

# Econometrics 1

## Lecture 6: Linear Regression (1)

### Linear regression with one regressor

---

黄嘉平

中国经济特区研究中心 讲师

办公室：文科楼2613

E-mail: [huangjp@szu.edu.cn](mailto:huangjp@szu.edu.cn)

Tel: (0755) 2695 0548

Website: <https://huangjp.com>

# The linear regression model

# Linear relationship between $X$ and $Y$

---

- A school district cuts the size of its elementary school classes. What is the effect on its students' test score?
- This question is about the unknown effect of changing one variable,  $X$  (class size), on another variable,  $Y$  (student test score).
- Linear regression (with one regressor) is a model investigating the linear relationship between  $X$  and  $Y$ .

# Class size and test score

---

- Relative change, or the effect of changing  $X$  on  $Y$ :

$$\beta_{ClassSize} = \frac{\text{change in } TestScore}{\text{change in } ClassSize} = \frac{\Delta TestScore}{\Delta ClassSize}$$

$$\Delta TestScore = \beta_{ClassSize} \times \Delta ClassSize$$

This is the definition of the slope of a straight line relating test scores and class size:

$$TestScore = \beta_0 + \beta_{ClassSize} \times ClassSize$$

# Incorporating other factors

---

- This relation may not hold for all districts. Therefore we must incorporate other factors influencing test scores.

$$TestScore = \beta_0 + \beta_{ClassSize} \times ClassSize + \text{other factors}$$

- In a more general expression, *ClassSize* becomes *X*, and *TestScore* becomes *Y*.

# The linear regression model

---

- The linear regression model with one regressor

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

dependent variable



coefficients

independent variable / regressor

error term

# The linear regression model

---

- The linear regression model with one regressor

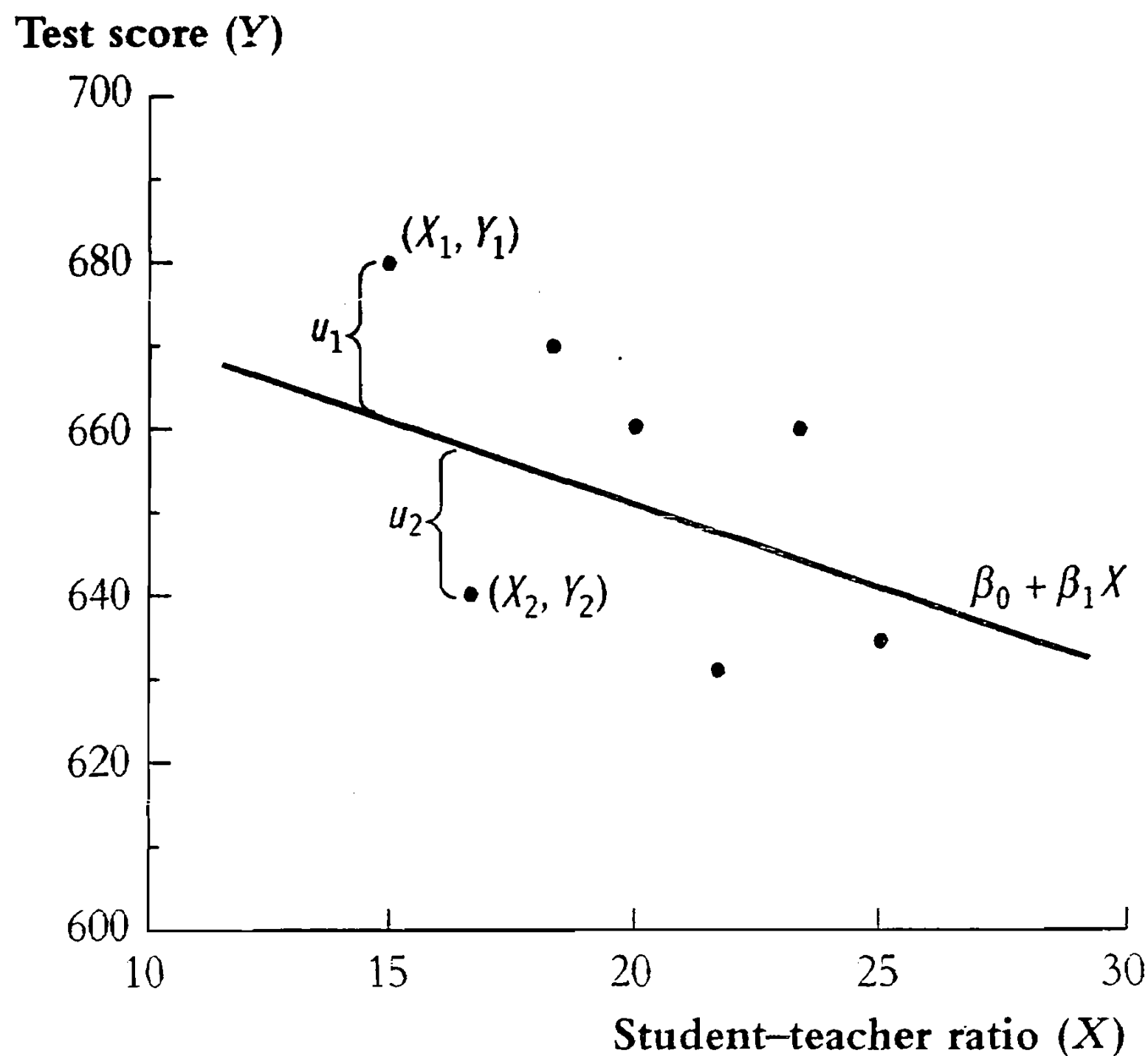
$$Y_i = \boxed{\beta_0 + \beta_1 X_i} + u_i$$



population regression line / population regression function

**FIGURE 4.1** Scatterplot of Test Score vs. Student-Teacher Ratio  
(Hypothetical Data)

The scatterplot shows hypothetical observations for seven school districts. The population regression line is  $\beta_0 + \beta_1 X$ . The vertical distance from the  $i^{\text{th}}$  point to the population regression line is  $Y_i - (\beta_0 + \beta_1 X_i)$ , which is the population error term  $u_i$  for the  $i^{\text{th}}$  observation.





# A test score data in California: the STAR dataset

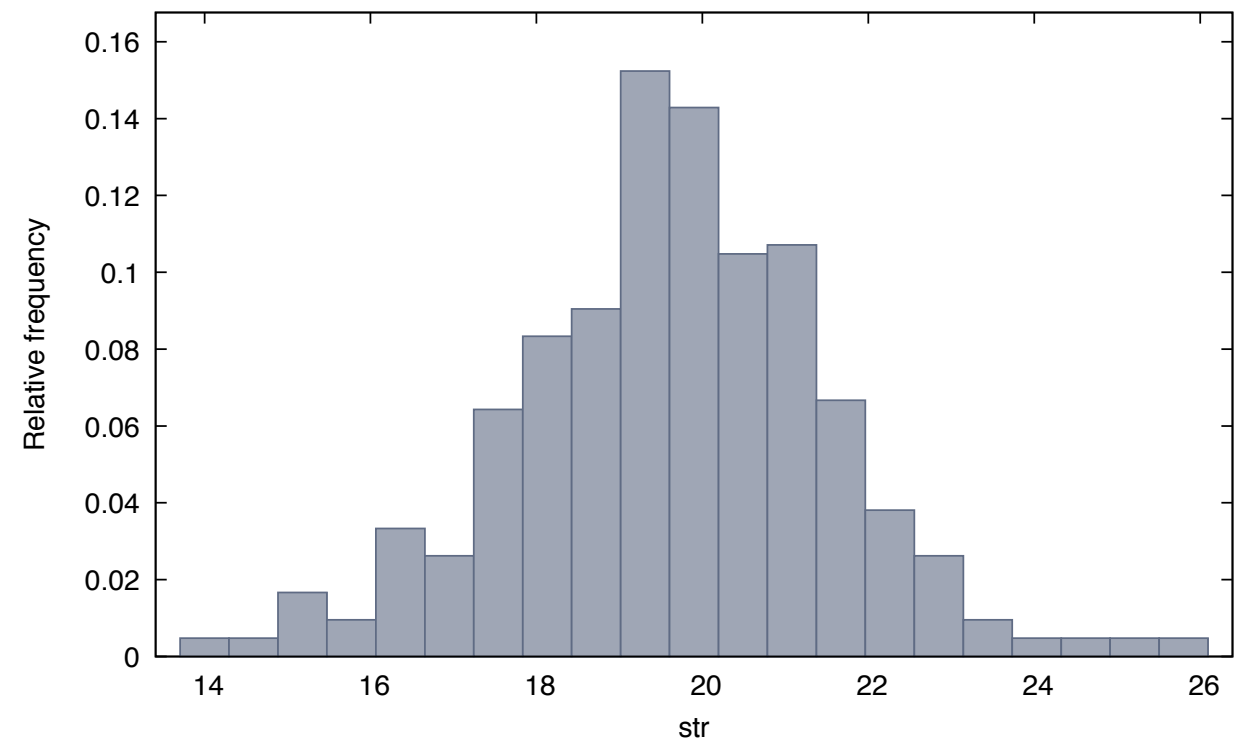
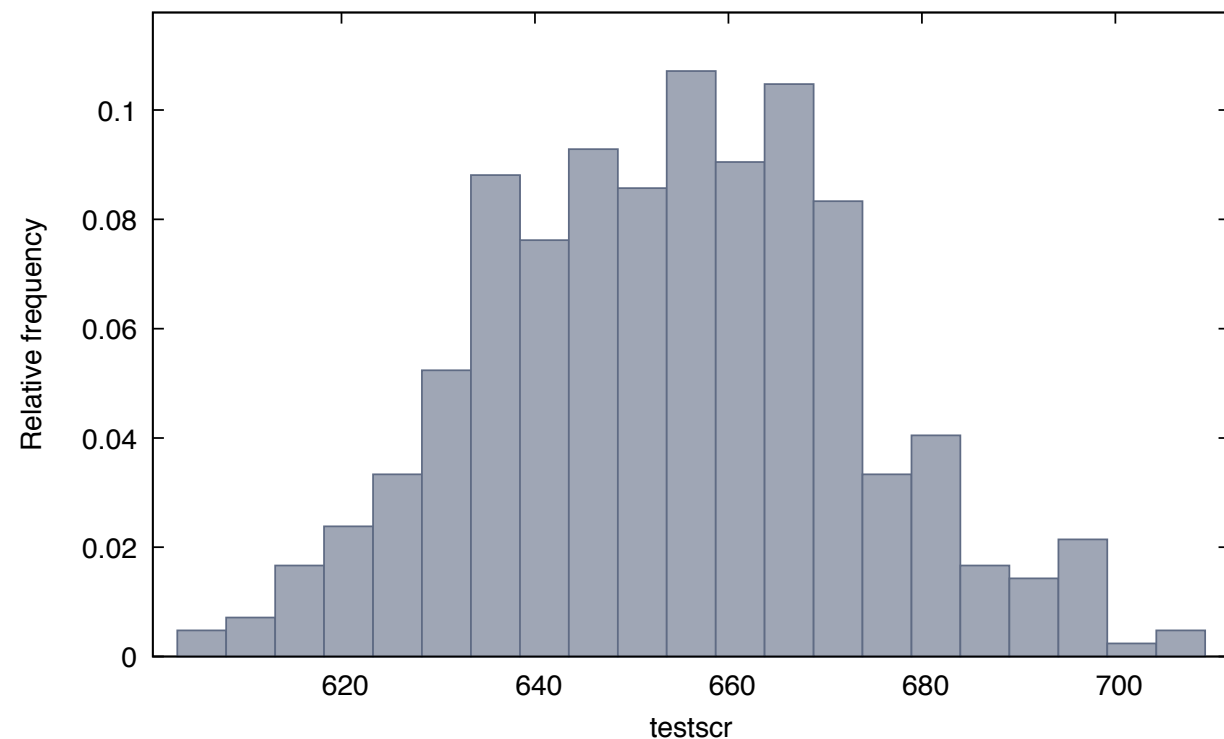
---

