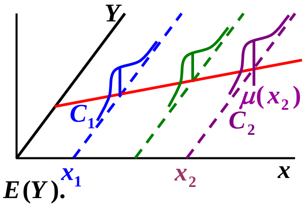
4.2　一元线性回归

4.2.1　基本概念

设随机变量 Y(因变量)和普通变量 x(自变量)之间存在着相关关系 .

$F(y|x)$表示当 x取确定的值 x时,所对应的Y的分布函数 .



考察Y的数学期望$E(Y)$.

$E(Y\mid x)=\mu_{Y|x}=\mu(x)$　Y关于x的回归函数

$E(Y\mid x)=\mu_{Y|x}=\mu(x)$

因为对随机变量 η,当 $c=E(\eta)$时, $E[(\eta-c)^2]$达到最小.

所以在一切 x 的函数中以回归函数 $\mu(x)$ 作为 Y 的近似,均方误差 $E[(Y-\mu(x))^2]$为最小.

实际问题中的 $\mu(x)$一般未知 .

回归分析的任务——根据试验数据估计回归函数;讨论回归函数中参数的点估计、区间估计;对回归函数中的参数或者回归函数本身进行假设检验;利用回归函数进行预测与控制等等.

问题的一般提法

对 x 的一组不完全相同的值 $x_1, x_2,\cdots,x_n$, 设 $Y_1,Y_2,\cdots,Y_n$ 分别是在 $x_1, x_2,\cdots,x_n$ 处对 Y 的独立观察结果.

称 $(x_1,Y_1),(x_2,Y_2),\cdots,(x_n,Y_n)$ 是一个样本.

对应的样本值记为

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

利用样本来估计 Y 关于 x 的回归函数 $\mu(x)$.

求解步骤

1.推测回归函数的形式

方法一　根据专业知识或者经验公式确定;

方法二　作散点图观察.

例 4.2.1　考查硫酸铜在水中的溶解度 y 与温度 x 的关系时，做了 9 组实验，其数据见表

温度 x/℃	0	10	20	30	40	50	60	70	80
溶解度 y/g	14.0	17.5	21.2	26.1	29.2	33.3	40.0	48.0	54.8

这里 x 是自变量，y 是随机变量，

散点图可以帮助人们粗略了解用什么形式的函数估计回归函数

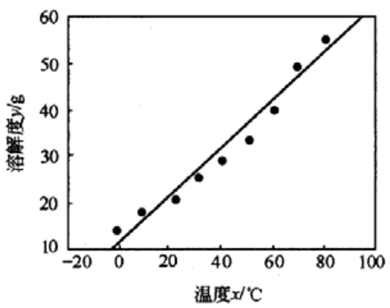


图 4.1　数据和拟合直线

观察散点图,$\mu(x)$具有线性函数$\beta_0+\beta_1 x$的形式.

2.建立回归模型

$\mu(x)=\beta_0+\beta_1 x$　**一元线性回归问题**

记 $\varepsilon=Y-(\beta_0+\beta_1 x)$, 那么

$$\boxed{\begin{aligned} Y&=\boxed{\beta_0+\beta_1 x}+\boxed{\varepsilon},\\ E(\varepsilon)&=0, D(\varepsilon)=\sigma^2<\infty. \end{aligned}}\quad(4.2.1)$$

一元线性回归模型

x的线性函数　　随机误差

β_0，β_1 称为模型参数，是未知的，

x 是可控制或可观测的非随机变量

ε 是不可观测的随机变量，Y 是可观测的随机变量

易知 $E(Y)=\beta_0+\beta_1 x$，$D(Y)=\sigma^2$，

称由式（4.2.1）确定的模型为一元线性回归模型，

固定的未知参数 β_1 称为回归系数，预报变量

x 也称为回归因子，$\tilde{y}=\beta_0+\beta_1 x$ 称为回归方程.

设由 n 组样本值 $(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$ 做出 β_0 和 β_1 的估计于是得到理论回归方程 $\tilde{y}=\beta_0+\beta_1 x$ 的一个估计 $\hat{y}=\hat{\beta}_0+\hat{\beta}_1 x$，

称为经验回归方程，对应的直线称为回归直线，

每代入一个 x 值，就得到对应的 y 值的预报.

4.2.2 最小二乘估计及其性质

4.2.2.1 β_0，β_1 的最小二乘估计

设给定了满足式（4.2.1）的 n 组观测值，则有

$$y_i=\beta_0+\beta_1 x_i+\varepsilon_i,\ i=1,2,\cdots,n$$

直观上的想法是画出一条直线，使它尽可能

靠近坐标平面上的这 n 个点

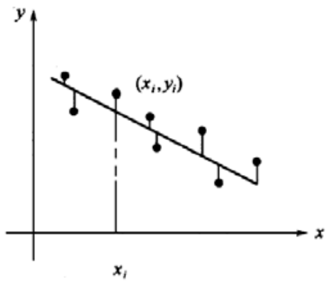


图 4.2 拟合直线示意图

记 $Q=Q(\beta_0,\beta_1)=\sum_{i=1}^{n}\varepsilon_i^2=\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2$ （4.2.2）

