Advances in Applied Mathematics 应用数学进展, 2021, 10(1), 92-97
Published Online January 2021 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2021.101011

Hans汉斯

一类改进广义岭估计的实例研究

刘金灵

云南财经大学统计与数学学院，云南 昆明
Email: helen_092020@163.com

收稿日期：2020年12月14日；录用日期：2021年1月3日；发布日期：2021年1月19日

摘　要

本文对广义岭回归方法进行了改进，并在真实数据中进行了实例验证。改进广义岭估计主要对广义岭估计在线性回归模型中当存在若干较大异常值影响模型精度的情况进行了修正，加入适当的修正参数，使模型达到对数据更精确的拟合以及预测作用，并对中俄贸易数据进行了实例验证。

关键词

多重共线性，改进广义岭估计，异常值处理，修正参数

Improved Generalized Ridge Regression and Its Application

Jinling Liu

School of Statistics and Mathematics, Yunnan University of Finance and Economics, Kunming Yunnan
Email: helen_092020@163.com

Received: Dec. 14th, 2020; accepted: Jan. 3rd, 2021; published: Jan. 19th, 2021

Abstract

In this paper, the generalized ridge regression method is improved and verified by real data. The improved generalized ridge regression mainly corrects the generalized ridge estimation in the linear regression model when there are some large outliers that affect the accuracy of the model. Appropriate correction parameters are added to make the model achieve more accurate data fitting and prediction effects. Finally, we verified it in trade data.

Keywords

Multicollinearity, Improved Generalized Ridge Regression, Outlier Handling,

文章引用: 刘金灵. 一类改进广义岭估计的实例研究[J]. 应用数学进展, 2021, 10(1): 92-97.
DOI: 10.12677/aam.2021.101011

Correction Parameters

1. 引言

本文主要提出改进广义岭估计。广义岭估计是 Horel 和 Kennard 提出的对于一般岭估计岭参数设置改进的一种方法；Hawkins 和 Yin 提出了针对实际情况中样本量 n 小于未知参数的 p 形成降秩矩阵的岭回归的快速迭代算法；由于实际研究中还没有一种能够确定岭参数的完美的方法，或多或少会造成估计误差不稳定，游、王和刘(2002) [1]提出了广义岭估计的直接解法，跳过岭参数的估计，直接求得具有最小均方误差的解；凌和叶对平衡损失下的双 h 岭估计的优良性进行了验证，给出了风险函数的表达式。改进岭估计在基于前人对该领域研究的基础上，适当改进，对于实际数据若干异常值较大影响回归结果的问题提出新的方法予以解决，将大大提高模型准确度。

2. 模型建立及数据预处理

2.1. 模型建立

在将中蒙俄 2008~2018 的贸易数据整理，所有变量重新命名。定义 Y_1 和 Y_2 为因变量，X_1、X_2、X_3、X_4、X_5、X_7、X_8、X_{10}、X_{12} 和 X_{14} 是对应 Y_1 的自变量；X_1、X_2、X_3、X_4、X_6、X_7、X_9、X_{11}、X_{13} 和 X_{15} 是对应 Y_2 的自变量。所有数据均能在统计局官网获得，这里不予介绍。建立如下线性模型：

$$Y_1 = X'\beta_1 + \varepsilon,\ \varepsilon \sim N\left(0, \sigma^2 I\right),$$

$$Y_2 = X''\beta_2 + \varepsilon,\ \varepsilon \sim N\left(0, \sigma^2 I\right),$$

其中：$X' = \left[1\ X_1 \sim X_{14}\right]$，$X'' = \left[1\ X_1 \sim X_{15}\right]$，$\beta_1, \beta_2 \in R^{11}$。

2.2. 多重共线性诊断

如果数据设计阵存在多重共线性问题，就不能简单的用最小二乘估计对其进行参数估计，虽然看似是对数据进行了最小均方误差的拟合，但是有时可能导致拟合的结果完全和事实背离，造成数据分析无效等后果。为了使数据得到更好的处理，我们将引用岭估计、主成分回归和逐步回归等方法才能对数据进行更加合理的处理，所以进行多重共线性诊断是十分必要的，这关系到模型的合理性以及对数据的解释程度。

