Sales Prediction Based on ARIMA Time Series and Multifactorial Linear Model

Ziru Zhang*

Shandong University of Finance and Economics, Jinan, China

* Corresponding Author Email: 20190911612@mail.sdufe.edu.cn

Abstract. Contemporarily, sales forecasting based on historical data is one of the key topics of research. Researchers have found that different analytical forecasting models can predict the future sales volume to a certain extent based on different actual conditions. However, it is also very important to compare the prediction results of these models, and a reasonable choice of models can help managers to make more scientific decisions. Therefore, in this paper, the historical sales data of a store on the Kaggle is utilized to make predictions based on different models, and investigate the impact of linear regression model and time series model on sales prediction. In this paper, multiple factorial linear regression model and ARIMA model in time series analysis will be adopted based on R to build prediction models to forecast sales volume, respectively. According to the analysis, the root mean square error of the prediction results under the multiple factorial linear regression model and the ARIMA model were 21.135 and 23.221, respectively. The study surface that there were differences in the degree of fit of the prediction results by different pairs of models. Therefore, the choice of the researcher needs to select the appropriate model for prediction according to the specific object of analysis, combined with the data characteristics to achieve the expected results. These results shed on guiding further exploration of differences in predictions from different models.

Keywords: Linear regression; time series; sales forecasting.

1. Introduction

Sales forecasting plays an important role in business development. Regardless of the size of a business or the number of salespeople, accurate sales forecasting can have a significant impact on all aspects of sales management, including planning, budgeting, and determining sales. In most densely populated cities, where the number of stores is growing, sound sales forecasting helps stores develop scientific and effective sales strategies to increase store revenues and reduce unnecessary losses.

In store sales, the amount of inventory in a store is closely related to sales, and the role of forecast sales cannot be underestimated. Too much forecast quantity and the store will incur losses due to backlogged merchandise or perishable merchandise. Forecast too few quantities and items that sell well are quickly sold out, easily leading to lost revenue and customer dissatisfaction. Therefore, accurate forecasting can help ensure that retailers have enough of the right products at the right time to please their customers [1].

There are various methods of forecasting based on historical data or values, which have been refined and optimized through the development of science and technology. In terms of sales forecasting, statistical analysis techniques for forecasting sales, including time series, linear regression and other methodologies. Both multiple factorial linear regression models and time series models in forecasting models are widely used in many scientific fields, e.g., economics, society, and medicine.

Linear regression is widely used in practice and can even be naturally adapted to complex prediction tasks. For example, Bianco, et al. forecasted electricity consumption using linear regression models [2]; Swain, et al. predicted precipitation based on multiple factorial linear regression [3]. In the field of sales, Puteri and Safitri applied linear regression for analysis of car sales [4]. Gopalakrishnan, et al. predicted sales value of online shopping with linear regression [5].

Time series are primarily concerned with finding correlations inherent in the values of historical data series. For example, Ramos et al. leveraged the performance of state space and ARIMA models

for consumer retail sales forecasting [8]; Permatasari et al. reported on sales forecasting using ARIMA [9].

The core difference between time series and regression analysis lies in the assumptions made about the data. To be specific, regression analysis assumes that each sample data point is independent; whereas time series uses correlations between data to make predictions. The assumptions of both can have different effects on the results. In practice, the usage of different methods can produce different sizes of errors in the forecasting results. The question of which model is more suitable for forecasting sales is of great interest to managers in terms of how science should manage the business.

In this paper, based on the idea of statistics and R programming language, this study establishes prediction models for forecasting sales of multiple factorial linear regression and time series, respectively. Afterwards, the prediction results of both according to the root mean square error (RMSE) are compared and their advantages and disadvantages are analyzed. Eventually, the conclusions are drawn for the store sales to forecasting to provide some reference [6].

