Introduction

In this project, we seeked to determine whether if we controlled for the 4 environmental variables, there would be an association between the 20 genetic variables and the outcome variable. To do so, we first isolated the environmental variables and created a model. Then, we seeked to determine the contribution of genetic variables.

The background of this research is based on gene-gene and gene-environment interactions and the effect of genes/environment on outcomes. We are using quantitative analysis through the use of regression models to explore the various relationships relating to this.

We found Y is associated with G variables $G5$ and G17.

We also found there are no associations of Y with G x G variables. There is an association of Y with the G x E variables G5 x E4.

Methods

We used RStudio to help generate all our models in this project. Firstly, we fitted a model using only the environmental variables to explain the outcome. Table 1 shows the output of this. Then, after controlling for the environmental variables, we assessed for the genetic variables. We modeled all interaction terms up to the second order, created a residual plot (Table 2) and modeled a box-cox transformation (Table 3). Our box-cox model helped us generate our λ value by determining the λ value where the log-Likelihood value reaches its peak. In my case, it was λ=0.5.

Then, I calculated the adjusted R^2 value after the transformation (Table 4) and created a New Residual Plot with the transformation (Table 5). I then performed a stepwise regression and generated a Model Summary (Table 6). This gave me 5 different possible models I could use to eventually help give me my final model. The goal is to determine what variables and interactions have a main effect that should be used in the final model. Then, I used information from Table 7 to confirm that the variables we chose from the model summary all have a significant main effect on the data by analyzing the p-values. Then, I used model 8 to help generate my final model, telling me about the interactions and variables used at a minimum t-value being the absolute value of 1, also looking at p-values. I then created my final model and performed an analysis of it (Table 9).

I didn't have any missing data in my data file. However, if there was missing data, I would use the data that was given to create the best possible multiple regression model, which I would then use to fill in the missing data.

We considered all interactions up to the 2nd order in our experiment. The specific interaction we used is between E4:G5, which is associated with the square root of the outcome variable.

Multiple comparison issues can occur when we perform multiple statistical tests on the same data set or when we compare multiple groups. This would increase the chances of making a type I error (false positives). To combat this, we used Bonferroni correction, which adjusts the significance level (α) by dividing it by the number of comparisons that we make. This would maintain the probability of a Type I error when conducting multiple tests.

Results

Our final model is $Y^{1/2} = \beta_0 + \beta_1 G_{17} + \beta_2 E_2 + \beta_3 E_4 + \beta_4 E_4 G_5 + \epsilon$. The G17, E2, E4, and G5 variables are associated with Y. We found the λ value of $\frac{1}{2}$ through the use of the box-cox model. This can be confirmed in the New Residual Plot (Table 5), where we have a flat ellipse, while we didn't in the original one (Table 2). After this transformation, our adjusted R^2 value is approximately 0.51. From our model summary (Table 6), we can conclude that the fourth model is the best because it has a low BIC, and while its BIC isn't quite as low as the fifth, the decrease from the fourth to the fifth is significantly smaller than the other decreases. Then, when we estimate the p-values of the variables in that model ($E2$, $E4$, $G5$, $G17$]) in Table 7, we find that all the p-values are close to 0. This confirms that the variables we chose from the model summary all have a significant main effect on the data and those are the variables that we use as candidate variables for inclusion in our model. Figure 8 shows that E2 and E4 have the strongest relationship in the model b/c their P-values are the only ones less than an alpha value of 0.05. The variables and interactions from table 8 is what we will use to come up with our final model of $Y^{1/2} = \beta_0 + \beta_1 G_{17} + \beta_2 E_2 + \beta_3 E_4 + \beta_4 E_4 G_5 + \epsilon$. In our analysis of our final model, we see that all the Pr(>|t|) values are less than an alpha value 0.05, so this confirms that the final model is good.

Discussion and Conclusions

Overall, we did this experiment to come up with our final model that best models the data given, involving genetic variables, environmental variables, and an outcome variable.
We found that there are associations between the outcome variable and certain genetic variables (G17 and G5) after controlling for the environmental variables.
A possible limitation to the experiment include model assumptions. We assumed a linear relationship between the variables, whereas possibly a more complex non-linear model could have been better for my data set.
Another limitation could be the presence of "noise" in the data, where the model captures fluctuations in the sample data set that may not generalize well in the overall population.

