AMS 578 Regression Analysis: Multiple Regression Computing Project

Michael Tartamella

05/03/2022

Abstract

Every day, people all around the world may struggle with mental illness. These types of illnesses are often difficult to fully understand, yet so many individuals are suffering from them daily. These issues are often seen especially in college, where students are put under an immense amount of pressure to succeed, which can be very mentally demanding. This can ultimately lead them to having issues such as depression, anxiety, stress, and much more (McLafferty, 2017). This makes it imperative for researchers to try to understand what factors potentially give rise to these type of illnesses. This field of study is difficult for researchers to fully grasp. In this paper, I plan to hopefully shed some light on the subject of analyzing this type of data. Hopefully, through my efforts and the efforts of other researchers we collectively begin to better understand the components which create mental illness.

Introduction

The purpose of the exploratory analysis in this paper is to discover and replicate the model used in order to generate our outcome variable Y. In this project I will be performing a lot of major analyzes on our data in order to bring to light the interactions between our response variable vs. all the different interactions possible between our environmental variables and gene variables. These interactions will include response vs. environmental variables, genes variable, gene-environment variables, gene-gene interactions variables, and possibly interactions which consist of up to four variable interactions among our variables. One Method we will use is Multiple Linear Regression. The inspiration for this report comes from the paper "Influence of Life Stress on Depression: Moderation by Polymorphism in the 5-HTT Gene", by Avshalom Caspi, where data similar to mine was analyzed in order to determine the affect of environment factors and genetic factors on depression (Caspi, 2013). Although, the findings from that study were unable to be replicated from other researchers in the field.

Variables

$Y -$ Response/outcome variable

E_1 to $E_8 -$ Environmental variables for participants

G_1 to $G_{30} -$ Genetic independent indicator/dummy variables

Methodology and Data

Our data was synthetically generated and consists of a total of 2675 entries of data. These entries of data are complete meaning there is no presence of missing data. Each entry consists of an outcome variable, eight environmental variables, and thirty indicator variables. Our goal is to generate a model that best estimates our outcome variable Y. There may also be the presence of up to four-way interactions variables. There is a total of 240 possible gene-environment variables, $\binom{30}{2} = 435$ possible gene-gene interaction variables, and up to any possible combination of four-different variables. The TA used a model to generate our data, therefore I will attempt to best replicate the model used by the TA.

There exists many different statistical methods in which we could use to analyze this data set. However, in order to best accomplish our task of generating a suitable model to estimate our response variable, I will analyze the data using multiple linear regression with non-indicator and indicator variables. Multiple linear regression uses a regression model to analyze aspects between variables. This strategy utilizes more than one independent variable in which we believe affects one dependent variable (Montgomery, 2000, pg. 67). We will be using this method to test some important characteristics of our data such as normality assumption, analysis of variance, plotting residuals, normal QQ plot, as well as the correlation between our variables. Using the software language R, I plan to accomplish

these tasks and uncover a model which best fits our data. Additionally, since we have 38 variables, we will consider the Bonferroni-corrected p-value, meaning that when I say that a particular variable is significant, that means that the p-value multiplied by 38, is still less than 0.01. As a final note regarding our methodology we may also implement the use of box cox or log transforming our data, which we may utilize in order to reduce the skewness of our data. This can help us get a better understanding of the relationship between our variables.

Analysis of Data

Our first step is to activate the various R packages used in the project and read in our data from the excel file, as seen in Appendix A (1).

Now, I will analyze the data for any missing values by using the summary function, seen in Appendix A (2)

This function has shown me that the data set is complete and we will not need to impute any data or remove any entries from the data set. Next, I plan to analyze a regression model one by one using our dependent variable Y vs each of our eight environmental variables.

After analyzing the results found in Appendix A (3), we see that the environmental variables, E_1, E_6, and E_8, appear to significantly affect the value of our response variable Y.

Next, I will preform a T-test comparing the expected value of Y given that one of our genetic indicator variables is not included vs. expected value of Y given that one of our genetic indicator variables is included. We will perform this for all thirty genetic/indicator variables.

After analyzing the results found in Appendix A (4), we determine that the only indicator variable which appears to significantly affect the expected value of our outcome variable Y is indicator G_{21}.

I will also check to see if raising our environmental variables to the power of 2 or taking the square root of our environmental variables affects the significance of any of our environmental variables.

Observing the results found in Appendix A (5), it is clear that the act of squaring or taking the square root of any of our environmental variables did not affect which variables appeared to significantly affect our response variable Y.

Before trying to model the data only exclusively using these significant variables. I believe it is essential to check normality assumptions about our model as a whole utilizing every variable, in order to determine if our data needs to be transformed in any way. Various tests and plots in order to verify assumptions can be seen in Appendix A (6).

