

4 – LINEAR REGRESSION I ONE REGRESSOR

University of
Massachusetts
Amherst BE REVOLUTIONARY™



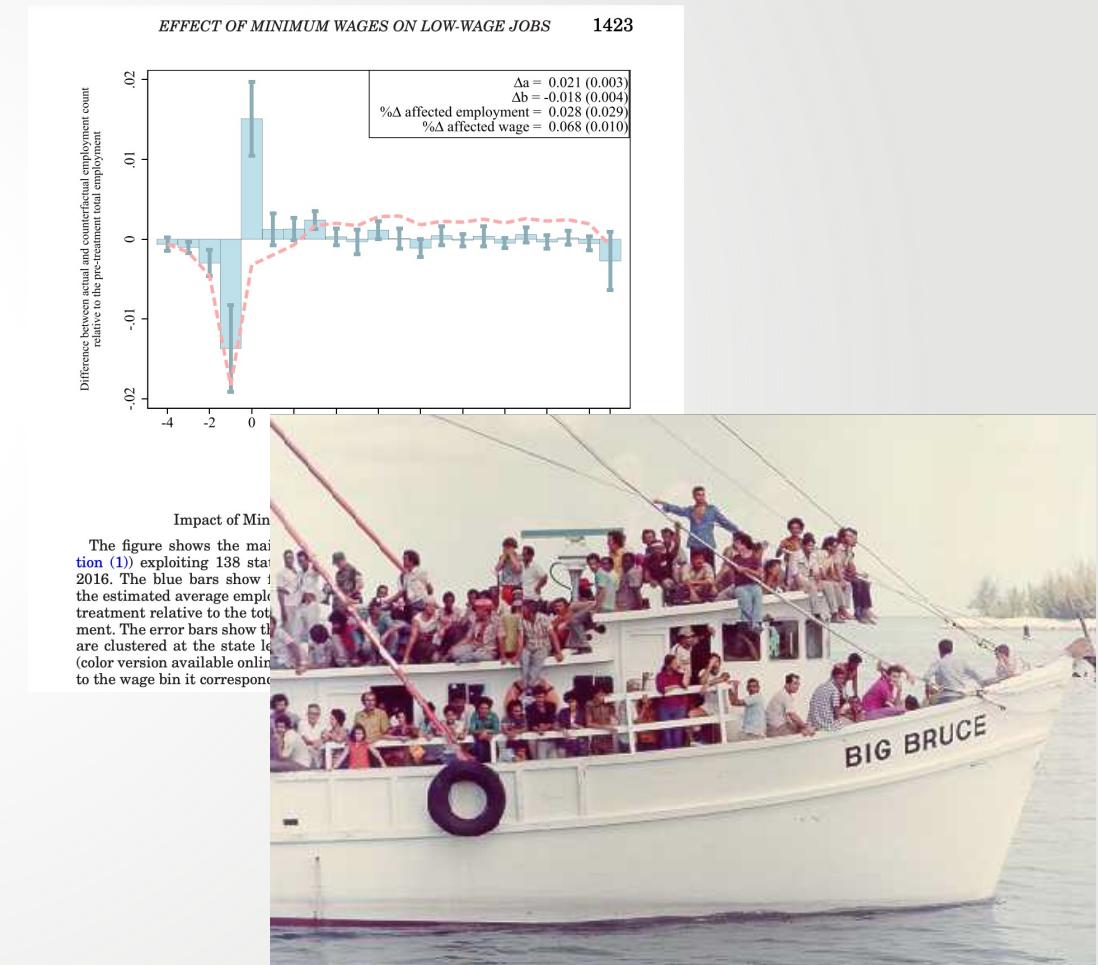
SECTION 4 – LINEAR REGRESSION, PART 1

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

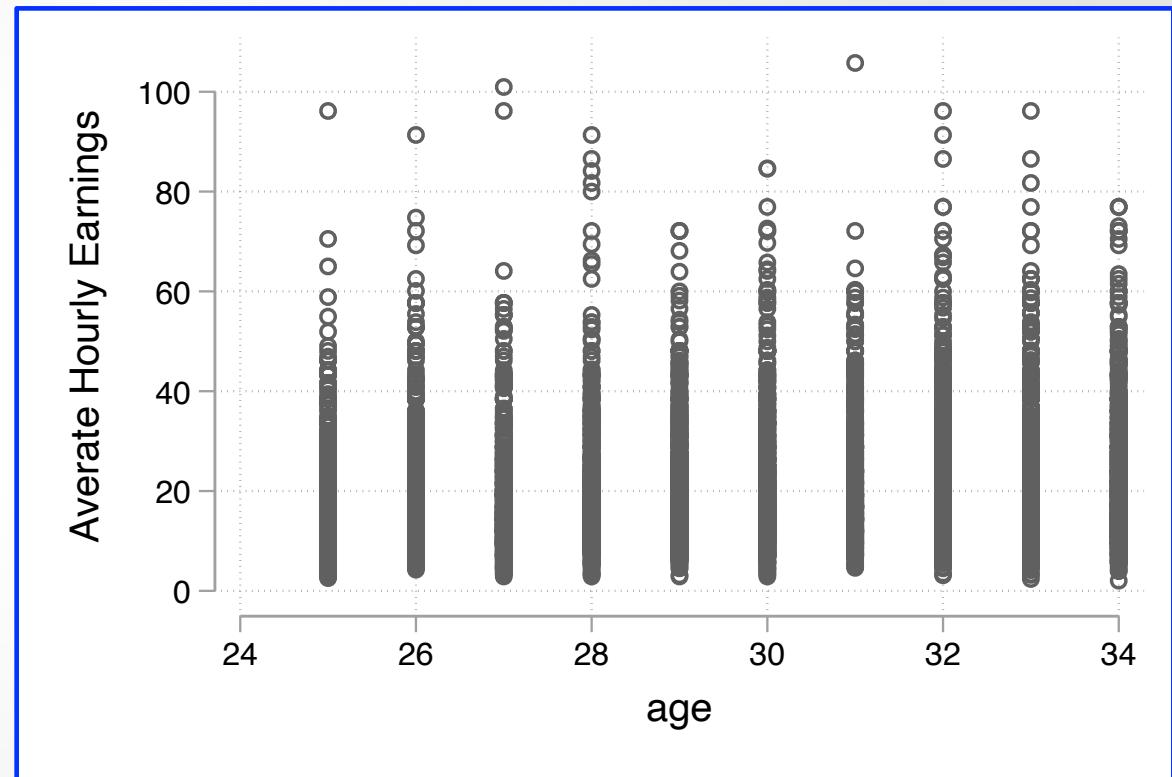
- Does a minimum wage decrease employment?
- Do increases in government spending boost or harm growth?
- Does immigration lower wages for native workers?
- Is a recession likely in the next year?
-



4.1 THE LINEAR REGRESSION MODEL

CONDITIONAL EXPECTATIONS

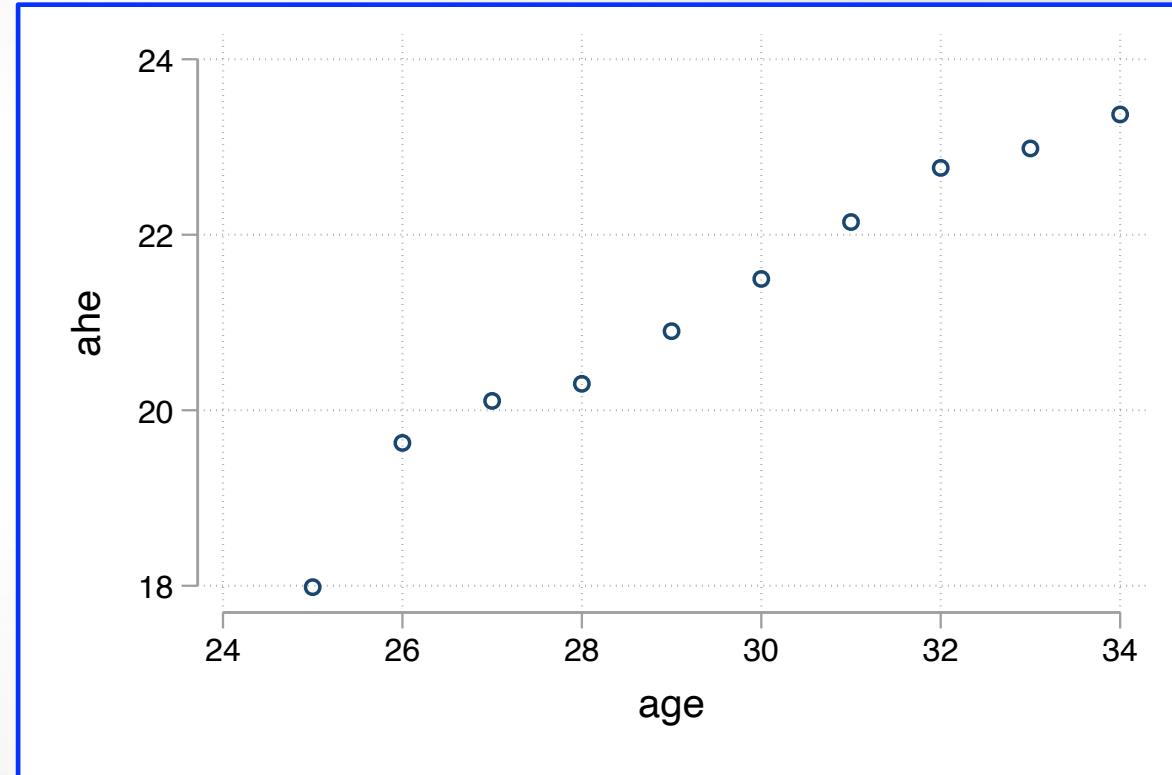
- How do average earnings vary with age?
- Conditional Expectation:
 $E(AHE|AGE = x)$
- Scatterplot of CPS data
 - STATA: “scatter ahe age”
- Imagine it was the population, not a sample.



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.



