Introduction

In this project, I was given a file that contained one dependent variable and twenty-four independent variables, each with 1236 observations. The dataset provided was complete. The independent variables were 4 environmental variables (E1 to E4) and 20 genetic variables (G1 to G20). The environmental variables are random variables and the genetic variables are 0-1 indicator variables. There are 80 gene-environment variables and 190 gene-gene interaction variables. There are a certain number of environmental variables that are associated with the outcome variable in my dataset. My objective is to use RStudio to discover if there are associations of the outcome variable with one or more of the genetic variables while controlling for the environmental variables, and to find the best association of independent variables to our dependent variable. The basis of our model comes from the paper by Caspi et al. that reports findings of gene-environment interactions. Similar to his project, we are using multiple regression models to analyze our synthetic data, in hopes to locate the association. When finding my results, I will be aware of the possibility of Type I or II errors. My research question is, are there any associations of Y with G, GxE, or GxG variables? Additionally, what is the function that the TA used to generate my data, including which GxE interactions?

Methods

After importing my dataset of independent and dependent variables onto RStudio, I modeled my data only using the environment variables as seen in Appendix 1. I retrieved the adjusted R squared. For the next linear model shown in Appendix 2, I included the genetic variables while controlling for the environmental variables to analyze their contribution. I plotted the residual plot to see if our model was adequate. After looking at the residual plot, our model didn't seem to be adequate, so then I used the Box-Cox transformation to help in transforming the dependent variable. After viewing the graph in Appendix 3, I chose a lambda, which I would use as the exponent for my transformation. I calculated the adjusted R squared of the raw original linear model and the transformed model in Appendix 4. I plotted the residuals, and as the plot resembled a flat ellipse more than the previous residual plot and was patternless, it was clear that the transformed model had better properties than the raw model. As shown in Appendix 5, I used the Leaps package, and proceeded to use the regsubset() to perform stepwise regression. Then, using the Knitr package seen in Appendix 6, I am able to retrieve proposed models to describe my data and their respective adjusted R squared values. Here I can see five good models that may represent my data. Then, by looking at the increases in adjusted R squared and changes in the BIC, we can pick the most optimal model. Afterwards, I made sure that any main effects that were significant are in the model, shown in Appendix 7. I found the significant variables while disregarding the variables interactions with each other.. In Appendix 8, I checked which variables had a t-value greater than 2, using the assumption that the model only has second order interactions. As seen in Appendix 9, I finally put our model together, and our goal was to find a better association than simply including all of the variables. After looking at the summary and ANOVA tables, if the adjusted R squared is higher, my model was correct.

Results

From the initial fitted linear model with only the environmental variables, the adjusted R squared was 0.3192173. After including all of the environmental and genetic variables, the adjusted R squared was 0.3399061. Using the Box-Cox transformation, I found lambda to be 0.3 and transformed the Y to Y^0.3. After this transformation, the adjusted R squared was 0.404001. The five models that I got from kable() from the Knitr package are seen in Appendix 6, and there you can see which variables are included and their adjusted R squared values. In the summary table, there is an obvious increase of approximately 0.0087 in the adjusted R squared from the first model to the second model. The other models have slighter increases in the adjusted R squared. Looking at the Bayesian Information Criterion (BIC), the decrease from the second to the third model is much smaller than the other decreases. Because of these observations, the best fit would look like $Y^{0.3} = \beta_0 + \beta_1 E4 + \beta_2 G8G13 + \epsilon$. I found any main effects that are significant in Appendix 7; the results are shown in the table. I changed the limit of the significance level from 0.001 to 0.03 with the objective of including all of the variables that I chose for this model, which meant that it showed me the variables where Pr(>|t|) was less than 0.03. In Appendix 8, you can see that E4 and G8:G13 had a t-value greater than 2. Finally, I created a model with E4 and G8:G13, and found that it had an adjusted R squared of 0.3832. This was higher than the original adjusted R squared of 0.3192173 when I just included the environmental variables in the model. Earlier I had to change the level of significance from 0.001 to 0.03 to accommodate for some genetic variables missing in the table. At the 0.001 level of significance, E4 was the only variable that remained. So, I decided to make sure that the model with only E4 wasn't optimal, and that I had chosen correctly when including the genetic variables. In Appendix 10, I completed the same steps as mentioned in the methods and as shown in Appendices 7 through 9. My final adjusted R squared for the E4 model was 0.3745, which was lower than if I were to include G8:G13 in my model. Thus, the final model that I propose for my dataset is $Y^{0.3} = 21.0739 + 5.3727E4 + 4.7738G8G13 + \epsilon$, where ϵ is the error term, and the estimates for the intercept and variables can be found in Appendix 9. In the ANOVA table for the final model, both F values are higher than $F_{0.01,1,1233}$, which is around 6.655. This means that the variables' coefficients aren't zero and they belong in the model.

