多元线性回归分析及其实际应用

田　兵
（包头师范学院《阴山学刊》编辑部，内蒙古 包头 014030）

摘　要：本文主要介绍了多元线性回归分析的数学模型。同时结合实例演示了应用 R 软件实现多元线性回归的过程。

关键词：多元线性回归；数学模型；估计；回归系数；显著性检验，逐步回归
中图分类号：O212　**文献标识码**：A　**文章编号**：1004－1869(2011)01－0016－04

回归分析是研究统计规律的方法之一。在回归分析中我们把所关心的一些指标称为因变量，通常用 Y 来表示；影响因变量的变量称为自变量，用 $X_1, X_2, \cdots, X_p$ 来表示。回归分析研究的主要问题是：确定 Y 与 $X_1, X_2, \cdots, X_p$ 间的定量关系表达式，这种表达式称为回归方程；对求得的回归方程的可信度进行检验；判断自变量对 Y 有无影响；利用所求得的回归方程进行预测和控制。

在解决实际问题时，我们通常从最简单而又最普遍的回归模型——线性回归模型入手。

1　线性回归的数学模型

设变量 Y 与变量 $X_1, X_2, \cdots, X_p$ 间有线性关系

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon, \quad (1)$$

其中 $\varepsilon \sim N(0, \sigma^2)$，$\beta_0, \beta_1, \cdots, \beta_p$ 和 σ^2 是未知参数，$p > 2$ 称模型(1) 为多元线性回归模型。

设 $x_{i1}, x_{i2}, \cdots, x_{ip}, i = 1, 2, \cdots, n$ 是 $(X_1, X_2, \cdots, X_p, Y)$ 是的 n 次独立观测值，则多元线性模型(1) 可表示为

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p X_{ip} + \varepsilon_i, i = 1, 2, \cdots, n$$

其中 $\varepsilon_i \in N(0, \sigma^2)$，且独立同分布。

若令

$$Y = \begin{Bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{Bmatrix}, \beta = \begin{Bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{Bmatrix}, X = \begin{Bmatrix} 1 & x_1 & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{1p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{Bmatrix},$$

$$\varepsilon = \begin{Bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{Bmatrix},$$

则多元线性模型模型可用矩阵的形式表示：

$$Y = X\beta + \varepsilon,$$

其中 Y 是由因变量构成的 n 维向量，X 是 $n \times (p+1)$ 阶矩阵，β 是 $p+1$ 维向量，ε 是 n 维误差向量，并满足

$$E(\varepsilon) = 0, Var(\varepsilon) = \sigma^2 I_n$$

2　回归系数的估计

求参数 β 的估计值 $\hat{\beta}$，就是求最小二乘函数

$$Q(\beta) = (Y - X\beta)^T (Y - X\beta)$$

达到最小的 β 值

可以证明的最小二乘估计

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

从而可得回归方程为

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_p X_p$$

称 $\hat{\varepsilon} = Y - X\hat{\beta}$ 为残差向量。通常取 $\sigma^2 = \hat{\varepsilon}^T \hat{\varepsilon}/(n-p-1)$ 为 σ^2 的估计，也称是 σ^2 的最小二乘估计。

3　显著性检验

检验有两种，一种是回归系数的显著性检验，简单地说就是检验某个变量 X_i 的系数是否为 0；另一个检验是回归方程的显著性检验，简单地说就是检

收稿日期：2010－07－20
作者简介：田　兵(1982－)，男，山西五台人，在读硕士，研究方向：概率中的敏感问题。

验该组数据是否适用于线性方程做回归。

3.1 回归系数的显著性检验

$H_{j0}:\beta_j=0, H_{j1}:\beta_j\neq 0, j=0,1,2,\cdots,p$

当 H_{j0} 成立时,统计量

$$T_j=\frac{\hat{\beta}}{\hat{\sigma}\sqrt{c_{jj}}}\sim t(n-p-1), j=0,1,2,\cdots,p$$

其中 c_{jj} 是 $C=(X^TX)^{-1}$ 的对角线上第 j 个元素。对于给定的显著性水平 α,检验的拒绝域为

$$|T_j|>t_{\alpha/2}(n-p-1), j=0,1,2,\cdots,p$$

3.2 回归方程的显著性检验

$H_0:\beta_0=\beta_1=\cdots=\beta_p=0, H_1:\beta_0,\beta_1,\cdots,\beta_p$ 不全为0。当 H_0 成立时,统计量

$$F=\frac{SS_R/p}{SS_E/(n-p-1)}\sim F(p,n-p-1),$$

其中

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

$$\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i, \hat{y}_i=\hat{\beta}_0+\hat{\beta}_1x_{i1}+\cdots+\hat{\beta}_px_{ip}$$