The first model assumption to check is the residuals having a normal distribution. As seen from the plot, the residual plot of the full model shows no significant pattern, which leads to the confirmation of the normal distribution assumption for the residuals. There is no

significant violation according to the residual plot. The next model assumption to check is that the residuals are uncorrelated. Again, from our residual plot, we do not see any sort of pattern in the plot as it seems that the residuals are distributed completely randomly in these plots. Proving that there is no correlation between the residuals. The next, and final, model assumptions to check is that the residuals have constant variance. By looking at the residual vs. fitted values plot in Appendix A (6), there does not seem to be any noticeable pattern in the plot suggesting that this assumption is fulfilled. Also, observe the QQ plot in Appendix A (6) to see that the residuals fit the QQ line very well, meaning that our model seems to fit a normal distribution well. Next assumption to check is the error term having a zero mean. As you can see above, the mean of the residuals is an extremely low number, -1.298201e-16, which can be approximated to zero, confirming this assumption.

While every assumption has been fulfilled in regard to our data fitting a normal approximation, I will still check to see if it is necessary to preform a box-cox transformation on our data. I perform this test in Appendix A (7).

As we can see in the plot found in the Appendix, since our model is within a 95% confidence interval of optimal $\lambda = 1$, we conclude that a box-cox transformation is not necessary. Our data fits the model well enough that a transformation of our data is not necessary in order to generate valid results. We have concluded that the data is clean and we do not have any outliers that are skewing our model. Now I must specify a plan in order to detect the interactions between our numerous variables. I could construct a model which measures all possible regressions and interactions. This method would not be very efficient as we have so many variables in data, that we are sure to run into multiple issues regarding multicollinearity and lack of significance. There also exists stepwise regression methods. I could use forward selection which involves starting with no variables in the model and adding variables one at a time, checking the significance of the model after each time a variables is introduced. This method could be very useful. Due to the vastness of the size of our data, it would be easier to start with an empty model and add variables to see if they are significant. Another possible method is backwards elimination which begins with a model which contains all variables and aims to delete variables one by one until a suitable model is acquired. It is unrealistic to go start with such a large model and remove only one variable at a time, using data of this size. What I plan to use is the type of stepwise regression forward regression and rely heavily on the Bayesian information criterion (BIC). This strategy aims to discover a simple model which aims to narrow the model down to only the essential variables which have a great impact on our data. When selecting variables for our model, it is possible to increase likelihood through the addition of new parameters, however these new parameters often result in overfitting. The method Bayesian information criterion solves this issue by creating a penalty term based on the number of parameters in the model. The formula used is $BIC = -2ln(L) + pln(n)$, where L is maximized value of the likelihood function for the given model, p represents number of free parameters, and n represents number of data points. This concept of introducing a penalty term based on number of parameters is also seen in the Akaike Information criterion, however for BIC the penalty term has a larger impact on the suggested model (What is Bayesian Information, 2019). We will now apply forward selection to our model and analyze the BIC value of the suggested models, as seen in Appendix B (1).

For the Bayesian information criterion value a lower value estimates a better model for our data. Notice how for the model with the lowest BIC value, the adjusted R squared is the highest value out of all the models in the table. Therefore, if we were to choose a model strictly based on the Bayesian information criterion, It would be $Y = \alpha + \beta_1 E_6 + \beta_2 E_8 + \beta_3 E_1 E_8 + \beta_4 E_1 G_{21} + \beta_5 E_6 E_8$. Lets denote this as model 1.

However, I believe it is essential to analyze the significance of interactions that may or may not appear in the models displayed in the table.

As expected analyzing this table in Appendix B (2), we see that the only variables in our data which have significant p-values and high t values are E_1, E_6, E_8, and G_{21}. Next, I plan to check the significance of variables which include interactions with at most one other variable.

Observing the table in Appendix B (3), it seems that when we include interactions with at most one other variables we begin to lose significance among some of our variables. In this model we only see significance from variables E_6 and E_8. The environment-environment interaction terms and gene-environment terms observed in the table have t-values lower than two. Therefore, these interaction terms lack significance and should not be included in our model. Next I will investigate interaction with up to two other variables.

Analyzing this table in Appendix B (4), shows how when we consider a model with interactions with up to two other variables we begin to lose significance among all our variables. We see every t-value is lower than two and our p-values are not that significant. Finally, we must investigate the possibility of 4-way interaction terms.

Similar to our situation regarding three-way interaction terms observing table in Appendix B (5), shows us that when we consider four-way interactions between variables we lose significance. The table shows no variables with significant t-values or significant p-values.

