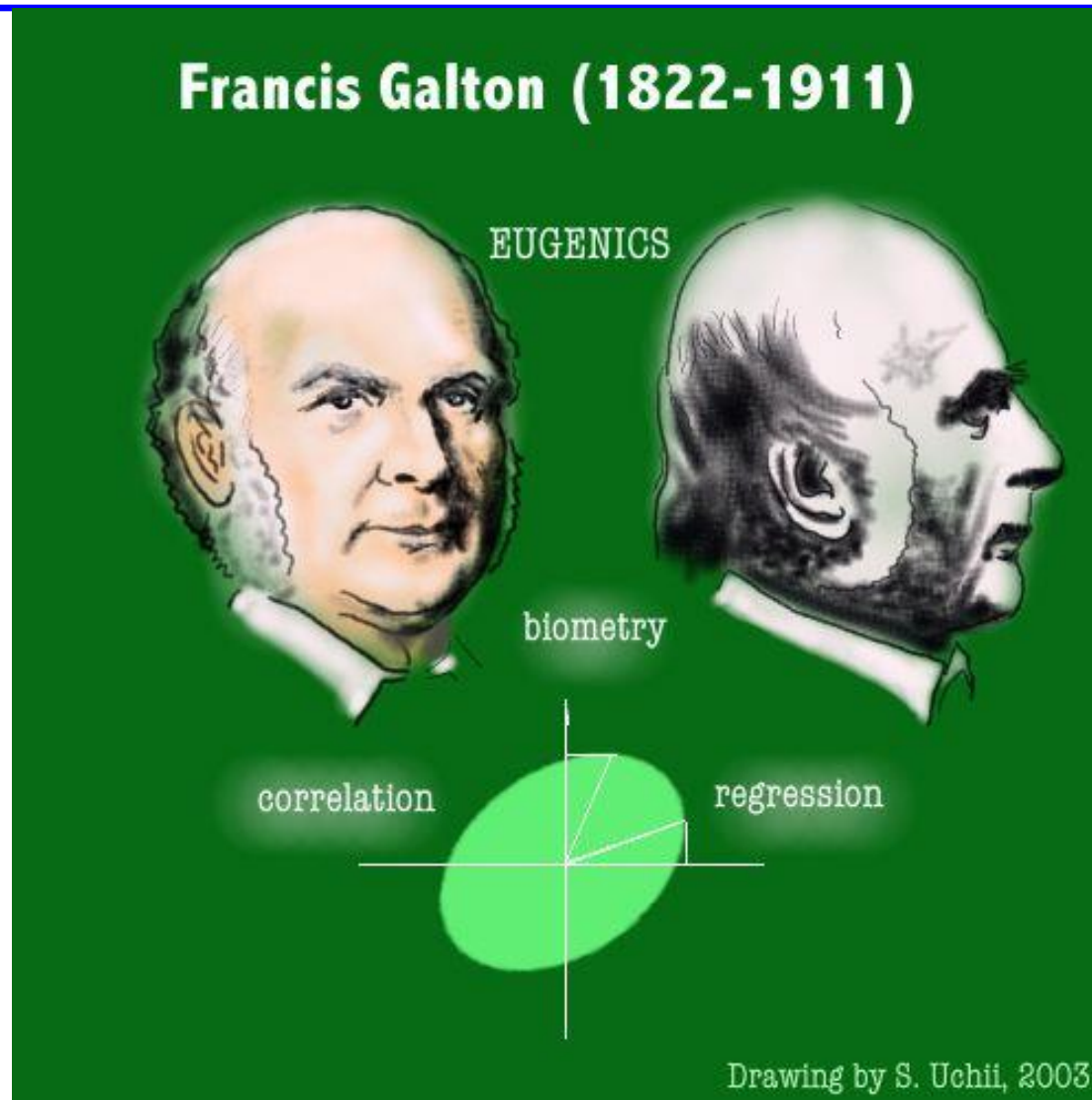


Topic 4: Simple Linear Regression-Estimation

Part A

Who Invented Regression ?

Econ 3334



1) Linear Regression Model

Econ 3334

- Here we are interested in estimating a linear equation that describes the relationship between 2 variables.
- A line is defined by its slope and intercept:

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

- E.g.,

$$\text{Quantity} = \beta_0 + \beta_1 \text{Price}$$

$$\text{Wage} = \beta_0 + \beta_1 \text{Education}$$

- If we can gather data on Quantity and Price, then we have information to use to estimate a linear statistical model.

1) Linear Regression Model

Econ 3334

- The basic linear single variable model is

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

- This equation says that Y_i is equal to $\beta_0 + \beta_1 X_i$ plus some errors.
- I.e., X_i does not perfectly explain each Y_i , rather Y_i is explained by X_i with some errors.
- Interpretation of the model: keeping everything else constant, increasing one unit X_i , Y_i will increase by β_1 .

1) Linear Regression Model

Econ 3334

- The basic linear single variable model is

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

- You have iid data (sample) :