Discussion/Conclusion

After testing to see which environmental, genetic variables, or interactions of variables fit my dataset the best, my final model is $Y^{0.3} = 21.0739 + 5.3727E4 + 4.7738G8G13 + \epsilon$. The goal of my project was to find the association of the independent variables with the dependent variables. I have found that there is an association between the dependent variable and E4, and the interaction of G8 and G13. In Appendix 9, I laid out my final model along with some tables to verify the strength of my model. As my final adjusted R squared value of 0.3832 was greater than the initial raw model adjusted R squared value of 0.3192173, my proposed model is definitely a more accurate model than the initial. I included ϵ in my final model to account for the possible statistical error. During statistical analysis, there is always the possibility of Type I or Type II error, either of which could skew the results in any way. My model is the accurate model, but it's important to account for the possible error term.

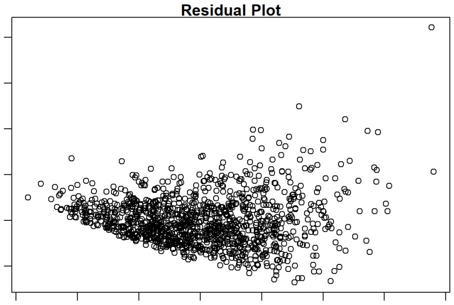
Appendix

Appendix 1 - Importing Data & Linear Model with Environmental Variables

```
library(readr)
Dataset <- read_csv("C:/Users/Angel/OneDrive/Desktop/P2_995021.csv")
View(Dataset)
M_E <- lm(Y ~ E1+E2+E3+E4, data=Dataset)
summary(M_E)
Call:
lm(formula = Y ~ E1 + E2 + E3 + E4, data = Dataset)

Residuals:
     Min       1Q   Median       3Q      Max 
-2576559  -793513  -207800   549651 10780993 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -2234220     341790  -6.537 9.18e-11 ***
E1            -10430      17957  -0.581    0.561    
E2              6749      17366   0.389    0.698    
E3             27221      17814   1.528    0.127    
E4            421503      17528  24.048  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1252000 on 1231 degrees of freedom
Multiple R-squared:  0.3214,	Adjusted R-squared:  0.3192 
F-statistic: 145.8 on 4 and 1231 DF,  p-value: < 2.2e-16
summary(M_E)$adj.r.squared
[1] 0.3192173
```

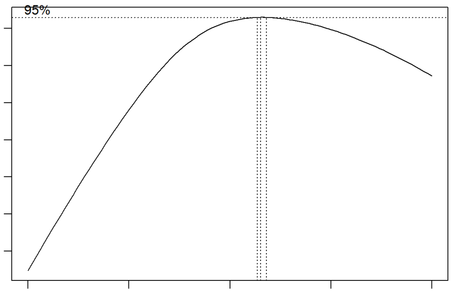
Appendix 2 - Model with all Independent Variables & Residual Plot

```
M_raw <- 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=Dataset)
par(mar=c(1,1,1,1))
plot(resid(M_raw) ~ fitted(M_raw), main='Residual Plot')
```



Appendix 3 - Using Box-Cox to choose a transformation

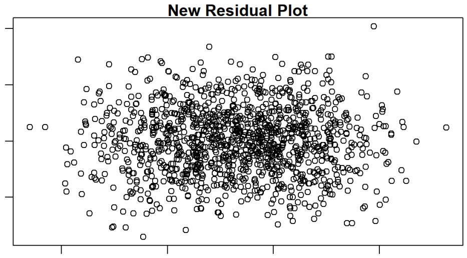
```
library(MASS)
boxcox(M_raw)
```



```
M_trans <- lm( I(Y^.3) ~ (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=Dataset)
```

Appendix 4 - Original and Transformed Model Adjusted R Squared Values & Residual Plot

```
summary(M_raw)$adj.r.square
[1] 0.3399061
summary(M_trans)$adj.r.square
[1] 0.404001
plot(resid(M_trans) ~ fitted(M_trans), main='New Residual Plot')
```



Appendix 5 - Leaps Package & Stepwise Regression

```
install.packages(“leaps”)
library(leaps)
M <- regsubsets( model.matrix(M_trans)[,-1], I((Dataset$Y)^.3), nbest = 1 , nvmax=5, method = 'forward', intercept = TRUE )
temp <- summary(M)
```

Appendix 6 - Knitr Package & Proposed Fitted Models

```
install.packages("knitr")
library(knitr)
Var <- colnames(model.matrix(M_trans))
M_select <- apply(temp$which, 1, function(x) paste0(Var[x], collapse='+'))
kable(data.frame(cbind( model = M_select, adjR2 = temp$adjr2, BIC = temp$bic)), caption='Model Summary')
```

```
Table: Model Summary

|model                                    |adjR2            |BIC              |
|:-----------------------------------------|:-----------------|:-----------------|
|(Intercept)+E4                           |0.374496965722079 |-566.692029125612 |
|(Intercept)+E4+G8:G13                    |0.383172388834281 |-577.837131008061 |
|(Intercept)+E4+E3:G2+G8:G13              |0.386017345630092 |-577.434250576287 |
|(Intercept)+E4+E3:G2+G2:G14+G8:G13       |0.388791072081885 |-576.914670160888 |
|(Intercept)+E4+E3:G2+G2:G14+G4:G14+G8:G13 |0.391539275432258 |-576.369512681074 |
```

