Nicholas Barrett

ID 110355429

ASM578 Final Project

Introduction

This report is the summary of a model building analysis for the observations of several environmental and genetic indicator variables. The task is to find a model that the teaching assistant used to generate the dependent variable. The dataset is an analogue of the data analyzed by Caspi et al. for studying the effects of Gene by Environment interaction in the development of clinical depression. The relationship between the genes involved in serotonin regulation, and environmental variables is crucial to understand for the prediction of depression. This is partially because not all people who encounter similar stressful life events become clinically depressed, but some do, and genetic interactions are thought to be predictive for this. These Gene by Environment or GxE interactions as well as GxG and ExE interactions were analyzed in this report. The data provided also contained missing values that had to be evaluated and imputed.

Methodology

To process the data and the missing values, first the data was sorted and merged by their index values. The missing values were then counted, and their locations noted and compared. The interaction effects of missing values were also documented and evaluated, as well as the correlation between the missing values and each indicator variable.

For imputation, each incomplete variable is imputed by a separate model created by classification and regression trees (CART) by Gibb's sampling. The implementation of the method being used is the MICE package in R. This will generate multiple imputations for each column based on models of the other columns, which are then evaluated and pooled after analysis for a robust imputation. The number and significance of outliers was small, so they remained in the model.

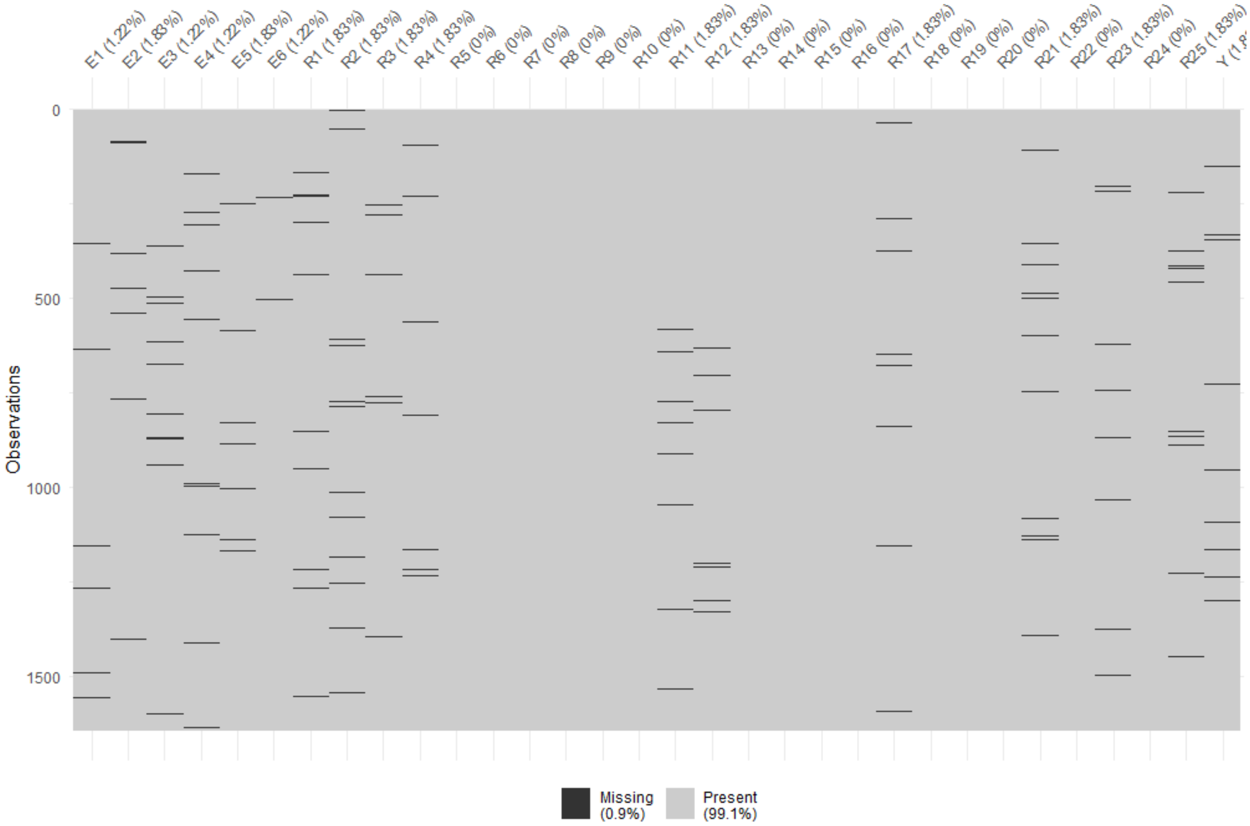
Several transformations of the dependent and independent variables were conducted, including the box cox transformations. The dependent variable was normal and linear as well as all the continuous independent variables. The most significant one that was included in the final model was the log of the dependent variable.

Up to three-way interactions between all variables were evaluated using forward selection. This was done by building a linear model, and then using the regsubsets function in R from the Leaps package. The process involved evaluating all models starting at 0 variables and adding the best predictor from the set of all three-way interaction terms, based on a F statistic, and then iterate. A LASSO based approach was also tested but yielded worse results than the forward selection method.

The Bayesian information criterion, as well as the adjusted R squared were gathered for model comparison. The models generated by this process were then subject to backwards selection using an F to keep of 16, starting at the best model with 10 covariates. The models with interaction terms were subjected to ANOVA comparison to the same model including the interaction terms and each of the individual terms from that interaction. The final model was selected based on an F to enter of 16 for each variable/interaction during backwards selection, as well as a competitive R squared and BIC value.

Results

There were either 20 or 30 missing values for each variable with missing values, with 13 of the variables having no missing values. The distribution is shown in the figure below. The interactions between the missing are also displayed on the figure below.



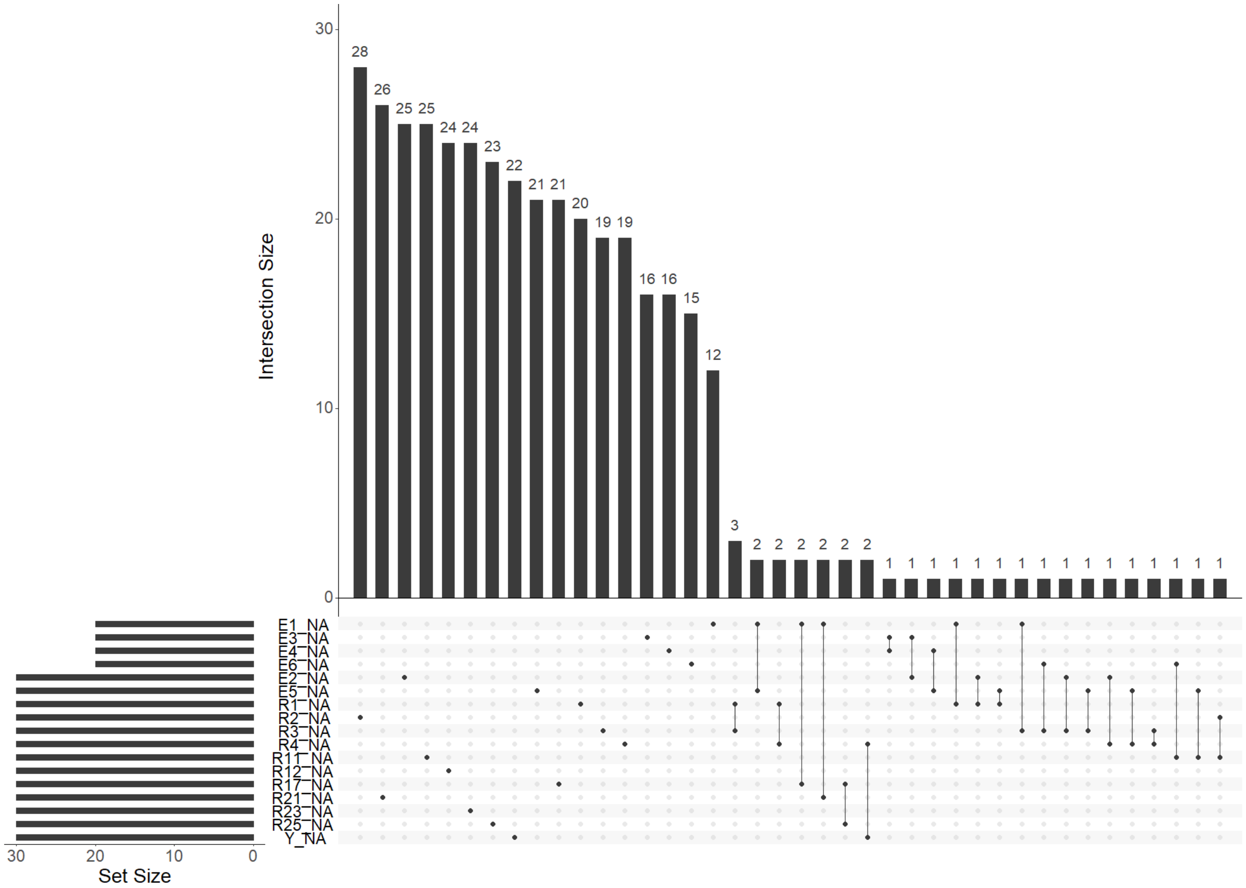
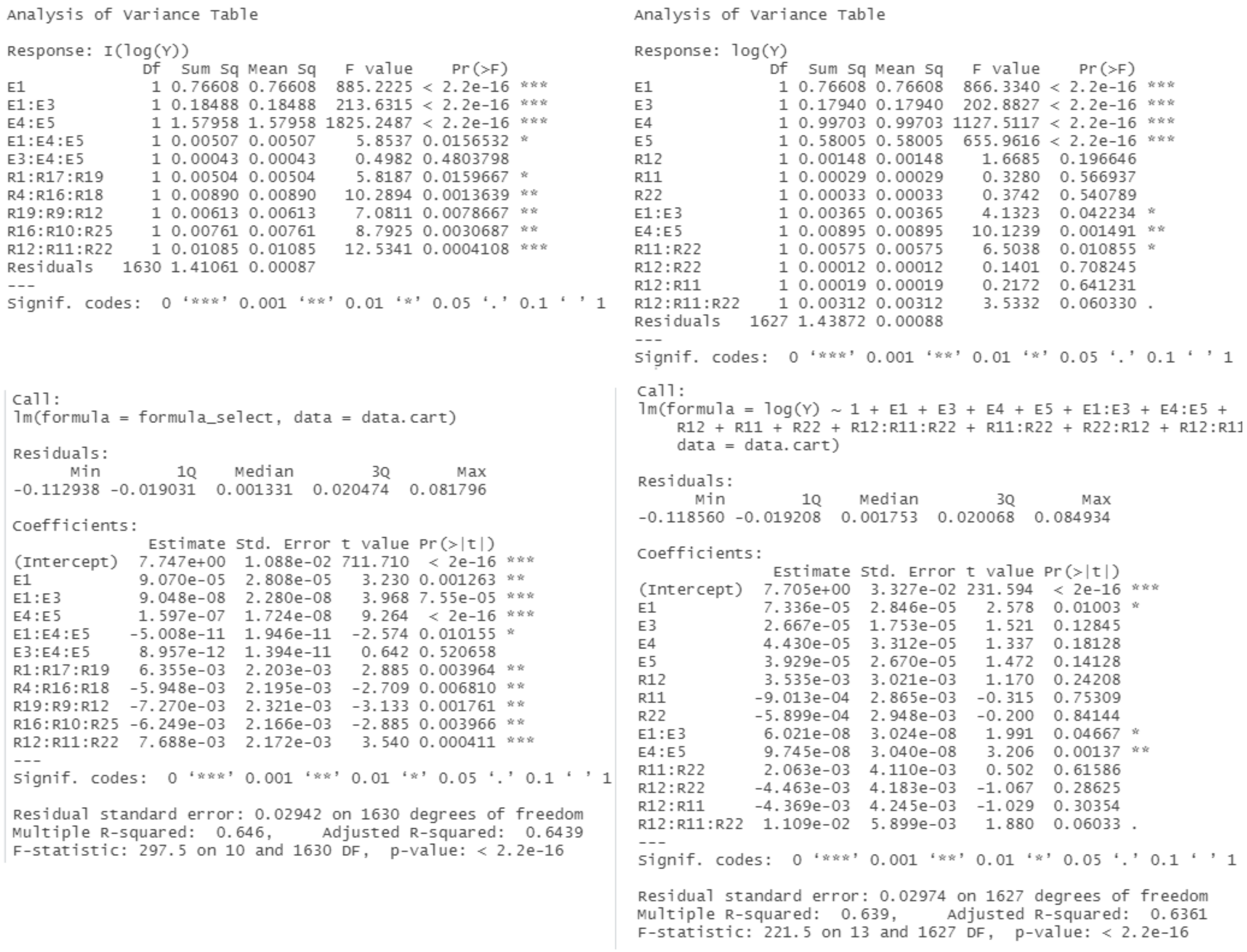


Figure 1: Missing Values and Interactions

There is no apparent pattern or connection between the missing values, with very few missing values in the same row, represented here by a line connecting two rows. The plots and statistics of the imputed values can be found in the appendix, as well as individual interactions between indicator variables and missing values. The highest correlation between an indicator variable and missing values was ~2%. These missing values were imputed by CART and Gibbs sampling described in the methods section for a total of 1641 observations.

