Data Analysis - Multiple Regression
Bushra Paracha

Introduction: This assignment assigns us with a data file that we use to observe the GxE interactions as given by Caspi et al. and others. The observations on the interactions can then provide information on how genotypes respond to specific environmental variations differently. Depending on the data given, a transformation may also be required in order to adequately achieve an accurate function.

Methods: For this assignment R was used to analyze the data given. Using summary() function first a summary of the data was created. Then, to find more specific values such as adjusted *R*^2 Value, we used summary()$adj.r.squared. We also use lm() to evaluate the genetic variables up to the 2nd order interactions. After creating the model, we use plot() to create a residual plot of these interactions. We will need to do a transformation on the dependent variables if the residual plot does not display a flat eclipse. For the transformation we will use the boxcox() function and when performing regression we will first install the leaps package in order to utilize its regsubset() function. Afterwards we will install the knitr package in order to produce a model summary which includes the adjusted R value and the BIC values. In order to pinpoint which independent variables could possibly be used in the final model we observe the changes in the adjusted R^2 values. When it comes to choosing significant variables, we can use the Kable() function to produce a table that allows us to see which variables have a significant main effect for the function.After choosing the necessary variables, we can generate the final model's summary and ANOVA table using the pureErrorAnova() function.

Results:

After using the summary() function it was found that the data assigned had 4 environment variables, 20 genetic variables and one dependent variable. The data did not have any missing values and the R^2 value is approximately 0.3399061. This value shows that the independent variables do not account for much variation. After finding summary statistics, we integrate the genetic variables and ensure that we only have up to 2nd-order interactions to make sure we generate a residual plot of the variables and show that we need to apply a transformation to a dependent variable due to its heteroscedastic residuals. We then use Box-Cox transformation and notice that λ= 0.5, which means that our dependent variable would benefit from applying the square root transformation to it. After this, we get a more acceptable residual plot (Figure 1). After getting the residual plot we find out that the adjusted R value for the transformed data is 0.3951426. Afterwards, we look at using stepwise regression to construct a model summary to observe which variables could be included in the final function. (Table 1). Based on table 1 starting with a single variable E4, the model achieves an adjR2 of 0.3689. Adding the interaction term G8:G13, the adjusted R2 increases to 0.3780, suggesting this interaction contributes meaningfully to explaining the variability in the dependent variable. Further inclusion of E3:G2, G2:G14, and G4:G14 results in incremental improvements, with the highest adjR2 of 0.3867 for the most complex model. This trend indicates that the inclusion of additional terms is improving the model, but the rate of improvement diminishes with added complexity. we can see that the first-to-second model shows a noticeable increase in the adjusted R values. It is also important to note that there is an increase from the second model to the fifth model. However, it is still unclear whether they are significant or not. Another thing to note is the BIC values. Lower BIC values indicate a better model, considering both fit and simplicity. The BIC starts at -555.81 for the simplest model and decreases as more predictors and interaction terms are added. The model with the lowest BIC is the one including (Intercept)+E4+G8:G13, with a value of -567.49. While the most complex model slightly improves adjR2, its BIC increases to -566.66, suggesting it is overfitted compared to the simpler model (Intercept)+E4+G8:G13. The BIC values gradually decreased, with an exception from the third model to the fourth model. After much deliberation the second model is recommended due to its optimal balance of model fit and complexity, as indicated by the lowest BIC. The interaction term G8:G13 is particularly important in explaining the variability in the dependent variable. The variables chosen are E4, G8 and G13. A table can be further created to select the rows where the p-values are either less than or equal to 0.001, indicating high significance (Table 2). From Table 2, we can see that all of the variables described earlier have high significance in the model. Thus, the final model that was decided on was: $Y1/2 = \beta0 + \beta1E4 + \beta2G8G13+ e$. Based on Figure 2, this final function produced an adjusted R value of about 0.378. On Table 3, the ANOVA table also indicates that 2 all the variables chosen are highly significant predictors in the model since all their p-values are less than 0.001. The G8:G13 interaction was significantly associated with the transformed dependent variable.

