**AMS 315 Data Analysis Project 2**

Young Seok Seo (109803869)

May 9th, 2019

***Introduction***

I received a data set which is contained one dependent variable, Y, and twenty four independent variables. The independent variable data sets are in the columns with the name of E1-E4, 4 environmental variables, and G1-G20, 20 gene variables. In this report, I will find a gene-environmental interaction through determining the correlation between the independent and dependent variables. Since there are no negative or zero and missing values, I will put the given dataset into R directly.

***Methodology***

Correlation :

Using the R, I found a correlation between the independent, environmental and dependent variables. I combined two different kinds of variable versus dependent variable. (see Table 1). As a result of Correlation Coefficient, I found that E1 with 0.27, E3 with 0.43, G7 with 0.30 and G17 with 0.34 are shown to be significant when compared to dependent variable, Y.

Box-Cox Transformation :

I can estimate by performing a function called boxcox as a parameter in a non-linear model. In my case, by using the R program, I got a nonlinear regression function of dependent variable (Y), Y^ with = 0.5. This result gives that the transformation for dependent variable is Y^0.5, and the boxcox table that I got from the R program supports the result.

Stepwise Regression :

After doing Box-Cox Transformation, Stepwise Regression method helps to find the important independent variables. I used the R code leaps and use regsubset() to get result.

As I put table called “model summary” I would like to choose 3rd model. There is a table below, and I compare the adjR^2 and Bayesian Information Criterion(BIC) between each equation candidates.

***Results***

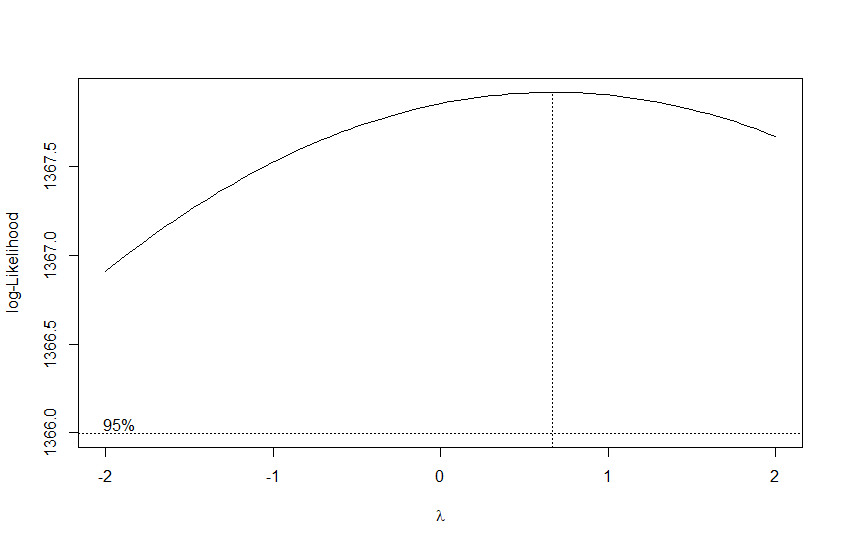


Figure: Box-Cox Analysis for Y

Adjusted R^2 using just the environmental variable and the original outcome was 0.2424. Also, at the Box-Cox Transformation method, the dependent variable (Y), Y^ with = 0.5. I set transform the model to Y^0.5 then adjR^2 value is 0.4750983, which value is quite adjoin to the

adjR^2 of genetic environmental variable, 0.4745053. Moreover, New Residual Plot become more accurate too.

