Connor Eng

111612084

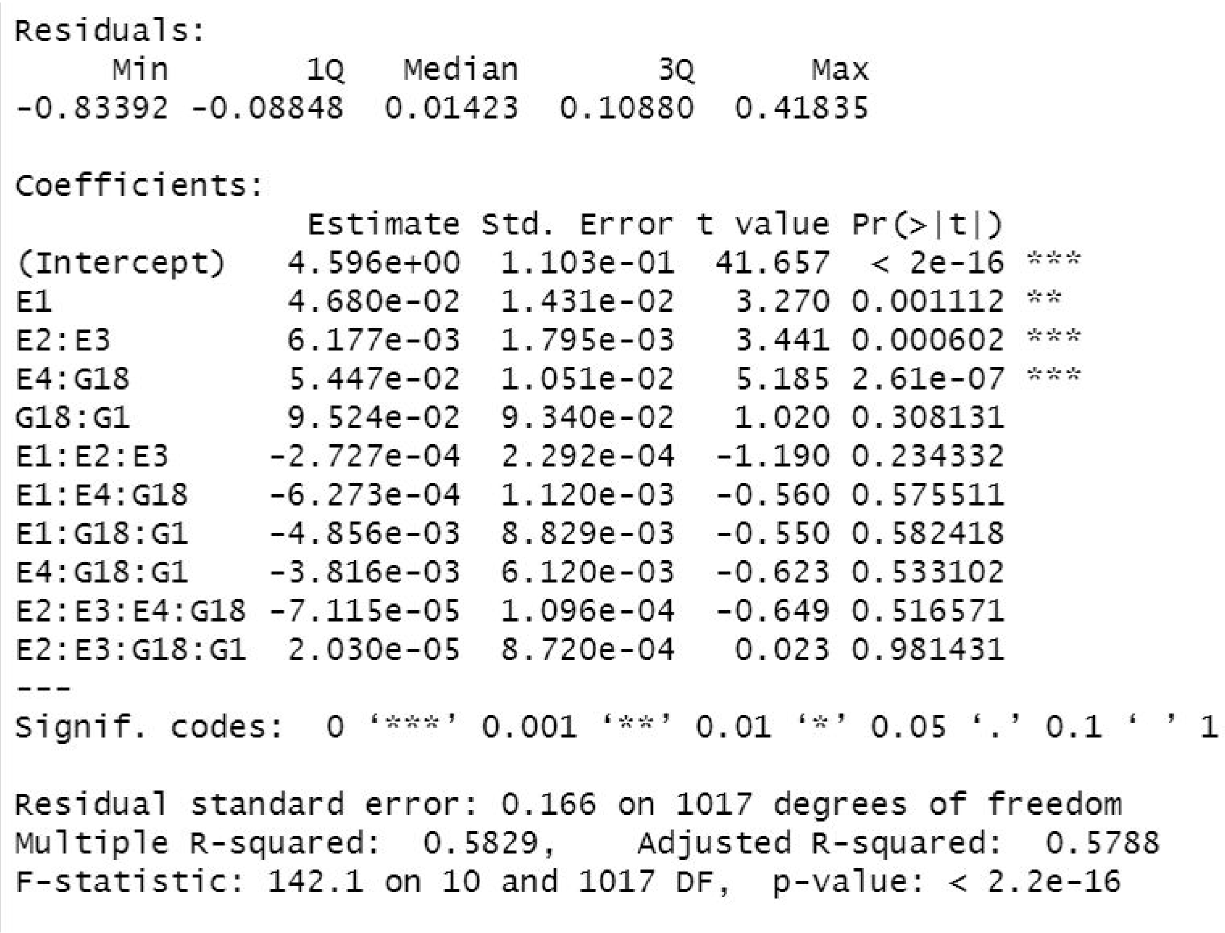
AMS 315 Project 2 Report

Introduction:

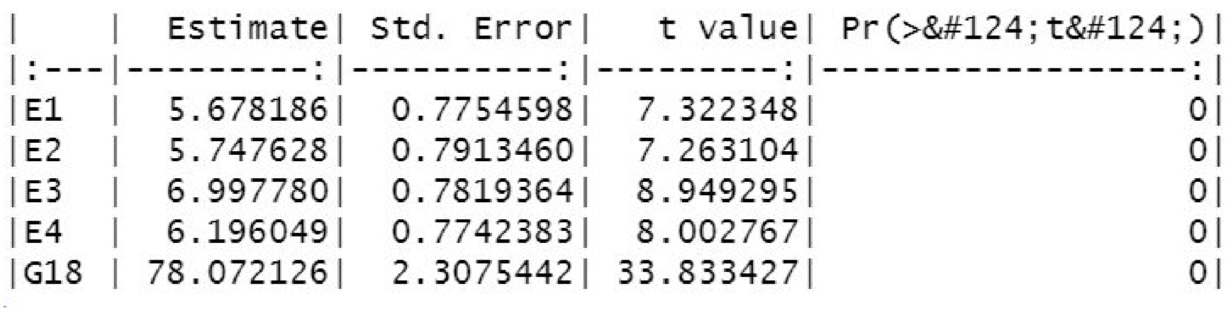
This assignment includes 1028 observations in my given data set. The data set also includes a single dependent variable, which is the Y column and 24 total independent variable, which include E1-E4 and G1-G20. The values names "E" are the environmental variables and the values named "G" are gene variables. In my file, not all of the variables will have a significant impact, therefore those values will not be in my function.

Method:

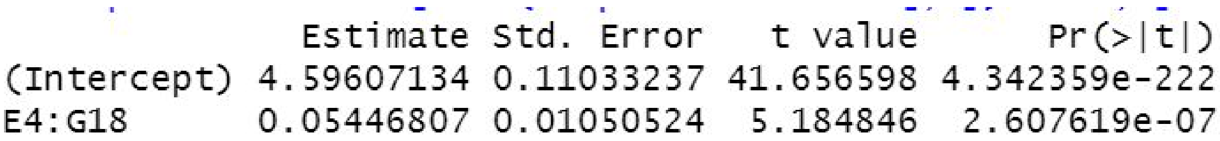
The first step I took in this project was to open up my given file to download into RStudio. Next, I created a model of only the environmental variables using the function lm() to help me evaluate the effects. The adjusted R^2 value for just the environmental variables was found to be 0.0887. After, I created a model using the lm() function that contained both environmental and gene variables and assumed that there will only be up to the second order interactions. Using this model, I created a residual plot to visually see the spread of the data points. After seeing the spread of the data, I was not 100% sure that it has heteroscedasticity, therefore I used the function boxcox() to see if my data needed to be transformed. The boxcox graph showed me that there is not need for transformation because the graph plateaus where lambda equals 1. I then calculated the adjusted R^2 value to be 0.579583. Next, I used the library leap, regsubsets, colnames, and apply function to help create the k-table of model, adjusted R^2, and BIC. This table showed me that the adjusted R^2 values increased but started to plateau down the table and the BIC values decreased and also started to plateau going down the table. To find the main significant values, I used the lm() function of just the Y values because my data file did not require transformation. This table told me that E1 to E4 and G18 were all significant variables because their t-values were high and generally above the value of 4. I then wanted to create a table for the second interaction to find values that are usable for second step regression. The second table showed me that the value E4:G18 was significant because the t-value was high at 6.33 and the Pr was low at 0. Finally, I used the lm() function on the significant values to test the second stage.



