Maize Analysis English

Miguel Barros (pg42877), Tiago Machado (pg42884), Tiago Silva (pg42885)

01/03/2021

Introduction

CME - Ear length in cm
NTF - Total number of leaves
ARF - Leaf area cm^2
ALP - Plant height in cm
PMG - Thousand grain weight in g
ALC - Harvest site altitude in m
CDE - Endosperm color
NTN - Total number of nodes
FAS - Fasciation

The dataset used consists of mean values from regional corn populations, obtained over 3 decades through morphological characterization activities of the Portuguese Plant Germplasm Bank (BPGV). The variables in this dataset correspond to descriptors used for plant characterization before harvest, which corresponds to grain maturation, with the exception of the ALC descriptor, which corresponds to a passport descriptor obtained by technicians who collect this germplasm. We can classify the descriptors obtained from the characterization effort into 3 components: vegetative (**ALP**, **NTN**), foliar (**NTF**, **ARF**), productive (**CME**, **PMG**), and intrinsic (**FAS**, **CDE**).

Thus, for performing linear regression, we selected **ALP** as the dependent variable, since morphological studies have proven an interdependence between the various types of descriptors presented here. Therefore, we propose to demonstrate the association between these variables through the creation of a linear regression model for plant height using this dataset.

Import and Factorization

```
#Import the dataset
library(readxl)
data_milho <- read_excel("data_milho.xlsx")

#Factorize the variables
data_milho$CDE <-factor(data_milho$CDE,c("1", "2", "3", "4", "5"),labels=c("White", "Cream", "Light-yel

data_milho$FAS <-factor(data_milho$FAS,c("0", "1"),labels=c("Absent", "Present"))
```

Preliminary Analysis

```
#Preliminary dataset information
str(data_milho)
```

```
## tibble [566 x 9] (S3: tbl_df/tbl/data.frame)
##  $ CME: num [1:566] 8 10 10 11 11 10 13 7 9 11 ...
##  $ NTF: num [1:566] 5 8 8 7 8 7 6 7 9 6 ...
##  $ ARF: num [1:566] 30 60 60 57.8 66 ...
##  $ ALP: num [1:566] 92 98 100 105 108 108 108 110 116 117 ...
##  $ PMG: num [1:566] 240 270 230 300 335 250 370 250 210 325 ...
##  $ ALC: num [1:566] 95 100 110 760 400 400 380 950 175 110 ...
##  $ CDE: Factor w/ 5 levels "White","Cream",..: 4 1 1 3 3 4 5 4 3 1 ...
##  $ NTN: num [1:566] 6 4 4 6 6 6 5 7 8 7 ...
##  $ FAS: Factor w/ 2 levels "Absent","Present": 1 1 1 1 1 1 1 1 1 1 ...
```

```
#Does our dataset have missing data?
any(is.na(data_milho))
```

```
## [1] FALSE
```

The dataset contains data from 566 populations, distributed across 9 variables explained above, with 6 continuous variables and 2 categorical variables. Within the categorical variables, the **CDE** variable has 5 levels (endosperm colors) and the **FAS** variable is a binary variable (Presence/Absence of fasciation).
The dataset does not contain missing data (NA).

Exploratory Data Analysis

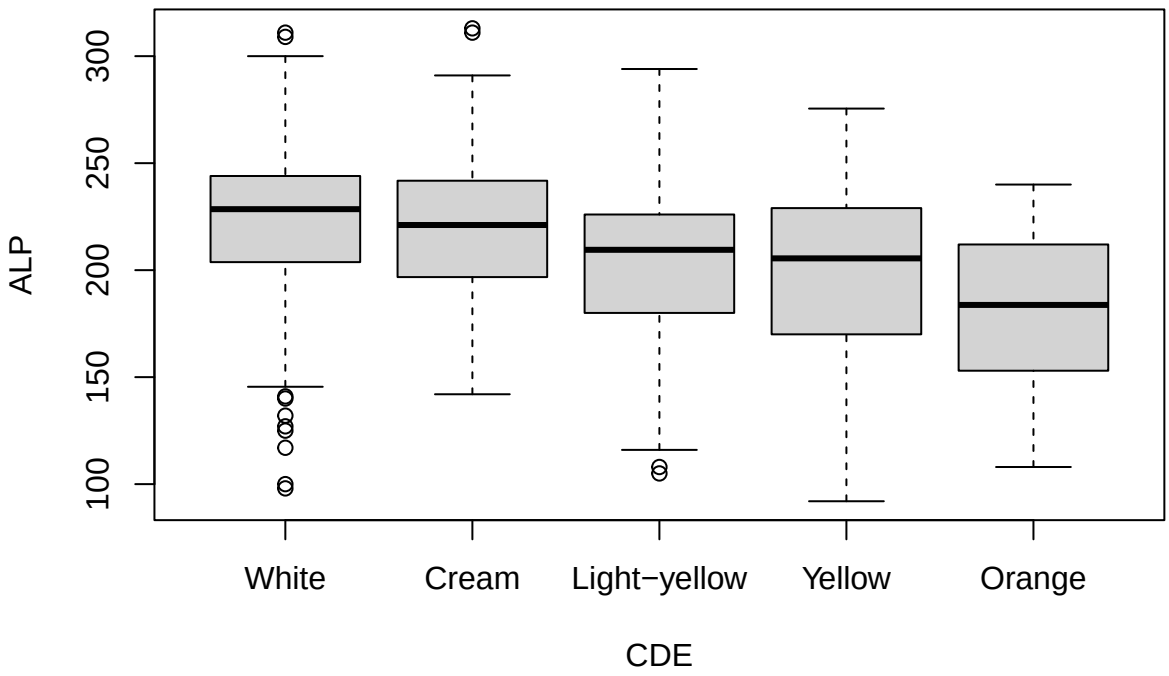
```
library(summarytools)
dfSummary(data_milho, na.col = F, valid.col = F)
```

```
## Data Frame Summary
## data_milho
## Dimensions: 566 x 9
## Duplicates: 1
##
## ------------------------------------------------------------------------------------------
## No   Variable    Stats / Values              Freqs (% of Valid)    Graph
## ---- ----------- --------------------------- --------------------- ----------------------
## 1    CME         Mean (sd) : 15 (2.1)        103 distinct values             :
##      [numeric]   min < med < max:                                             :
##                  7 < 15 < 22                                               . : .
##                  IQR (CV) : 2.1 (0.1)                                    . : : :
##                                                                        . . : : : : :
##
## 2    NTF         Mean (sd) : 10.2 (1.5)      56 distinct values              :
##      [numeric]   min < med < max:                                            : .
##                  5 < 10 < 14                                               . : : .
##                  IQR (CV) : 2 (0.1)                                      . : : : :
```

