Matthew Abruzzese Ott (111520701)
Professor Stephen Finch

AMS 315 Project 2

Introduction

The objective of Project Two is to estimate the multiple regression function that was used to generate the given data set. The data set contains twenty-four independent variables; four environmental variables (E1-E4), and twenty gene indicator variables (G1-G20), as well as one dependent variable, Y. Second order interactions of variables (for example, E1*E3, G2*G13, E2*G10) may be included in the multiple regression model of the data. The statistical packages R and SPSS will be used during this data processing.

Methodology

For all code, statistical summaries and tables, plots, and graphs, see the technical appendix. First, the data file containing the values for the dependent and independent variables was converted into a ".txt" file and imported into R. To see how the environmental variables affected the dependent variable Y, multiple regression was conducted with regards to these four independent variables, and the summary statistics were generated. The scatter and residual plots of Y versus each of the environmental variables were also generated using R and SPSS. Next, to consider the relation between Y and all of the independent variables, the multiple regression of E1-E4 and G1-G20 on Y was conducted allowing for second order interactions between the independent variables. The residual plot was generated, and it displayed that a transformation to the dependent variable was needed. Therefore, a Box-Cox Transformation was conducted to identify the proper transformation. To analyze the transformation, the residual plot for the regression of the independent variables (allowing for second order interactions) on the transformed dependent variable was generated. Next, stepwise regression was conducted to determine which independent variables were important and had a strong effect on the transformed dependent variable. A summary of potential models with their respective adjusted R-squared and BIC values was then created. These models were assessed, and analysis was carried out to identify which main effects (single independent variables) and second order interactions of independent variables were significant in determining the transformed dependent variable (using significance level $\alpha = 0.01$). From these analyses, a final model for the data was created. Upon conducting the final model's multiple regression, its summary statistics, ANOVA table, and residual plot were generated.

Results

From the analysis displayed in the technical appendix, the Box-Cox Transformation revealed that the dependent variable needed to be transformed from Y to sqrt(Y) because the approximate lambda value was 0.5. The independent variables that were significant (had very low p-values and large t-values), and therefore considered for the final model, were E1 and G11*G15. The multiple regression function that was used to generate the given data set is of the form $Y=B_0+B_1(X_1)+B_2(X_2)$. With respect to the data, the final model of the dependent variable is expressed as sqrt(Y)=(-228.55)+160.01(E1)+178.75(G11*G15). The 99% confidence intervals for B_0, B_1, and B_2 are [-346.5697 , -110.5338], [147.4731 , 172.5495], and [98.05107 , 259.4439] respectively. It can be observed (see ANOVA table) that the F-values of E1 and G11*G15 are rather large (F-values >= 6.655042), supporting the decision to reject the null hypothesis that B_i = 0 (i=1,2). The adjusted R-squared value was 0.4704, so approximately 47% of the variation of the dependent variable was explained. The ANOVA table of the final model shows that both independent variables have a significant association with the dependent variable (p-values = 0 < 0.01). The validity of this model can be verified qualitatively through its residual plot.

Conclusion

According to the data analysis, it was observed that there was a notable relationship between the independent variables E1 and G11*G15, and the dependent variable sqrt(Y). From this, it was concluded that the model sqrt(Y)=(-228.55)+160.01(E1)+178.75(G11*G15) properly describes the correlation between the dependent and independent variables. As displayed in the model, only environmental main effects and genetic second order interactions were significant in determining the dependent variable. No genetic main effects, environmental second order interactions, or gene-environment interactions were present in the model, and therefore were not significant in determining the dependent variable.

