Rahul Ranjan

AMS 315 Project 2

Professor Finch

Introduction

This project will focus on the topic of GxE interaction. This is an interaction between the environmental genetic variables that produce phenotypes of organisms. In this interaction, multiple genotypes respond to variation in the environment. It is known that most of this environmental variation is associated with the outcome variable Y. This is also the case for genetic variables and the outcome variable Y. However, the question to be answered is whether Y is associated with a single genetic variable or multiple genetic variables after controlling and conditioning for the environmental variables. I was given a few environmental variables and many genetic variables in an excel to determine whether there is an association with Y after conditioning on the environmental variables. (These variables had many unique vales) From here, I will determine, from the given data, the function used to generate the values for the outcome, environmental, and genetic variables.

Methodology

I performed all my statistical calculations through a coding platform known as R. Firstly, after reading the data from excel in R, I controlled for the environmental variables E1, E2, E3, and E4 by generating a regression model for them. Now since I have something to compare to, I generated a model for the four environmental variables and twenty genetic variables. The environmental and genetic variables were raised to the second power since I assumed up to 2nd interactions. I also generated a residual plot for this model. There was a notable increase in the adjusted R-squared value when I included all the genetic variables. The residual plot for the model containing environmental and genetic variables, appeared to closely resemble a “flat ellipse”. However, I was determined to find a higher R-squared value through a transformation of the data. I transformed the outcome variable by using the transformation log(Y). This generated a higher R-squared value along with a more pattern-less residual plot. (In depth analysis is in results section below) Next up, I performed stepwise regression on the transformed data. (The transformed data still followed the 2nd interaction principle) I generated a model summary that showed potential models/functions that were used to generate my data. The model with a large increase in R-squared value from the model on the previous row from it and the smallest decline in BIC value compared to the model on the next row, will be the one I choose. The variables in this model will be my “candidate variables” and I will make sure the significant main effects are a part of this model. From here, I generated a table showing all the significant coefficients. (These coefficients were shown on the table because they met the requirement of having a Pr(>|t|) < 0.001) If variables from the chosen model from before are not in this table, then they do not have significant main effect. I moved on to generating a table containing any 2nd interaction variables. This table was generated by setting a requirement for any 2nd interaction variables to also have Pr(>|t|) < 0.001. Any significant 2nd interaction variables will appear in this table and I recorded the estimate values. These vales will be used as numerical substitutes for β_i in out final model/function. This is the methodology I used to complete this project, and after careful analysis of my results, I chose my final model.

