第5章　回归分析

5.1 一元线性回归

晁颖
数学与统计学院
Email: yingchao1993@xjtu.edu.cn

变量之间关系的分类

确定性关系（函数关系）

圆的面积与圆的半径：$A=\pi r^2$；

欧姆定律：$V=IR$；

牛顿第二定律：$F=ma$.

……

相关关系

父辈的身高与子辈的身高之间的关系；

人的年龄与血压之间的关系.

……

变量之间相关关系体现在两个方面：

(1) 变量之间存在某种联系；

(2) 自变量取某一值时，因变量的取值具有**随机性**；

回归分析是研究**随机因变量**与**可控自变量**之间**相关关系**的一种**统计方法**；主要分为：**一元回归分析(线性、非线性)**和**多元回归分析**

本节首先介绍**一元线性回归**.

主要内容

- 一元线性回归模型
- 一元线性回归模型的参数估计
- 参数估计量的概率分布
- 一元线性回归的假设检验
- 预测

主要内容

- 一元线性回归模型
- 一元线性回归模型的参数估计
- 参数估计量的概率分布
- 一元线性回归的假设检验
- 预测

5.1.1 一元线性回归模型

- 设随机变量y与可控变量x之间有相关关系，即当自变量x取定值时，y有一个确定的(条件)分布与之对应.

若y的数学期望存在，则它的值随x取值而定，因此是x的函数，记为$\mu(x)=E\,[y|x]$，称$\mu(x)$为y关于x的回归函数.

- 若y满足关系

$$y=\mu(x)+\epsilon$$

其中$\epsilon \sim N(0,\sigma^2)$，称该关系式为回归模型.

- 回归分析的基本任务：用试验数据来推断回归函数；

5.1.1 一元线性回归模型

- 若$\mu(x) = a + bx, a, b \in R$, 则回归模型如下：

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

其中a, b和σ^2都是与x无关的未知参数, 称该模型为一元线性回归模型，a, b称为回归系数.

- 当x取n个不全相同的数x_i时，对y依次作独立观测(试验)，得 n 对试验数据：

$$(x_1, y_1),\quad (x_2, y_2),\quad \cdots,\quad (x_n, y_n)$$

设它们满足关系式:

$$\left.\begin{array}{l} y_i = a + bx_i + \varepsilon_i\,,\quad i = 1, \cdots, n \\ \varepsilon_1, \cdots, \varepsilon_n \text{ 为} i.i.d \text{且} \varepsilon_1 \sim N(0,\ \sigma^2) \end{array}\right\}$$

其中a, b, σ^2为未知参数，称上述模型为线性模型.

- 针对线性模型，主要研究下列问题:
  - 用n对试验数据(x_i, y_i)，对a,b和σ^2作估计；
  - 对回归系数b作假设检验；
  - 对y作预测.

主要内容

- 一元线性回归模型
- 一元线性回归模型的参数估计
- 参数估计量的概率分布
- 一元线性回归的假设检验
- 预测

5.1.2 一元线性回归模型的参数估计

1. a, b的最小二乘估计

已知变量 x, y 的 n 对试验数据 $(x_i, y_i)(i=1,\cdots,n)$，其中 x_i不全相同. 作离差平方和

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

选择参数a, b使$Q(a, b)$达到最小，这种方法称为最小二乘法;

用最小二乘法求得参数的估计量称为参数的最小二乘估计.

最小二乘法主要思想：通过确定未知参数，来使得真实值和预测值的误差（也称残差）平方和最小.

取Q关于a、b的一阶偏导数，并令它们等于零

$$\left.\begin{aligned} \frac{\partial Q}{\partial a} &= -2\sum_{i=1}^{n}(y_i - a - bx_i) = 0 \\ \frac{\partial Q}{\partial b} &= -2\sum_{i=1}^{n}(y_i - a - bx_i)\,x_i = 0 \end{aligned}\right\}$$

经整理，得

$$\left.\begin{aligned} &na + n\bar{x}b = n\bar{y} \\ &n\bar{x}a + \sum_{i=1}^{n}x_i^2\,b = \sum_{i=1}^{n}x_iy_i \end{aligned}\right\}$$

其中$\bar{x} = \frac{1}{n}\sum_{i=1}^{n}x_i$，$\bar{y} = \frac{1}{n}\sum_{i=1}^{n}y_i$，上式称为正规方程组.

