Econometrics 1 *Applied Econometrics with R*

Lecture 7: Review of Statistics (2) and Linear Regression

黄嘉平

中国经济特区研究中心 讲师

办公室：文科楼1726

E-mail: huangjp@szu.edu.cn

Tel: (0755) 2695 0548

Office hour: Mon./Tue. 13:00-14:00

Hypothesis testing of population means

- Practice the command `t.test` in hypothesis testing.

- A sample of data:
  1.95, 0.31, 0.47, 1.54, 1.64,
  2.99, 0.53, 1.21, 0.83, 1.45,
  3.46, 2.23, 1.17, 1.16, 0.36,
  1.76, 0.19, 0.43, 1.78, 1.56

- Test the hypothesis

$$H_0 : \mathrm{E}(Y) = 1 \qquad H_1 : \mathrm{E}(Y) \neq 1$$

(*t*-statistic, standard error, *p*-value)

Hypothesis testing of population means

- The sample is generated from an *F* distribution with d.f. = (3, 6)
- Search "F distribution" in wikipedia, and find the theoretical mean and variance of $F_{m,n}$.
- Redo your hypothesis testing with these new information. Compare the *p*-values obtained from `t.test`, large-sample formulas with unknown/known population variance. What have you learned?

p-value for large samples

- The p-value when the population mean is unknown

$$p\text{-value} = 2\Phi\left(-\left|\frac{\overline{Y}^{act} - \mu_{Y,0}}{s_Y/\sqrt{n}}\right|\right) = 2\Phi\left(-\left|\frac{\overline{Y}^{act} - \mu_{Y,0}}{SE(\overline{Y})}\right|\right)$$

- The p-value when the population mean is known

$$\begin{aligned} p\text{-value} &= \Pr_{H_0}\left[\left|\frac{\overline{Y} - \mu_{Y,0}}{\sigma_Y/\sqrt{n}}\right| > \left|\frac{\overline{Y}^{act} - \mu_{Y,0}}{\sigma_Y/\sqrt{n}}\right|\right] \\ &= 2\Phi\left(-\left|\frac{\overline{Y}^{act} - \mu_{Y,0}}{\sigma_Y/\sqrt{n}}\right|\right) \end{aligned}$$

```r
y <- c(1.95, 0.31, 0.47, 1.54, 1.64,
       2.99, 0.53, 1.21, 0.83, 1.45,
       3.46, 2.23, 1.17, 1.16, 0.36,
       1.76, 0.19, 0.43, 1.78, 1.56)
mu0 <- 1
ty <- t.test(y, mu = mu0)

# theoretical moments for F distribution with d.f. = (3,6)
d1 <- 3
d2 <- 6
pmean <- d2 / (d2 - 2)
pvar <- 2 * d2^2 * (d1 + d2 - 2) / (d1 * (d2 - 2)^2 * (d2 - 4))

# p-values under large sample assumption
estimate <- mean(y)
se <- sd(y) / sqrt(length(y))
tstat <- (estimate - mu0) / se

pvalue_un <- 2*pnorm(- abs(tstat))
pvalue_kn <- 2*pnorm(- abs((estimate - mu0) /
                       sqrt(pvar / length(y))))
```

A summary

- Sample size = 20. Sample estimate is 1.351.
- Population distribution is not normal, and population variance is known.
- `t.test` (which use Student *t* distribution for *t*-statistic and assume the population distribution is normal) gives *p*-value = 0.0920
- Large-sample formulas with unknown (known) population variance gives *p*-value = 0.0759 (0.4933)

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?

Sources and types of data

- Sources
  - Experimental data versus observational data
- Types
  - Cross-sectional data
  - Time series data
  - Panel data (longitudinal data)

Cross-sectional data

TABLE 1.1 A Cross-Sectional Data Set on Wages and Other Individual Characteristics

obsno	wage	educ	exper	female	married
1	3.10	11	2	1	0
2	3.24	12	22	1	1
3	3.00	11	2	0	0
4	6.00	8	44	0	1
5	5.30	12	7	0	1
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
525	11.56	16	5	0	1
526	3.50	14	5	1	0

Time series data

TABLE 1.3 Minimum Wage, Unemployment, and Related Data for Puerto Rico

obsno	year	avgmin	avgcov	prunemp	prgnp
1	1950	0.20	20.1	15.4	878.7
2	1951	0.21	20.7	16.0	925.0
3	1952	0.23	22.6	14.8	1015.9
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
37	1986	3.35	58.1	18.9	4281.6
38	1987	3.35	58.2	16.8	4496.7

Panel data

TABLE 1.5 A Two-Year Panel Data Set on City Crime Statistics

obsno	city	year	murders	population	unem	police
1	1	1986	5	350000	8.7	440
2	1	1990	8	359200	7.2	471
3	2	1986	2	64300	5.4	75
4	2	1990	1	65100	5.5	75
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
297	149	1986	10	260700	9.6	286
298	149	1990	6	245000	9.8	334
299	150	1986	25	543000	4.3	520
300	150	1990	32	546200	5.2	493

