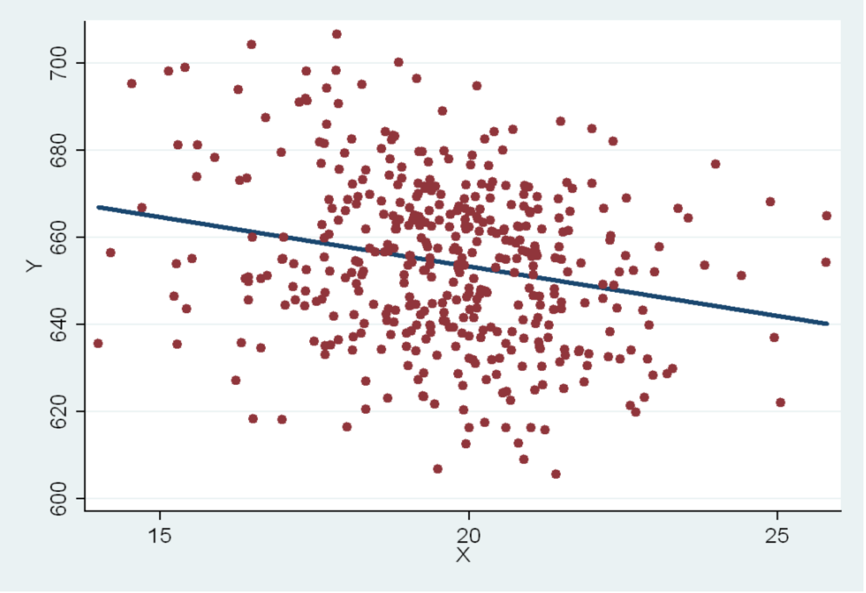
EC 3303: Econometrics I

Linear Regression with One Regressor (Part 1)



Kelvin Seah

AY 2022/2023, Semester 2

Outline

- Linear regression model
- OLS estimator and the sample regression line
- Measures of fit
- Least squares assumptions
- Sampling distribution of the OLS estimator

Today

Back to policy question:

Principal: "If I reduce class size by one student, what will the effect on student performance be?"

Introduction

- Question involves identifying the unknown effect of changing one variable, X, on another variable, Y.

- Introduce the linear regression model relating one variable, X, to another variable, Y…

  - model postulates a linear relationship between X and Y.

Linear Regression Model

- Restate policy question:
  - If class size is changed by a certain amount, what would we expect the change in test scores to be?

$$\beta_{classsize} = \frac{\Delta TestScore}{\Delta ClassSize} \qquad (1)$$

  - If we know $\beta_{classsize}$, will know how decreasing class size by one student will change test scores in the district.
  - (1) is the slope of a straight line relating test scores and class size.

- Straight line relating Y to X has form:
$$Y = mX + C$$

- here, the straight line:

$$TestScore = \beta_0 + \beta_{ClassSize} \times ClassSize \qquad (2)$$

where $\beta_{ClassSize}$ is the slope, β_0 is the intercept

- If we know $\beta_{ClassSize}$ & β_0, can:
  - determine the change in test scores at a district associated with a change in class size
  - predict the test score itself for a given class size
- But something is wrong!

$$TestScore = \beta_0 + \beta_{ClassSize} \times ClassSize \quad (2)$$

- Class size is just one factor potentially affecting test scores. Two districts with the same class size may have different test scores for other reasons.
- Other factors affecting test score may include:
  - Teacher quality
  - Student demographics (wealth, proportion of immigrants)
  - Luck…
- (2) will not hold exactly for all districts; (2) is a relationship that holds ***on average*** across the population of districts.

- A version of the relationship between test scores and class size ***that holds for each district*** must incorporate the other factors influencing test scores.