- The file `caschool.xlsx`
- The California Standardized Testing and Reporting (STAR) dataset (1998-1999).
- Average test scores on 420 districts in California.
- For details, see `californiatestscores.docx`

# Average test score v.s. student-teacher ratio

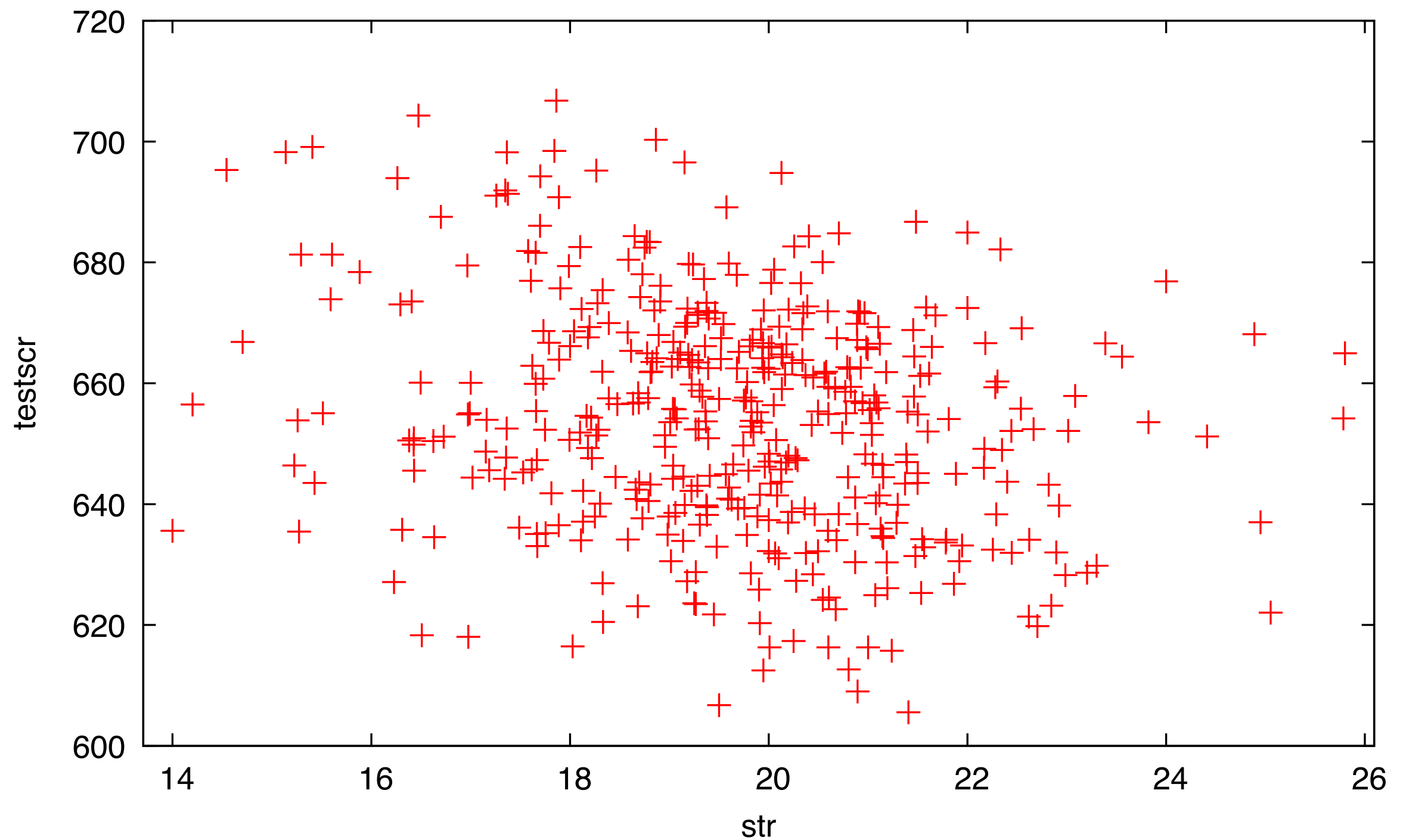
---

- “testscr”: the average test score (of reading and math)
- “str”: the student-teacher ratio (No. of student / No. of teachers)



# Average test score v.s. student-teacher ratio

---



Estimation

# Estimating the coefficients

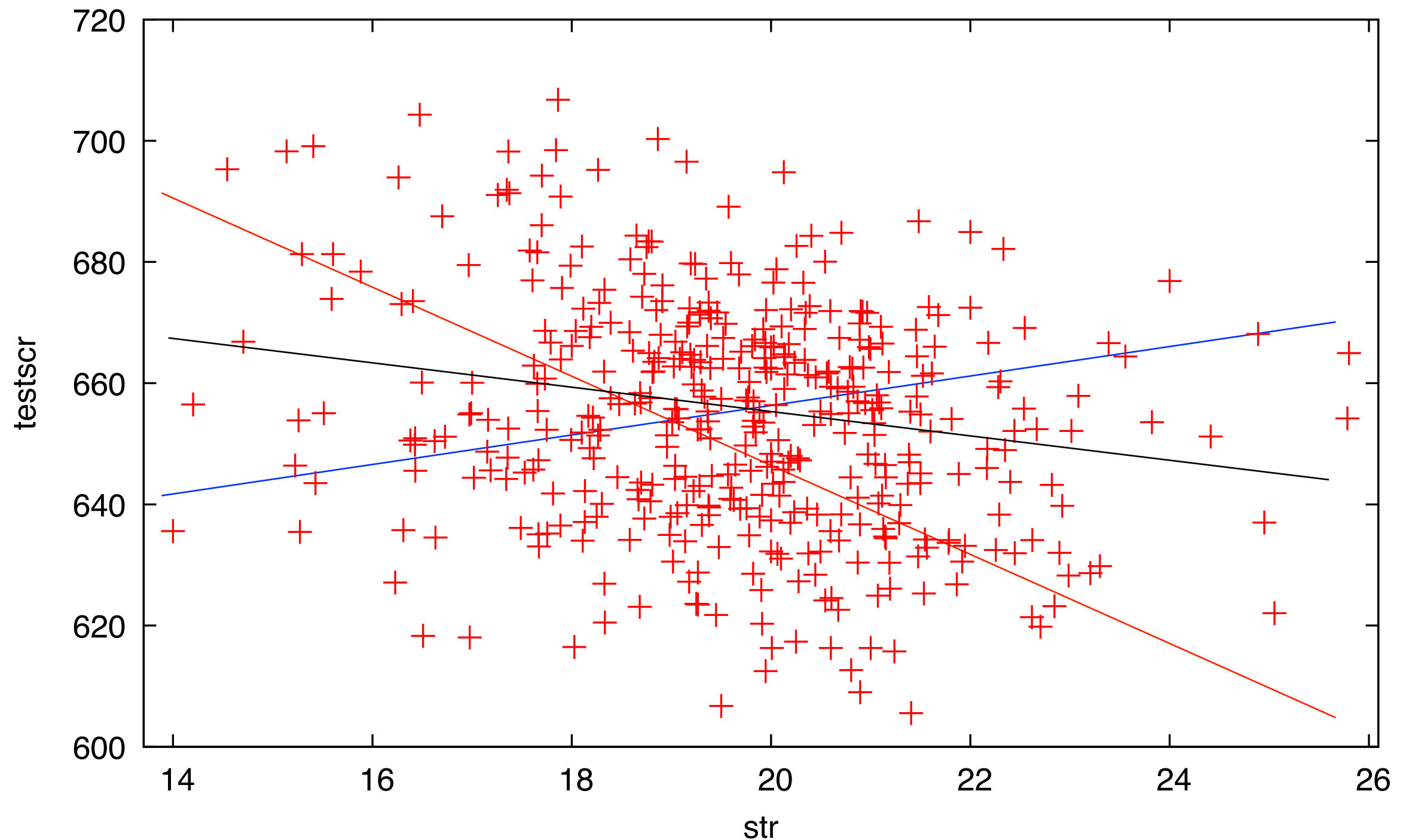
---

- $\bar{Y}$  is an estimator of the population mean.
- Similarly, we need estimators of the coefficients  $\beta_0$  and  $\beta_1$ .
- The ordinary least squares (OLS) estimators  $\hat{\beta}_0$  and  $\hat{\beta}_1$  are the ones that minimize

