第五章 回归分析

§5.1 一元线性回归

§5.2 假设检验及预测

理学院

- 本次课教学目的：
  - 变量之间的关系以及回归分析中，找出回归方程式，检验方程有效与否，当方程有效时对Y的值作预测与控制。
- 重点难点：
  - 回归分析的参数估计
  - 参数检验及预测

§5.1 回归与相关的概念

在现实问题中处于同一个过程中的一些变量往往是相互依赖和相互制约的，各种变量间的关系大致可分为两类：

- 确定性关系 ——函数关系；
- 非确定性关系 ——相关关系；

一类是完全确定性的关系，又称函数关系，可以用精确的数学表达式来表示，即当变量x的值取定后，变量y有唯一确定的值与之对应。

如长方形的面积(S) 与 长(a)和宽(b)的关系： $S=ab$。它们之间的关系是确定性的，只要知道了其中两个变量的值就可以精确地计算出另一个变量的值，这类变量间的关系称为函数关系。

另一类是非确定性关系，不能用精确的数学公式来表示，当变量x的值取定后，y有若干种可能取值。

如人的身高与体重的关系，食品价格与需求量的关系等等，这些变量间都存在着十分密切的关系，但不能由一个或几个变量的值精确地求出另一个变量的值。统计学中把这些变量间的关系称为相关关系，把存在相关关系的变量称为相关变量。

相关关系表现为这些变量之间有一定的依赖关系，但这种关系并不完全确定，它们之间的关系不能精确地用函数表示出来。

在一定范围内，对一个变量的任意数值（X_i），虽然没有另一个变量的确定数值y_i与之对应，但是却有一个特定y_i的条件概率分布与之对应，这种变量的不确定关系，称为相关关系。

研究一个随机变量与一个(或几个) 变量之间的相关关系的统计方法称为回归分析. 只有一个自变量的回归分析叫做一元回归分析。多于一个自变量的回归分析叫做多元回归分析.

回归分析主要包括三方面内容:

(1) 提供建立有相关关系的变量之间的数学关系式(通常称之为经验公式)的一般方法;

(2) 判别所建立的经验公式是否有效，并从影响随机变量的诸变量中判别哪些变量的影响是显著的，哪些是不显著的;

(3) 利用所得到的经验公式进行预测和控制.

5.2.1数学模型

对于两个相关变量，一个变量用x表示，另一个变量用y表示，如果通过试验或调查获得两个变量的n对观测值：

(x_1, y_1)，(x_2, y_2)，……，(x_n, y_n).

为了直观地看出x和y间的变化趋势，可将每一对观测值在平面直角坐标系中描点，作出散点图（见图5-1）。

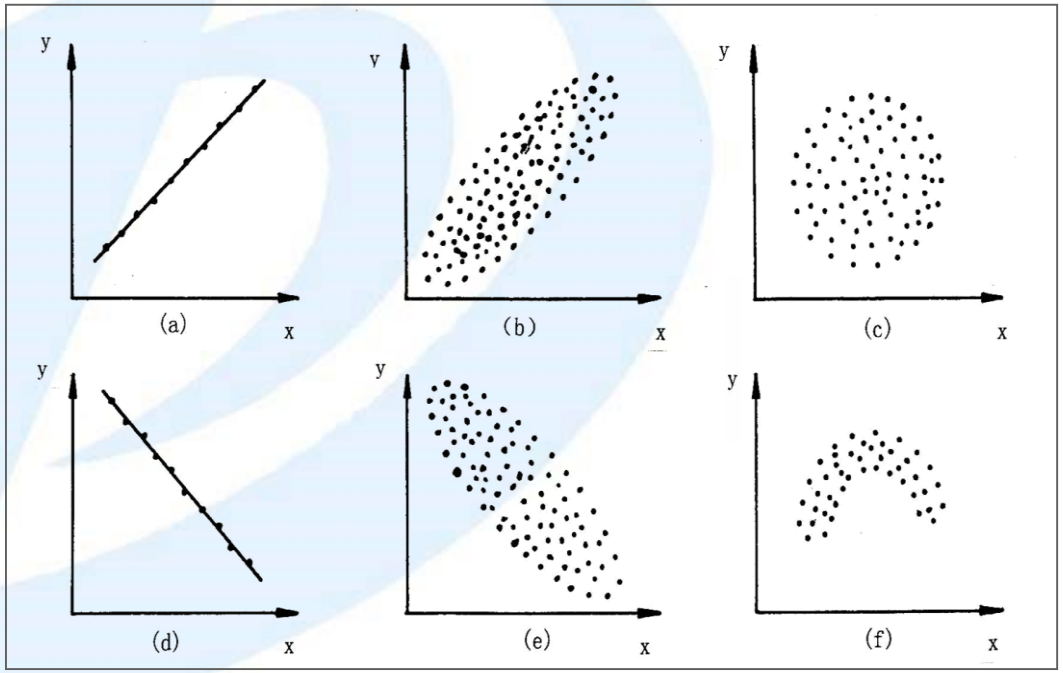


图5-1 x与y的关系散点图

由散点图（图5-1）可以看出：

① 两个变量间有关或无关；若有关，两个变量间关系类型，是直线型还是曲线型；

② 两个变量间直线关系的性质（是正相关还是负相关）和程度（是相关密切还是不密切）；

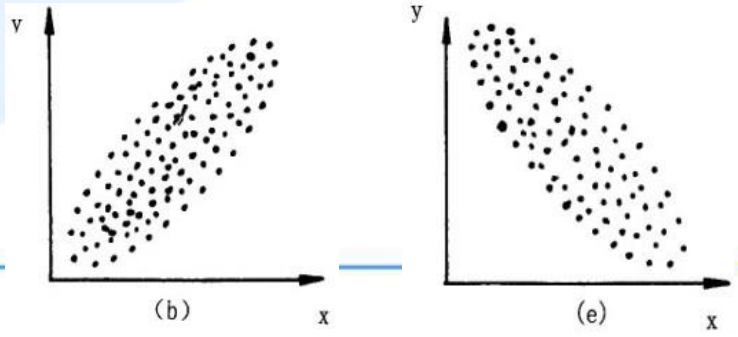
散点图可直观地、定性地表示了两个变量之间的关系。为了探讨它们之间的规律性，还必须根据观测值将其内在关系定量地表达出来。

