Research on the Models for Forecast of Tax Revenue of Wenzhou City

Hanqi Xie*

School of Business, Nankai University, Tianjin, 300110, China

*Corresponding author: 2013306@mail.nankai.edu.cn

Abstract. Tax revenue is a vital economic indicator that reflects the level of economic development, and tax revenue forecasting plays an important role in financial budgeting. Previous studies have demonstrated various influencing factors of tax revenue and proposed many feasible ways to forecast tax revenue. However, it is acknowledged that tax revenue of different region might bear different relation to influencing factors. In this research, tax revenue forecast of Wenzhou City is studied based on multiple linear regression model and MLP neural network model. The data for this research are collected from the website of Bureau of Statistics of Wenzhou, compiled in the 2022 statistical yearbook of Wenzhou. With multiple linear regression model, it is discovered that Value-added of the primary industry, Value-added of the tertiary industry, Investment in fixed assets, Total retail sales of consumer goods are significant for prediction. Comparing the forecasting outcomes of the two methods, the MLP neural network model appears to have better goodness of fit.

Keywords: Tax revenue; forecast; linear regression; MLP neural network.

1. Introduction

The analysis and forecasting of the financial and economic situation have become an important basis for the conception and formulation of fiscal policies [1]. Fiscal revenue is an important economic indicator that reflects the status of development of a country's economy, and is the financial guarantee and prerequisite for the realization of the salary payment and normal operation of government organizations and public institutions and the development of economic and social undertakings [2]. Tax forecasting is the first step in budgeting not only for Chinese government, but also for the governments of other countries around the world [3]. This paper is aimed at using appropriate models to forecast the tax revenue of Wenzhou, a city of Zhejiang Province of China, and comparing the forecasting effects of different models.

Regression models are often used for tax revenue forecast. Huang designed a set of local public finance budget revenue forecasting methods for Changzhou City based on a combined forecasting model, which is composed of two forecasting models: multiple linear regression model and quadratic moving average model [1]. Yang and Wen used a multiple linear regression model to study the possible factors that are related to the fiscal revenue of Hebei Province, for example, the gross regional product (GDP), the land purchase cost of real estate development enterprises, and the assets of industrial enterprises above designated size [2]. Liu used a variety of models such as linear regression, neural networks and time series to forecast various types of tax revenues, and compared the prediction effects of different models [3].

Xu mainly selected five variables, gross domestic product, tax level, and the three industries as explanatory variables, and used linear regression model to analyze their influence on the financial revenue of China [4]. Bai selected the added value of agriculture, the added value of the secondary industry (including industry and construction), the added value of the tertiary industry, the number of employees, and other income levels as explanatory variables, and used multiple linear regression model to analyzed their impact on financial revenue [5].

Starting from the analysis of the scale of fiscal deficit in Gansu Province, Wang and Guo scientifically forecast the future fiscal revenue of Gansu Province, using the moving average model, simple linear regression model and exponential smoothing model [6]. Li selected the relevant data of Gansu Province to explore the key factors affecting the fiscal revenue of Gansu Province. After

variable screening, this paper made a fitting estimate of the relation between main factors of economy and fiscal revenue according to semi-parametric regression theory [7].

The screening of factors is an important part of building regression models. Grey correlation analysis was usually introduced to solve the problem of predictor screening [8, 9]. In several papers, stepwise regression method was taken to screen factors [2, 4]. Bai used Ridge Regression method to eliminate multicollinearity and screen factors [5]. Based on lasso, adaptive lasso and SCAD methods, Li screened 39 economic factors affecting the fiscal revenue of Gansu Province respectively, and drew a comparison between the screening outcomes with respect to the three models [7].

Besides regression models, error correction model, grey model and neural network model are also used in relevant researches. With the fifteen years of data of thirty provinces that starts from 1994, Sun and Tong analyzed the long-term stable relationship between tax revenue and several predictors, applying error correction models [10]. Gu et al. proposed an improved BP algorithm combining conjugate gradient and adaptive variable step length, and then established the model for tax revenue prediction, using the improved neural network model [11]. Li proposed a tax forecasting research method based on grey algorithm model and neural network, and compared the results of these two models [12].

In summary, considering the explanatory variables are multiple, stepwise regression method is used to build multiple linear regression model. MLP neural network is also used to forecast the tax revenue of Wenzhou, and the forecasting effects of these two models are compared to find out which model is more suitable under the circumstance of this research.