对 Y_1、Y_2 设计阵进行诊断

对矩阵 Z_1 和 Z_2 进行中心化和标准化：

$$Z_1 = \left[\left(\frac{X_{1ij} - \mu_j}{\sigma_j}\right)_{ij}\right], Z_2 = \left[\left(\frac{X_{2ij} - \mu_j}{\sigma_j}\right)_{ij}\right],$$

其中 μ_j 和 σ_j 分别表示矩阵 Z 的第 j 列的均值和方差。计算 $Z_1'Z_1$ 和 $Z_2'Z_2$ 的特征值。

引入条件数公式 $k = \lambda_1 / \lambda_{p-1}$，$k$ 的值越大，说明存在的共线性越严重。

$$k_1 = \frac{\lambda_1}{\lambda_{p-1}} = \frac{1313.9}{2.9} \approx 453.81 > 1000,$$

$$k_2 = \frac{\lambda_1}{\lambda_{p-1}} = \frac{1301.2}{0.9} \approx 1408.03 > 1000,$$

从 k_1 和 k_2 的取值来看，不管是 Y_1 还是 Y_2 的设计阵，都存在略严重的多重共线性。此时如果要对 Y_1 和 Y_2 分别进行线性拟合，就不能再使用最小二乘估计了，处理此类问题常见的方法有常见的一般岭估计、逐步回归分析、主成分回归。

3. 理论推导

3.1. 改进广义岭估计理论推导

线性模型拟合中，对于参数的估计最常用的的是最小二乘估计，但是在遇到设计阵存在多重共线性的时候，往往最小二乘估计不是最好的，所以 A.E.Hoerl 提出了岭估计，引入岭参数 $kI_{p\times p}$ 来解决多重共线性问题，后来广义岭估计的提出又定义了岭参数 $k_1 \sim k_p$ 可以不相同来减小拟合结果的均方误差。叶、朱(1998) [2]提出了奇异值分解的解法。本节在前人建立的广义岭估计的基础上又提出对于若干异常值较大情况的一种广义岭估计的修正方法，改进广义岭估计。先对广义岭估计进行回顾，

岭估计参数 $\hat{\beta}$ 的估计方法：

$$\hat{\beta}(k) = \left(X'X + kI_{p\times p}\right)^{-1} X'Y = \left(X'X + kI_{p\times p}\right)^{-1} \left(X'X\right)\hat{\beta},$$

建立典则方程：

$$Y = \alpha_0 \mathbf{1} + Z\alpha + \varepsilon,$$

典则回归系数的岭估计：

$$\hat{\alpha}(k) = \left(\Lambda + kI_{p\times p}\right)^{-1} Z'Y = \left(\Lambda + kI\right)^{-1} \Lambda\hat{\alpha},$$

其中 $\hat{\alpha} = \left(\Lambda\right)^{-1} Z'Y$，$\Lambda = \phi'\left(X'X\right)\phi = \mathrm{diag}\left(\lambda_1, \cdots, \lambda_p\right)$，$\phi = \left(\varphi_1, \cdots, \varphi_p\right)$ 是 λ_i 对应的特征根，$Z = X\phi$，$\hat{\alpha} = \phi'\hat{\beta}$。定义 $K = \mathrm{diag}\left(k_1, \cdots, k_p\right)$，$k_i$ 是不必全相等的岭参数，于是定义该模型广义岭估计的典则参数和平差参数估计如下：

$$\begin{aligned} \hat{\alpha}(K) &= \left(Z'Z + K\right)^{-1} Z'Y = \left(\Lambda + K\right)^{-1} Z'Y = \left(\Lambda + K\right)^{-1} \Lambda\hat{\alpha} = \left(\Lambda + K\right)^{-1} C, \\ \hat{\beta}(K) &= \phi\hat{\alpha}(K) = \phi\left(\Lambda + K\right)^{-1} \Lambda\phi'^{\hat{\beta}} = \phi\left(\Lambda + K\right)^{-1} \phi' X'Y = \left(X'X + \phi K\phi'\right) X'Y, \end{aligned} \tag{A}$$

当 $K = kI_{p\times p}$ 时，广义岭估计就简化成为了一般的岭估计。

由于实际数据或许存在若干较大异常值情况，故提出修正参数对广义岭估计进行改进：

$$\gamma_{ij} = 1 - I\left\{\left|x_{ij} - \overline{x}_j\right| > Q_{ja}\right\}\left(x_{ij} - \overline{x_j}\right),$$

