§7.1　一元线性回归

一、变量间的两类关系

19世纪，英国生物学家兼统计学家高尔顿研究了父与子身高的遗传问题. 他观察了1078对数据 (x, y)，发现这些数据在直角坐标系上在一条直线附近，并求得直线方程：

$$\hat{y} = 33.73 + 0.516x \quad (\text{英寸}=2.54\text{cm})$$

这表明：(1) 父亲身高增加1个单位其儿子的平均身高增加0.516单位.

(2) 高个子父辈有生高个子儿子的趋势，但是一群高个子父辈的儿子们的平均身高要低于父辈

的平均高度．比如：

$$x = 80, \text{那么 } \hat{y} = 75.01.$$

（3）低个子父辈的儿子们虽为低个子，但其平均身高要比父辈的平均高度高一些．比如：

$$x = 60, \text{那么 } \hat{y} = 64.69$$

这便是子代的平均身高有向中心回归的意思，使得一段时间内人的身高相对稳定. 以后回归分析的思想渗透到数理统计的其他分支中, 随着计算机的发展和各种统计包的出现，回归分析的应用越来越广泛.

回归分析处理的是变量与变量之间的关系．变量之间的关系主要有两类：

（1）确定性关系（函数关系）$y = f(x)$，

圆的面积与半径，欧姆定律$\mathrm{V} = \mathrm{I}\,\mathrm{R}$等.

（2）统计相关关系

身高与体重, 消费与收入, 商品需求量与价格等.

变量的相关关系不能用函数关系来表示，但在平均意义下有一定的定量关系式. 例如：身高 1.7 米的人，他们的平均体重是一定的，即平均体重是身高的函数，寻找这种定量表达式就是回归分析的主要任务.

回归分析就是研究变量之间相关关系的一门科学，它通过大量的观察数据或试验数据，去寻找隐藏在数据背后的相关关系，给出它们的函数表达式——回归函数的估计.

二　一元线性回归模型

设变量 y 与 x 之间有相关关系，称 x 为自变量(预报变量)，称 y 为因变量(响应变量). 在知道 x 的取值后，y 的取值并不是确定的, 它是一个随机变量, 因此它有一个分布. 设其密度函数 $p(y|x)$，我们关心的是 y 的均值 $E(y|x)$，它是 x 的函数.

$$f(x)=\int_{-\infty}^{+\infty} yp(y|x)\mathrm{d}y$$

这个函数是确定的.

这便是y 关于x 的理论回归函数——条件期望. 也就是我们要求的回归函数的表达式.

以上的讨论是在x 与y 都是随机变量的场合, 这是一类回归问题, 还有一类回归问题：自变量x 是可控变量(一般变量), 只有因变量y (响应变量)是随机变量，它们之间的相关关系表示为

$$y=f(x)+\varepsilon$$

其中ε是随机误差，一般假设$\varepsilon \sim N(0,\sigma^2)$. 由于

ε是随机变量，因此y 是随机变量．本章主要研究这一种回归分析.

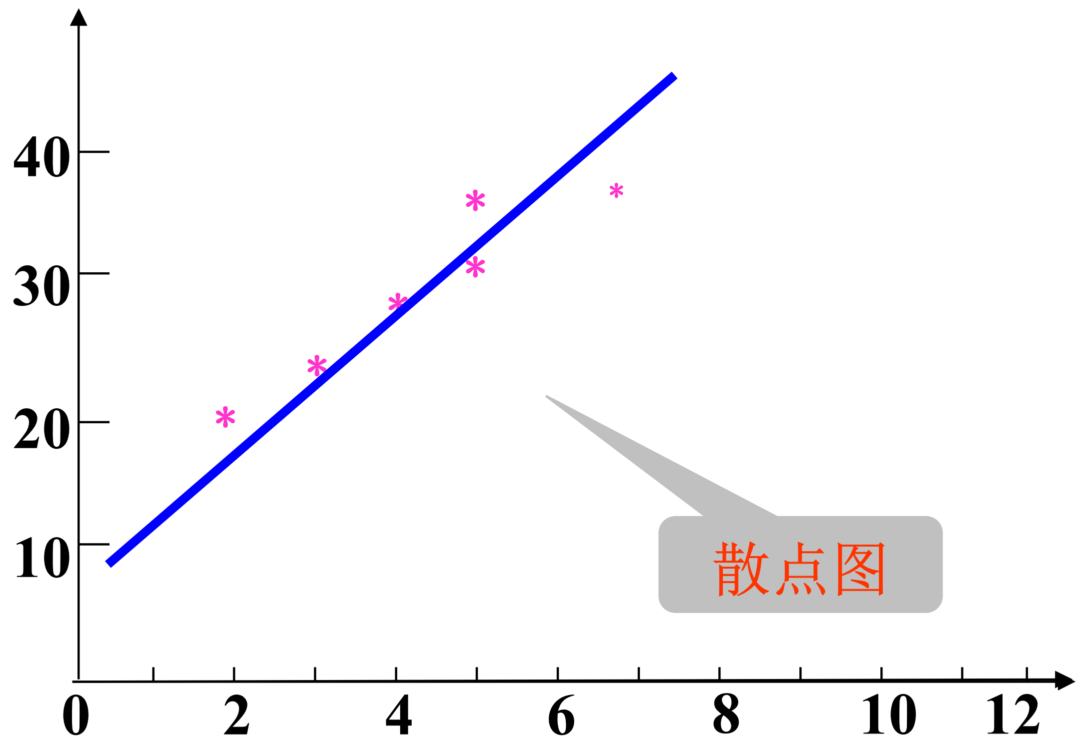
进行回归分析首先是回归函数的选择，当只有一个自变量时, 通常可采用画散点图的方法来选择.

