应用回归分析

上海财经大学　统计与管理学院

SHANGHAI UNIVERSITY OF FINANCE & ECONOMICS
上海财经大学
1917

第二章 简单线形回归

章节概括：

- 简单线形模型
- 最小二乘法
- 参数估计
- 方差分析
- 置信区间估计和检验

简单线形模型

- 均值函数和方差函数

$$\begin{aligned} \mathrm{E}(Y|X=x) &= \beta_0+\beta_1 x \\ \mathrm{Var}(Y|X=x) &= \sigma^2 \end{aligned}$$

- 参数： $\beta_0, \beta_1, \sigma^2$
- 未知：估计、检验
- 统计误差（statistical error)

$$e_i = y_i - \mathrm{E}(Y|X=x_i)$$

- 独立、期望为零 $\mathrm{E}(e_i|X_i)=0$
- 正态分布 （小样本）

结距与斜率

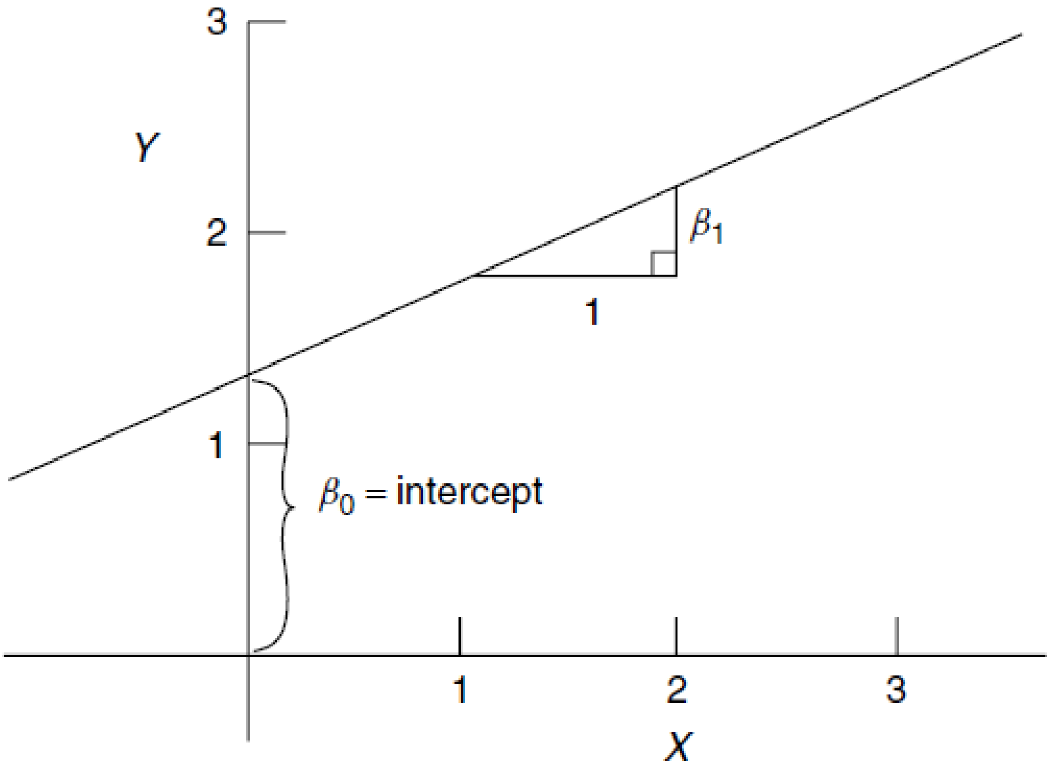


FIG. 2.1 Equation of a straight line $E(Y|X = x) = \beta_0 + \beta_1 x$.

