ADSC2020 Project Report: Crop Recommendation

Ivan Leonychev T00727314

2024-04-05

Contents

Data overview, research objective and hypothesis

Main hypothesis for the research are as follows:

- There is at least one significant interaction term between ratio of microelements in soil.
- There is a high correlation between response variable and rainfall variable.

The dataset "Crop_recommendation" was sourced from Kaggle [URL: https://www.kaggle.com/datasets/atharvaingle/crop-recommendation-dataset]. It comprises data on different types of crops' humidity depending on a set of selected factors. We will use **glimpse** function to get overview over variables in datasets: their data types and values.

```
## Rows: 2,200
## Columns: 8
## $ N           <int> 90, 85, 60, 74, 78, 69, 69, 94, 89, 68, 91, 90, 78, 93, 94~
## $ P           <int> 42, 58, 55, 35, 42, 37, 55, 53, 54, 58, 53, 46, 58, 56, 50~
## $ K           <int> 43, 41, 44, 40, 42, 42, 38, 40, 38, 38, 40, 42, 44, 36, 37~
## $ temperature <dbl> 20.87974, 21.77046, 23.00446, 26.49110, 20.13017, 23.05805~
## $ humidity    <dbl> 82.00274, 80.31964, 82.32076, 80.15836, 81.60487, 83.37012~
## $ ph          <dbl> 6.502985, 7.038096, 7.840207, 6.980401, 7.628473, 7.073454~
## $ rainfall    <dbl> 202.9355, 226.6555, 263.9642, 242.8640, 262.7173, 251.0550~
## $ label       <chr> "rice", "rice", "rice", "rice", "rice", "rice", "rice", "r~
```

The variables in the dataset as follows:

- N - ratio of Nitrogen content in soil
- P - ratio of Phosphorous content in soil
- K - ratio of Potassium content in soil
- temperature - temperature in degree Celsius
- humidity - relative humidity in % (response variable)
- ph - ph value of the soil
- rainfall - rainfall in mm
- label - crop type

In order to efficiently analyse data, we will use a subset of the dataset with label = "rice". We will also remove column label as we no more need it.

Models selection

Creating models

First, we need to select a proper model for the analysis.

Model 1:

```
##
## Call:
## lm(formula = humidity ~ 1 + N + P + K + N:P + N:K + P:K, data = rice)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.60790 -1.22387 -0.05568  1.16990  2.78277
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 90.617661  19.541205   4.637 1.15e-05 ***
## N           -0.183356   0.183455  -0.999    0.320
## P           -0.045889   0.303696  -0.151    0.880
## K            0.004750   0.442875   0.011    0.991
## N:P          0.001596   0.001658   0.963    0.338
## N:K          0.002207   0.003988   0.553    0.581
## P:K         -0.002523   0.006393  -0.395    0.694
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.411 on 93 degrees of freedom
## Multiple R-squared:  0.07023,    Adjusted R-squared:  0.01024
## F-statistic: 1.171 on 6 and 93 DF,  p-value: 0.3287
```

The linear regression model doesn't show a strong fit to the data with an adjusted R-squared of only 0.01, indicating that almost none of the variability in the response variable is explained by the explanatory variables. None of the variables and model overall are statistically significant.

Model 2:

```
##
## Call:
## lm(formula = humidity ~ 1 + temperature + ph + rainfall + temperature:rainfall,
##     data = rice)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5605 -1.1320 -0.0629  1.1979  2.6461
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          62.110929  11.867188   5.234 9.91e-07 ***
## temperature           0.784350   0.487597   1.609   0.1110
## ph                   -0.034258   0.185401  -0.185   0.8538
## rainfall              0.084249   0.048300   1.744   0.0843 .
## temperature:rainfall -0.003237   0.002020  -1.603   0.1124
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.407 on 95 degrees of freedom
## Multiple R-squared:  0.05618,    Adjusted R-squared:  0.01644
## F-statistic: 1.414 on 4 and 95 DF,  p-value: 0.2353
```

The linear regression model doesn't show a strong fit to the data with an adjusted R-squared of only 0.01, indicating that almost none of the variability in the response variable is explained by the explanatory variables. None of the variables and model overall are statistically significant. Though it's statistical significance better than previous model.

Model 3:

```
##
## Call:
## lm(formula = humidity ~ 1 + N + P + K + temperature + ph + rainfall,
##     data = rice)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6104 -1.0785 -0.1105  1.0251  2.6898
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 81.068662   3.280893  24.709   <2e-16 ***
## N           -0.018355   0.012167  -1.509   0.1348
## P           -0.024240   0.017886  -1.355   0.1786
## K            0.047835   0.049213   0.972   0.3336
## temperature  0.022465   0.069493   0.323   0.7472
## ph          -0.068201   0.183073  -0.373   0.7103
## rainfall     0.007717   0.004143   1.863   0.0657 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.395 on 93 degrees of freedom
## Multiple R-squared:  0.09133,    Adjusted R-squared:  0.03271
## F-statistic: 1.558 on 6 and 93 DF,  p-value: 0.1683
```

The linear regression model doesn't show a strong fit to the data with an adjusted R-squared of only 0.03, indicating that almost none of the variability in the response variable is explained by the explanatory variables. None of the variables and model overall are statistically significant. Though it's statistical significance is better than both previous model.

