

Application of Linear Regression Analysis in Predicting the Index of An Electronic Store

Chunyue Lu*

School of China World Academy, Suzhou, China

*Corresponding author: luchunyue@cwacs.cn

Abstract. This paper illustrates how linear regression analysis can be used to establish a suitable mathematical model for the return on investment (ROI) of the store from June to December in 2020. What's more, preliminary analysis of the known data will be explained. Then, the ROI of the store from January to May is estimated by using the model, and the limitations and uncertainties of using the analytical method will be stated by reflecting the process of the analysis. The paper aims to prove the feasibility and rationality of using the method of linear regression analysis to build a model for the prediction of future figures and show exactly how linear regression analysis can be exactly applied in real-world cases. Also, the result shows that the method is effective and reasonable, while there are still errors and uncertainties of the estimation, which makes merchants must consider about all factors when making decisions.

Keywords: Linear regression analysis; ROI; model prediction.

1. Introduction

The paper is mainly based on the analysis of the data of the advertising inputs and outputs of a firm from June to December in 2020. Firstly, the article introduces the background and meanings of the research topic. Then, the theories and methods which are used during the analysis are explained, and they are put into real cases to preliminary analyze the data. Next, a mathematical model is built for the analysis. After testing the effectiveness of the mathematical model, it is used to make a short-term estimation about the future ROI of a merchant working on e-commerce. The predicted figures play an essential role in deciding which month to invest more fund in advertising.

E-commerce is a kind of internet activity based on earning profits from the sale of goods or services within a certain cycle of financial or commercial transactions which are carried out using internet technologies. [1] The last 2 decades have witnessed the rapid development of e-commerce. Taobao, for example, the largest e-commerce company in China, offers research services for over 300 million consumers, bringing 10 billion search queries and subsequent page views (PVs) daily, and providing advertisers with sufficient opportunities to advertise their commodities online. The returns of advertising can be divided into two types: direct returns and indirect returns. The direct returns of advertising are the enhanced impressions of consumers to the goods and the enterprise, increasing number of clicks, and rising conversion rate of potential consumers into actual consumers which are embodied by the advertising page views (PVs). Indirect returns of advertising are online organic flow and offline word-of-mouth marketing. Modeling and quantification of the causal relationship between the overall advertising return and budget can enable the advertisers to spend their money more judiciously. [2]

In productive industries, product strategy is the core of 4P's price strategy of marketing mix (product, price, promotion and place), which is also the basis of distribution strategy and promotion strategy. Making the correct product strategy can let enterprises allocate human, financial, and material resources more effectively and reduce the opportunity cost. Therefore, the firms can timely provide the products that consumers need, and quickly deliver the products to the consumers. In this way, the enterprises can continuously improve their market competitiveness. The most effective evaluations of the quality of product strategies are the quantity of benefits gained by the firms, the efficiency of resource allocation and the extent of improvement of the market competitiveness of the enterprise. These three aspects can not only test the correctness of the product strategy but also reflect

the level of core competitiveness of the enterprise itself. [3] When making decisions on which project to invest in with limited funds, the advice on product strategy from economics is to invest in projects which has the highest rate of return on investment. [4] Thereby, it is very important and meaningful to study the rule of the operation of the ROI system. Also, by combining the realistic situation with the ROI system, enterprises can formulate a regulation system changing with the market. Overall, decision making on product strategy based on the ROI is essential to firms, which is also a research topic worth studying.

