Research on Vegetable Sales and Pricing Based on ARIMA Model and Regression Analysis

Wenshuo An [1, #, *], Shuqi Wang [2, #], Junhuan Li [2, #], Yingchu Liu [2, #]

[1] School of Pharmaceutical Sciences, Capital Medical University, Beijing, China, 100069

[2] School of Basic Medical Sciences, Capital Medical University, Beijing, China, 100069

* Corresponding Author Email: 13468311827@163.com

[#]These authors contributed equally.

Abstract. Vegetable products are perishable, so the supermarket's decision to replenish and price vegetables daily and sell defective products at a discount has a great impact on the profitability of the supermarket. This study is based on vegetable sales data. The aim is to analyze the relationship between total sales volume and cost markup pricing, study the seasonal changes in daily replenishment volume and its relationship with sales volume, determine the daily replenishment volume and pricing strategy for the next week, and explore the relationship between summer sales volume and average profit margin. Firstly, the sales data are classified by season and discounted or not, and the relationship between total sales and pricing is analyzed by replacing cost-plus pricing with average profit margin. Secondly, the daily replenishment of vegetables was correlated with the total sales, and the ARIMA time series model was used to predict the total daily replenishment in the coming week. Secondly, the daily replenishment volume of vegetables was correlated with the total sales volume, and the ARIMA time series model was used to predict the daily replenishment volume for the next week. Finally, a relationship model between sales volume and average profit margin was obtained through regression analysis, providing decision-makers with specific predictions on the total amount of daily vegetable restocking and pricing strategies for the next week, as well as a deeper understanding of the relationship between sales volume and average profit margin. Through the data analysis and modeling of this study, specific predictions on the daily replenishment volume and pricing strategy of vegetables for the next week have been provided to decision-makers, as well as a deep understanding of the relationship between sales volume and average profit margin. Through the data analysis and modeling of this study, the specific decision-making methods and promotion strategies of replenishment pricing were obtained, which provided a useful reference and guidance for the decision-making of vegetable supply and marketing.

Keywords: Vegetable sales, Data analysis, Pricing strategy.

1. Introduction

Supermarkets consider that vegetables are perishable products when selling raw, fresh, and vegetable products. Therefore, most vegetable products can only be sold on the same day and cannot be sold every other day. Moreover, the quality of vegetables may deteriorate over time on the same day, which affects sales effectiveness. Therefore, supermarkets will make purchasing decisions for the vegetables to be sold on the same day from 3:00 to 4:00 in the morning based on the supply and demand situation of various products in previous years, as well as factors such as seasons, weather, holidays, etc., without determining the specific items and purchase prices. In determining the selling price of vegetables, the "cost plus pricing" method is generally used, and supermarkets will offer discounts for goods with transportation losses and poor quality[1].The daily purchase volume and selling price of vegetables need to be determined by market demand analysis, which is related to customer demand and supplier supply. From the perspective of customers, sales volume is related to time. From the perspective of suppliers, the variety of supply is more abundant from April to October.Due to the limited sales space of supermarkets, a reasonable sales mix is particularly important for their profitability. In order to give a reasonable pricing and replenishment strategy, this paper uses a prediction algorithm based on vegetable sales data (data source http://cumcm.cnki.net),

firstly divides all sales into discounted and undiscounted, and then obtains the relationship between the total sales volume of six categories of vegetables and the average interest rate according to the four seasons, and uses regression analysis to draw a scatter plot to obtain the relationship between the total sales volume of each vegetable category and cost-plus pricing, and then uses the ARIMA time series model. The relationship between the total sales volume and the average profit margin of each category of vegetables in summer was obtained by using regression analysis, and finally the relationship between the daily replenishment volume and the total sales volume of each category of vegetables and the average profit margin was obtained by using the relationship between the daily replenishment volume and the total sales volume of each category of vegetables and the average profit margin from July 1 to 7, and the pricing strategy was obtained [2-4].

