Vivianne Huang
AMS 315 Data Project #2
November 22, 2022

"Influence of Life Stress on Depression: Moderation by a Polymorphism in the 5-HTT Gene" by Caspi et al. discusses gene-environment interaction. The paper uses multiple regression techniques to help determine the findings of these interactions. In this report, I will be conducting a similar experiment but with synthetic data. I will be determining what function best matches my dataset. With this, I have to take into consideration four environmental variables, 20 gene indicator variables, 80 gene-environment variables, and 190 gene-gene interaction variables.

For my project, I am using R to perform multiple regression techniques. The first thing I would do is to import my dataset using "read.csv()" and name it "dataproj". My dataset contains 1,034 observations and 25 variables. I have one Y variable, four environmental variables (E1 through E4), and 20 genetic variables (G1 through G20). As a starting point, I am going to first use all the environmental variables to fit them with my Y variable using "lm()". I would then be able to see the adjusted R-squared value for this model. Next, I would use all the environmental and genetic variables and square the values to include possible second-order interactions in the new model. I would then create a residual plot with this model to determine whether the model seems adequate. To test out other options, I am going to use the Box-Cox transformation command from the "MASS" package to determine whether or not a transformation may be useful for my dataset. I can reaffirm my decision by looking at the adjusted R-squared values of the different transformations. The next step of my project involves "regsubset()" and "kable()" from the packages "leaps" and "knitr", respectively, to help me create a model summary with different possible models and their adjusted R-squared and the Bayesian Information Criterion (BIC). This table is useful to determine what variables might be in my best fit model for my dataset. I continue to use "lm()" with all the environmental and genetic variables and use "kable()" to show the significant variables and their respective significant coefficients where the p-value is less than or equal to 0.001. Using those variables, I am going to conduct another "lm()" and "kable()" with the sum of the variables squared to help determine if my model has any second-order interactions. This time, it will take into account the t-value of the variables and

possible interactions to return ones that are greater than or equal to 4. For my final model, labeled “M_final”, it is going to include the variables and interaction terms, if any, and compare to my Y variable, or its transformation (if any). My final fitted model will include the intercept and coefficients from my “summary(M_final)”.

Based on my analysis, the fitted model for my dataset is Y = 9.9601*E1 + 10.3169*E2 + 15.1134*E3 + 14.6635*E4 + 13.7231*G10 + 8.2202. As mentioned before, my dataset is complete with 1,034 observations and 25 variables. When I use the initial model with just the four environmental variables, I get an adjusted R-squared value of 0.6807. The next model with the squared sum of all the environmental and genetic variables in relation to my Y variable has an adjusted R-squared value of 0.6992782. When I created a residual model of this model, the plot returned a relatively flat ellipse, which means the model appears to be “adequate”. The residual plot can be seen below in Figure 1. I wanted to make sure that my dataset did not need a transformation for my Y variable, so I tested Y^0.8 based on the Box-Cox figure (Figure 2). The adjusted R-squared value for this model is 0.6995498. As there is not a huge difference in R-squared values between my original and my transformed data, I decided that I will not be using a transformation. After making this decision, I created the “Model Summary” table, which can be found below in Table 1. From the table, I decided that the third model, “(Intercept)+E1:E2+E2:G10+E3:E4”, is a great candidate for what my actual model would be. “M_main” is the model that relates all the environmental and genetic variables with the Y variable. Using “kable()”, I was able to make a table of significant coefficients with the significant variables. The result can be found below (Table 2), where I filtered it by p-value of less than or equal to 0.001. As I had predicted, the variables from the third model are the significant variables for my fitted model for my dataset. After that, I set up another model with just E1, E2, E3, E4, and G10 and squared the sum to determine whether or not my model has interactions. This time I looked at t-values greater than or equal to four and nothing popped up on my table. I lowered the threshold to one so I can see the different variables and interactions. Both of the interactions shown had t-values that were less than two and p-values greater than 0.1. This means that they are not statistically significant and that my actual fitted model most likely does not contain second-order interactions. I set my final fitted model, M_final, as E1 + E2 + E3 + E4 + G10. The summary and ANOVA table of M_final is shown below (Table 3). To check my work, I can look at the p-values of the coefficients and all of them were less than 0.001, meaning

that they are statistically significant. This model also supports my hypothesis that the third model from the R-squared/BIC table relates to my fitted model, making me more confident in my final fitted model. Thus, with all this evidence, the final model is $Y = 9.9601*E1 + 10.3169*E2 + 15.1134*E3 + 14.6635*E4 + 13.7231*G10 + 8.2202$ with an adjusted R-squared value of 0.6853.