二、最小二乘估计 -- 图示

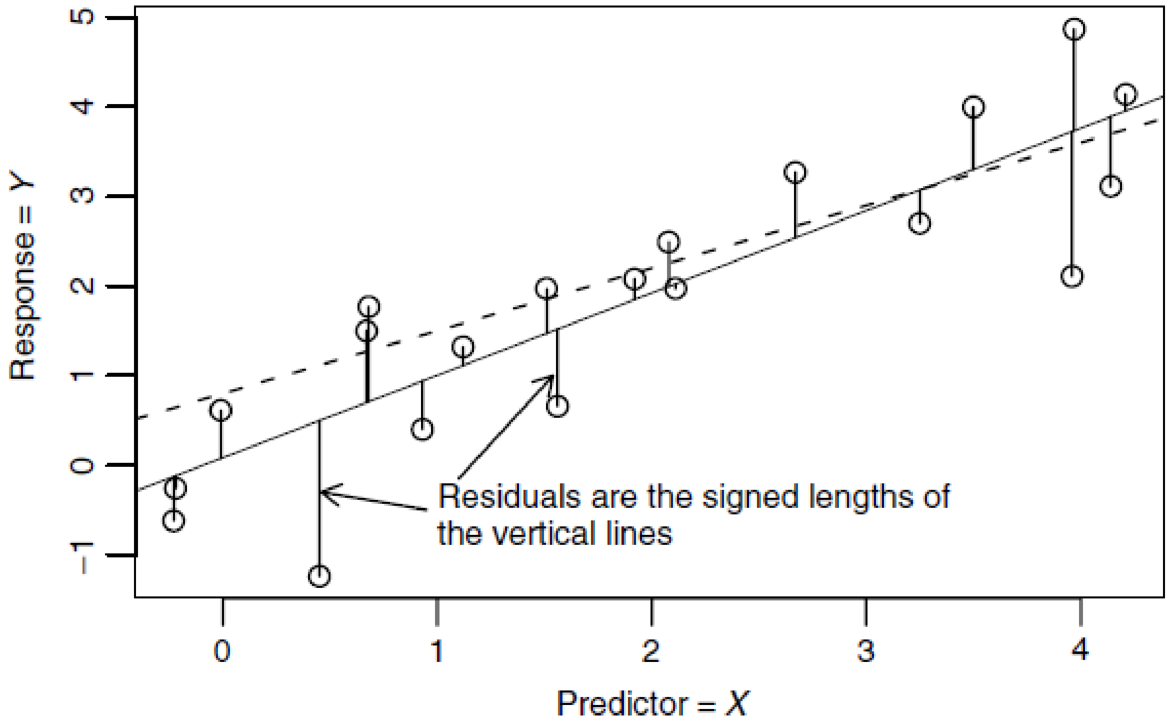


FIG. 2.3 A schematic plot for OLS fitting. Each data point is indicated by a small circle, and the solid line is a candidate OLS line given by a particular choice of slope and intercept. The solid vertical lines between the points and the solid line are the residuals. Points below the line have negative residuals, while points above the line have positive residuals.

OLS

- **最小二乘估计 (ordinary least squares, OLS)**
- 残差平方和 (residual sum of squares, RSS)

$$RSS(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

- 残差平方和最小化 → 最小二乘估计
- 估计（随机性）与 参数的区别

OLS

$$\frac{\partial RSS(\beta_0, \beta_1)}{\beta_0} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x) = 0$$

$$\frac{\partial RSS(\beta_0, \beta_1)}{\beta_1} = -2\sum_{i=1}^{n}x_i(y_i - \beta_0 - \beta_1 x) = 0$$

$$\beta_0 n + \beta_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \qquad SXX = \textstyle\sum_{i=1}^{n}(x_i - \bar{x})^2 \displaystyle = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

$$\beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \qquad SXY = \textstyle\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \displaystyle = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$$

$$\hat{\beta}_0 = \bar{y} - \beta_1 \bar{x}, \quad \hat{\beta}_1 = \frac{SXY}{SXX}$$

OLS

- 参数 vs 参数估计 / 统计量
- Parameter vs estimate / statistic

$$\beta_0, \beta_1 \qquad \text{vs} \qquad \hat{\beta}_0, \hat{\beta}_1$$

- 拟合值 $\hat{y}_i = \hat{\mathrm{E}}(Y|X = x_i) = \hat{\beta}_0 + \hat{\beta}_1 x_i$

- 统计误差与残差

$$\begin{aligned} e_i &= y_i - \beta_0 - \beta_1 x_i, i = 1, \cdots, n \\ \hat{e}_i &= y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i, i = 1, \cdots, n \end{aligned}$$

统计量、R演示

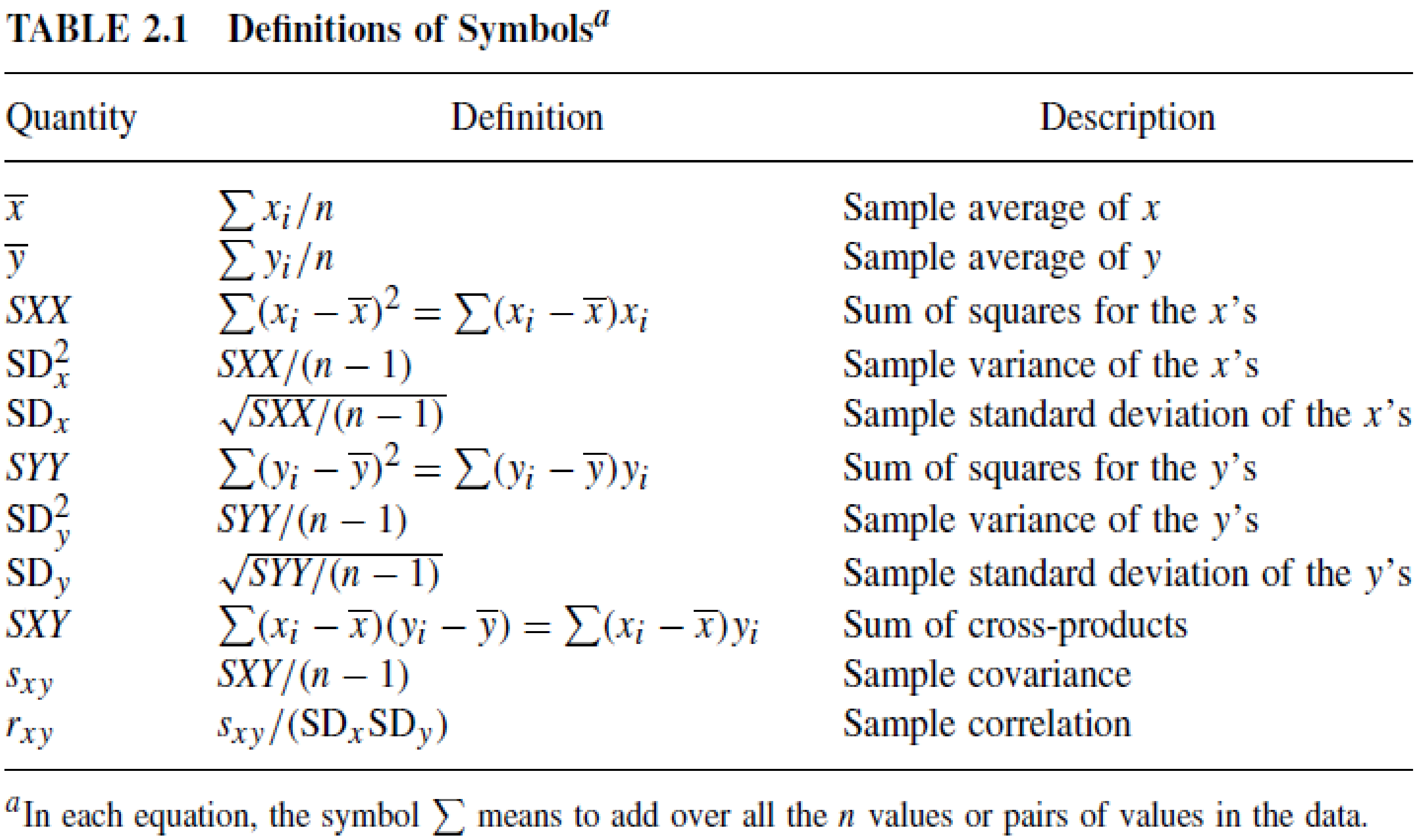


TABLE 2.1 Definitions of Symbols[a]

