Influence of Life Stress on Depression: Moderation by a Polymorphism in the 5-HTT Gene

Avshalom Caspi, et al.

Reported by **Shreyan Wankavala, SBU ID: 112634232**

Introduction:

The objective of the study was to test why stressful experiences lead to depression in some individuals but not in others. In the paper '*Influence of Life Stress on Depression: Moderation by a Polymorphism in the 5-HTT Gene*', written by Caspi et al., this question was examined by analyzing the behavioral genetics of certain individuals. The risk of depression after a stressful event is heightened among people with an elevated genetic risk and reduced for those at a lower genetic risk. What is not known however was whether or not specific genes are responsible for elevating or diminishing the effects of stressful events. We know that a number of environmental factors have an association with the outcome or dependent variable Y, but what we need to find is if there are associations of the Y with one or more of the genetic variables, once the environmental factors are controlled. This way, we can see if any of the genetic variables have a correlation to the findings proving or disproving that specific genes are responsible for elevating or diminishing the effects of stressful events. In my statistical research, my goal was to analyze a data set generated by a TA to find associations with genetic variables after the environmental variables had been controlled. These specific variables would then be used to estimate the function used by the TA to generate the data. The function would also point out which individual genetic variables were most significant in their associations with the outcome of the experiment.

Methods:

To analyze the data, I used RStudio. This is because it provided me with the methods I needed as well as with an easy to use interface. My first step was to begin modeling just the environment variables. This was done by assigning a variable to the entire data set, and then by assigning a model to just E1, E2, E3, and E4 with the dependent variable still being Y. The summary() function gave the adjusted R^2 value for this model, which was about 0.6494. After that, I switched to modeling just the genetic variables. This was done by assigning a model to just G1-G20. Using the same function, I found that the adjusted R^2 value for this model was about 0.7294. The genetic model was plotted using a plot function (seen in Figure 1), and it was seen that it could have been much more adequate, as in the shape of a "flat ellipse".

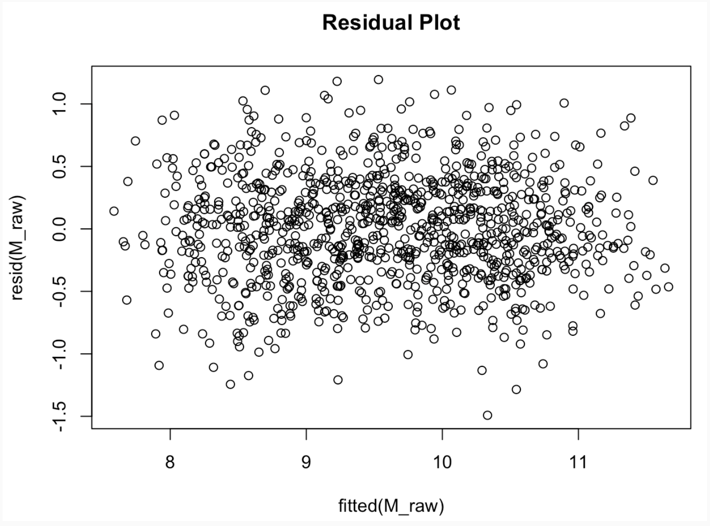
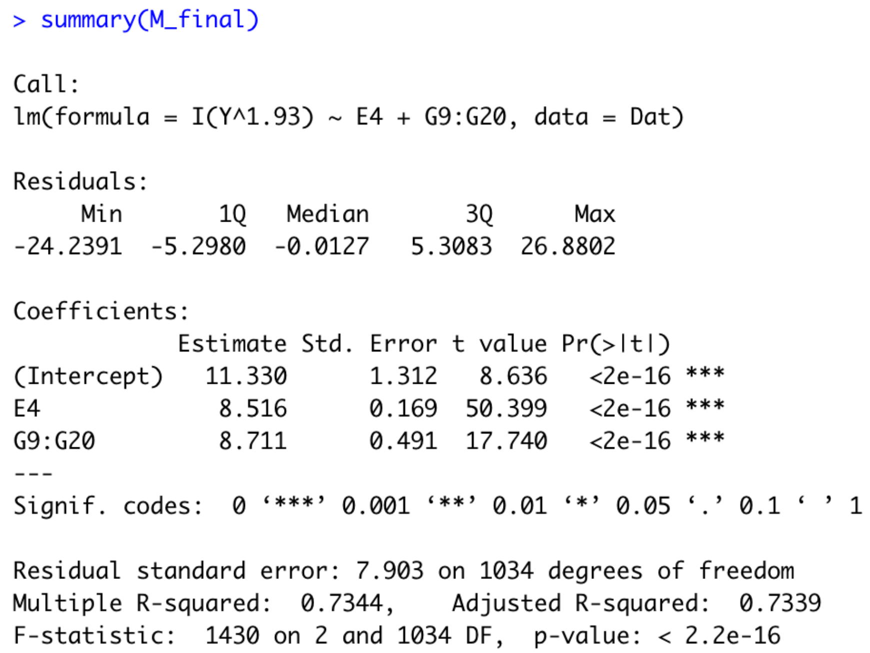


Figure 1

Transforming the dependent variable for this data set using the Box-Cox transformation would greatly improve my plot results, as well as increase my model's adjusted R^2 value. The boxcox() function was used from the MASS package in R which returned a graph. In this graph, I chose the number 1.93 for the λ value which maximized the log-likelihood function that was given. With this number, I then transformed the model and found a new adjusted R^2 value of about 0.732. The new plot also looked more like a flat ellipse, which showed that the new data was much more adequate than before. The next step was to use stepwise regression on the model. The reason for this was to see how significant the effects of the environmental and genetic variables were on the Y. The leaps() and regsubset() methods in R helped to propose models which I analyzed to see which variables had the greatest change in their adjusted R^2 values and Bayesian Information Criterion or BIC value. A significant coefficient summary was made from this model, and I saw that the variables E4, G9, and G20 were candidates to include in my final model. Finally, a summary of just these variables was made.