The best models produced containing 10 variables of up to 2- and 3-way interactions were compared to those same models but with each interaction variable listed individually.



```
Analysis of Variance Table

Response: I(log(Y))
              Df  Sum Sq Mean Sq   F value    Pr(>F)    
E1             1 0.76608 0.76608  885.2225 < 2.2e-16 ***
E1:E3          1 0.18488 0.18488  213.6315 < 2.2e-16 ***
E4:E5          1 1.57958 1.57958 1825.2487 < 2.2e-16 ***
E1:E4:E5       1 0.00507 0.00507    5.8537 0.0156532 *  
E3:E4:E5       1 0.00043 0.00043    0.4982 0.4803798    
R1:R17:R19     1 0.00504 0.00504    5.8187 0.0159667 *  
R4:R16:R18     1 0.00890 0.00890   10.2894 0.0013639 ** 
R19:R9:R12     1 0.00613 0.00613    7.0811 0.0078667 ** 
R16:R10:R25    1 0.00761 0.00761    8.7925 0.0030687 ** 
R12:R11:R22    1 0.01085 0.01085   12.5341 0.0004108 ***
Residuals   1630 1.41061 0.00087                        
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

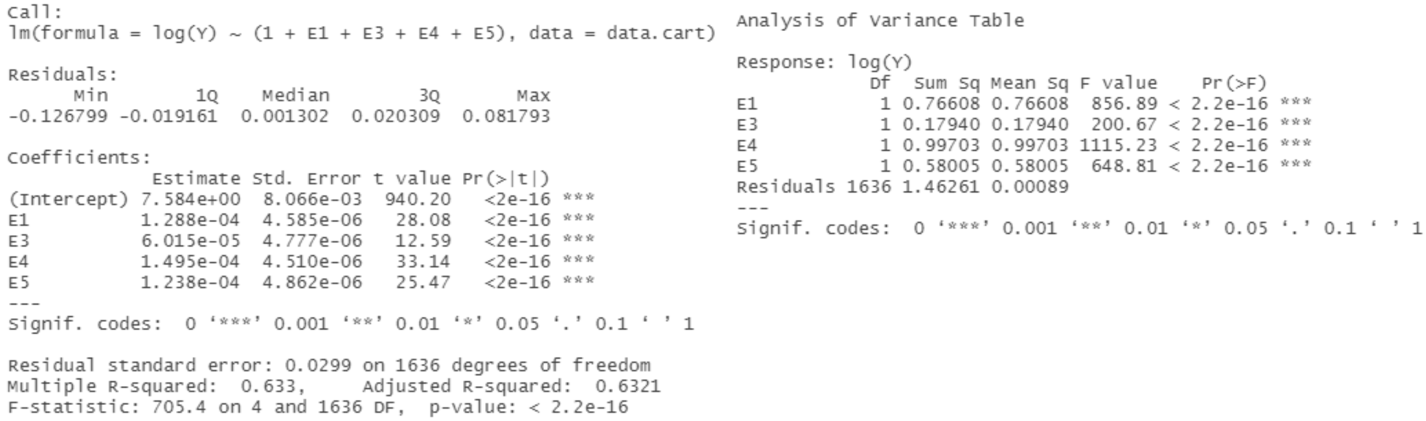
```
Call:
lm(formula = formula_select, data = data.cart)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.112938 -0.019031  0.001331  0.020474  0.081796 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  7.747e+00  1.088e-02 711.710  < 2e-16 ***
E1           9.070e-05  2.808e-05   3.230 0.001263 ** 
E1:E3        9.048e-08  2.280e-08   3.968 7.55e-05 ***
E4:E5        1.597e-07  1.724e-08   9.264  < 2e-16 ***
E1:E4:E5    -5.008e-11  1.946e-11  -2.574 0.010155 *  
E3:E4:E5     8.957e-12  1.394e-11   0.642 0.520658    
R1:R17:R19   6.355e-03  2.203e-03   2.885 0.003964 ** 
R4:R16:R18  -5.948e-03  2.195e-03  -2.709 0.006810 ** 
R19:R9:R12  -7.270e-03  2.321e-03  -3.133 0.001761 ** 
R16:R10:R25 -6.249e-03  2.166e-03  -2.885 0.003966 ** 
R12:R11:R22  7.688e-03  2.172e-03   3.540 0.000411 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02942 on 1630 degrees of freedom
Multiple R-squared:  0.646,	Adjusted R-squared:  0.6439 
F-statistic: 297.5 on 10 and 1630 DF,  p-value: < 2.2e-16
```

```
Analysis of Variance Table

Response: log(Y)
              Df  Sum Sq Mean Sq   F value    Pr(>F)    
E1             1 0.76608 0.76608  866.3340 < 2.2e-16 ***
E3             1 0.17940 0.17940  202.8827 < 2.2e-16 ***
E4             1 0.99703 0.99703 1127.5117 < 2.2e-16 ***
E5             1 0.58005 0.58005  655.9616 < 2.2e-16 ***
R12            1 0.00148 0.00148    1.6685  0.196646    
R11            1 0.00029 0.00029    0.3280  0.566937    
R22            1 0.00033 0.00033    0.3742  0.540789    
E1:E3          1 0.00365 0.00365    4.1323  0.042234 *  
E4:E5          1 0.00895 0.00895   10.1239  0.001491 ** 
R11:R22        1 0.00575 0.00575    6.5038  0.010855 *  
R12:R22        1 0.00012 0.00012    0.1401  0.708245    
R12:R11        1 0.00019 0.00019    0.2172  0.641231    
R12:R11:R22    1 0.00312 0.00312    3.5332  0.060330 .  
Residuals   1627 1.43872 0.00088                        
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Call:
lm(formula = log(Y) ~ 1 + E1 + E3 + E4 + E5 + E1:E3 + E4:E5 + 
    R12 + R11 + R22 + R12:R11:R22 + R11:R22 + R22:R12 + R12:R11
    data = data.cart)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.118560 -0.019208  0.001753  0.020068  0.084934 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  7.705e+00  3.327e-02 231.594  < 2e-16 ***
E1           7.336e-05  2.846e-05   2.578  0.01003 *  
E3           2.667e-05  1.753e-05   1.521  0.12845    
E4           4.430e-05  3.312e-05   1.337  0.18128    
E5           3.929e-05  2.670e-05   1.472  0.14128    
R12          3.535e-03  3.021e-03   1.170  0.24208    
R11         -9.013e-04  2.865e-03  -0.315  0.75309    
R22         -5.899e-04  2.948e-03  -0.200  0.84144    
E1:E3        6.021e-08  3.024e-08   1.991  0.04667 *  
E4:E5        9.745e-08  3.040e-08   3.206  0.00137 ** 
R11:R22      2.063e-03  4.110e-03   0.502  0.61586    
R12:R22     -4.463e-03  4.183e-03  -1.067  0.28625    
R12:R11     -4.369e-03  4.245e-03  -1.029  0.30354    
R12:R11:R22  1.109e-02  5.899e-03   1.880  0.06033 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02974 on 1627 degrees of freedom
Multiple R-squared:  0.639,	Adjusted R-squared:  0.6361 
F-statistic: 221.5 on 13 and 1627 DF,  p-value: < 2.2e-16
```

Table 1: ANOVA Model Comparison

This table contains the ANOVA comparison between the best 10 covariate model on the left and its hierarchical model on the right with the breakdown of the interaction terms. The breakdown of variables for the model on the right comes from the variables with F statistics of 12 or greater from the model on the left. The ANOVA on the right shows that the only significant variables are E1, E3, E4, E5, and the intercept, based on the F statistic of 16 or greater, and the P-values. Multiple ANOVA comparisons were made between the nonhierarchical models including some of the other significant interaction terms. In all cases the analysis points to E1, E3, E4, E5 based on F statistics; these extra ANOVA tables can be found in the appendix. Constructing the model with those parameters gives the final model. The R-squared values for these models are within .1% of each other and the BIC is -6838 for the model on the left and -6783 for the model on the right. The R-squared value and BIC of all possible interaction models is attached in the appendix.



```
Call:
lm(formula = log(Y) ~ (1 + E1 + E3 + E4 + E5), data = data.cart)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.126799 -0.019161  0.001302  0.020309  0.081793 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 7.584e+00  8.066e-03  940.20   <2e-16 ***
E1          1.288e-04  4.585e-06   28.08   <2e-16 ***
E3          6.015e-05  4.777e-06   12.59   <2e-16 ***
E4          1.495e-04  4.510e-06   33.14   <2e-16 ***
E5          1.238e-04  4.862e-06   25.47   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0299 on 1636 degrees of freedom
Multiple R-squared:  0.633,	Adjusted R-squared:  0.6321 
F-statistic: 705.4 on 4 and 1636 DF,  p-value: < 2.2e-16
```

```
Analysis of Variance Table

Response: log(Y)
            Df  Sum Sq Mean Sq F value    Pr(>F)    
E1           1 0.76608 0.76608  856.89 < 2.2e-16 ***
E3           1 0.17940 0.17940  200.67 < 2.2e-16 ***
E4           1 0.99703 0.99703 1115.23 < 2.2e-16 ***
E5           1 0.58005 0.58005  648.81 < 2.2e-16 ***
Residuals 1636 1.46261 0.00089                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

$$Log_e(Y) = 7.78 + \ 0.00013 * E1 + \ 0.000063 * E3 + \ 0.00015 * E4 + \ 0.00012 * E5$$

Table 2: Final Model and Final Model ANOVA

The p-values and the F statistics for these variables are very robust, gaining a lot of significance but losing no predictive power as the R-squared for the final model is comparable to the two higher interaction models presented earlier. The residual standard error is low, and the standard error of each coefficient is at least an order of magnitude less than the coefficient. Additionally, the BIC of the final model is -6823, also comparable to the previous two.

Models with other transformations were also explored including the boxcox transformation and other possible exponential effects of independent variables and are reported in the appendix. The results of these models had similar variables after the same variable selection method. The largest difference between the Log transformed model and all the others was a 6-8 order of magnitude drop in residual standard error. The boxcox transformation-based model is provided in the appendix, and produced the same variables as produced by the log transformation.

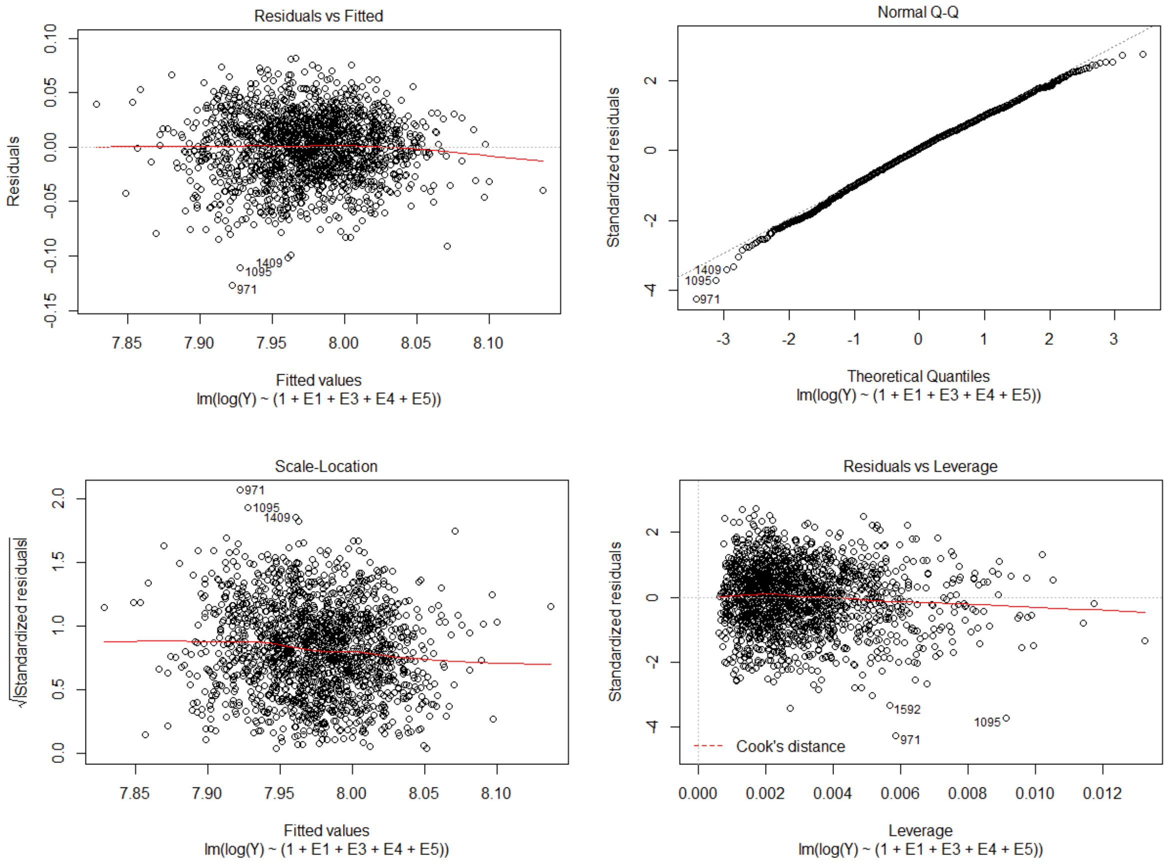


Figure 2: Final Model Residuals

The plots provided for the analysis of the final model show good results. The residuals vs fitted plot shows little overall trend with an even distribution and tight cluster, meaning there are few outliers. The Normal QQ plot also shows strong evidence for the normality of the residuals. The scale-location plot shows little change along the x axis implying homoscedasticity of the residuals. The residuals vs leverage plot shows all points within cook's distance, indicating no outliers.