• 因为x_i不全相同，正规方程组的系数行列式

$$\begin{vmatrix} n & n\bar{x} \\ n\bar{x} & \sum_{i=1}^{n} x_i^2 \end{vmatrix} = n\left[\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right] = n\sum_{i=1}^{n}(x_i - \bar{x})^2 \neq 0$$

所以正规方程组有惟一的一组解.

• 解方程组得 a、b的最小二乘估计为

$$\left.\begin{aligned} \hat{b} &= \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ \hat{a} &= \bar{y} - \hat{b}\bar{x} \end{aligned}\right\}$$

其中$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

$$\left.\begin{array}{l} y_i = a + bx_i + \varepsilon_i\ ,\quad i = 1,\cdots,n \\ \varepsilon_1,\cdots,\varepsilon_n \text{ 为 } i.i.d \text{ 且 } \varepsilon_1 \sim N(0,\ \sigma^2) \end{array}\right\}$$

其中a,b,σ^2为未知参数，称上述模型为线性模型.

注:

(1)在用最小二乘法求参数a, b的估计时，并不需要$\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_n$为$i.i.d.$且$\varepsilon_1 \sim N(0,\sigma^2)$这一条件.

(2) 对线性模型，若用极大似然法求参数a, b的极大似然估计, 则$\hat{a}_L, \hat{b}_L$与a,b的最小二乘估计$\hat{a}$, $\hat{b}$相同(留作练习).

$$L(a,b)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(y_i-a-bx_i)^2}{2\sigma^2}}=(2\pi\sigma)^{-\frac{n}{2}}e^{-\sum_{i=1}^{n}\frac{(y_i-a-bx_i)^2}{2\sigma^2}}$$

$$\ln L(a,b)=-\frac{n}{2}\ln 2\pi\sigma-\sum_{i=1}^{n}\frac{(y_i-a-bx_i)^2}{2\sigma^2}$$

令 $\frac{\partial \ln L(a,b)}{\partial a}=0$, $\frac{\partial \ln L(a,b)}{\partial b}=0$. 得似然方程组:

$$\begin{cases}-2\sum_{i=1}^{n}(y_i-a-bx_i)=0\\ -2\sum_{i=1}^{n}(y_i-a-bx_i)x_i=0\end{cases}\Rightarrow \hat{b}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\quad \hat{a}=\bar{y}-\hat{b}\bar{x}$$

取$\hat{a}+\hat{b}x$作为回归函数$\mu(x)=a+bx$的估计，称$\hat{a}+\hat{b}x$为经验回归函数；

称方程

$$\hat{y}=\hat{a}+\hat{b}x$$

为 y 关于 x 的经验回归(直线)方程，称$\hat{a}$, $\hat{b}$为经验回归系数.

将 $\hat{a}=\bar{y}-\hat{b}\bar{x}$代入，经验回归方程可改写为

$$\hat{y}=\bar{y}+\hat{b}(x-\bar{x})$$

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

2. σ^2的估计

因为$\sigma^2=D(\varepsilon)=E(\varepsilon^2)$，所以用$\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2$作为$\sigma^2$的矩估计.

由于$\varepsilon_i = y_i - a - bx_i$是未知的，以$\hat{a},\hat{b}$替换未知参数 a, b，

得到σ^2的形式上的矩估计

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{a} - \hat{b}x_i)^2$$

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

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

因为

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

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

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

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

所以

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

例5.1.1 某建材实验室在作陶粒混凝土强度实验中，考察每立方米混凝土的水泥用量 x 对 28 天后的混凝土抗压强度 y 的影响. 测得如下数据

水泥用量 x/kg	150	160	170	180	190	200	210	220	230	240	250	260
抗压强度 y/MPa	5.58	5.72	6.04	6.34	6.68	6.99	7.27	7.59	7.86	8.10	8.47	8.80

求 y 对 x 的经验回归函数，并计算 σ^2 的估计值.

