Corrected Principal Component Regression and Its Application in China's Urban Employment Demand

Ying-ying Zhang[1], Jing-yi OuYang[2]

1) Department of Statistics and Actuarial Science, Chongqing University, Chongqing, China
2) Department of Statistics and Actuarial Science, Chongqing University, Chongqing, China

***Abstract*—In this paper, we use the principal component regression to study the 16 influence factors of urban employment demand. In theory, we derive the mathematical models of corrected principal component regression. Ordinary principal component regression is the dependent variable doing multiple linear regression with the first several principal components, and corrected principal component regression is the dependent variable doing multiple linear regression with several principal components. In the empirical analysis, first of all, the ordinary multivariate linear regression does not pass the variable significance test, the reason is the multicollinearity between the independent variables. Stepwise regression and principal component regression both can eliminate the multicollinearity. For the data set in this paper, among the significant regressions, from the aspect of prediction within the sample, stepwise regression is the best, the principal component regression is more and more good with the increase of the number of principal components; From the aspect of prediction outside the sample, under the criterion of average minimum of the absolute value of the relative error in the prediction, the corrected principal component regression with the 1, 3, 4, 5, 6, and 8 principal components is one of the best in all kinds of regression methods. Finally, we obtain the optimal regression equation, and give some economic explanations of the model.**

***Keywords*—Urban employment demand, corrected principal component regression, stepwise regression, multicollinearity, prediction outside the sample**

修正主成分回归在我国城镇就业需求中的应用

张应应 1　欧阳静怡 2

1)　重庆大学统计与精算学系，重庆，中国

摘　要　本文利用主成分回归研究城镇就业需求的 16 个影响因素。在理论上，我们推导了修正主成分回归的数学模型。普通主成分回归是因变量对前几个主成分做多元线性回归，而修正的主成分回归是因变量对某几个主成分做多元线性回归。在实证分析中，首先，普通的多元线性回归未通过变量的显著性检验，究其原因是自变量之间存在多重共线性。逐步回归和主成分回归都可以消除多重共线性。对本文研究的数据集来说，在显著的回归当中，从样本内预测来看，逐步回归是最好的，主成分回归随着主成分的个数的增加越来越好；从样本外预测来看，在预测相对误差的绝对值的平均值最小的准则下，取第 1、3、4、5、6、8 个主成分的修正主成分回归是各种回归方法中最好的，由此得到了最优的回归方程，并给出了模型的经济解释。

关键词　城镇就业需求，修正主成分回归，逐步回归，多重共线性，样本外预测

1. 引言

经济增长、充分就业、稳定物价、国际收支平衡是国家宏观经济政策的四大目标，国家之所以将就业问题列入宏观调控的四大目标之一，是因为就业问题是关系到人民切身利益和社会稳定的大事情。实现充分就业，不仅是实现经济可持续增长的必要条件，也是构建和谐社会的重要要求。我国经济的快速增长和劳动力就业情况主要体现为三方面：第一，人口的快速增长和年龄结构的转变为我国经济持续发展提供了丰富的劳动力市场；第二，过于丰富的劳动力资源是我国最大的竞争优势，但是同时也带来了

巨大的就业压力；第三，我国经济的高速增长并没有带来就业的实质增长，引用欧洲经济学家对我国就业现状的看法是“发达国家只需要保持 2%-3%的经济增长速度就能保持就业的整体稳定，但是在中国，需要将经济增长速度保持在 8%-9%才能够满足国内的就业。”失业问题，会引发一系列经济和社会矛盾，造成社会不稳定。所以，研究城镇就业需求成为必要，本文以城镇登记就业人数为研究对象，在城镇就业需求影响因素分析的基础上，构建城镇就业需求拟合模型，选择最优回归方程。

城镇就业需求（Y）的影响因素有：经济活动人口（X_1）、城镇居民人均可支配收入(X_2)、居民消费价格总指数(CPI)(X_3)、商品零售价格总指数(X_4)、城镇人口占总人口比重(X_5)、国内生产总值(GDP)(X_6)、第一产业产值(X_7)、第二产业产值(X_8)、第三产业产值(X_9)、进口总额(X_{10})、出口总额(X_{11})、通货膨胀率(X_{12})、货币供应量M2(X_{13})、M2 同比增长率(X_{14})、财政支出(X_{15})和全社会固定资产投资(X_{16})。

2 主成分回归

2.1. 普通主成分回归[1]