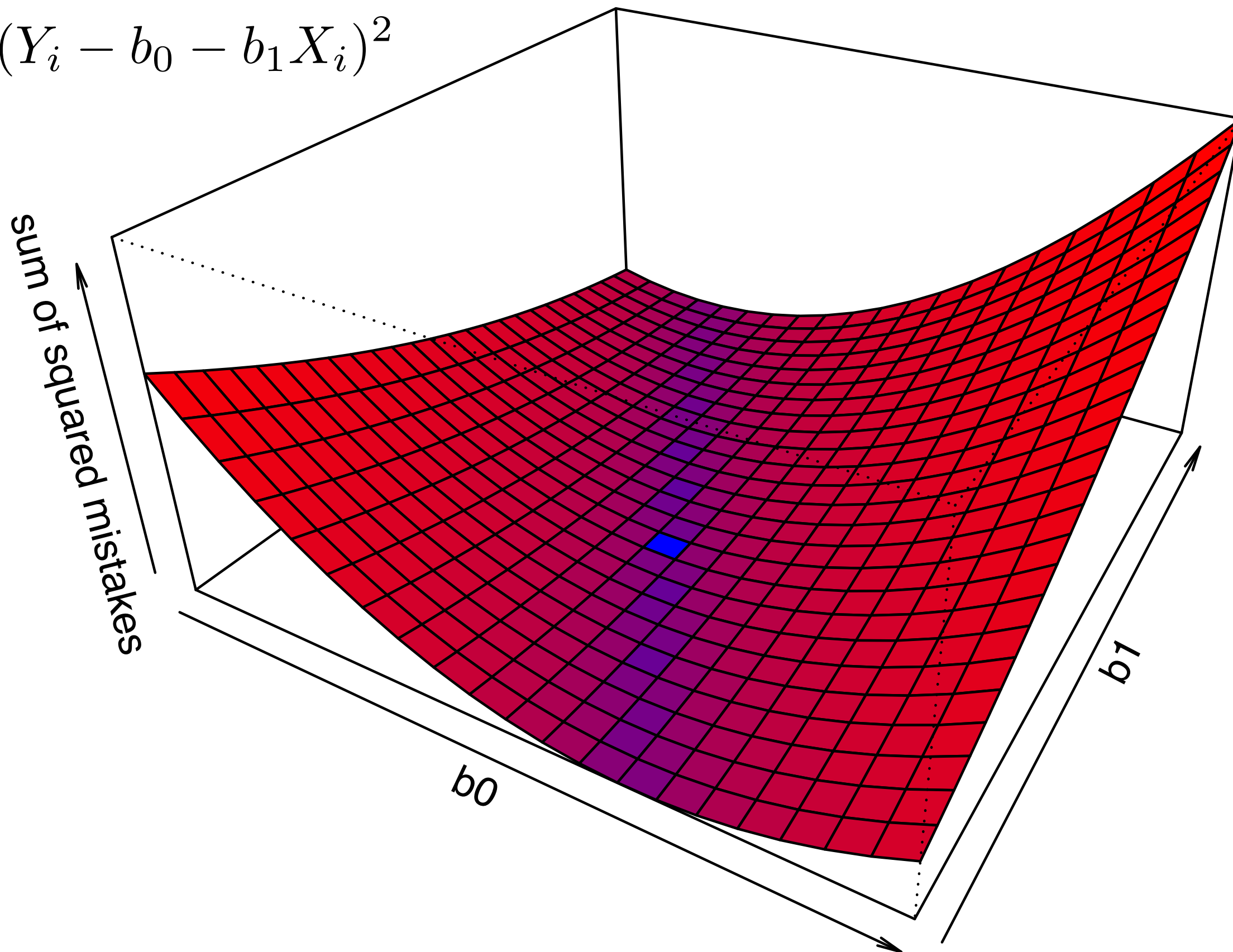
$$\sum_{i=1}^n (Y_i - b_0 - b_1 X_i)^2$$

# How to determine the sample regression line $\hat{\beta}_0 + \hat{\beta}_1 X$ ?

---



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



# The OLS estimator, predicted values, and residuals

---

- The OLS estimators of the slope and the intercept are

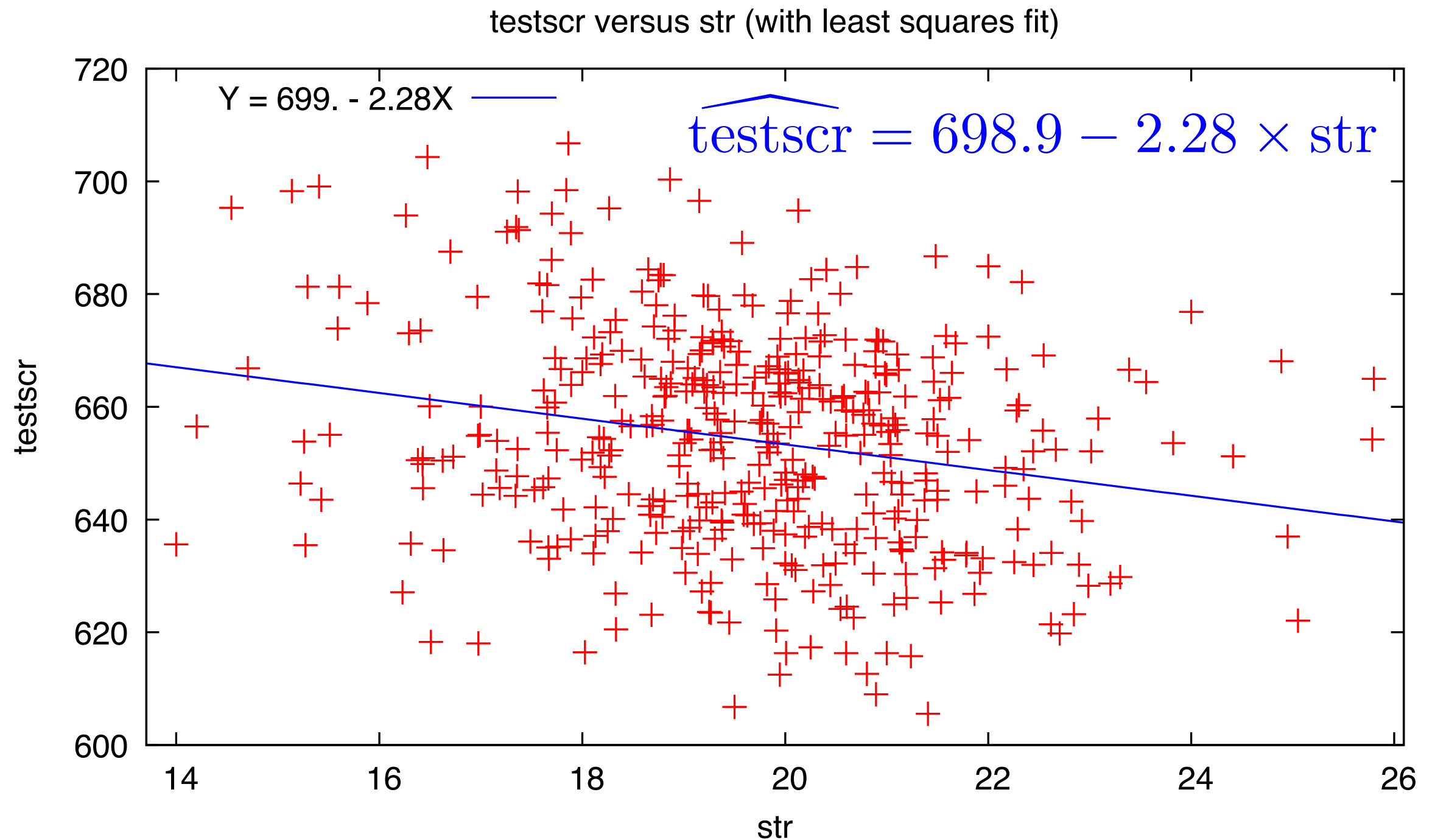
$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{s_{XY}}{s_X^2}$$

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

- The OLS predicted value:  $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$
  - The residuals:  $\hat{u}_i = Y_i - \hat{Y}_i$
- sample regression line/  
sample regression function



# Average test score v.s. student-teacher ratio



# Why use the OLS estimator

---

- OLS is the dominating method used in practice.
- Under certain assumptions, the OLS estimator is *unbiased* and *consistent*.
- With some further assumptions, the OLS estimator is also *efficient* among a class of unbiased estimators.

⇒ Gauss-Markov Theorem (Section 5.5)

For the definitions of unbiasedness, consistency, and efficiency, read Chapter 3.

# Measures of fit

# The $R^2$

---

- The  $R^2$  — correlation of determination, the fraction of the sample variance of  $Y_i$  explained by  $X_i$ .
- Recall that  $Y_i = \hat{Y}_i + \hat{u}_i$

