·科研方法专题·

岭回归分析

胡良平[1,2]*

(1. 军事科学院研究生院,北京 100850;
2. 世界中医药学会联合会临床科研统计学专业委员会,北京 100029
* 通信作者:胡良平,E-mail:lphu812@sina.com)

【摘要】 本文目的是介绍岭回归分析的概念、作用以及用软件实现计算的方法。先介绍有关的基本概念,再介绍基本原理和实施步骤,最后,通过一个实例并基于 SAS 软件演示如何实施岭回归分析。结果表明:对于计量因变量,要想构建高质量的多重线性回归模型,较好的建模策略是基于派生变量建立初步模型,在此基础上,再进行岭回归分析。

【关键词】 多重共线性;派生变量;标准化回归系数;岭回归分析;岭迹
中图分类号:R195.1 文献标识码:A doi:10.11886/j.issn.1007-3256.2018.03.001

Ridge regression analysis

Hu Liangping[1,2]*

(1. *Graduate School, Academy of Military Sciences PLA China, Beijing* 100850, *China*;
2. *Specialty Committee of Clinical Scientific Research Statistics of World Federation of Chinese Medicine Societies, Beijing* 100029, *China*
* *Corresponding author: Hu Liangping, E-mail: lphu812@sina.com*)

【**Abstract**】 The aim of this paper was to introduce the concepts and functions and the calculation methods by using the statistical software of the ridge regression analysis. Firstly, the basic concepts of the ridge regression analysis were introduced. Secondly, the basic principle and implementation steps of the ridge regression analysis were described. Finally, the ridge regression analysis was demonstrated through a real example based on the SAS software. The results were shown that the best modeling approach on the multiple linear regression can be achieved by following the two steps bellow: ① constructing the basic multiple linear model based on the derived variables; ② modeling the multiple linear ridge regression model based on the initial model mentioned before.

【**Keywords**】 Multicollinearity; Derived variable; Standardized regression coefficient; Ridge regression analysis; Ridge trace

1 概 述

1.1 何为岭回归分析

岭回归分析是在构建多重线性回归模型时,对基于“最小二乘原理”推导出的估计回归系数的计算公式作一下校正,使回归系数更稳定。

1.2 岭回归分析应用的场合

当自变量之间存在较强的多重共线性时,求得的多重线性回归模型很不稳定;尤其是某些自变量前回归系数的正负号与实际问题的专业背景不吻合。此时,岭回归分析有可能较好地解决前述提及的问题。

1.3 岭回归分析的原理

多重线性回归方程的回归系数可以表示为

$$\beta=(XX)^{-1}XY \quad (1)$$

其中 X 为自变量的 $n\times m$ 阶矩阵,X' 为 X 的转置,$(X'X)$ 为对称的 $m\times m$ 方阵,$(X'X)^{-1}$ 为 $X'X$ 的逆矩阵,Y 为因变量的 $n\times 1$ 向量,β 为待估参数,即回归系数的 $m\times 1$ 向量。这里的 n 为观测数,m 为待估计的回归系数(含截距项)的个数。当 $X'X$ 矩阵不满秩或至少有一个特征根很小(即接近于零)时,则矩阵 X 为病态矩阵,此时用最小二乘法进行回归系数的估计就会产生较大的偏差,使 $\hat{\beta}$ 很不稳定,在具体取值上与真值有较大的偏差,甚至有时会出现系数的正负号与实际不符的情况。为此,Hoerl 和 Kennard 于 1970 年提出了岭估计(Ridge estimate)方法,即对多重线性回归模型回归系数的估计方法如下[1]:

$$\beta(k)=(X'X+kI_m)^{-1}X'Y \quad (2)$$

即在矩阵 $X'X$ 的主对角线元素上加上一个非负因子 k,其中 I_m 为 m 阶单位矩阵,$k>0$ 称为岭参数或偏参数。如果 k 取与试验数据 Y 无关的常数,则 $\beta(k)$ 为线性估计,否则 $\beta(k)$ 为非线性估计。取不同的 k,得到不同的岭估计。故上式实际定义了一个很大的估计类。特别当 $k=0$ 时,$\beta(k)=(X'X)X'Y$ 就

项目基金:国家高技术研究发展计划课题资助(2015AA020102)

