Econometrics 1 *Applied Econometrics with R*

Lecture 7: Linear Regression

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
Tel: (0755) 2695 0548
Website: https://huangjp.com

Econometrics is the *science* and *art* of using economic theory and statistical techniques to analyze economic data.

Typical questions considered by econometricians

- Does reducing class size improve elementary school education?
- Is there racial discrimination in the market for home loans?
- How much do cigarette taxes reduce smoking?
- What will the rate of inflation be next year?

Linear regression with one regressor

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*:

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

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

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

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

coefficients

error term

dependent variable

independent variable / regressor

The linear regression model

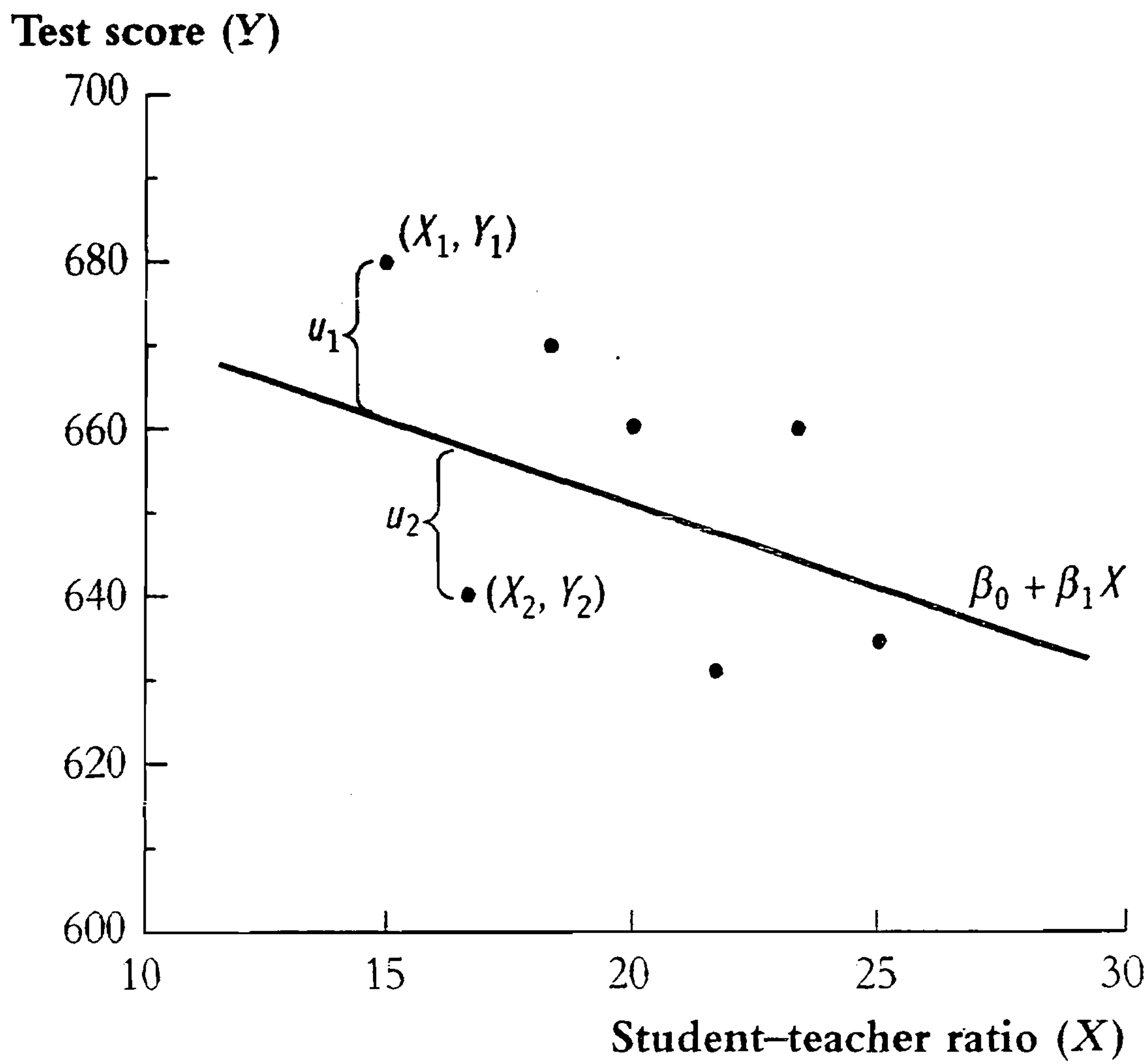
- The linear regression model with one regressor

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

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^{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^{th} observation.