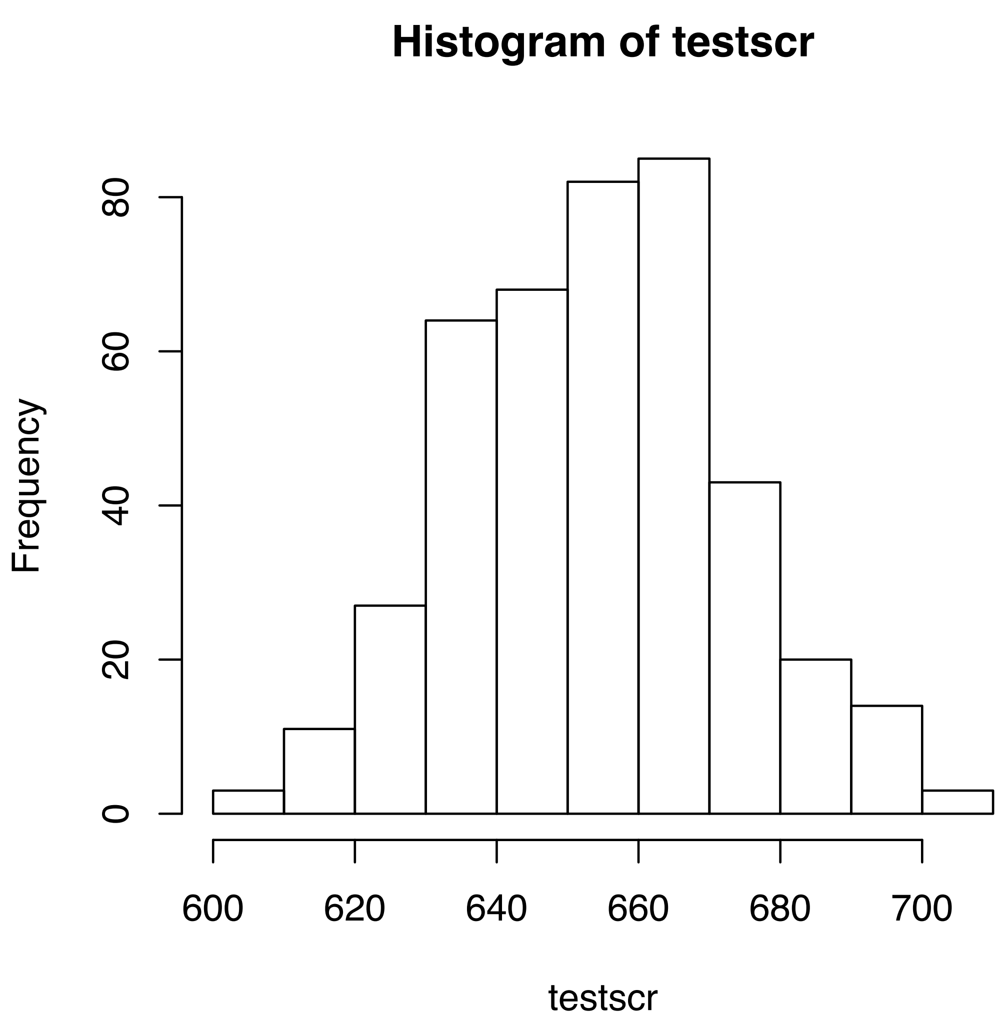
Linear regression

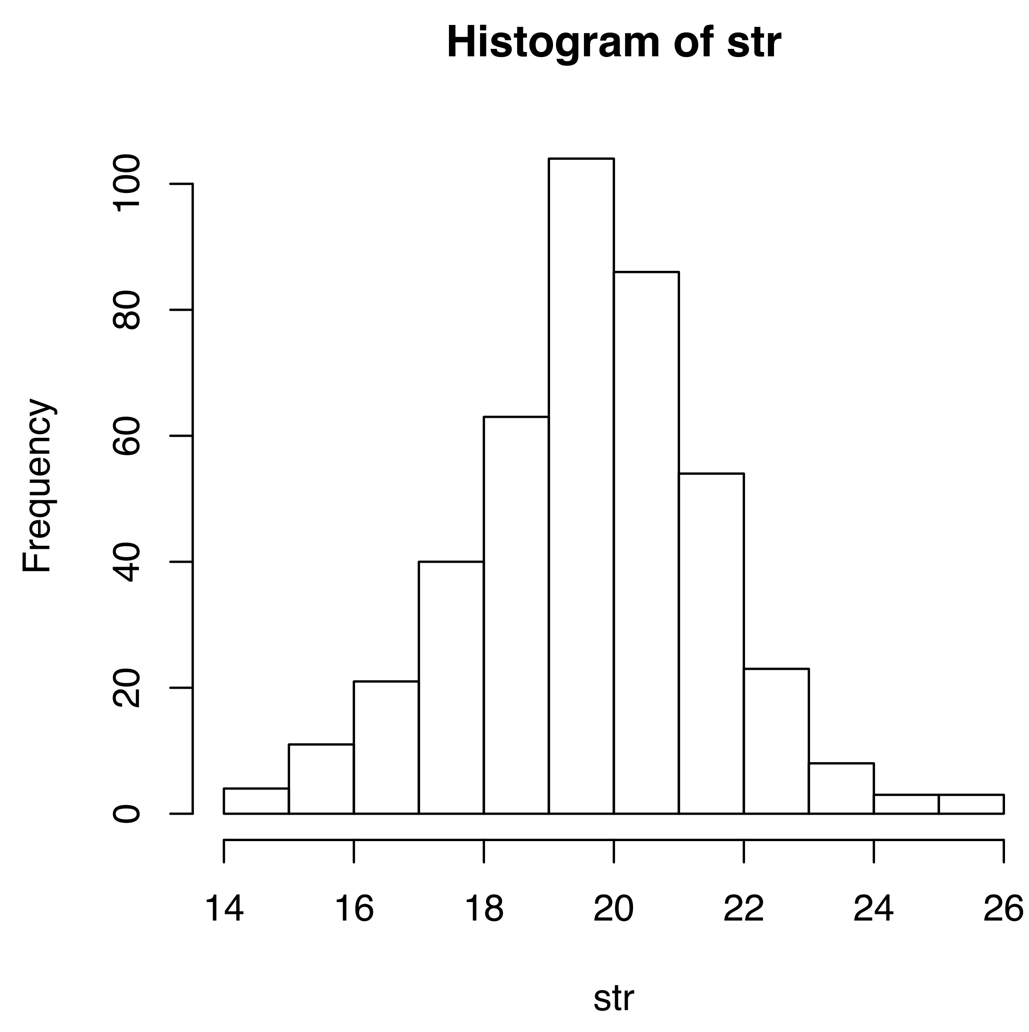
A test score data in California