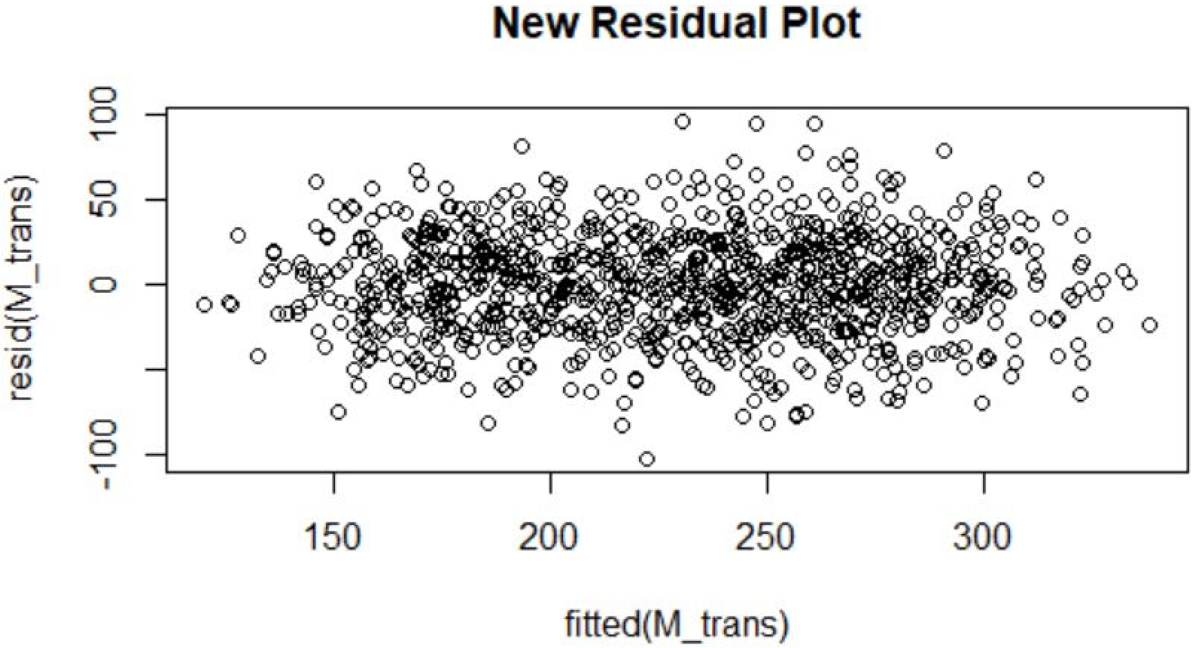
```
Residuals:
     Min       1Q   Median       3Q      Max 
-0.83392 -0.08848  0.01423  0.10880  0.41835 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)   4.596e+00  1.103e-01  41.657  < 2e-16 ***
E1            4.680e-02  1.431e-02   3.270 0.001112 ** 
E2:E3         6.177e-03  1.795e-03   3.441 0.000602 ***
E4:G18        5.447e-02  1.051e-02   5.185 2.61e-07 ***
G18:G1        9.524e-02  9.340e-02   1.020 0.308131    
E1:E2:E3     -2.727e-04  2.292e-04  -1.190 0.234332    
E1:E4:G18    -6.273e-04  1.120e-03  -0.560 0.575511    
E1:G18:G1    -4.856e-03  8.829e-03  -0.550 0.582418    
E4:G18:G1    -3.816e-03  6.120e-03  -0.623 0.533102    
E2:E3:E4:G18 -7.115e-05  1.096e-04  -0.649 0.516571    
E2:E3:G18:G1  2.030e-05  8.720e-04   0.023 0.981431    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.166 on 1017 degrees of freedom
Multiple R-squared:  0.5829,	Adjusted R-squared:  0.5788 
F-statistic: 142.1 on 10 and 1017 DF,  p-value: < 2.2e-16
```

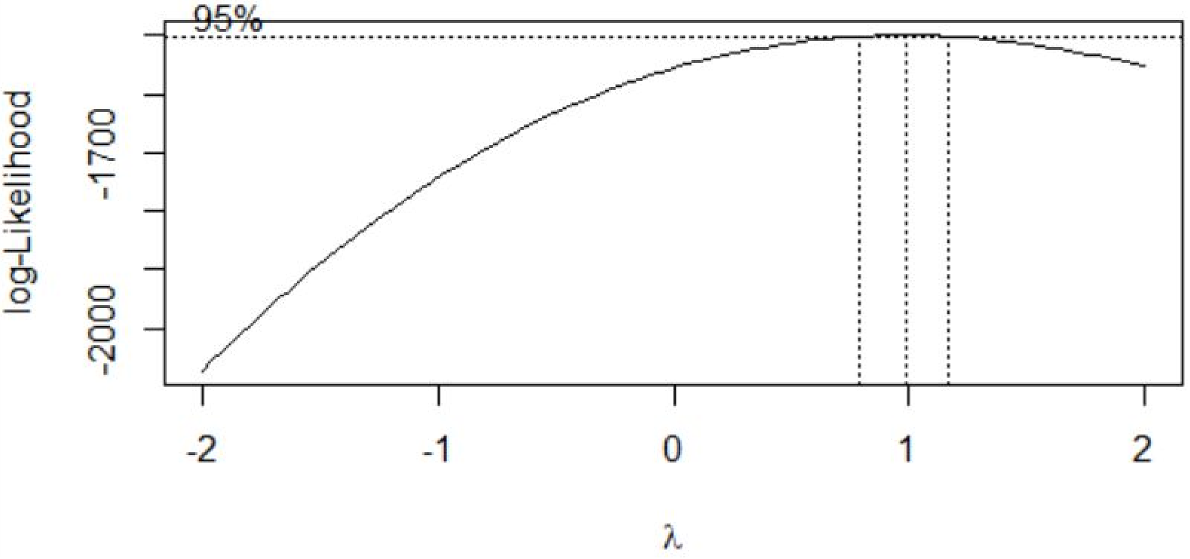


```
|    |  Estimate| Std. Error|   t value| Pr(>&#124;t&#124;)|
|:---|---------:|----------:|---------:|------------------:|
|E1  |  5.678186|  0.7754598|  7.322348|                  0|
|E2  |  5.747628|  0.7913460|  7.263104|                  0|
|E3  |  6.997780|  0.7819364|  8.949295|                  0|
|E4  |  6.196049|  0.7742383|  8.002767|                  0|
|G18 | 78.072126|  2.3075442| 33.833427|                  0|
```



```
              Estimate Std. Error   t value      Pr(>|t|)
(Intercept) 4.59607134 0.11033237 41.656598 4.342359e-222
E4:G18      0.05446807 0.01050524  5.184846  2.607619e-07
```


