AMS 578
Regression Analysis
Spring 2022
Multiple Regression Computing Project

Brandon D’Anna 108470479

May 3, 2022

1. Introduction

Background

Regression theory plays an instrumental role in genetic research. With the advancement of technologies to better sequence genetic indicators, large datasets can be leveraged to analyze predictive power of numerous genetic markers on various outcomes from the presence of certain diseases to physical attributes of an individual. Furthermore, environmental conditions can influence the presence of certain genes.

Depression is one of the most common mental illnesses, causing intense feeling of sadness, loss of interest and sometimes thoughts of death or suicide. The motivation for this report stems from a study in which it was hypothesized that gene-environment interactions play a role in predicting depression in individuals (Caspi et al. 2003). The data for this report are based off of that used by Caspi. The findings were later proven to be a Type I error by Risch et al. (2009).

Data Summary

The data provided were synthetic, having sample size n = 2867 with one dependent variable Y, 8 environmental independent variables, E_i, $i = 1,..., 8$ and 30 gene independent variables, G_i, $i = 1,..., 30$. The environmental variables are continuous while the gene variables are indicator variables, taking on value one or zero, corresponding to whether the gene is present in the individual.

Hypotheses

If there exists genetic associations with the outcome variable after controlling for environmental association. Additionally, at least one of the environmental variables influences the value of the dependent variable. Lastly, there may exist interaction terms of two, three, or four factors between the environmental and genetic variables where the effect of the environmental variable depends on the presence of the genetic variable. As an increased number of interaction terms are considered, it is important to account for the potential increase in multicollinearity in the masking of potentially significant variables.

2. Methods

R statistical package was used to perform the data analysis on the synthetic data set. Various libraries were used including MASS, ggplot2, and olsrr. The data set was checked for missing data, with none found.

Firstly, the normality assumption of the dependent variable, Y was assessed. A Q-Q plot was created and suggested the data were skewed (Figure 2.1). As a result, a Box-Cox transformation of Y was implemented, resulting in an optimal lambda of 14/33. Since this value is fairly close to 0.5, lambda was assigned 0.5. This allows for ease of interpretation as this value equates to a transformation of $\sqrt{Y}$. The same normality assessment was performed on the transformed variable and the normality assumption was upheld (Figure 2.1).

