ENVIRONMENT GENE INTERACTION

REGRESSION ANALYSIS

Emily L Thresher
EMILY.THRESHER@STONYBROOK.EDU

INTRODUCTION

With the advent of modern medicine, the use of statistical analyses has become essential. The complexities of disease become accessible. One such application is found when analyzing the relationships between genetic and environmental factors. Environmental factors can be anything from air pollution to apples; an agent from which a person can be exposed. Genetic factors are more difficult to detect but can also be quantified by assessing whether or not an individual possesses a certain biomarker. Almost every disease can be found to result from these two variables. In other words, an individual's response to environmental variables is affected by their genes.

One such study that examined this interaction analyzed these relationships concerning an individual's susceptibility to depression. In this study, entitled "Influence of Life Stress on Depression: Moderation by a Polymorphism in the 5-HTT Gene" and written by Caspi, et all, the researchers were particularly interested in why some people were more apt to exhibit depressive symptoms after being exposed to a stressful event. The genetic factor that was examined closely was the 5-HTT gene that is responsible for the transportation of serotonin, which is responsible for many effects throughout the body. The results of this study not only found a link between variations in the gene and exposure to trauma but also supports the notion of gene-environment interaction.

To simulate this real-life situation, a data set was produced that mimicked the variables of the aforementioned study. The data was produced from a predetermined equation and the objective of this analysis was to determine this equation. A number of methods were used in this process including simple random imputation, correlation tests, step-wise regression and transformations, all of which were performed in R. For the code utilized during this analysis, see the Appendix.1.

METHODOLOGY

Each set of variables, response, environmental and genetic markers, was provided in a separate frame, sorted by ID. Thus, the first step was to combine these data frames using the merge function. Once this was completed, each variable was analyzed to determine the number of missing values and to get a general idea of the distributions of each. The table shown below denotes a simplistic overview of the data.

Variable Category	Type	No. of Different Factors	Sample Size (per factor)	Missing Values (Y/N)
Response	Quantitative (Continuous)	1	1000	Yes
Environmental	Quantitative (Continuous)	6	1000	Yes
Genetic	Binary	25	1000	Yes

Additional measures were taken to assess the distributions of the environmental and response variables. To start, a Shapiro-Wilk Test was used to assess normality. The test, founded using Monte Carlo methods, was done on each of the continuous variable. The full results, i.e. the p-values, can be seen in Appendix.3. However, the conclusions from these iterations is that all of them are shown to be approximately normal at a significance level of $\alpha = .01$. Thus, it appears there may not be a cause to apply transformations to the environmental variables. To further assess the normality of the response variable, several plots were created to visualize the distribution. Based on the histogram and QQ-Plot, Appendix.4 and Appendix.5, respectively, it is reasonable to assume that the response variable is normally distributed.

As one can see from the table shown previously, every variable had missing values. Thus, it was essential to find a method to remedy this or methods of regression would be difficult to enact. Many methods can be used in this case. One such method is to simply eliminate any subject that contained missing values in any of the variables. This is not ideal, as it would have eliminated 463 observations, roughly 46% of the total observations. While, being left with 537 observations may appear to be enough to carry out any analyses, it can cause important relationships to be missed. Thus, a more sophisticated method needed to be chosen. For this particular situation, a simple random imputation method was employed. Simple random imputation involves creating a new vector based on the observations already recorded within the individual variable. This vector is then randomly inputted in for any missing values. This method provides a more accurate imputation than methods such as inputting the median or mean.

To begin assessing the relationships between the variables, correlation analyses were conducted. In order to ease the process in term of coding, the ID variable was dropped from the dataset going forward. The first correlation investigation was a Pearson correlation calculation that was used to create a correlation matrix, Appendix.6. From this matrix, it is clear that variables E3, E5 and E6

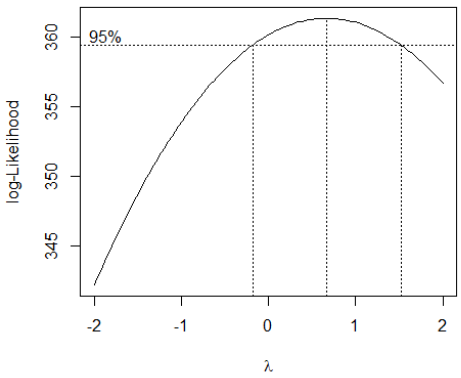
have the most significant correlation with the response. Indicated by the color, it should be noted that these variables have a direct or positive relationship.