- Lumping all these other factors, the relationship that holds for each district is:

$$TestScore = \beta_0 + \beta_{ClassSize} \times ClassSize + other\ factors \qquad (3)$$

$$TestScore = \beta_0 + \beta_{ClassSize} \times ClassSize + other\ factors \quad (3)$$

- express (3) using general notation:

Suppose we have n districts. Let Y_i be the test score in the i^{th} district, X_i be the class size in the i^{th} district, u_i denote the other factors influencing the test score in the i^{th} district. Then,

$$Y_i = \beta_0 + \beta_1 X_i + u_i \quad (4)$$

for each district ($i = 1, \ldots, n$)

Population Linear Regression Model

$$Y_i = \beta_0 + \beta_1 X_i + u_i \quad (4)$$

n observations, (X_i, Y_i), $i = 1,.., n$

Y: dependent variable

X: independent variable / regressor

$\beta_0 + \beta_1 X$: population regression line / function

β_0: intercept

β_1: slope

u_i: population error

More terminology

$$Y_i = \beta_0 + \beta_1 X_i + u_i \quad (4)$$

- $\beta_0 + \beta_1 X$ is a straight line that describes how Y changes as X changes.

- $\beta_0 + \beta_1 X$ tells us the relationship that holds between Y and X ***on average*** over the population.

- β_0 & β_1 are ***coefficients*** / ***parameters*** of the population regression line.

- β_1 is the expected effect on Y of a unit change in X .

- β_0 is the value of the population regression line when $X = 0$ (may or may not have real-world meaning).

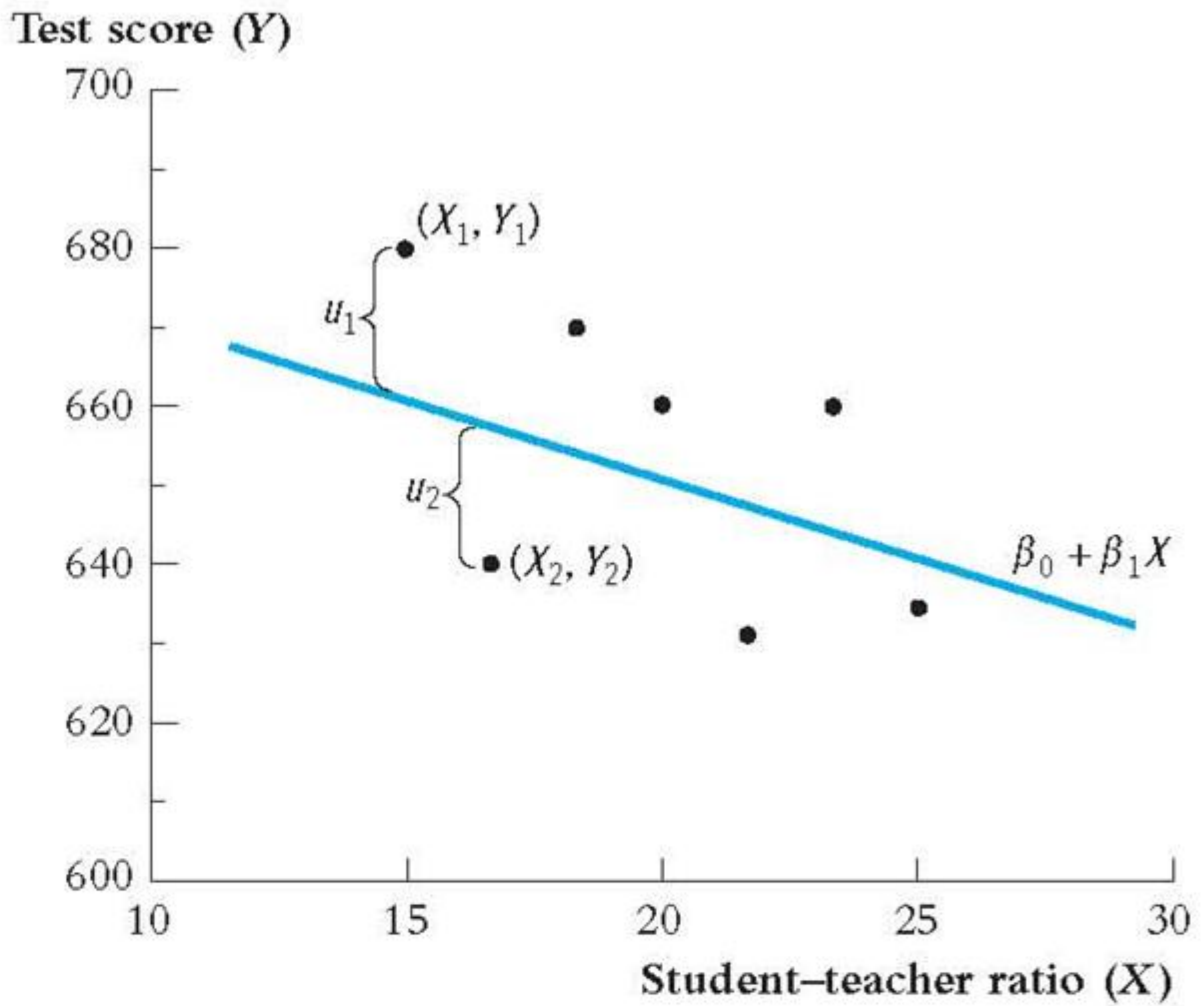
$$Y_i = \beta_0 + \beta_1 X_i + u_i \quad (4)$$