Quantity	Definition	Description
$\overline{x}$	$\sum x_i/n$	Sample average of x
$\overline{y}$	$\sum y_i/n$	Sample average of y
SXX	$\sum(x_i-\overline{x})^2=\sum(x_i-\overline{x})x_i$	Sum of squares for the x's
SD_x^2	$SXX/(n-1)$	Sample variance of the x's
SD_x	$\sqrt{SXX/(n-1)}$	Sample standard deviation of the x's
SYY	$\sum(y_i-\overline{y})^2=\sum(y_i-\overline{y})y_i$	Sum of squares for the y's
SD_y^2	$SYY/(n-1)$	Sample variance of the y's
SD_y	$\sqrt{SYY/(n-1)}$	Sample standard deviation of the y's
SXY	$\sum(x_i-\overline{x})(y_i-\overline{y})=\sum(x_i-\overline{x})y_i$	Sum of cross-products
s_{xy}	$SXY/(n-1)$	Sample covariance
r_{xy}	$s_{xy}/(\mathrm{SD}_x\mathrm{SD}_y)$	Sample correlation

[a] In each equation, the symbol $\sum$ means to add over all the n values or pairs of values in the data.

Forbes Data

$$\hat{\beta}_1 = \frac{SXY}{SXX} = r_{xy}\frac{\mathrm{SD}_y}{\mathrm{SD}_x} = r_{xy}\left(\frac{SYY}{SXX}\right)^2$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1\overline{x}$$

$$\overline{x} = 202.95294 \quad SXX = 530.78235 \quad SXY = 475.31224$$
$$\overline{y} = 139.60529 \quad SYY = 427.79402$$

$$\hat{\beta}_1 = \frac{SXY}{SXX} = 0.895$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1\overline{x} = -42.138$$

$$\widehat{\mathrm{E}}(Lpres|Temp) = -42.138 + 0.895Temp$$
$$= 139.606 + 0.895(Temp - 202.953)$$

估计方差

- 方差约为统计误差平方的平均
- 方差估计为残差平方的平均

$$\hat{\sigma}^2 = \frac{RSS}{n-2}$$

- 自由度 (Degree of Freedom) = df = n-2
- n vs (n-2)
- RSS：residual mean square 残差均方

$$RSS = SYY - \frac{SXY^2}{SXX} = SYY - \hat{\beta}_1^2 SXX$$

Forbe’s Data

- $$\begin{aligned} RSS &= 427.79402 - \frac{475.31224^2}{530.78235} \\ &= 2.15493 \\ \sigma^2 &= \frac{2.15493}{17-2} = 0.14366 \end{aligned}$$

- 标准误差(standard error)

$$\hat{\sigma} = \sqrt{0.14366} = 0.37903$$

方差估计的统计性质

- 卡方分布

$$(n-2)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2)$$

- 无偏性

$$\mathrm{E}(\hat{\sigma}^2) = \sigma^2$$

最小二乘法估计的统计性质

- 样本数据 --〉 参数估计
- 设 $c_i = (x_i - \overline{x})/SXX$

$$\hat{\beta}_1 = \sum \left(\frac{x_i - \overline{x}}{SXX} \right) y_i = \sum c_i y_i$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

- 无偏性

$$E(\hat{\beta}_0) = \beta_0$$

$$E(\hat{\beta}_1) = \beta_1$$

最小二乘法估计的统计性质

- 假设 $\mathrm{Var}(e_i) = \sigma^2, i = 1, \ldots, \quad n,$

  $\mathrm{Cov}(e_i, e_j) = 0, i \neq j,$

- 则

$$\mathrm{Var}(\hat{\beta}_1) = \sigma^2 \frac{1}{SXX}$$

$$\mathrm{Var}(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{SXX} \right)$$

$$\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1) = -\sigma^2 \frac{\overline{x}}{SXX}$$

$$\rho(\hat{\beta}_0, \hat{\beta}_1) = \frac{-\overline{x}}{\sqrt{SXX/n + \overline{x}^2}} = \frac{-\overline{x}}{\sqrt{(n-1)\mathrm{SD}_x^2/n + \overline{x}^2}}$$

估计最优性

- Gauss-Markov Theorem 高斯-马尔科夫定理
- 在所有无偏的线形组合（ys）估计中，最小二乘估计有最小方差。

- 误差服从正态分布，

$$e_i \sim \mathrm{NID}(0, \sigma^2) \qquad i = 1, \ldots, n$$

- OLS亦为最大似然估计，服从正态分布
- 大数，亦服从正态分布

估计的方差的估计

- $$\widehat{\mathrm{Var}}(\hat{\beta}_1) = \hat{\sigma}^2 \frac{1}{SXX}$$

  $$\widehat{\mathrm{Var}}(\hat{\beta}_0) = \hat{\sigma}^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{SXX} \right)$$

  $$\mathrm{se}(\hat{\beta}_1) = \sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_1)}$$

方差分析

- Analysis of Variance (ANOVA)

