

STAT 420 Project - Crop Yield

Ziquan Wang and Othon Almanza

2025-12-08

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 analyzed numeric and categorical predictors to uncover patterns that correlate with yield mass. The insight gained from a crop yield analysis can be beneficial in numerous ways: 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.

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

Methods

Data Cleaning

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. Observations with null values were also removed. Curiously, no data for the year 2003 was provided.

Additional datasets were identified and joined with the core data to supplement the analysis. Country-level gross domestic product (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 to match the core dataset (e.g. renaming ‘The Bahamas’ to ‘Bahamas’).

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

Our first step in data exploration involved using `cor()` and `pairs()` to identify any trends and obvious collinearity issues (see figure 5 of the appendix). Right away, we found patterns emerging. A strong correlation is

seen between GDP and yield, indicating that economic prosperity likely does affect crop output positively. A strong inverse correlation was found between temperature and GDP, illustrating known differences in economic development between the Global North and the Global South. We also see crop yields trending upward over the years, likely due to advances in technology and farming techniques.

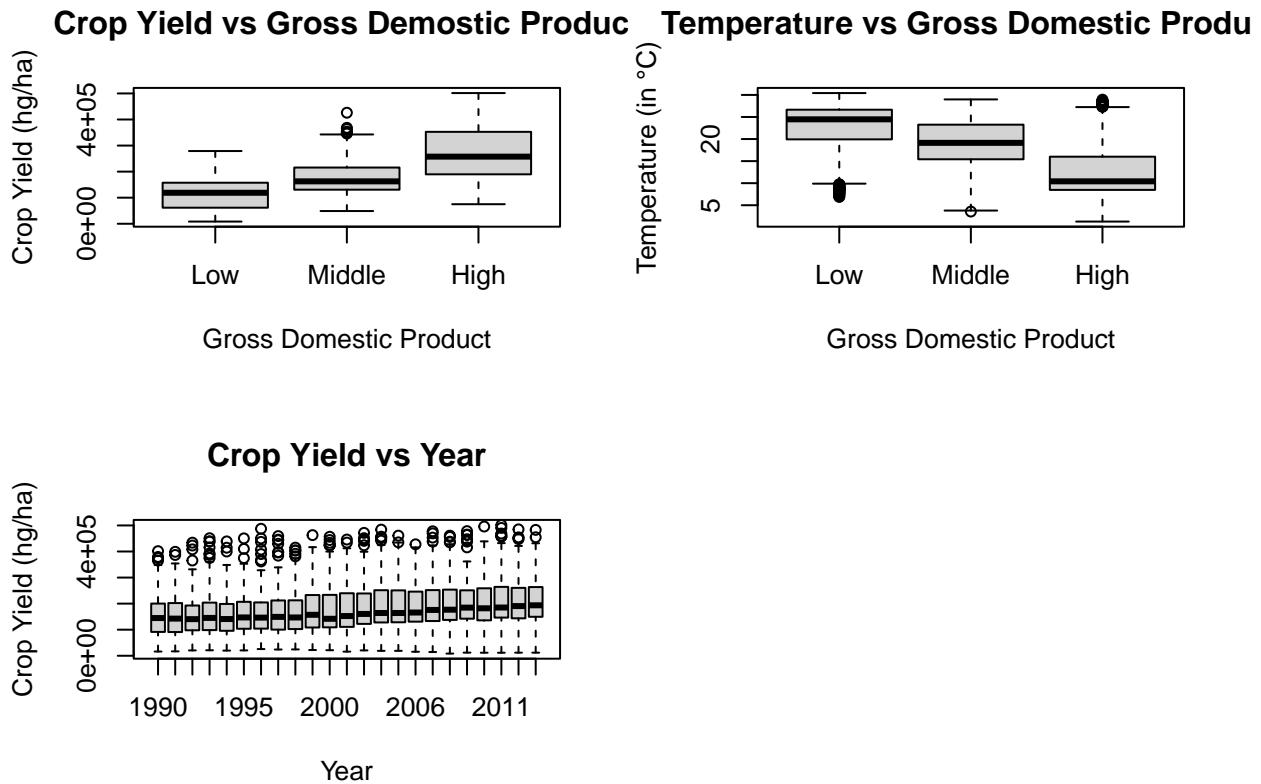


Figure 1: Trends and patterns box plots

Additionally, pairs() uncovered a pattern between pesticides and yield output that might benefit from one or more predictor-level transformations. Temperature might also benefit from a transformation.

