Project 3: Variable Selection and Regression Analysis on Maize Flowering Time

Rahrooh, Allen
Department of Statistics & Data Science
University of Central Florida
Orlando, United States of America
rahrooan@knights.ucf.edu

***Abstract*— Maize also known as corn has been used as a crop for over 10,000 years originating in southern Mexico. Past researchers have used various regression techniques such as lasso and elastic net regression to see if certain markers have a correlation with the male flowering time using variable selection. I have chosen to explore variable selection and then do a multiple logistic and linear regression to see if any correlation arises from the selected predictors and the male flowering time.**

I. Introduction & Background

Maize has been used for the past 10,000 years as a main source of food for South American countries originating from Southern Mexico. Maize is also known as corn in other parts of the world and in the United States of America it can be grown across the country but the majority of corn production comes from the heartland region which includes Illinois, Iowa, Indiana, South Dakota, Nebraska, Kentucky, Ohio, and Missouri. The maize crop has many other uses besides food as recent research and development has used maize to produce ethanol, which can be used to power gas operated machines. Scientists [1] [2] are interested in studying the various flowering times of maize crops. Past research has used regression methods such as lasso and elastic net regression to study the genome data such as maize flowering time with the associated genome codes. For my analysis of the maize data provided by [2] I will first perform a variable selection by taking subsets of the data to see which subset has the most significant variables and from there I will eliminate useless variables and then perform a regression analysis and report my findings.

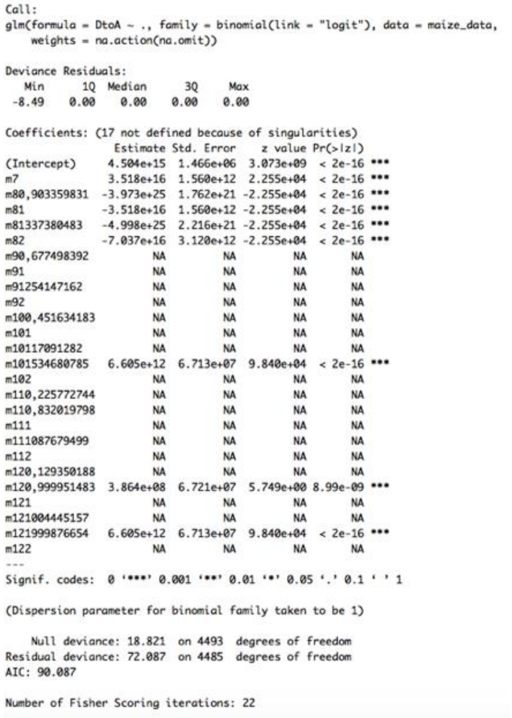
II. Methods

A. Variable Selection

To begin the variable selection I started by importing the maize data into Microsoft Excel and formatting the semicolon delimited data. Then I imported the data into R Studio using the fread command since the dimension of the data was 4494 x 7394. The dataset included some NA values that I omitted so the regression analysis would not give an error since regression takes in numerical data and not character data for the predictors. Also I omitted the geno_code, pop, and entry columns since they did not contain information useful for the regression analysis. I decided to take three subsets one from the beginning, one from the middle, and one from the end of the predictors m1 – m7389, which present the single-nucleotide polymorphisms (SNP) markers of each geno-code. I used a logistic regression model (1) to select the best subset with the male flowering time (DtoA) being the response variable and the selected SNP markers being the predictors.

$$Y = \frac{e^{\beta_0+\beta_1 x_1+\cdots\,\beta_n x_n}}{1+\,e^{\beta_0+\beta_1 x_1+\cdots\,\beta_n x_n}} \qquad (1)$$

In (1) the $\beta's$ represent the weights per each predictor x. I decided to go with the beginning subset because it had the most significant predictors as shown in Fig. 1.