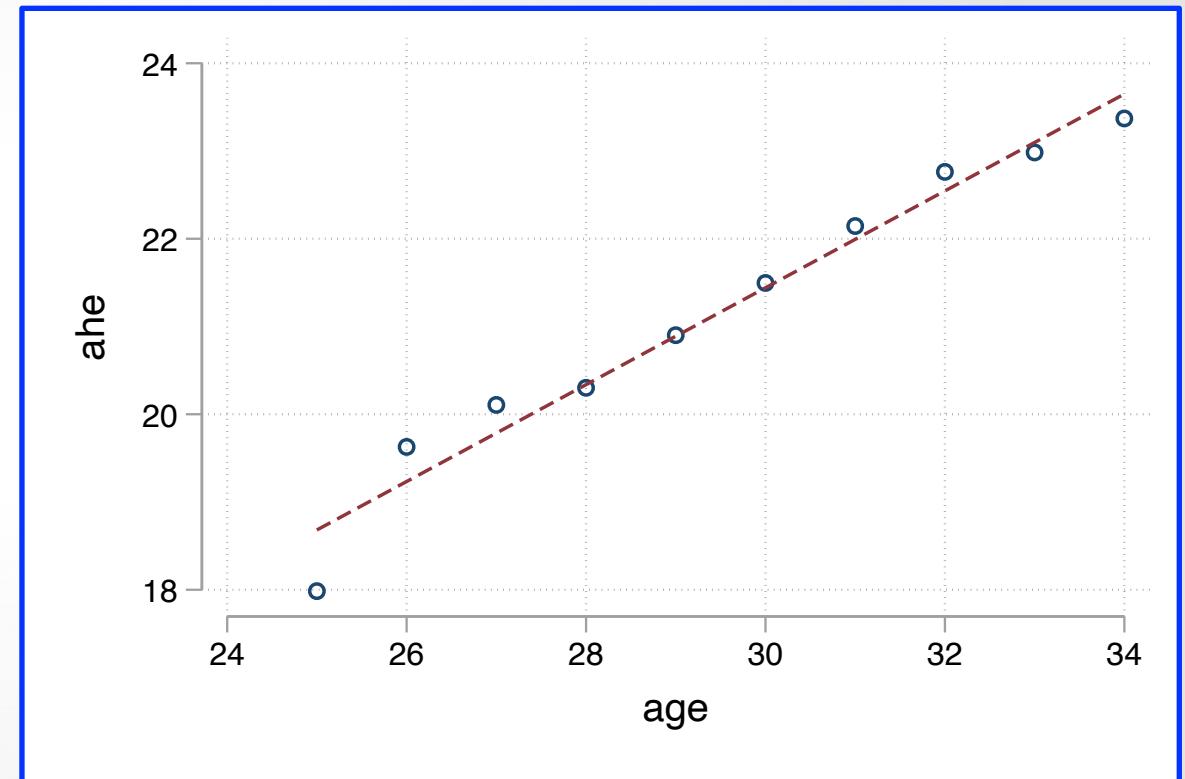
CONDITIONAL EXPECTATIONS

- 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.



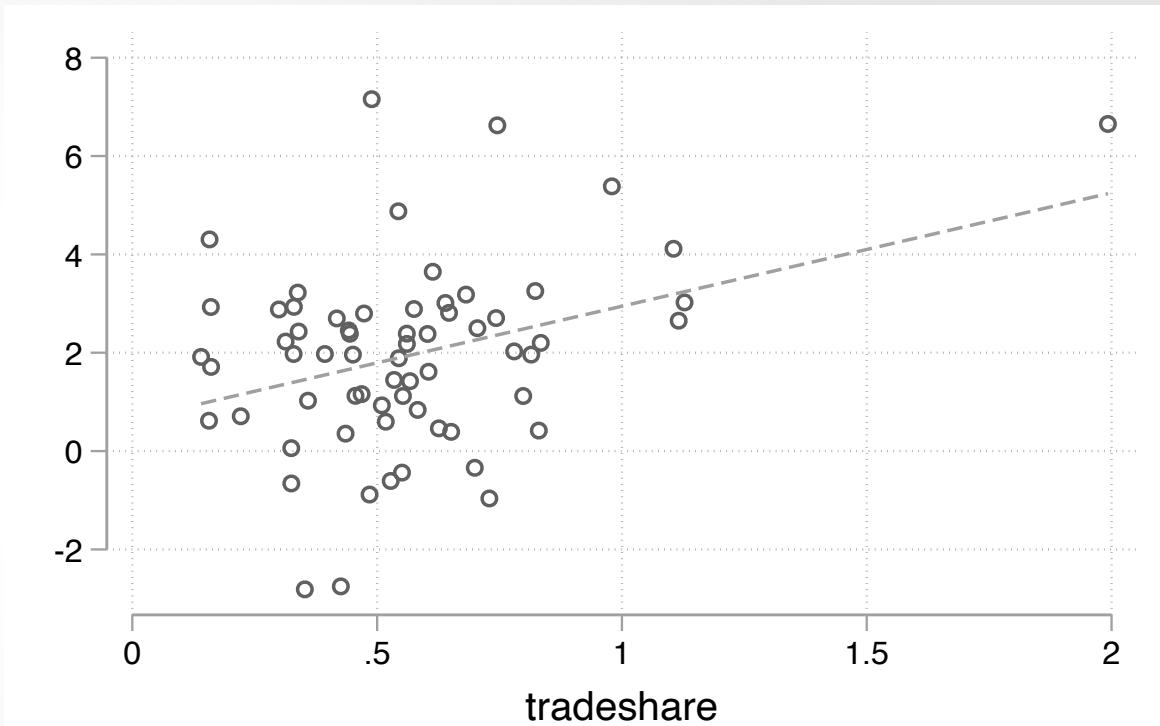
CONDITIONAL EXPECTATIONS

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

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

- For an individual country i :

$$\begin{aligned}\text{growth}_i &= E(\text{growth}|\text{trade}_i) + u_i \\ &= \beta_0 + \beta_1 \text{trade}_i + u_i\end{aligned}$$



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

THE LINEAR REGRESSION MODEL

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

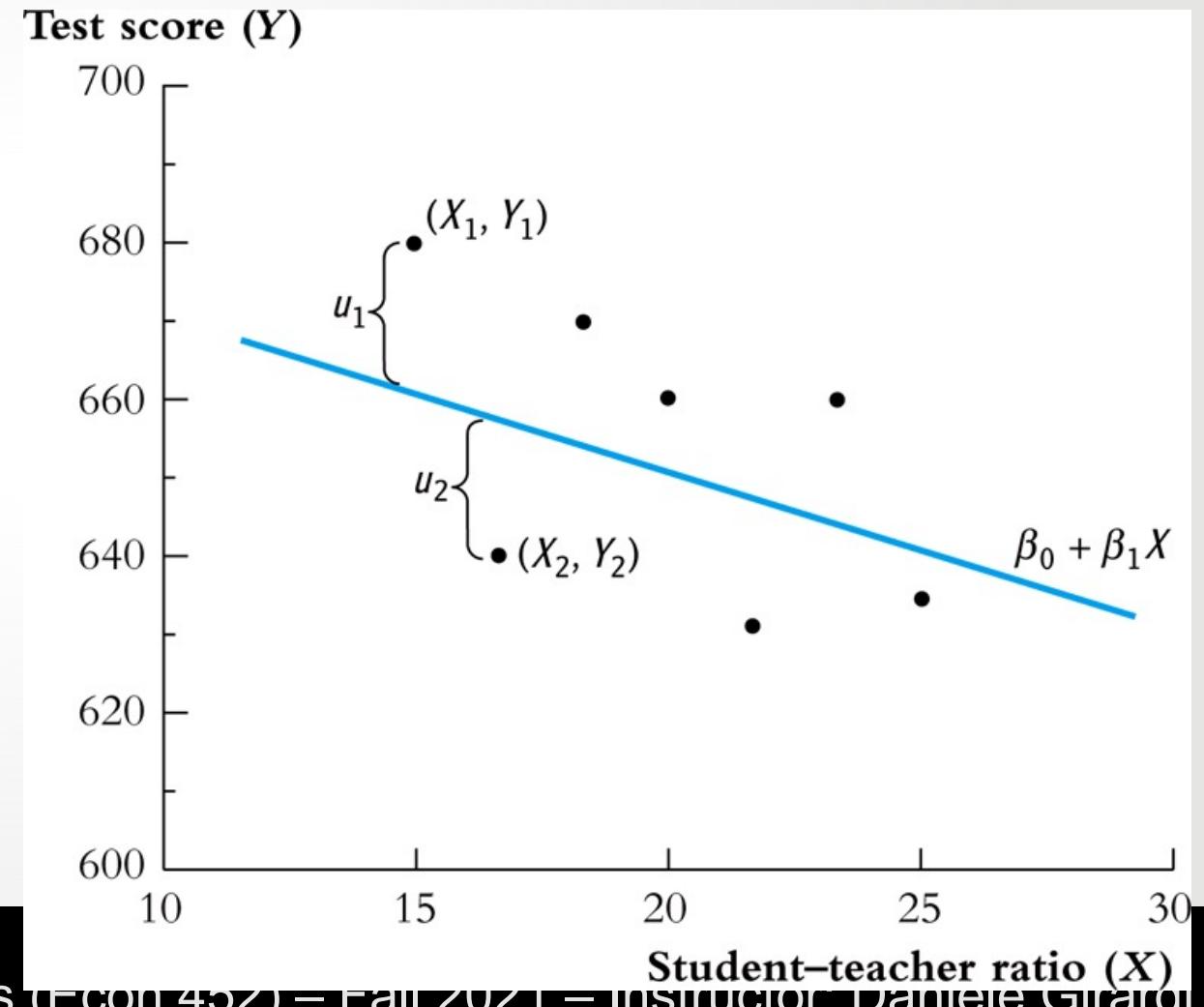
- Y = dependent variable
- X = independent variable (or regressor)
- $\beta_0 + \beta_1 X_i$ = population regression function
- β_0 = population intercept
- β_1 = population slope
- u_i = population error term

