International Journal of Engineering, Science and Mathematics
Vol. 7,Issue 9, September 2018,
ISSN: 2320-0294 Impact Factor: 6.765
Journal Homepage: http://www.ijesm.co.in, Email: ijesmj@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gage as well as in Cabell's Directories of Publishing Opportunities, U.S.A

THE ANALYSIS OF THE CIVIL AVIATION PASSENGER TRAFFIC BASED ON THE IMPROVED MULTIPLE LINEAR REGRESSION MODEL

Fenghua Xu[*]
Liguo Huang[**]

Abstract

By weakening the correlation between variables,the general multivariate linear regression model is improved.Based on the improved multivariate linear regression analysis, a regression model of civil aviation passenger volume is2154 established.And the regression model of civil aviation passenger volume gives the influence of different influencing factors on civil aviation passenger volume, and forecasts the passenger volume of civil aviation.The improved general multivariate linear regression model can weaken the influence of multicollinearity regression model and can got more reasonable results for economic structural analysis.

Keywords:

Multicollinearity;
Multiple linear regression Model;
Total sum of squares;
The civil aviation passenger traffic;
Forecast.

Author correspondence:

Fenghua Xu,
College of Science, Binzhou University, Binzhou, China

1. Introduction

In the economic problems, there are many independent variables involved. It is difficult to find a set of independent variables which are not related to each other, because they have a significant influence on the dependent variables. Objectively speaking, there is a certain correlation between these influencing factorswhen the economic phenomenon involves multiple influencing factors, and sometimes there is a serious multicollinearity linearity. When there is multiple collinearity between independent variables, the regression results may have the following effects: increase the variance of the least squares estimator, the parameter estimates are unstable, sensitive to sample changes, the test reliability is reduced, and generate Discard True Error. As the variance of the parameter estimator increases, the t-test value will become smaller in the case of a significant test, which may make some significant parameter test results become inconspicuous, thus discarding the important variables. In order to solve the problem of multiple collinearity among the independent variables, statisticians have put forward some effective methods in recent decades: Ridge regression method, principal component analysis

[*]College of Science, Binzhou University, Binzhou, China
[**]College of Science, Binzhou University, Binzhou, China

method, partial least squares method [1], but these methods do not fundamentally eliminate the multiple collinearity effect between variables. In recent years, several researchers have improved the multivariate linear regression model, see the relevant research [2-6].

At present, as the rapid growth of China's income, civil aviation industry is booming.In order to make accurate assessment and forecast the civil aviation business volume, the change trend and cause of civil aviation passenger volume become the main concern of the airline. The influencing factors of China's civil aviation passenger volume can be found in the relevant research [7-9]. There is a serious multiple collinearity among the influencing factors of civil aviation passenger volume problem. Zhang Yan [10] and other civil aviation passenger volume regression models do not take into account the multiple collinearity between variables, making the economic significance of some variables not reasonably explained. In this paper, an improved multivariate linear regression model is proposed to reduce the effect of multiple collinearity on regression model by eliminating overlapping information between variables. The result of the regression model of civil aviation passenger volume is more reasonable, and the result is more accurate when the economic forecast is made.

2. The Improved Multiple Linear Regression Model

2.1. Ordinary multiple linear regression model

Multiple linear regression is an important method in multivariate statistical analysis. It is widely used in the research of society, economy, technology and many natural sciences. It is the most basic method to study the uncertain relationship (correlation) between a variable (dependent variable) and multiple factors (independent variables).

Multivariate linear regression model with random variable y and independent variable $x_1, x_2, \cdots, x_p$ is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$$

where $\beta_0, \beta_1, \mathrm{L}\ , \beta_p$ are the unknown parameters, ε is a random error, and $\varepsilon \sim N(0, \sigma^2 I_n)$.

