Introduction

The purpose of this experiment is to determine which variables were used in a given model. A database is given containing a dependent variable y and 24 independent variables consisting of 4 environmental variables E and 20 genetic variables G. We are tasked with using multiple regression techniques to determine the original model. This experiment is based on the Caspi et al. research regarding gene and environment interactions. More specifically, it discusses whether a specific genotype 5-HTTLPR has any significant association with depression and any significant interaction effect between the gene and stressful life events on depression. This experiment models the same methods but instead uses synthetic data.

Methods

The dataset was analyzed using R and R Studio. The first thing to do was to fit a model using only the 4 environmental variables. The resulting adjusted R squared value for the model is 0.508288. Next, the 20 genetic variables were added in order to test their influence. The residual plot (Figure 1) demonstrates a relatively patternless ellipse. This suggests that the model is sufficient and that no dependent variable transformation is required. Next, we attempted to narrow down the impactful variables. The R package ‘leaps’ was used to perform stepwise regression. More specifically, subsets regression (regsubset) was used on the model, which resulted in the proposed models (Table 1). The third model was selected, which contained variables E1, E2, E4, and G20. In order to verify which variables are viable, we looked at the main effects of the variables on the model. From each variable, those with a p-value less than 0.001 were selected and shown in Table 2. The variables E1, E2, and G20 were then selected.

After testing all of the variables, the best model consisted of only E1, E2, and G20. Thus the final model is estimated to be: Y = $\beta_0 + \beta_1 E_1 + \beta_2 E_2 + \beta_3 G_{20} + Z$

Results

The final model, Y = $\beta_0 + \beta_1 E_1 + \beta_2 E_2 + \beta_3 G_{20} + Z$, produced an adjusted R squared value of 0.5168. The analysis of variance table of that final model is shown in Table 3. Regarding the dependent variable y, no transformation was chosen because there was a lack of strong reason to choose another. As shown in the Box-Cox transformation (Figure 2) on the model containing all variables, the maximum log-likelihood is when lambda = 1, meaning y^1 is a sufficient value. However, anywhere from 0.5 to 1.5 seemed to yield a high log-likelihood value as well. Thus other models were tested using y^0.5 and y^1.5. All models resulted in the same significant variables E1, E2, and G20. Table 4 shows the resulting summaries for both y^0.5 and y^1.5. There was no significant change compared to y^1, so it was left as it was. Regarding the significant variables chosen, the third model from Table 1 was used because there was no significant increase in adjusted R squared nor was there a significant decrease in BIC from the third to fourth model. The third model consisted of E1, E2, E4, and G20. These variables were put into a regression model without their interaction terms, which resulted in only E1, E2, and G20 remaining significant at p value < 0.001, which is why E4 was discarded. In addition, the main effects model (Table 2) only resulted in E1, E2, and G20. E4 had a high p value as well as a low t value, so it was discarded. After deciding on only E1, E2, and G20, we tested for interactions between them. Both two-way interactions and three-way interactions were tested, and none of them resulted in any significance. Thus it resulted in the final fitted function Y = $\beta_0 + \beta_1 E_1 + \beta_2 E_2 + \beta_3 G_{20} + Z$.

Discussion

One potential issue that could affect the results is the transformation of the dependent variable. While ultimately no transformation was selected, there is a possibility that the original model did include a transformation that was not detected using the methods in this experiment. An argument could be made for using y^0.5, as the residual plot may have slightly less of a pattern compared to normal. It is difficult to judge the difference in pattern through just looking at it, so more testing is recommended. In addition to 0.5, any value between 0.5 and 1 may have been used, such as y^0.75 or y^0.9. There was no significant increase in adjusted r squared values between the tested options, so there was no clear “best model”.

Conclusion

The resulting final function implies that there are no significant environmental-environmental interactions, environmental-gene interactions, or gene-gene interactions. There are significant associations with two environmental variables E1 and E2. After the environmental variables were controlled, there was one significant genetic variable G20 associated with the dependent variable.

Appendix

Figure 1 - Residual Plot of Model including all environmental and genetic variables