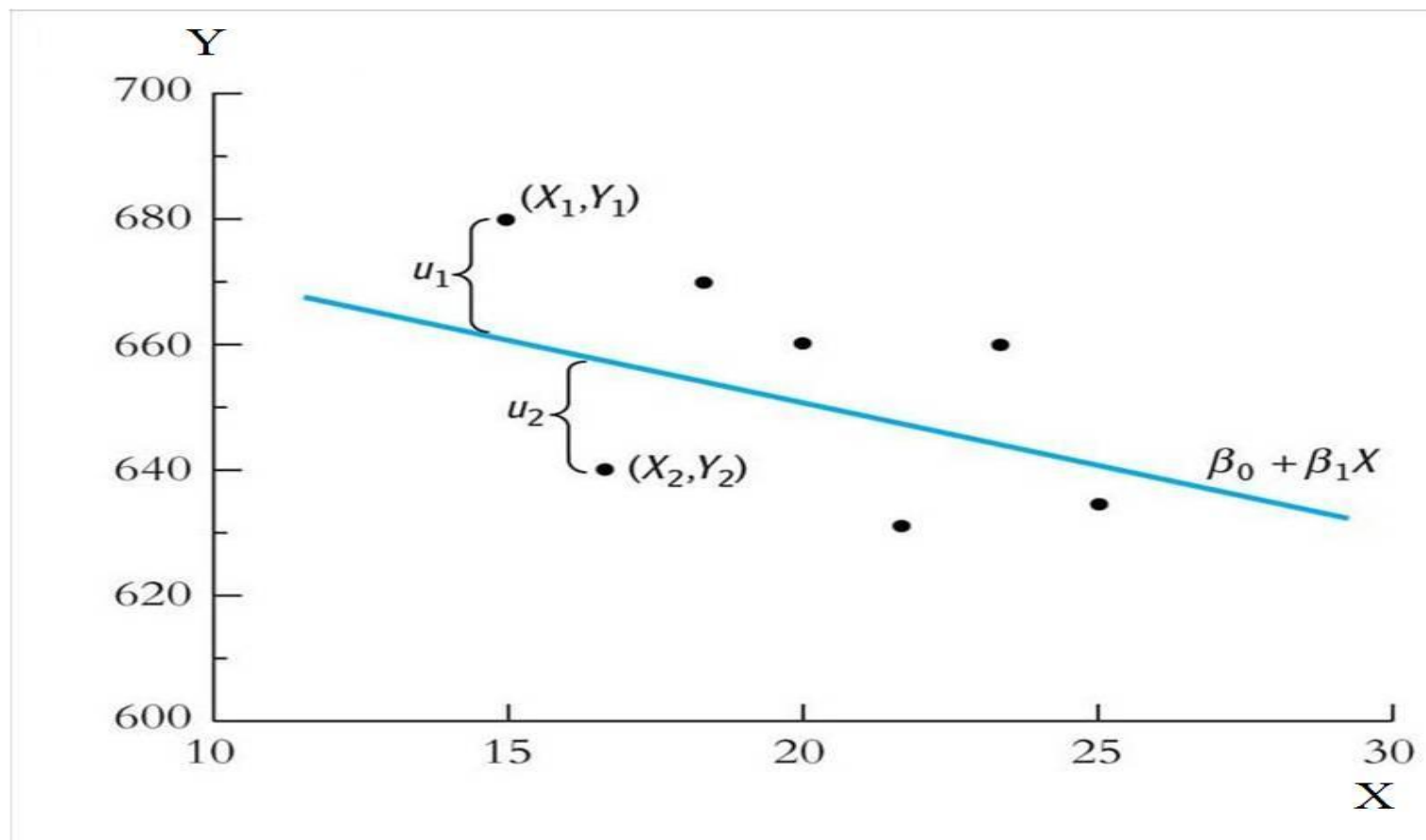
$$\{(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)\}$$

- Y_i is the dependent variable (random variable)
- X_i is the independent (or explanatory) variable (random variable)
- β_0 is the intercept or constant (unknown parameter, non-random)
- β_1 is the slope (unknown parameter, non-random)

1) Linear Regression Model

Econ 3334

- Imagine that you know true β_0 and β_1 , then

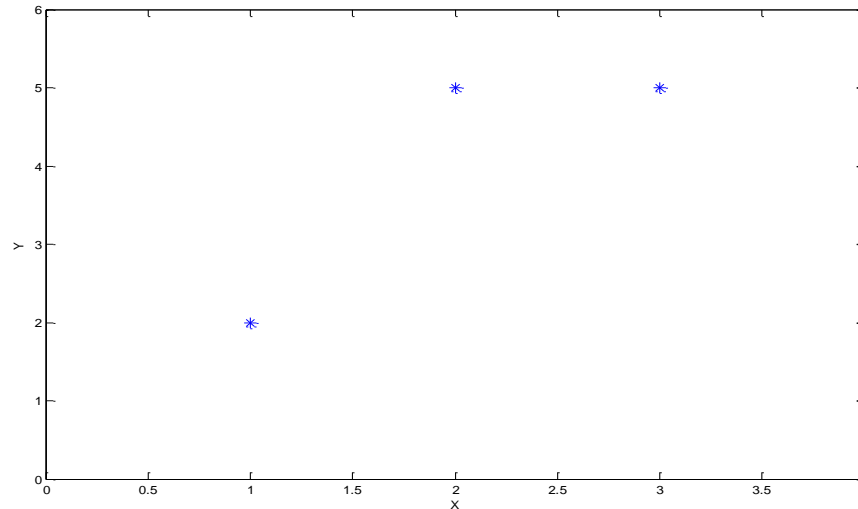


1) Linear Regression Model

Econ 3334

➤ E.g.

X_i	Y_i
1	2
2	5
3	5



➤ There is positive correlation between X_i and Y_i :