In linear regression analysis, the task is to estimate the unknown parameters in the above equation and make statistical inferences on the regression function. The commonly used parameter estimation method is the least squares method. Under Gauss-Markov conditions, the least squares estimate is a linear unbiased estimate, and it can be shown that the least squares estimate has the smallest variance among all linear unbiased estimators. In multiple regression analysis applications, multiple dependencies often exist between variables. When there is a serious multiple correlation in the independent variable, if the least square method is still used to fit the regression model, the accuracy and reliability of the model can no longer be guaranteed. The regression coefficients are susceptible to large rounding errors and increase the sampling variability of the estimated values. In practice, when there are multiple correlations between independent variables, there are many anomalies in the regression results, which makes the inexperienced analysts very confused. In order to eliminate the multiple collinearity in the system, the principal component analysis method is often used, but the principal component extracted by principal component analysis can better generalize the information in the system of independent variables, but it often brings a lot of useless noises, which lacks the ability to interpret the dependent variables.

2.2. Improved multiple linear regression model

2.2.1.Relevant definitions and conclusions

Definition 1: The observation value of n group sample set $(x_{ik}, x_{jk})(k=1,2,\cdots,n)$ is (x_i, x_j), $\bar{x}_i$, $\bar{x}_j$ is the mean value of the corresponding variable, the simple correlation coefficient of x_i and x_j is

$$r_{ij} = \frac{Cov(x_i, x_j)}{\sqrt{D(x_i)D(x_j)}} = \frac{\sum_{k=1}^{n}(x_{ik}-\bar{x}_i)(x_{jk}-\bar{x}_j)}{\sqrt{\sum_{k=1}^{n}(x_{ik}-\bar{x}_i)^2\sum_{k=1}^{n}(x_{jk}-\bar{x}_j)^2}}.$$

If $r_{ij}=0$, it is to say that x_i and x_j are irrelevant; if $|r_{ij}|\geq 0.8$, it is considered highly relevant; if $0.3\leq|r_{ij}|<0.5$, it is considered to be of low relevance; if $0<|r_{ij}|<0.3$, it indicates that the correlation between the two variables is extremely weak and can be considered irrelevant in practical applications.

Conclusion: For the independent variable x_i and x_j, if the simple correlation coefficient of x_i and x_j is r_{ij}, then $x_j - r_{ij}\frac{\sqrt{\sum_{k=1}^{n}(x_{ik}-\bar{x}_i)^2}}{\sqrt{\sum_{k=1}^{n}(x_{jk}-\bar{x}_j)^2}}x_i$ is not related to x_i.

prove: make $x_j' = x_j - r_{ij}\frac{\sqrt{\sum_{k=1}^{n}(x_{ik}-\bar{x}_i)^2}}{\sqrt{\sum_{k=1}^{n}(x_{jk}-\bar{x}_j)^2}}x_i$, due to

$$Cov(x_i, x_j') = Cov\left(x_i, x_j - r_{ij}\frac{\sqrt{\sum_{k=1}^{n}(x_{ik}-\bar{x}_i)^2}}{\sqrt{\sum_{k=1}^{n}(x_{jk}-\bar{x}_j)^2}}x_i\right)$$

$$= Cov(x_i, x_j) - r_{ij}\frac{\sqrt{\sum_{k=1}^{n}(x_{ik}-\bar{x}_i)^2}}{\sqrt{\sum_{k=1}^{n}(x_{jk}-\bar{x}_j)^2}}Cov(x_i, x_i) = 0,$$

so $x_j - r_{ij}\frac{\sqrt{\sum_{k=1}^{n}(x_{jk}-\bar{x}_i)^2}}{\sqrt{\sum_{k=1}^{n}(x_{ik}-\bar{x}_j)^2}}x_i$ versus x_i is irrelevant.

2.2.2. Modelling

According to [11], the use of data standard processing will result in the loss of information of indicators, and the data is processed by means of mean-value method. The so-called mean is to use the mean value of data to remove the original data, that $x_{ij}' = \frac{x_{ij}}{\bar{x}_j}$. Using the mean value method to deal with the data does not cause the variable variation degree of information loss, at the same time the relevant information utilization degree of the variables and the use of standardization are the same.

The improved multiple linear regression model is as follows:

Step 1 Assumes that the data matrix is averaged, and calculating sample correlation coefficient matrix $r = (r_{ij})$, which r_{ij} represents the simple correlation coefficient of the independent variables x_i and x_j .

Step2 It find the independent variable with multiple collinearity problem, in the sample correlation coefficient matrix, there is a multiplicity of collinearity between the variables $x_1, x_2, \cdots, x_p$.