称 $Q(\beta_0,\beta_1)$ 为偏差平方和

用最小二乘法就是选择 β_0，β_1 的估计 $\hat{\beta}_0$，$\hat{\beta}_1$ 使得

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

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

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

令 $\frac{\partial Q}{\partial \beta_i}=0$，$i=0$，1，并用 $\hat{\beta}_0$，$\hat{\beta}_1$ 取代 β_0，β_1 得

$$\begin{cases}\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1 x_i)=0\\ \sum_{i=1}^{n}x_i(y_i-\hat{\beta}_0-\hat{\beta}_1 x_i)=0\end{cases}$$

正规方程组 $\begin{cases}n\hat{\beta}_0+\hat{\beta}_1\sum_{i=1}^{n}x_i=\sum_{i=1}^{n}y_i\\ \hat{\beta}_0\sum_{i=1}^{n}x_i+\hat{\beta}_1\sum_{i=1}^{n}x_i^2=\sum_{i=1}^{n}x_i y_i\end{cases}$

解正规方程组得

$$\hat{\beta}_1=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}=S_{xy}/S_{xx}$$

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

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

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

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

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

显然$\hat{\beta}_0$，$\hat{\beta}_1$ 使偏差平方和 Q 达到最小值，称为参数 β_0，β_1 的最小二乘估计（Least Squares Estimation），简称 LS 估计.

经验回归方程为

$$\hat{y}=\hat{\beta}_0+\hat{\beta}_1 x=\bar{y}+\hat{\beta}_1\ (x-\bar{x})$$

$\bar{y}=\hat{\beta}_0+\hat{\beta}_1\bar{x}$，此式表明，经验回归直线总通过散点图的几何重心$(\bar{x},\bar{y})$.

例 4.2.2 （续例 4.2.1）求例 4.2.1 的经验回归方程

解 此处 $n=9$，$(x_i,y_i)(i=1,2,\cdots,9)$ 值见表 4.1，计算得出

$$\sum_{i=1}^{9}x_i=360,\sum_{i=1}^{9}x_i^2=20400$$

$$\sum_{i=1}^{9}y_i=284.1,\sum_{i=1}^{9}y_i^2=10501.47$$

$$\sum_{i=1}^{9}x_iy_i=14359,\bar{x}=40,\bar{y}=31.57$$

$$S_{xx}=\sum_{i=1}^{9}x_i^2-\frac{1}{9}\Big(\sum_{i=1}^{9}x_i\Big)^2=6000$$

$$S_{xy}=\sum_{i=1}^{9}x_iy_i-\frac{1}{9}\Big(\sum_{i=1}^{9}x_i\Big)\Big(\sum_{i=1}^{9}y_i\Big)=2995$$

$$S_{yy}=\sum_{i=1}^{9}y_i^2-\frac{1}{9}\Big(\sum_{i=1}^{9}y_i\Big)^2=1533.38$$

$$\hat{\beta}_1=S_{xy}/S_{xx}=0.499$$

$$\hat{\beta}_0=\bar{y}-\hat{\beta}_1\bar{x}=11.61$$

故所求经验回归方程为 $\hat{y}=11.61+0.499x$

回归系数$\hat{\beta}_1=0.499$，它的单位是 g/℃，其意义为每提高 1℃ 的温度，平均可以提高溶解度 0.499g/℃.

对每组(x_i,y_i)，可以求出拟合值$\hat{y}_i$ 以及残差 $y_i-\hat{y}_i$.

$$\hat{y}_i=\bar{y}+\hat{\beta}_1(x_i-\bar{x})$$

$$y_i-\hat{y}_i=(y_i-\bar{y})-\hat{\beta}_1(x_i-\bar{x})$$

$$\sum_{i=1}^{n}(y_i-\hat{y}_i)=\sum_{i=1}^{n}(y_i-\bar{y})-\hat{\beta}_1\sum_{i=1}^{n}(x_i-\bar{x})=0$$

这说明残差之和为零，但在实际计算中，由于有舍入误差，残差之和可能不为零.

4.2.2.2 最小二乘估计的性质

定理 4.2.1 $\hat{\beta}_0$，$\hat{\beta}_1$ 分别是 β_0 和 β_1 的无偏估计，且 $\mathrm{Cov}(\bar{Y},\hat{\beta}_1)=0$，$\mathrm{Cov}(\hat{\beta}_0,\hat{\beta}_1)=-\frac{\bar{x}}{S_{xx}}\sigma^2$，$D(\hat{\beta}_0)=\sigma^2\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)$，$D(\hat{\beta}_1)=\frac{\sigma^2}{S_{xx}}$.