Appendix 7 - Significant Main Effects

```
M_main <- lm( I(Y^.3) ~
E1+E2+E3+E4+G1+G2+G3+G4+G5+G6+G7+G8+G9+G10+G11+G12+G13+G14+G15+G16+G17+G18+G19+G
20, data=Dataset)
temp <- summary(M_main)
kable(temp$coefficients[ abs(temp$coefficients[,4]) <= 0.03, ], caption='Sig Coefficients')
```

```
Table: Sig Coefficients

|            |  Estimate| Std. Error|   t value| Pr(>&#124;t&#124;)|
|:-----------|---------:|----------:|---------:|------------------:|
|(Intercept) | 15.646622|  4.1396880|  3.779662|          0.0001647|
|E4          |  5.378011|  0.1981571| 27.140141|          0.0000000|
|G8          |  2.380077|  0.8221494|  2.894944|          0.0038603|
|G13         |  1.849405|  0.8300565|  2.228047|          0.0260605|
```

Appendix 8 - Significant Variables Included in Model

```
M_2stage <- lm( I(Y^.3) ~ (E4+G8+G13)^2, data=Dataset)
temp <- summary(M_2stage)
kable(temp$coefficients[ abs(temp$coefficients[,3]) >= 2, ])
```

```
|            |  Estimate| Std. Error|   t value| Pr(>&#124;t&#124;)|
|:-----------|---------:|----------:|---------:|------------------:|
|(Intercept) | 20.646644|  3.0421184|  6.786930|          0.0000000|
|E4          |  5.387877|  0.3109697| 17.326053|          0.0000000|
|G8:G13      |  4.114844|  1.6706623|  2.463001|          0.0139142|
```

Appendix 9 - Final Model with Summary Table and ANOVA Table

```
M_final <- lm(I(Y^.3) ~ E4 + G8:G13, data = Dataset)
summary(M_final)
Call:
lm(formula = I(Y^0.3) ~ E4 + G8:G13, data = Dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-41.377  -9.620  -0.238   9.392  44.946

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  21.0739     1.9205  10.973  < 2e-16 ***
E4            5.3727     0.1964  27.358  < 2e-16 ***
G8:G13        4.7738     1.1142   4.284 1.98e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.04 on 1233 degrees of freedom
Multiple R-squared:  0.3842,	Adjusted R-squared:  0.3832
F-statistic: 384.6 on 2 and 1233 DF,  p-value: < 2.2e-16
anova(M_final)
Analysis of Variance Table

Response: I(Y^0.3)
            Df Sum Sq Mean Sq F value    Pr(>F)
E4           1 147931  147931 750.824 < 2.2e-16 ***
G8:G13       1   3617    3617  18.356 1.975e-05 ***
Residuals 1233 242931     197
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Appendix 10 - Appendices 7-9 Repeated, Only Using E4 in the Model 0.001 Sig Level

```
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(>&#124;t&#124;)|
|:-----------|---------:|----------:|---------:|------------------:|
|(Intercept) | 15.646622|  4.1396880|  3.779662|          0.0001647|
|E4          |  5.378011|  0.1981571| 27.140141|          0.0000000|


M_2stage <- lm( I(Y^.3) ~ (E4+G8+G13)^2, data=Dataset)
temp <- summary(M_2stage)
kable(temp$coefficients[ abs(temp$coefficients[,3]) >= 2, ])
```

Appendix 10 Continued - Appendices 7-9 Repeated, Only Using E4 in the Model 0.001 Sig Level

```
|           |  Estimate| Std. Error|  t value| Pr(>&#124;t&#124;)|
|:----------|---------:|----------:|--------:|------------------:|
|(Intercept) | 21.718405|  1.9280076| 11.26469|                 0|
|E4         |  5.380854|  0.1977492| 27.21049|                 0|
M_final <- lm(I(Y^.3) ~ E4, data = Dataset)
summary(M_final)

Call:
lm(formula = I(Y^0.3) ~ E4, data = Dataset)

Residuals:
    Min      1Q  Median      3Q     Max 
-40.771  -9.713   0.270   9.406  48.971 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  21.7184     1.9280   11.27   <2e-16 ***
E4            5.3809     0.1977   27.21   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.13 on 1234 degrees of freedom
Multiple R-squared:  0.375,	Adjusted R-squared:  0.3745 
F-statistic: 740.4 on 1 and 1234 DF,  p-value: < 2.2e-16
```