# §8.2 回归分析

## 一、回归分析的基本思想

十九世纪，英国生物学家兼统计学家高尔顿研究发现：$\hat{y} = 33.73 + 0.516x$

其中$x$表示父亲身高，$y$表示成年儿子的身高（单位：英寸，1英寸=2.54厘米）。这表明子代的平均高度有向中心回归的意思，使得一段时间内人的身高相对稳定。之后回归分析的思想渗透到了数理统计的其它分支中。

- 回归分析处理的是变量与变量间的关系。变量间常见的关系有两类：确定性关系与相关关系。

- 变量间的相关关系不能用完全确切的函数形式表示，但在平均意义下有一定的定量关系表达式，寻找这种定量关系表达式就是回归分析的主要任务。

- 回归分析便是研究变量间相关关系的一门学科。它通过对客观事物中变量的大量观察或试验获得的数据，去寻找隐藏在数据背后的相关关系，给出它们的表达形式——回归函数的估计。

**回归分析**——处理变量之间的相关关系的一种最常用的数理统计方法.

1.建立变量之间的相关关系式——回归模型

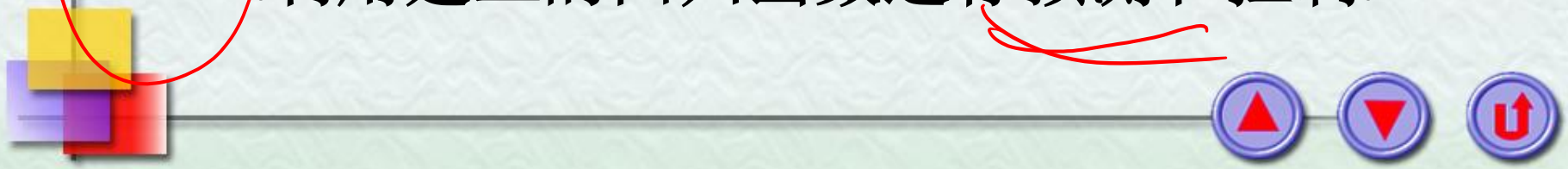
因变量　　自变量　　随机误差

$$Y = f(X_1, X_2, \ldots, X_p) + \varepsilon \qquad \varepsilon \sim N(0, \sigma^2)$$

回归函数

2.对建立的回归函数作显著性检验.

3.利用建立的回归函数进行预测和控制.

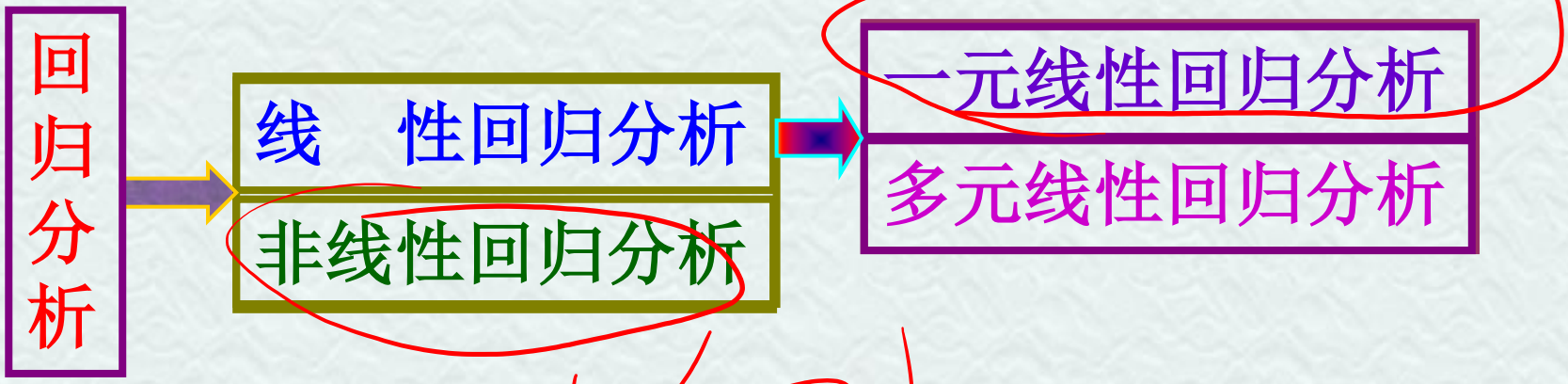


$$Y = f(X_1, X_2, \ldots, X_p) + \varepsilon$$

研究两个变量之间相关关系称为一元回归分析；

研究多个变量间的相关关系称为多元回归分析.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$



进行回归分析首先是回归函数形式的选择。当只有一个自变量时，通常可采用画散点图的方法进行选择。

例1　合金的强度$y$ ($\times 10^7$Pa) 与合金中碳的含量$x$ (%) 有关。为研究两个变量间的关系，首先是收集数据，我们把收集到的数据记为$(x_i,y_i),i=1,2,\ldots,n$。本例中，我们收集到12组数据，列于表1中

