An ARIMA-based study of pricing and replenishment decisions for vegetable commodities

Zhiyue Wang*

School of engineering, China University of Petroleum, Beijing, Karamay, China

*Corresponding author: catboy0068@gmail.com

Abstract. Fresh produce superstores face many challenges as an important part of the livelihood industry. The short freshness period and variable quality make their daily sales efficiency crucial. A simple, efficient and applicable decision model for supermarkets is beneficial to further promote the healthy and sustainable development of the fresh food market and the overall livelihood industry. In order to predict the daily replenishment and pricing strategy of each vegetable category in the coming week, the key lies in analyzing the supermarket's historical data. Since the sales volume shows strong cyclical fluctuations, the time series model ARIMA is chosen for this purpose to observe the model training and predict the pricing of different categories of products in the next 7 days. Due to the cost-plus pricing mechanism of the superstore, after processing the data, we selected simple linear regression model, multiple linear regression model and nonlinear regression model to fit the data, and opted for multiple linear regression to get the sales volume on the cost-plus pricing of the fitting formula, taking into account the discount rate of the different types of vegetables, we corrected the sales volume of the six types of vegetables, and ultimately obtained the superstore's daily replenishment and pricing strategy for the next seven days.

Keywords: ARIMA; multiple linear regression; vegetable commodities.

1. Introduction

Fresh food superstores, as the core of the livelihood industry, have numerous operational challenges. The short freshness period and volatile character require superstores to maintain an efficient daily sales and stocking rhythm. Meanwhile, replenishment decisions are complicated by the fact that superstores' decisions are often made during nighttime stocking hours, when information about varieties and categories of vegetables and prices is not always clear. At the same time, merchants need to weigh "cost-plus pricing" strategies and choose the most attractive mix of products within a limited space to meet the changing needs of consumers. Gross daily profit and cost of sales are the primary considerations for fresh food superstores, which often offer discounts on damaged or poor-quality items to boost sales. However, due to the complex historical sales data of unknown product categories and prices, merchants are faced with the problem of how to make appropriate replenishment decisions, the traditional "trial and error" method is very costly, and it is difficult to maintain the long-term operation of the livelihood of the industry [1].

Therefore, fresh food superstores need to consider a variety of factors in the business process to ensure that the sales strategy not only meets the needs of consumers, but also ensures the profitability of the business, in order to realize the healthy and high-quality development of the fresh food market and the overall livelihood of the industry [2].

2. Supermarkets' daily replenishment and pricing strategies for the coming week

First, the historical sales data were analyzed, and the relationship between the total sales volume and the unit price of vegetable categories was fitted by establishing a regression model. For judging the goodness of the model fitting, we adopted a total of three models, namely, simple linear regression model, polynomial regression model and nonlinear regression model, which were selected on the basis of their merits, respectively.

At the same time, we adopted the time series analysis model ARIMA to optimally forecast the prices of different categories of vegetables for the next seven days. The fitting function of sales volume and unit price was utilized to predict the unit price and hence the sales volume of the product for the next 7 days. At the same time, the discount rate of different individual products is used to get the average discount rate of different categories of vegetables, and the daily replenishment is obtained by using the number of discounts and the number of sales to get the daily replenishment, so that the daily replenishment and pricing strategy for the week of the superstore for the next 7 days is obtained.

2.1. Fit models

2.1.1 One-dimensional linear regression model

Suppose there is such a relationship between x and y:

$$\widehat{y_i} = ax_i + b \tag{1}$$

Where $\widehat{y_i}$ is the predicted value and the difference between the predicted value and the actual value, is the residual, i.e.:

$$\text{error} = y - \hat{y} \tag{2}$$

The sum of squares of the residuals is:

$$SSE = \sum_{i=1}^{n} (y_i - \widehat{y_i})^2 = \sum_{i=1}^{n} [y_i - (ax_i + b)]^2 \tag{3}$$

For the evaluation of the model, we want to get a prediction that is as small as possible and close to the actual value, that is, the smaller the residuals, the better. When we get a matrix A[a,b] with the smallest sum of squared residuals, we can consider that we have found the simple linear regression model with the best prediction to some extent.

2.1.2 Multiple linear regression model

The multinomial ground regression model is as follows:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_m x_i^m + \varepsilon_i (i = 1,2, \ldots, n) \tag{4}$$

It can be represented by the design matrix X, the response vector $\vec{y}$, the vector parameter $\vec{\beta}$, and the random error vector $\vec{\varepsilon}$. X and $\vec{y}$ in row i are the values of x and y for the ith data sample. The model can then be written as a system of linear equations.