In conclusion, my fitted model for my dataset contains both environmental and genetic variables but no second-order interactions. One possible limitation for my procedure was that the farthest I went was second-order interactions. It is possible that a third-order or even higher order interaction could have fit into my model and made it better. From my procedure, however, I can confidently say that my final model, $Y = 9.9601*E1 + 10.3169*E2 + 15.1134*E3 + 14.6635*E4 + 13.7231*G10 + 8.2202$ with an adjusted R-squared value of 0.6853, is highly reasonable.

Figure 1

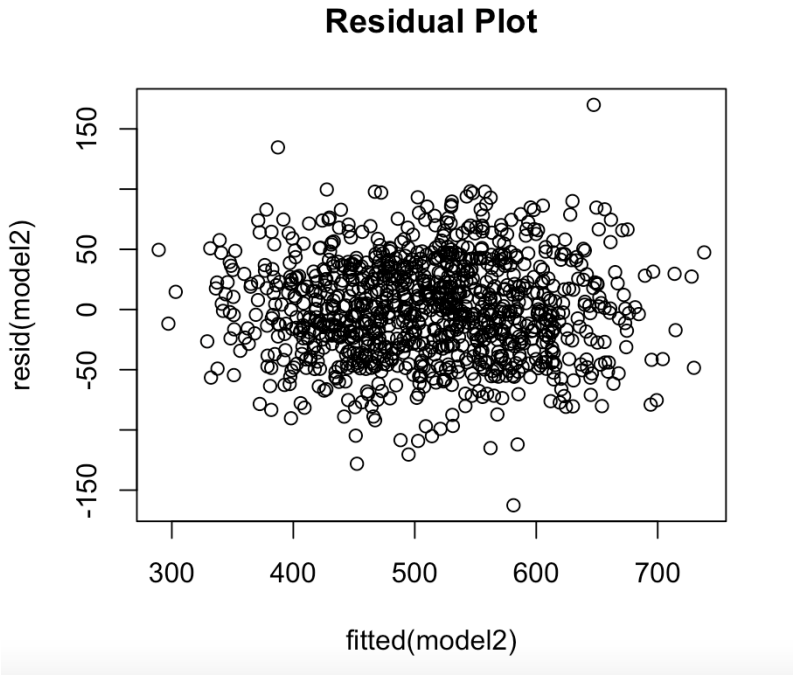

Figure 2

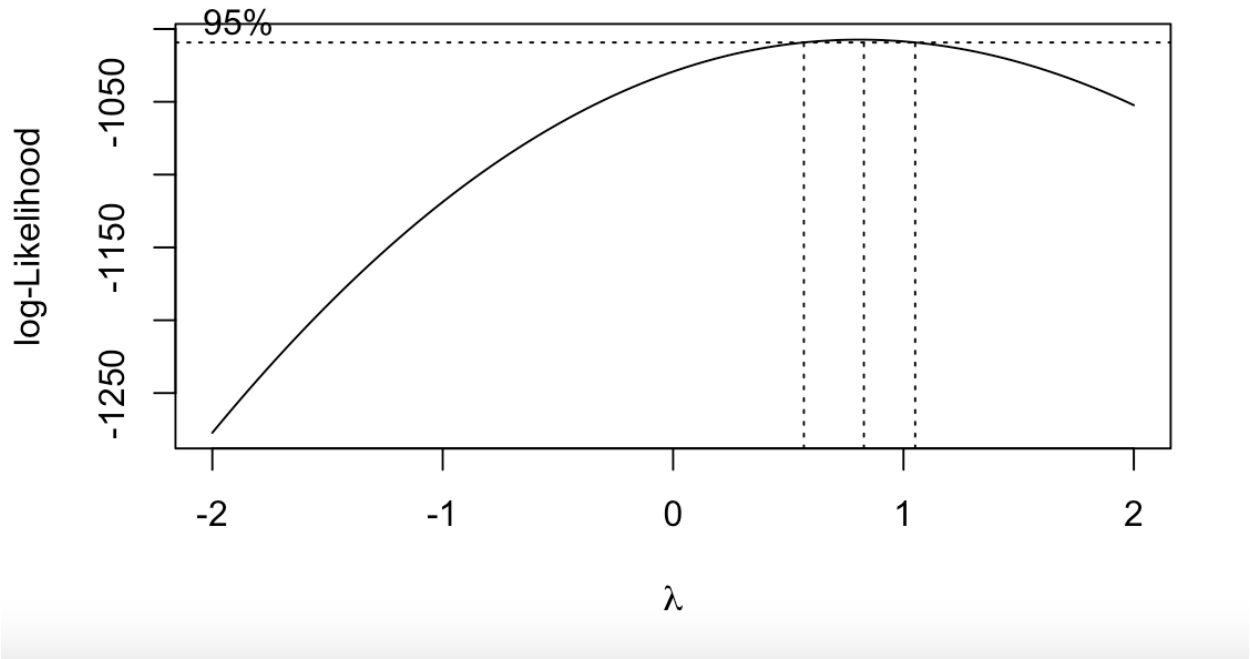

Table 1

model	adjR2	BIC
(Intercept)+E3:E4	0.445732006497428	-597.289676027052
(Intercept)+E1:E2+E3:E4	0.653005070454402	-1075.61255162122
(Intercept)+E1:E2+E2:G10+E3:E4	0.658256512335117	-1085.44299748881
(Intercept)+E1:E2+E2:G10+E2:G11+E3:E4	0.660611144603401	-1086.65515479363
(Intercept)+E1:E2+E2:G10+E2:G11+E3:E4+G2:G8	0.66260099587638	-1086.79954011048

Table 2

Table: Sig Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
E1	9.97906	0.5504260	18.129705	0.00e+00
E2	10.37681	0.5411443	19.175682	0.00e+00
E3	15.09874	0.5438416	27.763125	0.00e+00
E4	14.64652	0.5480630	26.724159	0.00e+00
G10	13.81820	3.4557291	3.998636	6.84e-05

Table 3

```
Call:
lm(formula = Y ~ (E1 + E2 + E3 + E4 + G10), data = dataproj)

Residuals:
     Min       1Q   Median       3Q      Max
-160.811  -33.945   -0.325   33.658  189.902

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   8.2202    10.8970   0.754    0.451
E1            9.9601     0.5462  18.236  < 2e-16 ***
E2           10.3169     0.5356  19.263  < 2e-16 ***
E3           15.1134     0.5397  28.003  < 2e-16 ***
E4           14.6635     0.5437  26.970  < 2e-16 ***
G10          13.7231     3.4274   4.004 6.68e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 50.47 on 1028 degrees of freedom
Multiple R-squared:  0.6868,    Adjusted R-squared:  0.6853
F-statistic: 450.9 on 5 and 1028 DF,  p-value: < 2.2e-16
```

Technical Appendix

```
> dataproj <- read.csv("~/Downloads/Project 2 Student Data Files/941886_project2.csv")
> model1 <- lm(Y ~ E1 + E2 + E3 + E4, data = dataproj)
> summary(model1)

Call:
lm(formula = Y ~ E1 + E2 + E3 + E4, data = dataproj)

Residuals:
     Min       1Q   Median       3Q      Max
-164.714  -35.090    0.507   34.613  185.526

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.4061    10.9625   0.949    0.343