Conclusions and Discussion

The model selected is robust and produces accurate predictions with an adjusted R-squared of .63 and a BIC of -6823. It is comparable in effectiveness to many more complicated models after eliminating the potential interaction terms. This model suggests no significant relation between GxE, GxG or ExE interactions and the outcome variable Y, for the data evaluated in this report.

One limitation of this report is the exclusion of possible four term interactions, which were possible to use by the TA for our dependent variable generation. It is also possible that there was a pattern in the missing data that could have made a different imputation technique more valid. Another limitation is the potential for models larger then 10 variables/interaction terms, as the F = 16 to keep began at 10 covariates.

References

Caspi A, Sugden K, Moffitt TE, et al. Influence of life stress on depression: moderation by a polymorphism in the5-HTTgene.Science. 2003;301(5631)

van Buuren S, Groothuis-Oudshoorn K (2011). “mice: Multivariate Imputation by Chained Equations in R.” *Journal of Statistical Software*, **45**(3), 1-67. https://www.jstatsoft.org/v45/i03/.

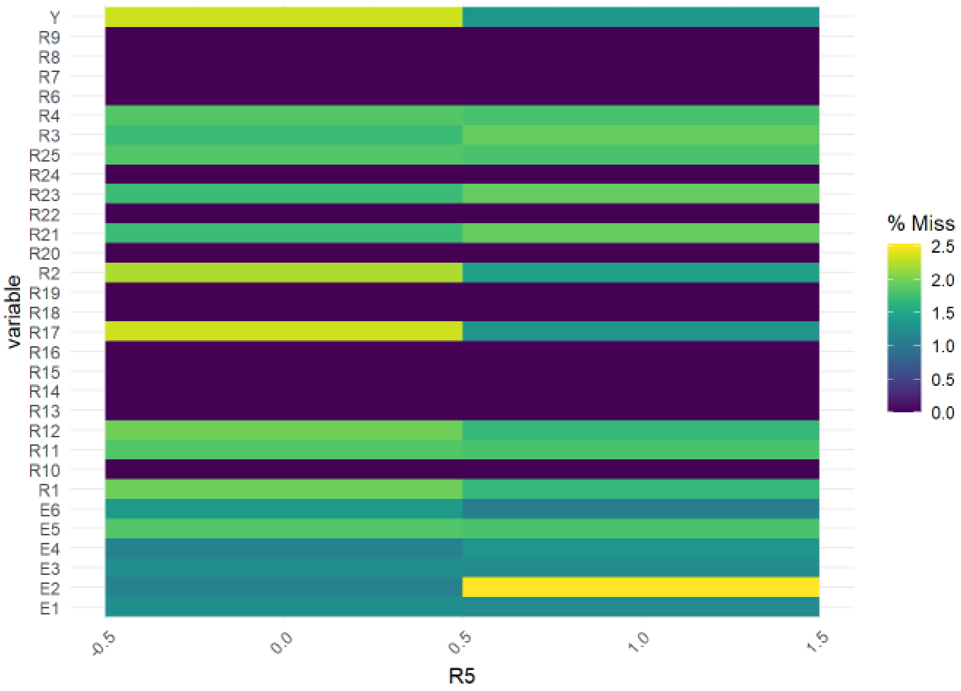
R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/

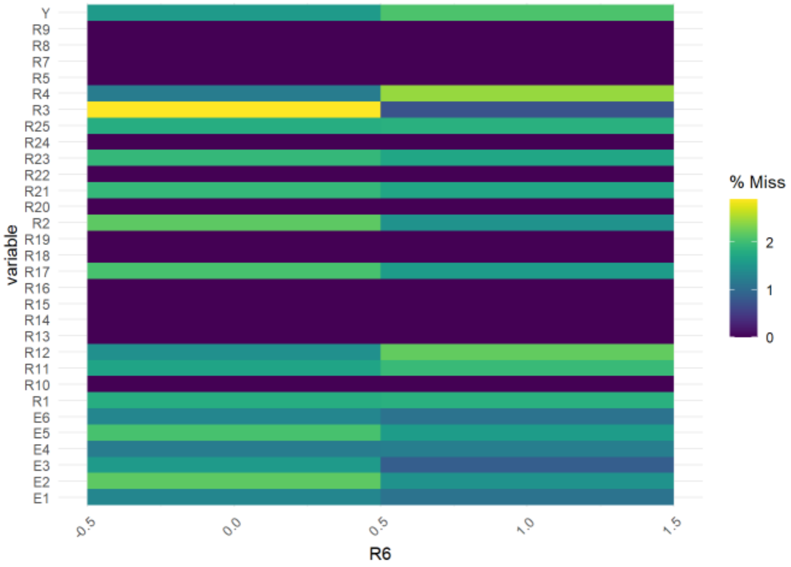
Lumley, Thomas., Miller, Alan. (2020). Leaps: Regression Subset Selection. R package Version 3.1

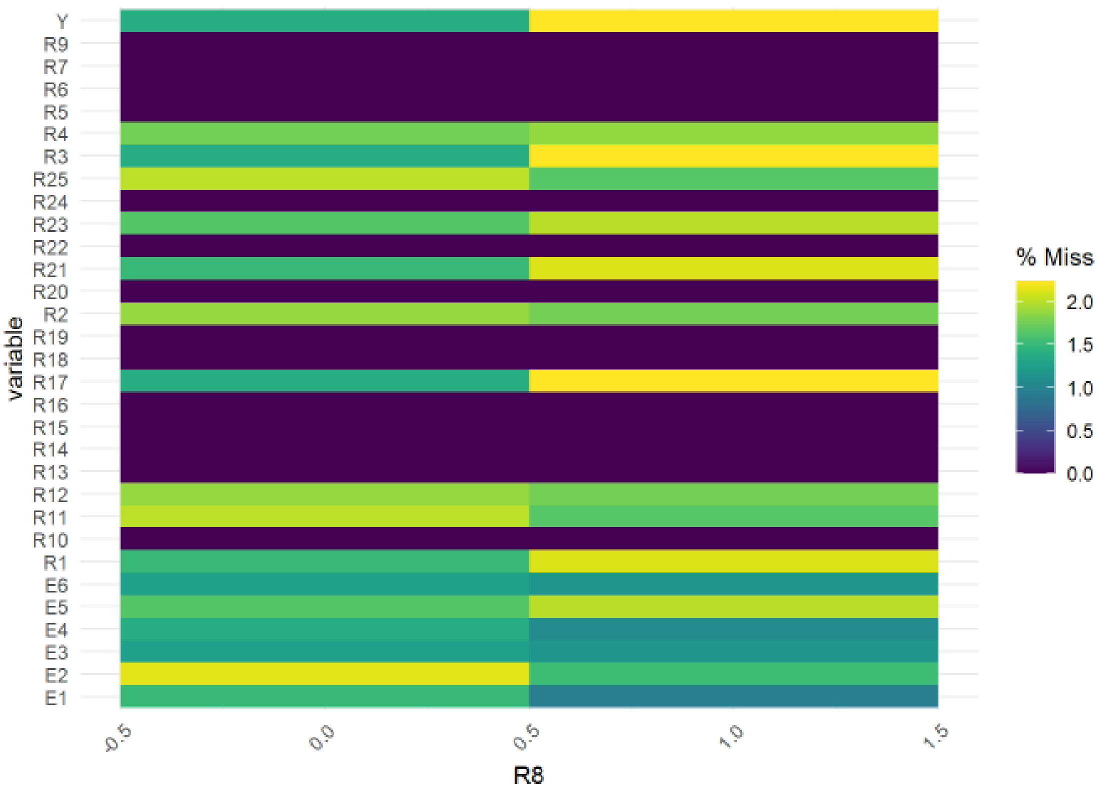
Montgomery, D.C.; Peck, E.A.; and Vining, G.G. (2012). Introduction to Linear Regression Analysis; Fifth Edition. Hoboken, N.J.: John Wiley and Sons.

Appendix

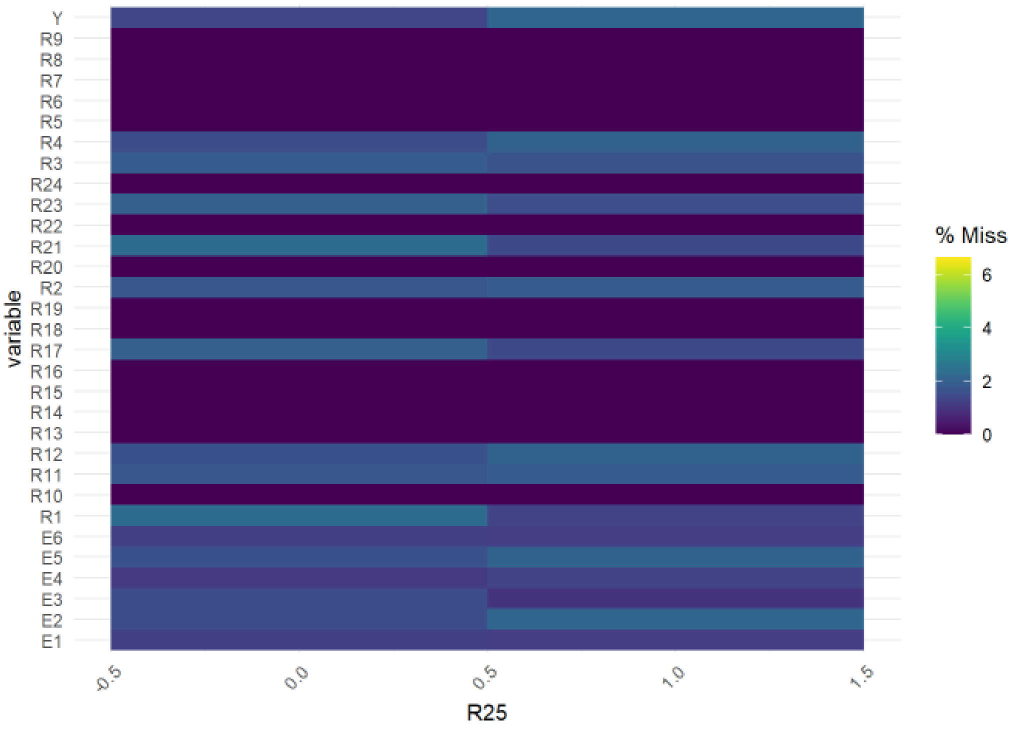
Missing Data by Indicator Variable,







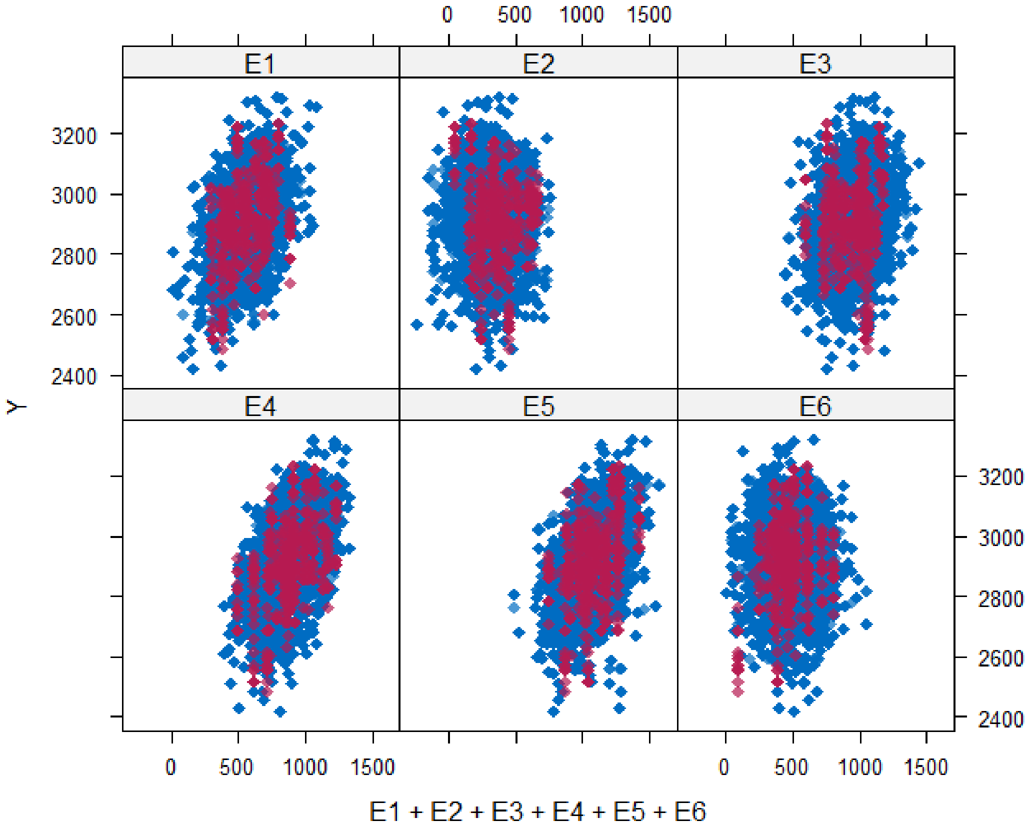
Y
R9
R7
R6
R5
R4
R3
R25
R24
R23
R22
R21
R20
R2
R19
R18
R17
R16
R15
R14
R13
R12
R11
R10
R1
E6
E5
E4
E3
E2
E1
variable
-0.5
0.0
0.5
1.0
1.5
R8
% Miss
2.0
1.5
1.0
0.5
0.0

