Crop yield predictions by regression statistical models

Manoj Kumar Beck

Division of Agricultural Physics

ICAR-IARI, New Delhi

Beck.manoj05a@gmail.com

Ananta vashisth

Division of Agricultural Physics

ICAR-IARI, New Delhi

ananta.iari@gmail.com

***Abstract***

Regression analysis is a statistical technique used to estimate the relationship between variables with a cause-and-effect connection. Predicting crop yield in the current growing season is crucial for farmers, policy makers, and food processing plants to make informed decisions and maximize income. Reliable prediction tools offer valuable insights for managing agricultural products and guide the adoption of appropriate strategies. However, it is important to consider the limitations and assumptions associated with regression models, such as linearity assumptions and the presence of outliers, to ensure their proper application. Research contributes to the existing knowledge on regression modeling for prediction tasks in the agricultural domain. By evaluating different regression algorithms and their performance metrics, practitioners and researchers gain a solid foundation for selecting and implementing suitable regression models in predictive analytics. Our results demonstrate that regression models are powerful tools for accurate predictions of target variables. Performance evaluation metrics such as mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), and R-squared (R²) coefficient provide a comprehensive assessment of the model's predictive ability and overall fit. The challenge in creating forecasting models lies in selecting the appropriate independent variables, which requires a deep understanding of the research subject. Regression analysis provides precise quantitative information that managers can rely on for decision-making and problem-solving. It offers a structured approach to data analysis, emphasizing its importance in research and decision-making processes. In this, regression analysis plays a significant role in predicting crop yield and enabling informed decision-making in the agricultural sector. By considering the limitations and assumptions of regression models, practitioners can leverage their predictive capabilities to generate reliable estimates. The comprehensive evaluation of different regression algorithms and their performance metrics presented in this research serves as a valuable resource for practitioners and researchers seeking to enhance their predictive analytics capabilities in the agricultural domain.

**Keywords**: Regression analysis; prediction tool; crop; coefficient; data

1. **Introduction**

Regression is a statistical technique used to model the relationship between a dependent variable and one or more independent variables. It aims to understand how changes in the independent variables are associated with changes in the dependent variable. It is widely used in various fields, including economics, finance, social sciences, and machine learning. In marketing applications, the dependent variable is typically the goal we are trying to attain, while the independent variables are the tools we might use to get there. Few other techniques are as insightful as regression analysis. Regression analysis has the following main advantages:

1. State whether a link between independent and dependent variables is meaningful.

2. Describe how powerfully certain independent variables have an impact on the dependent variable.

3. Assume the future.

Agricultural researchers can benefit in a variety of ways from understanding how independent variables affect dependent variables. For instance, knowing that promotional activities considerably enhance agricultural yield can assist target spending. In agriculture, understanding the relative strength of effects can be helpful in determining issues like whether productions are more dependent on weather or other inputs. Additionally, regression analysis enables us to examine the effects of factors assessed at various scales, such as the impact of changing weather patterns, various crop kinds, seasons, sowing dates, and spatial distribution.

Based on soil weather and crop factors, regression models for estimating the yield of crops like cotton, wheat, and maize. There is a correlation between a few (linear and nonlinear) methods for predicting crop yield [1]. The comparison is performed using the best property subset for each method identified using percentage split validation and an entire algorithm from the preparation dataset. The technique builds the models using the oldest samples and then searches the training datasets for the best attribute subset. Unseen samples make up the test datasets, where performance is assessed. The most well-known information-driven techniques for agricultural yield prediction, including regression trees, stepwise linear regression, multiple linear regression, and neural networks, were evaluated. Taking into account that the Ordinary Least Square (OLS) estimation is a widely used method for crop yield prediction, [2] researchers take the linear regression model into consideration. With a greater R', the autoregressive model outperformed the OLS in this case. The study found that, ignoring temperature, NDVI and precipitation contributed more to the com yield in Iowa.

Regression models, time series models, and probabilistic models are only a few of the forecasting methods covered by Ramasubramanian academics in the field of agriculture. In a regression model, there are three models: a multiple linear regression model based on weather indicators for agricultural pest count; a logistic regression model for qualitative response variables; and a multiple linear regression model for crop yield forecasts. For the production of crops, an exponential smoothing model and an auto regressive integrated moving average model are both utilized in time series models. Markov Chain Model is used in probabilistic models to anticipate agricultural yield.