若两个相关变量y(因变量)与x(自变量)间的关系是直线关系，那么，根据n对观测值所描出的散点图，如图5-1（b）和图5-1（e）所示。

由于因变量y的实际观测值总是带有随机误差，因而因变量y的实际观测值y_i可用自变量x的实际观测值x_i表示为：

$$\begin{cases} y_i = \alpha + \beta x_i + \varepsilon_i \\ \varepsilon_i \sim N(0, \sigma^2) \end{cases} \quad i = 1,2,\cdots,n, \text{各}\varepsilon_i\text{独立}.$$

其中α, β, σ^2是与x无关的未知参数,这就是直线回归的数学模型,则称之为一元线性回归模型.



例1 某工厂在分析产量与成本关系时，选取10个生产小组作样本，收得数据下如表：

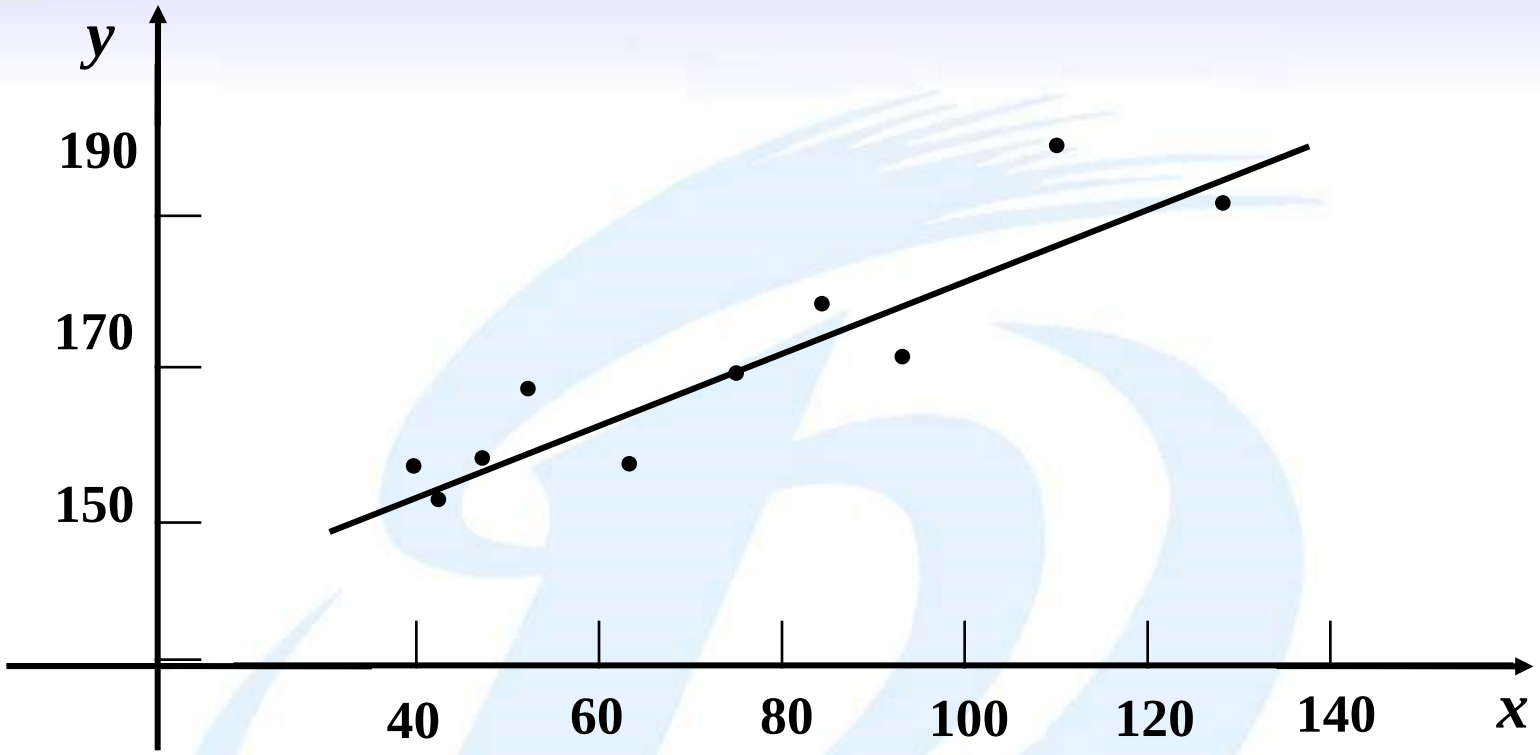
产量x(千件)	40	42	48	55	65	79	88	100	120	140
成本y(千元)	150	140	152	160	150	162	175	165	190	185

求y与x之间的关系，这里x是自变量，y是随机变量.

解： 把每对观测值$(x_i, y_i)(i=1, 2, \ldots, 10)$看成是平面直角坐标系中的点. 并描出相应的点.这个图称为散点图，如下图1. 由散点图，可以观察散点的分布规律，

这些散点分布在一条直线附近，但不全在一条直线上。

因此可以认为y 与x 之间的关系由两部分组成,一部分是由于x的变化引起的y的线性变化部分，记为$a+bx$，另一部分是由其它一切随机因素引起的，记为ε ，即



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

其中a, b 是与x 无关的未知参数,ε 是不可观测的随机变量,

且假定$E\varepsilon = 0, D\varepsilon = s^2$ (未知).

思考

在一元线性回归分析中，需要讨论以下三个问题：

(1) 如何由样本(x_i, y_i) $(i=1,2,\ldots,n)$求出a, b, s^2的估计$\hat{a}, \hat{b}, \hat{\sigma}^2$，并建立方程

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

称之为y关于x的一元线性经验回归方程，也简称为一元线性回归方程. $\hat{a}, \hat{b}$称为回归系数。

(2) 如何对所建立的回归方程进行可信度检验.

