

Multiple Linear Regression in Financial Prediction and Evaluation

Ruiqi Jiang

*School of Mathematic Science, Nanjing Normal University, Nanjing, China
06230307@njnu.edu.cn*

Abstract: In order to manage and utilize assets more effectively, prediction and evaluation of expenditure is becoming increasingly important. Scientific financial management facilitates the realization of savings and investments for financial health and sustainability, both for individuals and for businesses. This paper focuses on what role does multiple linear regression play in some economic activities. In order to present the issue, several core theories of multiple linear regression are illustrated. Furthermore, attention have also been paid on how the regression theories apply to realistic problems. To be more specific, some examples from different fields are given to manifest a widespread application, indicating the powerful and technical capacity shown by multiple linear regression. Conclusion can be drawn that the method does improve efficiency and accuracy of prediction and evaluation a lot in general. Thus, this work is significantly beneficial for obtaining information in this field and gaining a deeper understand of how regression act when it comes to more complex situations, which is obviously common in daily lives.

Keywords: Multiple linear regression, Regression analysis, Finance, Cost prediction.

1. Introduction

Multiple linear regression, as a classical method, has been widely utilized since entering the 21st century. Nowadays, applications can be found by continuous crossover with other subjects. With its advantages of strong interpretability and high computing efficiency, multiple linear regression is still a core tool for explanatory modeling in data science [1].

For example, multiple linear regression can be closely integrated with the economic field to study and predict a range of consumption and investment activities. Analysis of energy consumption can be seen in Ref. [2]. In this article, authors point out that energy consumption has become a crucial factor for not only users but also manufacturers and policy makers. Through a newly-built energy consumption predicting model based on multiple linear regression, the consumption is forecast with a better reliability, providing valuable information for improving. In Ref. [3], author notices problems of enterprise environmental management. Therefore, a cost benefit estimation method is proposed on basis of multiple linear regression to improve the accuracy and speed of estimation results. In medical field, regression is used as a vital tool as well. To make the cost of medical insurance more reasonable, a prediction method using regression is illustrated [4]. Under ideal circumstances, this model is of great importance in satisfying both public and organizations. Online finance has been a hit in the past several decades, however, a lot of problems exist. By applying multiple linear regression to e-commerce field, authors are able to figure out issues like financing difficulties [5]. What is more, the

model can also present some advantages and disadvantages from different perspectives. Cost evaluation is an important method of cost control of power transmission and transformation projects. Through regression analysis, cost evaluation quality can be hugely improved by provided references, allowing for more profitable investments [6].

In this paper, some basic theories of multiple linear regression are introduced at beginning. Then it is followed by a simple example of the previous theories. Examples of applications are also included in the main body part, presenting how various of multiple linear regression models and analysis approaches apply to housing price, medical cost and bank performance respectively. In the end, general characteristics of multiple linear regression can be concluded from a range of perspectives.

2. Method and Theory

2.1. Multiple Linear Regression

Multiple linear regression is widely used as a statistical tool [7]. What really counts is its ability to describe the relationship of a specific dependent variable Y and multiple independent variables X_1, X_2, \dots, X_p . This method can improve simple linear regression by combining multiple variables, allowing for a more complicated and general analysis of the factors influencing the dependent variable. Mathematically, it can be expressed by the following formula

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \varepsilon. \quad (1)$$

Variable Y on the left side of equation is predicted from p predictor variables X_1, X_2, \dots, X_p . The relationship between Y and X_1, X_2, \dots, X_p is linear with coefficients $\beta_0, \beta_1, \dots, \beta_p$, and ε is the error term. The major task is to obtain least squares estimates $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p$. To do the estimation, one can use matrix to calculate by least squares is a convenient approach.