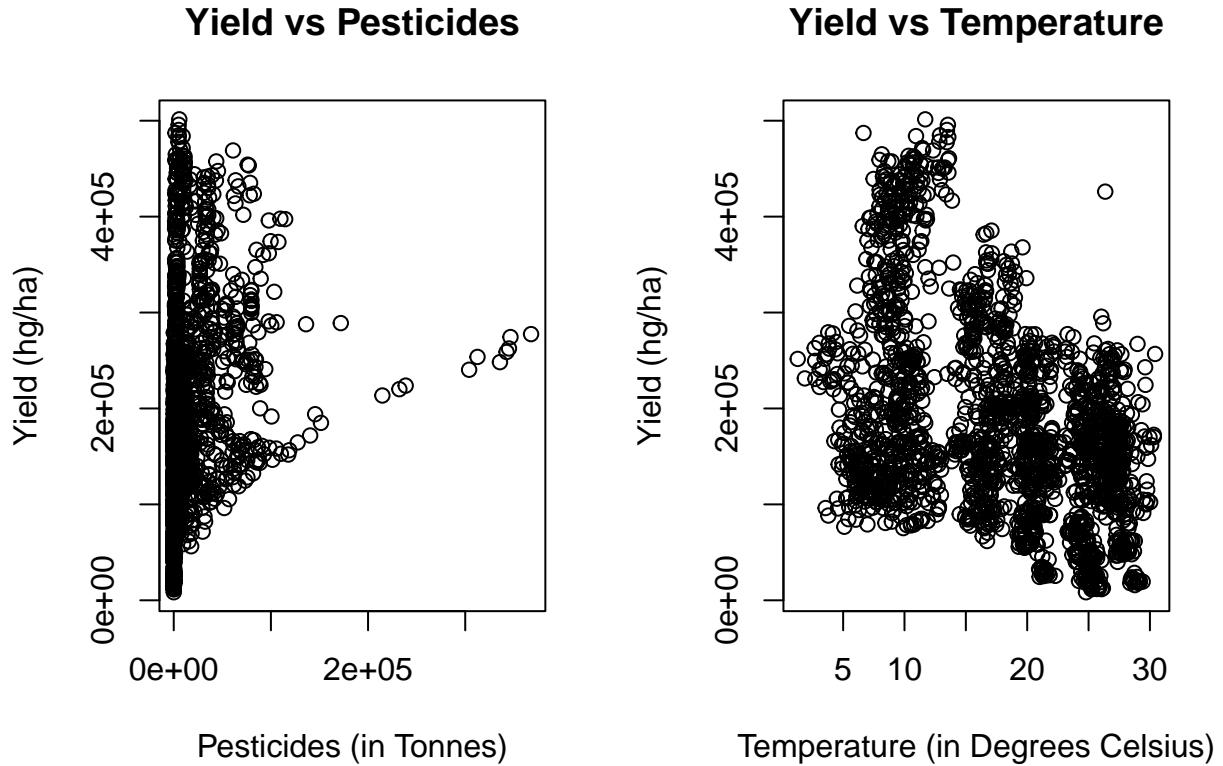


Figure 2: Candidates for predictor transformations

Full Linear Model

After exploring the cleaned data, we created a full additive model using yield as a function of the remaining predictors. This achieved a baseline adjusted r-squared of 0.701. The summary output is shown below.

```
##
## Call:
## lm(formula = yield_hg_ha ~ ., data = df_potatoes)
##
## Residuals:
##    Min     1Q   Median     3Q    Max 
## -172744 -35526   -1720   34694  217250 
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)             -4.220e+06  3.470e+05 -12.161 < 2e-16 ***
## year                      2.307e+03  1.735e+02  13.294 < 2e-16 ***
## rain_mm                   7.978e+00  3.009e+00   2.651 0.00808 ** 
## pesticides_t                 5.770e-01  4.787e-02  12.052 < 2e-16 ***
## temp_c                     -2.173e+03  4.189e+02  -5.187 2.35e-07 ***
## gdp_tierLow                -9.424e+04  6.359e+03 -14.820 < 2e-16 ***
## gdp_tierMiddle              -5.124e+04  4.781e+03 -10.719 < 2e-16 ***
```

```

## subregionCaribbean      -1.381e+05  1.314e+04 -10.515 < 2e-16 ***
## subregionCentral America -1.232e+05  1.128e+04 -10.926 < 2e-16 ***
## subregionCentral Asia   -1.652e+05  1.313e+04 -12.578 < 2e-16 ***
## subregionEastern Africa -1.627e+05  1.161e+04 -14.014 < 2e-16 ***
## subregionEastern Asia   -1.059e+05  1.446e+04 -7.323 3.47e-13 ***
## subregionEastern Europe -1.920e+05  1.021e+04 -18.814 < 2e-16 ***
## subregionMelanesia     -2.261e+05  1.707e+04 -13.248 < 2e-16 ***
## subregionMiddle Africa -2.177e+05  1.365e+04 -15.945 < 2e-16 ***
## subregionNorthern Africa -1.181e+05  1.179e+04 -10.012 < 2e-16 ***
## subregionNorthern America -2.868e+05  1.460e+04 -19.649 < 2e-16 ***
## subregionNorthern Europe -1.180e+05  9.601e+03 -12.292 < 2e-16 ***
## subregionSouth-eastern Asia -1.458e+05  1.416e+04 -10.294 < 2e-16 ***
## subregionSouth America   -1.848e+05  1.077e+04 -17.148 < 2e-16 ***
## subregionSouthern Africa -9.758e+04  1.240e+04 -7.868 5.74e-15 ***
## subregionSouthern Asia   -1.272e+05  1.265e+04 -10.050 < 2e-16 ***
## subregionSouthern Europe -1.889e+05  9.291e+03 -20.330 < 2e-16 ***
## subregionWestern Africa   -1.272e+05  1.402e+04 -9.071 < 2e-16 ***
## subregionWestern Asia     -1.400e+05  1.062e+04 -13.181 < 2e-16 ***
## subregionWestern Europe  -1.080e+04  9.773e+03 -1.105  0.26944
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 55280 on 2065 degrees of freedom
## Multiple R-squared:  0.7046, Adjusted R-squared:  0.701
## F-statistic:    197 on 25 and 2065 DF,  p-value: < 2.2e-16

```

With an r-squared of 0.7046, our baseline model shows a strong correlation between our predictors and the yield response. This coefficient of determination indicates that 70.46% of the observed variation in yield can be explained by the full additive linear model.

We can also see patterns emerge from the model's coefficients. First, we observe that year is positively correlated with crop yield. When all other predictors are equal, each additional year results in a yield increase of 2,300 hg/ha. We see that rain has a small but positive correlation with crop yield, with an increase of 7.9 hg/ha for every additional millimeter of rainfall. Pesticides are also correlated with increased crop yield, with an increase in 0.6 hg/ha for each additional tonne of pesticides. Temperature reverses these trends, with yield decreases of 2,200 hg/ha for every increase in degrees Celsius, illustrating potato's suitability for cooler climates. The GDP tiers reinforce what we had seen during the data exploration process: The base factor level of high-GDP corresponds with the highest crop yields. The intercept is reduced for middle-GDP observations and reduced even further for low-GDP observations. When we look at the subregion factor, we see that there are small but statistically significant variations among the different subregions. The notable exception to this is Western Europe, which has a high p-value of 0.269. This may be due to the inclusion of an Eastern Europe subregion, which likely shares many commonalities with Western Europe regarding climate conditions, economic output, and farming practices. This high p-value indicates the statistical insignificance on the potato yield considering the difference among these two regions as the variable.

In the appendix, we have outputted diagnostics for the full additive model. In figure 6, we see that variance is not equally distributed, but it is not grossly violated either. The residuals seem to follow a relatively normal distribution. Normality in the Q-Q plot looks acceptable. The model could benefit from a Box-Cox transformation. When we run a Breusch-Pagan test of homoscedasticity and a Shapiro-Wilk test of normality, the results suggest that we must reject the assumptions of equal variance and normality. We will try to reach a model that improves on these diagnostics.

We can further validate the full model using leave-one-Out cross-validation (LOOCV). Using hat values to conduct LOOCV, we attain a root mean square error (RMSE) of 73,268.77 hg/ha. This is only slightly

larger than the RMSE of the full additive model, 73,154.80 hg/ha, indicating a favorable result. The LOOCV RMSE is necessarily larger than the model's RMSE, but in this case only slightly so.

Other Methods

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. Including subregions was thus taken as a good compromise to build a stronger model without needing to classify crop yields at the country level.

Having identified pesticides as a predictor that might benefit from a transformation, we first observe that log transforming pesticides seems to improve its linear relationship with yield size. We can see the twice transformed pesticides predictor scattered more evenly around the red line after squaring the log transformation.

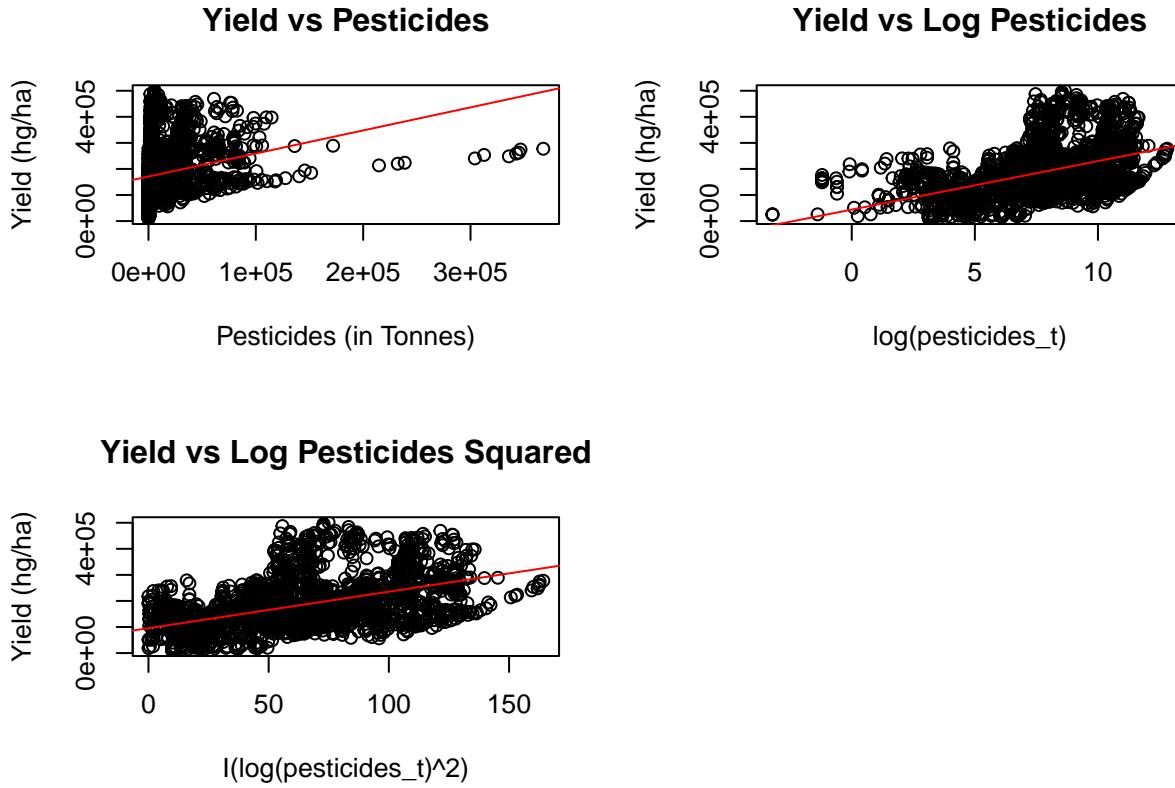


Figure 3: Predictor transformations on pesticides

A quadratic transformation was applied to temperature). The result is less significant than the transformation of pesticides predictor. However, the transformed temperature predictor scattered evenly around the red line as well (see figure 7 in the appendix).