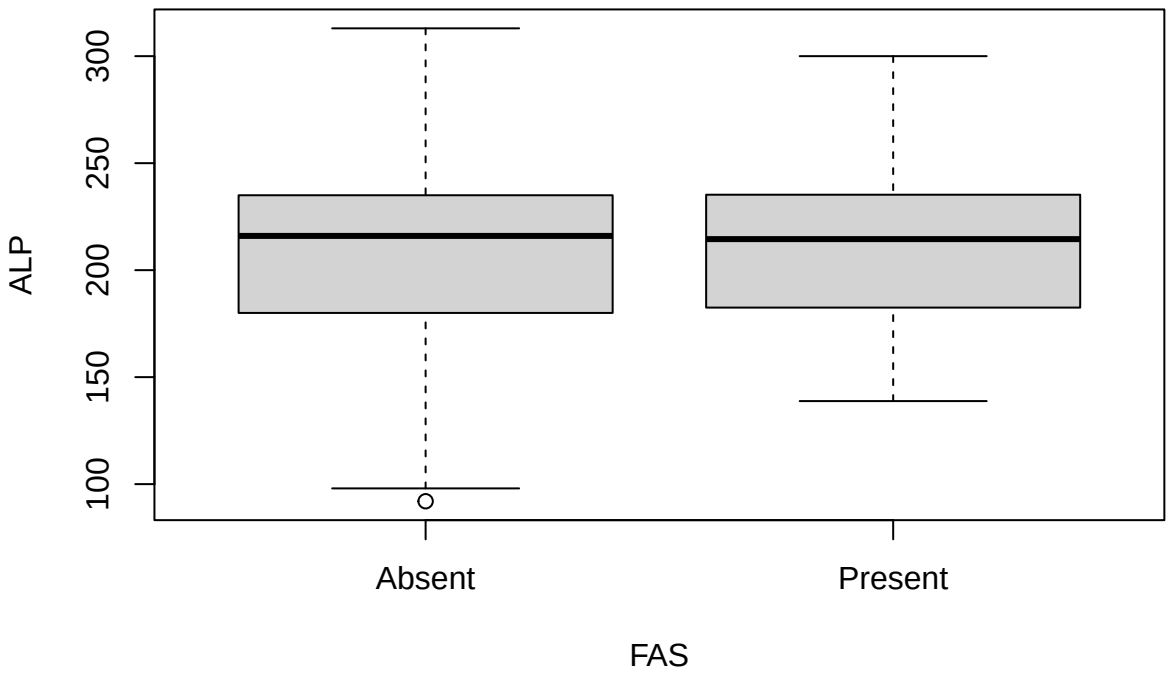
```
##                                                                    . : : : : : : :
##
## 3    ARF          Mean (sd) : 115.2 (28)      241 distinct values           :
##      [numeric]    min < med < max:                                          : :
##                   30 < 114.4 < 189                                        : : : :
##                   IQR (CV) : 38.9 (0.2)                                 . : : : : :
##                                                                       . : : : : : : .
##
## 4    ALP          Mean (sd) : 208 (39.7)      264 distinct values             : .
##      [numeric]    min < med < max:                                            : :
##                   92 < 215.5 < 313                                      .     : :
##                   IQR (CV) : 55 (0.2)                                 . : : : : : :
##                                                                 . . : : : : : : : : . .
##
## 5    PMG          Mean (sd) : 302.5 (60.8)    118 distinct values           :
##      [numeric]    min < med < max:                                        : :
##                   112 < 310 < 525                                       . : : .
##                   IQR (CV) : 77 (0.2)                                   : : : :
##                                                                       . : : : : .
##
## 6    ALC          Mean (sd) : 404.2 (268.5)   155 distinct values   :
##      [numeric]    min < med < max:                                  :     : .
##                   5 < 400 < 1150                                    : . : : : .
##                   IQR (CV) : 407.5 (0.7)                            : : : : : : :
##                                                                     : : : : : : : : : . .
##
## 7    CDE          1. White                    148 (26.1%)           IIIII
##      [factor]     2. Cream                    108 (19.1%)           III
##                   3. Light-yellow              94 (16.6%)           III
##                   4. Yellow                   194 (34.3%)           IIIIII
##                   5. Orange                    22 ( 3.9%)
##
## 8    NTN          Mean (sd) : 9.6 (1.5)       47 distinct values             :
##      [numeric]    min < med < max:                                         : : :
##                   4 < 10 < 13                                            . : : :
##                   IQR (CV) : 2 (0.2)                                     : : : : :
##                                                                      . . : : : : : : .
##
## 9    FAS          1. Absent                   491 (86.7%)           IIIIIIIIIIIIIIII
##      [factor]     2. Present                   75 (13.3%)           II
## ----------------------------------------------------------------------------------------
```