Results

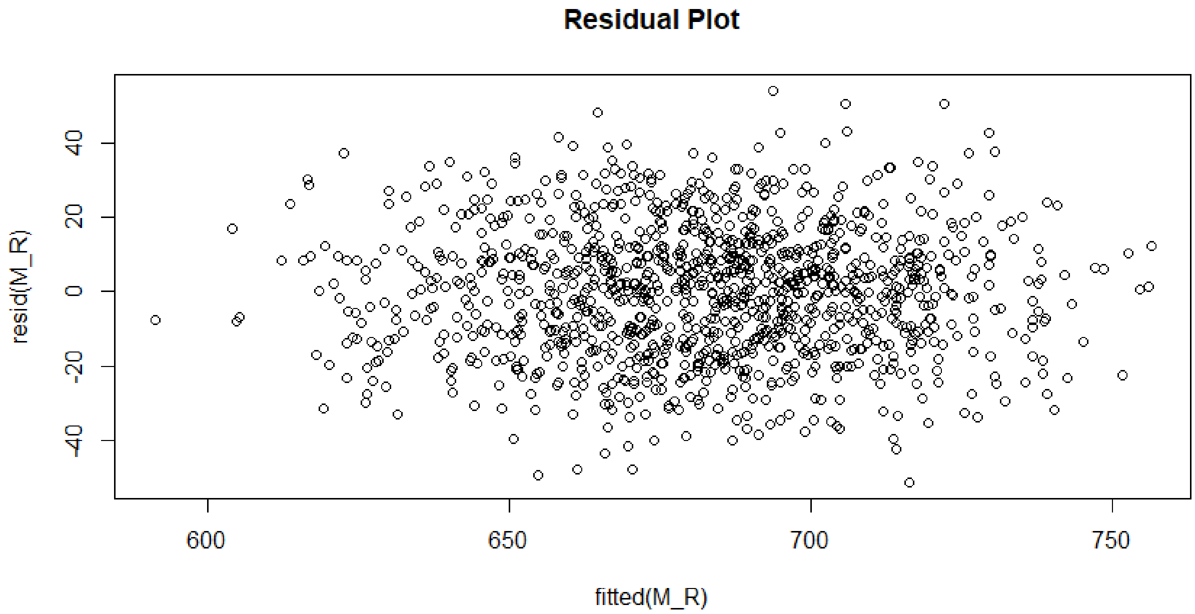
The total number of observations in the data set given to me was 1194. The adjusted R-squared value for the regression model containing the four environmental variables was 0.5174 and the p-value was 2.2 e-16. (Environmental variables were E1, E2, E3, E4 and Outcome variable was Y) The r value was 0.7193. This r-squared value tells us that 51.74% of the dependent variables' variance is explained by the independent variable. This value is moderate and could potentially be increased. The r value of 0.7283 tells us that there exists a moderately strong relationship between the dependent and independent variables. The adjusted R-squared value for the model containing all the environmental and genetic variables was 0.6149. (Environmental and genetic variables raised to 2nd power for 2nd interactions) This value is a notable increase from the previous R-squared value we had when controlling for the environmental variables. This shows that the twenty genetic variables that were included were significant. Now, 61.49% of the dependent variables' variance is explained by the independent variable. The residual plot for this model showed a pattern-less spread of the dots across the horizontal axis. It closely resembled a flat ellipse. I still decided to perform transformation on my outcome/dependent variable Y in order to potentially achieve a higher R-squared value for the data. I ended up choosing the log(Y) transformation since the transformations of $Y^2, Y^3,\ and\ y^{0.5}$ had slightly lower R-squared values upon generating models for them. The log transformation followed the same 2nd interaction procedure as the original data by raising the environmental and genetic variables to the 2nd power. The adjusted R-squared value for this transformed data was 0.6158 which is a slight increase from the original model, but nevertheless, an improvement. The residual plot for the transformed data was like the original residual plot, but the points were slightly more spread out from one another across the horizontal axis. The residual plot for this data was also slightly more pattern less and random then the other transformations I tried. Therefore, the lo(Y) transformation seems to be appropriate. Afterwards, I generated a model summary table for potential models that generated my data. I chose the model in the third row of this model summary table since it had one of the largest R-squared increases when compared to the other models in this summary. (0.0606 R-squared increase from 2nd model which was the second highest in this table) It also had the lowest decline in BIC when compared to other models in this summary. (15.3 BIC decline from 3rd model to 4th model) Therefore, my selected/candidate variables were E1, E2, E3, G5, and G16 after splitting the E1:E3, E2:E3, and G5:G16 interactions. (The model was initially (Intercept)+E1:E3+E2:E3+G5:G16 with an adjusted R-squared value 0.5676 and BIC of -975.7342) From here, I wanted to make sure that the main effects that are significant are in my final model. I created a table, named "Sig Coefficients" which only displayed the variables that were significant. (Significant variables had a Pr(>|t|) < 0.001) All the variables displayed in this table were also my candidate variables from above. This means that my candidate variables have a significant main effect. I went on and used the second power in the model request to discover 2nd order interactions. The table, named “2nd Interactions” only displayed the intercept along with E2. (The criteria to be met was also Pr(>|t|) < 0.001) Unfortunately, despite having different potential models as my final model, none of them seem appropriate. This is due to there being no interaction between any environmental and genetic variables. (Only E2 and intercept was displayed in “2nd Interactions” table) I looked at 2nd interactions again, but this time by altering the Pr(>|t|) < 0.001 to Pr(>|t|) < 0.01. There were actual 2nd interactions such as “G1:G8” and “G3:G10”, but these were not significant since the Pr(>|t|) was considerably higher than 0.001. For clarification, I inserted the significant/candidate variables into a regression model. They were all significant since they each had three stars in the model summary. Therefore, I decided to use the E1, E3, G5, and G16 estimates from the “Sig Coefficients” table and the intercept and E2 estimates from the “2nd Interactions” table in order to create my final model. My final model is $\log{(Y)} = \beta_0 + \beta_1 E1 + \beta_2 E2 + \beta_3 E3 + \beta_4 G5 + \beta_5 G16$. My final model with parameters is $\log(Y) = 3.9747280 + 0.0086437E1 + 0.0690439E2 + 0.0170842E3 + 0.0183732G5 + 0.0209264G16$.

