**Mathematical Modeling and Algorithm Application**
ISSN: 3006-0842 | Vol. 6, No. 1, 2025

# Research on Heilongjiang Rice Industry Based on Multiple Linear Regression Analysis

**Rui Yang[1], Junqian Shao[1, *], Xueyan Song[2], Zidan Wang[1]**

[1] College of Information Engineering, Suihua University, Suihua 152061, China
[2]College of Automobile and Transportation, Shenyang Ligong University, Shenyang 110159, China
[*] **Corresponding author**: Junqian Shao (Email: junqianshao@126.com)

**Abstract:** Against the backdrop of a severe global food security situation, accurate prediction of crop yields is crucial for ensuring stable food supply and maintaining socio-economic stability. Heilongjiang Province is an important region for rice production in China, and its rice yield is of great significance to the stable supply of national grain. This study uses multiple linear regression and random forest regression models to conduct in-depth research and predictive analysis on rice production in various cities of Heilongjiang Province. The data is sourced from authoritative databases and field research, after data cleaning, which includes removing outliers, handling missing values, and correcting errors, and standardization, which normalizes data to a common scale (such as using z-score normalization or min-max scaling) to ensure consistency and comparability, models are constructed. The multiple linear regression model is based on the assumption of linear relationship, estimates parameters using the least squares method, and rigorously tests and evaluates them; The random forest regression model, as an optimization model, utilizes multiple decision trees for ensemble learning to improve prediction accuracy and reliability. Comparing the predicted values of the two models, it was found that they each have their own advantages, providing diverse and powerful support for predicting rice yield in Heilongjiang. This not only contributes to the scientific planning and sustainable development of the rice industry in Heilongjiang Province, but also has profound significance for ensuring national food security and promoting the modernization of agriculture.

**Keywords:** Multiple Linear Regression, Model Testing, Error Analysis, Random Forest Regression

## 1. Introduction

Food is of paramount importance to the people, and rice is an important grain crop. As an important rice producing area in China, Heilongjiang Province ranks among the top in rice production with advanced planting technology, vast planting area, and high-quality natural resources. Its outstanding yield ensures national food security. The Sanjiang Plain and Songnen Plain in the province are the main production areas for high-quality rice. From introducing dry cultivation and sparse planting technology to promoting intelligent agricultural machinery and green pest control technology, Heilongjiang continues to innovate and achieve "good rice production in cold regions with black soil". Internationally, rice research in different regions has different focuses: Japan develops precision agriculture technology, uses multiple linear regression modeling [1] to analyze environmental factors and construct rice growth models; Previous studies have grounded the use of multiple linear regression in agriculture and economics in econometric and resource-based theories. India, the US, and Europe have applied this method in resource optimization, economic impact analysis, and sustainable development respectively. However, these studies are mostly region - specific and lack comprehensiveness in considering influencing factors. This study focuses on Heilongjiang Province's rice industry [2]. By integrating multiple economic, environmental, and policy factors, it offers greater regional specificity and captures the industry's diverse characteristics. In the face of growing agricultural complexity due to globalization, climate change, etc, this multi - factor analysis provides more targeted insights, enhancing both theoretical knowledge and practical strategies in agricultural economics.

## 2. Data processing of multiple regression models

Data collection [3] is conducted through multiple channels, including authoritative platforms such as the Dongfang Aixue database, the Chinese single season rice planting dataset, the Chinese daily meteorological data set, and the Heilongjiang Rural Statistical Yearbook, to obtain macro data such as planting area, yield, and meteorology. At the same time, we conducted on-site research in major rice planting areas such as Yichun and Wuchang in Heilongjiang Province, communicated with multiple households, and collected first-hand information on seed variety selection and field management. The original data contains missing values and outliers. Taking the rice yield data from 2015 to 2024 as an example, some years have missing data in certain regions. For continuous variables with missing values, if the missing ratio is less than 10%, the mean imputation method is used; For missing values in time series data, linear interpolation is used to identify outliers using box plots and correct them; To eliminate the influence of dimensionality, the Z-score standardization method was adopted, and the standardization results are shown in Table 1.

