Economic Management Journal
June 2022, Volume 11 Issue 1, PP.1-8

Research on the Impact of Employment on GDP Based on Multiple Linear Regression Model

Wei Zheng, Yao Xu, Jun Yang, Shuhuan Yang

School of mathematics and computer science, Chuxiong Normal University, Chuxiong, 675000, China

†Email: zw@cxtc.edu.cn

Abstract

In order to study the impact of employed persons in various industries on regional GDP, based on the data of GDP in various regions and employed persons divided by industries in various regions in 2019, the employed persons are divided into seven categories, and the multiple linear regression model of GDP in various regions of China on employed persons in various industries is established by using the methods of multiple linear regression analysis and cluster analysis, It also analyzes the impact of employees in various industries on the GDP of various regions.

Keywords: GDP; Employees in Various Industries; Multiple Linear Regression

1 INTRODUCTION

In the early 1980s, China began to study the gross domestic product (GDP) indicators of the United Nations system of national accounts. In 1985, China began to establish a GDP accounting system. In 1993, the national income accounting was officially abolished, and GDP became the core index of national economic accounting[1].

The employment index reflects the actual utilization of all labor resources in a certain period of time. It is an important index to study China's basic national conditions and national strength. The reason why employed people are widely used as employment indicators in macroeconomic analysis is that they are closely related to the economic cycle and can reflect the changes of macroeconomic status and trend[2].When the economic operation is in good condition and the market demand is relatively strong, enterprises will increase employment and expand production scale, resulting in an increase in employment; When enterprises feel the decrease of market demand, they will reduce production and employment, resulting in the reduction or less increase of employment. Because of this close relationship between employment and macroeconomic trends, both government departments and economic analysts regard it as one of the most important indicators to analyze macroeconomic operation. In order to better analyze China's economic situation, a multiple linear regression model is established based on the relevant data of employees in various industries and regional GDP, so as to predict, control and analyze the regional GDP; At the same time, combined with the method of systematic clustering, this paper analyzes the situation of employees in various regions.

2 PRELIMINARY DATA ANALYSIS

2.1 Data Standardization

We study the impact of employment on GDP, so let GDP be the dependent variable y and manufacturing employment x_1, The number of employed persons in the construction industry is x_2, Wholesale and retail employees x_3, The number of employees in transportation, warehousing and postal industry is x_4, The number of employees in accommodation and catering industry is x_5, The number of employees in leasing and business services is x_6, The number of employed persons in residential service, repair and other service industries is x_7。Because the units between dependent variables and independent variables are different, the data we selected are standardized.

2.2 Correlation Analysis

The data of employment in various industries and GDP in various regions in 2019 are selected. In order to study the relationship between these eight variables, the standardized data are analyzed to preliminarily determine the degree and direction of correlation between these variables. The results shown in Table 1 below are obtained by using SPSS statistical software.

TABLE 1 CORRELATION COEFFICIENT

		GDP	manufacturing	construction	Wholesale and retail	Transportation, storage and postal services	Accommodation and catering	Leasing and business services	Residential services, repair and other services
GDP	**Pearson correlation**	1	.872**	.867**	.907**	.857**	.870**	.871**	.797**
	Significance (two tailed)		.000	.000	.000	.000	.000	.000	.000
manufacturing	**Pearson correlation**	.872**	1	.829**	.773**	.797**	.699**	.740**	.716**
	Significance (two tailed)	.000		.000	.000	.000	.000	.000	.000
Wholesale and retail	**Pearson correlation**	.867**	.829**	1	.767**	.840**	.707**	.797**	.792**
	Significance (two tailed)	.000	.000		.000	.000	.000	.000	.000
Wholesale and retail	**Pearson correlation**	.907**	.773**	.767**	1	.810**	.924**	.839**	.898**
	Significance (two tailed)	.000	.000	.000		.000	.000	.000	.000
Transportation, storage and postal services	**Pearson correlation**	.857**	.797**	.840**	.810**	1	.755**	.756**	.787**
	Significance (two tailed)	.000	.000	.000	.000		.000	.000	.000
Accommodation and catering	**Pearson correlation**	.870**	.699**	.707**	.924**	.755**	1	.654**	.892**
	Significance (two tailed)	.000	.000	.000	.000	.000		.000	.000
Leasing and business services	**Pearson correlation**	.871**	.740**	.797**	.839**	.756**	.654**	1	.656**
	Significance (two tailed)	.000	.000	.000	.000	.000	.000		.000
Residential services, repair and other services	**Pearson correlation**	.797**	.716**	.792**	.898**	.787**	.892**	.656**	1
	Significance (two tailed)	.000	.000	.000	.000	.000	.000	.000	
**. At the 0.01 level (two tailed), the correlation was significant.									