TEST SCORES VS STUDENT-TEACHER RATIO (HYPOTHETICAL DATA)

$$E(\text{TestScore}) = \beta_0 + \beta_1 \text{STR}$$



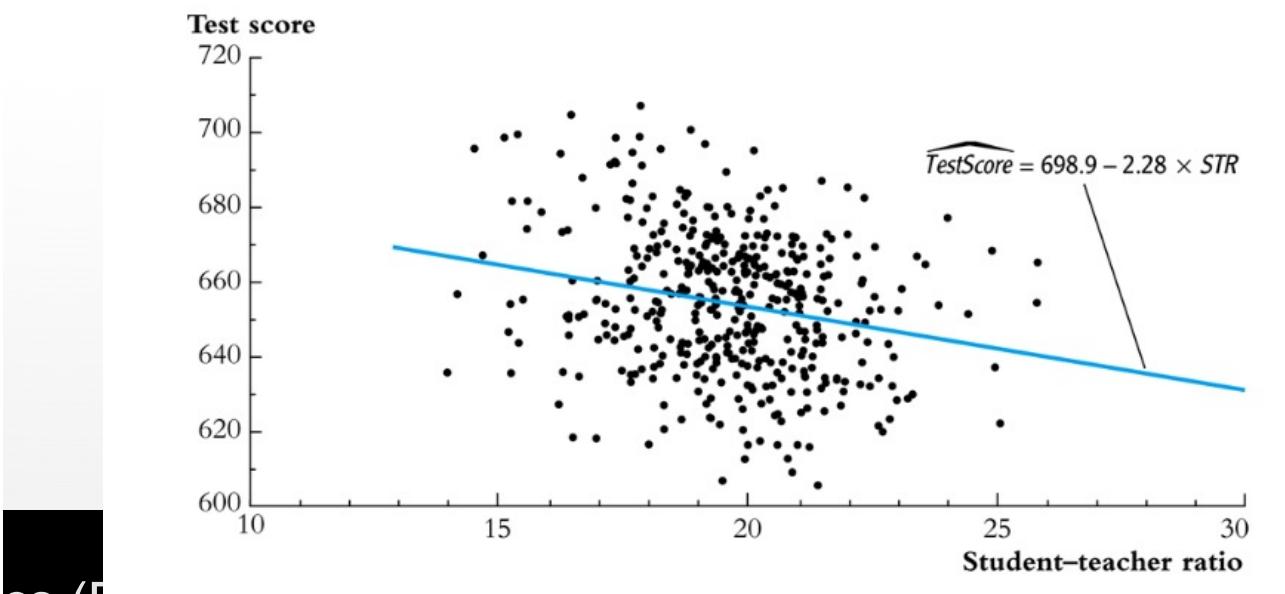
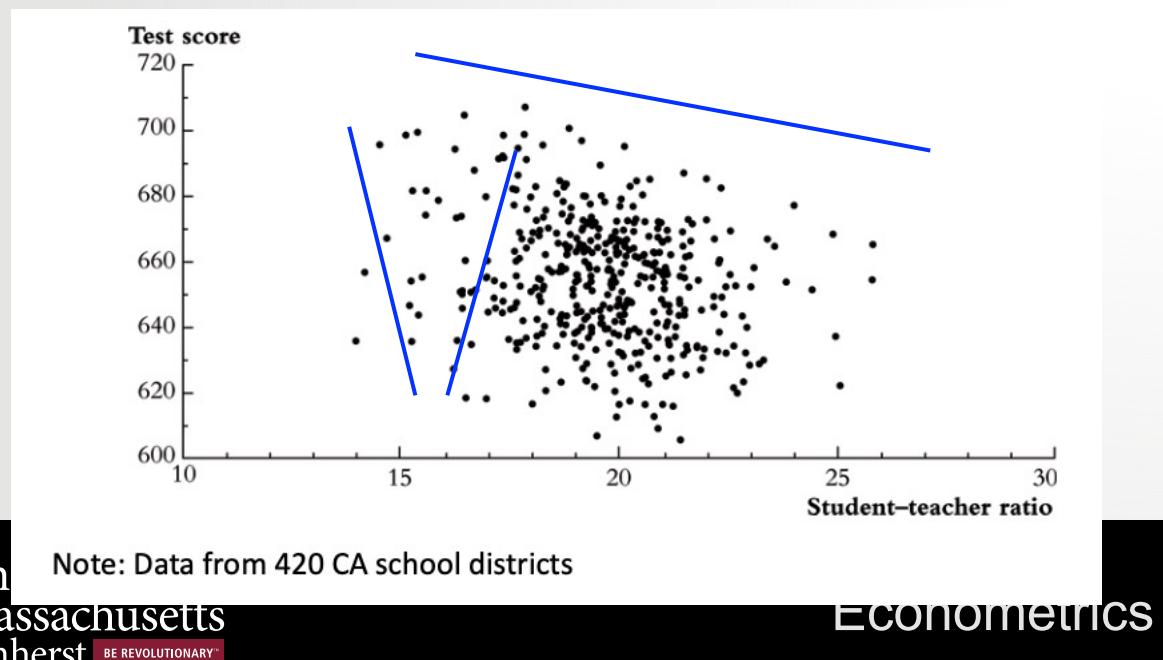
$$\text{TestScore}_i = \beta_0 + \beta_1 \text{STR}_i + u_i$$



4.2 ESTIMATION OF THE LINEAR REGRESSION MODEL

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.