解 由表所给的数据，得

$$n = 12\,, \quad \sum_{i=1}^{n} x_i = 2\,460\,, \quad \sum_{i=1}^{n} y_i = 85.44$$

$$\sum_{i=1}^{n} x_i^2 = 518600\,, \quad \sum_{i=1}^{n} y_i^2 = 621.064\,,$$

$$\sum_{i=1}^{n} x_i y_i = 17941.2$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 205\,, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = 7.12$$

$$\hat{b} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = 0.0298$$

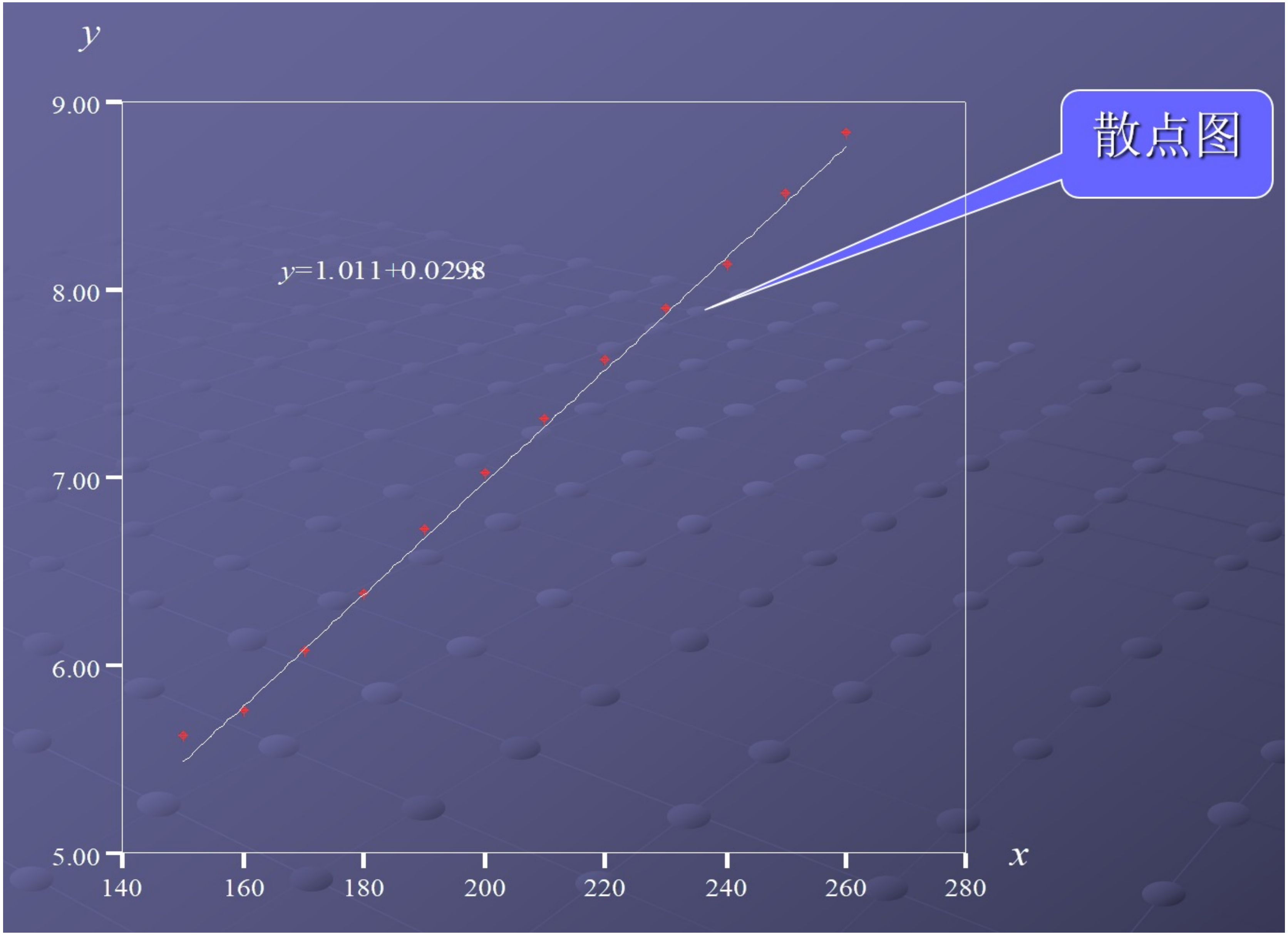
$$\hat{a} = \bar{y} - \hat{b}\bar{x} = 1.011$$

所以 y 对 x 的经验回归函数为

$$\hat{a}+\hat{b}x=1.011+0.0298x$$

σ^2的估计值为

$$\hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2-\frac{\hat{b}^2}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\approx 0.0012$$


y
9.00
8.00
7.00
6.00
5.00
y=1.011+0.0298
散点图
140
160
180
200
220
240
260
280
x

主要内容

- 一元线性回归模型
- 一元线性回归模型的参数估计
- **参数估计量的概率分布**
- 一元线性回归的假设检验
- 预测

5.1.3 参数估计量的概率分布

1. $\hat{b}$的分布

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

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

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

因为

$$\hat{b} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{l_{xy}}{l_{xx}} = \sum_{i=1}^{n} c_i y_i, \text{ 其中} c_i = \frac{(x_i-\bar{x})}{l_{xx}}$$

