多元线性回归

《美国数学建模竞赛》

完整课程请长按下方二维码

1.3 多元线性回归

- 人的体重与身高、胸围
- 血压值与年龄、性别、劳动强度、饮食习惯、吸烟状况、家族史
- 糖尿病人的血糖与胰岛素、糖化血红蛋白、血清总胆固醇、甘油三脂
- 射频治疗仪定向治疗脑肿瘤过程中，脑皮质的毁损半径与辐射的温度、与照射的时间

1.3 多元线性回归

多元回归模型：含两个以上解释变量的回归模型

多元线性回归模型：一个应变量与多个解释变量之间设定的是线性关系

多元线性回归模型一般形式为：

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_k X_k + \varepsilon$$

截距　　偏回归系数　　残差

1.3 多元线性回归

多元线性回归模型的假设:

$$Y = b_0 + b_1X_1 + b_2X_2 + \ldots + b_kX_k + u$$

解释变量 X_i 是确定性变量，不是随机变量；

解释变量之间互不相关，即无多重共线性。

随机误差项不存在序列相关关系

随机误差项与解释变量之间不相关

随机误差项服从0均值、同方差的正态分布

1.3 多元线性回归

多元模型的解析表达式：

$$Y = b_0 + b_1X_1 + b_2X_2 + \ldots + b_kX_k + \varepsilon$$

$$\begin{cases} Y_1 = b_0 + b_1X_{11} + b_2X_{21} + \\ Y_2 = b_0 + b_1X_{12} + b_2X_{22} + \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ Y_n = b_0 + b_1X_{1n} + b_2X_{2n} + \end{cases}$$

1.3 多元线性回归

多元模型的矩阵表达式：

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & X_{11} & X_{21} & \cdots & X_{k1} \\ 1 & X_{12} & X_{22} & \cdots & X_{k2} \\ \cdots & \cdots & \cdots & & \cdots \\ 1 & X_{1n} & X_{2n} & \cdots & X_{kn} \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ \cdots \\ b_k \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$$\boldsymbol{Y} = \boldsymbol{XB} + \boldsymbol{\varepsilon}$$

1.3 多元线性回归

参数值估计：最小二乘估计

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

$$=\sum_{i=1}^{n}\left(Y_i-\left(\hat{b}_0+\hat{b}_1X_{1i}+\cdots+\hat{b}_kX_{ki}\right)\right)^2$$

$$\leftarrow \begin{cases}\dfrac{\partial Q}{\partial \hat{b}_0}=0\\ \dfrac{\partial Q}{\partial \hat{b}_1}=0\\ \dfrac{\partial Q}{\partial \hat{b}_2}=0\\ \cdots\cdots\cdots\cdots\\ \dfrac{\partial Q}{\partial \hat{b}_k}=0\end{cases}$$

参数估计公式：

$$\hat{B}=(X^TX)^{-1}X^TY$$

1.3 多元线性回归

多元线性回归模型的检验

主要介绍：

- 拟合优度检验（判定系数）
- 回归方程的显著性检验（F—检验）
- 回归参数的显著性检验（t—检验）

1.3 多元线性回归

拟合优度检验

目的：构造一个不含单位，可以相互比较，而且能直观判断拟合优劣的指标。

判定系数的定义： $R^2 = \frac{SSR}{SST}$

意义：判定系数越大，自变量对因变量的解释程度越高，自变量引起的变动占总变动的百分比高。观察点在回归直线附近越密集。

取值范围：0-1

1.3 多元线性回归

回归方程的显著性检验

检验的目的

检验Y与解释变量x_1，x_2，……x_k之间的线性关系是否显著。

检验的步骤

第一步，提出假设：

$$\begin{cases} \text{原假设：} H_0\text{：} b_1=b_2=\ldots\ldots b_k=0 \\ \text{备择假设：} H_1\text{：} b_i\text{不全为}0 \quad (i=1, 2, \ldots, k) \end{cases}$$