(3) 若回归方程是可信的，如何用它进行预测和控制.

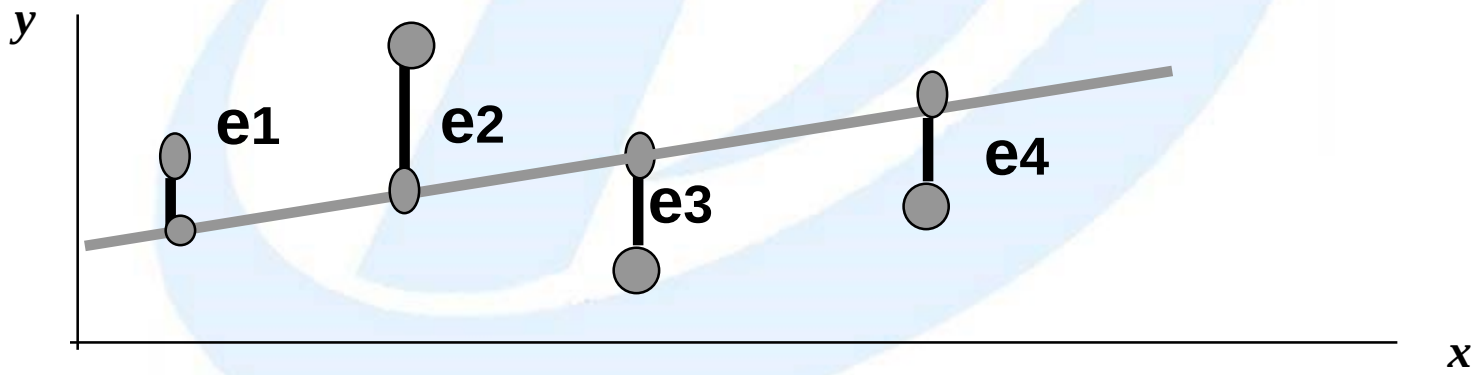
§5.2 参数检验及其预测

§5.2.1 未知参数的估计及性质

建立样本线性回归方程的方法——最小二乘法

实际观察值y_i与样本回归线上的点的距离的平方和最小

$$\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2=\sum_{i=1}^{n}e_i^2$$ 最小



最小二乘法 (Least squares estimate)

根据上面的分析，根据已知的散点 $(x_i, y_i)(i=1,2,\cdots,n)$ 我们可以得到一个回归函数 $y=\beta_0+\beta_1 x$，其中 β_0, β_1 待定。

令 $\varepsilon_i = y_i - (\beta_0 + \beta_1 x_i)$，这里 ε_i 表示当 $x = x_i$ 时，

的观测值 y_i 与 直线 $y=\beta_0+\beta_1 x$ 上的对应纵坐标 $y=\beta_0+\beta_1 x_i$

的偏差。这样，各个散点与直线的总的偏差的平方和为

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^{\,2} = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2$$

1. 未知参数的估计

现在我们用最小二乘法来估计未知参数α,β，记

$$Q(\alpha,\beta)=\sum_{i=1}^{n}(y_i-(\alpha+\beta x_i))^2$$

选择α,β的估计$\hat{\alpha},\hat{\beta}$使得

$$Q(\hat{\alpha},\hat{\beta})=\min_{\alpha,\beta}Q(\alpha,\beta)$$

令

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

$$\Rightarrow \begin{cases} n\alpha + \beta\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ \alpha\sum_{i=1}^{n} x_i + \beta\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \end{cases} \Rightarrow \begin{cases} \hat{\beta} = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \end{cases}$$

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

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

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

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

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

$$\begin{cases} \hat{\beta} = \dfrac{l_{xy}}{l_{xx}}, \\ \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}. \end{cases}$$

于是得y 对x 的经验回归方程 $\hat{y} = \hat{\alpha} + \hat{\beta}x.$

以上是一元线性回归的最小二乘法参数估计,简称为LS估计.

由经验回归方程，我们有

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

它的图形称为回归直线。回归直线必过点 $(0, \hat{\alpha})$和散点图的几何重心$(\bar{x}, \bar{y})$。

把x_i代入回归方程, 有

$$\hat{y}_i = \bar{y} + \hat{\beta}(x_i - \bar{x}).$$

$\hat{y}_i$是 $\tilde{y}_i = \alpha + \beta x_i$ 的估计，通常称 $\hat{y}_i$为回归值.