Step3 According to the correlation coefficient between the independent variables and the respective economic meaning of the choice need to remove the relevant information variables. It may be assumed that the correlation coefficient between x_1 and y is the largest, Among the independent variables $x_1, x_2, \cdots, x_p$, the correlation coefficient between x_2 and x_1 is the largest. Then put x_1 as a basic quantity, and let $x_2^{'} = x_2 - r_{12} \frac{\sqrt{\sum_{k=1}^{n}(x_{2k} - \bar{x}_2)^2}}{\sqrt{\sum_{k=1}^{n}(x_{1k} - \bar{x}_1)^2}} x_1$, then $x_2^{'}$ is treated with a mean value.

Step4 The least squares regression analysis is made by using the processed variables and the dependent variable y , and the regression model is tested statistically. If the regression coefficients can be verified by means of a significant test, and all of them can get a reasonable economic explanation, the algorithm terminates; otherwise, Step 3 is to continue the regression analysis.

3. Model Simulation

In order to study the changing trend of passenger traffic volume in China and its causes, we use passenger traffic volume as the dependent variable y (10,000 people), national income x_1 (100 million yuan), household consumption level x_2 (yuan), civil aviation route mileage x_3 (10,000 kilometers), railway the passenger traffic x_4 (10,000 people) and the number of tourists coming to China x_5 (10,000 people) are independent variables. According to the Summary of China Statistics 2017, the statistics for 2001-2016 are obtained. See Table 1 in detail.

Table 1.The data related to the civil aviation passenger traffic

Year	x_1	x_2	x_3	x_4	x_5	y
2001	109276.20	3987.00	155.40	105155.00	1122.64	7524.00
2002	120480.40	4301.00	163.77	105606.00	1343.95	8594.00
2003	136576.30	4606.00	174.95	97260.00	1140.29	8759.00
2004	161415.40	5138.00	204.94	111764.00	1693.25	12123.00
2005	185998.90	5771.00	199.85	115583.00	2025.51	13827.00
2006	219028.50	6416.00	211.35	125655.80	2221.03	15967.84
2007	270844.00	7572.00	234.30	135670.00	2610.97	18576.21
2008	321500.50	8707.00	246.18	146192.85	2432.53	19251.12
2009	348498.50	9514.00	234.51	152451.19	2193.75	23051.64
2010	411265.20	10919.00	276.51	167609.02	2612.69	26769.14
2011	484753.20	13134.00	349.06	186226.07	2711.20	29316.66

2012	539116.50	14699.00	328.01	189336.90	2719.15	31936.05
2013	590422.40	16190.00	410.60	210596.90	2629.00	35396.63
2014	644791.10	17778.00	463.72	230460.00	2636.10	39194.88
2015	686449.60	19397.00	531.72	253484.00	2598.50	43618.00
2016	740598.70	21258.00	634.81	281405.23	2813.00	48796.05

3.1.Simulation by the general multiple linear regression model

Before regression analysis, average the data. It is assumed that the variables after the mean are represented by x_1' , x_2' , x_3' , x_4' , x_5' and y' . Calculate the simple correlation coefficient between variables, get the correlation coefficient matrix, see Table 2.

Table 2. Correlation coefficient matrix

	x_1'	x_2'	x_3'	x_4'	x_5'	y'
x_1'	1	0.999	0.96	0.992	0.869	0.991
x_2'	0.999	1	0.965	0.993	0.855	0.989
x_3'	0.961	0.965	1	0.956	0.854	0.984
x_4'	0.992	0.993	0.956	1	0.866	0.985
x_5'	0.869	0.855	0.854	0.866	1	0.901
y'	0.991	0.989	0.984	0.985	0.901	1

It can be seen from the correlation coefficient matrix of Table 2 that the simple correlation coefficients of the independent variables x_1' and x_2' , x_3' , x_4' , have reached 0.9 or more, and there are serious multicollinearity between the independent variables.

3.1.1.Establish the regression model with all the independent variables

Through the SPSS software [12], the estimated values of the regression coefficients are: -1.64, -0.842, 2.237, -0.470, -0.196, 0.436.