The descriptive statistics for the descriptors used are shown above. Appealing to the Central Limit Theorem, we will assume normality of the distributions.

```
boxplot(ALP~CDE, data = data_milho)
```




```
boxplot(ALP~FAS, data = data_milho)
```




```
#ANOVA analysis between the dependent variable and the categorical variable CDE
#Verify homogeneity of variances
fligner.test(ALP~CDE, data = data_milho)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  ALP by CDE
## Fligner-Killeen:med chi-squared = 4.9117, df = 4, p-value = 0.2965
```

```
#Assuming normality
model = aov(ALP~CDE, data = data_milho)
summary(model)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## CDE           4  66381   16595    11.3 7.93e-09 ***
## Residuals   561 823721    1468
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
TukeyHSD(model)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
```

```
##
## Fit: aov(formula = ALP ~ CDE, data = data_milho)
##
## $CDE
##                            diff       lwr        upr     p adj
## Cream-White           -3.694645 -16.96579   9.576498 0.9413324
## Light-yellow-White   -19.989779 -33.82047  -6.159085 0.0008179
## Yellow-White         -21.281367 -32.72628  -9.836457 0.0000049
## Orange-White         -38.352641 -62.31410 -14.391179 0.0001375
## Light-yellow-Cream   -16.295134 -31.08727  -1.502997 0.0224832
## Yellow-Cream         -17.586722 -30.17661  -4.996836 0.0013770
## Orange-Cream         -34.657997 -59.18698 -10.129015 0.0011609
## Yellow-Light-yellow   -1.291588 -14.46998  11.886806 0.9988623
## Orange-Light-yellow  -18.362863 -43.19904   6.473316 0.2560756
## Orange-Yellow        -17.071275 -40.66224   6.519691 0.2769280
```

```
#ANOVA analysis between the dependent variable and the categorical variable FAS
#Verify homogeneity of variances
fligner.test(ALP~FAS, data = data_milho)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  ALP by FAS
## Fligner-Killeen:med chi-squared = 1.0155, df = 1, p-value = 0.3136
```

```
#Assuming normality
model = aov(ALP~FAS, data = data_milho)
summary(model)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## FAS           1    115   115.4   0.073  0.787
## Residuals   564 889986  1578.0
```

Since the dependent variable we intend to study corresponds to **ALP**, we chose to analyze the distribution of this variable against the categorical variables.
We observe the occurrence of outliers in the **ALP/CDE** box plots for the 'White', 'Cream', and 'Light-yellow' factors. Among all, the 'Orange' factor presents itself as the one with the lowest mean relative to **ALP**, also being the factor with the smallest number of individuals (3.9%).
The ANOVA analysis results indicate that the differences between the 'Light-Yellow', 'Yellow', and 'Orange' factors are not significant, as well as between the 'White' and 'Cream' factors, with all others being significant.
For the **FAS** variable, we verify that the ANOVA analysis does not reveal significant differences between its levels (Presence/Absence) and the dependent variable, which is observable in the **ALP/FAS** box plot.

