Research on Vegetable Product Pricing and Replenishment Decision Based on Linear Regression and ARIMA Model

Yu Chen*, Ziyan Hua, Yiwa Fu
School of Economics, Lanzhou University, Lanzhou, Gansu, 730000, China
* Corresponding Author Email: 13813904247@163.com

Abstract. With the rise of fresh food supermarkets, how to automatically price and restock vegetable products has become a major challenge for many businesses. To develop a restocking and pricing plan for vegetable products in supermarkets, we first calculate the cost profit margin, standardize the profit margin and total sales volume data, and then incorporate vegetable categories into a linear regression model using dummy variables. We found a significant negative correlation between profit margin and total sales volume, and obtained the product pricing strategy using the linear equation of profit margin. After conducting ADF testing, the number of autoregressive terms and the average number of sliding terms of the Time Series (ARIMA) model were determined based on the test set results, and the model was used to predict the total daily replenishment of each vegetable category in the next week. This study has reference value for vegetable restocking and pricing decisions.

Keywords: ARIMA Model, Linear Regression, Vegetable Restocking.

1. Introduction

The recent rise of fresh food superstores has placed new demands on product freshness, which has challenged the replenishment and pricing decisions for vegetable products [1]. Studies have considered commodity replenishment and pricing decisions under product upgrades [2], indefinite sales [3], and varying rates of product deterioration [4]. We established the commodity pricing strategy through a linear regression model, based on which we conducted an ADF test, which was passed with the help of the time series method in order to solve the prediction function of the sales volume of individual categories and predict the total amount of replenishment in a single day. Our study has reference value and significance for the replenishment and pricing decision of vegetable products. The data used in the article comes from the question c of the 2023 Chinese National Student Mathematical Modeling Contest.

2. Research on Vegetable Pricing Strategy

2.1. Construction and Improvement of Linear Regression Model

We analyze the pricing strategy based on the cost plus pricing method, which sets the price of each vegetable item based on its cost price and a certain proportion of profit [5]. Therefore, cost profit margin is the core explanatory variable of cost plus pricing strategy. The following will analyze the relationship between total sales and profit margin.

Firstly, standardize the profit margin data and total sales volume data to enhance their consistency, reliability, and comparability.

\[ y_i = \frac{Y_i - \bar{Y}}{\sigma_i}, \quad r_i = \frac{R_i - \bar{R}}{\delta_i} \quad (1) \]

Preliminary establishment of a linear regression model.

\[ y_i = \beta_0 + \alpha r_i \quad (2) \]

Considering that the total sales volume models for different categories of vegetables may have different intercepts, equation (2) directly performs linear regression analysis on the total sales volume and profit margin, and the results may differ significantly from the actual situation. Therefore, this
article incorporates category as a categorical variable into the model. Due to a total of six categories, five dummy variables need to be set [6], defined as follows.

\[
D_1 = \begin{cases} 
1, & \text{Flower and leaf vegetable} \\
0, & \text{Other} 
\end{cases} 
\quad 
D_2 = \begin{cases} 
1, & \text{Flourescent vegetable} \\
0, & \text{Other} 
\end{cases} 
\quad 
D_3 = \begin{cases} 
1, & \text{Aquatic rhizome vegetable} \\
0, & \text{Other} 
\end{cases} 
\quad 
D_4 = \begin{cases} 
1, & \text{Solanaceae vegetable} \\
0, & \text{Other} 
\end{cases} 
\quad 
D_5 = \begin{cases} 
1, & \text{Chili vegetable} \\
0, & \text{Other} 
\end{cases} 
\]  

(3)

This introduction method is equivalent to using different intercept terms in different vegetable categories, and the improved model is as follows.

\[
y_i = \beta_0 + \alpha r_i + \beta_1 D_1 + \beta_2 D_2 + \beta_3 D_3 + \beta_4 D_4 + \beta_5 D_5 
\]

(4)

