Statistical Methods II Final Project

Rebecca Stup

BTRY 6020 Statistical Methods II, Spring, 2025

**Introduction**

The dataset used in this analysis is publicly available on Kaggle (Attakorah 2024). It contains ten variables and a million observations. For the sake of computational efficiency, this analysis was performed using a subset of for which the crop in question was corn, which contained 166,824 observations. The dependent variable was the yield in tons per hectare, and all other variables were used as predictors. Full descriptions of the variables may be found in Table 1. The goal of this analysis is to predict corn yields based on the various environmental and management factors provided in the dataset.

|  |  |
| --- | --- |
| **Variable** | **Description** |
| Yield | Total crop yield (tons per hectare) |
| Region | The geographical region (North, East, South, West) |
| Soil Type | The type of soil (Clay, Sandy, Loam, Silt, Peaty, Chalky) |
| Crop | The species of crop (Wheat, Rice, Maize, Barley, Soybean, Cotton) |
| Rainfall | The total rainfall during the growing season (mm) |
| Temperature | The average temperature during the growing season (°C) |
| Fertilizer | A binary variable indicating whether the crop was fertilized |
| Irrigation | A binary variable indicating whether the crop was irrigated |
| Weather Condition | The predominant condition during the growing season |
| Days to Harvest | The length of the growing season (days) |

**Table 1.** Descriptions of the variables used in this analysis, adapted from (Attakorah 2024).

**Methodology**

Data were analyzed in R version 4.3.1 (R Core Team, 2023) using multiple linear regression. Categorical data were encoded as factors prior to analysis. A full model containing all possible covariates was fitted in order to verify whether the data met the assumptions required for multiple linear regression. Plots of the residuals were created using the full model to verify linearity, homoscedasticity, and normality. Independence of errors was verified with a Durbin-Watson test (package “lmtest”). The Variance Inflation Factor was used to test for multicollinearity (package “car”). No transformations were applied to the data.

Variables were selected using both a backwards stepwise selection procedure with BIC and a forward stepwise selection procedure with BIC. The covariates included in the true model chosen by these procedures were then tested for significance individually through t-tests and jointly with an F-test via a one way ANOVA. The R2, adjusted R2, and RMSE values were calculated to assess the performance of the model. The model was also cross-validated using K-fold cross-validation separated into ten folds (package “boot”).

**Results**

Based on the plot of the residuals and the Q-Q plot, the assumptions of linearity, homoscedasticity, and normality were determined to be met (Fig. 1; Fig, 2). The assumptions of independence of errors (DW = 1.9983; p = 0.3662) and lack of multicollinearity (VIF = 1 for all covariates) were also determined to be met. Both the forward and backwards stepwise selection procedures returned a model with rainfall, temperature, fertilization, and irrigation as the covariates (BIC = -231217.3 for both procedures). All covariates in the true model selected by these procedures were significant on their own and jointly (p < 2.2e-16 for all p values). This model fit the data well (R2 = 0.9131; adjusted R2 = 0.9131; RMSE = 0.4999846). The total MSE resulting from k-fold cross-validation of this model was 0.2500021. Coefficients and confidence intervals for each of the covariates used in the true model are presented in Table 2.

A black speckled object with numbers

AI-generated content may be incorrect.

**Fig. 1.** Plot of the residuals of the full model.

A graph with numbers and lines

AI-generated content may be incorrect.

**Fig. 1.** Q-Q residuals plot created from the full model.

|  |  |  |
| --- | --- | --- |
| Variable | Coefficient | 95% CI |
| Rainfall | 0.005 | [0.005, 0.005] |
| Temperature | 0.020 | [0.020, 0.020] |
| Fertilization | 1.498 | [1.493, 1.503] |
| Irrigation | 1.201 | [1.196, 1.205] |

**Table 2.** Coefficients and confidence intervals for each of the covariates used in the true model.

**Discussion**

Based on the coefficients of this model we can determine that provided all other factors in the model remain the same, then yield increases by 0.005 tons per hectare for each 1 mm increase in rainfall (95% CI [0.005, 0.005]), the use of fertilizer increases yield by 1.498 tons per hectare (95% CI [1.493, 1.503]), the use of irrigation increases yield by 1.201 tons per hectare (95% CI [1.196, 1.205]), and that the yield increases by 0.0020 tons per hectare for each 1 degree increase in temperature (95% CI [0.020, 0.020]). From a practical standpoint, the results of this analysis indicate that the most relevant factors when it comes to predicting yield are the total rainfall and average temperature in an area, as well as whether or not the field was irrigated or fertilized. All of these factors are positively correlated with yield. However, it would be very strange for a farmer not to fertilize their fields, so it can be safely assumed that the overwhelming majority of data points for this factor would be “true” for the majority of datasets. Also, it would likely be more useful to look at fertilization and irrigation rates than simply whether or not these practices were implemented. Unlike rainfall and average temperature, these factors are within farmers’ control, making them especially important. It is not possible with this particular dataset, but a model that predicts increases in crop yield based off of application rates would be the most useful from a management perspective. This would be especially useful if the model was capable of determining the rate at which the increase in yield begins to plateau, so that farmers could minimize costs by not applying extraneous fertilizer.

**Conclusion**

This analysis found that the optimal predictive model for corn yield was one that used total rainfall, average temperature, and whether or not the field was irrigated and fertilized as covariates. Each of these predictors was associated with an increase in yield. Future models should include fertilization and irrigation as continuous rather than categorical variables, so that specific application rates can be taken into account.

**References**:

Attakorah, S.O. (2024, September). Agriculture Crop Yield, Version 1. Retrieved May 13, 2025 from <https://www.kaggle.com/datasets/samuelotiattakorah/agriculture-crop-yield>.

R Core Team (2023). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.