Regression models are used to predict the values of the dependent variable based on the values of the independent variables. This predictive capability is valuable for forecasting future outcomes and making informed decisions. It explains the relationship between the dependent variable and the independent variables. They provide insights into the factors that influence the dependent variable and the direction and magnitude of their effects. Enable researchers and analysts to control for the effects of other variables and isolate the impact of specific factors. By accounting for confounding variables, regression models allow for a clearer understanding of causal relationships.

1. **Types of Regression Models**

There are various types of regression models, each suited for different scenarios and data types. Some commonly used regression models include:

1. Simple Linear Regression: This model involves a single independent variable and a linear relationship with the dependent variable. It assumes a straight-line relationship between the variables.

Simple linear regression is a statistical technique used to model the relationship between a single independent variable (X) and a dependent variable (Y). It assumes that there is a linear relationship between the two variables, meaning that a change in X is associated with a proportional change in Y. Simple linear regression can be expressed by the equation:

Y = β₀ + β₁X + ε

where:

- Y is the dependent variable

- X is the independent variable

- β₀ is the intercept (the value of Y when X is 0)

- β₁ is the slope (the change in Y for a one-unit change in X)

- ε is the error term (represents the random variability or unexplained factors)

1. Multiple Linear Regression: In this model, there are multiple independent variables influencing the dependent variable. It allows for the analysis of the combined effects of multiple factors. Multiple linear regression extends the concept of simple linear regression by considering the relationship between a dependent variable (Y) and two or more independent variables (X₁, X₂, ..., Xₚ). It assumes a linear relationship between the dependent variable and the independent variables, allowing for the analysis of multiple factors simultaneously. The multiple linear regression model can be expressed as:

Y = β₀ + β₁X₁ + β₂X₂ + ... + βₚXₚ + ε

where:

- Y is the dependent variable

- X₁, X₂, ..., Xₚ are the independent variables

- β₀ is the intercept (the value of Y when all X variables are 0)

- β₁, β₂, ..., βₚ are the slopes (representing the change in Y for a one-unit change in each X variable)

- ε is the error term (representing the random variability or unexplained factors)

1. Polynomial Regression: This model extends simple linear regression by including polynomial terms of the independent variable. It can capture nonlinear relationships between variables. Polynomial regression is an extension of linear regression that allows for modeling nonlinear relationships between the dependent variable (Y) and the independent variable (X). It involves fitting a polynomial function of the independent variable to the data. Polynomial regression can capture more complex patterns and curves in the data by including higher-order polynomial terms.

The polynomial regression model can be expressed as:

Y = β₀ + β₁X + β₂X² + ... + βₚXᵖ + ε

where:

- Y is the dependent variable

- X is the independent variable

- X², X³, ..., Xᵖ are the higher-order polynomial terms (squared, cubed, etc.)

- β₀, β₁, β₂, ..., βₚ are the coefficients for each term

- ε is the error term (representing the random variability or unexplained factors)

1. Logistic Regression: Logistic regression is a statistical modeling technique used to analyze the relationship between a binary or categorical dependent variable and one or more independent variables. Unlike linear regression, which is used for continuous dependent variables, logistic regression is specifically designed for predicting binary outcomes.

The logistic regression model uses the logistic function (also known as the sigmoid function) to transform the linear combination of the independent variables into a probability value between 0 and 1. This probability represents the likelihood of the dependent variable belonging to a specific category. The logistic regression model can be expressed as:

p = 1 / (1 + exp(-(β₀ + β₁X₁ + β₂X₂ + ... + βₚXₚ)))

where:

- p is the probability of the dependent variable belonging to a particular category

- β₀, β₁, β₂, ..., βₚ are the coefficients for each independent variable

- X₁, X₂, ..., Xₚ are the independent variables

5. Ridge regression: Ridge regression also known as Tikhonov regularization, adds a penalty term to the least squares objective function in linear regression. The penalty term is a scaled version of the sum of squared coefficients, multiplied by a regularization parameter (λ). The ridge regression objective function can be expressed as:

minimize: RSS + λ∑(βᵢ²)

where:

- RSS is the residual sum of squares, representing the difference between the predicted values and the actual values.