We created an additive model included all predictors and also incorporated predictor-level transformations for pesticides and temperature. We chose to keep only the transformed predictors in the model fitting because the clearly nonlinear relationship between yield and these transformed predictors. We attempted a backward BIC search to see if the model could be reduced from here, but perhaps unsurprisingly, no predictors were dropped. The Adjusted R-squared improved to 0.7457 against 0.701 before transformation. This indicated

the predictor transformation helped improving the model. The model returns an RMSE of 50,668.60 hg/ha and a LOOCV-RMSE of 51,263.41 hg/ha, an significant improvement from the full additive model.

Having chosen the predictor transformations to incorporate, we then created a model that created two-way interactions among rain, pesticides, temperature, and GDP. Year and region were kept in the model but not considered for two-way interactions because the subregion could explode the interactive term and make it hard to interpret and it is unlikely Year predictor interact with other predictors observing its relationship with other predictors the plot generated by Pairs() function. This created a large model appropriate for BIC backward search. The search was successful, and the model was reduced to a manageable size. The Adjusted R-squared improved to 0.7643 against 0.7457 before adding interaction terms. The model returns an RMSE of 48,705.68 hg/ha and a LOOCV-RMSE of 49,403.77 hg/ha, an improvement from the full additive transformed model. Additionally, variance inflation factors (VIF's) do not show problematic covariance within the model. Interestingly, the gdp_tier alone became insignificant and the interactive term with gdp_tier are significant, which indicated the gdp_tier's effect on the yield depends on other factors.

Finally, we used boxcox() to find an appropriate Box-Cox transformation on the response. A lambda value of 0.707 was identified and used. However this is where the complication happens. After the response transformation, the Adjusted R-squared dropped from 0.7643 to 0.7477 indicates less variance response transformed model explains. They have a disagreement on the significance of term $I(temp_c^2)$. In the response transformed model, the term $I(temp_c^2)$'s statistical level was downgraded.

We decided to leverage LASSO regression predictors selection technique to see which model LASSO regression agrees more on the term $I(temp_c^2)$. The LASSO regression was given the same predictor combination. When I choose the lambda value at 0.00375 reducing the betas of some predictors, the term $I(temp_c^2)$ appeared to be zero which indicated it is a weaker predictor than the remaining predictors. Thus we favor the response transformed model more to be selected as our final model.

LASSO Regression Coef. Shrink Chart

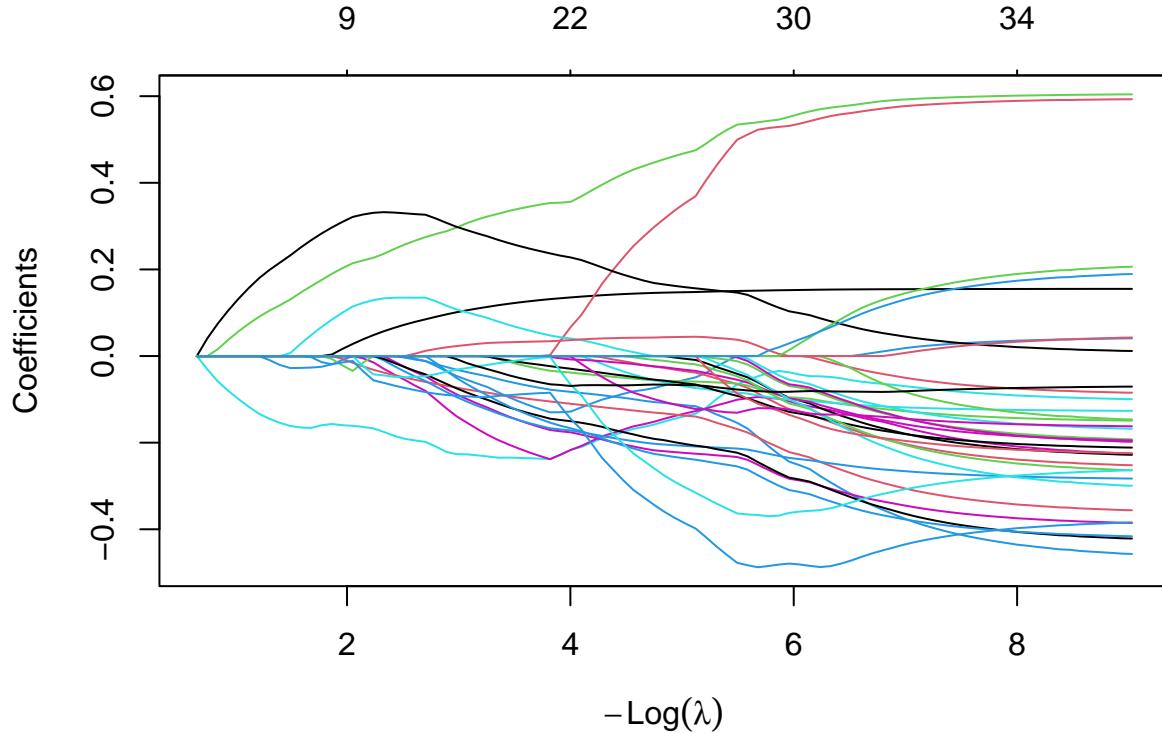


Figure 4: LASSO regression coefficient shrink

Additionally, by comparing the diagnostic plots, the response transformed model appeared to have more evenly distributed fitted vs residuals plot and its histogram of residuals appeared to be more following the normal distribution curve. With these observation and efforts, we chose response transformed model to be our final model.

Conclusion

Our final model achieves an r-squared of 0.7477, accounting for a larger portion of the response's observed variation compared to the full additive model's r-squared of 0.7046. We see improvements in other diagnostics as well: The model's RMSE is improved from 73,154.80 to 48,705.68. It also shows a favorable LOOCV-RMSE of 49,403.77, suggesting that the model is not compromised by under or over-fitting. Additionally, normality is considerably increased in the final model from a Shapiro-Wilk normality p-value of 0.003364 to a p-value of 0.0209, as seen under figure 11 of the appendix. We also see a more even distribution of the residuals compared to the full additive model. The final model indicates year, rain, pesticides and regions are individually significant predictor even at $p = 0.001$ significance level with regression coefficients of 44.27, 1.7 and 33 for numerical predictors.

This model reflects many of the trends that we uncovered previously, albeit with higher granularity resulting from interactions. The model still shows that crop yields are positively correlated with year, presumably reflecting advances in farming techniques and technology. Rain still has a slight but positive correlation with yield. Pesticides also remain positively correlated with yield, but interestingly, low-GDP countries see a diminishing return when compared to higher-income countries. Perhaps this is due to the use of lower-quality pesticides. We also see diminishing returns from pesticides when rainfall is considered. It could stand to reason that rainfall washes away pesticides, reducing their effectiveness.

Overall, We believe the updated model is a stronger model for prediction, as well as a sounder model for explanation.

Future Actions

Regarding the missing 2003 data, we could either find these missing data or interpolate the missing points to see how it affects the model.

Through research, LASSO regression algorithms were identified that are out of this class's scope to better identify the significance of predictors.

Appendix

Core Data Source

Crop yield data: <https://www.kaggle.com/datasets/patelris/crop-yield-prediction-dataset/discussion/293030>

Additional Data Sources

GDP data: <https://data.worldbank.org/indicator/NY.GDP.PCAP.CD>

UN geoscheme subregions: <https://unstats.un.org/unsd/methodology/m49/>

Data Exploration using Covariance and Pairs() function

```
round(cor(df_potatoes_num), 2)
```

```
##          year yield_hg_ha rain_mm pesticides_t temp_c gdp_tier subregion
## year      1.00     0.17    -0.02       0.07   0.01   0.00     0.02
## yield_hg_ha  0.17     1.00    -0.10       0.26  -0.42   0.63     0.19
## rain_mm     -0.02    -0.10     1.00       0.09   0.26  -0.18   -0.24
## pesticides_t  0.07     0.26     0.09       1.00  -0.06   0.21     0.09
## temp_c       0.01    -0.42     0.26      -0.06   1.00  -0.53   -0.08
## gdp_tier     0.00     0.63    -0.18       0.21  -0.53   1.00     0.21
## subregion    0.02     0.19    -0.24       0.09  -0.08   0.21    1.00
```

```
pairs(df_potatoes_num)
```