主成分分析[2][3][4]是通过线性变换，将原来的多个指标组合成相互不相关的少数几个能充分反映总体信息的指标，从而在不丢掉重要信息的前提下避开变量间共线性问题[5]，便于进一步分析。普通主成分回归的理论推导可参见[1]。

2.2. 修正主成分回归

普通主成分回归要求 $\hat{Y}$ 对 $Z_1^*,\cdots,Z_m^*$ 做多元线性回归，但是 $Z_1^*,\cdots,Z_m^*$ 中可能有变量不显著，这时应把不显著的变量删掉之后再做多元线性回归。令

$$I=\left\{i:i=1,\ldots,m,\ Z_i^*\text{ is significant}\right\}\subseteq\{1,\ldots,m\}$$

为 $Z_1^*,\cdots,Z_m^*$ 中显著变量的下标集合，也称为 $Z_1^*,\cdots,Z_m^*$ 中使用主成分的下标集合。我们想通过 $\left\{Z_i^*\right\}_{i\in I}$ 找出预测值 $\hat{Y}$ 与原变量 $X_1,X_2,\cdots,X_p$ 之间的回归方程。首先

$$\hat{Y}=\hat{\beta}_0^*+\sum_{i\in I}\hat{\beta}_i^*Z_i^*,$$

类似于普通主成分回归的推导，把 $\sum_{i=1}^{m}$ 换成 $\sum_{i\in I}$，得到

$$\hat{Y}=\hat{\beta}_0+\sum_{k=1}^{p}\hat{\beta}_kX_k,$$

其中

$$\hat{\beta}_0=\hat{\beta}_0^*-\sum_{k=1}^{p}\hat{\beta}_k\bar{x}_k,$$

$$\hat{\beta}_k=\sum_{i\in I}\frac{\hat{\beta}_i^*a_{ki}^*}{\sqrt{s_{kk}}},\quad k=1,2,\cdots,p.$$

我们编写了一个 R 函数 ComputeCoefficients()，它可以很方便地计算修正主成分回归中预测值 $\hat{Y}$ 关于原变量 $X_1,X_2,\cdots,X_p$ 的系数向量。

3. 实证分析

首先读入原始数据，原始数据共有 27 个样本，取前面的24个样本用于建立回归方程，最后3个样本用于预测[6]，可以用来检验回归模型的好坏。经初步的多元线性回归分析发现，在 16 个自变量中，并非全部自变量对因变量的影响都是显著的，究其原因，自变量间存在多重共线性。度量多重共线性严重程度的一个重要指标是矩阵 $\boldsymbol{X}^T\boldsymbol{X}$ 的条件数 $\kappa(\boldsymbol{X}^T\boldsymbol{X})$。若 $\kappa<100$，则认为多重共线性程度很小；若 $100\le\kappa\le1000$，则认为存在中等程度或较强的多重共线性；若 $\kappa>1000$，则认为存在严重的多重共线性。经计算 $\kappa=5793826$，故该数据集的自变量间存在严重的多重共线性。

逐步回归可以消除多重共线性。在 R 软件中，通过逐步回归，剔除掉部分变量后，剩下的所有变量均通过检验。可见采用逐步回归法来消除多重共线性是可以的，此时的回归方程为

$$\begin{aligned}Y=&-2.673\text{e+}03+2.451\text{e-}01X_1+1.328\text{e+}02X_5-2.087\text{e-}02X_8\\&+1.933\text{e-}02X_{11}+3.383\text{e-}02X_{13}-3.017\text{e-}02X_{16}.\end{aligned}$$

主成分回归也可以消除多重共线性。在 R 软件中利用 princomp()做主成分分析。前两个主成分的累积贡献率达到了 92.4%，一般的做法就是选取前两个主成分做主成分回归，但是 Z_2^* 不显著。由于我们的目标是选取一些主成分使得对因变量 Y 的预测效果好，所以我们让 Y 对前 1 个至前 16 个主成分做普通主成分回归。有些主成分不显著，我们就去掉相应的主成分再做修正主成分回归。随着主成分数变化的 $\hat{\sigma},R^2$ 及回归系数的显著性检验见表 1。在表 1 中我们只显示了显著的检验方法。表 1 中的 $\hat{\sigma}$ 表示残差的标准差，它越小越好；R^2 表示相关系数的平方，它越大越好。表 1 中还给出了普通的多元线性回归及逐步回归的结果。从表 1 中我们发现，逐步回归是显著的，主成分回归有 7