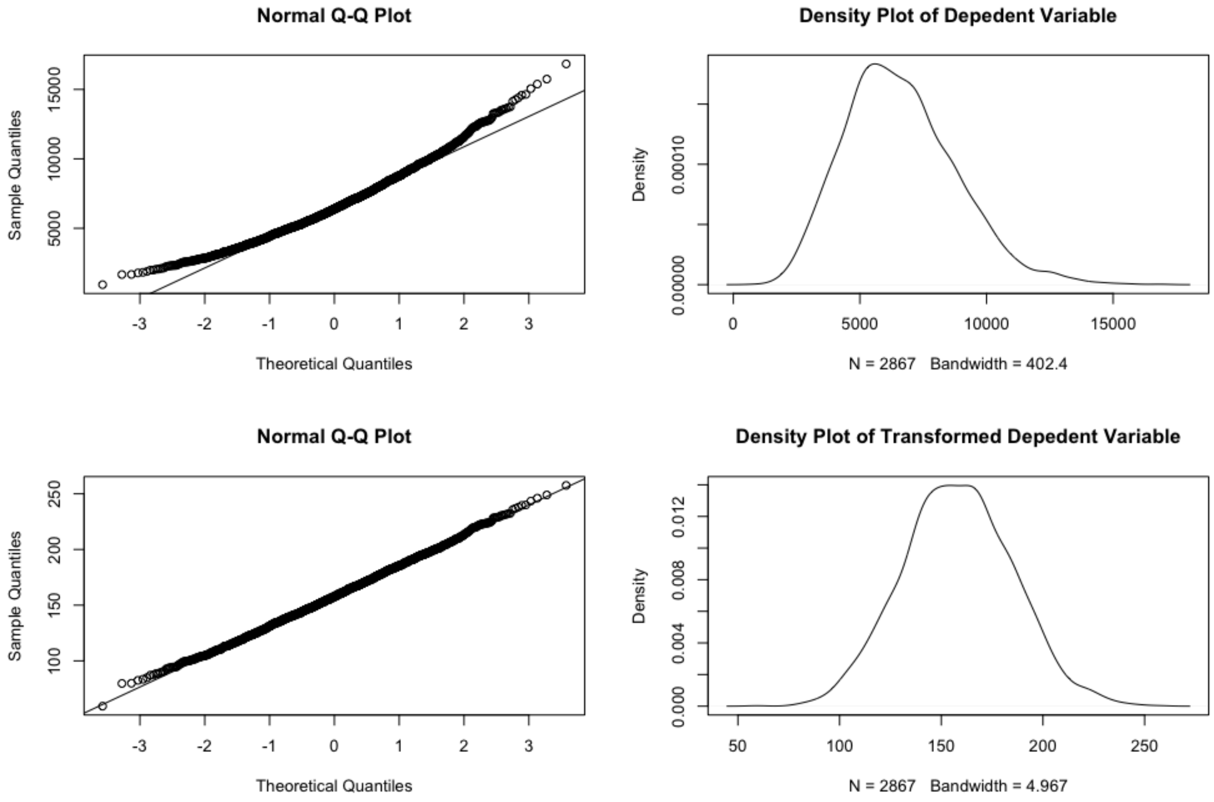


Figure 2.1 Q-Q plot (top left) and density plot (top right) of dependent variable Y

Q-Q plot (bottom left) and density plot (bottom right) of transformed dependent variable $\sqrt{Y}$

Bivariate association between Y and the independent variables was explored. Scatter plots were created to visualize any correlation between Y and the continuous environmental independent variables. Three of these plots, namely $E_1, E_3,$ *and* E_6 (Figure 2.2) suggested positive correlations. To confirm, tests for association using Pearson's product moment correlation coefficient were performed, $Corr(E_1,$ Y) = 0.399, $Corr(E_3,$ Y) = 0.323, and $Corr(E_6,$ Y) = 0.280, all with p values < $2.2 * 10^{-16}$.

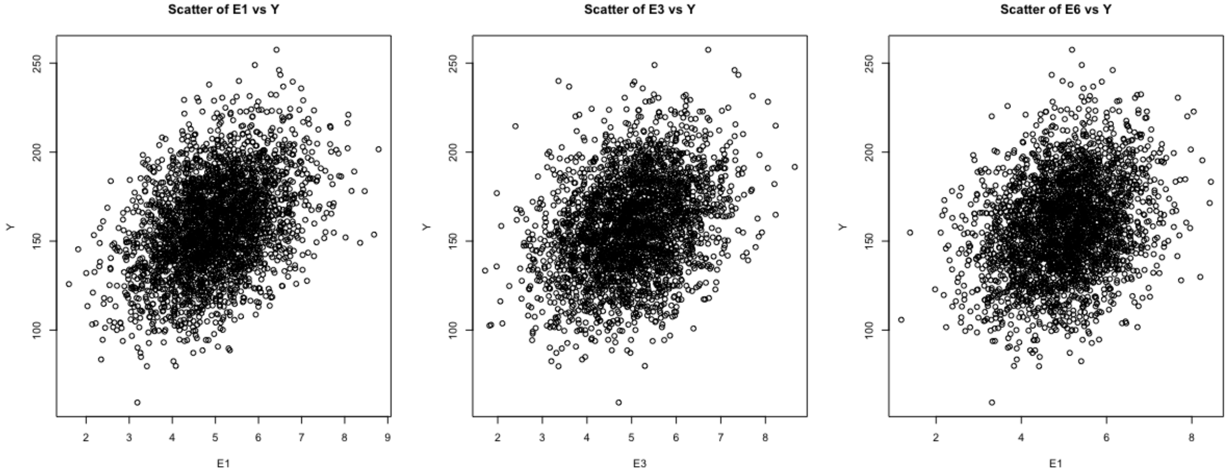


Figure 2.2 Scatter plots of Y against $E_1, E_3,$ *and* E_6 respectively

Each of the 30 genetic variables were subset into two groups, one where the gene was present, indicated by $G_i = 1, i = 1,..., 30$, and another where the gene was absent, indicated by $G_i = 0, i = 1,..., 30$. Two sample t-tests were performed on each group of subsets, testing the null hypothesis $\mu_1 = \mu_2$ against the two-sided alternative hypothesis. Of these 30 tests, the null hypothesis was rejected for three, namely, $G_2, G_{14},$ *and* G_{23} with p-values of $2.93 * 10^{-27}$, $6.11 * 10^{-17}$, and $7.79 * 10^{-29}$ respectively.