Conclusion:

While we were able to construct an estimated function for the dataset, there were some limitations along the way. For one, the adjusted R value was moderately low. This indicates that the model only explains a modest portion of the variability in the dependent variable. Other important predictors or interactions may have been omitted. Another limitation was selecting the significant variables that would end up in the final model. If we changed the values of what was considered significant, other predictor variables could have been eligible to be included in the model function. Furthermore, if we happen to overlook a variable, it could potentially lead to imprecise results.

Figure 1. Plotted graph of the transformed residual graph

```
> plot(resid(M_trans) ~ fitted(M_trans), main='New Residual Plot')
```

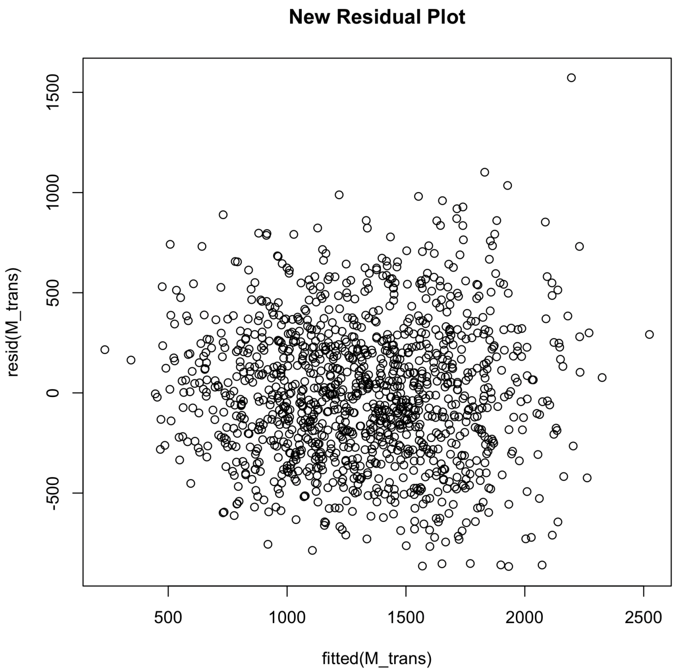


Figure 2. Boxcox plot

```
> boxcox(M_raw)
```

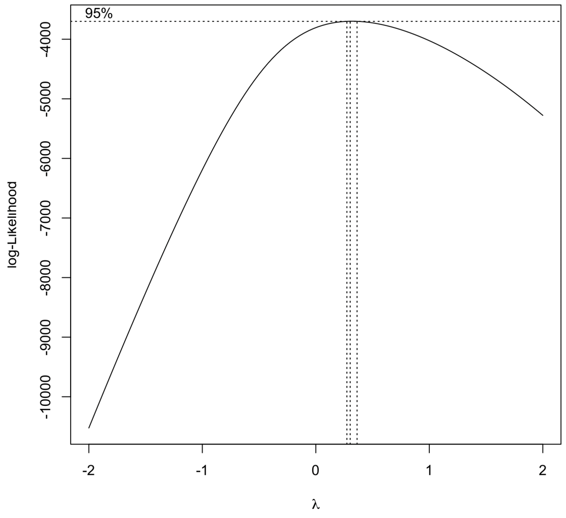
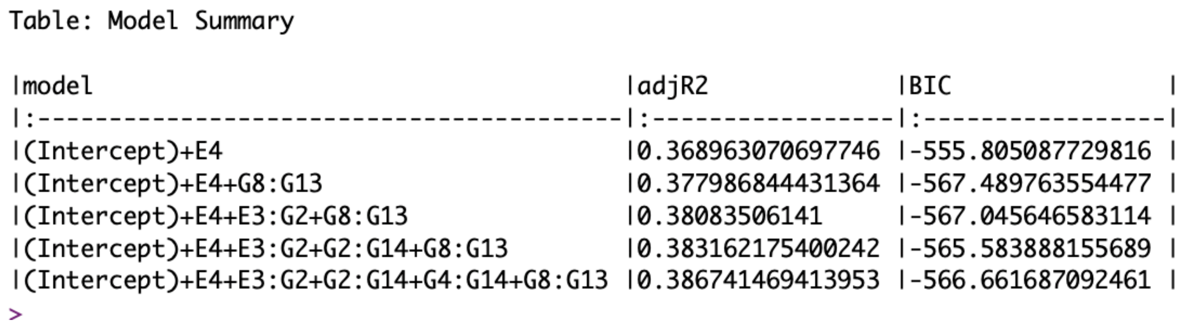