又因为$y_1,\cdots,y_n$相互独立且 $y_i \sim N(a+bx_i,\sigma^2)$，所以$\hat{b}$服从正态分布.

$$E\hat{b}=\sum_{i=1}^{n}c_iEy_i=\sum_{i=1}^{n}c_i(a+bx_i)=b\sum_{i=1}^{n}\frac{(x_i-\bar{x})x_i}{l_{xx}}=b\frac{l_{xx}}{l_{xx}}=b$$

$$D\hat{b}=\sum_{i=1}^{n}c_i^2Dy_i=\sigma^2\sum_{i=1}^{n}\frac{(x_i-\bar{x})^2}{l_{xx}^2}=\sigma^2\frac{l_{xx}}{l_{xx}^2}=\frac{\sigma^2}{l_{xx}}$$

故$\hat{b}$的分布为：

$$\hat{b}\sim N\left(b,\ \frac{\sigma^2}{l_{xx}}\right)$$

由此可见，$\hat{b}$是 b 的无偏估计.

2. $\hat{a}$ 的分布

$$\hat{a} = \bar{y} - \hat{b}\bar{x} = \sum_{i=1}^{n}\left[\frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{l_{xx}}\right] y_i$$

因此$\hat{a}$服从正态分布.

$$E\hat{a} = E\bar{y} - E\hat{b}\bar{x} = \frac{1}{n}\sum_{i=1}^{n}(a + bx_i) - b\bar{x} = a$$

$$D\hat{a} = \sum_{i=1}^{n}\left[\frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{l_{xx}}\right]^2 Dy_i = \left[\frac{1}{n} + \frac{\bar{x}^2}{l_{xx}}\right]\sigma^2$$

$$\sum_{i=1}^{n}\left[\frac{1}{n^2} - \frac{2}{n}\frac{(x_i-\bar{x})\bar{x}}{l_{xx}} + \frac{(x_i-\bar{x})^2(\bar{x})^2}{l_{xx}^2}\right]\sigma^2 = \left(\frac{1}{n} + \frac{l_{xx}}{l_{xx}^2}\bar{x}^2\right)\sigma^2$$

故

$$\hat{a} \sim N\left(a, \left[\frac{1}{n} + \frac{\bar{x}^2}{l_{xx}}\right]\sigma^2\right)$$

$\hat{a}$同样是 a 的无偏估计.

3. $\hat{\sigma}^2$的分布

记

$$\hat{y}_i = \hat{a} + \hat{b}x_i \ (i = 1, \ 2, \ \cdots, \ n)$$

$y_i - \hat{y}_i$称为x_i处的**残差.**

平方和

$$Q_e = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \hat{a} - \hat{b}x_i)^2$$

称为**残差平方和.**

$Q_e = Q(\hat{a}, \hat{b}) = Q_{min}$

$$Q_e = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \hat{a} - \hat{b}x_i)^2$$

定理5.1.1 对于线性模型(5.1.2)，有

(1) $\dfrac{Q_e}{\sigma^2} \sim \chi^2(n-2)$

(2) $\bar{y}$, $\hat{b}$, Q_e相互独立

由定理5.1.1可知，$E(Q_e/\sigma^2)=n-2$，令

$$\hat{\sigma}^{*2}=\frac{Q_e}{n-2}=\frac{1}{n-2}\sum_{i=1}^{n}(y_i-\hat{a}-\hat{b}x_i)^2$$

则$E(\hat{\sigma}^{*2})=\sigma^2$, 故$\hat{\sigma}^{*2}$为$\sigma^2$的无偏估计.

- σ^2的形式上的矩估计

$$\hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{a}-\hat{b}x_i)^2 = \frac{n-2}{n}\cdot\frac{Q_e}{n-2} = \frac{n-2}{n}\cdot\hat{\sigma}^{*2}$$

$$E(\hat{\sigma}^2)=E\left(\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{a}-\hat{b}x_i)^2\right)=E\left(\frac{Q_e}{n}\right)=\frac{n-2}{n}\sigma^2$$

故$\hat{\sigma}^2$ 是σ^2的渐近无偏估计.