E1           10.0057     0.5500  18.191   <2e-16 ***
E2           10.4058     0.5390  19.305   <2e-16 ***
E3           15.1700     0.5435  27.914   <2e-16 ***
E4           14.6681     0.5477  26.783   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 50.84 on 1029 degrees of freedom
Multiple R-squared:  0.6819,    Adjusted R-squared:  0.6807
F-statistic: 551.6 on 4 and 1029 DF,  p-value: < 2.2e-16

> model2 <- lm(Y ~ (E1 + E2 + E3 + E4 + G1 + G2 + G3 + G4 + G5 + G6 + G7 + G8 + G9 + G10 + G11 +
 G12 + G13 + G14 + G15 + G16 + G17 + G18 + G19 + G20)^2, data = dataproj)
> plot(resid(model2) ~ fitted(model2), main = 'Residual Plot')
> library(MASS)
> boxcox(model2)
> model3 <- lm( I(Y^.8) ~ (E1 + E2 + E3 + E4 + G1 + G2 + G3 + G4 + G5 + G6 + G7 + G8 + G9 + G10 +
 G11 + G12 + G13 + G14 + G15 + G16 + G17 + G18 + G19 + G20)^2, data = dataproj)
> summary(model2)$adj.r.square
[1] 0.6992782
> summary(model3)$adj.r.square
[1] 0.6995498
> plot(resid(model3) ~ fitted(model3), main = 'New Residual Plot')
> plot(resid(model2) ~ fitted(model2), main = 'Residual Plot')
```

```
> library(leaps)
> M <- regsubsets(model.matrix(model2)[,-1], I(dataproj$Y), nbest = 1, nvmax = 5, method = 'forwa
rd', intercept = TRUE)
> temp <- summary(M)
> library(knitr)
> vars <- colnames(model.matrix(model2))
> M_select <- apply(temp$which, 1, function(x) paste0(vars[x], collapse='+'))
> kable(data.frame(cbind(model = M_select, adjR2 = temp$adjr2, BIC = temp$bic, caption = 'Model S
ummary')))
```

```
|model                                        |adjR2             |BIC               |caption
 |
|:--------------------------------------------|:-----------------|:-----------------|:------------
-|
|(Intercept)+E3:E4                            |0.445732006497428 |-597.289676027052 |Model Summary
 |
|(Intercept)+E1:E2+E3:E4                      |0.653005070454402 |-1075.61255162122 |Model Summary
 |
|(Intercept)+E1:E2+E2:G10+E3:E4               |0.658256512335117 |-1085.44299748881 |Model Summary
 |
|(Intercept)+E1:E2+E2:G10+E2:G11+E3:E4        |0.660611144603401 |-1086.65515479363 |Model Summary
 |
|(Intercept)+E1:E2+E2:G10+E2:G11+E3:E4+G2:G8  |0.66260099587638  |-1086.79954011048 |Model Summary
 |
> M_main <- lm(Y ~ (E1 + E2 + E3 + E4 + G1 + G2 + G3 + G4 + G5 + G6 + G7 + G8 + G9 + G10 + G11 +
 G12 + G13 + G14 + G15 + G16 + G17 + G18 + G19 + G20), data = dataproj)
> temp <- summary(M_main)
> kable(temp$coefficients[abs(temp$coefficients[,4]) <= 0.001, ], caption = 'Sig Coefficients')
```