是β的最小二乘估计。当$k\in[0,+\infty)$时,对于每个i,$\beta(k)$的第i个分量$\beta_i(k)$的值均为k的函数,在直角坐标系中,点$[k,\beta_i(k)]$所构成的变化轨迹,称为岭迹。随着k的增大,$\beta_i(k)$的绝对值均趋于不断变小[由于自变量之间可能存在相关性,个别$\beta_i(k)$可能有小范围的向上波动或改变正、负号],若$k\to+\infty$,则$\beta(k)\to0$。

与最小二乘估计相比,岭估计是把$X'X$换成了$X'X+kI$得到的。从直观上讲,当X为病态时,$X'X$的特征根至少有一个非常接近于0,而$X'X+kI$的特征根则变成$\lambda_1+k,\cdots,\lambda_m+k$,它们中的某些接近于0的特征根就会得到改善,从而"打破"原来设计阵的复共线性,使得岭估计比最小二乘估计有更小的均方误差[$MSE(\hat{\beta}(k))<MSE(\beta)$]。

1.4 如何实施岭回归分析

如果在进行多重线性回归分析时,从专业上或通过共线性诊断得知自变量间存在多重共线性,那么可以考虑用岭回归分析进行参数估计,在进行岭估计时通常采用以下几步:

第1步,岭回归分析通常要先对X变量作中心化和标准化处理,以使不同自变量处于同样数量级上而便于比较。

第2步,确定k值。

引入岭估计的目的是减少均方误差,所以其关键点就是找出合适的k值,使量$MSE(\hat{\beta}(k))$达到最小。但是对k值的确定在应用中是十分困难的,主要是因为k的取值依赖于模型的未知参数β和σ^2,并且这种依赖关系没有显式表示。为此,统计学家们提出了十余种选k的方法,但是很难说它们中的哪一种是最优的。其中的主要方法如下:

(1)岭迹法

岭迹法主要是通过将$\beta(k)$的分量$\beta_i(k)$的岭迹画在同一幅图上,从图中选择尽可能小的k值,使得各回归系数的岭估计大体稳定,即各分量在图上的岭迹曲线趋于平行于X轴。选择k值的一般原则主要有:①各回归系数的岭估计基本稳定;②用最小二乘估计时符号不合理的回归系数,其岭估计的符号将变得合理;③回归系数的大小要与实际相符,即从专业上讲对因变量影响较大的自变量其系数的绝对值也较大;④均方误差增大不太多。

(2)方差膨胀因子法

方差膨胀因子c_{jj}度量了多重共线性的严重程度,一般当$c_{jj}>10$时,模型就有严重的多重共线性,如果计算岭估计$\beta_i(k)$的协方差阵为:

$$\mathrm{cov}(\hat{\beta}(k))=\sigma^2(X'X+kI)^{-1}X'X(X'X+kI)^{-1}=\sigma^2(c_{ij}(k))$$

上式中矩阵$C_{ij}(k)$的对角元素$c_{jj}(k)$就是岭估计的方差膨胀因子,不难看出,$c_{jj}(k)$随着k的增大而减小。应用方差膨胀因子选择k的经验做法是:选择k使所有方差膨胀因子$c_{jj}(k)\leq10$。

此外还有C_p准则法、Hoerl-Kennad公式法、Mcdorard-Gararneau法、双h公式法等,因SAS的REG过程主要可以运用岭迹法及方差膨胀因子法,所以其他的方法在此不作介绍,有兴趣的读者可以参考相关的文献。

第3步,根据岭迹图进行变量筛选及重新确定k值。

岭迹图不仅能够对k做出确定,还可以根据自变量的岭迹曲线对自变量进行筛选,也就是说可以根据自变量的岭迹曲线来判断该变量是否可以进入回归方程。把岭迹应用于回归分析中自变量的选择,其基本原则为:

(1)去掉岭回归系数比较稳定且绝对值比较小的自变量。这里岭回归系数可以直接比较大小,因为设计阵X是假定已经中心标准化了的。

(2)去掉岭回归系数不稳定但随着k值的增加迅速趋于零的自变量。

(3)去掉一个或若干个具有不稳定岭回归系数的自变量。如果不稳定的岭回归系数很多,究竟去掉几个,去掉哪几个,并无一般原则可遵循。这要结合已找出的复共线性关系以及去掉后重新进行岭回归分析的效果来决定。

