4 – LINEAR REGRESSION I
ONE REGRESSOR

University of Massachusetts Amherst
BE REVOLUTIONARY™

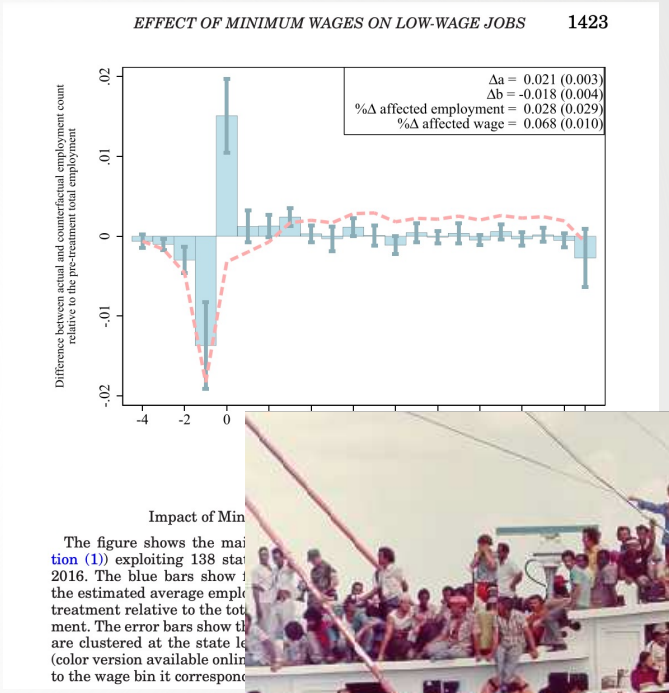
SECTION 4 – LINEAR REGRESSION, PART 1

THE PLAN

1. The Linear Regression Model
2. Estimation of the Linear Regression Model
3. Measures of Fit: R^2 and SER
4. Regression and Causality
5. Sampling Distribution of OLS Estimators
6. Hypothesis Tests
7. Confidence Intervals
8. Regression when X is a Binary Variable

LINEAR REGRESSION: OVERVIEW

- Does a minimum wage decrease employment?
- Do increases in government spending boost or harm growth?
- Does immigration lower wages for native workers?
- Is a recession likely in the next year?
- ….

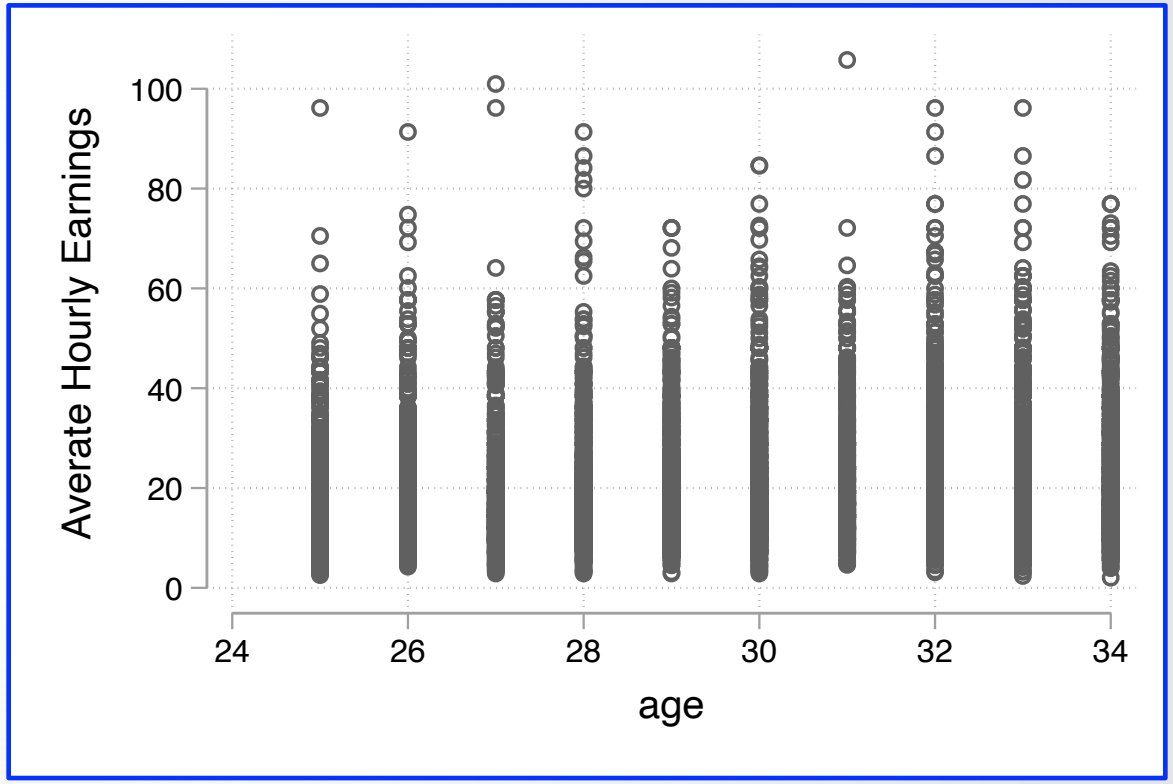


4.1 THE LINEAR REGRESSION MODEL

University of Massachusetts Amherst

CONDITIONAL EXPECTATIONS

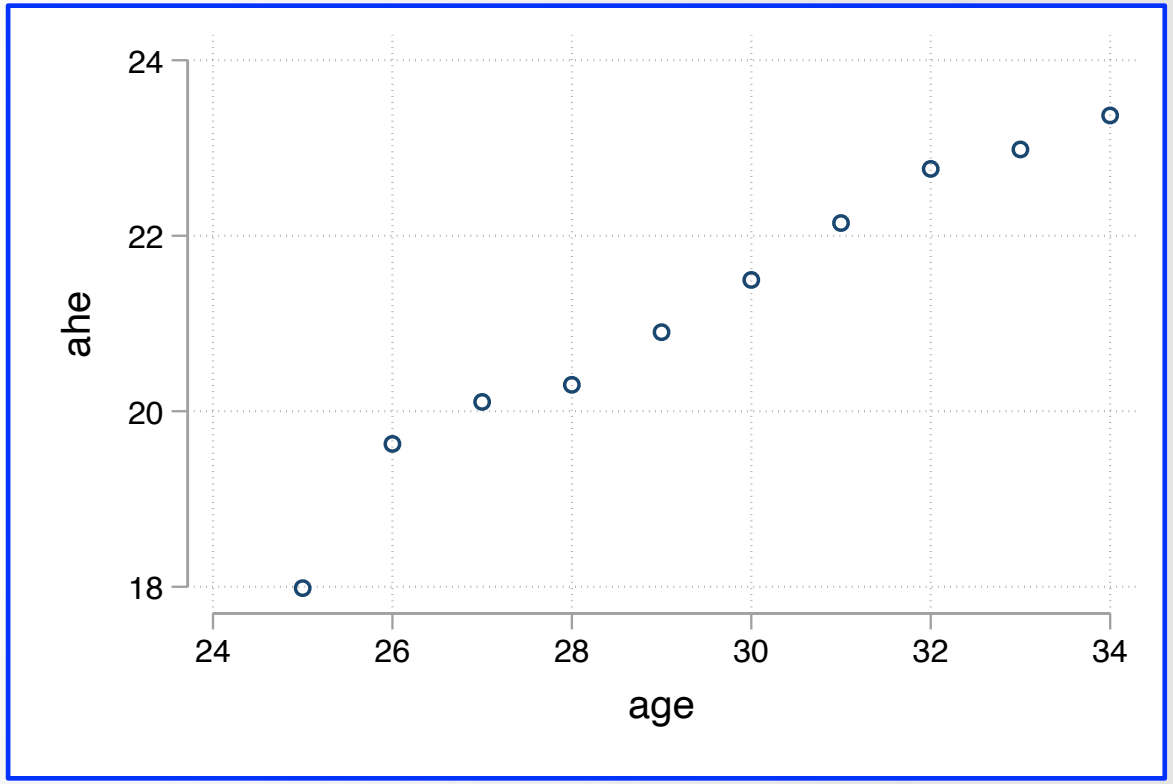
- How do average earnings vary with age?
- Conditional Expectation: $E(AHE|AGE = x)$
- Scatterplot of CPS data
  - STATA: "scatter ahe age"
- Imagine it was the population, not a sample.



