变量之间的关系一般可分为确定性与非确定性两种:

(1) 确定性关系是指变量之间的关系可以用函数关系来表达;

(2) 而非确定性的关系即所谓的相关关系. 回归分析是研究相关关系的一种数学工具.

一、回归的含义

1. 自变量与因变量的定义:

【补例1】人的体重Y与身高X之间存在着相关关系.

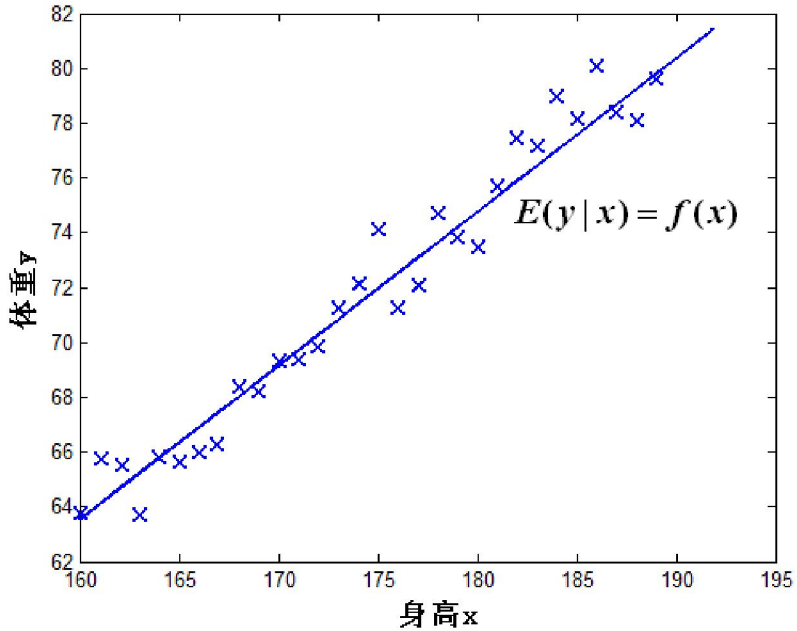

图4-2 三十个男子的身高和体重

由图4-2可作如下假设:

$$y=\beta_0+\beta_1 x+\varepsilon$$

设随机变量y与x之间存在着某种相关关系.这里:

①x是可控或可精确观察的变量,故把x看成普通变量,称x为自变量(预报变量、回归变量);

②y称为因变量(响应变量).

若对每一确定x值,$E(y|x)=f(x)$存在,则称$f(x)$为y关于x的回归函数.

【补例2】人的体重Y与身高X、性别S、地区C之间存在着相关关系.

因此还可作如下假设:

$$y=f(x,s,c)+\varepsilon$$

2. 回归分析的任务与目的:

任务:研究自变量之变动对因变量之变动的影响程度;

目的:根据自变量的变化来估计或预测因变量的变化情况.

3. 回归分析的内容:

(1) 确定回归模型;

(2) 根据样本估计未知参数;

(3) 检验回归方程与各自变量的显著性;

(4) 利用自变量的值来估计和预测因变量.

二、本节内容与思路

$$\left.\begin{array}{l}\begin{cases} y=\beta_0+\beta_1 x+\varepsilon \\ E\varepsilon=0, D\varepsilon=\sigma^2 \end{cases} \xRightarrow[(x_i,y_i)]{n\text{对独立观测值}} \text{⓪}\begin{cases}\hat{\beta}_0\\ \hat{\beta}_1\\ \hat{\sigma}^2\end{cases} \\ (\beta_0,\beta_1,\sigma^2\text{未知}) \\ \qquad\qquad\qquad \varepsilon \sim N(0,\sigma^2) \end{array}\right\} \Rightarrow \begin{cases} \text{①检验回归方程是否显著} \\ \quad (H_0:\beta_1=0;\ H_1:\beta_1\neq 0) \\ \text{②对}\beta_0,\beta_1,\sigma^2\text{进行区间估计} \\ \text{③预测与控制}\begin{cases} \xrightarrow{x=x_0} \begin{cases}\hat{y}_0 \\ [\hat{y}_0^{(1)},\hat{y}_0^{(2)}]\end{cases} \\ \xrightarrow{[y',y'']} [x',x''] \end{cases} \end{cases}$$

三、一元线性回归模型

下面的讨论中,自变量x为非随机变量,而因变量y为随机变量.