- u_i incorporates all of the factors responsible for the difference between the i^{th} district's test score and the value predicted by the population regression line.

- u_i contains all other factors (omitted factors) besides X that influence the value of Y, for a specific observation, i.

- u_i includes all the unique characteristics of the i^{th} district that affect the performance of its students.

Population regression model in a picture: Observations on *Y* and *X*; the population regression line; and the regression error



$$\beta_1 < 0$$

Districts with lower STR tend to have higher test scores.

- Interpretation of u_i.
- Why do the observations not fall exactly on the population regression line?

Questions to answer

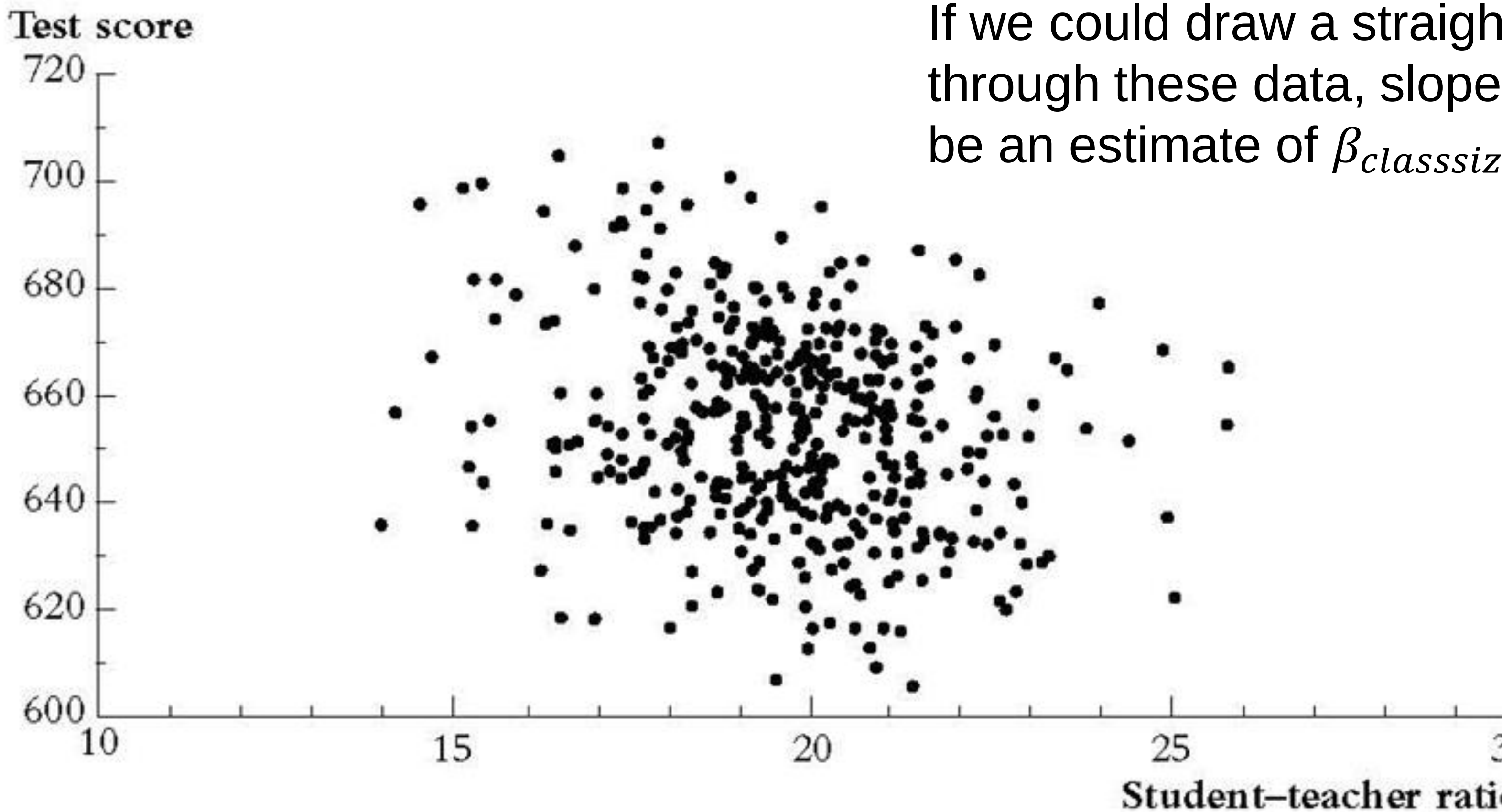
- Estimation:
  - We want to know the population value of β_1 (and β_0). But they are unknown parameters.
  - How do we use data to estimate the unknown slope β_1 and intercept β_0 of the population regression line?
  - Answer: ordinary least squares (OLS)
- Hypothesis testing:
  - How to test if the population slope is zero?
- Confidence intervals :
  - How to construct a confidence interval for the population slope?

