Modeling on Financial Revenue via Data Mining

Lijun Liu, Yongmei Ding* and Huan Wang
Institute of Systems Science, Wuhan University of Science and Technology, Wuhan, Hubei, 430065, China
dingyongmei@wust.edu.cn

Abstract. As a new first-tier city of core force in the central region, Wuhan plays a vital role in the development of the mid-regional economy of China, which arises our rich interest about the powerful factors for rapid economic improvement. In this paper, we select 12 factors of financial revenue from 2002 to 2018, and try to find out the significant elements through the Adaptive-Lasso algorithm, then predict the future financial revenue by combining with the grey and neural network model. Real example shows the model is efficient to provide a feasible way to predict the financial market in the future.

1. Introduction
The comprehensive reflection of the national economic regional level is reflected in financial revenue. The state and the government have macro-regulated the market economy according to financial revenue. Accurate financial revenue forecasting can not only help the relevant departments to effectively formulate fiscal policies, but also provide forward-looking assistance for strengthening the supervision of financial revenues. In order to ensure scientific and precise financial management and optimize the allocation of social resources, it is necessary to make effective and accurate predictions of financial revenue.

For a long time, the main prediction method of financial revenue include grey forecasting method, stepwise regression method and co-integration analysis method. Yuan R P et al. [1] used the grey model to predict the financial revenue of Fengtai Science and Technology Park and provided the basis management decision for the park. Zhang Q and Dong X [2] used the stepwise regression method to predict and analyze Wuhan financial revenue. Cao P P et al. [3] used co-integration analysis to find a long-term equilibrium relationship between total retail sales of social consumer goods, economic growth and financial revenue in Anhui Province. With the gradual deepening of the research, some scholars analyzed the linear correlation between financial revenue and influencing factors by establishing multiple linear regression models, using the least squares method for parameter estimation, and selecting variables according to stepwise regression method [4-5]. However, both the stepwise regression method and the least square method have an obvious shortcoming, that is the two methods can only achieve the local optimum rather than the global optimal. Therefore, Tibshirani R [6] proposed the Lasso algorithm, which not only reduces the dimensions of the data, but also predicts financial revenue more accurately. Some scholars use this method to estimate the parameters. In order to minimize the sum of squared residuals, the absolute values of the regression coefficients and the constraints less than the constants are set. When the variable coefficients are compressed to 0, it is equivalent to the variable selection [7-8].

As an important city of mid-China, Wuhan plays an vital role in the field of financial revenue. In this paper, we use the Adaptive-Lasso variable selection method to analyze influencing factors. Combining the grey forecasting model and neural network model to establish an effective and feasible financial
revenue prediction model with the relevant data in Wuhan over the years. Further predicting the financial revenue of Wuhan in 2019 and 2020 will provide a basis for macroeconomic regulation and control of the market economy.

2. Theoretical inquiry

2.1. Lasso regression

In 1996, Robert Tibshirani proposed Lasso regression which is a compression estimation of linear regression model. In order to reduce the coefficients to zero for variable selection, it takes the sum of squared residuals as the objective function and adds a penalty term to achieve the purpose of optimization, which is shown in (1).

\[
\hat{\beta} (\text{lasso}) = \arg\min_{\beta} \left\{ \frac{1}{2} \| y - \sum_{j=1}^{p} X_j \beta_j \|^2 + \lambda \sum_{j=1}^{p} | \beta_j | \right\}
\]  

(1)

Where \( \lambda \) is the non-negative regular parameter, and \( \sum_{j=1}^{p} | \beta_j | \) is the penalty.

2.2. Variable Selection via Adaptive-Lasso Process

It solves the shortcomings of least squares method and stepwise regression local optimal estimation in the section 2.1, but it is difficult to derive the evident parameter estimation. Zou\(^{(9)}\) adds different weights to the coefficient of the penalty item, which follows in (2).

\[
\hat{\beta}^{n(n)} = \arg\min_{\beta} \left\{ \frac{1}{2} \| y - \sum_{j=1}^{p} X_j \beta_j \|^2 + \lambda_n \sum_{j=1}^{p} \hat{\omega}_j | \beta_j | \right\}
\]  

(2)

Where \( \hat{\omega}_j = \frac{1}{| \hat{\beta}_j |} \left( y > 0, j = 1,2,\cdots, p \right) \), and \( \hat{\beta}_j \) is the different coefficients of different variables from the least square solution.