Hence, despite our exhaustive search regarding interactions between variables, it seems the best model to fit our data is one which does not contain any interaction terms. That being the model which looks like $Y = \alpha + \beta_1 E_1 + \beta_2 E_6 + \beta_3 E_8 + \beta_4 G_{21}$. Lets denote this as model 2

However, we are also told that it may be possible for our model to contain variables raised to the power of 2 or the square root of variables. Thus, we must investigate if a model with contains one of these situations fits our data better than our current model. I will accomplish this by including all possible variables that are theoretically able to appear in our model into our forward selection method. Results are seen in Appendix C (1).

Analyzing this Bayesian information criterion table, I choose to investigate the variables which appear in the fourth row because this row contains the lowest valued Bayesian information criterion value which indicates that it is the most accurate model to describe our data. Additionally, the adjusted R^2 increase from the model in the fourth row to the model in the fifth row is insignificant. Now similar to before I can determine if there exist any significant interactions between the variables in this chosen model, which are

$E_1, E_6^{.5}, E_8$, and G_{21}. Again similar to the previous strategy we analyze the possibility of significant interactions between these variables.

From this table in Appendix C (2), we see that the four variables we included appear very significant as their t-values are quite high and p-values are approximately zero. Next we examine possibility of interactions with one other variable.

Analyzing the outcome of this table in Appendix C (3), shows that when we consider two-way interactions, the t-value of most of our variables including interaction terms begin to plummet, meaning that we should not use these terms in our model. Now we must check possibility of three-way interaction terms.

Observing this table found in Appendix C (4), shows that once we consider the possibility of three-way interaction terms we not only see low t-values, but we also notice that our p-values are beginning to show a lack of significance. Hence, three-way interactions are not suitable for this model. Finally, we investigate the possibility of up to four way interactions in our model.

Similar to previous result, outcome from Appendix C (5), tells us we do not notice significance in our model when we consider the possibility of up to four-way interactions among our variables. Hence, we conclude using this method the best model to estimate our response variable would be $Y = \alpha + \beta_1 E_1 + \beta_2 E_6^{.5} + \beta_3 E_8 + \beta_4 G_{21}$. Lets denote this as model 3 and acknowledge this is the same model we chose from the Bayesian information criterion table before we begun analyzing further.

Conclusion

Throughout my investigations we are left with three viable models to represent the expected value of our response variable Y. Model 1 would be the results from our first table using the Bayesian information criterion table. This table gave us a model using the terms, $E_6, E_8, E_1E_8, E_1G_{21}$, and E_6E_8. Model 2 would be the results from our first investigation of the variables found in our model 1, which are E_1, E_6, E_8, and G_{21}. Our third viable model would be the results from investigating the variables found in our second Bayesian information criterion table, which considered the possibility of all theoretical variables which might be present in our model. This would be a model which includes variables $E_1, E_6^{.5}, E_8$, and G_{21}. Now we must compare the viability between these three models and conclude which one best fits our data. In order to accomplish this I will compare adjusted R^2 values, residual standard error, check for multicollinearity issues using the variance inflation factor, and checking the PRESS statistic among these three models.

Results from Appendix D (1), show that model 1 gives us Residual standard error of 13.22 and Adjusted R^2 of 0.4734339. Although this model has issues with multicollinearity see by examining the values we get from the vif function. The Press statistic for this model was 468223.

Results from Appendix D (2), show that model 2 gives us Residual standard ettor of 13.21 and Adjusted R^2 of 0.4741244. Notice, this model does not have issues regarding

multicollinearity seen by examining the low values we get from the vif function. The Press statistic for this model was 467425.1.

Results from Appendix D (3), show that model 3 gives us Residual standard error of 13.2 and Adjusted R^2 of 0.4743192. Notice, this model similar to the previous model does not have issues regarding multicollinearity seen by examining the low values we get from the vif function. The Press statistic for this model was 467249.9.

After comparing these results I determine that since model 3 displays the lowest Residual standard error, highest adjusted R^2, and lowest PRESS statistic out of all the models, I deem this model the best in terms of determining the value of our response variable Y. Lastly, I will check the necessary assumption regarding this model and formally write the model including all the coefficients. Various tests and plots in order to verify assumptions can be seen in Appendix D (4).

