

# ECON 7310 Elements of Econometrics

## Week 2: Linear Regression with One Regressor

David Du<sup>1</sup>

<sup>1</sup>University of Queensland

Draft

## Outline:

- ▶ The population linear regression model (LRM)
- ▶ The ordinary least squares (OLS) estimator and the sample regression line
- ▶ Measures of fit of the sample regression
- ▶ The least squares assumptions
- ▶ The sampling distribution of the OLS estimator

# Linear Regression

- ▶ Linear regression lets us estimate the slope of the population regression line.
- ▶ The slope of the population regression line is the expected effect on  $Y$  of a unit change in  $X$ .
- ▶ Ultimately our aim is to estimate the causal effect on  $Y$  of a unit change in  $X$  – but for now, just think of the problem of fitting a straight line to data on two variables,  $Y$  and  $X$ .

# Linear Regression

- ▶ The problem of statistical inference for linear regression is, at a general level, the same as for estimation of the mean or of the differences between two means.
- ▶ Statistical, or econometric, inference about the slope entails:
  - ▶ Estimation:  
How should we draw a line through the data to estimate the population slope? Answer: ordinary least squares (OLS).  
What are advantages and disadvantages of OLS?
  - ▶ Hypothesis testing:  
How to test if the slope is zero?
  - ▶ Confidence intervals:  
How to construct a confidence interval for the slope?

- ▶ The population regression line:

$$\text{Test Score} = \beta_0 + \beta_1 \text{ STR}$$

- ▶  $\beta_1$  = slope of population regression line  
= change in test score for a unit change in student-teacher ratio (STR)
- ▶ Why are  $\beta_0$  and  $\beta_1$  “population” parameters?
- ▶ We would like to know the population value of  $\beta_1$ .
- ▶ We don't know  $\beta_1$ , so must estimate it using data.

# The Population Linear Regression Model

Consider

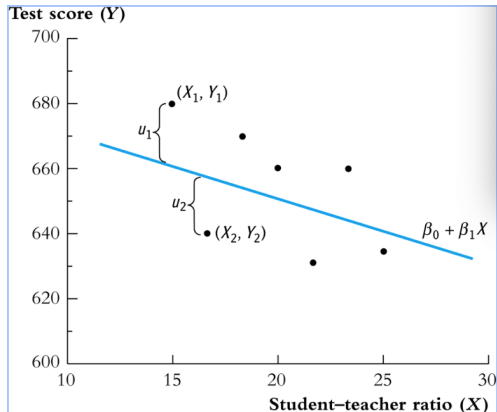
$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

for  $i = 1, \dots, n$

- ▶ We have  $n$  observations,  $(X_i, Y_i)$ ,  $i = 1, \dots, n$ .
- ▶  $X$  is the independent variable or regressor or right-hand-side variable
- ▶  $Y$  is the dependent variable or left-hand-side variable
- ▶  $\beta_0 = \textit{intercept}$
- ▶  $\beta_1 = \textit{slope}$
- ▶  $u_i$  = the regression error
- ▶ The regression error consists of omitted factors. In general, these omitted factors are other factors that influence  $Y$ , other than the variable  $X$ . The regression error also includes error in the measurement of  $Y$ .

# The population regression model in a picture

- Observations on  $Y$  and  $X$  ( $n = 7$ ); the population regression line; and the regression error (the “error term”):



# The Ordinary Least Squares Estimator (SW Section 4.2)

- ▶ How can we estimate  $\beta_0$  and  $\beta_1$  from data? Recall that was the least squares estimator of  $\mu_Y$ : solves,  $\bar{Y}$

$$\min_m \sum_{i=1}^n (Y_i - m)^2$$

- ▶ By analogy, we will focus on the least squares (“ordinary least squares” or “OLS”) estimator of the unknown parameters  $\beta_0$  and  $\beta_1$ . The OLS estimator solves,

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

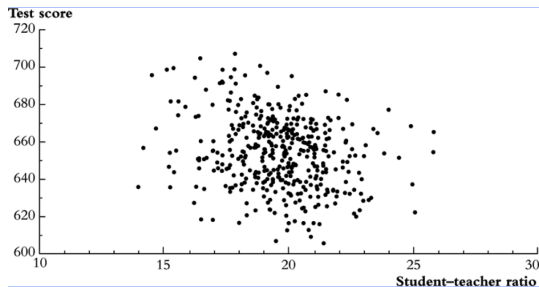
- ▶ In fact, we estimate the conditional expectation function  $E[Y|X]$  under the assumption that  $E[Y|X] = \beta_0 + \beta_1 X$



# Mechanics of OLS

- The population regression line:

$$\text{Test Score} = \beta_0 + \beta_1 \text{STR}$$



# Mechanics of OLS

- ▶ The OLS estimator minimizes the average squared difference between the actual values of  $Y_i$  and the prediction (“predicted value”) based on the estimated line.
- ▶ This minimization problem can be solved using calculus (Appendix 4.2).
- ▶ The result is the OLS estimators of  $\beta_0$  and  $\beta_1$ .