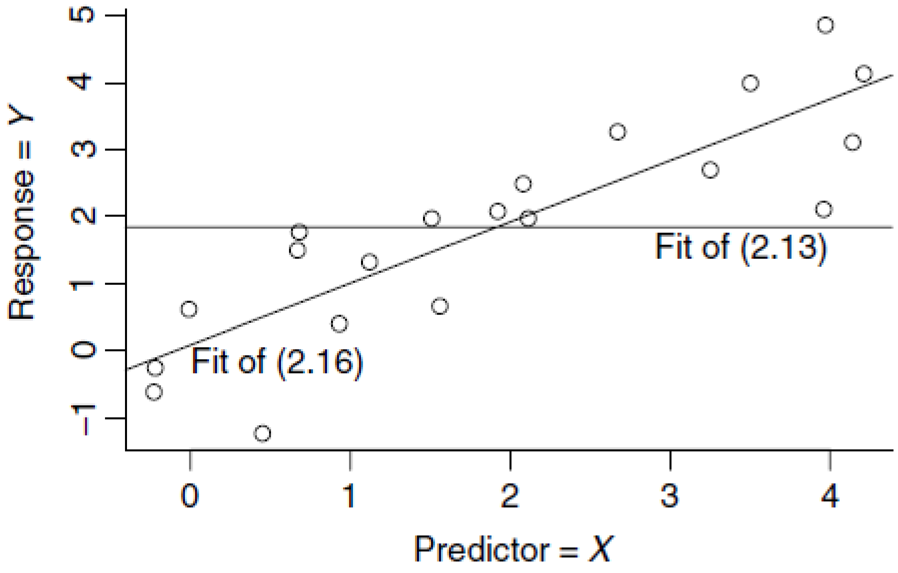


FIG. 2.4 Two mean functions compared by the analysis of variance.

ANOVA

- $\mathrm{E}(Y|X = x) = \beta_0$
- $\hat{\beta}_0 = \overline{y}$
- $\sum(y_i - \hat{\beta}_0)^2 = \sum(y_i - \overline{y})^2 = SYY$
- $n - 1$ df

- $\mathrm{E}(Y|X = x) = \beta_0 + \beta_1 x$
- $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}, \quad \hat{\beta}_1 = \dfrac{SXY}{SXX}$
- $RSS = SYY - \dfrac{(SXY)^2}{SXX}$
- $n - 2$ df

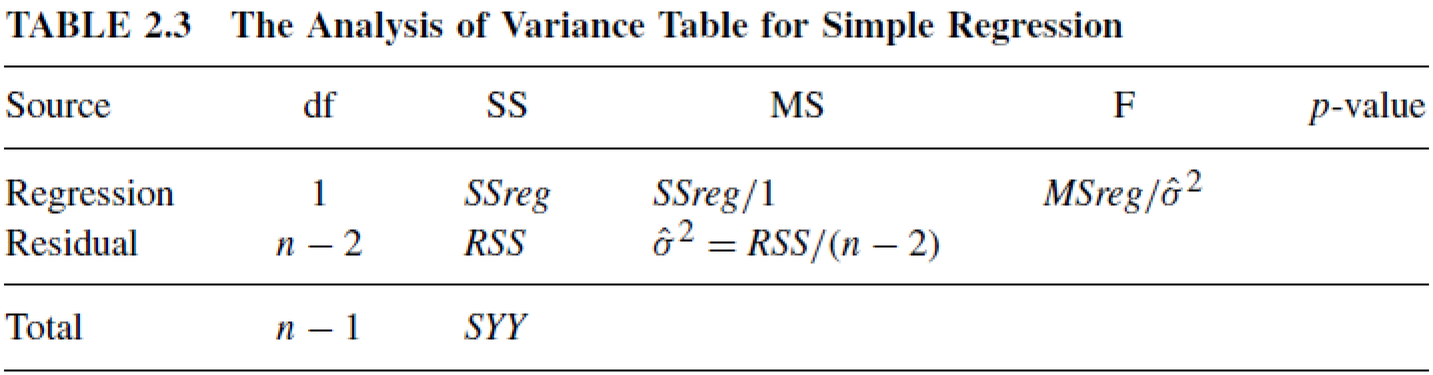


TABLE 2.3 The Analysis of Variance Table for Simple Regression

Source	df	SS	MS	F	p-value
Regression	1	$SSreg$	$SSreg/1$	$MSreg/\hat{\sigma}^2$	
Residual	$n - 2$	RSS	$\hat{\sigma}^2 = RSS/(n - 2)$		
Total	$n - 1$	SYY			

Forbe’s Data

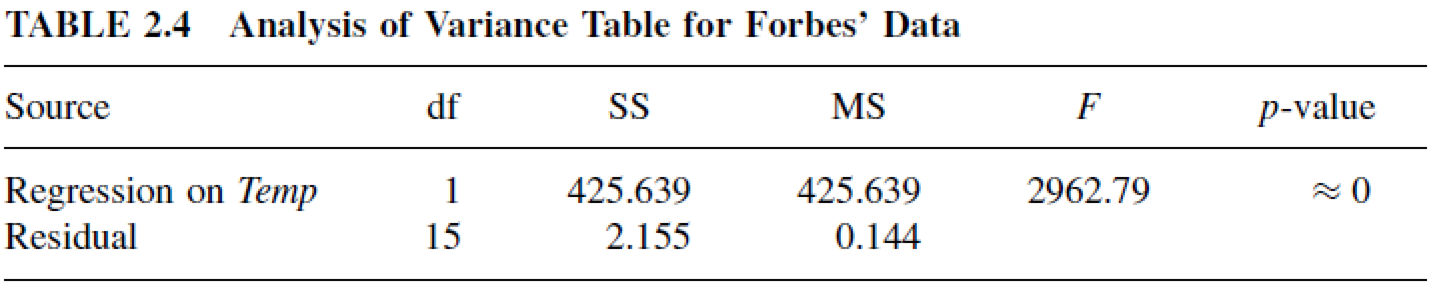


TABLE 2.4 Analysis of Variance Table for Forbes’ Data

Source	df	SS	MS	F	p-value
Regression on *Temp*	1	425.639	425.639	2962.79	≈ 0
Residual	15	2.155	0.144		