其中 $I\left\{x_{ij} > Q_{ja}\right\}$ 为示性函数，Q_{ja} 为对应设计阵 X 的第 j 列进行中心化并求绝对值后的 a%分位数，$\overline{x_j}$ 为对应设计阵 X 的第 j 列的均值。提出修正参数 γ_{ij}，将消除由于自变量中存在若干较大异常值对模型拟合的影响。

相应的，加入修正参数后的典则方程和平差方程如下：

$$\hat{\alpha}(K) = \left[\left(Z\gamma_{ij}\right)'\left(Z\gamma_{ij}\right) + K\right]^{-1} \left(Z\gamma_{ij}\right)' Y,$$

$$\hat{\beta}(K)=\left[\left(X\gamma_{ij}\right)'\left(X\gamma_{ij}\right)+\phi K\phi'\right]^{-1}\left(X\gamma_{ij}\right)'Y,$$

令 $ZR=Z\gamma_{ij}, XR=X\gamma_{ij}$ 得：

$$\hat{\alpha}(K)=\left[ZR'ZR+K\right]^{-1}ZR'Y,$$

$$\hat{\beta}(K)=\left[XR'XR+\phi K\phi'\right]^{-1}XR'Y,$$

易证 $\mathrm{MSE}\left[\hat{\alpha}(K)\right]=\mathrm{MSE}\left[\hat{\beta}(K)\right]$，这里略去证明过程，于是可以通过 $\hat{\alpha}(K)$ 来作为评价 $\hat{\beta}(K)$ 估计值好坏的指标。

$$\mathrm{MSE}\left[\hat{\alpha}(K)\right]=\mathrm{MSE}\left[\hat{\beta}(K)\right]=\sigma^2\sum_{i=1}^{p-1}\frac{\lambda_i}{\left(\lambda_i+k_i\right)^2}+\sum_{i=1}^{p-1}\frac{k_i^2\alpha_i^2}{\left(\lambda_i+k_i\right)^2}.$$

3.2. 改进广义岭估计解法介绍

对于岭参数的估计，将是建立合理的线性模型的至关重要的一点，虽然前人已经提出了各种确定岭参数的方法和准则，但是到目前为止，并没有一种公认的非常稳健的方法，所以本文将采用游、王和刘(2002) [1]提出的广义岭估计的直接解法 DSGRE (derct solution to generalized ridge estimate)，其内容如下：

由于 λ_i 和 α_i 与 k_i 没有关系且 α_i 是非随机的，为了求得最小的 $\mathrm{MSE}\left[\hat{\alpha}(K)\right]$，故对 $\mathrm{MSE}\left[\hat{\alpha}(K)\right]$ 关于 k_i 进行一阶求导：

$$f'(k_i)=\frac{\mathrm{dMSE}\left[\hat{\alpha}(K)\right]}{\mathrm{d}k_i}=-2\sigma^2\sum_{i=1}^{p-1}\frac{\lambda_i}{\left(\lambda_i+k_i\right)^3}+2\sum_{i=1}^{p-1}\frac{k_i\lambda_i\alpha_i^2}{\left(\lambda_i+k_i\right)^3},$$

令 $f'(k_i)=0$，

解得：

$$k_i=\frac{\sigma^2}{\alpha_i^2},\ i=1,2,\cdots,p-1,$$

由于真实的 σ^2 和 α_i^2 是未知的，故用 σ^2 的估计值 $\widehat{\sigma^2}$ 以及 α_i 的迭代值 $\widehat{\alpha}_i^{(j)}$ 来对 k_i 进行迭代求解：

$$\widehat{k}_i^{(j)}=\frac{\widehat{\sigma^2}}{\widehat{\alpha}_i^{(j)}\widehat{\alpha}_i^{(j)}},\tag{B}$$