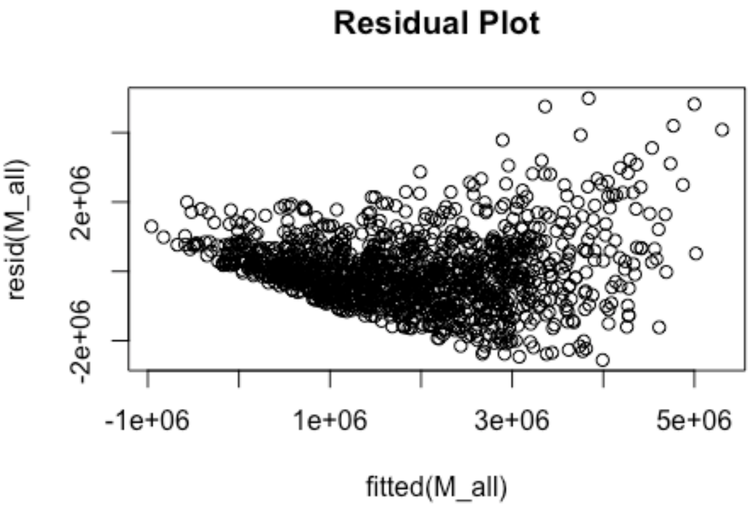
Throughout this data processing, multiple limitations were encountered. First, the scatter and residual plots of Y vs. E_i (i=1,2,3,4) revealed that E1 would be the most significant environmental variable in determining DV. After transforming the dependent variable into sqrt(Y), multiple transformations were conducted on E1 ($(E1)^2$, $(E1)^{1/2}$, ln(E1), log(E1)) to see if the adjusted R-squared value of the model would improve. However, it was never better than the original adjusted R-squared value, so E1 was used in the final model. After conducting the multiple regression of all independent variables on Y, allowing for third order interactions, the residual plot revealed that there were no third order interactions in my model. Therefore, only main effects and second order interactions of the independent variables should be considered. The final limitation occurred when deciding which significance level should be used to determine main effects, second order interactions, confidence intervals, and F-tests. It was deduced that a "safe" level of significance was $\alpha = 0.01$. At $\alpha = 0.001$ and $\alpha = 0.005$ no main effects nor second order interactions were displayed because the α-level was too low. If $\alpha = 0.05$ was used, then there was the risk of obtaining a false positive that the main effects and second order interactions that appeared to be significant in the final model were not actually significant because the α-level was too high.

Technical Appendix

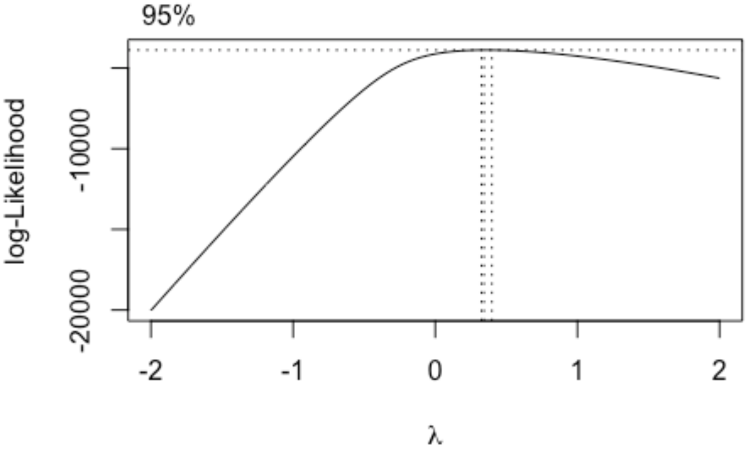
(referencing the HTML handout on Blackboard [by Songzhu Zheng])