2.3. Combined Model of Grey Prediction and Neural Network

The grey model is an way to predict filtered variables and obtain the predicted values for the next two years, which also has good prediction with less data volume. The neural network model is trained by historical data, and then we combine the grey prediction results into the neural network model for more accurate forecasting.

The step to establish a GM(1,1) model, first we supposed \( X^{(0)} = \left \{ X^{(0)}(i), i = 1,2,\ldots, n \right \} \), and accumulate once for \( X^{(0)} \), then get an accumulated sequence, which expressed as \( X^{(1)} = \left \{ X^{(1)}(k), k = 1,2,\ldots, n \right \} \). Second, establishing a first order linear differential equation for \( X^{(1)} \) as follows (3).

\[
\frac{dX^{(1)}}{dt} + aX^{(1)} = u
\]  

(3)

The above first order linear differential equation is GM (1,1) model. Solving the differential equation, the solution of the differential equation is the prediction model.

\[
\hat{X}^{(1)}(k+1) = [X^{(1)}(0) - \frac{\hat{u}}{a}] e^{-ak} + \frac{\hat{u}}{a}
\]  

(4)

The GM (1,1) model finally obtains a cumulative quantity, \( \hat{X}^{(1)}(k+1) \) is reduced to \( \hat{X}^{(0)}(k+1) \), which follows in (5).

\[
\hat{X}^{(0)}(k+1) = (e^{-\hat{a}} - 1) [X^{(0)}(n) - \frac{\hat{u}}{a}] e^{-\hat{a}k}
\]  

(5)
The posterior error ratio test model accuracy table is shown in Table 1.

| P  | C  | Model accuracy | P  | C  | Model accuracy |
|----|----|----------------|----|----|----------------|
| >0.95 | <0.35 | Perfect | >0.70 | <0.65 | Good |
| >0.80 | <0.5 | Great | <0.70 | >0.65 | Bad |

3. **Empirical process**

There are many factors affecting financial revenue (y), considering the relationship between financial revenue and energy consumption, population relations and consumption, we select 12 influencing factor of financial revenue from 2002 to 2018. The influence factor as follows, the number of social employees (x1), total wages of on-post staff and workers (x2), social consumable total retail sales (x3), urban per capita disposable income (x4), urban per capita consumption expenditure (x5), population at the year-end (x6), total social investment in fixed assets (x7), regional GDP (x8), primary industry output value (x9), tax (x10), CPI (x11), tertiary industry and secondary industry output value ratio (x12) and household consumption level (x13) data amount as an independent variable, analyze their impact on the financial revenue and its internal relationship.

3.1. **Adaptive-Lasso based on variable selection**

In the past, we used the Adaptive-Lasso method to select the variables affecting financial revenue. The coefficient of x2, x4, x6~x13 is 0 from Table 2, so these 10 factor can be screened out in the model. The selected variables are urban per capita consumption expenditure (x5), social consumable total retail sales (x3) and the number of social employees (x1).

Table 2. Coefficient table (financial revenue is dependent variable).

| x1 | x2 | x3 | x4 | x5 | x6 ~ x13 |
|----|----|----|----|----|----------|
| 77898.41519 | 0.00000 | -0.63782 | 0.00000 | 2166.21496 | 0.00000 |

Table 3 shows that the number of social employees (x1), total wages of on-post staff and workers (x2), urban per capita consumption expenditure (x5), total social investment in fixed assets (x7) and regional GDP (x8) are screened out.

Table 3. Coefficient table (tax is a dependent variable).

| x1 | x2 | x3 | x4 | x5 | x6 | x7 | x8 |
|----|----|----|----|----|----|----|----|
| -37457.6242 | -170.5929 | 0.0000 | 0.0000 | 646.0411 | 0.0000 | 0.0798 | 760.6234 |

In the same way, the household consumption level is taken as the dependent variable. As shown in Table 4, the two factors of urban per capita consumption expenditure (x5) and primary industry output value (x9) are selected.

