

计量经济学

第五讲：一元线性回归（一）

黄嘉平

工学博士 经济学博士
深圳大学中国经济特区研究中心 讲师

办公室	粤海校区汇文楼2613
E-mail	huangjp@szu.edu.cn
Website	https://huangjp.com

主要内容

- 线性回归模型
 - $Y_i = \beta_0 + \beta_1 X_i + u_i$
- 系数的估计
 - 最小二乘估计量 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 的表达式
- 拟合优度
 - R^2 与回归标准误
- 最小二乘假设
- OLS 估计量的抽样分布
 - 无偏性与一致性、渐进正态分布

回归分析的目的

- 因果推断 (causal inference)

通过数据估计某个变量 (X) 的变化给另一个变量 (Y) 带来的影响。回归分析是否能揭示因果关系取决于 X 的取值是否是随机分配 (或近似于随机分配) 。

- 预测 (prediction)

通过某个变量的观测值预测另一个变量的值。

线性回归模型

变量间的关系

- 某地区教育部门打算增加该地区小学教师数量。这样做会导致：
 - 教育的成本增加
 - 班级规模变小，从而促进教育质量的提升
- 为了准确衡量利弊，就有必要了解缩小班级规模对教育质量产生的影响。班级规模一般用人数衡量，而教育质量往往可以用考试成绩代表。
- 对以上问题更加具体的描述是：若班级规模平均减少了2个学生，则对该地区的标准化测试成绩（统一考试）有什么影响？

班级规模与考试成绩

- 班级规模 (ClassSize) 的变化对考试成绩 (TestScore) 产生的影响为

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

也可写成 $\Delta TestScore = \beta_{ClassSize} \times \Delta ClassSize$

- 若假设两者间的关系为线性，则 $\beta_{ClassSize}$ 为斜率，而班级规模和考试成绩间的关系可描述为

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

引入其他影响因素

- 很明显，考试成绩不是完全由班级规模决定的。其他影响考试成绩的因素包括：教师素质、教材质量、学生素质、学生家庭状况、考试时的临场发挥（等随机因素）。
- 我们把其他影响因素综合在一起，称为 other factors，加入关系式，便可得到

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

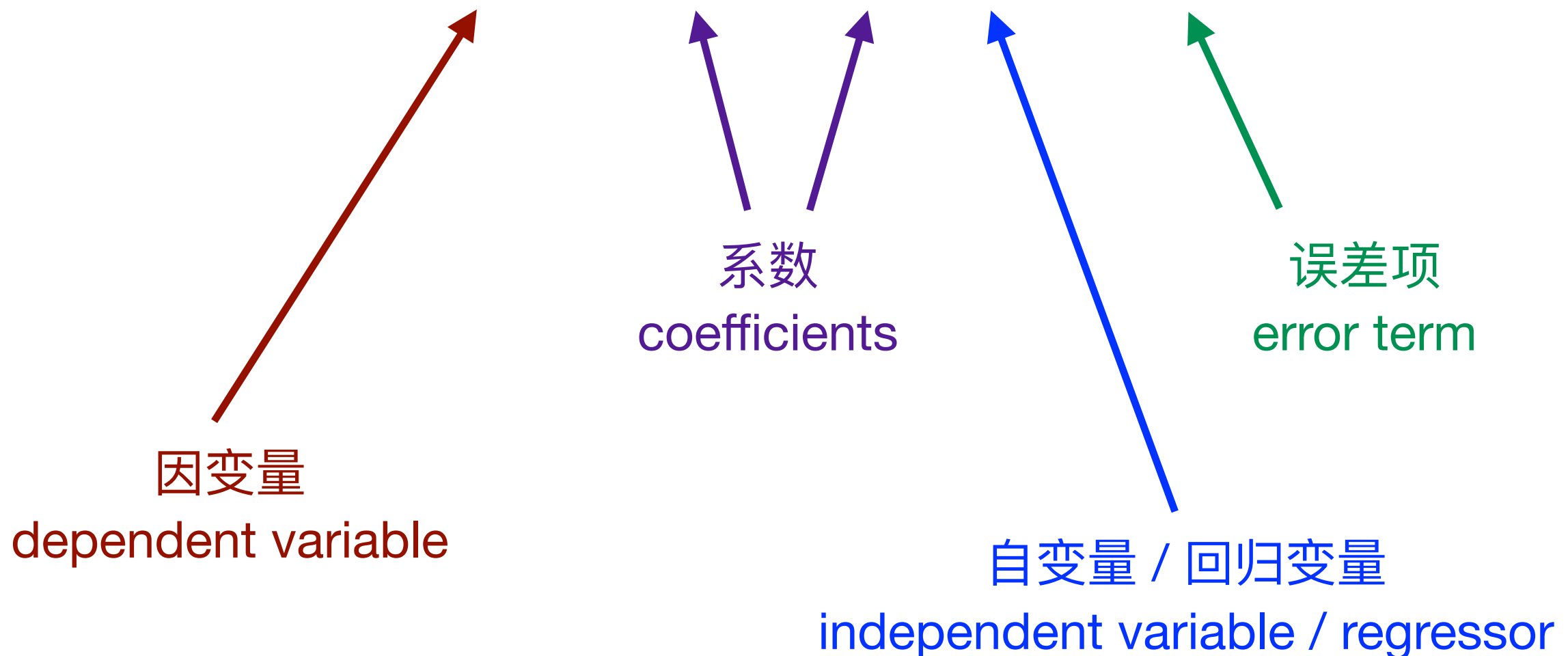
- 这种建模思路具有普遍性。我们可以用 X 替代班级规模，用 Y 替代考试成绩，从而得到线性模型的一般形式。

线性回归模型

Linear regression model

- 一元线性回归模型

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



线性回归模型

Linear regression model

- 一元线性回归模型

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



总体回归线 / 总体回归函数

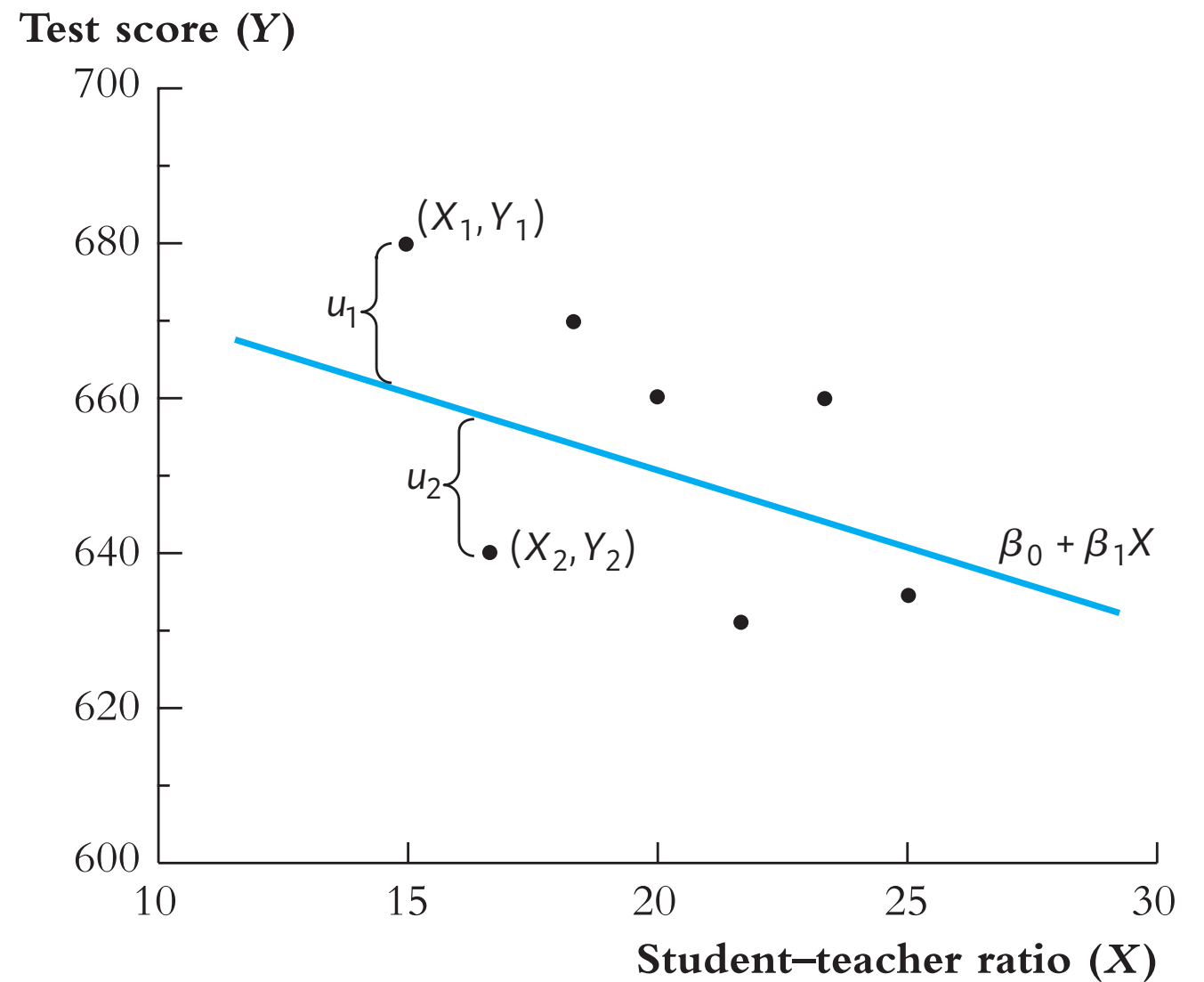
population regression line / population regression function

