

岭回归在修正多重共线性中的应用

王娟, 梁登星

(成都理工大学管理科学学院, 成都 610059)

摘要:以 2000-2010 年四川省 CPI 的数据及部分影响因素为基础,运用 SPSS17.0 对 CPI 数据建立多元线性回归模型,并基于岭回归对模型中的多重共线性进行修正,得到了修正后的模型,说明此方法具有一定的实用性。

关键词:岭回归;CPI;多重共线性

中图分类号:TB115

文献标志码:A

引言

1934 年,Frish R 提出了多重共线性,指在回归模型中,自变量之间彼此相关。在实际数据中,由于解释变量之间本身存在着各种联系,因此多重共线性是在进行线性回归时存在的问题。在数据的选取小于自变量的个数时,也会存在多重共线性。当存在共线性时,普通最小二乘估计仍是线性无偏估计,不再是有效估计,同时各种变量的显著性检验也会失效,模型的预测也会没有意义^[1]。处理多重共线性的方法主要有:去掉模型中一些不重要的变量,保留重要的变量;对已建立的模型进行主成分分析;岭回归。

CPI (Consumption Price Index) 指的是在一定时期内居民所消费商品及服务项目的价格水平变动趋势和变动程度。CPI 主要是由食品、烟酒及用品、居住、交通通讯、医疗保健个人用品、衣着、家庭设备及维修服务、娱乐教育文化用品及服务组成。从反应 CPI 的八大类别中选取了商品零售价格指数、工业品出厂价格指数、固定资产投资价格指数建立了多元线性回归模型^[2]。运用岭回归的方法,修正了解释变量之间的多重共线性,并得到了合适的回归模型。

1 多重共线性

在多元线性回归模型的经典假设中,最重要的一条假设是解释变量 x_1, x_2, \dots, x_k 之间互不相关,即不存在线性关系。如果存在某些常数 c_0, c_1, \dots, c_p ($p > 2$), 使得线性等式

$$c_1x_1 + c_2x_2 + \dots + c_px_p = c_0 \quad (1)$$

对数据近似成立,这表示,至少有一个 x_k , 可以由其它的变量决定

$$x_k \cong (c_0 - \sum_{j \neq k} c_j x_j) / c_k$$

则称自变量 x_1, x_2, \dots, x_p 之间存在线性关系,即模型当中存在多重共线性^[1]。

2 岭回归估计的定义

2.1 普通最小二乘法

多元回归方程写成矩阵的形式为:

$$Y = X\beta + e \quad (2)$$

其中 Y 和 e 为 $n \times 1$ 向量,其元素为

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + e_i$$

$$(i = 1, 2, \dots, n)$$

其中 y_i 和 e_i 为:

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, e = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

β 为长度为 $(p + 1) \times 1$ 的向量参数,它包含截距 β_0 ,

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$$

X 为 $n \times (p + 1)$ 矩阵:

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

β 的最小二乘估计 $\hat{\beta}$ 是使残差平方和为:

$$RSS(\beta) = \sum (y_i - x_i^T \beta)^2 = (Y - X\beta)^T (Y - X\beta) \quad (3)$$

(x_i^T 为 X 的第 i 行 $i = 1, 2, \dots, n$) 取最小值。

若 $(X^T X)^{-1}$ 存在, β 的最小二乘估计为:

$$\hat{\beta} = (X^T X)^{-1} X^T Y \quad (4)$$

估计量 $\hat{\beta}$ 依赖于统计量 $(X^T X)^{-1}$ 和 $X^T Y$ 。

当第 j 个自变量与其它自变量存在共线性时有:

$$\text{var}(\hat{\beta}_j) = \sigma^2 \left(\frac{1}{1 - R_j^2} \right) \left(\frac{1}{S X_j X_j} \right)$$

其中 R_j^2 为复相关系数的平方, $VIF_j = \frac{1}{1 - R_j^2}$ 为自变量 x_j 的方差扩大因子或方差膨大因子^[1]。

产生共线性后,模型中解释变量的系数估计的方差典型地增大,使得各变量的 t 统计量减小,使得某些重要的解释变量变得不显著。模型存在多重共线性时, t 检验和 F 检验失效,则预测会失败,因此得到的结果往往令人难以接受,结果没有意义。

2.2 岭回归

当自变量之间存在多重共线性时,即 $|X^T X| \approx 0$, 设想给 $X^T X$ 加一个正常数矩阵 kI ($k > 0, I$ 为单位矩阵)。得到岭回归的估计量为

$$\hat{\beta}(k) = (X^T X + kI)^{-1} X^T Y \quad (5)$$

显然,当 $k = 0$ 时,岭回归的估计量即为最小二乘估计;当 $k \rightarrow \infty$ 时,岭回归的估计量趋于 0,因此 k 不宜过大^[3]。

3 实证分析

通过参阅四川省统计年鉴,鉴于数据的可得性,选

取了 2000 - 2010 年的关于 CPI 及其相关影响因素的数据,居民消费价格总指数 (CPI) y , 商品零售价格指数 x_1 , 工业品出厂价格指数 x_2 , 固定资产投资指数 x_3 。