It can be found from table 1 that the correlation coefficients between regional GDP and employees in eight industries are large, and only the absolute values of the correlation coefficients between regional GDP and employees in residential service, repair and other service industries are between $0.5 \leq |r| < 0.8$, which is a significant correlation, and the absolute values of the other correlation coefficients are between $0.8 \leq |r| < 1$, which is a high correlation. However, by observing Table 1, we can also find that the correlation between two independent variables is also high, so we can preliminarily judge that there may be multicollinearity in this group of data.

At the same time, it is also found that the correlation coefficients between these variables are positive, indicating that they are positive correlation, that is, when one variable increases, another variable will also increase, but the degree of increase needs to be judged by the strength of its correlation.

In order to further study the correlation between these variables and predict and estimate the regional GDP, the following is a regression analysis of this group of data, establish a regression model, test and correct it.

3 PRELIMINARILY ESTABLISH MULTIPLE REGRESSION MODEL

3.1 Model Establishment

For datasets$\{(y_i, x_{i1}, x_{i2}, \cdots, x_{ip}): i = 1, \cdots, n\}$, if the dependent variable y_i and independent variable $x_{i1}, x_{i2}, \cdots, x_{ip}$ The relationship between IP addresses is as follows:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

Then the above formula is called the explained variable y about the explanatory variable $x_1, x_2, \cdots, x_p$, where β_0 is called the regression constant, $\beta_1, \beta_2, \cdots, \beta_p$ is called the regression coefficient[3].

In this set of data, the dependent variable y is the regional GDP and the independent variable x_1、x_2、x_3、x_4、x_5、x_6 and x_7 They are respectively employed in manufacturing industry, construction industry, wholesale and retail industry, transportation, warehousing and postal industry, accommodation and catering industry, leasing and business service industry, resident service, repair and other service industries. Therefore, a multiple linear regression model can be established:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_7 x_{i7}$$

Namely:

$$\begin{cases} y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_7 x_{17} \\ y_2 = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_7 x_{27} \\ \vdots \\ y_{31} = \beta_0 + \beta_1 x_{31\,1} + \beta_2 x_{31\,2} + \cdots + \beta_7 x_{31\,7} \end{cases}$$

The parameters in the above model can be estimated according to the least square principle. Through SPSS statistical software, we get the parameter estimation results shown in Table 2 below.

TABLE 2 PARAMETER ESTIMATES

Model		Non standardized coefficient		Standardization coefficient
		B	Standard error	Beta
1	**(constant)**	-3.512E-16	.036	
	Zscore(manufacturing)	.256	.075	.256
	Zscore(construction)	.148	.110	.148
	Zscore(Wholesale and retail)	-.181	.226	-.181
	Zscore: Transportation, storage and postal services	.076	.079	.076
	Zscore(Accommodation and catering)	.663	.127	.663
	Zscore(Leasing and business services)	.385	.125	.385
	Zscore(Residential services, repair and other services)	-.245	.126	-.245

The estimated values of the parameters can be known from table 2,Namely $\beta_0 = -3.512 \times 10^{-16}$, $\beta_1 = 0.256$, $\beta_2 = 0.148$, $\beta_3 = -0.181$, $\beta_4 = 0.076$, $\beta_5 = 0.663$, $\beta_6 = 0.385$, $\beta_7 = -0.245$. Therefore, the following multiple linear regression model can be obtained:

$$y_i = -3.512 \times 10^{-16} + 0.256x_{i1} + 0.148x_{i2} - 0.181x_{i3} + 0.076x_{i4} + 0.663x_{i5} + 0.385x_{i6} - 0.245x_{i7}$$

4 MODEL TEST

Statistical test of regression model generally includes significance test of regression equation, significance test of

regression coefficient, heteroscedasticity test, multicollinearity test, etc[3]. Next, we will test the model respectively.