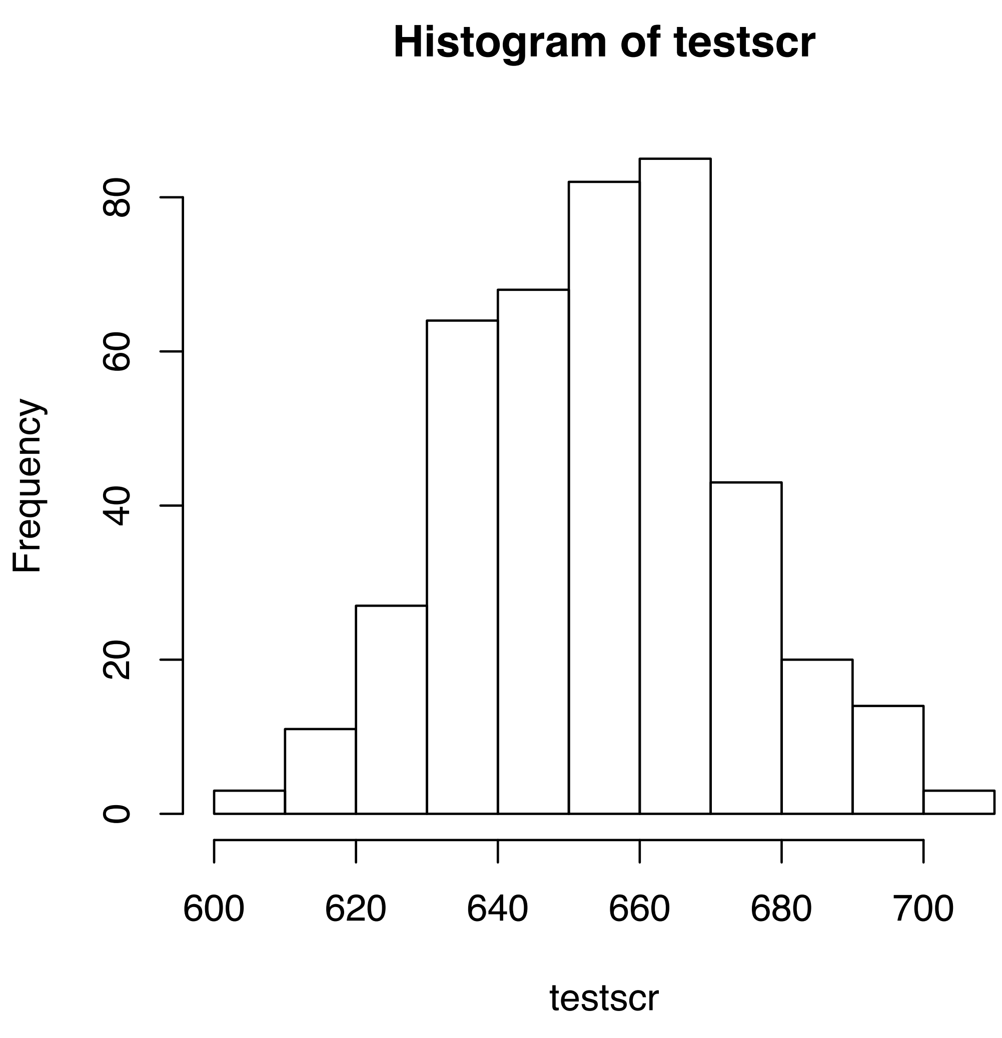


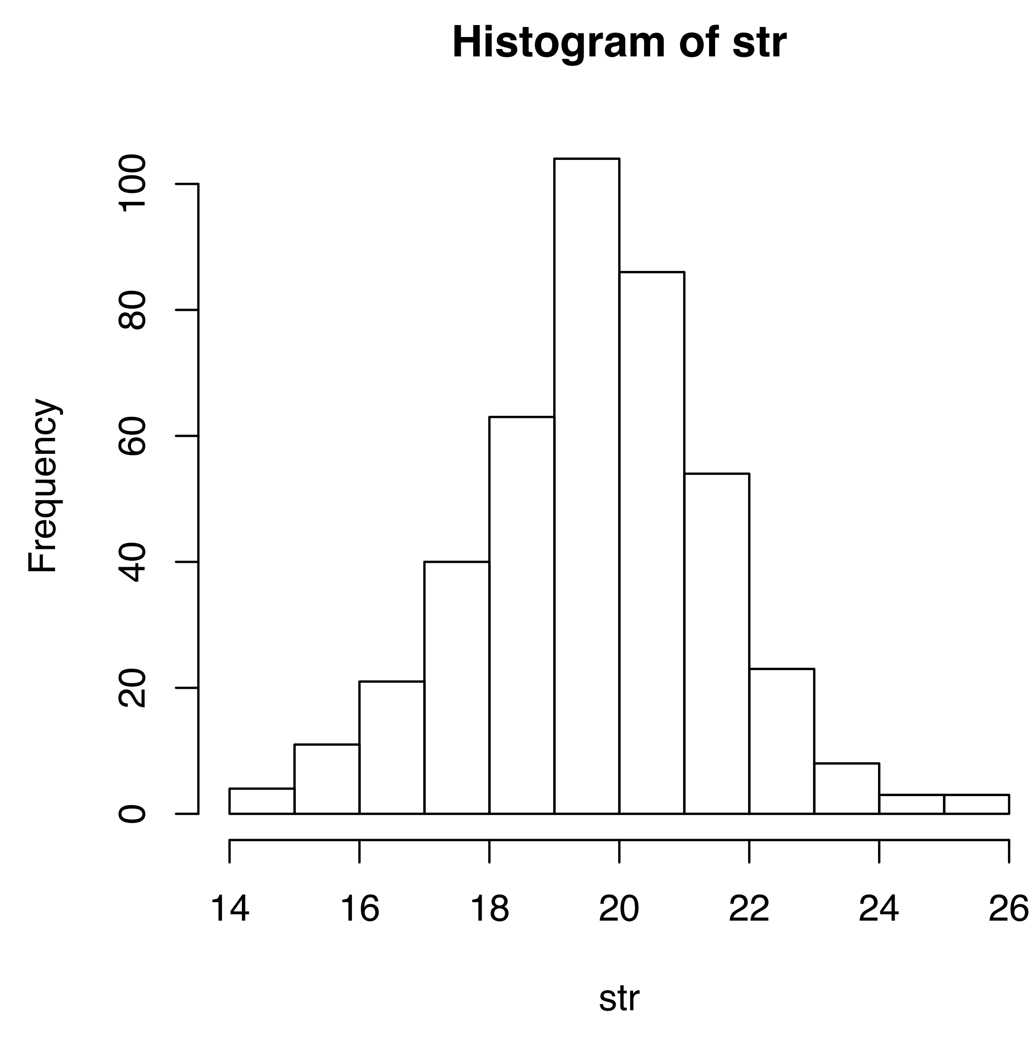
A test score data in California

- 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`

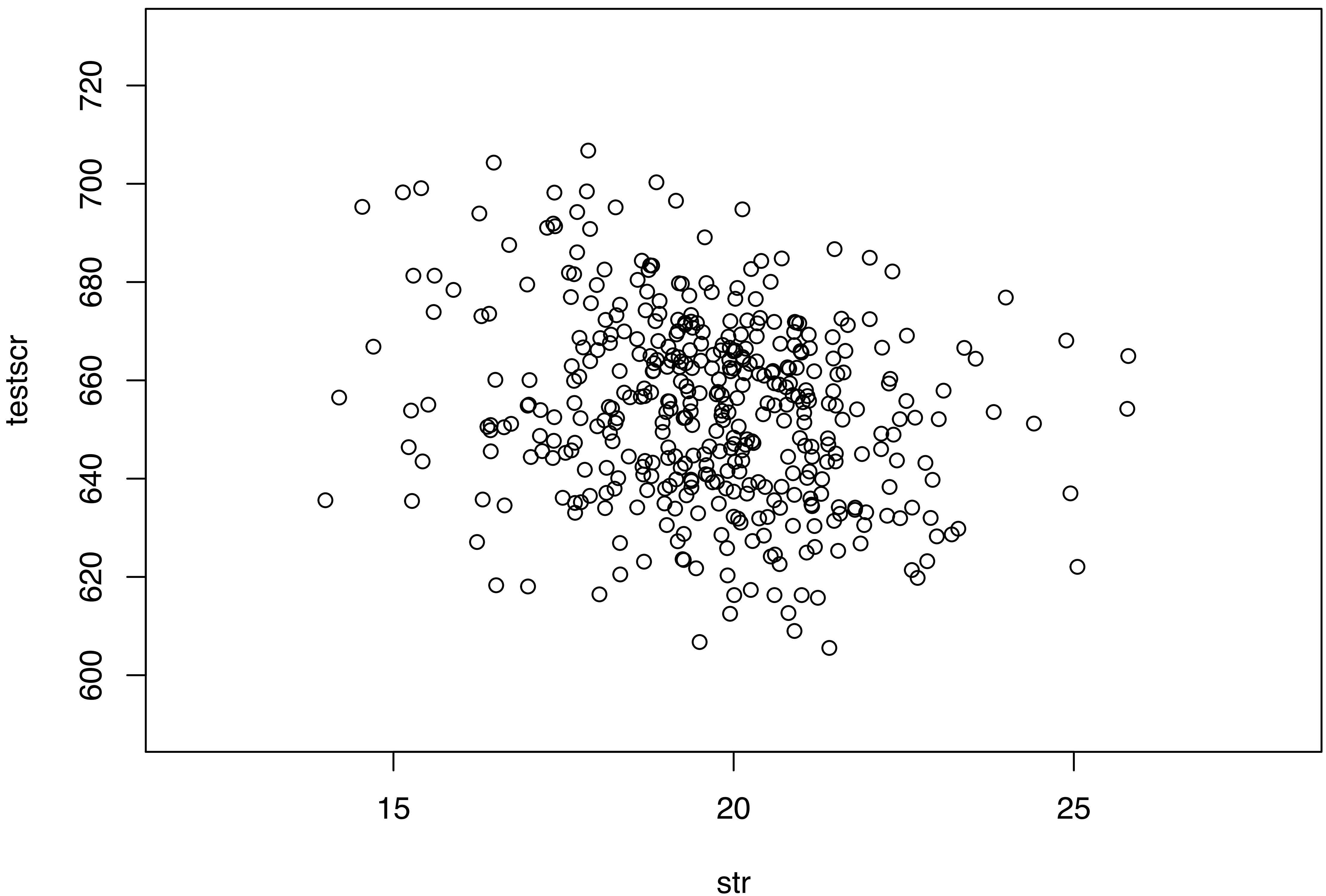
Average test score v.s. student-teacher ratio