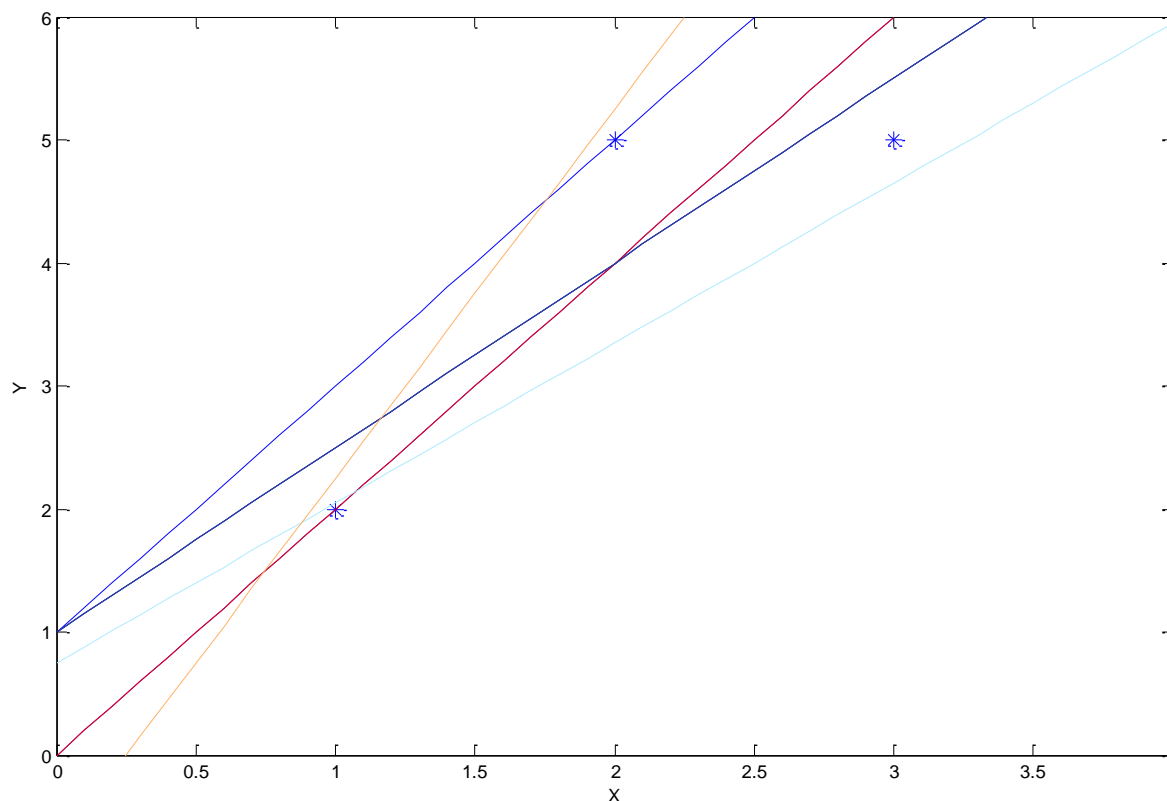
$$\text{Sample Correlation } (X, Y) = 0.866$$

➤ Our goal here is to come up with the equation of the line that links Y_i and X_i .

1) Linear Regression Model

Econ 3334

➤ Infinite ways to fit the line:

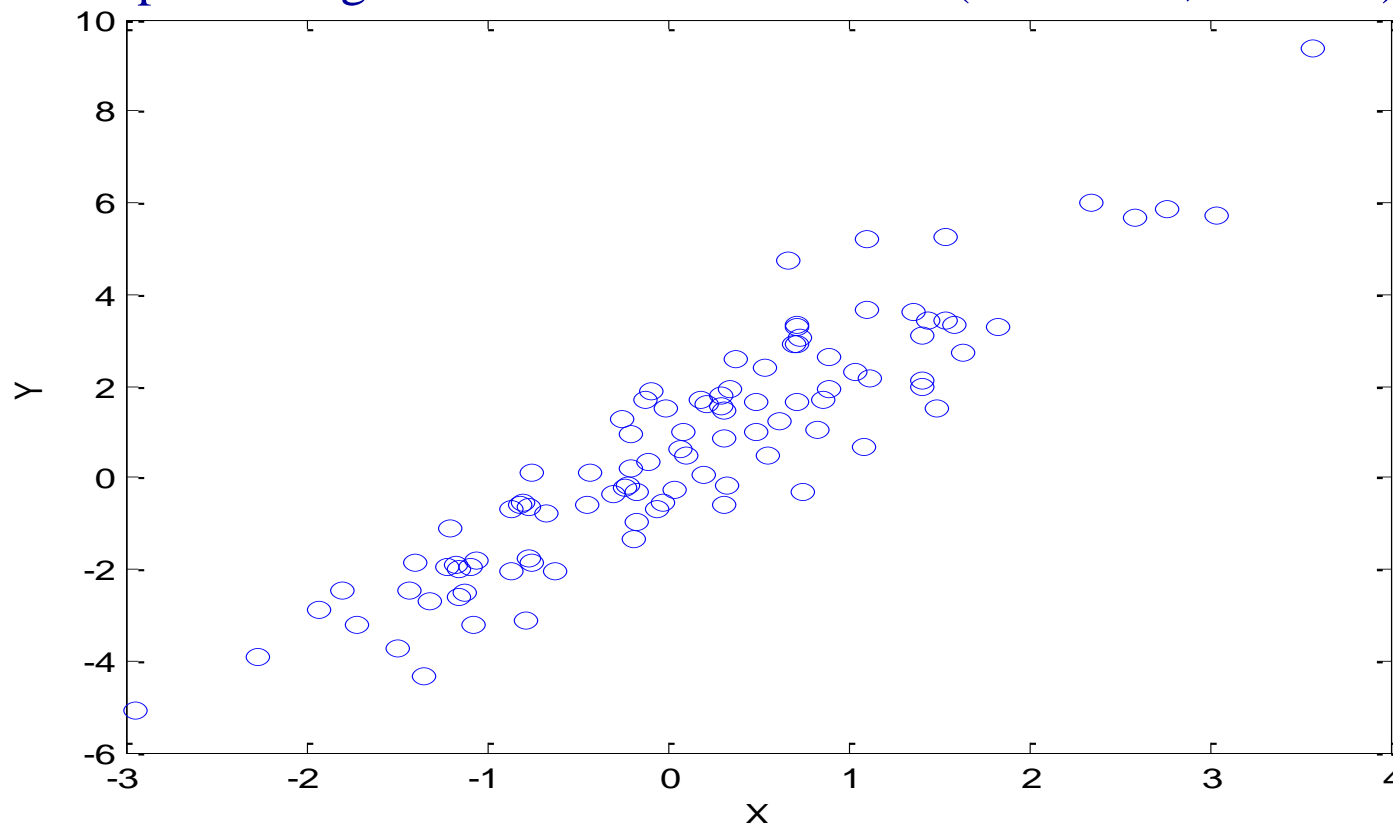


➤ which one is the best?

1) Linear Regression Model

Econ 3334

- A computer example: I ask my computer to generate the data:
 - Step 1: X_i follows $N(0,1)$, u_i follows $N(0,1)$, X_i and u_i are independent
 - Step 2: Y_i is generated as $Y_i = 0.5 + 2X_i + u_i$ ($\beta_0 = 0.5$, $\beta_1 = 2$)



- In reality, we don't know β_0 and β_1 . Need to find them.

