Computing Assignment Two

Vincent Forca (112620895)

May 7th, 2020

AMS 315
Data Analysis

Professor Stephen Finch

Stonybrook University

Spring 2020

AMS 315 computing assignment 2- Multiple Regression

Vincent Forca (112620895)
Professor Stephen Finch

INTRODUCTION

"This project is meant to expand on the concept in a paper by Capsi et al." (CH.12 lecture notes) The paper in question by Capsi et al "tested why stressful experiences lead to depression in some people but not others."(Influence of Life Stress on Depression: Moderation by a polymorphism in the 5-HTT Gene)

We were given a CSV file with 1094 observations containing one Dependent variable 'Y', and 24 Independent variables: 4 positive and continuous "environmental" variables named E1-E4, and 20 "indicator" variables named G1-G20. The environmental value represents an event that may happen in life; in the paper, these are described as, "stressful life events between ages 21 and 26". (CH.12 lecture notes). The "indicator" variables G1-G20 are referring to the gene that may put an individual at risk. These have a value 1 or 0 which indicate if the individual is "at risk" based on his genotype (value of 0 means the individual is not at risk). The dependent variable 'Y' is the outcome, dependent on the environmental and genetic data, that determines whether the participant is depressed or not (in the paper, Y is determined by age 26). The task of this assignment is to take the data inspired by *Influence of Life Stress on Depression: Moderation by a polymorphism in the 5-HTT Gene by Capsi et al,* and obtain the model that generated the data presented to us using stepwise regression in a statistical program of our choice. We will also be looking to find the environmental associations with the outcome, whether there are any gene-gene, gene-environmental, environmental-environmental interactions, and whether there is a genetic association with the outcome after controlling for the environmental variables.

METHODS AND MATERIALS

The statistical program used to achieve this outcome is called The R project for statistical Computing. *All codes utilized during this assignment was provided in the Multiple Regression Handout by Songzhu Zheng*. First, we fit a model using only the environmental data (E1, E2, E3, E4, E5). Then, we fit a model using all 24 environmental and indicator values. The model seemed accurate enough, but to confirm its accuracy we used a Box-Cox transformation to find that our estimated lambda was 1. From this point, we continued to stepwise regression, where our model summary gave us a selection: (Intercept)+E1:E4+E3:E4+G3:G8 (appendix #3). We then checked the main effects and confirmed that E1, E3, E4, G3 and G8 all had a considerable main effect (appendix #4). After this step, we checked for any second order interactions, and found there to be none. Finally, we extracted the coefficients to find the final model for our data set.

OUTCOMES

When we fit the first model using only the environmental data, our adjusted r-squared value was 0.3562, we also noted that E1, E3, and E4 have statistically significant P-values (appendix #1). After fitting the second model using all 24 environmental and indicator values, we obtained an increased adjusted r-squared value of 0.4764, we also noted that G3, and G8 has statistically significant P-values(appendix #2). We considered all interaction terms up to the second and third order {(.)^2 and (.)^3}, but we found that the highest adjusted r-squared value was given by the original data { (.) }. Using

a Box-Cox transformation to confirm lambda=1, we made no transformation to the dependent variable. After obtaining our proposed model (Intercept)+E1:E4+E3:E4+G3:G8, and confirming each variable had a main effect, we constructed the following table to assemble our final model (appendix #6):

	Estimate	Std. Error	**t value**	Pr(>|t|)
(Intercept)	**68.677268**	15.8284196	**4.338858**	1.565447e-05
E1	**4.808966**	0.4193454	**11.467792**	8.037752e-29
E3	**5.943418**	0.4160097	**14.286728**	1.449558e-42
E4	**8.451000**	0.4110064	**20.561726**	1.255978e-79
G3	**13.902624**	1.2301203	**11.301841**	4.404239e-28
G8	**13.346836**	1.2203896	**10.936536**	1.739709e-26

Given that there were no second order interactions (appendix #5), we used this information above to construct the model used to generate our data:

Y = **intercept** + b0**E1** + b1**E3** + b2**E4** + b3**G3** + b4**G8.**

After extracting the coefficients from the table:

Y = 68.677368 + 4.808966**E1** + 5.943418**E3** +8.451000**E4** + 13.902624**G3** + 13.346836**G8.**

This is the final model for the data.