## 表1 合金钢强度$y$与碳含量$x$的数据

| 序号 | $x$(%) | $y$ ($\times 10^7$Pa) | 序号 | $x$(%) | $y$ ($\times 10^7$Pa) |
|---|---|---|---|---|---|
| 1 | 0.10 | 42.0 | 7 | 0.16 | 49.0 |
| 2 | 0.11 | 43.0 | 8 | 0.17 | 53.0 |
| 3 | 0.12 | 45.0 | 9 | 0.18 | 50.0 |
| 4 | 0.13 | 45.0 | 10 | 0.20 | 55.0 |
| 5 | 0.14 | 45.0 | 11 | 0.21 | 55.0 |
| 6 | 0.15 | 47.5 | 12 | 0.23 | 60.0 |

为找出两个量间存在的回归函数的形式，可以画一张图：把每一对数$(x_i,y_i)$看成直角坐标系中的一个点，在图上画出$n$个点，称这张图为散点图，见图8.4.1

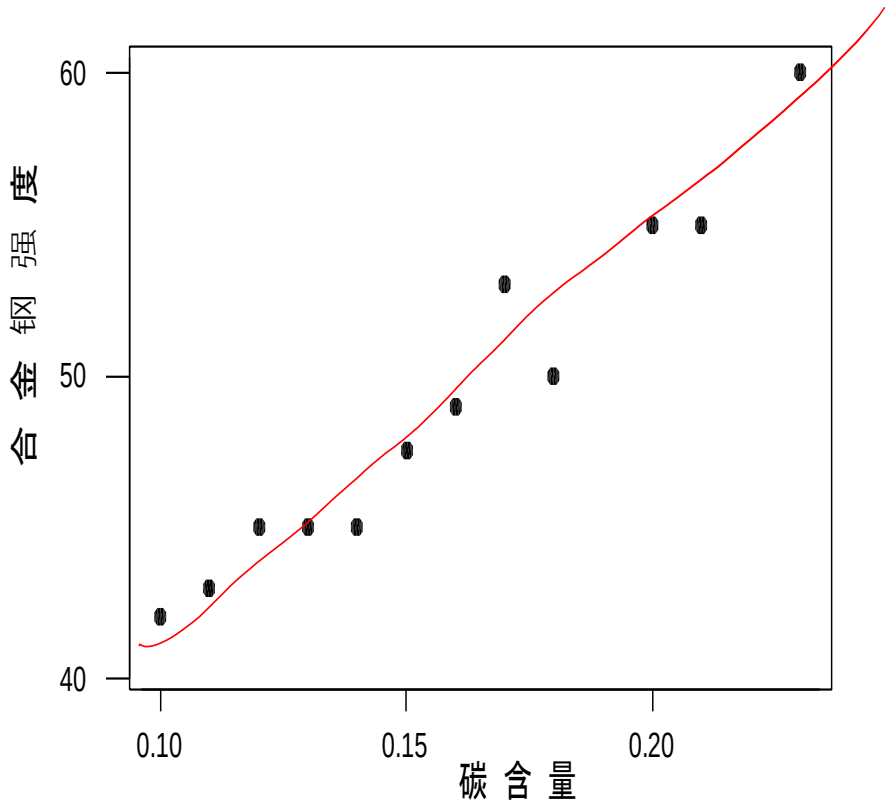



图 8.4.1 合金钢强度及碳含量的散点图

本节重点讨论一元线性回归，其内容为

（1）建立一元线性回归模型：

$$Y = a + bX + \varepsilon$$

主要任务估计回归系数$a, b$；

（2）对所建立的一元线性回归方程进行显著性检验，

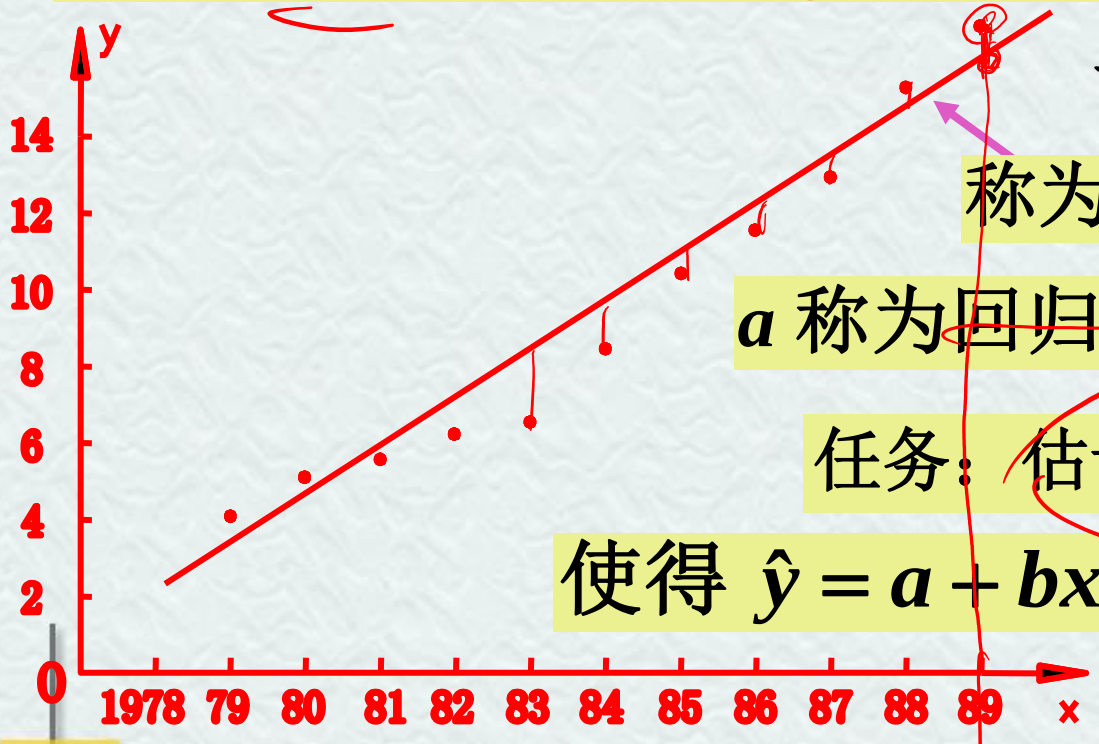
检验 $H_0 : b = 0$；

（3）利用一元线性回归方程进行预测和控制.

# 一、问题的提出

例1　由北京市城市居民家庭生活抽样调查，得1978至1989年人均收入与人均食品消费的12年数据。

