Application of Principal Component Regression Analysis in Economic Analysis

Chen Ming-ming[1,a], Ma Jing-lian[1,b]

[1]School of Economics and Management, Chang'an University, Xi'an 710064, China.

[a]changandx2013@126.com, [b]kunming85@126.com

Keywords: regression analysis; principal component analysis; principal component regression analysis; R software

Abstract. In regression analysis, when the independent variables appear multicollinearity, the general effect of the classical regression method for least square estimates of regression coefficients will be poor, but principal component analysis can overcome this deficiency effectively. In this paper, we combined principal component analysis with classical regression analysis. Firstly the principal component analysis was used to a group of sample data based on the introduction of two statistical methods and R software which can lower the dimension of high variant space, then classical regression analysis was used to the sample data to get the quantitative relationship between the variables, and finally compared the two results in order to explain principal component regression analysis is more accurate than the classical regression analysis.

Introduction

In practice, it is often several variables need to be considered at the same time. For example, we will meet the relationship between voltage, current and resistance in the circuit; in the process of steelmaking, we will encounter the relationship between the amount of carbon steel and physical properties, such as strength, elongation, etc.; in medical science, people often need to measure the height, weight, and study the relationship between blood pressure and age, these variables are mutually restricted. There are two types of relationships between variables, one kind is a complete set of relationships between variables and it can be expressed by functional relationship. The other is that there is a certain relationship between the variables, but because of the complexity of the situation can not be accurately determined, so that the relationship between the variables can not be expressed in the form of function. In order to study this kind of relationship, it is needed to obtain the observed data through a large number of experiments or observations, and to find the relationship between them by statistical method. The method of studying the statistical law of this kind is the regression analysis. Regression analysis is widely used in practice, it can be used in classification, forecasting or control. In the regression analysis, when the independent variables appear multiple linear regression, the regression coefficient of the classical regression method for the least squares estimation of the general effect will be poor and the principal component can effectively overcome this problem. The sample data were reduced by principal component analysis, then regression analysis was conducted to find out the quantitative relationship. This paper is combined with R software, the sample data were reduced by principal component analysis, then regression analysis was conducted to find out the quantitative relationship.

Description of principal component analysis and regression analysis

Principal component analysis. Principal component analysis is a statistical analysis method for the analysis of a few comprehensive indexes [1]. It was first proposed by Pearson in 1901, and was later developed by Hotelling in 1993. Principal component analysis is a method of reducing the number of variables intoa few principal components by means of dimension reduction techniques,

which can reflect most of the information of the original variables,which is usually expressed as a linear combination of the original variables.

The main function of principal component analysis in R is pricomp(), the use format [2,3]and parameters are described as follows princomp(formula, data=NULL, subset, na.action,…)

Where formula is a formula without the response variable, data is data frame, or princomp(x,cor=FALSE,scores=TRUE,covmat=NULL,subset=rep(TRUE, nrow(as.matrix(x))),…)

Where x is used for the principal component analysis of the data, in the formof a numerical matrix or data frame; cor is a logical variable, cor=TRUE represents the sample correlation matrix was used in principal component analysis,defaults to FALSE indicates the sample covariance matrix was used inprincipal component analysis; covmat is covariance matrix.

Regression analysis. Regression analysis is a statistical analysis method to determine the quantitative relationship between two or more variables. On the basisof the number of independent variables, regression analysis can be divided intosimple regression analysis and multiple regression analysis. According to the type of relationship between the independent variables and the dependent variable, it can be divided into linear regression and nonlinear regression analysis. In regression analysis, only one independent variable and one dependent variable, and the relationship between the two can be represented bya linear approximation, which is called one-dimensional linear regression. If theregression analysis includes two or more than two independent variables, and the dependent variable and the independent variable is linear relationship, the linear regression analysis is called multiple linear regression. The main function related to linear regression analysis in R is lm(formula, data=data.frame,subset,weights,ana.ction,…)

Where formula as the model formula, data.frame is the data frame, Subset as a selectable vector, weights is a selectable vector, which is expressed in the weight of data fitting. In addition, this paper also uses the function plot(), predict(), print(), summary(), and so on, their detailed usage can be seen in thereference.

Case study

Now consider a group of 1949-1959 years of economic analysis data, total imports of Y with 3 independent variables: the total output value of X1, the storage capacity of X2, an d the total consumption of X3, we analyze the classic regression analysis and the principal component regression analysis respectively, and compare the results of the two methods.

Classical regression analysis. The sample data is input in the form of data frame, and t he regression analysis is made by using the general linear regression method. The sample d ata is stored in sampledata, the model results are stored in lm.sol, and the output results ar e returned, R procedures are as follows.

```
m.sol=lm(y~x1+x2+x3,sampledata)
>summary(lm.sol)
Call:
lm(formula = y ~ x1 + x2 + x3, data = sampledata)
Residuals:    Min          1Q      Median          3Q         Max
-0.52367   -0.38953   0.05424      0.22644   0.78313
Coefficients:     Estimate  Std.  Error      t value      Pr(>|t|)
(Intercept)       -10.12799    1.21216    -8.355    6.9e-05  ***
x1                -0.05140     0.07028    -0.731    0.488344
x2                 0.58695     0.09462     6.203    0.000444  ***
x3                 0.28685     0.10221     2.807    0.026277  *
```

Residual standard error: 0.4889 on 7 degrees of freedom

Multiple R-squared: 0.9919, Adjusted R-squared: 0.9884

F-statistic: 285.6 on 3 and 7 DF, p-value: 1.112e-07

The regression equation can be obtained by the calculation results

$$Y = -10.12779 - 0.05140X_1 + 0.58695X_2 + 0.28685X_3 \quad (1)$$

The regression equation is not reasonable.