New Residual Plot
resid(M_trans)
100
50
0
-100
150
200
250
300
fitted(M_trans)


95%
log-Likelihood
-1700
-2000
-2
-1
0
1
2
λ

Result:

The model is Y = (4.59607134 + 0.04680 E1 + 0.006177 E2:E3 + 0.05447 E4:G18)^2. When constructing a 90% confidence interval, it is found that the variables E1, E2:E3, and E4:G18 are 0.001112, 0.000602, and 0.000000261. Since the all of the P-values of the variables are close to 0, we can conclude that we reject the null hypothesis at .1.

Conclusion:

In conclusion, I found that the model is Y = (4.59607134 + 0.04680 E1 + 0.006177 E2:E3 + 0.05447 E4:G18)^2 and that the null hypothesis is rejected. Since 0 was not a value in the confidence interval, and the P-values of the significant variables were very close to 0, this is why I can conclude the rejection. I can also conclude that the adjusted R^2 for the second step regression was valid because the value was 0.5788 which is large enough.

R Code:

```
data = read.csv("C:/Users/Connor/Documents/AMS 315/P2.csv", header=TRUE)
library(knitr)

M_E <- lm(Y ~ E1+E2+E3+E4, data=data)
summary(M_E)
```

```
M_raw <- lm( Y ~
(E1+E2+E3+E4+G1+G2+G3+G4+G5+G6+G7+G8+G9+G10+G11+G12+G13+G14+G15+G16
+G17+G18+G19+G20)^2, data=data)

plot(resid(M_raw) ~ fitted(M_raw), main='Residual Plot')

MASS::boxcox(M_raw)

M_trans <- lm( I((Y)) ~ (.)^2, data=data )

summary(M_raw)$adj.r.square
summary(M_trans)$adj.r.square

plot(resid(M_trans) ~ fitted(M_trans), main='New Residual Plot')

library(leaps)
M <- regsubsets( model.matrix(M_trans)[,-1], I(data$Y),
        nbest = 1 , nvmax=5,
        method = 'forward', intercept = TRUE )
temp <- summary(M)

Var <- colnames(model.matrix(M_trans))
```

```
M_select <- apply(temp$which, 1,
            function(x) paste0(Var[x], collapse='+'))
kable(data.frame(cbind( model = M_select, adjR2 = temp$adjr2, BIC = temp$bic)),
    caption='Model Summary')


M_main <- lm( I((Y)) ~ ., data=data)
temp <- summary(M_main)
kable(temp$coefficients[ abs(temp$coefficients[,4]) <= 0.001, ], caption='Sig Coefficients')


M_2nd <- lm( I((Y)) ~ (.)^2, data=data)
temp  <- summary(M_2nd)
kable(temp$coefficients[ abs(temp$coefficients[,4]) <= 0.1, ], caption='2nd Interaction')


M_2stage <- lm( I(log(Y)) ~ (E1+E2:E3+E4:G18+G1:G18)^2, data=data)
temp <- summary(M_2stage)
temp$coefficients[ abs(temp$coefficients[,3]) >= 4, ]
print(temp)
```