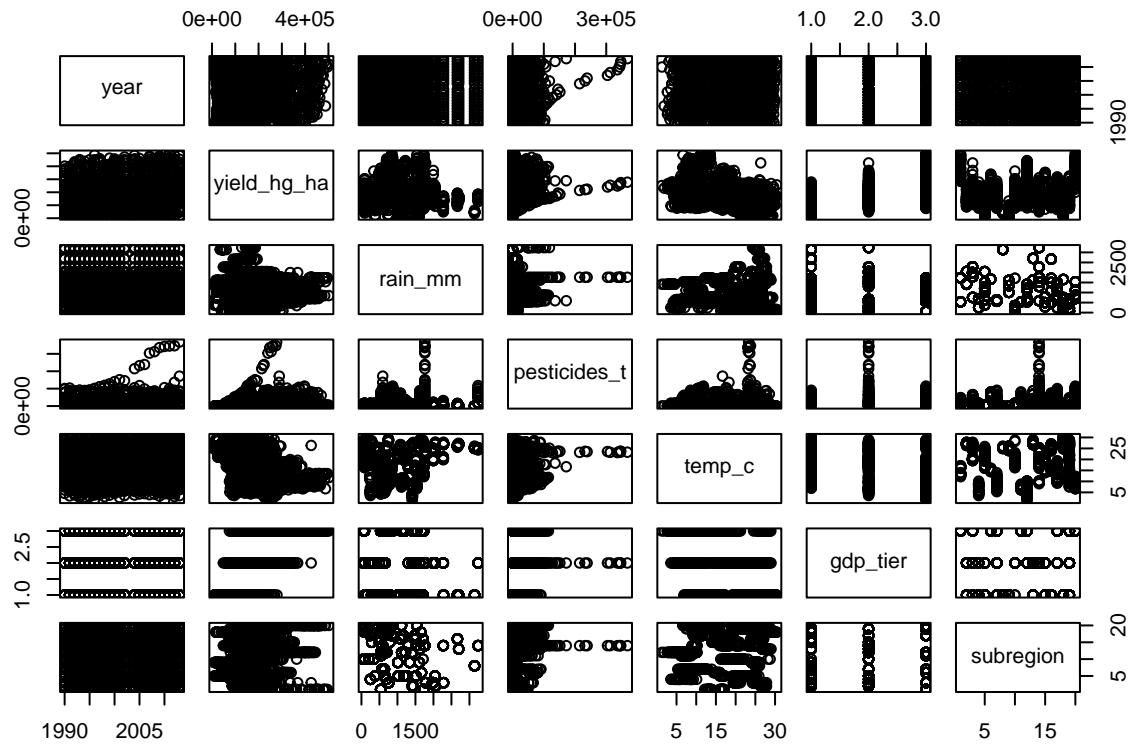


Figure 5: Pair scatterplots

Fitting and diagnosing the full model

```
add_model = lm(yield_hg_ha ~ ., data = df_potatoes)

diagnostic_plots(add_model)
```

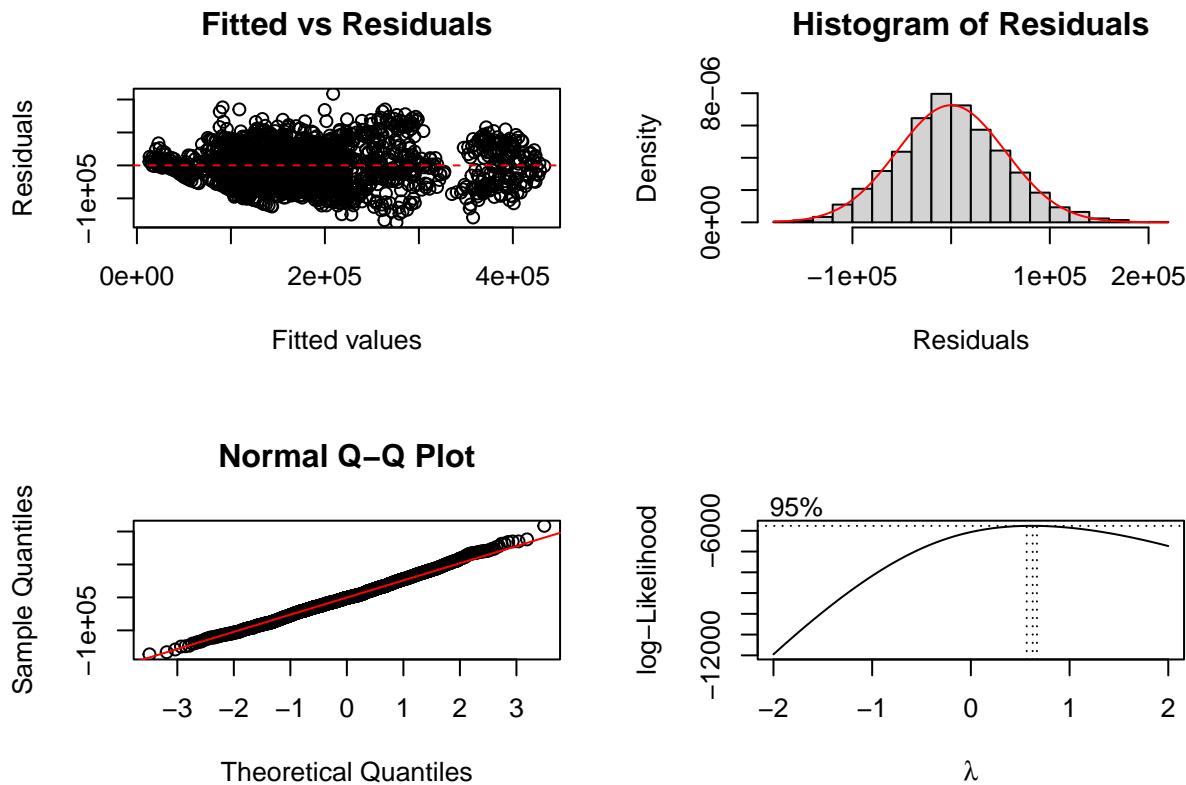


Figure 6: Diagnostic plots for full additive model

```
diagnostic_tests(add_model)

## === Breusch-Pagan Test (Homoscedasticity) ===
##
## studentized Breusch-Pagan test
##
## data: model
## BP = 523.89, df = 25, p-value < 2.2e-16
##
## Interpretation: p-value < 0.05 suggests heteroscedasticity.
##
##
## === Shapiro-Wilk Test (Normality of Residuals) ===
##
## Shapiro-Wilk normality test
##
## data: res
## W = 0.99765, p-value = 0.003364
```

```

## 
## Interpretation: p-value < 0.05 suggests non-normal residuals.
## 
## 
## === Variance Inflation Factor (VIF) ===
## 
##          GVIF Df GVIF^(1/(2*Df))
## year      1.011921  1     1.005943
## rain_mm   3.005037  1     1.733504
## pesticides_t 1.395090  1     1.181139
## temp_c    6.425042  1     2.534767
## gdp_tier  9.093351  2     1.736525
## subregion 123.163719 19    1.135044
## 
## Interpretation: VIF > 5 suggests moderate collinearity; VIF > 10 suggests high multicollinearity.

```

```
loocv(add_model)
```

```

## === LOOCV-RMSE ===
## 
##          RMSE LOOCV_RMSE
## 54938.25 55557.30
## 
## Interpretation: LOOCV-RMSE significantly larger than RMSE may indicate under or over-fitting

```

Testing subregion significance

```

add_nosubreg = lm(yield_hg_ha ~ . - subregion, data = df_potatoes)
anova(add_nosubreg, add_model)