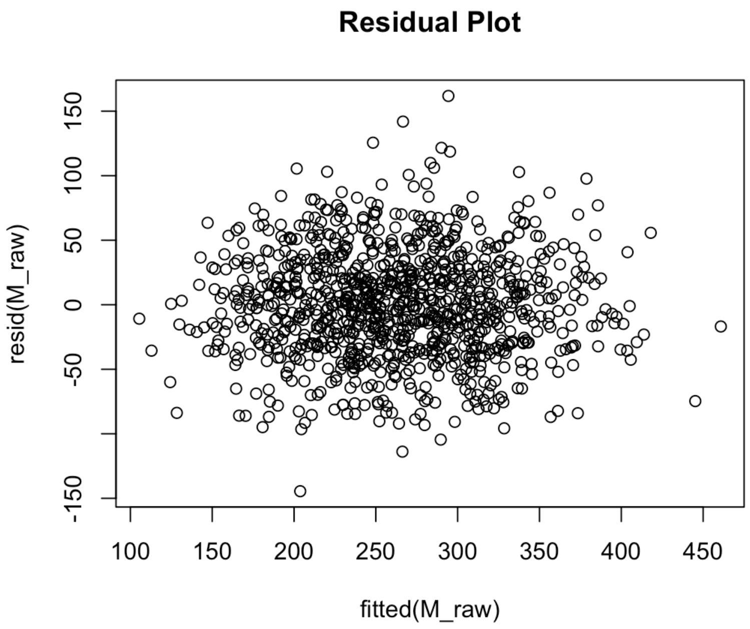


Figure 2 - Box-Cox transformation on raw model

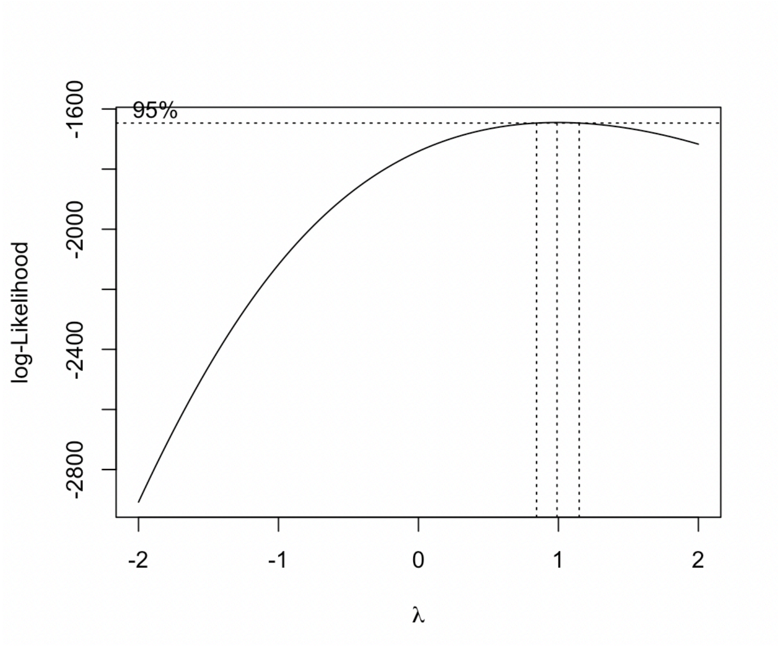
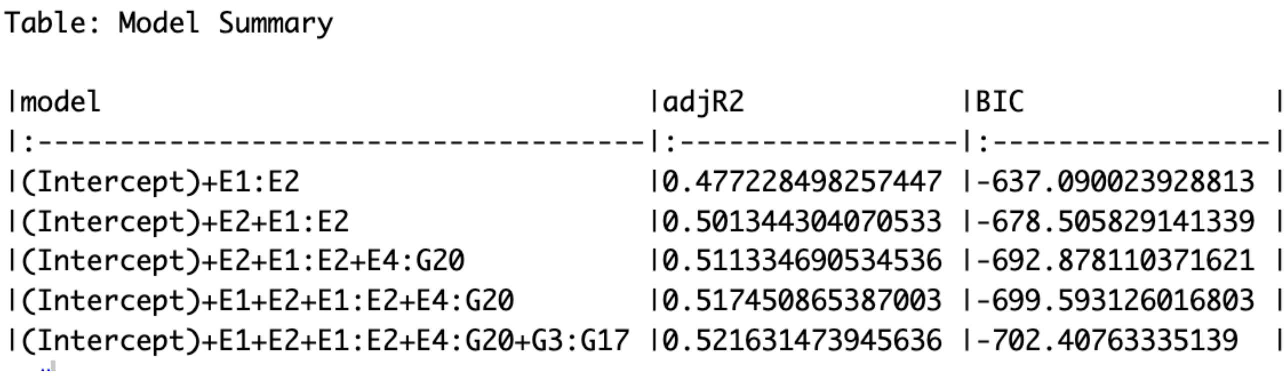
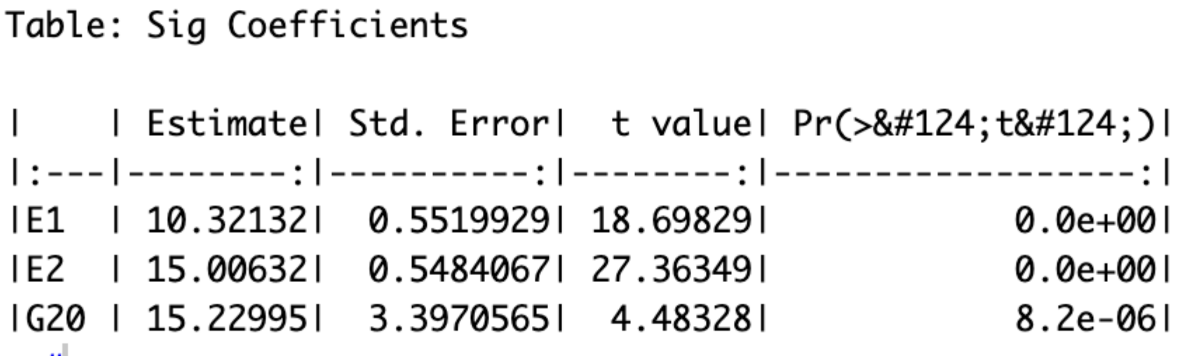