To illustrate, four matrixes are defined as the $(n \times 1)$ vector \mathbf{Y} , the $n \times (p + 1)$ matrix \mathbf{X} , the $(p + 1) \times 1$ vector unknown regression parameters $\boldsymbol{\beta}$ and the $(n \times 1)$ vector $\boldsymbol{\varepsilon}$ random errors like:

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}, \mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \quad (2)$$

and e_i = random fluctuation (or error) in Y_i , i.e.,

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_p X_{pi} + \varepsilon_i. \quad (3)$$

One can define the residual sum of squares (RSS) as

$$RSS = \sum_{i=1}^n \hat{e}_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (4)$$

Then the residual sum of squares as a function of $\boldsymbol{\beta}$ can be written in matrix form as

$$RSS(\boldsymbol{\beta}) = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{Y}'\mathbf{Y} + \boldsymbol{\beta}'(\mathbf{X}'\mathbf{X})\boldsymbol{\beta} + 2\mathbf{Y}'\mathbf{X}\boldsymbol{\beta} \quad (5)$$

In order to find the least squares estimates, differentiating the formula with respect to $\boldsymbol{\beta}$ and equating the result to zero and then canceling out the 2 common to both sides are needed. This gives the following matrix form of the normal equations $(\mathbf{X}'\mathbf{X})\boldsymbol{\beta} = \mathbf{X}'\mathbf{Y}$. The least squares estimates are given by $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$.

2.2. Example

Here is a simple example to illustrate the linear regression. Consider a range of samples shown in Table 1, according to Eq. (1), the model can be fitted as $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$.

Table 1: Data of samples for linear regression.

X_1	100	150	200	250
X_2	2	3	3	4
Y	300	400	500	600

Below the purpose lies in estimating $\beta_0, \beta_1, \beta_2$. Based on Eq. (2), one can define

$$\mathbf{X} = \begin{pmatrix} 1 & 100 & 2 \\ 1 & 150 & 3 \\ 1 & 200 & 3 \\ 1 & 250 & 4 \\ 1 & 300 & 4 \end{pmatrix}, \mathbf{Y} = \begin{pmatrix} 300 \\ 400 \\ 500 \\ 600 \\ 700 \end{pmatrix}, \hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} \quad (6)$$

The estimates of $\hat{\boldsymbol{\beta}}$ come out

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \begin{pmatrix} 12.8 & -0.08 & -2.4 \\ -0.08 & 0.0008 & 0.02 \\ -2.4 & 0.02 & 0.6 \end{pmatrix} \begin{pmatrix} 2500 \\ 575000 \\ 8850 \end{pmatrix} = \begin{pmatrix} 100 \\ 2 \\ -50 \end{pmatrix} \quad (7)$$

Therefore, results can be shown as $\hat{\beta}_0 = 100, \hat{\beta}_1 = 2, \hat{\beta}_2 = -50$, and the regression equation can be written as $Y = 100 + 2X_1 - 50X_2$.

3. Applications

3.1. Housing Prediction

In the Ref. [8], the authors proposed that previous study has identified three categories of characteristics that have a substantial impact on overall housing prices: house conditions, environmental factors, and transportation considerations.

Spearman's correlation coefficient has been considered as a nonparametric statistical approach to evaluate the monotonic association between two variables. Comparing to Pearson's correlation coefficient, the Spearman's measure of correlation does not require the data to follow a linear connection, giving it a benefit when dealing with nonlinear interactions. Furthermore, Spearman's correlation coefficient is insensitive to extremes and can deal with discontinuous data.

Based on the traditional model, a newly-built linear regression model can be set as:

$$y_i = \beta_0 + B'x_i + \varepsilon_i \quad (8)$$

where $y_i = (x_{i1}, \dots, x_{id}, \dots, x_{iD})'$ and x_i are the D -dimensional vector of output variables and the P -dimensional vector of fixed regressor values for i^{th} sample unit correspondingly. β_0 is a D -dimensional vector containing the intercepts for D response. B is a $P \times D$ vector matrix whose element β_{pd} is the regression coefficient of p^{th} regressor on the d^{th} response. ε_i represents the D -dimensional random vector of the error terms relating to the i^{th} observation. Noticeably, there is an assumption that the multiple linear regression model satisfies the equation

$$\varepsilon_i \sim \sum_{k=1}^K \prod_k MVN(v_k, \sum K), \quad (9)$$

with a condition that positive weights π_k and v_k are required to gain the relationship $\sum_{k=1}^K \pi_k v_k$.

The raw data is then standardized. In the tradition of regression analysis, the complete dataset is separated into two parts: training and testing. By comparing the regression results from the training set to the modeling results from the test set, the model's accuracy and efficiency may be successfully assessed. During the training procedure, the gradient descent optimizer is used to determine the ideal parameters. The progressive fall of the loss function indicates that the parameters in the model are becoming more suited for application.

