Regression Theory Group Project

AMS 578

Group 15：

Linna Hu (111783771)

Siao Sun (111572643)

May 5th, 2018

Table of Contents

[**1 Introduction**](#_57rviv6zgawr) **3**

[**2 Method**](#_sfq207fot3qc) **3**

[2.1 Merging and Cleaning the Dataset](#_4tnue5659cyw) 3

[2.2 Transformation of Y](#_hj4nj9ty4ltd) 4

[2.3 Variable Selection](#_oa8r0tlvlt91) 5

[2.4 Modeling](#_8vhoge1xdy8a) 5

[**3 Results**](#_mca7h2oiylcc) **6**

[3.1 Results of model](#_nz8ga9xx9bo6) 6

[3.2 Diagnosis](#_v7dns6lq3p79) 7

[**4 Conclusion & Discussion**](#_p8d6hpgirm0l) **8**

[**5 Appendix**](#_a8mmfslm756a) **9**

[5.1 R code](#_py0qvlm7b7pq) 9

# 

# **1 Introduction**

In the study of Caspi et al, he tested why stressful experiences lead to depression in some people but not in others. A function polymorphism in the promoter region of the serotonin transporter gene was found to moderate the influence of stressful life events on depression. His study provided evidence of a gene-by-environment interaction, in which an individual’s response to environmental insults were moderated by his or her genetic makeup.[[1]](#footnote-1)

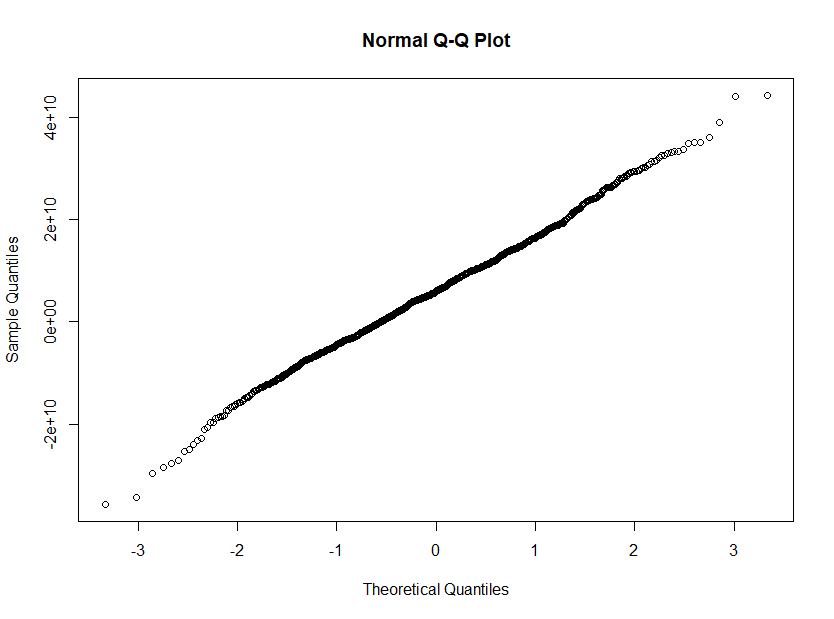
Similar to Caspi et al, to analyze a dataset containing observations of 1567 individuals, we determined to figure out whether there was a relationship between depression (), the dependent variables, and the independent variables, 6 environmental factors (- ) and 25 genetic factors (- ). According to the study, we also needed to find a relationship between Y and some interactions about and , such as , and etc. After a series of modeling, we got an optimal model for this dataset. This report illustrated the process of modeling and the final result of our model.

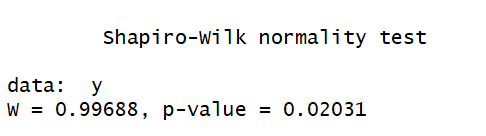
# **2 Method**

## 2.1 Merging and Cleaning the Dataset

When we got a new dataset and wanted to analyze it, the first important thing was to ensure the data was complete and without errors. So at first, we used R to import the dataset “IDEgroup15.csv”, “IDGgroup15.csv” and “IDYgroup15.csv”. In order to get a complete dataset, we merged all of the three datasets to a new dataset. Then we removed the observations that had at least one missing data. This process reduced the number of observations from 1567 to 1170 of 32 variables.