1. 总体模型:

$$\begin{cases} y=\beta_0+\beta_1 x+\varepsilon \\ E\varepsilon=0, D\varepsilon=\sigma^2<\infty\ (\sigma^2\text{未知}) \end{cases} \tag{4.3}$$

回归系数——固定的未知参数β_0,β_1.

$E(y|x)=\beta_0+\beta_1 x$ ——y对x的回归函数

$\hat{y}=\hat{\beta}_0+\hat{\beta}_1 x$ ——y对x的回归方程

(拟合方程或预报方程) (4.4)

回归直线——$\hat{y}=\hat{\beta}_0+\hat{\beta}_1 x$所表示的直线(拟合直线);

回归值——对固定的x,相应y的估计值$\hat{y}=\hat{\beta}_0+\hat{\beta}_1 x$(拟合值或预报值).

2. 样本模型:

假设有n组独立观测值$(x_i, y_i)\ (i=1,2,\cdots,n)$, 则由(4.3)有

$$\begin{cases} y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, & i=1,2,\cdots,n \\ E\varepsilon_i = 0,\ D\varepsilon_i = \sigma^2, & \text{且}\ \varepsilon_1,\varepsilon_2,\cdots,\varepsilon_n\text{相互独立} \end{cases} \tag{4.5}$$

四、最小二乘估计及统计性质

1. β_0, β_1的最小二乘估计:

(1) 最小二乘法的原理:

误差平方和: $Q(\beta_0,\beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$

$$\boxed{Q(\hat{\beta}_0,\hat{\beta}_1) = \min_{\beta_0,\beta_1} Q(\beta_0,\beta_1)}$$

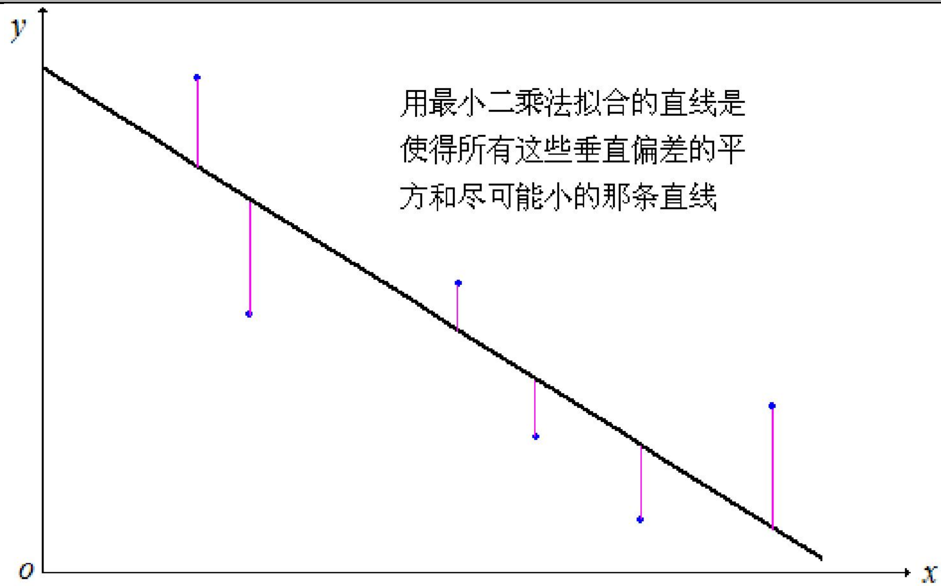

图4-5 最小二乘法拟合的直线

(2) β_0,β_1的最小二乘估计 (Least Squares Estimation, 简称LS估计):

$$\boxed{\hat{\beta}_1=\frac{L_{xy}}{L_{xx}},\quad \hat{\beta}_0=\bar{y}-\hat{\beta}_1\bar{x}} \qquad (4.11)(4.12)$$

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i,\quad \bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i,\quad L_{xx}=\sum_{i=1}^{n}(x_i-\bar{x})^2,\quad L_{xy}=\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}).$$

(3) y对x的回归方程:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \qquad —— \quad (4.13)$$

【★例4.1(P_{138})】为研究温度对某个化学过程生产量的影响,收集数据如下.求y对x的回归方程,并作拟合曲线图与观测数据的散点图.