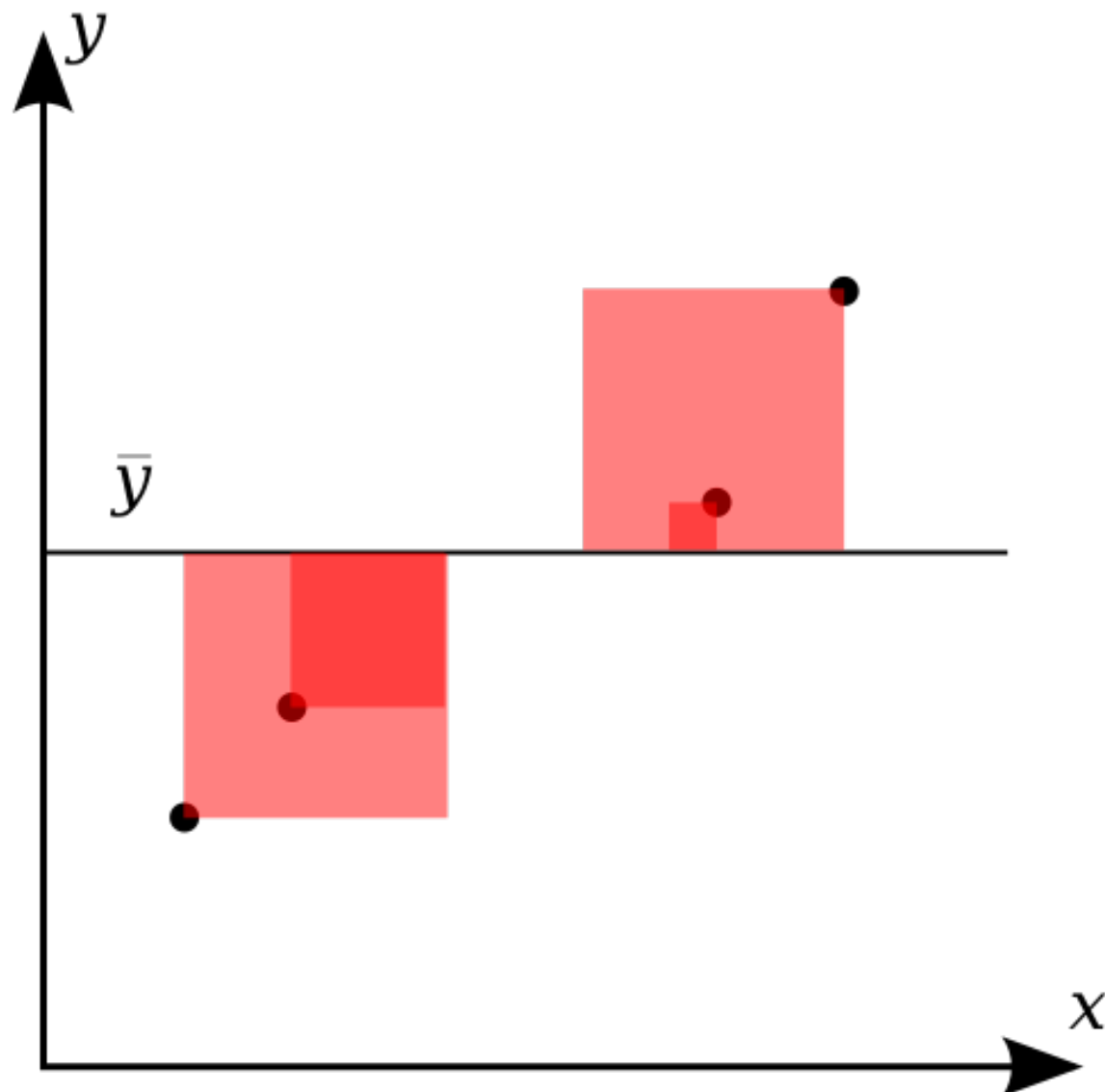
$$\begin{aligned} R^2 &= \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = \frac{ESS}{TSS} \quad \begin{array}{l} \text{(explained sum of squares)} \\ \text{(total sum of squares)} \end{array} \\ &= 1 - \frac{\sum_{i=1}^n \hat{u}_i^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = 1 - \frac{SSR}{TSS} \quad \text{(sum of squared residuals)} \end{aligned}$$

Read Appendix 4.3 if you want to know why the second equality holds.