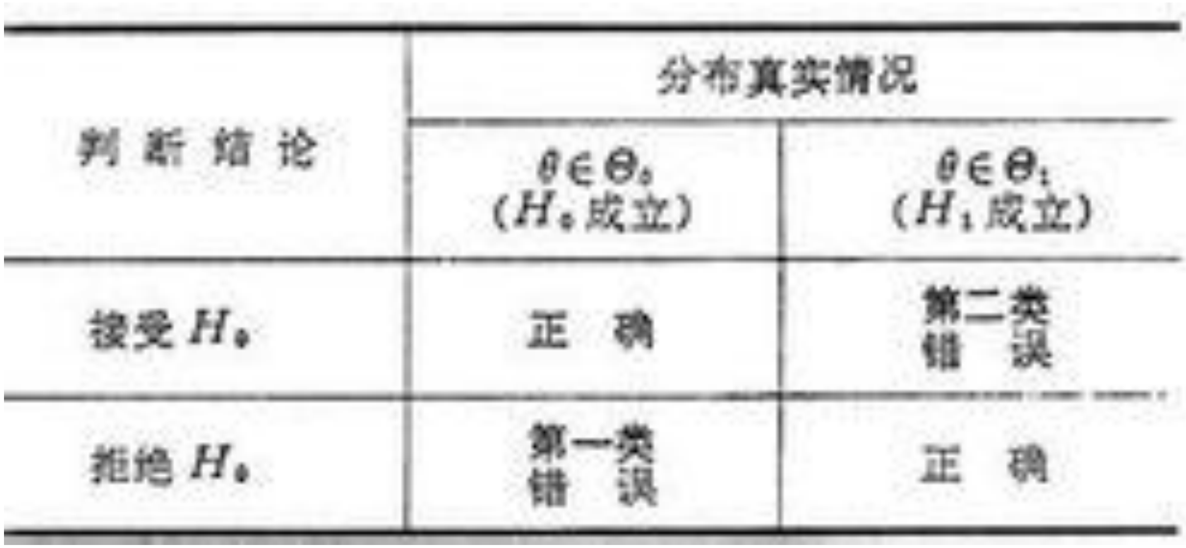
F检验

- $$\begin{aligned}\text{NH:}\quad & \mathrm{E}(Y|X=x)=\beta_0\\ \text{AH:}\quad & \mathrm{E}(Y|X=x)=\beta_0+\beta_1 x\end{aligned}$$
- 检验统计量 $$F=\frac{(SYY-RSS)/1}{\hat{\sigma}^2}=\frac{SSreg/1}{\hat{\sigma}^2}$$
- 大样本 / $$e_i \sim \mathrm{NID}(0,\sigma^2) \qquad i=1,\ldots,n$$
- 原假设成立下服从下 $F(1,n-2)$
- Forbe’s Data: $$F=\frac{425.639}{0.144}=2963$$

p-value解释

- P值就是当原假设为真时所得到的样本观察结果或更极端结果出现的概率
- P值越小，结果越显著, 拒绝原假设的理由越充分
- 显著性水平α
- 无对立假设

检验功效 Power



判断结论	分布真实情况	
	$\theta\in\Theta_0$ (H_0成立)	$\theta\in\Theta_1$ (H_1成立)
接受 H_0	正确	第二类错误
拒绝 H_0	第一类错误	正确

判定系数

- coefficient of determination
- $$R^2=\frac{SSreg}{SYY}=1-\frac{RSS}{SYY}=\frac{425.63910}{427.79402}=0.995$$
- 拟合优度越大,自变量对因变量的解释程度越高,自变量引起的变动占总变动的百分比高.观察点在回归直线附近越密集
- 取值范围:0-1
- $$R^2=\frac{SSreg}{SYY}=\frac{(SXY)^2}{SXX\times SYY}=r_{xy}^2$$

置信区间与假设检验

- $e_i \sim \text{NID}(0, \sigma^2) \qquad i = 1, \ldots, n$
- 截距
- 置信区间

$$-42.138 - 1.753(3.340) \le \beta_0 \le -42.136 + 1.753(3.340)$$

$$-47.993 \le \beta_0 \le -36.282$$

$$\hat{\beta}_0 - t(\alpha/2, n-2)\text{se}(\hat{\beta}_0) \le \beta_0 \le \hat{\beta}_0 + t(\alpha/2, n-2)\text{se}(\hat{\beta}_0)$$

- 假设检验

$$\begin{array}{lll} \text{NH:} & \beta_0 = \beta_0^*, & \beta_1 \text{ arbitrary} \\ \text{AH:} & \beta_0 \ne \beta_0^*, & \beta_1 \text{ arbitrary} \end{array}$$

$$t = \frac{\hat{\beta}_0 - \beta_0^*}{\text{se}(\hat{\beta}_0)} \qquad t = \frac{-42.138 - (-35)}{3.340} = 2.137$$

置信区间与假设检验

- 斜率
- 置信区间

$$\mathrm{se}(\hat{\beta}_1) = \hat{\sigma}/\sqrt{SXX}$$

$$0.8955 - 2.131(0.0164) \le \beta_1 \le 0.8955 + 2.131(0.0164)$$

$$0.867 \le \beta_1 \le 0.930$$

- 假设检验

$$\begin{aligned} \text{NH:}\quad & \beta_1 = 0 \\ \text{AH:}\quad & \beta_1 \neq 0 \end{aligned}$$

Ft. Collins data, $t = (0.20335 - 0)/0.1310 = 1.553$.