Afterwards, the trained parameters above are applied to the prediction. The author chooses 100 samples from the dataset for the model. Figure 1 shows an obvious conclusion, indicating that the model's anticipated outcomes are compatible with the trend of the actual values in general, and in the majority of cases, predicted values change in the same direction as the actual values. Some errors occur concurrently, although they are often considered acceptable. It can be inferred that the multiple regression model can forecast and evaluate property prices to a certain degree. However, the prediction has a limited accuracy.

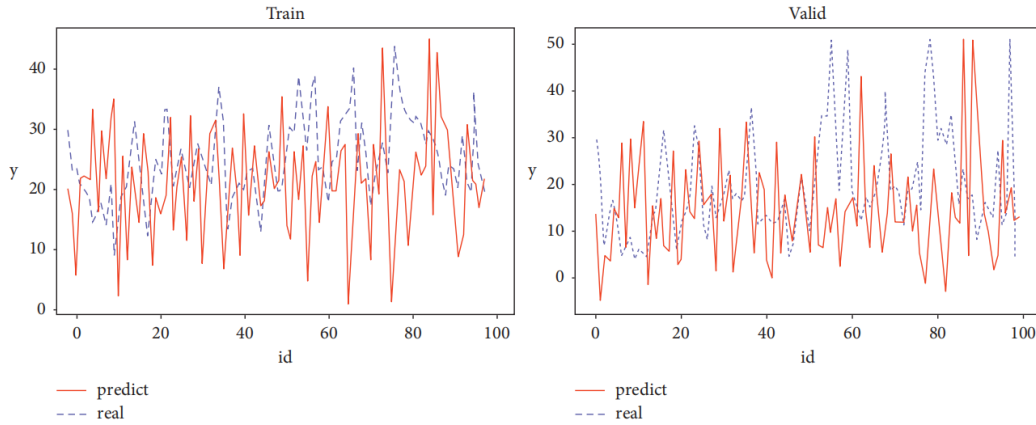


Figure 1: The real and prediction result [8].

3.2. Treatment Cost Prediction

By anticipating the expenditures of disease treatment, some organizations can effectively promote health insurance reform, rationalize medical resource allocation, and reduce patient burdens. In article [9], characteristics influencing treatment costs were collected from the electronic medical record, including the patient's age, gender, surgical history, treatment regimen, previous medical history, current admission status, smoking status, diabetes, and hypertension. The information provided above is divided into four sections: patient explanation, history of current illness, medical history, and personal history. To utilize the data better, preprocess the data is of great significance. Written textual descriptions (e.g., gender, medical history, etc.) in the electronic medical record are translated into numerical variables for use in model calculations.

To make the given data match better, the error term ε in Eq. (1) may be defined as a variable whose function is setting the coefficients of some low action variables into 0 in the iterative process. Therefore, it possesses the ability to deal with the situation where limited data appear. By comparison, local weighted LASSO (Least Absolute Shrinkage and Selection Operator) regression simplifies the model structure by compressing the coefficients of variables whose impact is relatively small to improve generalization. The specific error term is set as bellow:

$$\varepsilon = \min \frac{1}{m} \left[\lambda \sum_{i=1}^m |\beta_i| \right] \quad (10)$$

where m is the number of samples, λ is the regularization coefficients, β_i is the model parameter. Further, an upper limit is required to make the parameters easier to deal with $\sum_{i=1}^n |\beta_i| \leq \sigma$. One can enlarge or compress β with the value of σ changing.

In most cases, random variables represent normal distribution features, so the authors introduced weight function based on normal distribution as $\omega^{(i)} = \exp\left(-\frac{x^{(i)} - x}{2\sigma^2}\right)$. Thus, ε can be written as below, which allows for a local weighted way to promise the efficiency and feasibility of model:

$$\varepsilon = \min \frac{1}{m} \left[\lambda \sum_{i=1}^m \omega^{(i)} \beta_i \right] \quad (11)$$

Then comes the example data analysis. After the preprocess and classification, the data is regressed and fitted for different categories to acquire prediction situation of several models. To illustrate directly, linear regression, LASSO regression, neural network, locally weighted LASSO regression and Bayesian network fusion local weighted LASSO regression are included. Conclusions can be drawn by comparing values like accuracy in Table 2.