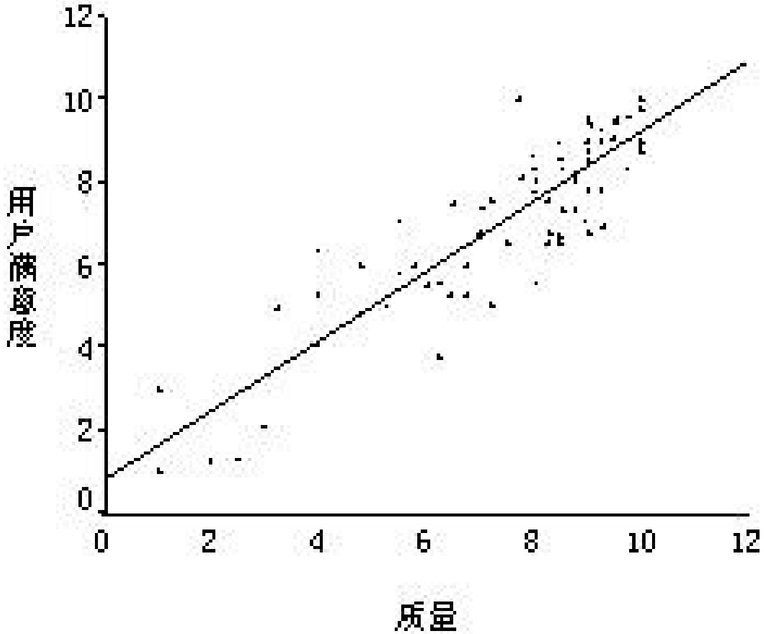
把每对观测数据(x_i, y_i)看成直角坐标系中的一个点. n对观测数据的构成的图叫散点图．见下图

