Statistics and Application 统计学与应用, 2023, 12(5), 1256-1263
Published Online October 2023 in Hans. https://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2023.125128

Hans汉斯

岭回归分析在研究城镇失业人数影响因素中的应用

宋玥璇，牟唯嫣

北京建筑大学理学院，北京

收稿日期：2023年9月12日；录用日期：2023年10月9日；发布日期：2023年10月16日

摘　要

江泽民同志曾说过：“就业是民生之本”。城镇失业问题一直是党和国家高度重视的问题之一，本文以来自国家统计年鉴2013~2022近十年的城镇失业人数为研究对象，选取了2013~2022年国民生产总值、全国财政支出、城镇居民消费水平指数为影响因素，采用最小二乘估计和岭估计方法，针对影响城镇失业人数的因素进行了研究。最终根据模型得出结论，国民生产总值和城镇居民消费水平指数对城镇失业人数有影响，并且城镇居民消费水平指数与城镇失业人数呈负相关。

关键词

最小二乘回归，岭回归，城镇失业人数，R语言

Application of Ridge Regression Analysis in the Study of Factors Affecting the Number of Unemployed in Urban Areas

Yuexuan Song, Weiyan Mu

School of Science, Beijing University of Civil Engineering and Architecture, Beijing

Received: Sep. 12th, 2023; accepted: Oct. 9th, 2023; published: Oct. 16th, 2023

Abstract

Comrade Zemin Jiang once said, “Employment is the basis of people’s livelihood”. Urban unemployment has always been one of the issues to which the Party and the State attach great importance. This paper takes the number of urban unemployed from the National Statistical Yearbook

文章引用: 宋玥璇, 牟唯嫣. 岭回归分析在研究城镇失业人数影响因素中的应用[J]. 统计学与应用, 2023, 12(5): 1256-1263. DOI: 10.12677/sa.2023.125128

2013~2022 in the past ten years as the object of study, selects the gross national product, the national financial expenditure, and the index of the consumption level of urban residents in the period of 2013~2022 as the influencing factors, and adopts the method of least squares estimation and ridge estimation, in order to research on the factors that affect the number of urban unemployed. Finally, according to the model, it is concluded that the GNP and the index of urban residents' consumption level have influence on the number of urban unemployed, and the index of urban residents' consumption level is negatively correlated with the number of urban unemployed.

Keywords

Least Squares Regression, Ridge Regression, Urban Unemployment, R Language

1. 引言

经过了 2020 年的新冠疫情之后，全球性金融经济危机爆发，中国企业也受到了影响，而随之出现的失业问题则成为当前中国面临的重大问题[1]。城镇失业人数的上升使得未就业人员面临巨大的压力，在对就业市场产生影响的同时，也是对国家稳定局面的冲击。党中央坚持以人民为中心的发展思想，为开创新的就业发展局面，需要充分发挥各项积极因素的作用，因此，完善这一方面的政府治理措施便显得尤为重要，对于提高就业质量、促进社会经济发展以及国家稳定具有十分重要的意义[2]。

国内生产总值(GDP)是指在一段时间内，一个国家或地区的经济中所生产出来的全部最终产品和劳务的价值，失业人数的上升通常反映着 GDP 的下降。政府在就业方面的投资性支出有助于缓解失业压力。通常来讲，居民消费水平指数在一定程度上也与失业情况有所关联。

本文针对城镇失业问题，基于国家统计年鉴上的数据对最小二乘估计和岭估计方法进行了分析，结果表明岭估计的回归效果较好，并得出结论，城镇失业人数主要受 GDP 和城镇居民消费水平指数影响。

2. 指标体系与数据来源

本文数据来源于国家统计年鉴 2013~2022 近十年的数据，所选变量为：城镇失业人数(万人)，国内生产总值(亿元)，国家财政支出(亿元)，城镇居民消费水平指数(以 1978 年为基期)。其中，被解释变量 Y 为城镇失业人数，解释变量 X1 为国内生产总值，X2 为国家财政支出，X3 为城镇居民消费水平指数。所选指标体系见表 1，原始数据见附录表 1。

Table 1. System of data indicators
表 1. 数据指标体系

变量代码	变量名称
Y	城镇失业人数(万人)
X1	国内生产总值(亿元)
X2	国家财政支出(亿元)
X3	城镇居民消费水平指数(以 1978 年为基期)