1.3 多元线性回归

第二步，计算统计量：

$$F=\frac{SSR/k}{SSE/(n-k-1)}\sim F(k,n-k-1)$$

第三步，查表，得： $\boldsymbol{F_\alpha=F_\alpha(k,n-k-1)}$

第四步，做检验：

检验法则

$$\begin{cases} \boldsymbol{F>F_\alpha} \longrightarrow \text{拒绝}H_0\text{，回归方程显著} \\ \boldsymbol{F<F_\alpha} \longrightarrow \text{接受}H_0\text{，回归方程不显著} \end{cases}$$

1.3 多元线性回归

回归系数的显著性检验

回归方程显著，并不意味着每个解释变量对因变量Y的影响都重要,因此需要进行检验。

1.3 多元线性回归

回归系数显著性的检验的步骤

第一步，提出假设：

原假设：H_0： $b_i=0$ （i=1，2，……k）

备择假设：H_1： $b_i\neq 0$ （i=1，2，……k）

第二步，构造并计算统计量：

$$T_i = \frac{\hat{b}_i}{s(\hat{b}_i)} \quad (i = 1,2,\cdots,k)$$

1.3 多元线性回归

第三步，查表得：$\boxed{t_{\alpha/2} = t_{\alpha/2}(n-k-1)}$

第四步，做检验：

检验法则

$|T_i| < t_{\frac{\alpha}{2}}$ → 接受H_0

$|T_i| > t_{\frac{\alpha}{2}}$ → 拒绝H_0

1.3 多元线性回归

例 某品种水稻糙米含镉量y(mg/kg)与地上部生物量x1(10g/盆)及土壤含镉量x2(100mg/kg)的8组观测值如表2.1。试建立多元线性回归模型。

x1	1.37	11.34	9.67	0.76	17.67	15.91	15.74	5.41
x2	9.08	1.89	3.06	10.2	0.05	0.73	1.03	6.25
y	4.93	1.86	2.33	5.78	0.06	0.43	0.87	3.86

1.3 多元线性回归

```
/*代码以及结果的解释见教材*/
data ex;
input  x1-x2 y@@;
cards;
…
;
proc reg;
model y=x1 x2;
run;
```

1.3 多元线性回归

回归方程显著性检验：

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	31.46291	31.46291	[illegible]	[illegible]
Error	6	0.38209	0.063[illegible]		
Corrected Total	7	31.84500			

拟合度很高

Root MSE	0.25235	R-Square	$\boxed{0.9880}$
Dependent Mean	2.51500	Adj R-Sq	0.9860
Coeff Var	10.03390		

由方差分析表可知，其F value=494.06,pr>F的值<0.0001，远小于0.05，故拒绝原假设，接受备择假设，认为y1与x1,x2之间具有显著性的线性关系；

1.3 多元线性回归

参数显著性检验：

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	3.61051	0.95915	3.76	0.0131
x1	1	-0.19828	0.05822	-3.41	0.0191
x2	1	0.20675	0.09769	2.12	0.0879

由参数估计 表可知，对自变量x_2检验t值分别为t=2.12,Pr>|t|的值=0.0879,大于0.05，因此，拒绝原假设认为x_2的系数应为0，说明x_2的系数没有通过检验。为此，需要在程序中model y1=x1 x2中去掉x_2

1.3 多元线性回归

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	5.62117	0.16580	33.90	<.0001
x1	1	-0.31911	0.01436	-22.23	<.0001

对常数检验t值分别为t=33.9、,Pr>|t|的值<0.0001,远小于0.05，说明截距项通过检验，估计值为5.62117，同理可知x_1的系数通过检验，估计值为-0.31911

回归方程：$y=-0.31911x_1+5.62117$

1.3 多元线性回归

许多实际问题中可能还会出现某几个变量的系数并没有通过检验，此时，可以在原程序中的model y1=x1-x2中去掉没用通过的变量,直到所有的系数均通过检验。或者使用逐步回归方法，让软件自动保留通过检验的变量。