Table 1 - Model Summary (proposed models)



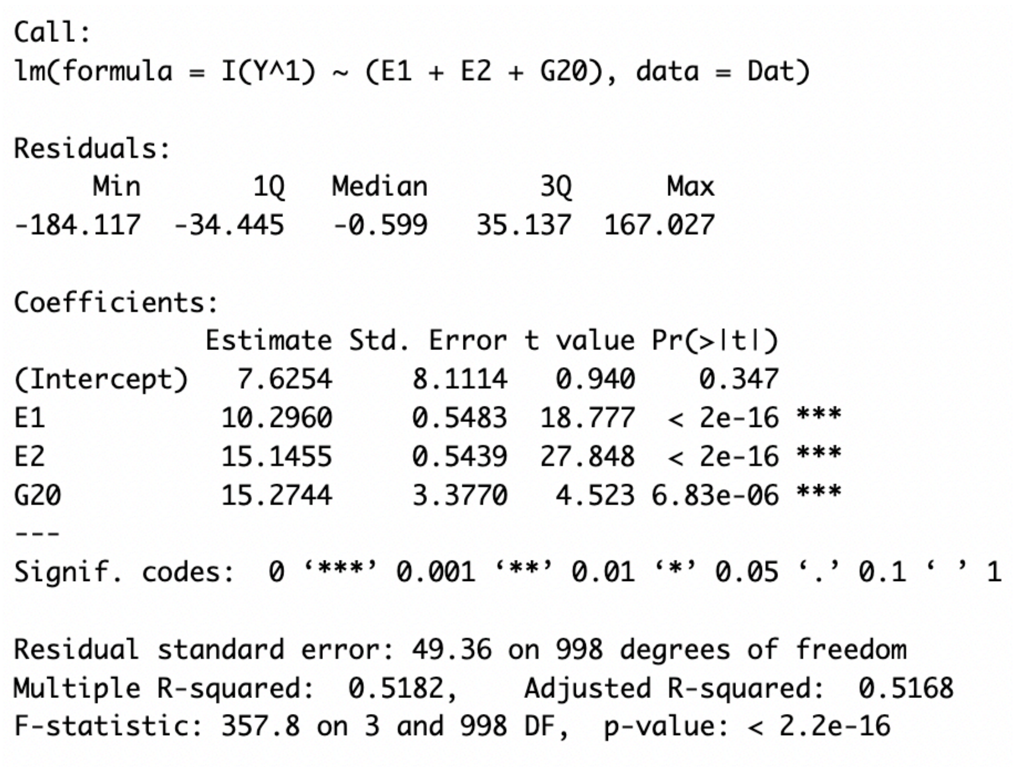
```
   Table: Model Summary

   |model                                 |adjR2             |BIC               |
   |:-------------------------------------|:-----------------|:-----------------|
1. |(Intercept)+E1:E2                     |0.477228498257447 |-637.090023928813 |
2. |(Intercept)+E2+E1:E2                  |0.501344304070533 |-678.505829141339 |
3. |(Intercept)+E2+E1:E2+E4:G20           |0.511334690534536 |-692.878110371621 |
4. |(Intercept)+E1+E2+E1:E2+E4:G20        |0.517450865387003 |-699.593126016803 |
5. |(Intercept)+E1+E2+E1:E2+E4:G20+G3:G17 |0.521631473945636 |-702.40763335139  |
```

Table 2 - Significant Main Effect



```
Table: Sig Coefficients

|    | Estimate| Std. Error|  t value| Pr(>&#124;t&#124;)|
|:---|--------:|----------:|--------:|------------------:|
|E1  | 10.32132|  0.5519929| 18.69829|            0.0e+00|
|E2  | 15.00632|  0.5484067| 27.36349|            0.0e+00|
|G20 | 15.22995|  3.3970565|  4.48328|            8.2e-06|
```