2.2. Model Regression Results

We used Model (3) to conduct regression analysis on the relevant data provided by the official Chinese National Student Mathematical Modeling Contest. The results of the model regression are shown in Table 1.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>T-value</th>
<th>P-value</th>
<th>VIF</th>
<th>R²</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-0.231</td>
<td>-1.909</td>
<td>0.058*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(r_i)</td>
<td>-0.113</td>
<td>-1.769</td>
<td>0.078*</td>
<td>1.017</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(D_1)</td>
<td>0.105</td>
<td>1.391</td>
<td>0.166</td>
<td>1.434</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(D_2)</td>
<td>0.236</td>
<td>3.629</td>
<td>0.000***</td>
<td>1.056</td>
<td>0.074</td>
<td>F=3.063 P=0.007***</td>
</tr>
<tr>
<td>(D_3)</td>
<td>0.062</td>
<td>0.899</td>
<td>0.370</td>
<td>1.174</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(D_4)</td>
<td>0.068</td>
<td>1.023</td>
<td>0.307</td>
<td>1.09</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(D_5)</td>
<td>0.091</td>
<td>1.252</td>
<td>0.212</td>
<td>1.326</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: ***, **, * respectively represent significance levels of 1%, 5%, and 10%, which are the same as the tables appearing later.

The significance P-value of the F-test is 0.007, showing significance at the 1% level, rejecting the original assumption that the regression coefficient is 0, so the model basically meets the requirements. And the VIF values are all less than 10, indicating that the multicollinearity problem is effectively avoided. It can be found that there is a negative correlation between profit margin and total sales volume at a significance level of 10%.

The formula obtained by linear regression is:

\[
y_i = -0.231 - 0.113\alpha r_i + 0.105D_1 + 0.236D_2 + 0.062D_3 + 0.068D_4 + 0.091D_5 
\]

(5)

Due to the fact that the profit margin is the ratio of the difference between the average pricing and average cost of each product and the average cost, the average cost of the product cannot be controlled by the quotient. Therefore, in the case where the wholesale price has been determined by the supplier, supermarkets can control pricing to adjust profit margins and achieve maximum profit. According to the above model, sales volume is a negative correlation function of profit margin, and total profit=sales volume×profit margin×cost. If a linear equation with profit margin is used to replace sales volume, then \(y = a\alpha r_i^2 + b\alpha r_i + c, \ a < 0\), the function image opens downwards, so there is a maximum value. The supermarket can determine the price of a single product based on this model to maximize revenue.
3. Research on Vegetable Replenishment Strategy

3.1. Construction and Improvement of ARIMA Model

Considering the dependence of vegetable sales volume on time series and the interference of random fluctuations, the ARIMA model is selected in this article. A time series is a set of variables sorted by time, typically predicting a potential process at a given sampling rate over equally spaced time periods. Time series data essentially reflects the trend of a certain variable or variables changing over time [7]. This model includes AR model and MA model. The AR model is used to describe the relationship between present value and lagged value, and can fully utilize historical data to predict future values. The MA model uses a linear combination of residual terms from past data to predict future residual terms [8]. The ARIMA model can be written as.

\[ (1 - \Phi(B) \cdot \theta(B)) \cdot x_t = \epsilon_t \]

Before establishing a time series model, it is necessary to verify whether the time series of sales volume for each vegetable category meets the requirements of stationarity [9]. We selected a difference order of 1 for ADF testing, and the results are shown in the Table 2.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Difference order</th>
<th>T-value</th>
<th>P-value</th>
<th>AIC</th>
<th>Critical value 1%</th>
<th>Critical value 5%</th>
<th>Critical value 10%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flower and leaf</td>
<td>1</td>
<td>-13.959</td>
<td>0.000***</td>
<td>10359.643</td>
<td>-3.437</td>
<td>-2.864</td>
<td>-2.568</td>
</tr>
<tr>
<td>Florescent</td>
<td>1</td>
<td>-10.515</td>
<td>0.000***</td>
<td>8097.909</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aquatic rhizome</td>
<td>1</td>
<td>-8.745</td>
<td>0.000***</td>
<td>8270.921</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Solanaceae</td>
<td>1</td>
<td>-9.85</td>
<td>0.000***</td>
<td>6666.257</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chili</td>
<td>1</td>
<td>-8.6</td>
<td>0.000***</td>
<td>9109.864</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Edible fungi</td>
<td>1</td>
<td>-8.743</td>
<td>0.000***</td>
<td>9122.466</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

From the table, it can be seen that through the ADF test, the P values corresponding to the time series of sales volume for the six vegetable categories are all close to 0, indicating that at a significance level of 1%, the time series of sales volume for each vegetable category are stable, and ARIMA models can be established.