Ordinary civil passenger traffic regression model (model 1) is

$$y' = -1.64 - 0.842x_1' + 2.237x_2' - 0.470x_3' - 0.196x_4' + 0.436x_5' \text{ .(1)}$$

The sample determination coefficient of this regression model $R^2 = 0.992$, the adjusted sample determination coefficient $R_a^2 = 0.984$, the fitting of the model is very high, and the regression equation can also pass the significance test. From the means of the mean regression coefficient, the most influential factor on the passenger traffic volume of the civil aviation is the consumption index, followed by the national income, and again the mileage of the civil aviation route. Every 1% increase in consumer index, civil aviation passenger traffic increases 2.237%, national income increases 1%, civil aviation passenger traffic decreases 0.842%. And the airline mileage per 1% increase, civil aviation passenger traffic reduced by 0.470%. From the qualitative analysis of economics, the national income and civil aviation passenger volume, airline mileage and civil aviation passenger volume should be positive correlation, negative regression coefficient can not be explained.

It can be seen that when there is multiple collinearity between variables, the regression coefficient estimates obtained by ordinary least squares estimation are very unstable, and the variance of regression coefficients accelerates with the increase of multicollinearity, which will cause regression. In the casewhen the equation is highly significant, some regression coefficients do not pass the significance test, and even the sign of the regression coefficient does not have a reasonable economic explanation.

3.1.2. Multiple linear regression model based on stepwise regression

The stepwise regression method is one of the most used variable selection methods at present, the basic idea of stepwise regression is to introduce the variable one by one, and the condition of introducing the variable is that the partial regression squared and the test are significant, and after each new variable is introduced, the selected variables are tested individually, and the insignificant variables are eliminated. This ensures that all the variables in the last obtained variable quantum set are significant. In this way, the "optimal" variable subset can be obtained after a number of steps.

Through the output of SPSS software, The optimal model of stepwise regression contains only the independent variable $x_1^{'}$, and the regression model (model 2) is

$y^{'} = 0.115 + 0.885 x_1^{'}$ (2)

The sample determination coefficient of the regression is $R^2 = 0.982$, the adjusted sample determination coefficient is $R_a^2 = 0.980$. Compared with the whole model, the fitting of the model is very good, and the regression equation can be verified by the significance. However, the model contains only one independent variable,it does not embody the influence of national income, airline mileage, railway passenger volume and inbound tourist arrivals on civil aviation passenger volume, andmakes the structural analysis of the passenger volume regression model become extremely difficult.

3.2.Simulation by theimproved multiple linear regression model

By the correlation coefficient matrix of Table 2, we can see that the national income and residents consumption level, the airline mileage, the railway passenger volume, and the number of inbound tourists have serious multiple collinearity, among which the simple correlation coefficient of national income and residents ' consumption level reaches 0.999. Therefore, the first consideration is to remove the information contained in national income from the consumption level of the residents, and the third step of modeling step in accordance with the improved multivariate linear regression model is to make $x_2^{''} = x_2^{'} - 0.8103 x_1^2$. The $x_2^{''}$ is then treated with a mean value, and the processing PostScript is $x_2^{'''}$.

A multivariate linear regression model of $y^{'}$ about independent variable $x_1^{'}$, $x_2^{'''}$, $x_3^{'}$, $x_4^{'}$, $x_5^{'}$ is established. Through SPSS software, the estimated values of regression coefficients for the improved multivariate linear regression model are: -0.623, 0 .816, 0.303, 0.096, -0.289, 0.696.

The mproved civil aviation passenger traffic regression model is

$y^{'} = -0.623 + 0.816 x_1^{'} + 0.303 x_2^{'''} + 0.096 x_3^{'} - 0.289 x_4^{'} + 0.696 x_5^{'}$.

Let $x_2^{'''}$ be reduced to a variable $x_2^{'}$, the civil passenger traffic regression model (model 3) becomes

$y^{'} = -0.623 + 1.6263 x_1^{'} + 0.303 x_2^{'} + 0.096 x_3^{'} - 0.289 x_4^{'} + 0.696 x_5^{'}$.(3)

The sample determination coefficient of this regression model $R^2 = 0.992$, the adjusted sample determination coefficient $R_a^2 = 0.984$, are consistent with the results of the general civil passenger traffic regression model, and the regression equation can also pass the significance test. From the mean regression coefficient, the most influential factor is national income, followed by the number of inbound tourists, again for the railway passenger volume. Among them, railway passenger volume increased by 1%, civil aviation passenger volume decreased by 0.289%, this also reflects, as China entered the era of high-speed rail, railway passenger volume of civil aviation passenger traffic has formed a greater impact.