Multiple regression of Y on the 38 independent variables was performed, using an overall alpha of 0.01. Bonferroni's inequality multiples each of the p-values by 38. Those that are less than 0.01 can be deemed significant. The full model had three environmental variable coefficients and three gene variable coefficients with significantly large t values. These were the same six variables mentioned above in the bivariate association analysis.

Forward stepwise selection was implemented to search for two factor interaction terms. The starting model was the one which consisted only of the intercept. The results did not show any significant two factor interaction terms. The AIC was used to validate the selection process. The AIC for the intercept only model was 18,924 and leveled out at 17,085 with the addition of two factor interaction terms not adding significant reduction. Based on this result, coupled with being underpowered due to sample size, three factor and four factor interaction terms were not considered.

$$AIC = 2K - 2ln(L)$$

The PRESS statistic was calculated for the various models, with the value for the final model being the smallest. This model yielded a PRESS statistic of 1,110,574 compared to the full model with a PRESS statistic of 1,126,233.

$$PRESS = \sum_{i=1}^{n} (\frac{r_i}{1-H_{ii}})^2$$

Residual analysis was performed to ensure the model assumptions were not violated. This analysis was performed using similar techniques to that for the dependent variable. Namely, a Q-Q plot and a density plot (Figure 2.3), both appearing normal. A scatter plot of residual values against fitted values appeared patternless.

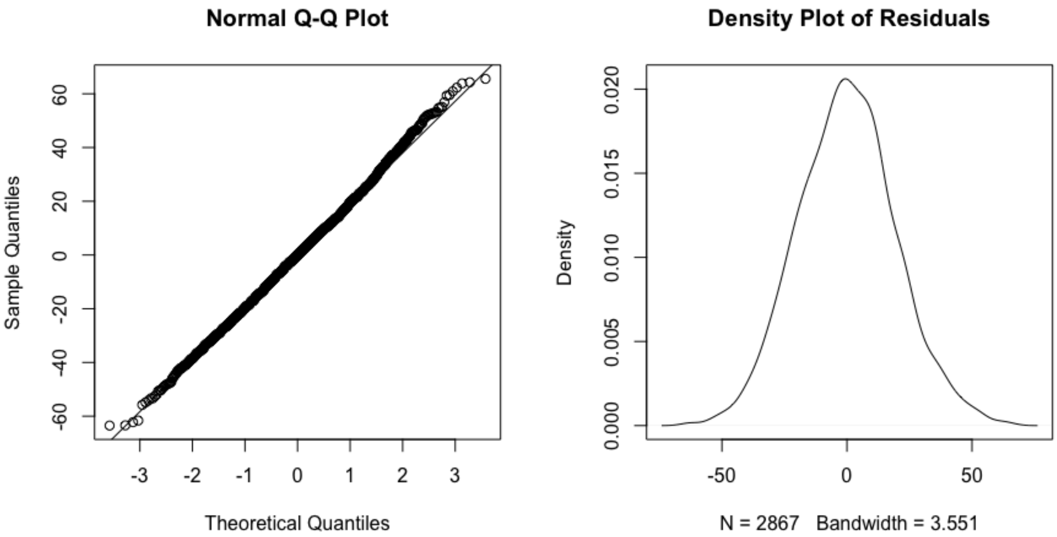


Figure 2.3 Q-Q plot (left) and density plot (right) of residuals

3. Results

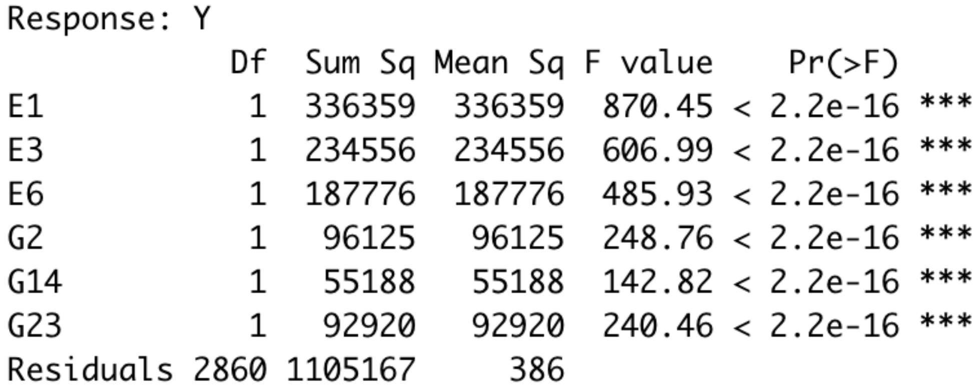
The final model shown below includes three continuous environmental variables, and three genetic variables, each of the latter taking on value 1 if the gene is present in the individual and value 0 if the gene is absent.