Now let, use a Both (Stepwise) Selection to identify the fourth model:

```
##         Step Df Deviance Resid. Df Resid. Dev      AIC
## 1              NA       NA        99   199.1687 70.89820
## 2          + N -1 5.957772        98   193.2109 69.86123
## 3 + rainfall -1 6.454122        97   186.7568 68.46371
```

```
##
## Call:
## lm(formula = humidity ~ 1 + N + rainfall, data = rice)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.42436 -1.20142 -0.05911  1.08660  2.78462
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 82.248284   1.313031  62.640   <2e-16 ***
## N           -0.021737   0.011718  -1.855   0.0666 .
## rainfall     0.007457   0.004073   1.831   0.0702 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.388 on 97 degrees of freedom
## Multiple R-squared:  0.06232,    Adjusted R-squared:  0.04298
## F-statistic: 3.223 on 2 and 97 DF,  p-value: 0.04412
```

This model still covers only 4% of variance in the data, but, contrary to previous models, it is statistically significant (p-value < 0.05).

Model selection

```
## Analysis of Variance Table
##
## Model 1: humidity ~ 1 + N + P + K + N:P + N:K + P:K
## Model 2: humidity ~ 1 + temperature + ph + rainfall + temperature:rainfall
## Model 3: humidity ~ 1 + N + P + K + temperature + ph + rainfall
## Model 4: humidity ~ 1 + N + rainfall
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     93 185.18
## 2     95 187.98 -2    -2.797 0.7023 0.4980
## 3     93 180.98  2     7.001 1.7580 0.1781
## 4     97 186.76 -4    -5.779 0.7256 0.5767
```

AIC:

```
##         df      AIC
## model_1  8 361.4045
## model_2  6 358.9037
## model_3  8 359.1082
## model_4  4 354.2514
```

BIC:

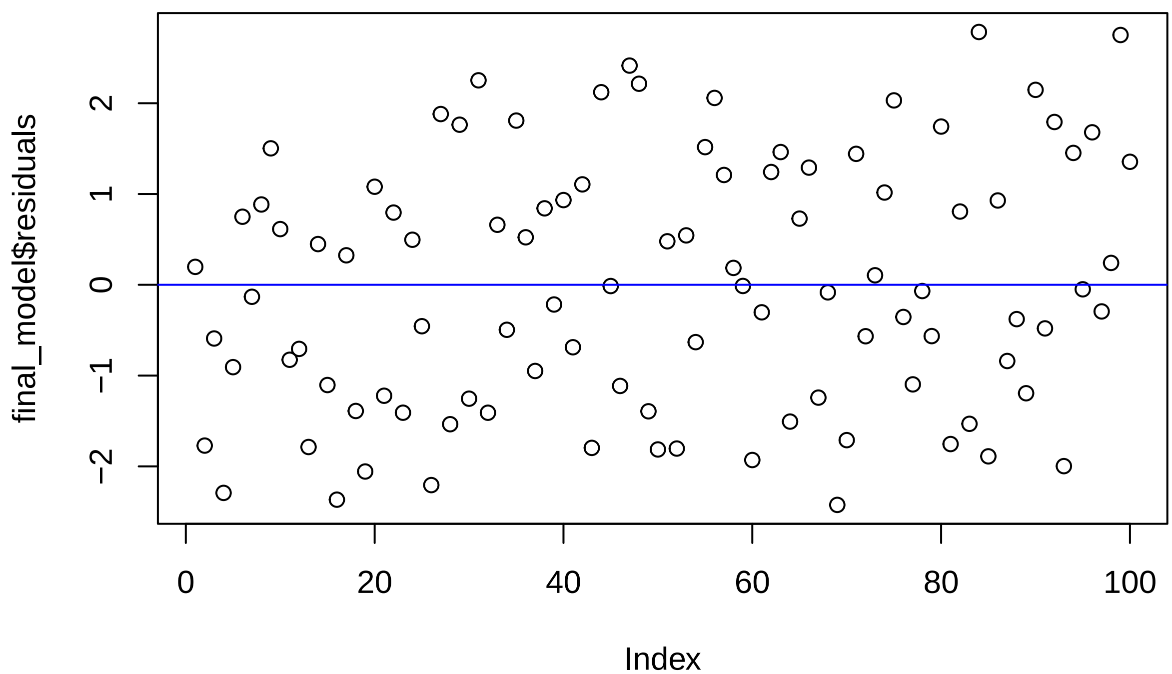
```
##         df      BIC
## model_1  8 382.2459
## model_2  6 374.5347
## model_3  8 379.9495
## model_4  4 364.6721
```