Table 4. Coefficient table (household consumption level is dependent variable).

| x1 | x2 | x3 | x4 | x5 | x6 | x7 | x8 | x9 | x11 | x12 |
|----|----|----|----|----|----|----|----|----|-----|-----|
| 0.0000 | 0.0000 | 0.0000 | 0.0712 | 0.0000 | 0.0000 | 0.6364 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |

It can be seen from Table 5 that the two variables of urban per capita disposable income (x4) and urban per capita consumption expenditure (x5) are selected among the factors affecting regional GDP.

Table 5. Coefficient table (regional GDP is a dependent variable).

| x1 | x2 | x3 | x4 | x5 | x6 | x7 |
|----|----|----|----|----|----|----|
| 0.0000 | 0.0000 | 0.2395 | 0.1653 | 0.0000 | 0.0000 | 0.0000 |

3.2 **Forecasting results**
3.2.1 Financial Revenue prediction. The financial revenue influencing factors are substituted into the grey model and artificial neural network. The neural network parameters are $10^7$, 10000, 3, respectively for the error precision, the number of learning and the number of neurons. The predicted values of $x_1$, $x_3$ and $x_5$ in 2019 and 2020 are obtained via grey model. The prediction accuracy of grey model is shown in Table 6.

Table 6. Prediction accuracy of grey model.

|          | $x_1$   | $x_3$   | $x_5$   |
|----------|---------|---------|---------|
| 2019 predicted value | 597.6880 | 85098400 | 35510.7800 |
| 2020 predicted value | 612.3367 | 96989230 | 39188.4500 |
| SSE      | 271.3039 | 868903.4000 | 5466426   |
| MARE     | 0.7564 % | 8.7163 %  | 3.4210 %  |
| Relative accuracy | 99.24363 % | 91.2837 % | 96.5790 % |
| The c value of posterior error ratio | 0.0379 | 0.0778 | 0.0440 |
| Prediction accuracy level | Perfect | Perfect | Perfect |

Through the neural network model, the predicted value of financial revenue in 2019 is 32,747,426 million yuan and the predicted value in 2020 is 33,384,371 million yuan. Figure 1 is a comparison of the financial revenue predicted values and true values based on the neural network model.

![Figure 1](image1.png)

Figure 1. True and predicted values of financial revenue.

3.2.2 Tax prediction. The influencing factors of tax are substituted into the grey prediction model and the artificial neural network. The neural network parameters are $10^7$, 10000, 5, respectively for the error precision, the number of learning and the number of neurons. The predicted values of the influence factors in 2019 and 2020 are obtained via grey model. The prediction accuracy of grey model is shown in Table 7.

Table 7. Prediction accuracy of grey model.

|          | $x_1$   | $x_2$   | $x_5$   | $x_7$   | $x_8$   |
|----------|---------|---------|---------|---------|---------|
| 2019 predicted value | 597.6880 | 104860.5000 | 35510.7800 | 12390.0400 | 18898.8400 |
| 2020 predicted value | 612.3367 | 118067.7000 | 39188.4500 | 14110.0300 | 21539.5700 |
| SSE      | 271.3039 | 124150223 | 5466426 | 14716421 | 8823714 |
| MARE     | 0.7564 % | 9.0187 %  | 3.4210 %  | 37.5096 % | 14.9289 % |
| Relative accuracy | 99.2436 % | 90.9813 % | 96.5790 % | 62.4904 % | 85.0710 % |
| The c value of posterior error ratio | 0.0379 | 0.0670 | 0.0440 | 0.2145 | 0.0978 |
| Prediction accuracy level | Perfect | Perfect | Perfect | Perfect | Perfect |

Based on the tax neural network prediction model, the predicted value in 2019 is 1,427,728,700 yuan and the predicted value in 2020 is 1,588,590,000 yuan. Figure 2 is a comparison of the tax predicted values and true values based on the neural network model.
3.2.3 Household Consumption level prediction. The selected influence factors are substituted into the grey prediction and artificial neural network. The neural network parameters are $10^{-7}$, 10000, 2, respectively for the error precision, the number of learning and the number of neurons. Through the grey model obtains the predicted values for the next two years. The prediction accuracy of grey model is shown in Table 8.