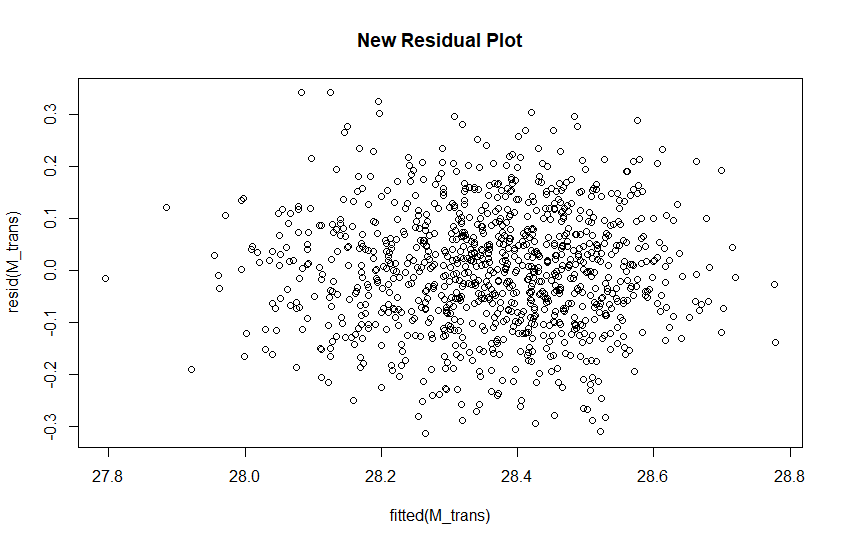
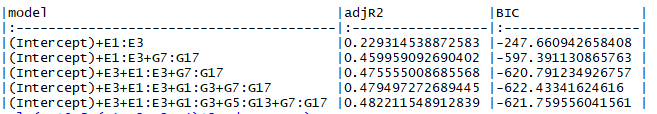


Table: Summary of Stepwise Selection



In the part of the Stepwise Regression, as shown in the table ,there is a obvious increase in adjR^2 from the 2nd to 3rd model. Meanwhile, difference adjR^2 between 3rd to 4th is smaller than that of 2nd to 3rd. Moreover, I uses to BIC values that consistently decrease. In BIC column, similar to adjR^2, difference of BIC value between 2nd to 3rd is larger than that of 3rd to 4th.

Consequently, I choose 3rd model to estimate the equation for my analysis. (Intercept) + E3 + E1:E3 + G7:G17 is used for my estimated equation. I consider the second order interaction by using the second power in the model request. In this regard, I obtained the following model result : Y^0.5 = 9.6638909 + 0.0183862E3 + 0.0001234E1E3 + 0.0008467G7G17

***Conclusion***

Based on all the results from the R program and my analysis, I can conclude that my model function is Y^0.5 = 9.6638909 + 0.0183862E3 + 0.0001234E1E3 + 0.0008467G7G17

with a model adjusted R-square value of 0.4750. This value shows that my model would be reasonable and significant.

***Appendix***

Table 1: correlation between the independent, environmental and dependent variables