2. The Model establishment and solution

2.1. The relationship between the total sales volume of various vegetable categories and cost markup pricing

Through the perspective analysis of the data (data source http://cumcm.cnki.net), it can be found that the total sales volume of each vegetable category is greatly affected by factors such as time, season, and whether it is discounted. Therefore, this paper divides the data into two categories: discounted and undiscounted, and divides the time into four categories according to seasons: spring, summer, autumn and winter, so as to weaken the impact of irrelevant factors on the total sales, so as to better analyze the relationship between the total sales volume of each category and cost-plus pricing. Considering that the correspondence between the pricing and sales volume of each single product in different categories of vegetables is quite different, therefore, in order to reduce the impact of the difference on the category relationship, this paper replaces the cost-plus pricing with the average profit margin of each single product, so as to make the data more standard.

The data were classified and summarized to obtain the total sales volume and average profit margin of each category of vegetables in the four seasons of spring, summer, autumn and winter under the premise of discount or no discount. Import the data into Python, use pandas, statsmodels. api, matplotlib. pyplot libraries, establish a fitted linear regression model, obtain scatter plots and regression lines, and analyze them. The result is shown in Figure 1.

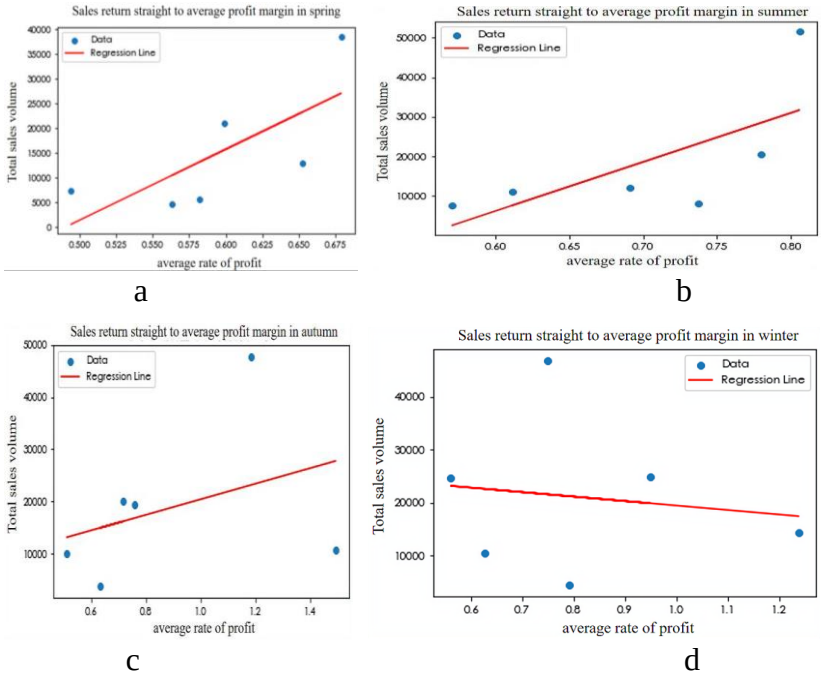


Figure 1. The relationship between the total sales volume and the average profit of each category in seasons without a discount

Under the premise of not discounting, according to the analysis of Figure 1, in spring, among the six categories of vegetables, mosaic is the category with the highest average profit margin, and because the total sales volume of mosaic and leaf is the highest, it is also the most profitable vegetable in spring among the six categories of vegetables, nightshade and aquatic rhizomes are categories with large average profit margin but small total sales volume, while peppers and edible fungi have larger total sales but lower average profit margins, and edible fungi have smaller total sales but higher average profit margins. The total sales volume and average profit margin of cauliflower were the lowest in the whole year, indicating that the cauliflower category was the least profitable in the spring throughout the year. In summer, the total sales volume and average profit margin of flowers and leaves are the highest for the whole year, indicating that flowers and leaves make the most profits in summer. In autumn, cauliflower, mosaic and edible mushrooms have the highest average profit margin among all categories throughout the year, and cauliflower belongs to the type of vegetables with high profit margins and low sales volume in autumn. In winter, edible fungi and flowering vegetables belong to the low profit and high sales type, while aquatic root and stem vegetables have a higher average profit margin, which may be related to the season. The relationship between total sales and average profit of each category in the four seasons of discount is shown in figure 2.