Correlations

```
#create dataset with only numeric variables
data_milho_num <- data_milho
data_milho_num$FAS <- NULL
data_milho_num$CDE <- NULL
```

```
#Correlations
library(GGally)
```

```
## Loading required package: ggplot2
```

```
library(ggplot2)
ggcorr(data_milho_num, geom="tile", label = T, label_alpha = F)
```

The correlations between continuous variables were computed using Pearson's correlation coefficient. Defining correlations above 0.5 as relevant, we observe the occurrence of strong and positive correlations between the following descriptors:

CME - **NTF**, **ARF**, **ALP** and **NTN**
NTF - **CME**, **ARF**, **ALP**, **NTN**
ARF - **CME**, **NTF**, **ALP**, **NTN**
ALP - **CME**, **NTF**, **ARF**, **NTN**

We also observe lighter and positive correlations between the **PMG** descriptor and the vegetative and foliar descriptors that are strongly correlated among themselves.
A slight negative correlation is also verified between the **ALC** descriptor and all others.

Linear Regression (Full-model)

```
#Complete fit
full.model <- lm(ALP ~., data = data_milho)
summary(full.model)
```

```
##
## Call:
## lm(formula = ALP ~ ., data = data_milho)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -74.130 -13.806  -0.116  14.842  64.623
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -2.034e+02  4.217e+01  -4.825 1.81e-06 ***
## CME              1.570e+01  2.847e+00   5.515 5.37e-08 ***
## NTF              2.278e+01  4.428e+00   5.146 3.70e-07 ***
## ARF             -1.417e+00  3.854e-01  -3.676 0.000259 ***
## PMG              7.776e-02  1.960e-02   3.968 8.21e-05 ***
## ALC              2.296e-03  3.840e-03   0.598 0.550045
## CDECream        -1.939e+00  2.942e+00  -0.659 0.510035
## CDELight-yellow  1.206e+00  3.428e+00   0.352 0.725112
## CDEYellow       -7.348e+00  2.627e+00  -2.797 0.005340 **
## CDEOrange       -1.470e+01  5.536e+00  -2.656 0.008138 **
## NTN              8.980e+00  8.830e-01  10.169  < 2e-16 ***
## FASPresent       4.470e+00  2.925e+00   1.528 0.127003
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.06 on 554 degrees of freedom
## Multiple R-squared:  0.6691, Adjusted R-squared:  0.6625
## F-statistic: 101.8 on 11 and 554 DF,  p-value: < 2.2e-16
```

```
#Test for multicollinearity
car::vif(full.model)
```

```
##           GVIF Df GVIF^(1/(2*Df))
## CME  39.436548  1        6.279853
## NTF  44.443851  1        6.666622
## ARF 123.331154  1       11.105456
## PMG   1.507722  1        1.227893
## ALC   1.129155  1        1.062617
## CDE   1.478180  4        1.050064
## NTN   1.984576  1        1.408750
## FAS   1.046837  1        1.023151
```

```
#Remove ARF
data_milho$ARF <- NULL

#New complete fit
```

```r
full.model <- lm(ALP ~., data = data_milho)
summary(full.model)
```

```
##
## Call:
## lm(formula = ALP ~ ., data = data_milho)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -84.138 -14.475   0.279  14.730  65.701
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -52.851483  10.111563  -5.227 2.45e-07 ***
## CME               5.455422   0.590183   9.244  < 2e-16 ***
## NTF               6.923725   1.006716   6.878 1.65e-11 ***
## PMG               0.072522   0.019767   3.669 0.000267 ***
## ALC               0.002880   0.003879   0.742 0.458230
## CDECream         -1.534346   2.972627  -0.516 0.605949
## CDELight-yellow   1.457687   3.466087   0.421 0.674241
## CDEYellow        -7.821781   2.653498  -2.948 0.003336 **
## CDEOrange       -12.639860   5.569023  -2.270 0.023610 *
## NTN               9.163944   0.891499  10.279  < 2e-16 ***
## FASPresent        5.150903   2.951549   1.745 0.081512 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.32 on 555 degrees of freedom
## Multiple R-squared:  0.661,  Adjusted R-squared:  0.6549
## F-statistic: 108.2 on 10 and 555 DF,  p-value: < 2.2e-16
```

```r
#New test for multicollinearity
car::vif(full.model)
```

```
##         GVIF Df GVIF^(1/(2*Df))
## CME 1.657244  1        1.287340
## NTF 2.247043  1        1.499014
## PMG 1.499747  1        1.224642
## ALC 1.127227  1        1.061709
## CDE 1.446683  4        1.047241
## NTN 1.978192  1        1.406482
## FAS 1.042638  1        1.021096
```

```r
#Diagnostic measures
#R² -> 1 (value tending to 1 implies that the predictor variables are well fitted in the model)
#RSE -> 0 (A smaller value implies a smaller model error)
```

Performing the complete fit of the dependent variable **ALP** to a linear model with the remaining 8 variables from the dataset, we verify that the coefficients of variables **CME**, **NTF**, **ARF**, **PMG**, **NTN**, and the intercept are recognized as significant at a significance level of 0, and the 'Yellow' and 'Orange' levels of variable **CDE** at a significance level of 0.001. Thus, for the intercept and predictor variables, we can reject H and accept the alternative hypothesis that there is a significant association between the predictor

variables and the dependent variable.