- 总体回归函数是总体中 X 和 Y 之间在平均意义上成立的关系。当给出 X 的值时，我们可以用它预测 Y 的取值。也可以写成 $E(Y | X) = \beta_0 + \beta_1 X$ 。

线性回归模型图示

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.



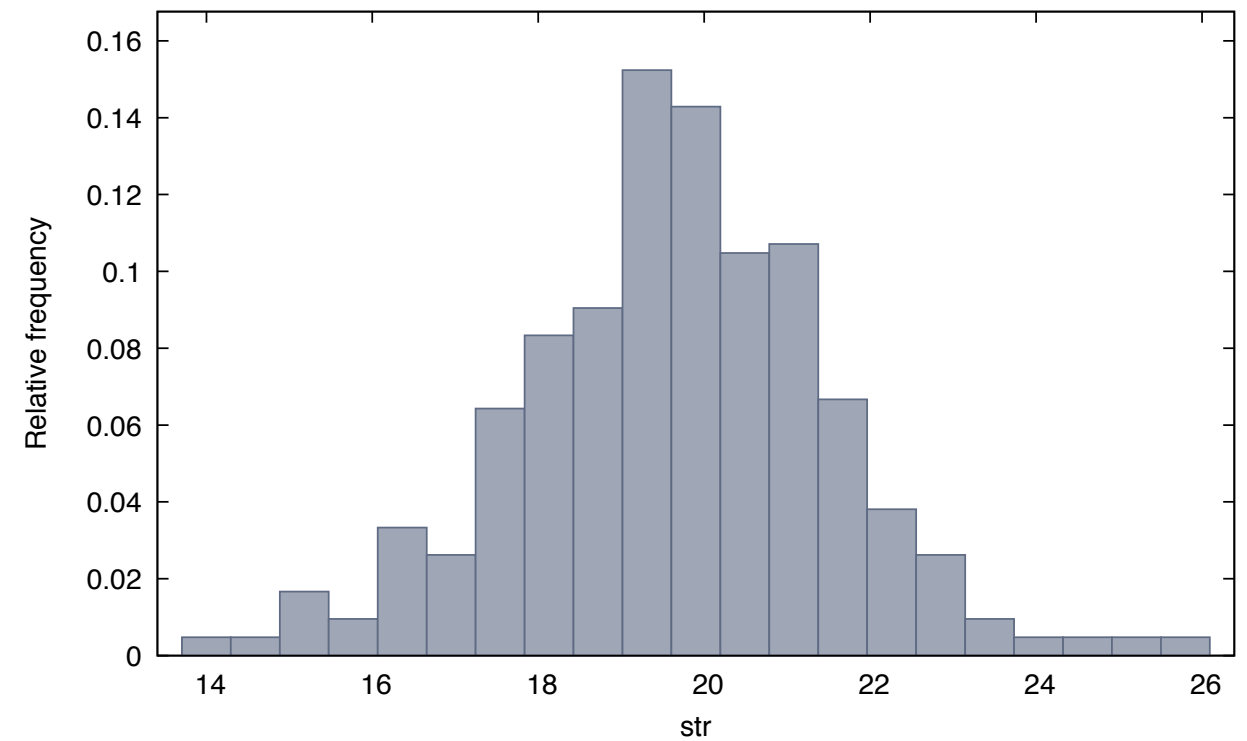
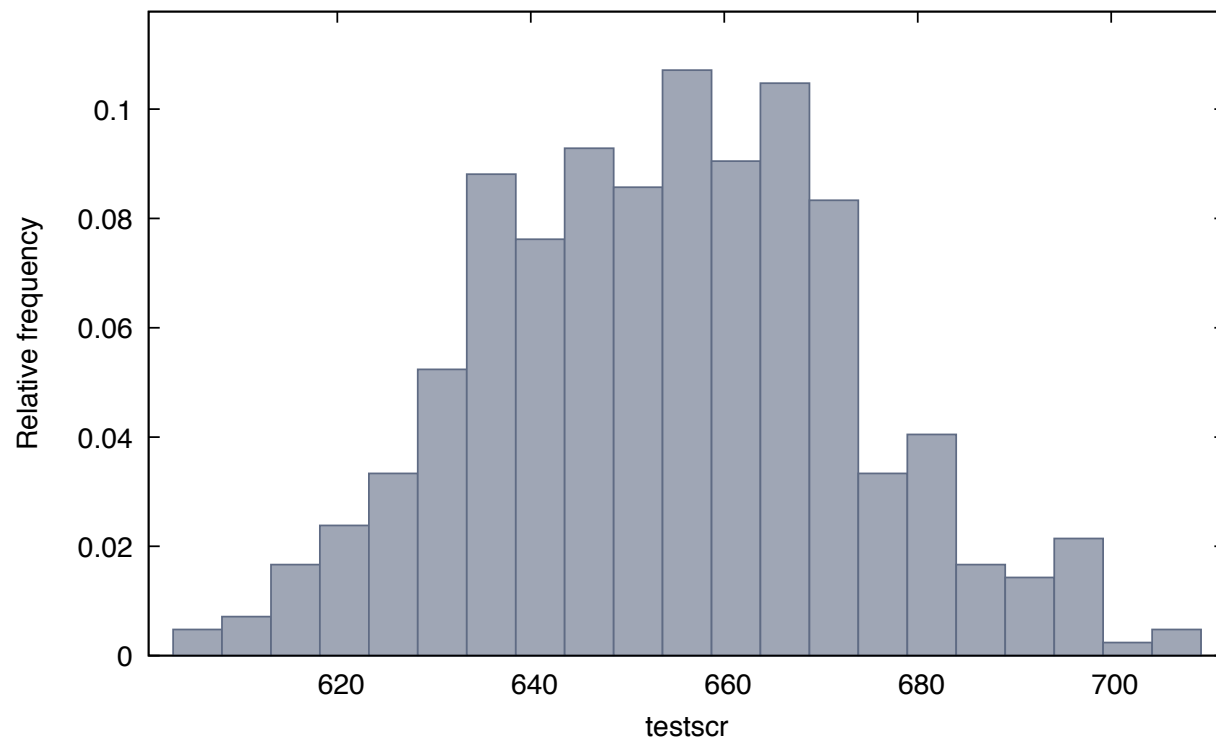
STAR 数据集

- STAR (California Standardized Testing and Reporting) 数据集是惯于美国加州小学教育的数据集，包含考试成绩、学校特征、学生背景等变量。本书中用到的数据涵盖1998-1999年间420个学区。
- 数据文件 `caschool.xlsx`
- 说明文件 `californiatestscores.docx`