组是显著的。在显著的回归当中，从样本内预测来看（$\hat{\sigma}$ 的取值），逐步回归是最好的，主成分回归随着主成分的个数的增加越来越好。

表 1 随着主成分数变化的 $\hat{\sigma}, R^2$ 及回归系数的显著性检验

方法	σ	R^2	是否显著
逐步回归: 1,5,8,11,13,16	88.2	0.9998	显著
主成分回归: 1 1	1363	0.9473	显著
主成分回归: 2.1 1,3	1119	0.9661	显著
主成分回归: 3.1 1,3,4	433.4	0.9952	显著
主成分回归: 4.1 1,3,4,5	286.4	0.998	显著
主成分回归: 5.1 1,3,4,5,6	142.3	0.9995	显著
主成分回归: 6.1 1,3,4,5,6,8	113.6	0.9997	显著
主成分回归: 7.1 1,3,4,5,6,8,12	102	0.9998	显著

但是对于样本外的预测哪一个最好呢？我们知道样本外的预测误差是实际中更为关心的问题。我们取最后 3 个样本（第 25，26，27 个样本，分别对应于 2009，2010，2011 三年的数据）用于预测，预测结果及相对误差的绝对值见图 1。第 25，26，27 个样本对应的实际 Y 值为 33322，34687，35914。在图 1 中，PCR1 表示 Y 对 Z_1^* 做主成分回归，PCR2.1 表示 Y 对 Z_1^*, Z_3^* 做主成分回归，2.1 对应的主成分可参见表 1。PCR3.1 至 PCR7.1 的解释类似地可参见表 1。SR 表示 Y 对 $X_1, X_5, X_8, X_{11}, X_{13}, X_{16}$ 做逐步回归。注意到绝对误差=预测值-实际 Y 值，相对误差=绝对误差/实际 Y 值。在图 1 中，PCR4.1 对 25 号样本的预测最好，其绝对误差及相对误差最小，PCR5.1 对 26，27 号样本的预测最好。在图 1 中红色实线、蓝色虚线、绿色点线、黑色点虚线分别画的是第 25 号样本、26 号样本、27 号样本、三个样本的平均值在各种回归方法下的预测相对误差的绝对值。图 1 中的纵坐标是相对误差的绝对值，它越小越好，横坐标是各种回归方法，其中 1 表示 PCR1，2 表示 PCR2.1，…，7 表示 PCR7.1，8 表示 SR。从图 1 我们发现，由 PCR6.1 计算的相对误差的绝对值在三个样本的平均值是 0.016316971，是最小的，故从预测相对误差的绝对值的平均值来说，PCR6.1 是各种回归方法中最好的。其实由 PCR5.1 计算的相对误差的绝对值在三个样本的平均值是 0.016317806，其所得平均值与 PCR6.1 的结果非常相近。

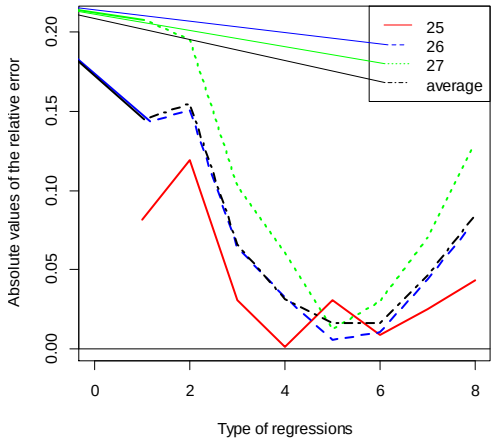

图 1 各种回归方法的预测相对误差的绝对值

PCR6.1 对应的回归方程为

$$Y = -4.605440e+03+1.916625e-01X_1+1.604503e-01X_2 -1.042150e+01X_3+9.717751e+00X_4+3.072439e+02X_5 +1.136851e-03X_6+2.926764e-03X_7-1.300085e-03X_8 +8.208277e-03X_9+3.319796e-02X_{10}+2.613773e-03X_{11} -1.254111e+01X_{12}+5.645425e-03X_{13}+2.001265e+01X_{14} -4.996409e-03X_{15}-1.057486e-02X_{16}.$$