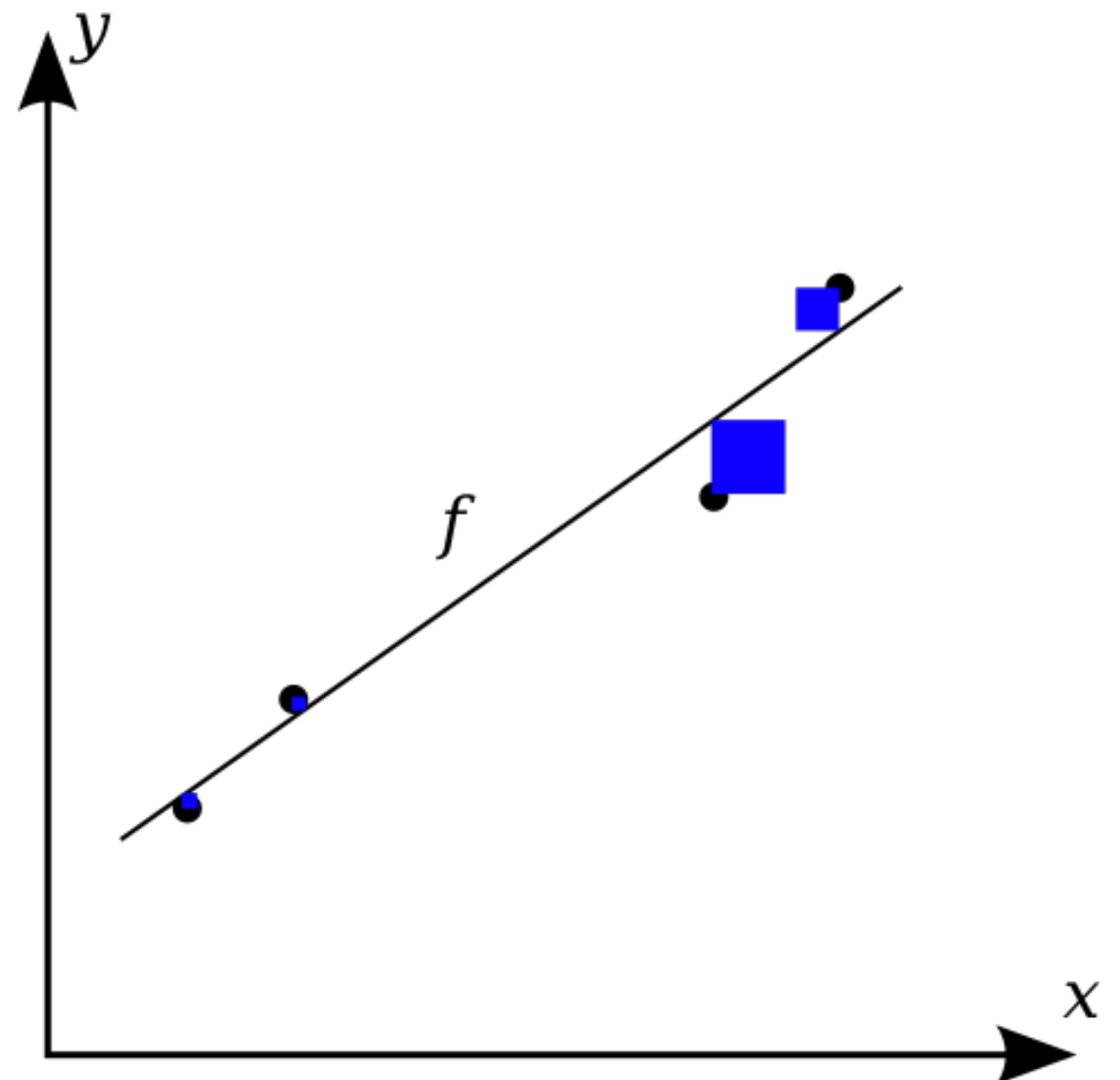
# A graphical explanation of SSR

---

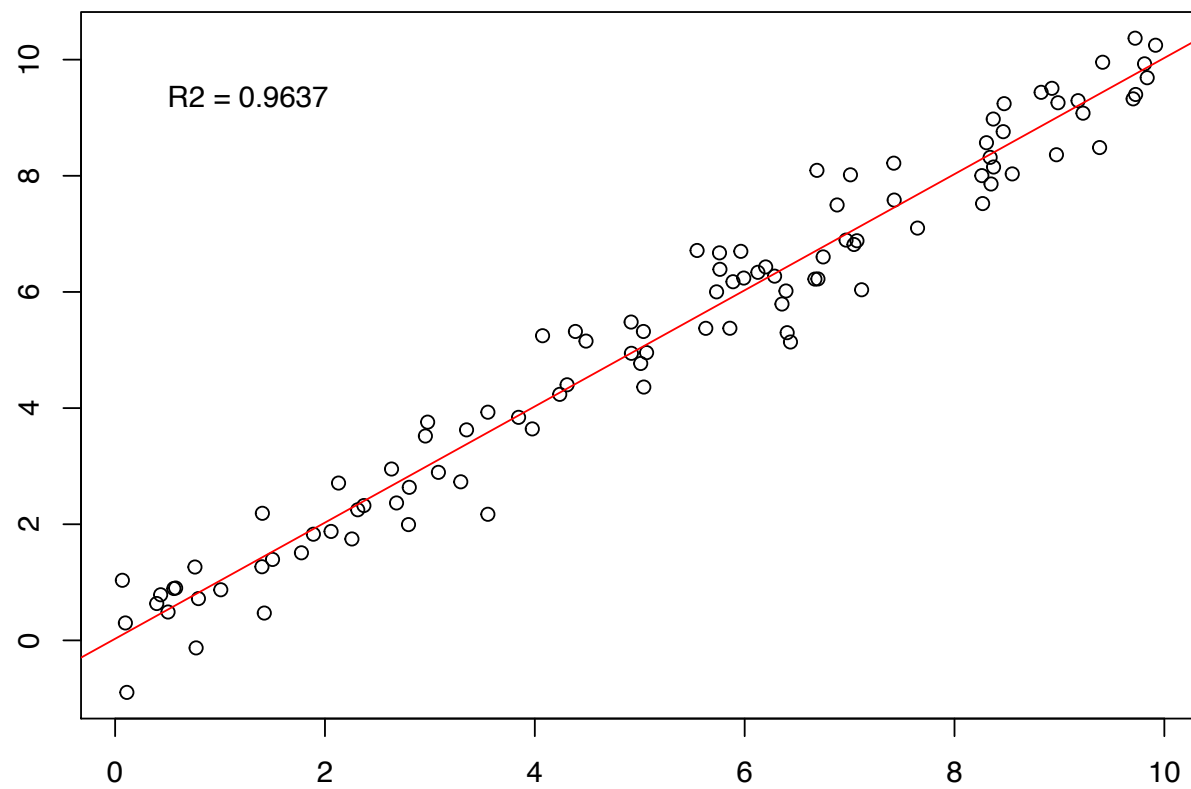
Simple average of  $Y_i$



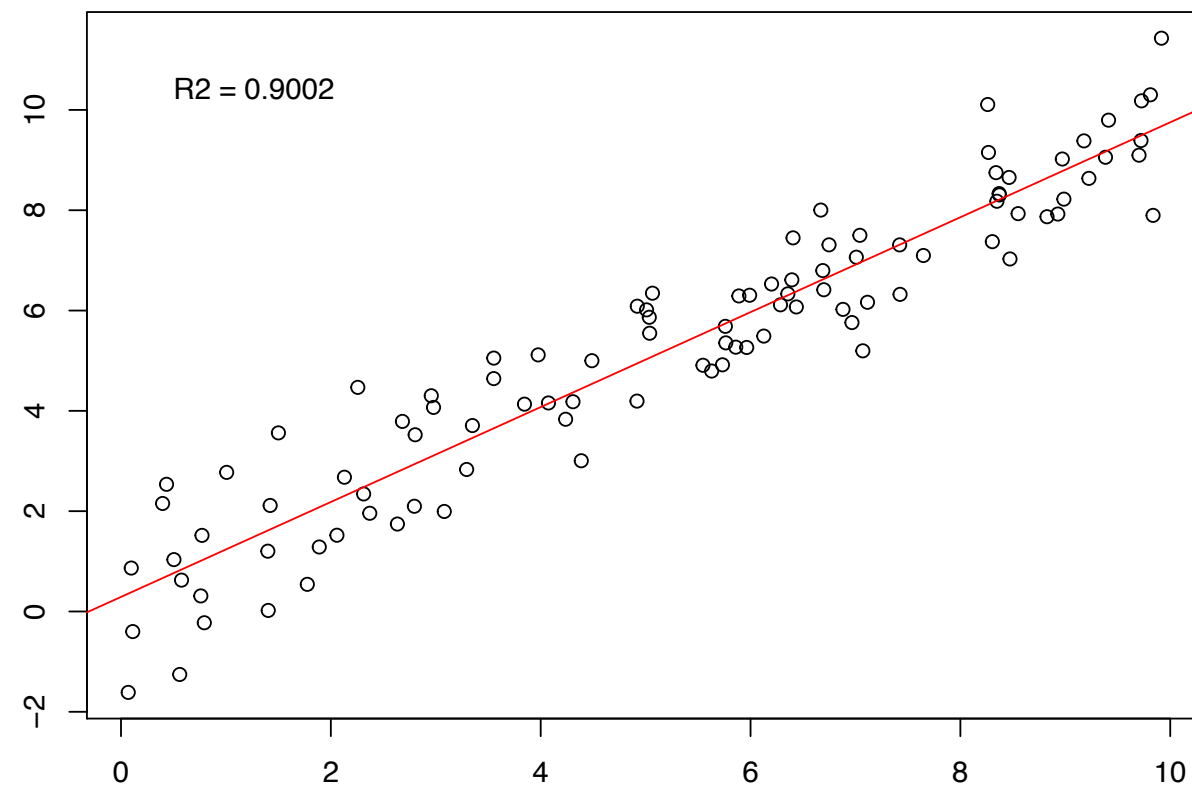
OLS regression



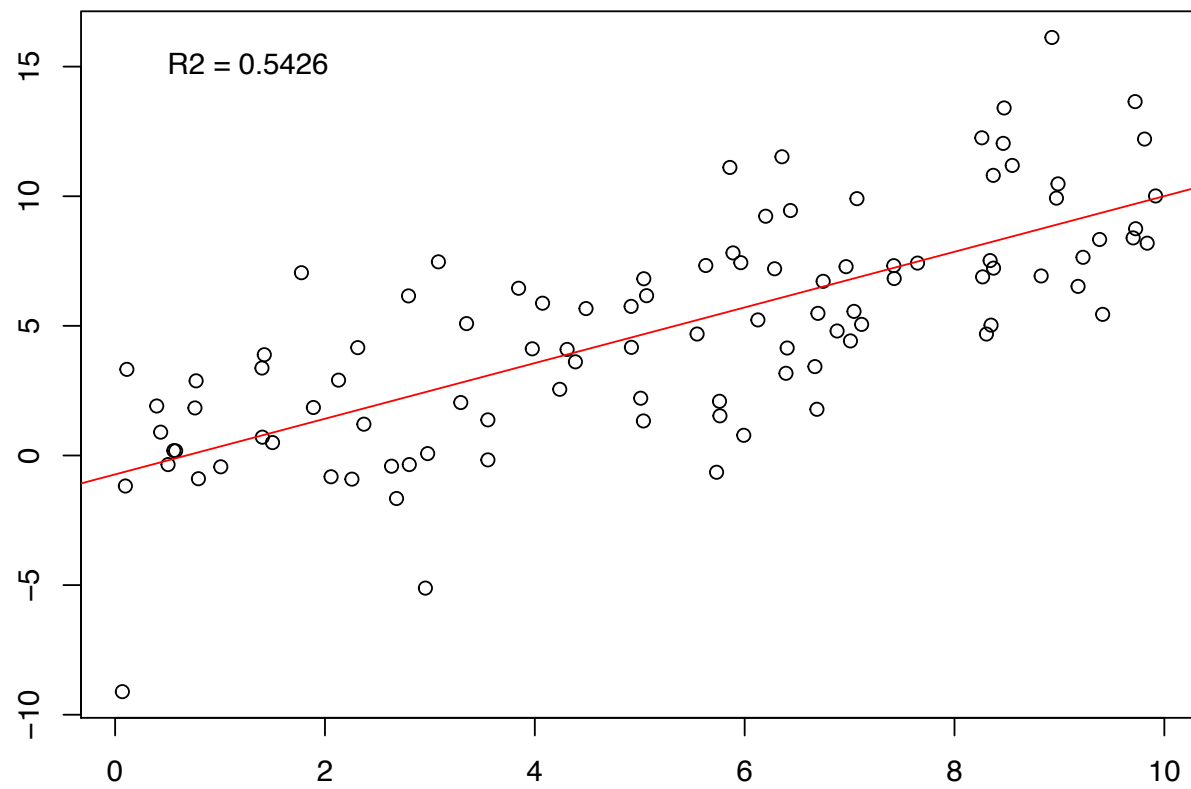
$Y \sim X + N(0, 0.25)$



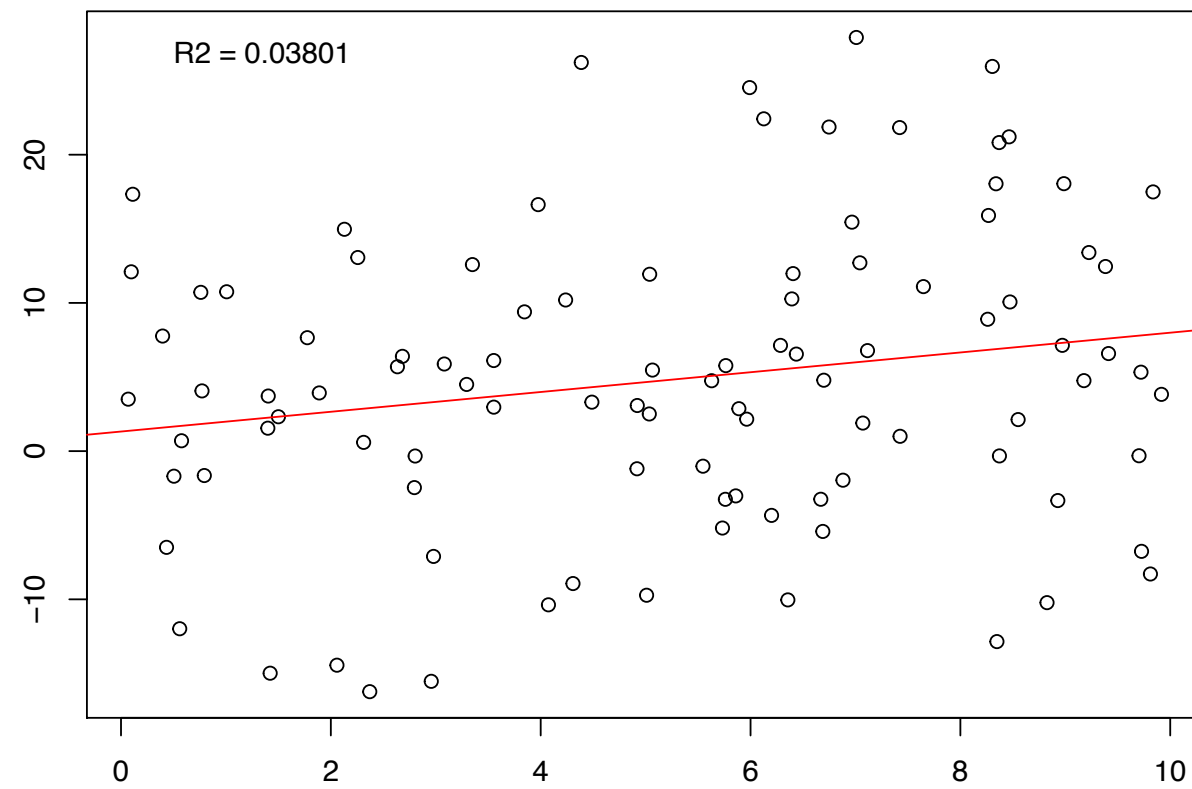
$Y \sim X + N(0, 1)$



$Y \sim X + N(0, 9)$



$Y \sim X + N(0, 100)$



# How to read $R^2$

---

- $R^2$  measures how well the OLS regression line fits the data.
- The value of  $R^2$  ranges between 0 and 1. A high  $R^2$  indicates that the regressor ( $X_i$ ) is good at predicting  $Y_i$ , while a low  $R^2$  indicates that the regressor ( $X_i$ ) is not very good at predicting  $Y_i$ .
- A low  $R^2$  does **not** imply that *this regression* is either “good” or “bad”, it **does** tell us that other important factors influence the dependent variable.

# The standard error of the regression

---

- The standard error of the regression (*SER*) is an estimator of the standard deviation of the regression error  $u_i$ .

$$SER = s_{\hat{u}}, \quad \text{where } s_{\hat{u}}^2 = \frac{1}{n-2} \sum_{i=1}^n \hat{u}_i^2 = \frac{SSR}{n-2}$$

- *SER* measures the magnitude of a typical deviation from the regression line.
- *SER* has the same units of the dependent variable.