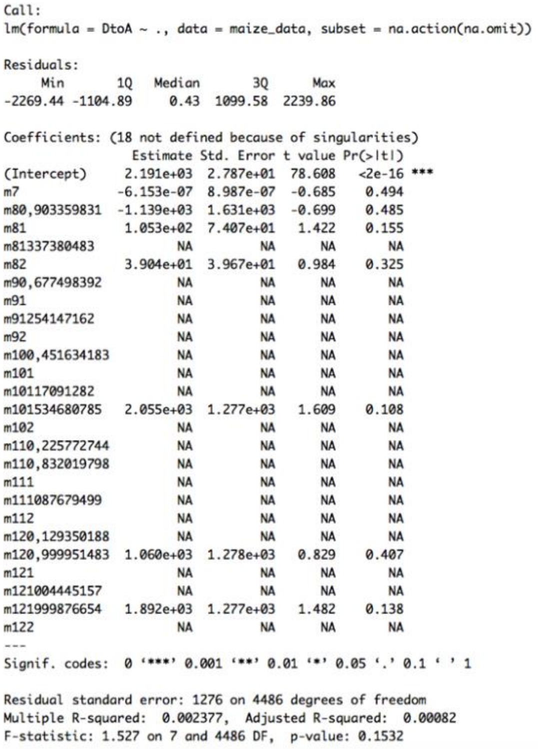
```
Call:
glm(formula = DtoA ~ ., family = binomial(link = "logit"), data = maize_data,
    weights = na.action(na.omit))

Deviance Residuals:
   Min      1Q  Median      3Q     Max
 -8.49    0.00    0.00    0.00    0.00

Coefficients: (17 not defined because of singularities)
                Estimate Std. Error   z value Pr(>|z|)
(Intercept)    4.504e+15  1.466e+06  3.073e+09  < 2e-16 ***
m7             3.518e+16  1.560e+12  2.255e+04  < 2e-16 ***
m80,903359831 -3.973e+25  1.762e+21 -2.255e+04  < 2e-16 ***
m81           -3.518e+16  1.560e+12 -2.255e+04  < 2e-16 ***
m81337380483  -4.998e+25  2.216e+21 -2.255e+04  < 2e-16 ***
m82           -7.037e+16  3.120e+12 -2.255e+04  < 2e-16 ***
m90,677498392         NA         NA         NA       NA
m91                   NA         NA         NA       NA
m91254147162          NA         NA         NA       NA
m92                   NA         NA         NA       NA
m100,451634183        NA         NA         NA       NA
m101                  NA         NA         NA       NA
m10117091282          NA         NA         NA       NA
m101534680785  6.605e+12  6.713e+07  9.840e+04  < 2e-16 ***
m102                  NA         NA         NA       NA
m110,225772744        NA         NA         NA       NA
m110,832019798        NA         NA         NA       NA
m111                  NA         NA         NA       NA
m111087679499         NA         NA         NA       NA
m112                  NA         NA         NA       NA
m120,129350188        NA         NA         NA       NA
m120,999951483 3.864e+08  6.721e+07  5.749e+00 8.99e-09 ***
m121                  NA         NA         NA       NA
m121004445157         NA         NA         NA       NA
m121999876654  6.605e+12  6.713e+07  9.840e+04  < 2e-16 ***
m122                  NA         NA         NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 18.821  on 4493  degrees of freedom
Residual deviance: 72.087  on 4485  degrees of freedom
AIC: 90.087

Number of Fisher Scoring iterations: 22
```

Fig. 1 Beginning Subset Using a Multiple Logistic Regression Model

I also ran a multiple linear regression model (2) to see if any other variables can be eliminated before performing the full regression analysis.

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_2 x_2 + \varepsilon \tag{2}$$

