Regression Analysis of Gene-Environment Interactions

Thomas Connolly
AMS 578

Abstract

The goal of this paper is to report my work on the assignment of using regression techniques to find the best fitting model to predict my outcome variable Y given my 31 regressor variables. This report includes an introduction to the problem at hand, a methods section consisting of the steps I took and techniques I used in order to find my final model, a results section consisting of the statement of my final model as well as well as the summary results of my model, and a conclusion section which sums up the project as well as provides final thoughts on the experience.

1. Introduction

The essence of this project surrounds a paper, *Influence of life stress on depression: moderation by a polymorphism in the 5-HTT gene* (Caspi et al. 2003) ,in which the authors use regression techniques to model depression based on certain factors. The paper discusses the development of depression and the role a person's genetic makeup along with the stressful situations a person has endured throughout their life as possible factors. It focuses on how a person deals with stressful situations in life and relates this to one's genetic makeup to test the relationship between these two things and the onset of depression. The authors of this paper found interactions between gene and environmental variables in their regression analysis. These gene-environment interactions have been scrutinized by other researchers and considered an error made by Caspi et al. In the paper *Interaction Between the Serotonin Transporter Gene(5-HTTLPR), Stressful Life Events, and Risk of Depression: A Meta-analysis* (Risch et al. 2009) the researchers deny this claim of gene-environment interactions and conclude that there is only correlation between the stressful events in one's life and their development of depression.

For my assignment I was given three data files. One consisting of my dependent continuous variable, labeled Y, one consisting of my six continuous environmental variables, labeled E1 through E6, and one consisting of my twenty-five binary genetic variables, labeled R1 through R25. Each data set contained missing values that needed to be filled which will be discussed later in the methods section. After merging the three data sets by the patient identifier number, the

number of observations should have been 1,617 per variable. Of the thirty-two total variables seventeen were missing observations. They were either missing twenty or thirty observations. In other words, variables were missing either 0%, 1.2%, or 1.9% of their data. Thirteen variables were missing 30 observations, four variables were missing 20 observations, and fifteen variables were missing no observations. In total my merged data set was missing 470 observations out of 51,744 total observations, about 0.9%.

2. Methods

2.1 Merging Data and Filling in Missing Values

My first step in analyzing my data sets was combining my three data sets into one for easier manipulation. I used the *merge()* function in R to perform this combining of data sets, and since each respective patient had a patient identifier, I used this ID number as the common value to merge and sort the combined data by. Since there are missing values in the data, I had to choose a method for dealing with these NA values. As an alternative to eliminating patients containing missing values, i.e., row wise deletion, I decided to use an R package called Amelia to perform imputation on the data set. The Amelia package performs multiple imputation using an EMB algorithm. This means it creates replicas of the inputted data set and then fills the missing values using a bootstrap based Expectation-Maximization algorithm. The Expectation- Maximization (EM) algorithm is a method for finding the maximum- likelihood estimates for a given data set when there are missing values. It involves generating expected values for missing data. Then based off of these values, it estimates the value which maximizes the likelihood function. This process is repeated until the sequence of new values converges (Honaker, King, and Blackwell 2019). This EM algorithm combined with a bootstrap approach, which is a method for resampling, is what makes up the EBM algorithm which the Amelia package uses to fill in missing data.

2.2 Goodness of Fit Criteria

Throughout my model building process, I used many criteria to compare models as well as assess how well each model fits my data set. Some of these important values I focused on were R-squared, adjusted R-squared, AIC, and BIC. In this section I will briefly introduce these concepts.

2.2.1. R-squared and Adjusted R-squared

R- squared and adjusted R-squared are goodness of fit criteria that represent the percentage of residuals of the data which are accounted for by a given model. R-squared values range from 0 to 1 and the larger the R-squared, the better the model accounts for the variation of the data thus making it a better fit. The formula for R-squared is:

$$R - squared = 1 - \frac{RSS}{TSS}$$

Where RSS is the residual sum of squares and TSS is the total sum of squares. The formula for adjusted R-squared is:

$$Adjusted\ R - squared = 1 - \frac{(1 - R^2)(n - 1)}{(n - k - 1)}$$

Where k is the number of variables in the model, n is the number of observations and $R^2$is the R-squared value. Both criteria represent how well a model fits a dataset but adjusted R-squared accounts for the number of parameters in your model, so it was more useful in my analysis.