## 2.2 Transformation of Y

In ordinary linear model, one of the basic assumption was that the response variable needed to be normally distributed. So, firstly we used Normal-QQ plot and Shapiro’s test to check the normality of Y. The result is as below:

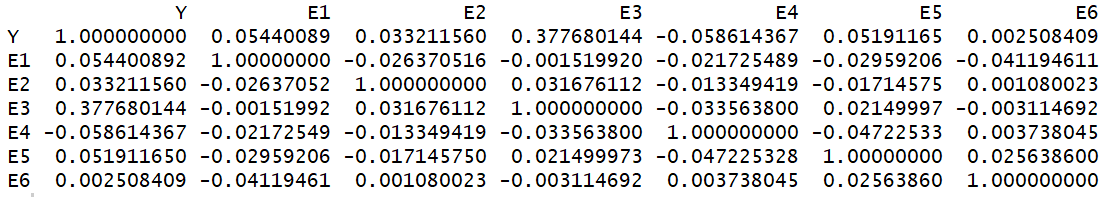


**Figure 1**: The Normal Q-Q Plot and Shapiro’s test of Y

From the figure of Normal-QQ plot, we could see the line was straight. However, in the Shapiro’s test, the p-value was smaller than . But we could not conclude that the data wasn’t normal. Because we all knew that if the sample size was large enough, the p-value of Shapiro’s test could be less than 0.05 even when the data was normal. So we were not sure if Y was normal or not. That’s why we did further checking.

First, we added a constant to make all the elements of Y bigger than 0, then we did the Box-cox Transformation to get the value of . In this case, the value of was equal to 1.0709, which was very close to 1. So we set and could think that our model accorded with a linear model.

## 2.3 Variable Selection

To narrow down the list of variables included in our model, we needed to find the correlation between Y and each E, Y and each G. For the independent variables - , we used a correlation test in R, then we chose the variables which had a correlation higher than 0.05 for Y. The result is as below:

**Table 1**: Correlation test between Y and E1 – E8

In this table, we could select , ,, which had strong relationships with Y.

For the binary independent variables G, we used t-test to find the significant terms for Y. The result is as below:

In this figure, we could see that ,and were significant terms for Y. By variable selection, we could fit out model more quickly and efficiently.

## 2.4 Modeling

After the first step, variable selection, we started to do the linear regression.

**Firstly**, we generated a model just with a intercept 1 and another linear model with all of two-way interaction terms. Then we did a stepwise regression for these two model to filter the variables which had multicollinearity.

Multicollinearity refers to a situation in which two or more explanatory variables in a multiple regression model are highly linearly related. The presence of multicollinearity will make the estimate of parameters unstable and less precise. That’s why we wanted to avoid multicollinearity when we built our model. To achieve the situation, a general approach was stepwise regression.

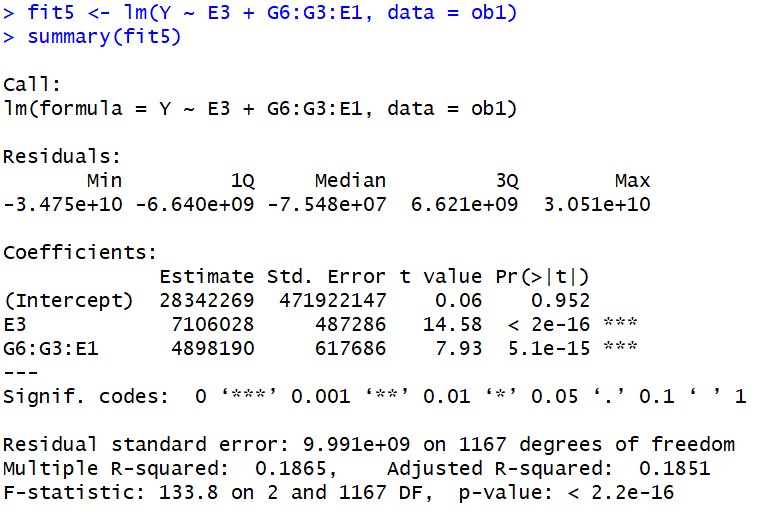
The tool we used to implement stepwise regression was function “stepAIC” in the R package “MASS”. It compared AIC of the model and chose the one with the minimum AIC. After stepwise regression, we observed that there were several variables and interactions whose p-values were smaller than 0.001. That meant they were significant and preferred. These variables turned out to be E1, E3, G3, G6, G20, G23, G25.