3.1 多重共线性的判断

建立模型如下:

$$y = \alpha x_1 + \beta x_2 + \gamma x_3 + \mu \quad (6)$$

利用 SPSS17.0 对模型进行回归分析,结果见表 1。

表 1 多元线性回归结果

	Unstandardized Coefficients			Collinearity Statistics	
	B	Std. Error	t	Tolerance	VIF
(Constant)	25.916	8.287	3.127		
x1	0.857	0.137	6.234	0.273	3.666
x2	0.006	0.08	0.08	0.306	3.265
x3	-0.106	0.063	-1.698	0.809	1.236

拟合度 $R^2 = 94.9\%$, 模型拟合较好。则初次拟合模型为:

$$y = 25.916 + 0.857x_1 + 0.006x_2 - 0.106x_3$$

但发现有些变量的 t 检验并不显著,如表 1 中 $VIF_{x_1} = 3.666, VIF_{x_2} = 3.265, VIF_{x_3} = 1.236$, 它们都大于 1, 则说明解释变量之间存在相互解释的关系。因此需进一步诊断是否模型中存在多重共线性,如表 2 所示,各个条件指数都很大,且特征矩阵中存在一些值为 0 的特征值,表明变量间存在多重共线性^[4]。

表 2 共线性诊断

Dimension	Eigenvalue	Condition Index	(Constant)	x 1	x 2	x 3
1	3.999	1	0	0	0	0
2	0.001	65.981	0.05	0	0.25	0.22
3	0	100.307	0.05	0	0.03	0.7
4	0	195.911	0.39	0.99	0.71	0.08

3.2 岭回归分析

岭回归中最重要的就是 k 值,理论上,存在 $k > 0$, 使得

$$MSE(\hat{\beta}(k)) < MSE(\hat{\beta}) \quad (7)$$

即存在 $k > 0$, 这里的 $k > 0$ 成为岭参数或偏参数。岭回归估计就是要减少均方误差,则此时岭回归估计要优于最小二乘估计^[5]。由于 $k > 0$, 设置 k 的范围为 $[0, 0.5]$, 每次增加步长为 0.02。得到估计值关于 k 的岭迹图^[6]。

由图 1 可以看出,当 k 从 0 变化到 0.5 时,各个自变量的回归系数有很大的变化幅度,其中商业零售价格指数由 1.0193 变化到 0.4944, 而工业品出厂价格指数和固定资产投资指数的回归系数变化趋势较为平缓。可以看到当 $k \in [0.15, 0.5]$, 各回归系数的岭迹图趋于稳定,取 $k = 0.15$ 进行岭回归。

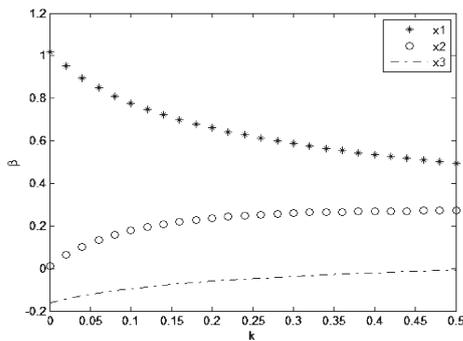


图1 岭迹图

得到回归方程为:

$$y = 0.7224x_1 + 0.2077x_2 - 0.0788x_3 \quad (8)$$

对比表1中的结果,解释变量 x_2 的 $t = 3.575 > t_{0.5}(3)$,则统计量是显著的,很好地修正了模型。

4 结束语

根据四川省2000-2010关于CPI的数据,用商品零售价格指数、工业品出厂价格指数和固定资产投资价格指数进行评估,建立了关于CPI的多元回归模型,发现选取的各要素之间存在多重共线性,通过岭回归的岭迹

图,选取合适的 k 值,对模型进行了修正,得到了合适的回归模型。

在实际经济数据中,很多要素之间都是相互影响相互制约的,因此在数据中也会存在一定的数量关系,通过岭回归的方法,使得模型得到了修正,从而能够更好地解释模型中的因变量。

参考文献:

- [1] Weisberg S. Applied Linear Regression[M]. 北京: 统计出版社, 1998.
- [2] 李生彪, 杨旭升. 基于多元回归模型的甘肃省CPI影响因素分析[J]. 甘肃科学学报, 2012(4): 152-155.
- [3] 张丹平. 基于岭回归方法的我国能源消费影响因素研究[J]. 统计与决策, 2012(21): 146-148.
- [4] 赖国毅, 陈超. SPSS17 中文版统计分析典型事例精粹[M]. 北京: 电子工业出版社, 2010.
- [5] 汪明瑾, 王静龙. 岭回归中确定K值的一种方法[J]. 应用统计概率, 2001(1): 7-13.
- [6] 程穆. 多重共线性的修正-岭回归的应用[J]. 中国商界, 2012(7): 138-139.

Application of Ridge Regression on Revising Multicollinearity

WANG Juan, LIANG Deng-xing

(College of Management Science, Chengdu University of Technology, Chengdu 610059, China)

Abstract: Based on the CPI data of Sichuan province from 2000 to 2010 and some influencing factors, the multiple linear regression model of CPI data is built by SPSS17.0. The multicollinearity in model is modified based on ridge regression, and an adjusted model is obtained. The result shows that this method has certain practicability.

Key words: ridge regression; CPI; multicollinearity