证 $\hat{\beta}_1=\frac{S_{xy}}{S_{xx}}=\sum_{i=1}^{n}\frac{(x_i-\bar{x})}{S_{xx}}Y_i$

$$E(\hat{\beta}_1)=E\left(\frac{\sum_{i=1}^{n}(x_i-\bar{x})Y_i}{S_{xx}}\right)=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(\beta_0+\beta_1x_i)}{S_{xx}}=\beta_1$$

$$E(\hat{\beta}_0)=E(\bar{Y}-\hat{\beta}_1\bar{x})=\frac{1}{n}\sum E(Y_i)-\bar{x}\beta_1=\frac{1}{n}\sum_{i=1}^{n}(\beta_0+\beta_1x_i)-\bar{x}\beta_1=\beta_0$$

故$\hat{\beta}_0$，$\hat{\beta}_1$ 分别是 β_0 与 β_1 的无偏估计.

$$\mathrm{Cov}(\bar{Y},\hat{\beta}_1) = \mathrm{Cov}\left(\bar{Y},\frac{\sum_{i=1}^{n}(x_i-\bar{x})Y_i}{S_{xx}}\right)$$
$$= \sum_{i=1}^{n}\frac{(x_i-\bar{x})}{S_{xx}}\mathrm{Cov}(\bar{Y},Y_i)$$
$$= \sum_{i=1}^{n}\frac{(x_i-\bar{x})}{S_{xx}}\times\frac{\sigma^2}{n} = 0.$$

$$\mathrm{D}(\hat{\beta}_1) = \mathrm{D}\left(\frac{\sum_{i=1}^{n}(x_i-\bar{x})Y_i}{S_{xx}}\right) = \frac{1}{S_{xx}^2}\sum_{i=1}^{n}(x_i-\bar{x})^2\sigma^2 = \frac{\sigma^2}{S_{xx}}$$

$$\mathrm{D}(\hat{\beta}_0) = \mathrm{D}(\bar{Y}-\hat{\beta}_1\bar{x})$$
$$= \mathrm{D}(\bar{Y}) + \bar{x}^2\mathrm{D}(\hat{\beta}_1) - 2\bar{x}\mathrm{Cov}(\bar{Y},\hat{\beta}_1)$$
$$= \frac{\sigma^2}{n} + \bar{x}^2\frac{\sigma^2}{S_{xx}} = \sigma^2\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right) = \sigma^2\frac{\sum_{i=1}^{n}x_i^2}{nS_{xx}}$$

$$\mathrm{Cov}(\hat{\beta}_0,\hat{\beta}_1) = \mathrm{Cov}(\bar{Y}-\hat{\beta}_1\bar{x},\hat{\beta}_1)$$
$$= \mathrm{Cov}(\bar{Y},\hat{\beta}_1) - \bar{x}\mathrm{Cov}(\hat{\beta}_1,\hat{\beta}_1) = -\frac{\bar{x}}{S_{xx}}\sigma^2$$

$$\mathrm{E}(\hat{Y}) = \mathrm{E}(\hat{\beta}_0+\hat{\beta}_1x) = \beta_0+\beta_1x$$

即经验回归方程是理论回归方程的无偏估计.

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \sum_{i=1}^{n}\frac{(x_i-\bar{x})}{S_{xx}}Y_i$$
$$\hat{\beta}_0 = \bar{Y}-\hat{\beta}_1\bar{x} = \sum_{i=1}^{n}\left[\frac{1}{n}-\frac{\bar{x}(x_i-\bar{x})}{S_{xx}}\right]Y_i$$

故最小二乘估计$\hat{\beta}_0$ 和$\hat{\beta}_1$，是 β_0，β_1 的线性无偏估计.
进一步还可证明,最小二乘估计$\hat{\beta}_0$，$\hat{\beta}_1$ 分别是 β_0 和 β_1 的最优线性无偏估计,即在形如 $\sum_{i=1}^{n}c_iY_i$ (c_i 为常数) 的估计中，它们分别是 β_0 和 β_1 的最小方差线性无偏估计.

4.2.2.3　σ^2 的无偏估计

$$\hat{\varepsilon}_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1x_i, i = 1,2,\cdots,n$$
$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2 = \frac{1}{n}RSS$$
(SSE)

称 $RSS = \sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2$ 为残差平方和或剩余平方和
(SSE)
(Residual sum of squares).

定理 4.2.2　在模型式（4.2.1）下，

有 $\mathrm{E}(\hat{\sigma}^2) = (n-2)\sigma^2/n$.

证　由于 $\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2 = \sum_{i=1}^{n}[Y_i-(\bar{Y}+\hat{\beta}_1(x_i-\bar{x})]^2$

$$= \sum_{i=1}^{n}(Y_i-\bar{Y})^2 - 2\hat{\beta}_1\sum_{i=1}^{n}(x_i-\bar{x})(Y_i-\bar{Y})$$
$$+ \hat{\beta}_1^2\sum_{i=1}^{n}(x_i-\bar{x})^2$$
$$= \sum_{i=1}^{n}(Y_i-\bar{Y})^2 - 2\hat{\beta}_1S_{xy} + \hat{\beta}_1^2S_{xx}$$
$$= \sum_{i=1}^{n}(Y_i-\bar{Y})^2 - \hat{\beta}_1^2S_{xx}$$