2.2.2. BIC and AIC

The Akaike information criterion(AIC) and the Bayesian information criterion(BIC) are goodness of fit criteria for model fitting based on the likelihood function. The rule of thumb for both being that the lower the BIC or AIC, the better the model. Both the AIC and BIC take into account the number of variables to prevent overfitting, so they were very useful in model comparison throughout my analysis. The formulas for each are as follows:

$$BIC = kln(n) - 2\ln(\hat{L}) \qquad AIC = 2k - 2\ln(\hat{L})$$

Where $\hat{L}$is the maximum value of the likelihood function, k is the number of parameters, and n is the number of observations.

2.3 Model Building

Once the data was merged and missing values were dealt with, I began my process of building the model which best fits my data. I began my work by checking the correlations between my dependent variable Y and the other 31 independent variables. Strong correlations, in my case

correlations greater than 0.1, represent a strong relationship between the two variables and such variables should be heavily considered in future analysis. I found that only environmental variables had correlations larger than 0.1, specifically variables E1, E2, E3, and E4. As a result, I had an early idea that my model may only be a combination of environmental variables. The resulting correlations between Y and each independent variable are given in the table below.

Correlations With Y

E1: .1098	E2: .2512	E3: .3085	E4: .4322	E5: .0120	E6: .0258	R1: .0287	R2: .0049
R3: .0285	R4: .0013	R5: .0074	R6: .0014	R7: .0090	R8: .0307	R9: .0318	R10: .0021
R11: .0057	R12: .0397	R13: .0080	R14: .0296	R15: .0351	R16: .0028	R17: .0142	R18: .0057
R19: .0171	R20: .0498	R21: .0111	R22: .0008	R23: .0230	R24: .0182	R25: .0390	Y: 1.0

I also wanted to consider the possibility of a transformation of my data, so I performed a Box-Cox transformation using the *boxcox()* function in R. This returned a lambda value of 1 which told me that it is likely that no transformation is necessary. I then began the process of fitting models using my data. I started with the model consisting of only environmental variables as regressors, *M_E <- lm(Y~ E1+E2+E3+E4+E5+E6, data=data).* I then used the *summary()* function in R on this model and found that this model had an R-squared value of .3852, an adjusted R-squared value of .3829, BIC value of 77516.63, and an AIC value of 77473.52. This summary function also told me, based on the p-values of each variable, that only E1, E2,E3, and E4 were significant in the model. This observation backed my initial analysis that perhaps only E1 through E4 are necessary in my final model. My next step was to repeat this process with the full linear model, *M1 <- lm(Y~(.), data = data).* I then got the summary statistics for this model. It had an R-squared of .3938, an adjusted R-squared of .3819, a BIC value of 77678.55, and an AIC of 77500.73. Also, in this model only E1 through E4 were highly significant based on p-value (p-value<=.01). The output of the summary function for the full linear model is given below.

```
Call:
lm(formula = Y ~ (.), data = data)

Residuals:
       Min         1Q     Median         3Q        Max
-1.981e+10 -3.897e+09 -2.993e+07  3.960e+09  1.744e+10

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.188e+10  1.822e+09  -6.517 9.60e-11 ***
E1           4.826e+06  9.448e+05   5.109 3.64e-07 ***
E2           1.331e+07  9.350e+05  14.232  < 2e-16 ***
E3           1.657e+07  9.843e+05  16.836  < 2e-16 ***
E4           2.330e+07  9.886e+05  23.570  < 2e-16 ***
E5          -3.717e+05  9.565e+05  -0.389   0.6976
E6          -8.404e+05  9.726e+05  -0.864   0.3876
R1          -4.124e+08  3.061e+08  -1.347   0.1782
R2           2.023e+08  3.068e+08   0.659   0.5098
R3          -1.628e+08  3.074e+08  -0.530   0.5964
R4          -3.568e+08  3.076e+08  -1.160   0.2463
R5          -3.677e+08  3.082e+08  -1.193   0.2330
R6           3.124e+08  3.073e+08   1.017   0.3095
R7          -1.257e+08  3.080e+08  -0.408   0.6834
R8           3.467e+08  3.073e+08   1.128   0.2595
R9          -1.615e+08  3.081e+08  -0.524   0.6002
R10          1.266e+07  3.077e+08   0.041   0.9672
R11         -5.486e+07  3.073e+08  -0.179   0.8583
R12         -2.649e+07  3.081e+08  -0.086   0.9315
R13          8.830e+07  3.061e+08   0.288   0.7730
R14          2.403e+08  3.078e+08   0.781   0.4352
R15          3.734e+08  3.075e+08   1.214   0.2248
R16         -4.821e+08  3.079e+08  -1.566   0.1176
R17          2.003e+07  3.058e+08   0.066   0.9478
R18          1.660e+08  3.089e+08   0.537   0.5911
R19          5.024e+08  3.077e+08   1.633   0.1027
R20          6.600e+08  3.065e+08   2.153   0.0315 *
R21          3.754e+06  3.066e+08   0.012   0.9902
R22         -6.684e+07  3.080e+08  -0.217   0.8282
R23         -1.682e+08  3.069e+08  -0.548   0.5837
R24          3.449e+08  3.064e+08   1.126   0.2604
R25         -2.167e+08  3.070e+08  -0.706   0.4803
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.121e+09 on 1585 degrees of freedom
Multiple R-squared:  0.3938,    Adjusted R-squared:  0.3819
F-statistic: 33.21 on 31 and 1585 DF,  p-value: < 2.2e-16
```

