Analysis of Vegetable Price Fluctuation Law and Causes based on Lasso Regression Model

Weige Yu\textsuperscript{1,2,a}, Chunjiang Zhao\textsuperscript{2,3,b}, Huarui Wu\textsuperscript{2,3,4} and Cheng Peng\textsuperscript{2,3,4}

\textsuperscript{1} School of Information Engineering, Capital Normal University
\textsuperscript{2} National Engineering Research Center for Information Technology in Agriculture
\textsuperscript{3} Key Laboratory of Agri-informatics, Ministry of Agriculture
\textsuperscript{4} Beijing Engineering Research Center of Agricultural Internet of Things

\textsuperscript{a}yuweige888@126.com; \textsuperscript{b}zhaocj@nercita.org.cn

Abstract. In recent years, the sharp price fluctuation of vegetables has attracted wide attention. In order to explore the causes of price fluctuation, the price fluctuation law of cucumber from 2010 to 2017 was deeply explored based on the empirical analysis of the price fluctuation law of vegetables. Most of the existing literatures use linear regression models to analyze the influence degree of various factors, but there are problems of multicollinearity among multiple variables and small influence degree of some variables, which affect the fitting accuracy. Lasso regression model is introduced to model and solve cucumber price data, and the influencing factors of price with smaller correlation are eliminated to obtain the main factors and their correlation. Compared with the least square method, the multicollinearity determination condition value is only 19.66 and the fitting coefficient reaches 0.8448, which proves that lasso regression model is suitable for vegetable price cause analysis and has better performance than traditional methods. It also provides basis for further vegetable price prediction.

1. Introduction

China is the country with the largest vegetable production and consumption in the world, and the vegetable industry is developing rapidly. In recent years, the vegetable market price fluctuates frequently, which has a great impact on people's lives, especially on low-income people. Farmers' untimely response to the market results in loss of benefits. Under this background, this paper deeply explores the law of vegetable price fluctuation and show the relationship between vegetable and influencing factors, which can be used to guide farmers to plan their planting rationally. It is of great significance to increase the economic benefits of vegetable farmers, reduce the living costs of urban residents and stabilize the prices of vegetables in the market.

At present, many scholars have done a lot of research on the law of vegetable price fluctuation and its causes. Literature\cite{1} analyzes the characteristics and causes of the price fluctuation of vegetables in China since the new century and puts forward the idea of stabilizing prices. The article also draws the conclusion that the price change direction of vegetable consumption market is not completely consistent with the production price change direction. Literature\cite{2} uses seasonal adjustment model and covariance analysis to analyze the degree of influence of factors affecting vegetable prices on price fluctuation. Literature\cite{3} starts from the circulation system, deeply studies the characteristics and causes of changes in vegetable prices in China, further studies the industrial problems behind
vegetable prices, and clarifies the influence of circulation factors. Literature[4] analyzes the law and influencing factors of price fluctuation of various vegetables, establishes a linear regression model, and obtains the correlation of various factors. Literature[5] explores the changes of vegetable daily price data and forecasts them through generalized regression neural network, which opens up new ideas for short-term price forecasting. Literature[6] combines time series decomposition and H-P filtering method to deeply analyze the law and influencing factors of vegetable price fluctuation in Beijing from 2003 to 2014, and establishes a linear regression model to analyze the influencing factors. However, neither the literature nor the literature[4] consider the problem of over-fitting and high correlation of the linear regression model, resulting in low interpretability of the results.

The above documents have all analyzed the composition of influencing factors of vegetable price from the change rule of vegetable price, but they have not quantified the influence degree of various factors on vegetable price. In addition, the existing researches mostly adopt multiple linear regression models to model vegetable prices, which often have multiple collinearities and over-fitting phenomena in the fitting process. Based on this, after deeply analyzing the advantages and disadvantages of lasso regression method, a lasso regression model was established to analyze the influencing factors of vegetable prices. The model has the characteristics of linear regression model describing the relationship between independent variables and dependent variables. At the same time, it effectively avoids the defects of over-fitting and high correlation and eliminates the problem of multicollinearity.

