STAT 420 Project - Crop Yield

Ziquan Wang and Othon Almanza

2025-12-07

```
diagnostic_plots = function(model) {

  # 2x2 layout
  par(mfrow = c(2, 2))

  fit = fitted(model)
  res = residuals(model)

  plot(fit, res,
       xlab = "Fitted values",
       ylab = "Residuals",
       main = "Fitted vs Residuals")
  abline(h = 0, lty = 2, col = "red")

  hist(res,
       breaks = 15,
       main = "Histogram of Residuals",
       xlab = "Residuals")

  qqnorm(res,
         main = "Normal Q-Q Plot")
  qqline(res, col = "red")

  b = boxcox(
    model,
    lambda = seq(-2, 2, 0.1),
    plotit = TRUE
  )

}

diagnostic_tests = function(model) {

  # Basic quantities
  res = residuals(model)
  n = length(res)
  p = length(coef(model))

  bp = bptest(model)

  shapiro = shapiro.test(res)
  vif_values = car::vif(model)
  cat("=== Breusch-Pagan Test (Homoscedasticity) ===\n")
  print(bp)
  cat("\nInterpretation: p-value < 0.05 suggests heteroscedasticity.\n\n")

  cat("=== Shapiro-Wilk Test (Normality of Residuals) ===\n")
  print(shapiro)
  cat("\nInterpretation: p-value < 0.05 suggests non-normal residuals.\n\n")

  cat("=== Variance Inflation Factor (VIF) ===\n")
  print(vif_values)
  cat("\nInterpretation: VIF > 5 suggests moderate collinearity; VIF > 10 suggests high multicollineari

  return(list(
    bp_test = bp,
    shapiro_test = shapiro,
    vif = vif_values,
    thresholds = list(
      vif = 5
    )
  ))
}
```

Introduction

Crop yield prediction is an important aspect of agricultural planning that has received greater interest with the advent of machine learning. Using a dataset of crop yields, we will analyze numeric and categorical predictors in hope of uncovering patterns that correlate with yield mass. The insight gained from a crop yield analysis can be beneficial in numerous way: It can assist in optimizing farming techniques to maximize crop output. It can also be used to predict crop yields within given conditions, assist in site suitability assessments, and prepare for long-term changes in climate patterns.

Methods

Data Cleaning # added your jouney as well

The dataset was acquired from Kaggle and was originally compiled form the Food and Agriculture Organization and the World Bank. Each observation represents a crop yield measured in hectograms per hectare, with accompanying data including the type of crop, country of production, average rainfall, pesticide tonnes, and average temperature.

The first step in cleaning the dataset was to remove an index column and identify duplicate observations. Duplicates were found in the dataset and removed. Further, there were duplicates in the data with different temperature values but identical values for all other fields. These instances were each aggregated into a single observation containing the average temperature for the duplicate records. Records with null values were also removed.

Additional datasets were identified and joined with the core data to supplement the analysis. Country-level GDP data was acquired from the World Bank since economic well-being could conceivably improve farming practices and efficiency. Countries were divided into low, middle, and high-income countries using quantile breaks to analyze as a categorical predictor. Additionally, we were concerned about the high granularity of countries as a predictor. To enhance the interpretability of other predictors, we instead categorized countries into subregions by joining this category from a United Nations geoscheme dataset. For the GDP and subregions joins, a few country names had to be manually renamed (e.g. renaming 'United Kingdom of Great Britain and Northern Ireland' to 'United Kingdom').

Finally, we decided to focus on a single crop for our analysis since different crops are affected differently by the predictors we analyzed. Potatoes had the highest number of observations in our dataset. As a staple in much of the world, it also had the highest geographic distribution. Our final steps in data cleaning involved narrowing the working data to the potato subset, removing the unused country and crop columns, and renaming the remaining columns for usability. The final output was written into its own .csv file.

Exploratory Data Analysis # added your jouney as well

Our first step in data exploration involved using pairs() and cor() to identify any trends and obvious collinearity issues. No collinearity issues were identified, but a pattern between pesticides and yield output was observed that might benefit from a predictor-level transformation.

We then created a full additive model using yield as a function of the remaining predictors. This achieved a baseline adjusted r-squared of 0.701. We also created a null model that omitted the subregions predictor to evaluate whether it was needed in our final model. This model resulted in a significantly lower adjusted r-squared of 0.464. An ANOVA test confirmed this trend. Using any reasonable significance level resulted in an outcome that rejects the null hypothesis in favor of including subregions. Inlucing subregions

Conlusion discuss on Sunday

Appendix