Comparing these results to the environmental only model, the R- squared decreased which in general would mean that this model is a better fit, but I attribute this to the increased number of variables. This observation is reflected in the adjusted R-squared which decreased, as well as in the BIC and AIC values which increased leading me to believe that this is not a better model. These observations also support my initial claim that perhaps my final model will only include environmental variables since including the genetic variables did not seem to improve the accuracy of the model. I also repeated this process with the model consisting of all terms including all possible two-way interactions. This model also left me with similar conclusions. The R-squared values only slightly increased, likely due to the large increase in the number of terms, and BIC and AIC values increased, suggesting a worse model fit. This suggested to me there might not be any two-way interactions in my final model. My next step would be to investigate the three-way and then four-way interactions, looking for the most significant variables. This was unfortunately not possible because of the large number of possible interaction terms for all variables. Because of this I was forced to consider a smaller subset of variables that I knew were strongly associated with the dependent variable Y.

At this point in my regression analysis, I was fairly certain that my final model would only utilize environmental variables, but I decided to now try backwards stepwise regression on the full linear model to see which variables this method deemed important. I used the R function *step(),*given by the following code: *step(M1, direction='backward', scope=formula(MI), trace=0).* In this code segment M1 is my full linear model and MI is my intercept only model. This function left me with the variables E1,E2,E3,E4,R16,R19,and R20. I figured that these genetic terms were most likely false positive, so I ran the model consisting of these six variables and found that only the environmental variables are significant once again. At this point I made the decision to only consider the environmental variables in the rest of my analysis, more specifically E1, E2, E3, and E4. These four variables are the only ones which have been consistently significant throughout all my attempts. With a subset of variables as small as this I was finally able to consider up to four-way interactions. So, I ran the model of all possible four-way interactions between these terms. I then used the *regsubets()* in R on the model consisting of all of the four-way interactions between E1 through E4. This function uses stepwise regression to select the best model based on the number of variables requested. I then compared the adjusted

R-squared and BIC of the best models consisting of eight or less variables. Upon comparing the results, the three-way interaction between E2,E3, and E4 was always the most significant term and the addition of the E1 term increased the R-squared while continuing to decrease the BIC value.

3. Results

The final model for my data is given by the equation below.

$$Y = \beta_0 + \beta_1 * E_1 + \beta_2 * E_2 * E_3 * E_4$$

Where $\beta_0 = 1.122e + 10$, $\beta_1 = 4.79e + 6$, and $\beta_2 = 3.751e + 1$. My final model has an adjusted R-squared value of .394, a BIC of 77461, and an AIC of 77440. The summary results of my final model are given below.