**Table 1** Standardized data of rice production related indicators from 2015 to 2023

| Project | Rice yield | Cultivated area | Fertilizer usage | Irrigation water volume | labor input | Agricultural machinery investment | Seed dosage | Pesticide dosage |
|---|---|---|---|---|---|---|---|---|
| 2015 | 0.7712 | -0.1494 | -0.9392 | -0.6171 | -0.9185 | -0.9336 | -0.9185 | -0.9185 |
| 2016 | 0.5360 | -0.1379 | -0.9490 | 2.6621 | -0.7724 | -0.8144 | -0.7724 | -0.7724 |
| 2017 | -0.1771 | -0.0983 | -0.7691 | -0.4999 | -0.6263 | -0.6953 | -0.6263 | -0.6263 |
| 2018 | -0.1649 | -0.4606 | -0.5889 | -0.4613 | -0.4802 | -0.4573 | -0.4802 | -0.4802 |
| 2019 | -0.0882 | -0.3491 | -0.4328 | -0.4327 | -0.3341 | -0.2782 | -0.3341 | -0.3341 |
| 2020 | 2.2919 | -0.2306 | -0.4525 | -0.4424 | -0.1880 | 0.0977 | -0.1880 | -0.8180 |
| 2021 | 0.1792 | -0.2421 | 0.1437 | -0.4138 | 0.0042 | 0.2769 | 0.0042 | 0.0042 |
| 2022 | 0.2014 | -0.6250 | -0.2226 | -0.4710 | 0.1503 | 0.4560 | 0.1503 | 0.1503 |
| 2023 | 0.4713 | -1.299 | 0.8601 | -0.3941 | 0.3604 | 0.7934 | 0.3604 | 0.3604 |

# 3. Modeling process of multiple linear regression

## 3.1. Hypothesis of Multiple Linear Regression Model

Assuming there is a linear relationship between rice yield Y and independent variables X1, X2,..., X10, establish an initial model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{10} X_{10} + \varepsilon \quad (1)$$

For farmers' income, considering its relationship with rice yield Y, sales price Pr (unit: yuan/ton), and production cost C (unit: yuan/hectare), assuming:

$$R = P_r \times Y - C \quad (2)$$

Further substituting the linear regression model of Y, we obtain:

$$R = P_r \times (\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_3 X_3 + \varepsilon) - C \quad (3)$$

The definition of arable land grain yield is the ratio of rice yield Y to arable land area X1, that is

$$P = \frac{Y}{X_1} = \frac{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{10} X_{10} + \varepsilon}{X_1} \quad (4)$$

## 3.2. Estimation of Parameters by Least Squares Method

Finding a set of regression coefficients that minimize the sum of squared residuals (SSE) and make the model's predicted values as close as possible to the actual observed values is also a key indicator for measuring the model's prediction error [4]. The sum of squared residuals is

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left[ y_i - \left( \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_{10} x_{i10} \right) \right]^2 \quad (5)$$

Calculate the partial derivatives of SSE separately and make them

$$\begin{cases} \dfrac{\partial SSE}{\partial \hat{\beta}_0} = -2 \sum_{i=1}^{n} \left[ y_i - \left( \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_{10} x_{i10} \right) \right] = 0 \\ \dfrac{\partial SSE}{\partial \hat{\beta}_1} = -2 \sum_{i=1}^{n} \left[ y_i - \left( \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_{10} x_{i10} \right) \right] = 0 \\ \cdots \\ \dfrac{\partial SSE}{\partial \hat{\beta}_j} = -2 \sum_{i=1}^{n} \left[ y_i - \left( \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_{10} x_{i10} \right) \right] = 0 \end{cases} \quad (6)$$