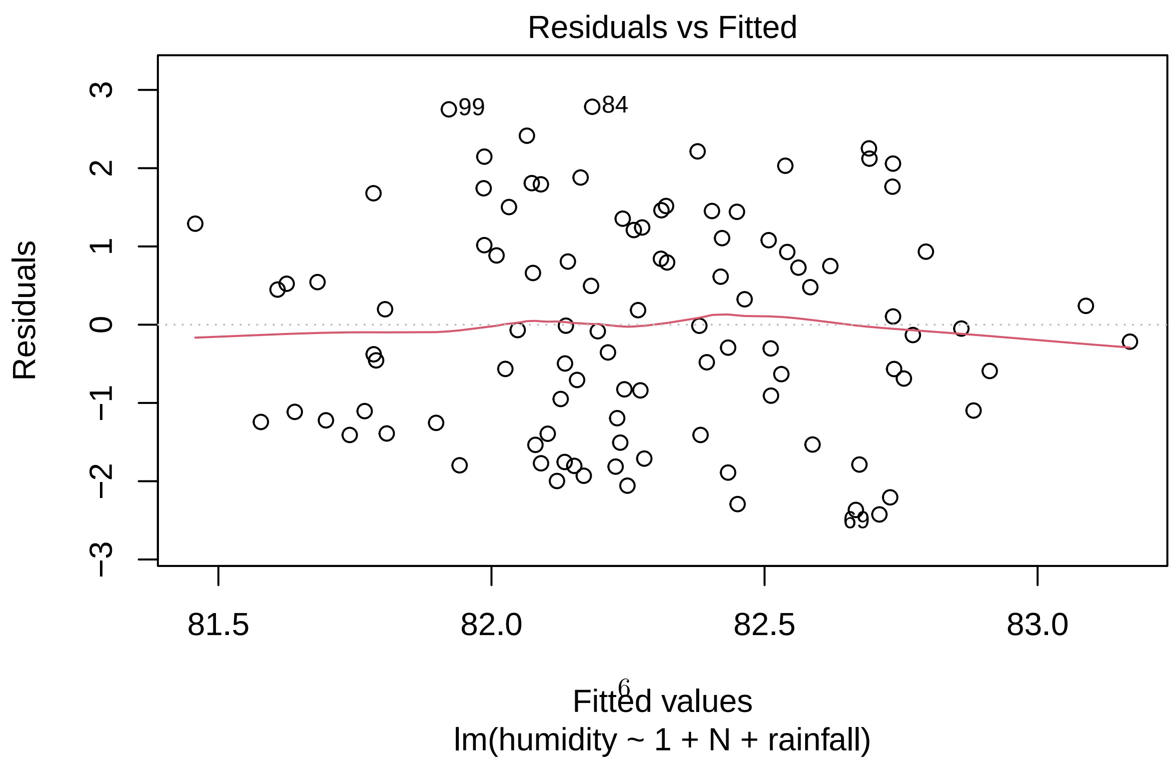
Model 4 is the best model of the three, because it has the lowest AIC and BIC values across three models. It also has highest Adjusted R-squared and and lowest p-value values.

