

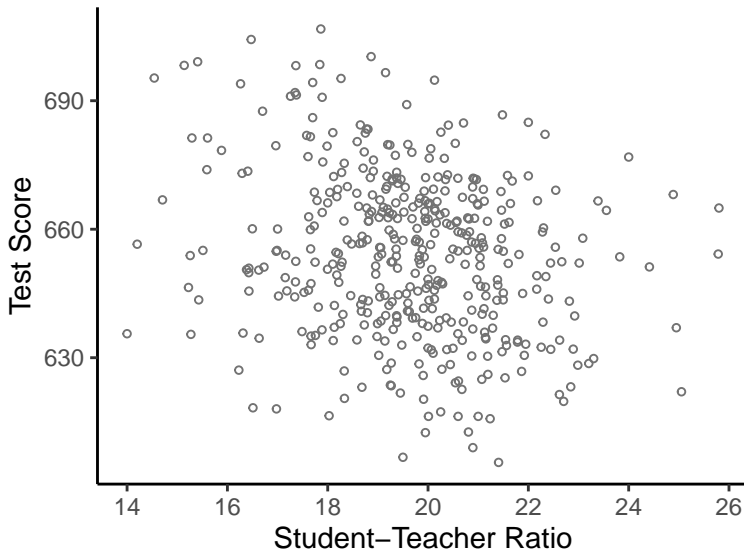
# ECON 340

## Economic Research Methods

Div Bhagia

Lecture 15: Ordinary Least Squares, Goodness of Fit

# Student-Teacher Ratio and Test Scores



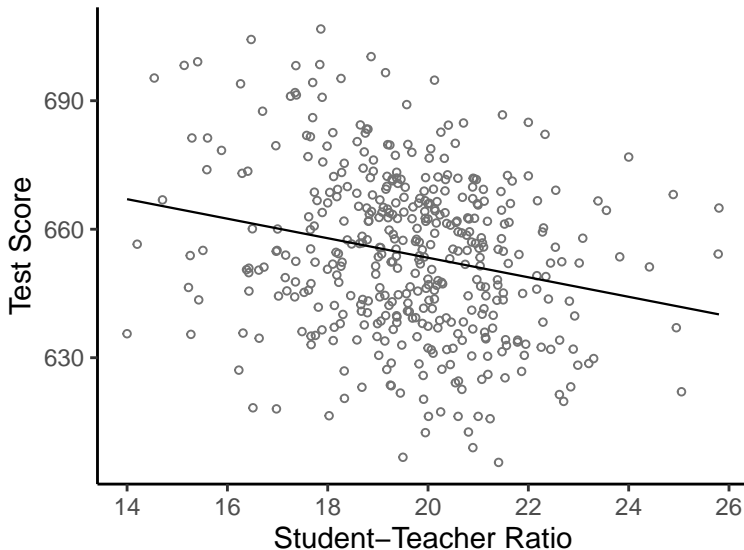
# Fitting a Line

- We are interested in the relationship between two variables  $X$  and  $Y$
- We start by assuming there is a linear relationship (with some error) between these variables in the population

