第 9 章 回归分析

9.1 内容提要

9.1.1 一元线性回归

1. 一元线性回归模型

在模型 $\begin{cases} y=\mu(x)+\varepsilon \\ E(\varepsilon)=0 \end{cases}$ 中, 如果 $\mu(x)$ 是 x 的线性函数, ε 服从正态分布, 则称该模型为一元线性回归模型, 它具有如下的形式

$$\begin{cases} y=ax+b+\varepsilon \\ \varepsilon \sim N(0,\sigma^2) \end{cases}$$

其中 a,b,σ^2 是与 x 无关的未知参数, a,b 称为回归系数. 称 $\hat{y}=ax+b$ 为一元线性理论回归模型, 或称 $\mu(x)=E(y)=ax+b$ 为 y 关于 x 的回归函数.

2. 未知参数的估计及统计性质

(1) 最小二乘法

构造如下的偏差平方和 $Q(a,b)=\sum_{i=1}^{n}(y_i-(a+bx_i))^2$, 最小二乘法就是选择 a, b 的估计 $\hat{a}$, $\hat{b}$ 使得 $Q(\hat{a},\hat{b})=\min\limits_{a,b} Q(a,b)$.

分别求 $Q(a,b)$ 关于 a, b 的偏导数, 并令它们等于零, 计算得到 a, b 的估计值:

$$\begin{cases} \hat{b}=\dfrac{\sum\limits_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2}=\dfrac{\sum\limits_{i=1}^{n}x_iy_i-n\overline{x}\cdot\overline{y}}{\sum\limits_{i=1}^{n}x_i^2-n\overline{x}^2}, \\ \hat{a}=\overline{y}-\hat{b}\overline{x}, \end{cases}$$

其中 $\overline{x}=\dfrac{1}{n}\sum\limits_{i=1}^{n}x_i$, $\overline{y}=\dfrac{1}{n}\sum\limits_{i=1}^{n}\overline{y}_i$. 由上式所确定的估计 $\hat{a}$, $\hat{b}$ 称为回归系数 a, b 的最小二乘估计, 该估计方法称为最小二乘法. 若记

$$L_{xx}=\sum_{i=1}^{n}(x_i-\overline{x})^2=\sum_{i=1}^{n}x_i^2-n\overline{x}^2,$$

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

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

(2) 最小二乘估计的性质

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

$Cov(\bar{y},\hat{b})=0$ 且 $\bar{y}$ 与 $\hat{b}$ 相互独立,

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

$Cov(\hat{a},\hat{b})=-\frac{\bar{x}}{l_{xx}}\sigma^2$.

(3) σ^2 的无偏估计

平方和 $S_e=\sum(y_i-\hat{y}_i)^2=\sum(y_i-a-\hat{b}x_i)^2$ 称为残差平方和或剩余平方和.

$E(S_e)=(n-2)\sigma^2$,

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

由此可以得到 σ^2 的一个无偏估计量: $\hat{\sigma}^2=\frac{S_e}{n-2}$.

3. 回归效果的显著性检验

(1) 平方和分解公式

称 $S_T=\sum(y_i-\bar{y})^2$ 为总偏差平方和, 称 $S_R=\sum(\hat{y}_i-\bar{y})^2$ 为回归平方和,

称 $S_e=\sum(y_i-\hat{y}_i)^2$ 为残差平方和,

且有平方和分解公式 $S_T=S_e+S_R$.

(2) 回归效果的显著性检验

F 检验法:

当原假设 H_0 为真时, $\frac{S_R}{\sigma^2}\sim\chi^2(1)$, $\frac{S_e}{\sigma^2}\sim\chi^2(n-2)$, 且 S_R 与 S_e 相互独立, 从而

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

对于给定的显著性水平 α，拒绝域为 $F=\dfrac{S_R}{S_e/(n-2)}\geq F_\alpha(1,n-2)$.

t 检验法：

由 $\hat{b}\sim N(b,\dfrac{\sigma^2}{l_{xx}})$ 知，$\dfrac{\hat{b}-b}{\sigma}\sqrt{l_{xx}}\sim N(0,1)$. 又由 $\dfrac{S_e}{\sigma^2}\sim\chi^2(n-2)$，且 $\hat{b}$ 与 S_e 相互独立，因此得到 $\dfrac{\hat{b}-b}{\sqrt{S_e/(n-2)}}\sqrt{l_{xx}}\sim t(n-2)$.

取检验统计量 $t=\dfrac{\hat{b}}{\sqrt{S_e/(n-2)}}\sqrt{l_{xx}}$，当原假设 H_0 为真时，$t\sim t(n-2)$. 对于给定的显著性水平 α，拒绝域为 $|t|\geq t_{\frac{\alpha}{2}}(n-2)$.

r 检验法：

x 与 y 的相关系数 $r=\dfrac{l_{xy}}{\sqrt{l_{xx}l_{yy}}}=\dfrac{\sum(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum(x_i-\overline{x})^2\sum(y_i-\overline{y})^2}}$，

由于 $r^2=\dfrac{S_R}{S_T}$，则

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

因此 F 检验的拒绝域 $F\geq F_\alpha(1,n-2)$ 等价于 $|r|\geq(\dfrac{n-2}{F_\alpha(1,n-2)}+1)^{-\frac{1}{2}}$.