例1 某公司的年科研经费与利润的关系如下表

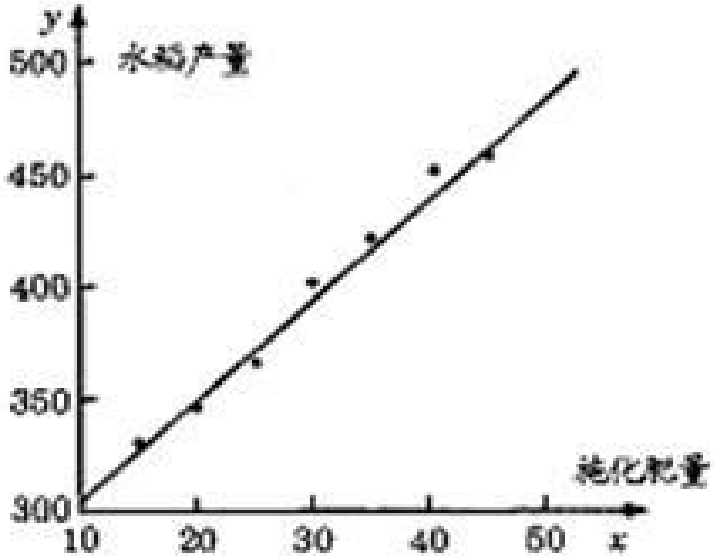
（单位：十万元）

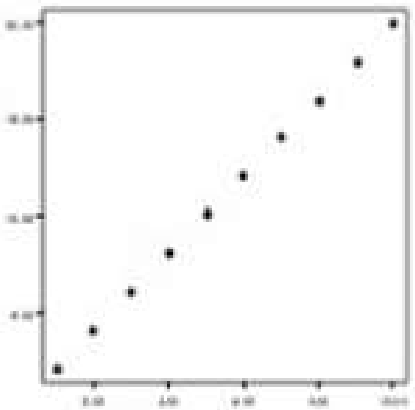
科研经费	2	3	5	4	5	7
利　润	20	25	34	30	31	35



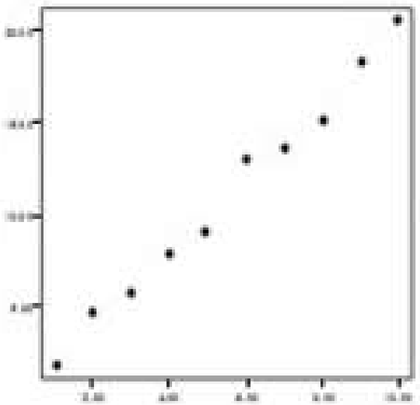


质量和用户满意度散点图

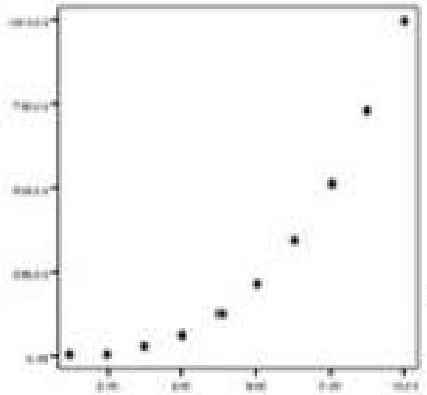




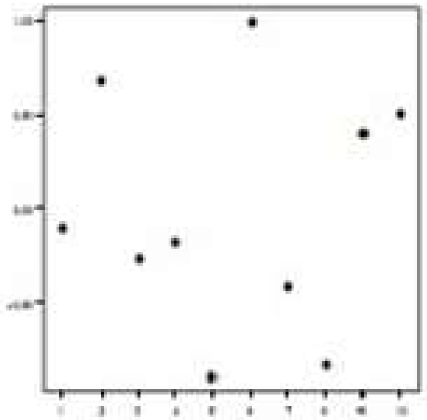
完全线性相关



线性相关



非线性相关



不相关

如果从散点图发现n个点$(x_1, y_1), (x_2, y_2), \cdots\cdots, (x_n, y_n)$在一条直线的附近，则说明两个变量之间是线性相关关系，这个相关关系可表示为

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

这里我们总是假定x为一般变量，是非随机变量，其值是可以精确测量或严格控制的，β_0, β_1是未知参数. β_1是直线的斜率，它表示x每增加一个单位, $E(y)$的增加量。ε是随机误差，一般假设

$$E(\varepsilon)=0, \quad D(\varepsilon)=\sigma^2,$$

在对未知参数进行估计或假设检验时，还需要假设

误差服从正态分布. 即

$$y=\beta_0+\beta_1 x+\varepsilon \sim N(\beta_0+\beta_1 x,\sigma^2).$$

在搜集数据时,通常要求独立进行，即假定$y_1, y_2,$ $\cdots, y_n$相互独立。由此得到一元线性回归模型.

$$\begin{cases} y_i=\beta_0+\beta_1 x_i+\varepsilon_i, \qquad i=1,2,\cdots,n \\ \text{诸}\varepsilon_i\text{相互独立，其分布为}N(0,\sigma^2). \end{cases}$$

根据n对数据(x_i, y_i),可以得到β_0, β_1的估计$\hat{\beta}_0, \hat{\beta}_1$, 称

$$\hat{y}=\hat{\beta}_0+\hat{\beta}_1 x$$

为y关于x的经验回归方程, 简称回归方程. 其图形称为回归直线. 给定$x=x_0$后,称$\hat{y}_0=\hat{\beta}_0+\hat{\beta}_1 x_0$为回归值.