x	-5	-4	-3	-2	-1	0	1	2	3	4	5
y	1	5	4	7	10	8	9	13	14	13	18

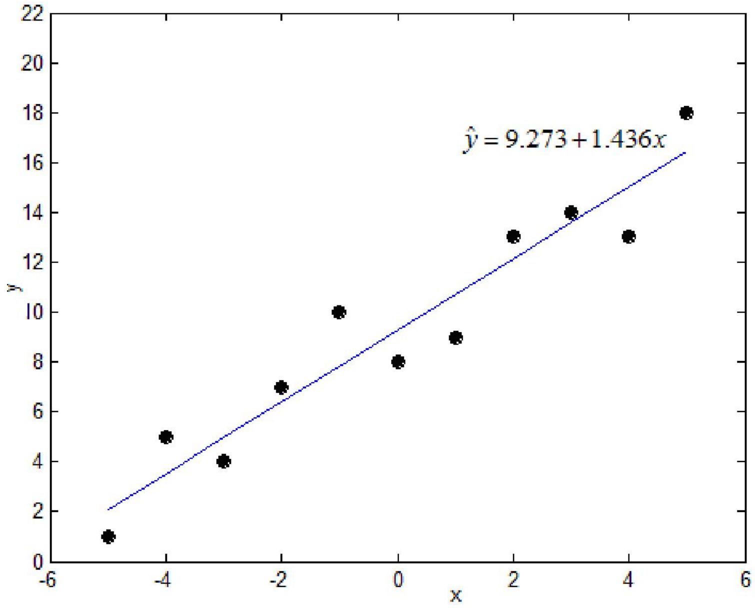

图4-4 数据散点图和拟合直线

2. LS估计$\hat{\beta}_0$和$\hat{\beta}_1$的性质：

(1) $\hat{\beta}_0$和$\hat{\beta}_1$是$y_1, y_2, \cdots, y_n$的线性函数.

(2) 定理4.1 $E(\hat{\beta}_0) = \beta_0$, $E(\hat{\beta}_1) = \beta_1$,

$$D(\hat{\beta}_0) = \left(\frac{1}{n} + \frac{\bar{x}^2}{L_{xx}}\right)\sigma^2, \qquad D(\hat{\beta}_1) = \frac{1}{L_{xx}}\sigma^2.$$

说明1：安排试验时应注意：

① $x_1, \cdots, x_n$可正可负时，尽可能使$\bar{x} = 0$，此时$D(\hat{\beta}_0)$最小；

② $x_1, \cdots, x_n$越分散越好，即使L_{xx}越大越好；

③ 试验次数n不能太小.

3. σ^2的无偏估计:

残差: $e_i = y_i - \hat{y}_i$

残差平方和: $Q_e = Q(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

定理4.2 $EQ_e = (n-2)\sigma^2 \quad \Rightarrow \quad E\left(\frac{Q_e}{n-2}\right) = \sigma^2.$

结论1: $\hat{\sigma}_e^2 = \frac{Q_e}{n-2}$(剩余方差、残差的方差)是$\sigma^2$的无偏估计.

五、回归方程的显著性检验和回归系数的置信区间

定理4.3: 若假定 $\varepsilon \sim N(0,\sigma^2)$, 则模型(4.5)可写成

$$\begin{cases} y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1,2,\cdots,n \\ \varepsilon_i \sim N(0,\sigma^2), \quad \text{且}\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_n\text{相互独立} \end{cases} \tag{4.17}$$

$$\left.\begin{array}{l} \hat\beta_0 \sim N\left(\beta_0, \left(\dfrac{1}{n} + \dfrac{\bar{x}^2}{L_{xx}}\right)\sigma^2\right) \\ \hat\beta_1 \sim N\left(\beta_1, \dfrac{1}{L_{xx}}\sigma^2\right) \\ \dfrac{Q_e}{\sigma^2} \sim \chi^2(n-2) \\ \bar{y}, \hat\beta_1, Q_e\text{相互独立}, \hat\beta_0\text{与}Q_e\text{独立} \end{array}\right\} \Rightarrow \begin{cases} \dfrac{(\hat\beta_0 - \beta_0)\Big/\sqrt{\dfrac{1}{n} + \dfrac{\bar{x}^2}{L_{xx}}}}{\sqrt{Q_e/(n-2)}} \sim t(n-2) \\ \dfrac{(\hat\beta_1 - \beta_1)\Big/\sqrt{1/L_{xx}}}{\sqrt{Q_e/(n-2)}} \sim t(n-2) \\ \dfrac{Q_e}{\sigma^2} \sim \chi^2(n-2) \end{cases}$$

