International Journal of Engineering and Management Research e-ISSN: 2250-0758 | p-ISSN: 2394-6962
Volume-10, Issue-3 (June 2020)
www.ijemr.net https://doi.org/10.31033/ijemr.10.3.10

An Empirical Study on the Change of Consumption Level of Chinese Residents

Wenxing Wang[1] and Xiangdan Wen[2]
[1]Student, Department of Mathematics, Yanbian University, Yanji 133002, Jilin Province, PR CHINA
[2]Professor, Department of Mathematics, Yanbian University, Yanji 133002, Jilin Province, PR CHINA

[2]Corresponding Author: xdwen0502@163.com

ABSTRACT

With the rapid development of Chinese economy since the reform and opening up, people's living standards have been improved, and people's consumption level has been gradually improved. Consumption plays an important role in stimulating economic growth. At present, China needs to adjust its economic structure and optimize its industrial structure. Therefore, it is very important to analyze the factors that affect the consumption level of Chinese residents and study the main factors for promoting the healthy and sustainable development of Chinese economy. Therefore, based on the statistical data from 1995 to 2018, this paper collects the variable data that affects the consumption level of residents, such as the freight volume of infrastructure railway and highway, the per capita disposable income of national residents, ordinary college students, the consumer price index of residents, the average real wage index and the gross domestic product. And through the establishment of multiple linear regression model and the stepwise regression, the paper also finds out the main factors influencing the consumption level of residents. Using R language and analyzing the results of the research, we can draw the conclusion that the national per capita disposable income, ordinary college students and consumer price index and GDP are the main factors that affect the consumption level of Chine.

Keywords-- Multiple Linear Regression, Stepwise Regression, Regression Diagnosis, Multiple Collinearity, R Language

I. INTRODUCTION

In the past decade, China's economy has been growing rapidly, with its GDP ranking second in the world, and the level of residents' consumption has been greatly improved. However, there are still some problems: People's domestic demand is insufficient, per capita GDP is still low, and the proportion of consumption in GDP [1] is still small. The effect of residents' consumption on economic growth is not obvious, especially in the vast rural areas. So how to improve the consumption level of residents has become crucial. And there are many related papers [2] devoted to study the influencing factors of residents' consumption level. These related papers are of great significance to the improvement of residents' consumption level, and provide some suggestions and countermeasures.

The consumption level of residents, simply speaking, refers to the satisfaction that residents can meet the needs of people's survival, development and enjoyment in the consumption process of material products and services. This level needs to be measured by the quantity and quality of the material products and services consumed. The consumption level of residents is calculated according to the standard of GDP, that is, the total consumption includes labor consumption. Its value is the ratio of the total consumption of residents in the GDP in the reporting period to the average population in the reporting period [3]. Therefore, this paper takes the consumption level Y of residents as the dependent variable, and establishes the multiple linear regression model with the transportation volume of infrastructurex_1, railway and highwayx_2, the per capita disposable income of residents in Chinax_3, college studentsx_4, the consumer price indexx_5, the average real wage index and the GDP$x_6$as the independent variables.

II. ESTABLISHMENT OF MULTIPLE LINEAR REGRESSION MODEL

2.1 Establishment of Multiple Linear Regression Model

Set variable Y and variable $x_1, x_2, \dots, x_p$ has the following linear relationship: $Y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_p x_p + \varepsilon$, where $\varepsilon \sim N(0, \sigma^2)$ and σ^2 are unknown parameters, $p \geq 2$, the model is called multiple linear regression model. In multiple linear regression, it is impossible to use graph to help judge whether the change is linear or not, so the significance test becomes very important. There are two kinds of significance tests: one is to test whether the regression coefficient is significant [4], in short, to test whether the regression coefficient corresponding to the variable is 0; the other is to test whether the regression equation is significant, roughly speaking, to test whether this group of sample data can use linear equation for regression prediction.