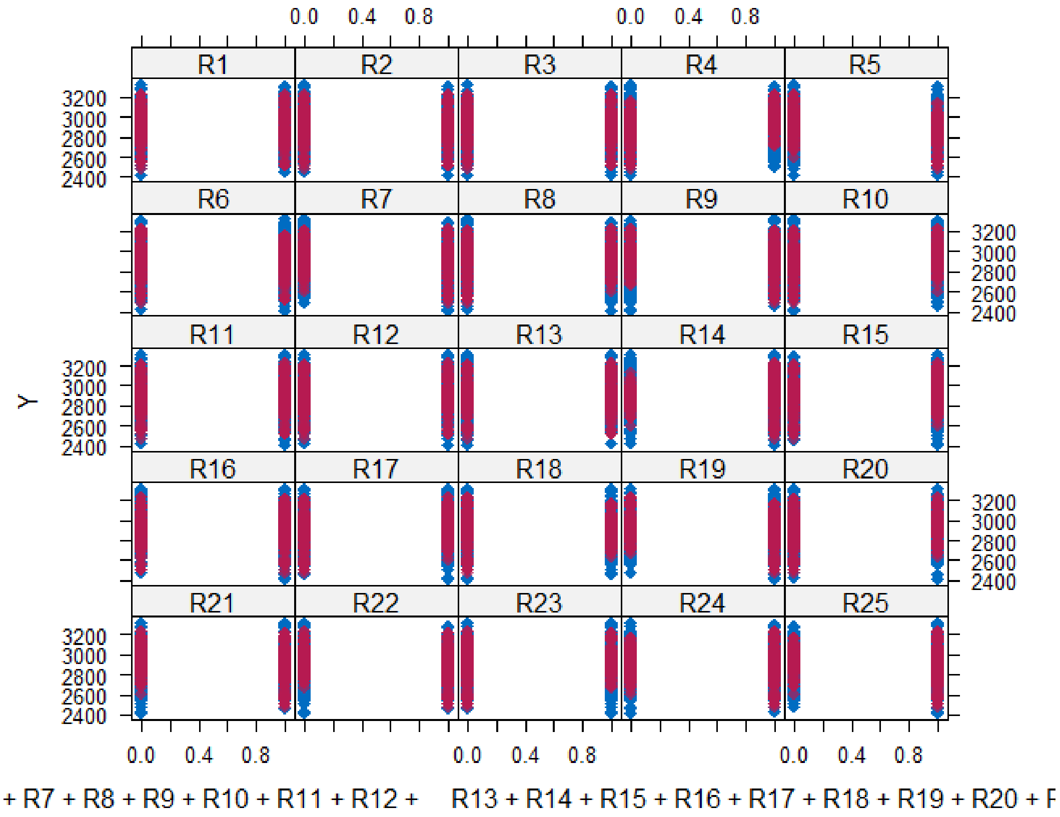


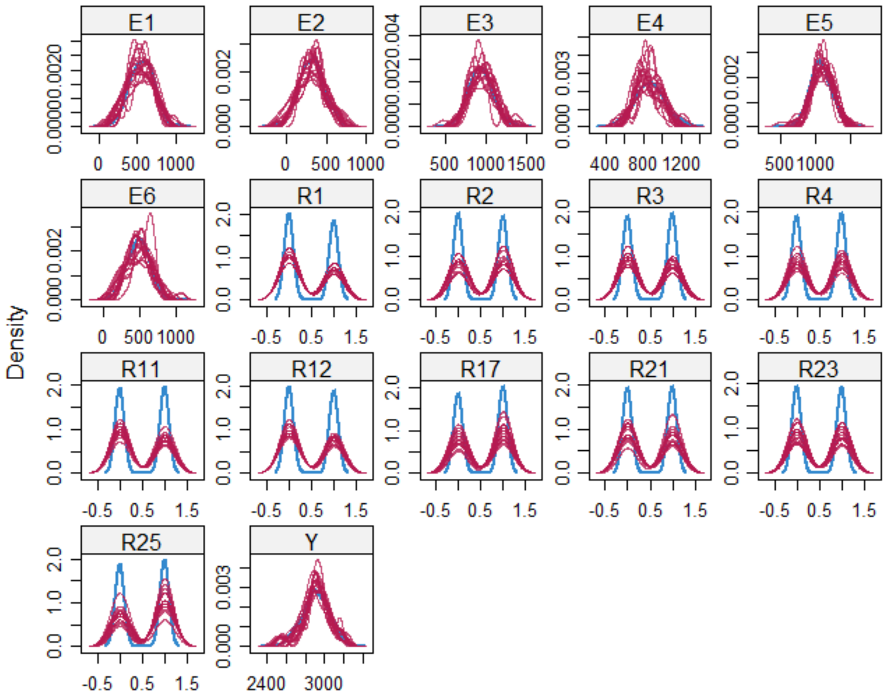
Y
R9
R8
R7
R6
R5
R4
R3
R24
R23
R22
R21
R20
R2
R19
R18
R17
R16
R15
R14
R13
R12
R11
R10
R1
E6
E5
E4
E3
E2
E1
variable
-0.5
0.0
0.5
1.0
1.5
R25
% Miss
6
4
2
0

These are some examples, but around 10 of the indicator variables showed effects similar to R8's plot above

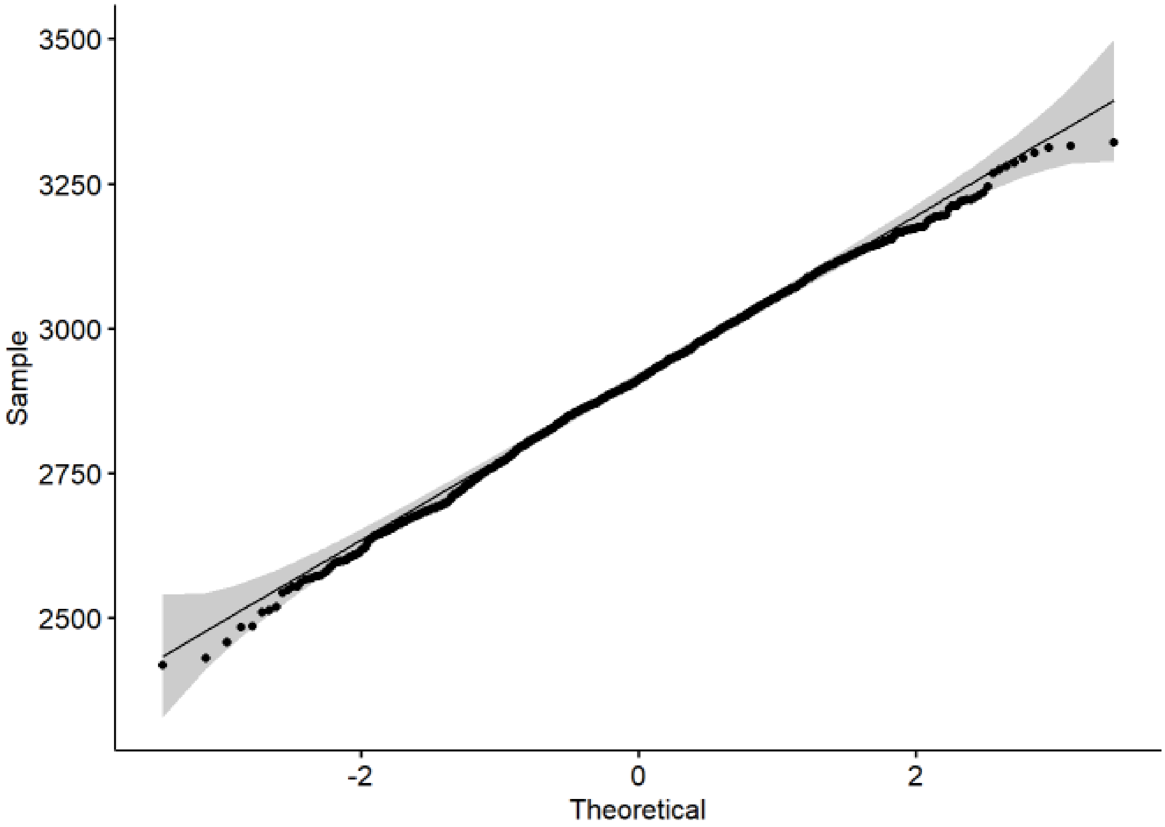
Imputed Data



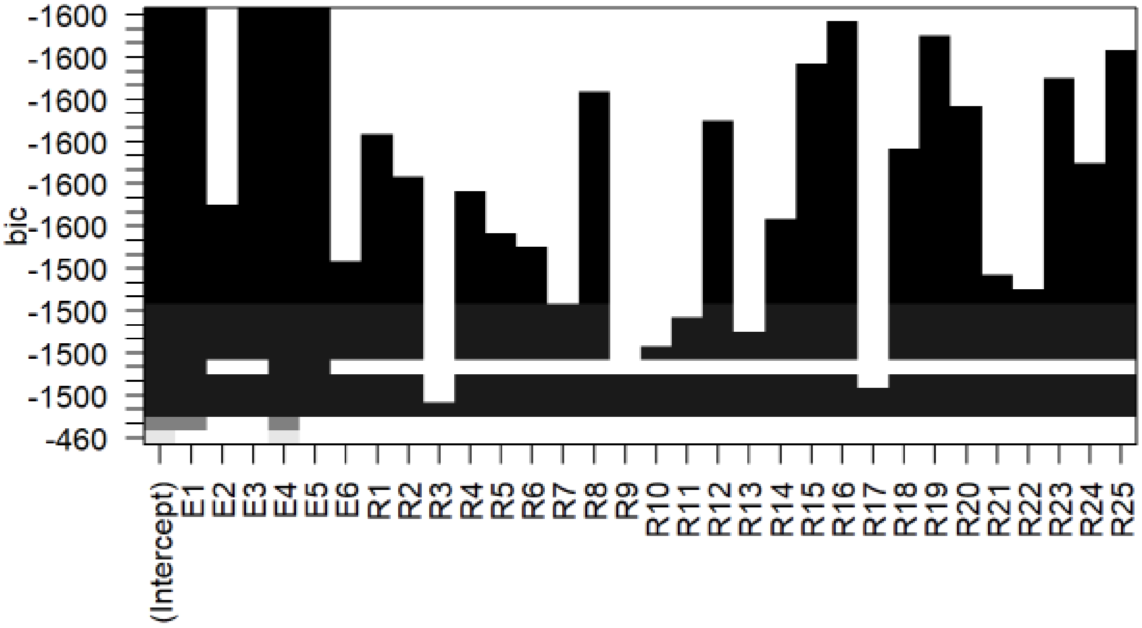

R1
R2
R3
R4
R5
R6
R7
R8
R9
R10
R11
R12
R13
R14
R15
R16
R17
R18
R19
R20
R21
R22
R23
R24
R25
Y
3200
3000
2800
2600
2400
0.0
0.4
0.8
+ R7 + R8 + R9 + R10 + R11 + R12 +
R13 + R14 + R15 + R16 + R17 + R18 + R19 + R20 + F

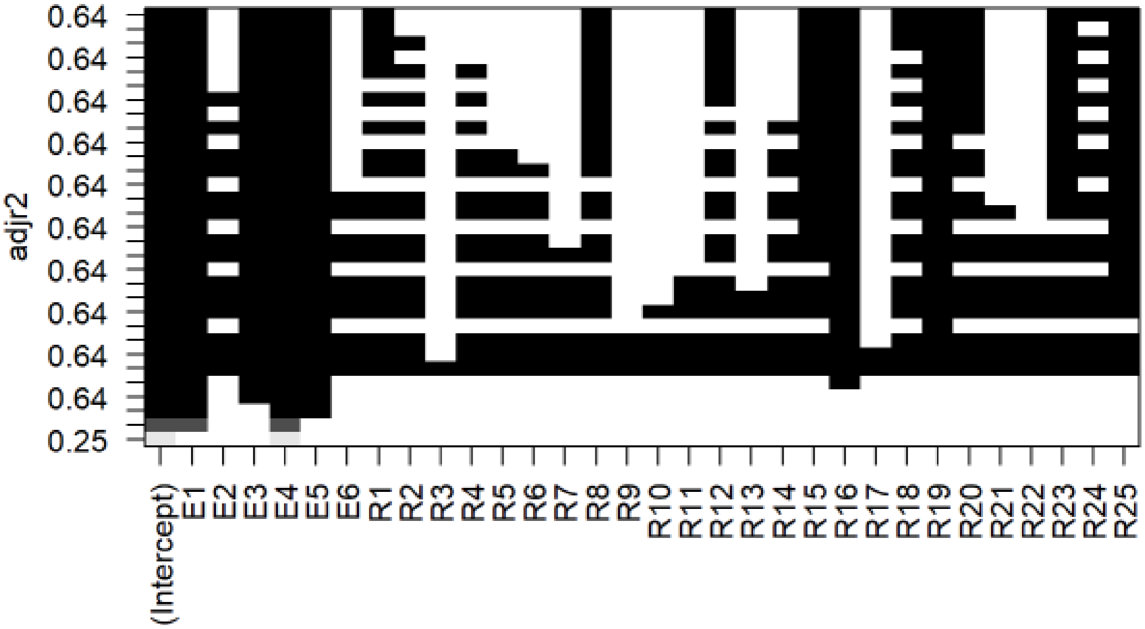

E1
E2
E3
E4
E5
E6
R1
R2
R3
R4
R11
R12
R17
R21
R23
R25
Y
Density
0 500 1000
500 1000 1500
400 800 1200
500 1000
-0.5 0.5 1.5
2400 3000
0.0000.0020
0.000 0.002
0.0000.0020.004
0.000 0.003
0.0 1.0 2.0