(一) 回归方程的显著性检验

检验y与 x 之间有无线性关系 $y=\beta_0+\beta_1 x+\varepsilon$, 即检验假设:

$$H_0:\beta_1=0;\quad H_1:\beta_1\neq 0 \tag{4.18}$$

1. 总离差平方和的分解式:

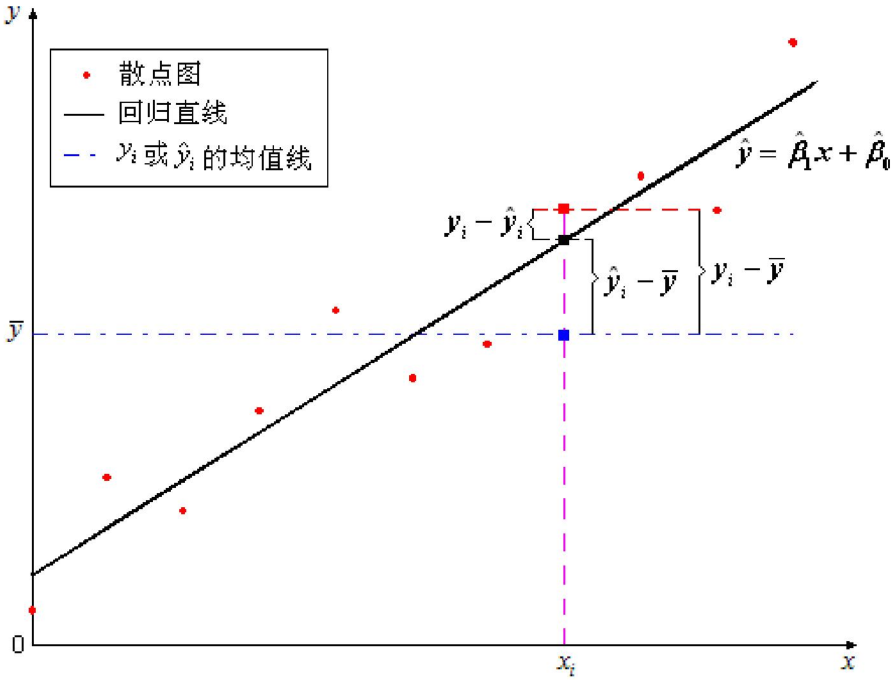

图4-7 总离差平方和分解式的几何意义

$$L_{yy}=\sum_{i=1}^{n}(y_i-\bar{y})^2$$ ——总离差平方和

$$Q_e=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$ ——残差平方和

$$U=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2=\hat{\beta}_1^2 L_{xx}=L_{xy}^2\big/L_{xx}$$ ——回归平方和

则 $$\boxed{L_{yy}=Q_e+U}$$

2. 分解式中各元素的统计意义：

$EQ_e=(n-2)\sigma^2$, $EU=\sigma^2+\beta_1^2L_{xx}$

Q_e——反映了误差引起数据$y_1,\cdots,y_n$的波动程度大小；

U——除反映了误差的作用外，还反映了回归因子x对y的线性影响.

3. 回归方程的显著性检验：

(1) F检验法

① 定理4.4 Q_e与U独立，且$U/\sigma^2 \overset{H_0\text{为真}}{\sim\sim\sim} \chi^2(1)$.

② 检验统计量：$\boxed{F=\dfrac{U}{Q_e/(n-2)}=\dfrac{\hat{\beta}_1^2L_{xx}}{Q_e/(n-2)} \overset{H_0\text{为真}}{\sim\sim\sim} F(1,n-2)}$.

③ $H_0:\beta_1=0$的拒绝域：$\boxed{F>F_{1-\alpha}(1,n-2)}$.

(2) t检验法:

① 检验统计量: $\boxed{T=\dfrac{\sqrt{L_{xx}}\hat{\beta}_1}{\sqrt{Q_e/(n-2)}} \overset{H_0\text{为真}}{\sim\sim\sim\sim} t(n-2)}$.

② $H_0:\beta_1=0$的拒绝域: $\boxed{|T|>t_{1-\alpha/2}(n-2)}$.

