0
20
40
x


$R^2 = 0.67$
20
10
0
-10
-20
-30
y
-40
-20
0
20
40
x


R2 = 1
20
10
0
-10
-20
y
-40
-20
0
20
40
x

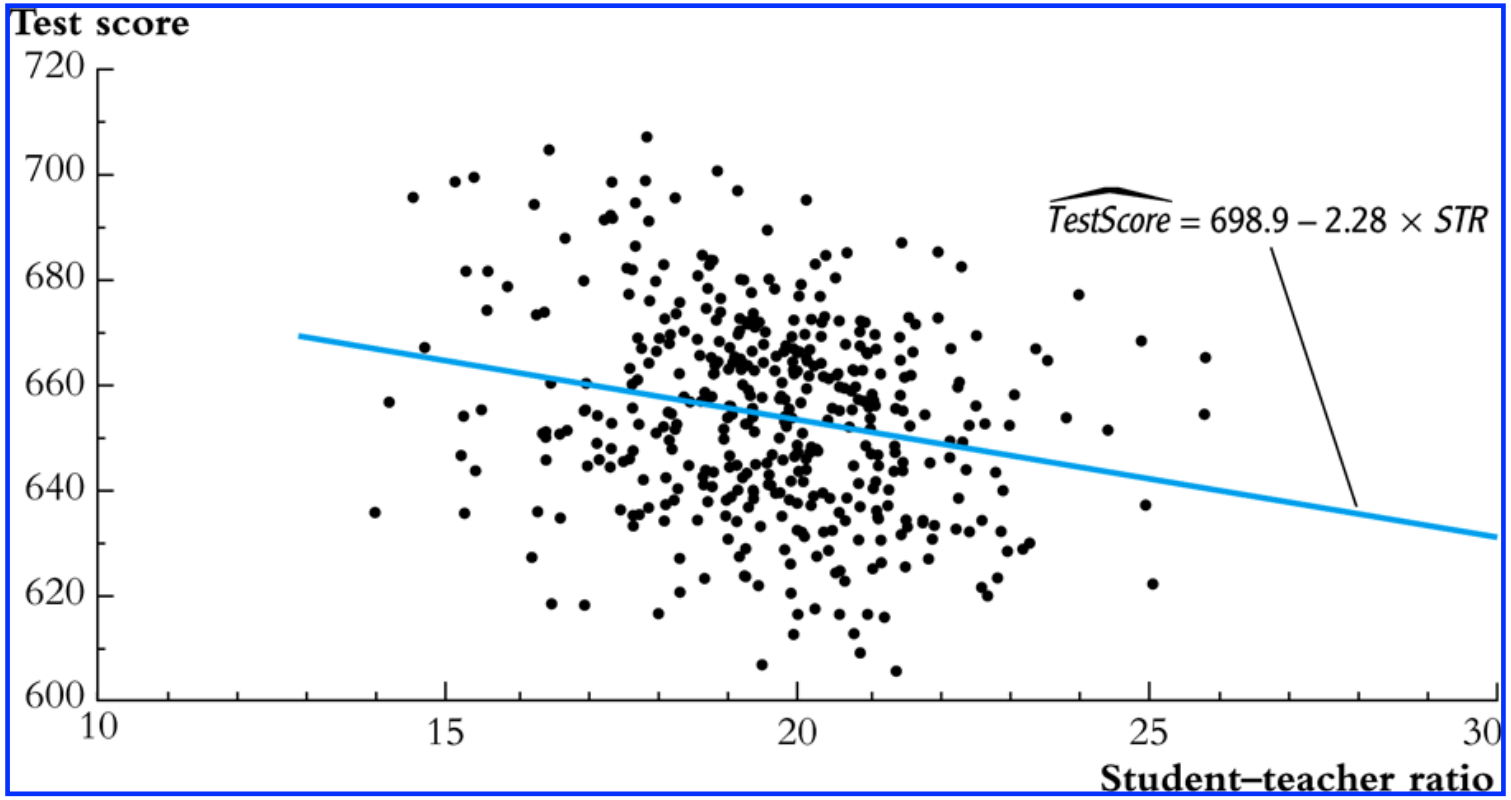
# TEST SCORES EXAMPLE




- Do you expect $R^2$ to be high or low?
- $R^2$= 0.05
- Pretty low

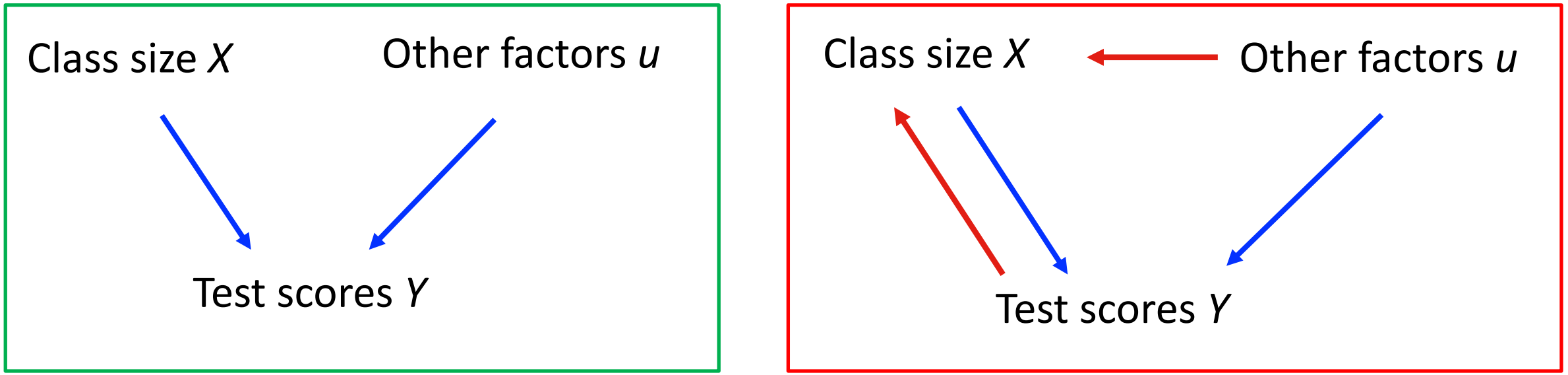
# Regression and causality

$$\widehat{TestScore} = 698.9 - 2.28 \times STR$$




- *Causal effect* of class size?
- Or captures something else?

# Causal relation between class size & test scores




- When one of the two red connections (or both) are present, the OLS coefficient is not guaranteed to capture a causal effect.

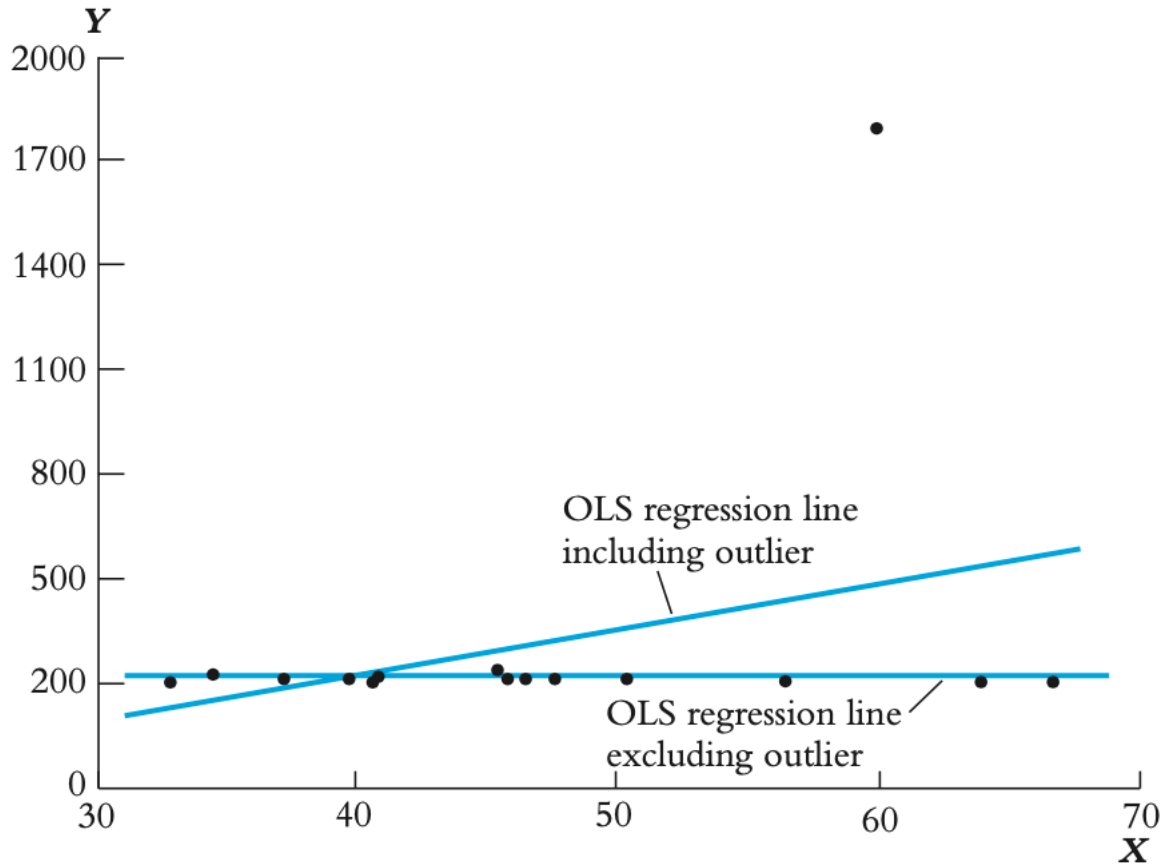
# Regression and causality

- When does $\beta_1$ measure the average *causal effect* of X on Y?
- X must be independent of other factors affecting Y
  → X must be independent of error term $u_i$ → $corr(X_i, u_i) = 0$
- True with experimental data
- Not always true with observational data

# 3 OLS ASSUMPTIONS FOR CAUSAL INFERENCE

1. The independent variable *X* is independent of the error term $u_i$.
   - $E(u_i|X = x) = 0; \quad corr(X_i, u_i) = 0$
2. $(X_i, Y_i)$, $i = 1, \ldots, n$, are i.i.d..
   - Meaning: the sample is random!
3. Large outliers in *X* or *Y* are rare.
   - Outliers can drive the OLS estimate of $\beta_1$ astray

# *OLS can be sensitive to an outlier:*




- *Is the lone point an outlier in X or Y?*
- Often data glitches.
- Or units with very specific characteristics that set them apart.

# Thank you for your attention