以年份作为$X$，人均收入（或食品消费）为$Y$




设：$\hat{y}=a+bx$

称为回归方程(或回归直线)

$a$ 称为回归常数，$b$ 称为回归系数

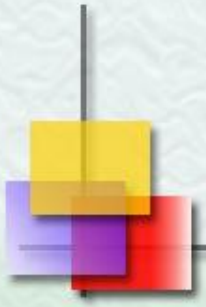
任务：估计参数 $a$ 和 $b$

使得 $\hat{y}=a+bx$ 与 $Y$ 拟合的最好.

最小二乘法

# 最小二乘法

- 1805 勒让德
- 1809 高斯



# 二、最小二乘法

观测值$y_i$与$x_i$之间的关系式：$y_i = a + bx_i + \varepsilon_i, i = 1, 2, \cdots, n.$

总的误差平方和

$$Q(a,b) = \sum_{i=1}^{n}[y_i - (a + bx_i)]^2$$

求使得$Q(a,b)$达到最小值的$(\hat{a},\hat{b})$作为回归系数$a,b$的估计值.

$$\begin{cases} \dfrac{\partial Q}{\partial a} = -2\sum_{i=1}^{n}[y_i - (a + bx_i)] = 0 \\ \dfrac{\partial Q}{\partial b} = -2\sum_{i=1}^{n}[y_i - (a + bx_i)]x_i = 0 \end{cases}$$

即

$$\begin{cases} na + nb\bar{x} = n\bar{y} \\ n\bar{x}a + b\sum_{i=1}^{n}x_i^2 = \sum_{i=1}^{n}x_i y_i \end{cases}$$