求解例1

例1 某工厂在分析产量与成本关系时，选取10个生产小组作样本，收得数据下如表：

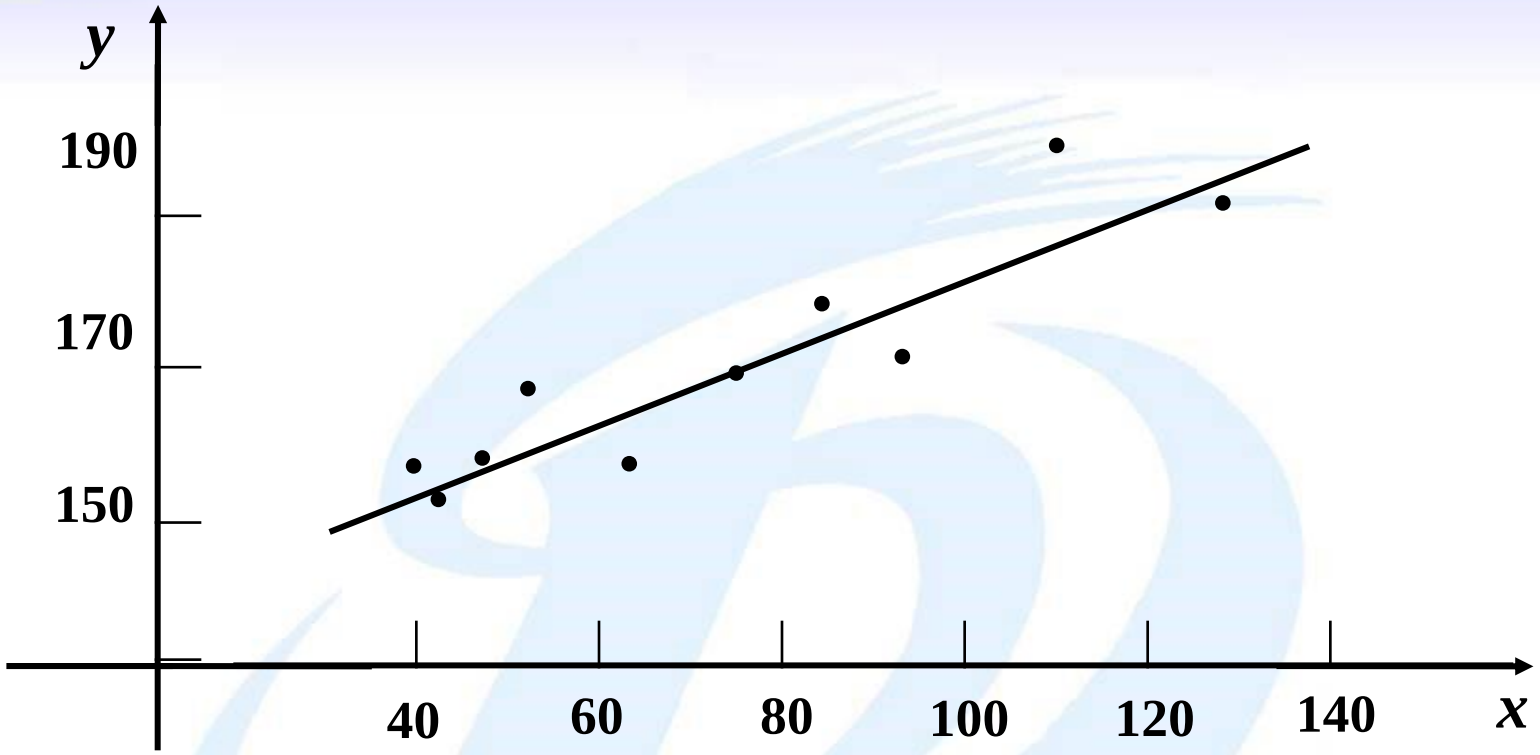
产量x(千件)	40	42	48	55	65	79	88	100	120	140
成本y(千元)	150	140	152	160	150	162	175	165	190	185

求y与x之间的关系，这里x是自变量，y是随机变量.

解： 把每对观测值$(x_i, y_i)(i=1, 2, \ldots, 10)$看成是平面直角坐标系中的点. 并描出相应的点 . 这个图称为散点图，如下图1. 由散点图，可以观察散点的分布规律，

这些散点分布在一条直线附近，但不全在一条直线上。

因此可以认为y 与x 之间的关系由两部分组成,一部分是由于x的变化引起的y的线性变化部分，记为$a+bx$，另一部分是由其它一切随机因素引起的，记为ε ，即



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

其中a, b 是与x 无关的未知参数,ε 是不可观测的随机变量,

且假定$E\varepsilon = 0, D\varepsilon = s^2$ (未知).

$$\begin{cases}\hat{\beta}=\dfrac{l_{xy}}{l_{xx}},\\ \hat{\alpha}=\bar{y}-\hat{\beta}\bar{x}.\end{cases}$$

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

例2 求例1中的经验回归方程

(40,150)(42,140)(48,152)(55,160)
(65,150)(79,162)(88,175)(100,165)
(120,190)(140,185)

解： 由例1中的几个样本点坐标，利用公式计算得

$$\sum x_i=777 \quad \sum x_i^2=70903 \quad \sum y_i=1629$$

$$\sum y_i^2=267723 \quad \sum x_i y_i=131124 \quad n=10$$

$$\bar{x}=77.7 \qquad \bar{y}=162.9$$

$$l_{xx}=\sum x_i^2-\frac{1}{n}(\sum x_i)^2=70903-\frac{1}{10}777^2=10530.1$$

$$l_{yy}=\sum y_i^2-\frac{1}{n}(\sum y_i)^2=267723-\frac{1}{10}1629^2=2358.9$$

$$l_{xy}=\sum x_i y_i-\frac{1}{n}(\sum x_i)(\sum y_i)=4550.7$$

由离差平方和可求得参数

$$\hat{\beta}=\frac{l_{xy}}{l_{xx}}=\frac{4550.7}{10530.1}=0.4322$$

$$\hat{\alpha}=\bar{y}-\hat{\beta}\bar{x}=169.2-0.4322\times 77.7=129.3181$$

故所求经验回归方程为

$$\hat{y}=129.3181+0.4322x$$

这里$\hat{\beta}=0.4322$，表示产量每增加1千件，成本平均增加0.4322千元。

回归参数实际意义

2. 最小二乘估计的性质

定理1

(1) $\hat{\beta} \sim N(\beta, \frac{\sigma^2}{l_{xx}})$　　(2) $\hat{\alpha} \sim N(\alpha, (\frac{1}{n} + \frac{\bar{x}^2}{l_{xx}})\sigma^2)$

(3) $cov(\hat{\alpha}, \hat{\beta}) = -\frac{\bar{x}}{l_{xx}}\sigma^2$　　(4) $cov(\bar{y}, \hat{\beta}) = 0$

(5) $\hat{y} = \hat{\alpha} + \hat{\beta}x \sim N\left(\alpha + \beta x, \left(\frac{1}{n} + \frac{(x-\bar{x})^2}{l_{xx}}\right)\sigma^2\right)$

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

$$\begin{cases} y_i = \alpha + \beta x_i + \varepsilon_i \\ \varepsilon_i \sim N(0, \sigma^2) \end{cases} \quad i = 1,2,\cdots,n, \text{各}\varepsilon_i\text{独立.}$$

记 $l_{xy} = \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^{n} (X_i - \bar{X}) Y_i$

$$\hat{\beta} = \frac{l_{xy}}{l_{xx}} = \frac{1}{l_{xx}} \sum_{i=1}^{n} (X_i - \bar{X}) Y_i$$

