Partial Least Squares Modeling and Its Multi-collinearity Analysis

Lixia Mao

Xi'an Railway Vocational & Technical Institute, Shaanxi, Xi'an, 710026

Keywords: partial least squares; modeling; multi-collinearity; regression analysis

Abstract: In mathematics problems, multiple regression analysis often encounters multiple collinear problems, which makes the multiple correlation problems between variables become serious, but this problem is ubiquitous, and the phenomenon of multi-collinearity will affect the estimation of the parameter values, making the model's error larger, thus destroying the stability of the model. Therefore, eliminating multi-collinearity has become the most critical issue. Modeling by partial least squares regression, and verifying the theory of partial least squares, screening the original independent variables in the least squares regression model, and a model is established to solve the practical problems in real life.

Partial least squares is a diversified statistical analysis method. It is also the latest data analysis method in recent years. Its research object is the linear regression of multi-dependent variables to multiple independent variables, and modeling the variables. Special attention should be paid to the fact that when the multicollinear correlation occurs within the variable, the partial least squares modeling method can solve this problem. The partial least squares regression can solve the problem that the number of samples is less than the number of variables, so It is called regression analysis.

1. Multi-collinear processing

One way that is frequently used is to remove unimportant collinear variables, select variables by means of multiple regression analysis, or use stepwise regression to filter variables, but in theory, these screening methods are aimed at the data without collinearity. If the multicollinearity is very complicated and there is still no suitable method to deal with it. To delete part of the multicollinearity variable will increase the error of the model and damage the information of the system itself, the risk of damage will increase continuously; increasing the capacity of the sample can reduce the damage caused by the multi-collinear model to a certain extent, but it will not be feasible due to time or cost constraints. If multiple collinearity exists, you can build a multi-collinear combination, construct a new variable instead of the old variable in the regression equation, or convert the function of the equation into a differential form, which can reduce the multi-collinearity to a large extent. There are also ridge regression, principal component regression and partial least squares regression. The best way to solve multiple collinearity is partial least squares regression [1].

In the case of severe multicollinearity, the use of ridge regression, principal component regression and partial least squares regression is more effective than ordinary regression models in modeling and extracting components. Different regression models have different test methods, the principal component will find the correlation between the component and the variable in the independent variable system; and the partial least squares will find the strongest molecule related to the variable in the system. Therefore, according to the magnitude of the square root error, it can be concluded that the partial minimum regression is better than the other two regressions.

2. Partial least squares modeling

In order to effectively solve the multiple correlations problem, we used the principal component regression method to extract the required components from the independent variables. However, it can't guarantee that the variables have the strongest ability, because the principal component

  DOI: 10.25236/iccse.18.054

regression method doesn't consider the role of the variable extraction components. Therefore, the partial least squares regression can solve the problem of variable extraction components more than the principal component regression. We suppose that there are n independent variables {p1, p2, p3, pn} and m dependent variables {q1, q2, q3, qm} and study the relationship between independent variables and dependent variables, then set up R assembly points, the independent variable data table p and and the dependent variable data table q. The partial least squares regression method needs to extract the corresponding components a1 and b1 in the independent variable p and the dependent variable q. The linear combination of a1 is p1, p2, p3, pn, and the linear combination of b1 is q1, q2. Q3, qm. According to the relevant requirements of the regression analysis method, we need focus on two points when extracting components. The a1 and b1 components need to carry the information in their own data tables, and the degree of association between a1 and b1 should be maximized. As the first components a1 and b1 are extracted, the partial least squares regression analysis needs to perform the regression of p-a1 and q-b1, and then check the accuracy. If the regression equation can reach a reasonable precision value, stop the calculation; If it is not qualified[2], the residual information of p-a_1 and q-b_1 will be extracted twice, and will be recycled until the satisfactory accuracy is achieved. If at the end p extracts r components $a_1,a_2,a_3,\cdots,a_r$, partial least squares regression will be performed through q_k $(k=1,2,\cdots,x)$ to $a_1,a_2,a_3,\cdots,a_r$ and finally form the regression equation of q_k with respect to variables $x_1,x_2,x_3,\cdots,x_p$.

