矩阵法在多元线性回归模型上的应用研究

张林泉，廖红文 （广东女子职业技术学院信息资源中心，广东 广州 511450）

[摘要] 利用矩阵的形式对多元线性回归模型的进行估计、检验并分析了实现的具体步骤。结合实例建立了回归模型，利用该模型给出了回归系数的置信区间、因变量平均值的置信区间、因变量个别值的预测区间及多元线性估计回归模型的拟合图形。研究表明，该法对进一步研究算法与结论之间的关系有重要作用，有利于理解矩阵代数和多元线性回归模型的内在关系。

[关键词] 回归系数；回归标准差；置信区间；预测区间；矩阵表示
[中图分类号] O212.7 [文献标志码] A [文章编号] 1673-1409 (2014) 25-0016-03

一元线性回归模型中，仅仅考虑了一个解释变量（自变量）对因变量的影响，而社会经济问题所研究的变量往往是受多个因素影响，将把模型扩展到有多个解释变量影响因变量的情形。包含多个解释变量的回归模型，称为多元回归模型[1]。针对应用统计软件进行多元回归模型分析时，出现混乱及一知半解的情况，笔者主要以矩阵的形式研究实际数据，建立回归模型，该形式对深刻理解算法与结论之间的关系有重要作用。

1 多元线性回归模型及矩阵表示

1.1 多元线性总体回归模型

假定因变量Y与k个解释变量X_1，X_2，…，X_k具有线性相关关系，取n期观察值，则总体线性回归模型为：

$$y_i=\beta_0+\beta_1X_{i1}+\beta_2X_{i2}+\cdots+\beta_kX_{ik}+\mu_i \qquad i=1,2\cdots,n \tag{1}$$

式中，β_0是截距；β_1，β_2，…，β_k为偏回归系数。也就是说，多元线性回归模型是以多个解释变量的固定值为条件的回归分析，并且所获取的是多个自变量X值固定时Y的平均值或者为Y的平均响应。

1.2 多元线性样本回归模型

多元线性总体回归模型是无法得到的，只能用样本观察值进行估计．对应于式(1)的总体回归结构，多元线性样本回归模型为：

$$\hat{Y}_i=\hat{\beta}_0+\hat{\beta}_1X_{i1}+\hat{\beta}_2X_{i2}+\cdots+\hat{\beta}_kX_{ik} \qquad i=1,2,3,\cdots n \tag{2}$$

式中，$\hat{Y}_i$是总体均值$E(Y|X_{i1},X_{i2},\cdots,X_{ik})$的估计；$\hat{\beta}_j(j=1,2,\cdots,k)$是总体偏回归系数$\hat{\beta}_j$的估计。

1.3 多元线性回归模型的矩阵表示

取n次观测值$(Y_i,X_{i1},X_{i2},\cdots,X_{ik})$，$(i=1,2,\cdots,n)$代入式(1)中，得到$n$个随机方程，将这$n$个方程写成矩阵形式：

$$\begin{pmatrix}Y_1\\Y_2\\\vdots\\Y_n\end{pmatrix}_{n\times1}=\begin{pmatrix}1&X_{11}&X_{12}&\cdots&X_{1k}\\1&X_{21}&X_{22}&\cdots&X_{2k}\\\vdots&\vdots&\vdots&&\vdots\\1&X_{n1}&X_{n2}&\cdots&X_{nk}\end{pmatrix}_{n\times(k+1)}\begin{pmatrix}\beta_0\\\beta_1\\\vdots\\\beta_k\end{pmatrix}_{(k+1)\times1}+\begin{pmatrix}\mu_1\\\mu_2\\\vdots\\\mu_n\end{pmatrix}_{n\times1}$$

简记为：

$$Y=X\beta+\mu \tag{3}$$

[收稿日期] 2014-05-06
[基金项目] 广东省教育科学“十二五”规划2012年度项目（2012JK078）。
[作者简介] 张林泉（1965-），男，硕士，副研究员，现主要从事应用统计、计量经济学与数学方面的教学与研究工作。