2) Estimation: OLS

Econ 3334

- Linear regression model:

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

- We don't know the population parameter β_0 and β_1 , so we must estimate them with the data, what are the estimators?
- We use a method called OLS (ordinary least square).

2) Estimation: OLS

Econ 3334

- How can we estimate β_0 and β_1 from data?
- Recall that \bar{Y} solves

$$\min_b \sum_{i=1}^n (Y_i - b)^2$$

- By analogy, we will focus on the least squares (“ordinary least squares” or “*OLS*”) estimator of the unknown parameters β_0 and β_1 , which solves

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

2) Estimation: OLS

Econ 3334

- 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.

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

- This minimization problem can be solved using calculus.
- The result is the OLS estimators of β_0 and β_1 . We denote the estimators as $\hat{\beta}_0$ and $\hat{\beta}_1$.
- We will see that OLS estimator is a “good” estimator (unbiased, consistent, efficient).

- Suppose that we are estimating a linear model:

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

- The OLS estimator $(\hat{\beta}_0, \hat{\beta}_1)$ is to minimize the objective function:

$$\sum_{i=1}^n [Y_i - b_0 - X_i b_1]^2$$

- Using calculus, we solve this minimization problem by solving the following two first-order conditions:

$$\frac{\partial}{\partial b_0} \sum_{i=1}^n [Y_i - b_0 - X_i b_1]^2 = 0$$

$$\frac{\partial}{\partial b_1} \sum_{i=1}^n [Y_i - b_0 - X_i b_1]^2 = 0$$

$$\text{➤} \quad \frac{\partial}{\partial b_0} \sum_{i=1}^n [Y_i - b_0 - X_i b_1]^2 = -2 \sum_{i=1}^n [Y_i - b_0 - X_i b_1] = 0 \quad (1)$$

$$\frac{\partial}{\partial b_1} \sum_{i=1}^n [Y_i - b_0 - X_i b_1]^2 = -2 \sum_{i=1}^n [Y_i - b_0 - X_i b_1] \cdot X_i = 0 \quad (2)$$

➤ Let's first look at equation (1):

$$\begin{aligned} -2 \sum_{i=1}^n [Y_i - b_0 - X_i b_1] &= 0 \Rightarrow \sum_{i=1}^n Y_i - \sum_{i=1}^n b_0 - b_1 \sum_{i=1}^n X_i = 0 \\ \Rightarrow \left(\frac{1}{n} \sum_{i=1}^n Y_i \right) - \left(\frac{1}{n} \sum_{i=1}^n b_0 \right) - b_1 \left(\frac{1}{n} \sum_{i=1}^n X_i \right) &= 0 \\ \Rightarrow \bar{Y} - b_0 - b_1 \bar{X} = 0 \Rightarrow b_0 = \bar{Y} - b_1 \bar{X} \quad (3) \end{aligned}$$

- Now let's substitute equation (3) into equation (2).
Equation (2) becomes

$$\begin{aligned} & -2 \sum_{i=1}^n [Y_i - (\bar{Y} - b_1 \bar{X}) - X_i b_1] \cdot X_i = 0 \\ \Rightarrow & -2 \sum_{i=1}^n [Y_i - \bar{Y} - b_1(X_i - \bar{X})] \cdot X_i = 0 \\ \Rightarrow & \sum_{i=1}^n (Y_i - \bar{Y})X_i - b_1 \sum_{i=1}^n (X_i - \bar{X}) \cdot X_i = 0 \\ \Rightarrow & b_1 = \frac{\sum_{i=1}^n (Y_i - \bar{Y}) \cdot X_i}{\sum_{i=1}^n (X_i - \bar{X}) \cdot X_i}. \end{aligned}$$

➤ We can show that

$$b_1 = \frac{\sum_{i=1}^n (Y_i - \bar{Y}) \cdot X_i}{\sum_{i=1}^n (X_i - \bar{X}) \cdot X_i} = \frac{\sum_{i=1}^n (Y_i - \bar{Y}) \cdot (X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X}) \cdot (X_i - \bar{X})}$$

➤ To show this,

$$\begin{aligned} \text{(i)} \quad \sum_{i=1}^n (Y_i - \bar{Y}) \cdot (X_i - \bar{X}) &= \sum_{i=1}^n (Y_i - \bar{Y}) \cdot X_i - \sum_{i=1}^n (Y_i - \bar{Y}) \cdot \bar{X} \\ &= \sum_{i=1}^n (Y_i - \bar{Y}) \cdot X_i - \bar{X} \cdot \underbrace{\sum_{i=1}^n (Y_i - \bar{Y})}_{=0} = \sum_{i=1}^n (Y_i - \bar{Y}) \cdot X_i \end{aligned}$$

$$\begin{aligned} \text{(ii)} \quad \sum_{i=1}^n (X_i - \bar{X}) \cdot (X_i - \bar{X}) &= \sum_{i=1}^n (X_i - \bar{X}) \cdot X_i - \sum_{i=1}^n (X_i - \bar{X}) \cdot \bar{X} \\ &= \sum_{i=1}^n (X_i - \bar{X}) \cdot X_i - \bar{X} \cdot \underbrace{\sum_{i=1}^n (X_i - \bar{X})}_{=0} = \sum_{i=1}^n (X_i - \bar{X}) \cdot X_i. \end{aligned}$$

2) Estimation: OLS

Econ 3334

➤ The OLS estimator

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

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}. \quad (4.8)$$

- Here, true β_0 and β_1 are unknown, we want to use data to estimate them; it turns out $\hat{\beta}_0$ and $\hat{\beta}_1$ are good estimators.
- $\hat{\beta}_0$ and $\hat{\beta}_1$ are calculated from the data, thus they are random variables. Our hope is that these two random variables are “close” to the true β_0 and β_1 .