```

```

## Analysis of Variance Table
## 
## Model 1: yield_hg_ha ~ (year + rain_mm + pesticides_t + temp_c + gdp_tier +
## subregion) - subregion
## Model 2: yield_hg_ha ~ year + rain_mm + pesticides_t + temp_c + gdp_tier +
## subregion
##   Res.Df       RSS Df  Sum of Sq      F    Pr(>F)
## 1    2084 1.1418e+13
## 2    2065 6.3111e+12 19 5.1072e+12 87.952 < 2.2e-16 ***
## --- 
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Temperature Predictor transformation

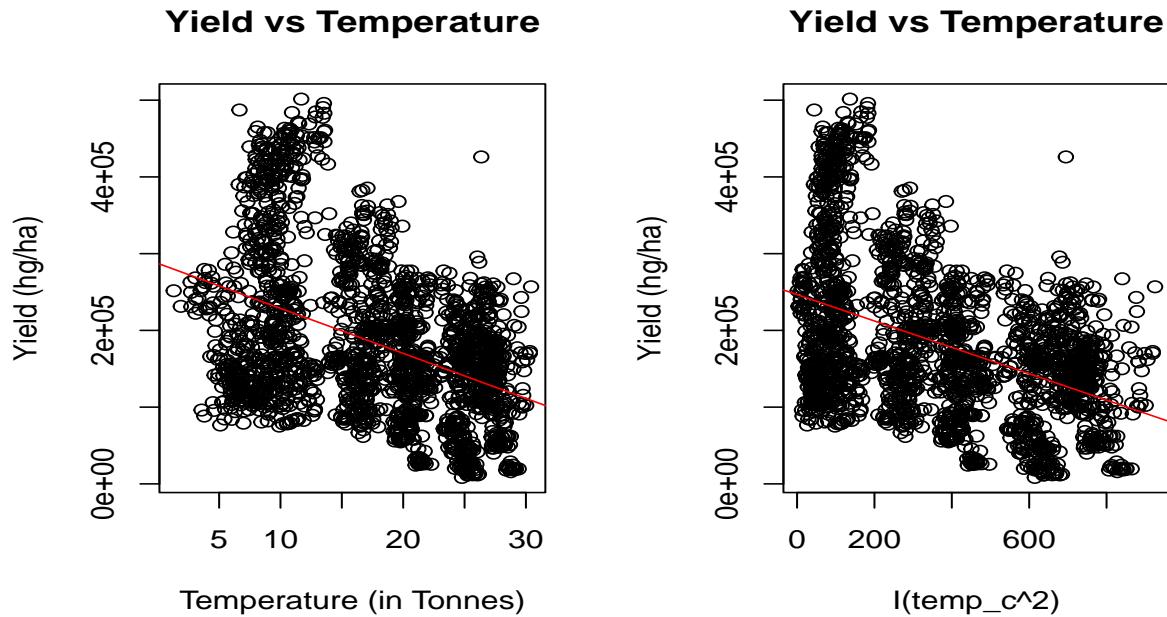


Figure 7: Predictor transformation on temperature

Fit the model with all transformed predictors and BIC model reduction

```

tran_model = lm(yield_hg_ha ~ year + rain_mm + I((log(pesticides_t))^2)
                + I(temp_c^2) + gdp_tier + subregion, data = df_potatoes)

summary(step(tran_model, direction = "backward", k=log(length(resid(tran_model))), trace = 0))

##
## Call:
## lm(formula = yield_hg_ha ~ year + rain_mm + I((log(pesticides_t))^2) +
##     I(temp_c^2) + gdp_tier + subregion, data = df_potatoes)
##
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -150402  -32220   -1718   31838  226544 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -4.041e+06  3.204e+05 -12.613 < 2e-16 ***
## year         2.166e+03  1.603e+02  13.513 < 2e-16 ***
## rain_mm      1.113e+01  2.773e+00   4.012 6.23e-05 ***
## I((log(pesticides_t))^2) 1.044e+03  4.705e+01  22.195 < 2e-16 ***
## I(temp_c^2)  -6.624e+01  1.033e+01  -6.410 1.79e-10 ***
## gdp_tierLow  -7.194e+04  6.025e+03 -11.941 < 2e-16 ***
## gdp_tierMiddle -5.063e+04  4.438e+03 -11.409 < 2e-16 ***
## subregionCaribbean -1.031e+05  1.243e+04  -8.295 < 2e-16 ***
## subregionCentral America -1.216e+05  1.049e+04 -11.592 < 2e-16 ***
## subregionCentral Asia   -1.449e+05  1.206e+04 -12.018 < 2e-16 ***

```

```

## subregionEastern Africa -1.368e+05 1.071e+04 -12.773 < 2e-16 ***
## subregionEastern Asia -1.157e+05 1.323e+04 -8.742 < 2e-16 ***
## subregionEastern Europe -1.896e+05 9.344e+03 -20.290 < 2e-16 ***
## subregionMelanesia -1.916e+05 1.582e+04 -12.113 < 2e-16 ***
## subregionMiddle Africa -1.773e+05 1.275e+04 -13.904 < 2e-16 ***
## subregionNorthern Africa -9.804e+04 1.088e+04 -9.009 < 2e-16 ***
## subregionNorthern America -2.914e+05 1.324e+04 -22.004 < 2e-16 ***
## subregionNorthern Europe -8.718e+04 8.697e+03 -10.024 < 2e-16 ***
## subregionSouth-eastern Asia -1.447e+05 1.340e+04 -10.804 < 2e-16 ***
## subregionSouth America -1.803e+05 9.903e+03 -18.211 < 2e-16 ***
## subregionSouthern Africa -6.564e+04 1.151e+04 -5.703 1.35e-08 ***
## subregionSouthern Asia -1.282e+05 1.181e+04 -10.851 < 2e-16 ***
## subregionSouthern Europe -1.749e+05 8.587e+03 -20.374 < 2e-16 ***
## subregionWestern Africa -7.798e+04 1.342e+04 -5.812 7.12e-09 ***
## subregionWestern Asia -9.977e+04 1.004e+04 -9.942 < 2e-16 ***
## subregionWestern Europe -3.597e+03 8.871e+03 -0.406 0.685
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 50990 on 2065 degrees of freedom
## Multiple R-squared: 0.7487, Adjusted R-squared: 0.7457
## F-statistic: 246.1 on 25 and 2065 DF, p-value: < 2.2e-16

```

```
loocv(tran_model)
```

```

## === LOOCV-RMSE ===
##
## RMSE LOOCV_RMSE
## 50668.60 51263.41
##
## Interpretation: LOOCV-RMSE significantly larger than RMSE may indicate under or over-fitting

```

Fit the model with interactive transformed predictors and BIC model reduction

```

int_model = lm(yield_hg_ha ~ year + (rain_mm + I((log(pesticides_t))^2) + I(temp_c^2)
                                         + gdp_tier)^2 + subregion , data = df_potatoes)

bic_model = step(int_model, direction = "backward", k=log(length(resid(int_model))), trace = 0)

summary(bic_model)

```

```

##
## Call:
## lm(formula = yield_hg_ha ~ year + rain_mm + I((log(pesticides_t))^2) +
##     I(temp_c^2) + gdp_tier + subregion + rain_mm:I((log(pesticides_t))^2) +
##     rain_mm:gdp_tier + I((log(pesticides_t))^2):I(temp_c^2) +
##     I((log(pesticides_t))^2):gdp_tier, data = df_potatoes)
##
## Residuals:
##      Min      1Q  Median      3Q      Max 
## -150312 -32010 -2149   31530  213912 
## 
```

