Stephen Hanna
(PROJECT PARTNER NAME OMITTED FOR PRIVACY)
Group 26: Project II

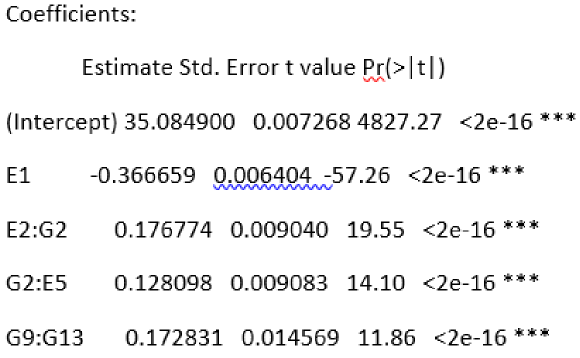
Introduction

For this project we were given an excel sheet with two thousand rows of data. Each row consisted of one dependent variable, Y, and twenty independent variables from E1 to E5 and G1 to G15. This data is used to find the gene-environment interaction where we determine what genes, environmental variables, or interactions between any of the independent variables lead to the dependent variable values. The goal for this project is to find the function used to generate our data.

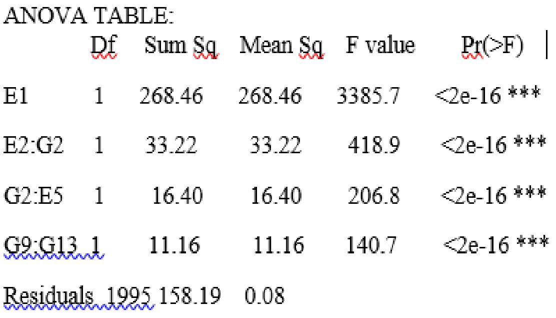
Methodology

We continued to work with the statistical package R for this project. We assigned our data from the excel sheet to R. First we tested for normality by using the Shapiro-Wilk Normality Test and found out that the data was not normal. Consequently, we used a boxcox transformation to make the dependent variable data normal. We proceeded to make a linear fit model that included all of the significant pairwise and triowise interactions. In order to find the significant correlations between independent variables and the dependent variable, a regression analysis was needed. Lasso was the chosen regression analysis because it penalizes covariates that don't have large enough coefficient, allowing for a simpler model to be chosen[1]. After transforming our independent variables to a matrix and our transformed dependent variables into a vector we used the lasso technique to help select the important independent variables. The important independent variables and interactions were those that had non zero coefficients in the lasso model and we implemented those variables into a separate linear model. The additional linear model was created so the p values could be measured and so the coefficients of those selected variables would no longer be penalized, possibly allowing for a better model.