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

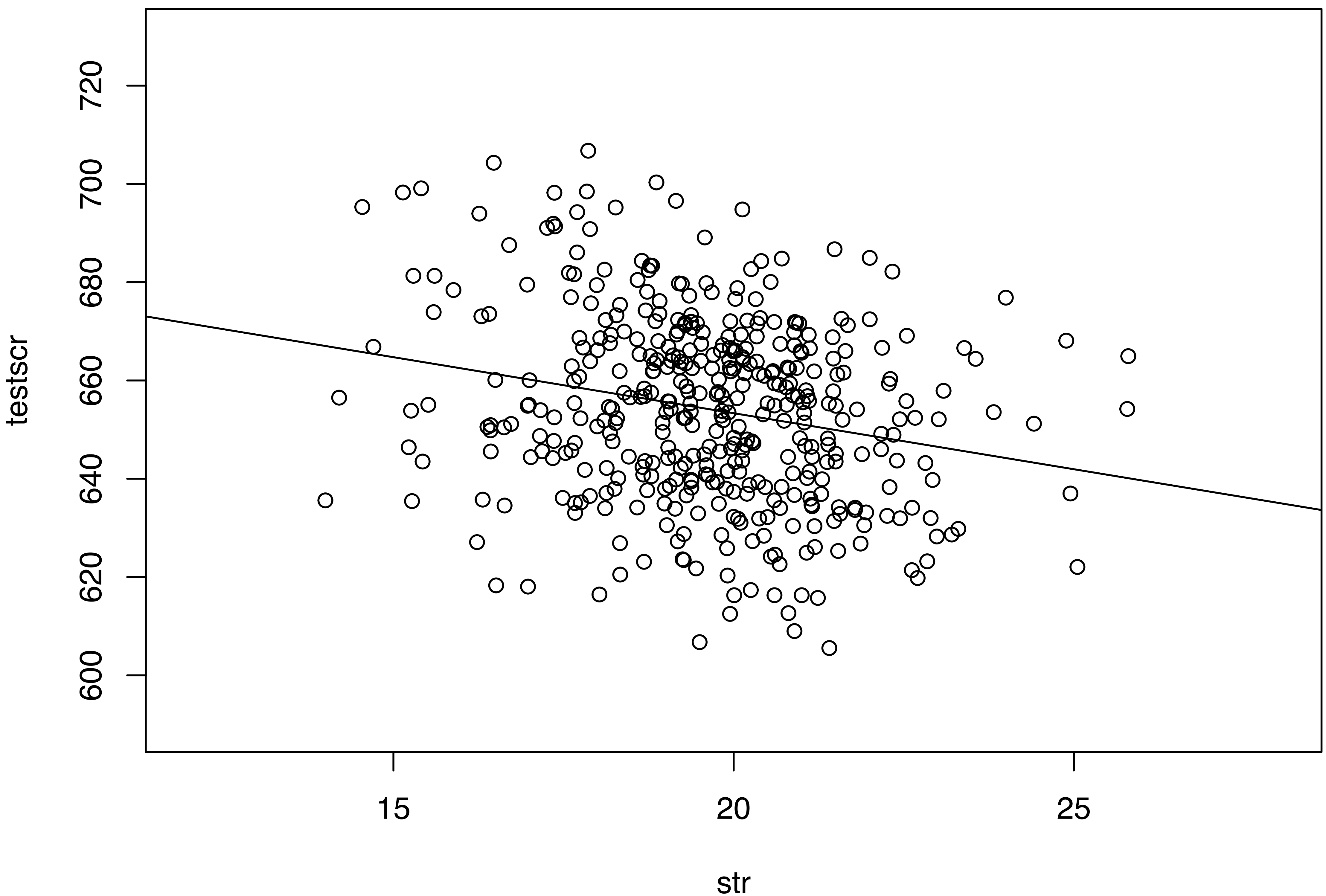




Average test score v.s. student-teacher ratio



Average test score v.s. student-teacher ratio

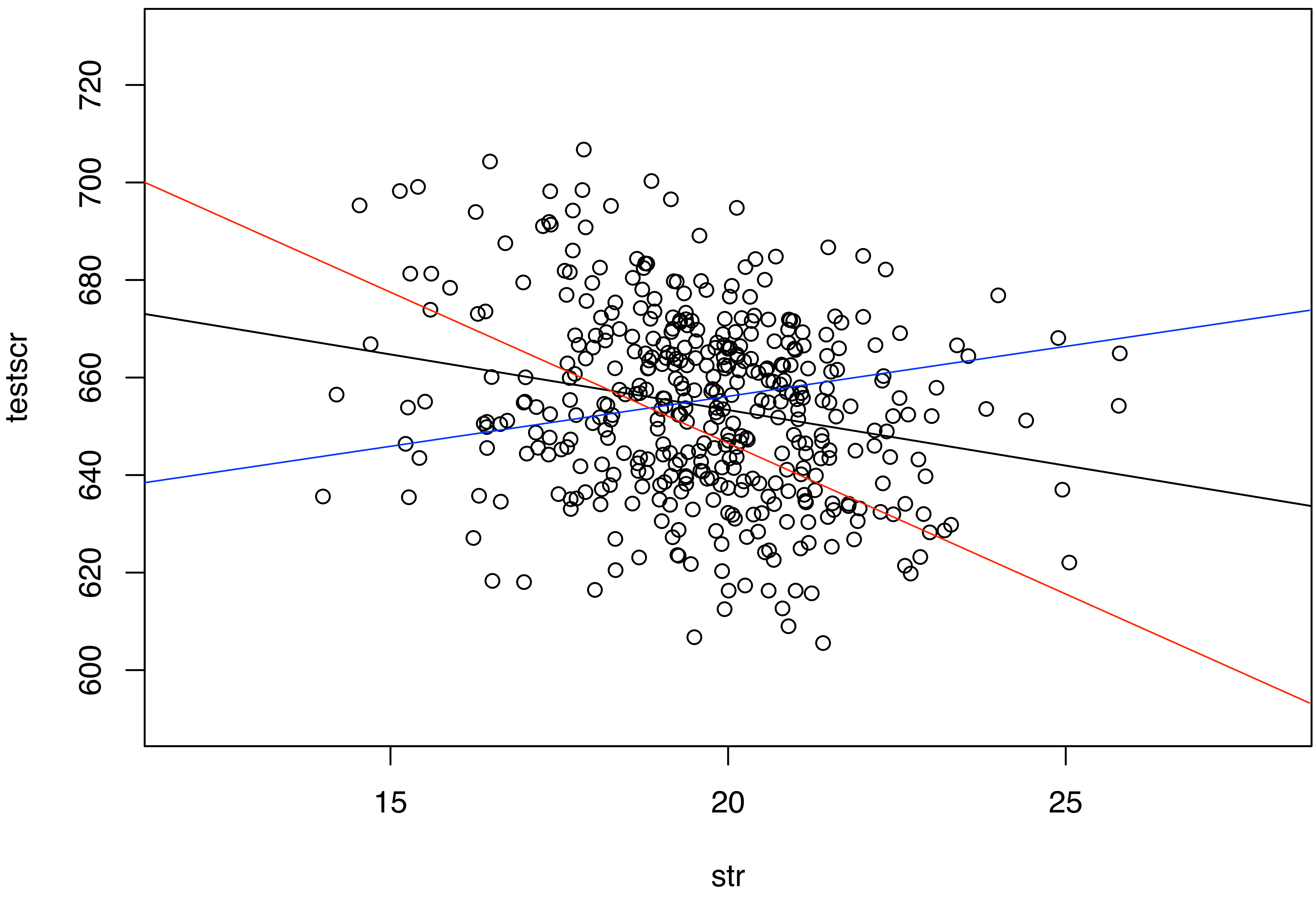


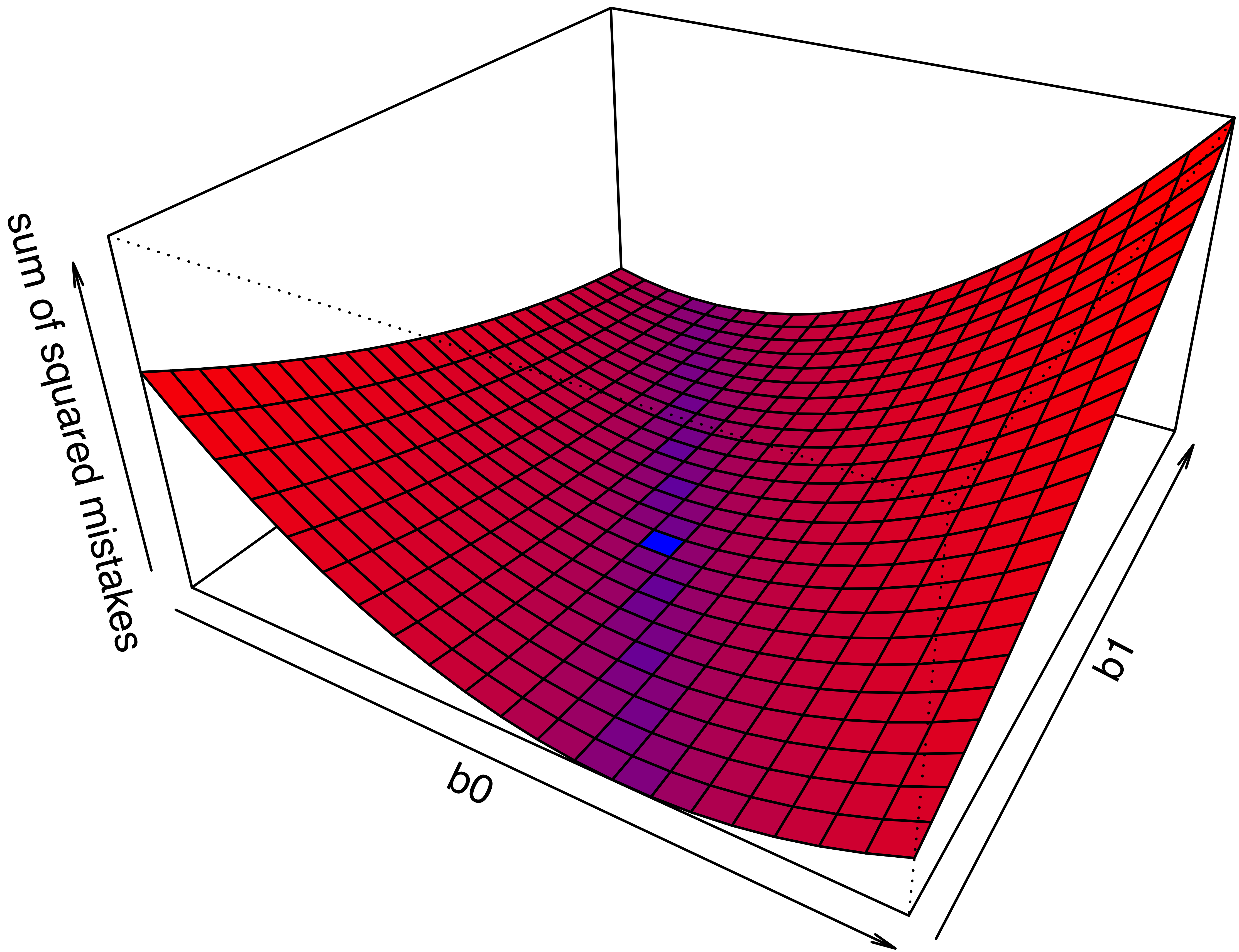
Estimating the coefficients

- $\overline{Y}$ is an estimator of the population mean.
- Similarly, we need estimators of the coefficients β_0 and β_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 of squared mistakes
b0
b1

Practice

- Import data from `caschool.xlsx`

- Take `str` as the independent variable (X) and `testscr` as the dependent variable (Y). Draw the histograms and the scatter plot.