3. 模型的建立与研究

含有 $p-1$ 个自变量的理论线性回归模型的一般形式为：

$$Y=\beta_0 X_0+\beta_1 X_1+\cdots+\beta_{p-1}X_{p-1}+e$$

如果对自变量 $x_1,\cdots,x_{p-1}$ 和因变量 Y 进行 n 次观察，则可以得到 n 组数据，并满足等式：

$$y_i=\beta_0+\beta_1 x_{i1}+\cdots+\beta_{p-1}x_{ip-1}+e_i \quad i=1,\cdots,n$$

记

$$y=\begin{pmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{pmatrix}, X=\begin{pmatrix} 1 & x_{11} & \cdots & x_{1,p-1} \\ 1 & x_{21} & \cdots & x_{2,p-1} \\ \cdots & \cdots & & \cdots \\ 1 & x_{n1} & \cdots & x_{n,p-1} \end{pmatrix}, \beta=\begin{pmatrix} \beta_0 \\ \beta_1 \\ \cdots \\ \beta_{p-1} \end{pmatrix}, e=\begin{pmatrix} e_0 \\ e_1 \\ \cdots \\ e_n \end{pmatrix}$$

如果满足 $r(X)=p, e_i\,(i=1,\cdots,n)$ 互不相关，均值皆为零，且有公共方差 σ^2，则线性回归模型可以写为：

$$y=X\beta+e, E(e)=0, Cov(e)=\sigma^2 I$$

本文主要应用最小二乘估计和岭估计方法，运用 R 语言及 SPSS 等软件，对数据进行统计分析。由于数据本身大小对分析结果可能造成影响，为消除变量间的量纲关系，使数据具有可比性，故对数据进行标准化处理，以便于对数据进行最小二乘估计与岭估计。标准化后的数据见附录表 2。

3.1. 相关性分析

Table 2. Correlation test
表 2. 相关性检验

		Y	X1	X2	X3
Y	皮尔逊相关性	1	0.747	0.714	0.594
	Sig.		0.007	0.010	0.035
	个案数	10	10	10	10
X1	皮尔逊相关性	0.747	1	0.970	0.963
	Sig.	0.007		0.000	0.000
	个案数	10	10	10	10
X2	皮尔逊相关性	0.714	0.970	1	0.974
	Sig.	0.010	0.000		0.000
	个案数	10	10	10	10
X3	皮尔逊相关性	0.594	0.963	0.974	1
	Sig.	0.035	0.000	0.000	
	个案数	10	10	10	10

分析表 2 可知，Y 与 X1、X2、X3 的相关系数均大于 0.5，相关性较强，且显著性 p 值均小于 0.05，也验证了 Y 与 X1、X2、X3 有显著相关性。尤其国内生产总值和国家财政支出对城镇失业人数的影响更

大。

3.2. 最小二乘估计

在模型的参数估计中，最常见的一种拟合准则是经典的最小二乘法[3]。对于线性模型 $y = X\beta + e$，其中 X 为设计矩阵，最小二乘法估计即是寻找 β 估计，使 $Q(\hat{\beta}) = \|y - X\hat{\beta}\|^2$ 达到最小。

3.2.1. 用最小二乘法对回归模型进行估计

由图 1 可知，估计的多元回归模型为：$y = 2.732\times10^{-8} + 1.732x_1 + 1.541x_2 - 2.574x_3$。

分析得，判定系数 R^2 为 0.8634，调整后的判定系数 R^2 为 0.7952，说明用 LS 方法估计的回归模型效果较好。整体检验的 p 值小于 0.05，整体的显著性检验通过，即至少存在一个自变量对因变量的影响显著，而参数的 t 检验只有 X_1，X_3 显著，造成这种结果的原因可能是因为存在多重共线性，因此还需要进一步检验。