|          | $x_5$     | $x_9$     | Relative accuracy | $c$ value of posterior error ratio | SSE        | Prediction accuracy | MARE       |
|----------|-----------|-----------|-------------------|-----------------------------------|-----------|---------------------|-----------|
| 2019     | 35510.7800| 507.4286  | 96.5790 %         | 0.0440                            | 5466426   | Perfect             | 3.4210 %  |
| 2020     | 39188.4500| 561.0204  | 85.7932 %         | 0.2023                            | 23092.6700| Perfect             | 14.2068 % |

Through the neural network prediction model of household consumption level, the predicted value of household consumption level in 2019 is 2785.0287 and the predicted value in 2020 is 2894.7607. Figure 3 is a comparison of the household consumption level predicted values and true values via the neural network model.

3.2.4 Regional GDP prediction. The selected influence factors are substituted into the grey model and artificial neural network. The neural network parameters are $10^{-7}$, 10000, 2, respectively for the error precision, the number of learning and the number of neurons. The prediction accuracy of grey model is shown in Table 9.

|          | $x_4$     | $x_5$     | Relative accuracy | $c$ value of posterior error ratio | SSE        | Prediction accuracy | MARE       |
|----------|-----------|-----------|-------------------|-----------------------------------|-----------|---------------------|-----------|
| 2019     | 55665.3200| 35510.7800| 94.9023 %         | 0.0514                            | 19047179  | Perfect             | 5.0977 %  |
| 2020     | 62052.9700| 39188.4500| 96.5790 %         | 0.0440                            | 5466426   | Perfect             | 3.4210 %  |
According to the neural network prediction model of the regional GDP, the predicted value of the regional GDP in 2019 is 156.085 billion yuan and the predicted value in 2020 is 15224.3304 billion yuan. Figure 4 is a comparison of the regional GDP predicted values and true values via the neural network model.

![Figure 4. True and predicted values of regional GDP.](image)

4. Conclusion
In the past, we used the least squares method to estimate the model parameters and used stepwise regression to selection variable. But both of them have an obvious shortcoming, that is their solutions can only be partially optimal. The Adaptive-Lasso method for variable selection makes up for this shortcoming because it’s very efficient in processing high-dimensional data and the computational speed is fast enough to produce a clear approximation path. We used grey model to predict a single variable because grey prediction has excellent performance for the prediction of less volume data, and then applying neural network to the final prediction. The more accurate conclusions have been achieved, which provide an effective way to financial analysis.

References
[1] Yuan, R.P., Li, J.X., Tian, Z.Y. (2016) Science Park Revenue Forecast Based on GM(1,1) Residual Model. Mathematics in Practice and Theory, 46: 109-114.
[2] Zhang, Q., Dong, X. (2019) Research on the Analysis and Forecast Model of Wuhan Financial Revenue. Modern Information Technology, 3: 87-88.
[3] CAO, P.P., LI, Y.H., GAO, F.W. (2018) Co-integration Analysis of Total Retail Sales of Social Consumer Goods, Economic Growth and Fiscal Revenue in Anhui Province. Journal of Suzhou University, 33: 40-43.
[4] Guo, S.H., Bao, W.J., Li, F. (2018) Empirical Study on the Influencing Factors of Tax Revenue in Yunnan Province—Based on Data from 1988-2016. Inquiry into Economic Issues, 3: 145-150.
[5] Wang, H., Song, X., Zhuang, Y. (2015) Empirical Analysis of the Main Factors Affecting Chinese Financial Revenue. New Economy, 10: 16-17.
[6] Tibshirani, R. (2011) Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society, 73(3): 267-288.
[7] Cao, F., Zhu, Y.Z. (2012) Lasso Method Based on Multicollinearity. Journal of Southern Yangtze University: Natural Science Edition, 11: 87-90.
[8] Wang, Z., Wu, Y., Zhao, L. (2010) A LASSO-Type Approach to Variable Selection and Estimation for Censored Regression Model. Chinese Journal of Applied Probability & Statistics, 26: 66-80.
[9] Zou, H. (2006) The adaptive lasso and its oracle properties. Journal of the American Statistical Association, 101: 1418-1429.