第 20 卷 第 6 期　　辽宁工程技术大学学报（自然科学版）　　2001 年 12 月
Vol.20, No.6　　Journal of Liaoning Technical University（Natural Science）　　Dec., 2001

文章编号：1008-0562(2001)03-0838-02

基于 MATLAB 的岭回归分析程序设计及其应用

曾繁会，吕渭济
（辽宁工程技术大学 工商管理学院，阜新 123000）

摘　要：岭回归分析是多元线性回归分析中的一种方法，在实际应用中经常遇到。本文通过设计 MATLAB 中的 Ridge1 函数程序，介绍如何利用 MATLAB 进行岭回归分析。

关键词：岭回归；MATLAB；岭回归分析；岭回归系数；程序设计
中图号：O 212.4　　文献标识码：A

0 引　言

岭回归亦称“脊回归估计”、“岭估计”，是一种改进最小二乘估计的方法，适用于自变量 $x_1,x_2,..,x_p$ 间相关性强时，或某些变量的变化范围太小时，也即线性回归模型中正规方程的系数矩阵 $X'X$ 接近奇异时的情形。在这种情况下，用传统的基于最小二乘法估计参数的多元线性回归、逐步回归等方法往往不能得到令人满意的结果，甚至有的结果与专业知识不一致，通常可以采用岭回归分析方法。

MATLAB（5.3 版本）的数值计算功能强大，又便于进行输出结果可视化的程序设计，它的统计工具箱（Statistic Toolbox）中的功能函数有 200 多个。MATLAB 的操作简便、可扩充性等诸多优点决定了它将在 SAS 等概率统计软件领域中占据及其重要的地位。笔者发现其中用于岭回归分析的函数程序 Ridge.m 中有几处欠佳，如其中的标准化处理、回归系数的求解。本文意在对其 Ridge.m 进行改进，并利用改进后的程序 Ridge1.m 作基于 MATLAB 的岭回归分析的应用。

1 岭回归程序(Ridge1.m)的设计

设有 p 个自变量 $x_1,x_2,..,xp$,一个因变量 y，进行 n 次统计得到数据表 $Xnxp,Ynx1$。对于岭参数 $k(k\in[0,1])$,估计岭回归系数的步骤如下：

(1) 将原始数据 $x_1,x_2,..,xp,y$ 作标准化变换，变换后均值为 0，标准差为 1。X,Y 矩阵分别变为 Z,Y。

(2) 对于 $k\in[0,1]$，求 $\sqrt{k(n-1)}$。

(3) 将标准化变换后的矩阵 $Znxp,Ynx1$ 添加伪样本数据变为 $Zplus(n+p)$xp,$Yplus(n+p)$x1.

(4) 利用 MATLAB 中 Regress 函数拟合过原点的多元线性回归方程，所估计出的回归系数即为岭回归系数 $\beta i(k)(i=1,2,..,p)$。

(5) 在应用程序中通过 MATLAB 的画图语句 Plot 绘出 βi(k) 随 k 变化的趋势，决定选择合适的 k 对应的 βi(k)作为最后的岭回归系数。

岭回归函数 Ridge.m 程序清单如下：

```
function [b,bint,r,rint,stats] = ridge1(Y,X,k)
[n,p] = size(X);
mx = mean (X); my = mean (Y); stdx =
std(X);stdy=std(Y);
idx = find(abs(stdx) < sqrt(eps));
MX = mx(ones(n,1),:);STDX = stdx(ones(n,1),:);
Z = (X - MX) ./ STDX;Y=(Y-my)./stdy;
pseudo = sqrt(k*(n-1)) * eye(p);
Zplus = [Z;pseudo];Yplus = [Y;zeros(p,1)];
 [b,bint,r,rint,stats] = regress(Yplus,Zplus);
```

注：在 Ridge1.m 中若用 Regress 求岭回归系数 $\beta i(k)$的同时也求出常数项，则可直接将其对应的回归模型用于经济预测及决策分析中。

应用程序设计

应用程序的数据表见[1]，福建省 1991 年 9 个地区的婴儿死亡率及相关指标。采用岭回归分析（Ridge1.m）来比较各种因素对婴儿死亡率的相关次序及数量程度。

程序中数据矩阵 x_9x_6，行指标为地区编号 1-9，列指标 $X=[x_1,x_2,..,x_6]$分别为 x_1：从事乡妇儿保工作年限 2 年以上的人员占乡妇儿保人员比重(%)；x_2：7 岁以下儿童系统管理率(%)；x_3：3 岁以下儿童生长发育监测率(%)；x_4：年人均收入（元）；x_5：文盲、半文盲率（%）；x_6：乡级妇儿保人员培训比例（%）；Y_{nx1}：婴儿死亡率(%)。程序清单如下：

```
x=[71.35 22.90 3.76 1158.18 12.20 55.87;
    67.92 34048 17.11 1494.38 19.82 56.60;
    79.38 24.91 33.60 691.56 16.17 92.78;
    87.97 10.18 0.73 923.04 12.15 24.66;
    59.03 7.71 3.58 696.92 13.50 61.81;
    55.23 22.94 1.34 1083.84 10.76 49.79;
    58.30 12.78 5.25 1180.36 9.58 57.02;
    67.43 9.59 2.92 797.72 16.82 38.29;
    76.63 15.12 2.55 919.49 17.79 32.07];
y=[28.46;27.76;26.02;33.29;40.84;44.50;28.09;46.24;
45.21];
x'*x；count=0;kvec=0.1:0.1:1;
for k=0.1:0.1:1
```