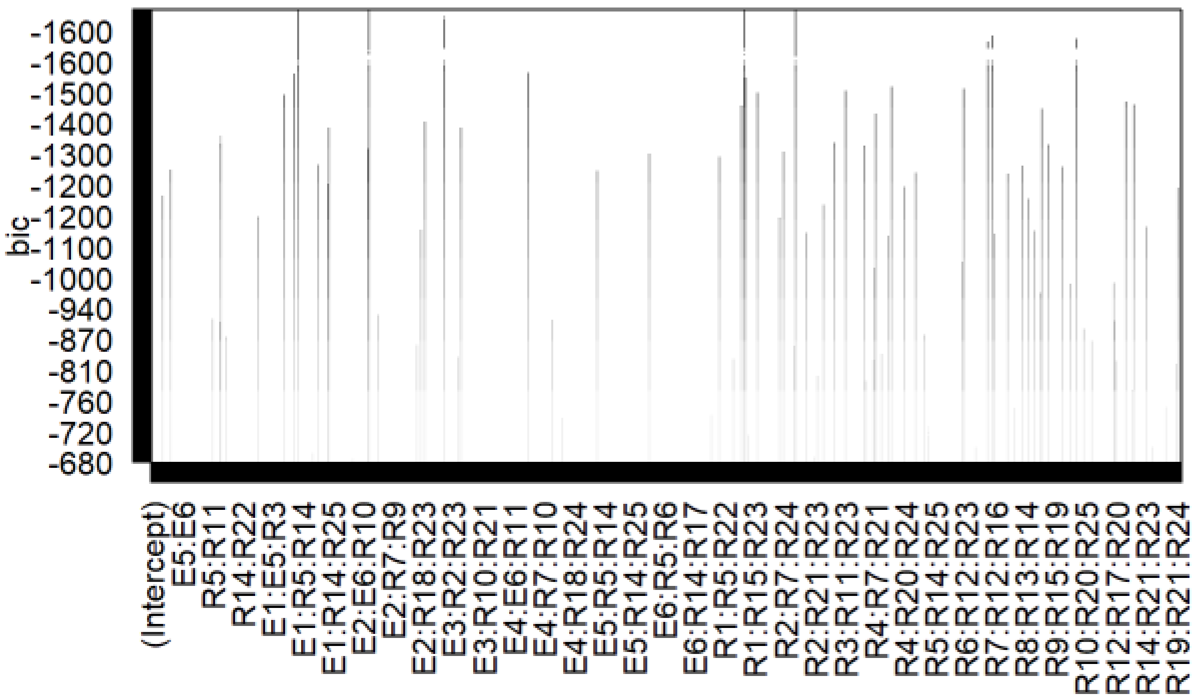
Y Linearity (untransformed)

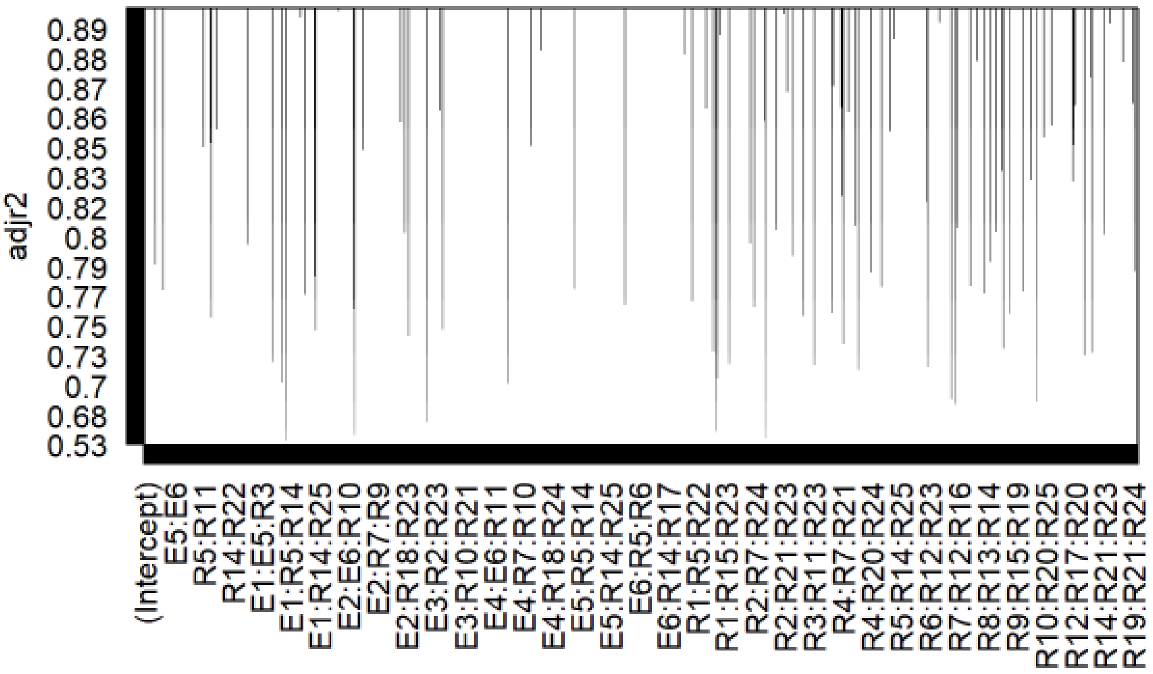


R-squared and BIC by model (Each row on graph is a model with black indicating an included variable)

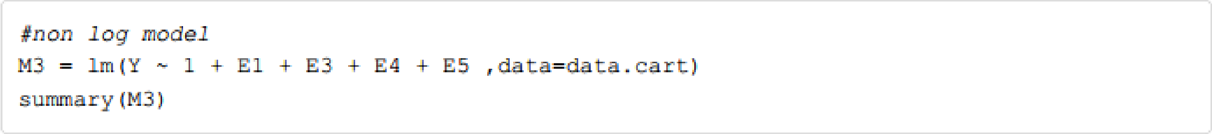




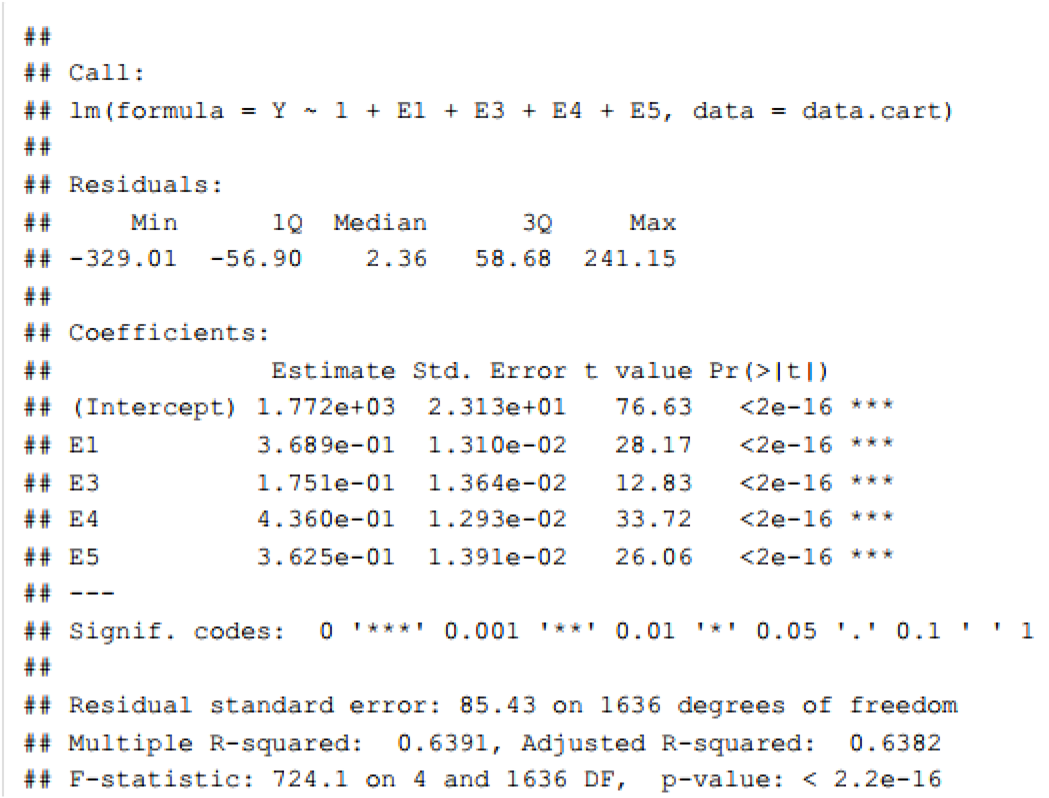

bic
-1600
-1600
-1500
-1400
-1300
-1200
-1200
-1100
-1000
-940
-870
-810
-760
-720
-680
(Intercept)
E5:E6
R5:R11
R14:R22
E1:E5:R3
E1:R5:R14
E1:R14:R25
E2:E6:R10
E2:R7:R9
E2:R18:R23
E3:R2:R23
E3:R10:R21
E4:E6:R11
E4:R7:R10
E4:R18:R24
E5:R5:R14
E5:R14:R25
E6:R5:R6
E6:R14:R17
R1:R5:R22
R1:R15:R23
R2:R7:R24
R2:R21:R23
R3:R11:R23
R4:R7:R21
R4:R20:R24
R5:R14:R25
R6:R12:R23
R7:R12:R16
R8:R13:R14
R9:R15:R19
R10:R20:R25
R12:R17:R20
R14:R21:R23
R19:R21:R24


adjr2
0.89
0.88
0.87
0.86
0.85
0.83
0.82
0.8
0.79
0.77
0.75
0.73
0.7
0.68
0.53
(Intercept)
E5:E6
R5:R11
R14:R22
E1:E5:R3
E1:R5:R14
E1:R14:R25
E2:E6:R10
E2:R7:R9
E2:R18:R23
E3:R2:R23
E3:R10:R21
E4:E6:R11
E4:R7:R10
E4:R18:R24
E5:R5:R14
E5:R14:R25
E6:R5:R6
E6:R14:R17
R1:R5:R22
R1:R15:R23
R2:R7:R24
R2:R21:R23
R3:R11:R23
R4:R7:R21
R4:R20:R24
R5:R14:R25
R6:R12:R23
R7:R12:R16
R8:R13:R14
R9:R15:R19
R10:R20:R25
R12:R17:R20
R14:R21:R23
R19:R21:R24

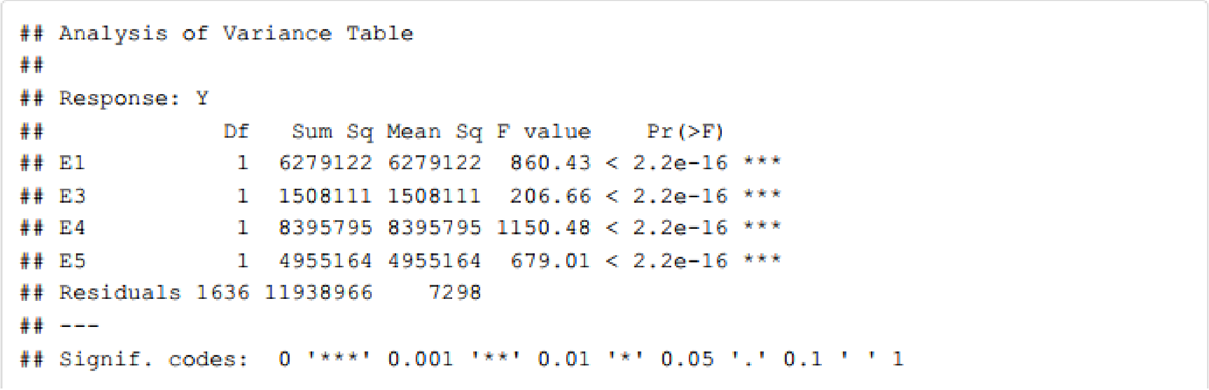
Non log model



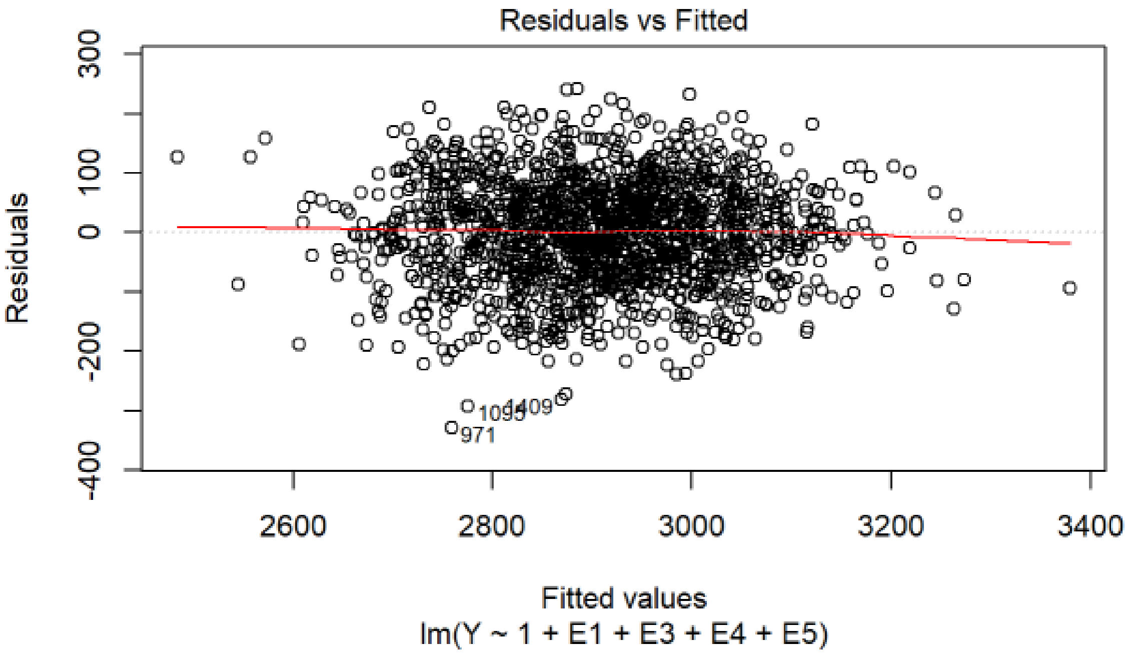
```
#non log model
M3 = lm(Y ~ 1 + E1 + E3 + E4 + E5 ,data=data.cart)
summary(M3)
```