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
```

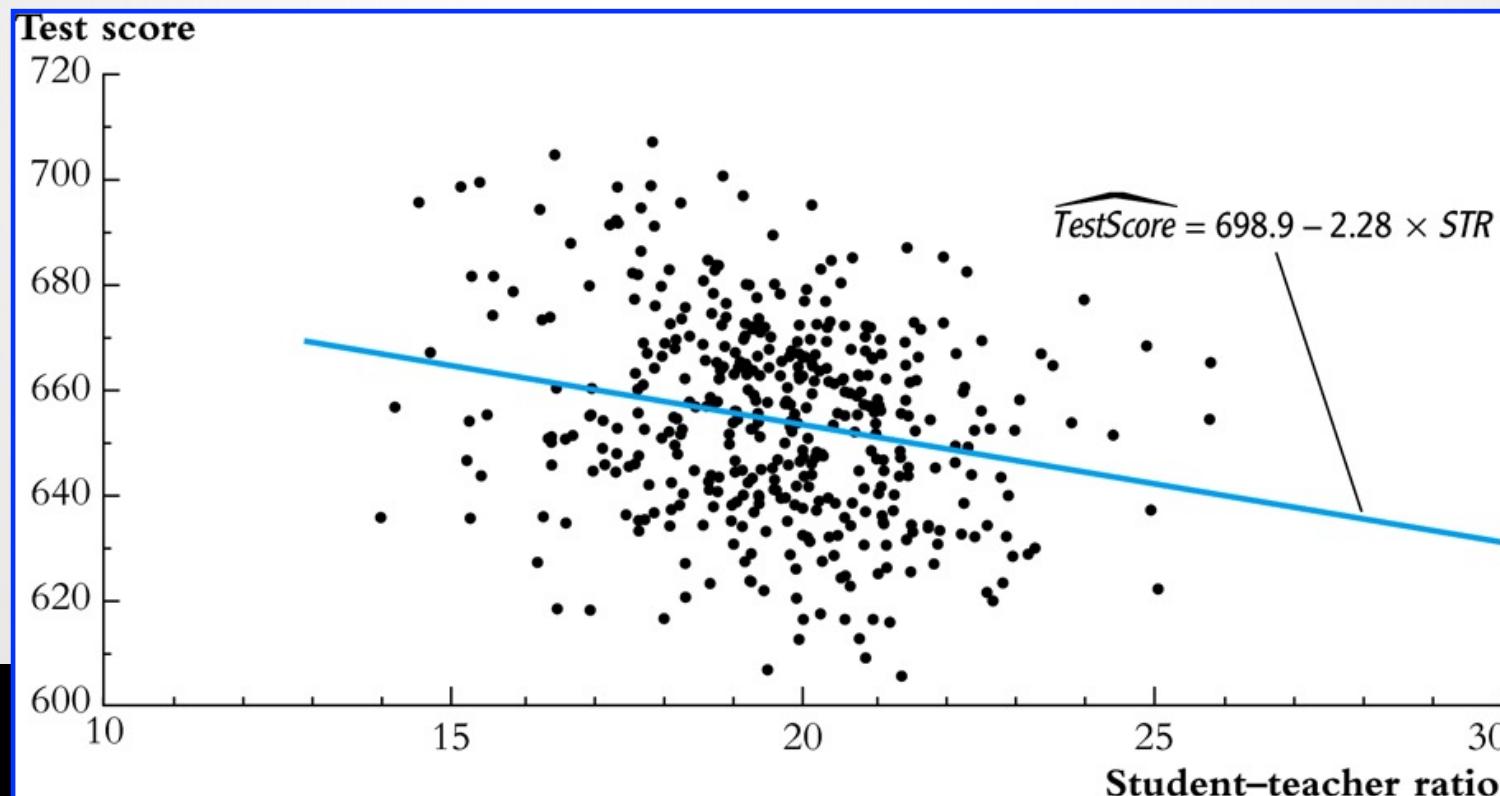
	Robust					
testscr	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	

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

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

OLS APPLICATION: CLASS SIZE & TEST SCORES

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



OLS APPLICATION: CLASS SIZE & TEST SCORES

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

- 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

	Average	Standard Deviation	Percentile						
			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

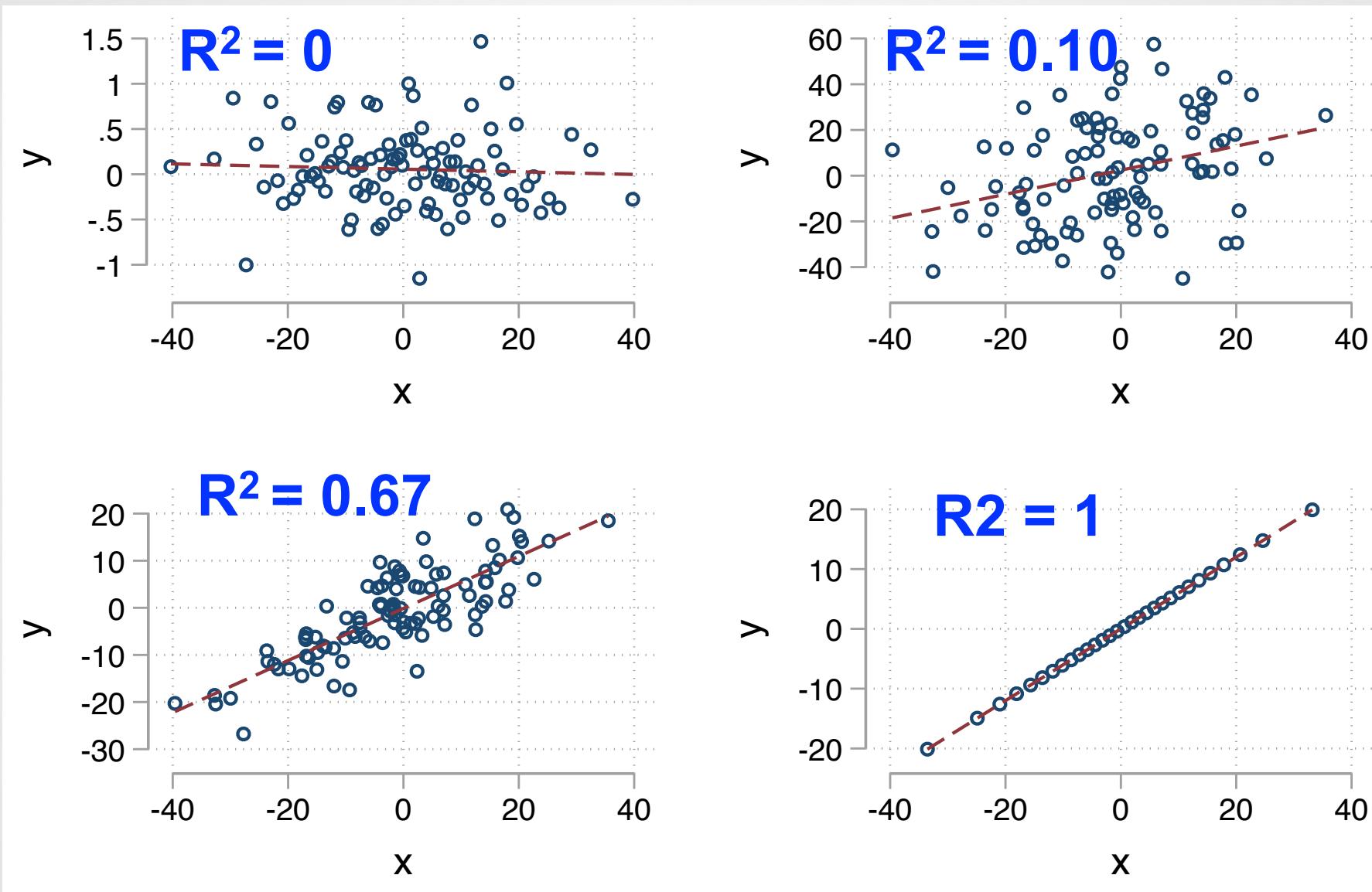
MEASURES OF FIT

How well does our regression line fit the data?

- R^2
- **SER:** Standard Error of the Regression

THE R²

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



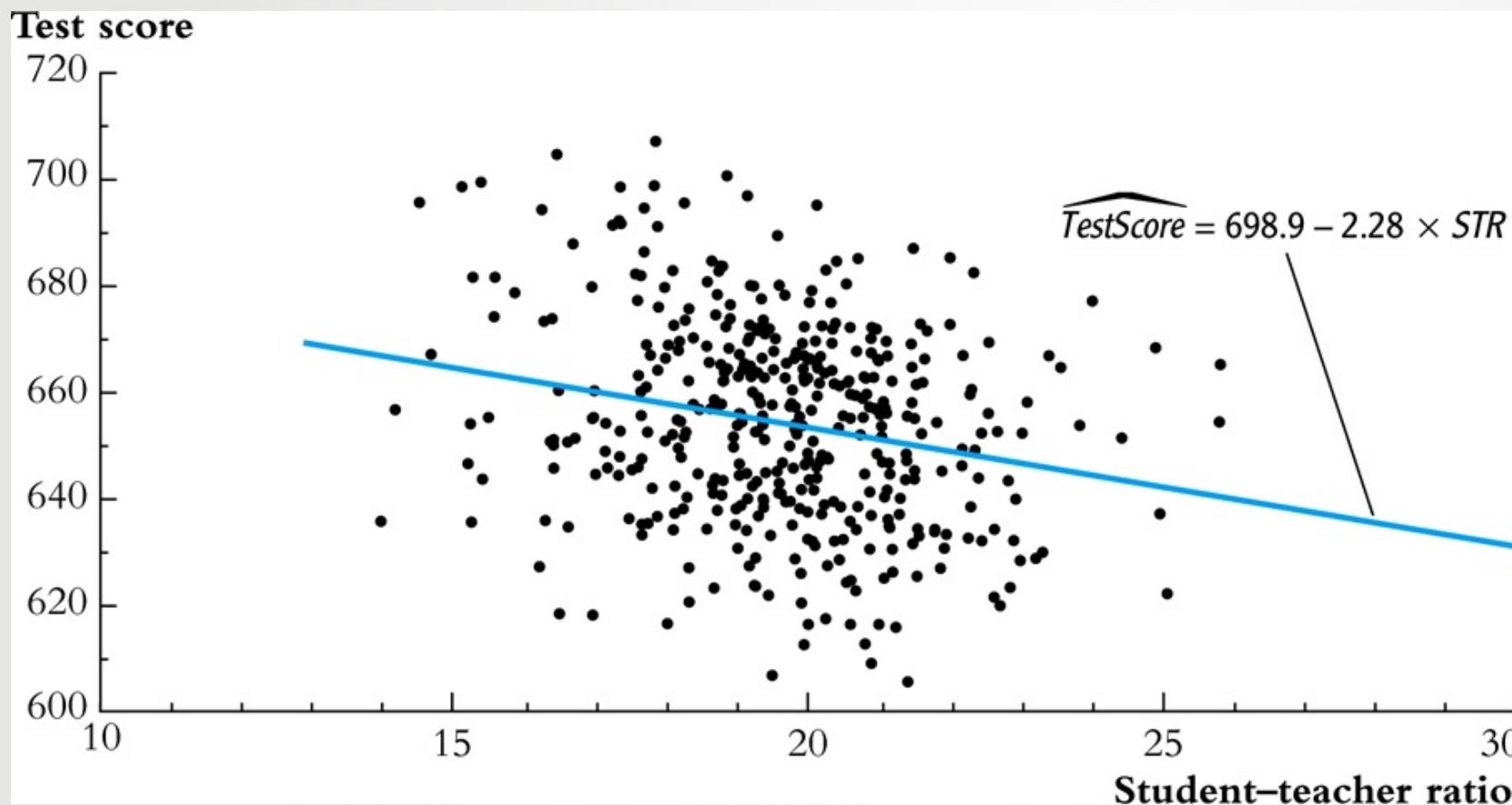
SER

- Standard Error of the Regression (SER):

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

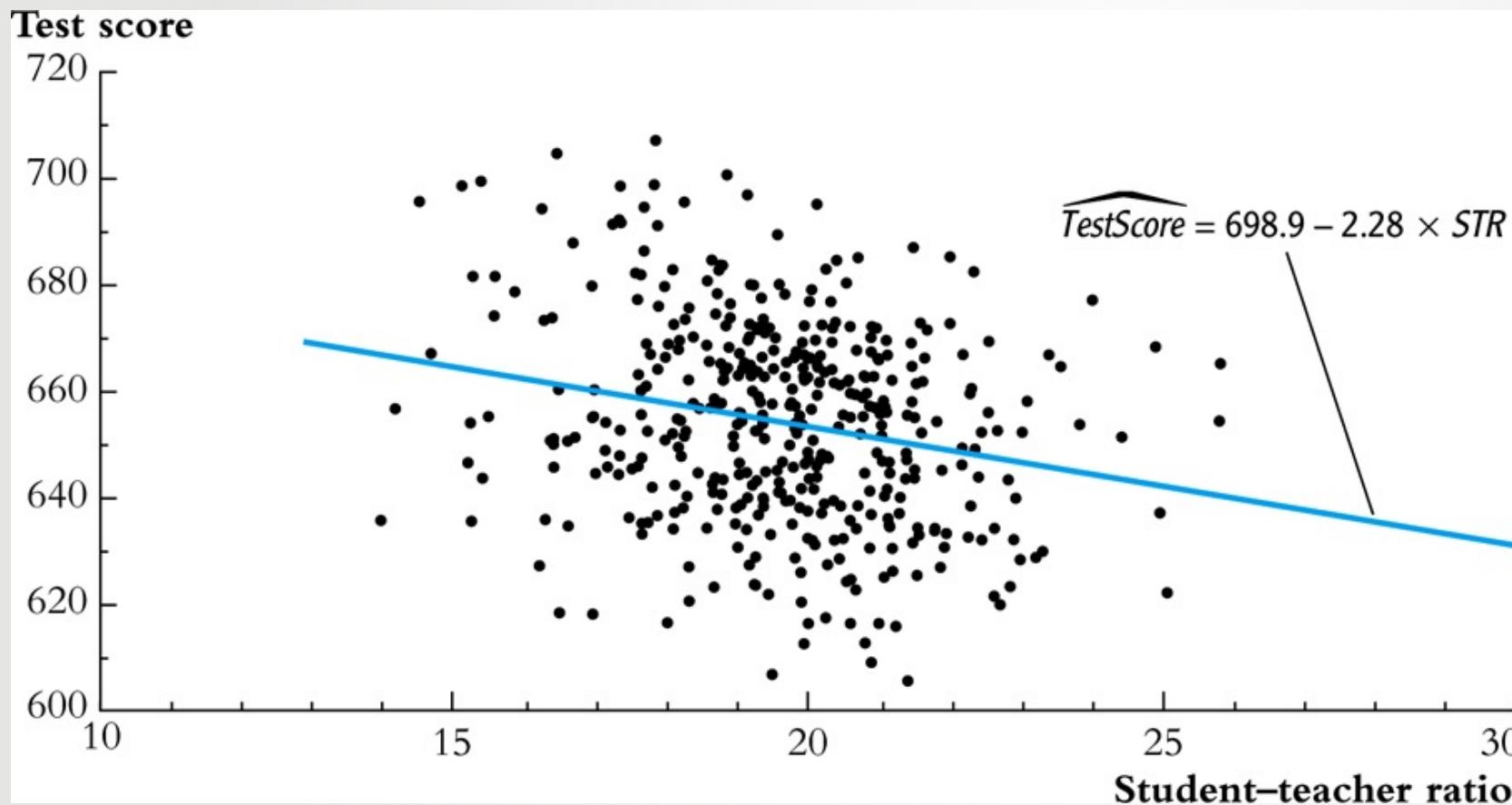
- Measures the “spread” of residuals around the regression lines.

TEST SCORES EXAMPLE



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

TEST SCORES EXAMPLE

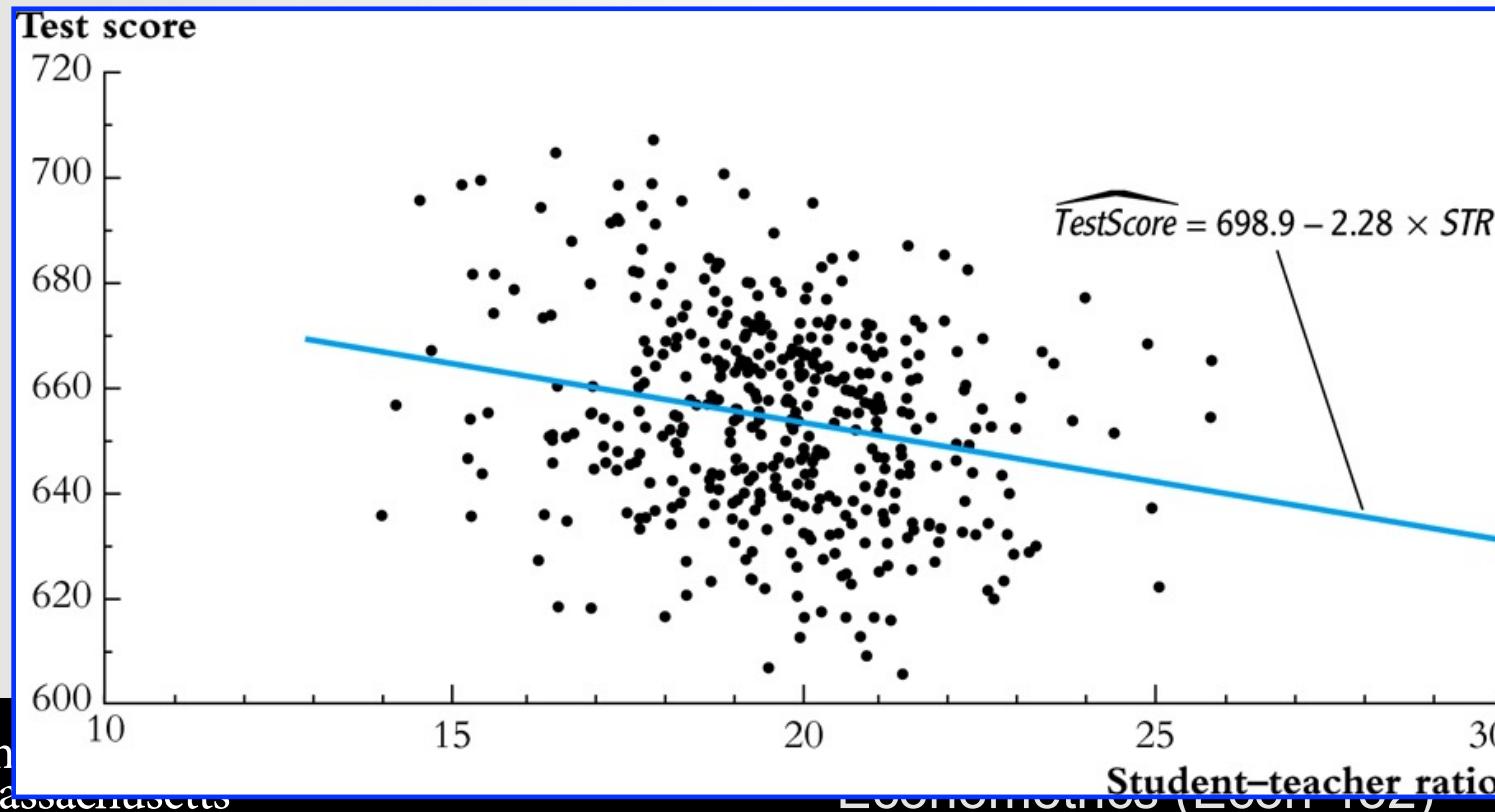


- $R^2= 0.05$
- $SER=18.6$

4.4 REGRESSION AND CAUSALITY

CLASS SIZE & TEST SCORES AGAIN

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



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

CAUSAL RELATIONS BETWEEN CLASS SIZE & TEST SCORES

Class size X

Other factors u

Test scores Y

Class size X ← Other factors u

Test scores Y

- 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 $u_i \rightarrow \text{corr}(X_i, u_i) = 0$
- 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:

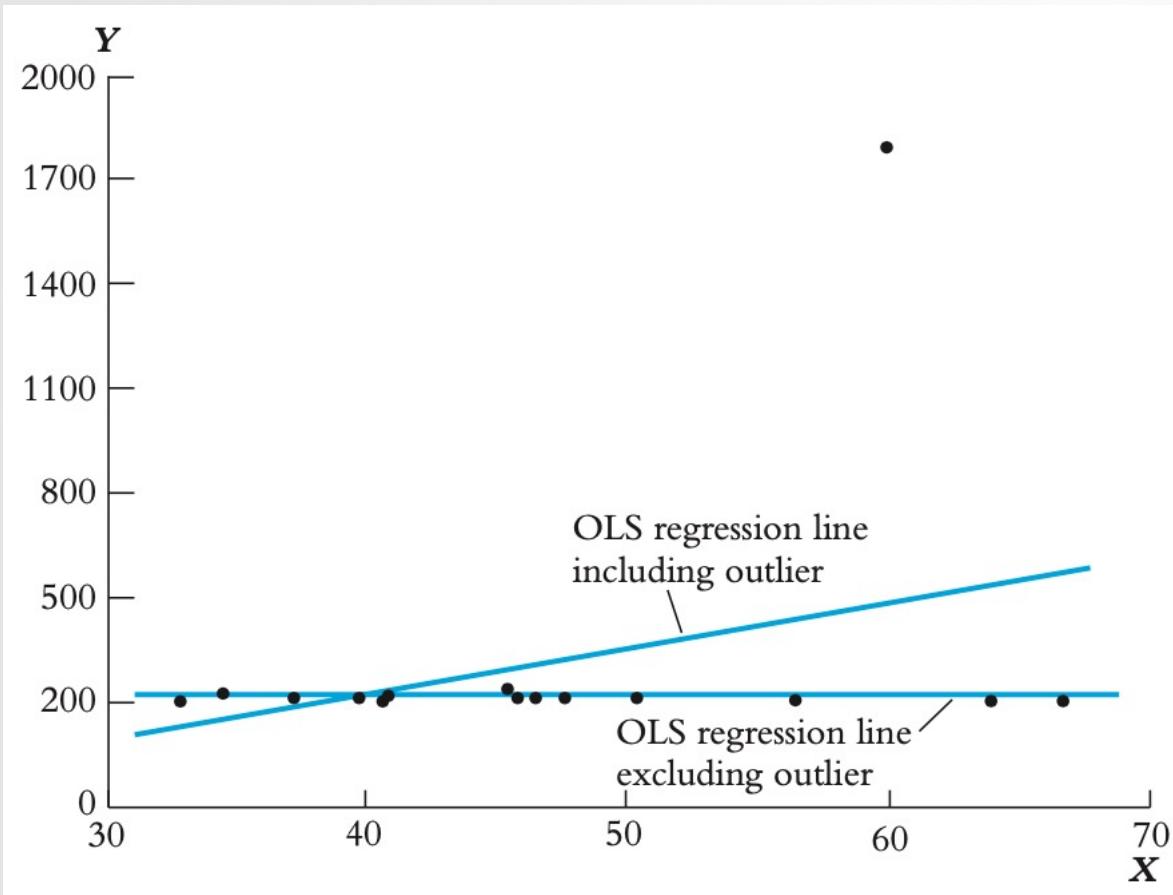
$$AHE_i = \beta_0 + \beta_1 Educ_i + u_i$$

- 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 .
 - $E(u_i|X = x) = 0; \ corr(X_i, u_i) = 0$
2. $(X_i, Y_i), i = 1, \dots, n$, 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.

SUMMING UP...

1. Linear regression model.
2. OLS estimator.
3. R^2 .
4. Assumptions for causal inference.
5. Sampling distribution.

4.5 SAMPLING DISTRIBUTION OF THE OLS ESTIMATOR

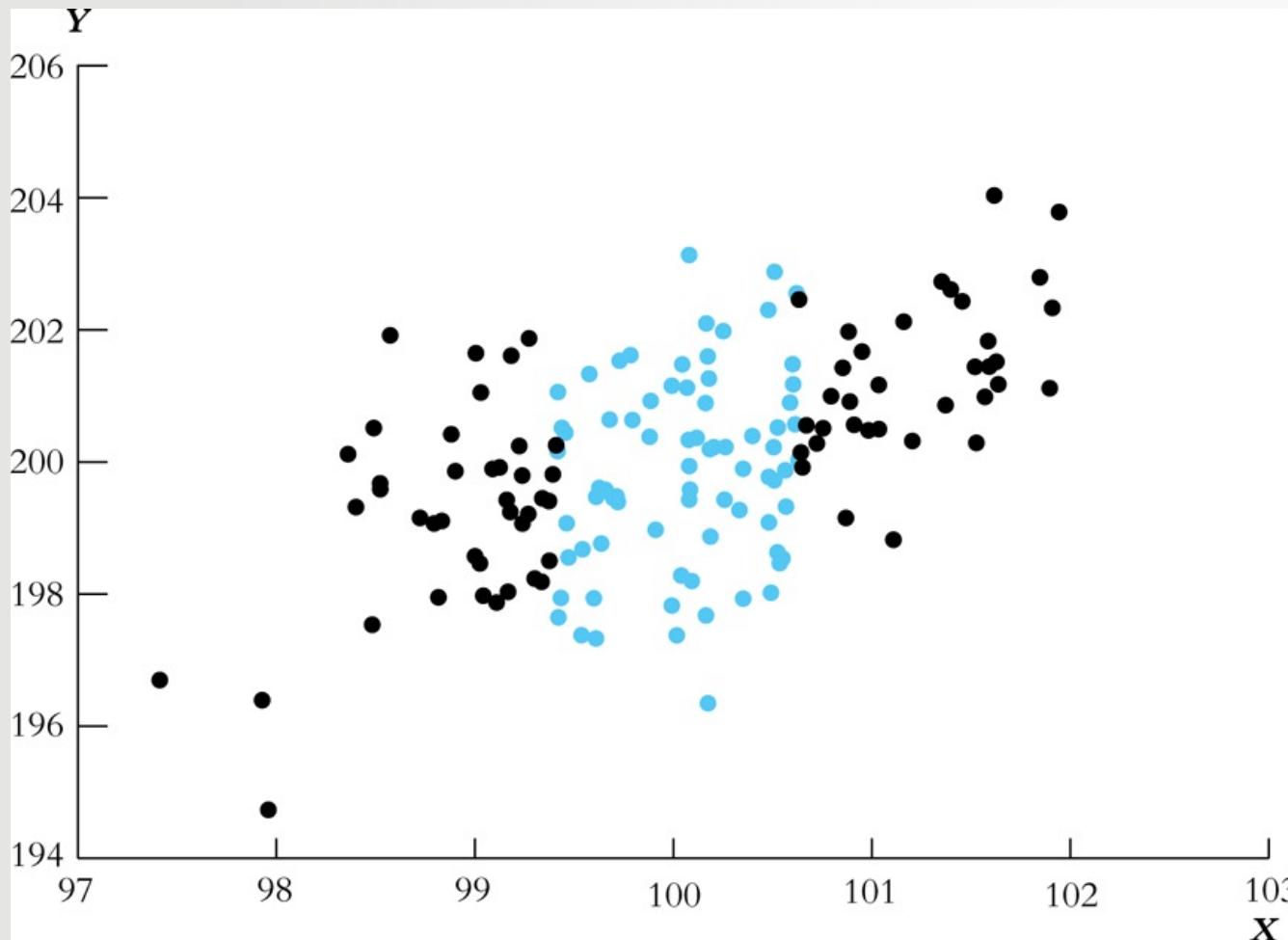
SAMPLING DISTRIBUTION OF $\hat{\beta}_0$ AND $\hat{\beta}_1$

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

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 $\text{var}(\hat{\beta}_1)$

4.6 HYPOTHESIS TESTS ABOUT REGRESSION COEFFICIENTS

HYPOTHESIS TESTS

- Null and alternative hypotheses:

$$H_0: \beta_1 = \beta_{1,0} \text{ vs. } H_1: \beta_1 \neq \beta_{1,0}$$

- Three steps for testing H_0 :
 1. Compute $\hat{\beta}_1$ and $SE(\hat{\beta}_1)$ using sample data.
 2. Compute the t-statistics
 3. Compute the p-value

HYPOTHESIS TESTS

- Null and alternative hypotheses:

$$H_0: \beta_1 = \beta_{1,0} \text{ vs. } H_1: \beta_1 \neq \beta_{1,0}$$

- Three steps for testing H_0 :
 1. Compute $\hat{\beta}_1$ and $SE(\hat{\beta}_1)$ using sample data.
 2. Compute the t-statistics
 3. Compute the p-value

THE STANDARD ERROR OF $\hat{\beta}_1$

- $SE(\hat{\beta}_1)$ is an estimator of $\sigma_{\hat{\beta}_1}$.
- $SE(\hat{\beta}_1) = \sqrt{\hat{\sigma}_{\hat{\beta}_1}^2} = \frac{1}{n} \times \frac{\frac{1}{n-2} \sum_{i=1}^n (X_i - \bar{X})^2 \hat{u}_i^2}{\left[\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \right]^2}$
- (complicated, but STATA will do it for you)
- Also called *robust* standard error.