主要内容

- 一元线性回归模型
- 一元线性回归模型的参数估计
- 参数估计量的概率分布
- 一元线性回归的假设检验
- 预测

5.1.4 一元线性回归的假设检验

1.T检验

H_0成立时 $\frac{\hat{b}}{\sigma}\sqrt{l_{xx}} \sim N(0,1)$

对线性模型(5.1.2)提出如下假设

$$H_0: b = 0$$

$\hat{b} \sim N\left(b, \frac{\sigma^2}{l_{xx}}\right)$

定理5.1.1

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

(2) $\bar{y}$, $\hat{b}$, Q_e相互独立

取

$$t = \frac{\hat{b}}{\hat{\sigma}^*}\sqrt{l_{xx}}$$

$$\frac{\frac{\hat{b}}{\sigma}\sqrt{l_{xx}}}{\sqrt{\frac{\hat{\sigma}^{*2}}{\sigma^2}}} = \frac{\frac{\hat{b}}{\sigma}\sqrt{l_{xx}}}{\sqrt{\frac{Q_e}{\sigma^2}\cdot\frac{1}{n-2}}} \sim N(0,1)$$

作为检验统计量；

当H_0成立时，由$\hat{b}$的分布及定理5.1.1知，$t \sim t(n-2)$.

故对给定的显著水平α，假设H_0的拒绝域为

$$W = \{\, |t| \geq t_{\alpha/2}(n-2) \,\}$$

注

(1) 理论上讲，要检验一元线性回归模型是否成立，需从以下几方面着手：

- 其一，当x取不同值时，检验 y是否服从正态分布，分布是否依赖于x，方差是否相同；
- 其二，当x取不同值时，检验$E(y\mid x)$是否为 x的线性函数；
- 最后，当x取不全相同的数值$x_1,\cdots,x_n$时，检验相应的$y_1,\cdots,y_n$

是否相互独立．

由于完成这一组检验工作相当繁琐且十分困难，因而把它简化成只检验回归系数b是否为零的这一项假设，实际上这是一步相当大的简化．

(2) 当H_0 ($b=0$)不成立时，认为所求的经验回归方程有意义；

当$H_0(b=0)$成立时，则认为所求的经验回归方程无实用价值；

如果出现后一种情况，那么就要从以下几方面找原因：

1) 影响 y 的数值除了 x 之外，可能还有其它变量；

2) y 与 x 有关系，但不是线性的；

3) y 与 x 无关.

例5.1.2 检验例5.1.1中的线性回归是否显著$(\alpha = 0.05)$?

解 已知

$$\hat{b} = 0.0298, \quad \hat{\sigma}^{*2} = \frac{n}{n-2}\hat{\sigma}^2 = 0.0014$$

$$l_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = 14300,$$

$$t = \frac{\hat{b}}{\hat{\sigma}^*}\sqrt{l_{xx}}$$

故

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{a} - \hat{b}x_i)^2$$

$$t = \frac{0.029\ 8}{\sqrt{0.001\ 4}}\sqrt{14\ 300} \approx 95.2402$$

查表得$t_{0.05/2}(12-2) = t_{0.025}(10) = 2.2281$,

因为$|t| = 95.2402 > 2.2281$,所以拒绝H_0,认为y与x之间线性回归关系是显著的.

判断经验回归方程是否具有使用价值，除了使用假设检验方法之外，还可以用经验相关系数或样本相关系数来衡量。

由等式

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

$$Q_e = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \hat{a} - \hat{b}x_i)^2$$

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

可见

(1) 当 $Q_e = 0$，即 $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 0$ 时，说明试验数据 (x_i, y_i) $(i = 1, 2, \cdots, n)$ 都在一条直线上 $(y = \hat{a} + \hat{b}x)$，此时可以认为变量 y 与 x 之间存在确定的线性函数关系．

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

(2) 当$\hat{b}^2\sum_{i=1}^{n}(x_i-\bar{x})^2=0$，即$\hat{b}^2=0$时，$Q_e=\sum_{i=1}^{n}(y_i-\bar{y})^2$达到最大值，此时，$x$ 的变化对 y 无影响，说明 y与 x之间无线性相关关系． Q_e可用来描述变量间线性相关的密切程度

将Q_e变形为

$$Q_e=\sum_{i=1}^{n}(y_i-\bar{y})^2-\hat{b}^2\sum_{i=1}^{n}(x_i-\bar{x})^2$$
$$=l_{yy}\left(1-\frac{\hat{b}^2 l_{xx}}{l_{yy}}\right)$$
$$=l_{yy}\left(1-\frac{l_{xy}^2}{l_{xx}l_{yy}}\right)$$

