Jingyan (Horatio) Xu AMS 315

Project 2 Report Due: 12/7/20

**Introduction**

The objective of this project is to find a linear model that correctly reflects our data. The project expands the concept from a paper by Caspi et al. that reports on the finding of a generic-environment interaction. Its use of multiple regression is a big part of our project. There are twenty-four independent variables with column names from E1 to E4 and G1 to G20, and unspecified rows of dependent variable Y. There are no missing values, so we won’t have to remove missing values like what we did for Project 1. While E1 to E4 are continuous and positive so they are simulation of environmental variables, G1 to G20 are indicator variables. The task is to find the model used by TAs that generated my data.

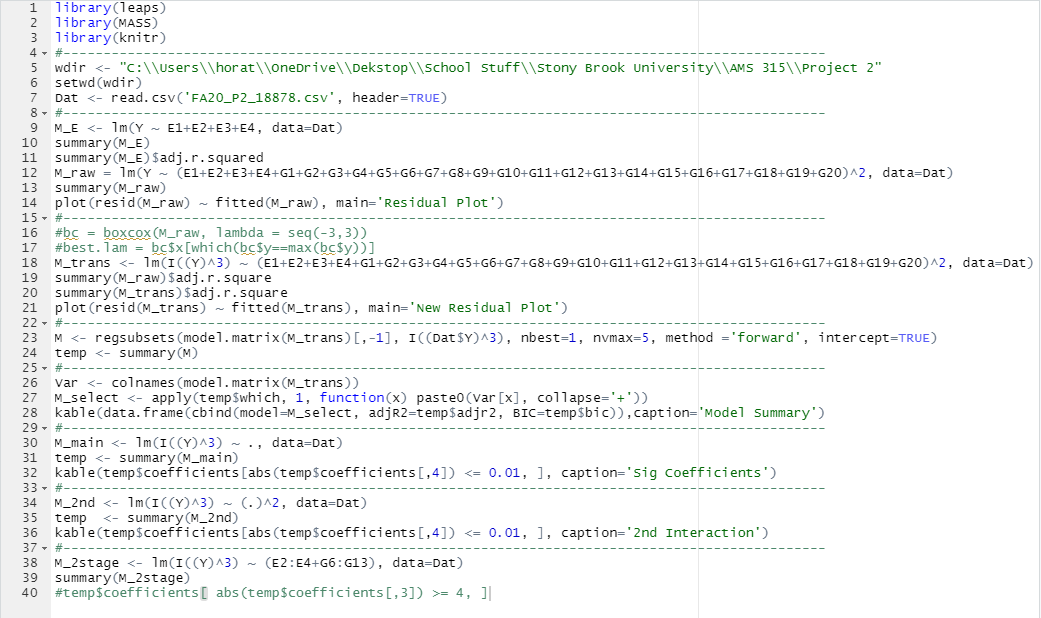
**Methodology**

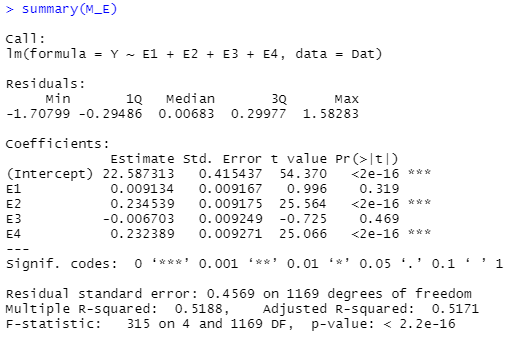
I chose to use R as my tool of statistical software. The data was processed in csv format so the first thing to do was to import the file into R. Next, I checked if there was a fit for model with only environmental variables. The adjusted R2 was 0.5171134 for this model. Additionally, I also tested a fit with the gene variables are also considered, with assumption that I only have up to 2nd order interaction in the model, and the resulted R2 for M\_raw is 0.5223124. I also plotted a residual plot to see if the model appears to be a flat ellipse. While my model did look quite adequate, I decided to do a transformation, nonetheless. I did not use box cox transformation because it doesn’t always give me the best result. I tried y2, y3, sqrt(y) and ln(y), and y3 gave the largest R2 result, which was the value I used to do transformation. The adjusted R2 was 0.5171134 when using just the environmental variables. The adjusted R2 with the genetic variables were also taken into the account is 0.5223124. The adjusted R2 after transformation was 0.5244032, which did not show a significant increase. This means that the model after the transformation did not necessarily have better