2. Model Formulation

2.1. Application of ROI

ROI is a criterion of judgement of the advertising of a firm, which is calculated by the ratio of the net profit and the total cost. It is often used to assess a company's ability to create market value. A relatively high ROI value often represents the evidence of a well-developed and well-managed company. [5] According to the comparison between the industry figures and gross profit, when the fund for advertisement is certain, the advertising of the store is in a benign state if the ROI is greater than 12. It means that if the firm invests larger quantity of capital for advertising, they will earn more revenue. However, when the store's ROI is below 12, the advertising is in a vicious state. If more money is invested for advertising, the store will be in loss. [6]

2.2. Linear Regression Analysis

Unitary linear regression analysis is a method used to establish the linear regression equation between X and Y according to the correlation between independent variable X and dependent variable Y. [7] Since the market is influenced by a lot of factors, using linear regression analysis is very suitable for the analysis of the data in the market. Besides, only when there is only one factor which influences the dependent variable most significantly compared to other factors, the special factor can be called independent variable which can be used for the analysis of linear regression. [6] Linear regression analysis has the advantage of producing the most accurate prediction model for linear data. The analysis makes it convenient for people to evaluate the data and has various methods which are easy for analyzers to apply. [8] It can offer a transparent and interpretable model illustrating the relationship between the interest rate and variables, which helps determine how the figures will likely fluctuate in response to various economic conditions. When a linear regression model describes the connection between the variables well, linear regression analysis can produce reliable predictions of the ROI. [9]

The mathematical model of unitary linear regression analysis: $y = ax + b$ (y is the dependent variable, x is the independent variable and a and b are parameters of the function.)

A and b are obtained from the following functions:

$$b = \frac{\sum Y_i}{n} - a \frac{\sum x_i}{n},$$
$$a = \frac{n \sum x_i Y_i - \sum x_i \sum Y_i}{n \sum x_i^2 - (\sum x_i)^2}.$$

Moreover, every time x goes up by an average of one unit, y goes up by an average of "a" units.

2.3. Multiple Linear Regression Analysis

The linear regression containing two or more independent variables is called multiple linear regression. Since the unit of each independent variable might be different, the coefficients before the independent variables cannot show how important the factors are. For example, there are lots of factors such as salary level, education level, occupation, region, and family burden which can affect the consumption level. However, it is impossible to measure the extent of impact on the consumption level as they are of different units. As a result, all variables including dependent variables should be firstly converted into standard scores, and then the linear regression is carried out. What's more, the

coefficients obtained this time can reflect the degrees of the impacts of the corresponding independent variables on the dependent variable. Meanwhile, the regression equation is called standard regression equation, and the coefficient of the regression is called standard regression coefficient. [6] The model is shown below:

$$Z = \beta_1 Z \cdot 1 + \beta_2 Z \cdot 2 + \dots + \beta_k Z \cdot k.$$

2.4. Application of Quantitative Analysis of Linear Regression Analysis

The following quantitative analysis is based on the investment of a store from June to December in 2020. The ROI figures are analyzed by using linear regression analysis and setting up a mathematical model for short-term prediction. Additionally, during the process of the analysis, it is assumed that there is no abnormal data among the figures above, which means that special factors of consumption such as the low seasons and high seasons of the product is not considered during the ideal analysis.

Table 1. Investment of a Store from June to December in 2020

Items	June	July	August	September	October	November	December
Advertising Input (yuan)	30000	10000	10000	35000	20000	50000	30000
Advertising Output (yuan)	360000	110000	120000	455000	180000	750000	360000
ROI	12	11	12	13	9	15	12
Gross Profit Rate	18%	18%	18%	18%	18%	18%	18%
Store Rate (discount 3%, freight 6%)	9%	9%	9%	9%	9%	9%	9%
Advertising Gross Margin (yuan)	64800	19800	21600	81900	32400	135000	64800
Cost (yuan)	32400	9900	10800	40950	16200	67500	32400
Remaining Gross Margin (yuan)	32400	9900	10800	40950	16200	67500	32400
Merchant State	profit	deficit	profit	profit	deficit	profit	profit

It can be calculated that the median of the advertising input is 30000, which is the middle number of the size ranking of advertising input. The modes of the series of figures are 10000 and 30000, which have the highest frequency of existence. Moreover, the average value of advertising input is 26428.57. By analyzing the figures of advertising output, it can be known that the median, mode and the average value are 360000, 360000 and 333571.43 respectively. The gross margin and the store rate stay at 18% and 9% severally. The promotion gross profit is equal to the sum of the store expenses and the residual gross profit. If labor costs are not considered, when the remaining gross profit is getting closer and keeping greater than the advertising input, the store is profiting. When the remaining gross profit is less than the advertising input, the store is losing money. Since when the value of ROI is greater than 12, the advertising input can still be increased, September and November are the 2 months that advertisers should invest more money in advertising.