Table: Sig Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
E1	9.97906	0.5504260	18.129705	0.00e+00
E2	10.37681	0.5411443	19.175682	0.00e+00
E3	15.09874	0.5438416	27.763125	0.00e+00
E4	14.64652	0.5480630	26.724159	0.00e+00
G10	13.81820	3.4557291	3.998636	6.84e-05

```
> M_main2 <- lm(Y ~ (E1 + E2 + E3 + E4 + G10)^2, data = dataproj)
> temp <- summary(M_main2)
> kable(temp$coefficients[abs(temp$coefficients[,3]) >= 4, ])
```

	x
Estimate	13.3798020
Std. Error	3.1817312
t value	4.2051956
Pr(>|t|)	0.0000284

```
> kable(temp$coefficients[ abs(temp$coefficients[,3]) >= 1, ])
```

	Estimate	Std. Error	t value	Pr(>|t|)
E1	9.9306831	3.2487188	3.056800	0.0022955
E2	6.5257349	3.4245835	1.905556	0.0569894
E3	12.4458827	3.3310448	3.736330	0.0001971
E4	13.3798020	3.1817312	4.205196	0.0000284
E2:E3	0.2922618	0.1896486	1.541070	0.1236105
E2:G10	1.4035966	1.1746485	1.194908	0.2324014

```
> M_final <- lm(Y ~ (E1 + E2 + E3 + E4 + G10), data = dataproj)
> summary(M_final)
```

```
> M_final <- lm(Y ~ (E1 + E2 + E3 + E4 + G10), data = dataproj)
> summary(M_final)

Call:
lm(formula = Y ~ (E1 + E2 + E3 + E4 + G10), data = dataproj)

Residuals:
     Min       1Q   Median       3Q      Max
-160.811  -33.945   -0.325   33.658  189.902

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   8.2202    10.8970   0.754    0.451
E1            9.9601     0.5462  18.236  < 2e-16 ***
E2           10.3169     0.5356  19.263  < 2e-16 ***
E3           15.1134     0.5397  28.003  < 2e-16 ***
E4           14.6635     0.5437  26.970  < 2e-16 ***
G10          13.7231     3.4274   4.004 6.68e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 50.47 on 1028 degrees of freedom
Multiple R-squared:  0.6868,    Adjusted R-squared:  0.6853
F-statistic: 450.9 on 5 and 1028 DF,  p-value: < 2.2e-16

> anova(M_final)
Analysis of Variance Table

Response: Y
            Df  Sum Sq Mean Sq F value    Pr(>F)
E1           1  838569  838569 329.233 < 2.2e-16 ***
E2           1 1045560 1045560 410.500 < 2.2e-16 ***
E3           1 1963703 1963703 770.974 < 2.2e-16 ***
E4           1 1853803 1853803 727.826 < 2.2e-16 ***
G10          1   40834   40834  16.032 6.676e-05 ***
Residuals 1028 2618359    2547
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

References

https://blackboard.stonybrook.edu/bbcswebdav/pid-2061759-dt-content-rid-18474001_1/courses/1228-AMS-315-SEC01-88750/AMS-315-Multiple-Regression-Handout-Updated-F22.html