# OLS regression in gretl

---

- From the menu:

> Model > Ordinary least squares >

- Scripts:

**ols** testscr **const** str

dependent variable

regressors

# Regression results in gretl

Model 1: OLS, using observations 1–420

Dependent variable: testscr

	coefficient	std. error	t-ratio	p-value	
const	698.933	9.46749	73.82	6.57e-242	***
str	-2.27981	0.479826	-4.751	2.78e-06	***

Mean dependent var	654.1565	S.D. dependent var	19.05335
Sum squared resid	144315.5	S.E. of regression	18.58097
R-squared	0.051240	Adjusted R-squared	0.048970
F(1, 418)	22.57511	P-value(F)	2.78e-06
Log-likelihood	-1822.250	Akaike criterion	3648.499
Schwarz criterion	3656.580	Hannan-Quinn	3651.693

The least square assumptions

# The least squares assumptions

---

For the linear regression model

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

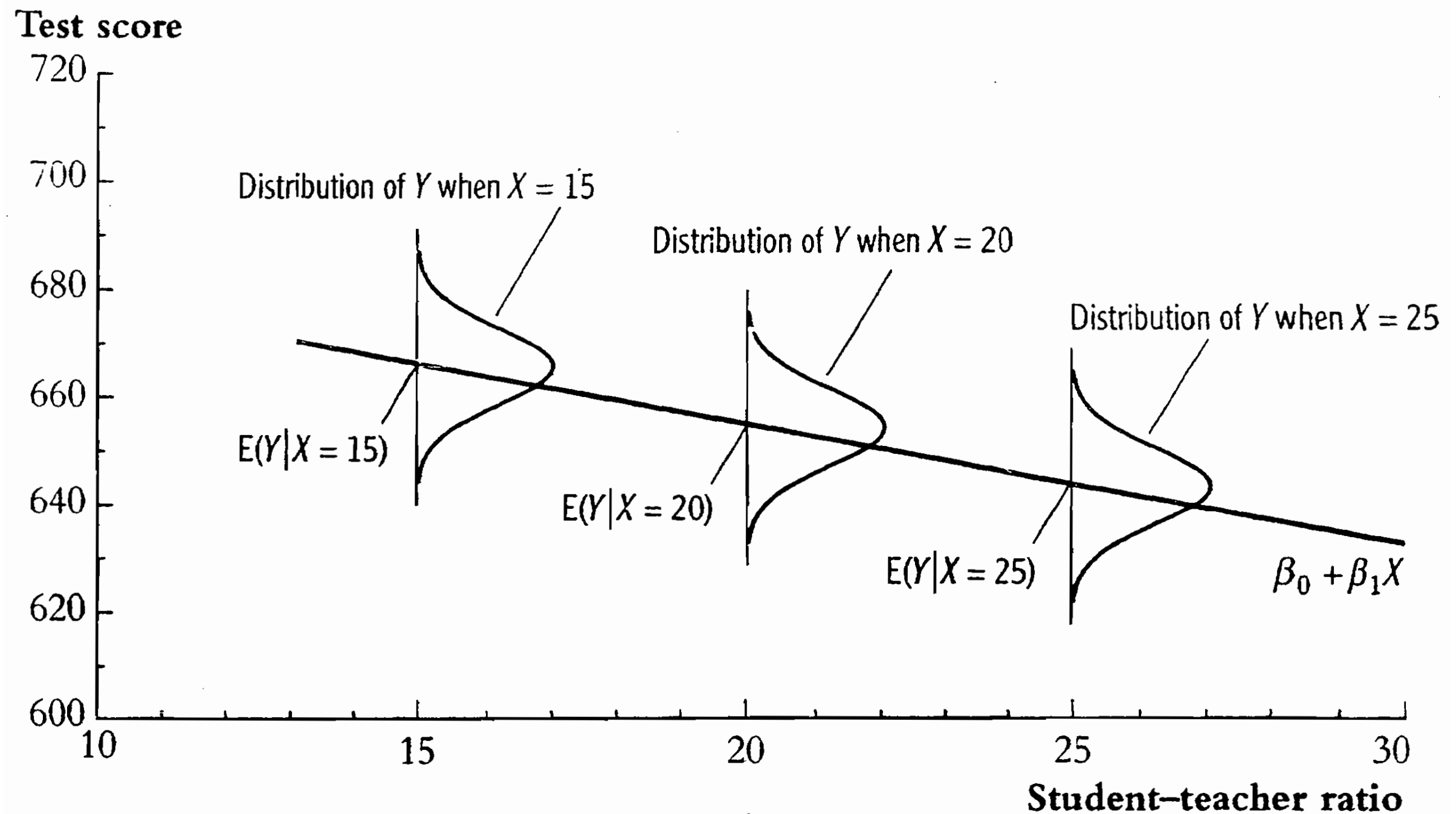
it is assumed that:

1. The error term  $u_i$  has conditional mean zero given  $X_i$ :

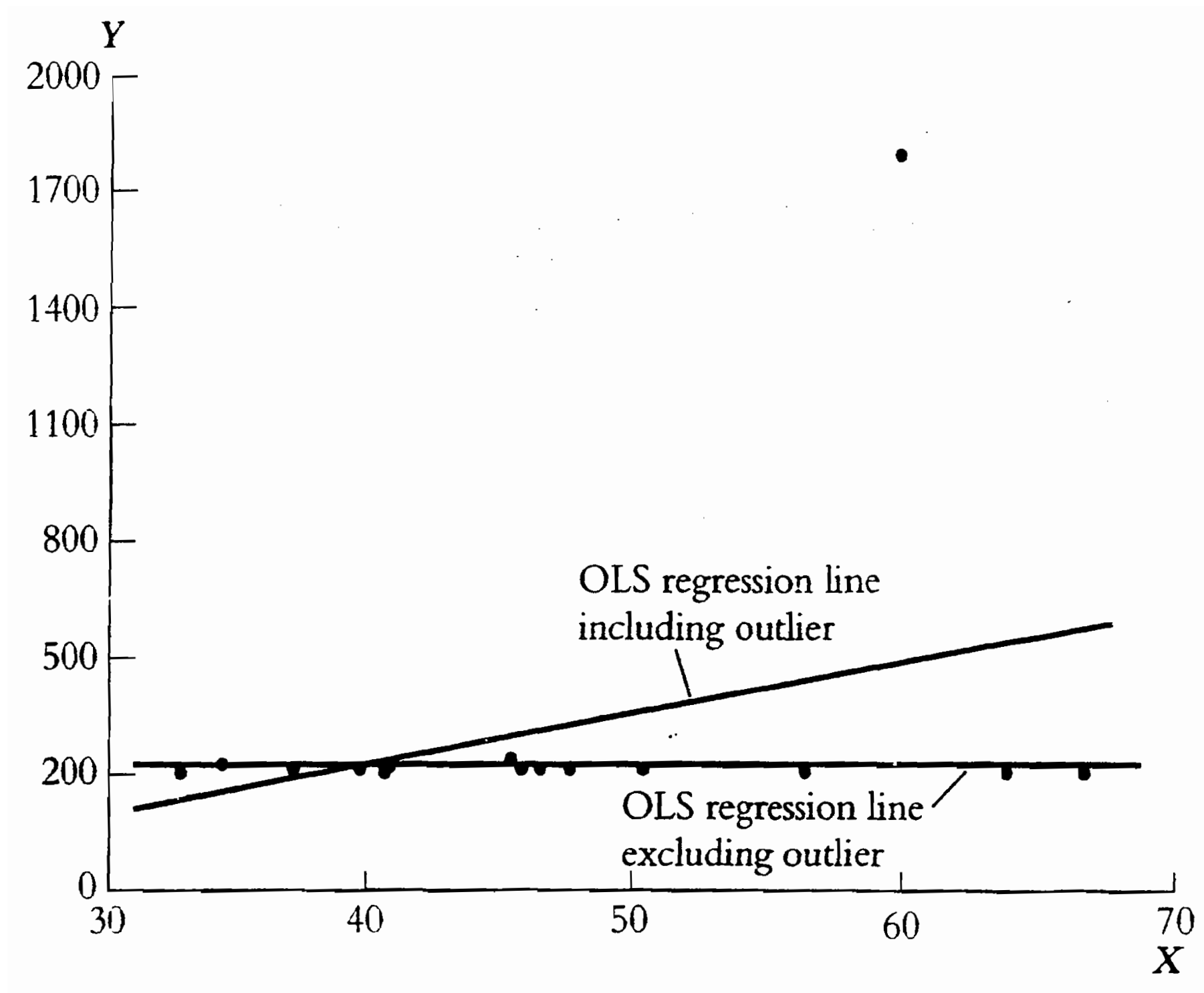
$$E(u_i \mid X_i) = 0 \quad (\Rightarrow \text{corr}(X_i, u_i) = 0)$$

2.  $(X_i, Y_i)$ ,  $i = 1, \dots, n$ , are i.i.d. draws from their joint distribution;  
and
3. Large outliers are unlikely:  $X_i$  and  $Y_i$  have nonzero finite fourth moments.

# Implication of $E(u_i | X_i) = 0$



# Linear regression is sensitive to outliers



# Hypothesis tests and confidence intervals

# Large-sample distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$

---

If the least square assumptions hold, then in large samples  $\hat{\beta}_0$  and  $\hat{\beta}_1$  have a jointly normal sampling distribution.