Source: Current Population Survey, March 2015 edition

CONDITIONAL EXPECTATIONS

- Calculate 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.

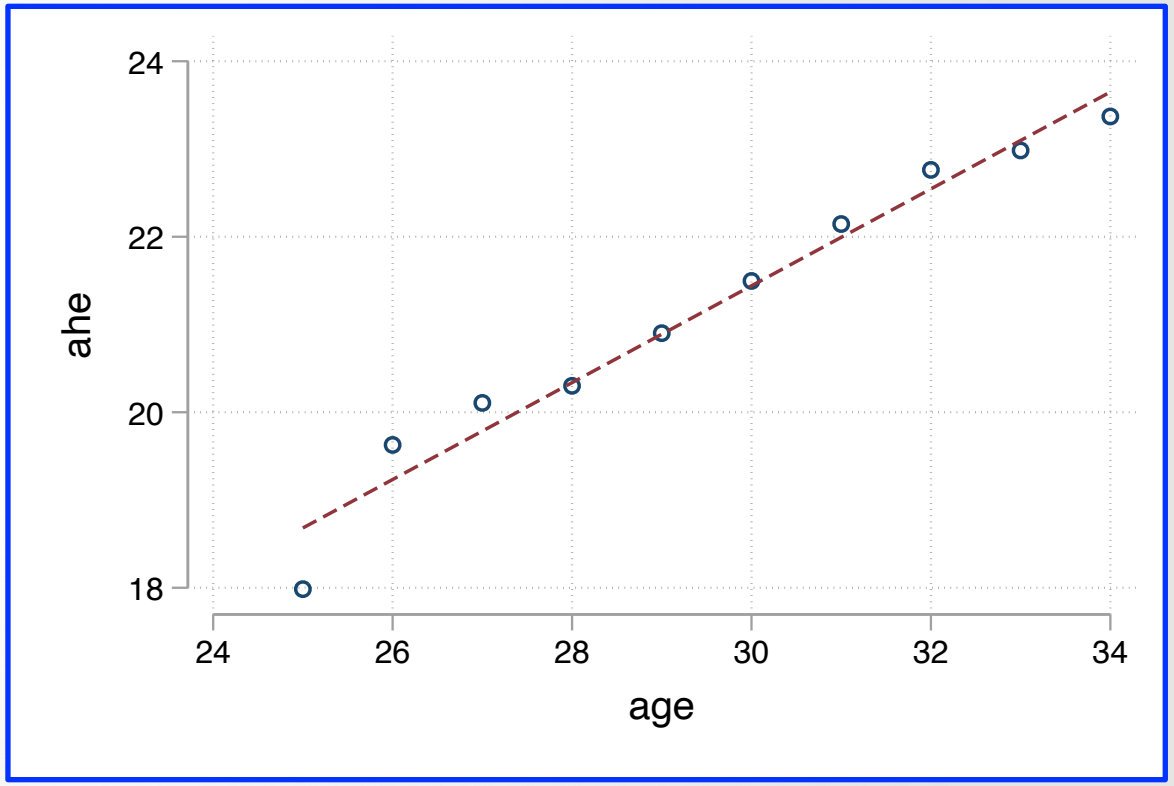


CONDITIONAL EXPECTATIONS

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

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

- Works with continuous X.
- β_1 gives an "average effect"
- Can be used to *predict* earnings based on age.



CONDITIONAL EXPECTATIONS

- A continuous variable:
  GDP growth vs. trade openness
- Again, assume population data.
- Linear conditional expectation:

$$E(growth|trade) = \beta_0 + \beta_1 trade$$

- For an individual country *i*:

$$growth_i = E(growth|trade_i) + u_i$$

$$= \beta_0 + \beta_1 trade_i + u_i$$

Source: Beck & Loayza (2000) "Finance and the Sources of Growth", Journal of Financial Economics

THE LINEAR REGRESSION MODEL

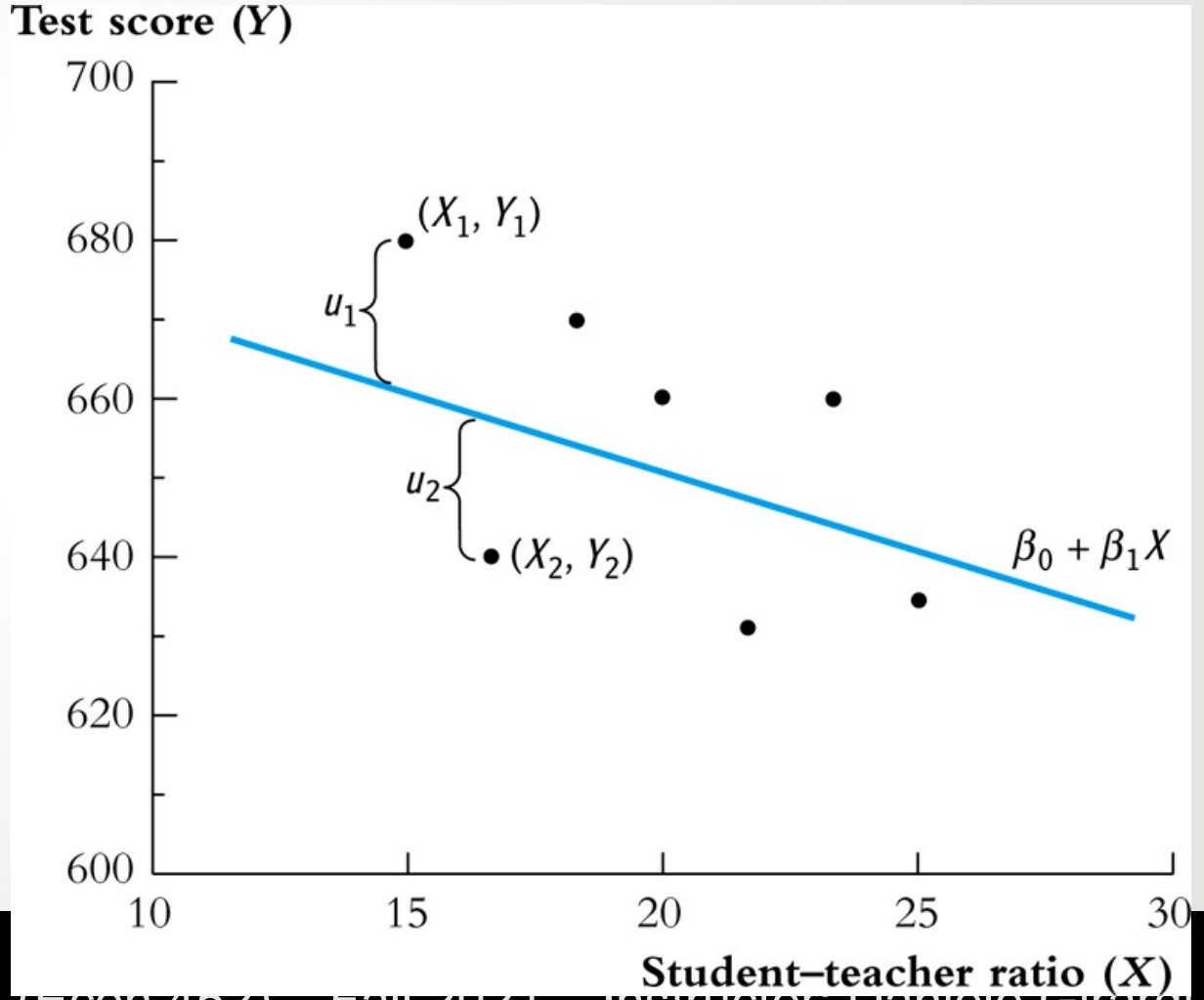
$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