式中，Y 为 n 阶因变量观测值向量；X 为 $n\times(k+1)$ 阶解释变量观测值矩阵；μ 为 n 阶随机扰动项向量；β 为 $(k+1)$ 阶总体回归参数向量。

相应的多元线性样本回归模型可用矩阵表示如下：

$$\hat{Y}=X\hat{\beta} \tag{4}$$

式中，$\hat{Y}$ 为 n 阶因变量回归拟合值向量；$\hat{\beta}$ 为 $(k+1)$ 阶回归参数 β 估计值向量；对于多元线性回归模型，只要解释变量个数多于 3 个，计算公式将会非常复杂，因此必须借助于矩阵代数来计算。

2　偏回归系数的最小二乘估计

对于多元线性总体回归模型式中的参数，即偏回归系数，可以用 OLS 法进行参数估计，OLS 法的一些性质对多元线性回归模型同样适用。总体回归模型的样本回归模型如下[2]：

$$Q=\sum_{i=1}^{n}e_i^2=e^{\mathrm{T}}e=(Y-X\hat{\beta})^{\mathrm{T}}(Y-X\hat{\beta})$$

即求解方程组：

$$\begin{cases}\dfrac{\partial}{\partial\hat{\beta}}(Y-X\hat{\beta})^{\mathrm{T}}(Y-X\hat{\beta})=0\\ \dfrac{\partial}{\partial\hat{\beta}}(Y^{\mathrm{T}}Y-\hat{\beta}^{\mathrm{T}}X^{\mathrm{T}}Y-Y^{\mathrm{T}}X\hat{\beta}+\hat{\beta}^{\mathrm{T}}X^{\mathrm{T}}X\hat{\beta})=0\\ \dfrac{\partial}{\partial\hat{\beta}}(Y^{\mathrm{T}}Y-2Y^{\mathrm{T}}X\hat{\beta}+\hat{\beta}^{\mathrm{T}}X^{\mathrm{T}}X\hat{\beta})=0\\ -X^{\mathrm{T}}Y+X^{\mathrm{T}}X\hat{\beta}=0\end{cases}$$

得到：

$$X^{\mathrm{T}}Y=X^{\mathrm{T}}X\hat{\beta}$$

于是：

$$\hat{\beta}=(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y \tag{5}$$

3　实例分析

例 1　$X=\begin{pmatrix}1 & 27.4 & 2.45\\ 1 & 18.0 & 3.254\\ \vdots & \vdots & \vdots\\ 1 & 37.0 & 2.605\end{pmatrix}$　$Y=\begin{pmatrix}1.62\\ 1.20\\ \vdots\\ 2.12\end{pmatrix}$　$\beta=\begin{pmatrix}\beta_1\\ \beta_2\\ \beta_3\end{pmatrix}$　$u=\begin{pmatrix}u_1\\ u_2\\ \vdots\\ u_{15}\end{pmatrix}$

3.1　回归系数

估计回归模型[3] 为 $Y=X\hat{\beta}+\hat{u}=\hat{Y}+\hat{u}$，根据(5) 估计 $\hat{\beta}$，可得：

$$\hat{\beta}=(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y=\begin{pmatrix}1.2463 & 0.0021 & -0.4157\\ 0.0021 & 0.0008 & -0.0070\\ -0.4157 & -0.0070 & 0.1977\end{pmatrix}\begin{pmatrix}22.59\\ 647.107\\ 70.9662\end{pmatrix}=\begin{pmatrix}0.0345\\ 0.0496\\ 0.0920\end{pmatrix}$$

3.2　回归标准差计算

根据估计回归模型[3] 为 $Y=X\hat{\beta}+\hat{u}=\hat{Y}+\hat{u}$ 可得：

$$\hat{u}=Y-X\hat{\beta}=\begin{pmatrix}0.0010\\ -0.0267\\ \vdots\\ 0.0106\end{pmatrix}_{15\times1}\qquad s^2=\hat{u}^{\mathrm{T}}\hat{u}/(n-k-1)=0.00047403$$