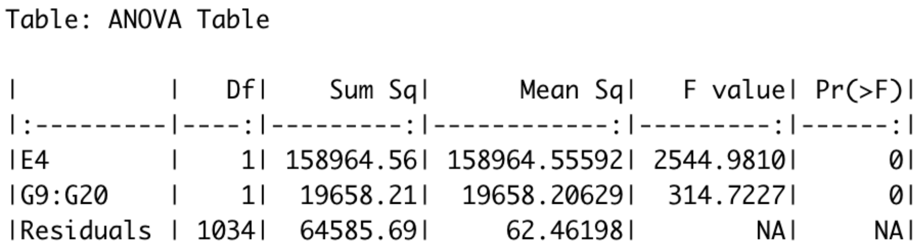
```
> summary(M_final)

Call:
lm(formula = I(Y^1.93) ~ E4 + G9:G20, data = Dat)

Residuals:
     Min       1Q   Median       3Q      Max 
-24.2391  -5.2980  -0.0127   5.3083  26.8802 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   11.330      1.312   8.636   <2e-16 ***
E4             8.516      0.169  50.399   <2e-16 ***
G9:G20         8.711      0.491  17.740   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.903 on 1034 degrees of freedom
Multiple R-squared:  0.7344,	Adjusted R-squared:  0.7339 
F-statistic:  1430 on 2 and 1034 DF,  p-value: < 2.2e-16
```

Figure 2

In Figure 2, we see that there are three stars on E4 and G9:G20. This means that these particular variables are significant, and should remain in my final model. They are also significant on the 0.001 level.

Results:

After performing my analysis of the data, I found an estimate of the function the TA used to generate the data set. This function is $Y^{1.93} = \beta_0 + \beta_1 E_4 + \beta_2 G_9 G_{20} + \varepsilon$. There is a transformation of Y to the power of 1.93 (as found by the boxcox function), as well as the inclusion of the intercept, E4, and G9:G20 as found by my models. My analysis found that there are in fact associations with the genetic variables after the environmental variables had been controlled. In addition, the G9:G20 interaction was significantly associated with the square root of the outcome variable (t-value 7.63). On top of this, the p-value of our final data remained less than 0.05, which shows that the independent and dependent variables were strongly correlated. The F-statistic is 1430 on 2 and 1034 degrees of freedom, and the residual standard error is 7.903 on 1034 degrees of freedom. Lastly, looking at Table 1, there is an ANOVA table for my



```
Table: ANOVA Table

|          |   Df|    Sum Sq|      Mean Sq|   F value| Pr(>F)|
|:---------|----:|---------:|------------:|---------:|------:|
|E4        |    1| 158964.56| 158964.55592| 2544.9810|      0|
|G9:G20    |    1|  19658.21|  19658.20629|  314.7227|      0|
|Residuals | 1034|  64585.69|     62.46198|        NA|     NA|
```

Table 1

final data values. The F value for the E4 variable was 2544.981 and 314.723 for G9:G20. To verify that my final model was correct, I put all of the variables I selected into a regression model without adding any of the other variables. The variables I put in were the same ones that remained significant at the 0.001 level and also had three stars next to them in the final model. In addition to this, all of the values chosen had an absolute t-value larger than 4, so there was no doubt as to whether or not any of them would be in the true model.

Conclusion and Discussion:

While in my data I did report that specific gene variables were significant in their effect on the dependent variable, this was indeed a false positive as was the case for Caspi et al. While the researchers confirmed that the results were calculated correctly, other papers of larger scale showed no relations that were reported by Caspi et al. as seen by Risch et al. 's work. Making a Type I error might be the result of the limitations of this particular procedure. According to the editorial '*Reporting Statistical Information in Medical Journal Articles'* by Peter Cummings, much of the information recorded can be omitted, while other parts serve little to no purpose in medical research like the topic of this study. For example, values like R^2 depend largely on the size of any biological effect of an exposure as well as the distribution of the exposure in the population. Because this information can be influenced by choice of study population and study design, it is not meaningful to include as it can be made smaller or larger by the researcher. In addition, it's highlighted what a study can really find if it is large enough, as seen by Caspi et al. where other larger studies do not find the same results that it did. It would be very interesting to see my results had more than just 4 environmental and 20 genetic variables were given for my independent variables. With what was given however, I was successful in my task of finding the original function used to generate the data set. Many other values were also successfully produced, like the R^2 value, F-values, and t-value for the final function.

Works Cited:

Influence of Life Stress on Depression: Moderation by a Polymorphism in ... https://www.science.org/doi/10.1126/science.1083968.

Peter Cummings, MD. “Reporting Statistical Information in Medical Journal Articles.” *Archives of Pediatrics & Adolescent Medicine*, JAMA Network, 1 Apr. 2003, https://jamanetwork.com/journals/jamapediatrics/fullarticle/481292.