- Y = dependent variable
- X = independent variable (or regressor)
- $\beta_0 + \beta_1 X_i$ = population regression function
- β_0 = population intercept
- β_1 = population slope
- u_i = population error term

TEST SCORES VS STUDENT-TEACHER RATIO (HYPOTHETICAL DATA)

$$E(TestScore) = \beta_0 + \beta_1 STR$$

$$TestScore_i = \beta_0 + \beta_1 STR_i + u_i$$

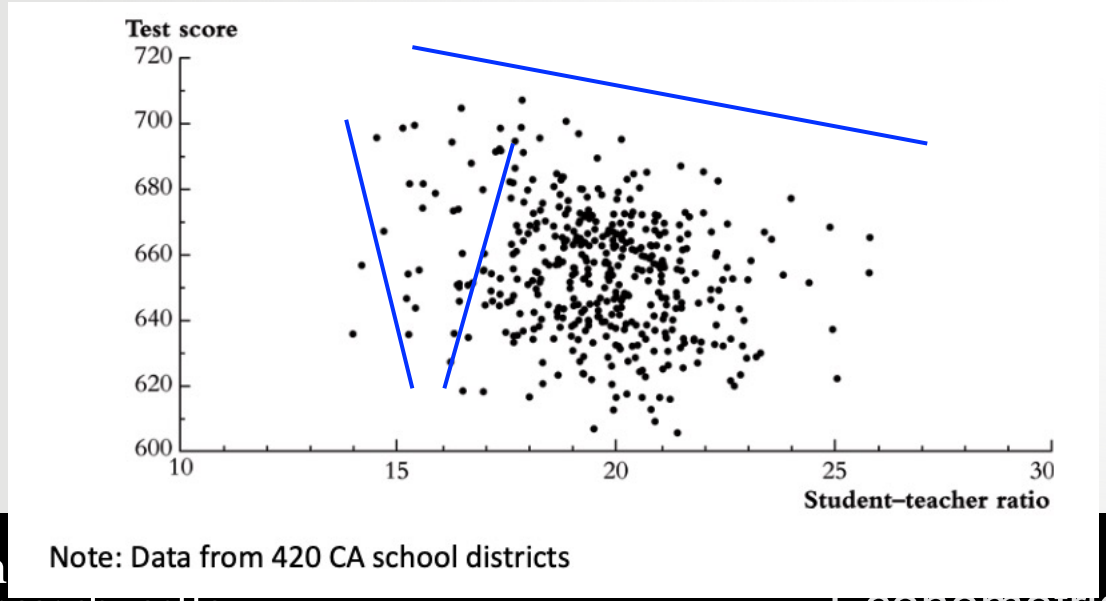


4.2 ESTIMATION OF THE LINEAR REGRESSION MODEL

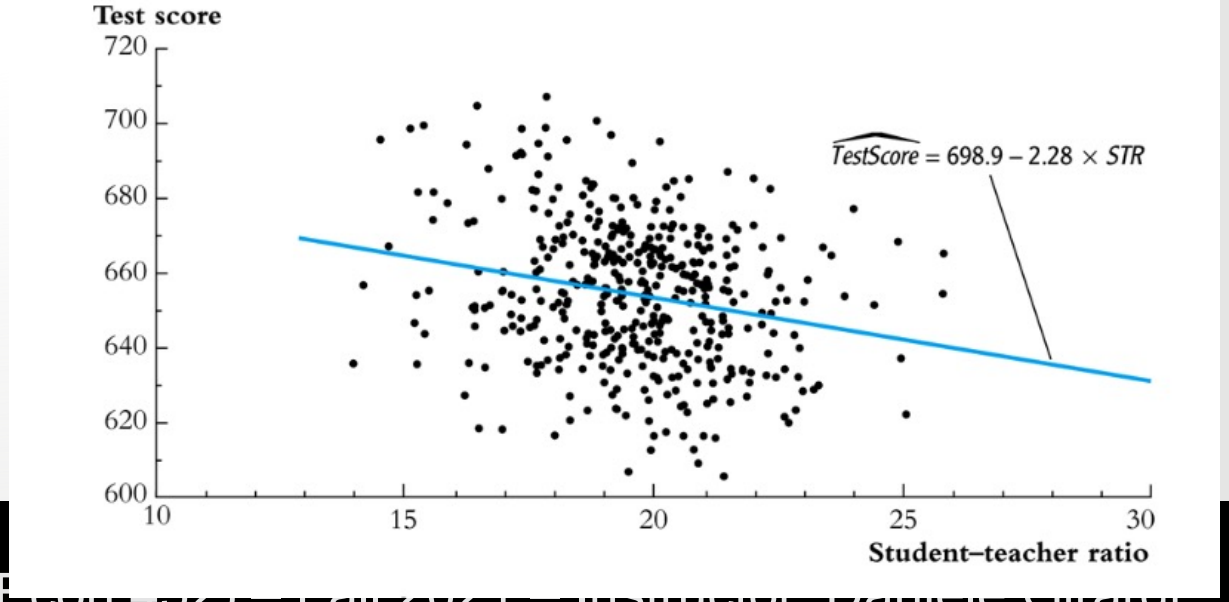
University of Massachusetts Amherst

ESTIMATING THE LINEAR REGRESSION MODEL

- We can *estimate* β_0 and β_1 from a sample.
- Choose β_0 & β_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 solution gives the Ordinary Least Squares (OLS) estimators:

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

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

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 β_0 and β_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 error term u_i)