$$Y = \beta_0 + \beta_1 X + u$$

- Fit a linear relationship between these two variables using sample data

# Student-Teacher Ratio and Test Scores

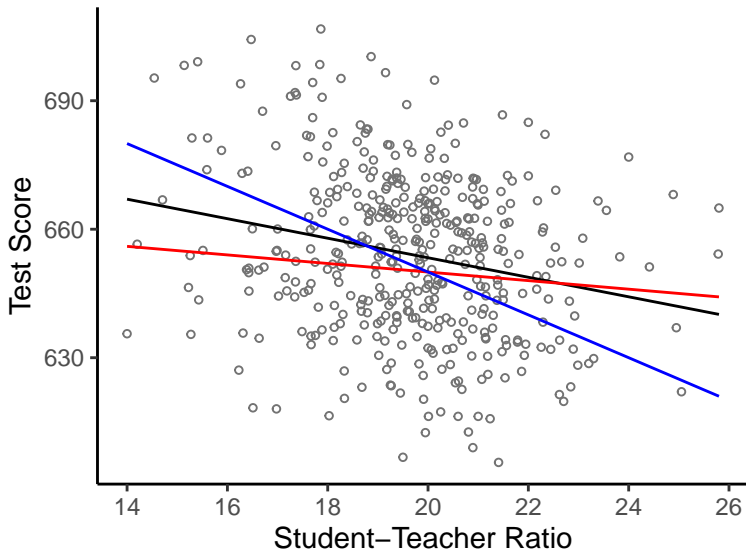


# Simple Linear Regression Model

$$Y = \beta_0 + \beta_1 X + u$$

- $Y$ : Dependent variable (outcome or response variable)
- $X$ : Independent variable (explanatory variable, regressor)
- $\beta_0, \beta_1$ : intercept and slope (population parameters)
- $u$ : mean zero error term,  $E(u) = 0$

# Which is the best line?



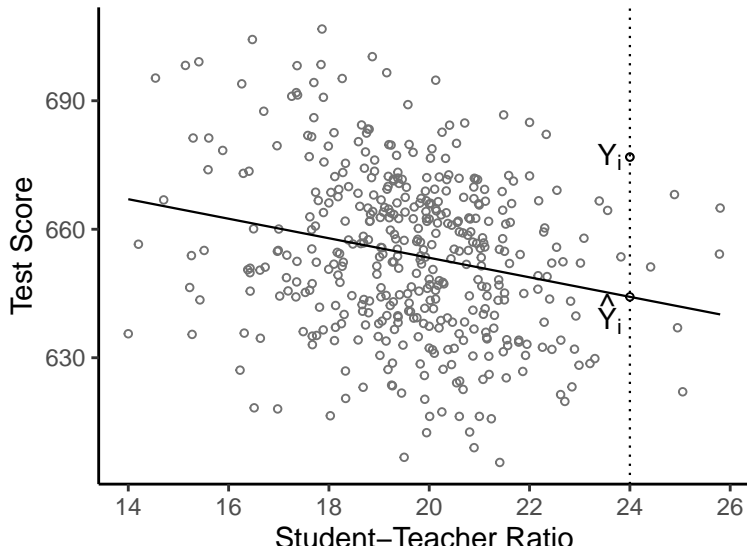
# Ordinary Least Squares (OLS)

- We observe  $Y_i$  and  $X_i$  for all individuals in our sample.
- Find  $\hat{\beta}_0$  and  $\hat{\beta}_1$  from sample data by minimizing the sum of squared residuals

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

- $\hat{\beta}_0$  and  $\hat{\beta}_1$  are called ordinary least squares (OLS) estimators

# Ordinary Least Squares (OLS)



Best line is the one that minimizes:

$$\sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$



# OLS Estimators

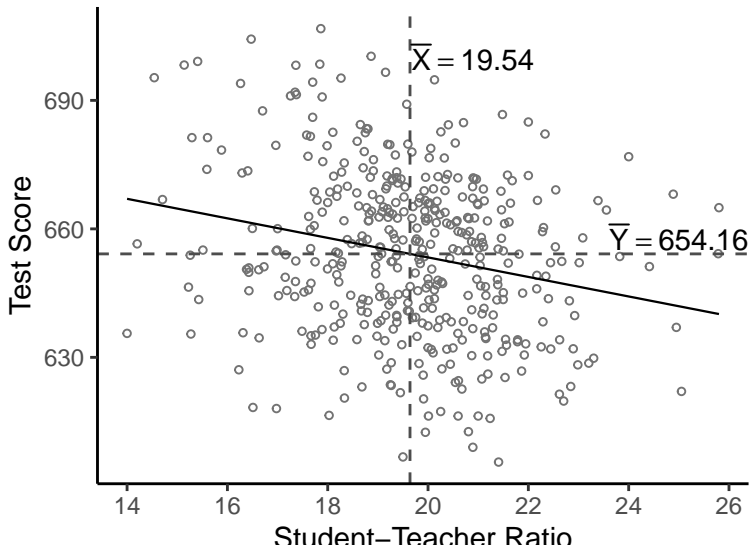
Some calculus reveals:

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

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

→ The best-fit line passes through the sample means!

