AMS 315 Second Computing Assignment - Formal Report

Name: Parv Joshi, ID: 112169570, net-id: pvjoshi File Used: P12_69570.csv

Introduction

The report describes the best transformed linear model of the dependent variable Y on 4 environmental variables (E1–E4) and 20 genetic variables (G1–G20). I used transformations of the variables and studied up to third order interactions (gene-environment, gene-gene, environment-environment). I used stepwise regression, with adjusted R^2 and BIC values as indicators to find the relevant variables and its interactions. The statistical programming language R was used for the study of data and produce the results. After plotting the graphs, I expected the independent variable to depend on at least one environmental variable, however, the dependence of genetic variables was questionable. The objective of this report is to study interactions of the independent variables, whether, it is gene-environment, gene-gene, or environment-environment, and find the best-transformed model that associates Y with only relevant variables, i.e., the model used to generate the data.

Methodology

I was given a data set with 1072 observations of Y, 4 environmental variables (E1–E4) and 20 genetic variables (G1–G20) with no missing data. I first found the adjusted R^2 for a regression model of Y on sum of environmental variables. I then graphed the residual vs. fitted plot of Y with different exponents of the sum all variables (environmental and genetic). I ended up using the exponent 2 as any higher exponent gave a flat line with zero residual suggesting an overfit. The result had a higher adjusted R^2 suggesting the relevance of second order interactions. I then used the box-cox transformation to transform Y. The Y-transformed model had a higher adjusted R^2. The New Residual vs. fitted plot was a flat ellipse, dense at the center, suggesting the variances of the residuals are heteroscedastic and hence the model is adequate. Using stepwise regression, I found the model summary with adjusted R^2 and Bayesian Information Criterion (BIC) values to check the relevance of the variables to the model. I picked the variables whose combination gave a highest adjusted R^2 with the least BIC value, which were G19, E2, E4, G2, G4. I then checked for first, second, and third order interactions of the variables to find the final model. As expected, no third order interaction was found. The final model yielded an adjusted R^2 of 0.5724. I ended up using α value (probability of type I error) 0.05 to reject the null that the coefficients of variable and variable interactions are zero, because most of the relevant variables found had a t-value between 2 and 3, with a p-value ranging from 0.05 to 0.01. I then cross-checked my result using confidence intervals of the variables and variable interactions at $\alpha = 0.05$. The intercept term had a very low t-value, a t-value of 0.4, however, I included it as it is important to the regression results. I also used regular regression to check relevance of independent variables. Genetic variable G19 showed highly significant in the Analysis of Variance table, however, it had a high p-value of 0.09 in the final model's summary. I ended up using E2, E4, G2, and G4 but not G19 as found initially.

Results

The final model found was

$$Y^{0.3434} = 56.8871 + 10.2680 \times E2 + 6.7590 \times E4 + 0.8517 \times E2 \times E4 - 16.6731 \times G2 \times G4$$

with an Adjusted R-squared value of 0.5724 (57.24% Variance Explained), with an F-statistic value of 96.59 on 15 as DF_1 and 1056 as DF_2. The p-value of the test H_0: β_i=0, where i is the coefficient of found variables and variable interactions, against H_1: β_i≠0 is less than 2.2×10^{-16}. This helps in claiming the model is adequate. The stepwise regression model summary, final model summary table, and the analysis of variance table for the final model is shown. Also, see the Residual vs. Fitted graphs for initial data and for the final model.

Conclusion

The p-value for regression analysis is less than 2.2×10^{-16} which indicates a highly significant association between of variables and variable interactions. The Adjusted R-squared reported was 0.5724, which is a 57.24 % of variance explained. As expected, the model did depend on the environmental variables. Furthermore, the model contained environment-environment and gene-gene interactions but not gene-environment interactions. One limitation of this analysis is that an α value (probability of type I error) of 0.05 was used which made the model moderately strong than extremely strong. A stronger model would have been easier to find with a greater number of observations. 1072 observations for four environmental and twenty genetic variables appeared less. Also, this project mainly studied only up to second order variable interactions, ignoring the possibility of higher order interactions.