$\hat{\beta}$是随机变量y_1、y_2、…、y_n的线性组合，而$y_i \sim N(\alpha + \beta x_i, \sigma^2)$，$x_i(i=1,2,\ldots,n)$具有非随机性，故$\hat{\beta}$也服从正态分布。

$$E\hat{\beta} = E\left(\frac{1}{l_{xx}} \sum_{i=1}^{n} (x_i - \bar{x})\, y_i\right)$$

$$= \frac{1}{l_{xx}} \sum_{i=1}^{n} (x_i - \bar{x})\, E(y_i)$$

$$= \frac{1}{l_{xx}} \sum_{i=1}^{n} (x_i - \bar{x})\, (\alpha + \beta x_i)$$

$$=\frac{\beta}{l_{xx}}\sum_{i=1}^{n}(x_i-\overline{x})\,x_i$$

$$=\frac{\beta}{l_{xx}}\sum_{i=1}^{n}(x_i-\overline{x})\,(x_i-\overline{x})$$

$$=\frac{\beta}{l_{xx}}\cdot l_{xx}=\beta$$

$$D\widehat{\beta}=D\left(\frac{1}{l_{xx}}\sum_{i=1}^{n}(x_i-\overline{x})\,y_i\right)$$

$$=\frac{1}{l_{xx}}\sum_{i=1}^{n}(x_i-\overline{x})\,D(y_i)$$

$$=\left(\frac{1}{l_{xx}}\right)^2\sum_{i=1}^{n}(x_i-\overline{x})^2\,D(y_i)$$

$$=\left(\frac{1}{l_{xx}}\right)^2\sum_{i=1}^{n}(x_i-\overline{x})^2\,\sigma^2=\frac{\sigma^2}{l_{xx}}$$

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

(2) $\hat{\alpha} \sim N(\alpha, (\frac{1}{n} + \frac{\bar{x}^2}{l_{xx}})\sigma^2)$

$$E\hat{\alpha} = \mathrm{E}(\bar{y} - \hat{\beta}\bar{x})$$
$$= \mathrm{E}(\frac{1}{n}\sum_{i=1}^{n}(\alpha + \beta x_i)) - E(\hat{\beta}\bar{x})$$
$$= \mathrm{E}\left(\frac{1}{n}(n\alpha + n\beta\bar{x})\right) - \beta\bar{x}$$
$$= \alpha + \beta\bar{x} - \beta\bar{x} = \alpha$$

(3) $cov(\hat{\alpha}, \hat{\beta}) = -\frac{\bar{x}}{l_{xx}}\sigma^2$

(4) $cov(\bar{y}, \hat{\beta}) = 0$

$$cov(\bar{y}, \hat{\beta}) = cov(\hat{\alpha} + \hat{\beta}\bar{x}, \hat{\beta})$$
$$= cov(\hat{\alpha}, \hat{\beta}) + cov(\hat{\beta}, \hat{\beta})\bar{x}$$

$$\begin{cases} y_i = \alpha + \beta x_i + \varepsilon_i \\ \varepsilon_i \sim N(0, \sigma^2) \end{cases} \quad i = 1,2,\cdots,n, \text{各}\varepsilon_i\text{独立}.$$

3. σ^2的无偏估计

记 $$S_e = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - \hat{a} - \hat{\beta} x_i)^2$$

$$= \sum (y_i - \bar{y} - \hat{\beta}(x_i - \bar{x}))^2.$$

称S_e为残差平方和或剩余平方和.

在一元线性回归程型下，有

定理 2 (1) $ES_e = (n-2)\sigma^2$ (2) $\dfrac{S_e}{n-2}$是 σ^2 的无偏估计。

(3) $\dfrac{S_e}{\sigma^2} \sim \chi^2(n-2)$, 且$S_e$与 $\bar{y}, \hat{\beta}$相互独立.

4. σ^2的计算

$$S_e = \sum (y_i - \bar{y} - \hat{\beta}(x_i - \bar{x}))^2.$$

为了计算无偏估计 $\hat{\sigma}^2$，将

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

代入式(5.33)，得

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{n-2}\sum_{i=1}^{n}[y_i - \bar{y} - \hat{b}(x_i - \bar{x})]^2 \\ &= \frac{n}{n-2}\left[\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 - \frac{2\hat{b}}{n}\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}) + \frac{\hat{b}^2}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right] \\ &= \frac{n}{n-2}\left[\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 - 2\hat{b}(\overline{xy} - \bar{x}\,\bar{y}) + \hat{b}^2 m_x^2\right]\end{aligned}$$

再注意到式(5.18)的第一式，有

$$\overline{xy} - \bar{x}\,\bar{y} = \hat{b}(\overline{x^2} - \bar{x}^2) = \hat{b}m_x^2$$

并引进计算器容易获得其值的二阶中心矩的记号

$$m_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2, \quad m_y = \sqrt{m_y^2} \tag{5.37}$$

则有

$$\hat{\sigma}^2 = \frac{n}{n-2}\left[m_y^2 - \hat{b}^2 m_x^2\right] \tag{5.38}$$

式(5.38)中的 $\hat{\sigma}^2$ 与 $\hat{a}, \hat{b}$ 也可利用表 5.1 及计算器的统计功能方便地获得.

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

5. 一元线性回归参数的计算

表 5.1　回归参数计算表

i	x_i	y_i	x_iy_i
1	x_1	y_1	x_1y_1
2	x_2	y_2	x_2y_2
⋮	⋮	⋮	⋮
n	x_n	y_n	x_ny_n
计算器计算	$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$	$\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i$	$\overline{xy}=\frac{1}{n}\sum_{i=1}^{n}x_iy_i$
	$m_x^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$	$m_y^2=\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2$	
回归参数	$\hat{b}=\frac{\overline{xy}-\bar{x}\,\bar{y}}{m_x^2}$	$\hat{a}=\bar{y}-\hat{b}\,\bar{x}$	$\hat{\sigma}^2=\frac{n}{n-2}\left(m_y^2-\hat{b}^2m_x^2\right)$

5. 一元线性回归参数的计算

例 5.2.1 用切削机床进行金属品加工时，为了适当地调整机床，应该测定刀具的磨损速度. 在一定时间(如每隔一小时)测量刀具的厚度，测得结果如下：

时间 x_i(h)	刀具厚度 y_i(cm)	时间 x_i(h)	刀具厚度 y_i(cm)	时间 x_i(h)	刀具厚度 y_i(cm)
0	30.0	6	27.5	12	26.1
1	29.1	7	27.2	13	25.7
2	28.4	8	27.0	14	25.3
3	28.1	9	26.8	15	24.8
4	28.0	10	26.5	16	24.0
5	27.7	11	26.3		

试求刀具厚度 Y 关于切削时间 x 的线性回归方程，并计算 σ^2 的估计值.