**Secondly**, we wanted to check if there were still strong correlations among these variables. So we generated a linear model between Y and the significant independent variables with two-way interaction terms. Then, we tested the multicollinearity for this model and noticed that the vif of almost all of the variables and interactions were bigger than 4, which indicated that there existed multicollinearity among the chosen variables. So we still needed to use stepwise regression to eliminate the multicollinearity when we considered three-way interaction terms. The function “vif” is in the R package “car”.

Afterwards we took three-way interaction terms into our model. We generated a new linear model with the significant variables and used stepwise regression to filtering the non-significant variables. Then we saw that only E3 and G6:G3:E1 were very significant.

# **3 Results**

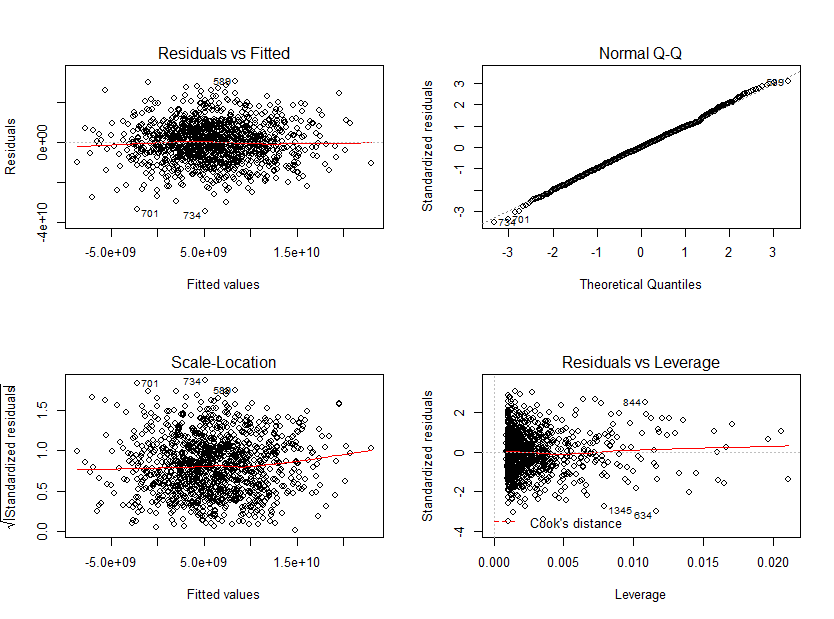
## 3.1 Results of model

According to stepwise regression, we found that all the independent variables and their two-way interactions were not significant when we considered the three-way interaction terms. In this situation, our final model was. The summary of our final model is as below:

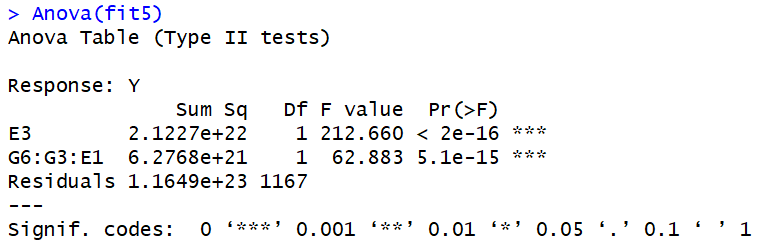
**Table 2:** Summary of stepwise model.

## 3.2 Diagnosis

We used “plot” in R to plot four graphs to test our final model. In the graphs “Residuals vs Fitted” and “Scale-Location” , points scattered randomly around y=0, which meant our regression model satisfied the assumption of linearity and [homoscedasticity](http://www.baidu.com/link?url=TbiB3LR4qjPcdQBuGweOxCVP6N6aq8v5Oz7nkyTkB87-KsTnWGA7F0-O4sjH2EjhbKAd28NIRCcosdzuYmC7iAkq-OEtk1ToJhypuTScTC3lpACgyYogaJUiN7BckwV4). The line in Normal-Q-Q plot was still straight, which was consistent with the assumption of normality.