properties than the model of the raw outcome variable. I next examined whether I could simplify the model and use stepwise regression. I attempted to find a regression model by forming a model summary. After thorough comparison, the 2nd model seemed to be the best model to use by assessing both the adjusted R2 and BIC values. In this case, there was an obvious increase in R2 from the 1st model to the 2nd model. The difference in R2 between the 2nd model and the 5th model, which had the largest adjusted R2 value, was only a little over 0.01, which I do not consider as a significant increase. In addition, the 2nd model has much less variables and thus make it easier to proceed with computation than other models. Although I have originally considered using the 4th model, as it was the best model to use since its BIC value was the most appropriate. However, not all the variables in the 3rd and 4th model have a significant main effect nor appear in the interaction table. The TAs and I had exhausted every method and concluded that the 4th model simply could not work. Consequently, the 2nd model became the next best option. Next, I wanted to also make sure main effects that are significant are in the model. After making a Significant Coefficients table using the command provided, I found that variable G6 and G13 do not have a significant main effect. I next consider the second order interactions by using the second power in the model request. Then, both variables appear in G6G13 interaction. The variables that I use as candidate variables for inclusion in my model are E2, E4, G6 and G13. After the 2nd stage process, I concluded that both E2:E4 and G6:G13 are quite likely to be in the model since they both have three stars. Therefore, the final model is Y3 = β0 + β1E2E4 + β2G6G13.

**Results**

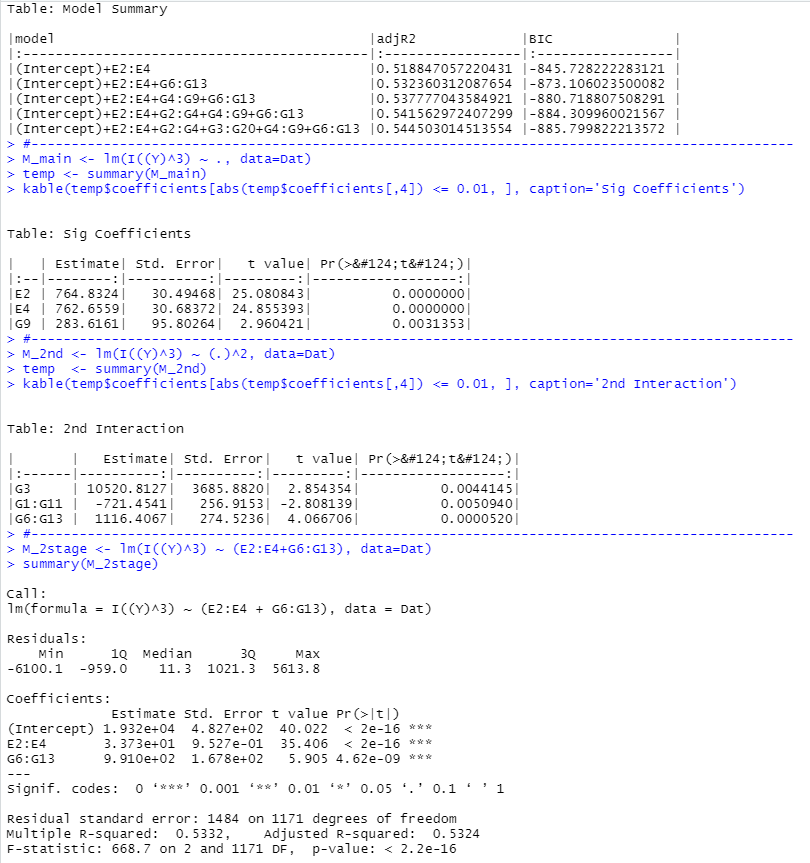
Based on the analysis of the data given, variables E2, E4, G6, and G13 are considered for the model. The final model with estimated parameter is: Y3 = 1.932\*104+ 33.73E2E4 + 991.0G6G13.

**Appendix**

****

****

****

****