2) Estimation: OLS

Econ 3334

➤ OLS predicted value and residuals

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

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i, \quad i = 1, \dots, n \quad (4.9)$$

$$\hat{u}_i = Y_i - \hat{Y}_i, \quad i = 1, \dots, n. \quad (4.10)$$

➤ Note that \hat{u}_i is different from u_i :

$$\begin{aligned} \hat{u}_i &= Y_i - \hat{Y}_i = \beta_0 + \beta_1 X_i + u_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i) \\ &= u_i + (\beta_0 - \hat{\beta}_0) + (\beta_1 - \hat{\beta}_1) X_i \end{aligned}$$

➤ \hat{u}_i is different from u_i because $\hat{\beta}_0$ and $\hat{\beta}_1$ are different from the true β_0 and β_1 .

THE OLS ESTIMATOR, PREDICTED VALUES, AND RESIDUALS

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

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

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}. \quad (4.8)$$

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

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i, \quad i = 1, \dots, n \quad (4.9)$$

$$\hat{u}_i = Y_i - \hat{Y}_i, \quad i = 1, \dots, n. \quad (4.10)$$

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 $Y_i, i = 1, \dots, n$. These are estimates of the unknown true population intercept (β_0), slope (β_1), and error term (u_i).

2) Estimation: OLS

Econ 3334

➤ A Numerical Example:

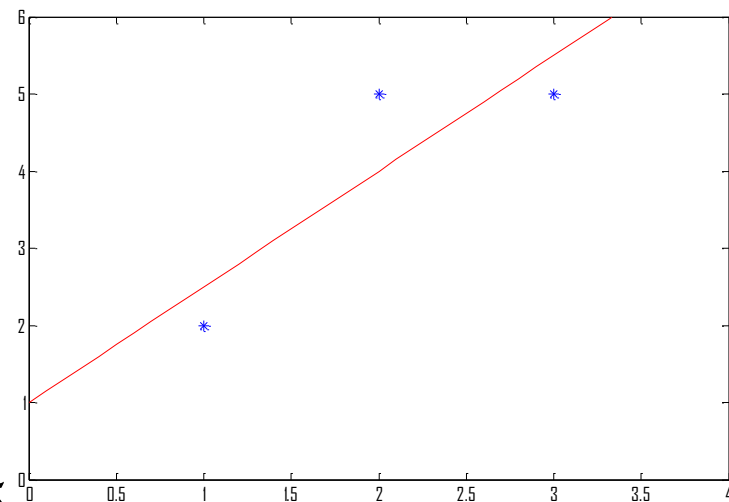
X_i	Y_i	\Rightarrow				
X_i	Y_i	$X_i - \bar{X}$	$Y_i - \bar{Y}$	$(X_i - \bar{X})(Y_i - \bar{Y})$	$(X_i - \bar{X})^2$	
1	2	-1	-2	2	1	
2	5	0	1	0	0	
3	5	1	1	1	1	

$$\hat{\beta}_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} = \frac{2 + 0 + 1}{1 + 0 + 1} = 1.5$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = 4 - 1.5 \cdot 2 = 1$$

➤ So the line that “best” fits the data is:

$$\hat{Y}_i = 1 + 1.5X_i$$



2) Estimation: OLS

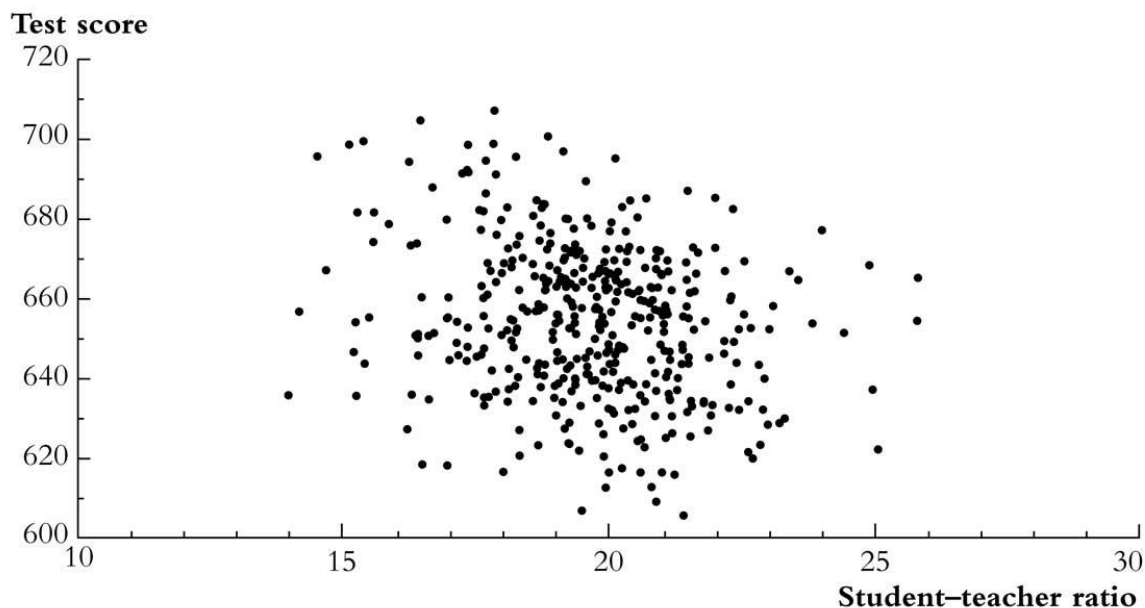
Econ 3334

- Example with a lot of data:
- What is the effect on test scores of reducing STR (student-teacher ratio) by one unit?