```
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

testscr	Robust					
	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

- $\hat{\beta}_1 = -2.28$ and $SE(\hat{\beta}_1) = 0.52$
- $\hat{\beta}_0 = 698.9$ and $SE(\hat{\beta}_1) = 10.36$

HYPOTHESIS TESTS

- Null and alternative hypotheses:

$$H_0: \beta_1 = \beta_{1,0} \text{ vs. } H_1: \beta_1 \neq \beta_{1,0}$$

- Three steps for testing H_0 :
 1. Compute $\hat{\beta}_1$ and $SE(\hat{\beta}_1)$ using sample data.
 2. Compute the t-statistics
 3. Compute the p-value

T-STATISTICS FOR OLS PARAMETERS

$$t = \frac{\text{estimator} - \text{hypothesized value}}{\text{standard error of estimator}}$$

$$t = \frac{\hat{\beta}_1 - \hat{\beta}_{1,0}}{SE(\hat{\beta}_1)}$$

- t has a *standard normal distribution* in large samples
- $t \sim N(0,1)$

```
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

testscr	Robust					
	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

- $\hat{\beta}_1 = -2.28$ and $SE(\hat{\beta}_1) = 0.52$ and $t = -4.39$
- $\hat{\beta}_0 = 698.9$ and $SE(\hat{\beta}_1) = 10.36$ and $t = 67.44$

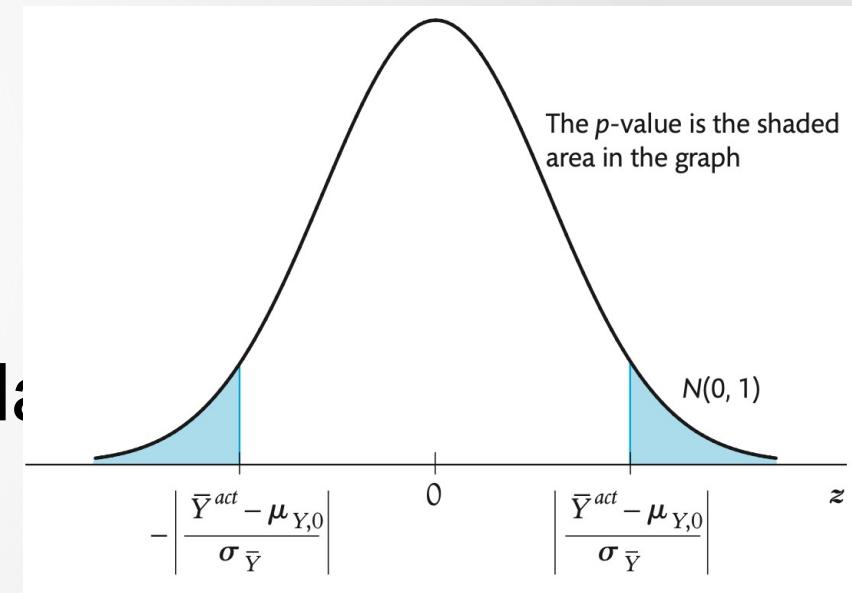
HYPOTHESIS TESTS

- Null and alternative hypotheses:

$$H_0: \beta_1 = \beta_{1,0} \text{ vs. } H_1: \beta_1 \neq$$

- Three steps for testing H_0 :

1. Compute $\hat{\beta}_1$ and $SE(\hat{\beta}_1)$ using sample data
2. Compute the t-statistics
3. Compute the p-value



COMPUTING THE P-VALUE

$$\bullet \text{ p-value} = Pr_{H_0} [|\hat{\beta}_1 - \beta_{1,0}| > |\hat{\beta}_1^{act} - \hat{\beta}_{1,0}|]$$

$$= Pr_{H_0} \left[\left| \frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)} \right| > \left| \frac{\hat{\beta}_1^{act} - \hat{\beta}_{1,0}}{SE(\hat{\beta}_1)} \right| \right]$$

$$= Pr_{H_0} (|t| > |t^{act}|)$$

$$= Pr_{H_0} (|Z| > |t^{act}|)$$

$$= 2\phi(-|t^{act}|)$$

```
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

testscr	Robust					
	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

- $\hat{\beta}_1 = -2.28$ and $SE(\hat{\beta}_1) = 0.52$ and $t = -4.39$ and $p < 0.001$
- $\hat{\beta}_0 = 698.9$ and $SE(\hat{\beta}_1) = 10.36$ and $t = 67.44$ and $p < 0.001$

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
<hr/>					
		Robust			
testscr	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
<hr/>					
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
<hr/>					

1. Compute $SE(\hat{\beta}_1) = 0.52$

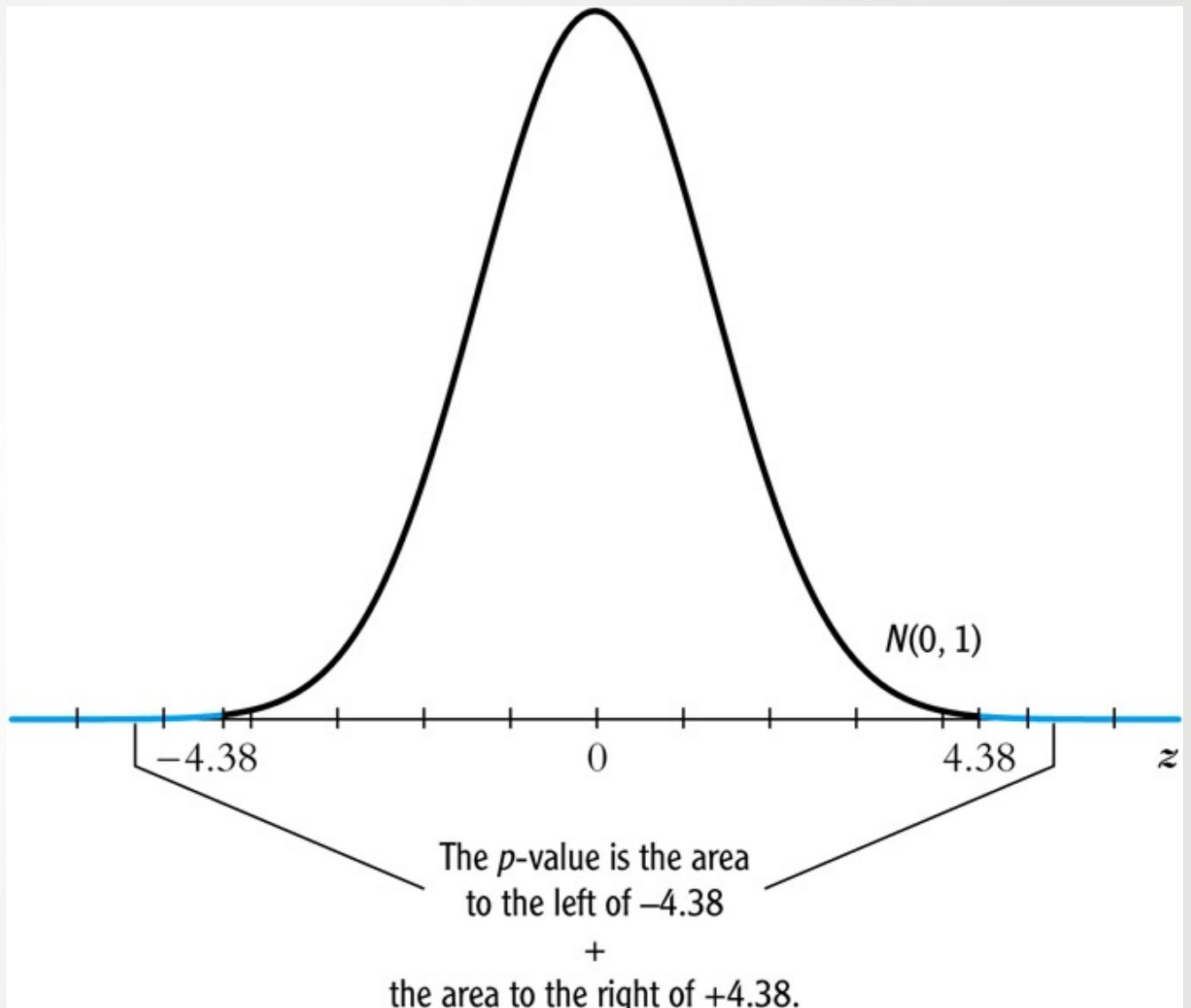
2. Compute the t-statistic:

$$t = \frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)} = \frac{-2.28 - 0}{0.52} = -4.39$$

3. Compute the p-value: $2\Phi(-|t|) = 2\Phi(-4.39) = 0.00001$

**p -value =0.00001 (or
 10^{-5})**

We can reject the null hypothesis: smaller classes do have higher test scores.