4.1 Significance Test

The above multiple linear regression model has been established. After obtaining the linear regression model, we also need to test the fitting effect between the obtained linear regression model and the actual observation data, and investigate the explanatory variable x as a whole $x_1, x_2, \cdots, x_p$ have a significant effect on the explained variable y [4].

Therefore, the significance test of the regression equation, i.e. F test, is carried out first, and the following assumptions are made:

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_7 = 0, \ H_1: \beta_i \neq 0, \exists i \in \{1, 2, \cdots, 7\}$$

The results of F-test in SPSS are shown in Table 3 below:

TABLE 3 ANALYSIS OF VARIANCE

Model		Sum of squares	freedom	freedom	F	Significance
1	**regression**	29.069	7	4.153	102.538	.000b
	residual	.931	23	.040		
	total	30.000	30			

From table 3, we can know that the F value is 102.538，At the significance level of $\alpha = 0.05$，the significance of F test of the model is $0 < \alpha = 0.05$，Therefore, if the model of the regression equation is significant, the original hypothesis should be rejected H_0，That is, at the significance level $\alpha = 0.05$，Dependent variable y and independent variable There is a significant linear relationship between x_1，x_2，$\cdots$，x_7.

After the significance test of the regression equation, we can know the dependent variable y and the independent variable x_1，x_2，$\cdots$，x_7 There is a significant linear relationship between them, but it is also necessary to test the significance of each independent variable of the model, Conduct t-test.

Make the following assumptions:

$$H_0: \beta_j = 0, H_1: \beta_j \neq 0, j = 1, 2, \cdots, 7$$

The results in Table 4 are obtained by t-test in SPSS:

TABLE 4 T-TEST

Model		t	Significance
1	**(constant)**	.000	1.000
	Zscore(manufacturing)	3.423	.002
	Zscore(construction)	1.339	.194
	Zscore(Wholesale and retail)	-.798	.433
	Zscore: Transportation, storage and postal services	.965	.345
	Zscore(Accommodation and catering)	5.208	.000
	Zscore(Leasing and business services)	3.068	.005
	Zscore(Residential services, repair and other services)	-1.944	.064

It can be seen from table 4 that in the t-test Under the significance level of $\alpha = 0.05$, The significance of x_2、x_3、x_4、x_7 is not significant，It means that the original hypothesis should be accepted, and the other independent variables are significant, the original hypothesis should be rejected, think β_j is not 0，That is, these independent variables have a significant impact on the dependent variables.

4.2 Multicollinearity Test

It is also necessary to test the multicollinearity between independent variables. If there is no multicollinearity between independent variables, it shows that the model is more feasible; If there is multicollinearity between independent variables, it is necessary to eliminate multicollinearity[5-8] .

The diagnosis results of multicollinearity of independent variables are shown in Table 5 below:

TABLE 5 MULTICOLLINEARITY DIAGNOSIS

Model		Collinearity statistics	
		tolerance	VIF
1	(constant)		
	Zscore(manufacturing)	.241	4.145
	Zscore(construction)	.111	9.010
	Zscore(Wholesale and retail)	.026	37.882
	Zscore: Transportation, storage and postal services	.218	4.579
	Zscore(Accommodation and catering)	.083	11.996
	Zscore(Leasing and business services)	.086	11.637
	Zscore(Residential services, repair and other services	.085	11.721

According to table 5, it is found that only the Vif values of the three independent variables of manufacturing employment, construction employment and transportation, warehousing and postal employment are less than 10, and the Vif values of the remaining wholesale and retail employment, accommodation and catering employment, leasing and business service employment, and resident service, repair and other service employment are greater than 10, It shows that there is multicollinearity among these variables, so it is necessary to eliminate the multicollinearity among these variables.

4.3 Heteroscedasticity Test

Now the residual diagram method is used to test whether the model has heteroscedasticity. The residual diagram analysis method is a more intuitive and simple analysis method. The results of heteroscedasticity diagnosis are shown in Figure 1 below.