Stepwise Regression Model Summary (Bolded is the Model Chosen)

Model	Adjusted R^2	BIC
(Intercept)+E2:E4	0.547343445922436	-836.73719360667
Intercept)+G19+E2:E4	0.566420029053594	-876.91980687873
(Intercept)+G19+E2:E4+G2:G4	**0.569018324712021**	**-877.38925461246**
Intercept)+G19+E2:E4+G2:G4+G3:G6	0.569870507809465	-873.53796032869
(Intercept)+G19+E2:E4+G2:G4+G3:G6+G7:G20	0.570658078905811	-869.53047726346

Final Model Summary Table

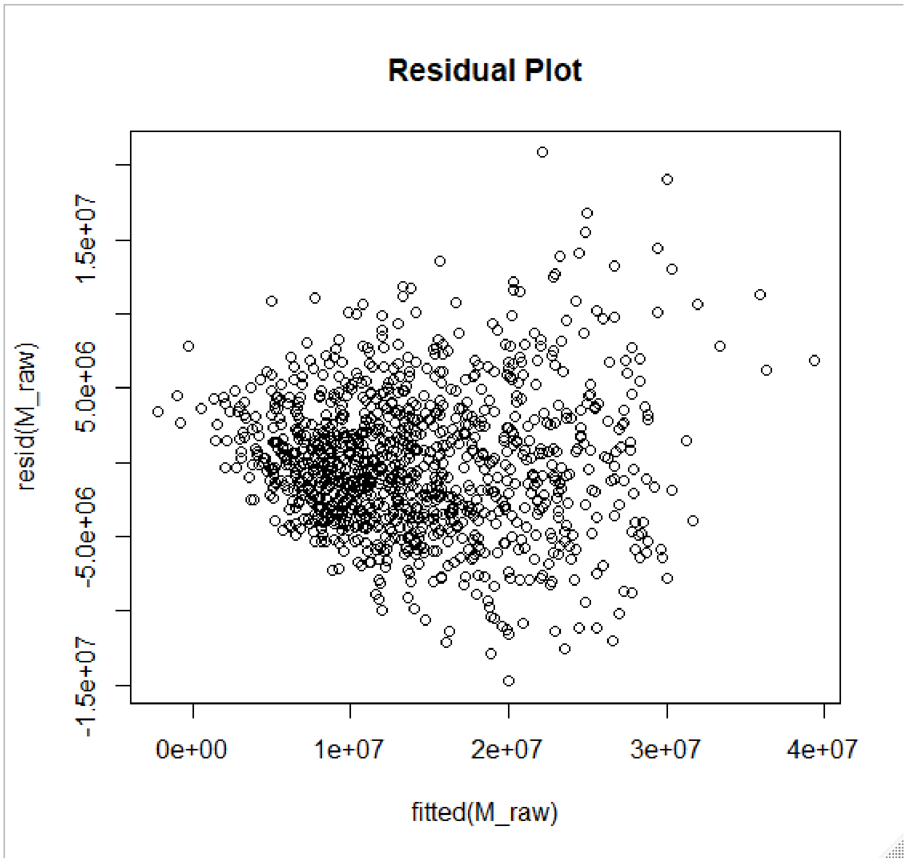
	Estimate	Standard Error	T value	P Value
E2	10.2679604	3.162086	3.247211	0.0012021
E4	6.7589594	3.164003	2.136205	0.0328921