- λ (lambda) is the regularization parameter that controls the amount of shrinkage applied to the coefficients.

- ∑(βᵢ²) is the sum of squared coefficients.

The ridge regression penalty term helps to reduce the magnitude of the coefficients, particularly those that are less important, without completely eliminating them. Ridge regression does not perform variable selection, as all variables remain in the model, although their coefficients are shrunk towards zero.

6. Lasso Regression: Lasso regression (short for "least absolute shrinkage and selection operator") is another regularization technique that adds a penalty term to the least squares objective function. However, unlike ridge regression, lasso regression uses the sum of the absolute values of the coefficients multiplied by the regularization parameter (λ) as the penalty term. The lasso regression objective function can be expressed as:

minimize: RSS + λ∑|βᵢ|

where:

- RSS is the residual sum of squares, representing the difference between the predicted values and the actual values.

- λ (lambda) is the regularization parameter that controls the amount of shrinkage applied to the coefficients.

- ∑|βᵢ| is the sum of absolute values of the coefficients.

Lasso regression has the property of performing both variable selection and regularization. It can drive the coefficients of irrelevant or less important variables to zero, effectively eliminating them from the model. Therefore, lasso regression can be useful in situations where feature selection is desirable.

3. Key Differences: The main differences between ridge regression and lasso regression are:

a) Penalty term: Ridge regression uses the sum of squared coefficients as the penalty term, while lasso regression uses the sum of absolute values of the coefficients.

b) Variable selection: Ridge regression does not perform variable selection; it keeps all variables in the model but shrinks their coefficients. Lasso regression, on the other hand, performs variable selection by driving the coefficients of less important variables to zero.

c) Solution: Ridge regression has a closed-form solution and can be solved analytically. Lasso regression does not have a closed-form solution, and various optimization algorithms like coordinate descent or least angle regression (LARS) are used to find the optimal solution.

d) Sparsity: Lasso regression can produce sparse models with fewer non-zero coefficients, making it useful when interpretability and feature selection are important. Ridge regression generally does not yield sparse models, as the coefficients are only shrunk towards zero, but not eliminated completely.

e) Tuning parameter: Both ridge regression and lasso regression have a regularization parameter (λ) that controls the amount of shrinkage. However, λ has a different impact on the coefficients in each method. In ridge regression, higher values of λ lead to more shrinkage of the coefficients, while in lasso regression, higher values of λ can drive more coefficients to zero.

7. Time series: In time series regression, the primary objective is to model and forecast the behavior of the dependent variable based on the values of the independent variables and their historical patterns. The independent variables can include lagged values of the dependent variable itself, as well as other exogenous variables that may influence the time series.

Data required for regression analysis

Estimate and Specify the model

To test the assumptions

Interpretation

Validation

Use the regression model

Fig. 1 Steps of regression analysis

The time series regression model can be expressed as:

Yₜ = β₀ + β₁X₁ₜ + β₂X₂ₜ + ... + βₚXₚₜ + εₜ

where:

- Yₜ represents the value of the dependent variable at time t

- X₁ₜ, X₂ₜ, ..., Xₚₜ represent the values of the independent variables at time t

- β₀, β₁, β₂, ..., βₚ are the coefficients associated with each independent variable

- εₜ represents the error term or residual at time t

**3. Yield prediction through regression model**

Design of the regression models used to estimate crop yield: In this work, the general approach taken to build the regression models for crop yield prediction is explained. Because it gives a predicted entity as a function of the dependent entities, regression analysis is frequently employed for prediction purposes. It provides correlations between independent and dependent variables in some circumstances [3]. The procedures for creating a regression model to forecast crop yield.

Start

Gathering and processing the raw data

Training and test the data set

Train the regression model and apply trained model on test set

Compute R2, RMSE and MPPE for model

Check if RMSE, MPPE is lowest and R2 is highest

No

Yes

Stop

Fig. 2: Regression methodology for crop yield prediction

A regression model as a tool for prediction and chose a few key elements in yield production [4]. The traits and information obtained together can be used as the input variables for the regression model. In a similar vein, the yield's best regression model is identified. To deal with potential root mean square and R2 statistics value estimates, each model is performed several times. The survey's best regression model is used to forecast the generation of the various crops throughout a range of years. The results show that the suggested regression model is an effective way to predict yield production. Based on the root mean square, R2 statistics, and percentage prediction error, the outcomes of several models are compared. The best model for predicting crop yields is one that has lower values for Root Mean Square, percentage prediction error and R2 statistics.