2. Methods

2.1. Data Sources

The data of this paper are collected from the website of Bureau of Statistics of Wenzhou, compiled in the 2022 statistical yearbook of Wenzhou.

2.2. Variable Selection

Variables and logograms thereof are shown in the Table 1. The dependent variable is tax revenue. The independent variables are chosen in reference to the research of Zhang [9]. The data are time series from 2002 to 2021, and the unit is ten thousand Yuan. The data set is divided into two parts, namely train set from 2002 to 2018 and test set from 2019 to 2021.

Table 1. Logograms of the 7 variables

Variables	Logograms
Tax revenue	Y
Value-added of the primary industry	X_1
Value-added of the secondary industry	X_2
Value-added of the tertiary industry	X_3
RMB household deposits balance	X_4
Investment in fixed assets	X_5
Total retail sales of consumer goods	X_6
Total value of imports and exports	X_7

2.3. Model Principle

2.3.1 Multiple linear regression model

In terms of simple linear regression, a main influential predictor is introduced to establish a regression equation so that the change of the dependent variable could be explained. In the study of practical problems, the dependent variable is often influenced by more than one predictor. Therefore, it is necessary to use two or more influencing predictors. This more complicated method is named

multiple regression. Furthermore, if a linear relation exists between multiple independent variables and dependent variables, the regression analysis becomes multiple linear regression. (1) represents the general form of a multiple linear regression model.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon. \quad (1)$$

In the formula (1), X_k represents the k^{th} independent variable; β_0 is a constant term; β_k is partial regression coefficients; ε is random errors.

To screen the variables, stepwise regression method is used. The process is that the independent variables are introduced into the regression equation respectively as long as the partial regression coefficients are significant after being tested. In the meanwhile, after a new independent variable is introduced, the old independent variables whose partial regression coefficients are not significant ought to be taken out of the equation. When no more new variables can be introduced and no more old variables can be eliminated, this process finally ends. The basic purpose of stepwise regression method lies in building an optimal multiple linear regression model.

2.3.2 MLP neural network model

Neural network is a kind of nonlinear model. It is aimed at imitating the working mechanism human brain and the interaction between the objects of the real world and biological nervous system. It is more suitable for the analysis and processing of some problems with very complicated information, unclear knowledge background and unclear reasoning rules.