Estimating β_1 & β_0

- Recall: to learn about the population mean, we used a random sample of data drawn from that population.
- Similarly, can learn about $\beta_{classsize}$ using a random sample of data.
- California Test Score data
  - Data set: test scores & class sizes in 1999 in 420 California school districts (caschool.dta)
  - test score is the district average of reading and math score for fifth graders.
  - goal: estimate the ***linear association*** between class size and test score.
  - Variables: testscr – test score

    str – student to teacher ratio

Scatterplot of Test Score vs. Student-Teacher Ratio (California School District Data)



If we could draw a straight line through these data, slope would be an estimate of $\beta_{classsize}$

Sample correlation is -0.23: weak negative linear relationship between test scores and class size.

Ordinary Least Squares (OLS) Estimator

- use the most common approach – choose the line that produces the "least squares" fit to the data:

  - use the *Ordinary Least Squares (OLS)* estimator.

- OLS chooses the estimators so that the estimated regression line comes as close as possible to the data points.

  - where "closeness" is measured by the sum of the squared mistakes made in predicting Y given X.

- Let b_0 and b_1 be some estimators of β_0 and β_1. Regression line based on these estimators is $b_0 + b_1 X$.

- value of Y_i predicted using this line is $b_0 + b_1 X_i$ for $i = 1, \dots n$.

- mistake in predicting the i^{th} observation is
$$Y_i - (b_0 + b_1 X_i)$$

- ***sum*** of the ***squared prediction mistakes*** over all n observations is
$$\sum_{i=1}^{n} [Y_i - (b_0 + b_1 X_i)]^2$$

- least squares method chooses the estimators which ***minimize the sum of squared mistakes***

$$\min_{b_0, b_1} \sum_{i=1}^{n} [Y_i - (b_0 + b_1 X_i)]^2 \qquad (5)$$

- estimators that solve (5) are called the ordinary least squares (OLS) estimators.

- minimization problem is solved using calculus.

How to derive the OLS estimators?

OLS Terminology

- $\hat{\beta}_0$: OLS estimator of β_0

- $\hat{\beta}_1$: OLS estimator of β_1

- $\hat{\beta}_0 + \hat{\beta}_1 X$: OLS regression line / function

- $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$: predicted value of Y_i given X_i, based on the OLS regression line

- $\hat{u}_i = Y_i - \hat{Y}_i$: residual for the i^{th} observation

OLS Notation & Terminology

Population	
β_0	intercept of population regression line
β_1	slope of population regression line
$\beta_0 + \beta_1 X$	population regression line
u_i	population error

Sample	
$\hat{\beta}_0$	intercept of OLS regression line
$\hat{\beta}_1$	slope of OLS regression line
$\hat{\beta}_0 + \hat{\beta}_1 X$	OLS regression line
$\hat{u}_i$	OLS residual