三　回归系数的最小二乘估计

一般采用最小二乘方法估计模型中的参数β_0,β_1. 令

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

$\hat{\beta}_0,\hat{\beta}_1$应该满足　　$Q(\hat{\beta}_0,\hat{\beta}_1)=\min\limits_{\beta_0,\beta_1} Q(\beta_0,\beta_1),$

得到的$\hat{\beta}_0,\hat{\beta}_1$称为$\beta_0,\beta_1$的最小二乘估计, 记为LSE. 令

$$\begin{cases}\dfrac{\partial Q}{\partial \hat{\beta}_0}=-2\sum\limits_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1 x_i)=0\\[2ex] \dfrac{\partial Q}{\partial \hat{\beta}_1}=-2\sum\limits_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1 x_i)x_i=0\end{cases}$$

这组方程称为正规方程组，经整理得

$$\begin{cases} n\hat{\beta}_0 + n\bar{x}\hat{\beta}_1 = n\bar{y} \\ n\bar{x}\hat{\beta}_0 + \sum_{i=1}^{n} x_i^2 \hat{\beta}_1 = \sum_{i=1}^{n} x_i y_i \end{cases}$$

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

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

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

求解正规方程组得 $$\begin{cases} \hat{\beta}_1 = l_{xy} / l_{xx} \\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \end{cases}$$

这就是参数的β_0,β_1的最小二乘估计. 记为**LSE.**

最小二乘估计$\hat{\beta}_0$，$\hat{\beta}_1$有下列性质.

定理7.1　在模型的假设条件下，有

（1）$\hat{\beta}_0 \sim N(\beta_0,(\frac{1}{n}+\frac{\bar{x}^2}{l_{xx}})\sigma^2)$，$\hat{\beta}_1 \sim N(\beta_1,\frac{\sigma^2}{l_{xx}})$;

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

（3）对给定的x_0,

$$\hat{y}_0=\hat{\beta}_0+\hat{\beta}_1x_0$$

$$\sim N(\beta_0+\beta_1x_0,(\frac{1}{n}+\frac{(x_0-\bar{x})^2}{l_{xx}})\sigma^2).$$

定理7.1表明

(1) $\hat{\beta}_0,\hat{\beta}_1$分别是$\beta_0,\beta_1$的无偏估计；

(2) $\hat{y}_0$是$E(y_0)=\beta_0+\beta_1 x_0$的无偏估计;

(3) 除$\bar{x}=0$外，$\hat{\beta}_0$与$\hat{\beta}_1$是相关的；

(4) 要提高$\hat{\beta}_0,\hat{\beta}_1$的估计精度，要求$n$，$l_{xx}$要大.

证明：利用$\sum_{i=1}^{n}(x_i-\bar{x})=0$，可将$\hat{\beta}_1$和$\hat{\beta}_0$改写为

$$\hat{\beta}_1=\sum\frac{x_i-\bar{x}}{l_{xx}}y_i,\quad \hat{\beta}_0=\sum\left[\frac{1}{n}-\frac{(x_i-\bar{x})\bar{x}}{l_{xx}}\right]y_i$$

它们都是独立正态变量$y_1,\cdots,y_n$的线性组合, 故都服从正态分布. 下面分别求它们的均值与方差.

$$E(\hat{\beta}_1)=\sum\frac{x_i-\overline{x}}{l_{xx}}E(y_i)=\sum\frac{x_i-\overline{x}}{l_{xx}}(\beta_0+\beta_1 x_i)=\beta_1\quad;$$

$$D(\hat{\beta}_1)=\sum(\frac{x_i-\overline{x}}{l_{xx}})^2D(y_i)=\sum(\frac{x_i-\overline{x}}{l_{xx}})^2\sigma^2=\frac{\sigma^2}{l_{xx}}\quad;$$

$$E(\hat{\beta}_0)=E(\overline{y})-E(\hat{\beta}_1\overline{x})=(\beta_0+\beta_1\overline{x})-\beta_1\overline{x}=\beta_0\quad;$$

$$D(\hat{\beta}_0)=\sum(\frac{1}{n}-\frac{(x_i-\overline{x})\overline{x}}{l_{xx}})^2D(y_i)=(\frac{1}{n}+\frac{\overline{x}^2}{l_{xx}})\sigma^2\quad.$$