FIGURE 4.2 Scatterplot of Test Score vs. Student-Teacher Ratio (California School District Data)

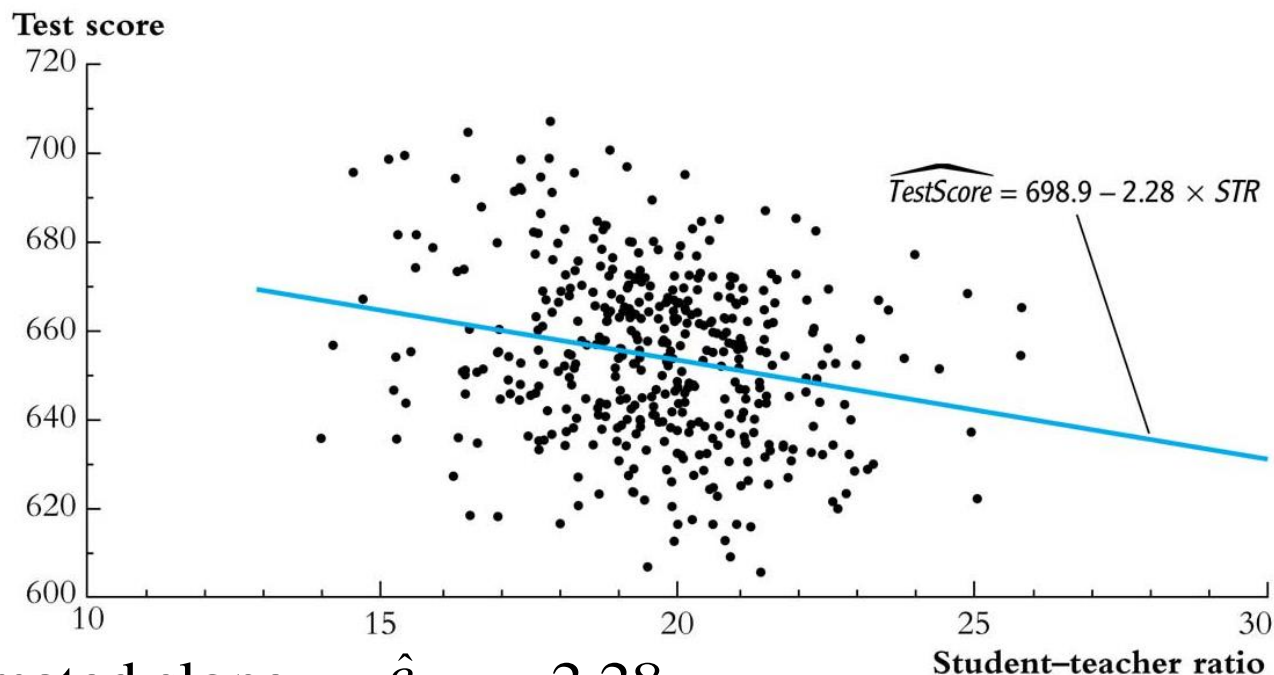
Data from 420 California school districts. There is a weak negative relationship between the student-teacher ratio and test scores: The sample correlation is -0.23 .



2) Estimation: OLS

Econ 3334

- Estimation results: (Stata will do OLS, so we don't have to calculate things by hand)



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

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

Estimated regression line: $\hat{TestScore} = 698.9 - 2.28 \times STR$

2) Estimation: Stata

Econ 3334

The screenshot displays the Stata/SE 12.1 [Results] window. The main area shows the Stata startup screen, which includes the Stata logo, the version number 12.1, and copyright information for StataCorp LP. It also displays the address: 4905 Lakeway Drive, College Station, Texas 77845 USA. Contact information for StataCorp is provided, including a phone number (800-STATA-PC, 979-696-4600) and a website (<http://www.stata.com>). The email address stata@stata.com is also listed. A single-user perpetual license is shown, with serial number 40120575044, licensed to Economics at HKUST. A note indicates that the maximum number of variables is 5000, which can be increased using the `/v#` option or `-set maxvar-`.

The interface includes a menu bar (File, Edit, Data, Graphics, Statistics, User, Window, Help) and a toolbar with various icons. On the left, there is a 'Review' panel with a 'Command' tab, which currently shows 'There are no items to show.' On the right, there is a 'Variables' panel, also showing 'There are no items to show.' Below the main window, there is a 'Properties' panel with tabs for 'Variables' and 'Data'. The 'Variables' tab is active, showing a list of variables with columns for Name, Label, Type, Format, Value Lab, and Notes. The 'Data' tab is also visible, showing fields for Filename, Label, Notes, Variables (0), Observations (0), and Size (0). At the bottom, there is a 'Command' window with a blue background and a cursor.

Stata/SE 12.1 - [Results]

File Edit Data Graphics Statistics User Window Help

Review

Command

There are no items to show.

(R)

12.1 Copyright 1985-2011 StataCorp LP
StataCorp
4905 Lakeway Drive
College Station, Texas 77845 USA
800-STATA-PC <http://www.stata.com>
979-696-4600 stata@stata.com
979-696-4601 (fax)

Statistics/Data Analysis

Special Edition

Single-user Stata perpetual license:
Serial number: 40120575044
Licensed to: Economics
HKUST

Notes:
1. (/v# option or -set maxvar-) 5000 maximum variables

Command

Variables

Variable Label

There are no items to show.

Properties

Variables

Name
Label
Type
Format
Value Lab
Notes

Data

Filename
Label
Notes
Variables 0
Observations 0
Size 0

Command

D:\program\Stata12

CAP NUM OVR

➤ Import data: File---->Import----->Excel spreadsheet

Stata/SE 12.1 - [Results]

File Edit Data Graphics Statistics User Window Help

Review T F X

Command ...

```
1 cd "D:\dro..."
2 cd
```

Statistics/Data Analysis

Special Edition

Single-user Stata perpe
Serial number:
Licensed to:

Notes:
1. (/v# option o

```
. cd "D:\dropbox\Dropbo
D:\dropbox\Dropbox\HKUS
. cd
D:\dropbox\Dropbox\HKUS
.
```

Command

Import Excel

Excel file: Browse...

Worksheet: Cell range: ...

☐ Import first row as variable names

☐ Import all data as strings

Variable case: preserve

Preview:

There are no items to show.

OK Cancel

Variables

Variable	Label
There are no items to show.	