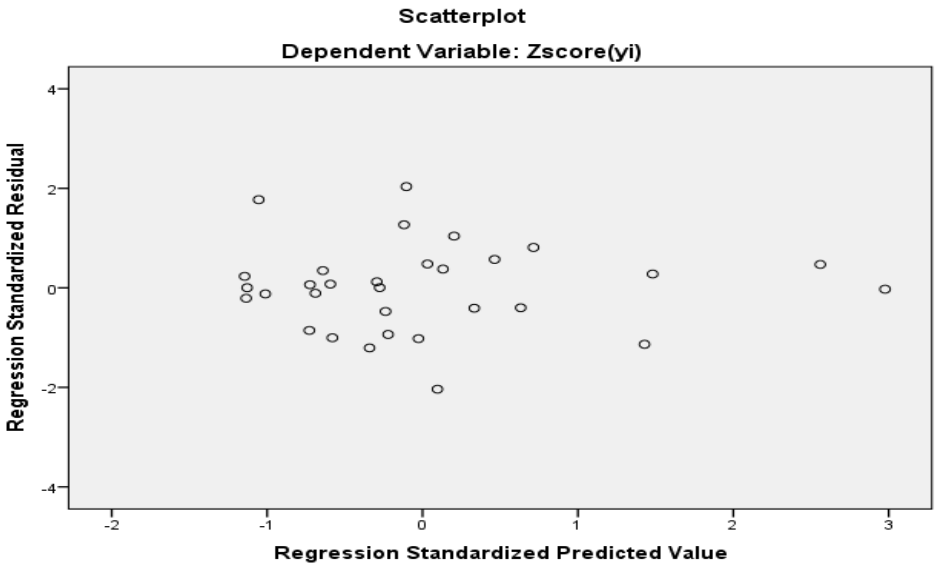

FIG. 1 SCATTER DIAGRAM

As shown in Figure 1, the 31 data points on the standardized residual diagram are scattered and distributed in the rectangular coordinate system. Observe the range of these values, and 95% of the data points fall in the value range of [- 2,2] and have no rules to follow, so there is no Heteroscedasticity in this group of data.

To sum up, in the significance test, multicollinearity test and heteroscedasticity test, it is found that the established multivariate linear regression model does not have heteroscedasticity, but in the t test, there are four independent variables that are not significant and multicollinearity, so we need to modify the model.

5 MODEL CORRECTION

5.1 Model Modification by Stepwise Regression Method

According to the information provided in Table 6, there is multicollinearity between independent variables, and the significance t test fails. Therefore, using the idea of stepwise regression, the above problems and multiple linear regression model are modified and tested.

TABLE 6 VARIABLE INPUT OR REMOVAL

Model	Input variables	Removed variables	method
1	Zscore(manufacturing)	.	Step (condition: probability of F to be input < =. 050, probability of F to be removed > =. 100).
2	Zscore(construction)	.	Step (condition: probability of F to be input < =. 050
3	Zscore(Wholesale and retail)	.	Step (condition: probability of F to be input < =. 050
4	Zscore: Transportation, storage and postal services	.	Step (condition: probability of F to be input < =. 050
5	Zscore(Accommodation and catering)	.	Step (condition: probability of F to be input < =. 050
6	Zscore(Leasing and business services)	.	Step (condition: probability of F to be input < =. 050
7	Zscore(Residential services, repair and other services)		Step (condition: probability of F to be input < =. 050
a. Dependent variable: zscore (Gross Regional Product)			

5.2 Fitting of the Introduced Model Under Stepwise Regression and F and T tests

With the gradual entry of variables, table 7 shows the fitting of the seven models formed in turn with the input of variables.

TABLE 7 MODEL FITTING

Model	R	R square	Adjusted R square	Error of standard estimation
1	.907a	.824	.817	.42724573
2	.947b	.896	.889	.33333424
3	.956c	.914	.905	.30841961
4	.964d	.930	.919	.28397716
5	.977e	.955	.945	.23351389
6	.984f	.968	.960	.20095429
7	.983g	.967	.960	.19920978

It can be found that the adjusted R-square of the seven models is increasing in turn. The fitting effects of the seven models are good, and the adjusted R-square of the seventh model is 0.983, which is better than other models.

Through the F-test under stepwise regression, the analysis results of variance of 7 models formed successively with the input of variables are obtained. It can be found that the significance is 0.000, which passes the F test.

The t-test under stepwise regression shows that the significance of the five independent variables input in the final model 7 obtained by stepwise regression is less than 0.05. Therefore, the t-significance test of model 7 passed.

5.3 Multicollinearity Yest Under Stepwise Regression

After multiple collinearity test, model 7 under final stepwise regression is obtained, Vif < 10. At this time, there is no multicollinearity. Then the final model is revised to model 7, and the final five variables are manufacturing; construction Residential services, repair and other services Accommodation and catering Leasing and business services.