By taking partial derivatives to obtain a system of equations for solving regression coefficients, the data is standardized and error processed using C language. The linear equation for the regression coefficients is obtained as follows:

$$Y = 456.3356 + 17.2115 X_1 - 1.2987 X_2 + \cdots + 0.000 X_{10} + \varepsilon \quad (7)$$

## 3.3. Multiple Linear Regression Model Testing

The goodness of fit test (R2-test) uses SST to reflect the total fluctuation of the dependent variable, SSR represents the fluctuation explained by the model, and the closer the R2 is to 1, the better the fit of the model to the data; Variable significance test (t-test) calculates the standard error of the regression coefficient, which measures the uncertainty of the regression coefficient. The t-test is used to test whether a single variable has a significant impact on the dependent variable; Overall significance test (F-test), the F-statistic is used to test the significance of the entire regression model.

## 3.4. Analysis of Inspection Numbers

In terms of model testing [5]: In the goodness of fit test, the closer the value is to 1, the better the fitting effect of the model on the data, indicating that the model can explain the relationship between the independent variable and the dependent variable well. The significance test of variables shows that only the t-value of coefficient 0 (10.0273) is relatively large, while the t-values of other coefficients are almost zero. As the absolute value of t increases, the independent variables become more significant, so only the constant term (coefficient 0) has a significant impact on the model, while the other independent variables have no significant impact. In the overall significance test (F-test), the F-value is relatively large. The F-test is used to determine whether the entire regression model is significant, that is, whether the independent variables as a whole have a significant impact on the dependent variable. The larger the F-value, the more likely it is to reject the null hypothesis that all independent variable coefficients are 0. This F-value indicates that the regression model is significant as a whole, although there may be differences in the significance of individual independent variables.

### 3.5. Model Optimization - Random Forest Regression Model

The random forest regression [6] model is based on decision tree ensemble learning. Randomly select multiple samples from the original training data to construct a decision tree, and randomly select some independent variables from all when splitting nodes. When predicting, take the average of the prediction results for a decision tree, and the formula is:

$$\hat{y} = \frac{1}{m}\sum_{j=1}^{m} y_j \tag{8}$$

By comparing the predicted values of multiple linear regression and random forest regression models, as shown in Figure 1:

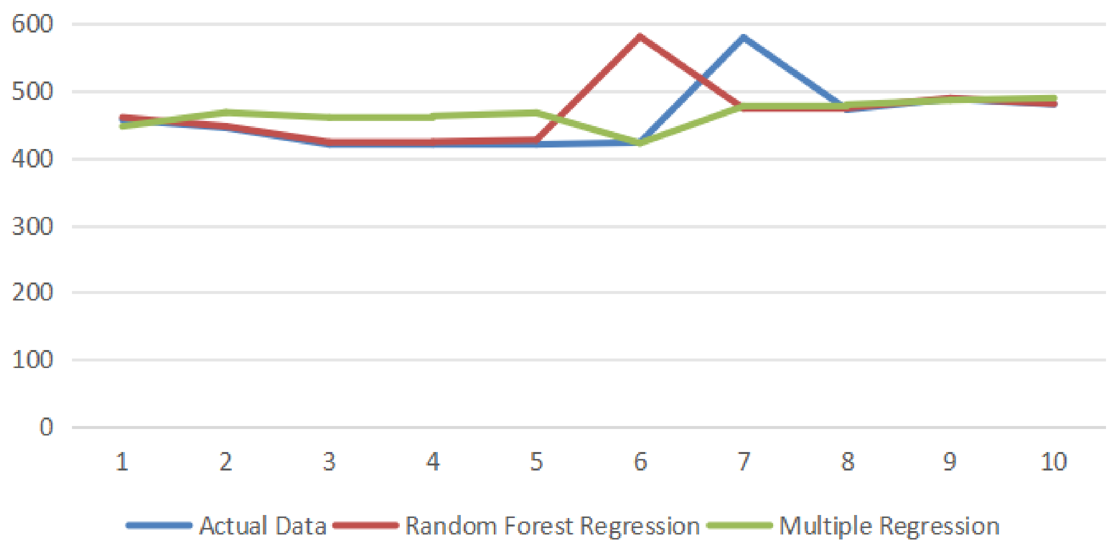


**Figure 1:** Line graph comparing raw data with predicted values from multiple regression and random forest models

## 4. Error analysis of multiple regression models

### 4.1. Error Types

Mean Square Error (MSE) [7] is the average of the sum of the squared differences between predicted and actual values. The smaller the MSE value, the higher the model's prediction accuracy; Mean Absolute Error (MAE) is calculated using the absolute value of error to avoid amplifying large errors and visually reflect the average deviation between predicted and actual values; Root Mean Square Error (RMSE) is the square root of MSE, which makes the error index consistent with the dimension of the dependent variable and reflects the average error amplitude between the predicted value and the actual value; Relative Error (RE) considers the actual value size and reflects the prediction accuracy of the model at different data scales by the ratio of the error to the actual value.

### 4.2. Error numerical analysis

The study compared and analyzed the predictive performance of multiple regression models and random forest regression models using four indicators: mean square error (MSE), mean absolute error (MAE), root mean square error (RMSE), and mean relative error (MRE). The results showed that the MSE (110.2019), MAE (8.1693), RMSE (10.4977), and MRE (0.0193) of the random forest regression model were all lower than the corresponding values of the multiple regression model (255.0632, 11.8179, 15.9707, 0.0272), indicating that its predicted values deviated less from the true values, had lower average errors, smaller fluctuations, and better relative errors. Based on the four indicators, the random forest regression model has better accuracy and stability in predicting rice yield in Heilongjiang province.

## 5. Model application and suggestions

### 5.1. Model Application

(1) Production forecasting and planning. Based on the established rice yield prediction model, agricultural departments and farmers can predict rice yield in advance based on the current year's arable land area planning, expected input of production materials (fertilizer, seeds, pesticides, etc.), and estimation of climate conditions (temperature, precipitation, sunshine duration). For example, before spring plowing, farmers can use models to predict yield based on their planned planting scale and input, in order to arrange production activities reasonably and reduce production risks.

(2) Benefit evaluation and decision-making. The farmer income model can help farmers evaluate the income situation under different planting strategies based on predicted rice yields, market price fluctuations, and production costs. This helps farmers to simulate the benefits of different options through models when faced with decisions such as seed variety selection, fertilizer usage adjustment[8], and whether to adopt new agricultural technologies, and choose the optimal production decision to achieve maximum benefits.

(3) Analysis of Land Use Efficiency. The model of cultivated land grain yield can be used to analyze the utilization efficiency of different regions and types of cultivated land. By comparing the actual yield with the model predicted yield, identifying farmland with lower yield, and analyzing influencing factors such as soil fertility and irrigation conditions, a scientific basis is provided to improve farmland utilization efficiency and guide land improvement and agricultural infrastructure construction[9].

## 5.2. Suggestions

(1) Optimize agricultural production inputs. Reasonably adjust agricultural production inputs based on the impact of various factors in the model on rice yield and farmers' income. For example, if the model shows that the marginal effect of fertilizer use on yield gradually decreases and may have a negative impact on the environment, then farmers should be encouraged to reduce fertilizer use and instead adopt green and environmentally friendly organic fertilizers or precision fertilization technologies to improve agricultural product quality and ecological benefits while ensuring yield.

(2) Strengthen the construction of agricultural infrastructure. In response to the key factors found in the model analysis that affect rice yield, such as insufficient irrigation water or inadequate irrigation facilities leading to limited yield, the government should increase investment in agricultural irrigation facilities [10], build efficient water-saving irrigation systems, ensure sufficient and reasonable water supply during rice growth, improve water resource utilization efficiency, and ensure stable growth of rice yield.