Properties

Variables

Name	Label	Type	Format	Value Lab	Notes
Data					
Filename	Label	Notes	Variables 0	Observati 0	Size 0

D:\dropbox\Dropbox\HKUST teaching 2014\ECON3334\Lecture\Topic 4\data

15:25
2014/2/21

24

➤ Import data: File---->Import----->Excel spreadsheet

Stata/SE 12.1 - [Results]

File Edit Data Graphics Statistics User Window Help

Review

Command

```
1 cd "D:\dro..."
2 cd
```

Statistics/Data Analysis

Special Edition

Single-user Stata perpetual license
Serial number:
Licensed to:

Notes:

```
1. (/v# option o...
. cd "D:\dropbox\Dropbox\HKUST teaching 2014\ECON3334\Lecture\Topic 4\data"
D:\dropbox\Dropbox\HKUST teaching 2014\ECON3334\Lecture\Topic 4\data
. cd
D:\dropbox\Dropbox\HKUST teaching 2014\ECON3334\Lecture\Topic 4\data
```

Import Excel

Excel file: D:\dropbox\Dropbox\HKUST teaching 2014\ECON3334\Lecture\Topic 4\caschool.xls [Browse...]

Worksheet: caschool Cell range: A1:B421

☒ Import first row as variable names
☐ Import all data as strings

Variable case: preserve

Preview: (showing rows 2-51 of 421)

	testscr	str
2	690.79999	17.88991
3	661.20001	21.524664
4	643.59998	18.697226
5	647.70001	17.357143
6	640.84998	18.671329
7	605.55005	21.40625
8	606.75	19.5

OK Cancel

Variables

Variable Label

There are no items to show.

Properties

Variables

Name	Label	Type	Format	Value Label	Notes

Data

Filename

Label	Notes	Variables	Observations	Size
		0	0	0

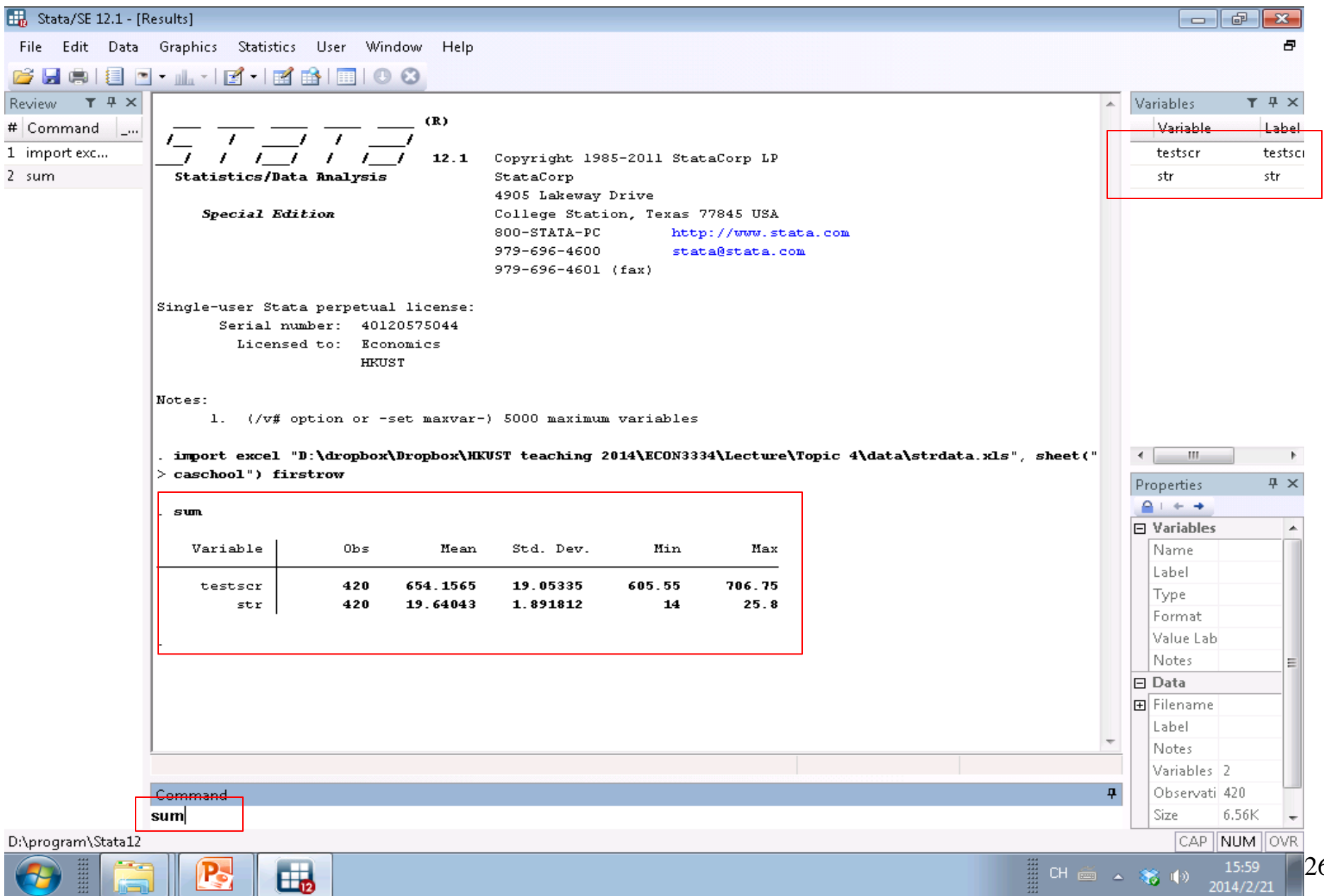
Command

D:\dropbox\Dropbox\HKUST teaching 2014\ECON3334\Lecture\Topic 4\data

CAP NUM OVR

15:28
2014/2/21

► Type sum (providing summary information of the data)



Stata/SE 12.1 - [Results]

File Edit Data Graphics Statistics User Window Help

Review

Command

```
1 import exc...
2 sum
```

(R)

12.1 Copyright 1985-2011 StataCorp LP
StataCorp
4905 Lakeway Drive
College Station, Texas 77845 USA
800-STATA-PC <http://www.stata.com>
979-696-4600 stata@stata.com
979-696-4601 (fax)

Statistics/Data Analysis

Special Edition

Single-user Stata perpetual license:
Serial number: 40120575044
Licensed to: Economics
HKUST

Notes:

```
1. (/v# option or -set maxvar-) 5000 maximum variables
```

```
. import excel "D:\dropbox\Dropbox\HKUST teaching 2014\ECON3334\Lecture\Topic 4\data\strdata.xls", sheet("caschool") firstrow
```