The large-sample distribution of  $\hat{\beta}_1$  is  $N(\beta_1, \sigma_{\hat{\beta}_1}^2)$ , where

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

The large-sample distribution of  $\hat{\beta}_0$  is  $N(\beta_0, \sigma_{\hat{\beta}_0}^2)$ , where

$$\sigma_{\hat{\beta}_0}^2 = \frac{1}{n} \frac{\text{var}(H_i u_i)}{[E(H_i^2)]^2}, \text{ where } H_i = 1 - \left[ \frac{\mu_X}{E(X_i^2)} \right] X_i$$



# Hypotheses concerning $\beta_1$

---

- Two-sided hypotheses

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

- The  $t$ -statistic

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

where

$$SE(\hat{\beta}_1) = \sqrt{\hat{\sigma}_{\hat{\beta}_1}^2}, \quad \hat{\sigma}_{\hat{\beta}_1}^2 = \frac{1}{n} \times \frac{\frac{1}{n-2} \sum_{i=1}^n (X_i - \bar{X})^2 \hat{u}_i^2}{\left[\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2\right]^2}$$

- The  $p$ -value

$$p\text{-value} = 2\Phi(-|t^{act}|)$$

## Confidence interval for $\beta_1$

---

- The 95% confidence interval for  $\beta_1$  is

$$[\hat{\beta}_1 - 1.96 SE(\hat{\beta}_1), \hat{\beta}_1 + 1.96 SE(\hat{\beta}_1)]$$

# Regression results in gretl

Model 1: OLS, using observations 1–420

Dependent variable: testscr

	coefficient	std. error	t-ratio	p-value	
const	698.933	9.46749	73.82	6.57e-242	***
str	-2.27981	0.479826	-4.751	2.78e-06	***

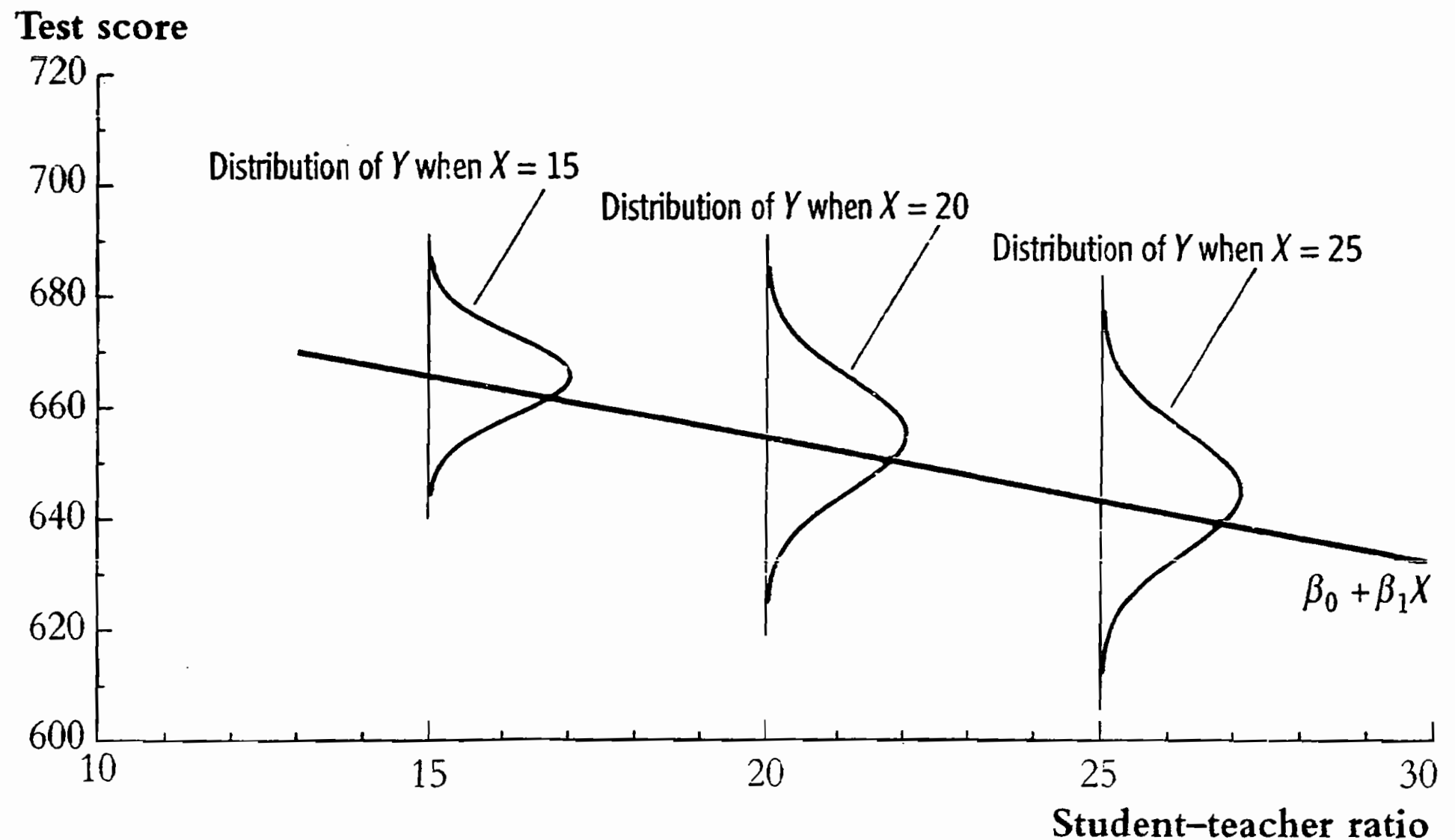
Mean dependent var	654.1565	S.D. dependent var	19.05335
Sum squared resid	144315.5	S.E. of regression	18.58097
R-squared	0.051240	Adjusted R-squared	0.048970
F(1, 418)	22.57511	P-value(F)	2.78e-06
Log-likelihood	-1822.250	Akaike criterion	3648.499
Schwarz criterion	3656.580	Hannan-Quinn	3651.693

# Heteroskedasticity and homoskedasticity

# An example of heteroskedasticity

**FIGURE 5.2** An Example of Heteroskedasticity

Like Figure 4.4, this shows the conditional distribution of test scores for three different class sizes. Unlike Figure 4.4, these distributions become more spread out (have a larger variance) for larger class sizes. Because the variance of the distribution of  $u$  given  $X$ ,  $\text{var}(u|X)$ , depends on  $X$ ,  $u$  is heteroskedastic.



# Definition

---

The error term  $u_i$  is *homoskedastic* if the variance of the conditional distribution of  $u_i$  given  $X_i$ ,

$$\text{var}(u_i \mid X_i = x)$$

is constant for  $i = 1, \dots, n$  and in particular does not depend on  $x$ .

Otherwise, the error term is *heteroskedastic*.

# Implications of homoskedasticity + least square assumptions

---

- The OLS estimators of coefficients are **efficient** among all estimators that are **linear** in  $Y_1, \dots, Y_n$ . [**BLUE**]
- The standard errors of  $\hat{\beta}_0$  and  $\hat{\beta}_1$  reduce to simpler form, e.g.,

$$SE(\hat{\beta}_1) = \sqrt{\tilde{\sigma}_{\hat{\beta}_1}^2}$$

where

$$\tilde{\sigma}_{\hat{\beta}_1}^2 = \frac{s_{\hat{u}}^2}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

# Standard errors of $\hat{\beta}_1$

---

- Homoskedasticity-only standard error

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

- Heteroskedasticity-robust standard error (HC1)

$$SE(\hat{\beta}_1) = \sqrt{\hat{\sigma}_{\hat{\beta}_1}^2}, \quad \hat{\sigma}_{\hat{\beta}_1}^2 = \frac{1}{n} \times \frac{\frac{1}{n-2} \sum_{i=1}^n (X_i - \bar{X})^2 \hat{u}_i^2}{\left[\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2\right]^2}$$



# In practice

---

- If the errors are heteroskedastic but the homoskedastic-only formulas are used
  - ⇒  $t$ -statistic does not have a standard normal distribution, even in large samples
- If the errors are homoskedastic but the heteroskedastic-robust formulas are used
  - ⇒ hypothesis tests and confidence intervals will be valid
- Always use heteroskedastic-robust standard errors

Practice in gretl

# Heteroskedasticity-robust estimation in gretl

---

- Settings for the whole script

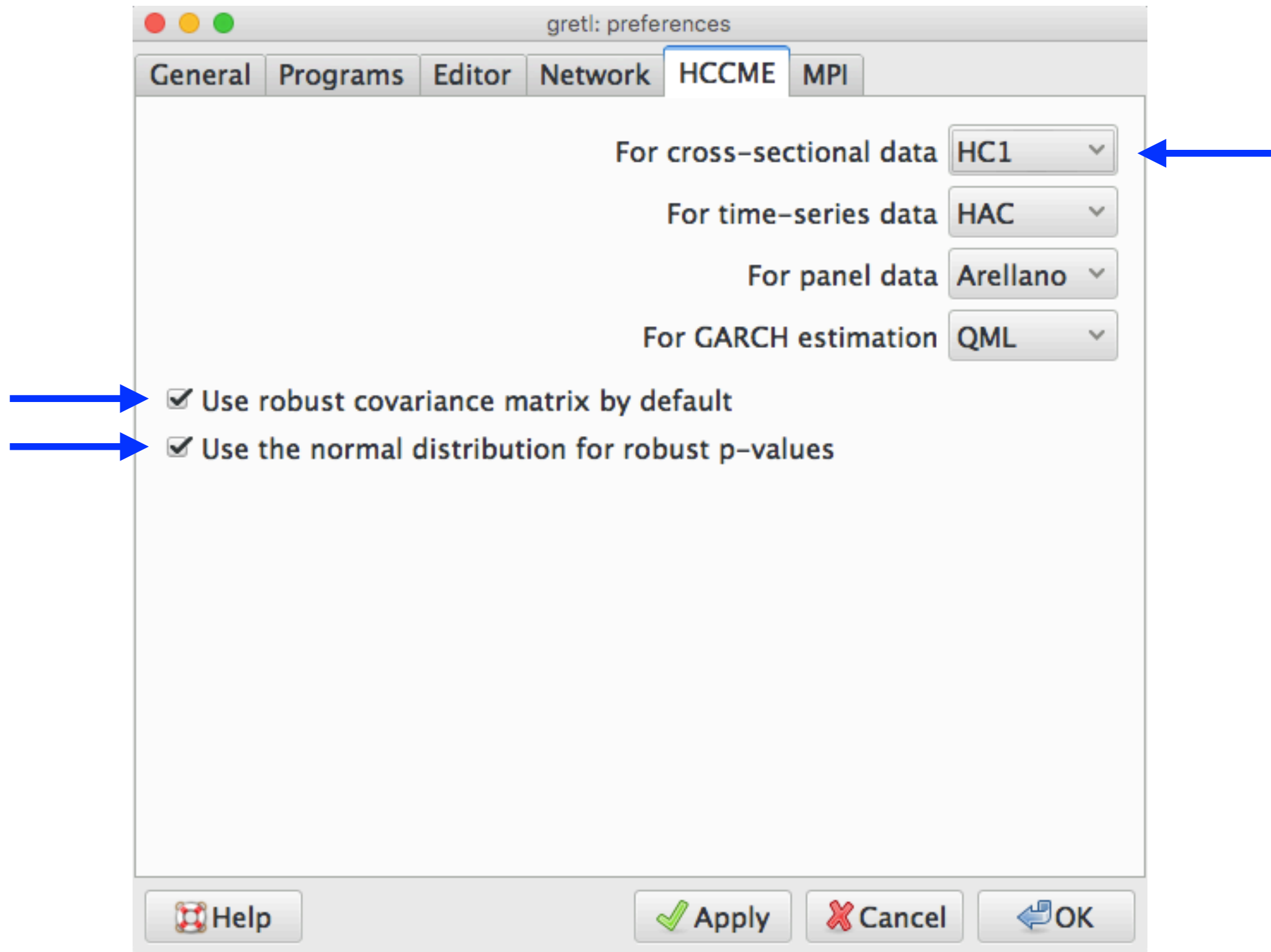
```
set force_hc on  
set hc_version 1  
      # 0 (the original White's) is the default  
set robust_z on
```