$$y_i = \beta_0 + \beta_{E1}X_{E1i} + \beta_{E3}X_{E3i} + \beta_{E6}X_{E6i} + \beta_{G2}X_{G2i} + \beta_{G14}X_{G14i} + \beta_{G23}X_{G23i}$$

$$y_i = 6.31 + 11.36X_{E1i} + 9.14X_{E3i} + 7.81X_{E6i} + 13.77X_{G2i} + 9.50X_{G14i} + 12.31X_{G23i}$$

No significant two-factor interaction terms were found. Therefore, three-factor and four-factor interaction terms were not evaluated. The initial scatter plots and subsequent regression analysis did not show any curvilinear relationship between the dependent variable Y and the continuous environmental variables. Therefore, higher order terms were not included in the final model.

Table 3.1 ANOVA Table for the final model with 3 environmental and 3 genetic variables



```
Response: Y
            Df  Sum Sq Mean Sq F value    Pr(>F)    
E1           1  336359  336359  870.45 < 2.2e-16 ***
E3           1  234556  234556  606.99 < 2.2e-16 ***
E6           1  187776  187776  485.93 < 2.2e-16 ***
G2           1   96125   96125  248.76 < 2.2e-16 ***
G14          1   55188   55188  142.82 < 2.2e-16 ***
G23          1   92920   92920  240.46 < 2.2e-16 ***
Residuals 2860 1105167     386
```

4. Conclusion(s)/Discussion

The final model included both environmental and genetic variables. However, no gene-environmental interactions were included. As with any study, a clear limitation is the sample size. While fairly large at n = 2867, it was not sufficient to avoid being underpowered to search for three factor and four factor interactions. Another limitation encountered in the analysis was computing power. With 2^{38} possibilities, it was not feasible to evaluate all possible regressions. There were 240 gene-environment interactions and $(30\,C\,2) = 435$ gene-gene interaction terms.

5. References

Caspi A, Sugden K, Moffitt TE, Taylor A, et al. 2003. Influence of Life Stress on Depression: Moderation by a Polymorphism in the 5-HTT Gene. Science. 301: 386-389.

Risch N, Herrell R, Lehner T, Liang KY, et al. 2009. Interaction Between the Serotonin Transporter Gene (5-HTTLPR), Stressful Life Events, and Risk of Depression. JAMA. 301(23): 2462-2471.