Conclusions

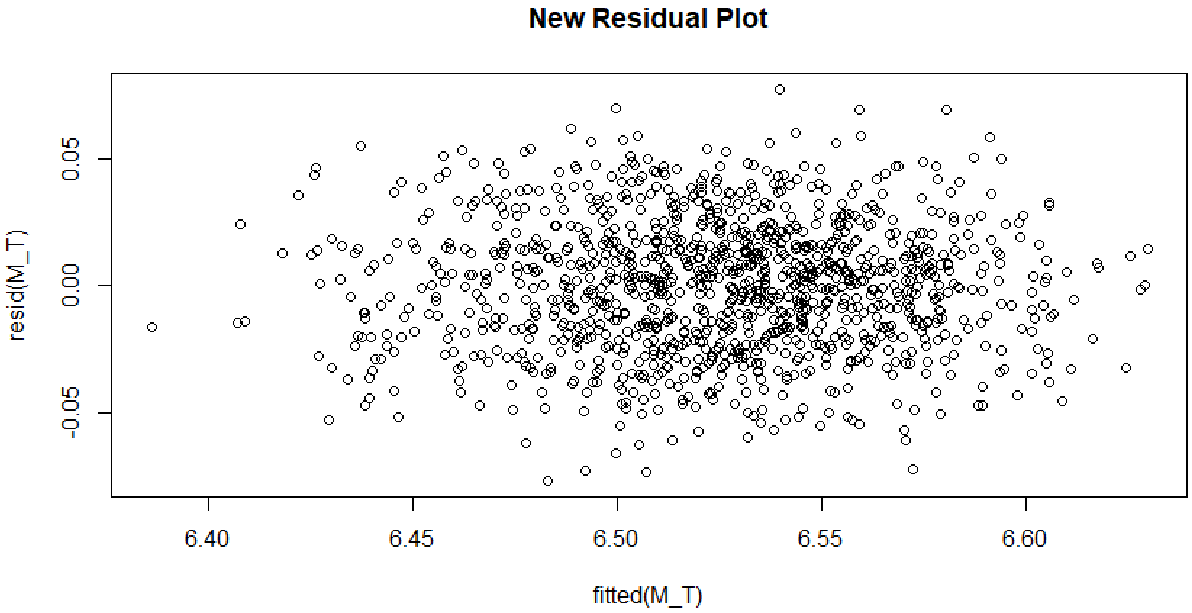
As stated in the introduction, it is known that environmental and genetic variables are associated with the outcome variable. The goal for this project was to discover any 2nd order interactions given the data, but none of them had a significant main effect. Consequentially, I had to rule them out of my true final model and this model contained separate environmental and genetic variables. It is quite a shame that things turned out this way, but I am confident that I just received data that did not support the possibility of interactions between environmental and genetic variables in association with the outcome variable. I am confident that data given to others most likely had 2nd order, and even 3rd order interactions between environmental and genetic variables.

Appendix

Residual Plot for Original Data (Y)



Residual Plot for Transformed Data (log(Y))



Model Summaries Table (Potential Models)

```
|model                                          |adjR2             |BIC               |
|:----------------------------------------------|:-----------------|:-----------------|
|(Intercept)+E2:E3                              |0.433243614034882 |-664.815041358357 |
|(Intercept)+E1:E3+E2:E3                        |0.506963836105872 |-825.112430225518 |
|(Intercept)+E1:E3+E2:E3+G5:G16                 |0.567604924581198 |-975.73424418182  |
|(Intercept)+E1:E3+E1:G16+E2:E3+G5:G16          |0.575621504291467 |-991.997432677435 |
|(Intercept)+E1:E3+E1:G16+E2:E3+E2:G5+G5:G16    |0.59401112402543  |-1038.81116680044 |
```

