AMS 578 Project Report

Yizhen Jia, student ID 111520996

Professor Stephen Finch

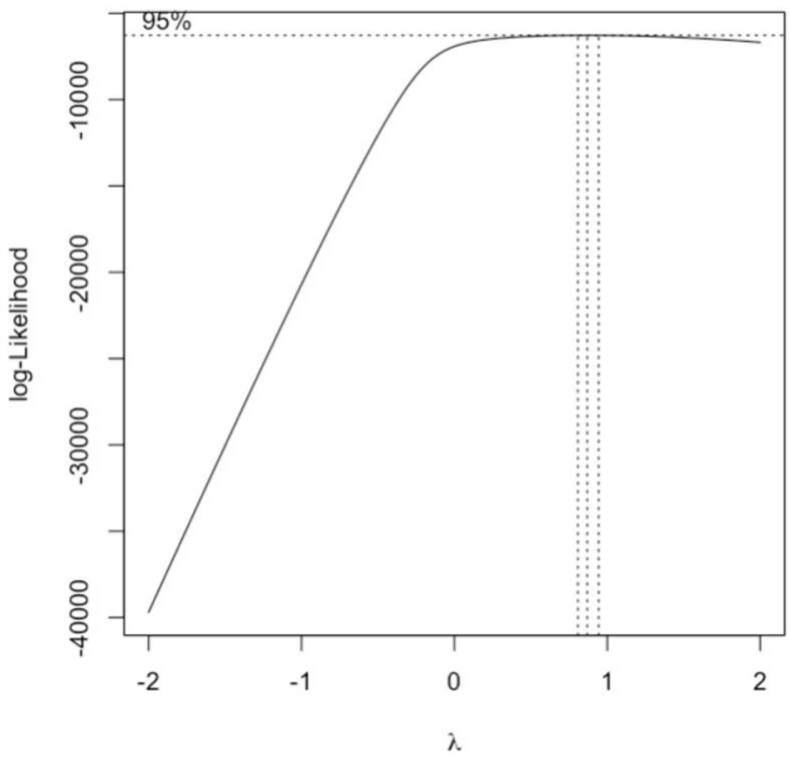
May 3, 2022

Introduction

In this report, I am provided a data set of 2505 observations with 39 variables, one dependent variable Y, 8 environment variables from E1 to E8, and 30 gene variables from G1 to G30. Our objective is to recover the function (containing some of these variables) that generates the data set. I first confirm that there is no missing value in the dataset, then calculate interactions between the genetic and environmental variables. To ensure that our procedure is correct, we will also use tables and figures to support our work.

Method

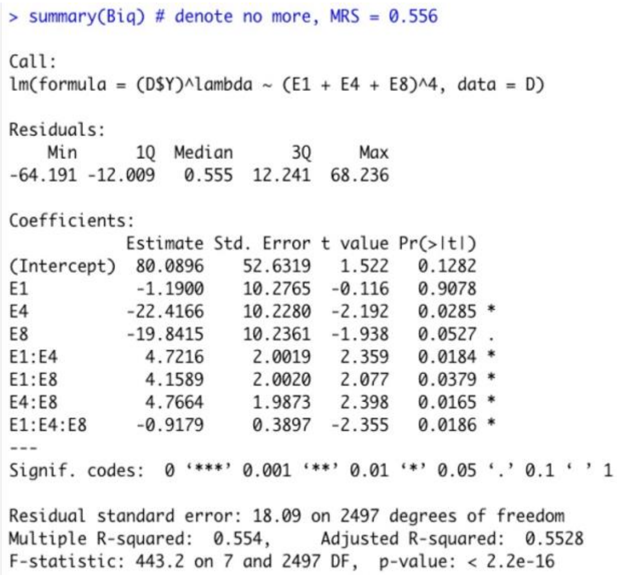
[Part (1) in code] I use package mice to ensure no missing value, and package MASS to support functions and datasets. [Part (2)] After I receive there are 2505 observations, [Part (3)] I apply the Boxcox method to the model, and I find out that λ is about 0.8686869. (If λ is 1, that it is not necessary to transform). Boxcox Plot:



In this case, the exponent of Y is 0.8686869. [Part (4)] Next, I use all 39 variables to fit linear model. [Part (5)] Finally, I use denoted variables with P value less than 10^-2 to fit bi-quadratic exponential model, then get the final model, follows residual plot.

Result

In part (4) in code, I denote E1 E4 and E8 which P value each of them is less than 10^-2. Put these three variables into bi-quadrature exponential model, get the following result:

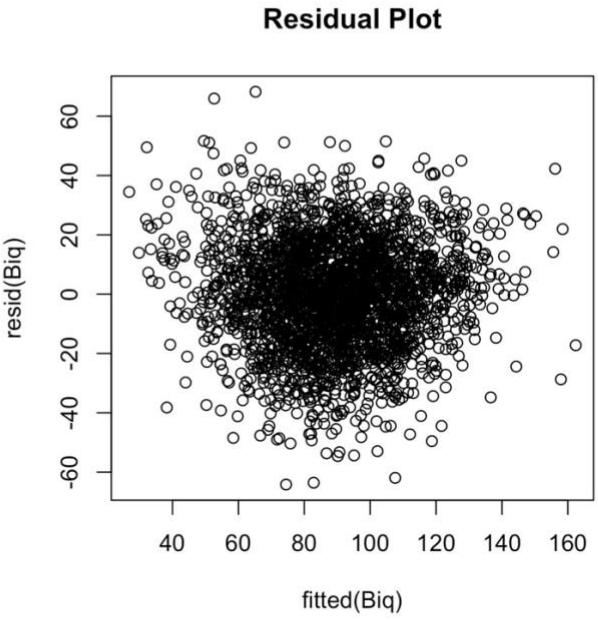


```
> summary(Biq) # denote no more, MRS = 0.556

Call:
lm(formula = (D$Y)^lambda ~ (E1 + E4 + E8)^4, data = D)

Residuals:
    Min      1Q  Median      3Q     Max 
-64.191 -12.009   0.555  12.241  68.236 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept)  80.0896    52.6319   1.522   0.1282  
E1           -1.1900    10.2765  -0.116   0.9078  
E4          -22.4166    10.2280  -2.192   0.0285 *
E8          -19.8415    10.2361  -1.938   0.0527 .
E1:E4         4.7216     2.0019   2.359   0.0184 *
E1:E8         4.1589     2.0020   2.077   0.0379 *
E4:E8         4.7664     1.9873   2.398   0.0165 *
E1:E4:E8     -0.9179     0.3897  -2.355   0.0186 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.09 on 2497 degrees of freedom
Multiple R-squared:  0.554,	Adjusted R-squared:  0.5528 
F-statistic: 443.2 on 7 and 2497 DF,  p-value: < 2.2e-16
```

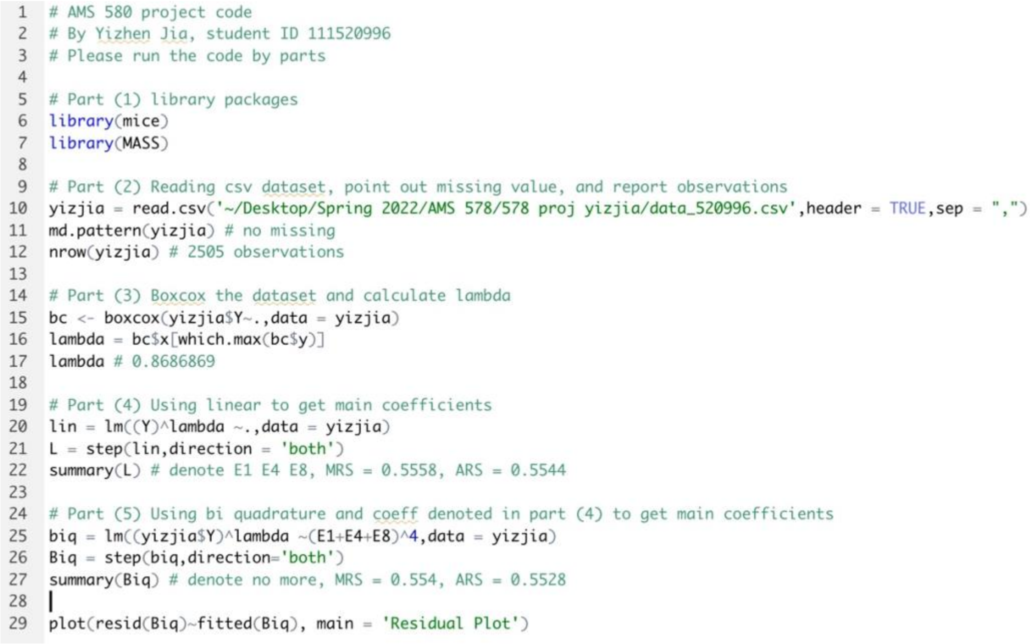
From is table, the transformed model is:

Y^0.8686869 = 80.0896 - 1.1900 E1 - 22.4166 E4 - 19.8415E8 + 4.7216 E1E4 + 4.1589 E1E8 + 4.7664 E4E8 - 0.9179 E1E4E8, with Multiple R-squared 0.554 and Adjusted R-squared 0.5528.

Residual Plot:



Code



```
# AMS 580 project code
# By Yizhen Jia, student ID 111520996
# Please run the code by parts

# Part (1) library packages
library(mice)
library(MASS)

# Part (2) Reading csv dataset, point out missing value, and report observations
yizjia = read.csv('~/Desktop/Spring 2022/AMS 578/578 proj yizjia/data_520996.csv',header = TRUE,sep = ",")
md.pattern(yizjia) # no missing
nrow(yizjia) # 2505 observations

# Part (3) Boxcox the dataset and calculate lambda
bc <- boxcox(yizjia$Y~.,data = yizjia)
lambda = bc$x[which.max(bc$y)]
lambda # 0.8686869

# Part (4) Using linear to get main coefficients
lin = lm((Y)^lambda ~.,data = yizjia)
L = step(lin,direction = 'both')
summary(L) # denote E1 E4 E8, MRS = 0.5558, ARS = 0.5544

# Part (5) Using bi quadrature and coeff denoted in part (4) to get main coefficients
biq = lm((yizjia$Y)^lambda ~(E1+E4+E8)^4,data = yizjia)
Biq = step(biq,direction='both')
summary(Biq) # denote no more, MRS = 0.554, ARS = 0.5528

plot(resid(Biq)~fitted(Biq), main = 'Residual Plot')
```

Conclusion

By the help of tools Boxcox transformation and stepwise regression, I can deduce the model to be Y^0.8686869 = 80.0896 - 1.1900 E1 - 22.4166 E4 - 19.8415E8 + 4.7216 E1E4 + 4.1589 E1E8 + 4.7664 E4E8 - 0.9179 E1E4E8. By the way, this is an approximated model because the more accurate model needs much more computing power. Multiple R-squared 0.554 and Adjusted R-squared 0.5528 suggest that the variation in the independent variables can explain about 55% of the variation in the dependent variable, which also suggest that there might be better transformation exist.

Thank you so much.