收稿日期：2001-01-09
作者简介：曾繁会（1970-），女，辽宁 阜新人，讲师，理学学士.本文编校：唐巧凤

```
    count=count+1;
    [b,bint,r,rint,stats]=ridge1(y,x,k);bb(:,count)=b;
     stats1(count,:)=stats;
end
bb'，stats1
plot(kvec',bb),xlabel('k'),ylabel('b','FontName','Symbol')
```

运行的部分结果如下：（*bb*矩阵各行分别为岭系数*k*=0.1,0.2,..0.9时的岭回归系数,stats1中各行分别为各行岭系数对应的R^2统计量和F以及*P*值。图1显示了岭系数$\beta_i(k)(i=1,2,..,6)$随k的变化情况。）

```
bb=[-0.492 0   -0.212 1   -0.229 4   -0.408 5
0.4557    -0.4636
    -0.400 1   -0.172 7   -0.270 2   -0.371 5
0.383 8    -0.378 0
    -0.341 9   -0.153 5   -0.272 8   -0.337 4
0.329 4    -0.334 4
    -0.299 6   -0.141 5   -0.266 3   -0.308 3
0.287 2   -0.304 9
    -0.267 0   -0.132 8   -0.257 1   -0.283 6
0.253 6    -0.282 6
    -0.240 9   -0.125 9   -0.247 3   -0.262 6
0.226 3    -0.264 6
    -0.219 5   -0.120 2   -0.237 8   -0.244 6
0.203 6    -0.249 4
    -0.201 6   -0.115 2   -0.228 6   -0.228 9
0.184 6    -0.236 4
    -0.186 5   -0.110 9   -0.220 0   -0.215 2
0.168 4   -0.224 9
    -0.173 4   -0.107 0   -0.212 0   -0.203 1
0.154 5   -0.214 7]
stats1 =
     [0.785 2     6.579 2     0.007 6
     0.705 4     4.310 9     0.028 0
     0.643 9     3.255 1     0.059 3
     0.594 2     2.635 5     0.097 9
     0.552 8     2.225 2     0.140 3
     0.517 7     1.932 0     0.184 1
     0.487 4     1.711 4     0.227 8
     0.460 9     1.538 8     0.270 3
     0.437 5     1.399 8     0.311 0
     0.416 6     1.285 2     0.349 6 ]
```

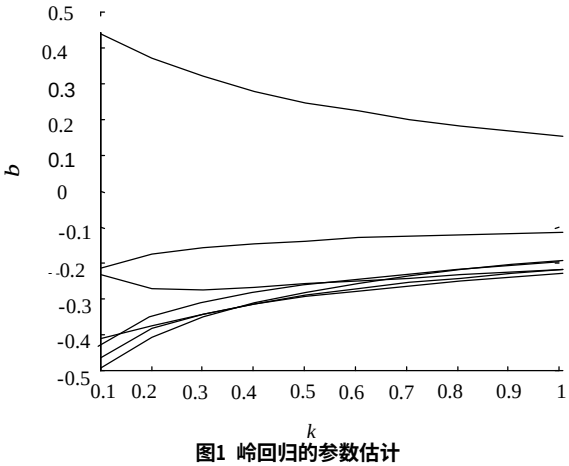

图1 岭回归的参数估计

Fig.1　the parametric estimation of ridge regression

从运行结果及图1可见，$k \geq 0.7$时每个变量相应的岭回归系数变化较为稳定，因而可选k=0.7，建立岭回归方程

$$y=-0.219\ 5x_1-0.120\ 2x_2-0.237\ 8x_3-0.244\ 6x_4+0.203\ 6x_5-0.249\ 4x_6$$

2　结　论

(1) x_5（文盲、半文盲率）与婴儿死亡率之间呈正相关，其它变量与婴儿死亡率呈负相关；(2) 对婴儿死亡率影响最大的是x_6（乡级妇儿保人员培训比例），其次是 x_4（年人均收入）、x_3（3岁以下儿童生长发育监测率）、x_1（从事乡妇儿保工作年限2年以上的人员占乡妇儿保人员比重）。

参考文献：

[1]　田　俊.1999岭回归分析的SAS程序设计[J].数理统计与管理 2000，18（3）：53-55.

[2]　李 涛. Matlab工具箱应用指南—应用数学篇[M].北京：电子工业出版社，2000. 63-64.

Program Design and the Use of Ridge Regression Based on MATLAB

ZENG Fan-hui,　LU Wei-ji

(LiaoNing Technical University, fuxin 123000,China)

Abstract： Ridge regression is a method for multi-variety linear regression analysis and often applied in practical data analysis. The presented paper gives a practical method to estimate coefficients of ridge regression equation with MATLAB by designing the program .

Key words： ridge regression; MATLAB; ridge regression analysis; coefficients of ridge regression; program design