Results

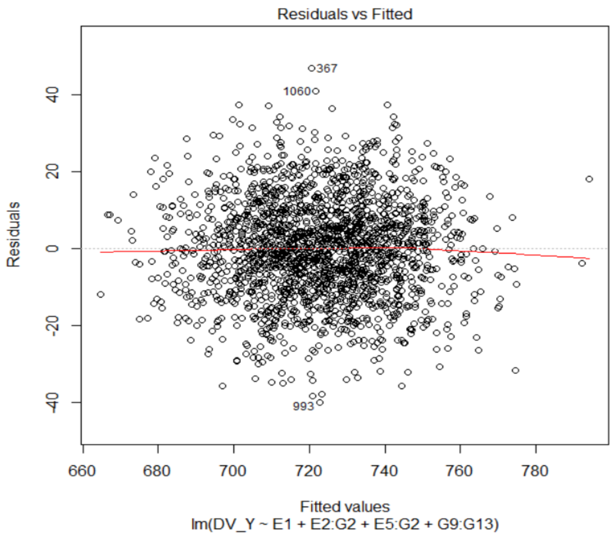


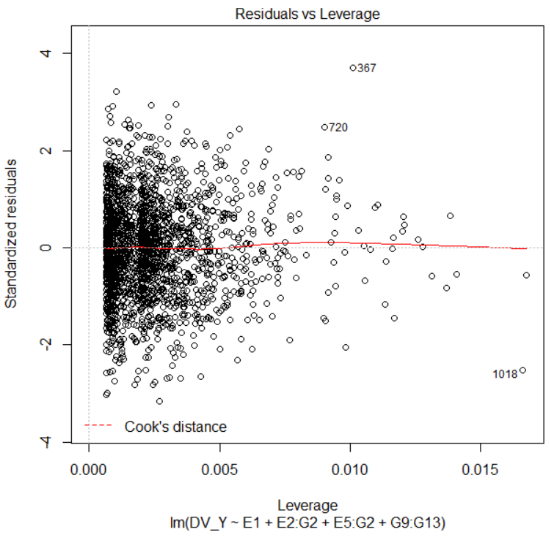
```
Coefficients:

        Estimate Std. Error t value Pr(>|t|)

(Intercept) 35.084900   0.007268 4827.27   <2e-16 ***

E1        -0.366659   0.006404  -57.26   <2e-16 ***

E2:G2        0.176774   0.009040   19.55   <2e-16 ***

G2:E5        0.128098   0.009083   14.10   <2e-16 ***

G9:G13        0.172831   0.014569   11.86   <2e-16 ***
```



```
ANOVA TABLE:
             Df   Sum Sq   Mean Sq   F value      Pr(>F)
E1            1   268.46    268.46    3385.7    <2e-16 ***
E2:G2         1    33.22     33.22     418.9    <2e-16 ***
G2:E5         1    16.40     16.40     206.8    <2e-16 ***
G9:G13        1    11.16     11.16     140.7    <2e-16 ***
Residuals  1995   158.19      0.08
```

[1] Lasso (statistics). (2018, April 03). Retrieved from https://en.wikipedia.org/wiki/Lasso_(statistics)





The analysis of variance table for our model is shown above. Based on this table we found that our model's equation is:

$$\mathrm{Y} = \left((-.36666 * \mathrm{E1} + .17677 * \mathrm{E2G2} + .12810 * \mathrm{G2E5} + .17283 * \mathrm{G9G13} + 35.085) * 3.1 + 1\right)^{\frac{1}{3.1}}$$

The 3.1 and +1 seen is to convert the values back from the boxcox transformation, so the model directly corresponds to the original dependent variable data set. We also have the Residuals vs Fitted graph that shows our data has a homogenous spread[2]. Since we cannot clearly see the dashed lines for Cook's distance in the Residuals vs Leverage plot we know that all of our cases are well within Cook's distance which suggests that outliers have no huge influence[3]. Our adjusted R^2 is .6748, which explains about 67.5% of the variance in the dependent variable.

Discussion

One of the issues with our methodology was that the boxcox transformation doesn't necessarily make the data normal. It did make data satisfy the shapiro test for normality, but that test is rather flawed. However, testing for normality for large data sets is often useless because as your sample size gets larger, it becomes easier for an awkward data point to throw off the normality of the entire set. Another problem was that several variables concluded to be significant by lasso were eliminated because of their lack of significance in the final linear model. This was primarily to satisfy the instructions of the assignment with regard to p value, but also because they didn't have large impact on the R^2 value, but it may be the case that they would play a role in, for example, disease formation if this was an epidemic data set. Despite the limitations, the model met other criteria like tests for a linear relationship in the residual vs fitted, tests for homoscedasticity in a scale location plot that is not in this report, tests for highly influential data points from residual vs leverage, and had a relatively large R^2 value.