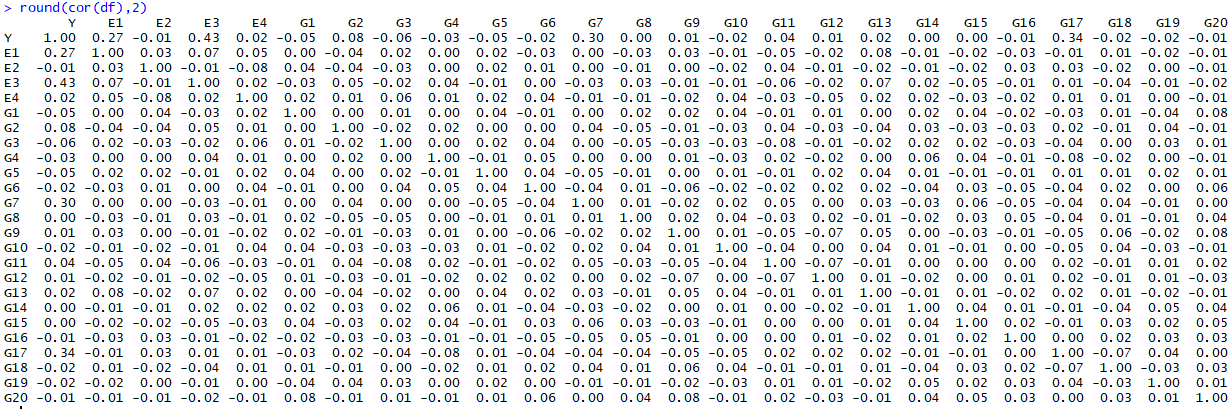


Figure 1: Residual by Regressors for Y

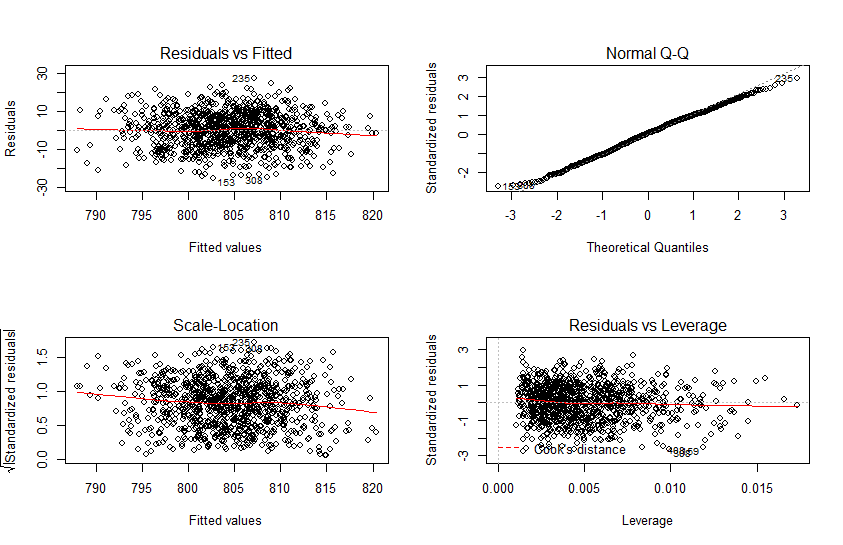
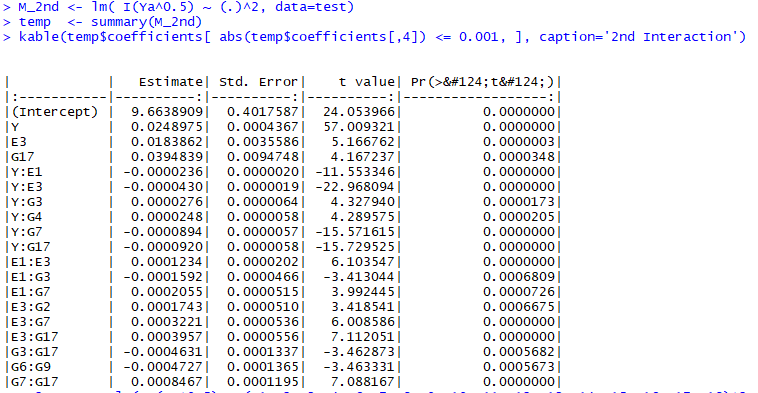


Figure 2: Adjusted R Square Value



Figure 3: estimated value for each coefficient



test <- read.csv('C:/Users/ryans/Desktop/AMS/Project2/P203869.csv', header=TRUE)

View(test)

M\_E <- lm(Y ~ E1+E2+E3+E4, data=test)

summary(M\_E)

ifelse(round(cor(df),2)>.3,round(cor(df),2),'') ## E3~G17 only have correlation with Y

round(cor(df),2)

M\_raw2 <- 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=test )

plot(resid(M\_raw2) ~ fitted(M\_raw2), main='Residual Plot')

library(MASS)

boxcox(M\_raw2)

M\_trans <- lm( I(Y^0.5) ~ (.)^2, data=test )

summary(M\_raw2)$adj.r.square;

summary(M\_trans)$adj.r.square

plot(resid(M\_trans) ~ fitted(M\_trans), main='New Residual Plot')

library(leaps)

Ya=test$Y

M <- regsubsets( model.matrix(M\_trans)[,-1], I(Ya^0.5), 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='+'))

library(MASS)

kable(data.frame(cbind( model = M\_select, adjR2 = temp$adjr2, BIC = temp$bic)),

caption='Model Summary')

lm(Ya^0.5~(E1+E2+E3+E4)^2, data=test)

M\_main <- lm( I(Ya^0.5) ~ (.), data=test)

# . here means include all variable from E1 to E5 and from G1 to G15 to the model

temp <- summary(M\_main)

kable(temp$coefficients[ abs(temp$coefficients[,4]) <= 0.001, ], caption='Sig Coefficients')

M\_2nd <- lm( I(Ya^0.5) ~ (.)^2, data=test)

temp <- summary(M\_2nd)

kable(temp$coefficients[ abs(temp$coefficients[,4]) <= 0.001, ], caption='2nd Interaction')

M\_2stage <- lm( I(Ya^0.5) ~ (E1+E3+G7+G17)^2, data=test)

temp <- summary(M\_2stage)

temp$coefficients[ abs(temp$coefficients[,3]) >= 4, ]