E2:E4	0.8517006	0.289096	2.946083	0.0032889
G2:G4	-16.6731345	6.917836	-2.410166	0.0161154

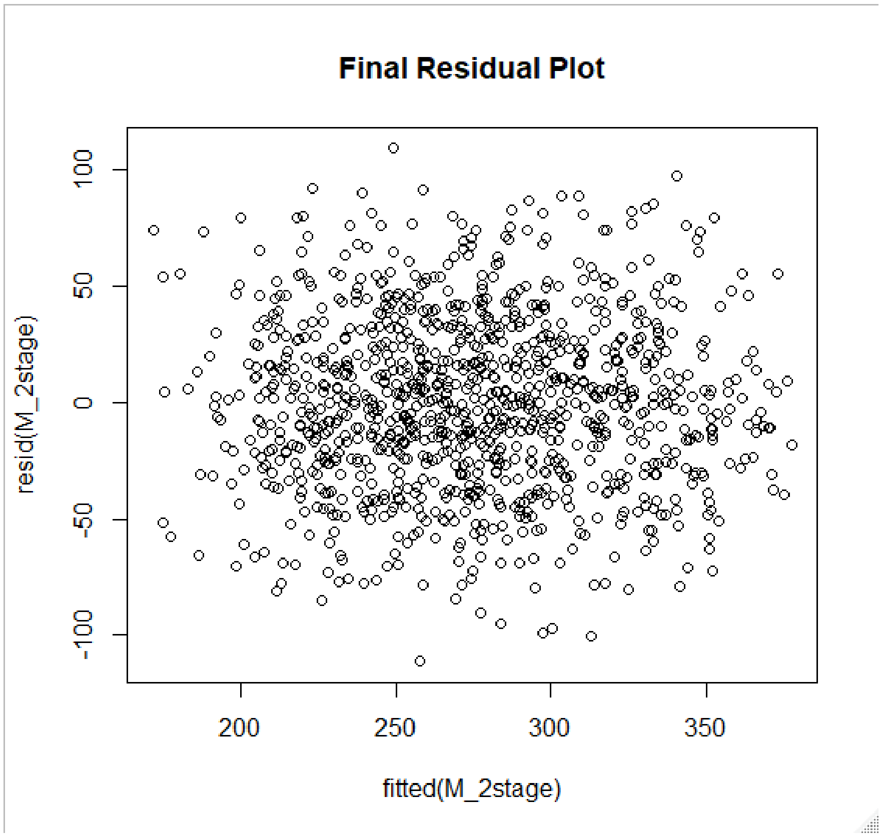
Analysis of Variance Table for Final Model (Bold: Chosen for Final Model since significant, Italic: Not Chosen although significant because high p-value in final interaction model summary)

Model	DF	Sum of Squares	Mean Square	F value	Pr(>F)
G19	*1*	*9.612881e+04*	*9.612881e+04*	*70.7860146*	*0.0000000*
E2	**1**	**1.017237e+06**	**1.017237e+06**	**749.0590335**	**0.0000000**
E4	**1**	**8.163817e+05**	**8.163817e+05**	**601.1559609|**	**0.0000000**
G2	1	6.319831e+03	6.319831e+03	4.6537107	0.0312115
G4	1	7.291001e+02	7.291001e+02	0.5368847	0.4638893
G19:E2	1	1.771321e+03	1.771321e+03	1.3043409	0.2536807
G19:E4	1	1.643566e+03	1.643566e+03	1.2102663	0.2715301
G19:G2	1	3.140977e+03	3.140977e+03	2.3129096	0.1286030
G19:G4	1	9.392409e+01	9.392409e+01	0.0691625	0.7926114
E2:E4	**1**	**1.159497e+04**	**1.159497e+04**	**8.5381429**	**0.0035519**
E2:G2	1	1.171064e+03	1.171064e+03	0.8623322	0.3532991
E2:G4	1	1.527516e+03	1.527516e+03	1.1248110	0.2891274
E4:G2	1	1.737821e+03	1.737821e+03	1.2796729	0.2582167
E4:G4	1	2.244226e+02	2.244226e+02	0.1652573	0.6844441
G2:G4	**1**	**7.888602e+03**	**7.888602e+03**	**5.8089007**	**0.0161154**
Residuals	1056	1.434069e+06	1.358020e+03	N/A	N/A