The difference between the two tests lies in the following two points: first, the original hypothesis is different from the alternative hypothesis. The regression

e-ISSN: 2250-0758 | p-ISSN: 2394-6962
Volume-10, Issue-3 (June 2020)
https://doi.org/10.31033/ijemr.10.3.10

coefficient tests whether a certain coefficient is 0, that is, the original hypothesis $h_0: \alpha_i = 0$, alternative hypothesis $h_1: \alpha_i \neq 0,\ i = 1,2,\dots,p$. The regression equation tests whether all coefficients are 0, and the original hypothesis is $h_0: \alpha_1 = \alpha_2 = \alpha_p = 0$, alternative hypothesis $h_1: \alpha_1, \alpha_2, \dots, \alpha_p$ is not all 0. The second is the difference of test statistics. The statistics of regression coefficient test are t statistics, while the statistics of regression equation test are F statistics. And

$$F = \frac{SSR/p}{SSE/(n-p-1)} \sim F(p, n-p-1)$$

Where SSR is the sum of regression squares, SSE is the sum of residual squares, and the sum of the two is the sum of total deviation squares. The square of correlation coefficient is defined as $R^2 = SSR/SST$, which can be used to evaluate the closeness of correlation between Y and independent variables.

2.2 Solution of Multiple Linear Regression Model

2.2.1 Relationship between Residents' Consumption Level and Various Indicators

In the statistical yearbook, statistical bulletin and some statistical tables[5] of China Statistics Bureau over the years, we found the required relevant data and collected the variable index data from 1995 to 2018 since the reform and opening up.

In order to make better regression analysis and comparative study, according to the collected data of various variables, this paper establishes the following multiple linear regression model: $Y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \alpha_4 x_4 + \alpha_5 x_5 + \alpha_6 x_6 + \varepsilon$, where $\alpha_i (i = 1,2,\dots,6)$is the variable index$x_i (i = 1,2,\dots,6)$the corresponding coefficient, α_0 is constant, $\varepsilon \sim N(0, \sigma^2)$ is random error. A multiple linear regression model is established with the consumption level of residents as the dependent variable and the railway freight volume of basic facilities, the per capita disposable income of the whole country, the college students, the consumer price index, the average real wage index and the GDP as the independent variables.

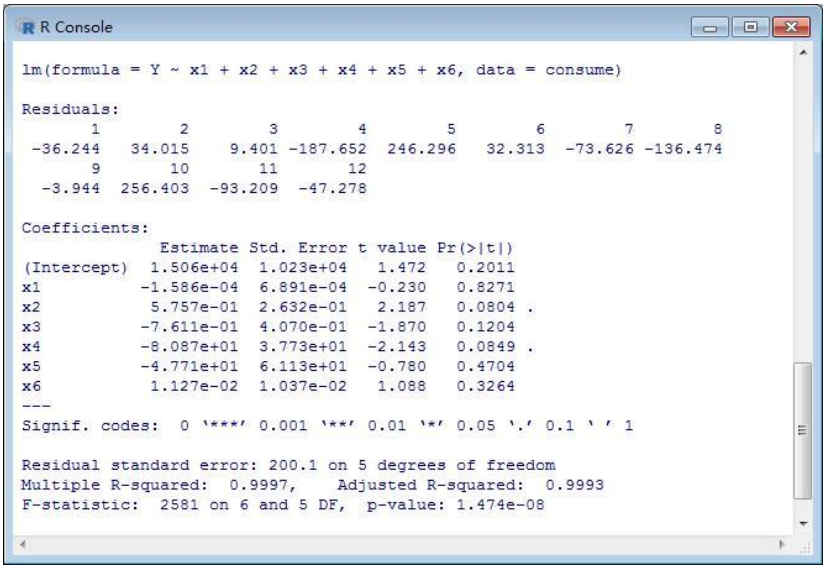

Figure 1fitting of multiple linear regression model

The results of R language show that the regression equation of the model is as follows:

$$Y = 15060 - 0.0001586x_1 + 0.5757x_2 - 0.7611x_3 - 80.87x_4 - 47.71x_5 + 0.01127x_6$$

In terms of the relationship between the variables, Y increases with the increase of x_2, x_6, which is consistent with the original conjecture of the actual situation. The level of residents' consumption increases with the increase of national per capita disposable income and GDP, while Y increases with the increase of x_1, x_3, x_4, x_5, which is consistent with the original conjecture. The consumption level of residents decreases with the increase of freight volume of infrastructure railway and highway, general college students, consumer price index and average real wage index. The coefficient before the freight volume of infrastructure railway and highway is small, so no matter what the change is, it affects the change of Y Very small. From the inspection results, there are the following points:

(1) Goodness of fit test (R^2 test): goodness of fit indicates the degree of fit between regression equation and sample data. $R^2$is used to express goodness of fit, and its

e-ISSN: 2250-0758 | p-ISSN: 2394-6962
Volume-10, Issue-3 (June 2020)
https://doi.org/10.31033/ijemr.10.3.10

maximum value is 1. The closer the value of R^2 is to 1, the higher the fitting degree of regression equation and sample data is [6]; otherwise, the closer the value of R^2 is to zero, the worse the fitting degree is. From the results of program running, we can see that $R^2 = 0.9997$, and the modified coefficient is $R^2 = 0.9993$. These two results indicate that the fitting degree of the model to the samples is very high.