In terms of the data used in this study, since the true values of the test dataset were not provided in the source data, the data from the test were not considered in the modeling, but the data from the train were divided into 80% as the training set and 20% as the test set to build the model. Multiple factorial linear regression method was used under the linear regression method through the store's historical sales data. Appropriate independent variables were selected for linear regression prediction of their sales quantities. By testing the results, the corresponding independent variables whose coefficients are significantly non-zero at the significance level of 0.05 are selected, multiple factorial linear regression expressions are established, and the root mean square error (RMSE) between their predicted results and the true values is calculated. Under the time series approach, the ARIMA model was selected to predict the historical data of the store, and firstly, the series smoothness was judged, i.e., the ADF test. If the series was not smooth, the series differencing was performed to make the differenced series smooth [7]. In addition, ACF and PACF plots are drawn based on the smooth series, and according to the results, the appropriate parameters p and q are estimated, and the ARIMA (p, d, q) model is built and the model fit test is performed [8]. Using the ARIMA model that passes the fit test, the series can be predicted and its corresponding root mean square error (RMSE) calculated [9]. By comparing the magnitude of the RMSE under the linear regression approach and the RMSE under the time series approach, it is possible to determine which model is closer for the predicted number of sales [10]. The rest part of the paper is organized as follows. The Sec. II will describe the data and introduce the methods of linear model and ARIMA model. The Sec. III will show the results of the two models and compare them.

2. Methodology

2.1. Data

The data used in this study is from Kaggle's supermarket sale, the Owner is Aung Pyae. the dataset is one of the historical sales data of the supermarket company, recorded for 3 months at 3 different branches. This data includes 1000 samples, each sample represents the sales of each order of a store separately. Changes in sales volume are the result of a combination of many variables. When forecasting sales volume, it is difficult to take into account all the factors that influence sales volume and the extent to which these factors affect the results. It is common to select the main influencing factors and achieve the desired results through analysis and optimization, i.e., the process of screening the main variables to build an optimal forecasting model. This data includes 17 variables, and 8 variables were finally selected for this study based on the relationship between each variable and the number of sales.

The dependent variable is quantity, which indicates the number of products purchased by customers. The independent variables include Branch (the branch of the supercenter. (There are 3 branches, A, B and C). City (the location of the supercenter). Customer type (customers with membership cards are recorded as members, those without membership cards are recorded as regular).

Gender (the gender type of the customer). Product Line (general item classification groups - Electronic Accessories, Fashion Accessories, Food & Beverage, Health & Beauty, Home & Lifestyle, Sports & Travel). Unit price (price per product, in US dollars). Since the source data is the historical data of each order in the store, the data is used by first sorting and aggregating the above variables into daily sales data by find date before performing subsequent analysis and forecasting.

2.2. Linear model

Let the dependent variable y and the independent variables x_1, x_2,⋯, x_k have n sets of actual observations (x_{i1}, x_{i2},⋯, x_{ik}) (i = 1,2,⋯,n), and the expression of the multiple factorial linear regression is:

$$y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} \tag{1}$$

Where $\varepsilon_i \sim N(0, \sigma^2)$. This paper established a multiple factorial linear regression model by R for the influence of many factors such as Branch, City, Customer type, Gender, Product line, Unit price, etc. [11]. After constructing the model, this paper analyzes the correlation coefficient and significance degree of the influencing factors by parameter estimation and data fitting methods are implmented to judge the credibility of parameter estimation and to make reasonable inferences about the sales trend of the store to some extent.

2.3. Arima model

The ARIMA Model -- Auto Regressive Integrate Moving Average Model is formulated as follows:

$$\nabla^d x_t = \phi_0 + \sum_{i=1}^{p} \phi_i x_{t-i} + \epsilon t - \sum_{j=1}^{q} \theta_j \epsilon_{t-j} \tag{2}$$

Here,

$$\begin{cases} \phi_p \neq 0, \theta_q \neq 0 \\ E(\epsilon_t \epsilon_s) = 0, s \neq t \\ E(x_s \epsilon_t) = 0, \forall s < t \\ \nabla^d = (1 - B)^d \end{cases} \tag{3}$$

The above equations shows that the essence of the ARIMA model is the combination of the difference operation and the ARMA model, and this relationship is significant. It indicates that any non-stationary series can be fitted to the ARMA model if the post-differential smoothness can be achieved by differencing of appropriate order. the analysis method of the ARMA model is very mature, which means that the analysis of the differential smooth series is also very reliable. The basic steps of modeling the ARIMA model are shown in Figure 1.

