# Prediction Model of Multiple Linear Regression Analysis in Grain Production

## Zhuoshi Li[1,a], Xuejun Cao[2,b], Xiaoqi Ding[3,c], Hang Chen[4,d]

[1,2,3] Jilin Agricultural University, Changchun, Jilin,130018, China

[4] Institute of Scientfic and Technical Information of Jilin, Jilin,130033, China

[a]leezs643@sina.com, [b] 499360221@qq.com, [c] leezs036@foxmail.com

**Abstract.** Food production as the basis for agricultural development, forecasting and analysis of the problem of food production is necessary. In this paper, 2004-2013 years, the national food production and related indicators of food production factors, namely as dependent and independent variables studied, to determine the multicollinearity between independent variables through diagnosis; using multiple linear regression analysis prediction model and finding the predicted results with the 2014's grain output value by comparing the prediction model error is small, relatively close to the actual food value.

## Introduction

In recent years, world grain production increase year by year developing rapidly. As the world's food problems of common concern important issues, predicted food production, analysis of food development, research the food problem into the future, not only help to improve food production, the most important is the development of humanity as a whole will have a very important practical meaning, the study is significant food production. By 2015, China as a developing country of nearly 1.4 billion population, food production problem has been actively concerned about the Chinese government. Since 2004, as the population grows, natural disasters occur frequently, and the destruction of the continuous improvement of food security issues can not be ignored even more[1].

Grain analysis and research in food production in favor of the deployment of the CPC Central Committee and the State Council to make reasonable adjustments to the relevant aspects of food production, food production forecast for ensuring effective and stable food supply capacity of the national economy is more positive strategic significance, effective food supply empowerment, social security will be enhanced national capacity, people's lives will be more stable, more happiness, China as a developing country, is the development of China, the development of agriculture is the basis of China's national economy sectors, food production is very prediction necessary[2].

The main purpose of this paper is to use the characteristics and predict the effect of multiple linear regression analysis model, by predicting the results of the national grain output.In order to keep abreast of relevant factors seriously affect food production, adjust to strive for higher grain production of  agriculture.

## Multiple correlation analysis

First of all,getting the data from the People's Republic of China National Bureau of Statistics network, choosing the real value of the national grain production from the statistics related data; then, a number of factors influence the selection of national food production,  the impact on national food production needs to be ignored not many of factors, the main selection of national food production significantly affect some of the relevant factors.Finally, let urban population $x_1$, rural population $x_2$, the effective irrigation area $x_3$, the use of pesticides $x_4$, and sown area of grain crops $x_5$ and grain disaster area $x_6$, the six influencing factors, showing various indicators for clarity, grain output $y$ .It is

easy to understand and calculate, the following Table 1 for the 2004-2013 years, the real country specific historical data and influencing factors of food production indicators.

Table 1  2004-2013 National food production index and influencing factors

| | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | $x_6$ | $y$ |
|---|---|---|---|---|---|---|---|
| 2004 | 54283 | 75705 | 54478.42 | 138.60 | 101606.03 | 37106 | 46946.95 |
| 2005 | 56212 | 74544 | 55029.34 | 145.99 | 104278.38 | 38818 | 48402.19 |
| 2006 | 58288 | 73160 | 55750.50 | 153.71 | 104958.00 | 41091 | 49804.23 |
| 2007 | 60633 | 71496 | 56518.34 | 162.28 | 105638.36 | 48992 | 50160.28 |
| 2008 | 62403 | 70399 | 58471.68 | 167.23 | 106792.65 | 39990 | 52870.92 |
| 2009 | 64512 | 68938 | 59261.40 | 170.90 | 108985.75 | 47214 | 53082.08 |
| 2010 | 66978 | 67113 | 60347.70 | 175.82 | 109876.09 | 37426 | 54647.71 |
| 2011 | 69079 | 65656 | 61681.56 | 178.70 | 110573.02 | 32471 | 57120.85 |
| 2012 | 71182 | 64222 | 62490.52 | 180.61 | 111204.59 | 24960 | 58957.97 |
| 2013 | 73111 | 62961 | 63473.30 | 183.72 | 111955.56 | 31350 | 60193.84 |

Using correlation coefficient, were multicollinearity analysis to determine multicollinearity, at present, the most common method is judged by using a variance inflation factor.First Independent variable $x_j$ variance inflation factor as to $VIF_j$ .The formula is $VIF_j = (1 - R_j^2)^{-1}$ .The larger value $VIF_j$, the greater $R^2$. Degree of regression equation fits the sample data points, the stronger the relationship of all the independent variables and the dependent variable also will be more closely.

Calculation formula is

$$R^2 = \frac{\sum(\hat{y} - \overline{y})^2}{\sum(y - \overline{y})^2} = 1 - \frac{\sum(y - \overline{y})^2}{\sum(y - \overline{y})^2} \tag{1}$$

Using EViews software to obtain the variance inflation factor. The following Table 2.

Table 2  Independent variable variance inflation factor

| Independent variables | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | $x_6$ |
|---|---|---|---|---|---|---|
| VIF | 2.5222 | 33.4961 | 67.9622 | 112.0939 | 75.9648 | 4.9941 |

From the Table 2. The vast majority of $VIF$ over 10. And the use of pesticides variance inflation factor $VIF$ =112.0939>100. This shows that there is a serious argument between linear correlation, and the maximum degree of influence the use of pesticides.