这就证明了(1). 进一步，利用诸y_i的独立性得

$$\mathrm{Cov}(\hat{\beta}_0,\hat{\beta}_1)=\mathrm{Cov}(\sum\left[\left(\frac{1}{n}-\frac{(x_i-\overline{x})\overline{x}}{l_{xx}}\right)y_i,\ \sum\frac{x_i-\overline{x}}{l_{xx}}y_i\right]$$

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

这就证明了(2). 下面证明(3), 注意到 $\hat{y}_0=\hat{\beta}_0+\hat{\beta}_1x_0$ 也是 $y_1,\cdots,y_n$的线性组合，故它也服从正态分布.

$$E(\hat{y}_0)=E(\hat{\beta}_0+\hat{\beta}_1x_0)=\beta_0+\beta_1x_0$$

$$D(\hat{y}_0)=D(\hat{\beta}_0+\hat{\beta}_1x_0)$$

$$=D(\hat{\beta}_0)+x_0^2D(\hat{\beta}_1)+2x_0\mathrm{Cov}(\hat{\beta}_0,\hat{\beta}_1)$$

$$(\frac{1}{n}+\frac{\bar{x}^2}{l_{xx}})\sigma^2+x_0^2\frac{\sigma^2}{l_{xx}}-2x_0\frac{\bar{x}}{l_{xx}}=(\frac{1}{n}+\frac{(x_0-\bar{x})^2}{l_{xx}})\sigma^2.$$

四　回归方程的显著性检验

从回归系数的最小二乘估计公式知，对任意给出的n 对数据(x_i, y_i),都可以求出$\hat{\beta}_0, \hat{\beta}_1$，从而得到回归方程到 $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$，但这样给出的回归方程不一定有意义(通过散点图来判断，毕竟带有主观性).

什么叫回归方程有意义呢？建立回归方程的目的是寻找y 的均值随x变化的规律，即寻找回归方程$E(y) = \beta_0 + \beta_1 x$. 如果$\beta_1 = 0$，那么不管$x$ 如何变化，$E(y)$不随x 的变化作线性变化，此时求得的回归方程没有意义, 称回归方程不显著. 如果$\beta_1 \neq 0$, 那么当

x 变化时, $E(y)$随 x 的变化作线性变化，此时求得的回归方程就有意义，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进行如下的显著性检验:

$$H_0:\beta_1=0 \Leftrightarrow H_1:\beta_1 \neq 0$$

拒绝表示回归方程是显著的.

（一）平方和分解式

运用方差分析的思想, 研究各y_i不同的原因, 记$\hat{y}_i=\hat{\beta}_0+\hat{\beta}_1 x_i$为回归值，$y_i-\hat{y}_i$为残差.

数据总的波动大小用总偏差平方和

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

来表示. 引起各y_i不同的原因主要有两方面：其一是H_0可能不真，$E(y)$随 x 的变化而变化，从而在每一个x_i处的观测值的回归值$\hat{y}_i$不同，其波动用回归平方和

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

表示；其二是其他因素(随机误差，x 对$E(y)$的非线性影响等). 这时可用残差平方和

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

表示. 由于$\hat{\beta}_0,\hat{\beta}_1$是正规方程的解，因此

$$\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1 x_i)=0 \Rightarrow \sum_{i=1}^{n}(y_i-\hat{y}_i)=0$$

$$\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1 x_i)x_i=0 \Rightarrow \sum_{i=1}^{n}(y_i-\hat{y}_i)x_i=0$$

利用 $\hat{y}_i=\hat{\beta}_0+\hat{\beta}_0 x_i=\overline{y}+\hat{\beta}_1(x_i-\overline{x})$,可得

$$\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\overline{y})=\sum_{i=1}^{n}(y_i-\hat{y}_i)\left[\hat{\beta}_1(x_i-\overline{x})\right]=0$$

从而 $$S_T=\sum_{i=1}^{n}(y_i-\overline{y})^2=\sum_{i=1}^{n}\left[(y_i-\hat{y}_i)+(\hat{y}_i-\overline{y})\right]^2$$

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

即

$$S_T = S_R + S_e$$

上式即为一元线性回归分析的平方和分解式.

S_R和S_e有以下的性质：

定理7.2　设 $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$，诸$\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$相互独立，且 $E(\varepsilon_i) = 0$，　$D(\varepsilon_i) = \sigma^2$，$i = 1, 2, \cdots, n$，沿用上面的记号，有

$$E(S_R) = \sigma^2 + \beta_1^2 l_{xx}, \qquad E(S_e) = (n-2)\sigma^2.$$

这说明 $\hat{\sigma}^2 = S_e/(n-2)$是$\sigma^2$的无偏估计.