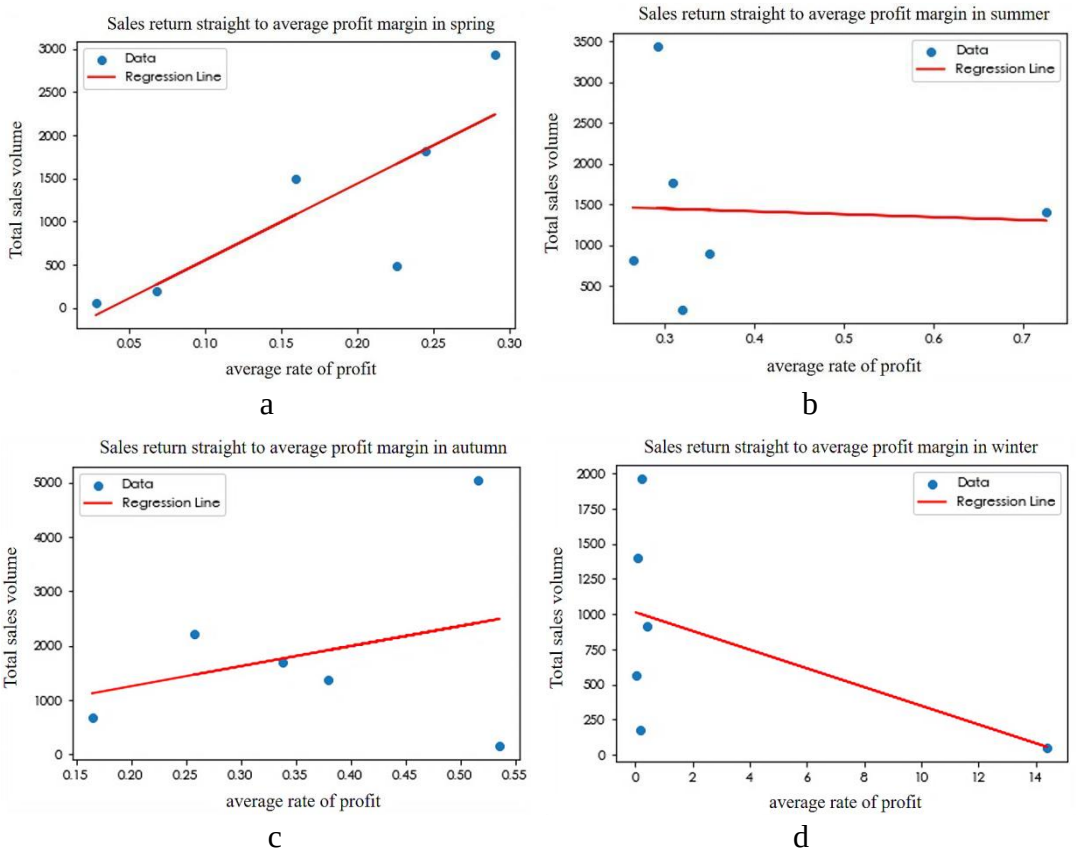


Figure 2. The relationship between total sales and average profit of each category in the four seasons of discount

In the case of discounting, the analysis of Figure 2 shows that the sales volume of nightshade and aquatic rhizomes in spring is very low, and the average profit margin is close to 0, indicating that in spring, these two types of vegetables are basically not profitable when they are sold at a discount after loss, and the profit margin of summer peppers is also very high when they are sold at a discount, and they can be profitable even if they are lost, and the loss of flowers and leaves is very large, and the average profit margin is low, and the loss and average profit margin of the remaining four types of vegetables are not high, and the average profit margin of mosaic and eggplant is higher in the autumn discounted goods. However, the loss of flowers and leaves is higher, the loss of eggplant is low, and in the winter discount goods, in addition to the average profit margin of eggplant is extremely high,

the remaining categories of vegetables are almost not profitable, therefore, this paper finds that the profit margin of eggplant is extremely high, and the purchase price of eggplant is only 0.01 yuan/kg in the purchase of a few days, which may be a record outlier, but this paper still considers a variety of other possible factors and analyze them. It may be due to the excessive inventory of eggplants from the supplier, which were supplied as a gift to supermarkets as a supplement to other vegetables. In order to record the supply data, they were sold and invoiced at an extremely low price.