(2) *P*value is the goodness of fit between the collected sample data and the original hypothesis. The larger the *P* value, the stronger the evidence to prove the original hypothesis. If we first determine a significance level α, when the*P* value is greater than α, then we should reject the original hypothesis h_0. That is to say, the variable corresponding to *P* value has no significant influence on the dependent variable. And the influence degree is: extremely significant, highly significant, general significant, not too significant, not significant. According to the results of significance test, when the significance level α is 0.05, none of the variables can pass the significance test, The p value corresponding to x_2, x_6 is a little larger than 0.05, which shows that 95% of the respondents believe that GDP has a little less significant impact on the consumption level of residents, and there are serious deficiencies. Therefore, they are not enough to affect the consumption level of residents in this model, and the model needs further optimization and improvement, so it needs gradual regression.

2.2.2 The Optimal Regression Equation of Residents' Consumption Level

In the actual production and life process, there are many factors affected by the dependent variables, so we need to select some variables to establish the regression model, which involves variable selection. If some independent variables with high degree of correlation with dependent variables are omitted in a regression equation, the established regression equation will certainly have a large deviation from the actual situation, but not too many independent variables, because it is not convenient to use, especially those independent variables with degree of correlation with dependent variables almost 0, at this time, it may reduce due to the residual square and SSE degrees of freedom Because of its small size, the estimation of σ^2 increases, so that the accuracy of regression equation for regression prediction is reduced, so it is particularly important to find an optimal regression equation [7].

In fact, there are many ways to find the optimal regression equation, such as: stepwise regression method, forward method, all subset regression method, backward method, etc. But the stepwise regression method is very important to eliminate the independent variables which are not highly correlated with the dependent variables, and the program is simple, so it is widely used. In this paper, the independent variables that affect the level of residents' consumption are the freight volume of infrastructure railway and highway, per capita disposable income of the whole country, college students, consumer price index, average real wage index, GDP, etc.