4.7 CONFIDENCE INTERVALS FOR REGRESSION COEFFICIENTS

CONFIDENCE INTERVAL FOR β_1

- **95% confidence interval:** a range of values that is 95% likely to include the “true” population coefficient β_1 .
- Includes all β_1 values that we *cannot* reject at the 5% significance level.
- 95% confidence interval for β_1 :

$$\hat{\beta}_1 - 1.96 * SE(\hat{\beta}_1) \leq \beta_1 \leq \hat{\beta}_1 + 1.96 * SE(\hat{\beta}_1)$$

CONFIDENCE INTERVAL FOR β_1

- 95% confidence interval for the effect of test scores.
- We estimated $\hat{\beta}_1 = -2.28$ and $SE(\hat{\beta}_1) = 0.52$
- So the true β_1 is 95% likely to be between:
 - $-2.28 - (1.96 \times 0.52) = -3.30$
 - $-2.28 + (1.96 \times 0.52) = -1.26$

```
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

testscr	Robust					
	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

- Confidence interval for β_1 : $[-3.30 \leq \beta_1 \leq -1.26]$

Confidence interval for β_1 (coefficient of STR):

$$[-3.30 \leq \beta_1 \leq -1.26]$$

YOUR TURN:

Can you compute a confidence interval for the average effect of a 3.5 increase in STR?

Confidence interval for β_1 (coefficient of STR):

$$[-3.30 \leq \beta_1 \leq -1.26]$$

Lower bound: $-3.30 * 3.5 = -11.55$

Upper bound: $-1.26 * 3.5 = -4.41$

Increasing STR by 3 students will decrease test scores by between 4.41 and 11.55 points.

CONFIDENCE INTERVAL FOR PREDICTED EFFECTS

- Confidence interval for the effect of a Δx change in X:

$$[(\hat{\beta}_1 \text{ lower bound}) \times \Delta x ; (\hat{\beta}_1 \text{ lower bound}) \times \Delta x]$$

$$[(\hat{\beta}_1 - 1.96 * SE(\hat{\beta}_1)) \times \Delta x ; \hat{\beta}_1 + 1.96 * SE(\hat{\beta}_1) \times \Delta x]$$

4.8 REGRESSION WHEN X IS A BINARY VARIABLE

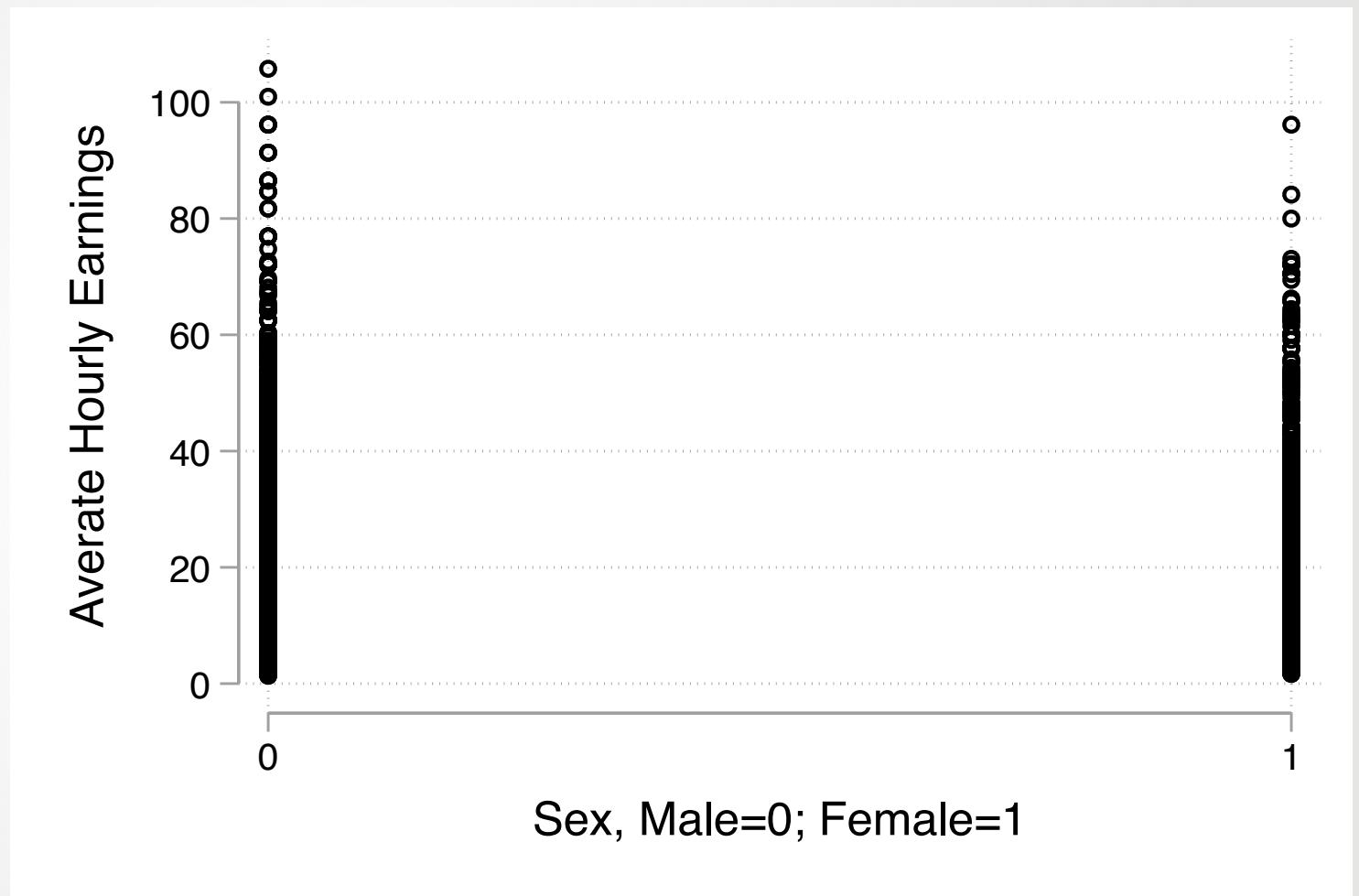
REGRESSION WHEN X IS BINARY

- **Binary (or *indicator* or *dummy*) variables**
 - Sex at birth (1 = female; 0 = male)
 - Urban or rural (1 = urban; 0 = rural)
 - Treatment or placebo
(1 = treatment; 0 = placebo)
 -



CPS 2015 data

scatter ahe female



REGRESSION WHEN X IS BINARY

$$E(Y|D) = \beta_0 + \beta_1 D$$

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

When $D_i = 0$:

$$E(Y|D = 0) = \beta_0 + \beta_1 \times 0 = \beta_0$$

When $D_i = 1$:

$$E(Y|D = 1) = \beta_0 + \beta_1 \times 1 = \beta_0 + \beta_1$$



$$\beta_1 = E(Y|D = 1) - E(Y|D = 0)$$

EXAMPLE: GENDER GAP IN EARNINGS

```
. reg ahe female, robust
```

Linear regression

Number of obs = 13,201
F(1, 13199) = 184.93
Prob > F = 0.0000
R-squared = 0.0131
Root MSE = 10.695

ahe	Robust					
	Coefficient	std. err.	t	P> t	[95% conf. interval]	
female	-2.495648	.1835205	-13.60	0.000	-2.855375	-2.135922
_cons	18.32845	.1300679	140.91	0.000	18.0735	18.5834

- AHE for men:
 $\beta_0 = 18.33$
- Difference between women and men:
 $\beta_1 = -2.50$
- AHE for women:
$$\begin{aligned}\beta_0 + \beta_1 &= \\ &= 18.33 - 2.50 \\ &= 15.83\end{aligned}$$

EXAMPLE: GENDER GAP IN EARNINGS