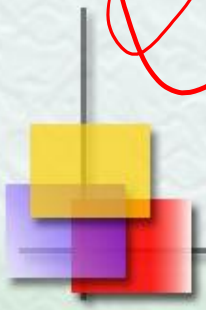
正规方程组

$$\begin{cases} na + nb\bar{x} = n\bar{y} \\ n\bar{x}a + b\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \end{cases}$$

$$\hat{b} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\,\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$

最小二乘估计（*LSE*）



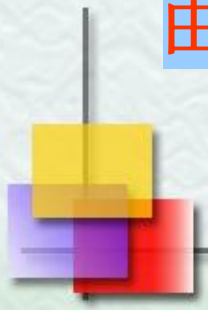
$$\text{记 } l_{xx}=\sum_{i=1}^{n}(x_i-\bar{x})^2=\sum_{i=1}^{n}x_i^2-n\bar{x}^2$$

$$l_{yy}=\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}y_i^2-n\bar{y}^2$$

$$l_{xy}=\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})=\sum_{i=1}^{n}x_iy_i-n\bar{x}\,\bar{y}$$

$$\text{则 } \hat{b}=\frac{l_{xy}}{l_{xx}},\quad \hat{a}=\bar{y}-\hat{b}\bar{x}=\bar{y}-\frac{l_{xy}}{l_{xx}}\bar{x}$$

由此知回归方程一定过点 $(\bar{x},\bar{y})$.



例2 合成纤维的强度 $y$ 与其拉伸倍数 $x$ 有关,测得

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $x_i$ | 2.0 | 2.5 | 2.7 | 3.5 | 4.0 | 4.5 | 5.2 | 6.3 | 7.1 | 8.0 | 9.0 | 10.0 |
| $y_i$ | 1.3 | 2.5 | 2.5 | 2.7 | 3.5 | 4.2 | 5.0 | 6.4 | 6.3 | 7.0 | 8.0 | 8.1 |

求 $y$ 对 $x$ 的一元线性回归方程.

解: $n=12,\ \sum_{i=1}^{n} x_i = 64.8\quad \sum_{i=1}^{n} y_i = 57.5\quad \bar{x}=5.4, \bar{y}\approx 4.79$

$$\sum_{i=1}^{n} x_i^2 = 428.18\quad \sum_{i=1}^{n} y_i^2 = 335.63\quad \sum_{i=1}^{n} x_i y_i = 378$$

$$l_{xx} = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 = 428.18 - 12\times 5.4\times 5.4 = 78.26$$

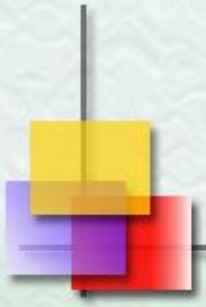
$$l_{xy}=\sum_{i=1}^{n}x_i y_i-n\bar{x}\,\bar{y}=378-5.4\times 57.5=67.5$$

$$\hat{b}=\frac{l_{xy}}{l_{xx}}=\frac{67.5}{78.26}\approx 0.8625$$

$$y=a+bx+\varepsilon$$

$$\hat{a}=\bar{y}-\hat{b}\bar{x}\approx 4.79-0.8625\times 5.4=0.1325$$

所以, $\hat{y}=0.1325+0.8625x$



例1续 找出例1人均生活费收入 $y$ 对时间 $x$ 的回归方程.

为简便计算,令:

$$x_i' = x_i - \bar{x} \qquad y_i' = y_i - \bar{y}$$

则 $$l_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i'^2$$

$$l_{xy} = \sum_{i=1}^{n} x_i' y_i'$$

| 时间$x_i$ | 收入$y_i$ | $x_i'$ | $y_i'$ | $x_i'^2$ | $y_i'^2$ | $x_i'y_i'$ |
|---|---|---|---|---|---|---|
| 1978 | 3.65 | –5.5 | –4.54 | 30.25 | 20.61 | 24.37 |
| 1979 | 4.15 | –4.5 | –4.04 | 20.25 | 16.32 | 18.18 |
| 1980 | 5.01 | –3.5 | –3.18 | 12.25 | 10.11 | 11.13 |
| 1981 | 5.14 | –2.5 | –3.05 | 6.25 | 9.30 | 7.625 |
| 1982 | 5.61 | –1.5 | –2.58 | 2.25 | 6.66 | 3.87 |
| 1983 | 5.91 | –0.5 | –2.28 | 0.25 | 5.20 | 1.14 |
| 1984 | 6.94 | 0.5 | –1.25 | 0.25 | 1.56 | –0.625 |
| 1985 | 9.08 | 1.5 | 0.89 | 2.25 | 0.79 | 1.395 |
| 1986 | 10.68 | 2.5 | 2.49 | 6.25 | 6.20 | 6.225 |
| 1987 | 11.82 | 3.5 | 3.63 | 12.25 | 13.18 | 12.705 |
| 1988 | 14.37 | 4.5 | 6.18 | 20.25 | 38.19 | 27.81 |
| 1989 | 15.97 | 5.5 | 7.78 | 30.25 | 60.53 | 42.79 |
| $\sum$ = 23802 | 98.33 | / | / | 143 | 188.65 | 157.145 |

$$n = 12,$$

$$\bar{x} = 1983.5,$$

$$\bar{y} = 8.194,$$

$$\hat{b} = \frac{l_{xy}}{l_{xx}} = \frac{\sum_{i=1}^{12} x_i' y_i'}{\sum_{i=1}^{12} x_i'^2}$$

$$= \frac{157.145}{143} = 1.099$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x} = -2171.67.$$

$$\hat{y} = -2171.67 + 1.099\,x.$$

Y　　X

例3 求北京城市居民人均食品支出对人均收入的回归方程.

| $x_i$ | $y_i$ | $x_i^2$ | $y_i^2$ | $x_i y_i$ |
|---|---|---|---|---|
| 3.65 | 2.11 | 13.32 | 4.45 | 7.70 |
| 4.15 | 2.37 | 17.22 | 5.62 | 9.84 |
| 5.01 | 2.72 | 25.10 | 7.34 | 13.58 |
| 5.14 | 2.95 | 26.42 | 8.70 | 15.16 |
| 5.61 | 3.18 | 31.47 | 10.11 | 17.84 |
| 5.91 | 3.38 | 34.93 | 11.42 | 19.98 |
| 6.94 | 3.79 | 48.16 | 14.36 | 26.30 |
| 9.08 | 4.67 | 82.45 | 21.81 | 42.40 |
| 10.68 | 5.43 | 114.06 | 29.48 | 57.93 |
| 11.82 | 6.05 | 139.71 | 36.60 | 71.51 |
| 14.37 | 7.43 | 206.50 | 55.20 | 106.77 |
| 15.97 | 8.41 | 225.04 | 70.73 | 134.31 |
| $\sum$ 98.33 | 52.48 | 994.38 | 275.82 | 523.38 |

$$\bar{x} = 8.194,\ \bar{y} = 4.373$$

$$l_{xx} = \sum_{i=1}^{12} x_i^2 - 12\bar{x}^2 = 188.648$$

$$l_{yy} = \sum_{i=1}^{n} y_i^2 - 12\bar{y}^2 = 46.31$$

$$l_{xy} = \sum_{i=1}^{n} x_i y_i - 12\bar{x}\,\bar{y} = 93.35$$

$$\hat{b} = l_{xy} / l_{xx} = 0.495$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x} = 0.317$$

$$\hat{y} = 0.317 + 0.495x$$

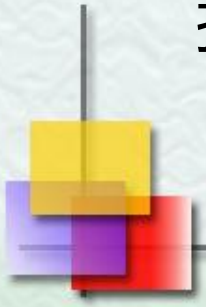
# 回归方程的显著性检验

在使用回归方程作进一步的分析以前，首先应对回归方程是否有意义进行判断。

如果$b=0$，那么不管$x$如何变化，$y$不随$x$的变化作线性变化，那么这时求得的一元线性回归方程就没有意义，称回归方程不显著。如果$b\neq 0$，$y$随$x$的变化作线性变化，称回归方程是显著的。

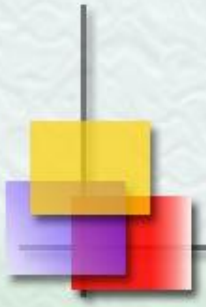
综上，对回归方程是否有意义作判断就是要作如下的显著性检验：$H_0$：$b=0$　　$vs$　　$H_1$：$b\neq 0$