The multicollinearity assumption was tested and it was discovered that the **ARF** variable had a GVIF > 10 value, which indicates a strong correlation of this variable with other independent variables, breaking this assumption. Therefore, the **ARF** variable was removed from the dataset and the model was recalculated. The new model contained the same significance levels for the variables that remained, with the R^2 value decreasing slightly to 0.661.

Stepwise Selection

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v lubridate 1.9.4     v tibble    3.2.1
## v purrr     1.1.0     v tidyr     1.3.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x tibble::view()  masks summarytools::view()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(caret)
```

```
## Loading required package: lattice
##
## Attaching package: 'caret'
##
## The following object is masked from 'package:purrr':
##
##     lift
```

```
library(leaps)
library(MASS)
```

```
##
## Attaching package: 'MASS'
##
## The following object is masked from 'package:dplyr':
##
##     select
```

```
step.model <- stepAIC(full.model, direction = "both", trace = FALSE)
summary(step.model)
```

```
##
## Call:
## lm(formula = ALP ~ CME + NTF + PMG + CDE + NTN + FAS, data = data_milho)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -84.425 -14.966   0.124  14.708  65.479
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -50.13497    9.42224  -5.321 1.50e-07 ***
## CME               5.43814    0.58949   9.225  < 2e-16 ***
## NTF               6.87414    1.00409   6.846 2.01e-11 ***
## PMG               0.07244    0.01976   3.666  0.00027 ***
## CDECream         -1.50961    2.97124  -0.508  0.61160
## CDELight-yellow   1.69147    3.45035   0.490  0.62416
## CDEYellow        -7.94778    2.64699  -3.003  0.00280 **
## CDEOrange       -13.29146    5.49718  -2.418  0.01593 *
## NTN               9.08775    0.88521  10.266  < 2e-16 ***
## FASPresent        5.10415    2.94968   1.730  0.08411 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.31 on 556 degrees of freedom
## Multiple R-squared:  0.6607, Adjusted R-squared:  0.6552
## F-statistic: 120.3 on 9 and 556 DF,  p-value: < 2.2e-16
```

```
step.modelf <- stepAIC(full.model, direction = "forward", trace = FALSE)
summary(step.modelf)
```

```
##
## Call:
## lm(formula = ALP ~ CME + NTF + PMG + ALC + CDE + NTN + FAS, data = data_milho)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -84.138 -14.475   0.279  14.730  65.701
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -52.851483  10.111563  -5.227 2.45e-07 ***
## CME               5.455422   0.590183   9.244  < 2e-16 ***
## NTF               6.923725   1.006716   6.878 1.65e-11 ***
## PMG               0.072522   0.019767   3.669 0.000267 ***
## ALC               0.002880   0.003879   0.742 0.458230
## CDECream         -1.534346   2.972627  -0.516 0.605949
## CDELight-yellow   1.457687   3.466087   0.421 0.674241
## CDEYellow        -7.821781   2.653498  -2.948 0.003336 **
## CDEOrange       -12.639860   5.569023  -2.270 0.023610 *
## NTN               9.163944   0.891499  10.279  < 2e-16 ***
## FASPresent        5.150903   2.951549   1.745 0.081512 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.32 on 555 degrees of freedom
## Multiple R-squared:  0.661,  Adjusted R-squared:  0.6549
## F-statistic: 108.2 on 10 and 555 DF,  p-value: < 2.2e-16
```

```r
step.modelb <- stepAIC(full.model, direction = "backward", trace = FALSE)
summary(step.modelb)
```

```
##
## Call:
## lm(formula = ALP ~ CME + NTF + PMG + CDE + NTN + FAS, data = data_milho)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -84.425 -14.966   0.124  14.708  65.479
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -50.13497    9.42224  -5.321 1.50e-07 ***
## CME               5.43814    0.58949   9.225  < 2e-16 ***
## NTF               6.87414    1.00409   6.846 2.01e-11 ***
## PMG               0.07244    0.01976   3.666  0.00027 ***
## CDECream         -1.50961    2.97124  -0.508  0.61160
## CDELight-yellow   1.69147    3.45035   0.490  0.62416
## CDEYellow        -7.94778    2.64699  -3.003  0.00280 **
## CDEOrange       -13.29146    5.49718  -2.418  0.01593 *
## NTN               9.08775    0.88521  10.266  < 2e-16 ***
## FASPresent        5.10415    2.94968   1.730  0.08411 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.31 on 556 degrees of freedom
## Multiple R-squared:  0.6607, Adjusted R-squared:  0.6552
## F-statistic: 120.3 on 9 and 556 DF,  p-value: < 2.2e-16
```