那么就可以用第 j 次 $\widehat{\alpha}_i^{(j)}$ 的迭代值来估计与之相同次数的岭参数 k_i，由式(A)可知，第 $j+1$ 次估计为：

$$\widehat{\alpha}_i^{(j+1)}=\frac{c_i}{\lambda_i+\widehat{k}_i^{(j)}},$$

带入式(B)可知：

$$\widehat{\alpha}_i^{(j+1)}=\frac{c_i}{\lambda_i+\dfrac{\widehat{\sigma^2}}{\widehat{\alpha}_i^{(j)}\widehat{\alpha}_i^{(j)}}}=\frac{c_i\widehat{\alpha}_i^{(j)}\widehat{\alpha}_i^{(j)}}{\lambda_i\widehat{\alpha}_i^{(j)}\widehat{\alpha}_i^{(j)}+\sigma^2}.\tag{C}$$

3.3. 解的存在性证明

薛、梁(2002) [3]给出了参数的迭代算法。结合已有的证明过程，对改进广义岭估计的解的存在性予以证明。如果 $\widehat{\alpha}_i^{(j+1)}$ 是一致收敛的，那么经过 $j+1$ 次迭代过程，必然能够求解出 $\widehat{\alpha}_i^{(j+1)}$ 的数值解，故下证

$\widehat{\alpha}_i^{(j+1)}$ 的一致收敛性：

证明：若 $c_i>0$，根据 $\widehat{\alpha}_i^{(j+1)}$ 的形式知：$\widehat{\alpha}_i^{(0)}=\dfrac{c_i}{\lambda_i}>0$，由于 $\sigma^2>0\Rightarrow\widehat{\alpha}_i^{(0)}>\widehat{\alpha}_i^{(1)}>\cdots>\widehat{\alpha}_i^{(\infty)}>0$，故 $\widehat{\alpha}_i^{(j+1)}$ 在 $j\in(0,\infty)$ 上市有界并且单调的，有 Dieichlet 判别法可知，$\widehat{\alpha}_i^{(j+1)}$ 在 $j\in(0,\infty)$ 上是一致收敛的。

若 $c_i<0$，根据 $\widehat{\alpha}_i^{(j+1)}$ 的形式知：$\widehat{\alpha}_i^{(0)}=\dfrac{c_i}{\lambda_i}<0$，由于 $\sigma^2>0\Rightarrow\widehat{\alpha}_i^{(0)}<\widehat{\alpha}_i^{(1)}<\cdots<\widehat{\alpha}_i^{(\infty)}<0$，故 $\widehat{\alpha}_i^{(j+1)}$ 在 $j\in(0,\infty)$ 上是有界并且单调的，有 Dieichlet 判别法可知，$\widehat{\alpha}_i^{(j+1)}$ 在 $j\in(0,\infty)$ 上是一致收敛的。

综上所知：$\widehat{\alpha}_i^{(j+1)}$ 在 $j\in(0,\infty)$ 上是一致收敛的，迭代式 $\widehat{\alpha}_i^{(j+1)}$ 必然有解。

于是，令 $\widehat{\alpha}_i^{(\infty)}=\widehat{\alpha}_i$，则式(C)可以写成如下形式：

$$\widehat{\alpha}_i=\frac{c_i\widehat{\alpha}_i^2}{\lambda_i\widehat{\alpha}_i^2+\sigma^2},\tag{D}$$

解式(D)得：

$$\widehat{\alpha}_i=\begin{cases}0, & c_i=0\text{或}c_i^2-4\lambda_i\sigma^2\geq 0,\\ \dfrac{c_i+\sqrt{c_i^2-4\lambda_i\sigma^2}}{2\lambda_i}, & c_i>0,c_i^2-4\lambda_i\sigma^2\geq 0,\\ \dfrac{c_i+\sqrt{c_i^2-4\lambda_i\sigma^2}}{2\lambda_i}, & c_i<0,c_i^2-4\lambda_i\sigma^2\geq 0.\end{cases}$$

又因为 $\hat{\beta}=\phi\hat{\alpha}$，故该方法直接给出了直接解改进广义岭估计参数的方式。本节将游、王和刘(2002) [1] 提出的 DSGRE 在改进广义岭估计中应用，推导过程验证，也是适用的。

4. 实例探究

基于第 3 节中对于改进广义岭估计的研究，本节将采用该方法对于中蒙俄贸易数据进行回归分析，在广义岭估计的基础上引入修正参数形成改进广义岭估计，训练相应的模型对总体数据进行更加精确地拟合。