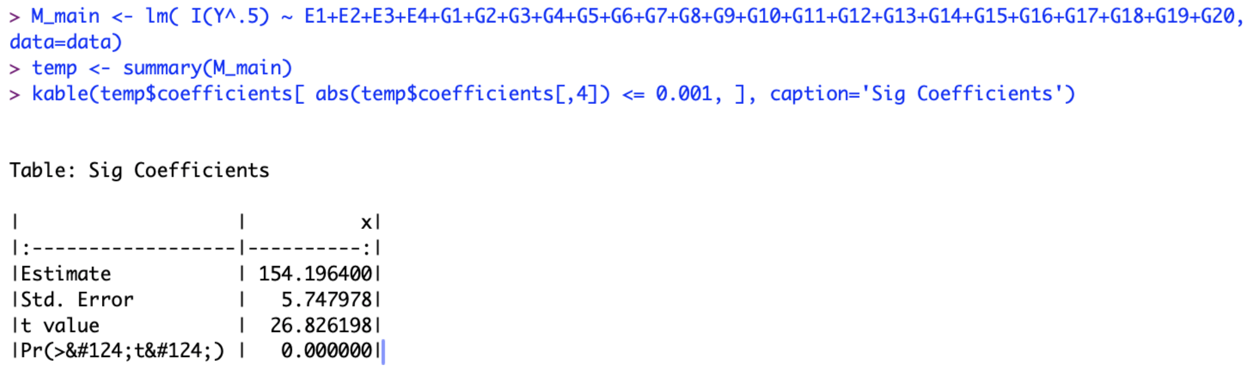
Table 1: Model summary using stepwise regression

```
 M <- regsubsets( model.matrix(M_trans)[,-1], I((data$Y)^.5),
+                 nbest = 1 , nvmax=5,
+                 method = 'forward', intercept = TRUE )
>
> temp <- summary(M)
> Var <- colnames(model.matrix(M_trans))
> M_select <- apply(temp$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                            |0.368963070697746 |-555.805087729816 |
|(Intercept)+E4+G8:G13                     |0.377986844431364 |-567.489763554477 |
|(Intercept)+E4+E3:G2+G8:G13               |0.38083506141     |-567.045646583114 |
|(Intercept)+E4+E3:G2+G2:G14+G8:G13        |0.383162175400242 |-565.583888155689 |
|(Intercept)+E4+E3:G2+G2:G14+G4:G14+G8:G13 |0.386741469413953 |-566.661687092461 |
>
```

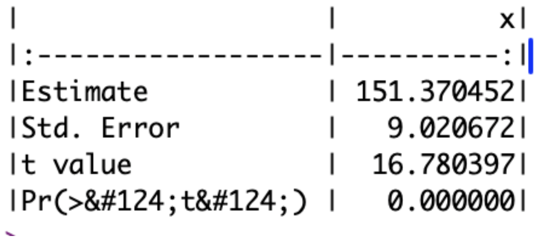
Table 2. Variables that have a significant main effect



```
> 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=data)
> temp <- summary(M_main)
> kable(temp$coefficients[ abs(temp$coefficients[,4]) <= 0.001, ], caption='Sig Coefficients')


Table: Sig Coefficients

|                   |          x|
|:------------------|----------:|
|Estimate           | 154.196400|
|Std. Error         |   5.747978|
|t value            |  26.826198|
|Pr(>&#124;t&#124;) |   0.000000|
```

Table 3. Variables that don't have a significant main effect

```
> M_2stage <- lm( I(Y^.5) ~ (E4+G8+G13)^2, data=Dat)
> temp <- summary(M_2stage)
> kable(temp$coefficients[ abs(temp$coefficients[,3]) >= 4, ])
```



```
|                   |          x|
|:------------------|----------:|
|Estimate           | 151.370452|
|Std. Error         |   9.020672|
|t value            |  16.780397|
|Pr(>&#124;t&#124;) |   0.000000|
```