Appendix

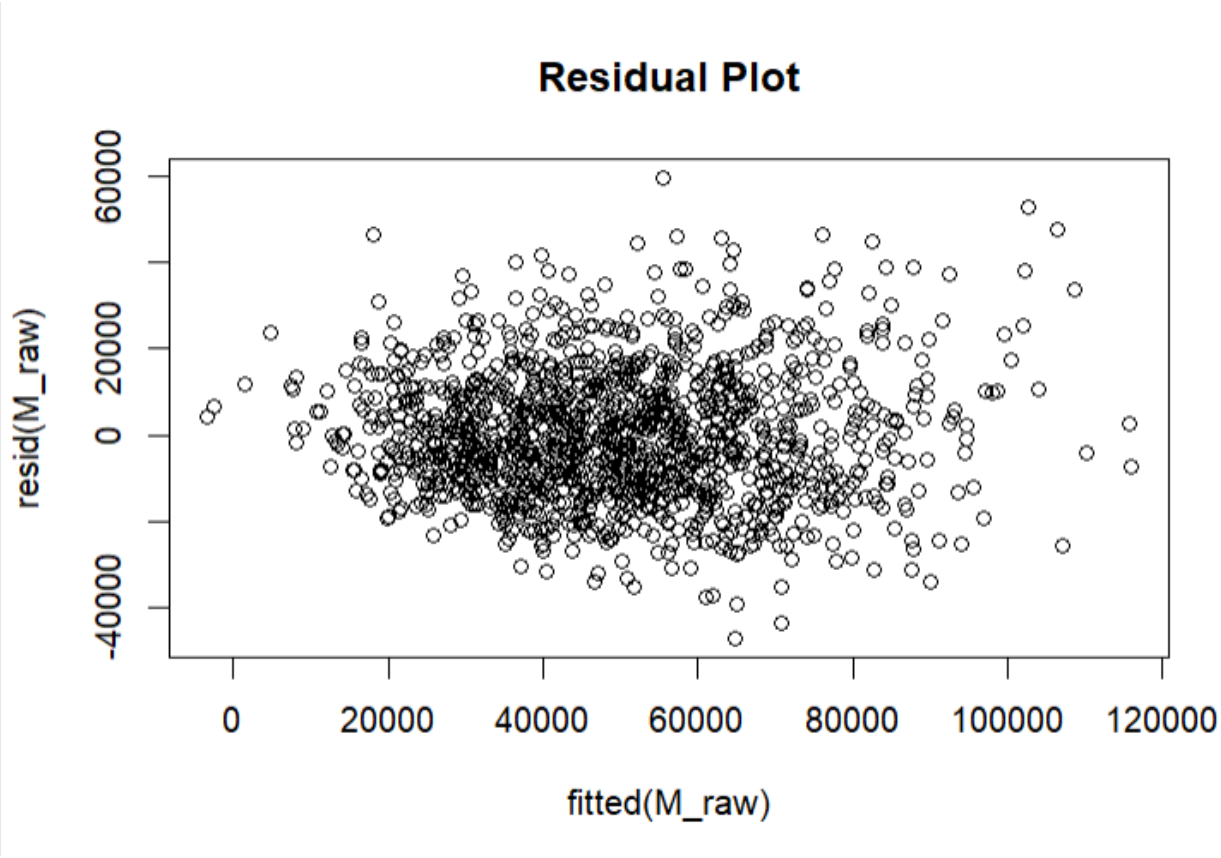
Code:

```
setwd ("C:/Users/arjun/OneDrive/Desktop/AMS 315 Project 2")
getwd()
Dat<-read.csv("P2_602237.csv")
M_E <- lm(Y ~ E1+E2+E3+E4, data=Dat)
summary(M_E)
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=Dat)
plot(resid(M_raw) ~ fitted(M_raw), main='Residual Plot')
library(MASS)
boxcox(M_raw)
M_trans <- lm( I(Y^.5) ~
(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=Dat )
summary(M_raw)$adj.r.square
summary(M_trans)$adj.r.square
plot(resid(M_trans) ~ fitted(M_trans), main='New Residual Plot')
#install.packages("leaps")
library(leaps)
M <- regsubsets( model.matrix(M_trans)[,-1], I((Dat$Y)^.5),
                 nbest = 1 , nvmax=5,
                 method = 'forward', intercept = TRUE )
temp <- summary(M)
#install.packages("knitr")
library(knitr)
Var <- colnames(model.matrix(M_trans))
M_select <- apply(temp$which, 1,
                  function(x) paste0(Var[x], collapse='+'))
kable(data.frame(cbind( model = M_select, adjR2 = temp$adjr2, BIC = temp$bic)),
     caption='Model Summary')
M_main <- lm( I(Y^.5) ~
E1+E2+E3+E4+G1+G2+G3+G4+G5+G6+G7+G8+G9+G10+G11+G12+G13+G14+G15+G16+
G17+G18+G19+G20, data=Dat)
temp <- summary(M_main)
kable(temp$coefficients[ abs(temp$coefficients[,4]) <= 0.001, ], caption='Sig Coefficients')
M_2stage <- lm( I(Y^.5) ~ (G17+E2+E4+G5)^2, data=Dat)
```