Table 2: Comparison of different models in accuracy

Model Name	Accuracy (%)	Mean Squared Error	R-square
Linear regression model	59.74	13.32	0.58
LASSO regression model	65.78	10.88	0.66
Neural network model	63.45	11.06	0.62
Locally weighted LASSO regression model	85.65	6.38	0.75
Bayesian network fusion local weighted LASSO regression model	89.14	5.36	0.81

From Table 2, accuracy of LASSO regression and Bayesian network fusion local weighted LASSO regression is apparently much higher than the rest of the models, which manifest that locally weighted LASSO regression does have advantages in generalization ability and validity. Furthermore, when combining with Bayesian network classification, the model can gain a more excellent prediction capacity during the experiment. The model is found valid in predicting the expense of treatment for a patient with a limited amount of data and is able to recommend an appropriate treatment plan for the patients.

3.3. Financial Performance of Bank

Banks are of great importance in economic growth, especially private banks. India's banking system has made remarkable achievements over the past few decades, particularly in the areas of risk management and capital adequacy. In order to assess the financial performance of a bank, it is usually analyzed using financial ratios. People also use the financial data of banks for the period 2006 to 2017 to assess banks' performance through three key indicators and uses multiple regression analysis to develop three models [10].

To begin with, Return on Assets (*ROA*), the ratio of market value of banks to book value of equity (*Tobin's Q*) and Return on Equity (*ROE*) are introduced as three dependent variables. Subsequently, five independent variables *Bank size*, Credit Risk (*CR*), Operational Efficiency (*OE*), Asset Management (*AE*), Debt Ratio (*DR*) are added as well.

Then, the authors consider the three dependent variables respectively with 3 different models. In each model, a null hypothesis that the single dependent variable is not influenced by the given 5 independent variables is presented. Log of Total Assets (*TA*) is used to represent *Bank size*. Total debt/*TA* represents Debt Ratio. Afterwards, do the multiple linear regression analysis and hypothesis test through calculated data.

In the first model, the regression model can be assumed as:

$$ROA = \beta_0 + \beta_1 CR + \beta_2 OE + \beta_3 AM + \beta_4 Bank\ size + \beta_5 DR + \varepsilon \quad (12)$$

The analysis results are presented in Table 3.

Table 3: Coefficients of the first model

Variables	Raw Coefficients		Standardized Coefficients	Significance Level
	B	Standard Error	β	
CR	-0.268	0.053	-0.622	0.000
OE	-1.268	0.429	-0.371	0.006
AM	22.206	6.047	0.378	0.001
Log (TA)	0.233	0.138	0.228	0.103
Total Debt/TA	-0.006	0.007	-0.122	0.362

The ANOVA (Analysis of Variance) result of the first model indicates that null hypothesis against alternative hypothesis. If the specific variable's Significance level is less than 0.05, it is assumed that the predictor variables do not have a significant impact on relevant variable in the model. Therefore, CR, OE and AM have significant influence on *ROA* while *Bank size* and Debt Ratio's impact on *ROA* is not apparent. In the second model, the regression model is the equation as below:

$$Tobin's Q = \beta_0 + \beta_1 CR + \beta_2 OE + \beta_3 AM + \beta_4 Bank\ size + \beta_5 DR + \varepsilon \quad (13)$$

The analysis results are shown in Table 4.

Table 4: Coefficients of the second model

Variables	Raw Coefficients		Standardized Coefficients	Significance Level
	B	Standard Error	β	
CR	0.004	0.258	0.003	0.987
OE	-2.212	2.098	-0.199	0.300
AM	-6.183	29.574	0.037	0.836
Log (TA)	-0.960	0.677	-0.288	0.016
Total Debt/TA	-0.104	0.033	-0.616	0.004

After doing the ANOVA, it can be seen that the null hypothesis against alternative hypothesis. The observed significance level (p-value) provides empirical evidence that the model achieves statistical significance at the predetermined 5% threshold. Therefore, from Table 4, on one hand, *Bank size* and DR have significant impact on *Tobin's Q*. On the other hand, the rest 3 variables show insignificant effect on *Tobin's Q*.