```
. ttest ahe, by(female) unequal unpaired
```

Two-sample t test with unequal variances

Group	Obs	Mean	Std. err.	Std. dev.	[95% conf. interval]	
0	7,694	18.32845	.1300665	11.40884	18.07348	18.58341
1	5,507	15.8328	.1294705	9.6079	15.57899	16.08661
Combined	13,201	17.28735	.0936909	10.76467	17.1037	17.471
diff		2.495648	.1835209		2.13592	2.855377

diff = mean(0) - mean(1) t = 13.5987
H0: diff = 0 Satterthwaite's degrees of freedom = 12855.9

- T-test for difference in means (see “review of statistics”).
- Regression does exactly the same thing!

REGRESSION WHEN X IS BINARY: SUMMING UP

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

- β_0 = mean of Y when $X=0$
- $\beta_0 + \beta_1$ = mean of Y when $X=1$
- β_1 = difference in group means: $E(Y|X = 1) - E(Y|X = 0)$
- T-stats, p-value, confidence intervals calculated as usual.
- Will give the same result as a t-test for difference in means.

4.9 HETEROSKEDASTICITY AND HOMOSKEDASTICITY

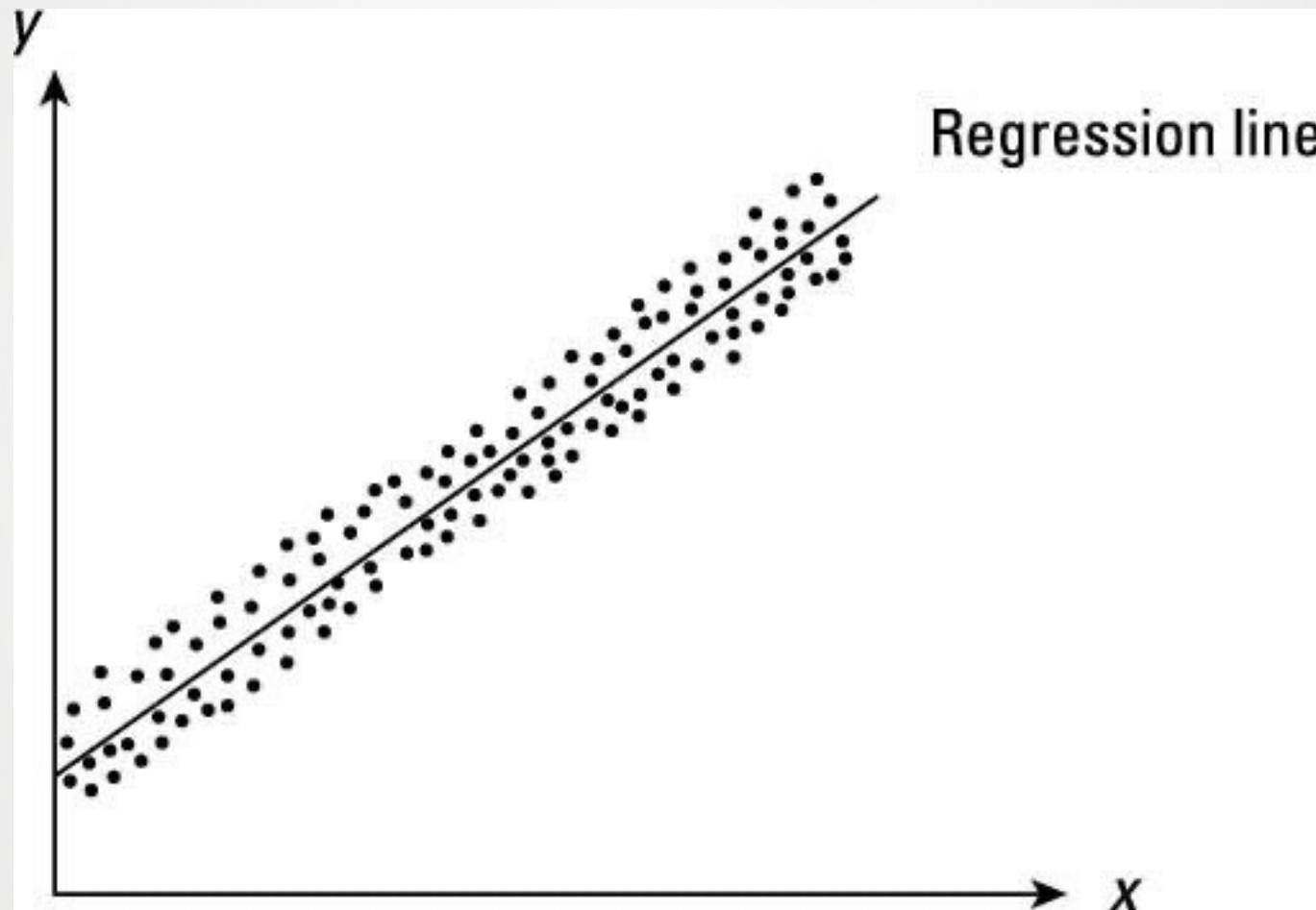
HETEROSKEDASTICITY & HOMOSKEDASTICITY

- Concerns the *conditional variance* of the error term:

$$\text{var}(u|X = x)$$

- Homoskedasticity: $\text{var}(u|X = x)$ is constant (does not depend on X).
- Heteroskedasticity: $\text{var}(u|X = x)$ varies with X.

HOMOSKEDASTICITY IN A PICTURE:



VARIANCE OF $\hat{\beta}_1$ UNDER HOMOSKEDASTICITY

- In general the variance of $\hat{\beta}_1$ is

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

- If u_i is homoscedastic we get a simpler formula:

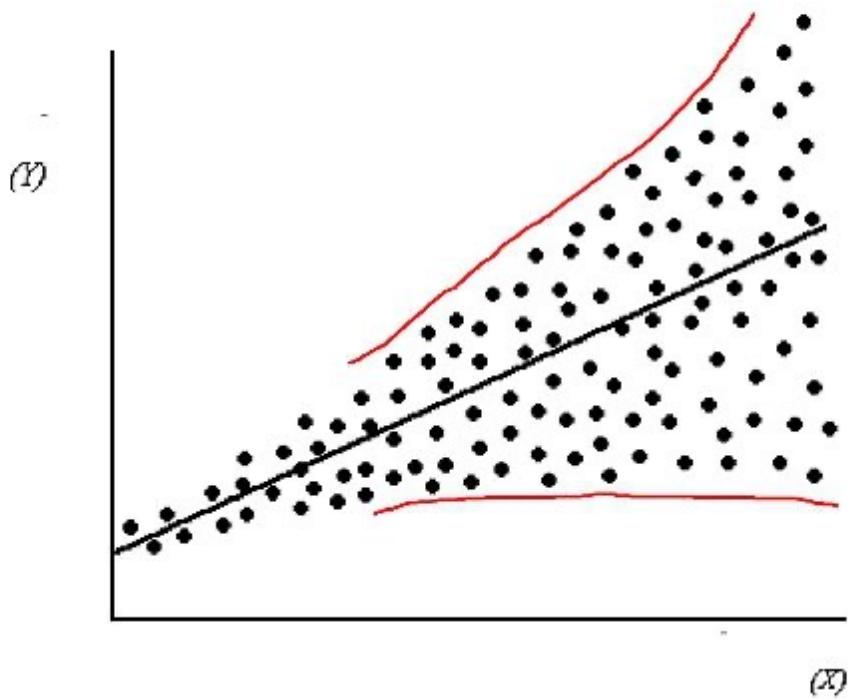
$$var(\hat{\beta}_1) = \frac{1}{n} \times \frac{var[u_i]}{[var(X_i)]^2}$$

- So the homoskedasticity-only standard error of $\hat{\beta}_1$ is also simpler:

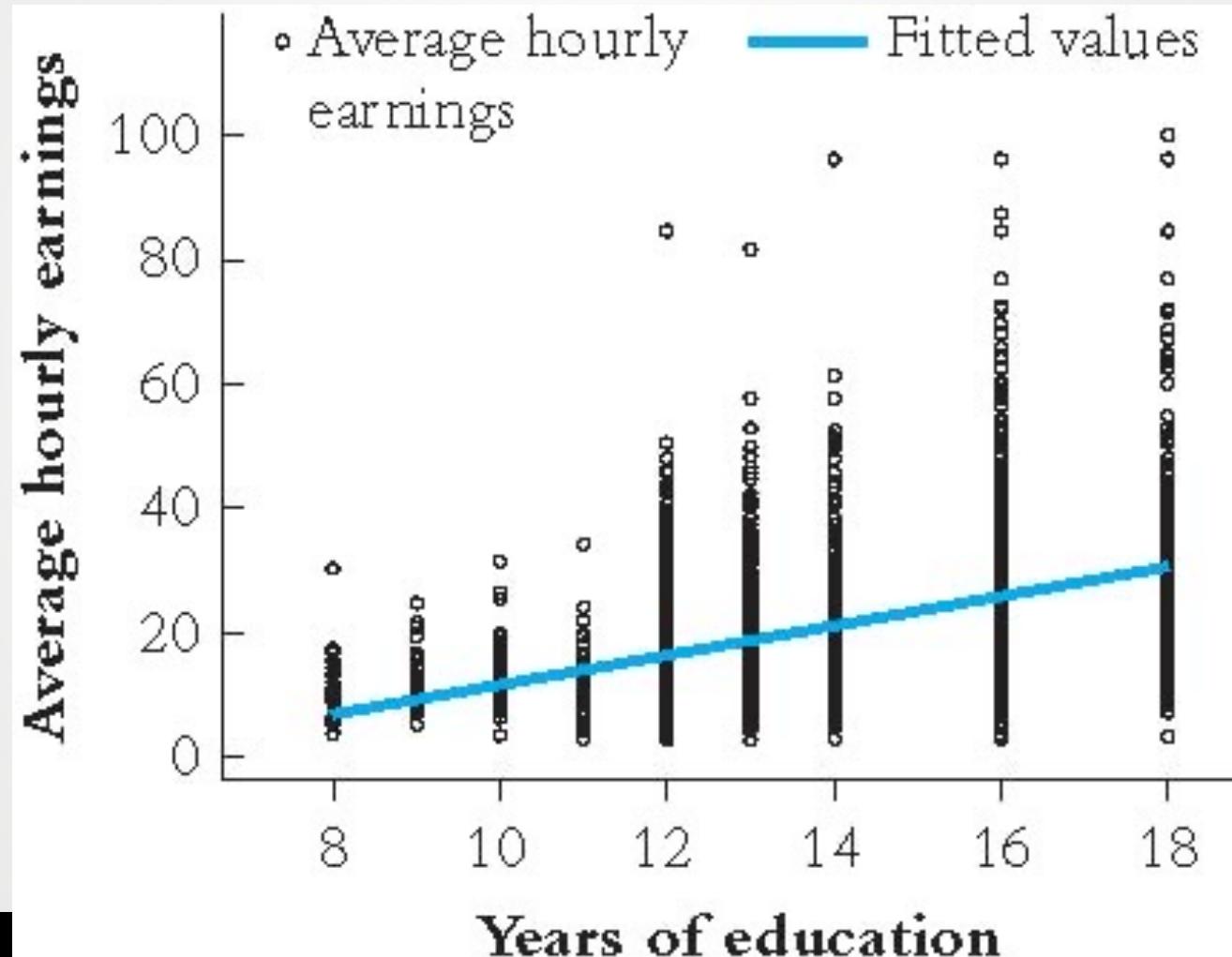
$$SE(\hat{\beta}_1) = \sqrt{\hat{\sigma}_{\hat{\beta}_1}^2} = \sqrt{\frac{\sigma_u^2}{\sum_{i=1}^n (X_i - \bar{X})^2}}$$

HETEROSKEDASTICITY IN A PICTURE:

Heteroskedasticity



REAL-DATA EXAMPLE OF HETROSKEDEASTICITY (CPS DATA)



SO, WHEN SHOULD YOU USE HOMOSKEDASTICITY-ONLY STANDARD ERRORS?

- Never.
- Our usual (heteroskedasticity-robust) SEs are always fine.
- Always use the *robust* option in STATA!

Command
`reg y x, robust`