(3) 回归系数的置信区间

当 σ^2 未知时，回归系数 a 的置信度为 $1-\alpha$ 的置信区间为：

$$(\hat{a}-t_{\frac{\alpha}{2}}(n-2)\sqrt{\frac{1}{n}+\frac{\overline{x}^2}{l_{xx}}}\sqrt{\frac{S_e}{n-2}},a+t_{\frac{\alpha}{2}}(n-2)\sqrt{\frac{1}{n}+\frac{\overline{x}^2}{l_{xx}}}\sqrt{\frac{S_e}{n-2}}).$$

当 σ^2 未知时，回归系数 b 的置信度为 $1-\alpha$ 的置信区间为：

$$(\hat{b}-t_{\frac{\alpha}{2}}(n-2)\frac{1}{\sqrt{l_{xx}}}\sqrt{\frac{S_e}{n-2}},b+t_{\frac{\alpha}{2}}(n-2)\frac{1}{\sqrt{l_{xx}}}\sqrt{\frac{S_e}{n-2}}).$$

(4) 预测

对于给定的 x_0，y_0 的置信度为 $1-\alpha$ 的置信区间为 $(\hat{y}_0-\delta(x_0),y_0+\delta(x_0))$，其中

$$\delta(x_0)=t_{\frac{\alpha}{2}}(n-2)S\sqrt{1+\frac{1}{n}+\frac{(x_0-\overline{x})^2}{l_{xx}}}\,.$$

(5) 控制

控制是预测的反问题, 即要求观察值 y 在某一区间 (y_1,y_2) 内取值时, 问应将 x 控制在什么范围内. 上述问题实际上就是要确定下列方程组的解:

$$\begin{cases}\hat{a}+\hat{b}x_1-\delta(x_1)=y_1\\ \hat{a}+\hat{b}x_1+\delta(x_2)=y_2\end{cases}.$$

9.1.2 多元线性回归

1. 多元线性回归模型

设随机变量 y 与 $m\ (m\geq 2)$ 个自变量 $x_1,x_2,\cdots,x_m$ 之间存在相关关系, 且有

$$\begin{cases}y=a+b_1x_1+b_2x_2+\cdots+b_mx_m+\varepsilon\\ \varepsilon\sim N(0,\sigma^2)\end{cases}$$

其中 $a,b_1,b_2,\cdots,b_m,\sigma^2$ 是与 $x_1,x_2,\cdots,x_m$ 无关的未知参数, ε 是不可观测的随机变量. 称上式为 m 元线性回归模型.

设有 n 组不同的样本观测值 $(x_{i1},x_{i2},\cdots,x_{im};y_i)(i=1,2,\cdots,n)$, 令

$$\mathbf{y}=\begin{pmatrix}y_1\\ y_2\\ \vdots\\ y_n\end{pmatrix},\ \mathbf{X}=\begin{pmatrix}1 & x_{11} & x_{12} & \cdots & x_{1m}\\ 1 & x_{21} & x_{22} & \cdots & x_{2m}\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ 1 & x_{n1} & x_{n2} & \cdots & x_{nm}\end{pmatrix},\ \boldsymbol{\beta}=\begin{pmatrix}a\\ b_1\\ \vdots\\ b_m\end{pmatrix},\ \boldsymbol{\varepsilon}=\begin{pmatrix}\varepsilon_1\\ \varepsilon_2\\ \vdots\\ \varepsilon_n\end{pmatrix}.$$

则回归模型可以写成矩阵形式 $\begin{cases}\mathbf{y}=\boldsymbol{\beta}\mathbf{X}+\boldsymbol{\varepsilon}\\ \boldsymbol{\varepsilon}\sim N(\mathbf{0},\sigma^2\ \ _n)\end{cases}$.

2. 未知参数的估计及统计性质

可以证明 $\boldsymbol{\beta}$ 的最小二乘估计为 $\hat{\boldsymbol{\beta}}=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.

参数 $\boldsymbol{\beta}$ 的最小二乘估计具有如下性质 $\hat{\boldsymbol{\beta}}\sim N(\boldsymbol{\beta},\sigma^2(\mathbf{X}'\mathbf{X})^{-1})$.

3. 回归效果的显著性检验

(1) 检验 y 与 $x_1,x_2,\cdots,x_m$ 之间是否有线性关系, 就是要检验假设:

$$H_0:b_1=b_2=\cdots=b_m=0,\ \leftrightarrow\ H_1:b_i(i=1,2,\cdots,m)\text{不全为}0\,.$$

(2) 在多元线性回归模型下有下列结论:

$$E(S_e)=(n-m-1)\sigma^2,$$

$\frac{S_e}{\sigma^2} \sim \chi^2(n-m-1)$，且 S_e 与 $\hat{b_1}, b_2, \cdots, b_m$ 相互独立，

当原假设 H_0 为真时，$\frac{S_R}{\sigma^2} \sim \chi^2(m)$，且 S_R 与 S_e 相互独立.