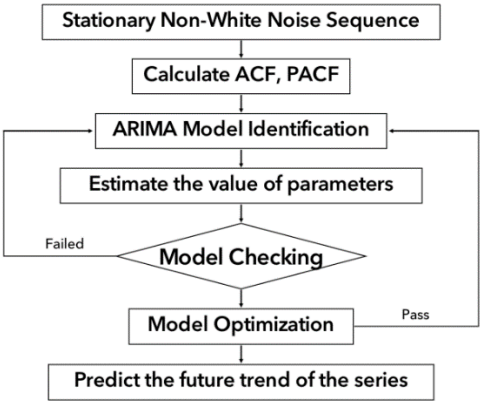


Fig. 1 Basic steps of modeling ARIMA model

3. Results & Discussion

3.1. Empirical analysis for linear model

A multiple factorial linear regression analysis was conducted using Branch, City, Customer type, Gender, Product line, and Unit price as independent variables and Quantity as dependent variable from the store's historical sales data. In multiple factorial linear regression, the meaning of the regression coefficient is the amount by which the dependent variable will increase if one predictor variable increases by one unit while the other predictor independent variables remain constant. By analyzing the software in R, the following results are obtained, as shown in Table. I.
On this basis, a linear regression equation can be obtained:

$$Y = 5.708 - 0.0805\, Unit.price + 0.0146\, cogs \quad (4)$$

The results show that the coefficient on unit.price is significantly non-zero at a significance level of 0.05, indicating that unit.price shows a linear correlation with Quantity when controlling for other variables that are constant. Overall, all independent variables explained 81.73% of the variance of Quantity.

Tab. 1 Results of MULTIPLE FACTORIAL linear regression model

Coefficients				
	Estimate	Std.	t value	Pr(>|t|)
Intercept	5.708	0.158	36.078	<2e-16
BranchB	-0.017	0.097	-0.171	0.864
BranchC	-0.086	0.098	-0.876	0.381
CityNaypyitaw	NA	NA	NA	NA
CityYangon	NA	NA	NA	NA
GenderMale	-0.040	0.080	-0.504	0.614
Customer.typeNormal	-0.061	0.080	-0.762	0.446
Product.lineFashion accessories	-0.146	0.136	-1.074	0.283
Product.lineFood and beverages	-0.076	0.136	-0.555	0.579
Product.lineHealth and beauty	-0.032	0.141	-0.230	0.818
Product.lineHome and lifestyle	-0.118	0.139	-0.848	0.397
Product.lineSports and travel	-0.068	0.138	-0.491	0.624
Unit.price	-0.080	0.002	-41.188	<2e-16
cogs	0.015	0.000	65.904	<2e-16
PaymentCredit card	-0.138	0.099	-1.393	0.164
PaymentEwallet	-0.002	0.096	-0.016	0.987
Residual standard error: 1.258 on 986 degrees of freedom				
Multiple R-squared:0.8173		Adjusted R-squared:0.8149		
F-statistic: 339.3 on 13 and 986 DF, p-value: < 2.2e-16				

3.2. Empirical analysis for ARIMA

Before performing time series analysis, the smoothness and pure randomness of the series are first tested, and these two tests are called preprocessing of the series. Once the series satisfies the condition of smooth non-white noise, the time series modeling and subsequent analysis can be performed.

The statistical test for smoothness is constructed mainly based on the relationship between a smooth series and a unit root. It is based on the theory that if the series is smooth, then all the characteristic roots of the series should be within the unit circle. The method of testing the smoothness of a series constructed based on this property is called the unit root test. Here, the ADF test is chosen to perform the unit root test. The ADF test was performed on this data and the results are shown in Table. II. The probability value for type 1 is greater than the criterion value of α=0.05 (p-value>0.05). The output indicates that the data is not stationary. Therefore, this data needs to be differentiated and

transformed so that the data becomes stationary. The ADF test was first performed and the results were differentiated as also in Table. II.

Alternative models are selected by analyzing the ACF and PACF plots: that is, determining whether the sample ACF and PACF are trailing or truncated, specifying the parameters p and q, where p is the highest order in the autoregressive polynomial and q is the highest term in the moving average polynomial, and thus determining the prediction model.

Tab. 2 ADF tests

adf.test (daily quantity)			adf.test (diff-daily quantity)		
lag	ADF	p_value	lag	ADF	p_value
0	-2.172	0.031	0	-14.57	0.01
1	-1.276	0.220	1	-10.28	0.01
2	-0.856	0.371	2	-9.18	0.01
3	-0.685	0.432	3	-8.47	0.01
Note: p_value = 0.01 means p_value <= 0.01					

To select the model, the parameters p, d and q in the ARIMA model need to be determined. Figure 2 shows the ACF and PACF plots respectively, according to which the ARIMA (4,1,2) model can be considered. To select the model, the parameters p, d and q in the ARIMA model need to be determined. Figure 2 shows the ACF and PACF plots respectively, according to which the ARIMA (4,1,2) model can be considered.