After assessing the variables to determine if there are any obvious candidate variables, the next step was to begin looking potential regression equations. To start, a linear regression model using solely the environmental factors and their interactions. Based on the results, the only environmental variables that were significant at an $\alpha = .01$ level, were E3, E5, and E6 which were the same variables indicated in the correlation matrix. This pattern continued throughout the process of determining the final model. The fit for this model was moderate at an adjusted-$R^2 = .6543$. To assess any potential multicollinearity, the variance inflation factor, VIF, was calculated and was shown to be less than 10 which indicates there is no such issue.

The main process for determining the final model was to run a step-wise regression of this new transformation while assessing any 2nd level interactions. The choice to not search out higher level interaction stemmed from the fact that anything higher than 2nd level produced significant computational errors in R. For each adjustment of the model, new residual plots were created as well as determining the significant variables and noting changes in the adjusted-R^2 values. The results of these processes are outlined below.

RESULTS

Even though the dependent variable was shown to be approximately normal through the previous measures, a Box Cox transformation was run. First, a raw model was produced using the second interaction of all variables, both environmental and genetic, was produced. The plot to the right shows this transformation. Based on the log-Likelihood, it is obvious that the parameter λ is not maximized at either 1 or 0. Therefore, neither a log transformation nor leaving Y as is seemed to be the best option. Thus, a test was run to find the maximum. In Appendix.7 one can see that the likelihood is maximized at $\lambda = .6667$.



After determining the transformation appropriate for the response variable, the interactions were taken into consideration. As mentioned previously, only 2nd level interactions were assessed. To accurately assess the significant variables within each interaction level, the models were passed through two stages of analysis. First, the main effects were chosen based on a significance level of

α = .01. Once these were determined, the interaction effects were judged based on the same criterion. Then, each variable that was found to be significant at these two points were assessed at the same level. For the interaction terms, each variable within the term was added into the model separately to determine if there was still significance.

The final model determined by utilizing the aforementioned analyses is shown below. The ANOVA table for this model also follows.

$$Y^{2/3} = 21.35876 + .23575E3 + .36796E5 + .37582E6 + .10678G17{:}G19$$

```
Analysis of Variance Table

Response: Y^(2/3)
           Df  Sum Sq Mean Sq F value    Pr(>F)
E3          1  72.063  72.063 373.931 < 2.2e-16 ***
E5          1 142.825 142.825 741.109 < 2.2e-16 ***
E6          1 140.146 140.146 727.209 < 2.2e-16 ***
G17:G19     1   2.201   2.201  11.423 0.0007537 ***
Residuals 995 191.754   0.193
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Residual standard error: 0.439 on 995 degrees of freedom
Multiple R-squared:  0.6507,   Adjusted R-squared:  0.6493
F-statistic: 463.4 on 4 and 995 DF,  p-value: < 2.2e-16
```

As one can see, the coefficients utilized in this equation were shown to be significant at the α = .01 level. Note, that the adjusted-R^2 is still only moderately strong. Despite moderate strength, this model produced an almost pattern-less residual plot as shown in Appendix.8. There does appear to be one outlier, however, looking at the plots shown in Appendix.9 it does not appear that the outlier had effects on the Scale-Location or Normal QQ-Plot. For comparison purposes, other residual plots of previously found models can be seen in the Appendix as well. Looking at these residual plots, it does appear that the regression model containing only the environmental variables had the most pattern-less residual plot.

LIMITATIONS

The majority of the limitations that arose during this examination were coding related. Being able to manipulate the dataset to fit the requirements of some useful packages posed a great challenge. Lasso techniques that would have been helpful proved too advanced in their prerequisites. One particular package was designed to analyze high level interaction terms using the elastic net lasso approach. Being able to do this would have surely improved the final model. Additionally, the

original step-wise regression technique was simple to code but it the program ran for approximately three hours before completing and coming to a result that was not sufficient. In the future, either a different statistical software would be utilized or deeper knowledge would be acquired before undertaking a similar task.

Another limitation, that relates to the first, is the inability to look at higher level interaction terms. Attempting to run the code written with anything higher than 2nd level resulted in AIC levels of infinity or coefficient estimates of NaN. This is obviously not useful. Further investigation would have to be done to determine whether or not these errors were caused by coding errors or due to the data not having significant higher level interactions. If the data did have such interactions, then that is one reason why the adjusted-R^2 was not stronger.