To further determine the number of autoregressive terms and the average number of sliding terms in the ARIMA model. Divide the sales volume data into an analysis set and a testing set. We conducted time series analysis on the sales volume of each vegetable category in the analysis set with an autoregressive coefficient of p=2 and a sliding term coefficient of q=3; And time series analysis with autoregressive coefficient p=3 and sliding term coefficient q=3. On this basis, we use these two models to predict the test set separately, compare their accuracy, and select the ARIMA model with higher prediction accuracy [10].

Finally, the test results of two time series models are as shown in Table 3.

<table>
<thead>
<tr>
<th>Category</th>
<th>Flower and leaf</th>
<th>Florescent</th>
<th>Aquatic rhizome</th>
<th>Solanaceae</th>
<th>Chili</th>
<th>Edible fungi</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSE(P=2)</td>
<td>121.2019</td>
<td>62.73284</td>
<td>98.94197</td>
<td>93.49547</td>
<td>244.663</td>
<td>110.6151</td>
</tr>
<tr>
<td>MSE(P=3)</td>
<td>1821.202</td>
<td>279.4827</td>
<td>575.5672</td>
<td>441.9646</td>
<td>1603.487</td>
<td>320.0182</td>
</tr>
</tbody>
</table>

We can find that the MSE predicted for the sales volume of each vegetable category in the test set at p=3 is smaller than the MSE predicted for the sales volume of each vegetable category in the test set at p=2. It can be concluded that the ARIMA model has better prediction performance at p=3, so the autoregressive term number p=3 was chosen.
3.2. Model Regression Results

Based on the above comparative analysis, we selected the ARIMA model with a regression number of \( p=3 \), and obtained the ARIMA model equation and time series diagram for the sales volume of each vegetable category as follows.

3.2.1 Flower and Leaf Vegetable Sales Volume

\[
y(t) = -0.035 + 1.577 y(t-1) - 1.398 y(t-2) + 0.338 y(t-3) - 2.066 \varepsilon(t-1) + 1.993 \varepsilon(t-2) - 0.768 \varepsilon(t-3)
\] (7)

The time series fitting diagram of flower and leaf vegetables is shown in Figure 1.

![Figure 1. Time series fitting diagram of flower and leaf vegetables (Unit: kg)](image)

3.2.2 Florescent Vegetable Sales Volume

\[
y(t) = -0.027 + 1.565 y(t-1) - 1.382 y(t-2) + 0.326 y(t-3) - 2.144 \varepsilon(t-1) + 2.052 \varepsilon(t-2) - 0.839 \varepsilon(t-3)
\] (8)

The time series fitting diagram of diagram of florescent vegetables is shown in Figure 2.

![Figure 2. Time series fitting diagram of florescent vegetables (Unit: kg)](image)
3.2.3 Aquatic Rhizome Vegetable Sales Volume

\[ y(t) = 0.008 + 1.391y(t-1) - 1.161y(t-2) + 0.156y(t-3) - 1.998\varepsilon(t-1) + 1.839\varepsilon(t-2) - 0.704\varepsilon(t-3) \]  

(9)

The time series fitting diagram of aquatic rhizome vegetables is shown in Figure 3.

![Figure 3. Time series fitting diagram of aquatic rhizome vegetables (Unit: kg)](image)

3.2.4 Solanaceae Vegetable Sales Volume

\[ y(t) = -0.009 + 1.418y(t-1) - 1.203y(t-2) + 0.175y(t-3) - 2.003\varepsilon(t-1) + 1.864\varepsilon(t-2) - 0.711\varepsilon(t-3) \]  

(10)

The time series fitting diagram of solanaceae vegetables is shown in Figure 4.