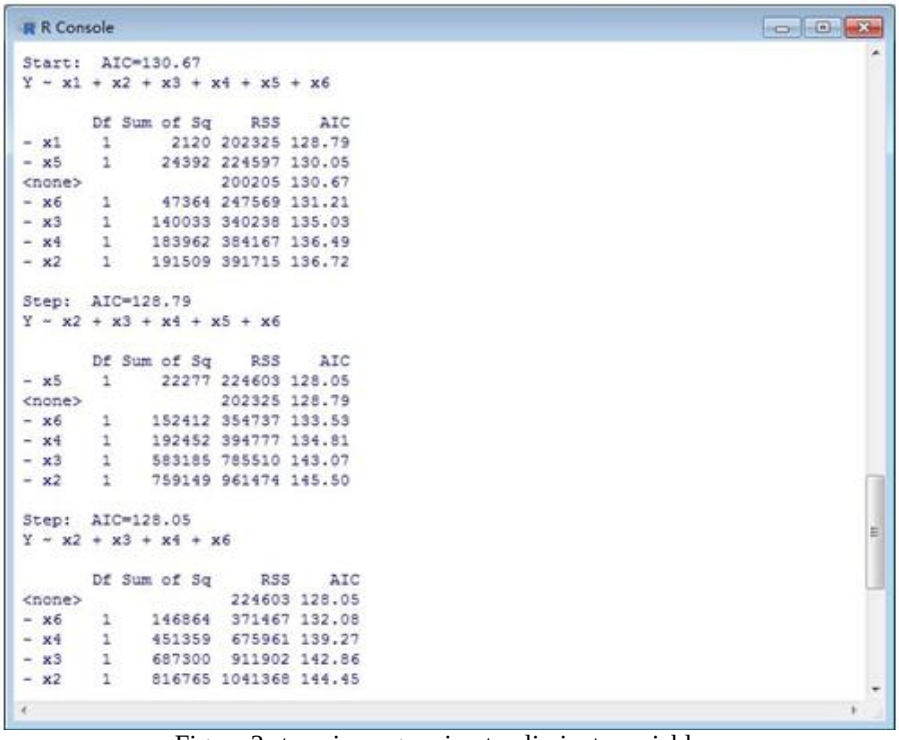

Figure 2 stepwise regression to eliminate variables

e-ISSN: 2250-0758 | p-ISSN: 2394-6962
Volume-10, Issue-3 (June 2020)
https://doi.org/10.31033/ijemr.10.3.10

It can be seen from the running results after stepwise regression with step() function that when these six variables are used as regression equation, the AIC value is one hundred and thirty point six seven 。 The data table then shows that if the variable x is removedx_1. The new AIC value[8] is obtained one hundred and twenty-eight point seven nine , next remove the variable x_5. The value of AIC is one hundred and twenty-eight point zero five In this case, the value of AIC is the lowest. In the new round of calculation, it is found that no matter which variable is removed, the AIC value will rise, so R language will terminate the calculation and get the optimal regression equation.

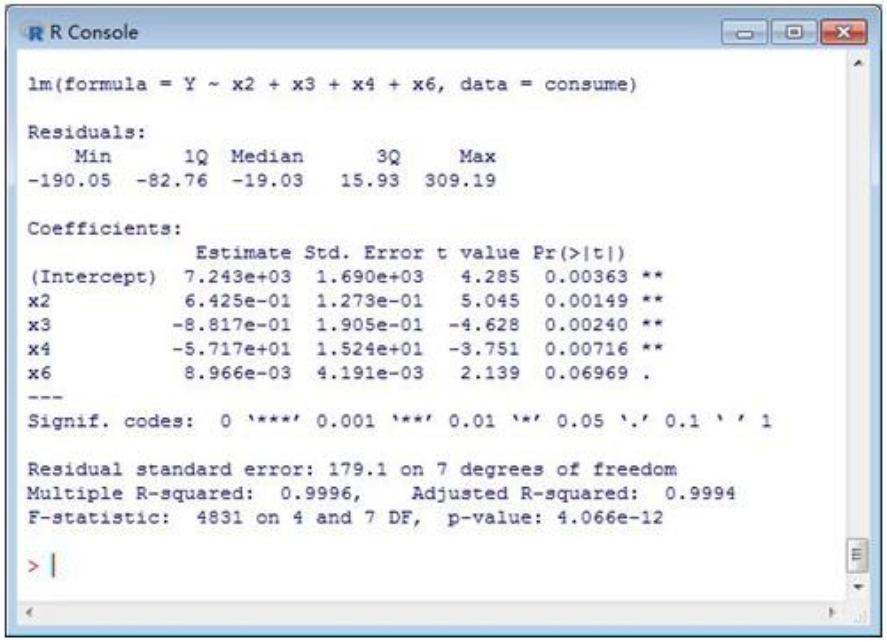

Figure 3 stepwise regression results

The data in Figure 3 shows that the significant level of the regression coefficient is significantly increased. When the significance level α is 0.05, x_2, x_3, x_4 can pass the significance test, and the corresponding P values are 0.00149,0.00240,0.00716, indicating that 95% of the respondents believe that the national per capita disposable income, general college students and consumer price index have a highly significant impact on the level of consumption, and the corresponding P value of x_6 is 0.06969 respectively, that is to say, 95% of the respondents believe that the GDP has a less significant impact on the level of consumption. Generally speaking, 95% of them think that the regression model is suitable for linear regression equation. So the optimal regression equation is:

$$Y = 7243 + 0.6425x_2 - 0.8817x_3 - 57.17x_4 + 0.008966x_6$$

The regression equation can be used to predict the multivariate linear regression equation when the coefficient of each variable is significant after the significance experience. For $(X = x_0, Y = y_0)$, the prediction interval with a confidence of $1 - \alpha$ is $[y_0 - l, y_0 + l]$, where l is the absolute difference. For example, the consumption level of residents can be predicted when $x_2 = 32000, x_3 = 3000, x_4 = 102, x_6 = 980000$.