证明 $$S_R=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2=\sum_{i=1}^{n}\left[\hat{\beta}_1(x_i-\bar{x})\right]^2=\hat{\beta}_1^{\,2}l_{xx}$$

$$E(S_R)=E(\hat{\beta}_1^{\,2})l_{xx}=(\beta_1^2+\frac{\sigma^2}{l_{xx}})l_{xx}=\sigma^2+\beta_1^2 l_{xx};$$

由于 $$E(S_T)=E\left[\sum_{i=1}^{n}(y_i-\bar{y})^2\right]=E\left[\sum_{i=1}^{n}y_i^2-n\bar{y}^2\right]$$

$$=\sum_{i=1}^{n}\left[(\beta_0+\beta_1x_i)^2+\sigma^2\right]-n\left[(\beta_0+\beta_1\bar{x})^2+\sigma^2/n\right]$$

$$=(n-1)\sigma^2+\beta_1^2 l_{xx}$$

从而 $$E(S_e)=E(S_T)-E(S_R)=(n-2)\sigma^2.$$

定理7.3 设 $y_i \sim N(\beta_0+\beta_1 x_i, \sigma^2)$，$i=1,\cdots,n$，且 $y_1, y_2, \cdots, y_n$相互独立, 在上面的记号下, 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）.

证明　取$n \times n$ 的正交矩阵A，具有下列形式

$$A=\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ \vdots & \vdots & & \vdots \\ a_{n-2,1} & a_{n-2,2} & \cdots & a_{n-2,n} \\ (x_1-\bar{x})/\sqrt{l_{xx}} & (x_2-\bar{x})/\sqrt{l_{xx}} & \cdots & (x_n-\bar{x})/\sqrt{l_{xx}} \\ 1/\sqrt{n} & 1/\sqrt{n} & \cdots & 1/\sqrt{n} \end{pmatrix}$$

注：这样的正交矩阵是存在的，记上述矩阵**A**的最后两行向量为α,β．易知α,β是2维线性空间$\boldsymbol{L}(\alpha,\beta)$的标准正交基，再取正交子空间$\boldsymbol{L}^{\perp}(\alpha,\beta)$的一组标准正交基作为**A**的前$\boldsymbol{n-2}$个行向量，即构成正交矩阵.

令
$$\boldsymbol{Z}=\begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{pmatrix}=\boldsymbol{AY}=\boldsymbol{A}\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

其中
$$z_{n-1}=\sum\frac{x_i-\bar{x}}{\sqrt{l_{xx}}}y_i=\frac{l_{xy}}{\sqrt{l_{xx}}}=\hat{\beta}_1\sqrt{l_{xx}};$$
$$z_n=\frac{1}{\sqrt{n}}\sum y_i=\sqrt{n}\,\bar{y}.$$

则Z仍然服从正态分布，其数学期望和方差分别为

$$E(z_1) = E\left(\sum a_{1j} y_j\right) = \sum a_{1j}(\beta_0 + \beta_1 x_j)$$

$$= (\beta_0 + \beta_1 \bar{x}) \sum a_{1j} + \beta_1 \sum a_{1j}(x_j - \bar{x}) = 0$$

同样得 $E(z_2) = \cdots = E(z_{n-2}) = 0.$

而 $E(z_{n-1}) = E(\hat{\beta}_1 \sqrt{l_{xx}}) = \beta_1 \sqrt{l_{xx}},$

$$E(z_n) = E(\sqrt{n}\bar{y}) = \sqrt{n}(\beta_0 + \beta_1 \bar{x}).$$

因而

$$E(Z) = \left(0, \cdots, 0, \sqrt{l_{xx}}\beta_1, \sqrt{n}(\beta_0 + \beta_1 \bar{x})\right)^T$$

$$D(Z) = D(AY) = AD(Z)A^T = \sigma^2 \mathrm{I}_n.$$

这表明$z_1,\cdots,z_n$相互独立, $z_1,\cdots,z_{n-2}$都服从$N(0,\sigma^2)$,

$z_{n-1}\sim N(\beta_1\sqrt{l_{xx}},\sigma^2)$, $\quad z_n\sim N(\sqrt{n}(\beta_0+\beta_1\bar{x}),\sigma^2)$.