[2] B. (n.d.). University of Virginia Library Research Data Services Sciences. Retrieved May 3, 2018, from http://data.library.virginia.edu/diagnostic-plots/

[3] B. (n.d.). University of Virginia Library Research Data Services Sciences. Retrieved May 3, 2018, from http://data.library.virginia.edu/diagnostic-plots/

Appendix

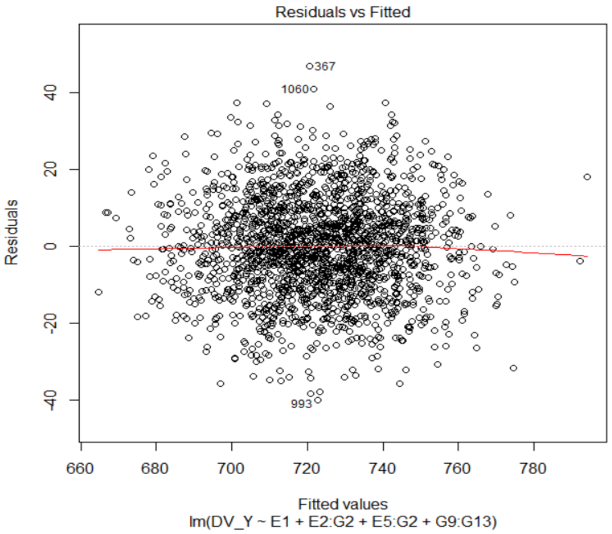
```
setwd("C:/Users/Stephen Hanna/Documents/Classes/Data analysis")
#Sets working directory to find the excel sheet with our values
MyData <-read.csv("Group_26.csv",header=T)
#Assigns excel sheet values to a data frame in R
MyData[1]<- NULL
#Removes first column since it is only composed of index values
shapiro.test(MyData[,1])
#Tests for normality
        Shapiro-Wilk normality test
data:  DV
W = 0.99844, p-value = 0.05962
#Since the p value is below .1, we reject the null hypothesis and conclude that the data is not normal
fit <- lm(DV_Y~(E1+E2+E3+E4+E5+G1+G2+G3+G4+G5+G6+G7+G8+G9+G10+G11+G12+G13+G14+G15)^3,data=MyData)
#Makes a linear model for the data, but includes all pairwise and triowise interactions.
bc <- boxcox(fit,lambda = seq(-6,6,0.1))
#Conducts BoxCox analysis to find lambda value for transformation
bc$x[which.max(bc$y)]
#Determines corresponding lambda value for the peak of the boxcox curve
[1] 3.1
MyData[,1] = ((MyData[,1]^.3.1)
#Conducts boxcox transformation on dependent variable values using lambda value
> shapiro.test(MyData[,1])
#Conducts another normality test for the dependent variable
```