- 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`

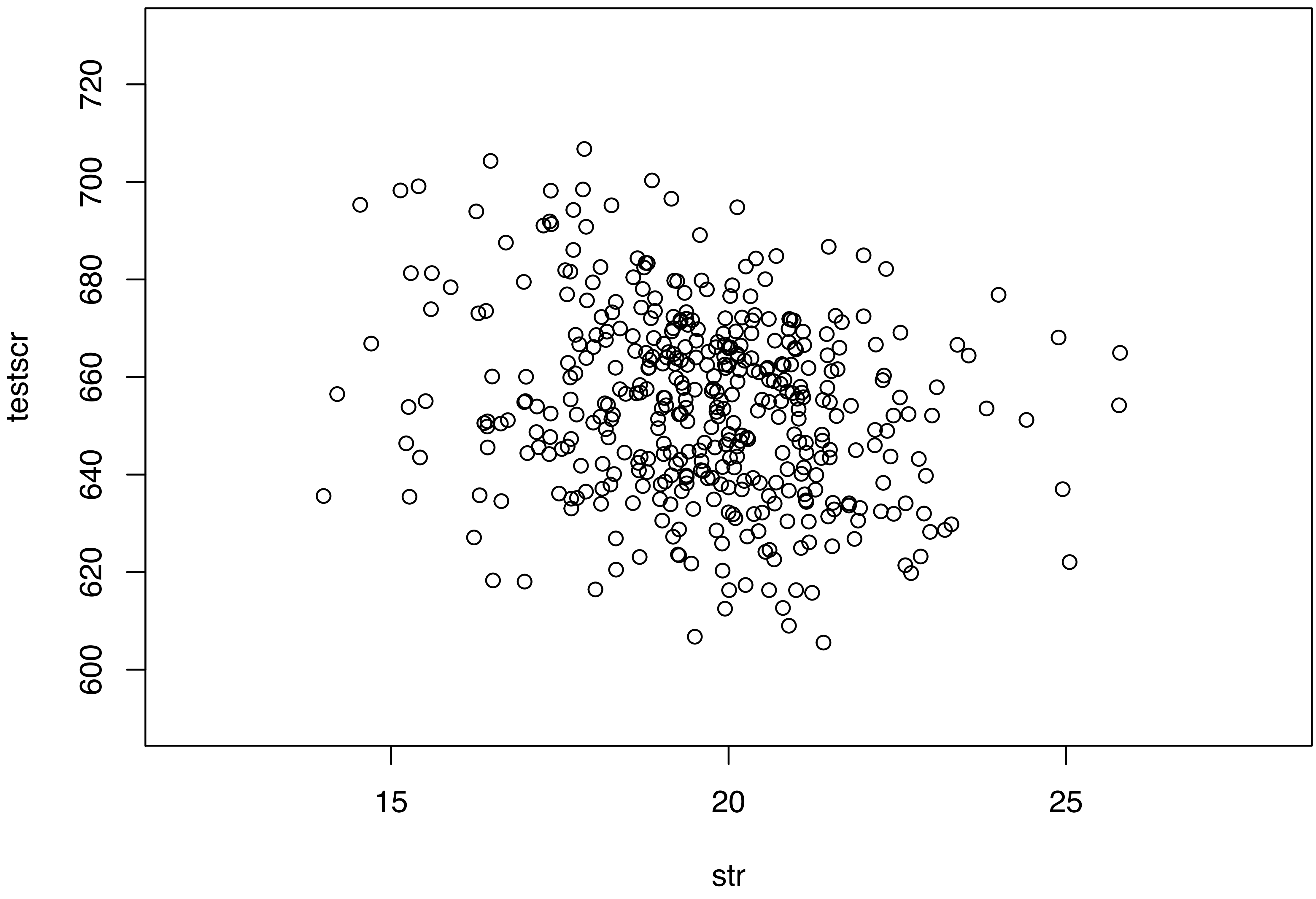
Average test score v.s. student-teacher ratio