6. Technical Appendix

R Code

```
df <- read.csv("/Users/brandondanna/Documents/AMS 578/Project/data_470479.csv")
#Check normality assumption of Y
#Adjust number of plots to display
par(mfrow=c(1,2))

#create Q-Q plot for Y
qqnorm(df$Y)

#add a straight diagonal line to the plot
qqline(df$Y)

#Create density plot of Y
plot(density(df$Y),main = "Density Plot of Depedent Variable")

#Y is skewed, use Box Cox transformation to find optimal lambda
require(MASS)
bc <- boxcox(df$Y ~ 1)
(lambda <- bc$x[which.max(bc$y)])
```

```
lambda <- 0.5

#Re-evaluate normality assumption
#Adjust number of plots to display
par(mfrow=c(1,2))

#create Q-Q plot for residuals
qqnorm(((df$Y^lambda-1)/lambda))

#add a straight diagonal line to the plot
qqline(((df$Y^lambda-1)/lambda))

#Create density plot of residuals
plot(density(((df$Y^lambda-1)/lambda)),main = "Density Plot of Transformed Depedent Variable")

df$Y = ((df$Y^lambda-1)/lambda)
#Plot Y against continuous environmental variables to determine if a linear relationship exists
require(ggplot2)

ggplot(df, aes(x=E1, y=Y)) + geom_point()
ggplot(df, aes(x=E3, y=Y)) + geom_point()
ggplot(df, aes(x=E6, y=Y)) + geom_point()

#Adjust number of plots to display
par(mfrow=c(1,3))
plot(df$E1,df$Y,xlab="E1",ylab="Y",main = "Scatter of E1 vs Y")
plot(df$E3,df$Y,xlab="E3",ylab="Y",main = "Scatter of E3 vs Y")
plot(df$E6,df$Y,xlab="E1",ylab="Y",main = "Scatter of E6 vs Y")
#E1, E3, E6
```

```
#Check for missing values
sum(is.na(df))
#No missing data was found

#Perform correlation test on each of the three environmental variables that showed a linear
association
cor.test(df$Y,df$E1) #0.39 with sig p-value
cor.test(df$Y,df$E3) #0.32 with sig p-value
cor.test(df$Y,df$E6) #0.27 with sig p-value

#Perform t-tests for mean of y conditioned on each indicator variable
for (i in colnames(df)[10:39]){
  print(t.test(subset(df, df[[i]] == 1,select=c(Y)),subset(df, df[[i]] == 0,select=c(Y)))$p.value)
}
#G2,G14,G21,G23,G24 appear to be significant between 0 and 1 indicator on/off

#Fit multiple regression of Y on all 38 IV
lm.fit.full <- lm(Y ~ E1 + E2 + E3 + E4 + E5 + E6 + E7 + E8 + G1 + G2 + G3 + G4 + G5 + G6
+ G7 + G8 + G9 + G10
                  + G11 + G12 + G13 + G14 + G15 + G16 + G17 + G18 + G19 + G20 + G21 + G22 +
G23 + G24
                  + G25 + G26 + G27 + G28 + G29 + G30,
                  data = df)
summary(lm.fit.full)

#Backward selection, check F stat for each of two models with one deleting variables
lm.fit.minus1 <- lm(Y ~ E1 + E3 + E6 + G2 + G6 + G14 + G23,
                  data = df)
lm.fit.minus2 <- lm(Y ~ E1 + E3 + E6 + G2 + G14 + G23,
                    data = df)
summary(lm.fit.minus2)
```

```
#Compare each iteration of two models
anova(lm.fit.minus1,lm.fit.minus2)

#Check all 2 way interaction terms
lm.fit.fullint <- lm(Y ~ (E1 + E2 + E3 + E4 + E5 + E6 + G1 + G2 + G3 + G4 + G5 + G6 + G7 +
G8 + G9 + G10
          + G11 + G12 + G13 + G14 + G15 + G16 + G17 + G18 + G19 + G20 + G21 + G22 +
G23 + G24
          + G25 + G26 + G27 + G28 + G29 + G30)^2,
          data = df)
summary(lm.fit.fullint)

#Stepwise forward selection to determine if there are any significant interaction terms
intercept_only <- lm(Y ~ 1, data = df)
all <- lm(Y ~.,data = df)
forward <- step(intercept_only,direction = "forward",scope = formula(lm.fit.fullint),trace = 0)
forward$anova

#Correlation matrix of the data set
cor(df)

#Calculate Mallow's CP
require(olsrr)
ols_mallows_cp(lm.fit.minus2,lm.fit.full)

#Create vector of residuals
res <- resid(lm.fit.minus2)

#produce residual vs. fitted plot
```

```
plot(fitted(lm.fit.minus2), res, main = "Residuals vs. Fitted Plot",xlab = "Fitted Values",ylab = "Residuals")

#add a horizontal line at 0
abline(0,0)

#Adjust number of plots to display
par(mfrow=c(1,2))

#create Q-Q plot for residuals
qqnorm(res)

#add a straight diagonal line to the plot
qqline(res)

#Create density plot of residuals
plot(density(res),main = "Density Plot of Residuals")

#create custom function to calculate the PRESS statistic
PRESS <- function(model) {
  i <- residuals(model)/(1 - lm.influence(model)$hat)
  sum(i^2)
}

#calculate PRESS for full model
PRESS(lm.fit.full)

#calculate PRESS for model 2
PRESS(lm.fit.minusint)

#calculate PRESS for final model
```

```
PRESS(lm.fit.minus2)

#Final model has smaller PRESS
```