Table 1 prediction range of residents' consumption level

fit	lwr	upr
28113.91	27509.71	28718.11

It can be seen from Table 1 after R language is run that $y_0 = 28113.91$, the corresponding confidence level of Y is zero point nine five The prediction interval of is[27509.71,28718.11].

III. MODEL TEST WITH REGRESSION DIAGNOSIS

Although the stepwise regression method has been used to select the independent variables that have significant influence on the dependent variable Y to enter

e-ISSN: 2250-0758 | p-ISSN: 2394-6962
Volume-10, Issue-3 (June 2020)
https://doi.org/10.31033/ijemr.10.3.10

the regression model, and the AIC criterion is used to select the optimal regression equation, but these are only to study the independent variables, without further analysis of some properties of the regression model, and without considering the problem of abnormal samples, the existence of abnormal samples will make the regression model unstable. The following are the main contents of regression diagnosis: (1) whether there are abnormal samples; (2) whether the linear model is appropriate; (3) whether the error term satisfies the independence, equal variance and normality; (4) whether there is a high correlation between independent variables, that is, whether there is a multiple collinearity problem; (5) whether the results of regression analysis rely heavily on some samples, That is, whether the regression model is stable.

When the least square method is used to calculate the regression model, the assumption of independence, equal variance and normality is made for the residual. However, whether the residuals of the regression model obtained from the N-group sample data of $p+1$ variables satisfy these three properties should be discussed. Before we discuss the problem of residual test, we first discuss the residual. There are three types of residuals: ordinary residuals, standardized residuals and biochemical residuals. In this paper, the residuals of the regression model obtained from 6 independent variables and 12 groups of sample data of one dependent variable of the actual index are presented.

The scatter plot with the residual ε_i as the ordinate, the fitted value y_i, or the corresponding data observation serial number i, or the data observation time as the abscissa is collectively referred to as the residual plot, which can be used to diagnose the regression model. In order to test whether the multiple linear regression model is suitable, we can use the scatter diagram of regression value and residual to test. The method is to draw the scatter diagram of regression value and common residual, or draw the scatter diagram of regression value and standard residual. There may be three different situations: normal situation, heteroscedasticity situation and nonlinear situation.

For the normal case, the residual ε_i has the same distribution regardless of the size of the regression value Y, and satisfies various assumptions of the model; for the heteroscedasticity case, it indicates that the size of the regression value is related to the volatility of the residual, that is, the assumption of equal variance is problematic; for the nonlinear case, the nonlinear model should be used. For the first case, if most of the points fall in the middle part, and only a few points fall outside, the corresponding samples of these points may have abnormal values.

3.1Model Checking by Residual Graph

Next, I take the residual graph of regression value and residual as an example for specific analysis. The results of R language operation are as follows:

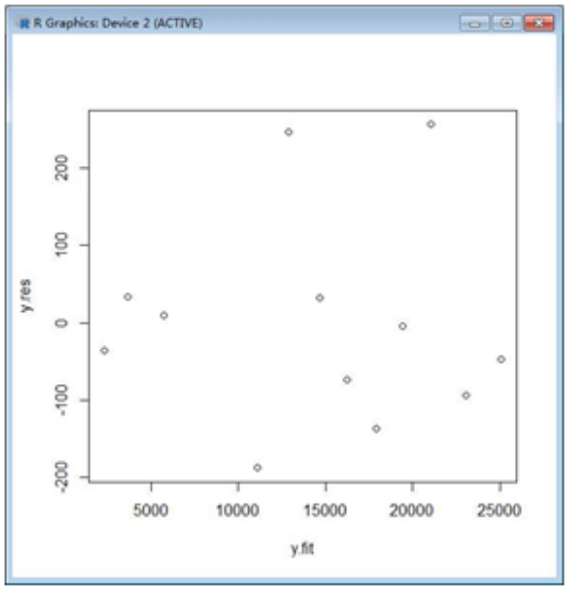

Figure 4 residual diagram

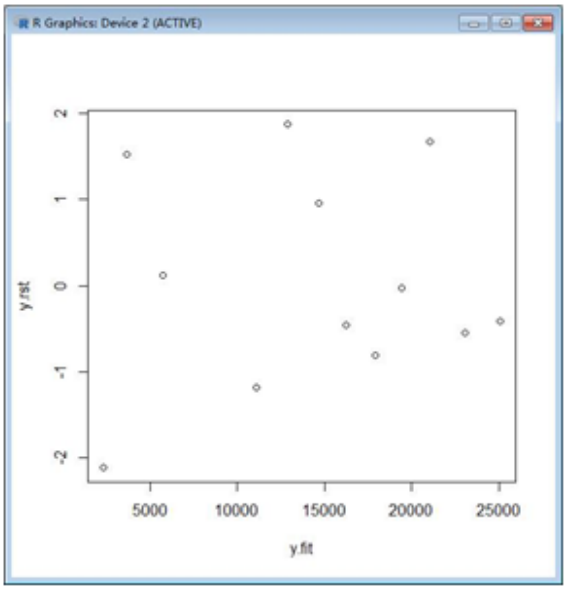

Figure 5 standardized residual diagram

It can be seen from the two figures that the residuals have the same distribution and satisfy various assumptions of the model. After careful analysis, when the assumption that the residual is normal distribution holds, the standardized residual should be approximately normal distribution. According to the nature of normal distribution, if the random variable $X \sim (0, \sigma^2)$, there is $P(\mu - 2\sigma < X < \mu + 2\sigma) = 0.954$. That is to say, for standardized residuals,

e-ISSN: 2250-0758 | p-ISSN: 2394-6962
Volume-10, Issue-3 (June 2020)
https://doi.org/10.31033/ijemr.10.3.10

95% of the sample points should fall in the interval [- 2, 2]. At the same time, the fitting value y_i and the residual ε_i are independent of each other, so they are different from the standardized residual $r_1, r_2, \ldots, r_n$ are also independent. Therefore, with the fitted value y_i as the abscissa and the standardized residual r_i as the ordinate, the point (y_i, r_i) on the plane should probably fall in the area of the horizontal band with a width of 4, and there is no trend. From this point of view, it is easier to diagnose whether the regression model has problems by standardizing the residual map. All the points in the graph are in the area of horizontal band with width of 4, and there is no trend, so the regression equation model is suitable and accurate.

3.2 Model Checking with Multicollinearity

When the independent variables are related to each other, the regression equation may be difficult to handle and understand. The effect of estimation may change the value, even the symbol, because of other independent variables in the model, so it is very important to know the relationship between independent variables when understanding the research. This tedious problem is usually called collinearity or multicollinearity [9]. If there is multicollinearity, the stability of the regression model is poor and the model is not accurate. Therefore, the next step is to test whether the regression model has multicollinearity.

If there are some constants b_0, b_1, b_2, which can make the linear equation $b_1x_1 + b_2x_2 = b_0$ hold for all data samples in the data, then the two independent variables x_1 and x_2 have exact collinearity. In actual production and life, precise collinearity is accidental. Therefore, when the equations are almost all true for the measurement data, it means that there is approximate collinearity [10]. The square of the sample correlation coefficient is a measure of the degree of collinearity between $x_1$and x_2. The exact collinearity is 1 for ${r_{12}}^2$; the non collinearity is 0 for ${r_{12}}^2$. When it is closer to 1, the approximate collinearity is stronger. Generally, if the word "approximate" is removed, when ${r_{12}}^2$ is large, x_1 and x_2 are collinear.

For p independent variables, if there are constants $b_0, b_1, b_2, \ldots, b_p$, so that $b_1x_1 + b_2x_2 + \cdots + b_px_p = b_0$ is approximately true, then these p variables have multicollinearity. The results show that the condition number $k = 235.8189$, so it is not considered that there is a large multicollinearity. Next, the eigenvalues of the matrix and the corresponding eigenvectors can be calculated to verify.

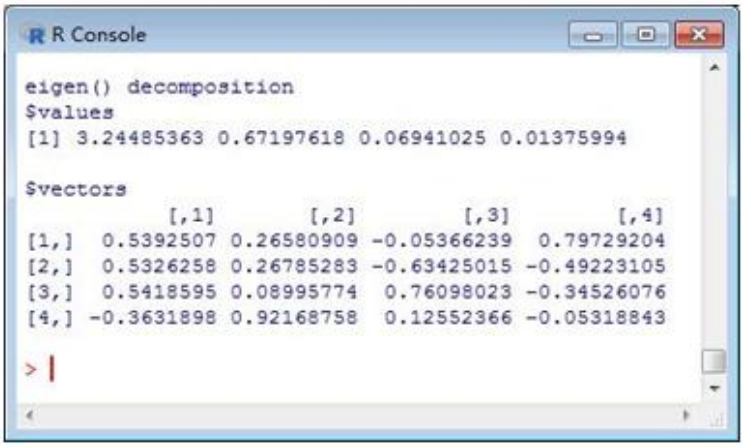

Figure 6 results of multicollinearity test

From the results of Figure 6 after running R language, it can be seen that there is no serious multicollinearity. There is no variable autocorrelation in this model, so this regression model is more appropriate and accurate.