考试成绩和学生教师比

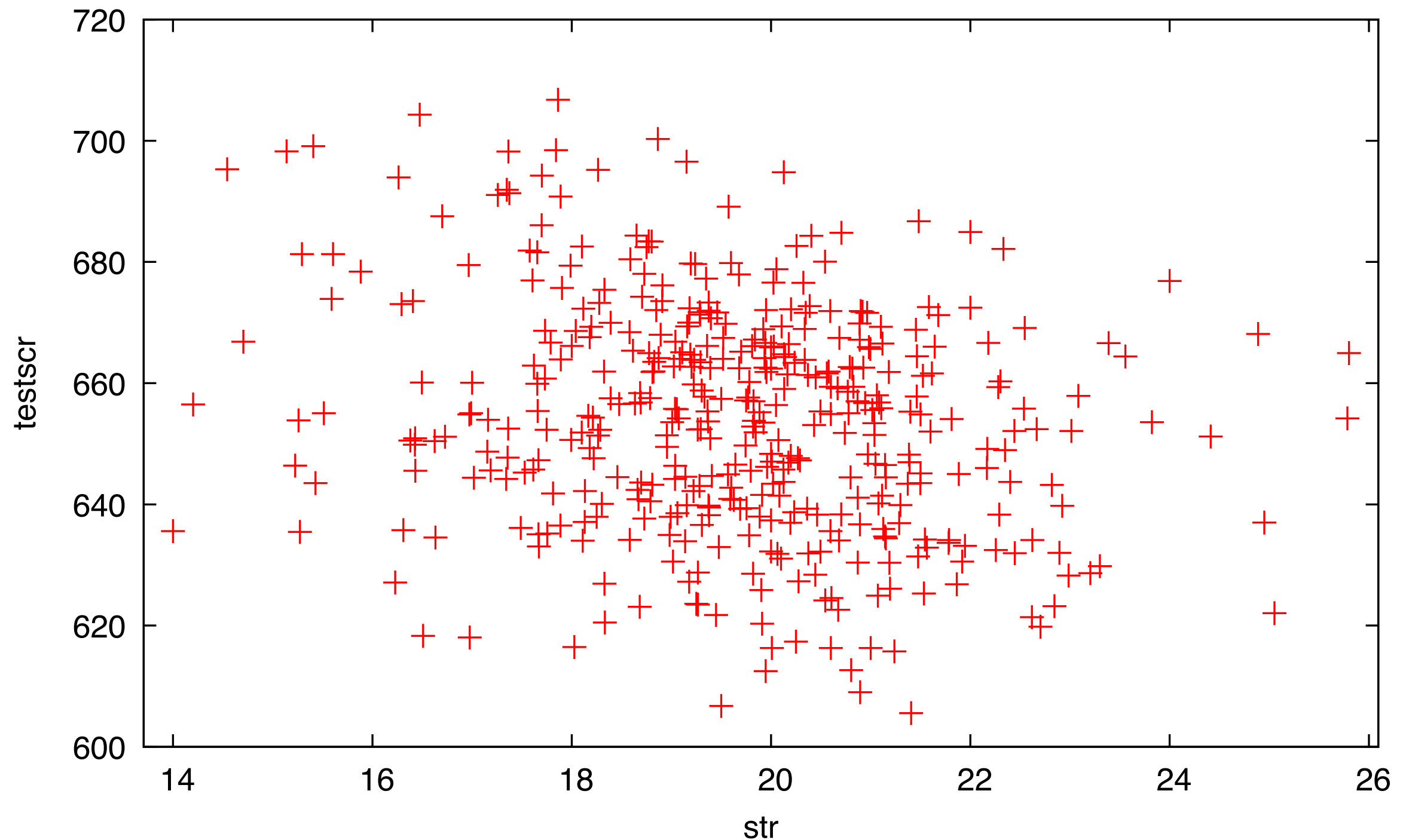
Average test score and student-teacher ratio

- “testscr”: 数学和语言考试的平均成绩
- “str”: 学生教师比 (学生人数 / 教师人数)



考试成绩和学生教师比

Average test score and student-teacher ratio



系数的估计

回归系数的估计

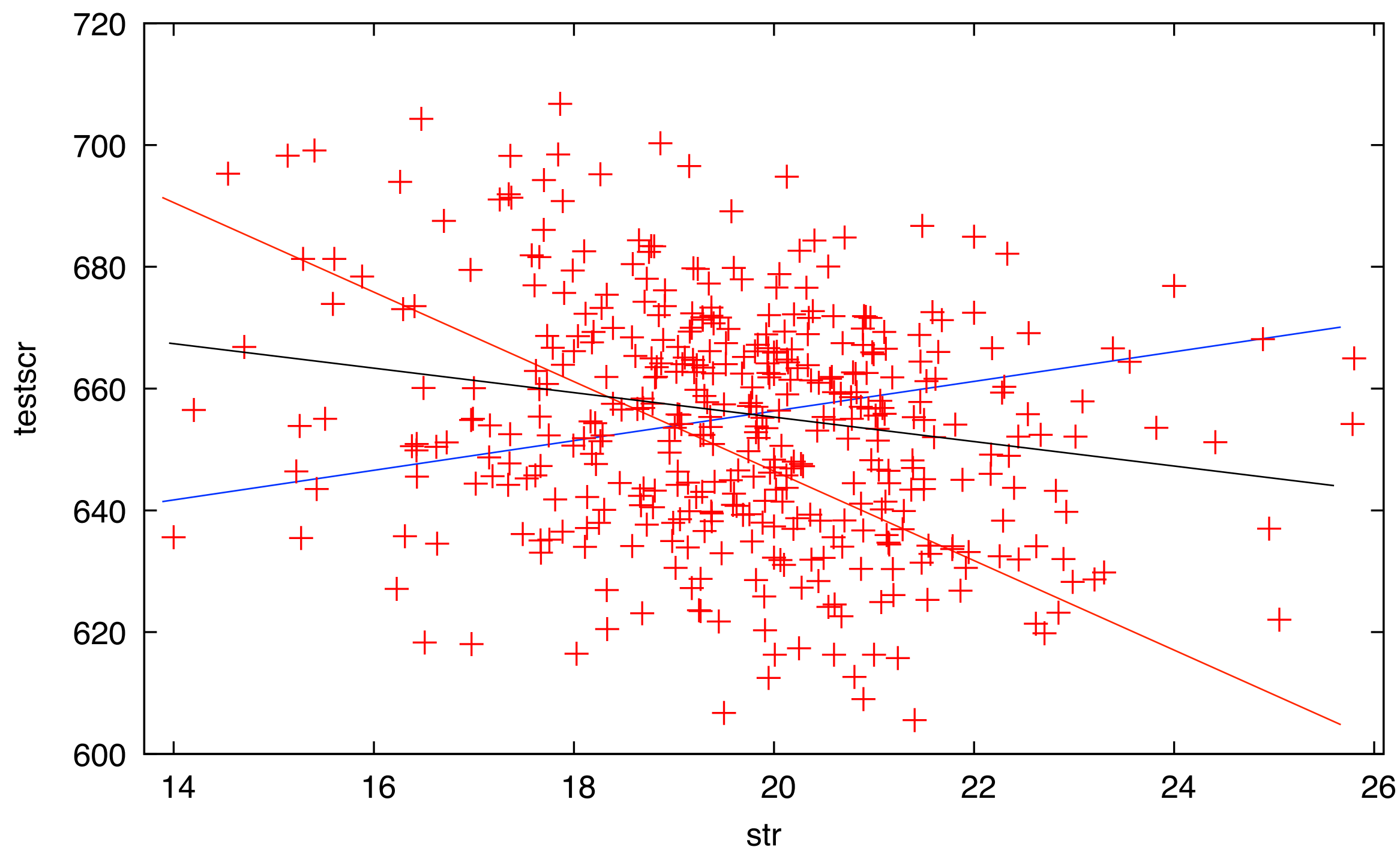
Estimating the regression coefficients

- \bar{Y} 是总体均值的估计量。
- 同样，在一元线性回归模型中， β_0 和 β_1 是我们假设的总体回归函数中的未知量，因此也需要利用数据对其进行估计。
- 令 b_0 和 b_1 分别表示 β_0 和 β_1 的某个估计量，则使观测误差的平方和