The first model assumption to check is the residuals having a normal distribution. As seen from the plot, the residual plot of the full model shows no significant pattern, which leads to the confirmation of the normal distribution assumption for the residuals. There is no significant violation according to the residual plot. The next model assumption to check is that the residuals are uncorrelated. Again, from our residual plot, we do not see any sort of pattern in the plot as it seems that the residuals are distributed completely randomly in these plots, proving that there is no correlation between the residuals. The next, and final, model assumptions to check is that the residuals have constant variance. By looking at the residual vs. fitted values plot above, there does not seem to be any noticeable pattern in the plot suggesting that this assumption is fulfilled. Also, observe the QQ plot in Appendix D (4) to see that the residuals fit the QQ line very well meaning that our model seems to fit a normal distribution well. Next assumption to check is the error term having a zero mean. As you can see above, the mean of the residuals is an extremely low number,1.665335e-16, which can be approximated to zero, confirming this assumption. Finally, all we must do now is include the coefficients for model 3, in order to formally write the model along with the parameter estimates. We will find these coefficients in Appendix D (5).

Therefore, my final model utilizing only Environmental variables is:

$$Y = -21.2038 + 3.6773E_1 + 32.6865E_6^{.5} + 9.2053E_8$$

Although, my final model period utilizing estimated parameter would be:

$$Y = -23.2720 + 3.6740E_1 + 32.7423E_6^{.5} + 9.1242E_8 + 6.2752G_{21}$$

Coefficients for both these models can be seen in Appendix D (5). Thus, we see Environmental variables E_1, E_6 and E_8 have an affect on our response variable variable. Additionally, genetic variable G_{21} also had an affect on our response variable. We also acknowledge that we did not discover any statistically significant interactions between any variables in our final model.

Biggest takeaways from our report is that our dependent variable Y, which represents some type of mental illness, is affected by Environmental variables 1,6, and 8 and is affected by the genetic variable 21. Additionally, our final model fulfills a common rule

when model building which is $n \geq 20p$. Meaning that since we have 2675 observations in our data and only four variable in our final model, our model satisfies $2675 \geq 20(5) = 100$. The limitations of my report would be the case where there exist interactions more than four-way interactions or possibly our variables needed to be raised to a power other than .5 or 2.

References

Caspi, A., et al. (2003, July 18). Influence of Life Stress on Depression: Moderation by a Polymorphism in the 5-HTT Gene. Sciencemag.Org. https://blackboard.stonybrook.edu/bbcswebdav/pid-1695608-dt-content-rid-12558879_1/courses/1224-AMS-578-SEC01-48530/Caspi_et_al._2003_Science.pdf

McLafferty, M. (2017, December 13). Mental health, behavioural problems and treatment seeking among students commencing university in Northern Ireland. Journals.Plos.Org. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0188785#pone.0188785.ref001

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). Introduction to Linear Regression Analysis. Wiley.

What is Bayesian Information Criterion (BIC)? - Analyttica Datalab. (2019, January 16). Medium. https://medium.com/@analyttica/what-is-bayesian-information-criterion-bic-b3396a894be6

Supplementary Material

Appendix A

A (1)

```
#Loading in necessary libraries/packages
library('readxl')
library('MASS')
library('leaps')
library('knitr')
library('car')
library('DAAG')

#Loading data set, setting variable names
projectdata <- read.csv("C:/Users/Micha/Desktop/data.csv")
names(projectdata)<-
c('Y','E1','E2','E3','E4','E5','E6','E7','E8','G1','G2','G3','G4','G5','G6','G7','G8','G9','G10','G11','G12','G13','G14','G15','G16','G17','G18','G19','G20','G21','G22','G23','G24','G25','G26','G27','G28','G29','G30')
attach(projectdata)
```