According to the above data analysis, it is found that when the gross profit margin and store rate do not change, ROI is equal to advertising output divided by advertising input. Furthermore, the advertising output is equal to the sum of the product of ROI and the advertising input and the floating-point number. The output of advertising is only related to the advertising input, so unitary linear analysis is applicative.

The model is illustrated below:

(Advertising output) $Y = (\text{ROI}) a \cdot (\text{Advertising input}) x + (\text{floating coefficient}) b$,

$$Y_i = ax_i + b.$$

Problems to be solved with unitary linear regression -- parameter estimation:

Estimation of ROI (a) and floating-point number (b); Estimation of difference of two squares.

The estimation of input-output coefficient (a) and floating-point number (b) (Using least square method):

$$Q(b, a) = \sum_{i=1}^n (y_i - b - ax_i)^2.$$

Estimate b and a, making Q(b, a) the smallest.

Least squares estimation of b and a:

$$b = \bar{y} - a\bar{x},$$

$$a = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}.$$

$$a = 15.5491, \quad b = -77369.91855.$$

By using residual sum of squares $\sum_{i=1}^n (y_i - \hat{y}_i)^2$ to estimate the variance σ^2 , $E(\sum_{i=1}^n (y_i - \hat{y}_i)^2) = (n - 2)\sigma^2$ can be proved to be correct. Therefore, $s^2 = \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y}_i)^2$ is the unbiased estimation of σ^2 . To conclude, the estimated unitary linear regression equation is $Y = 15.5491x - 77369.91855$. Moreover, the predicted variance is $s^2 = 1994046243$.

2.5. Model Test and Prediction

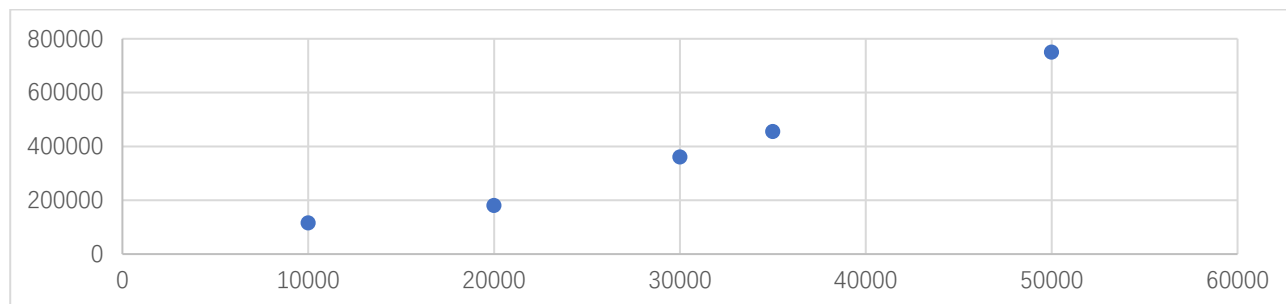


Fig. 1: Relationship between the Advertising input and the Advertising output

As what Figure 1 illustrates, the linear hypothesis is significant. In this case: several points on the scatter plot are almost in a straight line, showing a clear linear relationship. In parameter estimation, the univariate linear regression equation ($y = 15.5491x - 77369.91855$) also shows that there is a linear correlation between the figures. The meaning of the equation is that if the average promotion fee increases by 1 yuan, the average promotion output will increase by 15.5491 yuan.