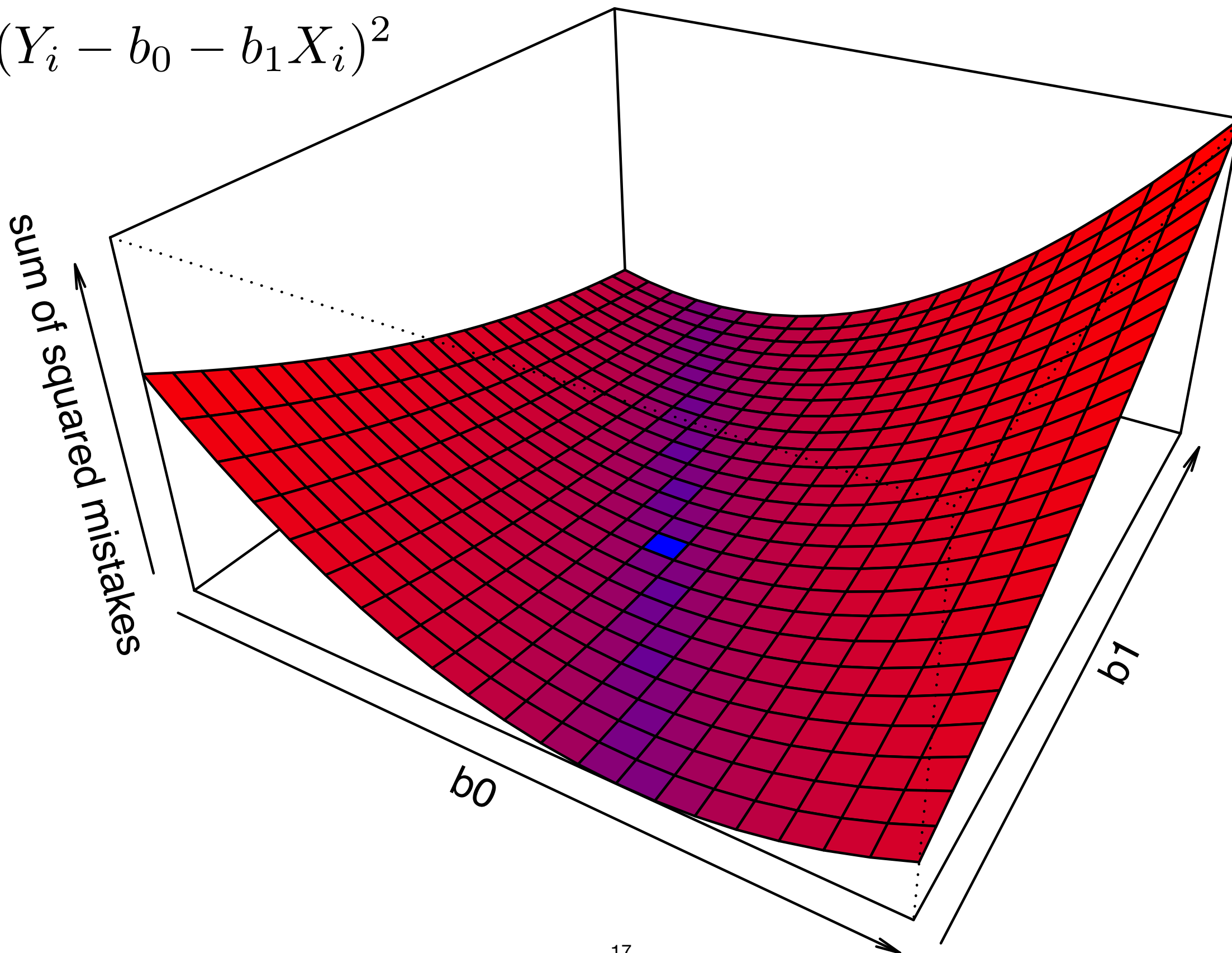
$$\sum_{i=1}^n (Y_i - b_0 - b_1 X_i)^2$$

取值最小的估计量被称为普通最小二乘（ordinary least squares, OLS）估计量，分别记为 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 。

散点图与回归曲线



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



最小二乘估计量的推导

对 $\sum_{i=1}^n (Y_i - b_0 - b_1 X_i)^2$ 求一阶偏导，可得一阶条件

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

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

将两式除以 n 并令其等于零，整理后可得一阶条件

$$\bar{Y} - b_0 - b_1 \bar{X} = 0$$

$$\frac{1}{n} \sum_{i=0}^n Y_i X_i - b_0 \bar{X} - b_1 \frac{1}{n} \sum_{i=1}^n X_i^2 = 0$$

最小二乘估计量 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 满足此条件。解方程组后可得

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{s_{XY}}{s_X^2}, \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

OLS 估计量、预测值和残差

OLS estimator, predicted values, and residuals

- 斜率 β_1 和截距 β_0 的 OLS 估计量分别为

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

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

- OLS 预测值 (predicted value) : $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$

- 残差: $\hat{u}_i = Y_i - \hat{Y}_i$

样本回归线 / 样本回归函数
sample regression line/
sample regression function

$\hat{\beta}_1$ 与相关系数

- 初学者经常会把回归分析和相关分析相混淆。
- 在线性回归模型中，

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

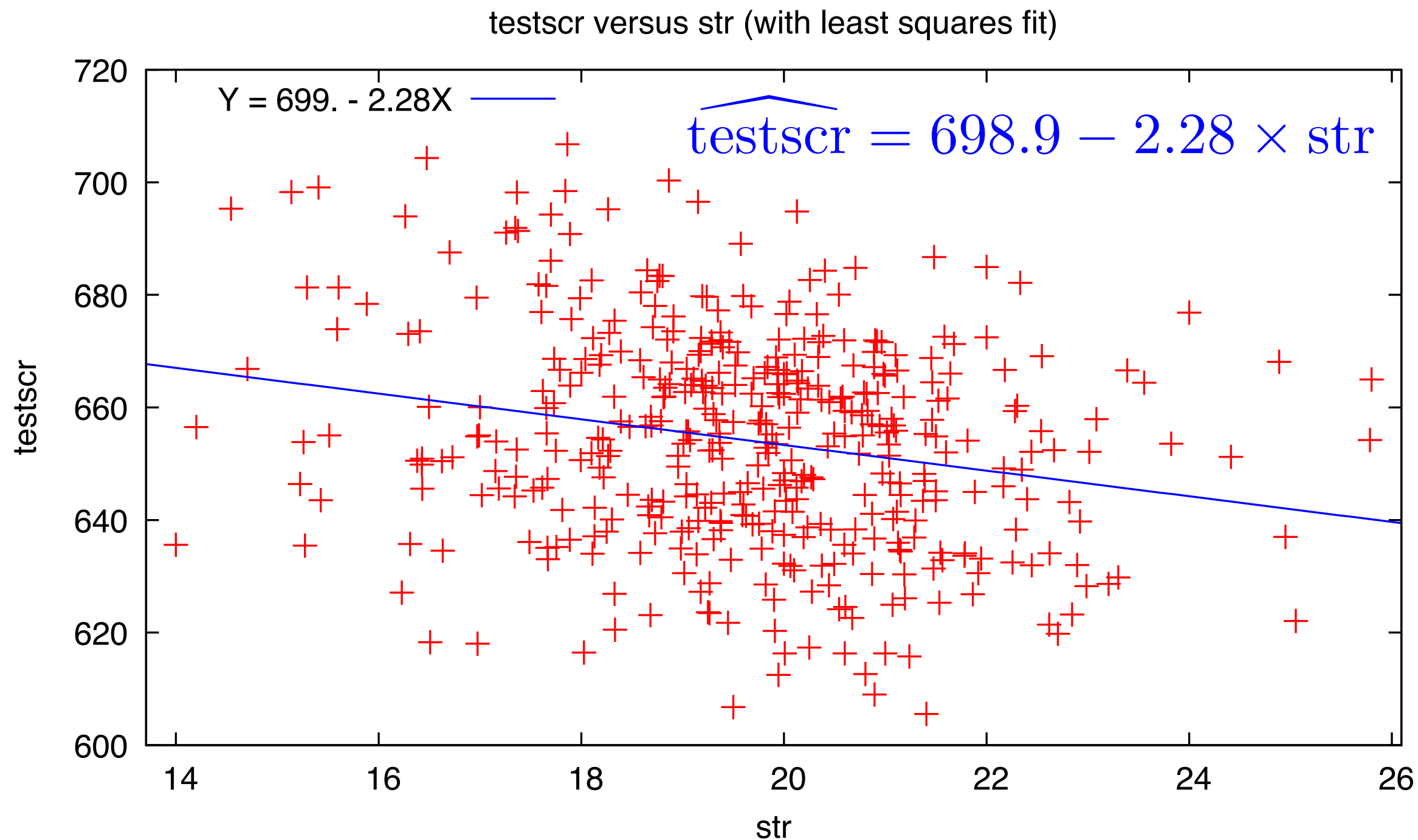
- 在统计学中，样本相关系数为

$$\text{corr}(X, Y) = \frac{s_{XY}}{s_X s_Y}$$

- 如果用 Y_i 回归 X_i ，则斜率的 OLS 估计量为 $\frac{s_{XY}}{s_Y^2}$ 。

考试成绩和学生教师比

Average test score and student-teacher ratio



如何理解回归结果

- 针对 STAR 数据的 OLS 回归结果为

$$\widehat{\text{testscr}} = 698.9 - 2.28 \times \text{str}$$

- 这里我们关注斜率的估计值 -2.28。该值表示当每个教师对应的学生人数增加一个时，学区考试成绩将平均下降 2.28 分。
- 截距的估计值 698.9 本身没有实际意义。但加入我们将学生教师比的均值带入，就能得到平均班级人数下的平均考试成绩，而在该计算中 698.9 的作用是不可或缺的。
- 斜率的估计值是大还是小呢？若将班级人数减小 2 人，则平均成绩上升 4.56 分。2 人在班级人数分布中的变化很明显；而 4.56 分在成绩分布中的变化并不明显，但也不是可以忽略的。