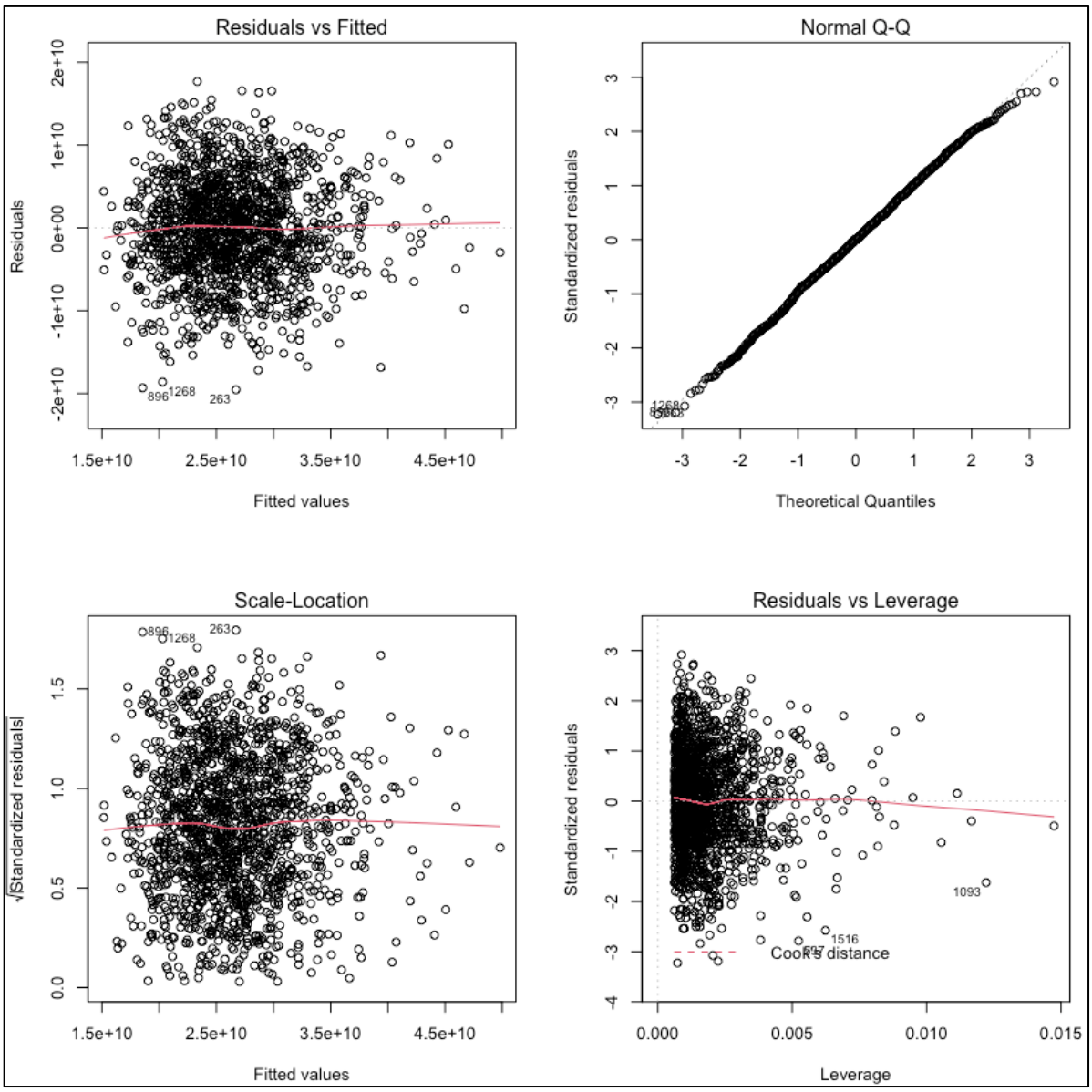
```
Call:
lm(formula = Y ~ E1 + E2:E3:E4, data = data)

Residuals:
       Min          1Q      Median          3Q         Max
-1.954e+10 -3.912e+09  3.282e+07  4.188e+09  1.768e+10

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.122e+10  7.946e+08  14.115  < 2e-16 ***
E1          4.790e+06  9.270e+05   5.168 2.67e-07 ***
E2:E3:E4    3.751e+01  1.174e+00  31.942  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.061e+09 on 1614 degrees of freedom
Multiple R-squared:  0.3947,    Adjusted R-squared:  0.394
F-statistic: 526.2 on 2 and 1614 DF,  p-value: < 2.2e-16
```