5.4 Establishment of Final Multiple Linear Regression Model

Based on the above analysis, the F test and t test are passed under the idea of stepwise regression, and there is no multicollinearity and heteroscedasticity. The final independent variables are: Wholesale and retail, manufacturing, construction, resident service and repair and other services, accommodation and catering, leasing and business services. Independent variable manufacturing; construction Residential services, repair and other services Accommodation and catering Leasing and business services are x_1, x_2, x_4, x_5, x_6, The dependent variable regional GDP is y。

According to the coefficient value given by model 7, the better fitting multiple regression model is:

$$y_i = -3.442 \times 10^{-16} + 0.254x_{i1} + 0.223x_{i2} - 0.299x_{i4} + 0.399x_{i5} + 0.310x_{i6}$$

Economic significance of the model: when other explanatory variables remain unchanged, the average regional GDP will increase by 25.4 million yuan for every 10000 people in manufacturing x_1); For every 10000 people in the construction industry (x_2), the average regional GDP will increase by 22.3 million yuan; For every 10000 people in residential service, repair and other service industries (x_4), the average regional GDP will decrease by 29.9 million yuan; For every 10000 people in the accommodation and catering industry (x_5), the regional GDP will increase by 39.9 million yuan on average; For every 10000 people in the leasing and business service industries x_6) respectively, the regional GDP will increase by an average of 31 million yuan.

6 CONCLUSION

To sum up, we first established a multiple linear regression model $y_i = -3.512 \times 10^{-16} + 0.256x_{i1} + 0.148x_{i2} - 0.181x_{i3} + 0.076x_{i4} + 0.663x_{i5} + 0.385x_{i6} - 0.245x_{i7}$, After testing, it is found that there is multicollinearity, and it is modified to obtain the modified model$y_i = -3.442 \times 10^{-16} + 0.254x_{i1} + 0.223x_{i2} - 0.299x_{i6} + 0.599x_{i4} + 0.31x_{i5}$。Assuming that the other variables remain unchanged, when the manufacturing employment increases by one unit on average, the regional GDP increases by 0.254 units on average. At the same time, we also found that the number of employees in the five industries of wholesale and retail, manufacturing, construction, residential service and repair and other services, accommodation and catering, leasing and business services have a significant impact on the regional GDP. In addition, we also conducted cluster analysis and found that the number of employed persons and regional GDP in coastal areas are generally higher than those in other inland areas.

ACKNOWLEDGMENT

Thank you for your valuable comments and suggestions. This study has obtained the 2020 Yunnan College Students' innovation and entrepreneurship training program (No.: 113912017); Chuxiong Normal University is supported by the school level general scientific research project (No.: XJYB2001).

REFERENCES

[1] YONG H Y, BAO G L. An Application of Combination ARMA Model in Trend Forecasting of GDP from 2006 to 2010 in Inner Mongolia[J]. Mathematics in Practice and Theory, 2008, 38 (21): 19－23.

[2] TIAN Z C, LIU M. Empirical analysis of GDP prediction based on improved GM (1,1) model [J]. Statistics & Decision, 2018, 34 (11): 83－85.

[3] Tang Niansheng, Li Huiqiong. Application of regression analysis [M] Beijing: Science Press, 2014:42, 10, 50,131-133, 78.

[4] He Xiaoqun Multivariate statistical analysis [M] Beijing: China Renmin University Press, 2019:62-69.

[5] LI H, TIAN Z C, LI S B, et al. The evaluation of the GDP in YiLi prefecture countries based on time series analysis[J]. Mathematics in Practice and Theory, 2017, 47 (13): 15-23.

[6] WANG S S, CHEN A, SU J, et al. Application of the combination prediction model in forecasting the GDP of China [J]. Journal of Shandong University (Natural Science), 2009, 44 (2): 56－59.

[7] Li Xiang, Zhu Yuchun Grey correlation analysis of rural residents' income and consumption structure [J] Statistical research, 2013 (1): 76-78.

[8] Song Baolin, Zhou Guofu, Zhang Chunhong, Chen Hanbin. An empirical study on the relationship between fiscal revenue, population agglomeration and regional economic growth [J] Statistics and decision making, 2020, 36 (03): 100-103.