2.2. Observing the seasonal changes in daily replenishment volume and its relationship with sales volume

Merchants do not know the specific items and purchase prices during the purchase transaction. The purchase prices of different types of vegetables and different items of the same category of vegetables vary greatly. Therefore, when determining the pricing strategy, this article cannot directly derive the specific selling price of a certain type of vegetable or a single item. In order to reduce the impact of the large difference in the purchase price of different single products, this paper simplifies the pricing strategy as the average profit margin to obtain a more standard analysis result [5-6].For the daily replenishment volume, considering the loss rate data of each single item, the possible loss caused by the transportation of vegetables will affect the sales volume of vegetables, therefore, this paper determines the relationship between the total daily replenishment and the total sales

$$N_X = N_B(1 - \alpha) \tag{1}$$

NX represents the total sales volume of vegetables, NB represents the daily replenishment volume, and α represents the average loss rate of various categories of vegetables.

Based on this relationship, this paper uses Excel data processing to obtain the change data of total daily replenishment over time.

After full observation and understanding of the data, this paper found that the harvest volume and average profit margin of each vegetable category show obvious cyclical and seasonal changes.We first used the grey prediction model [7] to depict the daily replenishment volume of six categories of vegetables (taking cauliflower as an example) in summer (Figure 3), which provides a better observation of data trends.

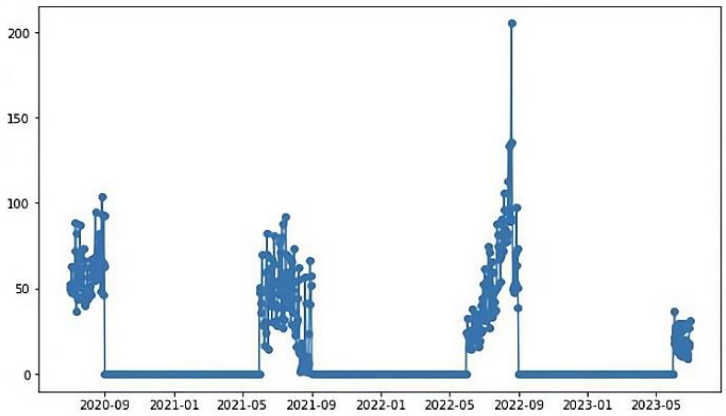


Figure 3. cauliflower gray forecast daily replenishment can be viewed

2.3. The relationship between the total sales of vegetables in summer and the average profit margin

The relationship between the total sales volume of vegetables and the average profit margin varies significantly with the seasons [8]. Therefore, in order to obtain the relationship between the total sales volume of each category and the average profit margin from July 1 to 7, 2023, we can first use regression analysis to obtain the relationship between the total sales volume of vegetables and the average profit margin in the summer, and then use the sales profit relationship obtained in the summer to predict the relationship between July 1 and 7.

Integrate the correspondence between the item numbers of the data and the categories, filter the dates in the integrated data, select sales records from June to August each year, sum up all sales records of each item within one day during the period of June to August, and then sum up the sales volume of various items under each category to obtain the total sales volume of each category for each day of June to August. Then, utilize the sales unit price and wholesale price to calculate the daily average profit margin of each item, and filter the sales dates to select sales records from June to August each year. Calculate the average profit margin of the items sold in each category within a day and obtain the average data. Using the pandas, nunpy, matplotlib. pyplot, and statsmodels. appi libraries in Python, import average data and perform regression analysis. Take cauliflower as an example, and get figure 4.

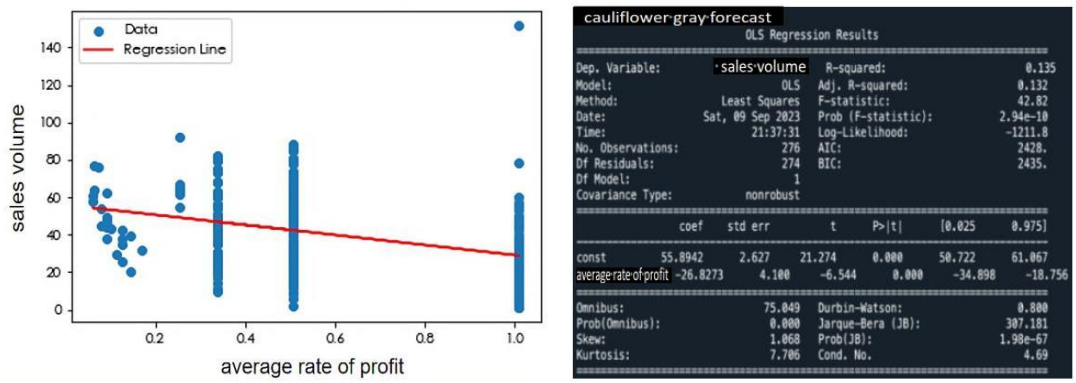


Figure 4. Summer cauliflower sales volume and average profit margin regression analysis