预测

- $$\tilde{y}_* = \hat{\beta}_0 + \hat{\beta}_1 x_*$$

- 新观测值的预测

$$\mathrm{Var}(\tilde{y}_* | x_*) = \sigma^2 + \sigma^2 \left(\frac{1}{n} + \frac{(x_* - \overline{x})^2}{SXX} \right)$$

$$\mathrm{sepred}(\tilde{y}_* | x_*) = \hat{\sigma} \left(1 + \frac{1}{n} + \frac{(x_* - \overline{x})^2}{SXX} \right)^{1/2}$$

- Pearson’s Data

$$\hat{\beta}_0 + \hat{\beta}_1 x_* \pm t(.025, 15)\mathrm{sepred}(\widetilde{Dheight}_* | Mheight_*)$$

预测

- 原有观测值的预测

$$\mathrm{sefit}(\tilde{y}_*|x_*) = \hat{\sigma}\left(\frac{1}{n} + \frac{(x_* - \overline{x})^2}{SXX}\right)^{1/2}$$

- 同时置信区间

$$(\hat{\beta}_0 + \hat{\beta}_1 x) - \mathrm{sefit}(\hat{y}|x)[2F(\alpha; 2, n-2)]^{1/2} \le y$$
$$\le (\hat{\beta}_0 + \hat{\beta}_1 x) + \mathrm{sefit}(\hat{y}|x)[2F(\alpha; 2, n-2)]^{1/2}$$

预测

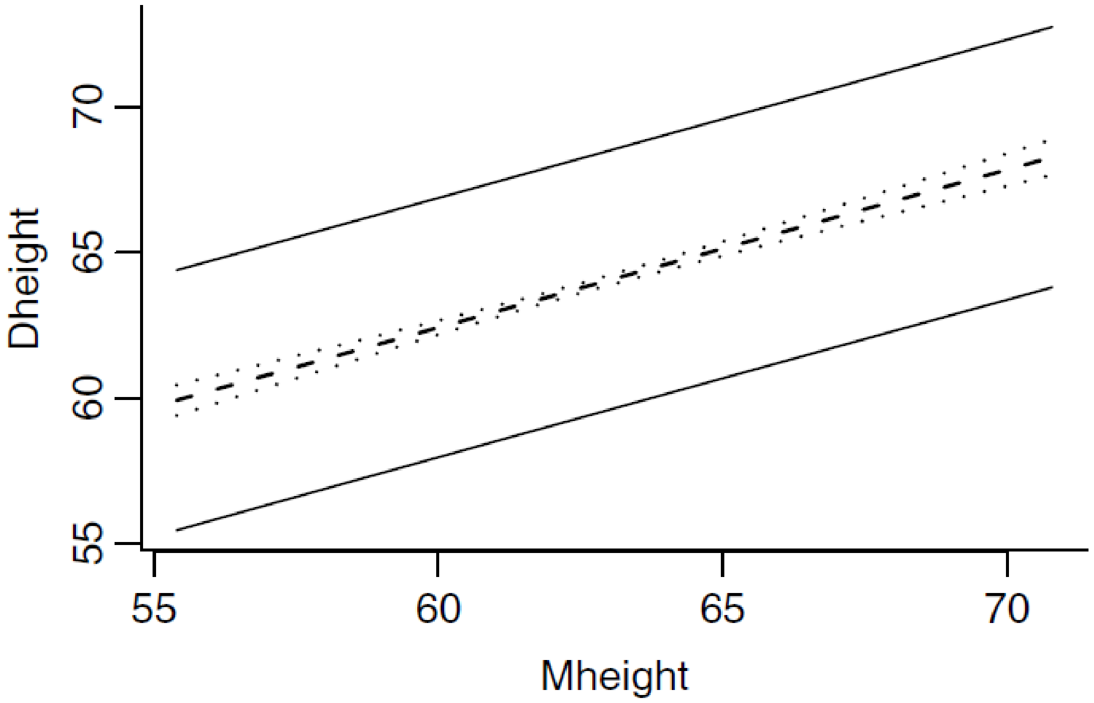


FIG. 2.5 Prediction intervals (solid lines) and intervals for fitted values (dashed lines) for the heights data.