**Figure 2**: Residuals

We also used ANOVA to test significant of variables. In the figure of ANOVA, we could see that both of these two terms were significant. We conclude that and could interpret the model well. The result is as below:

**Table 3**: ANOVA table.

# **4 Conclusion & Discussion**

In our study we first cleaned the dataset and eliminated all the incomplete data. The size of our dataset was reduced from 1567 to 1170 after that, with about 1/5 data eliminated, which was an acceptable proportion. Still we lost some information in the process. Some techniques might be used to complete the dataset so that we could fit better model for the data. Such techniques include maximum likelihood estimation, simple imputation, and multiple imputation. This is what can be improved of our study in the future.

We decided not to include four-way interactions because we found in our study that although adding four-way interaction terms could slightly improve the r - squared values of our model, it would make AIC values larger. So we decided to decrease the complexity of our model by excluding the four-way interactions.

To eliminate multicollinearity, we first applied correlation test and t-test to select variables, then did stepwise regression by AIC to the models which included two-way interactions and three-way interactions, and finally got our model. The r-squared value was 0.1851, which was acceptable. After diagnosis, we found that our model satisfied the assumptions of linear model. In the end, we concluded that this was our model for the data.

# **5 Appendix**

## 5.1 R code

# merge data to a new data

merge1 <- merge(Y[,2:3],E[,2:8],by="ID")

ob <- merge(merge1, G[,2:27], by="ID")

#remove NA

ob1 <- ob[complete.cases(ob), ]

# Transformation of Y

shapiro.test(y) #p-value= 0.02031<0.05

qqnorm(y)

y=ob1[,1] + 100000000000

summary(powerTransform(y)) # Est Power is 1.0709

## Variable selection

# choose the correlation bigger than 0.05

cor(ob1[,1:7]) #E1,E3,E4,E5

### t-test for Y vs.G: G6,G20,G22,

g <- NA

for (i in 1:25)

{

Y0 <- ob1$Y[ob1[,i+7]==0]

Y1 <- ob1$Y[ob1[,i+7]==1]

T <- t.test(Y0,Y1)

g[i] <- T$p.value

}

g<0.05

# fit the model

fit0 <- lm(Y ~ 1, data = ob1)

fit1 <- formula(lm(Y~(.)^2, data=ob1))

fit2 <- stepAIC(fit0, direction = "forward", scope=fit1)

summary(fit2)

# E3 G20 G6 G3 E1 G23 G25

fit3 <- formula(lm(Y ~ (E3 + G20 + G6 + G3 + E1 + G23 + G25)^3, data = ob1))

fits <- stepAIC(fit0, direction = "forward", scope=fit3)

summary(fits)

fit4 <- formula(lm(Y ~ (E1 + E3 + G3 + G6 + G20)^3, data = ob1))

fitss <- stepAIC(fit0, direction = "forward", scope=fit3)

summary(fitss)

fit5 <- lm(Y ~ E3 + G6:G3:E1, data = ob1)

summary(fit5)

#ANOVA

ANOVA(fit5)

#four-way interactions which is not good

fit6 <- formula(lm(Y~(E1 + E3 + G3 + G6 + G20)^4, data = ob1))

fit7 <- step(fit0, direction = "forward", scope = fit6)

summary(fit7)

fit8 <- lm(Y~E3+G20:G6:G3:E1, data = ob1)

summary(fit8)

#Residuals

par(mfrow=c(2,2))

plot(fit5)

1. Caspi, A., K. Sugden, T. E. Moffitt, A. Taylor, I. W. Craig, H. Harrington, J. McClay, J. Mill, J. Martin, A. Braithwaite, and R. Poulton. "Influence of Life Stress on Depression: Moderation by a Polymorphism in the 5-HTT Gene." Science 301 (2003): 386-389. [↑](#footnote-ref-1)