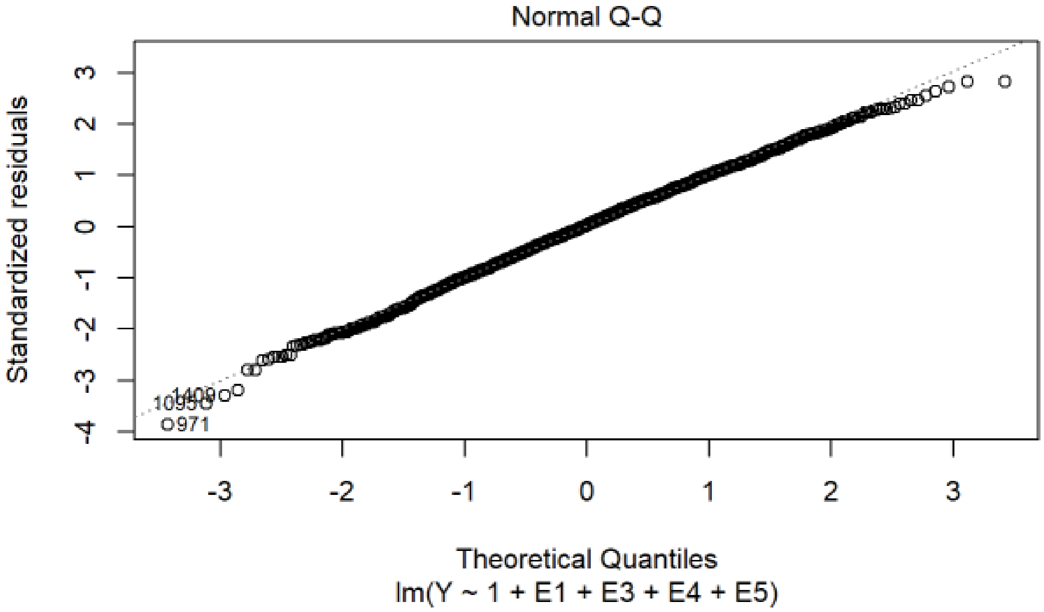


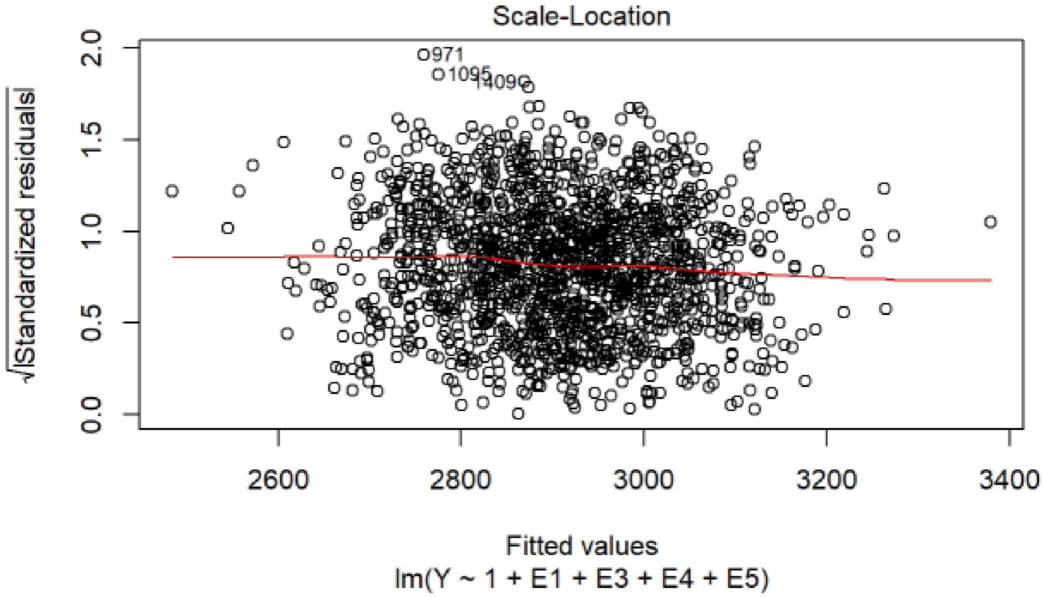
```
##
## Call:
## lm(formula = Y ~ 1 + E1 + E3 + E4 + E5, data = data.cart)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -329.01  -56.90    2.36   58.68  241.15
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.772e+03  2.313e+01   76.63   <2e-16 ***
## E1          3.689e-01  1.310e-02   28.17   <2e-16 ***
## E3          1.751e-01  1.364e-02   12.83   <2e-16 ***
## E4          4.360e-01  1.293e-02   33.72   <2e-16 ***
## E5          3.625e-01  1.391e-02   26.06   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 85.43 on 1636 degrees of freedom
## Multiple R-squared:  0.6391, Adjusted R-squared:  0.6382
## F-statistic: 724.1 on 4 and 1636 DF,  p-value: < 2.2e-16
```

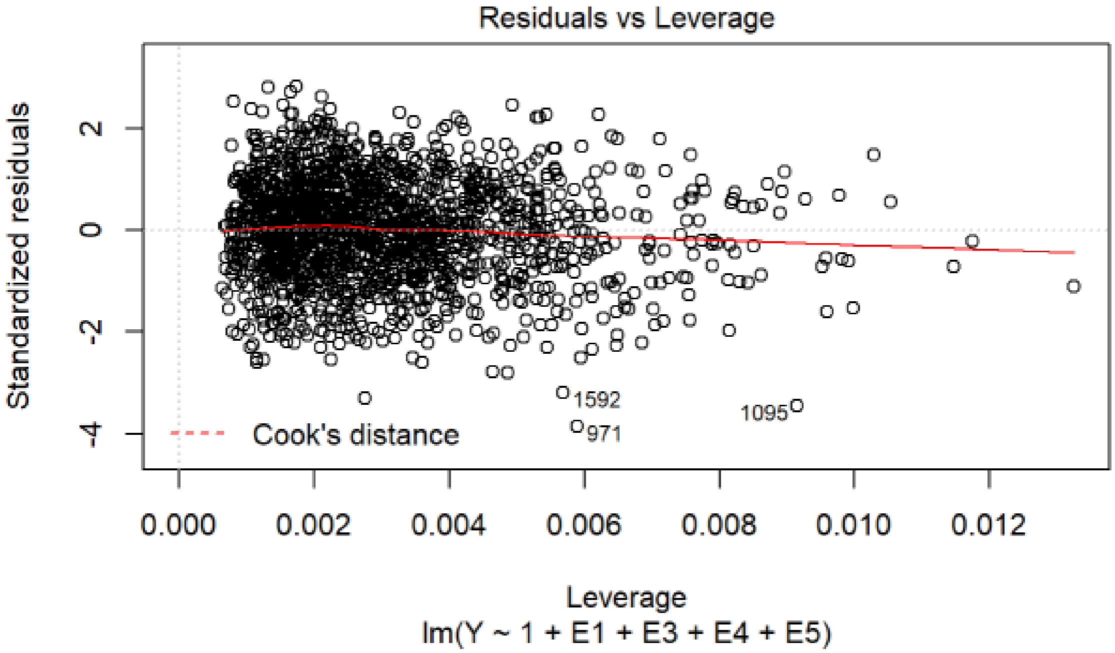


```
## Analysis of Variance Table
##
## Response: Y
##             Df   Sum Sq Mean Sq F value    Pr(>F)
## E1           1  6279122 6279122  860.43 < 2.2e-16 ***
## E3           1  1508111 1508111  206.66 < 2.2e-16 ***
## E4           1  8395795 8395795 1150.48 < 2.2e-16 ***
## E5           1  4955164 4955164  679.01 < 2.2e-16 ***
## Residuals 1636 11938966    7298
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```


Residuals vs Fitted
Residuals
300
100
0
-200
-400
2600
2800
3000
3200
3400
1095
1409
971
Fitted values
lm(Y ~ 1 + E1 + E3 + E4 + E5)

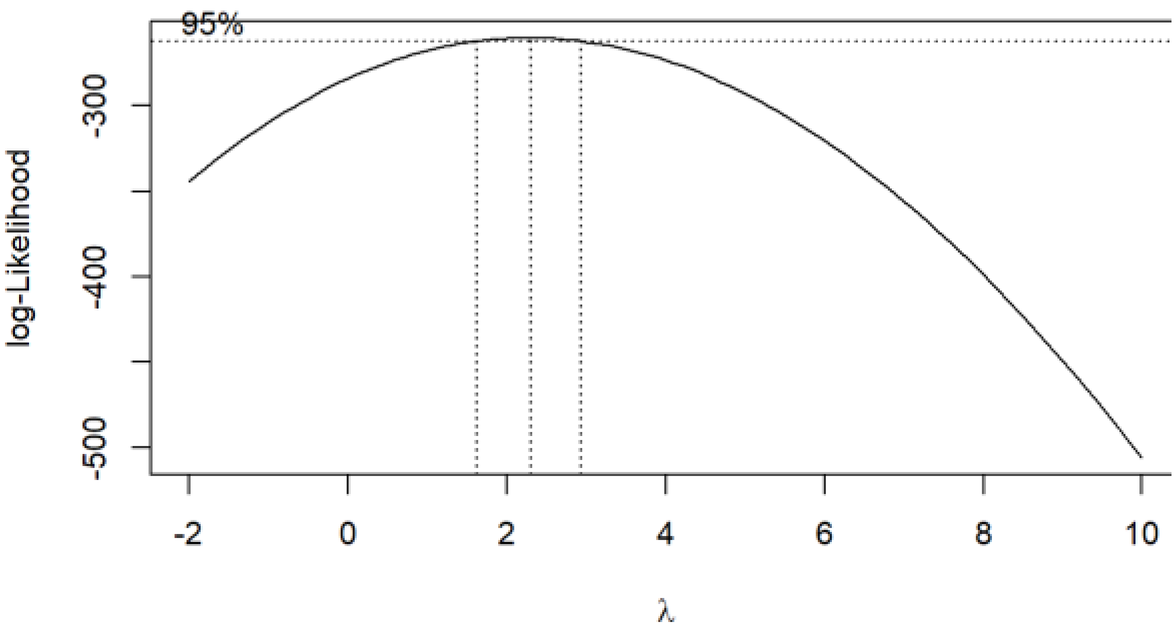

Normal Q-Q
Standardized residuals
3
2
1
0
-1
-2
-3
-4
-3
-2
-1
0
1
2
3
1409
1095
971
Theoretical Quantiles
lm(Y ~ 1 + E1 + E3 + E4 + E5)





Boxcox Transformation Model

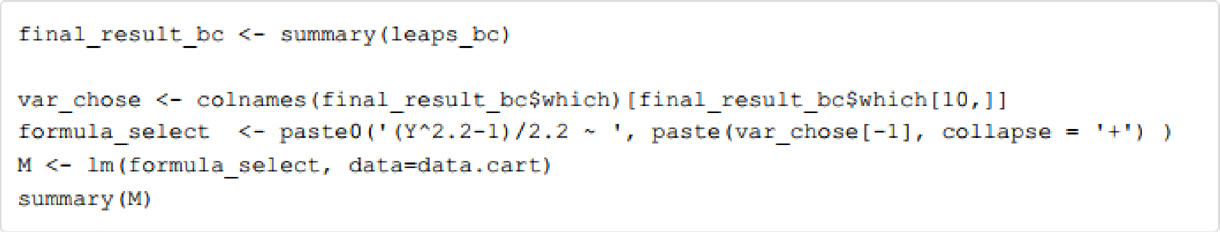
```
#box cox
bc <- boxcox(Y ~., data=test, lambda = seq(-2, 10, 1/10))
```



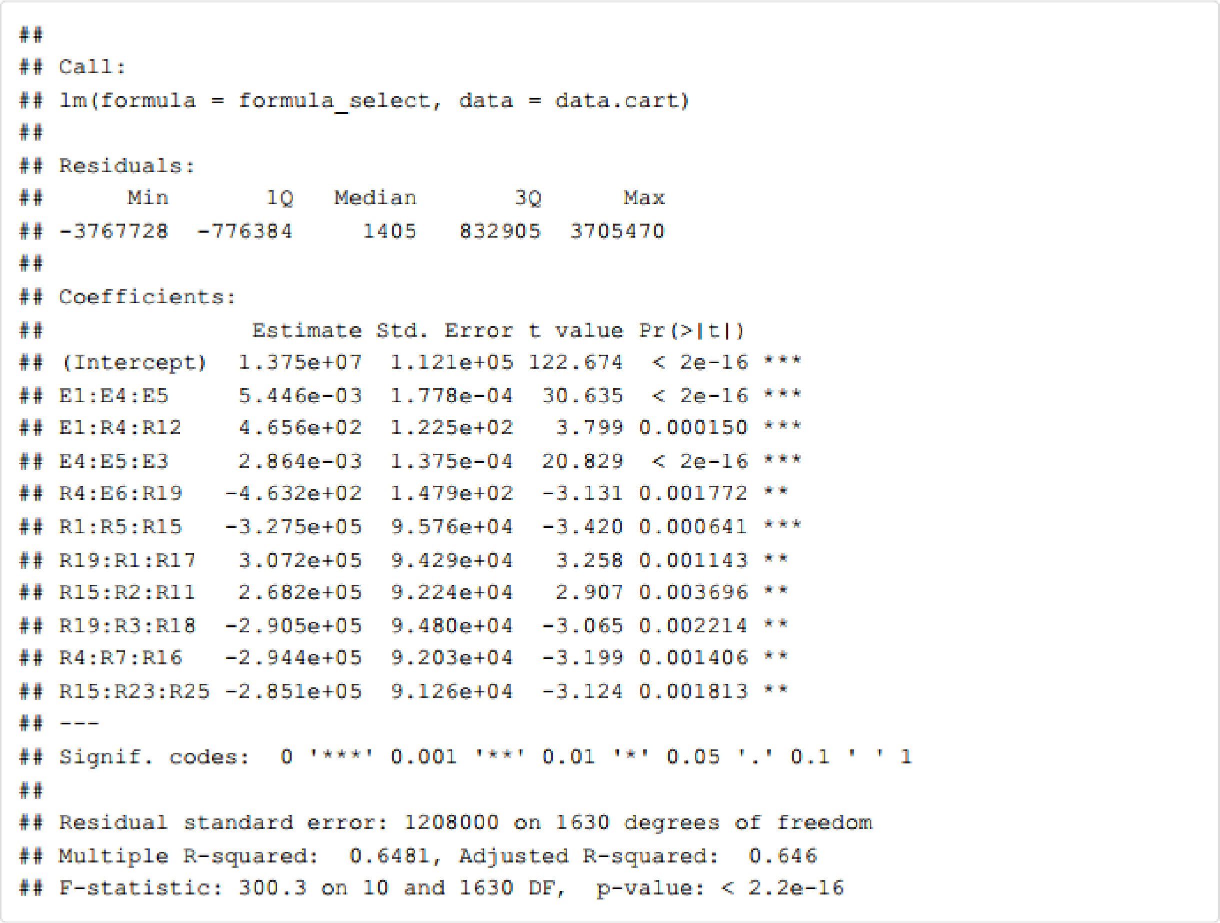
```
(lambda <- bc$x[which.max(bc$y)])
```