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_1^2 & \ldots & x_1^m \\ 1 & x_2 & x_2^2 & \ldots & x_2^m \\ 1 & x_3 & x_3^2 & \ldots & x_3^m \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \ldots & x_n^m \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{bmatrix} \tag{5}$$

When using a pure matrix representation, write it as:

$$\vec{y} = X\vec{\beta} + \vec{\varepsilon} \tag{6}$$

The vector of estimated polynomial regression coefficients is:

$$\hat{\vec{\beta}} = (X^\top X)^{-1} X^\top \vec{y} \tag{7}$$

Assuming that $m < n$ is necessary for the matrix to be invertible, then since X is a van der Merwe matrix, the invertibility condition is guaranteed to hold if all x_i values are different, which is the only least squares solution.

Significance test of linear relationship: The F-test can be used to test the significance of linear relationship for multiple linear regression equations. For multiple linear regression equations, the number of independent variables n=k, when the F value is larger indicates that the linear relationship is more significant, and vice versa the less significant.

Significance test of regression parameters: t-test can be used to test the significance of regression parameters of multiple linear regression equation. For the multiple linear regression equation, the number of independent variables is n, respectively, n regression parameters t-test, respectively, test whether each regression parameter has a significant effect on the regression equation.

2.1.3 Nonlinear regression model

Nonlinear regression is a form of regression analysis in which the observed data are modeled by a function that is a nonlinear combination of model parameters and depends on one or more independent variables. The data are fitted by means of successive approximations [3].

The underlying assumption of this process is that the model can be approximated by a linear function, i.e., a first-order Taylor series:

$$f(x_i, \beta) \approx f(x_i, 0) + \sum_j J_{ij}\beta_j \tag{8}$$

Where $J_{ij} = \frac{\partial f(x_i,\beta)}{\partial \beta_j}$, the resulting least squares estimator is given by the following equation.

$$\hat{\beta} \approx (J^T J)^{-1} J^T y \tag{9}$$

Calculate the nonlinear regression statistic and use it as a linear regression statistic, but use J instead of X in the equation. linear approximation introduces bias into the statistic. Therefore, more care than usual is needed in interpreting statistics obtained from nonlinear models.

By comparing the correlation coefficients of the three regression models for each category of vegetables: one-way linear regression, polynomial regression, and nonlinear regression, we can obtain the optimal regression model for each category of vegetables (as shown in Table 1) in order to improve the accuracy of the predictions.

Through the above analysis, we can get the optimal regression model for each category of vegetables. This will provide reliable sales prediction results for superstores, which will help in supply chain management and pricing decisions. Since different categories of vegetables have different sales characteristics, the optimal regression model can be used to predict their sales trends more accurately. Supermarkets can make personalized forecasts and decisions based on the optimal models for different categories to improve operational efficiency and market competitiveness.

Table 1. Regression equations for different categories

	Regression equation
Philodendron	Y=-2.29901*X^3+42.3073*X^2-258.70843*X^1+691.37341
Cauliflower	Y=-0.00274*X^3+-0.53358*X^2+8.91711*X^1+9.94957
Aquatic rhizomes	Y=-0.08896*X^3 + 2.73638*X^2-31.69253*X^1 + 164.44108
Eggplant	Y=0.11516*X^3 + -2.93300*X^2 + 21.37573*X^1 + -19.86860
Capsicum	Y=0.02319*X^3 + 0.32323*X^2-17.04740*X^1+176.90556
Edible fungi	Y=0.05951*X^3-2.23532*X^2+ 21.40945*X^1 + 10.66161

2.2. Predictive model

In this paper, a time series model is used to forecast different categories of vegetables for the next 7 days. The main principles of time series modeling are as follows. Time series modeling is a specialized approach to processing and forecasting data over time. This analysis examines the patterns, trends, and periodicity of the data, as well as other key factors affecting sales such as trends, seasonality, cyclicality, and irregular components. The ARIMA model predicts the future based on historical information about the data and assumes that the data at a point in time is not only influenced by its history but also by chance events. The model further assumes that the data fluctuates around a broad trend that is likely to change. The ARMA model belongs to one of the time series analyses [4]. Mathematical expressions for AR and MA models:

$$AR: Y_t = c + \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \cdots + \varphi_p Y_{t-p} + \xi_t$$

$$MA: Y_t = \mu + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q} \tag{10}$$

The ARMA(p,q) model contains p autoregressive terms and q moving average terms, and the ARMA(p,q) model can be expressed as when the difference is not considered:

$$X_t = c + \varepsilon_t + \sum_{i=1}^{p} \varphi_i X_{t-i} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} \quad (11)$$

The equation of the ARIMA model when considering differencing can be expressed as:

$$Y_t = c + \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \cdots + \varphi_p Y_{t-p} + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q} + \epsilon_t \quad (12)$$

In this equation: Y_t is the time series data we are considering. φ_1 through φ_p are the parameters of the AR model, θ_1 through θ_q are the parameters of the MA model, ϵ_t is the error term, and c is a constant term. p and q are the autoregressive and moving average orders of the model; and X_t is a smooth, normal, zero-mean time series.