回归模型的标准误差为：

$$\text{s. e.}=\sqrt{s^2}=\sqrt{0.00047403}=0.0218$$

3.3 标准差

标准差项报告了系数估计的标准差。标准差衡量了系数估计的统计可信性，标准差越大，估计中的统计干扰越大。参数估计向量 $\hat{\beta}$ 的方差-协方差矩阵：

$$s^2(\hat{\beta})=\mathrm{Cov}(\hat{\beta})=\hat{\sigma}^2(X^{\mathrm{T}}X)^{-1}=s^2(X^{\mathrm{T}}X)^{-1}=\begin{pmatrix}0.5908 & 0.0010 & -0.1970\\ 0.0010 & 0.0004 & -0.0033\\ -0.1970 & -0.0033 & 0.0937\end{pmatrix}$$

主对角线给出了各参数估计 $\hat{\beta}_j$ 的方差，其余部分给出了不同参数估计的协方差，称为参数估计向量的 $\hat{\beta}$ 方差-协方差矩阵。系数估计值的标准差是这个矩阵对角线元素的平方根。从上式矩阵的主对角线上提取 3 个方差，并计算标准差[3]：

$$\begin{pmatrix}Se^2(\hat{\beta}_0)\\ Se^2(\hat{\beta}_1)\\ Se^2(\hat{\beta}_2)\end{pmatrix}=\begin{pmatrix}0.5908\\ 0.0004\\ 0.0937\end{pmatrix}\qquad\begin{pmatrix}Se(\hat{\beta}_0)\\ Se(\hat{\beta}_0)\\ Se(\hat{\beta}_2)\end{pmatrix}=\begin{pmatrix}0.0243\\ 0.0006\\ 0.0097\end{pmatrix}$$

3.4 多元线性回归模型的检验

1) 拟合优度检验：

$$TSS=Y^{\mathrm{T}}Y-n\bar{Y}^2=39.4107-15\times1.506^2=5.3902$$

$$ESS=\hat{Y}^{\mathrm{T}}\hat{Y}-n\bar{Y}^2=39.4050-15\times1.506^2=5.3845$$

$$RSS=TSS-ESS=0.0057$$

多元判定系数：

$$R^2=\frac{ESS}{TSS}=\frac{5.3845}{5.3902}=0.9989$$

校正的判定系数：

$$\bar{R}^2=1-(1-R^2)\frac{(n-1)}{(n-k-1)}=1-(1-0.9989)\frac{14}{12}=0.9988$$

式中，TSS 为总离差平方和；ESS 为回归平方和；RSS 为残差平方和。

2) 回归方程的显著性检验(F- 检验)。假设 H_0：$\beta_j=0$，H_1：β_j 不全为 0，从而：

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

$$F=\frac{5.3845/2}{0.0057/12}=5679.4655$$

$$F_{0.05}(2,\ 12)=3.89\qquad F=5679.4655>3.89$$

结果表明回归函数存在线性关系。用统计量进一步检验 β_1 和 β_2 是不是为 0。

3) 回归系数的检验。假设 H_0：$\beta_j=0$，H_1：$\beta_j\neq0$，从而：

$$t=\frac{\hat{\beta}_j-\beta_j}{Se(\hat{\beta}_j)}\sim t(n-k-1)\qquad t_1=81.9242\qquad t_2=9.5021$$