$$\mathrm{E}(\hat{\sigma}^2) = \frac{1}{n}\mathrm{E}\left[\sum_{i=1}^{n}(Y_i-\bar{Y})^2 - \hat{\beta}_1^2S_{xx}\right]$$
$$= \frac{1}{n}\left[\mathrm{E}\left(\sum_{i=1}^{n}Y_i^2 - n\bar{Y}^2\right) - S_{xx}\mathrm{E}(\hat{\beta}_1^2)\right]$$
$$= \frac{1}{n}\left[\sum_{i=1}^{n}\mathrm{E}(Y_i^2) - n\mathrm{E}(\bar{Y}^2) - S_{xx}\mathrm{E}(\hat{\beta}_1^2)\right]$$
$$= \frac{1}{n}\left[\sum_{i=1}^{n}(\mathrm{D}(Y_i)+\mathrm{E}^2(Y_i))\right] - [\mathrm{D}(\bar{Y})+\mathrm{E}^2(\bar{Y})]$$
$$- \frac{S_{xx}}{n}[\mathrm{D}(\hat{\beta}_1)+\mathrm{E}^2(\hat{\beta}_1)]$$
$$= \frac{1}{n}\left[n\sigma^2 + \sum_{i=1}^{n}(\beta_0+\beta_1x_i)^2\right] - \left[\frac{\sigma^2}{n}+(\beta_0+\beta_1\bar{x})^2\right]$$
$$- \frac{S_{xx}}{n}\left[\frac{\sigma^2}{S_{xx}}+\beta_1^2\right] = \frac{n-2}{n}\sigma^2$$

记　$S^2 = n\hat{\sigma}^2/(n-2) = \overset{(SSE)}{RSS}/(n-2)$.

显然有 $E(S^2)=\sigma^2$，即 S^2 是 σ^2 的无偏估计.

4.2.3　相关系数与回归显著性检验

4.2.3.1　平方和分解公式

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

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

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

记　$\underset{(SSR)}{Reg.SS} = \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$

$$S_{yy}=Reg.SS+RSS$$

(SSY = SSR + SSE)

事实上，

$$\bar{\hat{y}}=\frac{1}{n}\sum_{i=1}^{n}\hat{y}_i=\frac{1}{n}\sum_{i=1}^{n}(\hat{\beta}_0+\hat{\beta}_1x_i)=\hat{\beta}_0+\hat{\beta}_1\bar{x}=\bar{y}.$$

$$\underset{(SSR)}{Reg.SS}=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\bar{\hat{y}})^2$$

就是 $\hat{y}_1$，$\hat{y}_2$，…，$\hat{y}_n$ 的偏差平方和，它描述了 $\hat{y}_1$，$\hat{y}_2$，…，$\hat{y}_n$ 的离散程度，其分散性来源于 x_1，x_2…，x_n 的分散性，且与直线的斜率 $\hat{\beta}_1$ 有关，

$$\underset{(SSR)}{Reg.SS}=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$$

$$=\sum_{i=1}^{n}(\hat{\beta}_0+\hat{\beta}_1x_i-\bar{y})^2$$

$$=\sum_{i=1}^{n}[\hat{\beta}_1(x_i-\bar{x})]^2$$

$$=\hat{\beta}_1^2S_{xx}$$

同样的一组 x_i（$i=1, 2, \cdots, n$），对于较陡的直线就会有较大 $\underset{(SSR)}{Reg.SS}$.

$\underset{(SSR)}{Reg.SS}$ 是回归直线上点的纵坐标的离差平方和，

称 $\underset{(SSR)}{Reg.SS}$ 为回归平方和（Regression sum of squares).

残差平方和

$$\underset{(SSE)}{RSS}=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=\sum_{i=1}^{n}[y_i-(\hat{\beta}_0+\hat{\beta}_1x_i)]^2$$

代表 y_i 与 $\hat{y}_i$ 的离差平方和，它是扣除了 x 对 y 的线性影响后剩余的平方和，特别当 x 与 Y 有密切线性关系时，$\underset{(SSE)}{RSS}$ 主要反映了随机误差的大小.

4.2.3.2　相关系数及其几何意义

定义 $r=S_{xy}/(\sqrt{S_{xx}}\sqrt{S_{yy}})$，称 r 为相关系数，

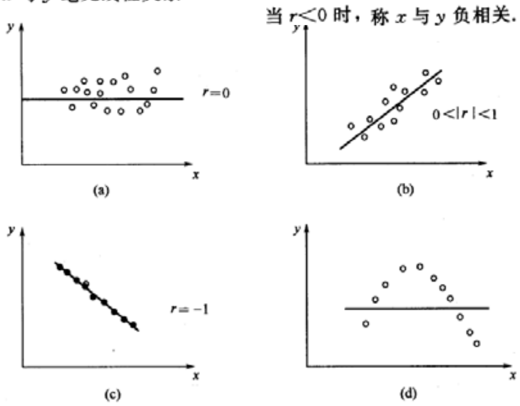
$$r^2=\frac{S_{xy}^2}{S_{xx}S_{yy}}=\frac{\hat{\beta}_1^2S_{xx}}{S_{yy}}=\frac{Reg.SS}{S_{yy}}\quad\left(\frac{SSR}{SSY}\right)$$

r^2 恰好代表了回归平方和占总离差平方和 S_{yy} 的比率.