【★例4.2(P153)】为研究温度对某个化学过程的生产量的影响,收集到如下数据:

x	-5	-4	-3	-2	-1	0	1	2	3	4	5
y	1	5	4	7	10	8	9	13	14	13	18

并用最小二乘法得到拟合直线为: $\hat{y}=1.436x+9.273$. 现要求在正态分布下分别用t检验法和F检验法, 检验回归方程效果是否显著 $(\alpha=0.05)$.

(3) r检验法

① 由 $r^2=\dfrac{U}{L_{yy}}=\dfrac{L_{xy}^2}{L_{xx}L_{yy}}\Rightarrow$ $$\boxed{r=\frac{L_{xy}}{\sqrt{L_{xx}L_{yy}}}=\frac{\sum\limits_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum\limits_{i=1}^{n}(x_i-\bar{x})^2\sum\limits_{i=1}^{n}(y_i-\bar{y})^2}}}$$

——样本相关系数 (4.23)

② $H_0:\beta_1=0$的拒绝域: $\boxed{|r|>r_{1-\alpha}}$.

其中$r_{1-\alpha}$通过查“相关系数检验临界值表”得到.

注意1: r检验法临界值可由F检验法推得.

说明2: ★若拒绝H_0, 则认为y与x存在线性关系, 所求回归方程有意义;

★否则此回归方程无意义. 此时, 可能有如下几种情况:

- x对y没有显著影响, 此时应丢掉x;
- x对y有显著影响, 但该影响不是线性的, 改用非线性回归;
- 除x外, 还有其它不可忽略的自变量对y有显著影响, 从而削弱了x对y的影响, 此时改用多元线性回归.

(二) 回归系数的置信区间(记$\hat{\sigma}_e = \sqrt{Q_e/(n-2)}$)

1. β_0的置信水平为$1-\alpha$的置信区间:

$$\boxed{\frac{(\hat{\beta}_0 - \beta_0)\Big/\sqrt{\dfrac{1}{n}+\dfrac{\bar{x}^2}{L_{xx}}}}{\sqrt{Q_e/(n-2)}} \sim t(n-2)} \Rightarrow \left[\hat{\beta}_0 \pm t_{1-\alpha/2}(n-2)\cdot\hat{\sigma}_e\cdot\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{L_{xx}}}\right].$$

2. β_1的置信水平为$1-\alpha$的置信区间:

$$\boxed{\frac{(\hat{\beta}_1 - \beta_1)\Big/\sqrt{1/L_{xx}}}{\sqrt{Q_e/(n-2)}} \sim t(n-2)} \Rightarrow \left[\hat{\beta}_1 \pm t_{1-\alpha/2}(n-2)\cdot\hat{\sigma}_e\cdot\sqrt{\frac{1}{L_{xx}}}\right].$$

3. σ^2的置信水平为$1-\alpha$的置信区间:

$$\boxed{\frac{Q_e}{\sigma^2} \sim \chi^2(n-2)} \Rightarrow \left[\frac{Q_e}{\chi^2_{1-\alpha/2}(n-2)}, \frac{Q_e}{\chi^2_{\alpha/2}(n-2)}\right].$$

【★例4.3(P_{156})】为研究温度对某个化学过程的生产量的影响,收集到如下数据:

x	-5	-4	-3	-2	-1	0	1	2	3	4	5
y	1	5	4	7	10	8	9	13	14	13	18

求回归系数β_0,β_1的置信区间($\alpha=0.05$).

六、预测与控制

设y与x满足模型$\begin{cases} y=\beta_0+\beta_1 x+\varepsilon \\ \varepsilon \sim N(0,\sigma^2) \end{cases}$(4.17). 并已得回归方程$\hat{y}=\hat{\beta}_1 x+\hat{\beta}_0$.

1. 预测: 对固定的x值预测相应的y值.

令x_0为x的一固定值, 且$\begin{cases} y_0=\beta_0+\beta_1 x_0+\varepsilon_0 \\ \varepsilon_0 \sim N(0,\sigma^2) \end{cases}$. 设$y_0,y_1,\cdots,y_n$相互独立,

求:

(1) y_0的预测值:

$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$, 且$\hat{y}_0$是$Ey_0$的无偏估计.

(2) y_0的置信水平为$1-\alpha$的预测区间:

$$\boxed{T = \frac{y_0 - \hat{y}_0}{\hat{\sigma}_e \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{L_{xx}}}} \sim t(n-2)} \Rightarrow [\hat{y}_0 \pm \delta(x_0)] \tag{4.31}$$

其中, $\delta(x_0) = \hat{\sigma}_e \cdot t_{1-\alpha/2}(n-2) \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{L_{xx}}}$.