第4步,对模型进行表达及作出专业结论。

在进行岭估计后,应根据所估计的参数写出回归方程,并结合专业知识判断方程中各自变量的系数及正负号是否符合实际情况。最后根据回归系数的大小来判断各自变量对因变量影响的大小及根据所求得的回归方程进行预测。

2 基于岭回归分析解决实际问题

2.1 问题与数据结构

沿用文献[2]中的"问题与数据",并基于派生变量得到的"最优回归模型"所决定的"数据集",来提出下面的"新问题":即"weight"的回归系数为"-88.00801",这个"负值"表明体重越重的人收缩压(SBP)越低,这似乎不符合临床专业知识。尽管计算出来的因变量的预测值在专业上都成立,而且,模型的残差方差=122.32418、$R^2=0.9931$,这些结

果都提示所构建的多重线性回归模型很好。但毕竟存在回归系数的正负号不符合专业知识的“严重瑕疵”,这是一个需要彻底解决的“疑难问题”!

2.2 所需要的 SAS 程序

解决上述提及的“疑难问题”的一个有效方法就是使用“岭回归分析”。所需要的 SAS 程序如下:

```
data a1;
  input id age height weight bmi sbp;
cards;
(此处输入文献[2]表 1 中 50 行 6 列数据)
;
run;
  /* 以上程序为了创建数据集 a1 */
data a2;
  set a1;
x1 = age * age; x2 = age * height; x3 = age * weight;
x4 = age * bmi; x5 = height * height; x6 = height * weight;
x7 = height * bmi; x8 = weight * weight; x9 = weight * bmi;
x10 = bmi * bmi;
run;
  /* 以上程序是在数据集 a1 基础上创建数据集 a2,它增添了 10 个派生变量 */
proc reg data = a2;
model sbp = age height weight bmi x1 - x10/noint selection = backward sls = 0.05;
quit;
  /* 以上程序是基于数据集 a2 拟合文献[2]中那个“最佳”回归模型 */
symbol1 v = x c = blue;
symbol2 v = circle c = black;
symbol3 v = square c = red;
symbol4 v = triangle c = green;
symbol5 v = dot c = yellow;
symbol6 v = # c = orange;
symbol7 v = % c = purple;
symbol8 v = $ c = blue;
legend1 mode = protect position = (bottom right inside)
across = 3 cborder = black offset = (0,0) label = (color = blue position = (top center) 'independent variables')
cframe = white;
  /* 以上程序为了设置绘图的基本条件 */
proc standard data = a2 m = 0 s = 1 out = a3;
run;
  /* 以上程序为了对数据集 a2 进行标准化变换 */
  /* 因前面用后退法筛选自变量,故仅保留有统计学意义的自变量,列在下面的 model 语句 */
proc reg data = a3 outest = b1 ridge = 0 to 0.1 by 0.01;
model sbp = age weight x3 x6 - x10/noint;
plot /ridgeplot vref = 0 lvref = 1 nomodel legend = legend1 nostat;
quit;
  /* 以上程序是基于标准化变换后的数据集 a3 进行岭回归分析并绘制出岭迹图 */
proc print data = b1;
run;
  /* 以上程序为了输出岭回归分析的计算结果 */
proc reg data = a2 outest = b2 ridge = 0 to 0.1 by 0.01;
model sbp = age weight x3 x6 - x10/noint;
quit;
  /* 以上程序是基于数据集 a2 进行岭回归分析以便获得与原变量对应的岭回归分析结果 */
  /* 其中,“ridge = 0 to 0.1 by 0.01”就是让岭参数 k 取一系列数值代入建模 */
proc print data = b2;
run;
  /* 以上程序是输出与原变量对应的岭回归分析结果 */
```