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.732e-08  1.431e-01   0.000   1.0000
X1           1.732e+00  6.608e-01   2.621   0.0396 *
X2           1.541e+00  7.791e-01   1.978   0.0954 .
X3          -2.574e+00  7.040e-01  -3.657   0.0106 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4526 on 6 degrees of freedom
Multiple R-squared:  0.8634,    Adjusted R-squared:  0.7952
F-statistic: 12.65 on 3 and 6 DF,  p-value: 0.005276
```

Figure 1. Least squares estimation results
图 1. 最小二乘估计结果

使用最小二乘估计需要进行回归模型的基本条件检验。

首先检验残差的正态性。由图 2 可知，大部分点都落在了直线附近，故满足正态性假设。此外，通过采用 Shapiro-Wilk 检验和 Kolmogorov-Smirnov 检验，得到的 p 值均大于 0.05，正态性检验通过。

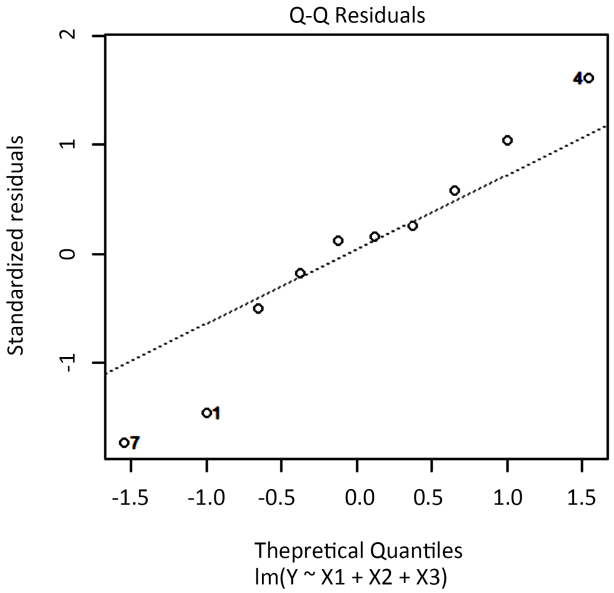

Figure 2. Q-Q plot of the residual distribution
图 2. 残差分布的 Q-Q 图

3.2.2. 多重共线性的检验

复共线性产生的原因是多方面的，一种是由于数据“收集”的局限性所致，原则上可以通过“收集”更多的数据来解决，但实现困难；另一种产生复共线性的原因是，自变量之间客观上就有近似的线性关系[4]。而最小二乘估计的性质理想与否与复共线性的存在与否息息相关，因此对数据进行复共线性的检验十分必要。本文采用的是扩大因子法(VIF)，通过计算扩大因子的值来判断是否存在多重共线性，通过计算得到三个指标的方差扩大因子均大于 10 (表 3)，故数据存在多重共线性。

Table 3. Value of variance expansion factor
表 3. 方差扩大因子的值

	VIF	Tolerance
X1	19.185737	0.052122
X2	26.674326	0.037489
X3	21.774650	0.045925

3.3. 岭估计

岭回归法(Ridge Regression)是通过放弃最小二乘法的无偏性，以损失部分信息、降低精度为代价来获得更实际和可靠性更强的回归系数。

当自变量间存在复共线性时，LS 估计的性质不够理想[4]。此时可以考虑采用岭估计，从某种意义上讲，岭估计是 LS 估计的改进[5]。岭估计的均值不等于待估参数，属于有偏估计的一种。

3.3.1. 方法原理

对于线性回归模型：$y=\alpha_0 1+X\beta+e, E(E)=0, Cov(e)=\sigma^2 I$ 。

回归系数 β 的岭估计定义为：$\hat{\beta}(k)=(X'X+kI)^{-1}X'y$ 。

与 LS 估计相比，岭估计将 $X'X$ 换成 $X'X+kI$，从而“打破”复共线性的影响，具有比 LS 更小的均方误差。

3.3.2. 用岭迹法对回归模型进行估计

对于标准化后的变量，运用 R 语言绘制岭迹图(图 3)，并得出最优岭参数为 $k=0.04$。根据选择的岭

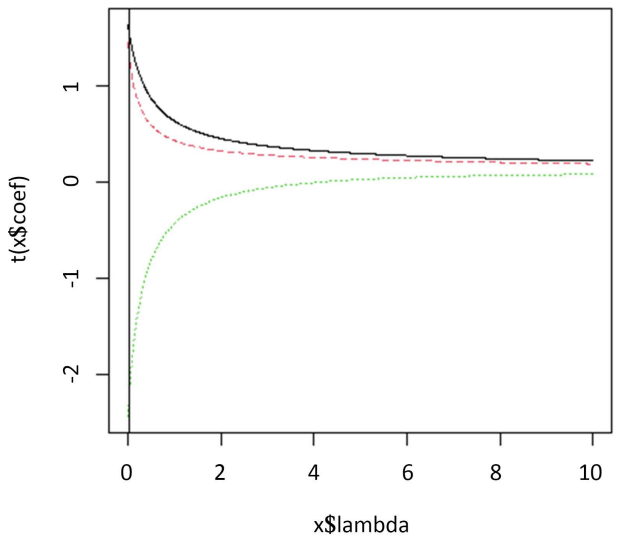

Figure 3. Mountain road map
图 3. 岭迹图

回归参数进行岭回归，得到结果见图 4。

采用岭回归估计的多元回归模型为：$y = 2.008\times10^{-8} + 1.008x_1 + 0.6934x_2 - 1.012x_3$。

```
Call:
linearRidge(formula = Y ~ X1 + X2 + X3, lambda = 0.04)


Coefficients:
             Estimate Scaled estimate Std. Error (scaled) t value (scaled)
(Intercept)  2.008e-08              NA                  NA               NA
X1           1.008e+00       3.025e+00           1.124e+00            2.693
X2           6.934e-01       2.080e+00           1.096e+00            1.897
X3          -1.012e+00      -3.036e+00           1.114e+00            2.725
            Pr(>|t|)
(Intercept)       NA
X1           0.00709 **
X2           0.05779 .
X3           0.00643 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Ridge parameter: 0.04