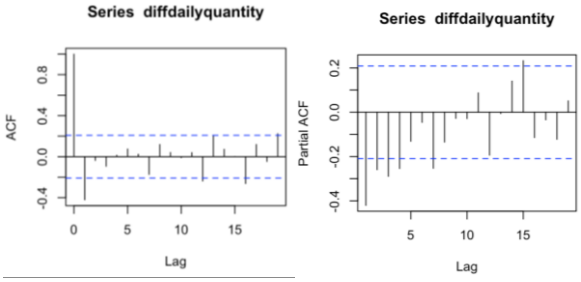


Fig. 2 ACP and PACF diagram.

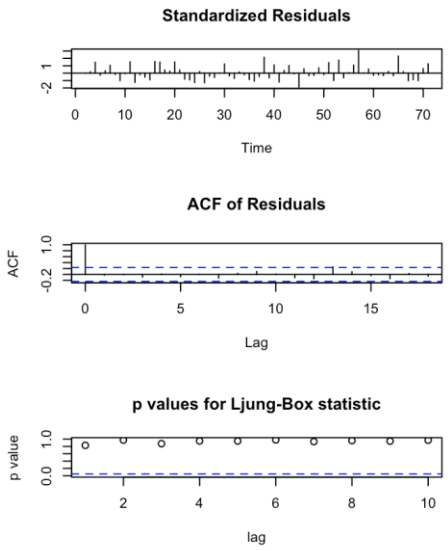


Fig. 3 Model fitting result.

Here, d=1 is specified, i.e., the first-order difference is applied to the function series. Therefore, the original model can be directly applied to the original series. The results of the model significance test in Figure 3 indicate that the residual series of the fitted model can be regarded as white noise

series. The p-value of the residual white noise test is greater than the significance level in Table III. It is considered that the residuals do not contain valid information. The fitted model holds significantly and the fitting effect is good. The data of the next 18 days (20% of the source data) are predicted by the forecast function in R, and the prediction results are illustrated in Figure 4.

Tab. 3 Residual white noise test results

Box-Pierce test
data: residual
X-squared = 6.1426, df = 6, p-value = 0.4074
X-squared = 11.106, df = 12, p-value = 0.5199

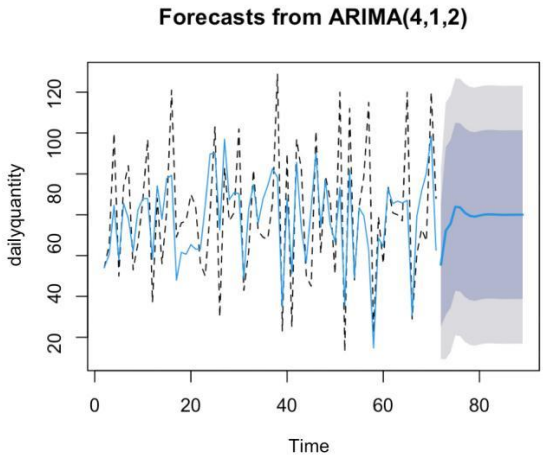


Fig. 4 Prediction results of ARIMA.

3.3. Comparison of results

The RMSE values under the two analysis methods were calculated using the accuracy function in R, and the results are summarized in Table. IV. In terms of the results, the RMSE of the predicted value under the multiple factorial linear regression model is 21.135. the RMSE under the time series ARIMA model is 23.221. indicating that the multiple factorial linear regression model is much closer for the predicted number of sales in the scenario of this data.

Tab. 4 RMSE results

Model	RMSE
linear model	21.13532
ARIMA	23.22054

3.4. Explanation

According to the analysis, factors have significant impacts on the predicted sales data due to the following reasons. In this study, the R2 value under the multiple factorial linear regression model is only 81.73%, which is not a very high degree of explanation for the explanatory variables and affects the accuracy of the forecast. In the time series analysis, the effect of specific external factors is not considered for the time being, highlighting the role of the time factor in forecasting. The main reason accounted for this is that the past of things continues into the future and does not change by leaps and bounds in market forecasting.

For implementation of the store historical sales data for future sales forecasting, one can combine a variety of analysis methods and models, not limited to multiple factorial linear regression models and time series ARIMA models. In addition, the actual situation of the market and society ought to be taken into consideration, and a reasonable forecast result for decision management should be chosen, which can improve forecast accuracy and reduce business losses.