【SAS 程序说明】在以上的 SAS 程序中,都用“/* …… */”注释语句作了说明。

2.3 SAS 输出结果及其解释

以上 SAS 程序输出的结果非常多,因篇幅所限,此处从略。下面仅摘要给出最主要的结果。

由绘制出的岭迹图(图略)可知:当岭参数 $k=0.01$ 时,全部 8 个自变量的回归系数就已经趋向于稳定状态,但结合数据集 b1 的输出结果可从数值清楚地看出:当岭参数 $k=0.08$ 时,全部 8 个自变量的回归系数的取值波动都在小数点后第 2 位上,因为 k 的取值越大,自变量的回归系数数值波动就会越小,但残差方差会越大。故要求不高时,本例可取 $k=0.01$;要求稍高一点时,可取 $k=0.08$。

(1)标准化条件下岭回归分析所得到的全部自变量的回归系数

由数据集 b1 可获得标准化条件下岭回归分析所得到的全部自变量的回归系数:

$k=0.01$ 和 $k=0.08$ 时，全部自变量的标准化 回归系数：

k	age	weight	x3	x6	x7	x8	x9	x10
0.01	1.36176	0.370	-1.24812	0.606	0.3183	-0.0352	-0.142	0.1000
0.08	0.57737	0.228	-0.26936	0.283	0.1905	0.0213	-0.048	-0.0693

（2）基于原变量的岭回归分析所得到的全部自变量的回归系数

k	age	weight	x3	x6	x7	x8	x9	x10
0.01	1.40070	0.1400	-0.006767	0.00287	0.01370	-0.000034	-0.00110	0.00592
0.08	0.59388	0.0865	-0.001460	0.00134	0.00820	0.000021	-0.00037	-0.00410

（3）基于原变量的常规多重线性回归分析所得到‘最佳结果’的全部自变量的回归系数

age	weight	x3	x6	x7	x8	x9	x10
1.82182	-88.00801	-0.0097	0.64569	4.32456	-0.05835	0.78530	-2.62458

【说明】比较上面的“（2）”与“（3）”中各自变量前的回归系数可知，引入派生变量并构建出“最优回归模型”后，weight 的回归系数为“-88.00801”［见上面的“（3）”标题之下第 2 行］，该负值不符合临床专业知识；经过岭回归分析后，得出：weight 的回归系数为“0.0865”［见上面的“（2）”标题之下 $k=0.08$ 的那一行］，这个系数是正值，与临床专业知识相符。由此可知，在创建多重线性回归模型的过程中，引入派生变量并采取多种方法进行自变量筛选，在获得“最佳多重线性回归模型”后，再采用岭回归分析进一步优化多重线性回归模型，这样在获得了较小残差方差的多重线性回归模型的基础上，又能使少数自变量回归系数的正负号不符合专业要求的问题得到合理解决。

参考文献

［1］ 茆诗松. 统计手册［M］. 北京：科学出版社，2006：499-502.
［2］ 胡良平. 主成分分析应用（Ⅰ）——主成分回归分析［J］. 四川精神卫生，2018，31（2）：128-132.

（收稿日期：2018-05-03）
（本文编辑：唐雪莉）

科研方法专题策划人——胡良平教授简介

胡良平，男，1955 年 8 月出生，教授，博士生导师，曾任军事医学科学院研究生部医学统计学教研室主任和生物医学统计学咨询中心主任、国际一般系统论研究会中国分会概率统计系统专业理事会常务理事和北京大学口腔医学院客座教授；现任世界中医药学会联合会临床科研统计学专业委员会会长、中国生物医学统计学会副会长，《中华医学杂志》等 10 余种杂志编委和国家食品药品监督管理局评审专家。主编统计学专著 45 部，参编统计学专著 10 部；发表第一作者学术论文220余篇，发表合作论文 130 余篇，获军队科技成果和省部级科技成果多项；参加并完成三项国家标准的撰写工作；参加三项国家科技重大专项课题研究工作。在从事统计学工作的 30 年中，为几千名研究生、医学科研人员、临床医生和杂志编辑讲授生物医学统计学，在全国各地作统计学学术报告 100 余场，举办数十期全国统计学培训班，培养多名统计学专业硕士和博士研究生。近几年来，参加国家级新药和医疗器械项目评审数十项、参加 100 多项全军重大重点课题的统计学检查工作。归纳并提炼出有利于透过现象看本质的“八性”和“八思维”的统计学思想，独创了逆向统计学教学法和三型理论。擅长于科研课题的研究设计、复杂科研资料的统计分析与 SAS 实现、各种层次的统计学教学培训和咨询工作。