2. Empirical Analysis of Vegetable Price Fluctuation Law

According to the cucumber price data from eight markets, including Beijing shunxin agricultural products wholesale market and Xinfadi agricultural products wholesale market from 2010 to 2017, the average exportable price trend chart is shown in figure. 1.

![Monthly price of cucumber from 2010 to 2017](image)

*Figure 1 Monthly Average Data*

From the price trend chart, it can be seen that cucumber price is highly correlated with time and fluctuates periodically and regularly every year. Vegetable prices are affected by the production period and consumers' holiday consumption habits and have obvious seasonal fluctuation characteristics. According to the theory of supply and demand in economics, when the supply decreases or the demand increases, the price will increase, while vice versa. Cucumber prefers warm environment but not cold. Summer is suitable for cucumber growth, and the yield increases greatly. In addition, there are many alternatives to choose from. The demand of the resident market decreases, and the price of cucumber falls to a low point. The data show that it conforms to the general law of supply and demand. On the other hand, the price increase during holidays is already a common phenomenon. It can be seen that the price of cucumbers will increase to varying degrees during major festivals such as Spring Festival. In addition to traditional festivals such as Spring Festival, Mid - Autumn Festival and National Day, holiday spending driven by the Internet such as "Double Eleven" is also occupying more and more important seats. With the combination of e-commerce and agricultural products
becoming more and more close, the successive traditional festivals and online holidays in the second half of the year continue to promote residents' demand for vegetables and drive the prices of vegetables to rise[6].

The average annual cucumber price data can be plotted into a line chart as shown in Figure 2.

![Figure 2 Annual Average Data](image)

It can be seen from the figure that the overall cucumber price change trend from 2010 to 2017 is fluctuating and rising. The reasons for the rise and fall of specific years and months are multiple and complex. According to the previous research experience[1-6], combined with the existing data, the possible factors affecting the price fluctuation of vegetables are classified into four categories, namely, supply factors, demand factors, circulation factors and comprehensive factors. Many influencing factors have different degrees of influence on prices, some of which are extremely small or even negligible. Some factors are easy to be ignored, but in fact they will have a greater impact on price fluctuations. Therefore, in the next step, lasso regression method will be used to model and analyze the relationship between influencing factors and prices, select the influencing factors with high correlation and eliminate the influencing factors with low correlation.

3. Methods

When we have data \((X, Y)\), where \(X = (x_{11}, x_{12}, \ldots, x_{1m})\) and \(Y\) are the values of independent and dependent variables respectively obtained in the I times experiment, we can determine the general linear regression model as follows:

\[
y = \beta_0 + \beta_1 x_1 + \ldots + \beta_m x_m + \varepsilon
\]  

(1)

Where \(m\) refers to the independent variable with \(m\) number, \(\beta_0, \beta_1, \ldots, \beta_m\) is the regression coefficient of the equation, and \(\varepsilon\) is the random error. The general solution of linear regression equation is the least square method, i.e. estimation of parameters using observed values.

\[
Q = \sum \varepsilon_k^2 = \sum \left( y_k - \left( \hat{\beta}_0 + \hat{\beta}_1 x_{1k} + \ldots + \hat{\beta}_m x_{mk} \right) \right)^2
\]  

(2)

The least square method is used to solve the problem. The method is usually to minimize the sum of squares of residuals and the \(Q\) value. Only the partial derivatives of the above formula \(\hat{\beta}_0 \hat{\beta}_1 \ldots \hat{\beta}_m\) need to be calculated respectively, and the corresponding regression coefficients can be obtained.

Lasso regression model belongs to a kind of linear regression model[7]. The commonly used method of linear regression models is the least square method, but there are often problems when dealing with models with multicollinearity. There are mainly two defects[8]: first, the prediction accuracy is not high. When the least square method is used to solve the problem, if the linear
correlation of independent variables is high, the variance is often large. The second is that the model does not have good interpretability. For the case with more independent variables, the estimated value has a large deviation from the actual situation. At the same time, the estimation results by the least square method will be affected by the multicollinearity between variables. This method cannot eliminate the collinearity and cannot reduce the dimension[9].