$0\leqslant r^2\leqslant 1$，即 $0\leqslant |r|\leqslant 1$.

x 与 y 毫无线性关系

当 $r>0$ 时，称 x 与 y 正相关
当 $r<0$ 时，称 x 与 y 负相关.



称 x 与 y 完全线性相关

x 与 y 间有非线性关系

从以上的讨论和相应的图示可以看出，相关系数 r 确实可以表示 x 与 y 之间的线性关系的密切程度，$|r|$ 越接近于 0，x 与 y 之间线性相关程度越差，$|r|$ 越大，越接近于 1，x 与 y 之间的线性关系程度越密切. 但还应指出，相关系数只表示 x 与 y 的线性关系的密切程度，当 r 很小甚至等于 0 时，也可能如图 4.3 (d) 所示，x 与 y 间有非线性关系.

4.2.3.3 线性回归的显著性检验

正态线性模型 $\begin{cases} Y_i=\beta_0+\beta_1 x_i+\varepsilon_i,\ i=1,2,\cdots,n \\ \varepsilon_i\sim N(0,\sigma^2),\ 且\ \varepsilon_1,\varepsilon_2,\cdots,\varepsilon_n\ 相互独立 \end{cases}$

可知 $Y_i\sim N(\beta_0+\beta_1 x_i,\sigma^2)$，$i=1,2,\cdots,n$，

且 $Y_1,Y_2,\cdots,Y_n$ 相互独立，

定理 4.2.3 在正态线性模型下

(1) $\hat\beta_0\sim N\left[\beta_0,\sigma^2\left(\frac{1}{n}+\frac{\bar x^2}{S_{xx}}\right)\right]$；$\hat\beta_1\sim N\left(\beta_1,\frac{\sigma^2}{S_{xx}}\right)$

$\mathrm{Cov}(\hat\beta_0,\hat\beta_1)=-\frac{\sigma^2\bar x}{S_{xx}}$

(2) $\mathrm{Cov}(\hat\varepsilon_i,\hat\beta_0)=0$，$\mathrm{Cov}(\hat\varepsilon_i,\hat\beta_1)=0$，$i=1,2,\cdots,n$

(3) $(n-2)S^2/\sigma^2\sim\chi^2(n-2)$，且 $\hat\beta_1,S^2,\bar Y$ 相互独立.

$$S^2=\underset{(SSE)}{RSS}/(n-2)$$

$$\begin{aligned}
&\mathrm{Cov}(\hat\varepsilon_i,\hat\beta_0)=\mathrm{Cov}(Y_i-\hat\beta_0-\hat\beta_1 x_i,\hat\beta_0)\\
&=\mathrm{Cov}(Y_i,\hat\beta_0)-\mathrm{Cov}(\hat\beta_0,\hat\beta_0)-x_i\mathrm{Cov}(\hat\beta_1,\hat\beta_0)\\
&=\mathrm{Cov}(Y_i,\bar Y-\hat\beta_1\bar x)-\sigma^2\left(\frac{1}{n}+\frac{\bar x^2}{S_{xx}}\right)+x_i\frac{\sigma^2\bar x}{S_{xx}}\\
&=\frac{1}{n}\sigma^2-\frac{\bar x}{S_{xx}}\mathrm{Cov}(Y_i,S_{xy})-\frac{1}{n}\sigma^2-\frac{\sigma^2\bar x^2}{S_{xx}}+x_i\frac{\sigma^2\bar x}{S_{xx}}\\
&=-\frac{\bar x}{S_{xx}}\mathrm{Cov}\left(Y_i,\sum_{i=1}^{n}(x_i-\bar x)Y_i\right)-\frac{\sigma^2\bar x^2}{S_{xx}}+x_i\frac{\sigma^2\bar x}{S_{xx}}\\
&=-\frac{\sigma^2\bar x(x_i-\bar x)}{S_{xx}}+\frac{\sigma^2\bar x(x_i-\bar x)}{S_{xx}}=0
\end{aligned}$$

回归效果的检验问题归结为对假设

$H_0:\beta_1=0$，$H_1:\beta_1\neq 0$ 进行检验.

若拒绝 H_0，就认为 x 与 Y 存在线性关系，

所求出的线性回归方程有意义，

若接受 H_0，则认为 x 与 Y 的关系不能用一元线性回归模型来描述，所得的回归方程也无意义.

(1) t 检验法.

$\hat\beta_1\sim N(\beta_1,\sigma^2/S_{xx})$，$(n-2)S^2/\sigma^2\sim\chi^2(n-2)$，

$\hat\beta_1$ 与 S^2 相互独立，则

$$\frac{\hat{\beta}_1-\beta_1}{S/\sqrt{S_{xx}}}\sim t(n-2)$$

当 H_0 成立时，$T=\hat{\beta}_1\sqrt{S_{xx}}/S\sim t(n-2)$

对于给定的检验水平 α，查 t 分位数表得临界值 $t_{1-\alpha/2}(n-2)$，使

$$P\{|T|\geqslant t_{1-\alpha/2}(n-2)\}=\alpha$$

当 $|T|$ 的值 $|t|\geqslant t_{1-\alpha/2}(n-2)$ 时拒绝 H_0，否则接受 H_0.