由于
$$\sum z_i^2=\sum y_i^2=S_T+n\bar{y}^2=S_e+S_R+n\bar{y}^2,$$

而
$$S_R=\hat{\beta}_1^2 l_{xx}=(\hat{\beta}_1\sqrt{l_{xx}})^2=z_{n-1}^2,\quad n\bar{y}^2=(\sqrt{n}\bar{y})^2=z_n^2.$$

于是有

$S_e=z_1^2+z_2^2+\cdots+z_{n-2}^2$,且$S_e$，$S_R$，$\bar{y}$三者相互独立.

$$S_e/\sigma^2=(z_1/\sigma)^2+(z_2/\sigma)^2+\cdots+(z_{n-2}/\sigma)^2\sim\chi^2(n-2);$$

在$\beta_1=0$ 时，有 $S_R/\sigma^2=(z_{n-1}/\sigma)^2\sim\chi^2(1)$;

定理证毕.

（二）检验统计量与拒绝域

和方差分析一样，可考虑采用F作为检验统计量

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

当H_0成立时，$F\sim F(1,n-2)$. 对于给定的显著性水平α，拒绝域为

$$W=\{F>F_\alpha(1,n-2)\}$$

整个检验可以列成方差分析表.

来源	平方和	自由度	均方和	F 比	临界值	显著性
回归	S_R	1	S_R	$F=\frac{S_R}{S_e/n-2}$	F_α	
残差	S_e	n-2	S_e/n-2			
总和	S_T	n-1				

注：(1)对$H_0:\beta_1=0$的检验也可以基于 t 分布进行，由于$\hat{\beta}_1\sim N(\beta_1,\sigma^2/l_{xx})$，$S_e/\sigma^2\sim\chi^2(n-2)$,且$\hat{\beta}_1$，$S_e$相互独立，因此$H_0$为真时，有 $t=\dfrac{\hat{\beta}_1}{\hat{\sigma}/\sqrt{l_{xx}}}\sim t(n-2)$

其中$\hat{\sigma}=\sqrt{S_e/(n-2)}$，拒绝域为

$$W=\left\{|t|>t_{\alpha/2}(n-2)\right\}$$

它与F检验等价.

(2)也可以利用相关系数检验 $H_0:\rho=0\Leftrightarrow H_1:\rho\neq 0$

利用样本相关系数$r=l_{xy}\big/\sqrt{l_{xx}}\sqrt{l_{yy}}$.　　拒绝域为

$$W=\left\{|r|>r_{\alpha/2}(n-2)\right\}\qquad (\text{注：}r^2=\frac{F}{F+(n-2)}).$$

四　估计与预测

当回归方程经过检验是显著的时，可以用来做估计和预测．这是两个不同的问题：

• 当$x=x_0$时，寻找 $E(y_0)=\beta_0+\beta_1 x_0$的点估计与区间估计，这是估计问题；(注：这里$E(y_0)$是常量).

• 当$x=x_0$时，y_0的观察值在什么范围内？由于y_0是随机变量，只能找一个区间，使得y_0落在该区间内的概率为$1-\alpha$，即要求δ,使$P\left(\left|y_0-\hat{y}_0\right|\le\delta\right)=1-\alpha$.

称区间$\left[\hat{y}_0-\delta,\ \hat{y}_0+\delta\right]$为$y_0$的概率为$1-\alpha$的预测区间,这是预测问题;.

（一）$E(y_0)$的估计

当$x=x_0$时，对应的因变量y_0是一个随机变量，其均值为$\beta_0+\beta_1 x_0$，故可将$\hat{y}_0=\hat{\beta}_0+\hat{\beta}_1 x_0$作为$E(y_0)$的点估计，且由于

$$E(\hat{y}_0)=E(\hat{\beta}_0+\hat{\beta}_1 x_0)=\beta_0+\beta_1 x_0=E(y_0)$$

因此该估计是无偏估计.

为得到$E(y_0)$的区间估计，我们根据定理5.1可得$\hat{y}_0$的分布

$$\hat{y}_0 \sim N(\beta_0+\beta_1 x_0,\left[\frac{1}{n}+\frac{(x_0-\overline{x})^2}{l_{xx}}\right]\sigma^2)$$

再利用定理5. 3知，　$S_e/\sigma^2 \sim \chi^2(n-2)$，且与

$\hat{y}_0 = \overline{y} + \hat{\beta}_1(x_0 - \overline{x})$相互独立，记$\hat{\sigma}^2 = S_e/(n-2)$.可得

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

于是$E(y_0)$的置信度为$1-\alpha$的置信区间为

$$\left[\hat{y}_0 - \delta_0,\ \hat{y}_0 + \delta_0\right]$$

其中 $\delta_0 = t_{\alpha/2}(n-2)\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{l_{xx}}}$.