Table 2. Estimation

Items	January	February	March	April	May
Advertising Input (yuan)	50000	30000	20000	30000	30000
Input-output Coefficient	15.5491	15.5491	15.5491	15.5491	15.5491
Estimated ROI	14	13	11.68	13	13
Advertising Output (yuan)	700086.7288	389104.0699	233612.7404	389104.0699	389104.0699

It can be seen from Table 2 that:

When the coefficient of ROI is unchanged, the actual predicted ROI is infinitely getting closer and closer to 12 to some extents. The overall prediction is credible and has guiding significance for store production.

Judging from the forecast results of this model, the actual predicted ROI in January, February, April and May is greater than 12. As a result, in these months, firms can add more fund for advertising. However, in March, the ROI 11.68 < 12. It shows that the advertising fee in March has reached its limitation, and it is inappropriate to add more money for advertising, or the firm will be in loss.

3. Limitations

The result is inaccurate since the analysis excludes most of the other important factors to establish the unitary linear regression model. For example, the low seasons and high seasons are not considered during the process of analyzing.

Linear regression analysis is not suitable for being used to analyze datasets with missing historical data that generally incorporates drastic fluctuation, or the analysis will be inaccurate. [10] That's because it is difficult to find the complexity and the long-term relevance from small quantity of data. Also, for data with a large time span, linear regression analysis is not a perfect tool for detecting the periodic changes and the changes of trends in the figures. Instead, linear regression analysis is suitable for the analysis of large-volume datasets in a short term. Under this circumstance, it can be a good guidance for the decision-makers in the merchants, which can help firms adjust the amount of investment according to the predicted variations of ROI and expand their market scales.

4. Conclusions

Merchants can increase or decrease the advertising input according to the predicted ROI (ROI > 12, increase the advertising input; ROI < 12, reduce the advertising input, when ROI is getting infinitely closer and closer to 12, stop reducing the advertising input). The application of ROI index in e-commerce stores is remarkable. The data fluctuation in a period plays a significant role in predicting the future figure tendency. Therefore, linear regression analysis has application value in data prediction and is easy and convenient to use. However, there are still limitations and uncertainties of the analysis. Analyzers should roundly consider all possible factors that could affect the data and try to analyze databases with a short time span and plenty of figures to ensure the accuracy of the analysis.

References

- [1] Anastasiia Kameneva. Development of E-commerce in China and in Russia. East China Normal University, 2018, 1-74.
- [2] Liu Hao, Li Yunze, Cao Qinyu, Qiu Guang, Chen Jiming. Estimating Individual Advertising Effect in E-Commerce. College of Control Science and Engineering, Zhejiang University, 2019, 1-7.
- [3] Liu Meili. Research on ROI of product development and operation. Inner Mongol University of Technology, 2007, 1-63.
- [4] William Murdoch, Stephen Polasky, Kerrie A. Wilson, Hugh P. Possingham, Peter Kareiva, Rebecca Shaw. Maximizing return on investment in conservation. Biological Conservation, 2007, 375-388.
- [5] Wang Weijia. Research on the improvement of return on investment of Company B. Dalian University of Technology, 2020, 1-43.
- [6] Li Xiaoyu. Application of linear regression analysis in predicting the index of an electronic store. Chinese Market, 2022, 185-187.
- [7] Zhao Yanjun, Pan Hongxia, Ma Chunmao, Liu Yongjiang. Failure Rate Prediction Based on Grey Linear Regression Combined Model. China Academic Journal Electronic Publishing House, 2014, 34(4): 664-667.
- [8] Matthew Avialdo Pratama, Wowon Priatna. Utilizing linear regression for predicting sales of top-performing products. Internal Journal of Information Technology and Computer Science Applications, 2023, 1(3): 174-180.

- [9] Qiu Tianqi. Predicting the Federal Funds Rate: A Linear Regression Analysis. School of Arts & Sciences, University of Rochester, 2023, 57-62.
- [10] J. Watada et al. Reconstruction of Flow Rate History Using Linear Regression. Institute of Hydrocarbon Recovery and Petroleum Engineering Department, University Technology Petronas, Malaysia, the authors, 2023, 22-35.