我们称 SS_R 为回归平方和,称 SS_E 为残差平方和。

对于给定的显著性水平 α,检验的拒绝域为

$$F>F_\alpha(p,n-p-1)$$

相关系数的平方定义为

$$R^2=\frac{SS_R}{SS_T}$$

用它来衡量 Y 与 $X_1,X_2,\cdots,X_p$ 之间相关的密切程度,其中 SS_T 为总体离差平方和,即

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

并且满足

$$SS_T=SS_E+SS_R$$

4 相关实例

研究同一地区土壤所含可给态磷的情况(Y),得到18组数据如表所示,表中 X_1 为土壤内所含无机磷浓度,X_2 为土壤内溶于 K_2CO_3 溶液并受溴化物水解的有机磷,X_3 为土壤内溶于 K_2CO_3 但不溶于溴化物水解的有机磷

表1:某地区土壤所含可给态磷的情况

序号	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
X1	0.4	0.4	3.1	0.6	4.7	1.7	9.4	10.1	11.6	12.6	10.9	23.1	23.1	21.6	23.1	1.9	26.8	29.9
X2	52	23	19	34	24	65	44	31	29	58	37	46	50	44	56	36	58	51
X3	158	163	37	157	59	123	46	117	173	112	111	114	134	73	168	143	202	124
Y	64	60	71	61	54	77	81	93	93	51	76	96	77	93	95	54	168	99

求出Y关于X的多元线性回归方程;对方程作显著性检验;对变量做逐步回归分析.

我们利用R软件解决上述问题,相应的R软件计算过程如下

```
earth <- data.frame(
X1 = c(0.4, 0.4, 3.1, 0.6, 4.7, 1.7, 9.4, 10.1, 11.6, 12.6, 10.9, 23.1, 23.1, 21.6, 23.1, 1.9, 26.8, 29.9),
X2 = c(52, 23, 19, 34, 24, 65, 44, 31, 29, 58, 37, 46, 50, 44, 56, 36, 58, 51),
X3 = c(158, 163, 37, 157, 59, 123, 46, 117, 173, 112, 111, 114, 134, 73, 168, 143, 202, 124),
Y = c(64, 60, 71, 61, 54, 77, 81, 93, 93, 51, 76, 96, 77, 93, 95, 54, 168, 99)
)
lm.can <- lm(Y ~ X1 + X2 + X3, data = earth);
summary(lm.can)
```

```
Call:
lm(formula = Y ~ X1 + X2 + X3, data = earth)
Residuals:
    Min      1Q  Median      3Q     Max
-28.349 -11.383  -2.659  12.095  48.807
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 43.65007   18.05442   2.418  0.02984 *
X1           1.78534    0.53977   3.308  0.00518 **
X2          -0.08329    0.42037  -0.198  0.84579
X3           0.16102    0.11158   1.443  0.17098
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 19.97 on 14 degrees of freedom
```

Multiple R - squared: 0.5493, Adjusted R - squared: 0.4527

F - statistic: 5.688 on 3 and 14 DF, p - value: 0.009227

在上述操作中,函数 lm( )表示作线性模型,其模型公式 Y ~ X1 + X2 + X3 表示 $Y = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon$, 函数 summary( )是提取模型的计算结果。

在计算结果的第一部分(Call)列出了相应的回归模型公式,第二部分(Residuals:)列出的是残差的最小值点、1/4 分位点、中位数点、3/4 分位点和最大值点。

在计算结果的第三部分(Coefficients:)中,Estimate 表示回归方程参数的估计,即 $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$.

Std. Error 便是回归参数的标准差,即 sd($\hat{\beta}_0$), sd($\hat{\beta}_1$), sd($\hat{\beta}_2$). t value 为 t 值, $Pr(>|t|)$ 表示 P 值,即概率值 $P\{T > |T_{值}|\}$。还有显著性标记:"***"说明极为显著,"**"说明高度显著,"*"说明显著,"."说明不太显著,没有标记则表示不显著。

从计算结果可以看出 X_1 对 Y 的影响是高度显著的; X_2 和 X_3 对 Y 的影响是不显著的。回归方程只有常数项和 X_1 的系数通过了检验。相应的得到回归方程为

$$\hat{Y} = 43.65007 + 1.79534X_1 - 0.08329X_2 + 0.16102X_3.$$

显然如果选择全部变量作回归方程,效果不好。我们通过逐步回归来获得 Y 关于 X 的"最优"回归方程。首先,我们用函数 step( )作逐步回归

```
> lm.step < - step(lm.can)
Start: AIC = 111.27
Y ~ X1 + X2 + X3
```