- Calculate the OLS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ using local grid search.

    1. Specify a set of possible values for (b_0, b_1)

    2. For each possible (b_0, b_1), compare $\sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2$

The OLS estimator, predicted values, and residuals

- The OLS estimators of the slope and the intercept are

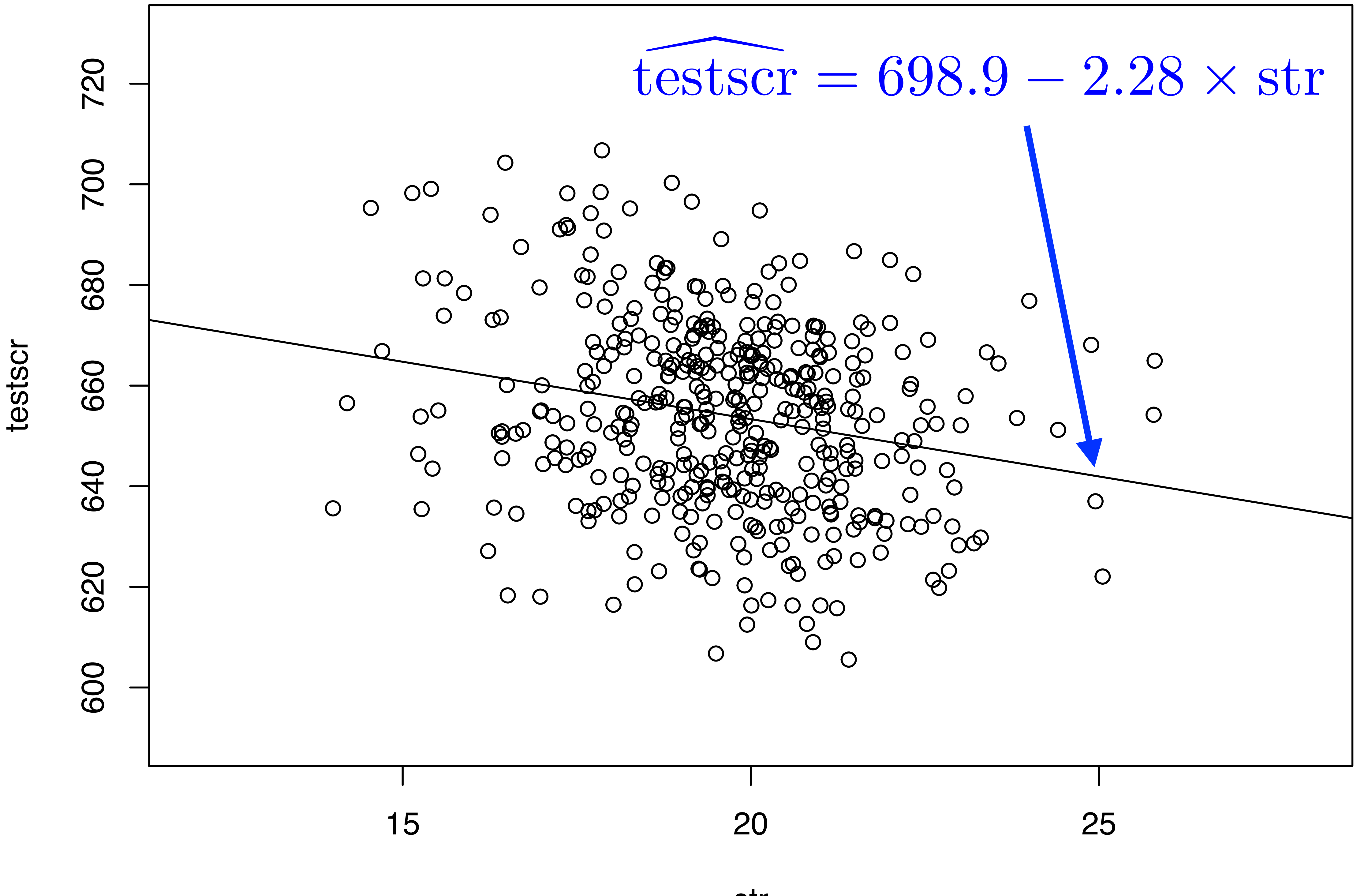
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{s_{XY}}{s_X^2}$$

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

- The OLS predicted value: $\hat{Y}_i = \boxed{\hat{\beta}_0 + \hat{\beta}_1 X_i}$
- 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.

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

A measure of fit

- The R^2 — correlation of determination, the fraction of the sample variance of Y_i explained by X_i.

- Recall that $Y_i = \hat{Y}_i + \hat{u}_i$

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2} = \frac{ESS}{TSS}$$

(**e**xplained **s**um of **s**quares)

(**t**otal **s**um of **s**quares)