拒绝$H_0$表示回归方程是显著的。



# 几种检验方法：

- T检验
- **F检验（平方和分解）**
- 相关系数的显著性检验



# 三 平方和分解公式及回归的效果检验



分析　回归估计方程 $\hat{y}=\hat{a}+\hat{b}x$

实际值 $y_i=a+bx_i+\varepsilon_i,\ i=1,\cdots,n.$

$$y_i=\hat{y}_i+(y_i-\hat{y}_i)$$

$\hat{y}_i=\hat{a}+\hat{b}x_i$　受 $x_i$ 影响

残差 $y_i-\hat{y}_i$　受其他因素影响

$$y_i-\bar{y}=(y_i-\hat{y}_i)+(\hat{y}_i-\bar{y})$$

$$l_{yy}=\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}[(y_i-\hat{y}_i)+(\hat{y}_i-\bar{y})]^2$$

$$=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y})=0$$

$$\text{而}\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y})=\sum_{i=1}^{n}(y_i-\hat{a}-\hat{b}x_i)(\hat{a}-\hat{b}x_i-\bar{y})$$

$$=(\hat{a}-\bar{y})\sum_{i=1}^{n}(y_i-\hat{a}-\hat{b}x_i)-\hat{b}\sum_{i=1}^{n}(y_i-\hat{a}-\hat{b}x_i)x_i$$

$$(\text{由正规方程组})=0$$

$$\begin{cases}\dfrac{\partial Q}{\partial a}=-2\sum\limits_{i=1}^{n}[y_i-(a+bx_i)]=0\\[2ex]\dfrac{\partial Q}{\partial b}=-2\sum\limits_{i=1}^{n}[y_i-(a+bx_i)]x_i=0\end{cases}$$

$$l_{yy}=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 = S_R+S_e=S_T.$$

回归平方和（regression sum of squares）

$$S_R=\sum_{i=1}^{n}[(\hat{a}+\hat{b}x_i)-(\hat{a}+\hat{b}\bar{x})]^2$$

回归方程过 $(\bar{x},\bar{y})$.

$$=\hat{b}^2\sum_{i=1}^{n}(x_i-\bar{x})^2 = \hat{b}^2 l_{xx}=\hat{b}l_{xy}=l_{xy}^2/l_{xx}$$

残差平方和（residual sum of squares）

$$S_e=l_{yy}-S_R=l_{yy}-\hat{b}l_{xy}$$

## 一元线性回归模型：

$$y_i = a + bx_i + \varepsilon_i, i = 1, \cdots, n.$$ $\varepsilon_i \sim N(0, \sigma^2)$且相互独立

判明 X 与 Y 之间是否有线性相关关系

线性回归显著性检验的 $F$ 检验法

检验 $H_0 : b = 0,\ H_1 : b \neq 0$

检验统计量 $F = \dfrac{S_R}{S_e/(n-2)} \sim F(1, n-2)\ (H_0 \text{ 为真})$

若 $H_0$ 为真, 则 $F$ 应该不大, 否则就拒绝 $H_0$.

拒绝域为: $\{F > F_{1-\alpha}(1, n-2)\}$ ★ 右侧检验

**进一步，有关$S_R$和$S_e$的分布，有如下定理。**

定理8.4.3　设$y_1, y_2, \ldots, y_n$相互独立，且

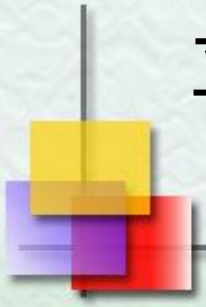
$$y_i \sim N(a+bx_i, \sigma^2), \quad i=1, \ldots, n,$$

则在上述记号下，有

(1) $S_e/\sigma^2 \sim \chi^2(n-2)$,

(2) 若$H_0$成立，则有$S_R/\sigma^2 \sim \chi^2(1)$

(3) $S_R$与$S_e$，$\bar{y}$ 独立（或$\hat{\beta}_1$ 与$S_e$，$\bar{y}$ 独立）。



例4 对例 2 所做的回归方程做 $F$ 检验, $\alpha = 0.05$.

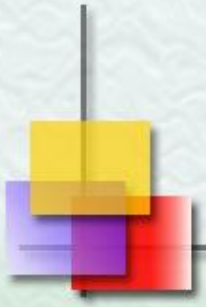
例2 合成纤维的强度与其拉伸倍数有关试验.