为什么使用 OLS 估计量

- OLS 估计是实证研究中最常用的方法。
- OLS 估计量具有理想的理论性质。在适当的假设条件下，OLS 估计量是无偏的且具有一致性。
- 在另外的附加条件下，可以证明 OLS 估计量是有效的 (BLUE) 。

⇒ 详见 5.5 节 Gauss-Markov 定理

拟合优度

R^2

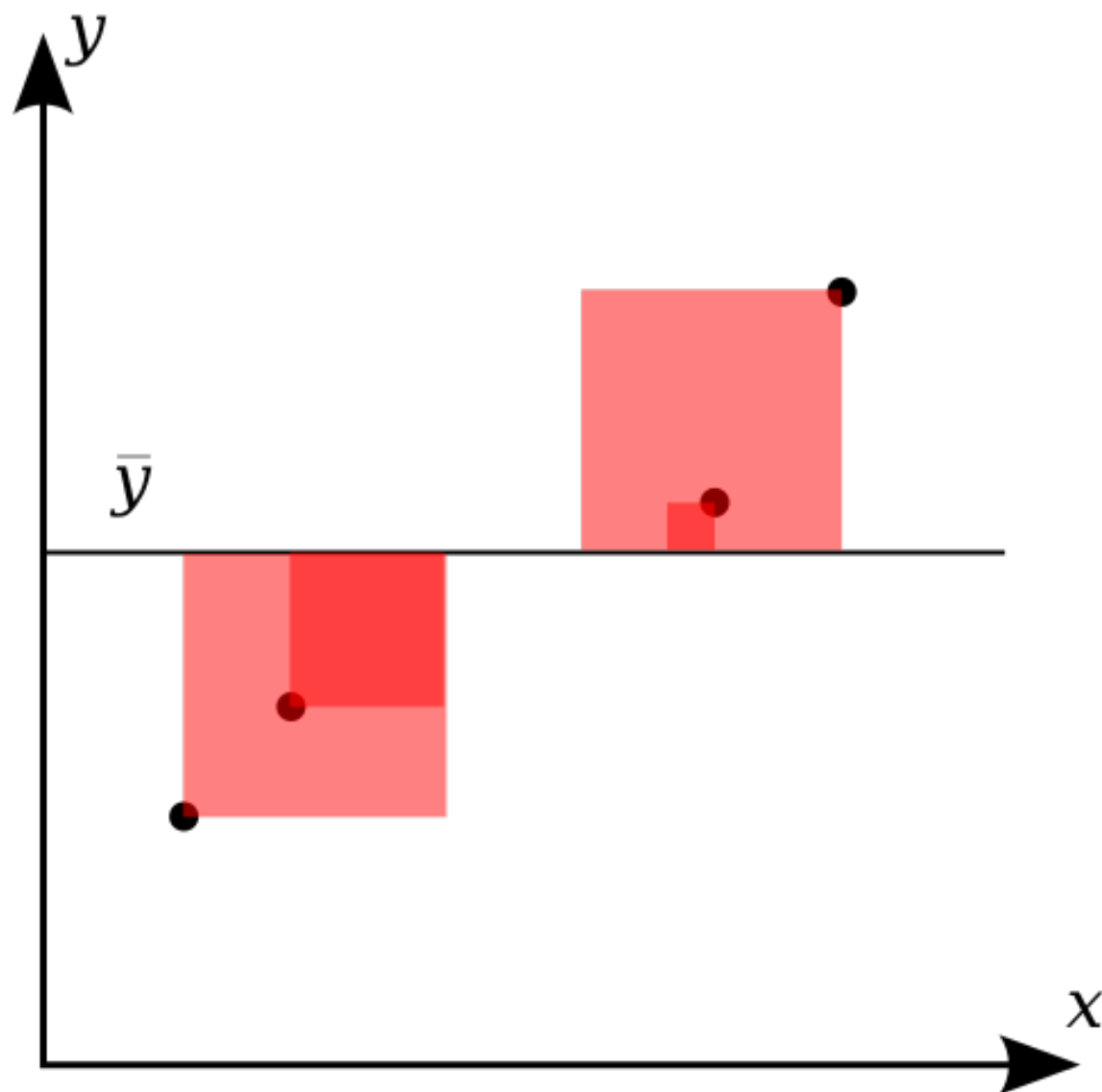
- 拟合优度是指回归线与数据的拟合 (fit) 效果。即回归变量解释了多少因变量的变化，或观测值是否紧密聚集在回归线周围。
- R^2 ，又称作**决定系数 (correlation of determination)**，是指可由 X_i 解释 (或预测) 的 Y_i 样本方差的比例。
- 从 $Y_i = \hat{Y}_i + \hat{u}_i$ 可定义 R^2 为 \hat{Y}_i 的样本方差和 Y_i 的样本方差之比，即

$$\begin{aligned} R^2 &= \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = \frac{ESS}{TSS} && \begin{array}{l} \text{(explained sum of squares, 被解释平方和)} \\ \text{(total sum of squares, 总平方和)} \end{array} \\ &= 1 - \frac{\sum_{i=1}^n \hat{u}_i^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = 1 - \frac{SSR}{TSS} && \text{(sum of squared residuals, 残差平方和)} \end{aligned}$$