The OLS Estimator, Predicted Values, and Residuals

The OLS estimators of the slope β_1 and the intercept β_0 are

Sample covariance between Y and X

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{s_{XY}}{s_X^2} \tag{4.7}$$

Sample variance of X

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1\overline{X}. \tag{4.8}$$

OLS regression line always passes through the point $(\overline{X}, \overline{Y})$

The OLS predicted values $\hat{Y}_i$ and residuals $\hat{u}_i$ are

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i,\ i = 1, \ldots, n \tag{4.9}$$

$$\hat{u}_i = Y_i - \hat{Y}_i,\ i = 1, \ldots, n. \tag{4.10}$$

The estimated intercept $(\hat{\beta}_0)$, slope $(\hat{\beta}_1)$, and residual $(\hat{u}_i)$ are computed from a sample of n observations of X_i and Y_i, $i = 1, \ldots, n$. These are estimates of the unknown true population intercept (β_0), slope (β_1), and error term (u_i).

California Test Score Data

Data Editor (Browse) - [caschool]

	district	testscr	str	avginc
1	Ackerman Elementary	664.15	19.93548	16.272
2	Adelanto Elementary	639.35	19.75422	8.384
3	Alexander Valley Union Elementary	660.05	17	18.326
4	Alisal Union Elementary	623.2	22.84553	7.983181
5	Allensworth Elementary	616.3	20.6	5.335
6	Alpine Union Elementary	667.15	20.89195	17.332
7	Alta Loma Elementary	664.4	23.55637	18.6202
8	Alta Vista Elementary	622.05	25.05263	9.63
9	Alta-Dutch Flat Union Elementary	673.55	18.90909	13.443
10	Alum Rock Union Elementary	630.35	21.18829	12.58158
11	Alview-Dairyland Union Elementary	640.5	18.79121	10.472
12	Alvina Elementary	655.05	16.99029	10.338
13	American Union Elementary	642.4	18.66072	7.385
14	Anaheim Elementary	633.15	21.94756	13.40063
15	Antelope Elementary	657.75	19.33333	11.426
16	Arcata Elementary	664.3	20.14178	11.834
17	Arcohe Union Elementary	646.2	19.96154	15.131
18	Arena Union Elementary	639.3	20.36254	10.035
19	Armona Union Elementary	633.9	19.14	9.082
20	Arvin Union Elementary	623.45	19.26667	7.305
21	Atwater Elementary	635.95	21.12796	10.676
22	Auburn Union Elementary	655.55	21.01796	16.272
23	Bakersfield City Elementary	630.55	19.01749	12.10913
24	Ballard Elementary	682.15	22.33333	21.957