In order to forecast the amount of rainfall in the region of Myanmar, researchers take into account the Polynomial Regression Model (1.1PR) [5]. The authors used second-arrange 1.1PR to develop a prediction forecast model with 15 predictors. A few tests have shown that the predicted precipitation swn is rather close to the actual characteristics. Four indicators were used to develop the SJ.1R expectation model. The model's outputs were put to use in the territories, such as in water management and repository control, harvest planning, and yield prediction. Helping with farm management and water management is the main goal of forecast model improvement. In the application of multiple polynomial regression, which only makes use of test data from 2006, all conceivable subsets of predictors have been investigated. The authors present evidence that 1.1PR outperforms MLR. Researchers expand on traditional regression neural networks [6]. The Vapor Pressure Deficit (VPD), CO2, radiation, and temperature are all taken into account in the construction of the Conformal Prediction (CF) framework for predicting tomato production in a greenhouse. This technique required the utilization of over 60,000 records.

There are three primary categories of statistical models: panel models based on spatio-temporal variation data, cross-section models based on spatial variation data, and time-series models based on time series data from a single point or region [7]. Recent literature has examined how well the three strategies indicated above perform, however the findings are not conclusive. For the several statistical models demonstrated that, based on data of crop yields, including maize, soybean, and cotton, as well as climate variables in America, the differences of results are not always clear [8]. However, the panel regression method to demonstrate that time-series models do not adequately capture the robustness of corn production responses to temperature in Sub-Saharan Africa [9]. Panel regression method's error is lower than that of the time-series model [10].

Time-series models make some fixed assumptions, such as the idea that the historical correlation between agricultural yields and climate will be constant in the future [11]. Time-series models hardly ever take climate change adaptations into account, with the exception of those for extreme weather events [12]. Cross-section models assess various geographical conditions while taking into account all of the adaptations in various areas and climatic variables [13]. The quantity of adaptations in panel models, which mix time data from various locations into one analysis, should be comparable to those in the other two methods. The exact amount of adaptation, however, depends on the annual fluctuations within sites as opposed to annual differences across sites [14].

Researchers demonstrated that the LASSO needs a particular circumstance in order to choose the ideal set of nonzero true coefficients, provided that these coefficients are bounded away from zero at a specific pace [15]. As a result, the LASSO generally lacks selection consistency, which is consistent with the findings of reference [16]. Due to the shrinkage effect, LASSO chooses more predictive variables than there are actual variables, yet it still manages to reach an optimality in the minimax sense, resulting in good prediction accuracy. For some precise minimax rates of the LASSO in high-dimensional models, see reference [17]. Reference [18] suggested using the two-stage adaptive Lasso for consistent model selection in linear and Gaussian graphical models under the restricted eigenvalue conditions, reference [19] demonstrated that marginal linear regression can be used to obtain a weight vector under partial orthogonality conditions even when p > n.

Logistic regression is becoming more popular as sophisticated statistical software for fast computers becomes more widely available. Due to this increased usage, readers, editors, and researchers must be aware of what to anticipate from articles that make use of logistic regression techniques. In educational research, particularly in higher education [20], logistic regression has become more prevalent [21]. The logistic regression method is useful for expressing and testing hypotheses concerning correlations between a categorical outcome variable and one or more continuous or categorical predictor variables. Assume that there is only one continuous predictor (a child's reading score on a standardized exam) and one dichotomous outcome variable (Y) in a linear regression study. The R2 score for either the entire model or for each predictor gave evidence of the goodness-of-fit of logistic models [22].

In one investigation, the findings were expressed as marginal probabilities [23]. Reference [24;25] have criticized the use of marginal probabilities because they do not correspond to a fixed change in the predicted probabilities that will happen if one predictor (for example, reading) changes discretely while other predictors are realized at a constant. In other words, the marginal likelihood associated with a 10-point shift in reading from, say, 60 to 70 points is different than the marginal probability associated with a change from 50 to 60 points.