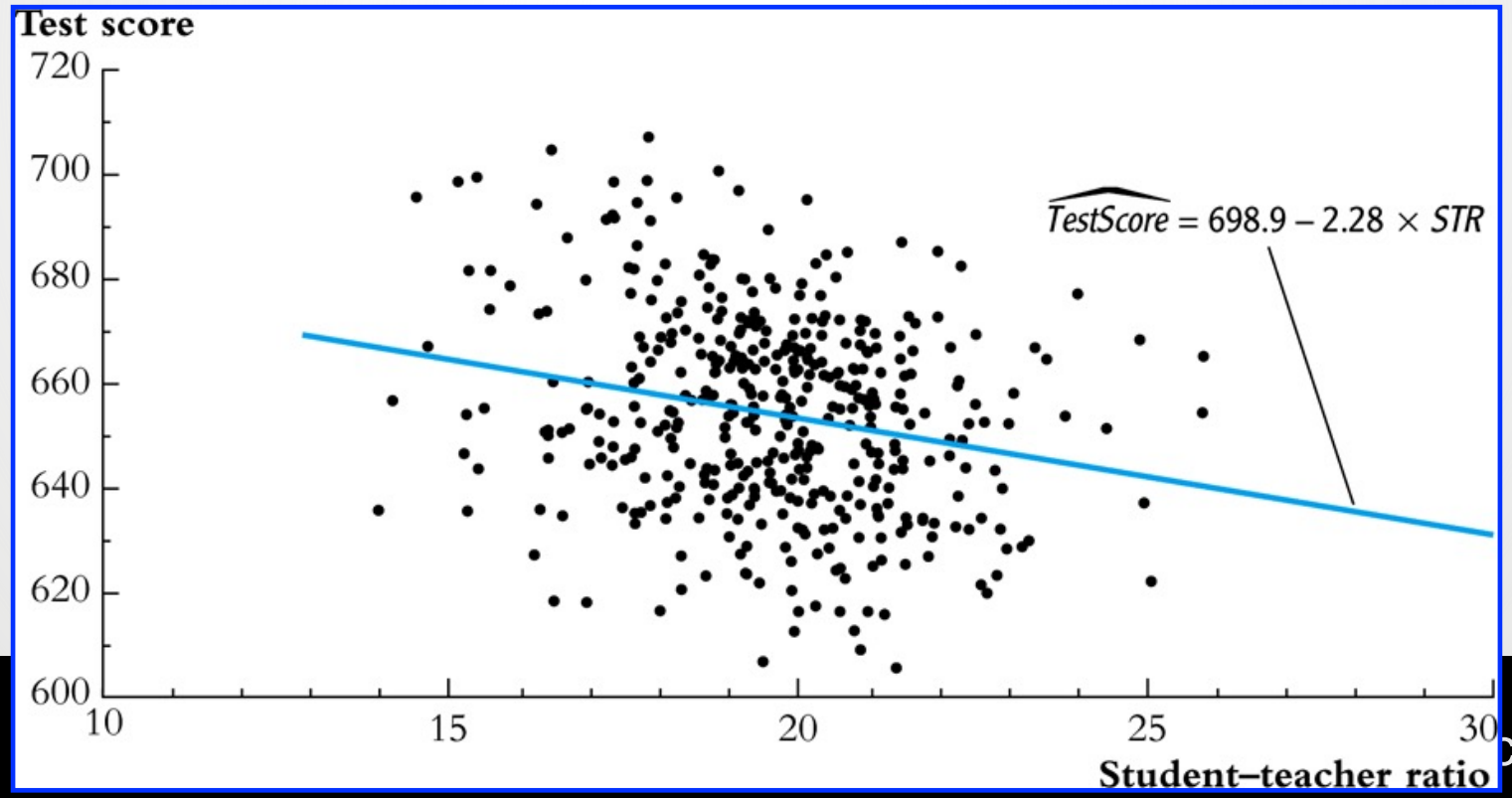
OLS REGRESSION IN STATA

```
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
------------------------------------------------------------------------------
```

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

OLS APPLICATION: CLASS SIZE & TEST SCORES

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



OLS APPLICATION: CLASS SIZE & TEST SCORES

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

- Is the estimated slope of -2.28 large or small?

Relatively small!

TABLE 4.1 Summary of the Distribution of Student–Teacher Ratios and Fifth-Grade Test Scores for 420 K–8 Districts in California in 1999

			Percentile						
	Average	Standard Deviation	10%	25%	40%	50% (median)	60%	75%	90%
Student–teacher ratio	19.6	1.9	17.3	18.6	19.3	19.7	20.1	20.9	21.9
Test score	654.2	19.1	630.4	640.0	649.1	654.5	659.4	666.7	679.1

4.3 MEASURES OF FIT: R^2 AND SER

University of Massachusetts Amherst

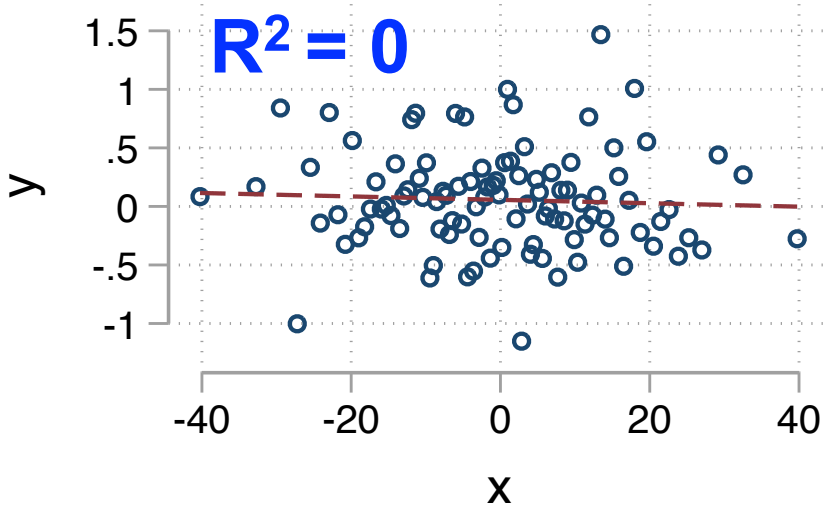
MEASURES OF FIT

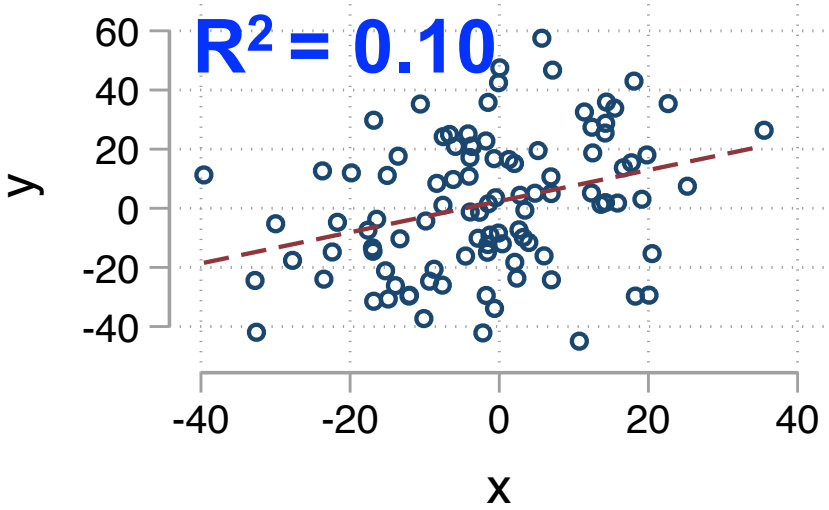
How well does our regression line fit the data?