Ready | Vars: 4 of 18 | Obs: 420 | Filter: Off | Mode: Browse

Example: OLS Regression – Stata command and output

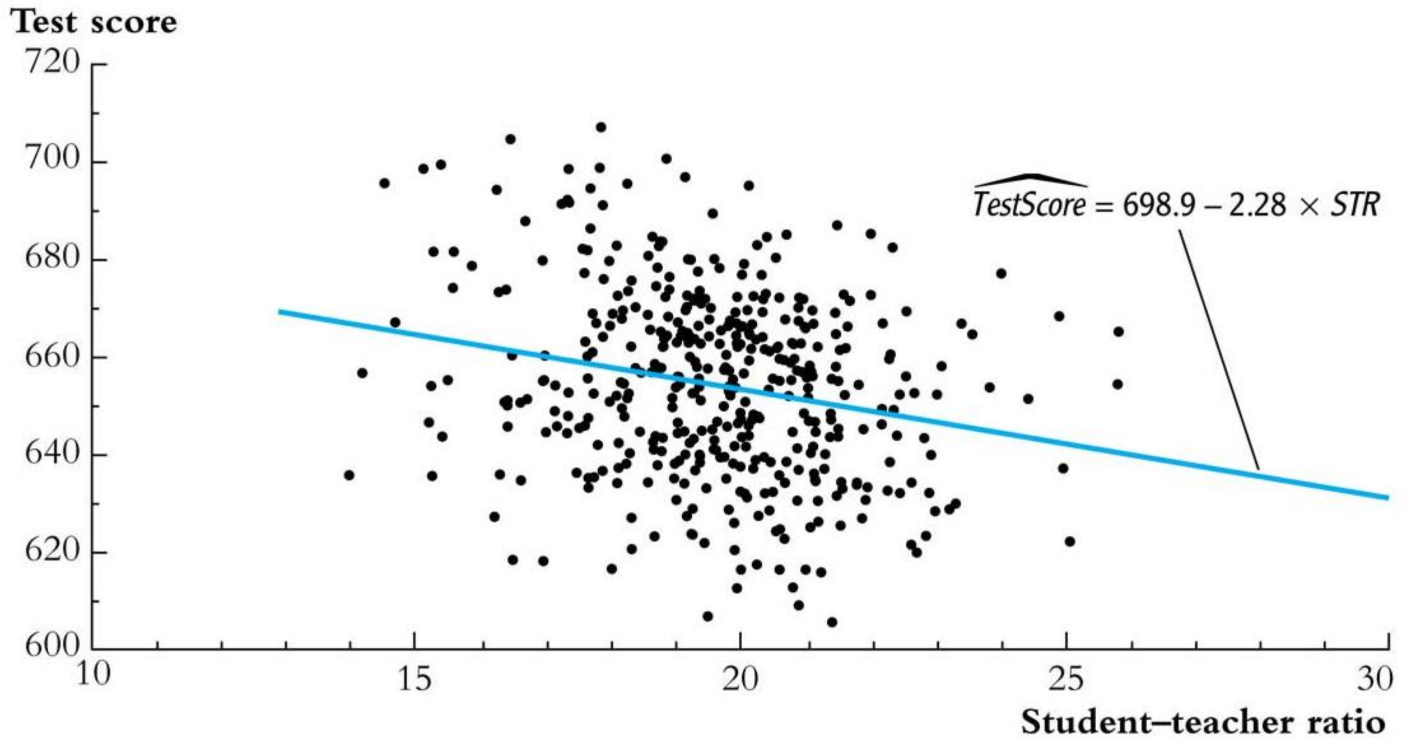
Using OLS to estimate a line relating STR to test scores with the 420 observations, we get:

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

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

(we'll discuss the rest of this output later)

Application to the California *Test Score – Class Size* data



Estimated slope = $\hat{\beta}_1$ = – 2.28

Estimated intercept = $\hat{\beta}_0$ = 698.9

Estimated regression line: 698.9 – 2.28×*STR*

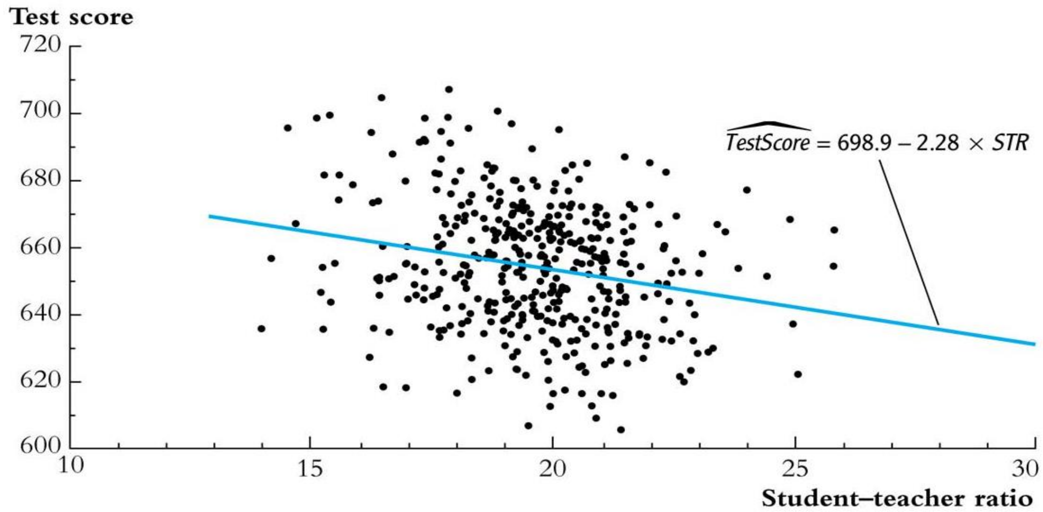
Interpretation

- OLS regression line for the 420 observations is:

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

- "^" over TestScore indicates it is the predicted value (of TestScore) based on the OLS regression line
- Districts with one more student per teacher ***on average*** have test scores that are 2.28 points lower.
- intercept (taken literally) means that, districts with zero students per teacher would have a predicted test score of 698.9.
- interpretation of the intercept makes no sense – it extrapolates the line outside the range of the data – here, the intercept is not economically meaningful.