In (2) the $\beta's$ represent the weights per each predictor x but as a linear function instead of a logistic function (1).



```
Call:
lm(formula = DtoA ~ ., data = maize_data, subset = na.action(na.omit))

Residuals:
     Min       1Q   Median       3Q      Max 
-2269.44 -1104.89     0.43  1099.58  2239.86 

Coefficients: (18 not defined because of singularities)
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     2.191e+03  2.787e+01  78.608   <2e-16 ***
m7             -6.153e-07  8.987e-07  -0.685    0.494    
m80,903359831  -1.139e+03  1.631e+03  -0.699    0.485    
m81             1.053e+02  7.407e+01   1.422    0.155    
m81337380483           NA         NA      NA       NA    
m82             3.904e+01  3.967e+01   0.984    0.325    
m90,677498392          NA         NA      NA       NA    
m91                    NA         NA      NA       NA    
m91254147162           NA         NA      NA       NA    
m92                    NA         NA      NA       NA    
m100,451634183         NA         NA      NA       NA    
m101                   NA         NA      NA       NA    
m10117091282           NA         NA      NA       NA    
m101534680785   2.055e+03  1.277e+03   1.609    0.108    
m102                   NA         NA      NA       NA    
m110,225772744         NA         NA      NA       NA    
m110,832019798         NA         NA      NA       NA    
m111                   NA         NA      NA       NA    
m111087679499          NA         NA      NA       NA    
m112                   NA         NA      NA       NA    
m120,129350188         NA         NA      NA       NA    
m120,999951483  1.060e+03  1.278e+03   0.829    0.407    
m121                   NA         NA      NA       NA    
m121004445157          NA         NA      NA       NA    
m121999876654   1.892e+03  1.277e+03   1.482    0.138    
m122                   NA         NA      NA       NA    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1276 on 4486 degrees of freedom
Multiple R-squared:  0.002377,	Adjusted R-squared:  0.00082 
F-statistic: 1.527 on 7 and 4486 DF,  p-value: 0.1532
```

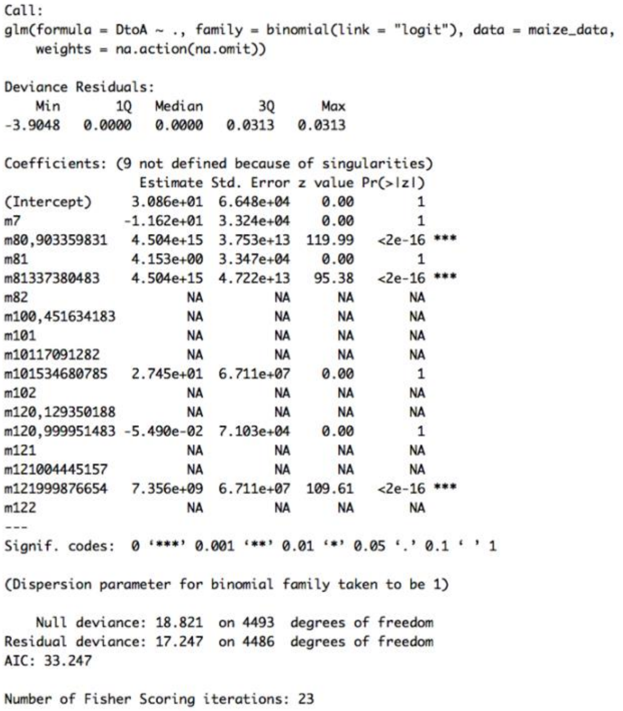
Fig. 2 Additional Variable Selection Using a Multiple Linear Regression Model

I chose the mulitple logistic and linear regression models to see if a simple regression model can produce the same results as a more complex regression such as lasso and elastic net regression.

B. *Regression Analysis*

From the two variable selection models of multiple logistic and linear regression I decide to get eliminate the predictors m10 and m11, since they produced NA values as shown in Fig. 1. So there is only four predictors left (m7, m8, m10, m12), which will be used for the full regression analysis. I then ran the multiple logistic and linear regression again with the variable selected data and reported the summary of each model statistics and corresponding graphs. The plots produced will be the residuals vs. fitted, normal Q-Q, scale-location, and residuals vs. leverage plots. The residuals vs. fitted plot shows if the residuals have a non-linear pattern. The normal Q-Q plot shows if the residuals are normally distributed. The scale-location plot shows if the residuals are spread equally along the ranges of the predictors (x_1,x_2, and x_n). The residuals vs. leverage plot shows if the data has any outliers using cook's distance (red dashed line).