Code and Output from R

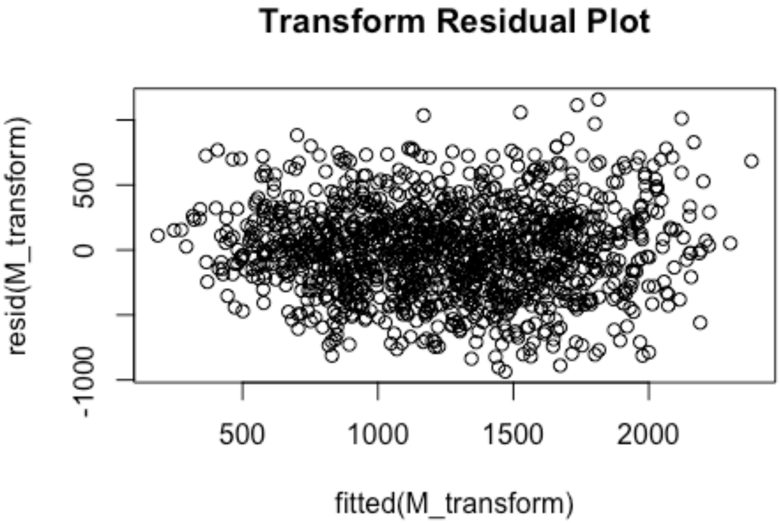
```
> setwd("/Users/mattabruzzeseott/Documents/AMS 315")
> Mydata <- read.table("P2_20701.txt", header = TRUE)
> M_Env <- lm(Y ~ E1+E2+E3+E4, data = Mydata)
> summary(M_Env)

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

Residuals:
    Min       1Q      Median      3Q       Max
-2703426  -738378  -131185   530476  6903025

Coefficients:
              Estimate   Std. Error   t value    Pr(>|t|)
(Intercept) -1811318    260904      -6.942     6.16e-12 ***
E1            421881     14489      29.117      < 2e-16 ***
E2             21252     14660       1.450       0.1474
E3             -1638     14394      -0.114       0.9094
E4            -36728     14565      -2.522       0.0118 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1185000 on 1258 degrees of freedom
Multiple R-squared:  0.4055,  Adjusted R-squared:  0.4036
F-statistic: 214.5 on 4 and 1258 DF,  p-value: < 2.2e-16

> M_all <- 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 = Mydata)
> plot(resid(M_all) ~ fitted(M_all), main='Residual Plot')
```



```
> library(MASS)
> boxcox(M_all)
```



```
> M_transform <- lm( I(sqrt(Y)) ~ (.)^2, data = Mydata)
> summary(M_transform)$adj.r.squared
[1] 0.4759795
> plot(resid(M_transform) ~ fitted(M_transform), main='Transform Residual Plot')
```



```
> library(leaps)
> M <- regsubsets( model.matrix(M_transform)[,-1], I(sqrt(Mydata$Y)), nbest = 1, nvmax = 5,
method = 'forward', intercept = TRUE)
> temp <- summary(M)
> library(knitr)
> Var <- colnames(model.matrix(M_transform))
> M_selected <- apply(temp$which, 1, function(x) paste0(Var[x], collapse = '+'))
> kable(data.frame(cbind( model = M_selected, adjR2 = temp$adjr2, BIC = temp$bic)),
caption='Model Summary')
```

```
                              Model Summary

|model                                            |adjR2              |BIC                |
|:----------------------------------------------- |:----------------- |:------------------|
|(Intercept)+E1                                   |0.457112741734882 |-758.226805911505 |
|(Intercept)+E1+G11:G15                           |0.470406313907362 |-783.399320268662 |
|(Intercept)+E1+G1:G13+G11:G15                    |0.474508038211148 |-787.080917321283 |
|(Intercept)+E1+G1:G13+G11:G15+G17:G18            |0.477400371468473 |-787.914062722654 |
|(Intercept)+E1+G1:G13+G7:G16+G11:G15+G17:G18 |0.479797052165871 |-787.582725687655 |
```

```
> M_mainE <- lm( I(sqrt(Y)) ~ ., data = Mydata)
> temp <- summary(M_mainE)
> kable(temp$coefficients[ abs(temp$coefficients[,4]) <= 0.01, ], caption = 'Sig. Coefficients')
```

```
                    Sig. Coefficients

|             |     Estimate|  Std. Error|      t value|       Pr(>|t|)|
|:----------- |------------:|-----------:|------------:|--------------:|
|(Intercept) |   -247.61198|    95.95956|    -2.580378|      0.0099835|
|E1           |   160.73534|     4.93782|   32.551882|      0.0000000|
|G15          |    60.21789|    23.23318|     2.591892|      0.0096573|
> M_2order <- lm( I(sqrt(Mydata$Y)) ~ (.)^2, data = Mydata)
> temp <- summary(M_2order)
> kable(temp$coefficients[ abs(temp$coefficients[,4]) <= 0.01, ], caption = 'Second Order
Interactions')
```