DISCUSSION

We first found the variable E1, E3, E4, G3 and G8 to have statistically significant P-values. This was the first clue that these variables would be meaningful to our final model. The proposed model from our stepwise regression table was the third model in the chart: (Intercept)+E1:E4+E3:E4+G3:G8. We chose this because the BIC decrease and the adjusted r-squared increase between this model and the fourth one seemed insignificant compared to the second and the third (see appendix #3). The 5 variables in our chosen model were the same 5 variables we found to have statistically significant P-values and to top it off, all 5 of these also had a significant main effect. We knew these 5 variables should be in our final model, and we found there to be no second order interactions (between E1, E3, E4, G3 and G5 - see appendix #4). This means that there were no gene-gene, gene-environmental, or environmental-environmental interactions present. One limitation encountered during this analysis is that we did not look for any third order interactions between E1, E3, E4, G3, and G8. Based on our outcome, we concluded that there are genetic associations (G3 and G8), and environmental associations (E1, E3, E4), with our outcome variable 'Y'.

Computer Data Appendix:

(code and results from R)

```
wdir <- "C:\\Users\\forca\\Documents\\AMS315"

> setwd(wdir)

> Proj2 <- read.csv('P2_20895.csv', header = TRUE)
```

(1) **MODEL WITH JUST ENVIRONMENTAL DATA***

> M_E <- lm(Y ~ E1+E2+E3+E4, data=Proj2)

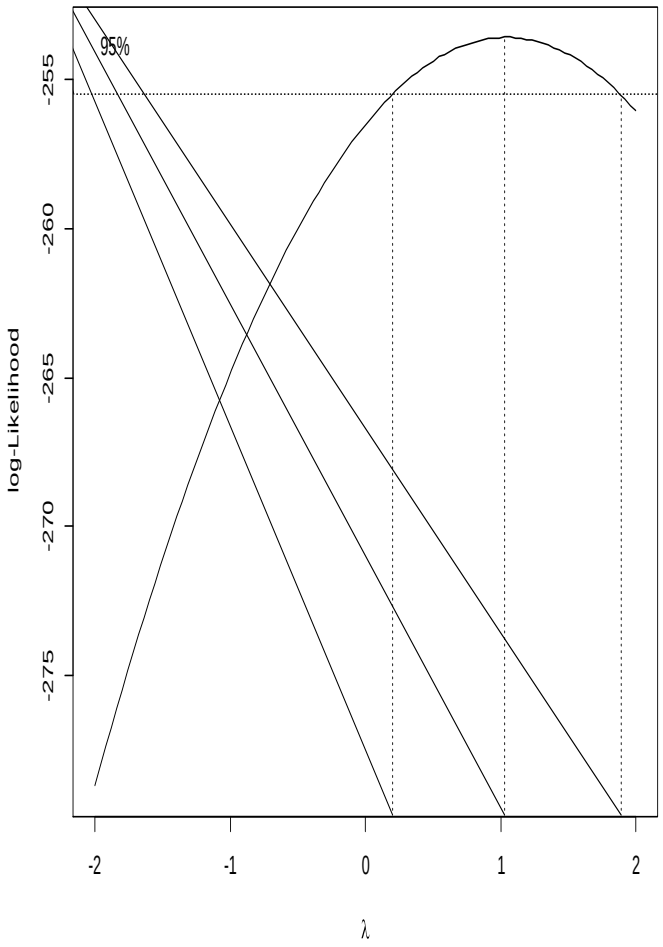
```
> summary(M_E)

Call:

lm(formula = Y ~ E1 + E2 + E3 + E4, data = Proj2)

Residuals:

   Min     1Q  Median     3Q    Max

-65.304 -15.220  -0.033  15.465  67.841

Coefficients:

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

(Intercept) 115.9836    20.3256   5.706 1.49e-08 ***

E1            4.7206     0.4654  10.142  < 2e-16 ***

E2           -0.6163     0.4585  -1.344    0.179

E3            5.6199     0.4607  12.199  < 2e-16 ***

E4            8.0941     0.4555  17.770  < 2e-16 ***

---

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 22.02 on 1089 degrees of freedom

Multiple R-squared:  0.3586,    Adjusted R-squared:  0.3562

F-statistic: 152.2 on 4 and 1089 DF,  p-value: < 2.2e-16
```

-> the statistically significant p-values (highlighted) show that E1, E3, and E4 are statistically significant in our final model.