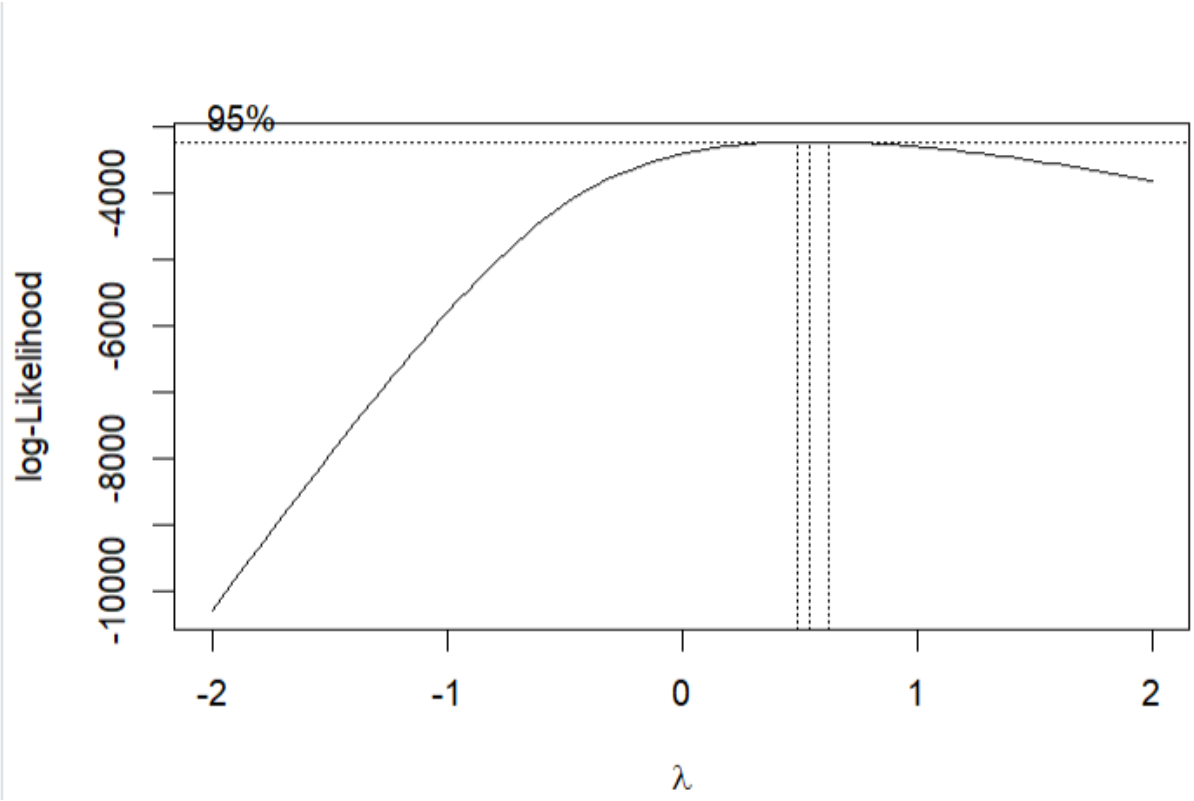
```
temp <- summary(M_2stage)
kable(temp$coefficients[ abs(temp$coefficients[,3]) >= 1, ])
M_final <- lm(I(Y^.5) ~ G17+E2+E4+E4:G5, data = Dat)
summary(M_final)
```

```
Call:
lm(formula = Y ~ E1 + E2 + E3 + E4, data = Dat)

Residuals:
   Min     1Q Median     3Q    Max
-45920 -12881  -2246  11372  83269

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -30124.96    3639.05  -8.278 3.07e-16 ***
E1             92.82     175.15   0.530    0.596
E2           4128.64     175.69  23.499  < 2e-16 ***
E3            -56.94     174.92  -0.326    0.745
E4           3845.43     176.76  21.755  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18310 on 1299 degrees of freedom
Multiple R-squared:  0.4339,    Adjusted R-squared:  0.4321
F-statistic: 248.9 on 4 and 1299 DF,  p-value: < 2.2e-16
```

- Table 1


Residual Plot
resid(M_raw)
60000
20000
0
-40000
0
20000
40000
60000
80000
100000
120000
fitted(M_raw)