因为分别大于临界值，结论是拒绝原假设 $\beta_j=0$，说明都是 Y 的重要解释变量，应保留在模型中。综上估计的回归模型为：

$$\hat{Y}_i=0.0345+0.0496X_{i1}+0.0920X_{i2}$$

4) 多元线性回归模型参数的区间估计。为了说明参数真实值的可能范围和可靠性，还需要在对参数点估计的基础上对多元线性回归模型参数作区间估计。

回归系数 β_j 在 $1-\alpha$ 显著水平下的置信区间为：

$$(\hat{\beta}_j-t_{\alpha/2}(n-k-1)\times Se(\hat{\beta}_j),\ \hat{\beta}_j+t_{\alpha/2}(n-k-1)\times Se(\hat{\beta}_j))$$

$$(\hat{\beta}_1-t_{0.025}(12)\times Se(\hat{\beta}_1)<\beta_1<\hat{\beta}_1+t_{0.025}(12)\times Se(\hat{\beta}_1))=(0.04828135<\beta_1<0.05091965)$$

$$(\hat{\beta}_2-t_{0.025}(12)\times Se(\hat{\beta}_2)<\beta_2<\hat{\beta}_2+t_{0.025}(12)\times Se(\hat{\beta}_2))=(0.07089742<\beta_2<0.1130842)$$

（下转第 33 页）

微间隙产生条件，并模拟微间隙随井深变化曲线，为后续施工提供技术参考。

3　结论与认识

1）开发了一套集固井设计、固井模拟、实时监控及事后分析评价等功能于一体的固井施工设计模拟与监控软件，有助于提高固井设计与分析水平和固井质量，缩短与国外同类软件的差距，并促进我国固井工程信息化建设。

2）综合考虑井眼质量、套管居中情况、泥浆性能、固井流体以及注水泥施工参数等因素的影响，开发了固井顶替效率模拟功能模块，为优化固井施工参数提供参考。

3）结合我国固井水泥车硬件采集装置的特点，开发了固井施工与监控功能模块，对注水泥施工流体密度、排量和压力参数进行实时监测、分析和辅助控制，有助于全面了解井下施工情况，及时科学决策以提高固井质量。

4）为了进一步提升软件水平，还应加强固井施工分析评价技术研究，建立固井施工分析评价专家系统，实现专家知识与经验的共享，朝向智能化方向发展[4]。

[参考文献]

[1] 黄志强．SQL Server 在固井数据管理中的应用 [J]．长江大学学报（自科版），2009，6 (1)：72-74.
[2] 黄志强．定向井下套管摩阻数值计算及其应用 [J]．石油地质与工程，2009，23 (4)：79-81.
[3] 黄志强．注水泥动态过程研究与计算机模拟 [J]．特种油气藏，2009，16 (3)：92-94.
[4] 徐璧华，何松，何可．高温超压地层固井平衡压力注水泥软件设计 [J]．西部探矿工程，2012 (7)：75-81.

[编辑]　洪云飞

（上接第 18 页）

3.5　多元线性回归模型的预测

利用估计模型进行点预测与区间预测[4]，其 $C=(1,\ X_{n+1,1},\ X_{n+1,2},\ \cdots,\ X_{n+1,k})$ 表示 $n+1$ 期解释变量，$n+1$ 期被解释变量 y_{n+1} 的点预测式为 $\hat{y}_{n+1}=C\hat{\beta}$，单个 y_{n+1} 的预测区间为 $(C\hat{\beta}-t_{\alpha/2}(n-k-1)\times s\sqrt{C(X^{\mathrm{T}}X)^{-1}C^{\mathrm{T}}+1},\ C\hat{\beta}+t_{\alpha/2}(n-k-1)\times s\sqrt{C(X^{\mathrm{T}}X)^{-1}C^{\mathrm{T}}+1})$。

$$C=(1,\ 22.17,\ 2.962)\qquad \hat{y}_{n+1}=C\hat{\beta}=(1,\ 22.17,\ 2.962)\begin{pmatrix}0.0345\\0.0496\\0.0920\end{pmatrix}=1.5058$$

$$t_{0.025}(12)\times s\sqrt{C(X^{\mathrm{T}}X)^{-1}C^{\mathrm{T}}+1}=0.048993343$$