4.1. 算法

① 对 X 的每一列进行中心化后再求绝对值得到 $A=\left(\left|x_{ij}-\overline{x_j}\right|\right)_{ij}$；

② 对 $\left|x_{ij}-\overline{x_j}\right|$ 按列进行排序，得到新的排序矩阵 B；

③ 找到排序矩阵每一列的 a%分位数 Q_{ja}，这里不妨取 $a=95$；

④ 通过判别函数将每一列小于对应这一列 a%分位数 Q_{ja} 的元素重置为 0，大于 Q_{ja} 的元素则保留得到矩阵 E；

⑤ A 中每个元素取绝对值得到矩阵 C，用 A 点除 C 得到矩阵 D；

⑥ 所以 $ZR=Z-E.*D$，$XR=X-E.*D$，其中的.*为 MATLAB 中的点乘语句；

⑦ 最后将 ZR 和 XR 进行标准中心化，任用 ZR 和 XR 表示。

4.2. 改进广义岭估计参数 $\hat{\beta}$

利用上述算法对相关参数直接求解：

$$c1_i=\left(51.0\ \ 118.2\ \ 147.7\ \ 103.8\ \ -113.4\ \ 126.7\ \ 45.8\ \ 15.1\ \ 22.0\ \ 14.6\right)^{\mathrm{T}},$$

$$c2_i = (59.21\ \ 36.1\ \ 142.3\ \ 125.9\ \ -48.8\ \ 125.9\ \ 37.4\ \ 13.2\ \ -28.5\ \ 35.9)^{\mathrm{T}},$$

$$\lambda 1_i = (5.0\ \ 44.7\ \ 80.6\ \ 101.2\ \ 196.0\ \ 270.6\ \ 309.3\ \ 330.7\ \ 623.8\ \ 1194.1)^{\mathrm{T}},$$

$$\lambda 2_i = (5.2\ \ 6.3\ \ 45.5\ \ 84.5\ \ 263.6\ \ 274.4\ \ 320.5\ \ 344.8\ \ 621.7\ \ 1193.2)^{\mathrm{T}}..$$

最后被估的回归系数为：

$$\widehat{\beta 1} = \phi\hat{\alpha} = (0\ \ -0.17\ \ -0.05\ \ -0.08\ \ 0.021\ \ -0.25\ \ 0.22\ \ 0.05\ \ 0\ \ 0\ \ 0)^{\mathrm{T}},$$

$$\widehat{\beta 2} = \phi\hat{\alpha} = (0\ \ -0.04\ \ 0.17\ \ -0.08\ \ 0.021\ \ 0.24\ \ 0.12\ \ -0.16\ \ 0.1\ \ 0.04\ \ 0\ \ 0\ \ 0)^{\mathrm{T}}.$$

从回归系数来看，对于响应变量 Y_1 来说， X_1 、 X_{10} 、 X_{12} 、 X_{14} 的系数为 0，与 Y_1 不存在线性关系，其余变量在线性回归中对 Y_1 起着决定性作用。对于响应变量 Y_2 来说 X_1 、 X_{11} 、 X_{13} 、 X_{15} 与 Y_2 几乎没有线性关系，其余变量决定 Y_2 的线性回归。

5. 结论

本文我们对广义岭估计进行了适当的改进，加入了修正参数，这使得岭估计在当数据存在部分异常点时，具有一定的稳健性。同时我们将中蒙俄的贸易数据引入到本文中，对改进广义岭估计进行了实例探究，从结果发现，改进广义岭估计能够估出线性回归系数。我们还将其结果与现有的岭估计、主成分回归、最小二乘法进行了比较，改进广义岭估计估计的结果具有最小的均方误差。岭估计、主成分回归、逐步回归是较为成熟的方法，其估计结果我们没有在文中给出。

参考文献

[1] 游扬声, 王新洲, 刘星. 广义岭估计的直接解法[J]. 武汉大学学报(信息科学版), 2002, 27(2): 175-178.

[2] 叶松林, 朱建军. 矩阵奇异值分解与广义岭估计及其在测量中的应用[J]. 中国有色金属学报, 1998(1): 160-164.

[3] 薛美玉, 梁飞豹. 广义岭估计参数的迭代算法[J]. 福州大学学报(自然科学版), 2002, 30(2): 167-171.