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

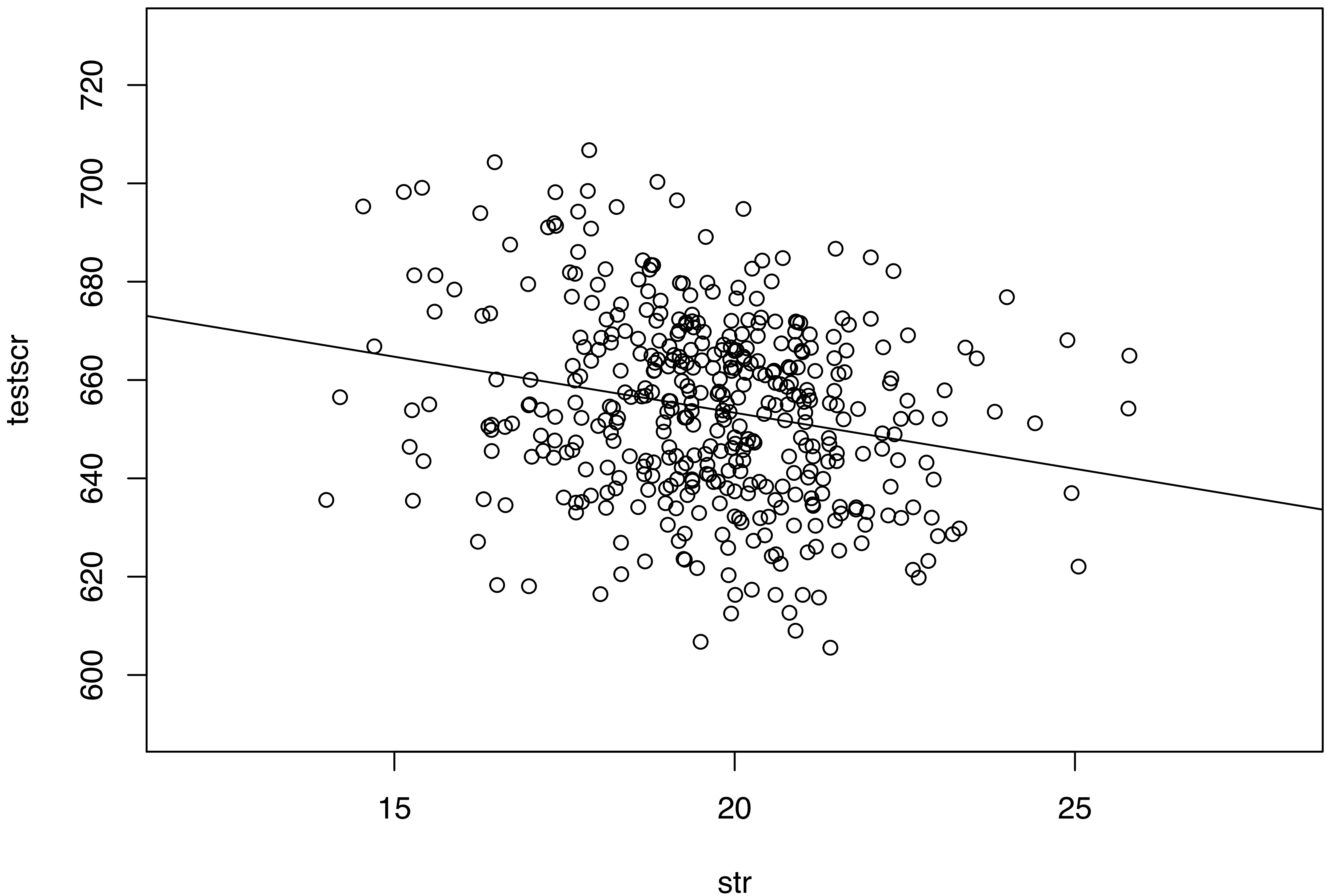




Average test score v.s. student-teacher ratio



Average test score v.s. student-teacher ratio