# OLS estimator, predicted values, and residuals

- ▶ The OLS estimators are

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

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

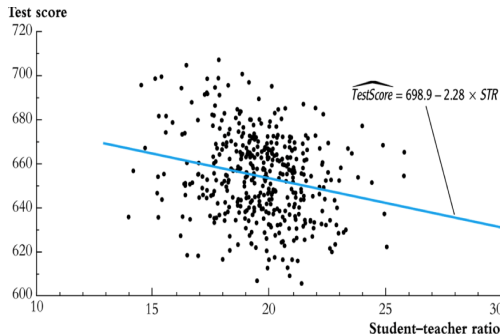
- ▶ The OLS predicted (fitted) values  $\hat{Y}_i$  and residuals  $\hat{u}_i$  are

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

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

- ▶ The estimated intercept,  $\hat{\beta}_0$ , and slope,  $\hat{\beta}_1$ , and residuals  $\hat{u}_i$  are computed from a sample of  $n$  observations  $(X_i, Y_i)$   $i = 1, \dots, n$ .
- ▶ These are estimates of the unknown population parameters  $\beta_0$  and  $\beta_1$ .

## Predicted values & residuals



- One of the districts in the data set is Antelope, CA, for which  $STR = 19.33$  and  $TestScore = 657.8$