Second Order Interactions

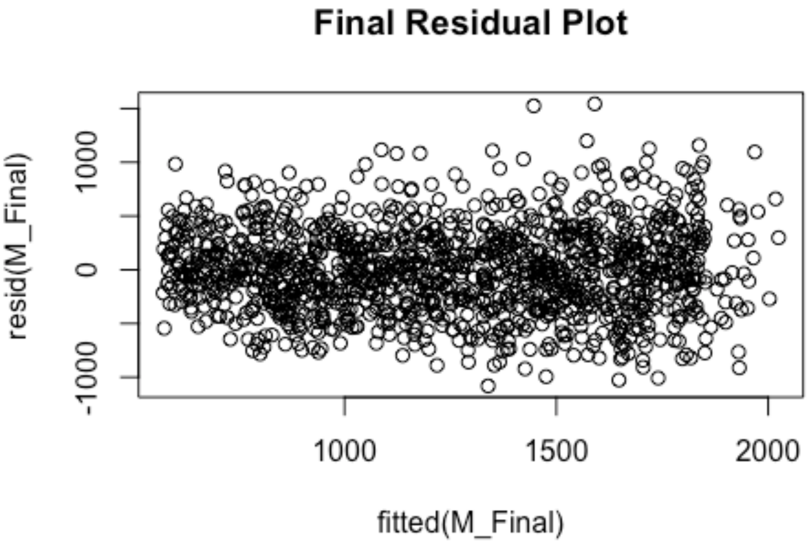
```
|             |  Estimate| Std. Error|      t value|            Pr(>|t|)|
|:------------|-----------:|------------:|------------:|--------------------:|
|E1           | 236.3413|   42.05675|  5.619580|        0.0000000|
|G8:G9        | 138.7508|   52.99915|  2.617982|        0.0089841|
|G11:G15      | 223.2003|   53.50799|  4.171346|        0.0000330|
```

```
> M_var <- lm( I(sqrt(Mydata$Y)) ~ (E1+G8+G9+G11+G15)^2, data = Mydata)
> temp <- summary(M_var)
> temp$coefficients[ abs(temp$coefficients[,3]) >= 4,]

             Estimate   Std. Error       t value           Pr(>|t|)
E1           168.8619     9.005014   18.751987    2.607448e-69
G11:G15      227.4030    47.206061    4.817242    1.634115e-06

> M_Final <- lm(sqrt(Y) ~ (E1+G11:G15), data = Mydata)
> summary(M_Final)

Call:
lm(formula = sqrt(Y) ~ (E1 + G11:G15), data = Mydata)

Residuals:
    Min      1Q    Median     3Q     Max
-1081.0  -277.2    -13.9     261.0  1543.4

Coefficients:
             Estimate     Std. Error     t value      Pr(>|t|)
(Intercept)  -228.55        45.75         -4.996     6.68e-07 ***
E1            160.01         4.86         32.922      < 2e-16 ***
G11:G15       178.75        31.28          5.714     1.37e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 397.4 on 1260 degrees of freedom
Multiple R-squared:  0.4712,  Adjusted R-squared:  0.4704
F-statistic: 561.5 on 2 and 1260 DF,  p-value: < 2.2e-16
```

```
> confint(M_Final, '(Intercept)', level = 0.99)
                    0.5 %        99.5 %
(Intercept)   -346.5697    -110.5338
> confint(M_Final, 'E1', level = 0.99)
         0.5 %       99.5 %
E1  147.4731   172.5495
> confint(M_Final, 'G11:G15', level = 0.99)
                 0.5 %        99.5 %
G11:G15   98.05107   259.4439

> kable(anova(M_Final), caption = 'ANOVA Table')
```