Final model: humidity ~ 1 + N + rainfall

Models diagnostics

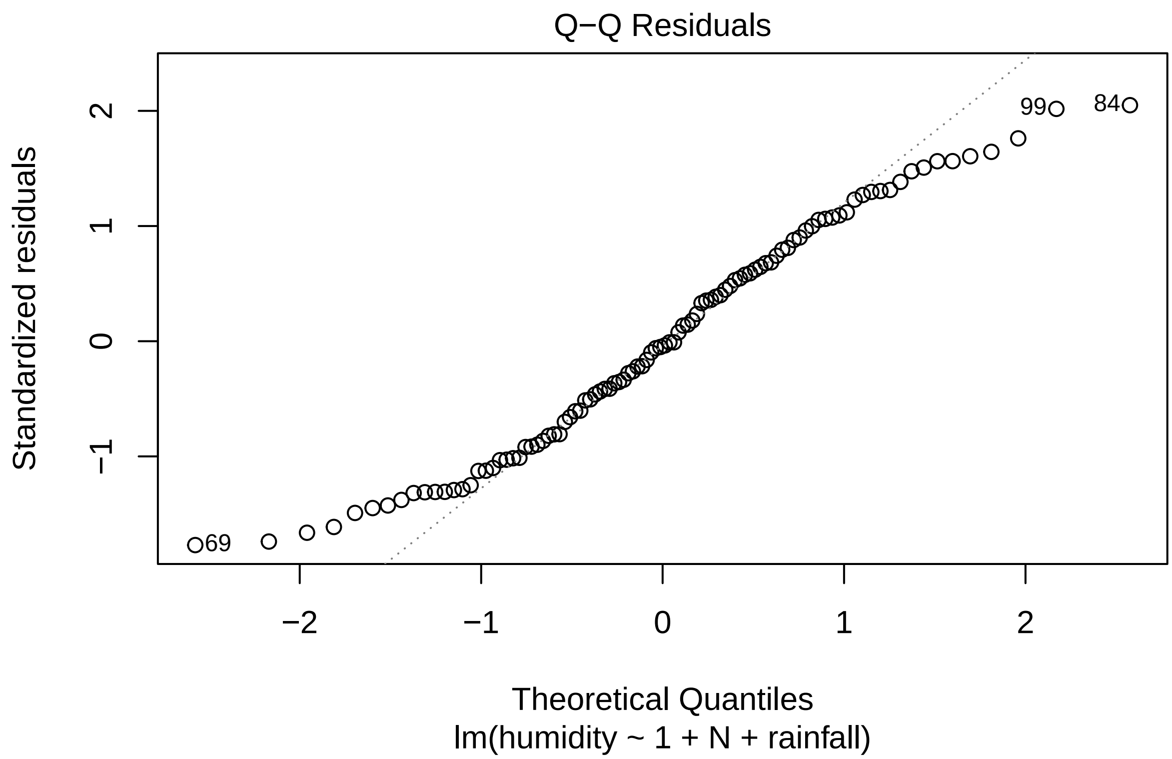
Linearity





Linearity appears to hold.

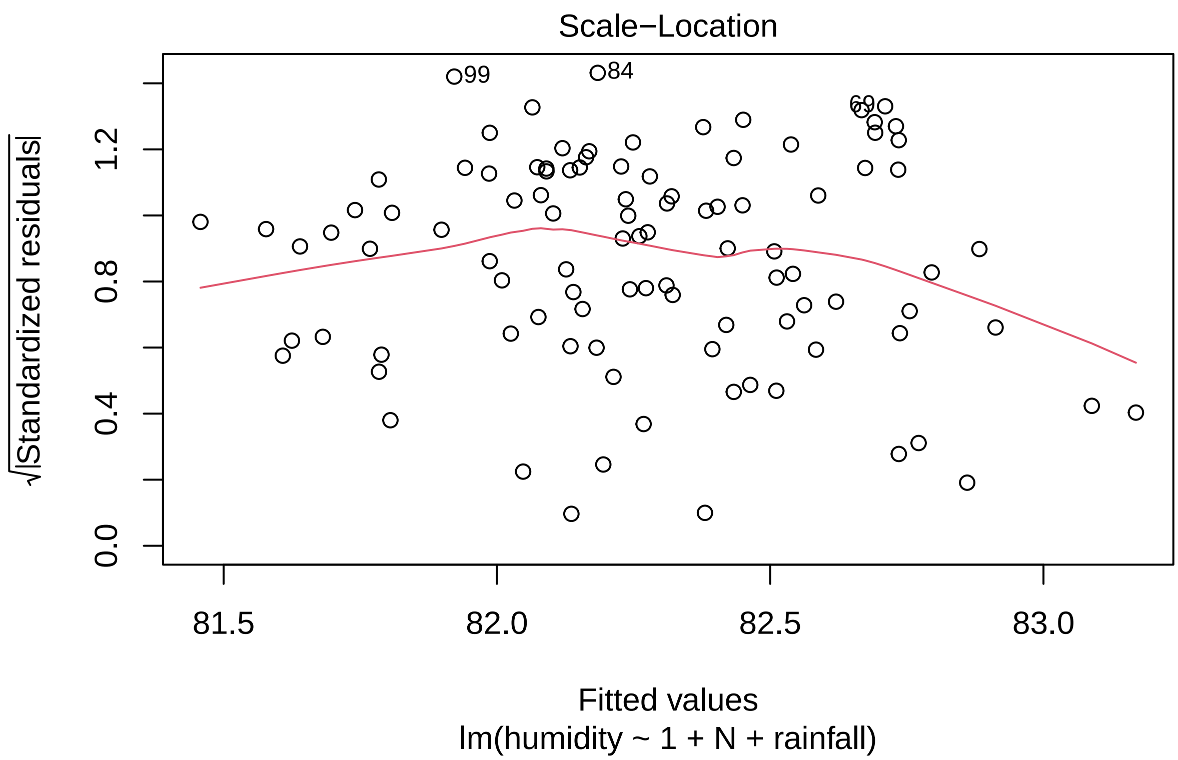
Normality




```
##
##  Shapiro-Wilk normality test
##
## data:  final_model.standard
## W = 0.96789, p-value = 0.01527
```

The **normality** assumption is violated. Reject the null hypothesis.