```r
AIC(step.model)
```

```
## [1] 5182.498
```

```r
AIC(step.modelf)
```

```
## [1] 5183.936
```

```r
AIC(step.modelb)
```

```
## [1] 5182.498
```

We performed the stepwise selection method from the complete model fit, using the 'both', 'forward', and 'backward' methods. After comparing the AIC values of the 3 selections, we opted for the 'backward' method (lowest AIC). In this method, the coefficients of variables **CME**, **NTF**, **PMG**, **NTN**, and the intercept are considered significant at a significance level of 0, the 'Yellow' and 'Orange' levels of variable **CDE** at a significance level of 0.001, and the **FAS** variable at a significance level of 0.05. Thus, for the intercept and predictor variables, we can reject H and accept the alternative hypothesis that there is a significant association between the predictor variables and the dependent variable.
The residual standard error takes the value of approximately 0.66, which we consider sufficiently significant to proceed with the regression.

Conclusions

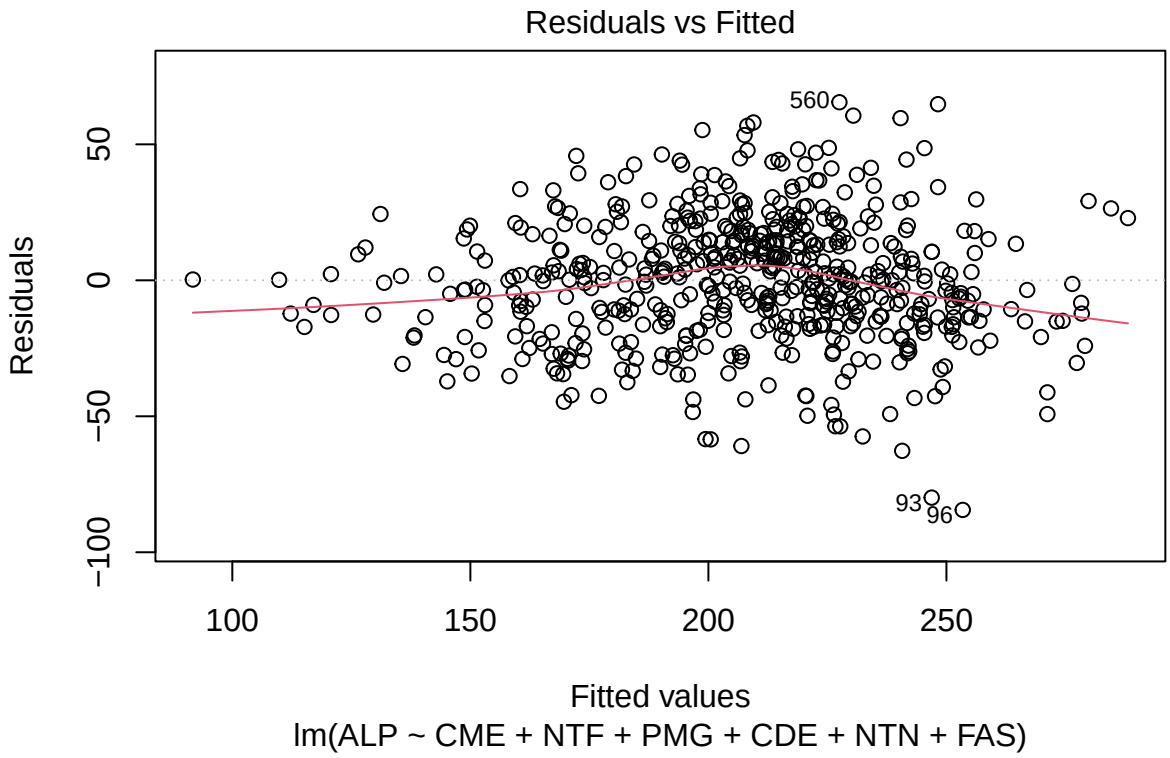
```
#Evaluate residual normalization
shapiro.test(residuals(step.modelb))
```

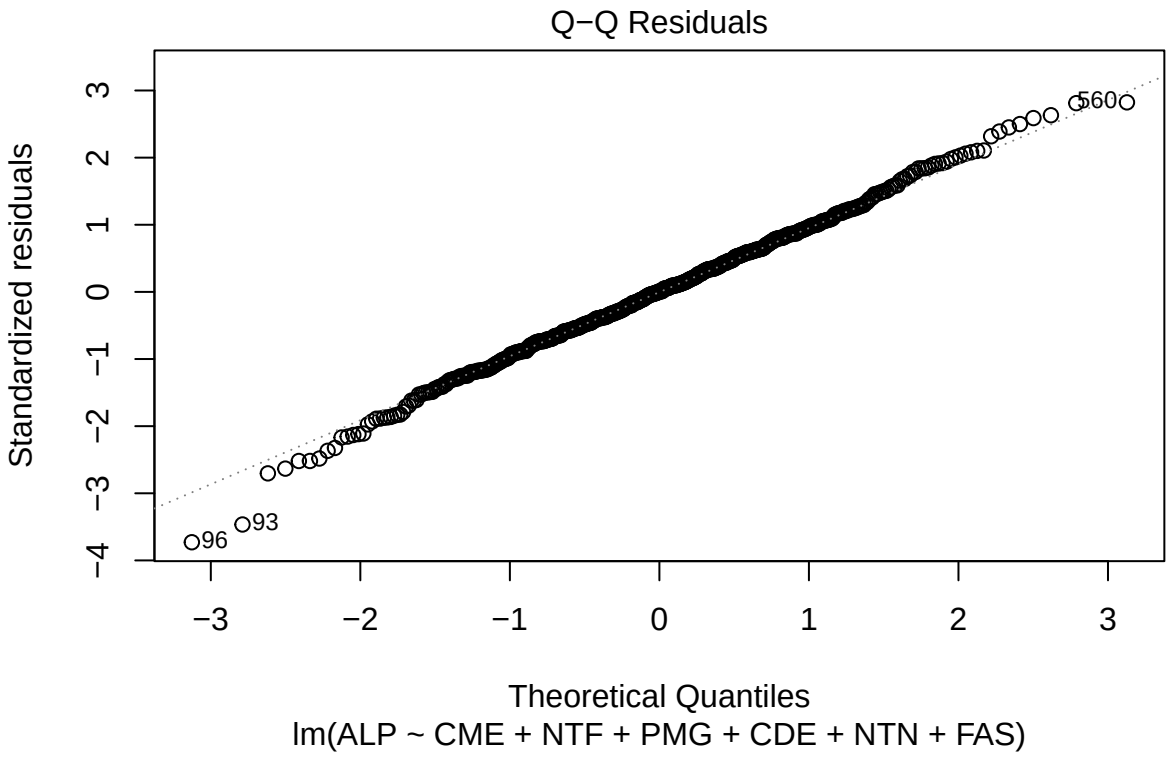
```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(step.modelb)
## W = 0.99714, p-value = 0.4295
```

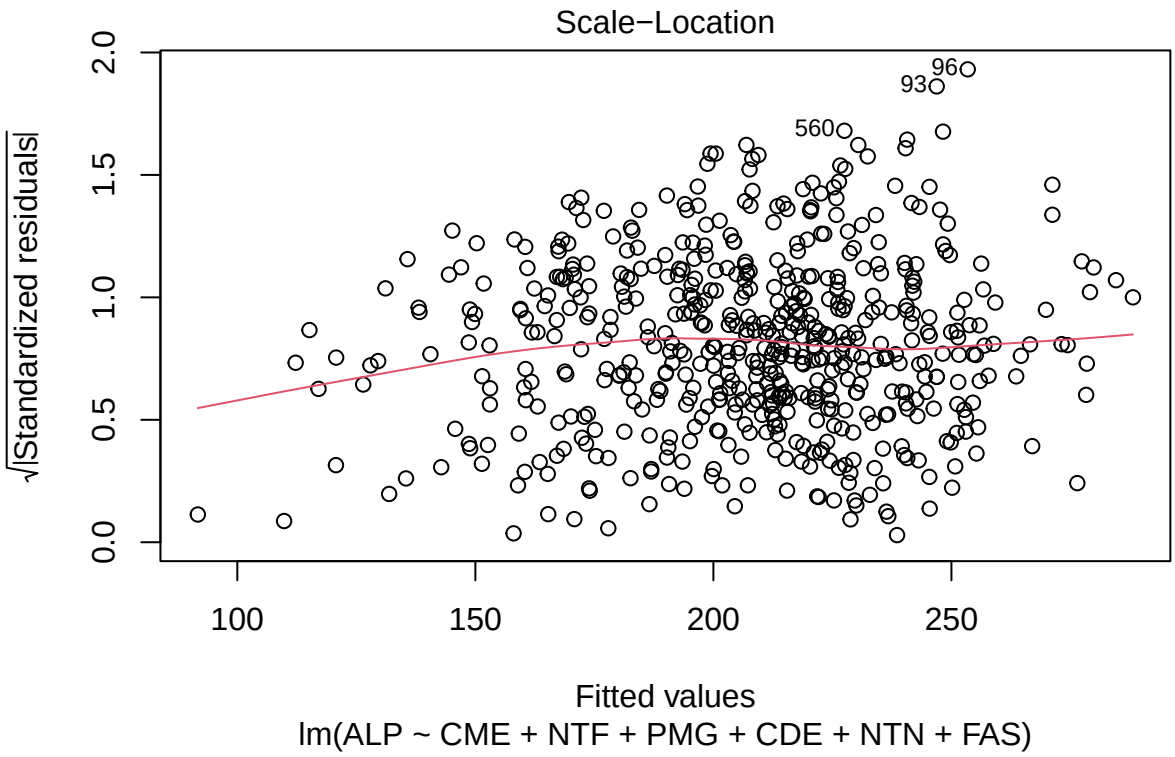
```
#Residual standard error
sigma(step.modelb)/mean(data_milho$ALP)
```

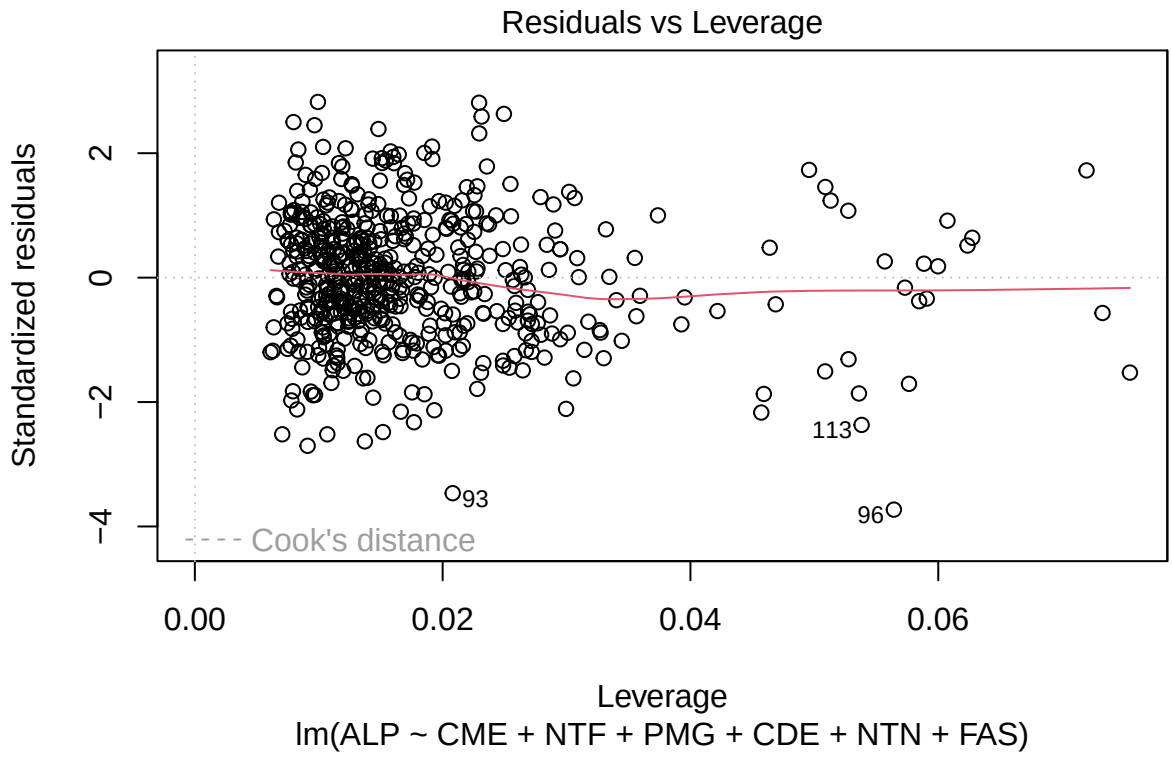
```
## [1] 0.1120357
```