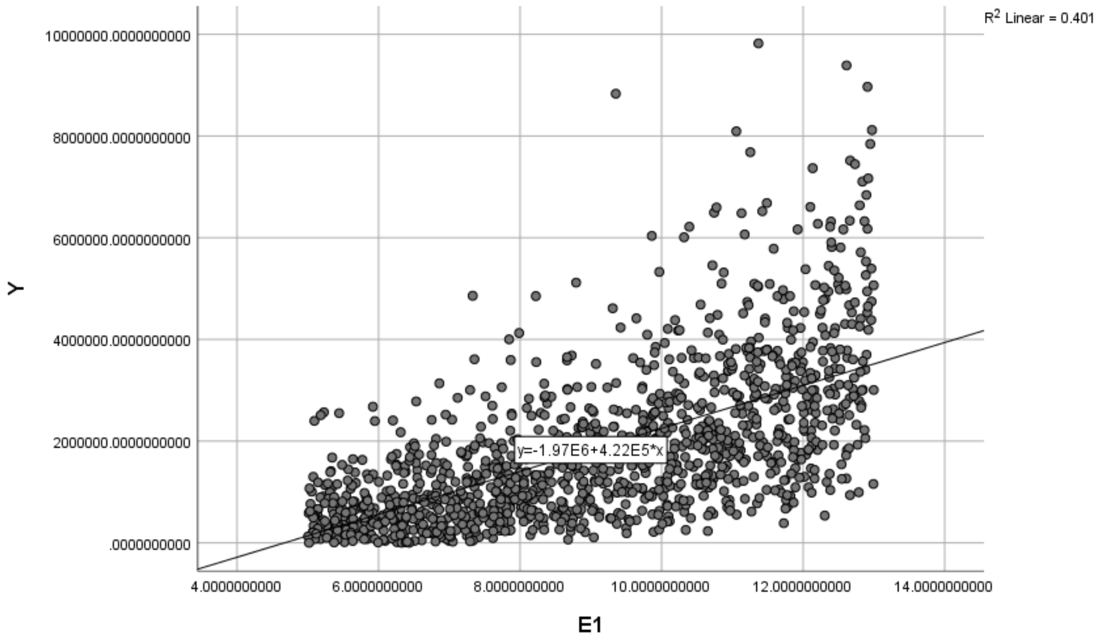
ANOVA Table

```
|              |    Df|       Sum Sq|        Mean Sq|       F value| Pr(>F)|
|:-------------|-----:|------------:|--------------:|-------------:|-------:|
|E1            |     1|  172163455| 172163454.9| 1090.30599|        0|
|G11:G15  |      1|      5156022|      5156022.4|    32.65293|        0|
|Residuals  |  1260| 198958784|        157903.8|             NA|     NA|

> qf(0.99, df1 = 1, df2 = 1260)
[1] 6.655042

> plot(resid(M_Final) ~ fitted(M_Final), main='Final Residual Plot')
```

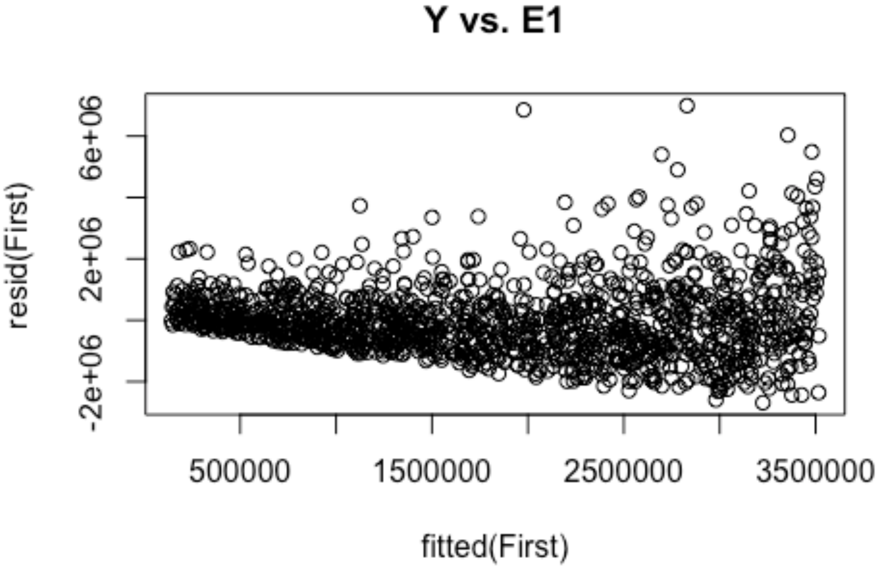


Scatter Plots (SPSS) and Residual Plots (R)