Predicted Values & Residuals



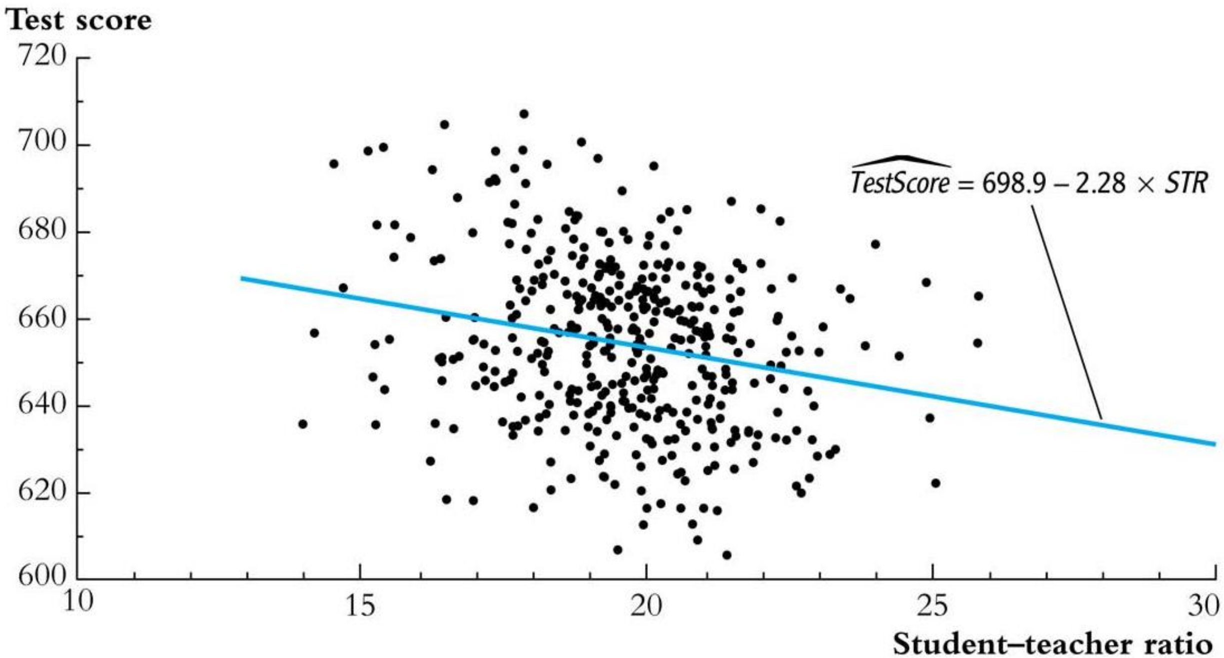
One of the districts in the data set is Antelope, CA, for which *STR* = 19.33 and *Test Score* = 657.8

Predicted value: $\hat{Y}_{Antelope}$ = 698.9 – 2.28×19.33 = 654.8

Residual: $\hat{u}_{Antelope}$ = 657.8 – 654.8 = 3.0

Predicted value will not be exactly right because of the other factors that determine a district's testscore

Predicted Values & Residuals



- Do not use the regression line for prediction far outside the range of values of the independent variable used to obtain the line.
- Such predictions are typically inaccurate:
  - what is the predicted test score for a district with $STR = 500$?

Announcement

- Tutorials begin next week (week 3, i.e. week starting 23 Jan) for students in the D groups.

- Note that 23 January (Monday) and 24 January (Tuesday) are public holidays. If your tutorial class happen to fall on these days, please attend any other tutorial slot in week 3 or week 4 as a make-up. Please keep your assigned tutor informed of your make-up tutorial attendance and participation when you meet your tutor during your second tutorial.

- For students in the E groups, your first tutorial will be in week 4.

- Tutorials are held at AS4-0110.

- The tutorial questions will be made available in Canvas Files --> Tutorials --> Tutorial 1.