Residual vs. Fitted graph for Initial Data



Residual vs. Fitted graph for Final Model



Technical Appendix

Code:

```
wdir -> "C:\Users\Parv\Documents\Spring 2020\AMS 315\Project 2"
setwd(wdir)

Data <- read.csv('P2_69570.csv', header = TRUE)

M_E <- lm(Y ~ E1+E2+E3+E4, data=Data)
summary(M_E)
summary(M_E)$adj.r.squared

#Assuming only up to 2nd order interactions
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=Data)
plot(resid(M_raw) ~ fitted(M_raw), main='Residual Plot')

#Using Box-Cox Transformation
library(MASS)
bc=boxcox(M_raw)
best.lam = bc$x[which(bc$y==max(bc$y))]
best.lam

M_trans <- lm( (Y^(best.lam)) ~
((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=Data)
# Here is the ADJ R^2 for Raw
summary(M_raw)$adj.r.square
# Here is the ADJ R^2 for Transformed
summary(M_trans)$adj.r.square
plot(resid(M_trans) ~ fitted(M_trans), main='New Residual Plot')
```

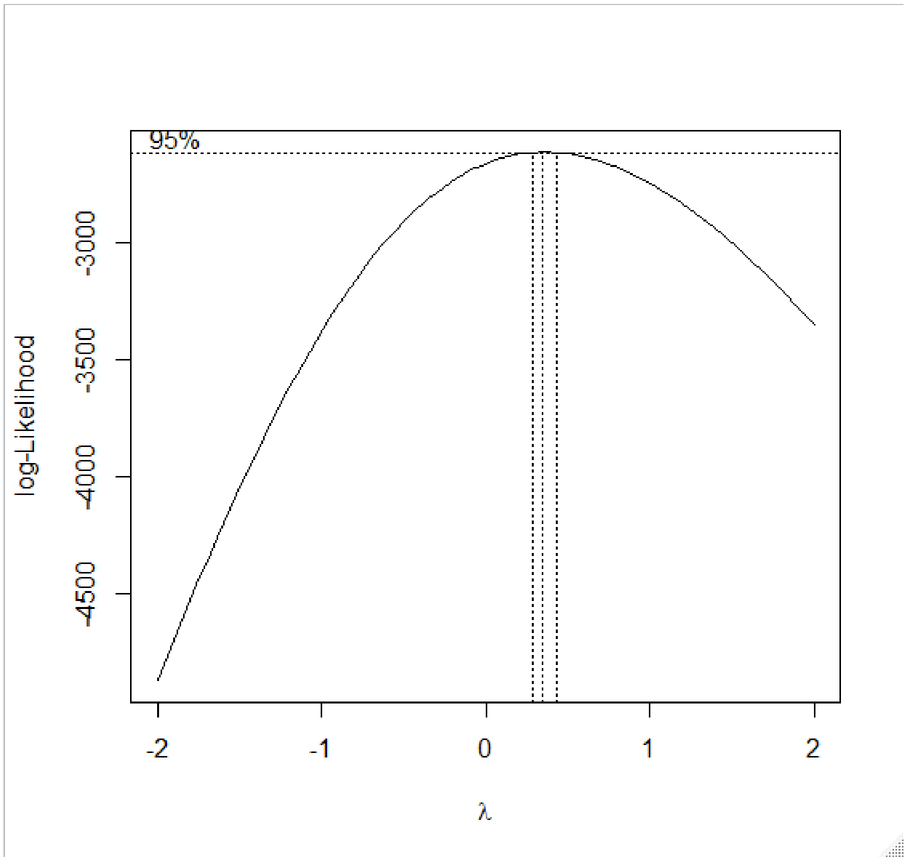
```
#Stepwise Regression
install.packages("leaps")
library(leaps)
M <- regsubsets( model.matrix(M_trans)[,-1], Data$Y^(best.lam), nbest = 1, nvmax=5, method = 'forward', intercept = TRUE )
temp <- summary(M)
temp


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( (Y^(best.lam)) ~ (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)
temp1 <- summary(M_main)
kable(temp1$coefficients[ abs(temp1$coefficients[,4]) <= 0.001, ], caption='Sig Coefficients')


M_2nd <- lm( (Y^(best.lam)) ~ (.)^2, data=Data)
temp2  <- summary(M_2nd)