(3) F 检验法

取检验统计量 $F = \frac{S_R / m}{S_e / (n-m-1)}$，当 H_0 为真时，$F \sim F(m, n-m-1)$. 因此，对于给定的显著性水平 α，拒绝域为 $F \geq F_\alpha(m, n-m-1)$.

9.1.3 可化为线性回归的曲线回归

1. 变量替换法

许多非线性模型可通过变量替换实现线性化，常见的变换如下所示：

原模型	变换函数	变换后模型
$y = \frac{1}{a+bx}$	$u = \frac{1}{y}$，$v = x$	$u = a + bv$
$y = \sqrt{a+bx}$	$u = y^2$，$v = x$	$u = a + bv$
$y = a + b_1 x + \cdots + b_m x^m$	$u = y$，$v_i = x^i$	$u = a + b_1 v_1 + \cdots + b_m v_m$
$y = a + b\ln x$	$u = y$，$v = \ln x$	$u = a + bv$
$y = cx^b$	$u = \ln y$，$v = \ln x$，$a = \ln c$	$u = a + bv$
$y = ce^{bx}$	$u = \ln y$，$v = x$，$a = \ln c$	$u = a + bv$

2. 判定系数

设 $(x_i, y_i)(i = 1, 2, \cdots, n)$ 为一组样本，通过回归分析后建立的曲线回归方程为 $\hat{y} = f(x)$，$\hat{y_1}, y_2, \cdots, y_n$ 为曲线回归方程用原始数据 $x_1, x_2, \cdots, x_n$ 算得回归值，则可以用判定系数 R^2 评价回归方程的拟合优劣程度，R^2 越接近于1，表明曲线拟合程度越好，其中判定系数为

$$R^2 = 1 - \frac{S_e}{S_T} = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \overline{y})^2}.$$

9.2 习题详解

1. 设 $(x_i, y_i)(i = 1, 2, \cdots, n)$ 是一组样本，$\hat{y_i} = a + \hat{b}x_i$ 是相应的线性回归方程，其中

$\hat{b}=\dfrac{l_{xy}}{l_{xx}}$，$\hat{a}=\overline{y}-\hat{b}\overline{x}$，试证下列恒等式：

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

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

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

(4) $S_e=\sum_{i=1}^{n}(y_i-\hat{a}-\hat{b}x_i)^2=\sum_{i=1}^{n}y_i^2-a\sum_{i=1}^{n}y_i-b\sum_{i=1}^{n}x_iy_i$.

解 (1) $\sum_{i=1}^{n}(y_i-\hat{y}_i)=\sum_{i=1}^{n}(y_i-a-\hat{b}x_i)=\sum_{i=1}^{n}(y_i-\overline{y}+\hat{b}\overline{x}-bx_i)$

$$=\sum_{i=1}^{n}y_i-n\overline{y}+\hat{b}\sum_{i=1}^{n}(\overline{x}-x_i)=0;$$

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

$$=\sum_{i=1}^{n}x_iy_i-\overline{y}\sum_{i=1}^{n}x_i-\hat{b}\sum_{i=1}^{n}(x_i-\overline{x})x_i$$

$$=\sum_{i=1}^{n}x_iy_i-\overline{y}\sum_{i=1}^{n}x_i-\hat{b}\sum_{i=1}^{n}(x_i-\overline{x})(x_i-\overline{x}+\overline{x})$$

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

$$=\sum_{i=1}^{n}x_i(y_i-\overline{y})-\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})=\overline{x}\sum_{i=1}^{n}(y_i-\overline{y})=0;$$

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

$$=\hat{a}\sum_{i=1}^{n}(y_i-y_i)+\hat{b}\sum_{i=1}^{n}(y_i-y_i)x_i=0;$$

(4) $S_e=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=\sum_{i=1}^{n}(y_i-a-\hat{b}x_i)^2=\sum_{i=1}^{n}(y_i^2-2\hat{a}\hat{y}_i-2\hat{b}\hat{x}_iy_i+(a+bx_i)^2)$

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

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

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

$$= \sum_{i=1}^{n} y_i^2 - 2\hat{a}\sum_{i=1}^{n} y_i - \hat{b}\sum_{i=1}^{n} x_i y_i + \sum_{i=1}^{n} \overline{y}^2 - b\overline{x}\sum_{i=1}^{n} y_i - b\overline{y}\sum_{i=1}^{n} x_i + nb\overline{xy}$$

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

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

$$= \sum_{i=1}^{n} y_i^2 - \hat{a}\sum_{i=1}^{n} y_i - \hat{b}\sum_{i=1}^{n} x_i y_i .$$

2. 假设回归直线过原点, 即一元线性回归模型为

$$y_i = bx_i + \varepsilon_i, i = 1, 2, \cdots, n,$$

$E(\varepsilon_i) = 0, D(\varepsilon_i) = \sigma^2$, 各观测值相互独立.

（1）写出 b 的最小二乘估计,并给出 σ^2 的无偏估计;

（2）对给定的 x_0, 其对应的因变量均值的估计为 $\hat{y}_0$,求 $D(\hat{y}_0)$.