说明3:

· $\hat{\sigma}_e$越小, 预测区间越窄, 预测就越精确;

· x_0越靠近$\bar{x}$, 预测精度也就越高.

(3) $y=\beta_0+\beta_1 x+\varepsilon$ 的预测区间:

$$\left[\hat{y}\pm\delta(x)\right] \tag{4.32}$$

特别, 当n很大而$|x-\bar{x}|$很小时, y的$1-\alpha$的预测区间近似为:

$$\left[\hat{y}\pm\hat{\sigma}_e u_{1-\alpha/2}\right] \tag{4.33}$$

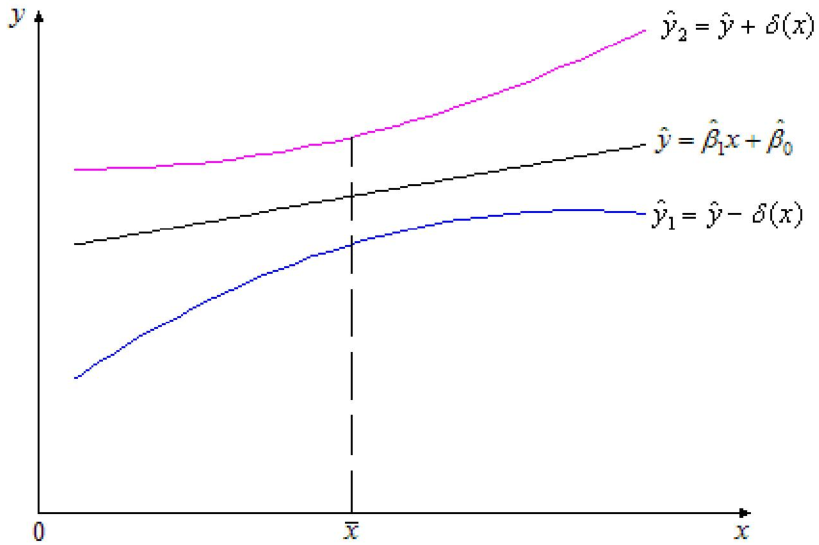

图4-8 y的预测带(4.32)

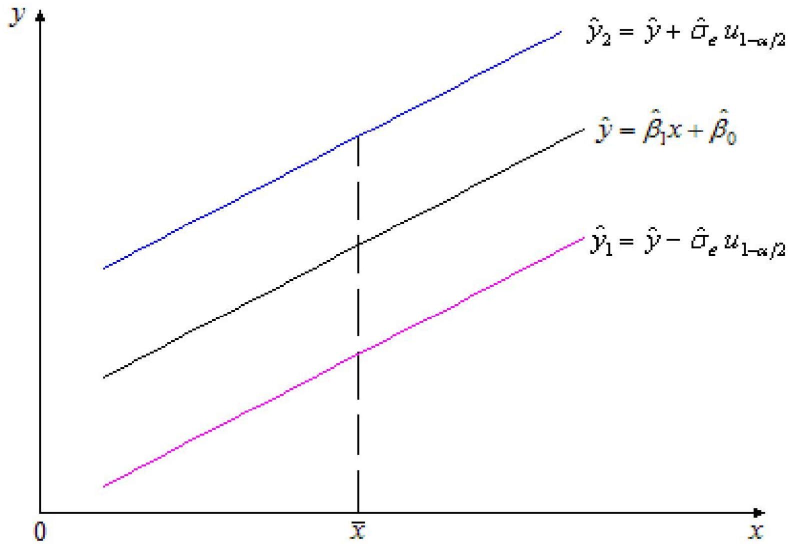

图4-9 y的近似预测带(4.33)

【★例4.4(P_{159})】为研究温度对某个化学过程的生产量的影响, 收集到如下数据:

x	-5	-4	-3	-2	-1	0	1	2	3	4	5
y	1	5	4	7	10	8	9	13	14	13	18

取$x_0=3$, 求y_0的预测值与置信水平为$1-\alpha=0.95$的预测区间.

2. 控制: 控制x的值以便把y的值控制在指定的范围内.

若要$y=\beta_0+\beta_1 x+\varepsilon$的值以$1-\alpha$的概率落在指定区间$(y',y'')$之内, 求自变量x的控制范围$(x',x'')$.