$n = 12, l_{xx} = 78.26, l_{xy} = 67.5.$ $\hat{y} = 0.1325 + 0.8625x.$

解: (1) $H_0 : b = 0$

(2)$H_0$为真时, $F = \dfrac{S_R}{S_e/(n-2)} \sim F(1, n-2)$

(3) 查表得$F_{1\text{-}0.05}(1,10) = 4.96$, 故拒绝域为[4. 96, ∞).



$S_T = S_R + S_e$

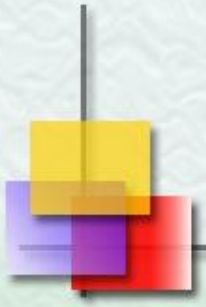
(4) 计算 $l_{yy} = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 = 335.63 - \frac{57.5^2}{12} \approx 60.1092$

则 $S_R = l_{xy}^2 / l_{xx} = 67.5^2 / 78.26 = 58.2194$

$S_e = l_{yy} - S_R = 60.1092 - 58.2194 = 1.8898$

$$f = \frac{S_R}{S_e/(n-2)} = \frac{58.2194}{1.8898/10} = 308.072 > 4.96$$

拒绝 $H_0$, 回归方程是显著的.



例5 对例 3所作的回归方程做 $F$ 检验, $\alpha = 0.05$.

(1) $H_0 : b = 0$

(2) $H_0$为真时， $F = \dfrac{S_R}{S_e/(n-2)} \sim F(1, n-2)$

(3) 给定$\alpha = 0.05$, 查 表得 $F_{0.95}(1,10) = 4.96$

$\alpha = 0.01$时, $F_{0.99}(1,10) = 10.04$

(4) $S_R = \hat{b} l_{xy} = 0.495 \times 93.35 = 46.21$

$$S_e = l_{yy} - S_R = 46.31 - 46.21 = 0.10$$

$$f = \frac{(n-2)S_R}{S_e} = 4621 > F_{0.99}(1,10)$$

拒绝$H_0$, 即回归效果是显著的.

**练习** 在合金钢强度的例1中，求回归方程，并作关于回归方程的显著性检验。

$$\hat{y} = 28.12 + 132.66x.$$

$$S_T = l_{yy} = 345.06 \qquad f_T = 11$$

$$S_R = \hat{b}^2 l_{xx} = 132.66^2 \times 0.0186 = 327.34, \qquad f_R = 1$$

$$S_e = S_T - S_R = 345.06 - 327.34 = 17.72 \qquad f_e = 10$$

| 来源 | 平方和 | 自由度 | 均方 | $F$比 |
|---|---|---|---|---|
| 回归 | $S_R$=327.34 | $f_A$=1 | $MS_A$=327.34 | 184.94 |
| 残差 | $S_e$=17. 72 | $f_e$=10 | $MS_e$= 1.77 | |
| 总和 | $S_T$=345.06 | $f_T$=11 | | |

若取$\alpha$=0.01，则$F_{0.99}(1,10)$ =10<$F$，因此在显著性水平0.01下回归方程是显著的。

残差平方和 $S_e = l_{yy} - S_R = l_{yy} - \hat{b} l_{xy}$

可以证明 $\dfrac{S_e}{\sigma^2} \sim \chi^2(n-2)$(定理8.5)

从而 $E(\dfrac{S_e}{\sigma^2}) = n-2, E(\dfrac{S_e}{n-2}) = \sigma^2.$

$\sigma^2$ 的无偏估计量为

$$\hat{\sigma}^2 = \frac{S_e}{n-2} = \frac{1}{n-2}\left(l_{yy} - \hat{b} l_{xy}\right).$$

# ***系数$b$的置信区间（补充）

$$Y = a + bx + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2).$$

$$\hat{b} \sim N(b, \sigma^2 / l_{xx}), \qquad \frac{(n-2)\hat{\sigma}^2}{\sigma^2} = \frac{S_e}{\sigma^2} \sim \chi^2(n-2).$$

并且$\hat{b}, S_e$相互独立,因此 $\dfrac{\hat{b} - b}{\hat{\sigma}} \sqrt{l_{xx}} \sim t(n-2).$

当回归效果显著时，对系数 $b$ 作区间估计.

系数 $b$ 的置信水平为 $1-\alpha$ 的置信区间为

$$\left( \hat{b} \pm t_{\alpha/2}(n-2) \times \frac{\hat{\sigma}}{\sqrt{l_{xx}}} \right).$$

作业： p.430-431 8.12，8.17