In the third model, authors consider the model as:

$$ROE = \beta_0 + \beta_1 CR + \beta_2 OE + \beta_3 AM + \beta_4 Bank\ size + \beta_5 DR + \varepsilon \quad (14)$$

The analysis results can be seen from Table 5.

Table 5: Coefficients of the third model.

Variables	Raw Coefficients		Standardized Coefficients	Significance Level
	B	Standard Error	β	
CR	-3.186	0.630	-0.622	0.000
OE	-3.123	5.129	-0.077	0.547
AM	263.533	72.287	0.377	0.001
Log (TA)	-4.990	1.654	0.410	0.005
Total Debt/TA	0.018	0.081	-0.030	0.823

Results can be seen from ANOVA of the third model that null hypothesis against alternative hypothesis. The significance of variables can be evaluated from Significance Level in Table 5 respectively, which indicates CR, AM and *Bank size* have significant influence on *ROE*. OE and DR have insignificant impact on *ROE*. Ultimately, the three models above derive the key factors affecting their financial performance, which provide a comprehensive view of the bank's operations and risk levels for bank managers and investors.

4. Conclusion

Considering the results, it is natural to find that multiple linear regression is a flexible method when combining with other techniques like neural network and Bayesian network. What really counts is that its massive potential value has been shown significant in numerous areas. The model is easy to understand and express because its coefficients visualize the relationship between the variables directly. By quantifying the effects of each factor, multiple linear regression can be clearly accomplished in complex scenarios, adapting to the needs of modern data analysis. However, it is not always the perfect choice in many circumstances. Due to its strict hypothesis, realistic data usually fail to meet the conditions of independent and identical distribution and linear relationship. What's more, some abnormal data may do extreme harm to the whole model, resulting in error and irrational consequences. This paper is limited in presenting the process of building the model specifically. In addition, some preprocessing of the selected data is neglected. For further development, future research may focus on how multiple linear regression can better connected with other powerful tools and deal with data whose relationship with other factors is nonlinear. When combining with artificial intelligence, the modeling process may be optimized automatically, lowering the expenditure and time consumption on adjusting the given environments.

References

- [1] Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). *Introduction to linear regression analysis*. John Wiley & Sons.
- [2] Chen, Y., Huang, M., & Tao, Y. (2022). Density-based clustering multiple linear regression model of energy consumption for electric vehicles. *Sustainable Energy Technologies and Assessments*, 53, 102614.
- [3] Wang, M. (2025). Study on cost benefit estimation of enterprise environmental management based on multiple linear regression. *International Journal of Manufacturing Technology and Management*, 39(1/2), 74-88.
- [4] Alzoubi, H. M., Sahawneh, N., AlHamad, A. Q., Malik, U., Majid, A., & Atta, A. (2022, October). Analysis of cost prediction in medical insurance using modern regression models. In *2022 International Conference on Cyber Resilience (ICCR)* (pp. 1-10). IEEE.
- [5] Wang, P., & Han, W. (2021). Construction of a new financial E-commerce model for small and medium-sized enterprise financing based on multiple linear logistic regression. *Journal of Organizational and End User Computing (JOEUC)*, 33(6), 1-18.
- [6] Ye, M., Chen, X., Chen, K., Liu, M., & Wu, H. (2023, March). Research on cost evaluation of power transmission and transformation project based on grey correlation and regression analysis. In *Second International Conference on Statistics, Applied Mathematics, and Computing Science (CSAMCS 2022)* (Vol. 12597, pp. 1049-1054). SPIE.
- [7] Aiken, L. S., West, S. G., Pitts, S. C., Baraldi, A. N., & Wurpts, I. C. (2012). Multiple linear regression. *Handbook of Psychology*, Second Edition, 2.
- [8] Zhang, Q. (2021). Housing price prediction based on multiple linear regression. *Scientific Programming*, 2021(1), 7678931.
- [9] Tong, L. L., Gu, J. B., Li, J. J., Liu, G. X., Jin, S. W., & Yan, A. Y. (2021). Application of Bayesian network and regression method in treatment cost prediction. *BMC Medical Informatics and Decision Making*, 21, 1-9.
- [10] Nataraja, N. S., Chilale, N. R., & Ganesh, L. (2018). Financial performance of private commercial banks in India: multiple regression analysis. *Academy of Accounting and Financial Studies Journal*, 22(2), 1-12.