解 填表并利用计算器计算得下表

i	x_i	y_i	x_iy_i
1	0	30.0	0.0
2	1	29.1	29.1
3	2	28.4	56.8
4	3	28.1	84.3
5	4	28.0	112.0
6	5	27.7	138.5

续表

i	x_i	y_i	x_iy_i
7	6	27.5	165.0
8	7	27.2	190.4
9	8	27.0	216.0
10	9	26.8	241.2
11	10	26.5	265.0
12	11	26.3	289.3
13	12	26.1	313.2
14	13	25.7	334.1
15	14	25.3	354.2
16	15	24.8	372.0
17	16	24.0	384.0
计算器计算	$\overline{x}=8$	$\overline{y}=26.97$	$\overline{xy}=208.54$
	$m_x=4.899$	$m_y=1.502$	
回归参数	$\hat{b}=-0.301$	$\hat{a}=29.38$	$\hat{\sigma}^2=0.0924$

表中回归参数的计算如下：

$$\hat{b}=\frac{\overline{xy}-\overline{x}\,\overline{y}}{m_x^2}=\frac{208.54-8\times 26.97}{4.899^2}=-0.301$$

$$\hat{a}=\overline{y}-\hat{b}\,\overline{x}=26.97-(-0.301)\times 8=29.38$$

$$\hat{\sigma}^2=\frac{17}{17-2}(m_y^2-\hat{b}^2m_x^2)=\frac{17}{15}(1.502^2-0.301^2\times 4.899^2)=0.0924$$

故所求回归直线方程为

$$\hat{y}=29.38-0.301x$$

线性回归模型为

$$Y=29.38-0.301x+\varepsilon,\quad \varepsilon\sim N(0,0.0924)$$

其回归直线如图 5.2 所示.

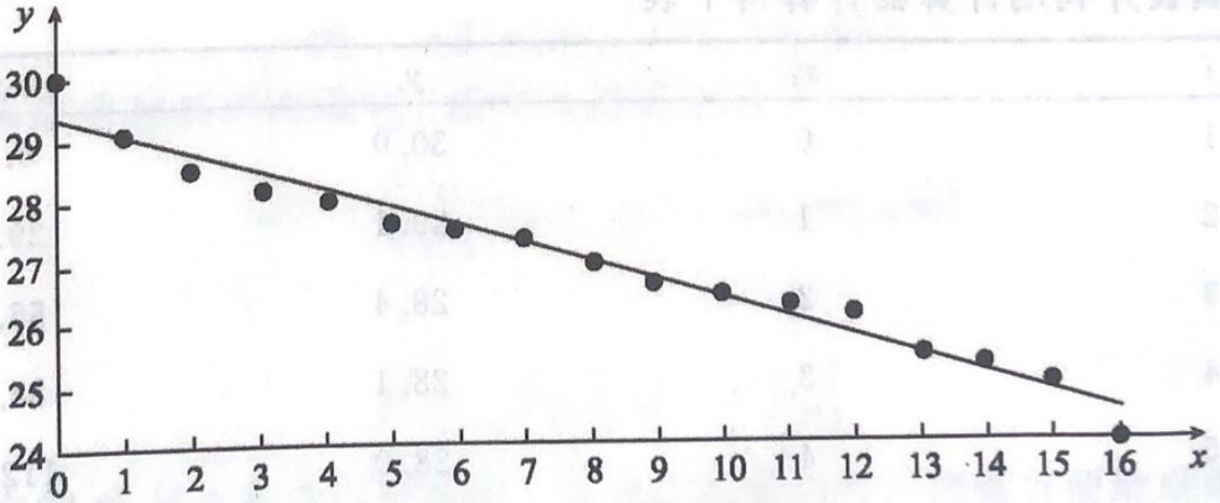


图 5.2　回归直线图

§5.2.2 回归效果的显著性检验

无论x, y是否具有线性相关关系，都能得到样本回归直线. 因此需要对得到的样本回归直线进行检验.

当$|b|$越大，y 随x 的变化就越大，b =0时，y 和x 间就不存在线性关系. 因此可设计统计假设为：

$$H_0: b=0,\ H_1: b\neq 0,$$

下面介绍三种不同的检验方法，它们本质上是相同的.

定理3

当$H_0: b=0$成立时，有 $\dfrac{S_R}{\sigma^2}\sim\chi^2(1)$;

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

且S_R与S_e是独立的。

$$\begin{cases} \hat{\beta} = \dfrac{l_{xy}}{l_{xx}}, \\ \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}. \end{cases}$$

$$\hat{y} = \hat{\alpha} + \hat{\beta}x.$$

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

$$S_R = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \hat{\beta}^2 l_{xx} = \frac{l_{xy}^2}{l_{xx}}$$

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

$$\frac{S_R}{l_{yy}} = \frac{\hat{\beta}^2 l_{xx}}{l_{yy}} = \left(\frac{l_{xy}}{\sqrt{l_{xx} \cdot l_{yy}}}\right)^2 = r^2$$

$$S_E = S_T - S_R = l_{yy} - r^2 \cdot l_{yy} = (1 - r^2) l_{yy}$$

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

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

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

(1) F 检验法

$$F=\frac{S_R}{S_e/(n-2)}\sim F(1,n-2),$$

原假设：$H_0:b=0$,

备择假设：$H_1:b\neq 0$.