III. RESULTS



```
Call:
glm(formula = DtoA ~ ., family = binomial(link = "logit"), data = maize_data, 
    weights = na.action(na.omit))

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-3.9048   0.0000   0.0000   0.0313   0.0313  

Coefficients: (9 not defined because of singularities)
                 Estimate Std. Error z value Pr(>|z|)    
(Intercept)     3.086e+01  6.648e+04    0.00        1    
m7             -1.162e+01  3.324e+04    0.00        1    
m80,903359831   4.504e+15  3.753e+13  119.99   <2e-16 ***
m81             4.153e+00  3.347e+04    0.00        1    
m81337380483    4.504e+15  4.722e+13   95.38   <2e-16 ***
m82                    NA         NA      NA       NA    
m100,451634183         NA         NA      NA       NA    
m101                   NA         NA      NA       NA    
m10117091282           NA         NA      NA       NA    
m101534680785   2.745e+01  6.711e+07    0.00        1    
m102                   NA         NA      NA       NA    
m120,129350188         NA         NA      NA       NA    
m120,999951483 -5.490e-02  7.103e+04    0.00        1    
m121                   NA         NA      NA       NA    
m121004445157          NA         NA      NA       NA    
m121999876654   7.356e+09  6.711e+07  109.61   <2e-16 ***
m122                   NA         NA      NA       NA    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 18.821  on 4493  degrees of freedom
Residual deviance: 17.247  on 4486  degrees of freedom
AIC: 33.247

Number of Fisher Scoring iterations: 23
```

Fig. 3 Summary of Multiple Logistic Regression after Variable Selection

From Fig. 3 we see that 3 out of the 4 selected predictors are significant with significant codes close to zero. Also the Akaike information criterion (AIC) value is 33.247, which is lower then the AIC before variable selection of 90.087 which represents the model has a better fit.

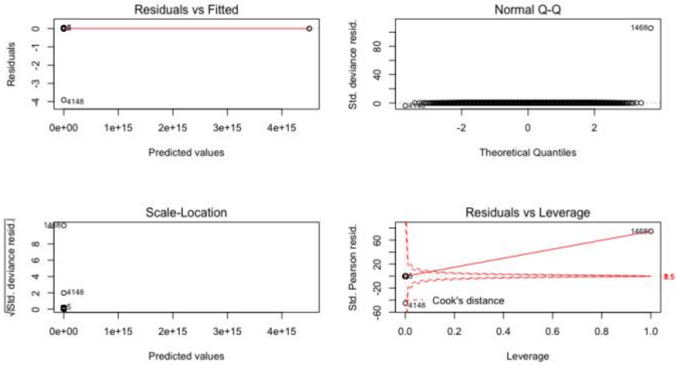
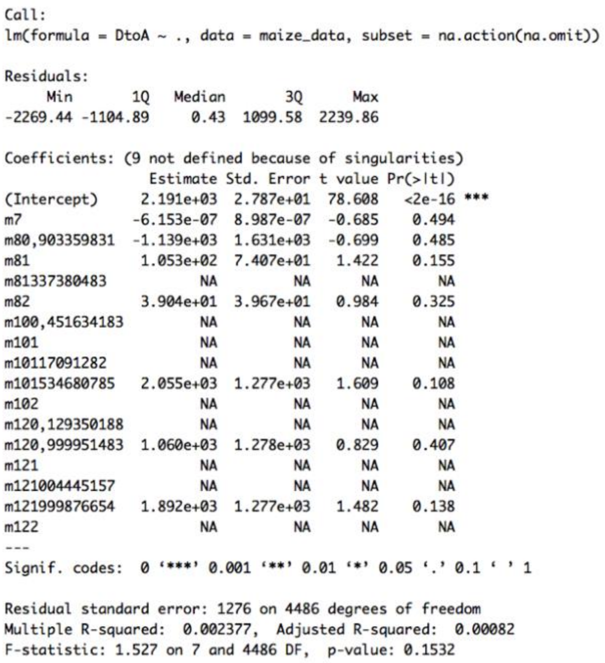