$\hat{b}=\frac{l_{xy}}{l_{xx}}$

记

$$r \triangleq \frac{l_{xy}}{\sqrt{l_{xx}l_{yy}}} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

不难验证

$$0 < Q_e < \sum_{i=1}^{n}(y_i-\bar{y})^2 \Leftrightarrow 0 < |r| < 1\ ;$$

$$Q_e = 0 \Leftrightarrow |r| = 1\ ;$$

$$Q_e = \sum_{i=1}^{n}(y_i-\bar{y})^2 \Leftrightarrow r = 0\ .$$

由此可见，r同样可以描述变量间线性相关的密切程度，并且它是无量纲的 .

在线性相关的研究中，通常都用 r 来衡量变量间线性相关的密切程度，r 称为经验相关系数或样本相关系数.

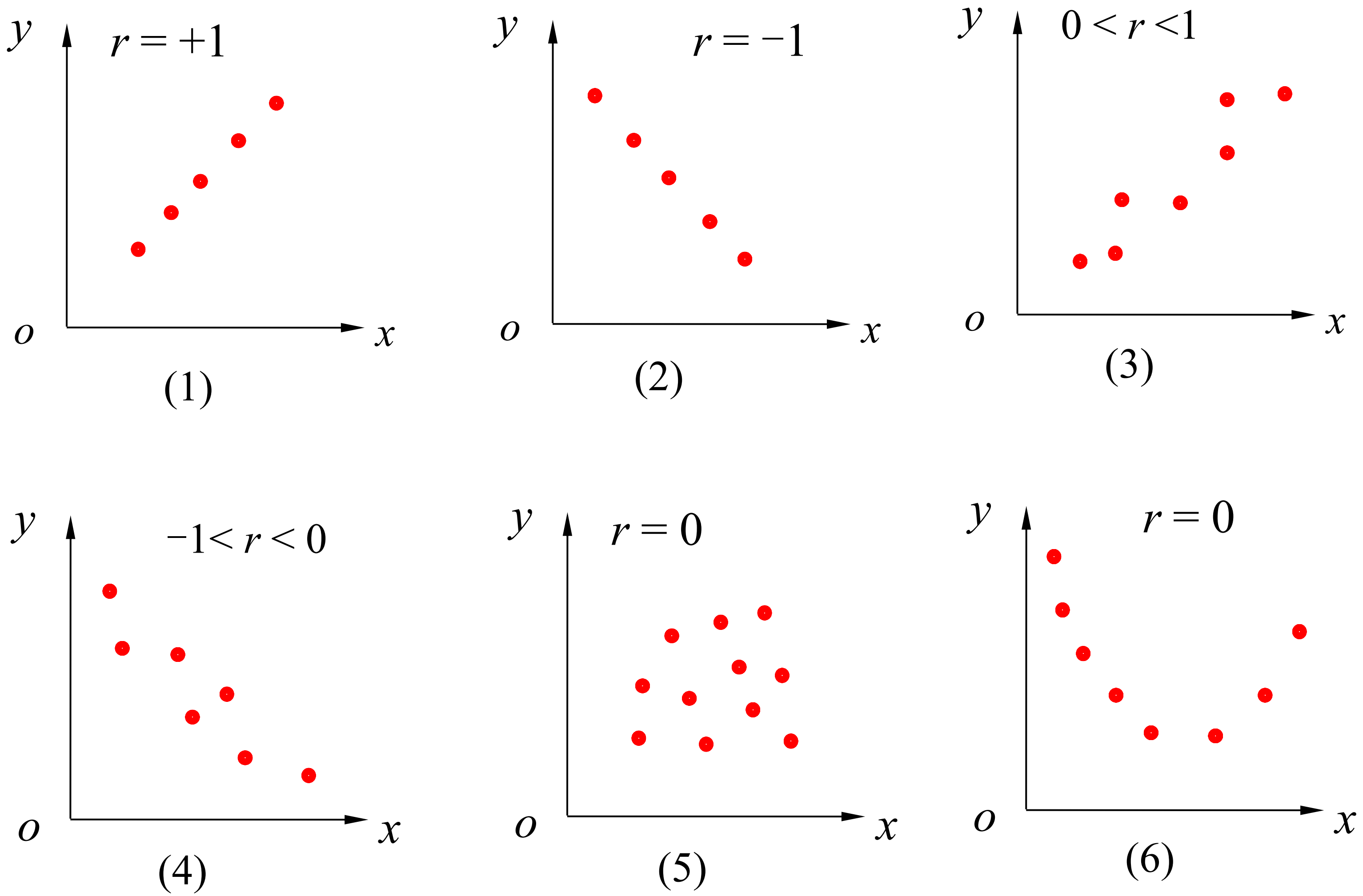


图 5.1.2

利用例5.1.1中的数据, 可以算得y与x之间的经验相关系数为

$$r=\frac{426}{\sqrt{14\ 300}\sqrt{12.7132}}=0.9991$$

$|r|$接近于1，这表明所配的经验回归方程是有意义的 .

主要内容

- 一元线性回归模型
- 一元线性回归模型的参数估计
- 参数估计量的概率分布
- 一元线性回归的假设检验
- 预测

5.1.5 预测

经过检验，如果认为线性回归方程是可信的，而且拟合得很好，则可以用它来预测.

预测可分为点预测和区间预测，即当 $x=x_0$时, 求因变量y的预测值和求 y 的具有给定的置信度的预测区间.

点预测：当 $x=x_0$ 时，就取 x_0 处的经验回归值

$$\hat{y}_0=\hat{a}+\hat{b}x_0=\bar{y}+\hat{b}(x_0-\bar{x})$$

作为 y 的预测值；

下面探讨求 y 的具有给定的置信度的预测区间问题.

当 $x=x_0$ 时， y 的真值 y_0 为

$$y_0=a+bx_0+\varepsilon_0，\ \varepsilon_0 \sim N(0,\sigma^2)$$

假定 y_0 与 $y_1,\cdots,y_n$ 相互独立 .

考虑 $y_0-\hat{y}_0$，由于 $y_0,\hat{y}_0$ 相互独立，并且都服从正态分布，因而 $y_0-\hat{y}_0$ 也服从正态分布. 由

$$E(y_0-\hat{y}_0)=Ey_0-E\hat{y}_0=a+bx_0-(a+bx_0)=0$$

$$\begin{aligned}
D(y_0-\hat{y}_0) &= D(y_0)+D(\hat{y}_0)\\
&= D(y_0)+D[\bar{y}+\hat{b}(x_0-\bar{x})]\\
&= D(y_0)+D(\bar{y})+D[\hat{b}(x_0-\bar{x})]\\