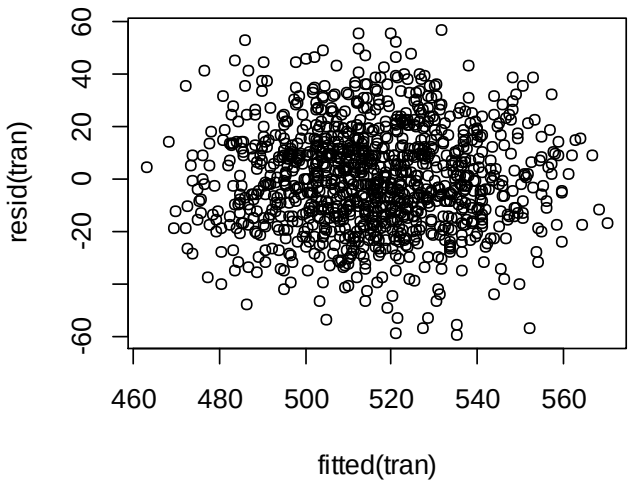
(2) ***MODEL WITH ALL ENVIRONMENTAL AND INDICATOR VALUES***

(AFTER BOXCOX FINDING LAMBDA=1)

```
tran <- lm(I((Y)) ~ (.), data=Proj2)

> summary(tran)

Call:

lm(formula = I((Y )) ~ (.), data = Proj2)

Residuals:

   Min     1Q  Median     3Q    Max

-59.549 -13.769  -0.571  12.828  56.245

Coefficients:

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

(Intercept) 82.77505   18.90574   4.325 1.66e-05 ***

E1           4.78101    0.42359  11.287  < 2e-16 ***

E2          -0.58485    0.41703  -1.402   0.1611

E3           5.91019    0.42008  14.069  < 2e-16 ***

E4           8.43589    0.41535  20.310  < 2e-16 ***

G1           0.38595    1.23671   0.312   0.7550

G2          -0.72438    1.23638  -0.586   0.5581

G3          13.83501    1.24714  11.093  < 2e-16 ***

G4          -1.14061    1.23784  -0.921   0.3570

G5          -1.28127    1.25127  -1.024   0.3061

G6           0.12370    1.23452   0.100   0.9202

G7           1.02118    1.23409   0.827   0.4082

G8          13.72471    1.23654  11.099  < 2e-16 ***

G9           1.53450    1.23547   1.242   0.2145

G10          2.13135    1.23054   1.732   0.0836 .

G11         -1.37869    1.24666  -1.106   0.2690

G12          0.25138    1.23835   0.203   0.8392
```



G13	0.36219	1.24165	0.292	0.7706
G14	1.44844	1.23729	1.171	0.2420
G15	-1.53157	1.22603	-1.249	0.2119
G16	-0.17252	1.22239	-0.141	0.8878
G17	-0.69740	1.23678	-0.564	0.5730
G18	-0.47591	1.23881	-0.384	0.7009
G19	0.04417	1.24615	0.035	0.9717
G20	1.15148	1.25376	0.918	0.3586



Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.86 on 1069 degrees of freedom

Multiple R-squared: 0.4879, ***Adjusted R-squared: 0.4764***

F-statistic: 42.43 on 24 and 1069 DF, p-value: < 2.2e-16

---

(3) ***STEPWISE REGRESSION MODEL CANDIDATE***

> Var <- colnames(model.matrix(trans))