```

## Coefficients:
## (Intercept)                               Estimate Std. Error t value Pr(>|t|)
## year                                     2.235e+03  1.547e+02 14.446 < 2e-16
## rain_mm                                    9.563e+01  8.128e+00 11.765 < 2e-16
## I((log(pesticides_t))^2)                  1.770e+03  1.046e+02 16.931 < 2e-16
## I(temp_c^2)                                5.201e+01  1.467e+01  3.545 0.000402
## gdp_tierLow                                1.602e+04  1.088e+04  1.473 0.140942
## gdp_tierMiddle                             -1.728e+04  1.034e+04 -1.672 0.094665
## subregionCaribbean                         -1.247e+05  1.272e+04 -9.804 < 2e-16
## subregionCentral America                   -1.105e+05  1.052e+04 -10.500 < 2e-16
## subregionCentral Asia                      -1.136e+05  1.232e+04 -9.224 < 2e-16
## subregionEastern Africa                   -1.375e+05  1.079e+04 -12.738 < 2e-16
## subregionEastern Asia                     -1.321e+05  1.403e+04 -9.417 < 2e-16
## subregionEastern Europe                  -1.581e+05  1.003e+04 -15.758 < 2e-16
## subregionMelanesia                        -2.035e+05  1.769e+04 -11.506 < 2e-16
## subregionMiddle Africa                   -1.963e+05  1.294e+04 -15.171 < 2e-16
## subregionNorthern Africa                 -7.850e+04  1.086e+04 -7.230 6.76e-13
## subregionNorthern America                -2.624e+05  1.366e+04 -19.216 < 2e-16
## subregionNorthern Europe                 -5.090e+04  9.086e+03 -5.602 2.40e-08
## subregionSouth-eastern Asia              -1.295e+05  1.361e+04 -9.516 < 2e-16
## subregionSouth America                  -1.566e+05  1.092e+04 -14.342 < 2e-16
## subregionSouthern Africa                -4.843e+04  1.149e+04 -4.214 2.62e-05
## subregionSouthern Asia                  -1.094e+05  1.180e+04 -9.268 < 2e-16
## subregionSouthern Europe                -1.548e+05  8.507e+03 -18.192 < 2e-16
## subregionWestern Africa                 -1.172e+05  1.362e+04 -8.602 < 2e-16
## subregionWestern Asia                  -7.338e+04  1.036e+04 -7.085 1.90e-12
## subregionWestern Europe                 1.150e+04  8.799e+03  1.307 0.191434
## rain_mm:I((log(pesticides_t))^2)        -4.132e-01  7.195e-02 -5.743 1.07e-08
## rain_mm:gdp_tierLow                    -8.040e+01  7.739e+00 -10.390 < 2e-16
## rain_mm:gdp_tierMiddle                 -5.484e+01  7.573e+00 -7.242 6.23e-13
## I((log(pesticides_t))^2):I(temp_c^2)    -8.346e-01  1.881e-01 -4.437 9.62e-06
## I((log(pesticides_t))^2):gdp_tierLow     -3.382e+02  1.115e+02 -3.034 0.002440
## I((log(pesticides_t))^2):gdp_tierMiddle   1.261e+02  1.113e+02  1.134 0.257124
##
## (Intercept)                               ***
## year                                     ***
## rain_mm                                    ***
## I((log(pesticides_t))^2)                  ***
## I(temp_c^2)                                ***
## gdp_tierLow                                .
## gdp_tierMiddle                             ***
## subregionCaribbean                         ***
## subregionCentral America                   ***
## subregionCentral Asia                      ***
## subregionEastern Africa                   ***
## subregionEastern Asia                     ***
## subregionEastern Europe                  ***
## subregionMelanesia                        ***
## subregionMiddle Africa                   ***
## subregionNorthern Africa                 ***
## subregionNorthern America                ***
## subregionNorthern Europe                 ***
## subregionSouth-eastern Asia              ***

```

```

## subregionSouth America      ***
## subregionSouthern Africa    ***
## subregionSouthern Asia      ***
## subregionSouthern Europe    ***
## subregionWestern Africa     ***
## subregionWestern Asia       ***
## subregionWestern Europe     ***
## rain_mm:I((log(pesticides_t))^2)   ***
## rain_mm:gdp_tierLow          ***
## rain_mm:gdp_tierMiddle        ***
## I((log(pesticides_t))^2):I(temp_c^2) *** 
## I((log(pesticides_t))^2):gdp_tierLow   **
## I((log(pesticides_t))^2):gdp_tierMiddle
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 49080 on 2059 degrees of freedom
## Multiple R-squared:  0.7678, Adjusted R-squared:  0.7643
## F-statistic: 219.7 on 31 and 2059 DF,  p-value: < 2.2e-16

```

```
car::vif(bic_model)
```

```

## there are higher-order terms (interactions) in this model
## consider setting type = 'predictor'; see ?vif

```

	GVIF	Df	GVIF^(1/(2*Df))
## year	1.020791	1	1.010342
## rain_mm	27.817550	1	5.274235
## I((log(pesticides_t))^2)	11.877943	1	3.446439
## I(temp_c^2)	12.169802	1	3.488524
## gdp_tier	226.266759	2	3.878423
## subregion	2399.835205	19	1.227304
## rain_mm:I((log(pesticides_t))^2)	17.738319	1	4.211688
## rain_mm:gdp_tier	219.693163	2	3.849942
## I((log(pesticides_t))^2):I(temp_c^2)	10.996190	1	3.316050
## I((log(pesticides_t))^2):gdp_tier	69.596834	2	2.888334

```
diagnostic_plots(bic_model)
```

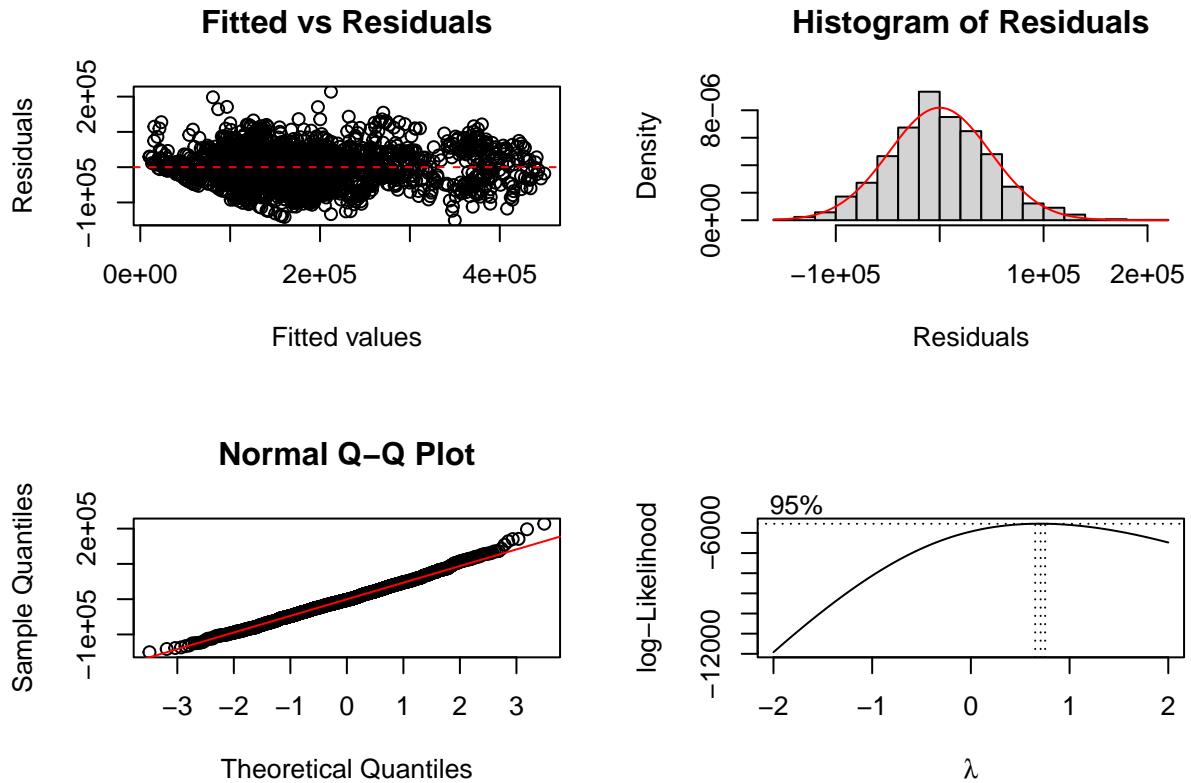


Figure 8: Diagnostic plots for BIC model

```
looocv(bic_model)
```

```
## === LOOCV-RMSE ===  
##  
##      RMSE LOOCV_RMSE  
##    48705.68    49403.77  
##  
## Interpretation: LOOCV-RMSE significantly larger than RMSE may indicate under or over-fitting
```

Using a Box-Cox transformation on the yield response