置信度为 95% 的年销量的预测区间为 (1.456853587，1.554840273)。

$E(y_{n+1})$ 的置信区间[5] 为 $(C\hat{\beta}-t_{\alpha/2}(n-k-1)\times s\sqrt{C(X^{\mathrm{T}}X)^{-1}C\mathrm{T}},\ C\hat{\beta}+t_{\alpha/2}(n-k-1)\times s\sqrt{C(X^{\mathrm{T}}X)^{-1}C^{\mathrm{T}}})$。

$$t_{0.025}(12)\times s\sqrt{C(X^{\mathrm{T}}X)^{-1}C^{\mathrm{T}}}=0.012248337$$

置信度为 95%的年销量的置信区间为 (1.493598593，1.518095267)。

4　结语

讨论了用矩阵的形式建立回归分析模型，并进行相应的显著性检验和多元线性回归模型的预测。结果表明，使用矩阵法对深入理解多元线性回归模型的原理与结论之间的关系至关重要。

[参考文献]

[1] 王维国．计量经济学 [M]．大连：东北财经大学出版社，2002.
[2] 徐国祥．统计学 [M]．上海：上海财经大学出版社，2007.
[3] 张晓峒．应用数量经济学 [M]．北京：机械工业出版社，2009：107-132.
[4] 庞皓．计量经济学 [M]．北京：科学出版社，2007：87-88.
[5] 默里．计量经济学：现代方法 [M]．北京：北京大学出版社，2009：212-266.

[编辑]　张涛

阅读此文的还阅读了：

1. 多元线性回归模型应用实证分析
2. 多元线性回归模型在钾盐含量预测中的应用
3. 基于多元线性回归的能耗模型
4. DE算法在多元线性回归模型参数估计中的应用
5. 外围股指对上证综指波动的多元线性回归模型
6. 多元线性回归分析的实例研究
7. 矩阵法在多元线性回归模型上的应用研究
8. MATLAB语言在多元线性回归中的应用
9. 系数为一般模糊数的多元线性回归模型
10. 多元线性回归分析在车内噪声预测的应用研究
11. 多元线性回归模型在物流需求预测中的应用
12. 多元线性回归在平均工资预测中的应用研究
13. 多元线性回归模型预测城市日水量
14. 矩阵法在多元线性回归模型上的应用研究
15. 多元线性回归分析在我国煤炭需求研究中的应用
16. 广义可加模型与经典线性回归模型的比较研究
17. 多元线性回归模型在钾盐含量预测中的应用
18. 基于多元线性回归的碳配额价格预测模型研究
19. 多元线性回归模型及股票板块指数预测
20. 多元线性回归模型预测2020年奥运会成绩
21. 多元线性回归在卫生检验中的应用
22. 应用多元线性回归与多元逐步回归模型研究骨肉瘤的预后影?…
23. 基于多元线性回归的拍照任务定价模型
24. 多元线性回归模型的研究和应用

32. 多元线性回归模型异方差检验研究
33. 多元线性回归的轿车产量预测模型研究
34. 基于多元线性回归的任务定价规律模型
35. 基于多元线性回归模型预测分析的实例研究
36. 基于多元线性回归的任务定价规律模型
37. 多元线性回归模型在钾盐含量预测中的应用
38. 试论多元线性回归模型在股票定价中的运用
39. 基于多元线性回归模型的灯柱沉降预测
40. 多元线性回归模型在ETC客户发展预测中的应用研究
41. 多元线性回归模型在物流成本预测中的应用
42. 多元线性回归模型的简单运用
43. 基于多元线性回归的房价预测模型
44. 多元线性回归统计预测模型的应用
45. 多元线性回归模型在警力资源配置中的应用
46. 多元线性回归中多重共线性的研究
47. 灰色多元线性回归模型及其应用
48. 基于多元线性回归的任务定价规律模型
49. 关于具有线性约束的多元线性回归模型的注记
50. 基于MATLAB的高校学费多元线性回归模型