IV. CONCLUSION

Through the above regression analysis, correlation test and regression diagnosis, this paper concludes that the main factors affecting the level of residents' consumption are the national per capita disposable income, ordinary college students, consumer price index, followed by GDP. Moreover, the consumption level of residents increases with the increase of per capita disposable income and GDP, which is in a positive proportion to the two independent variables. Although the freight volume of infrastructure railway and highway and the average real wage index are excluded, this does not mean that these two independent variables have no impact on the level of residents' consumption, but in this regression model, the impact of these two variables on the level of residents' consumption is not considered temporarily. Therefore, the multiple linear regression model established in this paper has some limitations and needs further improvement. To sum up, we get a quaternion linear regression equation without multicollinearity and heteroscedasticity, which is consistent

with the actual economic significance, so we can predict the future trend more accurately.

To sum up, one belt, one road is proposed. 1, we must persist in reform and opening up, speed up the construction of the whole belt and further promote economic development. Through the analysis of the above results, it can be found that accelerating economic growth and increasing GDP are of great significance for improving the level of consumption of residents. With the increase of GDP, the consumption level of residents is also improved; with the decrease of GDP, the consumption level of residents is also reduced. (2) We should further increase the per capita disposable income of residents. It can be seen from the above data that if the per capita disposable income is properly increased, the corresponding consumption level of residents will also be increased; if the per capita disposable income is properly reduced, the corresponding consumption level of residents will also be reduced. Therefore, it is necessary to formulate a suitable real average annual wage to improve the real average annual wage index, and then improve the per capita disposable income of residents, so as to improve the consumption level of residents, to promote economic growth and improve the index of people's happy life.

REFERENCES

[1] Lizhong Xie. (2020). The ideal level of China's urbanization in the future and the Chinese plan of Rural Governance. *Journal of Wuhan University (PHILOSOPHY AND SOCIAL SCIENCES EDITION), 73*(03), 159-168.

[2] Yang Liu. (2020). An empirical analysis on the change of consumption structure and upgrading of industrial structure of urban and rural residents. *Research on Commercial Economy*, *08*, 55-58.

[3] Qisen Jin. (2018). Suggestions on the revision of the personal income tax law in the new era. *China Township Enterprise Accounting*, *10*, 52-54.

[4] Yangyang Zou & Guoxun Jing. (2015). Experimental study on comprehensive indicators of driving fatigue based on physiological signals. *Journal of Safety and Environment*, *15*(03), 57-61.

[5] National Bureau of statistics. (2019). China statistical yearbook. Available at: http://www.stats.gov.cn/english/Statisticaldata/AnnualData.

[6] Kanghui Wang. (2020). Regression analysis of influencing factors of GDP growth rate in China. *China Management Informatization*, *23*(05), 171-175.

[7] Qianglong Zhang, Mingyi Yan, Zhongxiang Yu, & Wu Sen. (2020). Path analysis of body size and body weight of Euler sheep and construction of optimal regression model. *Qinghai Journal of Animal Husbandry and Veterinary Medicine*, *50*(01), 10-15.

[8] Wei Lin & Mei Liu. (2014). An empirical analysis of the impact of economic factors on tax growth in China: 1978-2011. *Journal of Chifeng University (NATURAL SCIENCE EDITION)*, *30*(03), 129-133.

[9] Deyan Li, Weifang Yang, Zhiyu Gao, Yuxiang Jiang, & Rongrong Li. (2020). Construction of multi factor weighted average temperature model under the influence of multiple collinearity. *Surveying and Mapping Science*, 1-9.

[10] Fang Liu & Fenyi Dong. (2020). Diagnosis and treatment of multicollinearity in Econometrics. *Journal of Zhongyuan Institute of Technology*, *31*(01), 44-48+55.