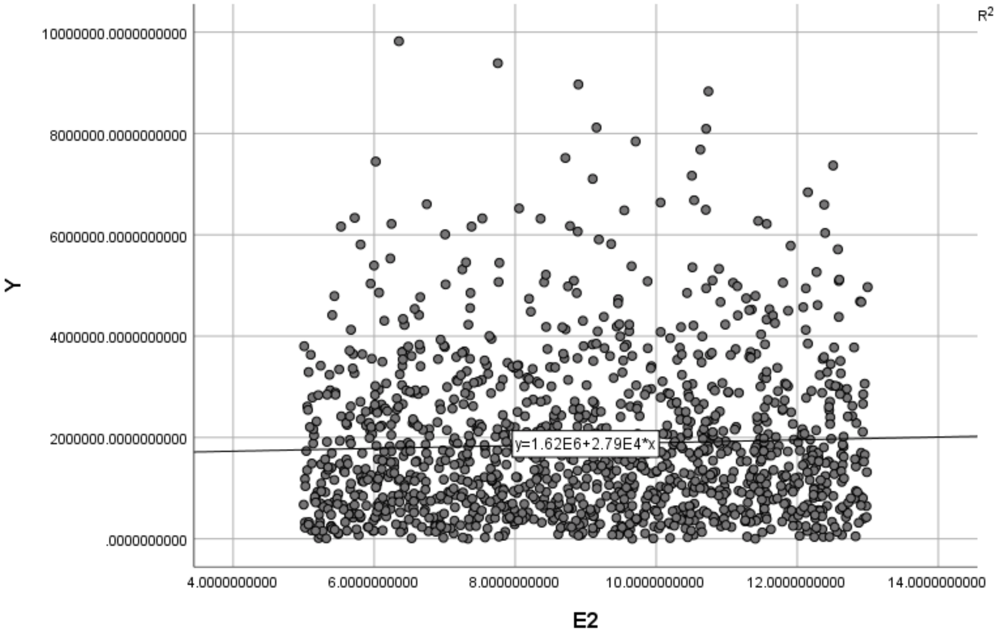
Y vs. E1



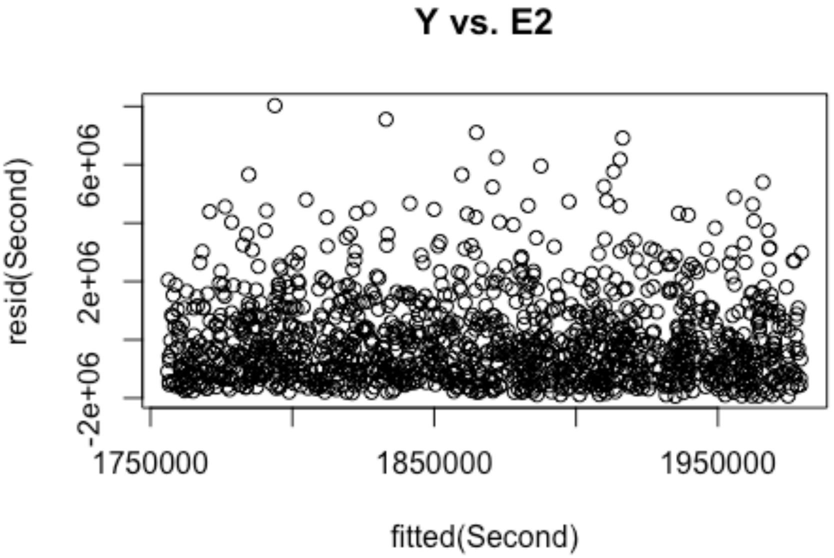
```
> First <- lm(Y ~ E1, data = Mydata)
> plot(resid(First) ~ fitted(First), main='Y vs. E1')
```