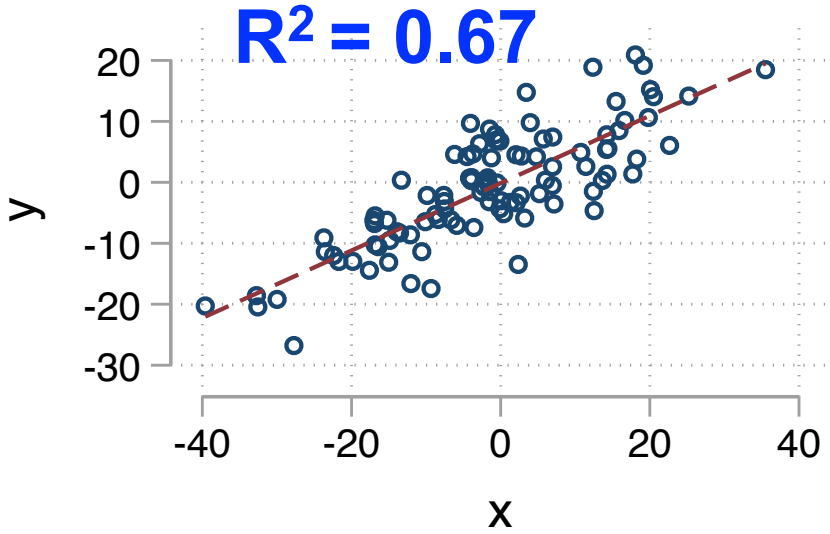
- $\mathbf{R^2}$
- **SER**: Standard Error of the Regression

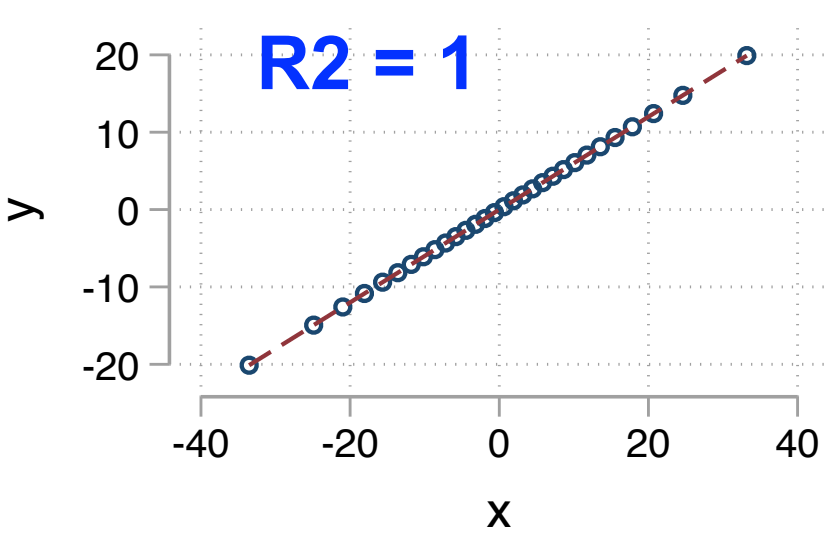
THE R^2

- $Y_i = \hat{Y}_i + \hat{u}_i$ = OLS prediction + OLS residual
- $\text{var}(Y_i) = \text{var}(\hat{Y}_i) + \text{var}(\hat{u}_i)$
- $R^2 = \dfrac{var(\hat{Y}_i)}{var(\hat{Y}_i)+var(\hat{u}_i)} = \dfrac{var(\hat{Y}_i)}{var(Y_i)} = \dfrac{\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2}{\sum_{i=1}^{n}(Y_i-\bar{Y})^2} = \dfrac{ESS}{TSS}$


$R^2 = 0$
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

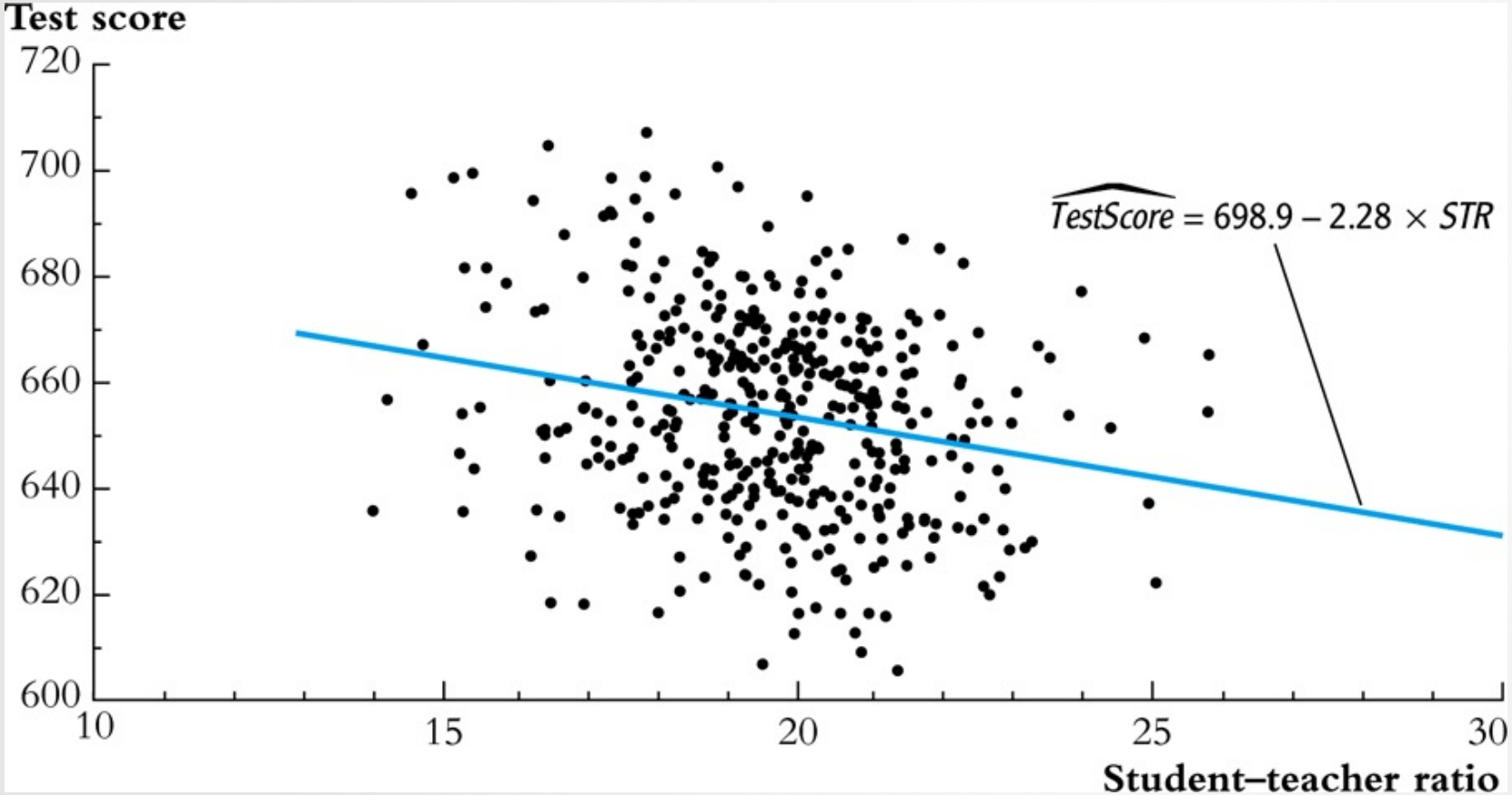
SER

- Standard Error of the Regression (SER):

$$SER = s_{\hat{u}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}\hat{u}_i^2}$$