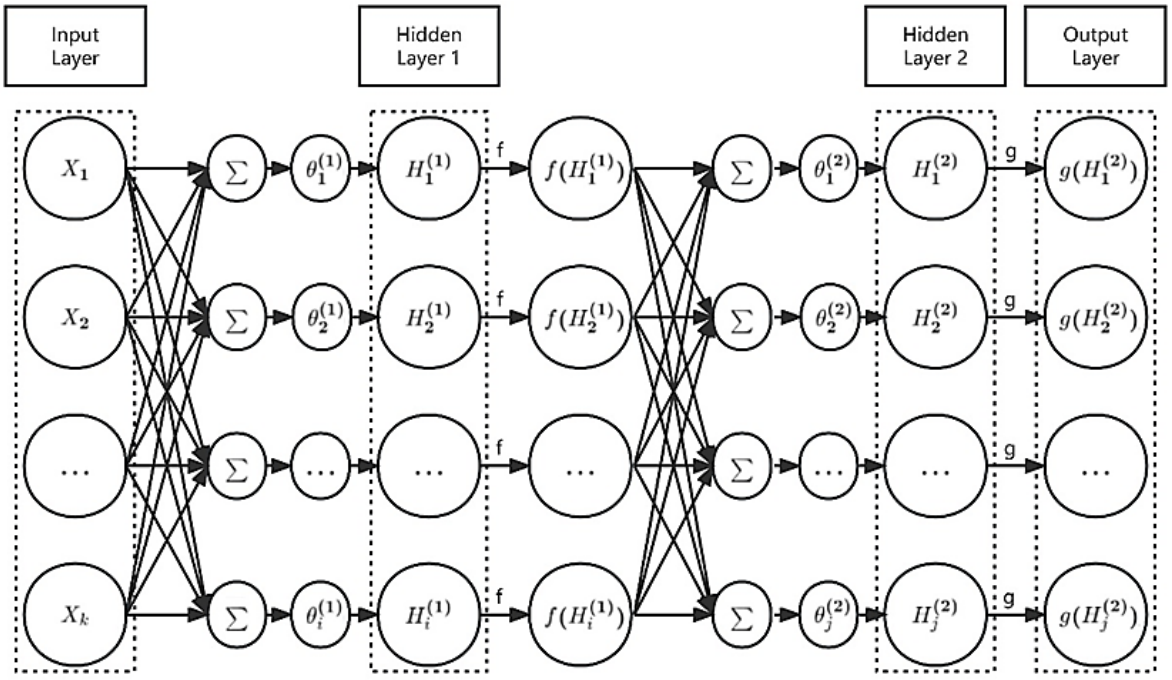

Fig. 1 MLP neural network structure

Fig. 1 manifests MLP neural network's basic structure. In Fig. 1, X_k represents the k^{th} independent variable; $H_i^{(l)}$ simulates the i^{th} neuron of the l^{th} layer, and $H_i^{(l)} = \sum_{n=1}^{k} W_{n,i}^{(l)} X_n - \theta_i^{(l)}$, in which $W_{n,i}^{(l)}$ represents the weight of X_n and $\theta_i^{(l)}$ represents the threshold of $H_i^{(l)}$; f and g are activation functions which are nonlinear differentiable non-decreasing S-shaped function whose range is limited, such as sigmoid function and tanh function. Loss function is defined to calculate the error between the output of model and the actual value. $W_{n,i}^{(l)}$ and $\theta_i^{(l)}$ are adjusted to minimize the loss function by means of gradient descent method. Gradient descent method can automatically determine the direction in which the loss function decrease fastest, but how long a stride taken in that direction should be, namely the learning rate, is determined manually.

2.3.3 Model evaluation criteria

Two indicators, called MAE and MAPE, are introduced to appraise the effect of forecasting. MAE is the abbreviation for Mean Absolute Error, and in the same way, MAPE for Mean Absolute Percentage Error. Both of them have a range of [0, +∞). The smaller MAE and MAPE are, the better the forecasting effect is. The definition of MAE is shown in the formula (2), and that of MAPE in the formula (3).

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i|. \tag{2}$$

$$\mathrm{MAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right|. \tag{3}$$

3. Results and Discussion

3.1. Descriptive Statistics

Table 2 contains the minimum, first quantile, median, mean, third quantile and maximum of all the variables.

Table 2. Variable descriptive statistics

	Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7
Min.	1159664	544913	5688237	4290376	7459660	2960230	4910886	345359
1^{st}Qu.	2425238	670866	11325372	8767804	15976016	6199150	8820546	1165929
Median	3244380	1071731	17817894	16822038	34796136	18253241	18599838	1995820
Mean	3185647	1032175	17487652	18980159	38227935	23151667	19526016	6220812
3^{rd}Qu.	3878599	1342978	22599460	28025480	52562075	39739243	30208160	12264041
Max.	5494709	1643076	31913198	42293949	92144259	58221234	38076696	24111937

3.2. Multiple Linear Regression

3.2.1 Correlation analysis

SPSS is used to conduct correlation analysis, and the result is shown in the Table 3. According to the Table 3, the selected independent variables are highly linearly correlated with Y, so it is appropriate to use Y and independent variables as multiple linear regression. Therefore, the model can be defined as follow:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \varepsilon. \tag{4}$$

Table 3. Variable correlation

	Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7
Y	1	.890**0	.920**0	.911**0	.924**0	.855**0	.894**0	.841**0
X_1	.890**0	1	.989**0	.988**0	.976**0	.979**0	.994**0	.867**0
X_2	.920**0	.989**0	1	.984**0	.974**0	.968**0	.989**0	.850**0
X_3	.911**0	.988**0	.984**0	1	.994**0	.989**0	.989**0	.921**0
X_4	.924**0	.976**0	.974**0	.994**0	1	.972**0	.971**0	.936**0
X_5	.855**0	.979**0	.968**0	.989**0	.972**0	1	.986**0	.900**0
X_6	.894**0	.994**0	.989**0	.989**0	.971**0	.986**0	1	.865**0
X_7	.841**0	.867**0	.850**0	.921**0	.936**0	.900**0	.865**0	1