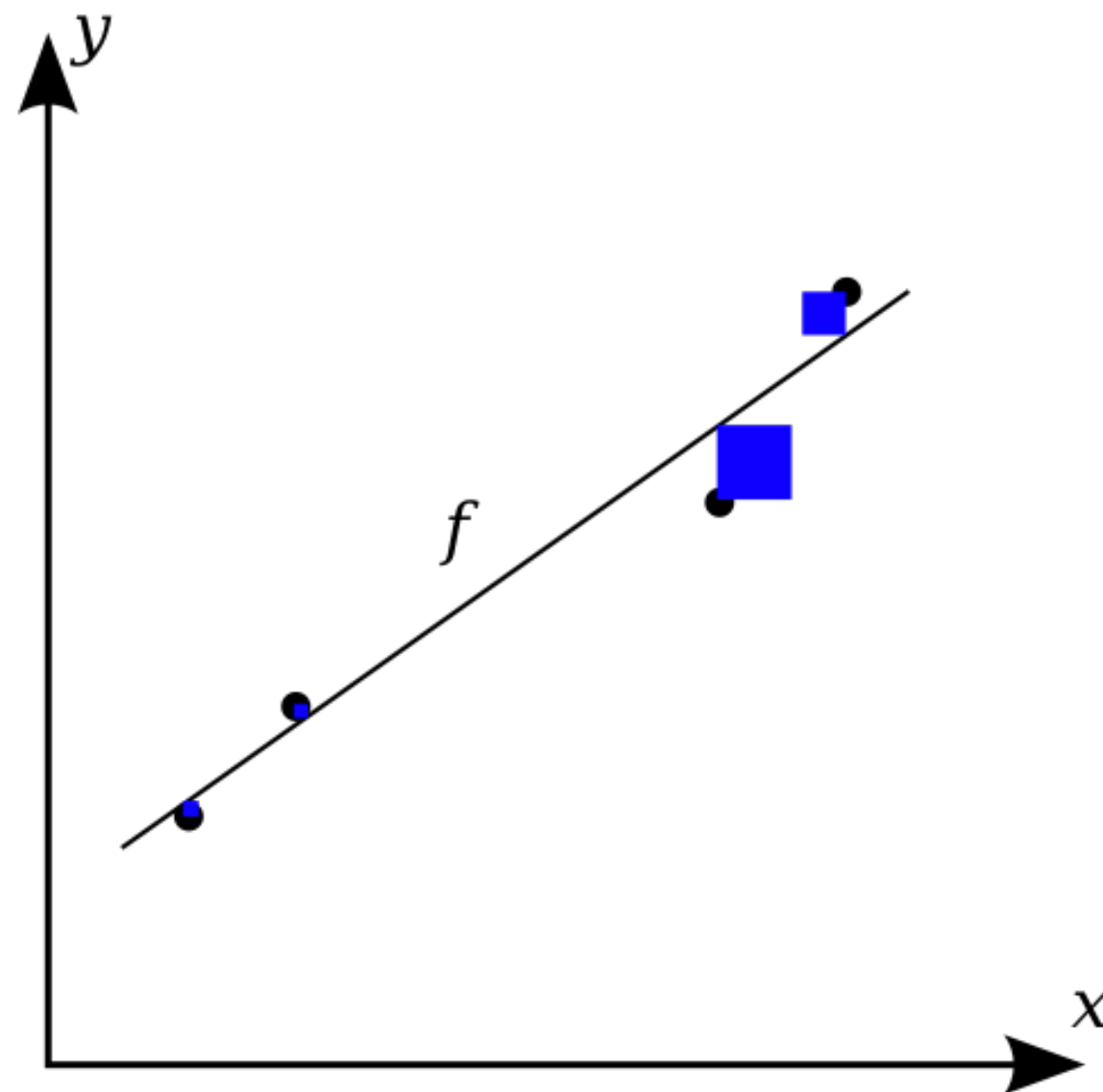
详细推到过程参见附录4.3

残差平方和的图示

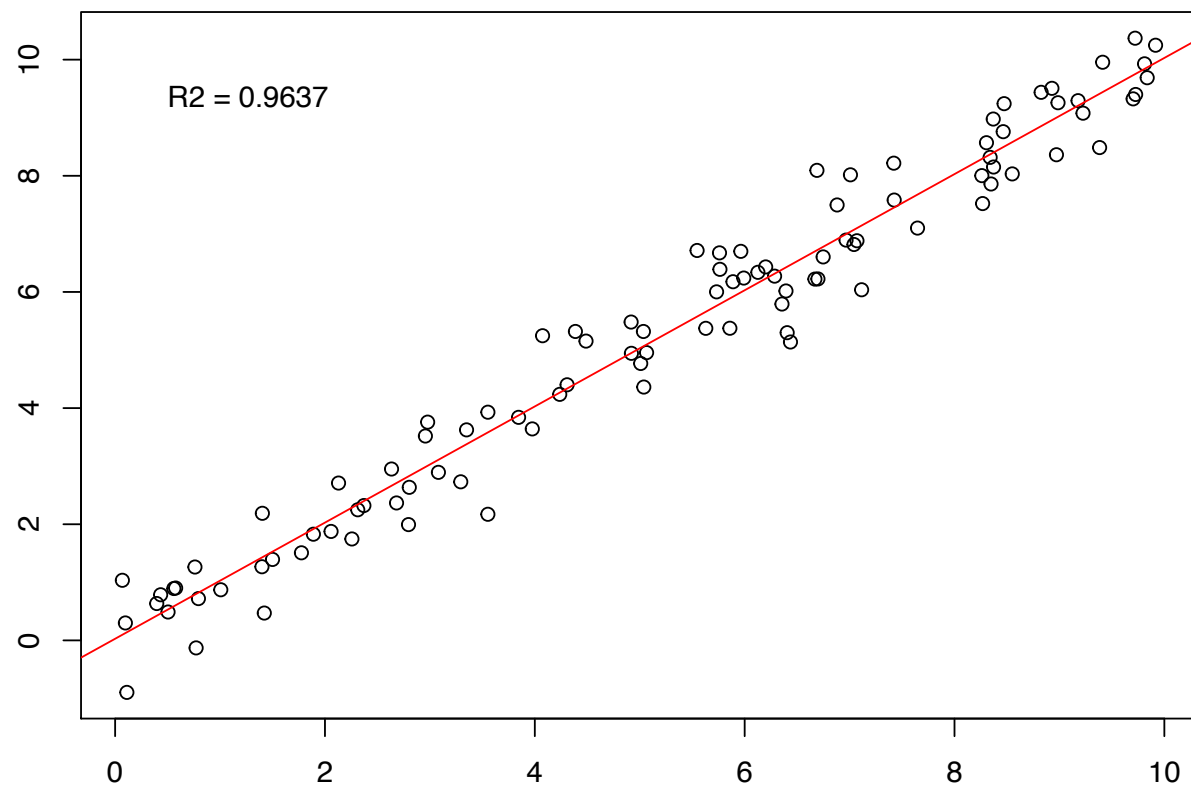
Y_i 的平均值



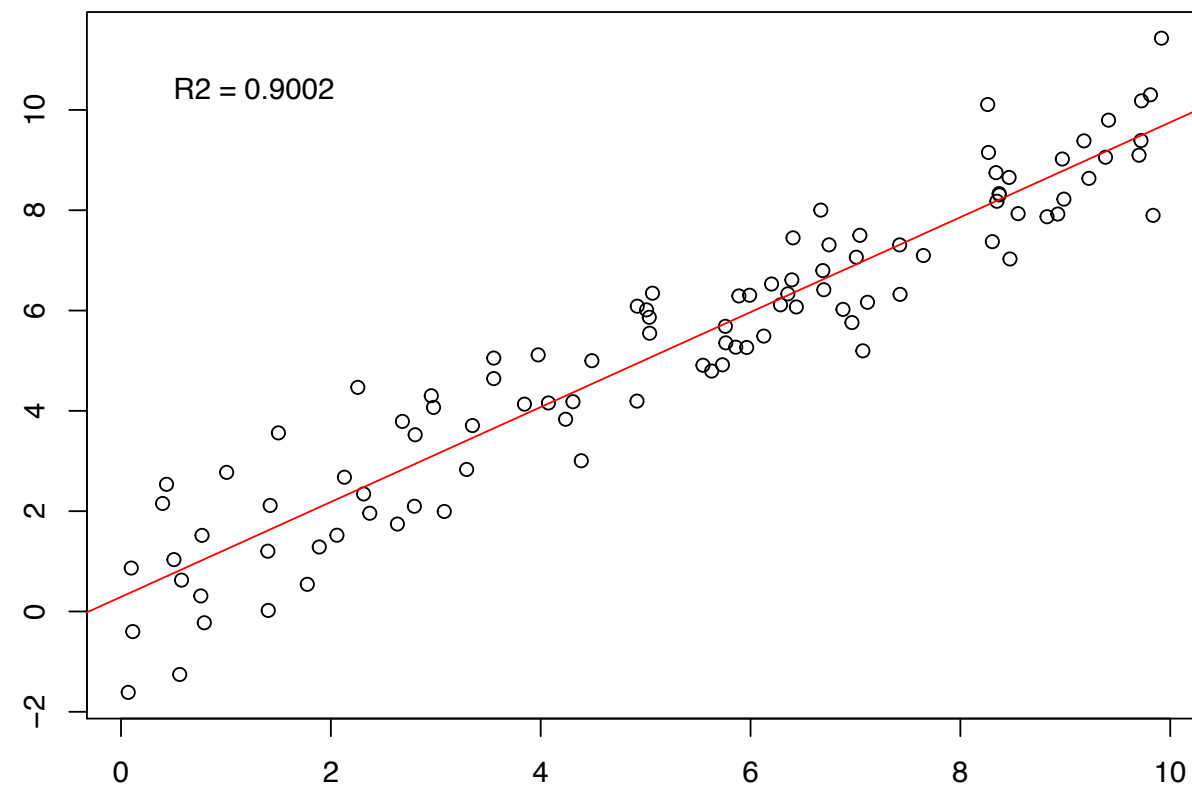
OLS 回归



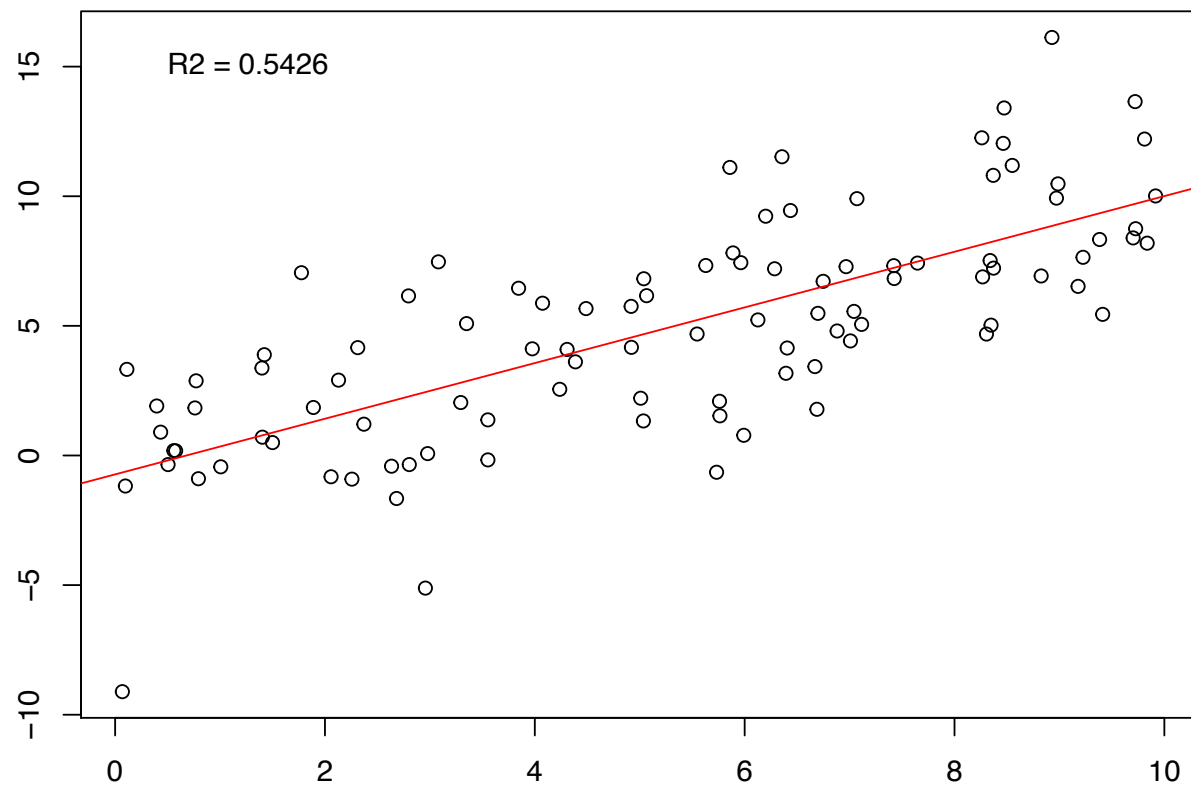
$Y \sim X + N(0, 0.25)$



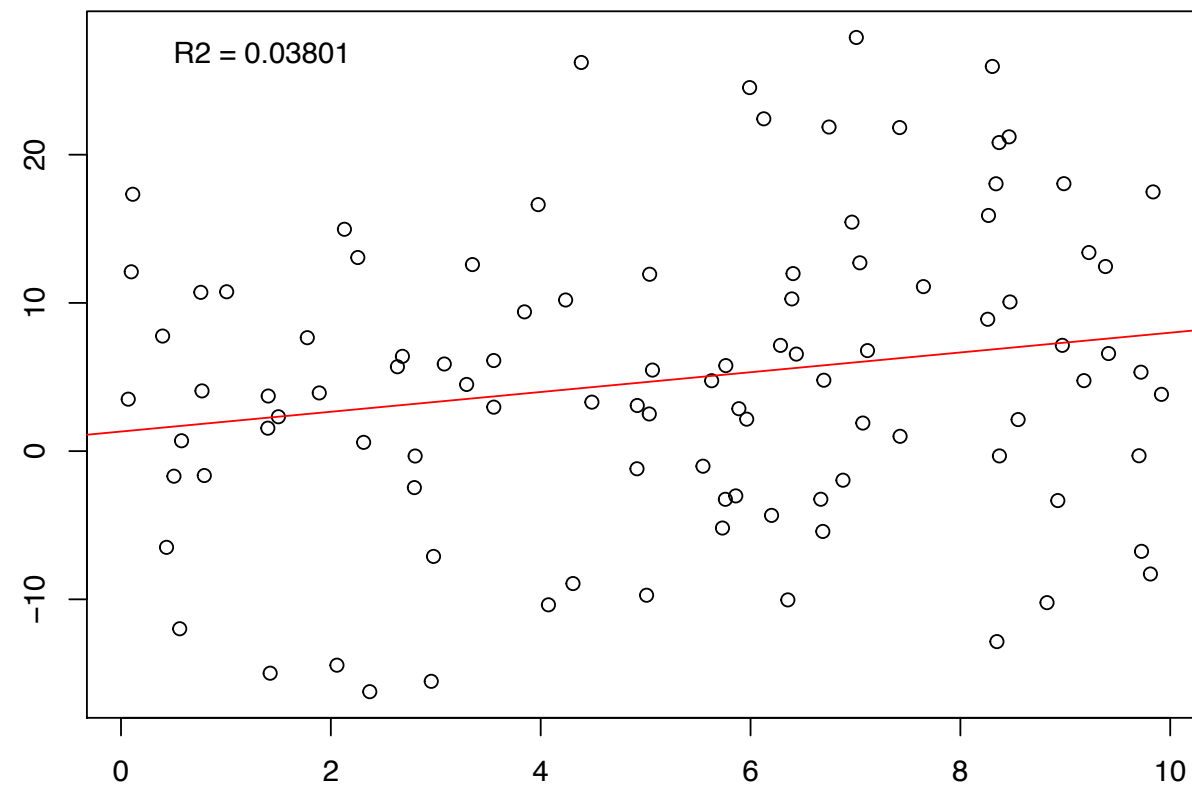
$Y \sim X + N(0, 1)$



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



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



回归标准误

Standard error of regression

- 回归标准误（standard error of regression, SER）是另一个衡量拟合优度的量，其定义为回归误差 u_i 的标准偏差的估计值。

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

- 从定义可以看出，SER 是用因变量单位度量的观测值在回归线附近的离散程度。
- SER 取值为非负实数，没有理论上限。

如何理解 R^2 和 SER

- R^2 衡量的是 OLS 回归线对数据的拟合程度，或者回归变量对因变量变化的解释能力。SER 衡量的是样本数据在回归线附近的离散程度。
- R^2 的取值范围在 0 和 1 之间，越接近 1 代表回归变量能够更好的解释因变量的变化。
- SER 取值越低，代表用回归变量预测因变量的准确度越高。
- R^2 的取值低（SER 的取值高），并不代表 **OLS 回归“好”或“不好”**，而是反映出存在影响因变量的其他重要因素（包含在残差项中）。

在 gretl 中进行 OLS 回归

- 在主程序窗中：

Model > Ordinary least squares >

- 在编程模式中：

ols testscr **const** str

因变量

回归变量

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + u_i \\ &= \beta_0 \mathbf{1} + \beta_1 X_i + u_i \end{aligned}$$

在 gretl 中进行 OLS 回归

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

最小二乘假设

区别因果推论和预测