```
boxcox_lambda(bic_model)
```

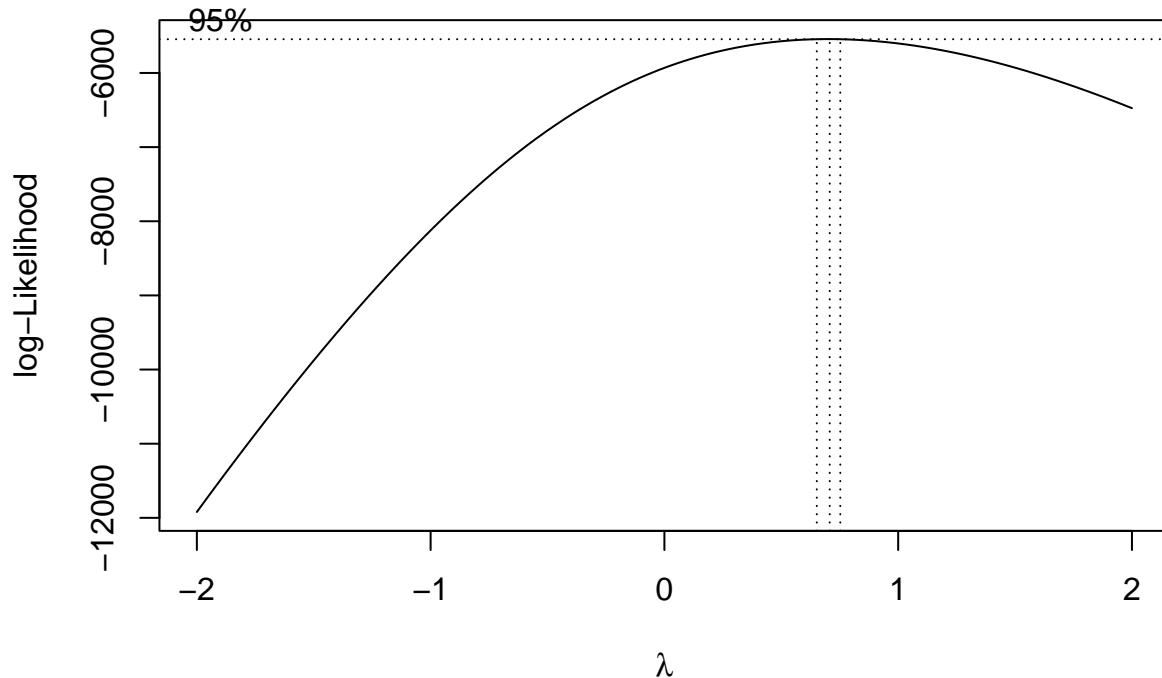


Figure 9: Lambda for Box-Cox transformation of yield response

```
## Box-Cox Lambda
##      0.7070707
```

Final model and diagnostics

```
final_model = lm(yield_hg_ha^.707 ~ year + rain_mm + I((log(pesticides_t))^2) + I(temp_c^2) +
                  gdp_tier + subregion + rain_mm:I((log(pesticides_t))^2) +
                  rain_mm:gdp_tier + I((log(pesticides_t))^2):I(temp_c^2) +
                  I((log(pesticides_t))^2):gdp_tier, data = df_potatoes)

summary(final_model)

##
```

```
## Call:
## lm(formula = yield_hg_ha^0.707 ~ year + rain_mm + I((log(pesticides_t))^2) +
##     I(temp_c^2) + gdp_tier + subregion + rain_mm:I((log(pesticides_t))^2) +
##     rain_mm:gdp_tier + I((log(pesticides_t))^2):I(temp_c^2) +
```

```

##      I((log(pesticides_t))^2):gdp_tier, data = df_potatoes)
##
## Residuals:
##      Min     1Q Median     3Q    Max
## -3496.5 -646.6   -16.6  664.8 3961.8
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -8.382e+04  6.450e+03 -12.996 < 2e-16
## year                      4.427e+01  3.219e+00  13.753 < 2e-16
## rain_mm                    1.726e+00  1.691e-01  10.209 < 2e-16
## I((log(pesticides_t))^2)   3.301e+01  2.175e+00  15.175 < 2e-16
## I(temp_c^2)                 6.042e-01  3.053e-01   1.979 0.047908
## gdp_tierLow                 6.794e+01  2.264e+02   0.300 0.764076
## gdp_tierMiddle              -4.796e+02  2.150e+02  -2.231 0.025818
## subregionCaribbean          -2.033e+03  2.646e+02  -7.683 2.39e-14
## subregionCentral America    -1.833e+03  2.189e+02  -8.375 < 2e-16
## subregionCentral Asia       -2.032e+03  2.563e+02  -7.929 3.58e-15
## subregionEastern Africa     -2.547e+03  2.246e+02  -11.342 < 2e-16
## subregionEastern Asia       -2.282e+03  2.919e+02  -7.819 8.45e-15
## subregionEastern Europe     -2.902e+03  2.087e+02  -13.906 < 2e-16
## subregionMelanesia          -3.964e+03  3.680e+02  -10.772 < 2e-16
## subregionMiddle Africa     -4.033e+03  2.692e+02  -14.981 < 2e-16
## subregionNorthern Africa   -1.253e+03  2.259e+02  -5.546 3.29e-08
## subregionNorthern America   -4.985e+03  2.841e+02  -17.544 < 2e-16
## subregionNorthern Europe   -8.798e+02  1.890e+02  -4.654 3.46e-06
## subregionSouth-eastern Asia -2.176e+03  2.832e+02  -7.684 2.36e-14
## subregionSouth America      -2.862e+03  2.272e+02  -12.597 < 2e-16
## subregionSouthern Africa    -6.761e+02  2.391e+02  -2.827 0.004737
## subregionSouthern Asia      -1.854e+03  2.455e+02  -7.553 6.34e-14
## subregionSouthern Europe    -2.773e+03  1.770e+02  -15.670 < 2e-16
## subregionWestern Africa     -2.037e+03  2.834e+02  -7.188 9.17e-13
## subregionWestern Asia       -1.176e+03  2.155e+02  -5.459 5.35e-08
## subregionWestern Europe     1.770e+02  1.831e+02   0.967 0.333638
## rain_mm:I((log(pesticides_t))^2) -7.792e-03  1.497e-03  -5.205 2.13e-07
## rain_mm:gdp_tierLow         -1.477e+00  1.610e-01  -9.175 < 2e-16
## rain_mm:gdp_tierMiddle      -9.518e-01  1.576e-01  -6.041 1.81e-09
## I((log(pesticides_t))^2):I(temp_c^2) -1.306e-02  3.914e-03  -3.336 0.000866
## I((log(pesticides_t))^2):gdp_tierLow -4.268e+00  2.319e+00  -1.841 0.065833
## I((log(pesticides_t))^2):gdp_tierMiddle 2.935e+00  2.315e+00   1.268 0.204914
##
## (Intercept) ***  

## year ***  

## rain_mm ***  

## I((log(pesticides_t))^2) ***  

## I(temp_c^2) *  

## gdp_tierLow  

## gdp_tierMiddle *  

## subregionCaribbean ***  

## subregionCentral America ***  

## subregionCentral Asia ***  

## subregionEastern Africa ***  

## subregionEastern Asia ***  

## subregionEastern Europe ***

```

```

## subregionMelanesia      ***
## subregionMiddle Africa  ***
## subregionNorthern Africa ***
## subregionNorthern America ***
## subregionNorthern Europe  ***
## subregionSouth-eastern Asia ***
## subregionSouth America    ***
## subregionSouthern Africa   **
## subregionSouthern Asia     ***
## subregionSouthern Europe    ***
## subregionWestern Africa    ***
## subregionWestern Asia     ***
## subregionWestern Europe    ***
## rain_mm:I((log(pesticides_t))^2)      ***
## rain_mm:gdp_tierLow        ***
## rain_mm:gdp_tierMiddle      ***
## I((log(pesticides_t))^2):I(temp_c^2)    ***
## I((log(pesticides_t))^2):gdp_tierLow    .
## I((log(pesticides_t))^2):gdp_tierMiddle .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1021 on 2059 degrees of freedom
## Multiple R-squared:  0.7514, Adjusted R-squared:  0.7477
## F-statistic: 200.8 on 31 and 2059 DF,  p-value: < 2.2e-16

```