Lasso (Least absolute shrinkage and selection operator; Tibshirani(1996)) regression method[10] is essentially a compressed estimation. First make \( z_1, z_2, ..., z_m \) the mean value 0, variance 1, dimensionless difference. Lasso estimation of the regression model is:

\[
(\hat{\alpha}, \hat{\beta}) = \arg \min_{(\alpha, \beta)} \sum_{i=1}^{n} (y_i - \alpha - \sum_{j=1}^{p} x_{ij} \beta_j)^2
\]

subject to \( \sum_{j=1}^{p} |\beta_j| \leq s \)  \( (3) \)

Among them, \( s \geq 0 \) is the harmonic parameter. For any \( S \) has the estimation of \( \hat{\alpha}, \hat{\beta} = \bar{y}, \) only the harmonic parameter \( S \) needs to be adjusted in order to reduce the overall regression coefficient. When the \( s^* = \sum_{j=1}^{p} |\beta_j|, s \leq s^0 \), the coefficients of some variables will be reduced, even close to 0 or equal to 0. These independent variables with little or no relation will be screened out, thus improving the accuracy and interpretability of the regression model. The Mallows Cp statistic[11] is generally used to solve the Lasso regression coefficient. This method can evaluate the advantages and disadvantages of the regression model. The method is defined as follows:

\[
C_p = \frac{SSE_p}{\hat{\sigma}^2} - n + 2p, \quad (4)
\]

\( n \) represents the total number of samples, \( p \) represents the number of independent variables in the subset regression model, \( SSE_p \) represents the sum of squares of residuals after regression of dependent variable \( Y \) by \( p \), and \( \hat{\sigma}^2 \) represents the prediction of the mean value of variances when all independent variables regression of dependent variable \( Y \). According to this, when the model[12] is obtained as \( C_p \) is the minimum value, the subset of variables that are best in the global range is obtained, and at the same time, the regression equation with the best effect is generated.

4. Analysis on the Causes of Vegetable Price Fluctuation Based on lasso Model

4.1 Establishment of lasso Regression Model

This paper selects the monthly price data of cucumbers in Beijing from 2010 to 2017 as the research object. Cucumber prices come from a total of 8962 price data on the website of AgSoSo (National Agricultural Science and Education Cloud Intelligent Search System) and the national agricultural products business information public service platform. The demand factor \( W \) (independent variable) and dependent variable \( Q \) that affect the price of cucumber are used as component residual plot[13]. The supply factor \( X \) (independent variable), dependent variable \( Q \), circulation factor \( Y \) (independent variable) and comprehensive factor \( Z \) (independent variable) which affect cucumber price are also made into component residual plot, which can analyze the linear correlation or nonlinear correlation between independent variable and dependent variable in regression model[13]. It is observed that the demand factor, supply factor, circulation factor and comprehensive factor all present linear relationship. The cucumber price fluctuation and influencing factor model can be designed as the following linear regression model:
Among them, Q represents the price of cucumber, W represents the demand factor affecting the price of cucumber, X represents the supply factor, Y represents the circulation factor, and Z represents the comprehensive factor. $\beta_1$ represents the regression coefficient of demand factor, $\delta_1$ represents the regression coefficient of supply factor, $\lambda_1$ represents the regression coefficient of circulation factor, $\xi_1$ represents the regression coefficient of comprehensive factor and $\epsilon_1$ represents random error $\epsilon_1 \sim N(0, \delta^2)$ of the established model.

Using the model (5) to model the acquired data, we can obtain the degree of influence of various factors on cucumber prices. The next section will show the difference between the general solution effect of using lasso regression method to solve the model and multiple linear regression model, proving that lasso regression model is effective in reducing the multicollinearity of data and removing the factors that have little correlation with cucumber price fluctuation.

4.2 lasso Regression Model Experiment and Analysis

In this paper, regression model (5) is used to model the experimental data. Lasso regression model was adopted to obtain the degree of influence of various factors on the fluctuation of cucumber price. Meanwhile, the experiment is compared with multiple linear regression model solved by the widely used least square method. The experimental program is programmed with R language.

Before the experiment, the data were processed into standardized data with a mean value of 0 and a variance of 1. The first two thirds of the data were taken as training sets to build the prediction model, and the last third was taken as test sets to evaluate the model performance. Firstly, a linear regression model is established to solve the problem.