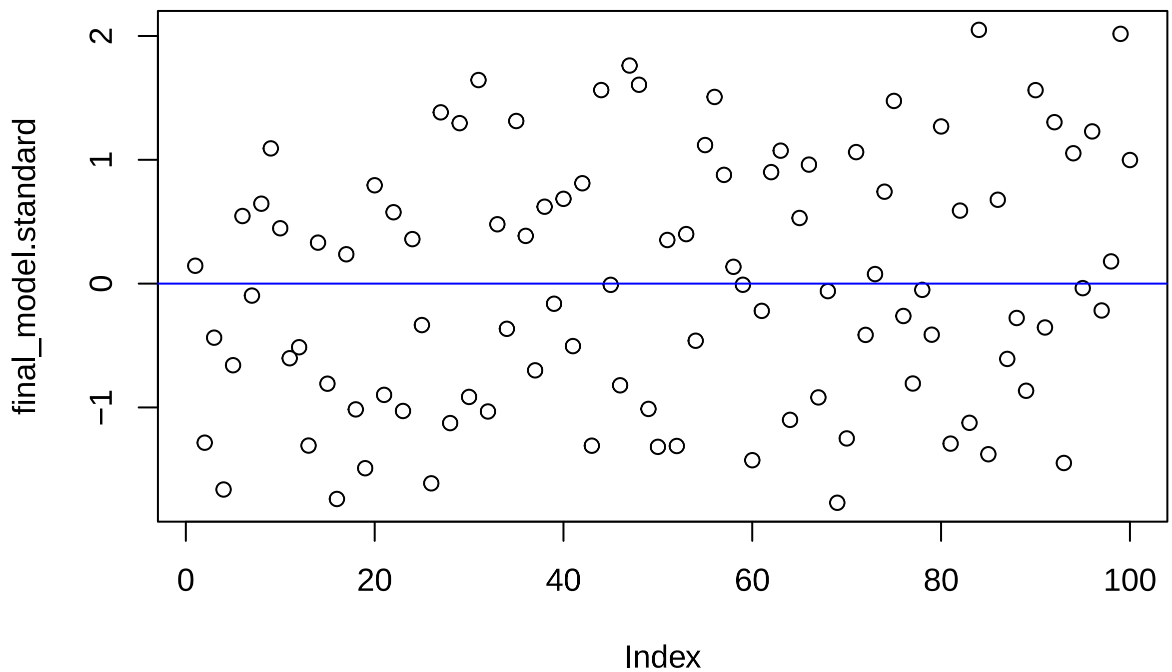
Homoscedasticity




```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.0003460939, Df = 1, p = 0.98516
```

Homoscedasticity appears to hold. Do not reject the null hypothesis.

Independence




```
##  lag Autocorrelation D-W Statistic p-value
##    1      -0.2357353      2.461433   0.016
##  Alternative hypothesis: rho != 0
```

Independence test failed. Reject the null hypothesis.

Model adjustments

Remove any possible **outliers**:

No outliers detected.

Box-Cox:

```
##
## Call:
## lm(formula = humidity ~ 1 + N + rainfall, data = rice)
##
## Residuals:
##         Min          1Q      Median          3Q         Max
## -9.698e-06 -3.912e-06  8.260e-08  4.312e-06  8.847e-06
##
## Coefficients:
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.481e-04  4.711e-06  31.433   <2e-16 ***
## N            7.756e-08  4.204e-08   1.845   0.0681 .
## rainfall    -2.711e-08  1.461e-08  -1.855   0.0666 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.978e-06 on 97 degrees of freedom
## Multiple R-squared:  0.06277,    Adjusted R-squared:  0.04345
## F-statistic: 3.248 on 2 and 97 DF,  p-value: 0.04311
```

Check the Box-Cox model:

```
##
##  Shapiro-Wilk normality test
##
## data:  BC_model.standart
## W = 0.96716, p-value = 0.01346
```

Box-Cox transformation didn't solve any any normality issues.

WLS:

```
##
## Call:
## lm(formula = humidity ~ 1 + N + rainfall, data = rice, weights = wt)
##
## Weighted Residuals:
##        Min         1Q     Median         3Q        Max
## -8.248e-06 -3.389e-06  3.520e-08  3.619e-06  7.854e-06
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.484e-04  4.703e-06  31.549   <2e-16 ***
## N            7.486e-08  4.178e-08   1.792   0.0763 .
## rainfall    -2.746e-08  1.464e-08  -1.876   0.0636 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.25e-06 on 97 degrees of freedom
## Multiple R-squared:  0.06184,    Adjusted R-squared:  0.04249
## F-statistic: 3.197 on 2 and 97 DF,  p-value: 0.04524
```

Check WLS model:

```
##
##  Shapiro-Wilk normality test
##
## data:  wls.model.standard
## W = 0.97015, p-value = 0.02271
```

Normality assumption is still violated but better now.