解 （1） $Q(b) = \sum_{i=1}^{n} (y_i - bx_i)^2, \frac{\partial Q}{\partial b} = -2\sum_{i=1}^{n} (y_i - bx_i)x_i = 0,$

$$\Rightarrow \sum_{i=1}^{n} x_i y_i - b\sum_{i=1}^{n} x_i^2 = 0 \quad \Rightarrow \hat{b} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}.$$

$$S_e = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2, E(S_e) = (n-1)\sigma^2, \Rightarrow \hat{\sigma}^2 = \frac{S_e}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n} (y_i - y_i)^2 .$$

（2） $D(\hat{y}_0) = D(\hat{b}x_0) = x_0^2 D(b) = x_0^2 D(\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2})$

$$= x_0^2 D(\sum_{i=1}^{n} \frac{x_i}{\sum_{i=1}^{n} x_i^2} y_i) = x_0^2 \frac{\sum_{i=1}^{n} x_i^2}{(\sum_{i=1}^{n} x_i^2)^2} D(y_i) = \frac{x_0^2 \sigma^2}{\sum_{i=1}^{n} x_i^2}.$$

3. 某建材实验室在做陶粒混凝土强度试验中, 考察每立方米混凝土的水泥用量 $x(kg)$ 对 28 天后的混凝土抗压强度 $y(kg/cm)$ 的影响, 则得如下数据

x	150	160	170	180	190	200	210	220	230	240	250	260
y	56.9	58.3	61.6	64.6	68.1	71.3	74.1	77.4	80.2	82.6	86.4	89.7

(1) 求 y 对 x 的线性回归方程, 并问: 每立方米混凝土中每增加 $1kg$ 水泥时, 可提高的抗压强度是多少?

(2) 检验回归效果的显著性 $(\alpha = 0.05)$;

(3) 求相关系数 r, 并求回归系数 b 的 95% 的置信区间;

(4) 求 $x_0 = 225(kg)$ 时, y_0 的预测值及 95% 的预测区间.

解 (1) $\overline{x} = 205, \overline{y} = 72.6, l_{xx} = 14300, l_{yy} = 1323.82, l_{xy} = 4347$. 则

$$\hat{b} = \frac{l_{xy}}{l_{xx}} = 0.304, \hat{a} = \overline{y} - b\overline{x} = 10.28, y = 10.28 + 0.304x.$$

因此, 每立方米混凝土中每增加 $1kg$ 水泥时, 抗压强度可提高约 $0.304kg/cm^2$.

(2) F 检验法: H_0: 没有线性相关性, 即 $b = 0$, H_1: 具有线性相关性, 即 $b \neq 0$.

在 H_0 下, $F = \frac{(n-2)S_R}{S_e} \sim F(1, n-2)$, 取显著性水平 $\alpha = 0.05$ 时,

拒绝域为: $F \geq F_{0.05}(1, n-2)$, $n = 12$.

计算得 $S_R = \hat{b}^2 l_{xx} = 1321.4272, S_e = l_{yy} - b^2 l_{xx} = 2.3928, F = 5522.514, F_{0.05}(1,10) = 4.96$.

$F \geq F_{0.05}(1,10)$, 拒绝原假设, 即回归效果显著.

(3) $r = \frac{l_{xy}}{\sqrt{l_{xx} l_{yy}}} = 0.999$, $\hat{b} \sim N(b, \frac{\sigma^2}{l_{xx}}) \Rightarrow \frac{\hat{b} - b}{\sigma} \sqrt{l_{xx}} \sim N(0,1)$,

$$\frac{S_e}{\sigma^2} \sim \chi^2(n-2), \Rightarrow \frac{\frac{\hat{b} - b}{\sigma} \sqrt{l_{xx}}}{\sqrt{\frac{S_e/\sigma^2}{n-2}}} = \frac{\hat{b} - b}{\sqrt{\frac{S_e}{n-2}}} \sqrt{l_{xx}} \sim t(n-2),$$

可得回归系数 b 的置信度为0. 95的置信区间为:

$$(\hat{b} - t_{0.025}(10) \frac{1}{\sqrt{l_{xx}}} \sqrt{\frac{S_e}{10}}, b + t_{0.025}(10) \frac{1}{\sqrt{l_{xx}}} \sqrt{\frac{S_e}{10}}).$$

代入数值计算得回归系数 b 的置信度为0.95的置信区间为：$(0.2949, 0.3131)$.

(4) $x_0 = 225kg$，$\hat{y_0} = a + \hat{b}x_0 = 78.68$，

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

可得 y_0 的置信度为95%的置信区间为 $(\hat{y_0} - \delta(x_0), y_0 + \delta(x_0))$，

其中 $\delta(x_0) = t_{0.025}(n-2)\sqrt{\frac{S_e}{n-2}}\sqrt{1+\frac{1}{n}+\frac{(x_0-\overline{x})^2}{l_{xx}}}$.

则 y_0 的置信度为95%的置信区间为 $(77.47, 79.89)$.

