DOI：10.3969/j.issn.1001-8972.2010.09.020

多元线性回归分析及其应用

林彬 湛江师范学院信科院 524048

Multiple linear regression analysis and its application

Lin Bin (Zhanjiang Normal College, Zhanjiang, 524048, China)

摘 要

本文研究了多元线性回归理论及应用，探讨了多元线性回归模型中未知参数的估计及其参数的检验问题，以实例进行了验证。

关键词

多元线性回归分析；回归模型；检验问题。

Abstract

We study the multiple linear regression theory and application, and estimation of unknown parameters and parameter testing problem in multivariate linear regression model. and then an example is solved to verify its effectiveness.

Key words

Multiple linear regression analysis; regression model; parameter testing problem

1 概述

回归分析是一种传统的应用性较强的科学方法，是现代应用统计学的一个重要的分支，在各个科学领域都得到了广泛的应用。它不仅能够把隐藏在大规模原始数据群体中的重要信息提炼出来，把握住数据群体的主要特征，从而得到变量间相关关系的数学表达式，利用概率统计知识对此关系进行分析，以判别其有效性，还可以利用关系式，由一个或多个变量值去预测和控制另一个因变量的取值，从而知道这种预测和控制达到的程度，并进行因素分析。

2 多元线性回归数学模型

设可预测的随机变量为y，它受到p个非随机因素$x_1,x_2,\cdots x_{p-1}$，x_p，和不可预测的随机因素ε的影响。多元线性回归数学模型为

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_{p-1}x_{p-1}+\beta_px_p+\varepsilon \quad (1)$$

$$\varepsilon\sim N(0,\sigma^2)$$

其中$\beta_0,\beta_1,\cdots,\beta_p$为回归系数

对y和$x_1,x_2,\cdots x_{p-1}$，x_p,分别进行n次独立观测，取得n组数据(样本)

$$y_i,x_{i1},x_{i2},\cdots,x_{ip-1}\quad(i=1,2,3,...,n)$$

则有

$$\begin{cases}y_1=\beta_0+\beta_1x_{11}+\beta_2x_{12}+\cdots+\beta_{p-1}x_{1p-1}+\varepsilon_1\\ y_2=\beta_0+\beta_1x_{21}+\beta_2x_{22}+\cdots+\beta_{p-1}x_{2p-1}+\varepsilon_2\\ \vdots\\ y_n=\beta_0+\beta_1x_{n1}+\beta_2x_{n2}+\cdots+\beta_{p-1}x_{np-1}+\varepsilon_n\end{cases} \quad (2)$$

其中$\varepsilon_1,\varepsilon_2,\cdots\varepsilon_n$相互独立，且服从$N(0,\sigma^2)$分布。

令

$$\mathbf{Y}=\begin{bmatrix}y_1\\y_2\\\vdots\\y_n\end{bmatrix},\quad \boldsymbol{\beta}=\begin{bmatrix}\beta_1\\\beta_2\\\vdots\\\beta_{p-1}\end{bmatrix},$$

$$\boldsymbol{\varepsilon}=\begin{bmatrix}\varepsilon_1\\\varepsilon_2\\\vdots\\\varepsilon_n\end{bmatrix},\quad \mathbf{x}=\begin{bmatrix}1&x_{11}&x_{12}&\cdots x_{1p-1}\\1&x_{21}&x_{22}&\cdots x_{2p-1}\\\vdots&\vdots&\vdots&\vdots\\1&x_{n1}&x_{n2}&\cdots x_{np-1}\end{bmatrix}$$

则式(2)用矩阵形式表示为

$$\begin{aligned}&\mathbf{y}=\mathbf{x}\boldsymbol{\beta}+\boldsymbol{\varepsilon}\\&\boldsymbol{\varepsilon}\sim N(0,\sigma^2I_n)\end{aligned} \quad (3)$$

3 模型参数β的最小二乘法估计与误差方差σ^2的估计

β的最小二乘法估计即选择β使误差项的平方和为最小值，这时β的值$\hat{\beta}$作为

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

β的点估计.

$$S(\boldsymbol{\beta})=\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon}=(\mathbf{y}-\mathbf{x}\boldsymbol{\beta})^T(\mathbf{y}-\mathbf{x}\boldsymbol{\beta}) \quad (4)$$

为了求β，由(4)式将S(β)对β求导，并令其为零，得

$$\frac{dS(\boldsymbol{\beta})}{d\boldsymbol{\beta}}=\frac{d[(\mathbf{y}-\mathbf{x}\boldsymbol{\beta})^T(\mathbf{y}-\mathbf{x}\boldsymbol{\beta})]}{d\boldsymbol{\beta}}=$$

$$\frac{d[(\mathbf{y}^T\mathbf{y}-\boldsymbol{\beta}^T\mathbf{x}^T\mathbf{y}-\mathbf{y}^T\mathbf{x}\boldsymbol{\beta}+\boldsymbol{\beta}^T\mathbf{x}^T\mathbf{x}\boldsymbol{\beta})]}{d\boldsymbol{\beta}}=0 \quad (5)$$