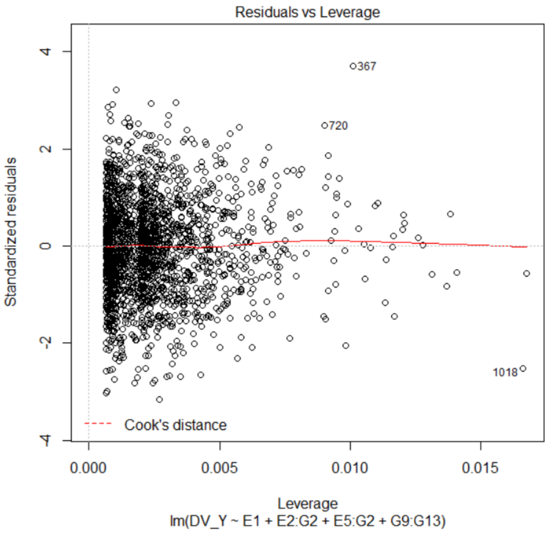
```
	Shapiro-Wilk normality test


data:  MyData[, 1]
W = 0.99893, p-value = 0.1089
#Since p is greater than .01 now, the dependent variable has been transformed into a normal distribution
x <- model.matrix(f,MyData[,-1])[,-1]
#Creates matrix with all independent variables and their associated values
y = as.matrix(MyData[,1])
#Creates vector of transformed dependent variable
cvfit = cv.glmnet(x, y)
#Creates a Lasso model
coef(cvfit, s = "lambda.1se")
#Shows coefficients for all variables, two way interactions, and three way interactions. All non-significant variables or interactions have a coefficient of zero.
fit <- lm(DV_Y~E1 + E1:G2 + E1:G11 + E2:G2 + E5:G2 + G9:G13 + E1:G8:G9 + E1:G8:G11 + E5:G2:G15,data=MyData)
#Creates linear model using variables or interactions with non zero coefficients from Lasso analysis

summary(fit)
#Creates table with relevant values from previous linear model

Call:
lm(formula = DV_Y ~ E1 + E1:G2 + E1:G11 + E2:G2 + E5:G2 + G9:G13 + 
    E1:G8:G9 + E1:G8:G11 + E5:G2:G15, data = MyData)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.89549 -0.18563  0.00331  0.18198  0.91483 
```

```
Coefficients:
            Estimate Std. Error         t value              Pr(>|t|)
(Intercept) 35.086362   0.007246 4842.240  < 2e-16 ***
E1          -0.328753   0.012261  -26.812  < 2e-16 ***
E1:G2       -0.025489   0.012778   -1.995  0.04621 *
E1:G11      -0.018457   0.015865   -1.163  0.24482
G2:E2        0.175919   0.008991   19.567  < 2e-16 ***
G2:E5        0.098591   0.013023    7.571 5.64e-14 ***
G9:G13       0.169871   0.014494   11.720  < 2e-16 ***
E1:G9:G8    -0.042587   0.016297   -2.613  0.00904 **
E1:G11:G8   -0.023761   0.019410   -1.224  0.22103
G2:E5:G15    0.056629   0.018070    3.134  0.00175 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2798 on 1990 degrees of freedom
Multiple R-squared:  0.6803,    Adjusted R-squared:  0.6788
F-statistic: 470.5 on 9 and 1990 DF,  p-value: < 2.2e-16
fit <- lm(DV_Y~E1+E2:G2+E5:G2+G9:G13,data=MyData)
#Creates linear model using only variables or interactions that had p values significantly below
.01 from last linear model
summary(fit)
#Creates table with relevant values from previous linear model
Call:
lm(formula = DV_Y ~ E1 + E2:G2 + E5:G2 + G9:G13, data = MyData)

Residuals:
    Min       1Q        Median      3Q          Max
```

```
-0.8942  -0.1870    0.0002    0.1823   1.0250


Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 35.084900   0.007268 4827.27   <2e-16 ***
E1          -0.366659   0.006404  -57.26   <2e-16 ***
E2:G2        0.176774   0.009040   19.55   <2e-16 ***
G2:E5        0.128098   0.009083   14.10   <2e-16 ***
G9:G13       0.172831   0.014569   11.86   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Residual standard error: 0.2816 on 1995 degrees of freedom
Multiple R-squared:  0.6755,    Adjusted R-squared:  0.6748
F-statistic:  1038 on 4 and 1995 DF,  p-value: < 2.2e-16
```

```
plot(fit)
#Plots several graphs related to the model. most important two are shown below.
```





```
Anova <- aov(DV_Y~E1+E2:G2+E5:G2+G9:G13,data=MyData)
#Assigns anova table of the model to a variable
```

```
summary(Anova)
```

```
#Creates summary of aforementioned anova analysis. Probability corresponding to f statistic is the most important.
```

```
            Df    Sum Sq    Mean Sq   F value    Pr(>F)
E1           1    268.46    268.46    3385.7    <2e-16 ***
E2:G2        1     33.22     33.22     418.9    <2e-16 ***
G2:E5        1     16.40     16.40     206.8    <2e-16 ***
G9:G13       1     11.16     11.16     140.7    <2e-16 ***
Residuals 1995    158.19      0.08
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```