残差图

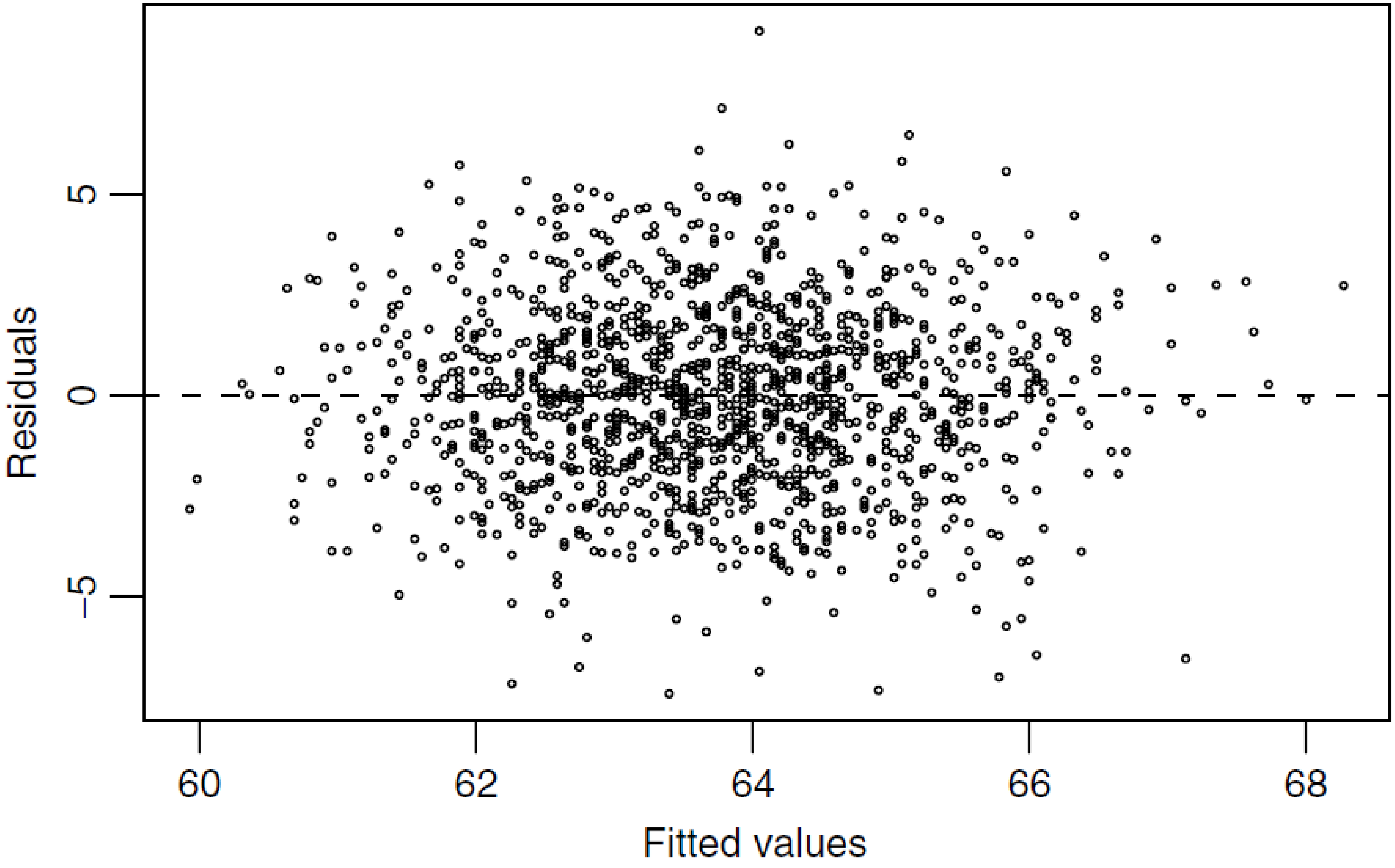
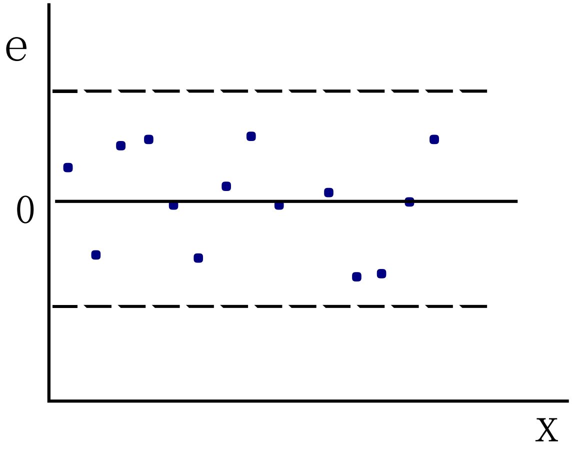
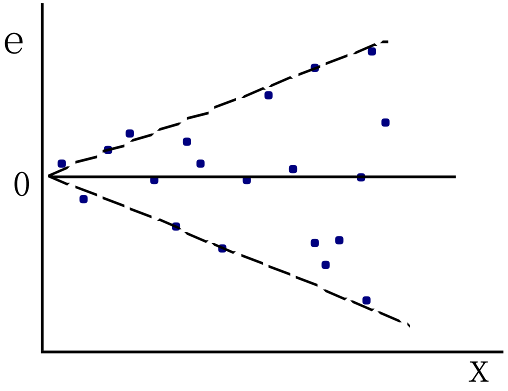


FIG. 2.6 Residuals versus fitted values for the heights data.

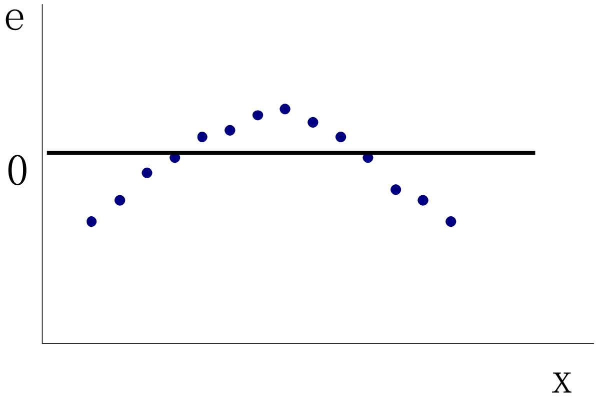
残差图



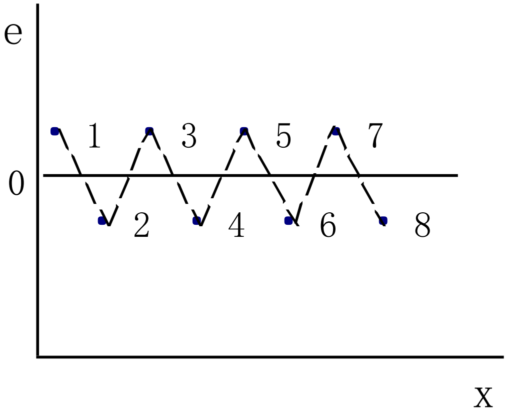
(a)



(b)



(c)



(d)