I then also looked at the diagnostic plots for my final model. The residual vs fitted plot shows no pattern or nonlinear relationship which is a good quality. The Normal Q-Q plot shows an approximately forty-five-degree line which indicates that the residuals are normally distributed, once again a good quality. The scale-location plot shows a random spread among the residuals, showing equal variance. The residuals vs leverage plot looks normal, with no overly influential points.



I also checked for multicollinearity among the variables within my final model using the *VIF()* function in R and both variables had VIF values of approximately 1 which shows there is no multicollinearity present. The analysis of variance, or ANOVA, table for my final model is also

given below.

```
Analysis of Variance Table

Response: Y
            Df     Sum Sq    Mean Sq  F value    Pr(>F)
E1           1 1.1815e+21 1.1815e+21   32.159  1.68e-08 ***
E2:E3:E4     1 3.7485e+22 3.7485e+22 1020.305 < 2.2e-16 ***
Residuals 1614 5.9297e+22 3.6739e+19
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

4. Discussion and Conclusion

The results of my regression analysis showed that only environmental factors are significant when it comes to modeling my dependent variable Y. This result denies the claims by Caspi et al. of gene-environment interactions and supports the claims of Risch et al. Choosing the best model for my data required the use of many regression techniques, such as forward and backward stepwise regression, and the analysis of many different goodness of fit criteria, such as R-squared, adjusted R-squared, BIC, AIC, and p-values. This project required me to utilize many of the tools I have learned throughout this course and many of my other statistics courses.

References

Caspi, A. (2003). Influence of Life Stress on Depression: Moderation by a Polymorphism in the 5-HTT Gene. *Science*, *301*(5631), 386–389. https://doi.org/10.1126/science.1083968

Honaker, J., King, G., & Blackwell, M. (2019, November 24). *Amelia II: A Program for Missing Data*. https://cran.r-project.org/web/packages/Amelia/vignettes/amelia.pdf.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2013). *Introduction to linear regression analysis* (5th ed.). Wiley-Blackwell.

Risch, N., Herrell, R., Lehner, T., Liang, K.-Y., Eaves, L., Hoh, J., … Merikangas, K. R. (2009). Interaction Between the Serotonin Transporter Gene (5-HTTLPR), Stressful Life Events, and Risk of Depression. *JAMA*, *301*(23), 2462. https://doi.org/10.1001/jama.2009.878

Appendix

R Code

```
### Code for AMS578 Project ###
library(corrplot)
library(leaps)
library(knitr)
library(regclass)
library(Amelia)
library(MASS)

# Reading in the 3 data files.
print(getwd())
setwd("/Users/thomasconnolly/Downloads")
print(getwd())
IDE <- read.csv(("IDEgroup477375.csv"), header =  TRUE)[,-1]
IDG <- read.csv(("IDGgroup477375.csv"), header =  TRUE)[,-1]
IDY <- read.csv(("IDYgroup477375.csv"), header =  TRUE)[,-1]

# Merging the 3 data variables by ID number.
data1 <- merge(IDE, IDG, by="ID")
data2 <- merge(data1, IDY, by="ID")
data2

# Computing Summary Statistics for the merged data, before imputation.
summary_before <-summary(data2[,-1])
summary_before

apply(data2[,-1], MARGIN = 2,FUN =  sd, na.rm = TRUE)
apply(data2, MARGIN = 2, FUN = length)
cor <- cor(data2[,-1])
corr_NA <- cor(na.omit(data2[,-1]))
corr_NA2 <- cor(data2[,-1],use = "complete.obs")