Predictions

10-fold cross validation:

```
## note: only 1 unique complexity parameters in default grid. Truncating the grid to 1 .
```

Model 1 Results:

```
##   mtry         RMSE   Rsquared          MAE       RMSESD RsquaredSD
## 1    2 5.437686e-06 0.05698832 4.605677e-06 8.144856e-07 0.06234699
## 2    4 5.536964e-06 0.06007050 4.726823e-06 8.731725e-07 0.06507508
## 3    6 5.507259e-06 0.06388138 4.697067e-06 8.818888e-07 0.07237397
##          MAESD
## 1 7.594284e-07
## 2 8.136570e-07
## 3 7.955420e-07
```

Model 2 Results:

```
##   mtry         RMSE  Rsquared          MAE       RMSESD RsquaredSD        MAESD
## 1    2 5.122516e-06 0.1061828 4.228335e-06 7.140244e-07 0.09603080 6.675204e-07
## 2    3 5.126188e-06 0.1022081 4.214234e-06 6.802424e-07 0.09569777 6.572659e-07
## 3    4 5.145016e-06 0.1018062 4.217493e-06 6.840845e-07 0.10647145 6.635788e-07
```

Model 3 Results:

```
##   mtry         RMSE  Rsquared          MAE       RMSESD RsquaredSD        MAESD
## 1    2 5.198158e-06 0.1395044 4.470256e-06 9.080144e-07  0.1416550 8.947543e-07
## 2    4 5.208132e-06 0.1347349 4.495322e-06 9.413069e-07  0.1391270 9.049484e-07
## 3    6 5.251142e-06 0.1443940 4.475630e-06 9.881184e-07  0.1669471 9.767888e-07
```

Model 4 Results:

```
##   mtry         RMSE  Rsquared          MAE      RMSESD RsquaredSD        MAESD
## 1    2 5.663512e-06 0.1132994 4.824915e-06 9.43983e-07 0.08887197 9.539881e-07
```

Model 4 has the lowest RMSE and MAE values while model 3 has the highest R-squared value. Based on these results, model 4 or 3 is probably the best for prediction. The other two models were outperformed in all categories.

Generalized linear models

We will use the same formulas for all our glm models and use gamma distribution with "log" link:

gamma_1 <- glm(humidity ~ N + P + K + N:P + N:K + P:K, family = Gamma(link="log"), data = rice)

gamma_2 <- glm(humidity ~ temperature + ph + rainfall + temperature:rainfall, family = Gamma(link="log"), data = rice)

gamma_3 <- glm(humidity ~ N + P + K + temperature + ph + rainfall, family = Gamma(link="log"), data = rice)

gamma_4 <- glm(humidity ~ N + rainfall, family = Gamma(link="log"), data = rice)

Selecting the best model:

Gamma_1 and Gamma_2:

```
## Analysis of Deviance Table
##
## Model 1: humidity ~ N + P + K + N:P + N:K + P:K
## Model 2: humidity ~ temperature + ph + rainfall + temperature:rainfall
##   Resid. Df Resid. Dev Df   Deviance Pr(>Chi)
## 1        93    0.10919
## 2        95    0.11085 -2 -0.0016531   0.4939
```

Gamma_2 is better.

Gamma_2 and Gamma_3:

```
## Analysis of Deviance Table
##
## Model 1: humidity ~ temperature + ph + rainfall + temperature:rainfall
## Model 2: humidity ~ N + P + K + temperature + ph + rainfall
##   Resid. Df Resid. Dev Df  Deviance Pr(>Chi)
## 1        95    0.11085
## 2        93    0.10668  2 0.0041654   0.1621
```

Gamma_3 is better.

Gamma_3 and Gamma_4:

```
## Analysis of Deviance Table
##
## Model 1: humidity ~ N + P + K + temperature + ph + rainfall
## Model 2: humidity ~ N + rainfall
##   Resid. Df Resid. Dev Df   Deviance Pr(>Chi)
## 1        93    0.10668
## 2        97    0.11013 -4 -0.0034443   0.5563
```

Model "logit 4" is the best model based on ANOVA comparison.