After regression analysis, we obtained the regression equation between sales volume and average profit margin

$$N_X = a_i P_i + b_i \tag{2}$$

Among them, ai represents the coefficient, bi represents the intercept, and Pi represents the average profit margin of various vegetable categories.

Substitute the Python output results into the above equation to obtain the regression equations for six types of vegetables (taking cauliflower as an example) as follows

Flower and vegetable category:

$$N_{x1} = -26.8273P_1 + 55.8942 \tag{3}$$

2.4. Using ARIMA time series model to predict daily replenishment volume

2.4.1 data processing

Corresponding the individual product loss rate and total sales volume to the vegetable number, classify and summarize them to obtain the total sales volume and individual product loss rate of each category of vegetables in summer (June August). Using the relationship between sales volume and daily replenishment volume, obtain the daily replenishment volume of each individual product on the corresponding date, and summarize them by category.

2.4.2 Model Introduction

We use ARIMA time series model [9], which is introduced below.

ARIMA (Differential Autoregressive Moving Average) is a statistical model commonly used for time series prediction and analysis. It can help us understand and predict features such as trends, seasonality, and periodicity in time series data. The following are the main components of the ARIMA model:

Differentiation: Differentiation is the first step in ARIMA models, used to make time series stationary. The meaning of a stationary time series is that the mean and variance remain constant over time, which contributes to the stability of the model. If the original time series is not stationary,

stationarity can be achieved through first-order differencing (subtracting the value of the previous time point) or higher-order differencing.

Autoregressive (AR): The AR part represents the correlation between the observed values at the current time point and those at several past time points. The AR (p) model uses observations from p past time points to predict the value at the current time point. The order p of the AR model is usually a pre-selected parameter.

Moving Average (MA): The MA part represents the correlation between the observed values at the current time point and the error terms at several past time points. The MA (q) model uses the error terms from q past time points to predict the value at the current time point. The order q of the MA model is usually a pre-selected parameter.

Integration: The integration part represents the number of differential operations performed to make the time series stationary, usually represented by a parameter called "d". If d=0, it means there is no need for differentiation; If d=1, it means performing first-order differencing; and so on.

Its mathematical expression is:

$$Y_t = c + \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \cdots + \varphi_p Y_{t-p} + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q} + \epsilon_t \quad (4)$$

The general steps for establishing an ARIMA model include:

Visualization: Firstly, visualize the features of the data by drawing time series plots and autocorrelation (ACF) and partial autocorrelation (PACF) plots to assist in the initial estimation of p, d, and q parameters.

Determine parameters: Determine the parameters of the ARIMA model based on visualization results and model evaluation methods (such as AIC, BIC, etc.).

Fitting model: Use determined parameters to fit the ARIMA model.

Model diagnosis: Diagnose the residuals of the model to ensure that they conform to the assumption of stationary white noise.

5)Prediction: Use a fitted ARIMA model to predict future time points.

The ARIMA model is an important tool in time series analysis, which can be used for various applications including economics, finance, weather forecasting, sales forecasting, and other fields. It performs well in many situations, but there are also some limitations, such as poor performance on nonlinear and non-stationary data.

2.4.3 Model application

After verification, the time series belongs to a stationary type and can be predicted using the ARIMA model.

Import the sales volume and average profit margin data into Python, use pandas, numpy, matplotlib. pyplot library, and ARIMA package to cycle and predict the cauliflower category, and obtain the predicted daily replenishment volume for the next seven days. (See Table 1 for specific results)

3. Results

3.1. Determine the daily replenishment quantity and pricing strategy for each vegetable category for the next week

For the daily replenishment total, we have obtained it through the observation of the grey prediction model and the specific prediction of the ARIMA model. For pricing strategy, we fully consider the impact of the loss rate of each individual product. Firstly, we map the daily replenishment volume to the sales volume data through its relationship with sales volume. Then, using the regression relationship between sales volume and average profit margin obtained earlier, we obtain the average profit margin of each category for the next seven days, which is the pricing strategy. Table 1 is obtained.