Variable	Obs	Mean	Std. Dev.	Min	Max
testscr	420	654.1565	19.05335	605.55	706.75
str	420	19.64043	1.891812	14	25.8

Command

```
sum
```

Properties

Variables

Name

Label

Type

Format

Value Lab

Notes

Data

Filename

Label

Notes

Variables 2

Observations 420

Size 6.56K

CAP NUM OVR

D:\program\Stata12

CH 15:59 2014/2/21

► Type reg testscr str

Stata/SE 12.1 - [Results]

File Edit Data Graphics Statistics User Window Help

Review

Command

```
1 import exc...
2 sum
3 reg testscr str
```

Single-user Stata perpetual license:
Serial number: 40120575044
Licensed to: Economics
HKUST

Notes:

- (/v# option or -set maxvar-) 5000 maximum variables

```
. import excel "D:\dropbox\Dropbox\HKUST teaching 2014\ECON3334\Lecture\Topic 4\data\strdata.xls", sheet("caschool") firstrow
. sum
```

Variable	Obs	Mean	Std. Dev.	Min	Max
testscr	420	654.1565	19.05335	605.55	706.75
str	420	19.64043	1.891812	14	25.8

```
. reg testscr str
```

Source	SS	df	MS	Number of obs =	420
Model	7794.11004	1	7794.11004	F(1, 418) =	22.58
Residual	144315.484	418	345.252353	Prob > F =	0.0000
Total	152109.594	419	363.030056	R-squared =	0.0512
				Adj R-squared =	0.0490
				Root MSE =	18.581

	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
testscr					
str	-2.279808	.4798256	-4.75	0.000	-3.22298 -1.336637
_cons	698.933	9.467491	73.82	0.000	680.3231 717.5428

beta1_hat & beta0_hat

Command

```
reg testscr str
```

D:\program\Stata12

CH 16:03 2014/2/21

27

2) Estimation: OLS

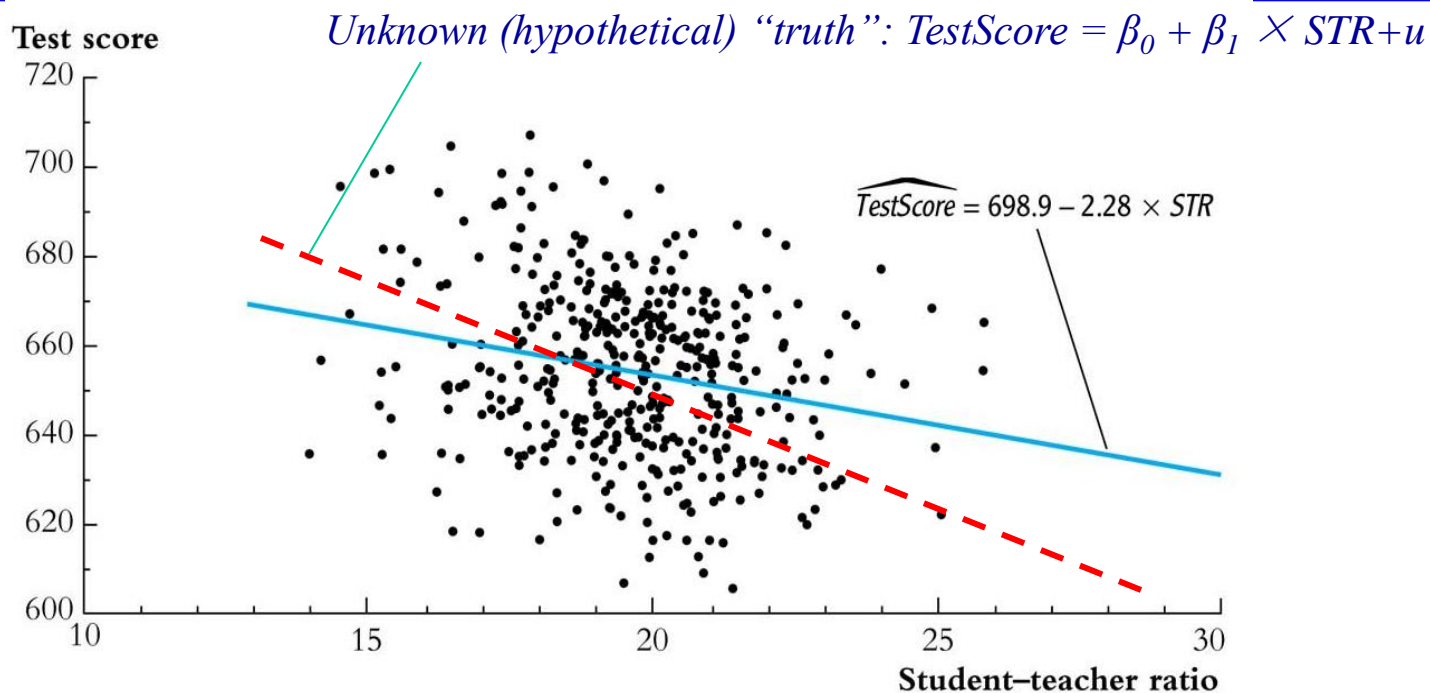
Econ 3334

$$\hat{T}estScore = 698.9 - 2.28 \times STR$$

- Districts with one more student per teacher on average have test scores that are 2.28 points lower.
- That is, $\frac{\Delta \text{Test score}}{\Delta STR} = -2.28$
- The intercept (taken literally) means that, according to this estimated line, districts with zero students per teacher would have a (predicted) test score of 698.9.
- This interpretation of the intercept makes no sense – it extrapolates the line outside the range of the data – here, the intercept is not economically meaningful.

2) Estimation: OLS

Econ 3334



One of the districts in the data set is Antelope, CA, for which $\text{STR} = 19.33$ and $\text{Test Score} = 657.8$

predicted value: $\hat{Y}_{\text{Antelope}} = 698.9 - 2.28 \times 19.33 = 654.8$

residual: $\hat{u}_{\text{Antelope}} = 657.8 - 654.8 = 3.0$