4. 假设 x 是一可控变量，y 是一随机变量且服从正态分布，现在不同的 x 值下，分别对 y 进行观测，得数据如下：

x	0.25	0.37	0.44	0.55	0.60	0.62	0.68	0.70	0.73
y	2.57	2.31	2.12	1.92	1.75	1.71	1.60	1.51	1.53
x	0.75	0.82	0.84	0.87	0.88	0.90	0.95	1.00	
y	1.41	1.33	1.31	1.25	1.20	1.19	1.15	1.00	

(1) 求 y 对 x 的线性回归方程，并求 $\sigma^2 = D(y)$ 的无偏估计；

(2) 求回归系数 a, b 的置信度为95%的置信区间；

(3) 检验线性回归效果的显著性 $(\alpha = 0.05)$；

(4) 求 y 的置信度为95%的置信区间.

(5) 为了把观测值 y 限制在区间 $(1.08, 1.68)$，需要把 x 的值限制在什么范围之内？

解 (1) $\hat{b} = \frac{l_{xy}}{l_{xx}}$，$\hat{a} = \overline{y} - \hat{b}\overline{x}$，$\Rightarrow \hat{y} = 3.0332 - 2.0698x$，$\hat{\sigma}^2 = \frac{S_e}{n-2} = 0.0019$.

(2) 由于 $t = \frac{\hat{a}-a}{\sqrt{\frac{1}{n}+\frac{\overline{x}^2}{l_{xx}}}\sqrt{\frac{S_e}{n-2}}} \sim t(n-2)$，

回归系数 a 的置信度为95%的置信区间为：

$$(\hat{a} - t_{0.025}(n-2)\sqrt{\frac{1}{n}+\frac{\overline{x}^2}{l_{xx}}}\sqrt{\frac{S_e}{n-2}}, a + t_{0.025}(n-2)\sqrt{\frac{1}{n}+\frac{\overline{x}^2}{l_{xx}}}\sqrt{\frac{S_e}{n-2}}),$$

代入数值计算得 $(2.9671, 3.1117)$.

由于 $t = \frac{\hat{b}-b}{\sqrt{\frac{S_e}{n-2}}}\sqrt{l_{xx}} \sim t(n-2)$，回归系数 b 的置信度为95%的置信区间为：

$$(\hat{\hat{b}}-t_{0.025}(n-2)\frac{1}{\sqrt{l_{xx}}}\sqrt{\frac{S_e}{n-2}},b+t_{0.025}(n-2)\frac{1}{\sqrt{l_{xx}}}\sqrt{\frac{S_e}{n-2}}),$$

代入数值计算得 $(-2.1711,-1.9625)$.

(3) $\frac{S_e}{\sigma^2}\sim\chi^2(n-2)$, 当原假设为真时, 有 $\frac{S_R}{\sigma^2}\sim\chi^2(1)$, 因此有

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

拒绝域为 $F\geq F_{0.05}(1,n-2)$, 因为 $F_{0.05}(1,15)=4.54$, $F\geq F_{0.05}(1,n-2)$, 所以线性效果显著.

(4) $t=\dfrac{\hat{y}-y}{S\sqrt{1+\frac{1}{n}+\frac{(x-\overline{x})^2}{l_{xx}}}}\sim t(n-2)$, $S=\sqrt{\frac{S_e}{n-2}}$,

对于给定的自变量 x, 可得因变量 y 的置信度为 95% 的置信区间为:

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

其中

$$\hat{\hat{y}}=a+\hat{b}x,\ \delta(x)=t_{0.025}(n-2)\sqrt{\frac{S_e}{n-2}}\sqrt{1+\frac{1}{n}+\frac{(x-\overline{x})^2}{l_{xx}}},$$

代入数值计算得

$$\delta(x)=0.1073\sqrt{0.7506+(x-0.7029)^2}.$$

(5) 该问题本质上就是要确定下列方程组的解 $\begin{cases}\hat{a}+\hat{b}x_1-\delta(x_1)=y_1\\ \hat{a}+\hat{b}x_2+\delta(x_2)=y_2\end{cases}$,

因为 $t_{0.025}(n-2)\approx u_{0.025}$, $\sqrt{1+\frac{1}{n}+\frac{(x-\overline{x})^2}{l_{xx}}}\approx 1$, 所以方程组变为

$$\begin{cases}\hat{a}+\hat{b}x_1-u_{0.025}S=y_1\\ \hat{a}+\hat{b}x_2+u_{0.025}S=y_2\end{cases},\text{解得}\begin{cases}x_1=\dfrac{1}{\hat{b}}(y_1+u_{0.025}S-\hat{a})\\ x_2=\dfrac{1}{\hat{b}}(y_2-u_{0.025}S-\hat{a})\end{cases},$$

代入计算得 x 的值应该限制在 $(0.7,0.9)$ 内.

5. 在回归分析中, 常对数据进行变换:

$$\tilde{y}_i=\frac{y_i-c_1}{d_1},\ \tilde{x}_i=\frac{x_i-c_2}{d_2},\ i=1,2,\cdots,n$$

其中 c_1, c_2, $d_1>0$, $d_2>0$ 是适当选取的常数.