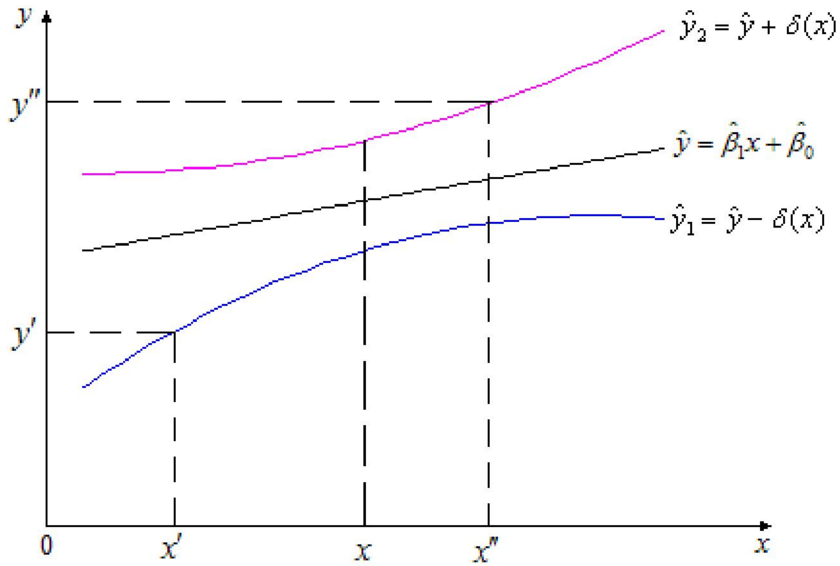

图4-10 x的控制范围

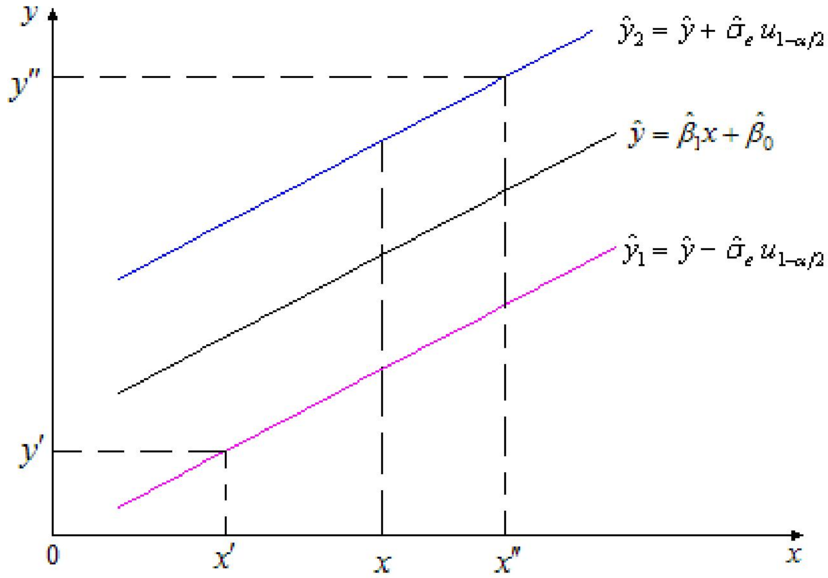

图4-11 x的近似控制范围 (n很大且$|x-\bar{x}|$很小)

对图4-11, 由$\begin{cases} y'=\hat{\beta}_1 x+\hat{\beta}_0-\hat{\sigma}_e u_{1-\alpha/2} \\ y''=\hat{\beta}_1 x+\hat{\beta}_0+\hat{\sigma}_e u_{1-\alpha/2} \end{cases} \Rightarrow \begin{cases} x'=\dfrac{1}{\hat{\beta}_1}(y'-\hat{\beta}_0+\hat{\sigma}_e u_{1-\alpha/2}) \\ x''=\dfrac{1}{\hat{\beta}_1}(y''-\hat{\beta}_0-\hat{\sigma}_e u_{1-\alpha/2}) \end{cases}$.

当$\hat{\beta}_1>0$时, x控制为(x',x''); 当$\hat{\beta}_1<0$时, x控制为(x'',x').

注意2: 为了实现上述控制, 必须使区间(y',y'')的长度大于$2\hat{\sigma}_e u_{1-\alpha/2}$.