kable(temp2$coefficients[ abs(temp2$coefficients[,4]) <= 0.01, ], caption='2nd Interaction')
#usage of p-value <= 0.01 is done instead of <= 0.001 since there are no variable or interactions with p-value <= 0.001
temp2
#temp2 gives p-values of all variables and its interactions


#I tried to check if there were any 3rd interactions, but I found none
```

```
M_3rd <- lm( (Y^(best.lam)) ~ (G19+E2+E4+G2+G4)^3, data=Data)
temp3  <- summary(M_3rd)
kable(temp3$coefficients[ abs(temp3$coefficients[,4]) <= 0.01, ], caption='3rd Interaction')
#usage of p-value <= 0.01 is done instead of <= 0.001 since there are no variable or interactions with p-value <= 0.001
temp3
#temp3 gives p-values of all variables and its interactions

M_2stage <- lm( (Y^(best.lam)) ~ (G19+E2+E4+G2+G4)^2, data=Data)
temp4 <- summary(M_2stage)
kable(temp4$coefficients[ abs(temp4$coefficients[,3]) >= 2, ], caption='M_2stage')
#usage of p-value <= 0.01 is done again
temp4
#temp4 gives p-values of all variables and its interactions

#Plotting Residual vs Fitted Final Model
plot(resid(M_2stage) ~ fitted(M_2stage), main='Final Residual Plot')

# Anova Table
kable(anova(M_2stage), caption='ANOVA Table')

# Result Check - Using Confidence Intervals
confint(M_2stage, level=1-0.05)

# Model Found: Y^(best.lam)) ~ 56.8871 + 10.2680*E2 + 6.7590*E4 + 0.8517*E2*E4 - 16.6731*G2*G4
# Here is a check of regular regression to confirm the relevant variables
M_check <- lm( (Y^(best.lam)) ~ (E2+E4+G19+G2+G4), data=Data)
check <- summary(M_check)
check
```