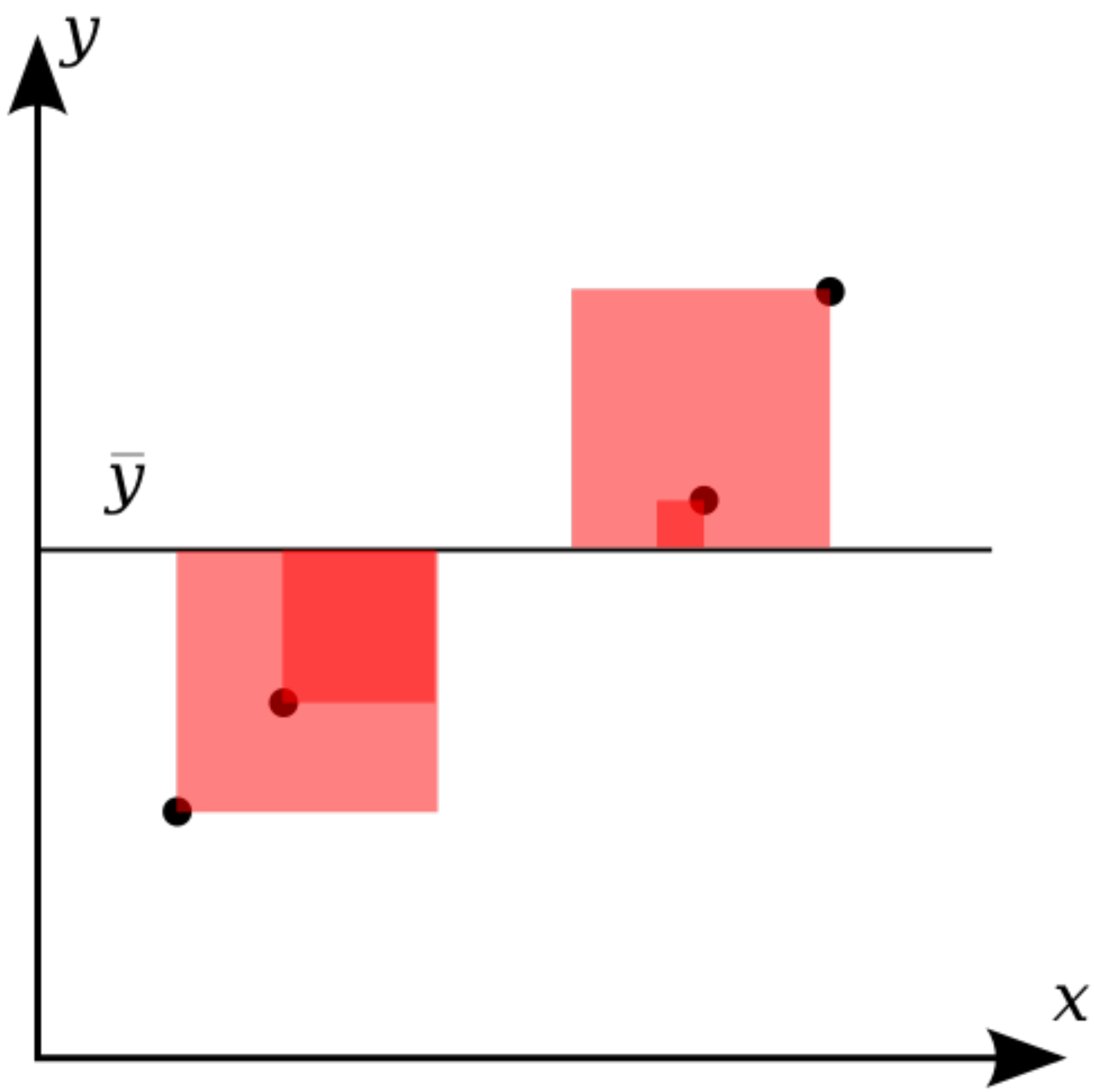
$$= 1 - \frac{\sum_{i=1}^{n} \hat{u}_i^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2} = 1 - \frac{SSR}{TSS}$$

(**s**um of **s**quared **r**esiduals)

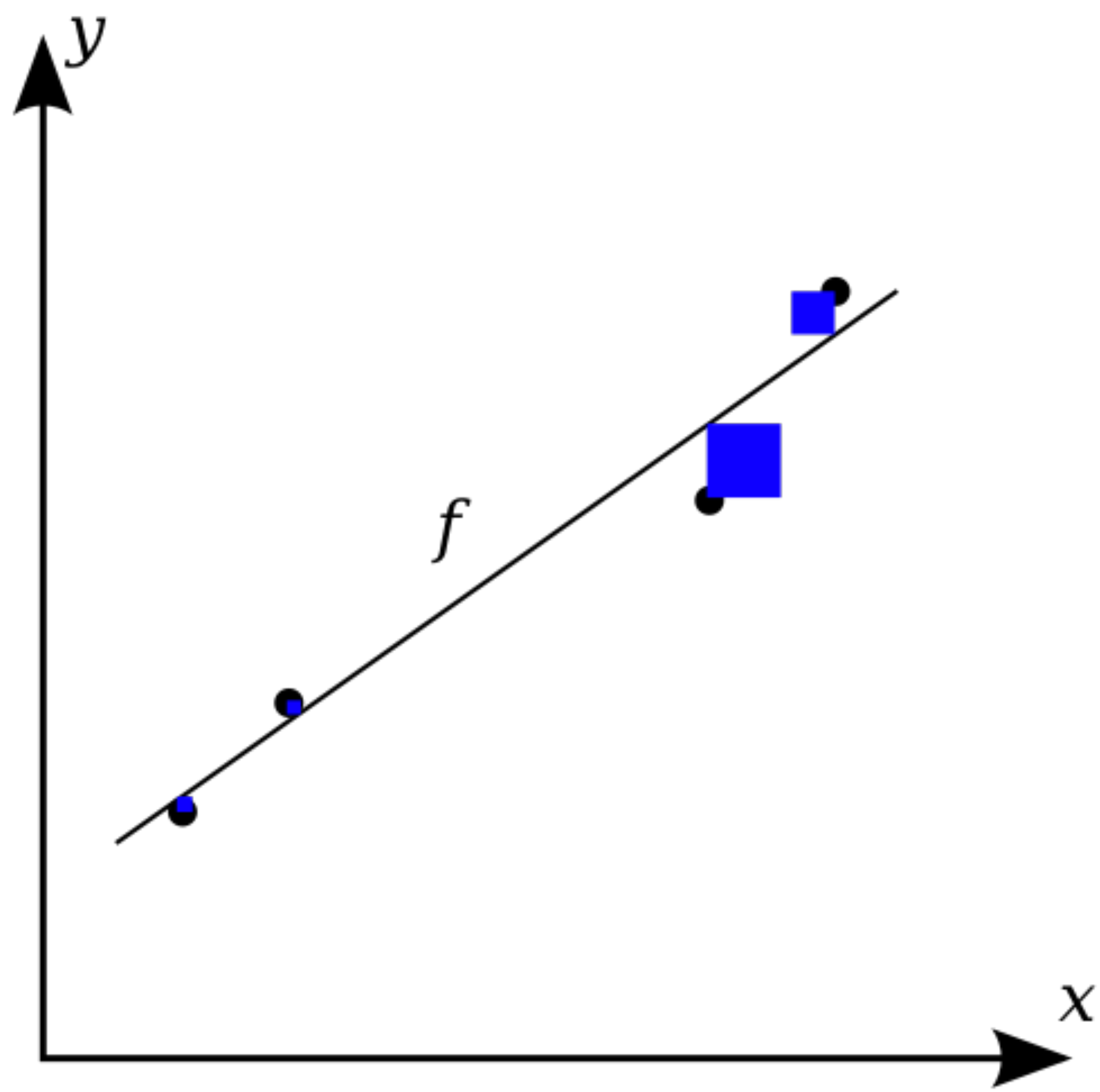
Read Appendix 4.3 if you want to know why the second equality holds.