In this problem, Y is imported, X_1 is the gross domestic product, and the corresponding regression coefficient is negative, that is, the higher the domestic total output value, the less the amount of its imports, which is inconsistent with the actual situation, the reason is that there is a total linear between three independent variables. The principal component analysis [4] of the variables is firstly carried out, then regression analysis was done.

Principal component regression analysis. Firstly, the principal component analysis was carried out.

```
>data.pr=princomp(~x1+x2+x3,sampledata,cor=TRUE)
>summary(data.pr,loadings=TRUE)
Importance of components:     Comp.1      Comp.2      Comp.3
Standard deviation          1.413915   0.9990767   0.0518737839
Proportion of Variance      0.666385   0.3327181   0.0008969632
Cumulative Proportion       0.666385   0.9991030   1.0000000000


Loadings:  Comp.1  Comp.2  Comp.3
x1         -0.706           0.707
x2                 -0.999
x3         -0.707          -0.707
```

From results, $\lambda_3 = 0.0518737839^2 \approx 0,$ so there is a collinearity between the variables. The contribution rate of the first two principal components has reached 99%, the first principal component is the total output value and the total consumption, so the first principal component is the production and marketing factor. The second main components are related to the amount of storage, called the storage factor.

Secondly, principal component regression analysis. Here, first calculate the prediction value of the principal components of the sample, and store the first and second principal components of the forecast values in the data frame. Then the principal component regression was conducted. R codes are as follows,

```
>pre=predict(data.pr);sampledata$z1=pre[,1];sampledata$z2=pre[,2]
>lm.sol=lm(y~z1+z2,sampledata);summary(lm.sol)
Call:
lm(formula = y ~ z1 + z2, data = sampledata)
Residuals:      Min          1Q      Median      3Q         Max
-0.89838    -0.26050   0.08435     0.35677   0.66863
Coefficients:  Estimate     Std. Error   t value   Pr(>|t|)
(Intercept)     21.8909      0.1658    132.006   1.21e-14 ***
z1               2.9892      0.1173    -25.486   6.02e-09 ***
z2              -0.8288      0.1660    -4.993    0.00106 **
```

Residual standard error: 0.55 on 8 degrees of freedom

Multiple R-squared: 0.9883, Adjusted R-squared: 0.9853

F-statistic: 337.2 on 2 and 8 DF, p-value: 1.888e-08

From the above results, the regression coefficient and regression equation are all tested, and the effect is remarkable, and the regression equation is obtained.

$$Y = 21.8909 + 2.9892Z_1^* - 0.8288Z_2^* \tag{2}$$

The relationship between the response variable and the principal component is obtained by the above equation, but the relationship between the response variable and the original independent variable is not expressed by the equation. To transform to find out the original regression coefficients were -9.13010782、0.07277981、0.60922012 and 0.10625939, that is the regression equation is

$$Y\text{=-9.13010782+0.07277981}X_1\text{+0.60922012}X_2\text{+0.10625939}X_3 \tag{3}$$

At this time, the coefficients of X, X_2, X_3 are positive, the regression equation is more accurate and reasonable than the classical regression analysis.

Conclusion

In order to overcome the shortcomings of classical regression analysis in dealing with the independent variables with collinearity, through the introduction of the basic ideas of classical regression analysis and principal component analysis. In this paper, we combine the two statistical analysis methods and through the R software programming, a set of sample data related to the economic analysis of the classic linear regression analysis, and then the principal component regression analysis, and finally compare the two methods to get the model. The findings show that the results obtained from the classical regression analysis are not reasonable, and the results obtained by the principal component regression analysis are more accurate and reasonable. Because of the principal component regression analysis has good properties in determining the quantitative relationship between the variables, which makes it is widely used in the problem of colinearity.

Acknowledgements

This work was financially supported by the Fundamental Research Funds for the Central Universities (CHD2009JC141), Fundamental Research Funds for the Central Universities (Y1117).

Literature References

[1] Zhang R.C., Multivariate statistical analysis, Science Press, Beijing, 2006, pp. 271-285.

[2] Xue Y., Chen L.P., Statistical modeling and R software, Tsinghua University press, Beijing, 2007, pp. 463-471.

[3] Li H.C., Zhu W.J., Shen Y.C., R Cookbook, Machinery Industry Press, Beijing, 2013, pp. 146-250.

[4] Ren X.S., Yu X.L., Multivariate statistical analysis, China Statistics Press, Beijing, 2011, pp. 316-324.