Relevant Output:

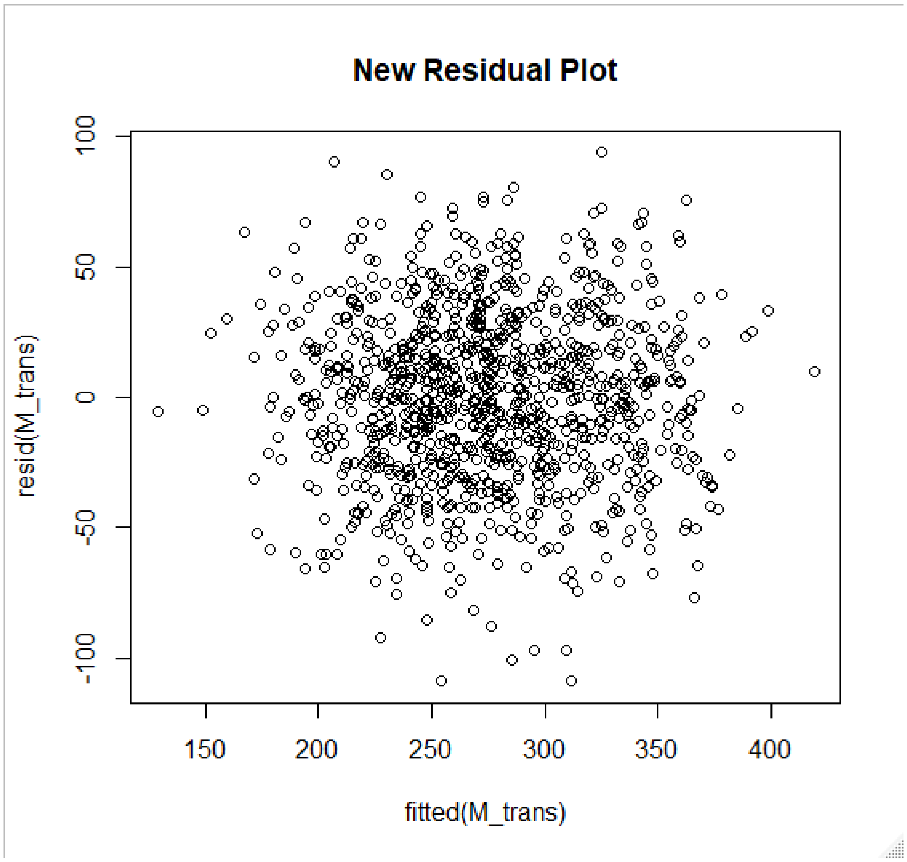
```
#Box-cox Lambda Value
> best.lam
[1] 0.3434343
```

```
#Box-cox Graph
```



```
#Comparing the Adj R^2 values
> summary(M_E)$adj.r.squared
[1] 0.5188648
> summary(M_raw)$adj.r.square
[1] 0.5368831
> # Here is the ADJ R^2 for Transformed
> summary(M_trans)$adj.r.square
[1] 0.5459324
```

```
#M_Trans Plot
```



```
#M_Select Model Summary

|model                                    |adjR2             |BIC
|
|:----------------------------------------|:-----------------|:--------------
---|
|(Intercept)+E2:E4                        |0.547343445922436 |-836.7371936066
79 |
|(Intercept)+G19+E2:E4                    |0.566420029053594 |-876.9198068787
36 |
|(Intercept)+G19+E2:E4+G2:G4              |0.569018324712021 |-877.3892546124
67 |
|(Intercept)+G19+E2:E4+G2:G4+G3:G6        |0.569870507809465 |-873.5379603286
96 |
|(Intercept)+G19+E2:E4+G2:G4+G3:G6+G7:G20 |0.570658078905811 |-869.5304772634
68 |
```

```
#M_Main Summary Table

 |    | Estimate| Std. Error|   t value| Pr(>&#124;t&#124;)|
 |:---|--------:|----------:|---------:|------------------:|
 |E2  | 14.27959|  0.5687841| 25.105461|                  0|
 |E4  | 13.83493|  0.5753004| 24.048187|                  0|
 |G19 | 18.75827|  2.8812628|  6.510434|                  0|
#M_2nd Summary Table
```

```
|         |  Estimate| Std. Error|   t value| Pr(>&#124;t&#124;)|
|:--------|---------:|----------:|---------:|-------------------:|
|E4:G12   | -6.009381|   1.894548| -3.171933|           0.0015743|
|G14:G19  | 23.032638|   8.714638|  2.642983|           0.0083847|
```

```
#M_3rd Summary Table
```

```
| Estimate| Std. Error| t value| Pr(>&#124;t&#124;)|
|--------:|----------:|-------:|-------------------:|
```