由(5)式可解出$\hat{\boldsymbol{\beta}}$

$$\hat{\boldsymbol{\beta}}=(\mathbf{x}^T\mathbf{x})^{-1}(\mathbf{x}^T\mathbf{y}) \quad (6)$$

对残差向量$\boldsymbol{\varepsilon}$

$$\hat{\boldsymbol{\varepsilon}}=\mathbf{y}-\hat{\mathbf{y}}=\mathbf{y}-\mathbf{x}\hat{\boldsymbol{\beta}}=[\mathbf{I}-\mathbf{x}(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T]\mathbf{y} \quad (7)$$

则残差平方和

$$\hat{\boldsymbol{\varepsilon}}^T\hat{\boldsymbol{\varepsilon}}=\hat{\boldsymbol{\varepsilon}}^T[\mathbf{I}-\mathbf{x}(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T]\mathbf{y}=\mathbf{y}^T\mathbf{y}-\hat{\boldsymbol{\beta}}^T\mathbf{x}^T\mathbf{y} \quad (8)$$

又因为$E(\mathbf{y})=\mathbf{x}\boldsymbol{\beta}$，因此

$$E(\hat{\boldsymbol{\varepsilon}}^T\hat{\boldsymbol{\varepsilon}})=\sigma^2(n-p)$$

$$\sigma^2=\frac{1}{n-p}(\hat{\boldsymbol{\varepsilon}}^T\hat{\boldsymbol{\varepsilon}}) \quad (9)$$

4 模型检验

多元线性回归数学模型建立后，是否与实际数据有较好的拟合度，其模型线性关系的显著性如何等，还需通过数理统计进行检验。常用的统计检验有R检验和F检验。

4.1 R检验

$$R=\sqrt{1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y}_i)^2}} \quad (10)$$

R是复相关系数，用于测定回归模型的拟合优度，R越大，说明Y与$x_1,x_2,\cdots x_{p-1}$的线性关系越显著，$\bar{y}_i$为y_i的平均值，R取值范围为$0<|R|\leq 1$。

4.2 F检验

$$F=\frac{U/m}{Q/(n-m-1)}\sim F(m,n-m-1) \quad (11)$$

式中$Q=\sum(y_i-\hat{y}_i)^2$，$U=\sum(\hat{y}_i-\bar{y}_i)^2$，

m为自变量个数,n为数据个数。

F服从F（m,n-m-1）分布，取显著性水平为α，如果$F>F_a$(m,n-m-1),表明回归模型显著，可从用于预测。反之，回归模型不能用于预测。

5 应用实例

某医院为了解病人对医院工作的满意程度Y和病人的年龄X_1，病情的严重程度X_2和病人的忧虑程度X_3之间的关系，随机调查了该医院的10位病人,得数据如表1所示。

表1 病人满意度的调查数据

年龄（x_1）	病情程度（x_2）	忧虑程度（x_3）	满意程度（y）
50	51	2.3	48
36	46	2.3	57
40	48	2.2	66
41	44	1.8	70
28	43	1.8	89
49	54	2.9	36
42	50	2.2	46
45	48	2.4	54
52	62	2.9	26
29	50	2.1	77

使用MATLAB语言编程并计算得下面结果:

RegCoff=
175.5249
-1.1713
-0.5117
-19.6453
R=0.9603
F=23.7098
FX=9.0886 0.4105 2.5260
TX=9.0224 0.8376 79.7754

从结果可以得出，回归模型为

$$y=175.5249-1.1713x_1-0.5117x_2-19.6453x_3$$

取α=0.05对方程和回归系数进行检验，查F分布表可得$F_{0.05}$（3.6）=4.76，$F_{0.05}$（1.6）=5.99

本例中的方程检验值F=23.7098>4.76，说明模型的回归效果高度显著。

F_1=9.0886>5.99，说明x_1显著。

F_2=0.4105<5.99，说明x_2很不显著。

F_3=2.5260<5.99，说明x_3不显著。

R为0.9603接近1，表明线性相关性较强。

在实际中，由于Y的影响因素还有很多，使Y与X关系更为复杂，而且记录数据的准确性，可靠性，异常数据等问题，将影响Y的预测分析。

参考文献

[1]梅长林，范金城.数据分析方法.高等教育出版社.2006,2.

[2]何晓群.现代统计分析与应用.中国人民大学出版社.2007,8.

作者简介

林彬(1962-)，男，湛江人，博士，讲师，研究方向：应用数学。

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net