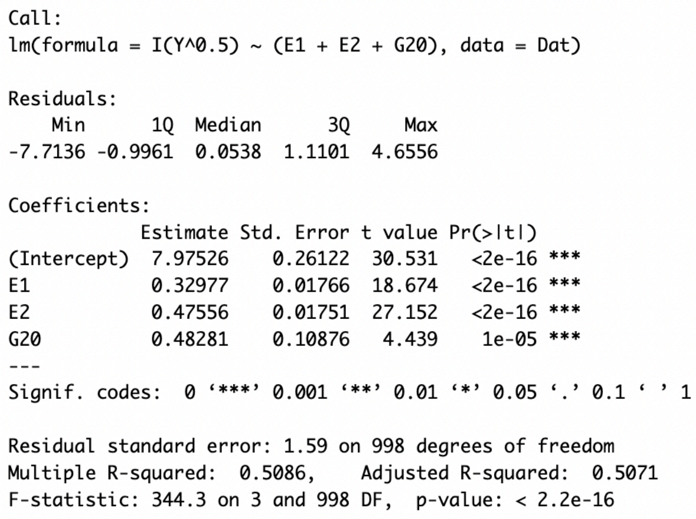
Table 3 - Final Model



```
Call:
lm(formula = I(Y^1) ~ (E1 + E2 + G20), data = Dat)

Residuals:
     Min       1Q   Median       3Q      Max 
-184.117  -34.445   -0.599   35.137  167.027 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   7.6254     8.1114   0.940    0.347    
E1           10.2960     0.5483  18.777  < 2e-16 ***
E2           15.1455     0.5439  27.848  < 2e-16 ***
G20          15.2744     3.3770   4.523 6.83e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 49.36 on 998 degrees of freedom
Multiple R-squared:  0.5182,	Adjusted R-squared:  0.5168 
F-statistic: 357.8 on 3 and 998 DF,  p-value: < 2.2e-16
```

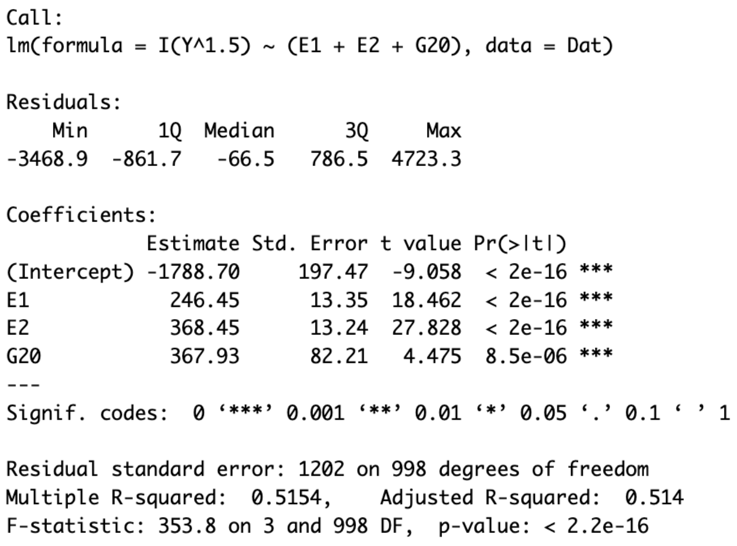
Table 4 - Comparing y^0.5 and y^1.5 respectively

y^0.5



```
Call:
lm(formula = I(Y^0.5) ~ (E1 + E2 + G20), data = Dat)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.7136 -0.9961  0.0538  1.1101  4.6556 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  7.97526    0.26122  30.531   <2e-16 ***
E1           0.32977    0.01766  18.674   <2e-16 ***
E2           0.47556    0.01751  27.152   <2e-16 ***
G20          0.48281    0.10876   4.439    1e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.59 on 998 degrees of freedom
Multiple R-squared:  0.5086,	Adjusted R-squared:  0.5071 
F-statistic: 344.3 on 3 and 998 DF,  p-value: < 2.2e-16
```

y^1.5



```
Call:
lm(formula = I(Y^1.5) ~ (E1 + E2 + G20), data = Dat)

Residuals:
    Min      1Q  Median      3Q     Max 
-3468.9  -861.7   -66.5   786.5  4723.3 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1788.70     197.47  -9.058  < 2e-16 ***
E1            246.45      13.35  18.462  < 2e-16 ***
E2            368.45      13.24  27.828  < 2e-16 ***
G20           367.93      82.21   4.475  8.5e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1202 on 998 degrees of freedom
Multiple R-squared:  0.5154,	Adjusted R-squared:  0.514 
F-statistic: 353.8 on 3 and 998 DF,  p-value: < 2.2e-16
```