```
plot(lasso_results$cv_fit)
```

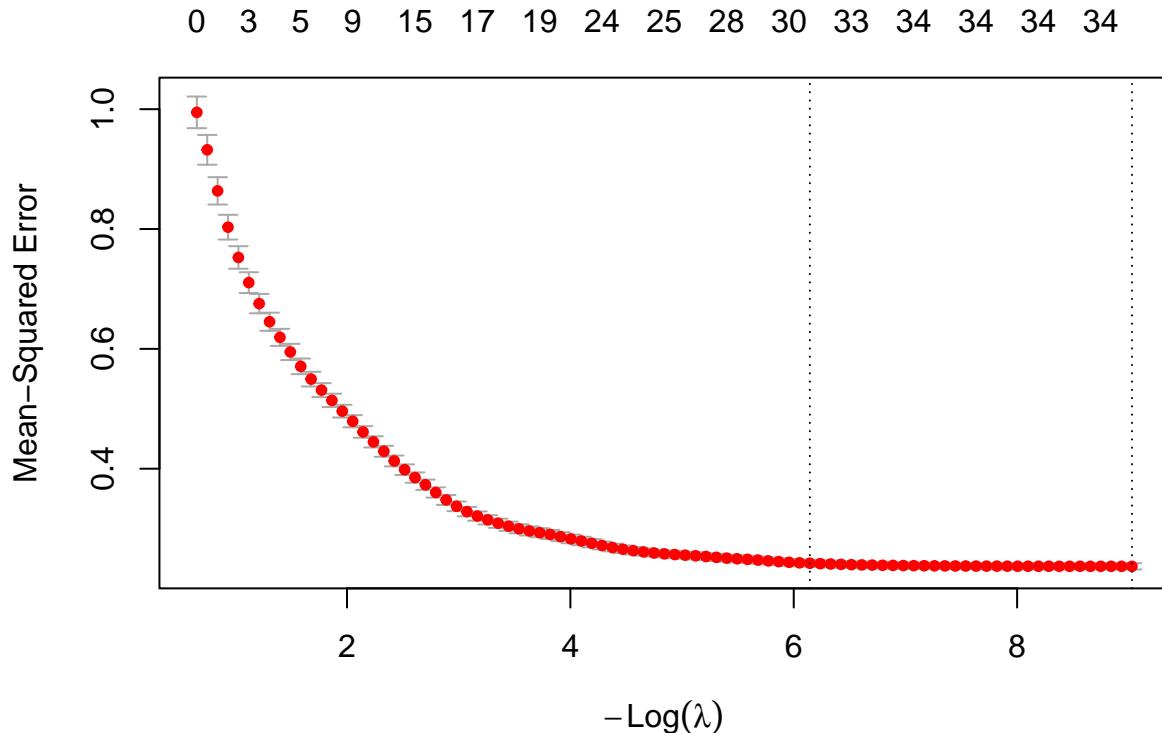


Figure 10: LASSO CV fit

```
coef_32para = coef(lasso_results$cv_fit, s = 0.00375)  
coef_32para
```

```
## 35 x 1 sparse Matrix of class "dgCMatrix"  
##                                     s=0.00375  
## (Intercept)                 7.379276e-16  
## year                      1.506103e-01  
## rain_mm                     5.114258e-01  
## I((log(pesticides_t))^2)   5.370514e-01  
## I(temp_c^2)                  .  
## gdp_tierLow                -5.570006e-02  
## gdp_tierMiddle              -9.756935e-02  
## subregionCaribbean          -4.902915e-02  
## subregionCentral America    -7.327127e-02  
## subregionCentral Asia       -6.813388e-02  
## subregionEastern Africa     -1.674821e-01  
## subregionEastern Asia       -8.668965e-02  
## subregionEastern Europe     -2.396401e-01  
## subregionMelanesia          -9.782349e-02  
## subregionMiddle Africa      -1.765153e-01
```

```

## subregionNorthern Africa -1.956955e-02
## subregionNorthern America -2.160331e-01
## subregionNorthern Europe -5.798260e-03
## subregionSouth-eastern Asia -6.128688e-02
## subregionSouth America -2.315481e-01
## subregionSouthern Africa 3.235528e-02
## subregionSouthern Asia -5.648322e-02
## subregionSouthern Europe -2.624204e-01
## subregionWestern Africa -4.356738e-02
## subregionWestern Asia -1.138788e-02
## subregionWestern Europe 1.412933e-01
## rain_mm:I((log(pesticides_t))^2) -9.821650e-02
## rain_mm:I(temp_c^2) .
## rain_mm:gdp_tierLow -4.830221e-01
## rain_mm:gdp_tierMiddle -3.658051e-01
## I((log(pesticides_t))^2):I(temp_c^2) -1.250883e-01
## I((log(pesticides_t))^2):gdp_tierLow -7.941060e-02
## I((log(pesticides_t))^2):gdp_tierMiddle .
## I(temp_c^2):gdp_tierLow .
## I(temp_c^2):gdp_tierMiddle .

```

```
diagnostic_plots(final_model)
```

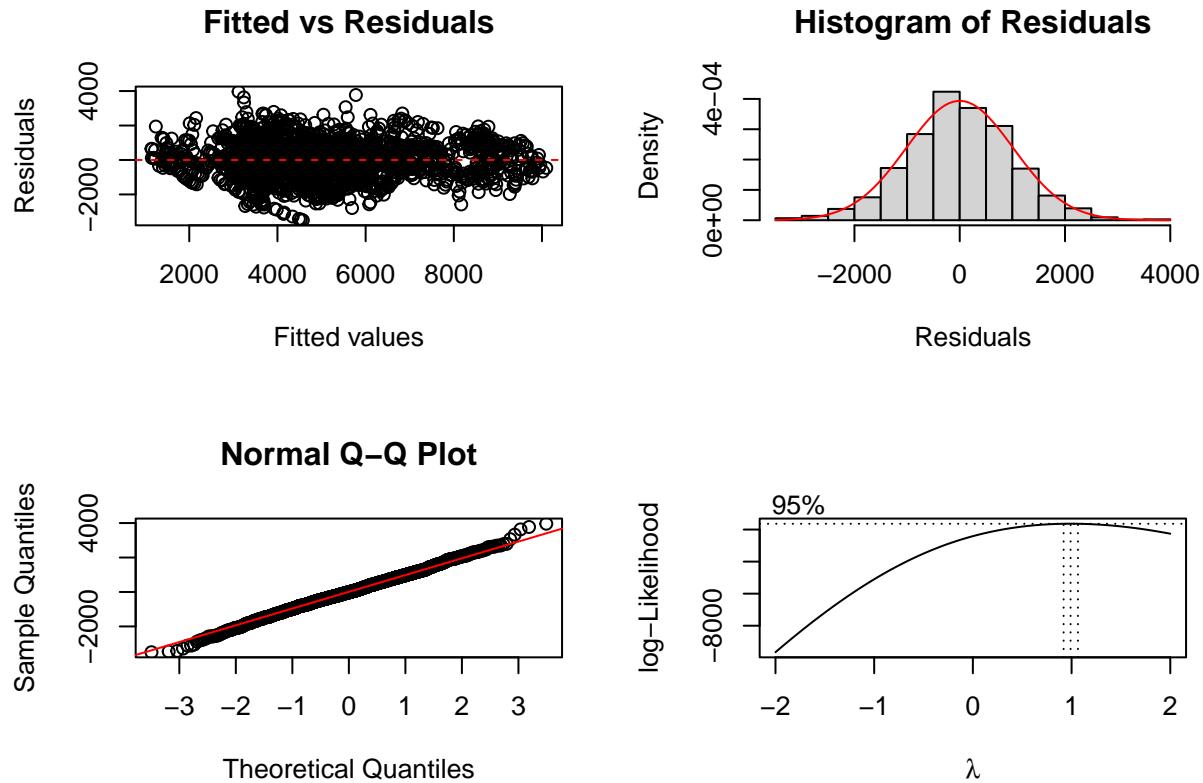


Figure 11: Diagnostic plots for final model

```
diagnostic_tests(final_model)
```

```
## there are higher-order terms (interactions) in this model
## consider setting type = 'predictor'; see ?vif

## === Breusch-Pagan Test (Homoscedasticity) ===
##
## studentized Breusch-Pagan test
##
## data: model
## BP = 498.43, df = 31, p-value < 2.2e-16
##
## Interpretation: p-value < 0.05 suggests heteroscedasticity.
##
##
## === Shapiro-Wilk Test (Normality of Residuals) ===
##
## Shapiro-Wilk normality test
##
```

```

## data: res
## W = 0.9982, p-value = 0.0209
##
## Interpretation: p-value < 0.05 suggests non-normal residuals.
##
##
## === Variance Inflation Factor (VIF) ===
##
##                                     GVIF Df GVIF^(1/(2*Df))
## year                           1.020791  1     1.010342
## rain_mm                         27.817550  1     5.274235
## I((log(pesticides_t))^2)        11.877943  1     3.446439
## I(temp_c^2)                     12.169802  1     3.488524
## gdp_tier                        226.266759  2     3.878423
## subregion                       2399.835205 19    1.227304
## rain_mm:I((log(pesticides_t))^2) 17.738319  1     4.211688
## rain_mm:gdp_tier                219.693163  2     3.849942
## I((log(pesticides_t))^2):I(temp_c^2) 10.996190  1     3.316050
## I((log(pesticides_t))^2):gdp_tier   69.596834  2     2.888334
##
## Interpretation: VIF > 5 suggests moderate collinearity; VIF > 10 suggests high multicollinearity.

```