# OLS line passes through the means



# Prediction and Residuals

OLS fitted line/predicted values:

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

Residuals/prediction error (the one we minimized):

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

# Prediction and Residuals

For our example, the fitted line:

$$\widehat{testscr} = 698.93 - 2.28 \cdot str$$

What is the predicted test score for a school with a student-teacher ratio of 24?

What is the prediction error for a school with a student-teacher ratio of 24 and average test score of 677?

## Goodness of Fit: The $R^2$

- $R$ -squared measures how well the OLS regression line fits the data
- $R$ -squared is the percent of sample variation in  $Y$  that is explained by  $X$

Note that:

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

$R^2$  is the ratio of sample variation of  $\hat{Y}_i$  to sample variation of  $Y_i$

# Goodness of Fit: The $R^2$

*Total Sum of Squares:*

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

*Explained Sum of Squares:*

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

*Residual Sum of Squares:*

$$RSS = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n \hat{u}_i^2$$

## Goodness of Fit: The $R^2$

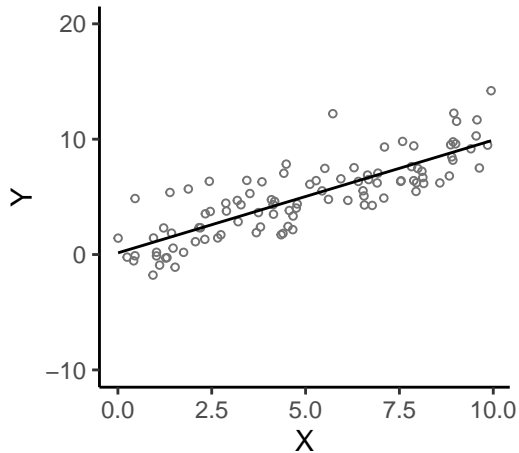
One can show that,  $TSS = ESS + RSS$ .

A measure of goodness of fit:

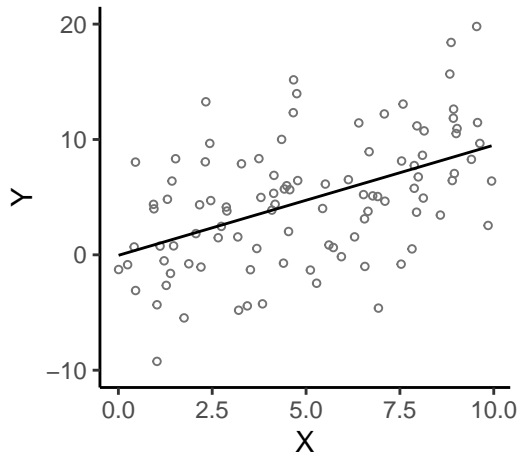
$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

# Goodness of Fit: The $R^2$

High  $R^2$



Low  $R^2$





## Goodness of Fit: The $R^2$

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

$R^2$  lies between 0 and 1

- If  $X$  explains no variation in  $Y$ ,  $\hat{\beta}_1 = 0$  and  $\hat{Y}_i = \hat{\beta}_0 = \bar{Y}$ . In which case,  $ESS = 0$  and hence  $R^2 = 0$ .
- On the other hand, if  $X$  explains all the variation in  $Y$ ,  $\hat{Y}_i = Y_i$  and  $RSS = 0$ . In which case,  $R^2 = 1$ .

# How to interpret the coefficients?

Fitted line:

$$\widehat{testscr} = 698.93 - 2.28 \cdot str$$

- Intercept: Predicted test score is 698.93 for a school with  $str = 0$ . (Doesn't always make sense!)
- Slope: One more student per teacher lowers the predicted test score by 2.28. How?

*Alternatively:* Schools in our sample that had one more student per teacher on average had an average test score that was 2.28 points lower.