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

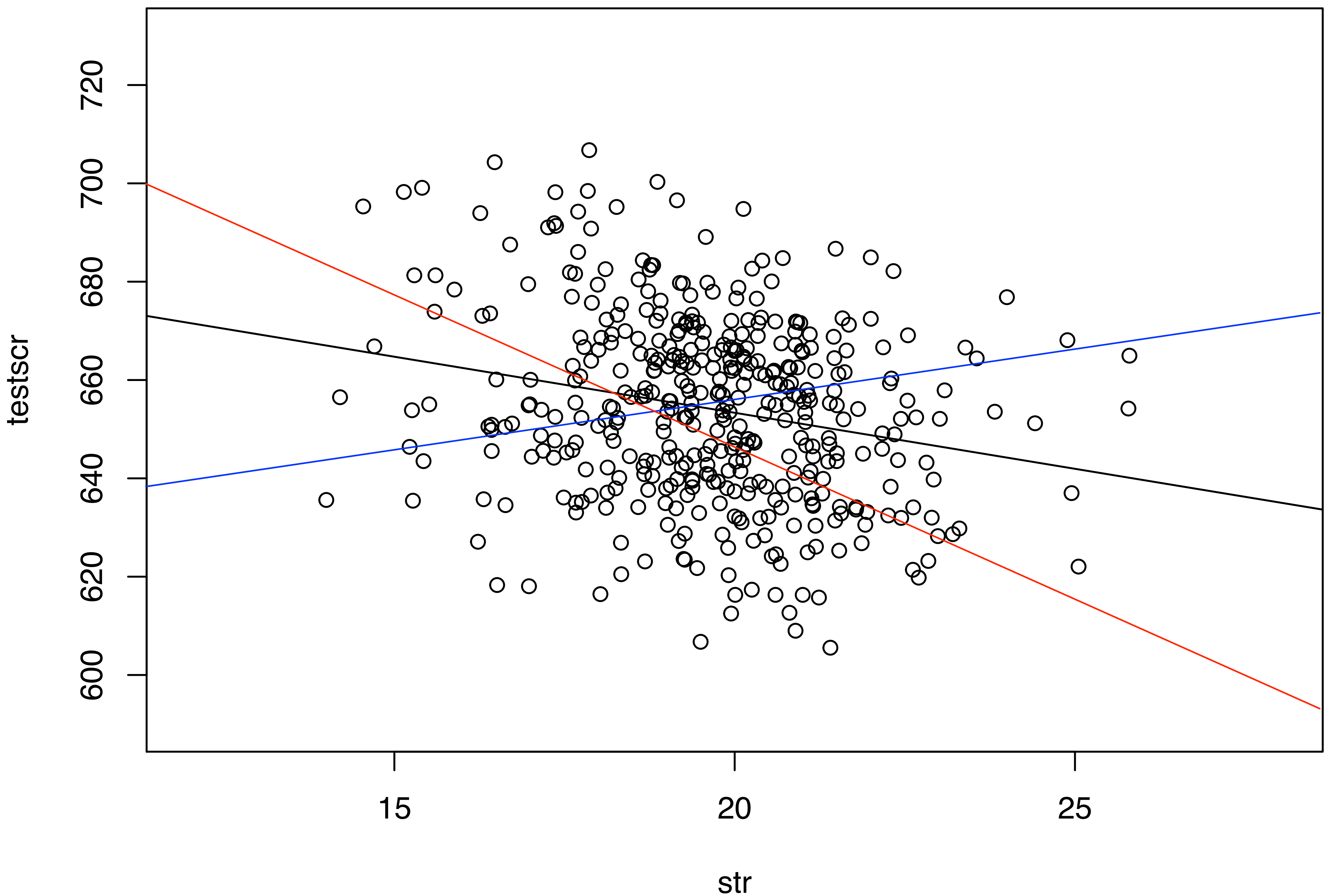
$$\text{TestScore} = \beta_0 + \beta_1 \times \text{ClassSize} + \text{other factors}$$