The experimental results obtained from the solution are shown in Table 1. The coefficient in the table is greater than 0, indicating that this item has a positive effect on the price of cucumber. The larger the value, the greater the effect on the price of cucumber. If the coefficient in the table is less than 0, it indicates that this item has a negative effect on the price of cucumber. The smaller the value, the greater the negative effect on the price of cucumber. The smaller the standard deviation in the table, the better the effect.

| Variable                                    | Regression coefficient | Standard deviation |
|----------------------------------------------|------------------------|--------------------|
| Per capita disposable income
  inhabitant                                    | 0.0862                 | 0.2259             |
| Consumption level of residents
  Inflation level                                | 0.5265                 | 0.0563             |
| Sowing area of cucumber
  Cucumber yield                                 | -0.0085                | 0.0181             |
| Index of agricultural means of production
  Price Index of Means of Production for
  Mechanized Farm Tools                          | 0.4582                 | 0.0324             |
| Affected area                                  | 0.1256                 | 0.2558             |
| Diesel price                                   | 0.0495                 | 0.0544             |
| Driver service fee                             | 0.0836                 | 0.0674             |
| seasonal variation                             | 0.6229                 | 0.0635             |
| Level of economic development                  | 0.0482                 | 0.0649             |
The model has the following disadvantages: when the multicollinearity is checked by the conditional discriminant method, the multicollinearity determination condition k value of the multicollinearity model solved by the least square method is 118.4072. At present, it is generally believed that when the k value is more than 30, the k value can be considered too large, which indicates that there is a serious collinearity problem [12, 14, 15]. When there is such a collinearity problem, it shows that there is a high correlation between the independent variables in the regression model, which will cause the confidence interval of the model parameters to be too large and the solution result of the model to be inaccurate. At the same time, it is difficult to explain a single regression coefficient, causing instability of the model [16].

In the experiment, the model (5) is further solved by lasso method, and redundant variables are eliminated by Mallows Cp method. The solution results of this model are shown in Table 2. The experimental results show that six variables are eliminated and the multicollinearity criterion value is 19.660. The multicollinearity problem is well optimized and the set of extracted variables is also better. The regression coefficient of a variable in Table 2 is greater than zero, which indicates that the variable plays a positive role in cucumber price, while it plays a negative role when it is less than zero. The coefficient of per capita disposable income, resident population, index of production data of mechanized farm tools, labor cost of drivers, and economic development level is 0, which is eliminated by lasso regression solution and Mallows Cp unified measurement method. It can be seen that lasso regression method can effectively solve the problem of multiple collinearities in the model, select the best variable set and obtain a regression model with better generalization ability.

| Variable                        | Regression coefficient |
|---------------------------------|------------------------|
| inhabitant                      | -0.1552                |
| Consumption level of residents  | 0.4679                 |
| Sowing area of cucumber         | -0.0081                |
| Cucumber yield                  | -0.0698                |
| Index of agricultural means of production | 0.4253             |
| Affected area                   | 0.0658                 |
| Diesel price                    | 0.0365                 |
| seasonal variation              | 0.5965                 |

We use the remaining one-third of the data set as the test set to test the multiple linear regression model and lasso regression model respectively, and observe the test error and model fitting performance. Table 3 shows the multicollinearity determination condition values, fitting errors and goodness of fit of the two methods. Obviously, the multicollinearity determination condition value of method 3 is only 19.660, and the fitting coefficient is 0.8448, which is better than that of method (2) on the whole, and the problem of multicollinearity is reasonably solved, and the fitting error is small, indicating that lasso regression method has better performance in the model of influencing cucumber price factors.

| Method                        | Multicollinearity determination condition value | regression error | regression coefficient |
|-------------------------------|-----------------------------------------------|------------------|------------------------|
| Least square method           | 118.4072                                      | 0.1169           | 0.7629                 |
| Lasso regression method       | 19.66                                         | 0.0903           | 0.8448                 |
5 Conclusion
According to the empirical analysis of the price fluctuation rule, we can draw a conclusion that the cucumber market price in Beijing has obvious seasonal characteristics, with high price in winter and low price in summer. Besides seasonal factors, it is also affected by supply factors, demand factors, circulation factors and other comprehensive factors. In order to determine the degree of influence of each influencing factor, lasso regression method is used to model each influencing factor and price. Compared with conventional multivariate linear model, it is proved to have better effect, mainly reflected in higher precision of regression coefficient of influencing factors and smaller linear correlation.