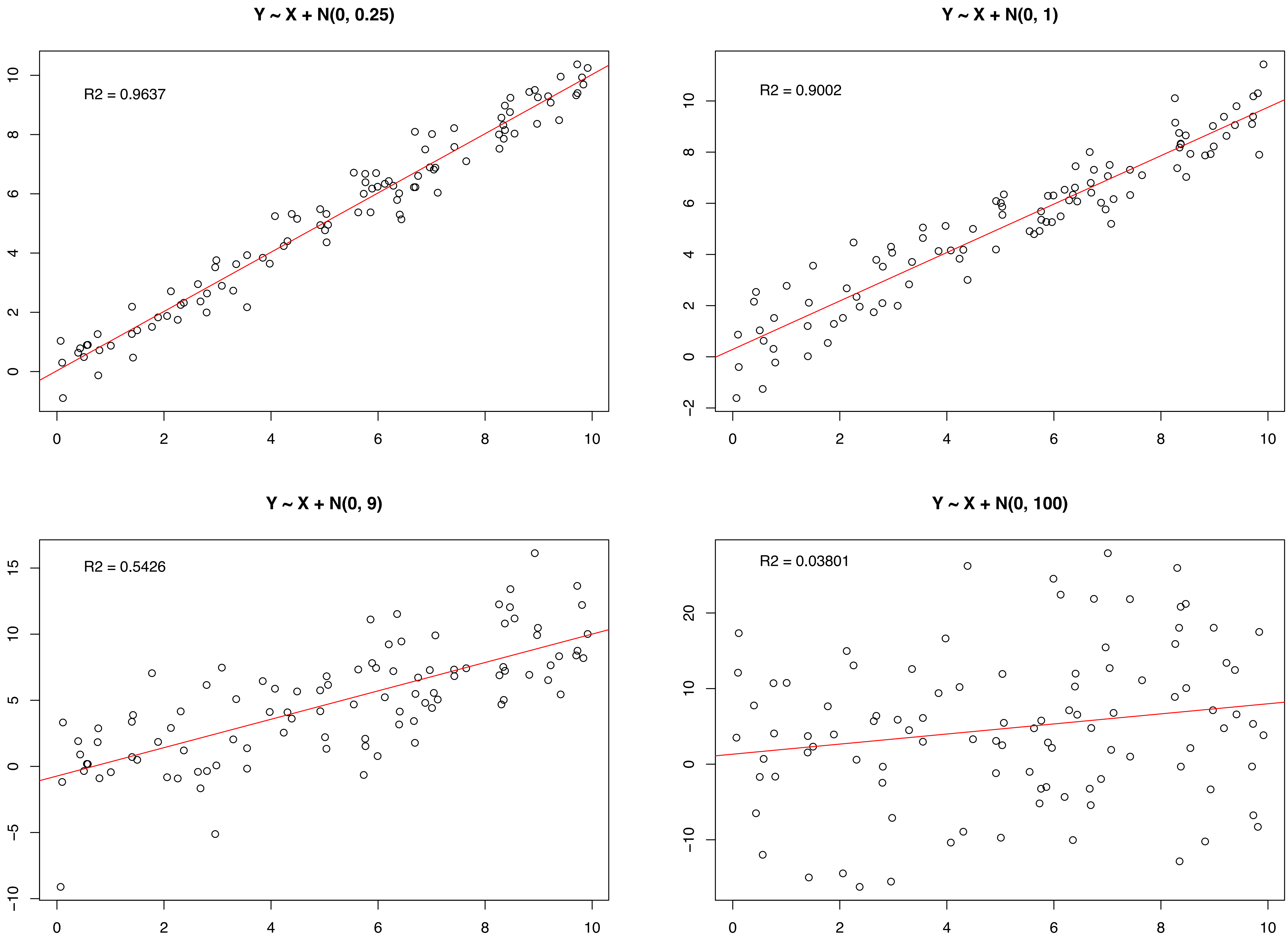
A graphical explanation of SSR

Simple average of Y_i



OLS regression




Y ~ X + N(0, 0.25)
R2 = 0.9637
Y ~ X + N(0, 1)
R2 = 0.9002
Y ~ X + N(0, 9)
R2 = 0.5426
Y ~ X + N(0, 100)
R2 = 0.03801

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.

Practice

- Use the formula to recalculate the OLS estimates of coefficients in `testscr` and `str` regression model.

- Use the formula to calculate the R^2 of this model, and give an explanation of your result.

The least squares assumptions

For the linear regression model

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

it is assumed that:

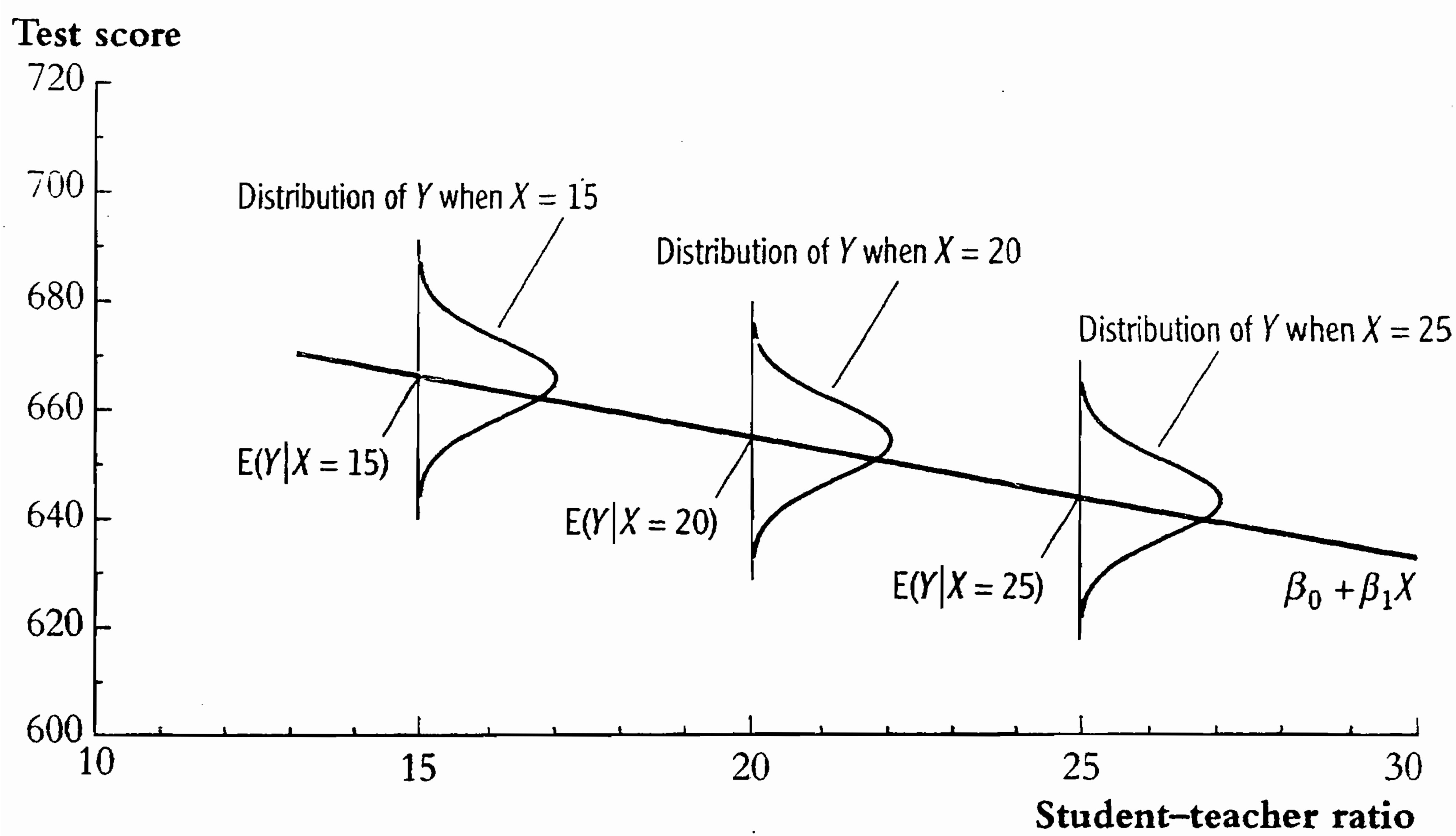
1. The error term u_i has conditional mean zero given X_i:

$$\mathrm{E}(u_i \mid X_i) = 0 \qquad (\Rightarrow \mathrm{corr}(X_i, u_i) = 0)$$

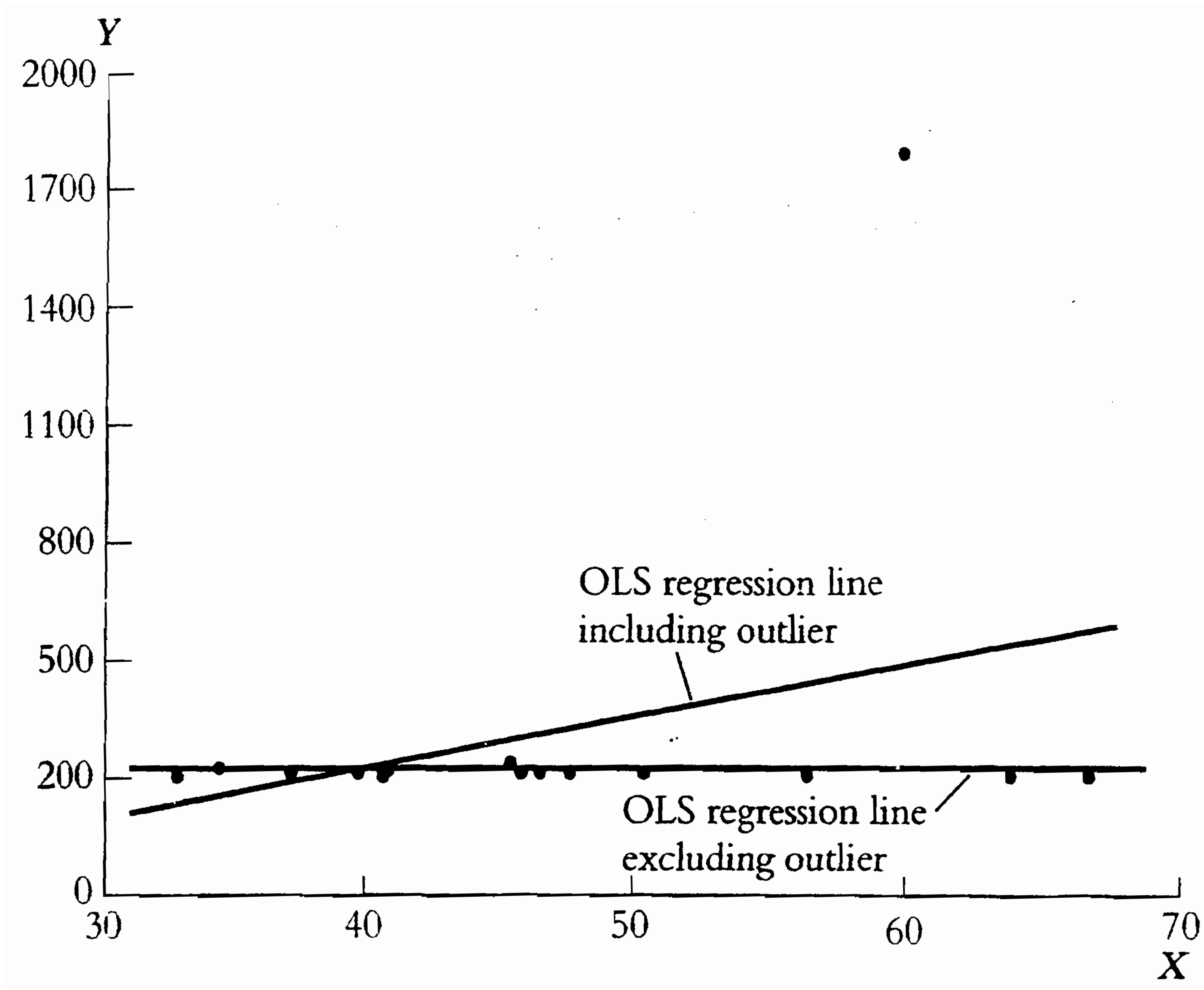
2. (X_i, Y_i), $i = 1, \ldots, n$, are i.i.d. draws from their joint distribution; and

3. Large outliers are unlikely: X_i and Y_i have nonzero finite fourth moments.

Implication of $\mathrm{E}(u_i \mid X_i) = 0$



Linear regression is sensitive to outliers



References

1. Stock, J. H. and Watson, M. M., *Introduction to Econometrics*, 3rd Edition, Pearson, 2012.