- OLS 估计值可以很好的回答关于**预测**的问题，例如当一个学区的平均班级规模为已知时，我们可以预测该区的平均考试成绩。
- 但是我们更关心的是**因果推论**，即班级规模的变化是否会，以及在哪种程度上，影响考试成绩。
- 在随机对照试验中，处理的因果效应可以表达为

$$E(Y | X = 1) - E(Y | X = 0)$$

并可以通过比较处理组和对照组的样本均值来估计。

- 当 X 取连续值时，在能够保证随机赋值的情况下，因果效应可以通过 OLS 回归估计。此时我们是在假设 β_1 具有因果效应，即当 X 增加一个单位时给 Y 带来的影响。

进行因果推断所需要的 OLS 假设

假设 1：给定 X_i 时误差项 u_i 的条件分布均值为零，即 $E(u_i | X_i) = 0$

- 误差项 u_i 包含了其他所有可能对因变量产生影响的因素。若假设 1 成立，则 $E(Y_i | X_i) = \beta_0 + \beta_1 X_i$ 。
- 假设 1 保证了在给定 X_i 的情况下，其他变量的变化在均值意义上不对 Y_i 产生影响。同时可导出 X_i 与 u_i 不相关 ($E(u_i | X_i) = 0 \Rightarrow \text{corr}(X_i, u_i) = 0$)，即 X_i 不受其他变量影响、同时 X_i 不会通过其他变量间接地影响 Y_i 。
- 在随机对照试验中，由于 X_i 是随机赋值的，其取值和其他变量独立，即可导出假设 1 (X_i 和 u_i 独立 $\Rightarrow E(u_i | X_i) = 0$)。
- 在观测数据中，我们无法期待 X_i 是随机赋值的。我们最多只能希望 X_i 就像是随机赋值的，即保证 $E(u_i | X_i) = 0$ 。
- 如果 Y_i 也会对 X_i 产生影响，则可导出 $\text{corr}(X_i, u_i) \neq 0$ ，进而得到 $E(u_i | X_i) \neq 0$ （具体见第12章）。因此假设 1 排除了反向因果关系。

进行因果推断所需要的 OLS 假设

假设 2: $(X_i, Y_i), i = 1, 2, \dots, n$ 为独立同分布

- 这个假设是说每一组观测值 (X_i, Y_i) 都服从同一个联合分布，且不同观测值之间相互独立。
- 如果样本是从一个总体中随机抽样得出的，就满足这个假设。
- 大多数抽样调查的数据满足这个假设。
- 有些数据不满足这个假设，例如：
 - 非随机对照试验：如果处理组和对照组是人为分类的，就可能出现选择偏差，也就难以满足同分布的假设。
 - 时间序列数据：宏观经济研究中常用的时间序列数据，是同一变量在不同时间点所观测到的数据，其生成服从特定的随机过程。这种数据容易产生自相关 (autocorrelation / serial correlation)，即 X_t 和 X_{t-1} 之间存在相关性。