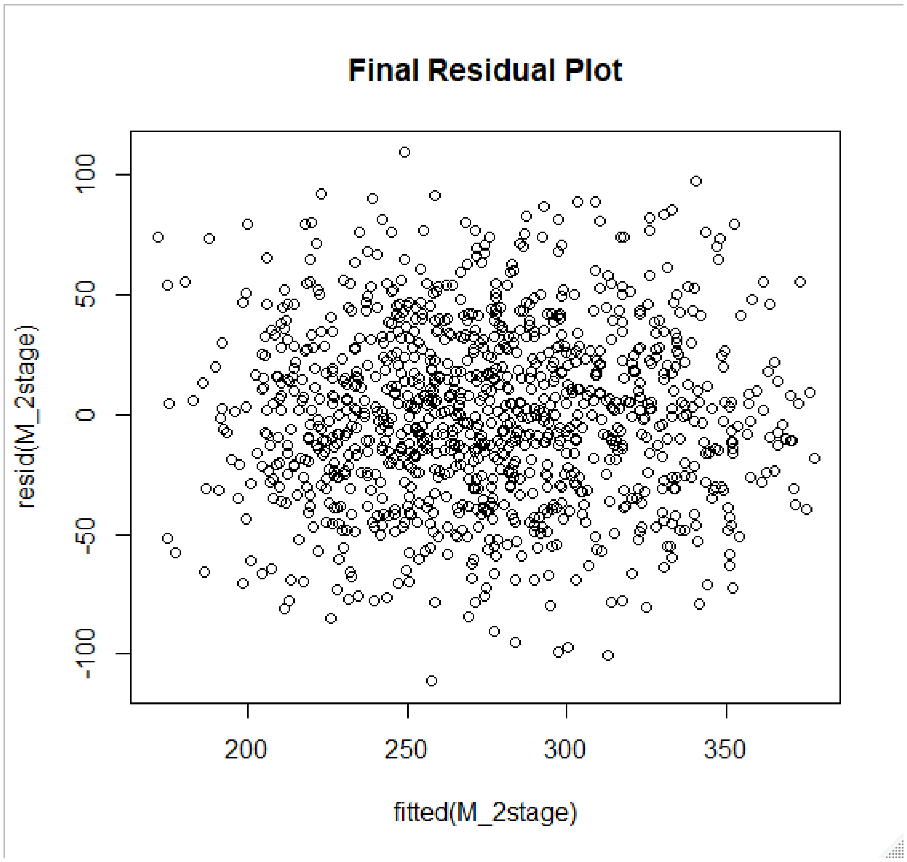
```
#M_2stage Table
```

```
|      |     Estimate| Std. Error|   t value| Pr(>&#124;t&#124;)|
|:-----|------------:|----------:|---------:|-------------------:|
|E2    |   10.2679604|   3.162086|  3.247211|           0.0012021|
|E4    |    6.7589594|   3.164003|  2.136205|           0.0328921|
|E2:E4 |    0.8517006|   0.289096|  2.946083|           0.0032889|
|G2:G4 |  -16.6731345|   6.917836| -2.410166|           0.0161154|
```

```
#M_2stage Summary Table
```

```
Call:
lm(formula = (Y^(best.lam)) ~ (G19 + E2 + E4 + G2 + G4)^2, data = Data)

Residuals:
     Min       1Q   Median       3Q      Max 
-111.501  -25.144   -0.176   24.836  109.417 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept)  56.8871    33.0085   1.723  0.08511 . 
G19          32.7101    19.2944   1.695  0.09031 . 
E2           10.2680     3.1621   3.247  0.00120 **
E4            6.7590     3.1640   2.136  0.03289 * 
G2           -3.8793    17.6113  -0.220  0.82570   
G4           26.3184    18.1736   1.448  0.14787   
G19:E2       -1.3846     1.4051  -0.985  0.32464   
G19:E4       -1.4857     1.4620  -1.016  0.30978   
G19:G2       12.3110     7.0458   1.747  0.08088 . 
G19:G4        1.2573     7.1611   0.176  0.86066   
E2:E4         0.8517     0.2891   2.946  0.00329 **
E2:G2        -1.2548     1.3238  -0.948  0.34340   
E2:G4        -1.5980     1.3923  -1.148  0.25136   
E4:G2         1.5110     1.3332   1.133  0.25730   
E4:G4        -0.3176     1.3683  -0.232  0.81649   
G2:G4       -16.6731     6.9178  -2.410  0.01612 * 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 36.85 on 1056 degrees of freedom
Multiple R-squared:  0.5784,	Adjusted R-squared:  0.5724 
F-statistic: 96.59 on 15 and 1056 DF,  p-value: < 2.2e-16
```

```
#M_2stage Plot
```



```
#M_2stage Anova Table
```

```
|           |   Df|        Sum Sq|       Mean Sq|      F value|    Pr(>F)|
|:----------|----:|-------------:|-------------:|------------:|---------:|
|G19        |    1|  9.612881e+04|  9.612881e+04|   70.7860146| 0.0000000|
|E2         |    1|  1.017237e+06|  1.017237e+06|  749.0590335| 0.0000000|
|E4         |    1|  8.163817e+05|  8.163817e+05|  601.1559609| 0.0000000|
|G2         |    1|  6.319831e+03|  6.319831e+03|    4.6537107| 0.0312115|
|G4         |    1|  7.291001e+02|  7.291001e+02|    0.5368847| 0.4638893|
|G19:E2     |    1|  1.771321e+03|  1.771321e+03|    1.3043409| 0.2536807|
|G19:E4     |    1|  1.643566e+03|  1.643566e+03|    1.2102663| 0.2715301|
|G19:G2     |    1|  3.140977e+03|  3.140977e+03|    2.3129096| 0.1286030|
|G19:G4     |    1|  9.392409e+01|  9.392409e+01|    0.0691625| 0.7926114|
|E2:E4      |    1|  1.159497e+04|  1.159497e+04|    8.5381429| 0.0035519|