**Multiple linear regression model**

According to multiple linear regression model features, analysis and forecast whether the problem of food production can take advantage of multiple linear regression model to predict. Table 1 in the 2004--2013 years of national food production and food production related factors entered into an Excel spreadsheet, drawing linear scatter plot displays the coefficient of determination and the regression equation, we can see from Figure 1, since there is a close between variables linear

relationship, in line with the conditions set up multiple regression equation, it is possible to build multiple linear regression prediction model, reasonably predictable [3].
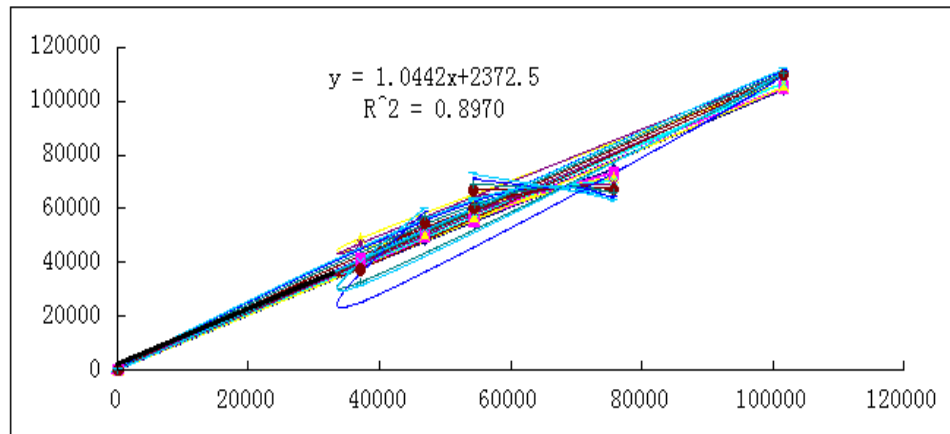


Figure 1 The linear regression equation of the food production and its related factors

From the figure 1, the independent variables and the dependent variable linear relationship exists between sex and meet specific conditions multivariate linear regression model[4], then you can prove using multiple linear regression modeling, multivariate linear regression model shows that:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_j x_j + e \qquad (2)$$

Then, the multiple linear regression model, the paper finds that the regression coefficient of multiple linear regression model, if the use of formula, not only responsible for the calculation, and prone to error, where the use of EViews software, you can use a simple operation directly to obtain multiple linear The regression equation. Obtain multiple linear regression coefficients of regression model $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6$ and the coefficient of $R^2$ [5].

$\beta_0$=189418.4 $\qquad$ $\beta_1$=0.003661 $\qquad$ $\beta_2$=-1.711254

$\beta_3$=0.003661 $\qquad$ $\beta_4$=-183.6932 $\qquad$ $\beta_5$=0.452071

$\beta_6$=-0.026848 $\qquad$ $R^2$=0.9961

Multiple linear regression model:

$$y = 189418.4 + 0.003661 x_1 - 1.711254 x_2 - 0.596015 x_3 - 183.6932 x_4 + 0.452071 x_5 - 0.026848 x_6$$

Table 3 Full-year 2014 food production and related factors in real value

| Symbol | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | $x_6$ | $y$ |
|---|---|---|---|---|---|---|---|
| The true value | 74916 | 61866 | 62466.67 | 185.37 | 112738.3 | 33198.67 | 60709.9 |

The table data substitute into multiple linear regression equation by equation (2)

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_j x_j + e$$
$$= 189418.4 + 0.003661 x_1 - 1.711254 x_2 - 0.596015 x_3 - 183.6932 x_4 + 0.452071 x_5 - 0.026848 x_6$$
$$= 62616.35$$

**Model checking**

Using the following methods for model checking [6]:

(1) Measuring regression level. It can be obtained from the data in the table for the goodness of fit

$$R^2 = \frac{\sum(\hat{y} - \bar{y})^2}{\sum(y - \bar{y})^2} = 1 - \frac{\sum(y - \bar{y})^2}{\sum(y - \bar{y})^2} = 0.998591, \quad \bar{R}^2 = 0.994362,$$

That the model fit well for sample;

(2)F-test.

The significance test whole regression equation of the significance, we can be often used F test

method.The comparison of $R^2$ substitute F statistics.

$$F = \frac{\sum (\hat{y} - \bar{y})^2 / k}{\sum (\hat{y} - \bar{y})^2 / n - k - 1} = \frac{R^2 / k}{(1 - R^2) / n - k - 1} = 236.1642 > F_\alpha$$

Therefore, the regression results are obvious, the regression equation was significant.

(3)Error estimates. Predicted value and the actual value of the difference is 19.0645 million tons, 3.14% relative error.

## Summary

According to multiple linear regression modeling method can predict the future of food production. The study found that they are able to solve multiple correlations between variables problem, with reliable performance problem solving, and the use of the problem model is more than the number of variables in the case of sample size. Many multivariate statistical regression analysis by multivariate linear regression analysis of integrated applications.

## Acknowledgements

## References

[1] Jianhua Wu,Forecast and analysis of China's grain output.Heilongjiang Grain and Oil Technology. Vol.1(2000),p.2-5.

[2] Chengjie Yin,Reflections on the improvement of the comprehensive grain production capacity. Agricultural Economy.Vol.1(2005),p.12-16.

[3] Zhenyou Wang and Lier Chen,Application of multivariate linear regression prediction model. Statistics and Decision . Vol.5(2008),p.46-47.

[4] Zhenhai Yang,Applied Mathematical Statistics,first ed.,Beijing University of Technology Press. Beijing,2005.

[5] Yanke Bao and Na Li,Matlab Mathematical Statistics and Data Processing, third ed.,Northeastern University Press,Shenyang,2008.

[6] Huixuan Gao,Several methods deal with the multiple linear regression of the independent variables collinearity.Statistics and Management.Vol.5(2000),p.20-22.