(二)y_0的预测区间

当$x = x_0$时，$y_0 = \beta_0 + \beta_1 x_0 + \varepsilon_0$是一个随机变量,通常假定$\varepsilon_0 \sim N(0, \sigma^2)$, 由于$y_0$与$\hat{y}_0$相互独立，因此

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

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

从而y_0的概率为$1-\alpha$ 的预测区间为

$$\left[\hat{y}_0-\delta,\ \hat{y}_0+\delta\right],$$

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

上面的预测区间要比$E(y_0)$的置信区间宽一些.

五　实例分析

例2 在动物学研究中，有时需要找出某种动物的体积与重量的关系．因为动物的重量相对而言容易测量，而测量体积比较困难，因此，人们希望用动物的总量去预测体积.

下面是**18**只某种动物的体积与重量数据，在这里, 动物重量被看作自变量, 用x表示，单位是**kg**, 动物体积则看作是因变量，用y表示，单位是**dm**立方. **18**组数据的数据列于下表．试建立y与x之间的线性回归模型，给出$x_0 = 17.6$**kg**时, 动物体积的预测值及概率为**95%**的预测区间.

表　18只某种动物重量 x 与体积 y 的数据

x	y	x	y	x	y
10.4	10.2	15.1	14.8	16.5	15.9
10.5	10.4	15.1	15.1	16.7	16.6
11.9	11.6	15.1	14.5	17.1	16.7
12.1	11.9	15.7	15.7	17.1	16.7
13.8	13.5	15.8	15.2	17.8	17.6
15.0	14.5	16.0	15.8	18.4	18.3

解：从散点图(略)知：这18个点基本在一条直线附近，这说明重量x与体积y之间存在着线性关系，下面求该线性回归方程. 计算过程如下表.

$\sum x_i = 270.1$	$n = 18$	$\sum y_i = 265.0$
$\overline{x} = 15.0056$		$\overline{y} = 14.7222$
$\sum x_i^2 = 4149.39$	$\sum x_i y_i = 4071.71$	$\sum y_i^2 = 4149.39$
$n\overline{x}^2 = 4053.0006$	$n\overline{x}\overline{y} = 3976.4722$	$n\overline{y}^2 = 3901.3889$
$l_{xx} = 96.3894$	$l_{xy} = 95.2378$	$l_{yy} = 94.7511$
$\hat{\beta}_1 = l_{xy}/l_{xx} = 0.9881$		$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1\overline{x} = -0.104$

由此给出回归方程为

$$\hat{y} = -0.1048 + 0.9881x$$

接下来是关于回归方程的显著性检验，经计算有

$$S_T = l_{yy} = 94.7511 \qquad f_T = 17,$$

$$S_R = \hat{\beta}_1^2 l_{xx} = \hat{\beta}_1 l_{xy} = 94.1090 \qquad f_R = 1,$$

$$S_e = S_T - S_R = 0.6421 \qquad f_e = 16.$$

显著性水平$\alpha = 0.01$，$F_{0.01}(1,16) = 8.53$，由于$F > F_\alpha$，因此回归方程是高度显著的，检验的方差分析表如下.

来源	平方和	自由度	均方和	F 比	临界值	显著性
回归	94.109	1	94.109	2346.9	8.53	**
残差	0.6421	16	0.0401			
总和	94.7511	17				

当$x_0 = 17.6$kg时，动物体积 y_0的预测值为

$$\hat{y}_0 = -0.1048 + 0.9881 \times 17.6 = 17.2858$$

由于

$$t_{0.025}(16) = 2.1199, \quad \text{又} \hat{\sigma} = \sqrt{0.0401} = 0.2002.$$

根据公式得

$$\delta = \hat{\sigma} t_{0.025}(16) \sqrt{1 + \frac{1}{18} + \frac{(x_0 - \overline{x})^2}{l_{xx}}} = 0.4776.$$

因此 y_0概率为95%的预测区间为

$$(\hat{y}_0 - \delta, \hat{y}_0 + \delta) = (16.8082,\ 17.7634).$$