CONCLUSION

Environmental-gene interactions play a crucial role in determining the onset of numerous diseases, as evidenced through extensive research. The results of the study conducted by Caspi et al. is just one of thousands of examples as to how our individual genetics, as well as our specific environment can come together to affect our well-being. Certain exposure to an outside agent combined with the smallest structural difference in one gene, can cause an individual to be greatly susceptible to disease. These conclusions are crucial in addressing societal issues that predispose individuals to disease and better understand the trends in mental and physical health that are prevalent across the globe.

References

1. https://www.r-bloggers.com/how-to-create-a-correlation-matrix-in-r/
2. https://www.niehs.nih.gov/health/topics/science/gene-env/index.cfm
3. https://www.itl.nist.gov/div898/handbook/prc/section2/prc213.htm
4. Influence of Life Stress on Depression: Moderation by a Polymorphism in the 5-HTT Gene BY AVSHALOM CASPI, KAREN SUGDEN, TERRIE E. MOFFITT, ALAN TAYLOR, IAN W. CRAIG, HONALEE HARRINGTON, JOSEPH MCCLAY, JONATHAN MILL, JUDY MARTIN, ANTONY BRAITHWAITE, RICHIE POULTON

APPENDIX

1.

```
E_Data <- read.csv(file = "C:/Users/emthr/Documents/Spring 2019/AMS 578/Group_6667_E.csv", header = TRUE)
G_Data <- read.csv(file = "C:/Users/emthr/Documents/Spring 2019/AMS 578/Group_6667_G.csv", header = TRUE)
Y_Data <- read.csv(file = "C:/Users/emthr/Documents/Spring 2019/AMS 578/Group_6667_Y.csv", header = TRUE)

my_temp_data <- merge(Y_Data,E_Data)
dataset <- merge(my_temp_data,G_Data)
completeData <- read.csv(file = "C:/Users/emthr/Documents/Spring 2019/AMS 578/Complete Data.csv", header = TRUE)

"dataset.mis <- prodNA(dataset, noNA = 0.1)
summary(dataset.mis)
dataset.mis <- subset(dataset.mis, select = -c(ID))
summary(dataset.mis)
imputed_Data <- mice(dataset.mis,m=5,maxit=50, method = 'pmm', seed = 500)
summary(imputed_Data)
completeData <- complete(imputed_Data,2)
describe(completeData)"
#write.csv(completeData, "Complete Data.csv")

attach(completeData)
shapiro.test(E1)$p.value
hist(E1)
shapiro.test(E2)$p.value
hist(E2)
shapiro.test(E3)$p.value
hist(E3)
shapiro.test(E4)$p.value
hist(E4)
shapiro.test(E5)$p.value
hist(E5)
shapiro.test(E6)$p.value
hist(E6)
shapiro.test(Y)$p.value
detach(completeData)
#Non significant at .01 sig level => normally distributed
```