```
## [1] 2.3
```

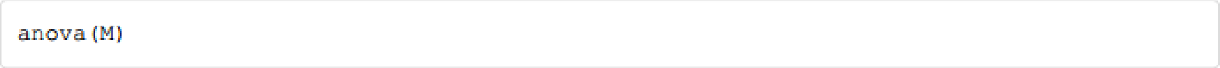
```
new_model <- lm(((Y^lambda-1)/lambda) ~(.)^3,data=test)
leaps_bc = regsubsets(model.matrix(new_model)[,-1], I((data.cart$Y^lambda-1)/lambda),
really.big = TRUE,
                              method = "forward", nbest = 1, intercept = TRUE, nvmax = 500)
```



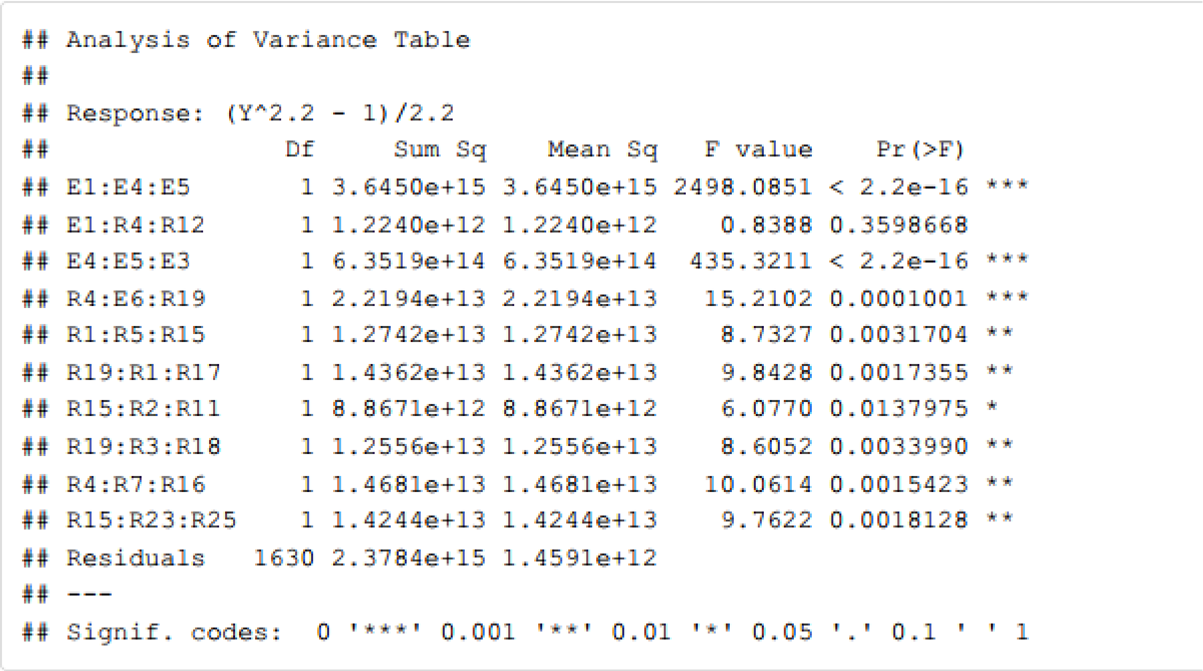
```
final_result_bc <- summary(leaps_bc)

var_chose <- colnames(final_result_bc$which)[final_result_bc$which[10,]]
formula_select  <- paste0('(Y^2.2-1)/2.2 ~ ', paste(var_chose[-1], collapse = '+') )
M <- lm(formula_select, data=data.cart)
summary(M)
```



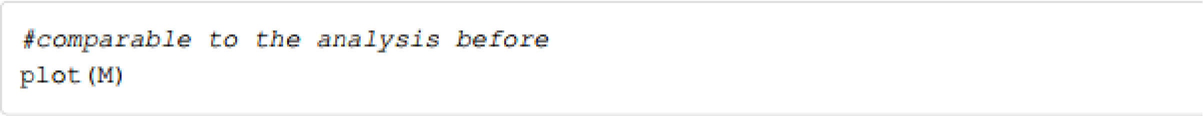
```
##
## Call:
## lm(formula = formula_select, data = data.cart)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3767728  -776384     1405   832905  3705470
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.375e+07  1.121e+05 122.674  < 2e-16 ***
## E1:E4:E5     5.446e-03  1.778e-04  30.635  < 2e-16 ***
## E1:R4:R12    4.656e+02  1.225e+02   3.799 0.000150 ***
## E4:E5:E3     2.864e-03  1.375e-04  20.829  < 2e-16 ***
## R4:E6:R19   -4.632e+02  1.479e+02  -3.131 0.001772 **
## R1:R5:R15   -3.275e+05  9.576e+04  -3.420 0.000641 ***
## R19:R1:R17   3.072e+05  9.429e+04   3.258 0.001143 **
## R15:R2:R11   2.682e+05  9.224e+04   2.907 0.003696 **
## R19:R3:R18  -2.905e+05  9.480e+04  -3.065 0.002214 **
## R4:R7:R16   -2.944e+05  9.203e+04  -3.199 0.001406 **
## R15:R23:R25 -2.851e+05  9.126e+04  -3.124 0.001813 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1208000 on 1630 degrees of freedom
## Multiple R-squared:  0.6481, Adjusted R-squared:  0.646
## F-statistic: 300.3 on 10 and 1630 DF,  p-value: < 2.2e-16
```



```
anova(M)
```

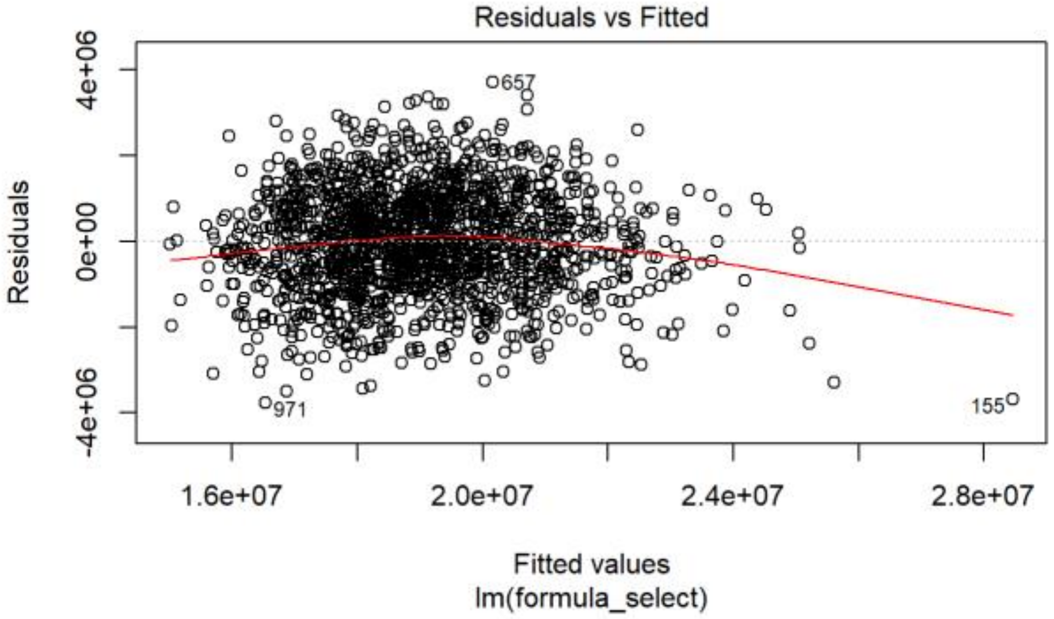


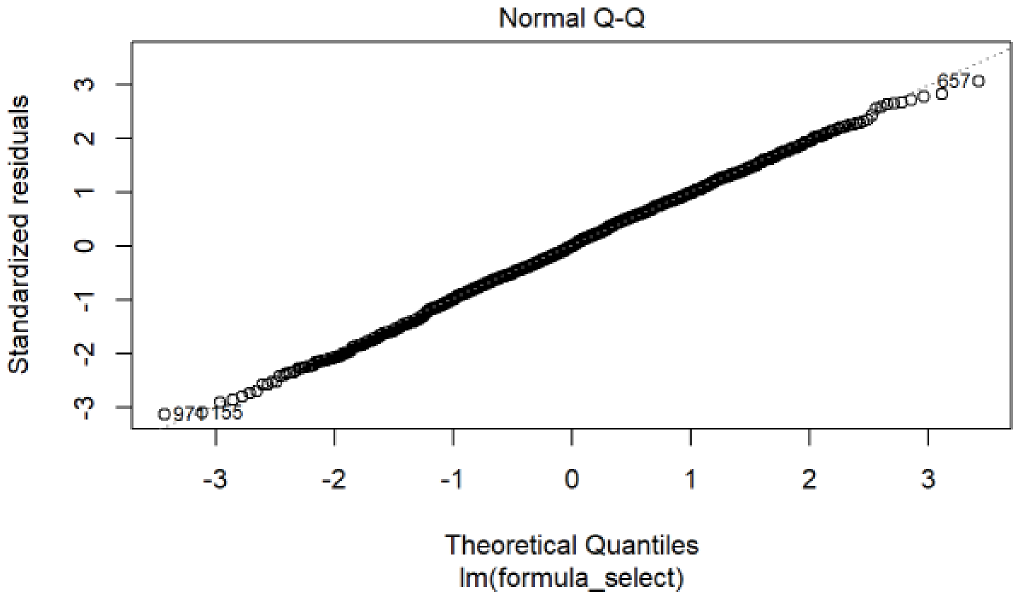
```
## Analysis of Variance Table
##
## Response: (Y^2.2 - 1)/2.2
##                Df     Sum Sq    Mean Sq   F value    Pr(>F)
## E1:E4:E5        1 3.6450e+15 3.6450e+15 2498.0851 < 2.2e-16 ***
## E1:R4:R12       1 1.2240e+12 1.2240e+12    0.8388 0.3598668
## E4:E5:E3        1 6.3519e+14 6.3519e+14  435.3211 < 2.2e-16 ***
## R4:E6:R19       1 2.2194e+13 2.2194e+13   15.2102 0.0001001 ***
## R1:R5:R15       1 1.2742e+13 1.2742e+13    8.7327 0.0031704 **
## R19:R1:R17      1 1.4362e+13 1.4362e+13    9.8428 0.0017355 **
## R15:R2:R11      1 8.8671e+12 8.8671e+12    6.0770 0.0137975 *
## R19:R3:R18      1 1.2556e+13 1.2556e+13    8.6052 0.0033990 **
## R4:R7:R16       1 1.4681e+13 1.4681e+13   10.0614 0.0015423 **
## R15:R23:R25     1 1.4244e+13 1.4244e+13    9.7622 0.0018128 **
## Residuals    1630 2.3784e+15 1.4591e+12
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