Technical Appendix

Figure 1. Residual plot for raw data

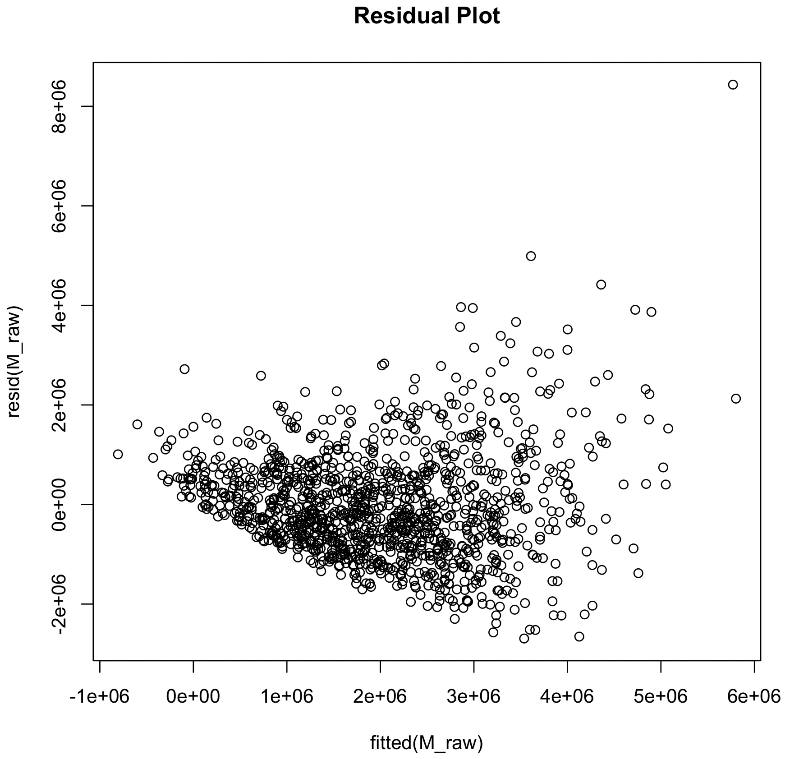


Figure 2. Summary of raw estimated model function

```
> summary(M_E)

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

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
```

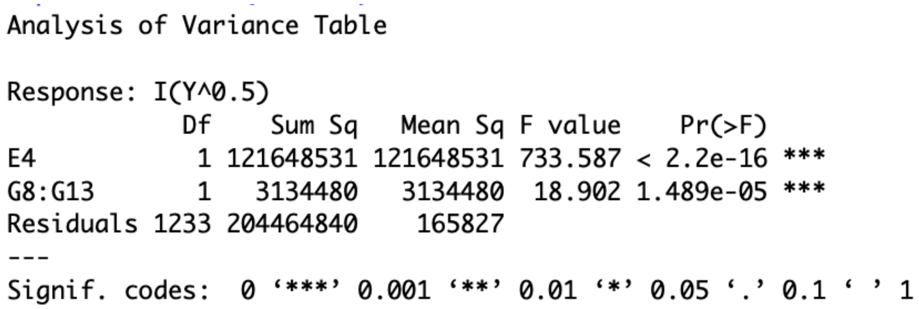
```
E4           421503      17528  24.048  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

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
```

Figure 3. Summary of transformed estimated model function

```
> M_final <- lm(I(Y^.5) ~ E4 + G8:G13, data = data)
> summary(M_final)

Call:
lm(formula = I(Y^0.5) ~ E4 + G8:G13, data = data)

Residuals:
    Min       1Q   Median       3Q      Max
-1104.84  -292.13   -32.99   258.65  1819.21

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -171.604     55.716  -3.080  0.00212 **
E4           154.063      5.697  27.041  < 2e-16 ***
G8:G13       140.541     32.326   4.348 1.49e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 407.2 on 1233 degrees of freedom
Multiple R-squared:  0.379,    Adjusted R-squared:  0.378
F-statistic: 376.2 on 2 and 1233 DF,  p-value: < 2.2e-16
```

Table 1. ANOVA Table of the estimated model function

> pureErrorAnova(M_final)



```
Analysis of Variance Table

Response: I(Y^0.5)
            Df    Sum Sq   Mean Sq F value    Pr(>F)    
E4           1 121648531 121648531 733.587 < 2.2e-16 ***
G8:G13       1   3134480   3134480  18.902 1.489e-05 ***
Residuals 1233 204464840    165827                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```