AIC and BIC:

```
##         df       AIC
## gamma_1  8 -2146.132
## gamma_2  6 -2148.629
## gamma_3  8 -2148.460
## gamma_4  4 -2153.282
```

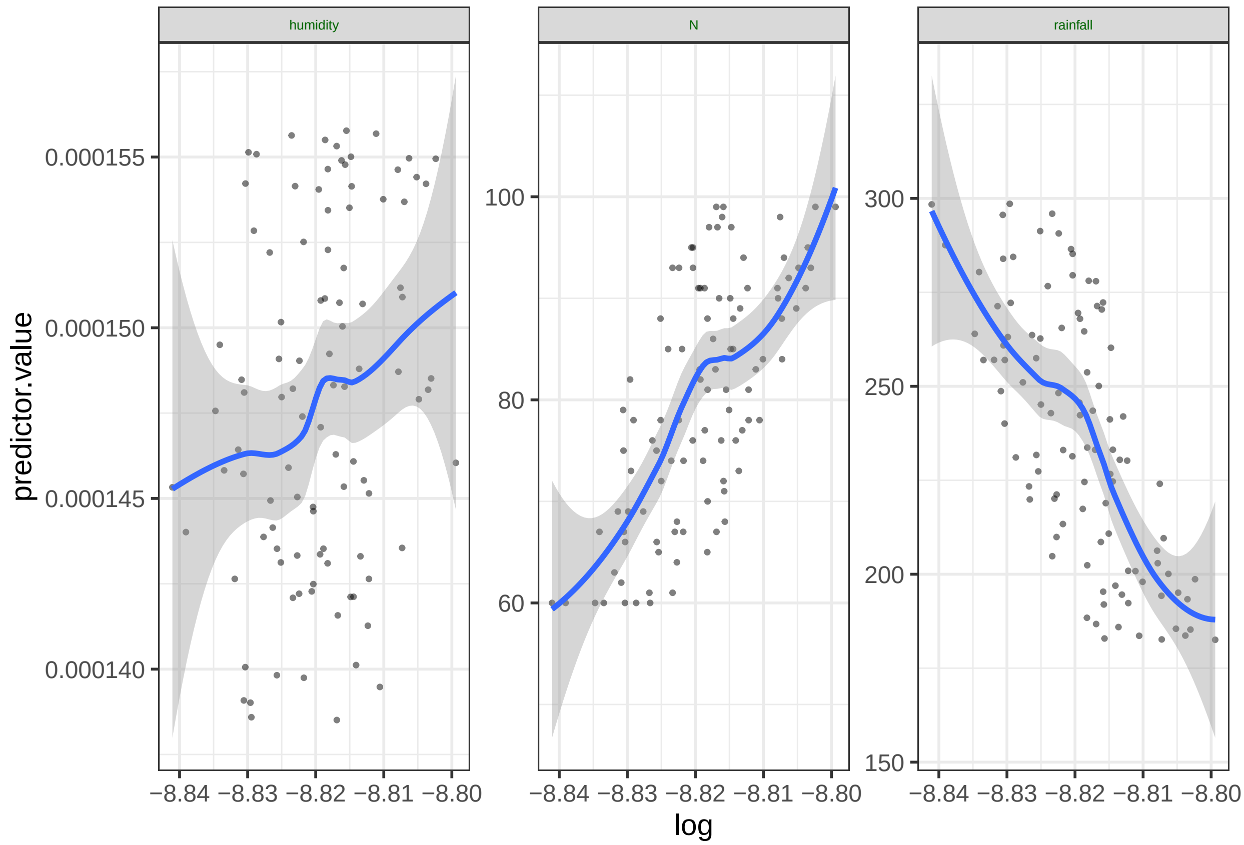
```
##         df       BIC
## gamma_1  8 -2125.290
## gamma_2  6 -2132.998
## gamma_3  8 -2127.618
## gamma_4  4 -2142.861
```

Gamma_2 is the best model based on both AIC and BIC analysis.

Diagnostics:

Linearity:

```
## `geom_smooth()` using formula = 'y ~ x'
```



Seems that all variables from the model are not **linear**. Humidity is the only variable that much more closer to linearity than others.

Multicollinearity:

```
##        N rainfall
## 1.002898 1.002898
```

There are **no problems** with multicollinearity, variables shouldn't be removed.

Conlusion

In conclusion, we can state the following:

- Linear regression models poorly fit the dataset, as no model covers much of the variance in the data, but the final model is statistically significant.
- There is no significant interaction term between any microelements variables: reject the first hypothesis.
- There is no high correlation coefficient between rainfall and humidity variables: reject the second hypothesis.
- Both linear model 3 and 4 should fit for predictions, though model 4 could be a bit better.