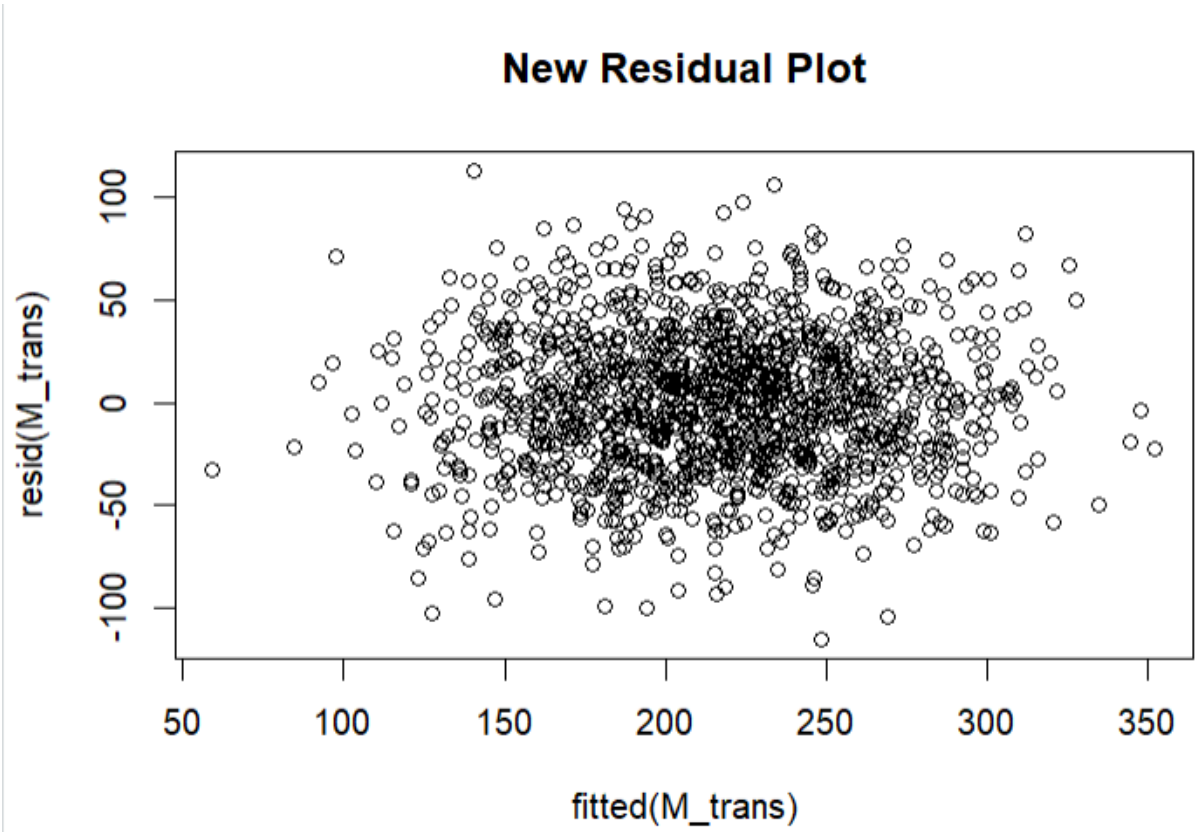
- Table 2



- Table 3

```
> summary(M_raw)$adj.r.square
[1] 0.5048645
> summary(M_trans)$adj.r.square
[1] 0.5132216
```

- Table 4



-

- Table 5

Table: Model Summary

model	adjR2	BIC
(Intercept)+E2:E4	0.423472289939325	-704.80913242359
(Intercept)+E2:E4+G5:G17	0.463586730756189	-792.68040630065
(Intercept)+E2:E4+E2:G5+G5:G17	0.467790113394727	-796.768413193033
(Intercept)+G17+E2:E4+E2:G5+G5:G17	0.477237549918984	-813.954342397763
(Intercept)+G17+E2:E4+E2:G5+G4:G11+G5:G17	0.480303223290919	-815.455038082408

- Table 6

Table: Sig Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
E2	9.616845	0.3869813	24.850927	0
E4	8.904064	0.3905407	22.799326	0
G5	17.486394	2.2856867	7.650390	0
G17	17.446541	2.2661562	7.698737	0

- Table 7

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	23.165613	16.8409344	1.375554	0.1691980
G17	15.626911	11.6026523	1.346839	0.1782679
E2	8.648913	1.4926208	5.794448	0.0000000
E4	8.360865	1.5759046	5.305439	0.0000001
E4:G5	1.080063	0.7869401	1.372485	0.1701506

- Table 8

```
Call:
lm(formula = I(Y^0.5) ~ G17 + E2 + E4 + E4:G5, data = Dat)

Residuals:
     Min       1Q   Median       3Q      Max
-144.187  -27.536    0.116   26.372  132.869

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  20.1373     5.7480   3.503 0.000475 ***
G17          17.6917     2.2452   7.880 6.90e-15 ***
E2            9.6922     0.3837  25.260  < 2e-16 ***
E4            7.9073     0.4034  19.603  < 2e-16 ***
E4:G5         1.6854     0.2202   7.653 3.81e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 39.96 on 1299 degrees of freedom
Multiple R-squared:  0.4942,    Adjusted R-squared:  0.4927
F-statistic: 317.4 on 4 and 1299 DF,  p-value: < 2.2e-16
```

- Table 9

Reference:

Multiple Regression Handout

-