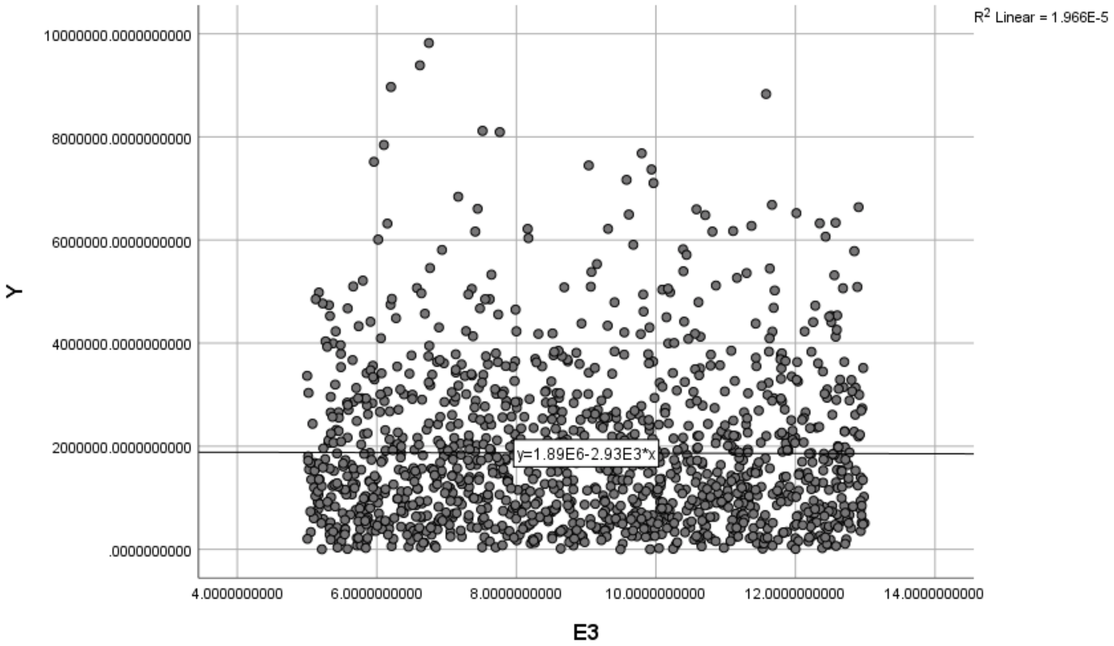
Y vs. E2



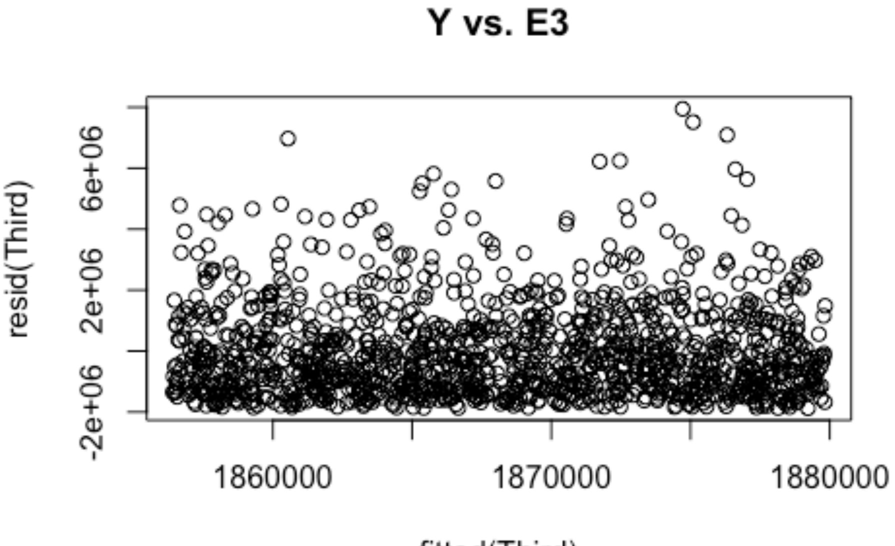
```
> Second <- lm(Y ~ E2, data = Mydata)
> plot(resid(Second) ~ fitted(Second), main='Y vs. E2')
```