3.2.2 Stepwise regression

Backward elimination is used to screen the independent variables. In the first model, all the independent variables are included, and after partial F test, X_2 has the largest p-value of 0.474. It is larger than the 0.05 significance level. This indicates that X_2 is not a significant independent variable. In the second model, X_2 is eliminated, and at this time, X_7 has the largest p-value of 0.249, being an insignificant independent variable. In the third model, X_2 and X_7 are eliminated, and X_4 has the largest p-value of 0.274. In the last model, when X_2, X_4 and X_7 are eliminated, each of the rest variables has a p-value smaller than significance level. Eventually, after variable screening, the regression equation is as follow:

$$\hat{y} = 3043170.352 - 7.129X_1 + 0.330X_3 - 0.186X_5 + 0.283X_6. \quad (5)$$

Multiple R^2 is 0.9461. Adjusted R^2 is 0.9282. These indicators mean that the regression equation is significant. In ANOVA, F-statistic is 52.71 and p-value is 0.000, also manifesting the significance of the regression equation and that X_1, X_3, X_5 and X_6 are highly linearly correlated with Y.

According to the result of Breusch-Godfrey test that the p-value is 0.04922. It is near the 0.05 significance level and larger than significance level of 0.01, it should not be rejected that the residuals are uncorrelated. Fig. 2 is the Q-Q plot of the residuals, which is nearly a straight line. The plot shows that the distribution of residuals is approximately normal distribution. Fig. 3 is the plot of residuals, judging from which the variances of residuals are basically homogeneous. Besides, the MAE is 182456.4, and the MAPE is 7.49095%.

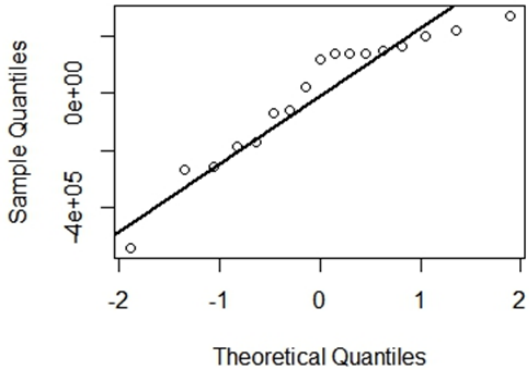

Fig. 2 Q-Q plot of residuals

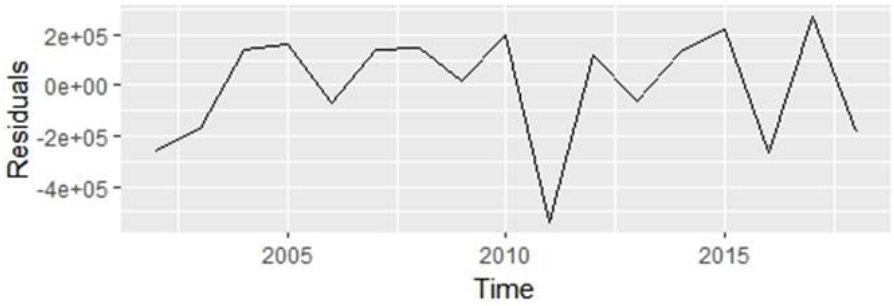

Fig. 3 Residuals plot

However, it must be mentioned that the coefficients of X_1 and X_5 are smaller than zero, which is obviously contradictory to the common sense. The reason might be multicollinearity of these variables. If the purpose of building multiple linear regression model consists in figuring out the impact of predictors on tax revenue, certain methods like ridge regression should be taken to eliminate multicollinearity. In this paper, the purpose is only to forecast the value of tax revenue, so multicollinearity is not a concern.

3.2.3 Forecasting effect

After using the multiple linear regression model to forecast the tax revenue from 2019 to 2021, the MAE is 405064, and the MAPE is 8.01446%, manifesting that forecasting effect of the model is not bad.

From 2002 to 2018, Fig. 4 shows the actual value of tax revenue. From 2019 to 2021, Fig. 4 shows the point forecast (the thinner curve), 80% confidence intervals, 95% confidence intervals and actual value (the thicker curve).