(3) Improving the technological literacy of farmers. By organizing agricultural technology training [11], expert lectures, and distributing technical manuals, farmers' mastery of advanced planting techniques and management experience can be improved. Enable farmers to better understand and apply model analysis results, such as controlling pesticide usage to reduce costs and environmental pollution, adjusting planting time and field management measures according to climate conditions, etc., thereby improving agricultural production efficiency and economic benefits.

# 6. Conclusions

This study collected, organized, and analyzed rice production data in Heilongjiang Province, and constructed statistical models for rice yield, farmer income, and cultivated land grain yield. The entire process from data preprocessing to model optimization was demonstrated in detail, and the model prediction and error analysis were implemented using C language. By studying the model results, clarify the impact mechanism of each factor on relevant indicators. The results show that the multiple linear regression model can effectively fit and predict to a certain extent, but there are still errors. Error analysis provides direction for model improvement. At the same time, targeted suggestions are proposed based on model analysis to provide decision-making basis for the sustainable development of the rice industry in Heilongjiang. Future research can expand data sources and time spans, introduce more influencing factors and advanced modeling methods, improve model accuracy and applicability, and better serve agricultural production.

# Acknowledgements

The authors gratefully acknowledge the financial support from Heilongjiang Province College Students Innovation Training Program Project (No. 202310236030) and Suihua College Student Innovation Training Program Project (No. shxy202489).

# References

[1] Lu Wenxi, Si Changliang, Cheng Weiguo, etc. Research on the Relationship between Rice Yield and Various Physiological Parameters Based on Multiple Linear Regression [J]. Water saving Irrigation, 2012, (12):37-39+42.

[2] Chai Ling. Research on Production Behavior and Influencing Factors of Rice Growers in Heilongjiang Province [D]. Northeast Agricultural University, 2017.

[3] Jia Ning, Zheng Chunjun. Exploration of Hybrid Education for Big Data Talent Cultivation under CDIO Concept: Taking Data Acquisition and Preprocessing Practice Course as an Example [J]. Computer Education, 2021, (08):35-39.

[4] Dong Qiaoling. Research on the Application of Total Least Squares Method in Multiple Linear Regression under Different Error Influence Models [D]. Taiyuan University of Technology, 2016.

[5] Yao Zengfu, Tang Huajun, Liu Xin. Research on the Threshold Effect between Factor Accumulation, Human Capital, and Agricultural Environmental Efficiency: Panel Threshold Model Test under Low Carbon Constraints [J]. Journal of Chongqing University (Social Sciences Edition), 2017, 23 (04): 26-36.

[6] Qi Hongyan. Research on Soil erodibility Factors of Soil Erosion in Jiangxi Province Based on Random Forest Regression [J]. Water Resources Science and Cold Region Engineering, 2025, 8 (03): 79-81.

[7] Ekermu Kerermu. Error Analysis and Countermeasures of Surface Water Monitoring Results [J]. Heilongjiang Environmental Bulletin, 2025, 38 (04): 72-74.

[8] Populus euphratica. Research on the Factors Influencing the Long term Value of Enterprises [D]. Sichuan University, 2021.

[9] Huang Jianbi, Ma Chunyan, Zeng Guang. Environmental regulation and efficiency of agricultural green technology innovation: based on the two-stage perspective of innovation [J]. Journal of China Agricultural University, 2025, 30 (05): 230-247.

[10] Wen xin. Application Analysis of Water saving Irrigation Technology in Greenhouse White Crispy Melon Planting in Qin'an County, Gansu Province - Taking the Irrigation Project of Qin'an County Facility Agriculture Production Base Construction Project as an Example [J]. Friends of Fruit Farmers, 2024, (04):80-83.

[11] Zhao Peng, Li Shijun. The impact of agricultural technology promotion on agricultural planting [J]. Cultivation and Cultivation, 2024, 44 (05): 149-151.