The model regression results show that among the demand factors, the correlation degree between per capita disposable income and inflation level and cucumber price is smaller, and the consumer level and resident population have greater influence on cucumber price. Among the supply factors, the index of the means of production of mechanized farm tools had almost no effect on the price of cucumber, so it was eliminated, while the sowing area, the yield of cucumber, the index of the means of agricultural production and the affected area all had an effect on the price of cucumber to some extent. Among the circulation factors, the correlation between the driver's labor cost and the price is not much excluded, while the price of diesel oil will affect the price of cucumber. Among the comprehensive factors, the level of economic development has little to do with the price of cucumber.

Based on cucumber price data and influencing factor data, this paper applies lasso regression method to analyze the influencing factors of vegetable price and more accurate results have been achieved. It provides new ideas and analysis methods for the analysis of influencing factors of vegetable price and provides basis and reference for further research on vegetable price prediction.

Acknowledgments
This work was supported by Beijing Municipal Natural Science Foundation (4172025, 4172026) and the Natural Science Foundation of China (41501418).

References
[1] Li Chongguang, Bao Yuze. Analysis on the Characteristics and Causes of Vegetable Price Fluctuation in China [J]. Chinese Vegetables, 2012 (9): 1 - 7.
[2] Song Changming, Xu Juan, Xiang Chaoyang. Analysis on the Causes of Vegetable Price Fluctuation Based on Time Series Decomposition [J]. statistics and decision, 2014 (3): 106 - 108.
[3] Zhou Zhenya, Li Jianping, Zhang Qing, et al. Analysis of Vegetable Price Problem and Its Causes in China [J]. Agricultural Economic Problems, 2012 (7): 91 - 95.
[4] Zhao Anping, Zhao Yousen, Wang Chuan. Influencing Factors and Evaluation of Vegetable Price Fluctuations in Beijing and Policy Recommendations [J]. Agricultural Modernization Research, 2012,33 (5): 000598 - 602.
[5] Chen Fang, Lou wengao. daily price forecast of vegetable market based on generalized regression neural network [J]. journal of Zhejiang agriculture, 2015,27 (7): 1253 - 1258.
[6] Xu Yiting, Mu Yueying, Zhao Yousen. Analysis of Vegetable Price Changes and Influencing Factors in Beijing [J]. Chinese Vegetables, 2017 (11): 62 - 69.
[7] Liu Yan. Mathematical Model of Multiple Linear Regression [J]. Journal of Shenyang Institute of Engineering ( Natural Science Edition ), 2005,1 (2): 128 - 129.
[8] Breiman L. Heuristics of instability and stabilization in model selection[J]. Ann Statist, 1996, 24(6):2350-2383.
[9] Li Haichao, Wang Kaijun. Mining Factors Affecting the Success of Online Loan Borrowing Based on LASSO Regression Model [J]. Computer System Application, 2017, 26 (7): 204 - 209.
[10] Tibshirani, Robert. Regression shrinkage and selection via the lasso: A retrospective[J]. Journal of the Royal Statistical Society: Series B Statistical Methodology, 2011, 73(3):273-282.
[11] Mallows C L. Some Comments on Cp[J]. Technometrics, 42(1):87-94.
[12] Wu Xizhi. Complex data statistical methods: R - based applications [M]. Renmin University of China Press, 2015.

[13] Fox J. Component+Residual (Partial Residual) Plots[J].

[14] Gao huixuan. several methods to deal with collinearity of independent variables in multiple linear REGression - use of reg and other process enhancement functions in --SAS/STAT software ( 6.12 ) [J]. mathematical statistics and management, 2000,19 (5): 49 - 460.

[15] He Xiaoqun, Liu Yiqing. Applied Regression Analysis [M]. Renmin University of China Press, 2001.

[16] Robertl.Kabacoff, Cabaco Fu, Wang Xiaoning, et al. R Language Practice [M]. People's Posts and Telecommunications Press, 2016.