```
#Diagnostic plots
# Test for homoscedasticity assumption
plot(step.modelb)
```




Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(ALP ~ CME + NTF + PMG + CDE + NTN + FAS)
560
93
96


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
96
93
560
100
150
200
250
Fitted values
lm(ALP ~ CME + NTF + PMG + CDE + NTN + FAS)




```
#Model confidence interval
confint(step.modelb)
```

```
##                         2.5 %      97.5 %
## (Intercept)     -68.64251284 -31.6274193
## CME               4.28025225   6.5960335
## NTF               4.90186319   8.8464149
## PMG               0.03363078   0.1112514
## CDECream         -7.34584060   4.3266130
## CDELight-yellow  -5.08584978   8.4687938
## CDEYellow       -13.14711014  -2.7484502
## CDEOrange       -24.08924496  -2.4936813
## NTN               7.34898060  10.8265189
## FASPresent       -0.68973826  10.8980377
```

```
#Model ANOVA
anova(step.modelb)
```

```
## Analysis of Variance Table
##
## Response: ALP
##            Df Sum Sq Mean Sq  F value    Pr(>F)
## CME         1 368547  368547 678.5021 < 2.2e-16 ***
## NTF         1 139824  139824 257.4184 < 2.2e-16 ***
## PMG         1   2536    2536   4.6687   0.03114 *
```

```
## CDE         4  19762    4940   9.0953 4.024e-07 ***
## NTN         1  55800   55800 102.7287 < 2.2e-16 ***
## FAS         1   1626    1626   2.9943   0.08411 .
## Residuals 556 302007     543
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The normality assumption of residuals was verified through a Shapiro-Wilk test, obtaining a p-value $>$ 0.05, thus proving the normality of residuals. The 95% confidence intervals for the intercept and predictor variables were computed.

The homoscedasticity assumption was tested using the residuals vs fitted values scatter plot. Since the dispersion does not present a conical shape, we consider this assumption satisfied.

The quality of the final linear model was assessed through the R-squared and Residual Standard Error metrics, obtaining a value of approximately 0.6689 for the first and an estimated error rate of 11% for the second.
Satisfying all requirements, we define as our final model **ALP** ~ **CME** + **NTF** + **PMG** + **CDE** + **NTN** + **FAS**, with the stepwise selection method only excluding the **ALC** variable. The ANOVA analysis of the final model supports these conclusions.

We conclude that from a set of morphological descriptors it is possible to establish linear regression models for plant height, with this model having utility from the perspective of eliminating missing data through its prediction using this model, on the datasets of regional corn populations conserved in the BPGV.