Fig. 4 Multiple Logistic Regression Plots

The residuals vs. fitted shows the residuals do not have a nonlinear pattern with the residuals being close to zero. The Q-Q plot shows that the residuals are normally distributed which they are not following more of an uniform distribution. The scale-location shows the spread of the residuals and it shows that there are some outliers, which can create a nonlinear model. The residuals vs. leverage shows cook's distance and that there is an outlier at observation 4148.



```
Call:
lm(formula = DtoA ~ ., data = maize_data, subset = na.action(na.omit))

Residuals:
     Min       1Q   Median       3Q      Max 
-2269.44 -1104.89     0.43  1099.58  2239.86 

Coefficients: (9 not defined because of singularities)
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     2.191e+03  2.787e+01  78.608   <2e-16 ***
m7             -6.153e-07  8.987e-07  -0.685    0.494    
m80,903359831  -1.139e+03  1.631e+03  -0.699    0.485    
m81             1.053e+02  7.407e+01   1.422    0.155    
m81337380483           NA         NA      NA       NA    
m82             3.904e+01  3.967e+01   0.984    0.325    
m100,451634183         NA         NA      NA       NA    
m101                   NA         NA      NA       NA    
m10117091282           NA         NA      NA       NA    
m101534680785   2.055e+03  1.277e+03   1.609    0.108    
m102                   NA         NA      NA       NA    
m120,129350188         NA         NA      NA       NA    
m120,999951483  1.060e+03  1.278e+03   0.829    0.407    
m121                   NA         NA      NA       NA    
m121004445157          NA         NA      NA       NA    
m121999876654   1.892e+03  1.277e+03   1.482    0.138    
m122                   NA         NA      NA       NA    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1276 on 4486 degrees of freedom
Multiple R-squared:  0.002377,	Adjusted R-squared:  0.00082 
F-statistic: 1.527 on 7 and 4486 DF,  p-value: 0.1532
```

Fig. 5 Summary of Multiple Linear Regression Model After Variable Selection

From Fig. 5 which represents the multiple linear regression model after variable selection shows that only the intercept is a significant predictor with a significant code close to zero. Also the Adjusted R-squared value is 0.00082, which is the same as before variable selection so that means the predictors are not positively linear correlated with the male flowering time since the highest Adjusted R-squared value is 1.0 and the value is close to zero.

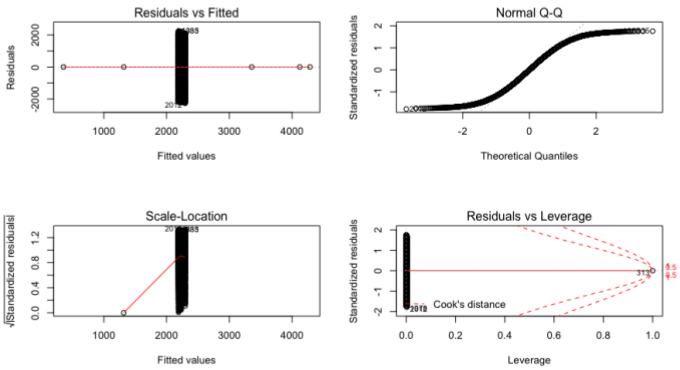


Fig. 6 Multiple Linear Regression Plots

The residuals vs. fitted show that the residuals have a non-linear pattern because a lot of residuals are at the fitted value of around 2000. The normal Q-Q plot shows that the data follows more of a logit function then a normal distribution. The scale-location plot shows the spread of the residuals and as in the residuals vs. fitted plot a majority of the residuals are at the fitted value of 2000. The residuals vs. leverage shows cook distance and that there is an outlier at observation 313.

IV. Discussion

We can see that for the multiple linear and logistic regression models are not well fitted for the maize dataset. Logistic regression works well for binary response variables where the male flowering time is a numerical variable. For the linear regression the SNP markers being the predictors only had three values assigned being 0,1 and 2 so that is why the adjusted R-squared was close to zero since multiple SNP markers can have multiple 0,1, or 2 values and the NA values were treated as zeros.

V. Future Work & Conclusions

Overall the multiple logistic and linear regression did not fit the maize data well with very low adjusted R-Squared of 0.00082, which shows very little correlation between the SNP markers and the male flowering time ,and lower AIC from 90.087 to 33.247. For future analysis I can look into forward and backward stepwise variable selection to identify other significant predictors from the maize dataset. I can also look at other regression methods such as lasso, ridge, and elastic net regression to see if the results are better than the multiple linear and logistic regression models.

Acknowledgment

I would like to thank the authors [1] for providing the reference text for the regression analysis of genome-wide association studies. Also [2] for providing the maize data to the public.

References

[1] Waldmann, Patrik, Gábor Mészáos, Birgit Gredler, Christian Fuerst, and Johann Sölkner. “Evaluation of the lasso and the elastic net in genome-wide association studies.” *Frontiers in genetics* 4 (2013): 270.

[2] Buckler et al. (2009), “The Genetic Architecture of Maize Flowering Time,” Science 325, 714-718.