Degrees of freedom: model 1.85 , variance 1.351 , residual 2.349
```

Figure 4. Ridge regression for standardized data
图 4. 标准化数据的岭回归

3.4. 最小二乘估计与岭估计的比较

比较两种方法预测的回归模型，解释变量的系数没有太大的差异，正负情况也相同。其中，自变量 X1 对因变量 Y 为正向影响，自变量 X3 对 Y 的影响为负向，而 X2 对 Y 的影响不显著，即城镇失业人数增多与国内生产总值的增加成正相关，而与城镇居民消费水平指数呈负相关。

3.4.1. 相关性比较

根据两种方法构建的回归方程分别计算预测的 Y 值，最小二乘法记为 Y1，岭回归法记为 Y2，分别与原始的 Y 值进行相关性比较。

Table 4. Correlation coefficient
表 4. 相关系数

	N	Correlation
Pair1 Y&Y1	10	0.503
Pair2 Y&Y2	10	0.782

根据 SPSS 计算 Y 与 Y1、Y2 的相关系数，由表 4 可得，Y 与 Y1 的相关系数为 0.503，Y 与 Y2 的相关系数为 0.782，得出结论：岭回归模型预测的值与实际值更接近。同时比较标准化后数据的平均值，实际平均值为−0.9081242，最小二乘模型得出的预测平均值为−0.46338259，岭回归模型得出的预测平均值为−0.75986041，也提示岭回归模型的预测效果更好。

3.4.2. 均方误差比较

从理论上来讲，岭估计应该比 LS 估计有更小的均方预测误差(MSE)，计算两种方法的 MSE 结果如图 5 所示。用岭迹法计算出来的均方预测误差小于 LS 估计计算出来的均方预测误差。

```
LS Mean Squared Error:  1.6782862185827476
RR Mean Squared Error:  1.3907372234926885
```

Figure 5. Comparison of mean square error between LS estimation and ridge estimation
图 5. LS 估计与岭估计的均方误差比较

4. 结论

通过上文对失业人数影响因素的探究可得出结论，城镇失业人数受 GDP 与城镇居民消费水平指数影响，且 GDP 与城镇失业人数呈正相关，而城镇居民消费水平指数与其呈负相关。事实上，随着 GDP 的增长，失业人数也一路上涨，高经济的增长并没有创造出令人期望的足够多的就业岗位。究其原因是多方面的，如：劳动力结构的变化，科技进步的影响使得越来越多的机器替代人类，以及区域性劳动力需求的影响等等。针对于城镇失业问题，本文的研究还不够全面，后续还可以选取更多的指标进行更细致的研究。

参考文献

[1] 牟雨慧. 基于 logistic 回归的山东省失业人员再就业影响因素研究[J]. 山东纺织经济, 2021(11): 25-29.

[2] 李祎涵. 数字经济对我国劳动力就业规模与结构的影响及对策研究[D]: [硕士学位论文]. 济南: 山东财经大学, 2023.

[3] 王福昌, 曹慧荣, 朱红霞. 经典最小二乘与全最小二乘法及其参数估计[J]. 统计与决策, 2009, 25(1): 16-17.

[4] 王松桂, 史建红, 尹素菊, 吴密霞, 编著. 线性模型引论[M]. 北京: 科学出版社, 2004.

[5] 李宏, 李建武, 莫荣, 等. 基于回归分析的失业预警建模实证研究[J]. 中国软科学, 2012(5): 138-147.

附　录

Table 1. Original data
附表 1. 原始数据

Year	Y	X1	X2	X3
2013	926	592,963.2	140,212.1	878.2
2014	952	643,563.1	151,785.56	930.3
2015	966	688,858.2	175,877.77	994.3
2016	982	746,395.1	187,755.21	1050.4
2017	972	832,035.9	203,085.49	1092
2018	974	919,281.1	220,904.13	1143
2019	945	986,515.2	238,858.37	1195.9
2020	1160	1013,567	245,679.03	1140.7
2021	1040	1149,237	245,673	1254
2022	1203	1210,207.2	260,609.17	1238.2

Table 2. Standardized data
附表 2. 标准化后的数据

Year	Y	X1	X2	X3
2013	−0.9081242	−1.341271	−1.58816672	−1.673295813
2014	−0.6335751	−1.1033867	−1.31313955	−1.26496461
2015	−0.4857409	−0.8904417	−0.74062169	−0.763367738
2016	−0.3167875	−0.6199446	−0.45837086	−0.323686731
2017	−0.4223834	−0.2173233	−0.09406808	0.002351235
2018	−0.4012642	0.1928409	0.32936711	0.402061242
2019	−0.7074921	0.5089272	0.75602465	0.816662406
2020	1.5628185	0.6361053	0.91810814	0.384035105
2021	0.2956684	1.2739279	0.91796485	1.272018316
2022	2.0168806	1.5605659	1.27290215	1.148186588