&= \sigma^2\left[1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{l_{xx}}\right]
\end{aligned}$$

$\hat{y}_0=\bar{y}+\hat{b}(x_0-\bar{x})$

$\bar{y}\sim N(a+b\bar{x},\frac{\sigma^2}{n})$ $\bar{y}$与$\hat{b}$独立

$\hat{b}\sim N(b,\frac{\sigma^2}{l_{xx}})$

$\Rightarrow \hat{y}_0\sim N(a+bx_0,\sigma^2[\frac{1}{n}+\frac{(x_0-\bar{x})^2}{l_{xx}}])$

可知

$$y_0-\hat{y}_0 \sim N\left(0,\ \sigma^2\left[1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{l_{xx}}\right]\right)$$

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

$$\frac{\dfrac{y_0-\hat{y}_0}{\sigma\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{l_{xx}}}} \sim N(0,1)}{\sqrt{\dfrac{(n-2)\hat{\sigma}^{*2}}{\sigma^2}\Big/(n-2)} \sim \chi^2(n-2)}$$

又$y_0-\hat{y}_0$与$\hat{\sigma}^{*2}$相互独立,

$$\downarrow \frac{Q_e}{n-2}$$

故

$$\frac{y_0-\hat{y}_0}{\hat{\sigma}^*\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{l_{xx}}}} \sim t(n-2)$$

给定置信度$1-\alpha$，查 $t(n-2)$分布表，得$t_{\alpha/2}(n-2)$，使得

$$P\left\{\frac{|y_0-\hat{y}_0|}{\hat{\sigma}^*\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{l_{xx}}}}<t_{\alpha/2}(n-2)\right\}=1-\alpha$$

即

$$P\{\hat{y}_0-\delta(x_0)<y_0<\hat{y}_0+\delta(x_0)\}=1-\alpha$$

其中

$$\delta(x_0)=t_{\alpha/2}(n-2)\hat{\sigma}*\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{l_{xx}}}$$