(1) 试建立由原始数据和变换后数据的最小二乘估计, 总平方和, 回归平方和以及残差平方和之间的关系;

(2) 证明:由原始数据和变换后数据得到的 F 检验统计量的值保持不变.

解 (1) $\tilde{y}_i=\dfrac{y_i-c_1}{d_1}\Rightarrow y_i=c_1+d_1\tilde{y}_i$, $\tilde{x}_i=\dfrac{x_i-c_2}{d_2}\Rightarrow x_i=c_2+d_2\tilde{x}_i$,

$$\tilde{y}_i-\overline{\tilde{y}}=\tilde{y}_i-\frac{1}{n}\sum_{i=1}^{n}\tilde{y}_i=\tilde{y}_i-\frac{1}{n}\sum_{i=1}^{n}\frac{y_i-c_1}{d_1}=\tilde{y}_i-\frac{1}{d_1}(\frac{1}{n}\sum_{i=1}^{n}y_i-c_1)=\tilde{y}_i+\frac{c_1}{d_1}-\frac{\overline{y}}{d_1}=\frac{y_i-\overline{y}}{d_1},$$

同理 $\tilde{x}_i-\overline{\tilde{x}}=\dfrac{x_i-\overline{x}}{d_2}$, 则 $l_{xy}=\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})=\sum_{i=1}^{n}d_2(\tilde{x}_i-\overline{\tilde{x}})d_1(\tilde{y}_i-\overline{\tilde{y}})=d_1d_2l_{\tilde{x}\tilde{y}}$,

同理得 $l_{xx}=d_2{}^2l_{\tilde{x}\tilde{x}}$, $l_{yy}=d_1{}^2l_{\tilde{y}\tilde{y}}$, 变换后的方程为 $\tilde{y}_i=\tilde{a}+\tilde{b}\tilde{x}_i+\varepsilon_i$,

进行最小二乘估计后 $\hat{\tilde{b}}=\dfrac{l_{\tilde{x}\tilde{y}}}{l_{\tilde{x}\tilde{x}}}=\dfrac{\frac{1}{d_1d_2}l_{xy}}{\frac{1}{d_2{}^2}l_{xx}}=\dfrac{d_2}{d_1}\dfrac{l_{xy}}{l_{xx}}=\dfrac{d_2}{d_1}\hat{b}$,

$$\hat{\tilde{a}}=\overline{\tilde{y}}-\hat{\tilde{b}}\overline{\tilde{x}}=\frac{1}{n}\sum_{i=1}^{n}\tilde{y}_i-\hat{\tilde{b}}\frac{1}{n}\sum_{i=1}^{n}\tilde{x}_i=\frac{1}{n}\sum_{i=1}^{n}\frac{y_i-c_1}{d_1}-\frac{d_2\hat{b}}{d_1}\frac{1}{n}\sum_{i=1}^{n}\frac{x_i-c_2}{d_2}$$

$$=\frac{1}{d_1}(\overline{y}-\hat{b}\overline{x}-c_1+\hat{b}c_2)=\frac{1}{d_1}(\hat{a}-c_1+\hat{b}c_2)=\frac{1}{d_1}\hat{a}-\frac{1}{d_1}(c_1-\hat{b}c_2),$$

$$S_R=\hat{b}^2l_{xx}=\hat{b}^2d_2{}^2l_{\tilde{x}\tilde{x}}=\frac{d_1{}^2}{d_2{}^2}\hat{\tilde{b}}^2d_2{}^2l_{\tilde{x}\tilde{x}}=d_1{}^2\hat{\tilde{b}}^2l_{\tilde{x}\tilde{x}}=d_1{}^2\tilde{S}_R,$$

$$S_e=l_{yy}-\hat{b}^2l_{xx}=d_1{}^2l_{\tilde{y}\tilde{y}}-\hat{b}^2d_2{}^2l_{\tilde{x}\tilde{x}}=d_1{}^2l_{\tilde{y}\tilde{y}}-\frac{d_1{}^2}{d_2{}^2}\hat{\tilde{b}}^2d_2{}^2l_{\tilde{x}\tilde{x}}=d_1{}^2(l_{\tilde{y}\tilde{y}}-\hat{\tilde{b}}^2l_{\tilde{x}\tilde{x}})=d_1{}^2\tilde{S}_e,$$

$$S_T=S_R+S_e=d_1{}^2\tilde{S}_R+d_1{}^2\tilde{S}_e=d_1{}^2(\tilde{S}_R+\tilde{S}_e)=d_1{}^2\tilde{S}_T.$$

(3) 证明: $\because F=\dfrac{(n-2)S_R}{S_e}\sim F(1,n-2)$,

$$\therefore\tilde{F}=\frac{(n-2)\tilde{S}_R}{\tilde{S}_e}=\frac{(n-2)\frac{1}{d_1{}^2}S_R}{\frac{1}{d_1{}^2}S_e}=\frac{(n-2)S_R}{S_e}=F,$$

F 检验统计量的值保持不变.

