***AMS 578***

***Regression Analysis, Spring 2020***

***Multiple Regression Computing Project***

Yunlong Pan 113061415

1.Introduction

From Caspi's paper, influence of Life Stress on Depression, we know why stressful experiences lead to depression in some people but not in others. This epidemiological study thus provides evidence of a gene-by-environment interaction, in which an individual’s response to environmental insults is moderated by his or her genetic makeup.

In this study, I will examine this evidence using three file data. One file contains the patient identifier and the dependent variable value. The second file contains the patient identifier and values of six environment variables called E1 to E6. The third file contains the patient identifier and the twenty independent indicator variables called G1 to G20. The final goal of this study is to find a best model of Y on environmental/gene variables.

In section 2, I will introduce the methods to deal with missing data , select important independent variables and estimate each variable’s coefficient. In section 3, I will discuss the result of each step and the final model. Section 4 talks about model strengths and limitations.

2.Methods

2.1 Missing values

I make summary statistics of my data, with 1415, the mean, median, standard deviation, lower quartile point, upper quartile point, minimum, maximum, and the number of missing values. There are some missing data which will make some trouble to get the model and my method dealing with them is Mice(Multivariate Imputation via Chained Equations). The package creates multiple imputations (replacement values) for multivariate missing data. The method is based on Fully Conditional Specification, where each incomplete variable is imputed by a separate model. The MICE algorithm can impute mixes of continuous, binary, unordered categorical and ordered categorical data. In addition, MICE can impute continuous two-level data, and maintain consistency between imputations by means of passive imputation.

2.2 Variable Selection

We calculate the correlations between the independent variables and the dependent variable. For each variable, I process Pearson's product-moment correlation test which include p-value, 95 percent confidence interval(CI). Basically, a Pearson product-moment correlation attempts to draw a line of best fit through the data of two variables, and the Pearson correlation coefficient, *r*, indicates how far away all these data points are to this line of best fit. We can get our important independent variables, which have high correlations, also, small p-value with dependent variable, Y. Therefore, my model will contain them and also their interactions(up to four-way interactions). The model after this step is complicacy with coefficients, .

**2.3 Model Validation**

I will use Box-Cox transformation and stepwise regression methods to estimate each coefficient. Before that suppose that we wish to transform y to correct nonnormality and non-constant variance. A useful class of transformations is the power transformation , where is a parameter to be determined. Box and Cox show how the parameters of the regression model and λ can be estimated simultaneously using the method of maximum likelihood. I will use Box-Cox transition to decide if I need to do a transformation to dependent variable,, and stepwise regression with criterion BIC to get the best coefficients of each important independent variables and interactions. Finally, I use Shapiro-Wilk’s method to conduct the residual diagnose and get a good result.

3. Results

Total 301 missing values in 1000 patients are fought in summary of three files. Table 1 presents the summary statistics of the 28 variables after ‘mice’ dealing with missing data.

Table 1 Summary

Table 2 presents the result correlations between the independent variables and the dependent variable. Because have p-value much smaller than , I choose as important independent variables in my model. As mentioned above, I will assume my model is: . Next task is to find the best coefficients.

Table 2



Figure 1 shows the best , in Figure. But also Figure 1 tells us, the 95% confidence interval is about from 0 to 6, which means whatever the value we choose between this CI it makes limiting difference to the model.



Figure 1

For a simpler model, I will choose . We can get a model :. For estimating each , I will use stepwise regression in R.

> model1<-lm(Y~(E1+G1+G16)^4,data=data\_complete)

# Call:

# lm(formula = Y ~ (E1 + G1 + G16)^4, data = data\_complete)

#

# Coefficients:

# (Intercept) E1 G1 G16 E1:G1

# 19.69120 0.05612 -4.16155 -3.25276 0.08838

# E1:G16 G1:G16 E1:G1:G16

# 0.07121 4.34972 -0.08466

>model1\_bestBIC<-stepAIC(model1,k=log(nrow(data\_complete)))

> summary(model1\_bestBIC)

#

# Call:

# lm(formula = Y ~ E1 + G1 + G16 + G1:G16, data = data\_complete)

#

# Residuals:

# Min 1Q Median 3Q Max

# -0.74397 -0.13188 -0.00122 0.14139 0.66312

#

# Coefficients:

# Estimate Std. Error t value Pr(>|t|)

# (Intercept) 15.966761 0.353075 45.222 < 2e-16 \*\*\*

# E1 0.130898 0.007057 18.549 < 2e-16 \*\*\*

# G1 0.242760 0.037350 6.500 1.27e-10 \*\*\*

# G16 0.293077 0.037612 7.792 1.65e-14 \*\*\*

# G1:G16 0.132039 0.041872 3.153 0.00166 \*\*

# ---

# Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

#

# Residual standard error: 0.212 on 995 degrees of freedom

# Multiple R-squared: 0.589, Adjusted R-squared: 0.5874

# F-statistic: 356.5 on 4 and 995 DF, p-value: < 2.2e-16

Finally, I use this model and data in file to get and use Shapiro-Wilk’s normality test to check residual .