|E2:G2      |    1|  1.171064e+03|  1.171064e+03|    0.8623322| 0.3532991|
|E2:G4      |    1|  1.527516e+03|  1.527516e+03|    1.1248110| 0.2891274|
|E4:G2      |    1|  1.737821e+03|  1.737821e+03|    1.2796729| 0.2582167|
|E4:G4      |    1|  2.244226e+02|  2.244226e+02|    0.1652573| 0.6844441|
|G2:G4      |    1|  7.888602e+03|  7.888602e+03|    5.8089007| 0.0161154|
|Residuals  | 1056|  1.434069e+06|  1.358020e+03|           NA|        NA|
```

```
#M_2stage Confidence Intervals (For checking if model is correct)


> confint(M_2stage, level=1-0.05)
```

```
                  2.5 %     97.5 %
(Intercept)  -7.8825797 121.656795
G19          -5.1496365  70.569856
E2            4.0632740  16.472647
E4            0.5505111  12.967408
G2          -38.4363284  30.677803
G4           -9.3421364  61.978937
G19:E2       -4.1417101   1.372488
G19:E4       -4.3544414   1.383120
G19:G2       -1.5143516  26.136338
G19:G4      -12.7943572  15.309022
E2:E4         0.2844327   1.418968
E2:G2        -3.8524302   1.342758
E2:G4        -4.3300534   1.134117
E4:G2        -1.1049456   4.126980
E4:G4        -3.0025452   2.367319
G2:G4       -30.2474026  -3.098866
```

```
#M_chcek (Regular regression to check relevant variables)

Call:
lm(formula = (Y^(best.lam)) ~ (E2 + E4 + G19 + G2 + G4), data = Data)

Residuals:
     Min       1Q   Median       3Q      Max 
-118.272  -25.126   -0.373   24.215  102.925 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  27.4235     7.6459   3.587  0.00035 ***
E2           14.1846     0.5620  25.238  < 2e-16 ***
E4           13.8314     0.5691  24.305  < 2e-16 ***
G19          19.2333     2.8465   6.757 2.32e-11 ***
G2           -5.8020     2.6916  -2.156  0.03134 *  
G4           -2.0885     2.8672  -0.728  0.46652    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 37.07 on 1066 degrees of freedom
Multiple R-squared:  0.5694,	Adjusted R-squared:  0.5673 
F-statistic: 281.9 on 5 and 1066 DF,  p-value: < 2.2e-16
```

References:

1. Class Handout: https://blackboard.stonybrook.edu/bbcswebdav/pid-5337763-dt-content-rid-41358946_1/courses/1204-AMS-315-SEC01-49021/Multiple%20Regression%20Handout%20S2020%282%29.html
2. Caspi_et_al._2003_Science Document: https://blackboard.stonybrook.edu/bbcswebdav/pid-5307826-dt-content-rid-40485137_1/courses/1204-AMS-315-SEC01-49021/Caspi_et_al._2003_Science.pdf
3. Risch_et_al._2009 Document: https://blackboard.stonybrook.edu/bbcswebdav/pid-5307826-dt-content-rid-40485138_1/courses/1204-AMS-315-SEC01-49021/risch_et_al._2009.pdf
4. Reporting Statistical Information in Medical Journal Articles Document: https://blackboard.stonybrook.edu/webapps/blackboard/content/listContent.jsp?course_id=_1204877_1&content_id=_5184750_1&mode=reset