Appendix

```
knitr::opts_chunk$set(echo = TRUE)
library(dplyr)
library(tidyr)
library(knitr)
library(caret)
library(car)
library(sjPlot)
library(kableExtra)
library(broom)
library(ggplot2)
library(MASS)

all_crops <- read.csv(file = "Crop_recommendation.csv", header = TRUE)

glimpse(all_crops)

rice <- all_crops %>%
  subset(label == 'rice') %>%
  subset(select = -label)

rice$temperature <- as.numeric(rice$temperature)
rice$humidity <- as.numeric(rice$humidity)
rice$ph <- as.numeric(rice$ph)
rice$rainfall <- as.numeric(rice$rainfall)

model_1 <- lm(humidity ~ 1 + N + P + K + N:P + N:K + P:K, data = rice)
summary(model_1)

model_2 <- lm(humidity ~ 1 + temperature + ph + rainfall + temperature:rainfall, data = rice)
summary(model_2)

model_3 <- lm(humidity ~ 1 + N + P + K + temperature + ph + rainfall, data = rice)
summary(model_3)

intercept_model <- lm(humidity ~ 1, data = rice)
full_model <- lm(humidity ~ .^2, data = rice)

both <- step(intercept_model, direction="both", scope=formula(full_model), trace=0)

both$anova

model_4 <- lm(humidity ~ 1 + N + rainfall, data = rice)
summary(model_4)

anova(model_1, model_2, model_3, model_4)

AIC(model_1, model_2, model_3, model_4)

BIC(model_1, model_2, model_3, model_4)

final_model <- lm(humidity ~ 1 + N + rainfall, data = rice)
```

```
plot(final_model$residuals)
abline(h = 0, col = "blue")

plot(final_model, 1)

plot(final_model, 2)

final_model.standard <- rstandard(final_model) #standardized residuals

shapiro.test(final_model.standard)
plot(final_model, 3)

ncvTest(final_model)

final_model.standard <- rstandard(final_model)
plot(final_model.standard)
abline(h=0,col="blue")

durbinWatsonTest(final_model)
rice$cooks <- cooks.distance(final_model)

rice_new <- rice %>%
  filter(cooks < 0.5)

bc <- boxcox(final_model)
lambda <- bc$x[which.max(bc$y)]

rice$humidity <- rice$humidity^lambda

BC_model <- lm(humidity ~ 1 + N + rainfall, data = rice)

summary(BC_model)
BC_model.standart <- rstandard(BC_model)

shapiro.test(BC_model.standart)
wt <- 1/lm(abs(final_model$residuals) ~ final_model$fitted.values)$fitted.values^2

wls.model <- lm(humidity ~ 1 + N + rainfall, data = rice, weights = wt)
summary(wls.model)

wls.model.standard <- rstandard(wls.model) #standardized residuals
# # null hypothesis of normality
shapiro.test(wls.model.standard)
set.seed(2020)

train.control <- trainControl(method = "cv", number = 10)

model1 <- train(humidity ~ 1 + N + P + K + N:P + N:K + P:K, data = rice,
                trControl = train.control)

model2 <- train(humidity ~ 1 + temperature + ph + rainfall + temperature:rainfall, data = rice,
                trControl = train.control)
```

```r
model3 <- train(humidity ~ 1 + N + P + K + temperature + ph + rainfall, data = rice,
                trControl = train.control)

model4 <- train(humidity ~ 1 + N + rainfall, data = rice,
                trControl = train.control)

model1$results

model2$results

model3$results

model4$results

gamma_1 <- glm(humidity ~ N + P + K + N:P + N:K + P:K,
              family = Gamma(link="log"), data = rice)

gamma_2 <- glm(humidity ~ temperature + ph + rainfall + temperature:rainfall,
              family = Gamma(link="log"), data = rice)

gamma_3 <- glm(humidity ~ N + P + K + temperature + ph + rainfall,
              family = Gamma(link="log"), data = rice)

gamma_4 <- glm(humidity ~ N + rainfall,
              family = Gamma(link="log"), data = rice)

## LRT ##

anova(gamma_1, gamma_2, test='LR') # 2 is better

anova(gamma_2, gamma_3, test='LR') # 3 is better

anova(gamma_3, gamma_4, test='LR') # 4 is better

AIC(gamma_1, gamma_2, gamma_3, gamma_4)

BIC(gamma_1, gamma_2, gamma_3, gamma_4)

# Model values #
probabilities <- predict(gamma_4, type = "response")
probabilities <- probabilities[1:100]

# Numeric variables #
mydata <- rice[1:100,] %>%
  na.omit() %>%
  dplyr::select_if(is.numeric) %>%
  dplyr::select(humidity, N, rainfall)
predictors <- colnames(mydata)

# transformed relationship #
mydata <- mydata %>%
  mutate(log = log(probabilities)) %>%
  gather(key = "predictors", value = "predictor.value", -log)
```

```r
# generate the plots #
LinCheck <- ggplot(mydata, mapping = aes(log, predictor.value))+
  geom_point(size = 0.5, alpha = 0.5) +
  geom_smooth(method = "loess") +
  theme_bw() +
  facet_wrap(~predictors, scales = "free_y") +
  theme(strip.text = element_text(
    size = 5, color = "dark green"))

LinCheck

vif(gamma_4)
```