	Df	Sum of Sq	RSS	AIC
- X2	1	15.7	5599.4	109.3
< none >			5583.7	111.3
- X3	1	830.6	6414.4	111.8
- X1	1	4363.4	9947.2	119.7

```
Step: AIC = 109.32
Y ~ X1 + X3
```

	Df	Sum of Sq	RSS	AIC
< none >			5599.4	109.3
- X3	1	833.2	6432.6	109.8
- X1	1	5169.5	10768.9	119.1

从程序运行结果来看,用全部变量作回归方程时,AIC 的值为 111.27。接下来显示的数据表明,如果去掉 X2 得到回归方程 AIC 的值为 109.3;如果去掉变量 X3,得到的回归方程 AIC 的值为 111.8。由于去掉变量 X2 可使 AIC 达到最小,故 R 软件自动去掉变量 X2,进行下一轮计算。

在下一轮计算中,无论去掉那一个变量,AIC 的值均会升高,因此 R 软件终止计算,得到"最优"的回归方程。

用 summary( )提取相关信息

```
> summary(lm.step)
Call:
lm(formula = Y ~ X1 + X3, data = earth)
Residuals:
    Min      1Q  Median      3Q     Max
-29.713 -11.324  -2.953  11.286  48.679
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 41.4794   13.8834   2.988  0.00920 **
X1           1.7374    0.4669   3.721  0.00205 **
X3           0.1548    0.1036   1.494  0.15592
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 19.32 on 15 degrees of freedom
Multiple R - squared: 0.5481, Adjusted R - squared: 0.4878
F - statistic: 9.095 on 2 and 15 DF, p - value: 0.002589
```

不难发现:回归系数检验的显著水平有了一定的提高,但 X3 系数检验的显著水平仍不理想。为了得到更好的结果,我们用函数 drop1( )来作更进一步的回归

```
> drop1(lm.step)
Single term deletions
Model:
Y ~ X1 + X3
```

	Df	Sum of Sq	RSS	AIC
< none >			5599.4	109.3
X1	1	5169.5	10768.9	119.1
X3	1	833.2	6432.6	109.8

从结果来看如果去掉变量 X3,AIC 的值会从 109.3 增加到 109.8,是增加的最少的。另外,除 AIC 准则外,残差的平方和也是逐步回归的重要指标之一。

残差的平方和越小,对应的方程拟合程度就越好。去掉 X3,残差的平方和上升 833.2,相比之下也是最少的。因此,从这两项指标来看,应该去掉 X3.

```
> lm.opt <- lm(Y ~ X1,data = earth);summary(lm.opt)
Call:
lm(formula = Y ~ X1, data = earth)
Residuals:
    Min      1Q  Median      3Q     Max
-31.486  -8.282  -1.674   5.623  59.337
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 59.2590  7.4200  7.986 5.67e-07***
X1           1.8434  0.4789  3.849  0.00142**
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 20.05 on 16 degrees of freedom
Multiple R-squared: 0.4808, Adjusted R-squared: 0.4484
F-statistic: 14.82 on 1 and 16 DF, p-value: 0.001417
```

明显可以看出:去掉变量 X2,X3 后,所有检验都是显著的。因此经过上面的计算,我们才得到真正的"最优"的回归方程:

$$\hat{Y} = 59.2590 + 1.8434X_1.$$

〔参考文献〕

[1]冯士雍,施锡铨.抽样调查——理论、方法与实践[M].上海:上海科学技术出版社,1996.

[2]薛毅,陈立萍.统计建模与 R 软件[M].北京:清华大学出版社,2007.

[3]王松桂,陈敏,陈立萍.线性统计模型[M].北京:高等教育出版社,1999.

[4]薛毅.数学建模基础[M].北京:工业大学出版社,2004.

[5]王学民.应用多元统计分析[M].上海:上海财经大学,2004.

Multiple linear Regression Analysis and Its Application

TIAN Bing

(Editor of Academic Journal , Baotou Teachers College ; Baotou 014030)

Abstract: In this article, we chiefly introduce the statistic idea and the mathematical model of multiple linear regression. We demonstrate the applied progress of multiple linear regression through using the software of R to solve an example

Key words: Multiple linear regression; Mathematical model ; Estimate; Regression coefficient; Significant test; Stepwise regression