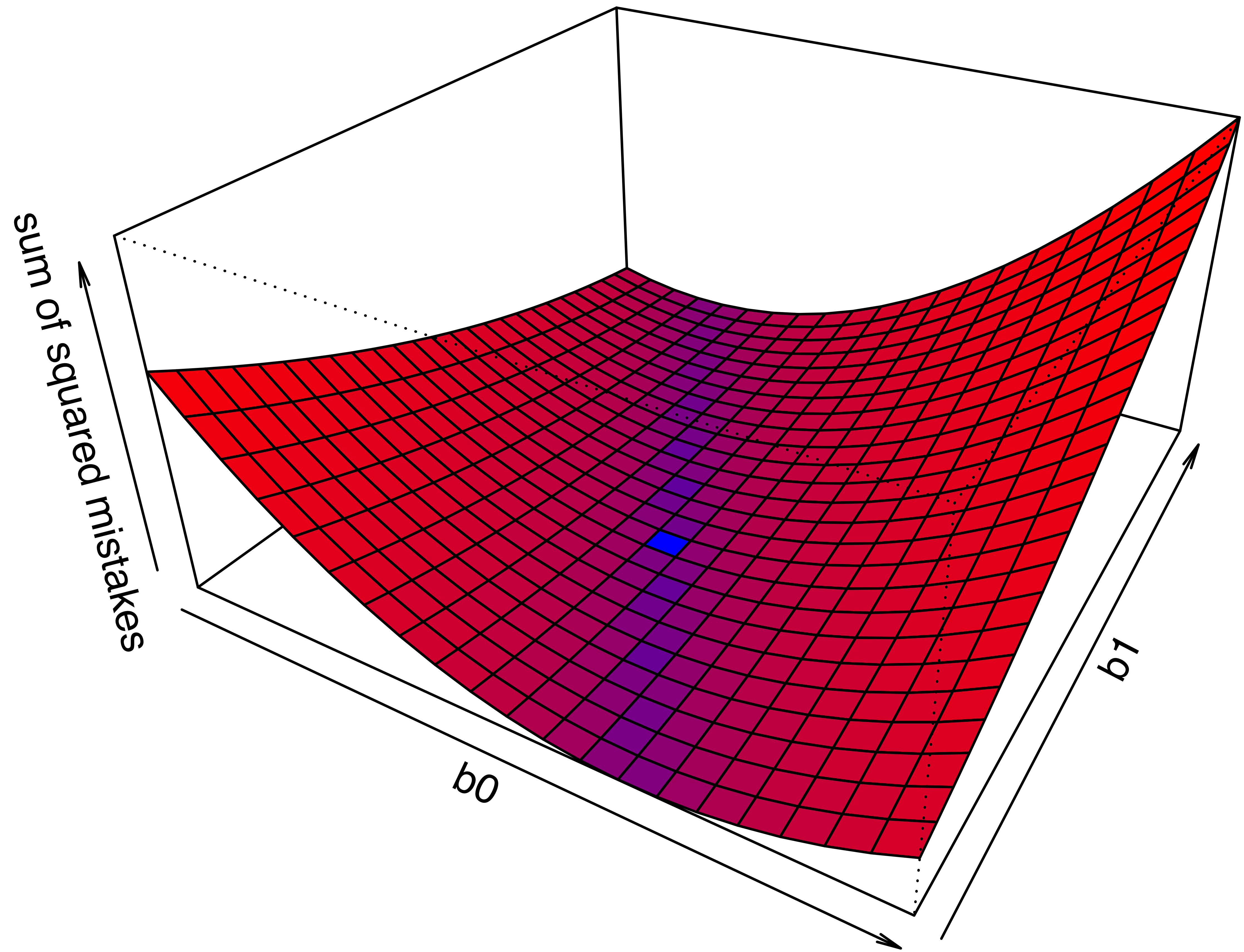
Estimating the coefficients

- $\overline{Y}$ 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).
- 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 estimators

- The OLS estimators of the slope and the intercept are

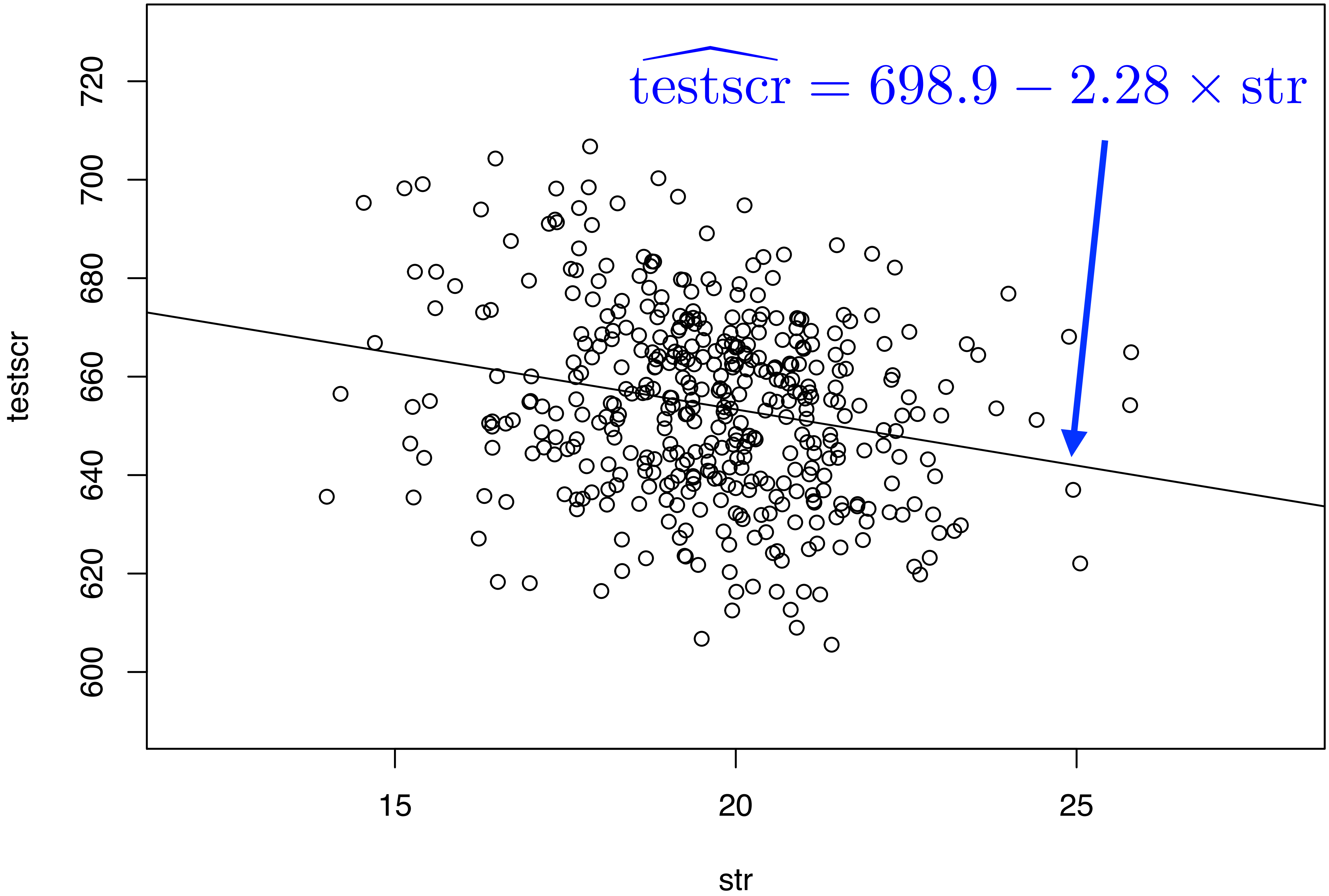
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{s_{XY}}{s_X^2}$$