3. Partial least squares regression calculation

First of all, the data will be normalized and the processed data form is $E_0=(E_0,\cdots E_p)_{mxp}$,$a_1$ is a vector, extracted by E_0, $\|w_1\|=1,\|a_1\|=1$. In partial least squares regression, the problem needs to be optimized. So it is concluded that:

$$\max_{w_1-c_1} <E_0w_1,F_0a_1>$$

$$s.t\begin{Bmatrix} w_1'w_1=1 \\ a_1'a_1=1 \end{Bmatrix}$$

Regression equation of E_0 and F_0 are:

$$E_0=t_1p'_1+E_1$$

$$F_0=t_1r'_1+F_1$$

In the multiple regression analysis of partial least square method, the method of sample detection is often used to determine whether the regression model can adapt to the prediction ability. Therefore, the samples need being processed step by step. First of all, the quantity and the square sum of the regression coefficients are obtained by building regression equation. All the data would be put into the regression equation as the experimental point and corresponding value would be obtained. If the two are equal, the regression equation can have better prediction effect. Otherwise, the regression equation will not be suitable for prediction.In the modeling of partial least squares, several components need to be selected for analysis. Each new component needs being tested whether it is helpful for the prediction ability of the model. By way of sampling test, all samples are divided into two parts. First, the components are used to form a regression equation. Then, the sample points that have been excluded are substituted into the regression equation to obtain the value of sample points. Repeated tests were performed to obtain the final predicted error sum. If the error $PRESS_h$ of the regression equation is less than (h-1) components in different components, the error SS_{h-1} of the regression equation will increase one component and improve the accuracy of prediction[3].

$$PRESS_{hj}=\sum_{i=1}^{n}(y_{ij}\hat{y}_{hj(-i)})^{2}$$

If the stability of the regression equation is not ideal, there will be a large error difference, and the change of sample points will be obvious, which will also increase the value of $PRESS_h$. Partial least squares is not only a kind of algorithm but also an idea of regression. In the actual application of testing models, variables does not need being tested and you just need to extract the ingredients, which can meet the relevant accuracy requirement and forecast whether tested model is reasonable, which can be judged and measured by partial least-squares regression.

4. Analysis of partial least squares multicollinearity

The main function of partial least square method is to inhibit multicollinearity. In order to prove the research on multicollinearity of partial least square (SMLS), the extracted components of SMLS were proved. First, mathematical method is used to prove proposition, such as to prove if there is interaction between α_1 and α_2 and the existence of $\alpha'_1\alpha_2=0$ [4].

$$\alpha'_1\alpha_{2=}\alpha'_1P_1\alpha_{2=}\alpha'_1\ (P_0-\alpha'_1\theta'_1)\ \alpha_2=[\alpha'_1P_1-\alpha'_1\alpha_1\theta'_1]\alpha_2=[\alpha'_1P_1-\alpha'_1\alpha_2\frac{\alpha'P_0}{\|\alpha_1\|^2}]\alpha_2=0,$$

The interaction between α_1 and α_2 was proved. Further analysis is conducted according to the mathematical method. If $\alpha_1,\alpha_2,\alpha_3,\cdots\alpha_h$ intersect, $\alpha_1,\alpha_2,\alpha_3,\cdots\alpha_{h+1}$ can be proved to intersect, which we can verify from.

$$\alpha'_h\alpha_{h+1}=\alpha'_hP_h\alpha_{h+1=}\alpha'_h\ (P_{h-1}-\alpha_h\theta'_h)\ \alpha_{h+1}=[\alpha'_hP_{h-1}-\alpha'_h\alpha_h\frac{a'_hP_{h-1}}{\|a_h\|^2}]\alpha_{h+1}=0.$$

Therefore, relationship between $\alpha_{h-1}\alpha_h=0,\alpha_1,\alpha_2,\alpha_3,\cdots\alpha_h$ also can be proven. In the analysis of the regression problem, the regression combination of these variables was carried out, and it was concluded that the components were mutual and there was no multicollinearity problem.

5. Conclusion

This paper focused on analysis of modeling process of the partial least square method, rearranged the solutions of partial least square method, analyzed multiple collinearity problem, and carried out the study using the real cases as the evaluation standard. The research founds that the partial least squares can solve the problem of multiple collinearity and have strong applicability to the solving process of complex relations

References

[1] Yang Chunhua, Yang Ling. Partial least square modeling and analysis of multicollinearity inhibition [J]. Journal of Huaqiao University (Natural Science), 2016, 37(4):523-526.

[2] Zhang Lijie. Research on the prediction of urban water consumption based on partial least square regression [J]. Bulletin of Science and Technology, 2012, 28(2):179-181.

[3] Xu Jia. Partial Least Squares Regression [J]. Arts and Sciences Navigation, 2017(7):3-3.

[4] Xiao Xuemeng, Zhang Yingying. Comparison of three regression methods in eliminating the multicollinearity and prediction [J]. Statistics and Decision, 2015(24):75-78.