由分位数定义

$$P\left\{\frac{S_R(n-2)}{S_e}\geq F_\alpha(1,n-2)\right\}=\alpha$$

对于样本观测值$(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$,算得$F$ 的观测值于是得线性回归效果的显著性的F检验法则：

- 若$F\geq F_a(1,\ n-2)$，则拒绝H_0，此时称为线性回归效果显著，即y 与x之间存在显著的线性相关关系；
- 若$F< F_a(1,\ n-2)$，则接受H_0，此时称为线性回归效果不显著，即y 与x之间没有显著的线性相关关系.

回归效果不显著的原因可能有以下几种：

①影响y取值的除x外，还有其它不可忽略的变量；

②y与x的关系不是线性的，而是其它非线性关系；

③y与x之间根本不存在任何关系。

需要进一步分析原因，分别处理。

$$\frac{S_R}{l_{yy}}=\frac{\hat{\beta}^2 l_{xx}}{l_{yy}}=\left(\frac{l_{xy}}{\sqrt{l_{xx}\cdot l_{yy}}}\right)^2=r^2$$

$$S_E=(1-r^2)l_{yy}$$

(2) r 检验法

因为 $$F=\frac{S_R}{S_e/(n-2)}=\frac{r^2(n-2)}{1-r^2}$$

$F\geq F_a(1,\ n-2)$等价于

$$|r|\geq\left[\frac{n-2}{F_\alpha(1,\ n-2)}+1\right]^{-\frac{1}{2}}\triangleq r_{n-2,\alpha}$$

这种利用相关系数的临界值表来检验方法称为r检验法，对于给定的显著性水平α，它的检验法则为：

若$|r| \geq r_{n-2,\alpha}$，则拒绝H_0, 否则接受H_0.

由此可见，F 检验法和r 检验法的拒绝域是相等的.

对于较大的n，r 检验可用下述变换方法。

Fisher证明了：当y 与x不相关时，对于较大的n($n \geq 50$)

$$Z = \frac{1}{2}\ln\frac{1+r}{1-r}$$

近似服从正态分布 $N(0, \frac{1}{n-3})$, $\sqrt{n-3}Z \overset{近似}{\sim} N(0,1)$

从而近似地有

$$P\{\sqrt{n-3}\,|\,Z\,| \geq u_{\alpha/2}\} = \alpha$$

若$\sqrt{n-3}\,|\,Z\,| \geq u_{\alpha/2}$，则拒绝$H_0$, 否则接受$H_0$.($n \geq 50$)

(3) t 检验法

由 $\hat{\beta} \sim N(\beta, \frac{\sigma^2}{l_{xx}})$ 有 $\frac{\hat{\beta}-\beta}{\sigma}\sqrt{l_{xx}} \sim N(0,1)$

$$t = \frac{\hat{\beta}-\beta}{\sqrt{S_e/(n-2)}}\sqrt{l_{xx}} \sim t(n-2)$$

在原假设成立时，有

$$t = \frac{\hat{\beta}}{\sqrt{S_e/(n-2)}}\sqrt{l_{xx}} \sim t(n-2)$$

对于给定的显著性水平a，若由样本观测值算得的

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

则拒绝H_0，否则接受H_0.

例2回归方程：$\hat{y}=129.3181+0.4322x$

例3　分别用F检验法，r检验法和t检验法，检验例2的线性回归效果是否显著($a=0.05$).

解：$l_{xx}=10530.1,\ l_{yy}=2358.9,\ \hat{\beta}=0.4322$

(1) F 检验法：

$$S_R=\hat{\beta}^2 l_{xx}=0.4322^2\times 10530.1=1966.9894$$

$$S_e=l_{yy}-S_R=2358.9-1969.9894=1391.9106$$

$$F=\frac{S_R}{S_e}(n-2)=\frac{1966.9894}{391.9106}\times 8=40.1520$$

查F分布表，得$F_\alpha(1,n-2)=F_{0.05}(1,8)=5.32$

因为$F=40.1520>5.32=F_{0.05}(1,8)$，故拒绝$H_0$，即认为线性回归效果显著。

$$S_E=S_T-S_R=l_{yy}-r^2\cdot l_{yy}$$

(2) r 检验法：

$$r^2 = \frac{S_R}{l_{yy}} = \frac{1966.9894}{2358.9} = 0.8339$$

$$|r| = 0.9132$$

对 $n = 10, \alpha = 0.05$, 查相关系数临界值表，得

$$r_{n-2,\alpha} = r_{8,\,0.05} = 0.6319 < 0.9132 = |r|$$

故拒绝H_0.

(3) t 检验法：

$$S = \sqrt{\frac{S_e}{n-2}} = \sqrt{\frac{391.9106}{8}} = 6.9992$$

$$t = \frac{\hat{\beta}\sqrt{l_{xx}}}{S} = \frac{0.4322 \times \sqrt{10530.1}}{6.9992} = 6.3365$$

对 $n = 10, \alpha = 0.05$, 查临界值表，得

$$t_{\alpha/2}(n-2) = t_{0.025}(8) = 2.3060 < 6.3365 = t$$

故拒绝假设H_0.

$$t = \frac{\hat{\beta}}{\sqrt{S_e/(n-2)}}\sqrt{l_{xx}} \sim t(n-2)$$

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

§5.2.3 回归系数的置信区间

因为 $\hat{\alpha} \sim N(\alpha, (\frac{1}{n} + \frac{\bar{x}^2}{l_{xx}})\sigma^2)$, 所以

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

$$\frac{\hat{\alpha} - \alpha}{\sigma\sqrt{1/n + \bar{x}^2/l_{xx}}} \sim N(0,1)$$

又 $\frac{S_e}{\sigma^2} \sim \chi^2(n-2)$, 且$\hat{\alpha}$与$S_e$相互独立,

$$t = \frac{\hat{\alpha} - \alpha}{\sigma\sqrt{1/n + \bar{x}^2/l_{xx}}\sqrt{S_e/\sigma^2(n-2)}} \sim t(n-2),$$