The first step is to test whether the time series is smooth or not. Through observation, we find that only aquatic root vegetables have a significance P-value of 0.139 when the difference is divided into 0th order, which is not significant and therefore the original hypothesis cannot be rejected, i.e., the series is an unsteady time series. Whereas, the significance P-value of the other categories of vegetable series at difference of order 0, 1 and 2 is less than 0.05, which is significant and hence the original hypothesis can be rejected, i.e., these series are smooth time series.

The results of these analyses provide important reference information for the superstores. For most categories of vegetables are stable in time, superstores can make operational decisions with more confidence.

By comparing the predicted and true selling prices, we find that the predicted and true values are very close in most cases. This shows that our prediction model has a good ability to predict the sales trend and price change of these vegetable categories. The predicted pricing of each category of vegetables for the next seven days can also be obtained.

2.3. Fitting the Call Prediction Model

By using the time series model ARIMA, the projected pricing for each type of product for the next 7 days is predicted, and the optimal fitting model is selected for different categories of products, which is brought into the projected pricing to arrive at the daily sales volume for different categories.

Considering the problem of the attrition rate of the goods, so the predicted daily replenishment = sales * (1 + attrition rate); so far, we get the daily replenishment and pricing strategy for the coming week as shown in Table 2.

Table 2. Daily replenishment and pricing for different categories for the coming week

Category	Strategic decision	1	2	3	4	5	6	7
Philodendron	Daily replenishment	44.22	46.04	47.55	48.87	49.94	50.80	51.45
	Pricing	11.26	10.85	10.46	10.06	9.67	9.28	8.89
Cauliflower	Daily replenishment	186.60	188.91	191.66	195.28	199.68	204.62	210.76
	Pricing	4.84	4.67	4.51	4.34	4.17	4.01	3.84
Aquatic rhizomes	Daily replenishment	2.51	3.64	7.77	8.18	12.52	16.35	19.75
	Pricing	16.01	15.61	15.16	14.70	14.24	13.77	13.31
Eggplant	Daily replenishment	21.09	19.99	18.90	17.86	16.91	16.06	15.33
	Pricing	8.77	9.06	9.37	9.67	9.98	10.29	10.59
Capsicum	Daily replenishment	85.96	86.82	88.08	89.47	90.92	92.42	93.95
	Pricing	7.28	7.19	7.06	6.92	6.78	6.64	6.49
Edible fungi	Daily replenishment	33.17	29.96	25.41	21.81	20.05	20.66	24.08
	Pricing	14.82	15.37	16.28	17.29	18.31	19.35	20.38

3. Summary

In this paper, in the fitting model, we consider the wear and tear of the vegetable merchant and the discounted goods and finally adopt the optimal fitting model - the nonlinear regression model, the decision boundary in the training process is easy to practice and understand the advantages.

At the same time, this paper takes the time series model to analyze the historical data while using the time series model to recognize the continuity of the development of things, you can use the past data to speculate on the development trend of things, the time series model has a simple and easy to use, easy to grasp, can make full use of the data of the original time series, the calculation speed is fast, the ability to dynamically determine the parameters of the model, and the accuracy of the advantages of a better. be integrated with multiple time series models (such as ARIMA, Prophet, etc.) or integrated w

The time series analysis can ith the prediction results of multiple machine learning models (such as Random Forest, XG-Boost, etc.). This can make full use of the advantages of each model and improve the accuracy and stability of the overall prediction.

Fusion of data from other related fields can also be considered. For example, weather data, population movement data, competitor data, etc. Fusing and analyzing these data with the superstore's sales data can reveal some hidden correlations and influencing factors, further improving prediction accuracy [5].

References

[1] Lou Pengkui. Research on Service Marketing Mix Strategy of Henan YF Superstore [D]. Henan University of Finance and Economics and Law, 2023.

[2] PAN Xiaofei, XIE Zhiheng, WANG Shuyun. Optimization decision of preservation efforts and pricing of fresh food superstores considering loss aversion [J]. Highway Transportation Science and Technology, 2022, 39(06): 177-185+190.

[3] Zhou Ji-Hsiang. Practical regression analysis [M]. Shanghai Science and Technology Press, 1990.

[4] Gao L. ARIMA-based demand forecasting model for small-lot material production [J]. Modern Information Technology, 2023, 7(15): 97-101.

[5] Shi Falei. Analysis of operating profit influencing factors of Carrefour supermarket chain [D]. Taiyuan University of Technology, 2018.