Table.1. Total daily replenishment volume and pricing strategy of cauliflower for July 1-7,2023

	Total daily replenishment in the next 7 days	Sales in the next 7 days	average rate of profit
July 1	25.0158	21.4780	1.2829
July 2	22.7645	19.5452	1.3549
July 3	21.9143	18.8152	1.3821
July 4	21.5932	18.5395	1.3924
July 5	21.4719	18.4353	1.3963
July 6	21.4261	18.3960	1.3978
July 7	21.4088	18.3811	1.3983

3.2. Comprehensive comparative analysis

In order to better observe the prediction results of the model and verify the summary of the seasonal and cyclical relationship between different categories of vegetables, this paper uses the ARIMA model again to predict the total sales of each category of vegetables (this paper is characterized by time and does not divide seasons), and the prediction results in the following table are obtained.

Table.2. Forecast sales of cauliflower in the next seven days

The first day in the future	The next day in the future	The next third day in the future	The fourth day in the future	The next fifth day in the future	The sixth day in the future	The seventh day in the future
19.3200	18.1253	17.9662	21.1620	22.3002	20.8008	19.7484

Through the analysis of tabular data, this paper finds that the sales volume of some categories of vegetables in the next day is greater than the total daily replenishment, and this paper assumes that vegetables will not be sold every other day, so the results of ARIMA time series prediction by season are slightly less reasonable, which reflects the importance of seasonal and cyclical factors in the vegetable supply and marketing problem [10].

4. Conclusions

The marketing strategy of daily purchase and pricing of vegetables by supermarkets has a significant impact on their profitability. In order to explore reasonable pricing and replenishment strategies, this paper obtains pricing and the relationship between pricing and total sales, the relationship between daily replenishment and total sales, and the relationship between sales and average profit margin by using ARIMA time series model and regression model based on the data provided by http://cumcm.cnki.net. The specific forecasting method of daily replenishment amount and pricing strategy in the coming week is obtained, and cauliflower is taken as an example to illustrate. From the above comparison, it can also be found that the prediction conclusion obtained in this paper is more realistic, the prediction is more accurate, and can be used in daily life, which has a strong reference value for the decision-making of supermarket purchase pricing. In the future, the contact of the model can be optimized according to the actual situation, and more influencing factors can be added to the model to make the prediction results more suitable for real life.

References

[1] Deng L. Study on Supermarket Replenishment and Pricing Based on Time Series Forecasting Algorithm and Nonlinear Multi-Objective Optimization Model [J]. Financial Engineering and Risk Management, 2024, 7(3):

[2] Nie Yuxuan. Research on automatic pricing and replenishment strategy of fresh commodities based on ARIMA prediction optimization model--taking vegetable commodities as an example [J]. Commercial Exhibition Economy, 2024, (05): 19-22.

[3] Ding Nan. Correlation analysis of vegetable price prediction based on time series[D]. Shanxi Agricultural University, 2022.

[4] Mao Lisa. Research on pricing strategy and production and marketing model of vegetable wholesale market from the perspective of supply chain [D]. Central South University of Forestry and Technology, 2022.

[5] Liu X, Li S. A study on vegetable pricing replenishment strategy based on seasonal time series [J]. The Frontiers of Society, Science and Technology, 2024, 6(4):

[6] Luo Qin. Analysis of vegetable price formation and influencing factors in Chongqing [J]. Rural Economy and Technology, 2021, 32(03): 169-172.

[7] Bo Z, Shuliang L, Wei M, et al. An improved gray prediction model for China's beef consumption forecasting. [J]. PloS one, 2019, 14(9): e0221333.

[8] Zhang Huiyuan. Analysis of vegetable price fluctuation and its causes in Beijing [J]. National Circulation Economy, 2021, (01): 21-23.

[9] Jatinder K, Singh K P, Sarbjit S. Autoregressive models in environmental forecasting time series: a theoretical and application review. [J]. Environmental science and pollution research international, 2023, 30(8): 19617-19641.

[10] Zhang Bi. Characteristics of vegetable price fluctuation in China: An empirical analysis based on PVAR model [J]. China Business Review, 2021, (1)