(2) F 检验法.

由于 $\dfrac{\hat{\beta}_1-\beta_1}{\sigma/\sqrt{S_{xx}}}\sim N(0,1)$，则 $\dfrac{(\hat{\beta}_1-\beta_1)^2}{\sigma^2/S_{xx}}\sim\chi^2(1)$.

又由于 $\hat{\beta}_1$ 与 S^2 相互独立，

$$而(n-2)S^2/\sigma^2\sim\chi^2(n-2),$$

故有 $\dfrac{(\hat{\beta}_1-\beta_1)^2}{S^2/S_{xx}}\sim F(1,n-2)$.

当 H_0 成立时，$F=\dfrac{\hat{\beta}_1^2}{\underset{(SSR)}{S^2/S_{xx}}}=\dfrac{S_{xy}^2/S_{xx}}{\underset{(SSE)}{RSS}/(n-2)}$

$$=\frac{Reg.SS}{\underset{(SSE)}{RSS}/(n-2)}\sim F(1,n-2)$$

对于给定的检验水平 α，查 F 分布表得临界值 $F_{1-\alpha}(1,n-2)$，使

$$P\{F\geqslant F_{1-\alpha}(1,n-2)\}=\alpha$$

当 F 值不小于 $F_{1-\alpha}(1,n-2)$ 时拒绝 H_0，否则接受 H_0.

表 4.2 方差分析表

方差来源	平方和	自由度	均方	F统计值
回归	$Reg.SS=\frac{S_{xy}^2}{S_{xx}}$ (SSR)	1	$Reg.SS$ (SSR)	$Reg.SS\Big/\frac{RSS}{n-2}$ ($SSR\Big/\frac{SSE}{n-2}$)
剩余	$RSS=\sum_{i=1}^{n}\hat{\varepsilon}_i^2$ (SSE)	$n-2$	$RSS/(n-2)$ ($SSE/(n-2)$)	
总和	$S_{yy}=\sum_{i=1}^{n}(y_i-\bar{y})^2$ (SST)	$n-1$		

$$F=\frac{(n-2)r^2}{(1-r^2)},\quad 其中\ r=\frac{S_{xy}}{(\sqrt{S_{xx}}\sqrt{S_{yy}})}$$

例 4.2.3 （续例 4.2.2）在正态分布下，分别用 t 检验法和 F 检验法检验例 4.2.2 的回归方程效果是否显著，($\alpha=0.01$).

解 $H_0:\beta_1=0$；$H_1:\beta_1\neq 0$

$S_{xx}=6000$，$S_{xy}=2995$，$\hat{\beta}_1=0.499$，

$S_{yy}=1533.38$，$n=9$

用 F 检验法检验：

$\underset{(SSR)}{Reg.SS}=S_{xy}^2/S_{xx}=2995^2/6000=1495.00$

$\underset{(SSE)}{RSS}=\underset{(SST)}{S_{yy}}-\underset{(SSR)}{Reg.SS}=1533.38-1495.00=38.38$ (SSR)

$$F=\frac{Reg.SS}{\underset{SSE}{RSS}/(n-2)}=\frac{1495.00}{38.38/7}=272.67$$

查 F 分布表得 $F_{1-\alpha}(1,n-2)=F_{0.99}(1,7)=12.2$，

由于 $272.67>12.2$，

拒绝 H_0，即认为回归方程的效果显著.

若取 $\alpha=0.005$，有 $F_{0.995}(1,7)=16.24$，仍拒绝 H_0，可见回归方程的效果高度显著.

用 t 检验法检验：

$$S=\sqrt{\underset{(SSE)}{RSS}/(n-2)}=\sqrt{38.38/7}=2.3416$$

$$T=\hat{\beta}_1\sqrt{S_{xx}}/S=0.499\times\sqrt{6000}/2.3416=16.6994$$

查 t 分布表得 $t_{1-\alpha/2}(n-2)=t_{0.995}(7)=3.4995$，

由于 $|16.6994|>3.4995$，

拒绝 H_0，与 F 检验法的检验结果一致.

4.2.3.4 回归系数的置信区间

$$\frac{\hat{\beta}_1-\beta_1}{S/\sqrt{S_{xx}}}\sim t(n-2),$$

β_1 的置信水平为 $1-\alpha$ 的置信区间为

$$\left[\hat{\beta}_1-t_{1-\alpha/2}(n-2)S/\sqrt{S_{xx}},\ \hat{\beta}_1+t_{1-\alpha/2}(n-2)S/\sqrt{S_{xx}}\right]$$

$$(n-2)S^2/\sigma^2\sim\chi^2(n-2)$$

β_0 的置信水平为 $1-\alpha$ 的置信区间为

$$\left[\hat{\beta}_0-t_{1-\alpha/2}(n-2)S\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}},\ \hat{\beta}_0+t_{1-\alpha/2}(n-2)S\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}}\right]$$

σ^2 的置信水平为 $1-\alpha$ 的置信区间为

$$\left[\frac{(n-2)S^2}{\chi^2_{1-\alpha/2}(n-2)}, \frac{(n-2)S^2}{\chi^2_{\alpha/2}(n-2)}\right]$$

例 4.2.4 求例 4.2.2 中回归系数 β_1 的置信区间($\alpha=0.01$)

解 $S_{xx}=6000, S_{xy}=2995, S_{yy}=1533.38, \hat{\beta}_1=0.499$,

$S=2.3416$.