残差

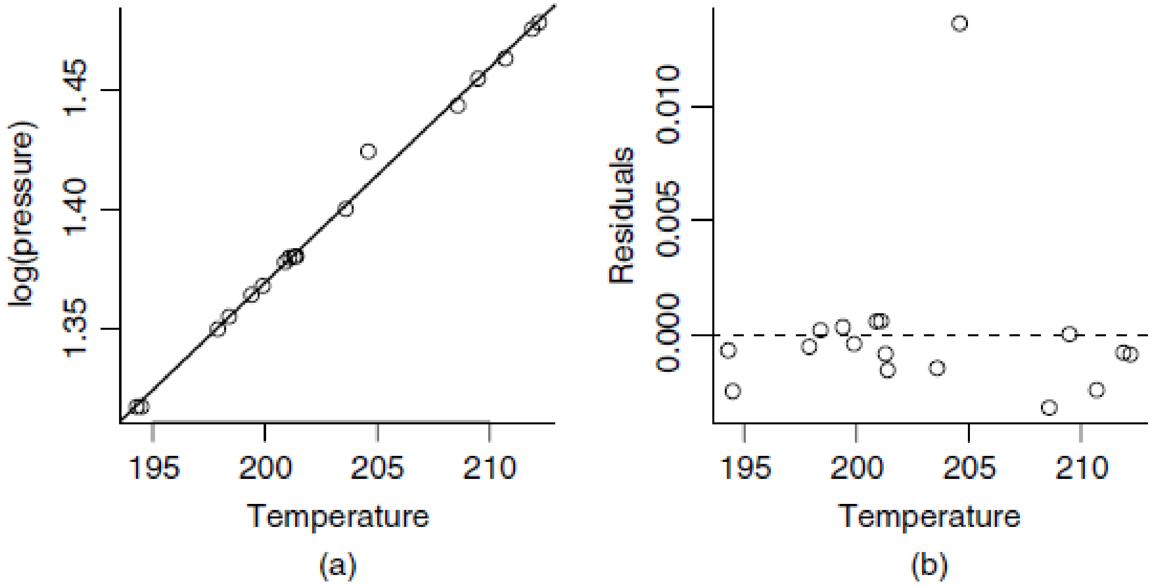


FIG. 1.4 (a) Scatterplot of Forbes' data. The line shown is the OLS line for the regression of log(*Pressure*) on *Temp*. (b) Residuals versus *Temp*.

残差

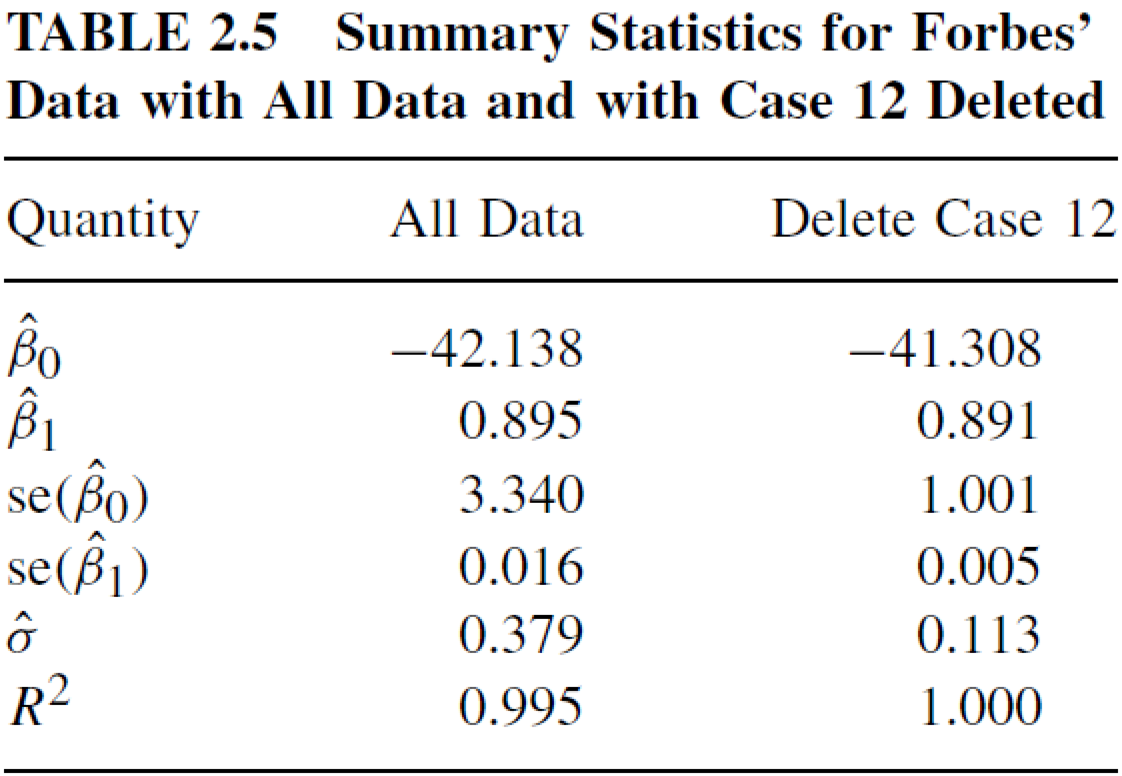


TABLE 2.5 Summary Statistics for Forbes' Data with All Data and with Case 12 Deleted

Quantity	All Data	Delete Case 12
$\hat{\beta}_0$	−42.138	−41.308
$\hat{\beta}_1$	0.895	0.891
$\mathrm{se}(\hat{\beta}_0)$	3.340	1.001
$\mathrm{se}(\hat{\beta}_1)$	0.016	0.005
$\hat{\sigma}$	0.379	0.113
R^2	0.995	1.000

Thank You !

SHANGHAI UNIVERSITY OF FINANCE & ECONOMICS

上海财经大学

1917