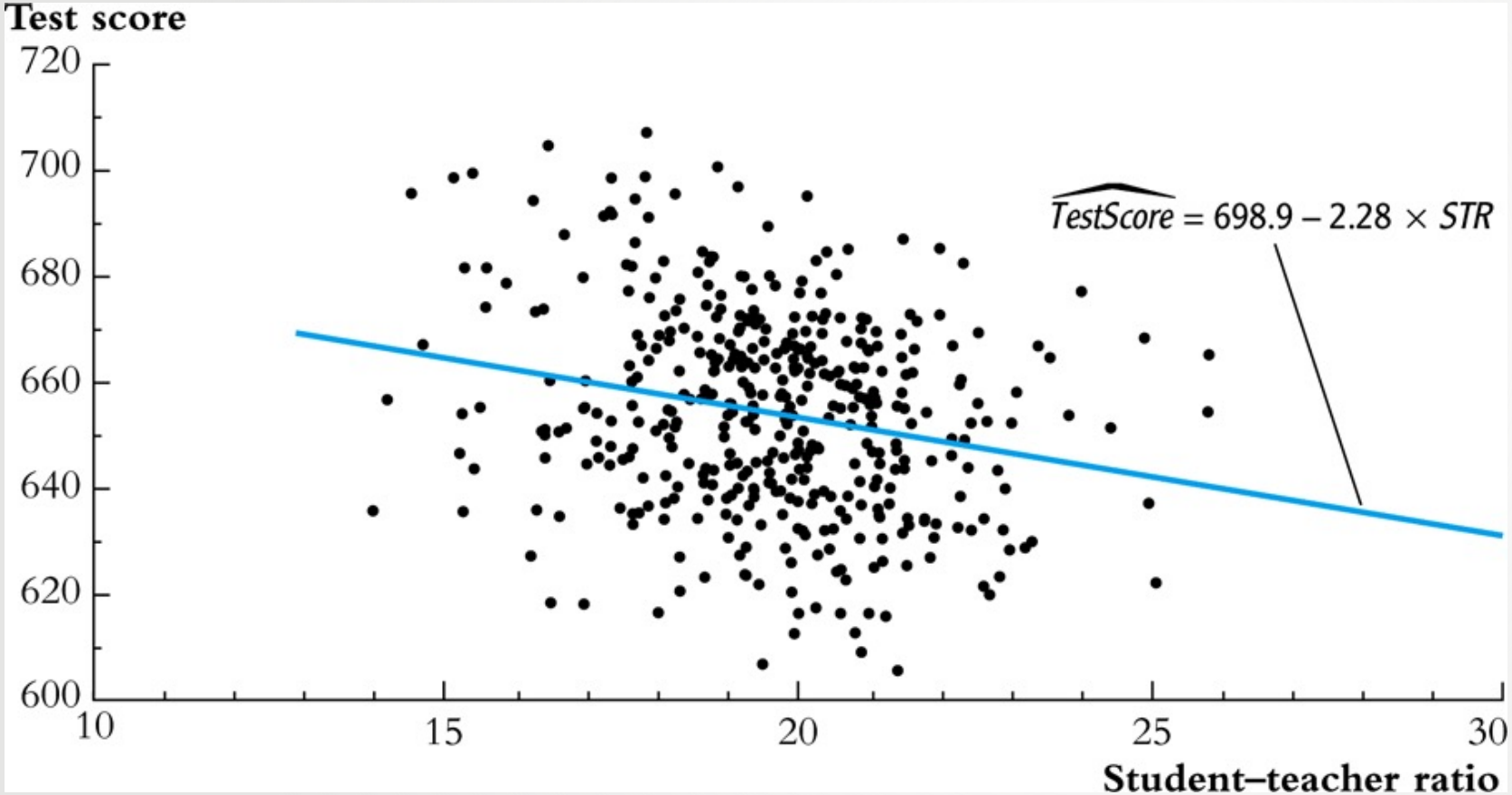
- Measures the “spread” of residuals around the regression lines.

TEST SCORES EXAMPLE



- Do you expect R^2 to be high or low?
- And the SER?