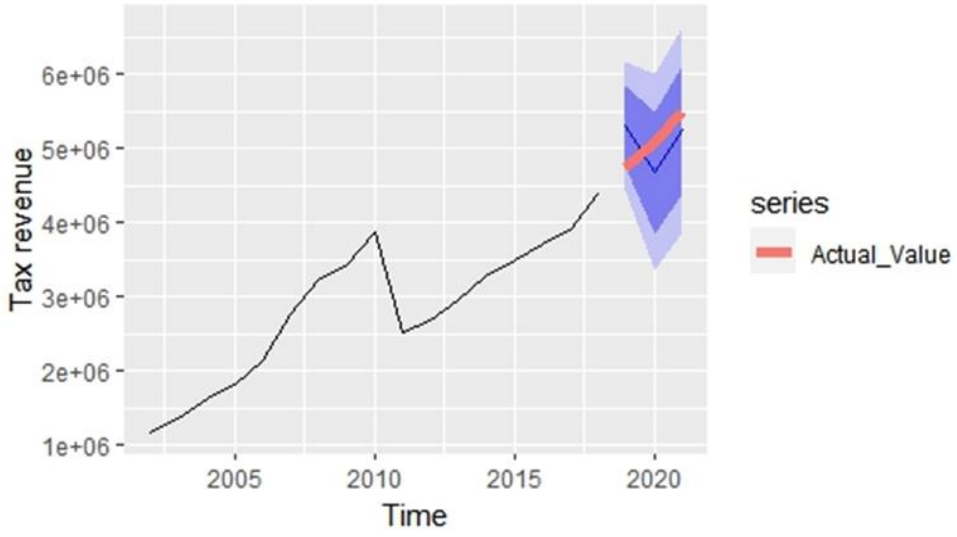

Fig. 4 Forecasts from multiple linear regression model

3.3. MLP Neural Network

SPSS is used to build the MLP neural network model. There are two hidden layers in the model, with 25 neurons in each layer. Learning rate is set at 0.4, and the termination rule is that the loss function has not decreased for last one hundred thousand steps.

In train set, the MAE is 2552, and the MAPE is 0.13916%, in contrast to the MAE of 182456.4 and MAPE of 7.49095% by regression model. As for test set, the MAE is 23419.33, and the MAPE is 0.44767%, in contrast to the MAE of 405064 and MAPE of 8.01446% by regression model.

4. Conclusion

In this research, two models are introduced in order to forecast the tax revenue of Wenzhou City. The multiple linear regression model, despite the possible multicollinearity, is agreeable in terms of forecasting, judging from indicators such as multiple R^2, adjusted R^2, MAE and MAPE. However, when compared with the MLP neural network model, the forecasting accuracy of the multiple linear regression model has obvious deficiency.

Due to the limited amount of data, both models are very likely to have errors. Nevertheless, the study still has some value and enlightenment. Most of the past research of tax revenue forecast were concentrated on linear regression model, while this research reflects the possibility that tax revenue might not be perfectly linearly correlated with predictors, reminding researchers that tools for study should be broadened.

References

[1] Huang Wenjun. Study and implementation of local public finance budget income forecasting. Xidian University, 2014.

[2] Yang Han, Wen Ge. Analysis of influencing factors of fiscal revenue in Hebei Province. Co-operative Economy & Science, 2022, (21): 181-183.

[3] Liu Xiang. Research and application of prediction and analysis method for fiscal and tax categories. University of Chinese Academy of Science, 2017.

[4] Xu Weiyi. The analysis of financial revenue growth influencing factors and countermeasures of China. Liaoning Normal University, 2007.

[5] Bai ping. The multiple linear regression model of the financial revenue of our country. Statistics & Decision, 2005, (10): 92-94.

[6] Wang Qi, Guo Shuang. Forecast analysis of fiscal revenue in Gansu Province. China Market, 2018, (28): 39-40.

[7] Li Ming. Influencing factors of financial revenue and forecast of financial revenue in Gansu Province. Shandong University, 2019.

[8] Zhang Rui, Lin Jianming, Peng Jichun. Gray Relational Analysis on the General Budget Revenue of Regional Finance: A Case Study of Hangzhou City. 2009 Second International Symposium on Knowledge Acquisition and Modeling, Wuhan, China, 2009, 300-303.

[9] Zhang Yong. Research on the Model of Tax Revenue Forecast of Jilin Province Based on Gray Correlation Analysis. 2014 Sixth International Conference on Intelligent Human-Machine Systems and Cybernetics, Hangzhou, China, 2014, 30-33.

[10] Sun Jian, Tong Jinzhi. A Research on Economic Factors Affecting China's Tax Growth Based on Panel Error Correction Model. 2011 Fourth International Conference on Intelligent Computation Technology and Automation, Shenzhen, China, 2011, 1005-1009.

[11] Gu Junhua, Song Lijuan, Song Hao, et al. Tax revenue forecasting model based on improved BP neural network. Journal of Hebei University of Technology, 2003, (01): 39-43.

[12] Li Peng. Study of tax revenue prediction based on BP neural network and grey model. Xidian University, 2011.