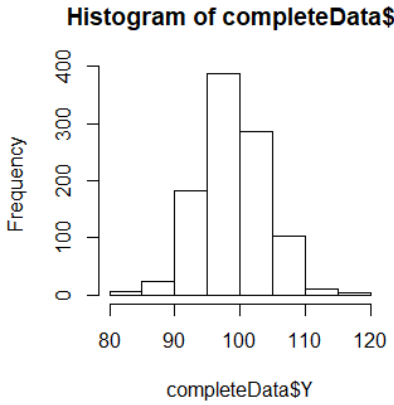
```
hist(completeData$Y)
qqnorm(completeData$Y)
qqline(completeData$Y, col = "red", lwd = 2)
car::qqPlot(completeData$Y)

mydata.cor <- cor(completeData)
corrplot(mydata.cor)
ModelE <- lm(Y^(2/3) ~ E1 * E2 * E3 * E4 * E5 * E6, data = completeData)
summary(ModelE)
VIF(ModelE) #less than 10 => no evidence of multicollinearity
car::ncvTest(ModelE)
plot(resid(ModelE) ~ fitted(ModelE), main = 'Residual Plot')
abline(0,0)

ModelComplete <- lm(Y ~ . , data=completeData)
summary(ModelComplete)
VIF(ModelComplete)
car::ncvTest(ModelComplete)
plot(resid(ModelComplete) ~ fitted(ModelComplete), main = 'Residual Plot')
abline(0,0)

ModelRaw <- lm(Y ~ (.)^2 , data = completeData)
summary(ModelRaw)
VIF(ModelRaw) #none
car::ncvTest(ModelRaw)
b <- boxcox(ModelRaw)
lambda <- b$x
lik <- b$y
bc <- cbind(lambda,lik)
sorted_bc <- bc[order(-lik),]
head(sorted_bc, n=10)
plot(resid(ModelRaw) ~ fitted(ModelRaw), main = 'Residual Plot')
abline(0,0)

ModelTransform <- lm((Y)^(2/3) ~ (.)^2 , data=completeData)
summary(ModelTransform)
VIF(ModelTransform) #none
```

```
car::ncvTest(ModelTransform)

b <- boxcox(ModelRaw)

lambda <- b$x

lik <- b$y

bc <- cbind(lambda,lik)

sorted_bc <- bc[order(-lik),]

head(sorted_bc, n=10)

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

abline(0,0)


Model <- regsubsets(Y^(2/3) ~ (.)^2 , data = completeData, nbest = 1, nvmax = 5, method = 'backward', intercept = TRUE)

temp <- summary(Model)

Var <- colnames(model.matrix(ModelTransform))

Model_Select <- apply(temp$which, 1,

            function(x) paste0(Var[x], collapse='+'))

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

    caption='Model Summary')


M_main <- lm( Y^(2/3) ~ . , data=completeData)

temp <- summary(M_main)

kable(temp$coefficients[ abs(temp$coefficients[,4]) <= 0.01, ], caption='Sig Coefficients')


M_2nd <- lm(Y^(2/3) ~ (.)^2 , data=completeData)

temp <- summary(M_2nd)

kable(temp$coefficients[ abs(temp$coefficients[,4])<= .01, ],caption = "2nd Interaction")


M_2stage <- lm(Y^(2/3) ~ (E3+E5+E6+G23+G12+E2+G22+G16+G4+G17+G25+G5+G9+G19+G10+G20+G14)^2,
data=completeData)

temp <- summary(M_2stage)

temp$coefficients[ abs(temp$coefficients[,3]) >= 2.5,]


FINALMODEL <- lm(Y^(2/3) ~ E3+E5+E6+G17:G19, data = completeData)

summary(FINALMODEL)

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

abline(0,0)

plot(FINALMODEL)

anova(FINALMODEL)
```

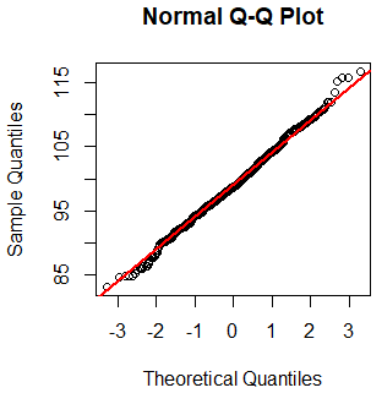
1.

2.
```
> shapiro.test(E1)$p.value

[1] 0.05112693
> shapiro.test(E2)$p.value
[1] 0.428478
> shapiro.test(E3)$p.value
[1] 0.4114093
> shapiro.test(E4)$p.value
[1] 0.3887547
> shapiro.test(E5)$p.value
[1] 0.103266
> shapiro.test(E6)$p.value
[1] 0.1297099
> shapiro.test(Y)$p.value
[1] 0.06536399
```

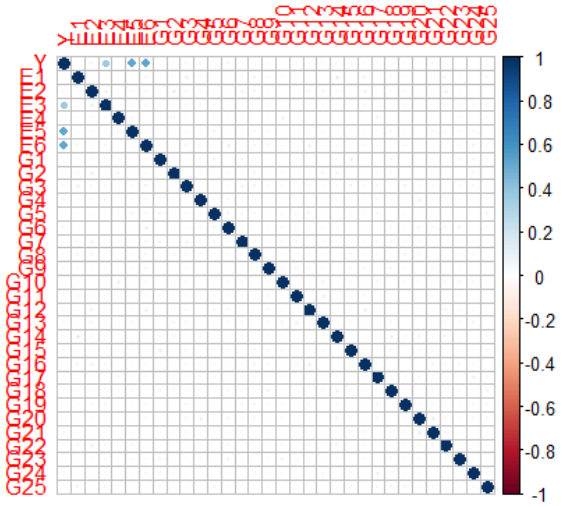
3.



4.