# Fixing data by filling in missing data using the Amelia package.
set.seed(123)
a.out <-
amelia(data2,noms=c('R1','R2','R3','R4','R5','R6','R7','R8','R9','R10','R11','R12','R13','R14','R15','
R16','R17','R18','R19','R20','R21','R22','R23','R24','R25'),idvars = 'ID',m=1)
summary(a.out)
a.out$imputations$imp1
data <- a.out$imputations$imp1[,-1]

#Box-Cox transformation
(boxcox(lm(Y-min(Y)+1 ~ ., data=data))) #lambda=1 so most likely no transformation needed
```

```
# Correlations
abs(cor(data$E1, data$Y))  #.1098 ####
abs(cor(data$E2, data$Y))  #.2512 ####
abs(cor(data$E3, data$Y))  #.3085 ####
abs(cor(data$E4, data$Y))  #.4322 ####
abs(cor(data$E5, data$Y))  #.0120
abs(cor(data$E6, data$Y))  #.0258
abs(cor(data$R1, data$Y))  #.0287
abs(cor(data$R2, data$Y))  #.0049
abs(cor(data$R3, data$Y))  #.0285
abs(cor(data$R4, data$Y))  #.0013
abs(cor(data$R5, data$Y))  #.0074
abs(cor(data$R6, data$Y))  #.0014
abs(cor(data$R7, data$Y))  #.0090
abs(cor(data$R8, data$Y))  #.0307
abs(cor(data$R9, data$Y))  #.0318
abs(cor(data$R10, data$Y)) #.0021
abs(cor(data$R11, data$Y)) #.0057
abs(cor(data$R12, data$Y)) #.0397
abs(cor(data$R13, data$Y)) #.0080
abs(cor(data$R14, data$Y)) #.0296
abs(cor(data$R15, data$Y)) #.0351
abs(cor(data$R16, data$Y)) #.0028
abs(cor(data$R17, data$Y)) #.0142
abs(cor(data$R18, data$Y)) #.0057
abs(cor(data$R19, data$Y)) #.0171
abs(cor(data$R20, data$Y)) #.0498
abs(cor(data$R21, data$Y)) #.0111
abs(cor(data$R22, data$Y)) #.0008
abs(cor(data$R23, data$Y)) #.0230
abs(cor(data$R24, data$Y)) #.0182
abs(cor(data$R25, data$Y)) #.0390
abs(cor(data$Y, data$Y))   #1.0
corrplot(cor(data))

### Model Fitting
# Model with only intercept
MI <- lm(Y~ 1, data=data)
summary(MI)

# Environment Only Model
M_E <- lm(Y~ E1+E2+E3+E4+E5+E6, data=data)
summary(M_E)
summary(M_E)$adj.r.squared #.383
BIC(M_E) #77516
```

```
AIC(M_E)#77473

# Full Linear Model
M1 <- lm(Y~(.), data = data)
summary(M1)
summary(M1)$coefficients[,4][summary(M1)$coefficients[,4]<0.01]
summary(M1)$adj.r.squared #.382
BIC(M1) #77678
AIC(M1) #77500
# The addition of R terms don't seem to help my model

# Full Model, Including All Possible Two-way Iteractions
M2 <- lm(Y~(.)^2, data = data)
summary(M2)
summary(M2)$coefficients[,4][summary(M2)$coefficients[,4]<0.001]
summary(M2)$adj.r.squared #.410
BIC(M2) #80476
AIC(M2) #77793

# Backwards Stepwise Regression On Full Linear Model
step(M1, direction='backward', scope=formula(MI), trace=0) #Outputs the following variables:
E1,E2,E3,E4,R16,R19,R20

S1 <- lm(Y~(E1+E2+E3+E4+R16+R19+R20), data=data)
summary(S1)
summary(S1)$adj.r.squared #.386
summary(S1)$coefficients[,4][summary(S1)$coefficients[,4]<0.001]
BIC(S1) #77514
AIC(S1) #77466

# Only E1,E2,E3,E4 seem to be significant
# Model of all four way interactions between E1,E2,E3,E4
S2 <-lm(Y~(E1+E2+E3+E4)^4+poly(E1,4)+poly(E2,4)+poly(E3,4)+poly(E4,4), data=data)
summary(S2)
summary(S2)$adj.r.squared #.400
summary(S2)$coefficients[,4][summary(S2)$coefficients[,4]<0.001]