$$\text{predicted value:} = 698.9 - 2.28 \times 19.33 = 654.8$$

$$\text{residual:} = 657.8 - 654.8 = 3.0$$

# OLS regression: Stata output

```
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					
<u>testscr</u>	<u>Coef.</u>	Std. Err.	t	P> t	[95% Conf. Interval]	
<u>str</u>	-2.279808	.5194892	-4.39	0.000	-3.300945	-1.258671
<u>_cons</u>	698.933	10.36436	67.44	0.000	678.5602	719.3057

*Test Score* = 698.9 - 2.28 × *STR*

(We'll discuss the rest of this output later.)

- ▶ Two regression statistics provide complementary measures of how well the regression line “fits” or explains the data:
- ▶ The **regression  $R^2$**  measures the fraction of the variance of  $Y$  that is explained by  $X$ ; it is unit free and ranges between zero (no fit) and one (perfect fit)
- ▶ The **standard error of the regression (SER)** measures the magnitude of a typical regression residual in the units of  $Y$ .

## Regression $R^2$

- ▶ The sample variance of  $Y_i = \frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y})^2$   
The sample variance of  $\hat{Y}_i = \frac{1}{n} \sum_{i=1}^n (\hat{Y}_i - \bar{\hat{Y}})^2$ , where in fact  $\bar{\hat{Y}} = \bar{Y}$ .  
 $R^2$  is simply the ratio of those two sample variances.
- ▶ Formally, we define  $R^2$  as follows (two equivalent definitions);

$$R^2 := \frac{\text{Explained Sum of Squares (ESS)}}{\text{Total Sum of Squares (TSS)}} = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$$

$$R^2 := 1 - \frac{\text{Residual Sum of Squares (RSS)}}{\text{Total Sum of Squares}} = 1 - \frac{\sum_{i=1}^n \hat{u}_i^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$$

- ▶  $R^2 = 0 \iff ESS = 0$  and  $R^2 = 1 \iff ESS = TSS$ . Also,  $0 \leq R^2 \leq 1$
- ▶ For regression with a single  $X$ ,  
 $R^2$  = the square of the sample correlation coefficient between  $X$  and  $Y$

# The Standard Error of the Regression (SER)

- ▶ The SER measures the spread of the distribution of  $u$ . The SER is (almost) the sample standard deviation of the OLS residuals:?

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

- ▶ The SER:
  - ▶ has the units of  $u_i$ , which are the units of  $Y_i$
  - ▶ measures the average “size” of the OLS residual (the average “mistake” made by the OLS regression line)
- ▶ The **root mean squared error (RMSE)** is closely related to the SER:

$$RMSE := \sqrt{\frac{1}{n} \sum_{i=1}^n \hat{u}_i^2}$$

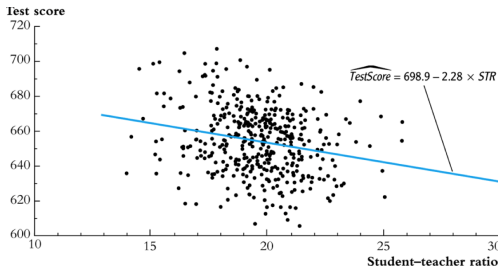
- ▶ When  $n$  is large,  $SER \approx RMSE$ .<sup>1</sup>

---

<sup>1</sup>Here,  $n - 2$  is the degrees of freedom – need to subtract 2 because there are two parameters to estimate. For details, see section 18.4.



## Example of the $R^2$ and the $SER$



- ▶  $TestScore = 698.9 - 2.28 \times STR$ ,  $R^2 = 0.05$ ,  $SER = 18.6$
- ▶  $STR$  explains only a small fraction of the variation in test scores.
  - ▶ Does this make sense?
  - ▶ Does this mean the  $STR$  is unimportant in a policy sense?

## Least Squares Assumptions (SW Section 4.4)

- ▶ What, in a precise sense, are the properties of the sampling distribution of the OLS estimator? When will it be unbiased? What is its variance?
- ▶ To answer these questions, we need to make some assumptions about how  $Y$  and  $X$  are related to each other, and about how they are collected (the sampling scheme)
- ▶ These assumptions – there are three – are known as the Least Squares Assumptions.

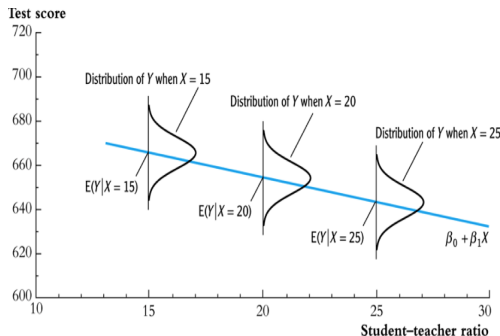
## Least Squares Assumptions (SW Section 4.4)

$$Y_i = \beta_0 + \beta_1 X_i + u_i, \quad i = 1, \dots, n$$

1. The conditional distribution of  $u$  given  $X$  has mean zero, that is,  $E(u|X = x) = 0$ .
  - ▶ This implies that OLS estimators are unbiased
2.  $(X_i, Y_i)$ ,  $i = 1, \dots, n$ , are i.i.d.
  - ▶ This is true if  $(X, Y)$  are collected by simple random sampling
  - ▶ This delivers the sampling distribution of  $\hat{\beta}_0$  and  $\hat{\beta}_1$
3. Large outliers in  $X$  and/or  $Y$  are rare.
  - ▶ Technically,  $X$  and  $Y$  have finite fourth moments
  - ▶ Outliers can result in meaningless values of  $\hat{\beta}_1$

## Least squares assumption #1: $E(u|X = x) = 0$ .

For any given value of  $X$ , the mean of  $u$  is zero:



Example:  $TestScore_i = \beta_0 + \beta_1 STR_i + u_i$ ,  $u_i$  = other factors

- ▶ What are some of these “other factors”?
- ▶ Is  $E(u|X = x) = 0$  plausible for these other factors?

## Least squares assumption #1: $E(u|X = x) = 0$ (continued)

- ▶ A benchmark for thinking about this assumption is to consider an ideal randomized controlled experiment:
- ▶  $X$  is randomly assigned to people (students randomly assigned to different size classes; patients randomly assigned to medical treatments). Randomization is done by computer – using no information about the individual.
- ▶ Because  $X$  is assigned randomly, all other individual characteristics – the things that make up  $u$  – are distributed independently of  $X$ , so  $u$  and  $X$  are independent
- ▶ Thus, in an ideal randomized controlled experiment,  $E(u|X = x) = 0$  (that is, LSA #1 holds)
- ▶ In actual experiments, or with observational data, we will need to think hard about whether  $E(u|X = x) = 0$  holds.

## Least squares assumption #2: $(X_i, Y_i), i = 1, \dots, n$ are i.i.d.

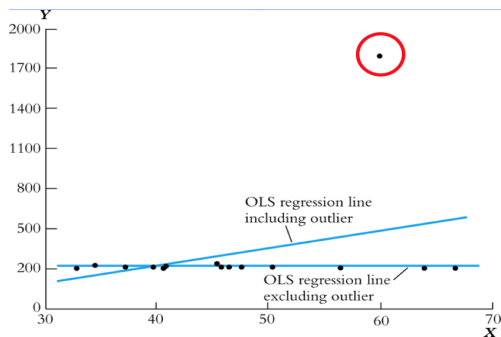
- ▶ This arises automatically if the entity (individual, district) is sampled by simple random sampling:
  - ▶ The entities are selected from the same population, so  $(X_i, Y_i)$  are identically distributed for all  $i = 1, \dots, n$ .
  - ▶ The entities are selected at random, so the values of  $(X, Y)$  for different entities are independently distributed.
- ▶ The main place we will encounter non-i.i.d. sampling is when data are recorded over time for the same entity (panel data and time series data)
  - we will deal with that complication when we cover panel data.

## Least squares assumption #3: Large outliers are rare

Technical statement:  $E(X^4) < \infty$  and  $E(Y^4) < \infty$

- ▶ A large outlier is an extreme value of  $X$  or  $Y$
- ▶ On a technical level, if  $X$  and  $Y$  are bounded, then they have finite fourth moments. (Standardized test scores automatically satisfy this; *STR*, family income, etc. satisfy this too.)
- ▶ The substance of this assumption is that a large outlier can strongly influence the results – so we need to rule out large outliers.
- ▶ Look at your data! If you have a large outlier, is it a typo? Does it belong in your data set? Why is it an outlier?

## OLS can be sensitive to an outlier:



- ▶ Is the lone point an outlier in  $X$  or  $Y$ ?
- ▶ In practice, outliers are often data glitches (coding or recording problems). Sometimes they are observations that really shouldn't be in your data set. Plot your data before running regressions!



## The Sampling Distribution of the OLS Estimator (SW Section 4.5)

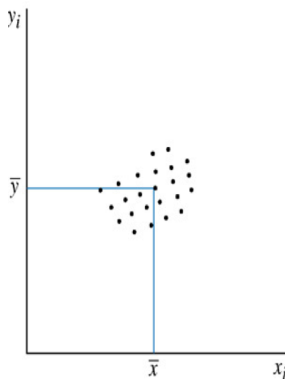
The OLS estimator is computed from a sample of data. A different sample yields a different value of  $\hat{\beta}_1$ . This is the source of the “sampling uncertainty” of  $\hat{\beta}_1$ . We want to:

- ▶ quantify the sampling uncertainty associated with
- ▶ use  $\hat{\beta}_1$  to test hypotheses such as  $\beta_1 = 0$
- ▶ construct a confidence interval for  $\beta_1$
- ▶ All these require figuring out the sampling distribution of the OLS estimator.

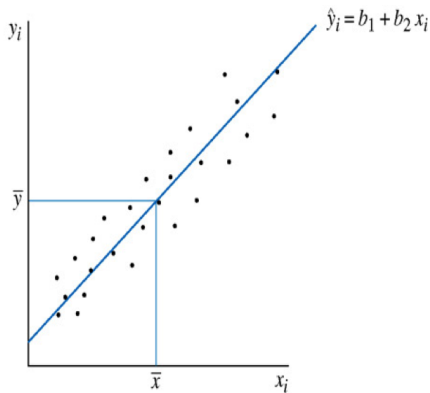
# Sampling Distribution of $\hat{\beta}_1$

- ▶ We can show that  $\hat{\beta}_1$  is unbiased, i.e.,  $E[\hat{\beta}_1] = \beta_1$ . Similarly for  $\hat{\beta}_0$ .
- ▶ We do not derive  $V(\hat{\beta}_1)$ , as it requires some tedious algebra. Moreover, we do not need to memorize the formula of it. Here, we just emphasize two aspects of  $V(\hat{\beta}_1)$ .
- ▶ First,  $V(\hat{\beta}_1)$  is inversely proportional to  $n$ , just like  $V(\bar{Y}_n)$ . Combining  $E[\hat{\beta}_1] = \beta_1$ , it is then suggested that  $\hat{\beta}_1 \xrightarrow{P} \beta_1$ , i.e.,  $\hat{\beta}_1$  is consistent. That is, as sample size grows,  $\hat{\beta}_1$  gets closer to  $\beta_1$ .
- ▶ Second,  $V(\hat{\beta}_1)$  is inversely proportional to the variance of  $X$ ; see the graphs below.

# Sampling Distribution of $\hat{\beta}_1$



Low  $x$  variation  
 $\Rightarrow$  low precision



High  $x$  variation  
 $\Rightarrow$  high precision

- Intuitively, if there is more variation in  $X$ , then there is more information in the data that you can use to fit the regression line.

# Sampling Distribution of $\hat{\beta}_1$

- ▶ The exact sampling distribution is complicated – it depends on the population distribution of  $(Y, X)$  – but when  $n$  is large we get some simple (and good) approximations:
- ▶ Let  $SE(\hat{\beta}_1)$  be the standard error (SE) of  $\hat{\beta}_1$ , i.e., a consistent estimator for the standard deviation of  $\hat{\beta}_1$  which is  $\sqrt{V(\hat{\beta}_1)}$
- ▶ Then, it turns out that

$$\frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} \underset{\text{approx}}{\sim} \mathcal{N}(0, 1)$$

- ▶ Using this approximate distribution, we can conduct statistical inference on  $\hat{\beta}_1$ , i.e., hypothesis testing, confidence interval  $\Rightarrow$  Ch5.