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

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

- The residuals: $\hat{u}_i = Y_i - \hat{Y}_i$

Average test score v.s. student-teacher ratio



A measure of fit

- The R^2 — 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)

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.

- 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:

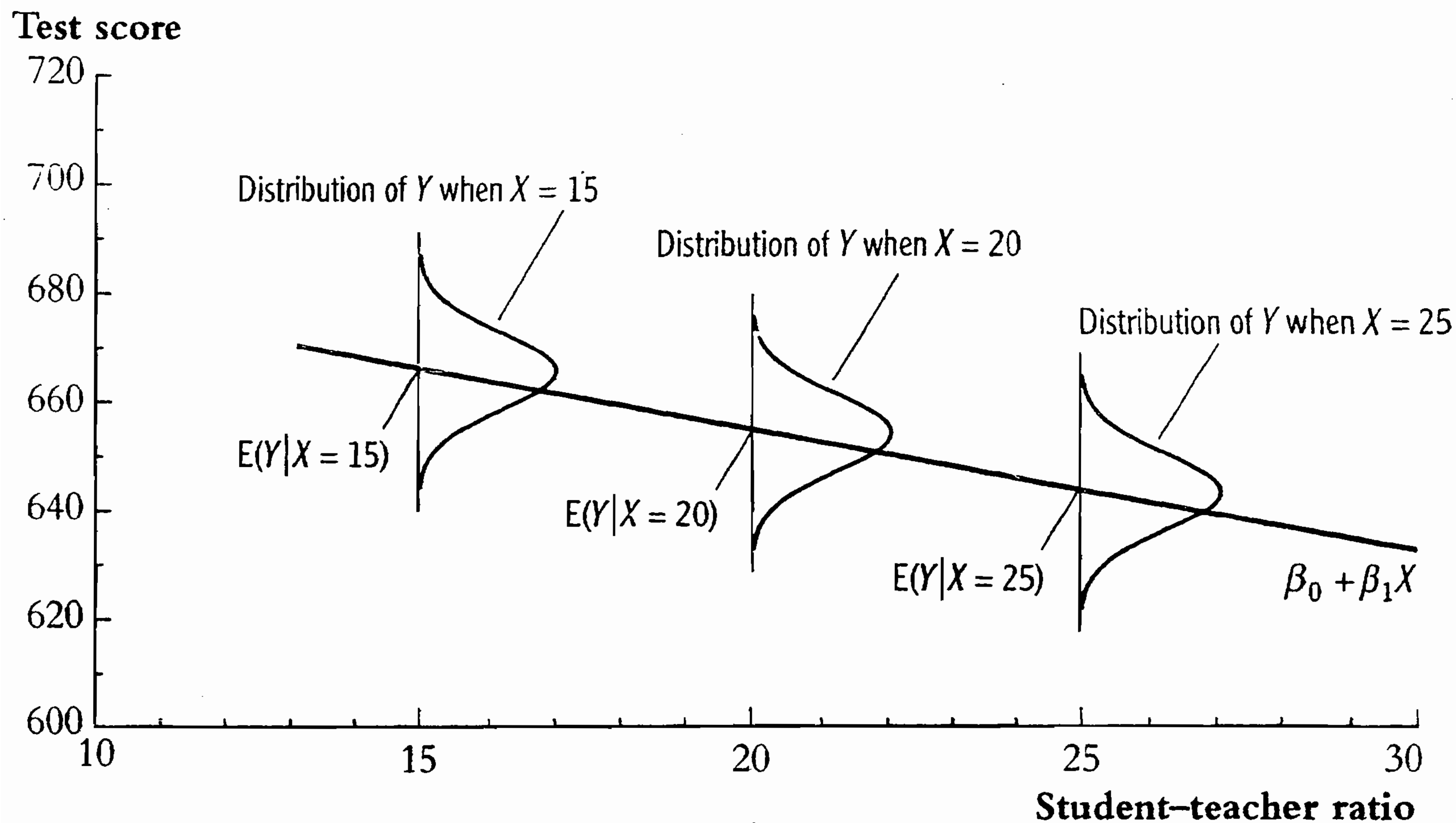
1. The error term u_i has conditional mean zero given X_i:

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

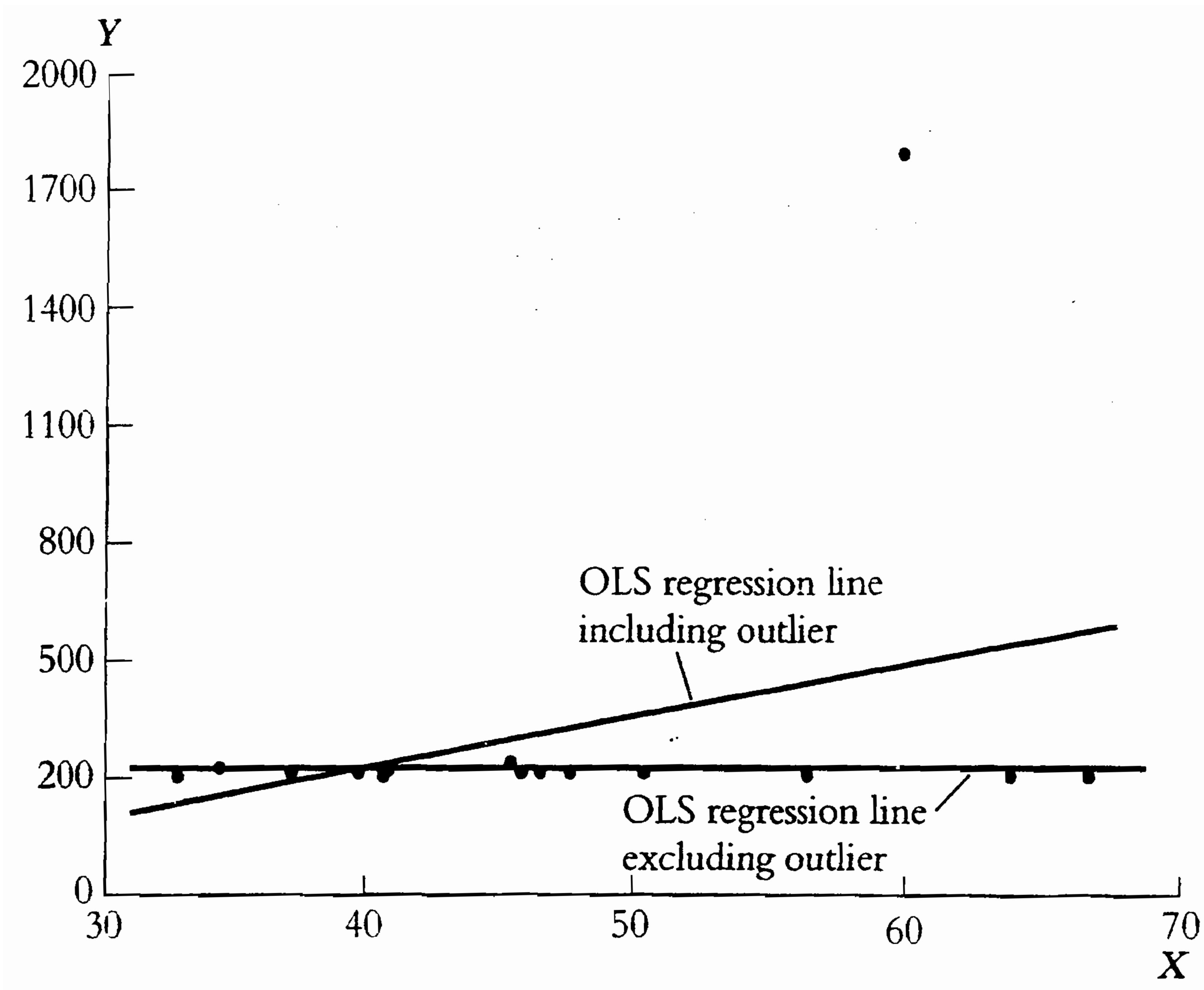
2. (X_i, Y_i), $i = 1, \ldots, n$, are i.i.d. draws from their joint distribution; and

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.