TEST SCORES EXAMPLE



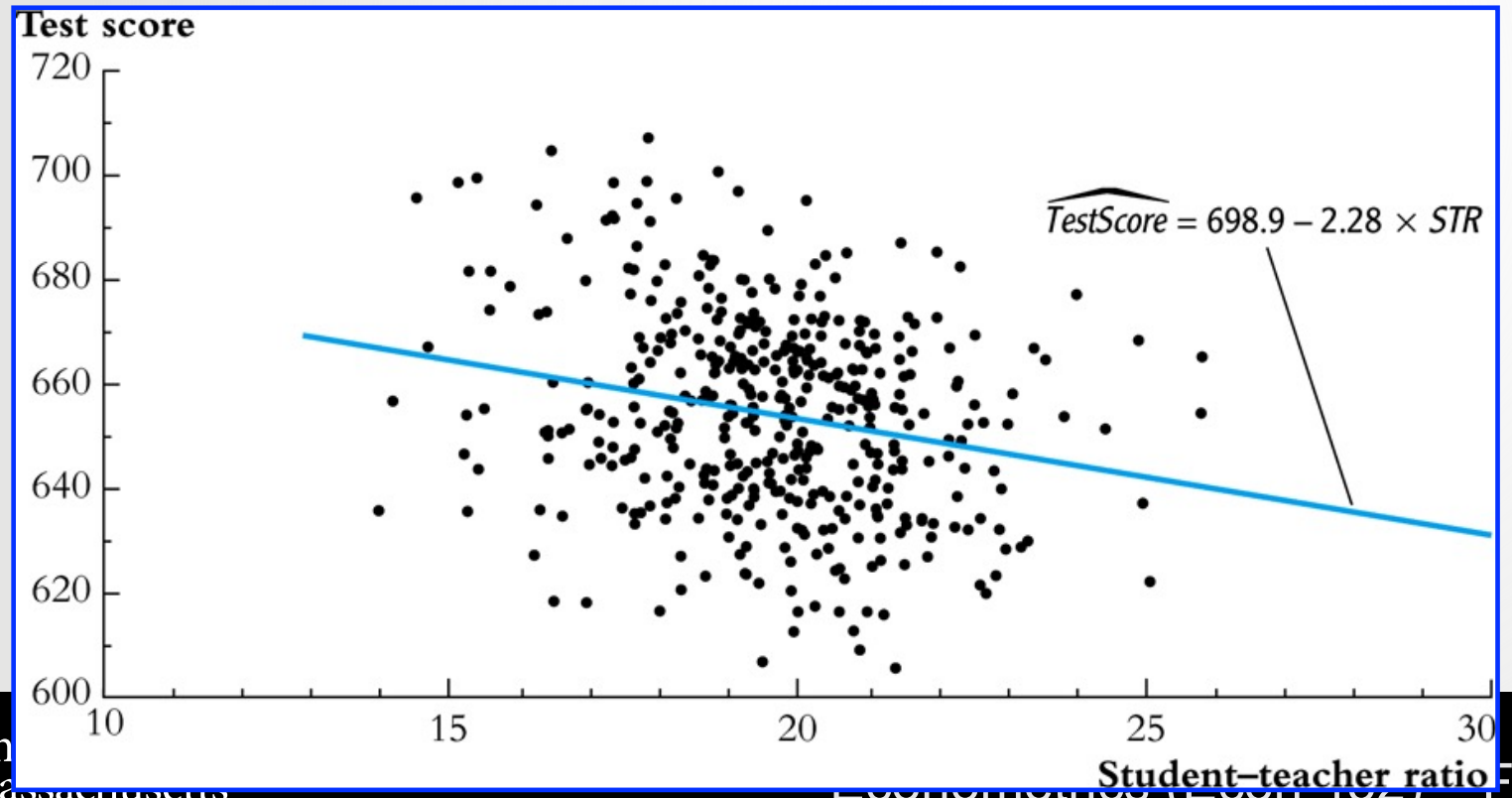
- R^2= 0.05
- SER=18.6

4.4 REGRESSION AND CAUSALITY

University of Massachusetts Amherst

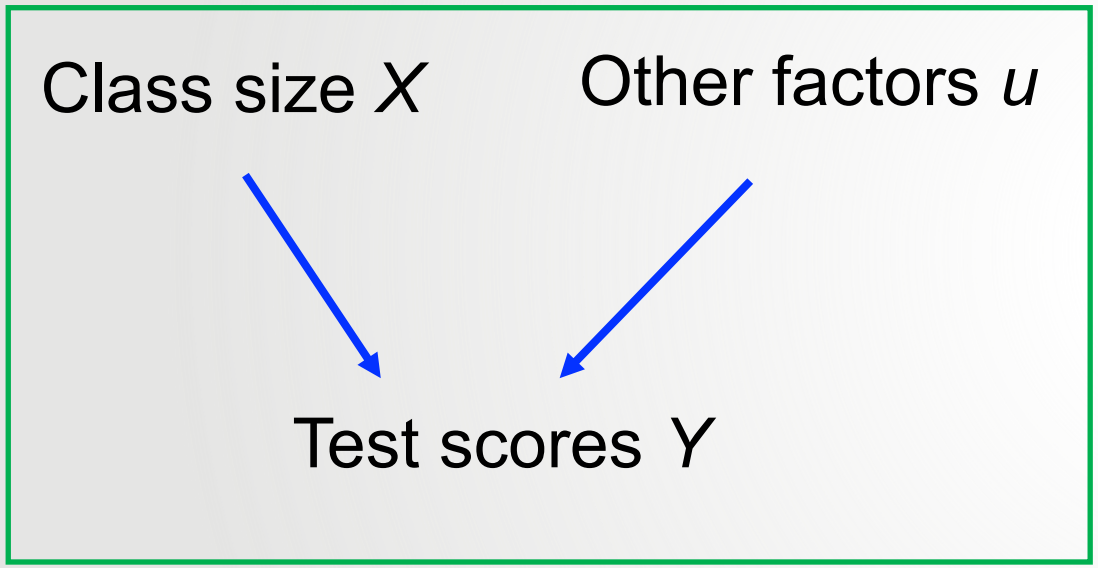
CLASS SIZE & TEST SCORES AGAIN

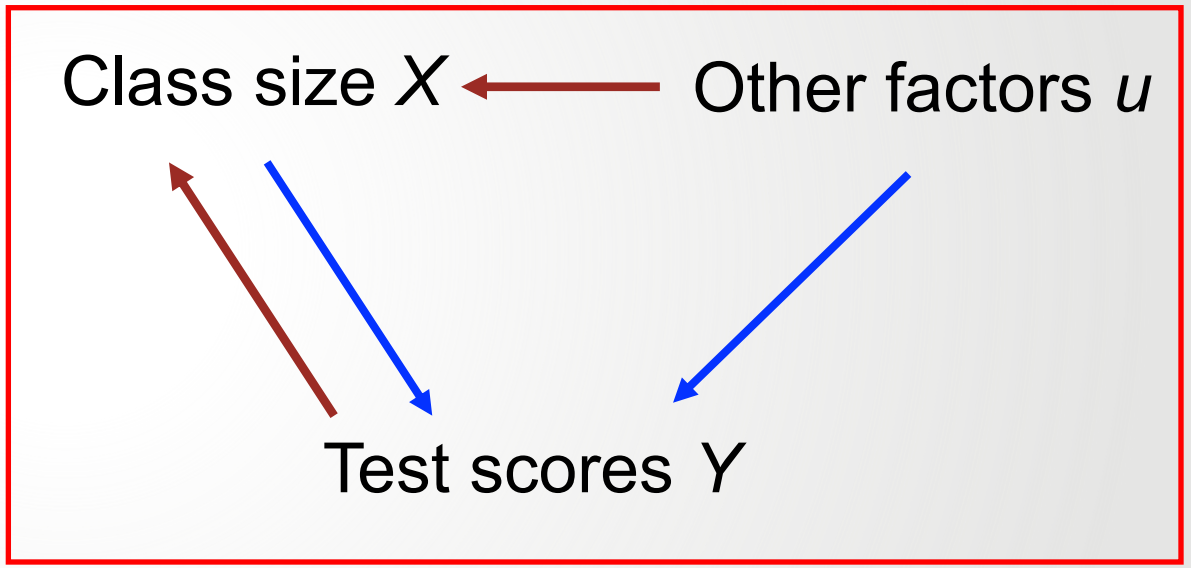
$$\widehat{TestScore} = 698.9 - 2.28 \times STR$$



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

CAUSAL RELATIONS 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 β_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

EXAMPLE: THE EFFECT OF EDUCATION ON EARNINGS

- To study the effect of formal education on earnings, you estimate:

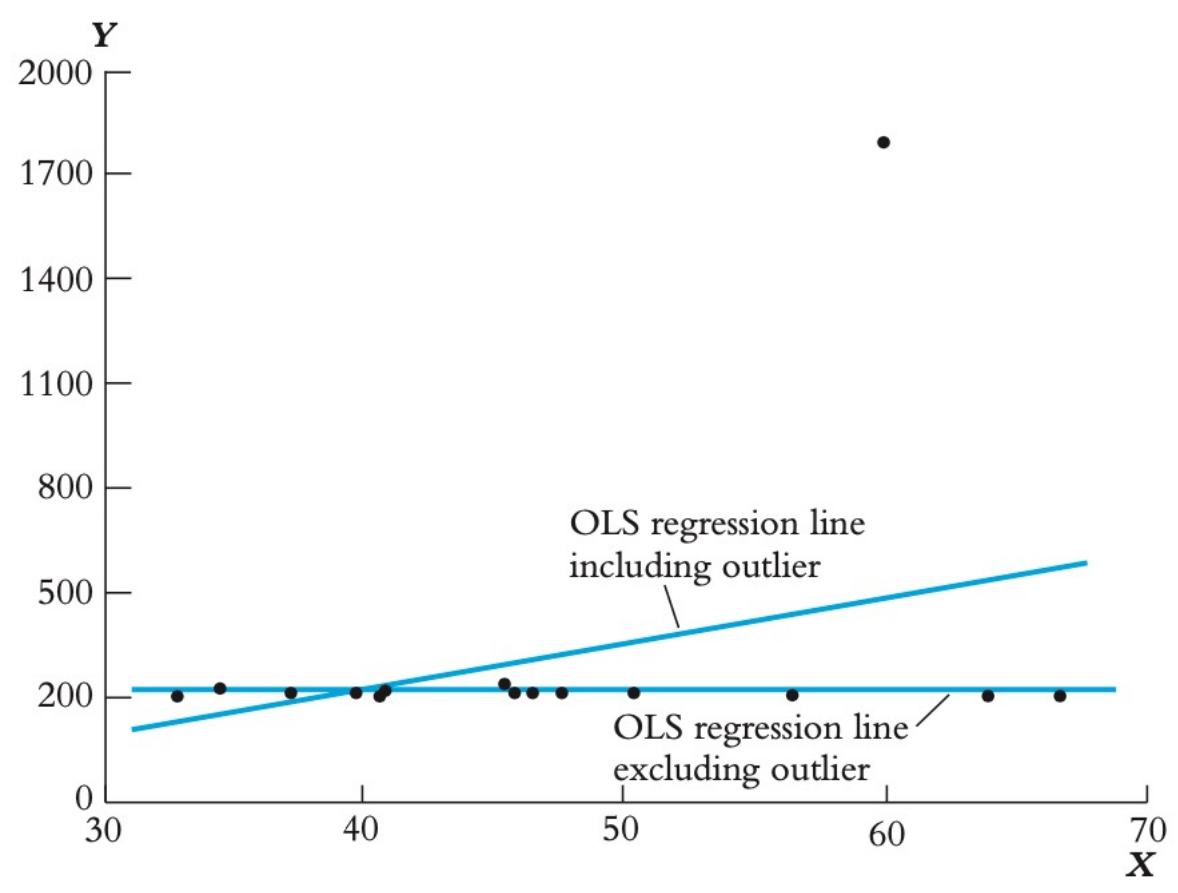
$$AHE_i = \beta_0 + \beta_1 Educ_i + u_i$$