# Function below is used for creating best subset of variables using forward stepwise regression.
# I ran this on the model consisting of all interactions up to four-way between E1,E2,E3,E4 and
# choose the best model by BIC and adjusted R-squared to be Y = E1 + E2:E3:E4
var <- colnames(model.matrix(S2))
sets <- regsubsets(model.matrix(S2)[,-
1],data$Y,nbest=1,nvmax=8,method='forward',intercept=TRUE)
best_subsets <- summary(sets)
subset_selection <-apply(best_subsets$which,1,function(x) paste0(var[x],collapse='+'))
```

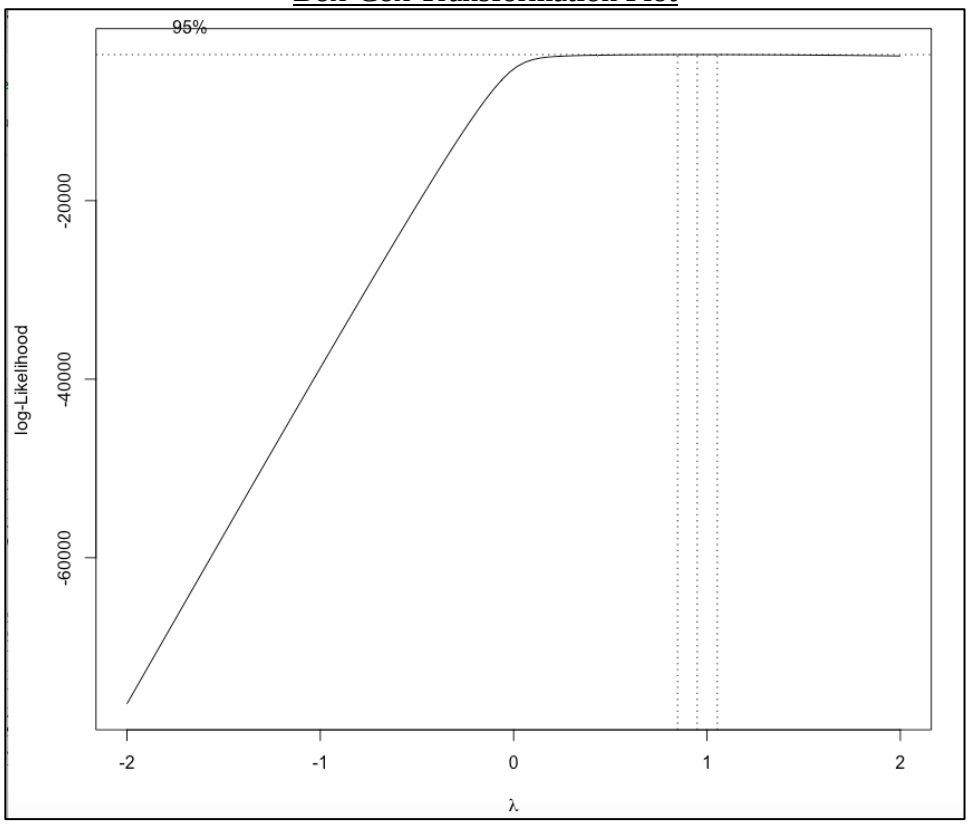
```
subsets <-kable(data.frame(cbind(model=subset_selection
,adjR2=best_subsets$adjr2,BIC=best_subsets$bic)),caption='Model Summary')
subsets
summary(regsubsets(Y~(E1+E2+E3+E4)^4+poly(E1,4)+poly(E2,4)+poly(E3,4)+poly(E4,4),
data=data, nvmax=5,method="forward"))

# Final Model and Results
final_model <- lm(Y~E1+E2:E3:E4, data = data)
summary(final_model)
BIC(final_model) #77461
AIC(final_model) #77440
summary(final_model)$adj.r.squared #.394
summary(final_model)$coefficients[,4][summary(final_model)$coefficients[,4]<0.001]

par(mfrow=c(2,2))
plot(final_model)
anova(final_model)
VIF(final_model)
```

Additional Images and Outputs

Box-Cox Transformation Plot



Correlation Plot (*corrplot()*)