> Y\_hat<-model1\_bestBIC$fitted.values

> res <- data\_complete$Y-Y\_hat

> shapiro.test(res)

#

# Shapiro-Wilk normality test

#

# data: res

# W = 0.99842, p-value = 0.5005

From the output, the p-value > 0.05 implying that the distribution of the data are not significantly different from normal distribution. In other words, we can assume the normality. After calculating, I get my final model as:



Figure 2

4. Discussion and Conclusion

To find an estimating function, the easiest way is to run a stepwise model selection of regression with all candidate variables. However, this is not feasible for a large pool of dependent variables. In this study, candidate pool includes 6 s, 20 s, 120 s, 190 s and so on.

Limitations of this analysis should also be considered. First, the s variable is dummy variable with 0 or 1. In my model, I just consider them the same as s. There may be an improvement if consider s differently. Second, during variable selection, especially test the correlation between all three-way interactions with dependent variable Y, I get too many variables have high correlation and low p-value. And I don’t know how to handle these interaction variables. I think include some terms in model will increase R-squared. Third, this final model doesn’t have significant environment-genes interactions. This only means they are less statistically important than corresponding non-cross terms. Obviously, for an more accurate result, further analysis is needed.

Appendix

Code in R

E1415 <- read.csv("/Users/panyl/Desktop/project\_578/E/AMS578\_E\_1415.csv")

G1415 <- read.csv("/Users/panyl/Desktop/project\_578/G/AMS578\_G\_1415.csv")

Y1415 <- read.csv("/Users/panyl/Desktop/project\_578/Y/AMS578\_Y\_1415.csv")

data1415 <- cbind(E1415,G1415,Y1415)

library('MASS')

data1415\_summary<-data.frame(

mean = sapply(data1415, mean,na.rm=T),

median = sapply(data1415,median,na.rm=T),

sd = sapply(data1415,sd,na.rm=T),

lower\_quartile=sapply(data1415,quantile,na.rm=T)[2,],

upper\_quartile=sapply(data1415,quantile,na.rm=T)[4,],

min = sapply(data1415,min,na.rm=T),

max = sapply(data1415, max,na.rm=T),

number\_missing = sapply(data1415,function(x)sum(is.na(x)))

)

data1415\_summary<-round(data1415\_summary, digits=2)

sum(data1415\_summary$number\_missing)

colnames(data1415\_summary)[0]<-'var'

write.table(data1415\_summary,file="data1415\_summary.csv",sep=',',col.names = T,row.names = T)

library('mice')

data\_complete <- complete(mice(data1415,printFlag = F))

data\_summary<-data.frame(

mean = sapply(data\_complete,mean),

median = sapply(data\_complete,median),

sd = sapply(data\_complete,sd),

lower\_quartile=sapply(data\_complete,quantile)[2,],

upper\_quartile=sapply(data\_complete,quantile)[4,],

min = sapply(data\_complete,min),

max = sapply(data\_complete, max),

number\_missing = sapply(data\_complete,function(x)sum(is.na(x)))

)

data\_summary<-round(data\_summary,digits = 2)

write.table(data\_summary,file="data\_summary.csv",sep=',',col.names = T,row.names = T)

# Get Table 1

model1\_box<-boxcox(Y~(E1+G1+G16),data=data\_complete,lambda = seq(-5,10,by=0.1))

model1\_cox<-cbind(model1\_box$y,model1\_box$x)

model1\_cox[model1\_cox[,1]==max(model1\_box$y)]

#

data\_cox=data\_complete

data\_cox$Y=data\_complete[,27]^(model1\_cox[model1\_cox[,1]==max(model1\_box$y)][2])

# head(data\_cox)

# correlation<-cor.test(data\_complete$Y,data\_complete[,1])

# cor.test(data\_complete$Y,data\_complete[,1])$estimate

# correlation

n=c()

m=c()

p=c()

for(i in 1:26){

n[i] = cor.test(data\_complete$Y,data\_complete[,i])$conf.int[c(1)]

}

for(i in 1:26){

m[i] = cor.test(data\_complete$Y,data\_complete[,i])$conf.int[c(2)]

}

for(i in 1:26){

p[i] = cor.test(data\_complete$Y,data\_complete[,i])$p.value

}

cr=c()

for(i in 1:26){

cr[i] = cor.test(data\_complete$Y,data\_complete[,i])$estimate

}

data\_cor<-data.frame(

var = colnames(data\_complete[-27]),

'CI\_lower' = round(n,digits = 4),

'CI\_upper'=round(m,digits = 4),

'p-value' = round(p,digits = 4),

'Cor' = round(cr,digits = 4)

)

data\_cor=round(data\_cor,digits = 4)

write.table(data\_cor,file = 'data\_cor.csv',sep=',') # Get Table 2

model1<-lm(Y~(E1+G1+G16)^4,data=data\_complete)

model1\_bestBIC<-stepAIC(model1,k=log(nrow(data\_complete)))

summary(model1\_bestBIC)

Y\_hat<-model1\_bestBIC$fitted.values

res <- data\_complete$Y-Y\_hat

shapiro.test(res)

plot(res) # Get Figure 2