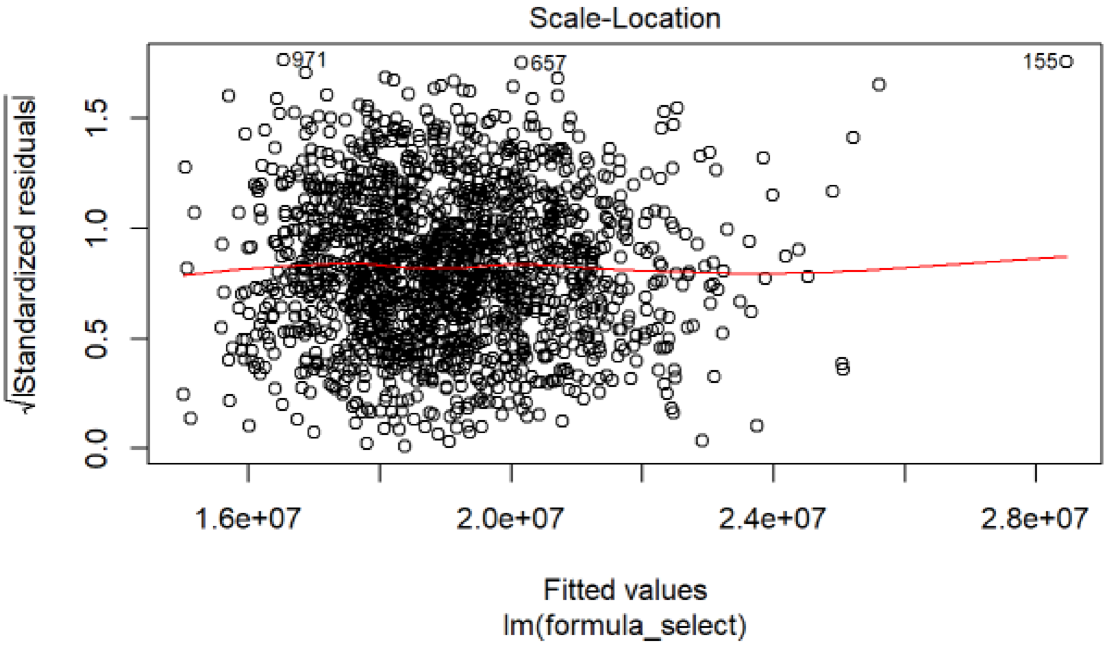


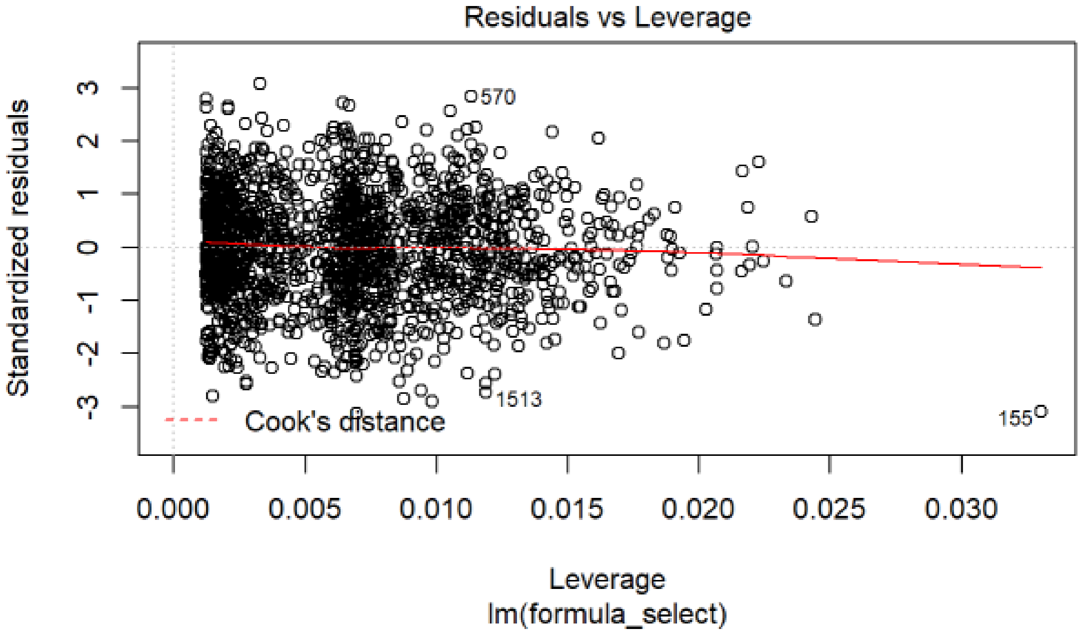
```
#comparable to the analysis before
plot(M)
```

Produces the same interaction terms as the non box cox transformation, which we know turn into the single terms after inspection by F test, residuals are comparable, but much larger


Residuals vs Fitted
Residuals
4e+06
0e+00
-4e+06
657
971
155
1.6e+07
2.0e+07
2.4e+07
2.8e+07
Fitted values
lm(formula_select)


Normal Q-Q
Standardized residuals
3
2
1
0
-1
-2
-3
657
971
155
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(formula_select)


Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
971
657
1550
1.6e+07
2.0e+07
2.4e+07
2.8e+07
Fitted values
lm(formula_select)


Residuals vs Leverage
Standardized residuals
3
2
1
0
-1
-2
-3
570
1513
155
Cook's distance
0.000
0.005
0.010
0.015
0.020
0.025
0.030
Leverage
lm(formula_select)

Full Code

```
library(naniar)
library(mice)
library(VIM)
library(readr)
library(ggpubr)
library(leaps)
library(lattice)
library(MASS)

IDEgroup <- read_csv("C:/Users/Nick/OneDrive/Documents/Spring 2021/AMS 578
Regr/Project/IDEgroup355429.csv")
IDGgroup <- read_csv("C:/Users/Nick/OneDrive/Documents/Spring 2021/AMS 578
Regr/Project/IDGgroup355429.csv")
IDYgroup <- read_csv("C:/Users/Nick/OneDrive/Documents/Spring 2021/AMS 578
Regr/Project/IDYgroup355429.csv")

IDE <- subset(IDEgroup[order(IDEgroup$ID),], select = -c(X1,ID))
IDG <- subset(IDGgroup[order(IDGgroup$ID),], select = -c(X1,ID))
IDY <- subset(IDYgroup[order(IDYgroup$ID),], select = -c(X1,ID))
dataset <- cbind(IDE,IDG,IDY)

sum.stat = summary(dataset)
sum.stat

#Missing Value Analysis
data.na =is.na(dataset)
row.sums = rep(NA,length(data.na[,1]))
for(i in 1:length(data.na[,1])){
 row.sums[i] = sum(data.na[i,])
}
col.sums = rep(NA,length(data.na[1,]))
col.sd = rep(NA,length(data.na[1,]))
for(i in 1:length(data.na[1,])){
  col.sums[i] = sum(data.na[,i])
  col.sd[i] = sd(dataset[,i],na.rm=TRUE)
}
col.sums = data.frame(col.sums,row.names = colnames(dataset))
col.sddf = data.frame(col.sd,  col.names = colnames(dataset))
col.sddf
col.sums

data.na.num = matrix(lapply(data.na, as.numeric),ncol= 32)

levelplot(t(data.na.num[1:250,]))
```

```
mat.cor = cor(dataset,use = "complete.obs")
mat.cor
mat.cor.e = cor(IDE,use = "complete.obs")
mat.cor.g = cor(IDG,use = "complete.obs")

vis_miss(dataset)
gg_miss_upset(subset(dataset,select=-
c(R5,R6,R7,R8,R9,R10,R13,R14,R15,R16,R18,R19,R20,R22,R24)),nsets=40,nintersects=50)
gg_miss_fct(dataset,fct=R5)
gg_miss_fct(dataset,fct=R6)
gg_miss_fct(dataset,fct=R8)
gg_miss_fct(dataset,fct=R25)


#Imputation

cart.impute<- mice(dataset, m=6, maxit = 10, method = 'cart', seed = 500)
summary(cart.impute)
#This generates 16 different data sets using the cart method,
#we will take the 15th but return to pool them all after
#model building
data.cart <- complete(cart.impute,5)


#Inspection
summary(data.cart)
#boxplot(log(data.cart$Y))
xyplot(cart.impute,Y~
R1+R2+R3+R4+R5+R6+R7+R8+R9+R10+R11+R12+R13+R14+R15+R16+R17+R18+R19+R
20+R21+R22+R23+R24+R25,pch=18,cex=1)
xyplot(cart.impute,Y~ E1+E2+E3+E4+E5+E6,pch=18,cex=1)
densityplot(cart.impute)


ggqqplot(data.cart$Y)

#model building

test = data.cart
M1 = lm( log(Y) ~(.)^3,data=data.cart)
leaps_int = regsubsets(model.matrix(M1)[,-1], I(log(data.cart$Y)),really.big = TRUE,
            method = "forward", nbest = 1, intercept = TRUE, nvmax = 500)

leaps_noint = regsubsets(Y~.,data=data.cart,nbest=1,nvmax = 500,method =
"exhaustive",intercept = TRUE)
```

```
final_result_int <- summary(leaps_int)


#No interactions plot
plot(leaps_noint,scale="bic")

plot(leaps_noint,scale="adjr2")

plot(leaps_int,scale="bic")

plot(leaps_int,scale="adjr2")

final_result_int$rsq
final_result_int$bic[1:20]
colnames(final_result_int$which)[final_result_int$which[10,]]

#look at models with 10 or less terms

var_chose <- colnames(final_result_int$which)[final_result_int$which[10,]]
formula_select  <- paste0('log(Y) ~ ', paste(var_chose[-1], collapse = '+') )
M <- lm(formula_select, data=data.cart)
summary(M)
anova(M)
#Keep terms with F > 16

#Hierarchical model
M2 = lm(log(Y) ~ 1 + E1+ E3 + E4 + E5 + E1:E3 + E4:E5 + R12 +R11 + R22 + R12:R11:R22 +
R11:R22 + R22:R12 + R12:R11 ,data=data.cart)
summary(M2)
anova(M2)


M3 = lm(log(Y) ~ (1 + E1 + E3 + E4 + E5 + R8:R25:R3),data=data.cart)
summary(M3)
anova(M3)
plot(M3)

#non log model
M3 = lm(Y ~ 1 + E1 + E3 + E4 + E5 ,data=data.cart)
summary(M3)
anova(M3)
plot(M3)

#
```

```
#box cox
bc <- boxcox(Y ~., data=test, lambda = seq(-2, 10, 1/10))
(lambda <- bc$x[which.max(bc$y)])
new_model <- lm(((Y^lambda-1)/lambda) ~(.)^3,data=test)
leaps_bc = regsubsets(model.matrix(new_model)[,-1], I((data.cart$Y^lambda-
1)/lambda),really.big = TRUE,
                 method = "forward", nbest = 1, intercept = TRUE, nvmax = 500)
final_result_bc <- summary(leaps_bc)

var_chose <- colnames(final_result_bc$which)[final_result_bc$which[10,]]
formula_select  <- paste0('(Y^2.2-1)/2.2 ~ ', paste(var_chose[-1], collapse = '+') )
M <- lm(formula_select, data=data.cart)
summary(M)
anova(M)
#comparable to the analysis before
plot(M)

#merging of the data
fit <- with(cart.impute, lm(log(Y) ~ E1 + E3 + E4 + E5 + 1))
combine <- pool(fit)
summary(combine)
#aov(combine)


#Here are some of the other methods i had tried
#
#null = lm(Y~1,data=data.cart)
#null
#full = lm(Y~.,data=data.cart)
#full
#step(null,scope=list(lower=null,upper=full),direction="forward")


#leaps
#summary.out <- summary(leaps)
#as.data.frame(summary.out$outmat)

#library(car)
#subsets(leaps,statistic="bic",max.size = 10,)


#library(MASS)
#fit.test = lm(Y~., data= data.cart)
#step <- stepAIC(fit.test,scope =list(upper=  . ~ .^2, lower= ~1),direction="both")
#step$anova
```

```
#library(rFSA)
#fsa.fit = FSA(Y~.,data=data.cart,fitfunc = lm,m=3,numrs = 50,criterion = BIC)
#print(fsa.fit)

#library(bestglm)
#bestglm(data.cart,IC="BIC",family=binomial)

#library(glmnet)
#f <- as.formula(Y ~ .^3)
#x <- model.matrix(f,data.cart)[,-1]
#x


#glmnet(x,data.cart$Y)
#M_LASS <- glmnet::cv.glmnet(x, data.cart$Y,nfolds = 5, alpha=1, grouped = TRUE)
#coef_select <- as.matrix(coef(M_LASS, s="lambda.1se"))
#cbind(rownames(coef_select)[coef_select > 0], coef_select[coef_select > 0])
#M_LASS$lambda.1se
#coef_select
```

Full code output provided in separate pdf document