查 t 分布表得 $t_{0.995}(7)=3.4995$,

β_1 的置信水平为 0.99 的置信区间为

$$[0.3932,\ 0.6048].$$

例 4.2.5 K. Pearson 收集大量父亲身高与儿子身高的资料，其中 10 对数据见表 4.3.

表 4.3 父子身高数据

父身高 x/in	60	62	64	65	66	67	68	70	72	74
子身高 y/in	63.6	65.2	66	65.5	66.9	67.1	67.4	68.3	70.1	70

注：1in=25.4mm.

现在的问题是：Pearson 的资料能证实 F. Galton 的论断吗?

解 根据资料检验 $H_0:\beta_1\geqslant 1$; $H_1:\beta_1<1$

若拒绝 H_0，则证实了 Galton 的论断.

$S_{xx}=171.6$, $S_{yy}=38.529$, $S_{xy}=79.72$,

$\bar{x}=66.8$, $\bar{y}=67.01$, $n=10$.

于是：$\hat{\beta}_1=\dfrac{S_{xy}}{S_{xx}}=0.4646$, $\hat{\beta}_0=\bar{y}-\hat{\beta}_1\bar{x}=35.977$

故回归方程为 $\hat{y}=35.977+0.4646x$

做回归方程的显著性检验

$$F=\frac{\hat{\beta}_1^2}{S^2/S_{xx}}=\frac{S_{xy}^2/S_{xx}}{\underset{(SSE)}{RSS}/(n-2)}=198.37$$

取 $\alpha=0.05$，查表得 $F_{0.95}(1,8)=5.32$,

易见回归效果显著，即儿子身高与父亲身高之间有密切的线性关系，即父高者，其子也高，父矮者，其子也矮.

当 $\beta_1\geqslant 1$ 成立，取 $\beta_1=1$，有 $T=\dfrac{\hat{\beta}_1-1}{S}\sqrt{S_{xx}}\sim t(n-2)$

计算得：$T=\dfrac{0.4646-1}{0.4321}\times\sqrt{171.6}=-16.232$.

取 $\alpha=0.05$，查表得 $t_{1-\alpha}(n-2)=t_{0.95}(8)=1.860$

此单侧检验的拒绝域应为 $\dfrac{\hat{\beta}_1-1}{S}\sqrt{S_{xx}}\leqslant t_\alpha(n-2)$

或 $-\dfrac{\hat{\beta}_1-1}{S}\sqrt{S_{xx}}\geqslant -t_\alpha(n-2)=t_{1-\alpha}(n-2)$

由于 $-16.232<-1.860$，拒绝 H_0，即认为 $\beta_1<1$，即父亲“很高”，儿子只是“较高”，父亲“很矮”，儿子只是“较矮”，Galton 断言得到证实.

4.2.4 预测与控制

4.2.4.1 预测

设 Y 与 x 满足正态线性模型式

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

由历史资料 $(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$

求得回归方程 $\hat{y}=\hat{\beta}_0+\hat{\beta}_1 x$.

今 x_0 表示 x 的某个固定值，$Y_0\sim N(\beta_0+\beta_1 x_0,\sigma^2)$.

Y_0 与 $\hat{\beta}_0$、$\hat{\beta}_1$ 和 S^2 相互独立.

$$\hat{\beta}_0+\hat{\beta}_1x_0\sim N\left[\beta_0+\beta_1 x_{01},\sigma^2\left(\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)\right].)\Big]$$

$Y_0-(\hat{\beta}_0+\hat{\beta}_1x_0)$服从正态分布，且

$$Y_0-(\hat{\beta}_0+\hat{\beta}_1x_0)=Y_0-\hat{Y}_0\sim N\left[0,\sigma^2\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)\right]$$

由$\hat{\varepsilon}_i$与$\hat{\beta}_0$，$\hat{\beta}_1$均相互独立，

故S^2与$Y_0-(\hat{\beta}_0+\hat{\beta}_1x_0)$相互独立，

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

因而有

$$T=\frac{Y_0-(\hat{\beta}_0+\hat{\beta}_1x_0)}{S\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}}}\sim t(n-2)$$

利用T可求出Y_0的置信水平为$1-\alpha$的预测区间为

$$\left[\hat{\beta}_0+\hat{\beta}_1x_0-t_{\alpha/2}(n-2)S\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}},\ \hat{\beta}_0+\hat{\beta}_1x_0+t_{\alpha/2}(n-2)S\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}}\right]$$

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

则Y_0的置信水平为$1-\alpha$的预测区间为

$$[\hat{\beta}_0+\hat{\beta}_1x_0-\delta(x_0),\ \hat{\beta}_0+\hat{\beta}_1x_0+\delta(x_0)]$$