- For single resression

```
ols yvar xvar --robust
```

(you still need to set the HC version)

# Settings in the preferences of gretl



# Regression results in gretl (homoskedasticity-only)

Model 1: OLS, using observations 1–420

Dependent variable: testscr

	coefficient	std. error	t-ratio	p-value	
const	698.933	9.46749	73.82	6.57e-242	***
str	-2.27981	0.479826	-4.751	2.78e-06	***
Mean dependent var	654.1565	S.D. dependent var	19.05335		
Sum squared resid	144315.5	S.E. of regression	18.58097		
R-squared	0.051240	Adjusted R-squared	0.048970		
F(1, 418)	22.57511	P-value(F)	2.78e-06		
Log-likelihood	-1822.250	Akaike criterion	3648.499		
Schwarz criterion	3656.580	Hannan-Quinn	3651.693		

# Regression results in gretl (heteroskedasticity-robust with normal distribution)

---

Model 1: OLS, using observations 1–420

Dependent variable: testscr

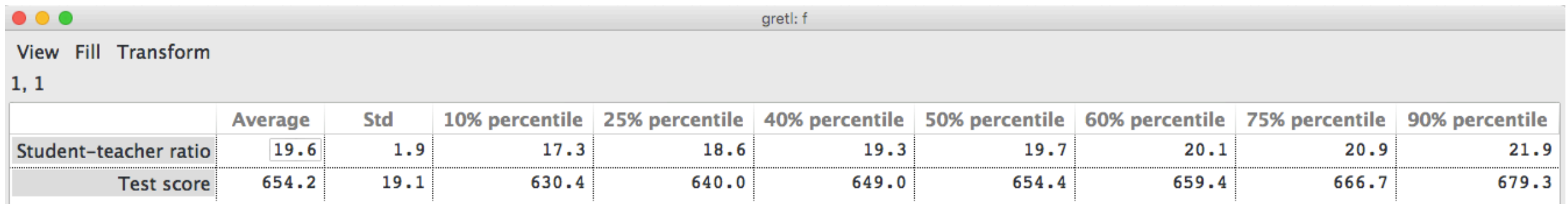
Heteroskedasticity-robust standard errors, variant HC1

	coefficient	std. error	z	p-value	
const	698.933	10.3644	67.44	0.0000	***
str	−2.27981	0.519489	−4.389	1.14e−05	***

Mean dependent var	654.1565	S.D. dependent var	19.05335
Sum squared resid	144315.5	S.E. of regression	18.58097
R-squared	0.051240	Adjusted R-squared	0.048970
F(1, 418)	19.25943	P-value(F)	0.000014
Log-likelihood	−1822.250	Akaike criterion	3648.499
Schwarz criterion	3656.580	Hannan-Quinn	3651.693

# Exercises

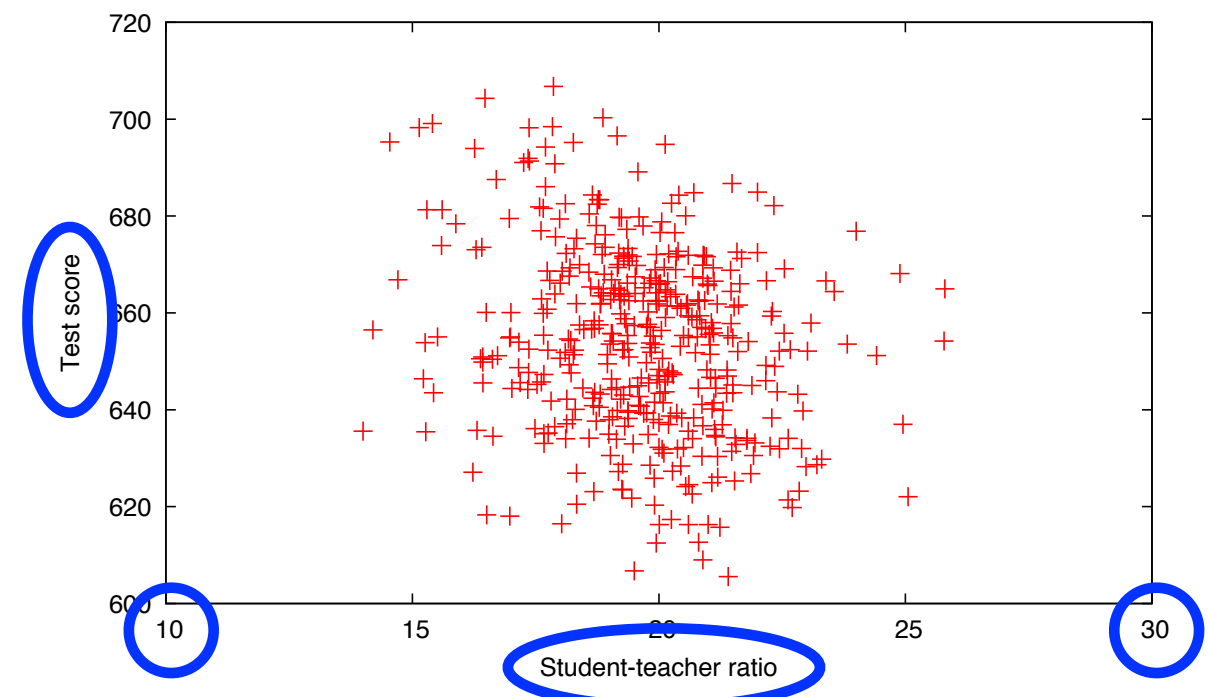
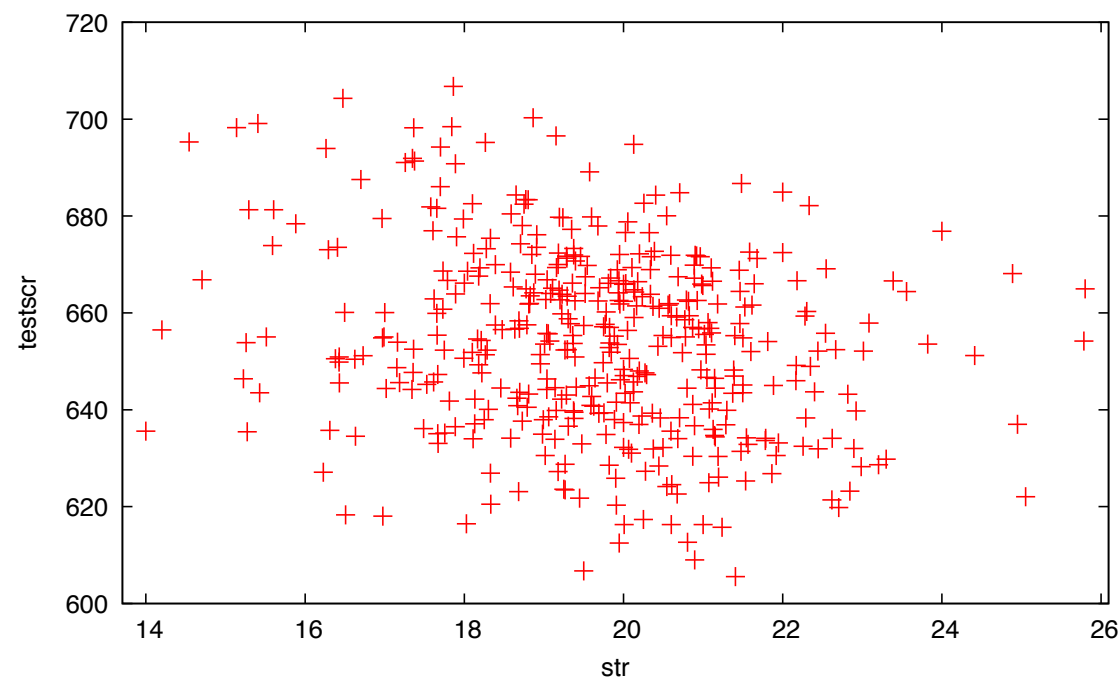
1. Reproduce Table 4.1 using matrix.



The screenshot shows the gretl software window with a menu bar (View, Fill, Transform) and a status bar (1, 1). Below the menu is a table with the following data:

	Average	Std	10% percentile	25% percentile	40% percentile	50% percentile	60% percentile	75% percentile	90% percentile
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	649.0	654.4	659.4	666.7	679.3

2. Learn command `gnuplot` (or `plot`) and reproduce Figure 4.2 with appropriate titles and ranges of axes.



`gnuplot testscr str --output=display --fit=none`

## Exercises (cont.)

---

3. Find the relation and differences among  $\hat{\beta}_1$  ,  $r_{XY}$  , and  $R^2$  by solving Exercises 4.9 and 4.12.
4. Learn Section 5.3.



# References

---

1. Stock, J. H. and Watson, M. M., *Introduction to Econometrics*, 3rd Edition, Pearson, 2012.
2. *Gretl User's Guide*