注意，在上面的回归方程中，X, Y 均是原始变量，没有做过标准化。

这里我们解释一下为什么自变量 X_3，X_8，X_{12}，X_{15}，X_{16} 的系数是负的？

X_3 是居民消费价格总指数(CPI)。CPI 是衡量通货膨胀率的重要指标，由通货膨胀率和就业率的反向变化关系（见后面的解释）可得 CPI 与就业也呈反向变化。

X_8 是第二产业产值。第二产业资本替代劳动，制约了其吸纳劳动力的能力，目前经济结构逐渐地脱离劳动密集型产业进入资金密集型产业，增加相同资金带来的劳动就业的增长比过去减少了。另外，我国增长最快的投资基本上是资本、技术密集型的，对劳动力有挤出作用。统计分析表明，2000—2005 年全部国有及规模以上非国有企业人均固定资产从 93162 元增长到 129729 元，年均增长 7.85%；国有和国有控股企业人均固定资产从 123150 元增长到 262102 元，年均增长 22.57%。值得注意的是，国有及国有控股企业的人均固定资产净值的绝对额和增长速度要大于全部国有及规模以上非国有企业，这意味着国有及国有控股企业更加偏好采取资本密集的生产技术，并且这种偏好

还在加强。由此表明，对于相同数量的资本，如果投入到国有及国有控股企业，其所吸纳的劳动力远低于普通服务业和民营企业。因此，第二产业产值的增加反而会导致就业的减少。

X_{12} 是通货膨胀率。一般来说，通货膨胀率与就业率是呈正向变化的，但当通货膨胀较高时，货币相对贬值，购买力下降，消费者效费需求降低。此时企业有两种方法，一是保持较高价格，一是降低价格吸引顾客。而这两种方法都只会导致收入的减少，收入和支出不成比例，有些企业可能会选择降低产品质量，但不利于以后的发展，因而大部分企业选择裁员，社会就业降低。

X_{15} 是财政支出。政府支出有两种形式，一种是政府购买，一种是政府的转移支付，转移支付以提高某些群体的收入。政府转移支付增加，失业的人有了经济来源，则不愿去找工作，而且工资低的劳动者也宁愿失业，来领取高于其工资的救济金。因此政府转移支付的增加反而会造成就业的降低。

X_{16} 是全社会固定资产投资。由上述第二产业产值与就业关系的分析可知，就业的降低是由于固定资产投资的挤出效应。一般来说，随着固定资产余额的增加就业率会相应提高。但 2005 年后我国固定资产投资对就业的拉动作用明显下降。主要原因是经济结构逐渐脱离劳动密集型产业转向资金密集型产业，挤出了劳动。因此，随着我国全社会固定资产投资的增加，就业有所下降。

4. 总结

本文利用主成分回归研究城镇就业需求的 16 个影响因素。在理论上，我们推导了修正主成分回归的数学模型。普通主成分回归是因变量对前几个主成分做多元线性回归，而修正的主成分回归是因变量对某几个主成分做多元线性回归。在实证分析中，首先，普通的多元线性回归未通过变量的显著性检验，究其原因是自变量之间存在多重共线性。逐步回归和主成分回归都可以消除多重共线性。对本文研究的数据集来说，逐步回归是显著的，主成分回归有 7 组是显著的。在显著的回归当中，从样本内预测来看，逐步回归是最好的，主成分回归随着主成分的个数的增加越来越好。对于样本外预测，我们取最后 3 个样本（第 25，26，27 个样本，分别对应于 2009，2010，2011 三年的数据）用于预测，得到了在各种显著的回归方法下的预测结果、绝对误差和相对误差。在预测相对误差的绝对值的平均值最小的准则下，取第 1、3、4、5、6、8 个主成分的修正主成分回归 PCR6.1 是各种回归方法中最好的，由此得到了最优的回归方程，并给出了模型的经济解释。

参考文献(References)

[1] Y. Xue and L.P. Chen, *Statistical Modeling and R Software*. Beijing: Tsinghua University Press, 2009.

[2] H. Yang, Q.S. Liu, and B. Zhong, *Mathematical Statistics*. Beijing: Higher Education Press, 2004.

[3] X.M. Wang, *Applied Multivariate Analysis*. Third Edition. Shanghai: Shanghai University of Finance and Economics Press, 2009.

[4] R.A. Johnson and D.W. Wichern (authors), X. Lu (translator). *Applied Multivariate Statistical Analysis*. Fourth Edition. Beijing: Tsinghua University Press, 2001.

[5] S.G. Wang, M. Chen, and L.P. Chen, *Linear Statistical Model: Linear Regression and Variance Analysis*. Beijing: Higher Education Press, 1999.

[6] X.D. Xiang and F. Song, "Forecasting of the urban registered unemployment rate in Fujian province based on kernel principal component analysis and weighted support vector machine," *Systems Engineering - Theory & Practice*, no. 1, pp. 73-79, 2009.