It should be noted that since there are many models for the time series for the selected data (e.g., exponential smoothing models), the researcher can choose the appropriate model to fit the forecast according to the specific data situation. Besides, because of the characteristics of time series analysis mentioned above, there are often large deviations when there are large changes in the external world. On this basis, the time series model is more suitable for short- and medium-term forecasting. Second, in multiple factorial linear regression, the choice of specific factors is more subjective, which can affect the diversity and unpredictability of factors, making the regression analysis limited.

4. Conclusion

In summary, this paper investigates sales forecasting based on linear regression and time series models. It is found that different analytical models have different results on sales volume forecasting and different magnitudes of forecasting errors. Based on the analysis, the two forecasting methods, multiple factorial linear regression model and time series model, are mainly summarized, and the basic principles of the two methods are introduced in detail, and the sales data are forecasted separately. In the study of this paper, the final fitting effect of both is good, and the model is reasonable and effective through the test analysis, which has some practical significance.

The multiple factorial linear regression model measures the effects of changes in multiple independent variables on the dependent variable, and the model is simple and easy to analyze, which is suitable for practical economic problems influenced by multiple factors. Nevertheless, the choice of specific factors is more subjective, which affects the diversity and unpredictability of factors and makes the regression analysis limited. Time series analysis temporarily disregards the influence of specific external factors, highlighting the role of the time factor in forecasting. The main reason is that the past of things will continue into the future and will not change by leaps and bounds in market forecasting. Moreover, there will often be large deviations when large changes occur in the outside world.

In this study, the multiple factorial regression model is relatively more accurate in prediction, but it is also too complicated and cumbersome. The time series model, on the other hand, has a relatively large error compared to the multiple factorial regression model, but the method is simpler and easier to master and learn. Time prediction models can be used in issues that do not strictly require accurate data and where the accuracy of the predicted values does not have a significant impact. If the two prediction models can be well integrated, it will expand the application of the model and improve the prediction accuracy at the same time. In the future, some machine learning scenarios can be adopted for prediction to achieve better prediction results. Overall, these results offer a guideline for the forecasting of sales under the linear model and the Arima model.

References

[1] G. Kechyn, et al. "Sales forecasting using WaveNet within the framework of the Kaggle competition." arXiv preprint arXiv:1803.04037, 2018.

[2] V. Bianco, O. Manca, and S. Nardini, "Electricity consumption forecasting in Italy using linear regression models." Energy, vol. 34.9, 2009, pp. 1413-1421.

[3] S. Swain, P. Patel, and S. Nandi, "A multiple linear regression model for precipitation forecasting over Cuttack district, Odisha, India." 2017 2nd international conference for convergence in technology (I2CT). IEEE, 2017.

[4] C. K. Puteri, and L. N. Safitri. "Analysis of linear regression on used car sales in Indonesia." Journal of Physics: Conference Series. Vol. 1469. No. 1. IOP Publishing, 2020.

[5] T. Gopalakrishnan, R. Choudhary, and S. Prasad, "Prediction of Sales Value in Online shopping using Linear Regression." 2018 4th International Conference on Computing Communication and Automation (ICCCA). IEEE, 2018.

[6] B. M. Pavlyshenko, "Linear, machine learning and probabilistic approaches for time series analysis." 2016 IEEE First International Conference on Data Stream Mining & Processing (DSMP). IEEE, 2016.

[7] P. Ramos, N. Santos, and R. Rebelo. "Performance of state space and ARIMA models for consumer retail sales forecasting." Robotics and computer-integrated manufacturing, vol. 34, 2015, pp. 151-163.

[8] T. C. Lin, M. Pourahmadi, and A. Schick. "Regression models with time series errors." Journal of time series analysis, vol. 20.4, 1999, pp. 425-433.

[9] C. I. Permatasari, W. Sutopo, and M. Hisjam. "Sales forecasting newspaper with ARIMA: A case study." AIP Conference Proceedings. Vol. 1931. No. 1. AIP Publishing LLC, 2018.

[10] T. Fang, and R. Lahdelma. "Evaluation of a multiple linear regression model and SARIMA model in forecasting heat demand for district heating system." Applied energy, vol. 179, 2016, pp. 544-552.

[11] James, Gareth, et al. An introduction to statistical learning. Vol. 112. New York: springer, 2013.