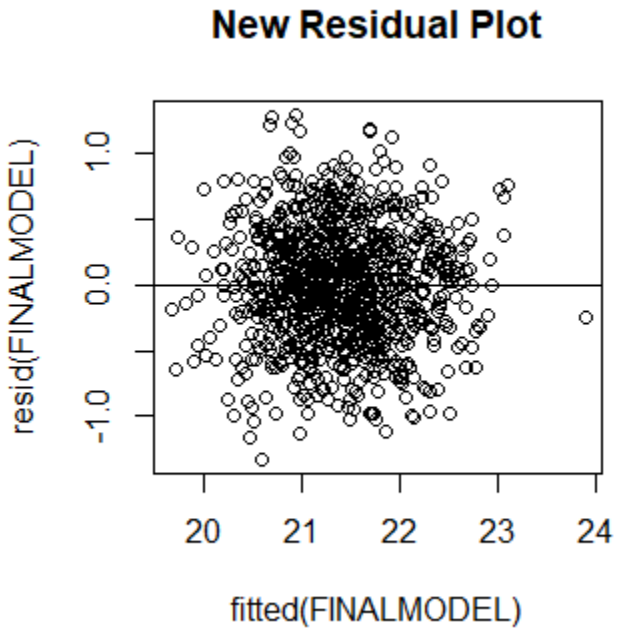


5.

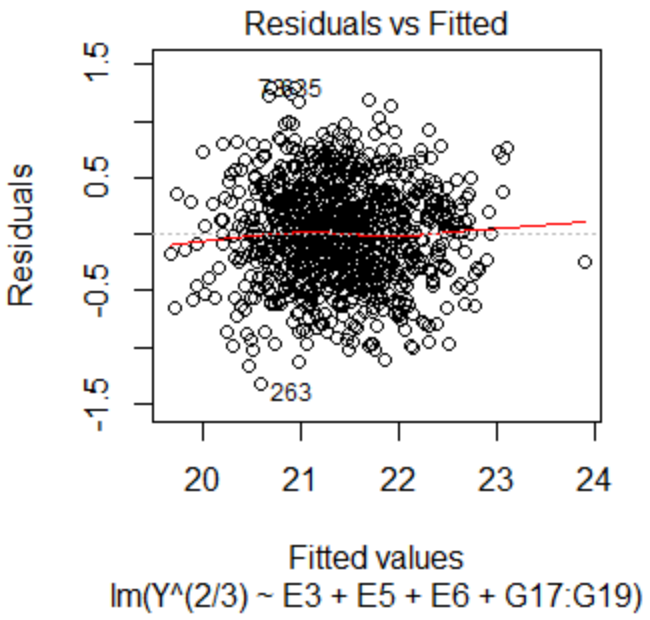


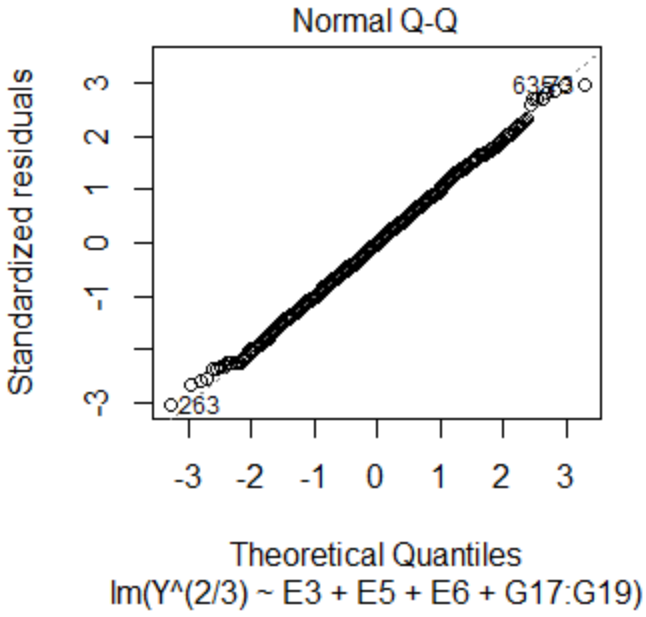
6.
```
          lambda      lik
 [1,] 0.6666667 361.3821
 [2,] 0.7070707 361.3779
 [3,] 0.6262626 361.3775
 [4,] 0.7474747 361.3651
 [5,] 0.5858586 361.3642
 [6,] 0.7878788 361.3436
 [7,] 0.5454545 361.3423
 [8,] 0.8282828 361.3134
 [9,] 0.5050505 361.3116
[10,] 0.8686869 361.2745
```

7.



8.


Residuals vs Fitted
Residuals
1.5
0.5
-0.5
-1.5
263
20
21
22
23
24
Fitted values
lm(Y^(2/3) ~ E3 + E5 + E6 + G17:G19)


Normal Q-Q
Standardized residuals
3
2
1
0
-1
-3
263
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(Y^(2/3) ~ E3 + E5 + E6 + G17:G19)