6. 测得一组弹簧形变 $x(cm)$ 和相应的外力 $y(N)$ 数据如下:

x	1	1.2	1.4	1.6	1.8	2.0	2.2	2.4	2.8	3.0
y	3.08	3.76	4.31	5.02	5.51	6.25	6.74	7.40	8.54	9.24

由胡克定理知 $y=kx$，若假定 $y=kx+\varepsilon$，$\varepsilon\sim N(0,\sigma^2)$，试估计 k，并在 $x=2.6cm$ 处给出相应的外力 y 的 95% 预测区间.

解 由第 2 题结论得 $\hat{k}=\dfrac{\sum\limits_{i=1}^{10}x_iy_i}{\sum\limits_{i=1}^{10}x_i^2}=0.3245$，$y_0$ 的置信度为 95% 的置信区间为：

$$(\hat{y_0}-\delta(x_0),y_0+\delta(x_0)),$$

其中 $x_0=2.6$，$\hat{y}_0=0.8437$，$\delta(x_0)=t_{0.025}(n-2)\sqrt{\dfrac{S_e}{n-2}}\sqrt{1+\dfrac{1}{n}+\dfrac{(x_0-\overline{x})^2}{l_{xx}}}=0.0431,$

代入计算得到 y_0 的预测区间为 $(0.8006,0.8868)$.

7. 我们知道营业税税收总额 y 与社会商品零售总额 x 有关，为能从社会商品零售总额去预测税收总额，需要了解两者之间的关系. 现收集了如下九组数据(单位：亿元)

序号	社会商品零售总额 x	营业税税收总额
1	142.08	3.93
2	177.30	5.96
3	204.68	7.85
4	242.68	9.82
5	316.24	12.50
6	341.99	15.55
7	332.69	15.79
8	389.29	16.39
9	453.40	18.45

(1) 画出散点图；

(2) 建立一元线性回归方程，并作显著性检验 $(\alpha=0.05)$，列出方差分析表；

(3) 若已知某年社会商品零售额为 300 亿元，试给出营业税税收总额的概率为 0.95 的预测区间；

(4) 若已知回归直线过原点，试求回归方程，并在显著性水平 0.05 下作显著性检验.

解 (1) 图略.

(2)

回归统计	
Multiple R	0.981069208
R Square	0.96249679
Adjusted R Square	0.957139189
标准误差	1.06405519
观测值	9

	df	SS	MS	F	Significance F
回归分析	1	203.4029	203.4029	179.6507	3.01722E-06
残差	7	7.925494	1.132213		
总计	8	211.3284			

	Coefficients	标准误差	t Stat	P-value	Lower 95%
Intercept	-2.25822	1.107518	-2.03899	0.080833	-4.877080404
X Variable 1	0.048672	0.003631	13.40338	3.02E-06	0.040085199

	Upper 95%	下限 95.0%	上限 95.0%
Intercept	0.360646595	-4.877080404	0.360646595
X Variable 1	0.057258584	0.040085199	0.057258584

$\hat{b}=\dfrac{l_{xy}}{l_{xx}}=0.0487$，$\hat{a}=\overline{y}-\hat{b}\overline{x}=-2.26$，$\Rightarrow \hat{y}=-2.26+0.0487x$，

$F=\dfrac{S_R}{S_e/(n-2)}\sim F(1,n-2)$，拒绝域$F\geq F_{0.05}(1,n-2)$，$F=179.65$，$F_{0.05}(1,7)=5.59$，

即表明回归效果显著.

（3）给出预测区间：$(\hat{y}_0-\delta(x_0),y_0+\delta(x_0))$，其中$\hat{y}_0=12.35$，

$$\delta(x_0)=t_{0.025}(n-2)\sqrt{\frac{S_e}{n-2}}\sqrt{1+\frac{1}{n}+\frac{(x_0-\overline{x})^2}{l_{xx}}}=2.6555,$$

则预测区间$(9.688,14.999)$.

(4)

Multiple R	0.99568
R Square	0.99138
Adjusted R Square	0.86638
标准误差	1.256615
观测值	9

	df	SS	MS	F	Significance F
回归分析	1	1452.8	1452.8	920.029	1.09201E-08
残差	8	12.63265	1.579081		
总计	9	1465.433			

	Coefficients	标准误差	t Stat	P-value	Lower 95%
Intercept	0	#N/A	#N/A	#N/A	#N/A
X Variable 1	0.041658	0.001373	30.33198	1.52E-09	0.038490601

Upper 95%	下限 95.0%	上限 95.0%
#N/A	#N/A	#N/A
0.04482469	0.0384906	0.04482469

$\hat{b}=\dfrac{\sum_{i=1}^{n}x_iy_i}{\sum_{i=1}^{n}x_i^2}=0.0417$, 回归方程为 $\hat{y}=0.0417x$, 由上表格得知回归性显著.