> M_select <- apply(temp$which, 1, function(x) paste0(Var[x], collapse='+'))

kable(data.frame(cbind(model = M_select, adjR2 = temp$adjr2, BIC = temp$bic)), + caption='Model Summary')

model	adjR2	BIC
(Intercept)+E3:E4	0.287160014815579	-357.323330558124
(Intercept)+E3:E4+G3:G8	0.387997148874093	-518.184937144158
(Intercept)+E1:E4+E3:E4+G3:G8	***0.455508455546389***	***-640.062224359159***
(Intercept)+E1:E4+E3:E4+G3:G8+G3:G10	0.460291276644037	-643.720924304892
(Intercept)+E1:E4+E1:G8+E3:E4+G3:G8+G3:G10	0.468221497821803	-653.922359410227

➔ Here, we chose the third row, because the adjusted R-squared barley increases, and the BIC value barley decreases between model 3 and model 4, which may not be statistically significant.

---

(4) ****CHECKING MAIN EFFECT****

```
> M_main <- lm( I(Y) ~ ., data=Proj2)
> tempp <- summary(M_main)
> kable(tempp$coefficients[ abs(tempp$coefficients[,4]) <= 0.001, ], caption='Sig Coefficients')
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	82.775047	18.9057367	4.378303	1.31e-05
E1	4.781014	0.4235900	11.286890	0.00e+00
E3	5.910185	0.4200755	14.069342	0.00e+00
E4	8.435889	0.4153490	20.310364	0.00e+00
G3	13.835012	1.2471361	11.093426	0.00e+00
G8	13.724712	1.2365405	11.099282	0.00e+00

➔ All significant variables have a main effect.

---

(5) ***Checking second order interactions***

```
> M_2nd <- lm(Y~(.)^2, data=Proj2)
> temppp <- summary(M_2nd)
> kable(temppp$coefficients[ abs(temppp$coefficients[,4]) <= 0.001, ], caption='2nd Interaction')
```

	x
Estimate	3.6511413
Std. Error	0.9984177
t value	3.6569277
Pr(>|t|)	0.0002721

➔ There were no second order interactions between the variables.

(6) *** TABLE FOR FINAL MODEL***

> M_2stage <- lm(Y ~ (E1+E3+E4+G3+G8), data=Proj2)

> temp <- summary(M_2stage)

> temp$coefficients[abs(temp$coefficients[,3]) >= 4,]

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

(Intercept) --|68.677268 |15.8284196| 4.338858 | 1.565447e-05

E1 ---------| 4.808966 | 0.4193454 | 11.467792 |8.037752e-29

E3 ---------| 5.943418 |0.4160097 | 14.286728 |1.449558e-42

E4 ---------| 8.451000 | 0.4110064 |20.561726 |1.255978e-79

G3 ---------|13.902624 |1.2301203 | 11.301841 |4.404239e-28

G8 ---------| 13.346836|1.2203896 |10.936536 |1.739709e-26

➔ These are the coefficients used in our final model:

Y = **intercept** + b0**E1** + b1**E3** + b2**E4** + b3**G3** + b4**G8**

Y = 68.677368 + 4.808966**E1** + 5.943418**E3** +8.451000**E4** + 13.902624**G3** + 13.346836**G8**

End Computer data Appendix

REFERENCES

I. **Influence of Life Stress on Depression: Moderation by a polymorphism in the 5-HTT Gene**
Avshalom Caspi, *et* al.
Science **301**, 386 (2003);
DOI: 10.1126/science. 1083968
https://blackboard.stonybrook.edu/bbcswebdav/pid-5307826-dt-content-rid-40485137_1/courses/1204-AMS-315-SEC01-49021/Caspi_et_al._2003_Science.pdf

II. **Chapter 12 Lecture Notes**
https://blackboard.stonybrook.edu/bbcswebdav/pid-5307826-dt-content-rid-40485135_1/xid-40485135_1

III. **Multiple Regression handout by Songzhu Zheng**
https://blackboard.stonybrook.edu/bbcswebdav/pid-5337763-dt-content-rid-41358946_1/courses/1204-AMS-315-SEC01-49021/Multiple%20Regression%20Handout%20S2020%282%29.html

---

- End of Computing Assignment Two -