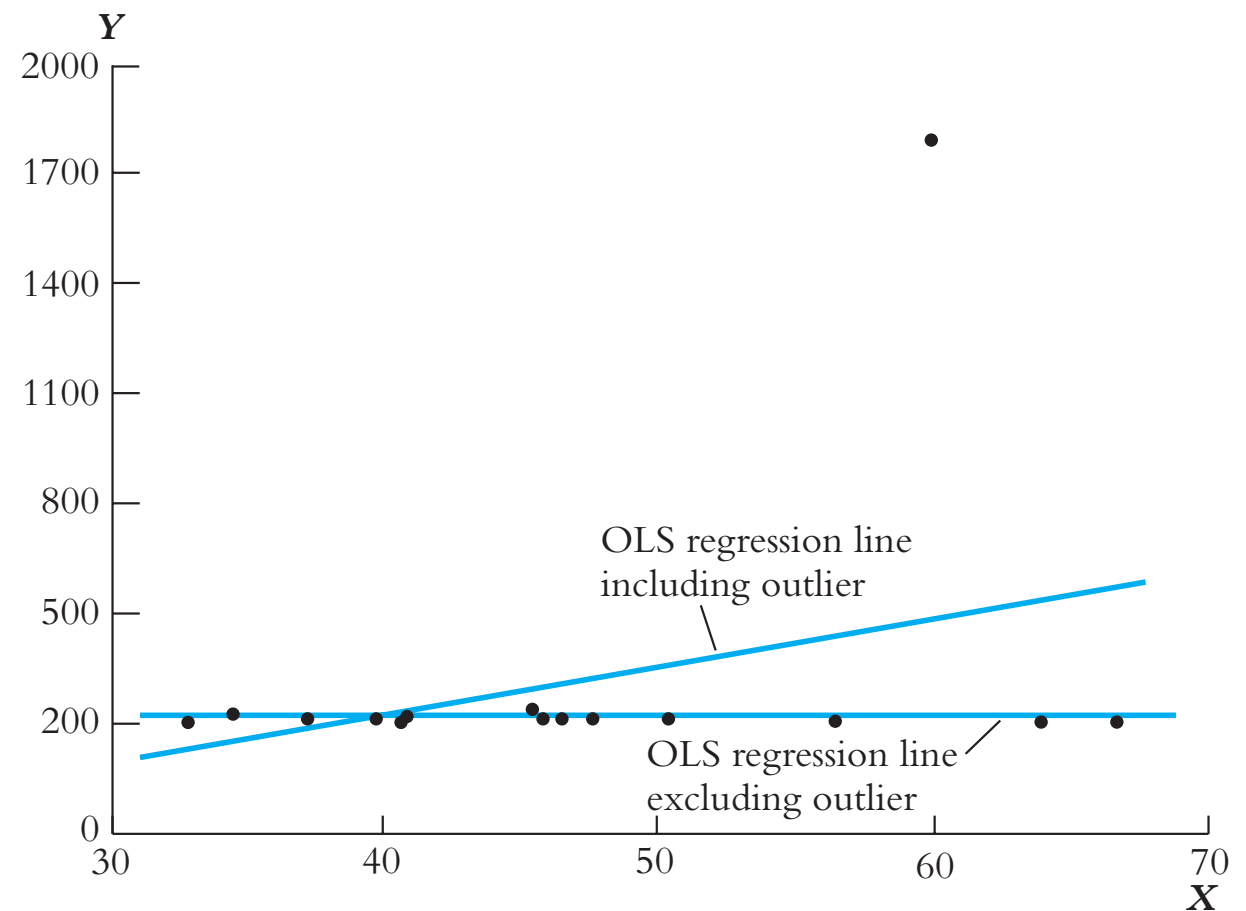
进行因果推断所需要的 OLS 假设

假设 3：不太可能出现大异常值

- OLS 估计对大异常值异常敏感。在大异常值存在的情况下获得的系数有可能偏离真实值。

异常值可能源于数据收集和整理过程中的失误。

如果能确认是观测值本身不正确，应尽量修正，无法修正时可删除该观测值。



- 不存在大异常值的数学描述是变量的四阶矩非零且有限，即 $0 < E(X_i^4) < \infty$, $0 < E(Y_i^4) < \infty$ 。正态分布满足此条件。

进行因果推断所需要的 OLS 假设

总结

1. 给定 X_i 时误差项 u_i 的条件分布均值为零： $E(u_i | X_i) = 0$
 2. $(X_i, Y_i), i = 1, 2, \dots, n$ 为独立同分布
 3. 不太可能出现大异常值： $0 < E(X_i^4) < \infty, 0 < E(Y_i^4) < \infty$
- 以上假设成立时，在大样本条件下，OLS 估计量具有一致性且其抽样分布为正态分布。
 - 以上假设在实际中是否成立，关系到 OLS 估计结果是否可以解释为因果关系（假设1），是否应该对回归方法进行修正（假设2），以及研究者的认真程度（假设3）。

OLS 假设的其他表述方式

- In Baltagi (2011), *Econometrics*, 5ed, Springer.
 - X_i 为非随机变量时:
 - (i) $E(u_i) = 0$; (ii) $\text{var}(u_i) = \sigma_u^2$; (iii) $E(u_i u_j) = 0$; (iv) 不存在大异常值
 - (i), (ii), (iii) 也可以简写成 $u_i \sim \text{IID}(0, \sigma_u^2)$
 - X_i 为随机变量时:
 - (i') $E(u_i | X_i) = 0$; (ii') $\text{var}(u_i | X_i) = \sigma_u^2$; (iii') $E(u_i u_j | X_i) = 0$; (iv)
- In Stock and Watson (2011), 3rd.
 - 假设 1 \Leftrightarrow (i'); 假设 2 \Rightarrow (ii') and (iii'); 假设 3 \Rightarrow (iv)

OLS 估计量的抽样分布

\bar{Y} 的抽样分布

- 令 μ_Y 和 σ_Y^2 分别表示 Y_i 的期望值和方差。

- \bar{Y} 的期望值和方差分别为

$$E(\bar{Y}) = \mu_Y, \quad \text{var}(\bar{Y}) = \frac{\sigma_Y^2}{n}$$

\bar{Y} 是无偏的

- 当总体服从正态分布时, \bar{Y} 也服从正态分布。

$$Y_i \sim N(\mu_Y, \sigma_Y^2) \quad \Rightarrow \quad \bar{Y} \sim N(\mu_Y, \sigma_Y^2/n)$$

- 在大样本条件下:

大数定律: $\bar{Y} \xrightarrow{p} \mu_Y$ \bar{Y} 是一致的

中心极限定理: \bar{Y} 的分布近似于正态分布 $N(\mu_Y, \sigma_Y^2/n)$ 。

$\hat{\beta}_0$ 和 $\hat{\beta}_1$ 的抽样分布

- OLS 估计量 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 是随机变量。
- 在 OLS 假设下, $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 分别是 β_0 和 β_1 的无偏估计量:

$$E(\hat{\beta}_0) = \beta_0, \quad E(\hat{\beta}_1) = \beta_1$$

- 在 OLS 假设下, 如果满足大样本条件, 则 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 服从联合正态抽样分布。 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 的边缘分布也都是正态分布:

$$\hat{\beta}_0 \sim N(\beta_0, \sigma_{\hat{\beta}_0}^2), \quad \sigma_{\hat{\beta}_0}^2 = \frac{1}{n} \frac{\text{var}(H_i u_i)}{[E(H_i^2)]^2}, \quad \text{其中 } H_i = 1 - \left[\frac{\mu_X}{E(X_i^2)} \right] X_i$$

$$\hat{\beta}_1 \sim N(\beta_1, \sigma_{\hat{\beta}_1}^2), \quad \sigma_{\hat{\beta}_1}^2 = \frac{1}{n} \frac{\text{var}[(X_i - \mu_X)u_i]}{[\text{var}(X_i)]^2}$$

$\hat{\beta}_0$ 和 $\hat{\beta}_1$ 是一致估计量
推导过程见附录4.3

X_i 的方差越大, 则 $\hat{\beta}_1$ 的方差越小

FIGURE 4.5 The Variance of $\hat{\beta}_1$ and the Variance of X

The colored dots represent a set of X_i 's with a small variance. The black dots represent a set of X_i 's with a large variance. The regression line can be estimated more accurately with the black dots than with the colored dots.