![Figure 4. Time series fitting diagram of solanaceae vegetables (Unit: kg)](image)

3.2.5 Chili Vegetable Sales Volume

\[ y(t) = 0.031 + 0.057y(t-1) + 0.278y(t-2) - 0.384y(t-3) - 0.528\varepsilon(t-1) - 0.572\varepsilon(t-2) + 0.372\varepsilon(t-3) \]  

(11)

The time series fitting diagram of chili vegetables is shown in Figure 5.
3.2.6 Edible Fungi Vegetable Sales Volume

\[ y(t) = 0.013 + 1.594y(t - 1) - 1.409y(t - 2) + 0.341y(t - 3) - 2.159\varepsilon(t - 1) + 2.067\varepsilon(t - 2) - 0.828\varepsilon(t - 3) \]

The time series fitting diagram of edible fungi vegetables is shown in Figure 6.

Based on the six time series diagrams mentioned above, we can find that the ARIMA model for sales volume of each vegetable category fits well. Based on this, we can use the above model to predict the total daily replenishment amount of each vegetable category in the next week, and obtain the following Table 4.
### Table 4. Forecast of daily replenishment volume for the next week (Unit: kg)

<table>
<thead>
<tr>
<th>Date</th>
<th>Flower and leaf</th>
<th>Florescent</th>
<th>Aquatic rhizome</th>
<th>Solanaceae</th>
<th>Chili</th>
<th>Edible fungi</th>
</tr>
</thead>
<tbody>
<tr>
<td>2023/7/1</td>
<td>150.1155</td>
<td>23.46342</td>
<td>18.94723</td>
<td>27.24441</td>
<td>96.03845</td>
<td>52.75308</td>
</tr>
<tr>
<td>2023/7/2</td>
<td>147.2591</td>
<td>20.46147</td>
<td>17.28075</td>
<td>26.12772</td>
<td>96.01595</td>
<td>58.05215</td>
</tr>
<tr>
<td>2023/7/3</td>
<td>132.8004</td>
<td>17.4789</td>
<td>16.14583</td>
<td>21.29584</td>
<td>83.64385</td>
<td>55.95641</td>
</tr>
<tr>
<td>2023/7/4</td>
<td>120.6598</td>
<td>17.43422</td>
<td>16.43489</td>
<td>16.25509</td>
<td>71.61557</td>
<td>49.67185</td>
</tr>
<tr>
<td>2023/7/5</td>
<td>120.7254</td>
<td>15.36115</td>
<td>17.89913</td>
<td>14.71136</td>
<td>68.63478</td>
<td>44.41698</td>
</tr>
<tr>
<td>2023/7/6</td>
<td>132.8627</td>
<td>17.08465</td>
<td>19.42771</td>
<td>17.72168</td>
<td>76.77211</td>
<td>44.17373</td>
</tr>
<tr>
<td>2023/7/7</td>
<td>147.7713</td>
<td>19.20193</td>
<td>19.90521</td>
<td>22.94688</td>
<td>89.77887</td>
<td>49.04236</td>
</tr>
</tbody>
</table>

### 4. Conclusions

This research focuses on developing a comprehensive pricing and replenishment strategy for vegetable products in fresh food supermarkets. The study employs a combination of Linear Regression and ARIMA models to address the challenges associated with automatic pricing and restocking in the context of the rising trend of fresh food supermarkets.

The Linear Regression model incorporates dummy variables for different vegetable categories, revealing a significant negative correlation between profit margin and total sales volume. This correlation is used to formulate a pricing strategy, allowing supermarkets to adjust prices for individual products to maximize revenue.

The ARIMA model, chosen due to its effectiveness in handling time series data, is applied to predict the total daily replenishment of each vegetable category for the upcoming week. ADF testing ensures the stability of the time series data, and the comparison of two ARIMA models determines the optimal autoregressive term number.

The research concludes with a comprehensive forecasting table, providing insights into the daily replenishment volume for each vegetable category in the next week. The findings of this study offer valuable reference points for supermarkets in making informed decisions regarding vegetable restocking and pricing.

### References