- Where *Educ* = years of completed education.
- What could cause correlation between $Educ$ and u_i?

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..
   - BUT some violations of independence can be dealt with (time-series and panel data).
3. Large outliers in *X* and/or *Y* are rare.
   - Outliers can drive the OLS estimate of β_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.

4.5 SAMPLING DISTRIBUTION OF THE OLS ESTIMATOR

University of Massachusetts Amherst

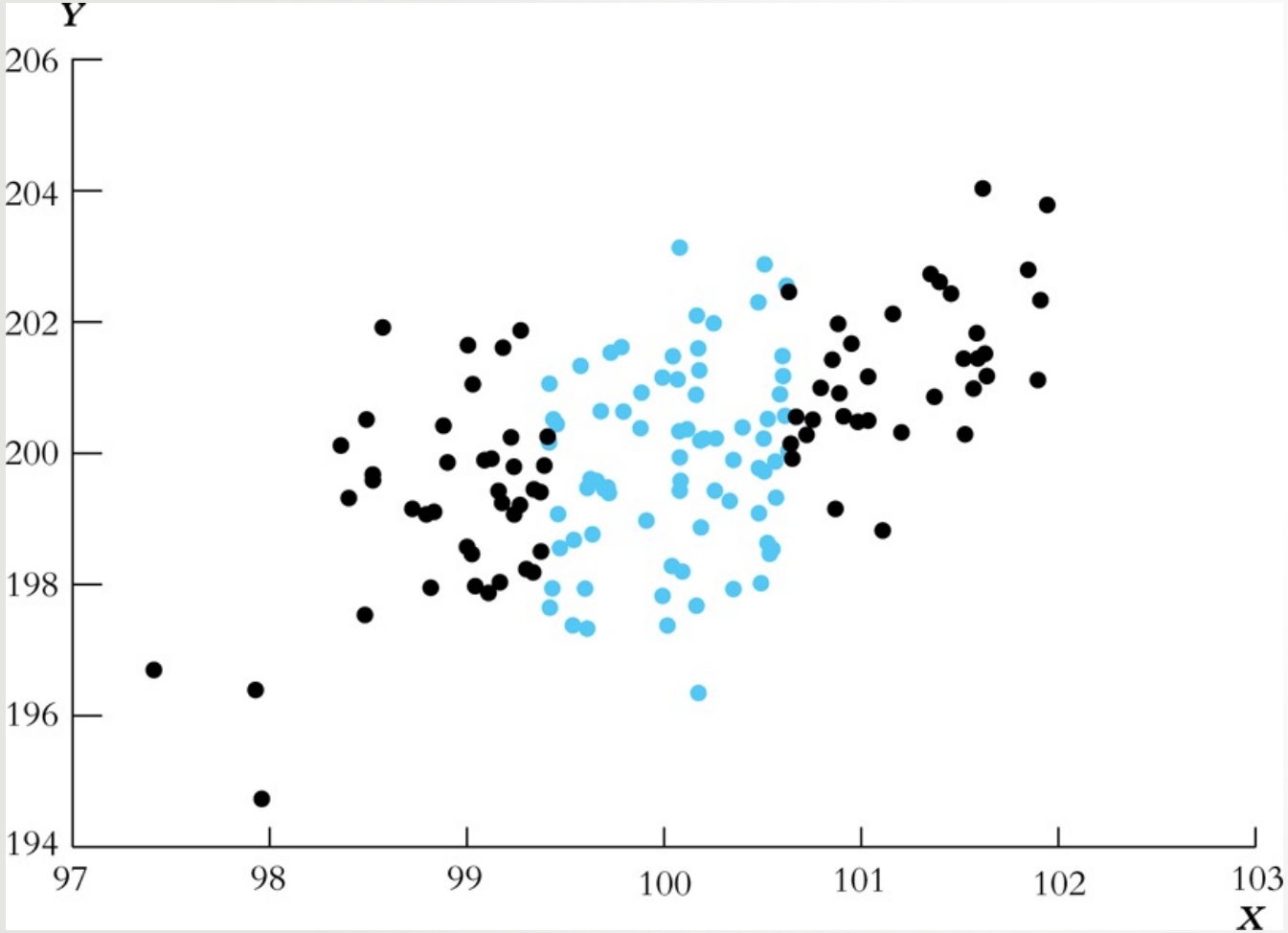
SAMPLING DISTRIBUTION OF $\widehat{\beta}_0$ AND $\widehat{\beta}_1$

- $\widehat{\beta}_0$ and $\widehat{\beta}_1$ = random variables (*why?)*
- $\mathrm{E}(\widehat{\beta}_0) = \beta_0$ and $\mathrm{E}(\widehat{\beta}_1) = \beta_1$
- $\widehat{\beta}_0$ and $\widehat{\beta}_1$ tend to be normally distributed in large samples.
- $\widehat{\beta}_0 \sim N(\beta_0, \sigma^2_{\widehat{\beta}_0})$ and $\widehat{\beta}_1 \sim N(\beta_1, \sigma^2_{\widehat{\beta}_1})$
- What determines $\sigma^2_{\widehat{\beta}_0}$ and $\sigma^2_{\widehat{\beta}_1}$?

THE VARIANCE OF THE OLS ESTIMATOR

$$\text{var}(\hat{\beta}_1) = \frac{1}{n} \times \frac{\text{var}[(X_i - \mu_x)u_i]}{(\sigma_X^2)^2}.$$

THE VARIANCE OF THE OLS ESTIMATOR



- The number of black and blue dots is the same and they come from the same joint distribution.
- Using which would you get a more accurate regression line?
- Increasing the spread of X decreases $var(\hat{\beta}_1)$