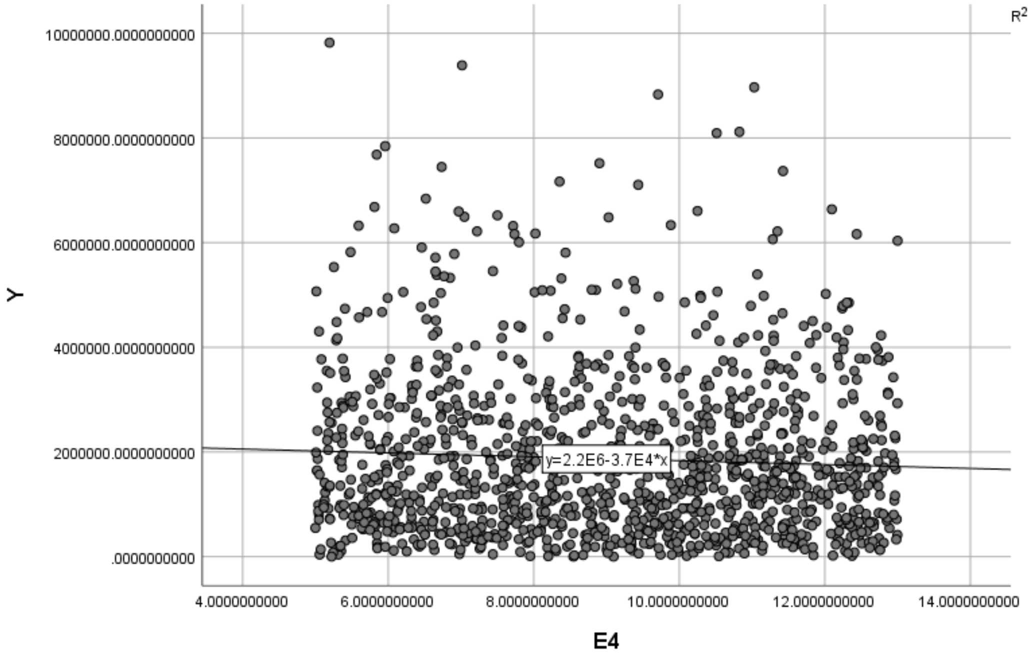
Y vs. E3



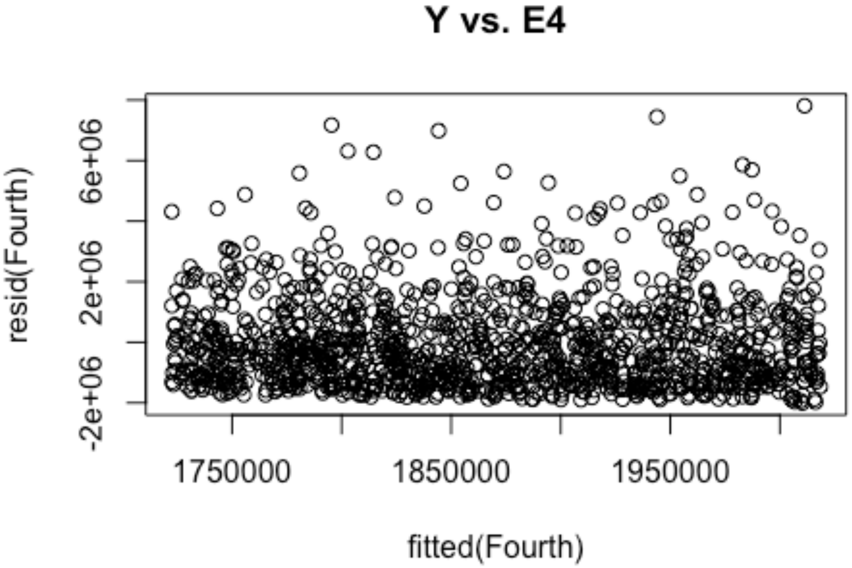
```
> Third <- lm(Y ~ E3, data = Mydata)
> plot(resid(Third) ~ fitted(Third), main='Y vs. E3')
```



Y vs. E4



```
> Fourth <- lm(Y ~ E4, data = Mydata)
> plot(resid(Fourth) ~ fitted(Fourth), main='Y vs. E4')
```


Y vs. E4
resid(Fourth)
-2e+06
2e+06
6e+06
fitted(Fourth)