8. 在林业工程中, 需要知道树干的体积 y 与树干直径 x_1 和树干高度 x_2 之间的关系, 下表给出了一组树干的体积, 直径和高度的观测值:

序号	直径	树高	体积
1	8.4	71	10.4
2	8.7	66	10.4
3	8.9	64	10.3
4	10.6	73	16.5
5	10.8	82	18.9
6	10.9	84	19.8
7	11.1	67	15.7
8	11.1	76	18.3
9	11.2	81	22.7
10	11.3	76	20
11	11.4	80	24.3
12	11.5	77	21.1
13	11.5	77	21.5
14	11.8	70	21.4
15	12.1	76	19.2
16	13	75	22.3
17	13	86	33.9
18	13.4	87	27.5
19	13.8	72	25.8

20	13.9	65	25
21	14.1	79	34.6
22	14.3	81	31.8
23	14.6	75	36.7
24	16.1	73	38.4
25	16.4	78	42.7
26	17.4	82	55.5
27	17.6	83	55.8
28	18	81	58.4
29	18.1	81	51.6
30	17.1	81	51.1
31	20.7	88	77.1

试求 y 对 x_1 和 x_2 的回归方程, 并作显著性检验.

解 设 $y=a+b_1x_1+b_2x_2+\varepsilon$, 则要使 $Q(a,b_1,b_2)=\sum_{i=1}^{n}(y_i-a-b_1x_{1i}-b_2x_{2i})^2$ 最小, 则

$$\frac{\partial Q(a,b_1,b_2)}{\partial a}=-2\sum_{i=1}^{n}(y_i-a-b_1x_{1i}-b_2x_{2i})=0,$$

$$\frac{\partial Q(a,b_1,b_2)}{\partial b_1}=-2\sum_{i=1}^{n}(y_i-a-b_1x_{1i}-b_2x_{2i})x_{1i}=0,$$

$$\frac{\partial Q(a,b_1,b_2)}{\partial b_2}=-2\sum_{i=1}^{n}(y_i-a-b_1x_{1i}-b_2x_{2i})x_{2i}=0,$$

计算可得 $\hat{a}=-52.83$, $\hat{b}_1=4.48$, $\hat{b}_2=0.298$, 则 $\hat{y}=-52.83+4.48x_1+0.298x_2$,

取检验统计量 $F=\dfrac{S_R/m}{S_e/(n-m-1)}$,

当 H_0 为真时 $F=\dfrac{S_R/m}{S_e/(n-m-1)}=\dfrac{S_R/2}{S_e/28}=\dfrac{14S_R}{S_e}\sim F_{0.05}(2,28)$,

拒绝域为 $F\geq F_{0.05}(2,28)$, 因为 $F_{0.05}(2,28)=3.34$, $F\geq 3.34$. 即表明回归性显著.

9. 对于如下一组数据

x	2	3	4	5	6	7	8	9
y	6.42	8.20	9.58	9.50	9.70	10.00	9.93	9.99
x	10	11	12	13	14	15	16	
y	10.49	10.59	10.60	10.80	10.60	10.90	10.76	

试分别按(1) $y=a+\dfrac{b}{x}$, (2) $y=ae^{\frac{b}{x}}$ 来建立 y 对 x 的回归方程, 并用判定系数 R^2 指出哪一种相关较好.

解 (1) $y=a+\dfrac{b}{x}$, 令 $y=u$, $\dfrac{1}{x}=v$, 则 $u=a+bv$, 则 $\hat{a}=0.0823$, $\hat{b}=0.1312$,

$$\Rightarrow \hat{y}=0.0823+\frac{0.1312}{x},\ \Rightarrow R_1^{\,2}=1-\frac{\sum(y_i-\hat{y}_i)^2}{\sum(y_i-\overline{y})^2}.$$

(2) $y=ae^{\frac{b}{x}}$, 令 $u=\ln y$, $v=x$, $c=\ln a$, 则 $\hat{b}=-1.1107$, $\hat{c}=2.4578$,

$$\Rightarrow \hat{a}=11.6791,\ \Rightarrow \hat{y}=11.6791e^{-\frac{1.1107}{x}},\ R_2^{\,2}=1-\frac{\sum(y_i-\hat{y}_i)^2}{\sum(y_i-\overline{y})^2}.$$

因为 $R_1^{\,2}<R_2^{\,2}$, 所以方程 (2) 比方程 (1) 好.

10. 某研究机构对 200 北京鸭进行试验, 得到鸭的周龄 x 与平均日增重 y 的数据如下

x	1	2	3	4	5	6	7	8	9
y	21.9	47.1	61.9	70.8	72.8	66.4	50.3	25.3	3.2

试求回归方程 $y=a+b_1x+b_2x^2$, 并检验回归效果的显著性.

解 $y=a+b_1x+b_2x^2$, 令 $u=y$, $v_1=x$, $v_2=x^2$, 原方程可化为 $u=a+b_1v_1+b_2v_2$, 代入数据可得 $\hat{a}=-8.3515$, $\hat{b_1}=34.8267$, $b_2=-3.7623$, 则回归方程为

$$\hat{y}=-8.3515+34.8267x-3.7623x^2.$$

计算得到 $R^2=1-\dfrac{\sum(y_i-\hat{y}_i)^2}{\sum(y_i-\overline{y})^2}=0.993743525$, 表明回归效果显著.