The fundamental idea behind the hole-drilling approach is to identify the changes in stress states that take place when a structural member with residual stresses is subjected to a drilling operation. There are numerous articles devoted to this method that provide detailed descriptions of the process [26-29]. In a metallurgical facility, the hole-drilling method was used to determine the residual stresses in the case of the transverse beam supporting the casting ladle. A polynomial regression model has multiple power terms that follow one another. The highest order word as well as all other terms, important or not, will be included in each model. As a specific instance of multiple linear regression, polynomial regression can be viewed as such. Polynomial models are a reliable and adaptable method for fitting curves. Ordinary least squares analysis is the regression analysis technique that is most frequently employed. The "best fit" line is drawn through all of the available data points in this procedure, and parameter values are selected to reduce the error sum of squares. A regression model must meet certain assumptions in order to be fit. The errors must be assumed to be uncorrelated random variables with a mean of zero and constant variance in order to estimate the model's parameters. The errors must be regularly distributed in order to perform hypothesis tests and interval estimates. Advanced statistical tests can be applied to determine whether or not these premises hold true for a particular regression equation.

1. **A look at statistical models in the future**

In this portion of the viewpoint on the future development of statistical models, appropriate solutions have been proposed for the four problems mentioned above.

(1). The accuracy of meteorological data interpolation and statistical data should be carefully reviewed, and detailed information about the effects of climate change on the agricultural system should be gathered. For instance, it helps to understand how adaptations might be used efficiently by using accurate and detailed data. The data include phenology yield and management information at county and station scales, homogenization data of meteorological stations, precise interpolators for weather data, continuous long-term crop growth, and so on. These thorough and precise SHI Review of statistical models for determining how the climate affects crop production. To investigate how the climate affects crop yields under various adaptations, 575 geographical data at different scales can be used. likewise, it makes sense to arrange data with similar characters together. In order to properly investigate how the climate influence crop yields, it is important to select the critical growth seasons for various crops.

(2). Crop yields should be eliminated as non-climatic trends, therefore distributions of yield data and climate data, such as hetero skedasticity of yield changes, spatial variations, etc., should be thoroughly taken into account. Future studies may employ higher order polynomials fitting, initial differences, or the addition of time variables to eliminate the effects of method advancements.

(3). Analyzing climate variables with minimal correlations among themselves can help to mitigate the consequences of the colinearity problem. Using crop growth records data from agrometeorological stations is important in order to study the effects of different weather factors on crop growth and development in different growing periods. Understanding how agricultural yields respond to climate variables is important for minimizing colinearity problems. Additionally, the colinearity problem can be resolved by combining crop models and statistical models to examine yield responses to climate change.

(4). To attempt to solve the issue of failing to take adaptations into account, we should classify yield data according to various adaptations based on gathering more precise information on management, weather, and yield. We should then utilize time-series models, cross-section models, and panel models to examine the yield responses to climate change. Combining crop models is an excellent option for investigating how the climate affects crop yields for various adaptations.

Reference:

1. Sanchez, AG., J.F. Solis and W.O. Bustamante, 2014. Attribute selection impact on linear and nonlinear regression models for crop yield prediction. Sci. World J., 2014: 1-10.
2. Zhang T, Zhu J, Wassmann R, 2010. Responses of rice yields to recent climate change in China: An empirical assessment based on long-term observations at different spatial scales (1981–2005). Agricultural and Forest Meteorology, 150(7/8): 1128–1137.
3. Alan, O.S., 1993. An introduction to regression analysis. Master Thesis, University of Chicago Law School, Chicago, Illinois.
4. Shastry, A., Sanjay, H.A. and Bhanusree, E. 2017. Prediction of Crop Yield Using Regression Techniques. International J ownal of Soft Computing 12 (2): 96-102.
5. Zaw, W.T. and T .T. N aing, 2009. Modeling of rainfall prediction over Myanmar using polynomial regress10n. Proceedings of the International Conference on Computer Engineering and Technology, January 22-24, 2009, IEEE, New York, USA, ISBN: 978-0-7695-3521-0, pp: 316-320.
6. Qaddourn, K. and EL. Hines, 2012. Reliable yield prediction with regress10n neural networks. Proceedings of the 12th WSEAS Intermtional Conference on Signal Processing, Computational Geometry and Artificial Intelligence, August 21-23, 2012, WSEAS Press, Turkey, Istanbul,-pp: I.
7. Lobell D, Burke M, 2010. On the use of statistical models to predict crop yield responses to climate change. Agricultural and Forest Meteorology, 150(11): 1443–1452.
8. Schlenker W, Roberts M J, 2009. Nonlinear temperature effects indicate severe damages to US crop yields under climate change. Proceedings of the National Academy of Sciences, 106(37): 15594.
9. Schlenker W, Lobell D B, 2010. Robust negative impacts of climate change on African agriculture. Environmental Research Letters, 5: 014010.
10. Lobell D, Burke M, 2009. Climate Change and Food Security: Adapting Agriculture to a Warmer World: Springer Verlag.
11. Lobell D, Schlenker W, Costa-Roberts J, 2011a. Climate trends and global crop production since 1980. Science, 333: 616.
12. Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. Journal of Machine Learning Reserach, 7, 2541-2563.
13. Meinshausen, N. and Yu, B. (2009). Lasso-type recovery of sparse representation for high-dimensional data. The Annals of Statistics, 37, 246.270.
14. Leng, C., Lin, Y. and Wahba, G. (2006). A note on the lasso and related procedures in model selection. Statistica Sinica, 16, 1273-1284.
15. Zou, H. (2006). The adaptive lasso and its oracle properties. Journal of the American Statistical Association, 101, 1418-1429.
16. Raskutti, G., Wainwright, M. J. and Yu, B. (2011). Minimax rates of estimation for high-dimensional linear regression over Lq-balls. IEEE Transactions on Information Theory, 57, 6979-6994.
17. Zhang, T. (2009). Some sharp performance bounds for least squares regression with L1 regularization. The Annals of Statistics, 37, 2109-2144.
18. Zhou, S., van de Geer, S. and B¨ulmann, P. (2009). Adaptive lasso for high dimensional regression and gaussian graphical modeling. arXiv preprint arXiv:0903.2515.
19. Huang, J., Ma, S. and Zhang, C.-H. (2008). Adaptive lasso for sparse high-dimensional regression models. Statistica Sinica, 18, 1603-1618.
20. Peng, C. Y., & So, T. S. H. (2002a). Modeling strategies in logistic regression. Journal of Modern Applied Statistical Methods, 14, 147–156.
21. Chuang, H. L. (1997). High school youth’s dropout and re-enrollment behavior. Economics of Education Review, 16(2), 171–186.
22. Trusty, J. (2000). High educational expectations and low achievement: Stability of educational goals across adolescence. The Journal of Educational Research, 93(6), 356–365.
23. McNeal, R. B., Jr. (1998). High school extracurricular activities: Closed structures and stratifying patterns of participation. The Journal of Educational Research, 91(3), 183–191.
24. Long, J. S. (1997). Regression models for categorical and limited dependent variables. Thousand Oaks, CA: Sage.
25. Peng, C. Y., So, T. S., Stage, F. K., & St. John, E. P. (2002). The use and interpretation of logistic regression in higher education journals: 1988–1999. Research in Higher Education, 43, 259–293.
26. Frankovský, P., Kostelníková, A., Šarga, P., 2010. The use of strain-gage method and PhotoStress method in determining residual stresses of steel console. Metalurgija. Vol. 49, no. 2, p. 208-212.
27. Ostertag, O., Sivák, P., 2010. Degradation processes and fatigue life prediction (in Slovak), Typopress Košice, Slovakia, ISBN 978-80-553-0486-1.
28. TrebuĖa, F., *et al*., 2009. The solution for life extension of casting ladle supporting structure on the continuous casting machine 2 while taking advantage of until now available conclusions and proposals for life extension of the casting ladle supporting structure, final report, Technical University of Košice, Košice, p. 130.
29. TrebuĖa, F., Masláková, K., Frankovský, P., 2011. “Residual stress measurements,” Modelling of Mechanical and Mechatronical Systems, Proceedings of the 4th international conference, HerĐany, Slovakia, pp. 487-491.