由此得回归系数α 的 1− α 的置信区间为

$$P\left\{|t| < t_{\frac{\alpha}{2}}\right\} = 1 - \alpha$$

$$\left(\hat{\alpha} \pm \sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{l_{xx}}\right)\frac{S_e}{n-2}}\, t_{\alpha/2}(n-2)\right)$$

同样可得β 的1—α置信区间为

$$t = \frac{\hat{\beta} - \beta}{\sqrt{S_e/(n-2)}}\sqrt{l_{xx}} \sim t(n-2)$$

$$\left(\hat{\beta} \pm \sqrt{\frac{S_e}{(n-2)l_{xx}}}\, t_{\alpha/2}(n-2)\right)$$

例2回归方程：$\hat{y}=129.3181+0.4322x$

例4 求例2中回归系数a,b的95%置信区间。

解： $\hat{\alpha}=129.3181,\ \hat{\beta}=0.4322,\ l_{xx}=105301,$

$\sqrt{S_e/(n-2)}=6.9992,\ \bar{x}=77.7$

$$\left(\hat{\alpha}\pm\sqrt{\left(\frac{1}{n}+\frac{\bar{x}^2}{l_{xx}}\right)\frac{S_e}{n-2}}\,t_{1-\alpha/2}(n-2)\right)$$

$$=(116.0740,\ 142.5622)$$

$$\left(\hat{\beta}\pm\sqrt{\frac{S_e}{(n-2)l_{xx}}}\,t_{1-\alpha/2}(n-2)\right)$$

$$=(0.2749,\ 0.5893)$$

§5.2.4 回归直线预测

所谓预测，就是对于给定的值$x=x_0$，预测对应的y_0的估计值及y_0的取值范围，即给出y_0的区间估计。

(1) 点预测(点估计)

在获得经验回归方程后，对给定的$x=x_0$，要预测y_0的取值, 自然会将x_0代入经验回归方程，得

$$\hat{y}_0 = \hat{\alpha} + \hat{\beta}x_0$$

并用$\hat{y}_0$作为y_0的预测值(即估计值). 这种作法是合理的, 因为

$$y_0 = \alpha + \beta x_0 + \varepsilon_0$$

$$E\hat{y}_0 = E(\hat{\alpha} + \hat{\beta}x_0) = \alpha + \beta x_0 = Ey_0$$

且是Ey_0的无偏估计。

(2) 预测区间

所谓预测区间，就是对给定的x_0，求y_0的$1-a$置信区间。

定理4 在一元线性回归模型中，设y_0，y_1，…，y_n相互独立，则

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

P106定理5.3.1

根据定理4及t分布的分位数，得

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

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

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

所以y_0的$1-a$置信区间为

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

由于x_0的任意性，若把x_0换为x，则相应的可写为

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

其中 $\delta(x)=t_{\alpha/2}(n-2)\sqrt{\frac{S_e}{n-2}}\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{l_{xx}}}$

这时，我们有两条曲线：

$$\hat{y}^*=\hat{y}+\delta(x)\qquad \hat{y}_*=\hat{y}-\delta(x)$$

他们之间的部分就是含回归方程 $\hat{y}=\hat{\alpha}+\hat{\beta}x$ 的带域，且在 $x=x_0$处最窄。

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

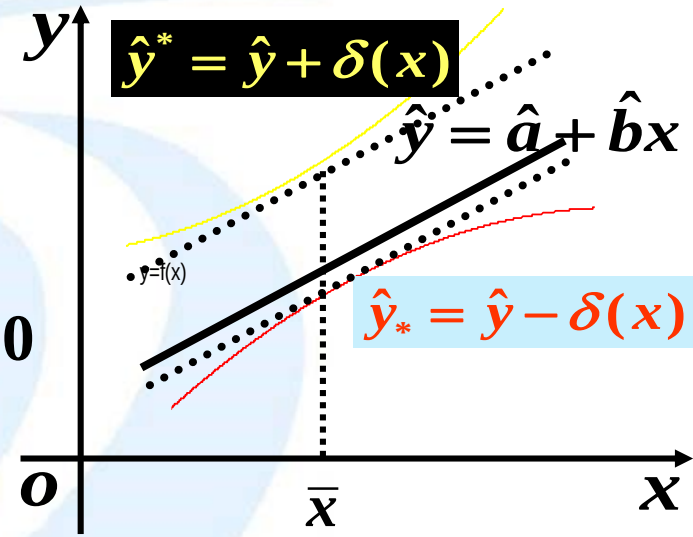
例5 在例1中，取x_0=90，求y_0的预测值及95%的预测区间。

解： 根据例1中所求的线性回归直线，如下图，可得

$$\hat{y} = \hat{\alpha} + \hat{\beta}x = 129.3181 + 0.4322x$$

当x_0=90，y_0的预测值为

$$\hat{y}_0 = \hat{\alpha} + \hat{\beta}x_0 = 129.3181 + 0.4322 \times 90$$
$$= 168.2161$$



$\bar{x} = 77.7,\ l_{xx} = 10530.1,\ \sqrt{S_e/(n-2)} = 6.9992,\ t_{0.025}(8) = 2.3060$

y_0的95%预测区间为

$$(\hat{y}_0 - \delta(x_0),\ \hat{y}_0 + \delta(x_0)) = (\hat{y}_0 \pm t_{\alpha/2}(n-2)\sqrt{\frac{S_e}{n-2}}\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{l_{xx}}})$$

$$= (151.1780,\ 185.2542)$$

特别，当n很大且x在$\bar{x}$附近取值时，我们有

$$t_{\alpha/2}(n-2) \approx u_{\alpha/2},\ \sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{l_{xx}}} \approx 1$$

于是y的$1-\alpha$预测区间为

$$(\hat{y}-\delta(x),\ \hat{y}+\delta(x)) = (\hat{y} \pm u_{\alpha/2}\sqrt{\frac{S_e}{n-2}})$$

这时的预测带域是平行于回归直线的两条平行线之间的部分，如上图虚线所示。这样作使预测工作得到大大的简化。

Southwest Petroleum University
西南石油大学

Thank you!