对固定的x值预测相应的Y值

Y的置信水平为$1-\alpha$的预测区间为

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

其中，$\hat{y}=\hat{\beta}_0+\hat{\beta}_1x$，

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

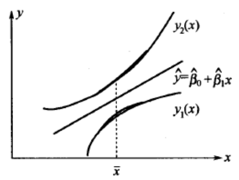


图 4.4 带形置信区域

$$y_1(x)=\hat{y}-\delta(x)$$
$$y_2(x)=\hat{y}+\delta(x)$$

当x离$\bar{x}$不太远，且n比较大时，

$t_{\alpha/2}(n-2)\approx u_{1-\alpha/2}$，$\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}}\approx1$，

此时$\delta(x)\approx Su_{1-\alpha/2}$，

Y的置信水平为$1-\alpha$的预测区间近似为

$$[\hat{y}-Su_{1-\alpha/2},\ \hat{y}+Su_{1-\alpha/2}]$$

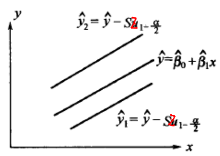


图 4.5 近似预测区间示意

在进行预测时，还要注意一点，x_0一定要落在已有的x的数据范围内部，对超出原观测范围的x_0进行预测常常是没有意义的.

例 4.2.6 在例 4.2.1 中取$x_0=45$，求Y_0的预测值与置信水平为 0.95 的预测区间.

解 由例 4.2.2 知经验回归方程为

$$\hat{y}=11.61+0.499x$$

$x_0=45$时，Y的预测值为$\hat{y}_0=34.065$，

$\bar{x}=40$，$S_{xx}=6000$，$S=2.3416$，$t_{0.025}(7)=2.3646$.

$x_0=45$时，Y的置信水平为 0.95 的预测区间为

$$[28.218,39.912].$$

4.2.4.2 控制

要求观测值y在区间(y',y'')内取值时，那么x应控制在什么范围内?

即要求求出相应的x',x''，当$x'<x<x''$时，以至少$1-\alpha$的置信水平使x所相应的观测值y落在(y',y'')内

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

且要求$y''-y'\geqslant2\delta(x)$，

若 $\begin{cases}\hat{y}-\delta(x)=y'\\ \hat{y}+\delta(x)=y''\end{cases}$ 有解x'和x''，

即$\hat{y}-\delta(x')=y'$且$\hat{y}+\delta(x'')=y''$，则(x',x'')就是所求的x的控制区间.

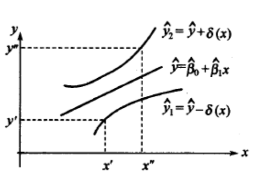


图 4.6　控制区间

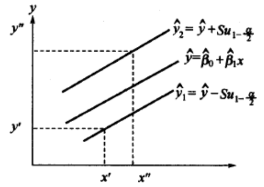


图 4.7　简化后的控制区间（示意图）

若要将 y 控制在 y',y'' 内，置信水平为 $1-\alpha$，求 x 的控制区间.

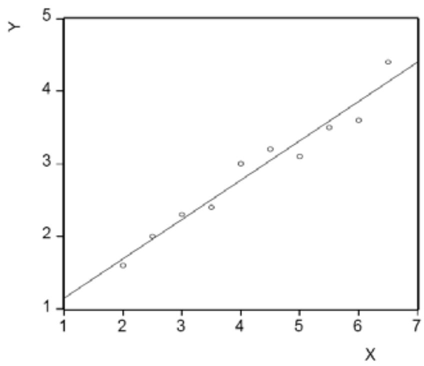
解方程组 $\begin{cases} y'=\hat{y}-Su_{1-\alpha/2} \\ y''=\hat{y}+Su_{1-\alpha/2} \end{cases}$，　$\begin{cases} y'=\hat{\beta}_0+\hat{\beta}_1x-Su_{1-\alpha/2} \\ y''=\hat{\beta}_0+\hat{\beta}_1x+Su_{1-\alpha/2} \end{cases}$

可得 x' 和 x''. 要注意的是，要有 $x'<x''$ 的解，对于 $\hat{\beta}_1>0$ 的线性回归方程 $\hat{y}=\hat{\beta}_0+\hat{\beta}_1x$，应满足 $y''-y'>2Su_{1-\alpha/2}$

- 作业：P114
  2，3，4

例：　消费支出与可支配收入的观测值

消费支出Y（千元）	可支配收入X（千元）
1.6	2.0
2.0	2.5
2.3	3.0
2.4	3.5
3.0	4.0
3.2	4.5
3.1	5.0
3.5	5.5
3.6	6.0
4.4	6.5



观测值的散点图及其拟合直线

一元线性回归分析结果输出

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.977 [a]	.954	.948	.1918

a　Predictors: (Constant), X

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	6.055	1	6.055	164.655	.000
	Residual	.294	8	3.677E-02		
	Total	6.349	9			

a　Predictors: (Constant), X

b　Dependent Variable: Y

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	.607	.189		3.206	.013
	X	.542	.042	.977	12.832	.000

a　Dependent Variable: Y