故 y_0的置信度为$1-\alpha$的预测区间是

$$(\hat{y}_0-\delta(x_0),\ \hat{y}_0+\delta(x_0))$$

- 让$\delta(x_0)$中的x_0变动，并记$\hat{y}=\hat{a}+\hat{b}x$, 则在x处，y的预测下限为$y_1(x)=\hat{y}-\delta(x)$, 预测上限为$y_2(x)=\hat{y}+\delta(x)$

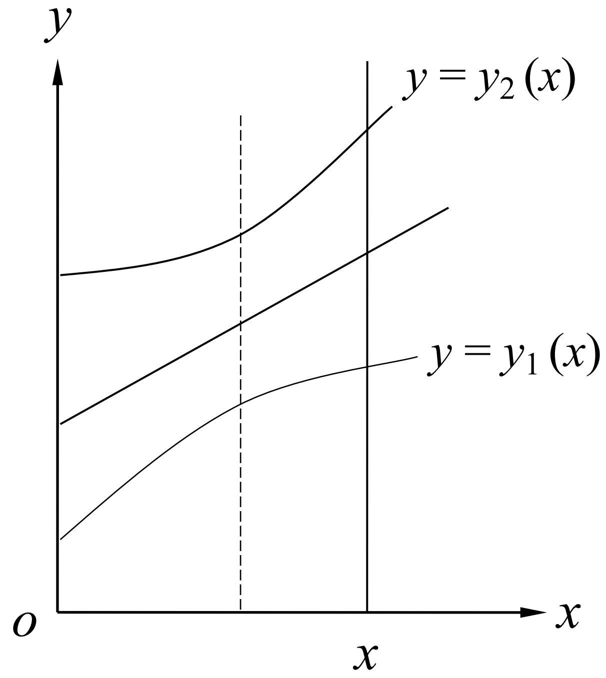


图 5.1.3

- 当n很大且 x 在$\bar{x}$附近时，有

$$\delta(x) \approx u_{\alpha/2}\hat{\sigma}^*$$

- 此时，y的预测下限 $y_1(x)$ 和预测上限 $y_2(x)$ 近似为

$$y_1(x) \approx \hat{a} + \hat{b}x - u_{\alpha/2}\hat{\sigma}^*$$

$$y_2(x) \approx \hat{a} + \hat{b}x + u_{\alpha/2}\hat{\sigma}^*$$

例5.1.3 在例5.1.1中，对x=195处的y进行预测$(1-\alpha=95\%)$.

解　将$x=195$代入经验回归函数，得 y 的点预值

$\hat{y}=1.011+0.0298\times195=6.822,$

由$n=12$及$\delta(x)$的表达式，可算得

$$\delta(195)=t_{\frac{\alpha}{2}}(n-2)\hat{\sigma}*\sqrt{1+\frac{1}{n}+\frac{(195-\bar{x})^2}{l_{xx}}}$$

$$=2.228\ 1\times\sqrt{0.001\ 4}\times\sqrt{1+\frac{1}{12}+\frac{(195-205)^2}{14\ 300}}$$

$$=0.087$$

故 y 的置信度为95%的预测区间是

$$(\hat{y}\mp\delta(195))=(6.735,\ \ 6.909)$$

本讲小结

- **一般回归模型：** $y=E(y\,|\,x)+\varepsilon$，其中ε为具有零均值、有限方差σ^2的随机变量．
- **一元线性回归模型**——$y=a+bx+\varepsilon$， $\varepsilon\sim N(0,\sigma^2)$

对上述模型，研究了下列问题：

- 用n对试验数据(x_i,y_i)，对a，b和σ^2作估计

$$\hat{b}=\frac{l_{xy}}{l_{xx}},\quad \hat{a}=\bar{y}-\hat{b}\bar{x},\quad \hat{\sigma}^2=\frac{1}{n}l_{yy}-\frac{\hat{b}^2}{n}l_{xx}$$

- 对$b=0$作假设检验

检验统计量为$t=\frac{\hat{b}}{\hat{\sigma}*}\sqrt{l_{xx}}$， 拒绝域为$W=\left\{|t|\geq t_{\frac{\alpha}{2}}(n-2)\right\}$

- 对y作预测

点预测 $\hat{y}_0=\hat{a}+\hat{b}x_0=\bar{y}+\hat{b}(x_0-\bar{x})$

区间预测$(\hat{y}_0\mp\delta(x_0))$，$\delta(x_0)=t_{\alpha/2}(n-2)\hat{\sigma}*\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{l_{xx}}}$