3.3. Comparative analysis of forecast results

Table 3. Comparison of prediction results of three models

Year	**Actual Passenger volume**	**model 1 （Forecast value）**	**Error**	**model 2 （Forecast value）**	**Error**	**model 3 （Forecast value）**	**Error**
2014	39194.88	37954.62	3.16%	37694.53	3.83%	38152.63	2.66%
2015	43618.00	42369.66	2.86%	41753.78	4.27%	42953.06	1.52%
2016	48796.05	47553.18	2.56%	46234.24	5.25%	47986.79	1.66%
Average error			2.86%		4.45%		1.95%

From the above results, it can be seen that the regression model of passenger traffic volume and the improved model of air passenger traffic return have little difference in predicting ability, which is more accurate than the multivariate linear regression model based on stepwise regression method.

4. Conclusion

In the modeling of economic problems, there are multiple collinearity in the variables involved, which requires the improvement of the ordinary multivariate linear regression model. From the case of civil aviation passenger traffic, the regression model of ordinary civil aviation passenger volume although some regression coefficients do not have a reasonable economic explanation, they fit the historical data well, the sample decision coefficient is $R^2 = 0.992$, so long as the correlation type of the independent variable is kept unchanged in the future period, the better prediction result can be obtained. But not the economic structure of cooperation analysis. Based on the stepwise regression method, the prediction precision and economic structure analysis results are not satisfactory in the civil aviation passenger volume problem. The improved regression model of civil aviation passenger volume, both the prediction precision and the economic structure of the analysis results are very ideal.

Acknowledgments

This work is supported by NSFC 51674038, NSF of Shandong Province undergrant numbers ZR2017MA019, ZR2014AL003, Science and Technology Project of Shandong Province under grant J14li54. Also the authors would like to thank the editor and the anonymous referee for their valuable comments and suggestions, which helped us to improve this paper greatly.

References

[1] He Xiaoqun, Liu Wenqing. “Applied regression analysis,”*Beijing: China Renmin University Press*, pp. 171-200, 2011.

[2] Guo Liang. “Application of improved multiple linear regression model in production forecasting of an oilfield,”*Journal of Xidian University(Social Science Edition)*,vol. 19, pp. 71-75, 2009.

[3] Chu Bin, Fan Dongming. “Improved Multiple Linear Regression Model and Its Application,”*Surveying and Mapping Engineering*, vol. 23, pp. 63-66, 2014.

[4] Sun Caiyun, Wang Wei. “An Improved Linear Regression Model,”*Journal of North China Institute of Science and Technology*,vol. 61, pp. 80-83,2009.

[5] Xue Xiangyang. “An improved linear regression prediction model,”*Science Technology and Engineering*,vol. 10, pp. 2970-2973, 2010.

[6] Sun Caiyun, Wang Wei. “An improved linear regression model,”*Computer Applications and IT Technology*, vol. 140, pp. 49-51, 2011.

[7] Li Cuiping. “Analysis of Factors Affecting Passenger Traffic Volume in China,”*Science & Technology and Industry*, vol. 11, pp. 59-61, 2011.

[8] Ji Yuezhi, Deng Bo, Qin Xiwen.“Analysis of Passenger Traffic Volume and Related Factors in Civil Aviation,”*Mathematics in Practice and Theory*,vol. 42, pp. 175-183,2012.

[9] Peng Linan. “Correlation Analysis and Empirical Research on Factors Affecting Passenger Traffic Volume in Civil Aviation,”*China Market*, vol. 35, pp. 160-162, 2014.

[10] Zhang Yan,Miao Gang,Li Yingke.“Multivariate Linear Regression Analysis of Passenger Traffic Volume in Civil Aviation,”*Journal of Sichuan Armed Forces*,vol. 33, pp. 81-85,2012.

[11] Ye Shuangfeng. “Improvement on the Comprehensive Evaluation of Principal Component Analysis,”*Mathematical Statistics and Management*, vol. 20, pp. 52-61, 2001.

[12] Xue Wei. “Statistical Analysis and Application of SPSS,”*Beijing: China Renmin University Press*, pp. 247-272, 2010.