“Sig Coefficients” Table (Candidate Variables)

```
|            |  Estimate| Std. Error|   t value| Pr(>&#124;t&#124;)|
|:-----------|---------:|----------:|---------:|------------------:|
|(Intercept) | 5.6404262|  0.0274511| 205.47205|                  0|
|E1          | 0.0086437|  0.0006029|  14.33666|                  0|
|E2          | 0.0130394|  0.0006116|  21.31863|                  0|
|E3          | 0.0170842|  0.0006006|  28.44438|                  0|
|G5          | 0.0183732|  0.0017775|  10.33685|                  0|
|G16         | 0.0209264|  0.0017941|  11.66376|                  0|
```

“2nd Interactions” Table (2nd Order Interactions)

```
|            |  Estimate| Std. Error|  t value| Pr(>&#124;t&#124;)|
|:-----------|---------:|----------:|--------:|------------------:|
|(Intercept) | 3.9747280|  0.6067002| 6.551388|          0.0000000|
|E2          | 0.0690439|  0.0191767| 3.600407|          0.0003352|
```

R Code

```
> proj2 <- read.csv('P2_48179.csv', header = TRUE)
> M_E <- lm(Y ~ E1+E2+E3+E4, data = proj2)
> summary(M_E)
> M_R <- 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 = proj2 )
> plot(resid(M_R) ~ fitted(M_R), main ='Residual Plot')
> M_T <- lm( I(log(Y)) ~ (.)^2, data = proj2 )
> summary(M_R)$adj.r.square;
[1] 0.6149263
> summary(M_T)$adj.r.square
[1] 0.6158233
> plot(resid(M_T) ~ fitted(M_T), main ='New Residual Plot')
> library(leaps)
> Mod <- regsubsets( model.matrix(M_T)[,-1], I(log(proj2$Y)), nbest = 1 , nvmax=5, method = 'forward', intercept = TRUE )
> temp <- summary(Mod)
> Var <- colnames(model.matrix(M_T))
> Mod_S <- apply(temp$which, 1, function(x) paste0(Var[x], collapse='+'))
> library(knitr)
> kable(data.frame(cbind( model = Mod_S, adjR2 = temp$adjr2, BIC = temp$bic)), caption='Model Summary')
> M_M <- lm( I(log(Y)) ~ ., data = proj2)
> temp <- summary(M_M)
> kable(temp$coefficients[ abs(temp$coefficients[,4]) <= 0.001, ], caption ='Significant Coefficients')
> M_2nd <- lm( I(log(Y)) ~ (.)^2, data = proj2)
> temp  <- summary(M_2nd)
> kable(temp$coefficients[ abs(temp$coefficients[,4]) <= 0.001, ], caption='2nd Interaction')
> kable(temp$coefficients[ abs(temp$coefficients[,4]) <= 0.01, ], caption='2nd Interaction')
```

References

Caspi, A., Sugden, K., Moffitt, T., Taylor, A., Craig, I., Harrington, H., McClay, J., Mill, J., Martin, J., Braithwaite, A. and Poulton, R., 2003. *Influence Of Life Stress On Depression: Moderation By A Polymorphism In The 5-HTT Gene*. [online] Available at: <https://science.sciencemag.org/content/301/5631/386.full> [Accessed 25 April 2020].

En.wikipedia.org. 2020. *Gene–Environment Interaction*. [online] Available at: <https://en.wikipedia.org/wiki/Gene%E2%80%93environment_interaction> [Accessed 26 April 2020].

Heath, A., Phil, D. and Nelson, E., 2003. *Effects Of The Interaction Between Genotype And Environment Research Into The Genetic Epidemiology Of Alcohol Dependence*. [online] Available at: <https://pubs.niaaa.nih.gov/publications/arh26-3/193-201.htm> [Accessed 27 April 2020].

Risch, N., Herrell, R., Lehner, T., Liang, K., Eaves, L., Hoh, J., Griem, A., Kovacs, M., Ott, J. and Merikangas, K., 2009. *Interaction Between The Serotonin Transporter Gene (5-HTTLPR), Stressful Life Events, And Risk Of Depression*. [online] JAMA Network. Available at: <https://jamanetwork.com/journals/jama/article-abstract/184107> [Accessed 25 April 2020].

Statistics How To. 2015. *Residual Plot: Definition And Examples - Statistics How To*. [online] Available at: <https://www.statisticshowto.com/residual-plot/> [Accessed 27 April 2020].