A (2)

```
summary(projectdata)

##        Y                E1              E2              E3
##  Min.   : 57.05   Min.   :1.809   Min.   :1.099   Min.   :1.622
##  1st Qu.:103.36   1st Qu.:4.296   1st Qu.:4.316   1st Qu.:4.320
##  Median :115.94   Median :4.944   Median :5.019   Median :5.042
##  Mean   :115.77   Mean   :4.980   Mean   :5.004   Mean   :5.012
##  3rd Qu.:128.18   3rd Qu.:5.648   3rd Qu.:5.673   3rd Qu.:5.715
```

```
##  Max.   :170.84   Max.   :8.244   Max.   :8.428   Max.   :8.371
##        E4              E5              E6              E7
##  Min.   :1.659   Min.   :1.005   Min.   :1.183   Min.   :1.189
##  1st Qu.:4.351   1st Qu.:4.337   1st Qu.:4.313   1st Qu.:4.292
##  Median :5.023   Median :5.025   Median :4.987   Median :4.988
##  Mean   :5.018   Mean   :5.013   Mean   :4.984   Mean   :4.982
##  3rd Qu.:5.669   3rd Qu.:5.686   3rd Qu.:5.673   3rd Qu.:5.660
##  Max.   :8.714   Max.   :8.512   Max.   :8.874   Max.   :9.033
##        E8              G1              G2              G3
##  Min.   :1.867   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:4.348   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :4.978   Median :0.0000   Median :0.0000   Median :0.0000
##  Mean   :5.004   Mean   :0.3873   Mean   :0.2531   Mean   :0.2994
##  3rd Qu.:5.705   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :8.277   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##        G4              G5              G6              G7
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000   Median :0.000   Median :0.0000
##  Mean   :0.2927   Mean   :0.3839   Mean   :0.268   Mean   :0.3312
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.000   Max.   :1.0000
##        G8              G9              G10             G11
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000   Median :0.0000   Median :0.0000
##  Mean   :0.3234   Mean   :0.2464   Mean   :0.2583   Mean   :0.3637
##  3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##       G12             G13             G14             G15
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000   Median :0.0000   Median :0.0000
##  Mean   :0.2841   Mean   :0.2901   Mean   :0.3148   Mean   :0.3503
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##       G16             G17             G18             G19
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000   Median :0.0000   Median :0.0000
##  Mean   :0.3006   Mean   :0.2583   Mean   :0.2594   Mean   :0.3993
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##       G20             G21             G22             G23
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000   Median :0.0000   Median :0.0000
##  Mean   :0.3944   Mean   :0.3772   Mean   :0.3581   Mean   :0.3103
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
```

```
##       G24               G25               G26               G27         
##  Min.   :0.0000   Min.   :0.000   Min.   :0.0000   Min.   :0.0000  
##  1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:0.0000  
##  Median :0.0000   Median :0.000   Median :0.0000   Median :0.0000  
##  Mean   :0.3241   Mean   :0.385   Mean   :0.3821   Mean   :0.2841  
##  3rd Qu.:1.0000   3rd Qu.:1.000   3rd Qu.:1.0000   3rd Qu.:1.0000  
##  Max.   :1.0000   Max.   :1.000   Max.   :1.0000   Max.   :1.0000  
##       G28               G29               G30         
##  Min.   :0.0000   Min.   :0.000   Min.   :0.0000  
##  1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:0.0000  
##  Median :0.0000   Median :0.000   Median :0.0000  
##  Mean   :0.3159   Mean   :0.326   Mean   :0.3787  
##  3rd Qu.:1.0000   3rd Qu.:1.000   3rd Qu.:1.0000  
##  Max.   :1.0000   Max.   :1.000   Max.   :1.0000
```

A (3)

```
#Regression model Y vs E1
mod.E1 <- lm(Y ~ E1, data = projectdata)
summary(mod.E1)

##
## Call:
## lm(formula = Y ~ E1, data = projectdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -57.655 -12.488   0.356  12.293  51.660
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   97.424      1.757   55.44   <2e-16 ***
## E1             3.684      0.346   10.65   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.84 on 2673 degrees of freedom
## Multiple R-squared:  0.04069,    Adjusted R-squared:  0.04033
## F-statistic: 113.4 on 1 and 2673 DF,  p-value: < 2.2e-16

#Regression model Y vs E2
mod.E2 <- lm(Y ~ E2, data = projectdata)
summary(mod.E2)

##
## Call:
## lm(formula = Y ~ E2, data = projectdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -58.389 -12.440   0.168  12.615  55.234
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 119.2976     1.7683  67.464   <2e-16 ***
## E2           -0.7048     0.3463  -2.035   0.0419 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.2 on 2673 degrees of freedom
## Multiple R-squared:  0.001547,   Adjusted R-squared:  0.001174
## F-statistic: 4.142 on 1 and 2673 DF,  p-value: 0.04193
```

```
#Regression model Y vs E3
mod.E3 <- lm(Y ~ E3, data = projectdata)
summary(mod.E3)
```

```
##
## Call:
## lm(formula = Y ~ E3, data = projectdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -58.456 -12.284   0.132  12.503  55.559
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 118.6613     1.7673  67.142   <2e-16 ***
## E3           -0.5768     0.3456  -1.669   0.0952 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.21 on 2673 degrees of freedom
## Multiple R-squared:  0.001041,   Adjusted R-squared:  0.0006673
## F-statistic: 2.785 on 1 and 2673 DF,  p-value: 0.09524
```

```
#Regression model Y vs E4
mod.E4 <- lm(Y ~ E4, data = projectdata)
summary(mod.E4)
```

```
##
## Call:
## lm(formula = Y ~ E4, data = projectdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -58.674 -12.395   0.139  12.420  55.089
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 115.55379    1.79326  64.438   <2e-16 ***
## E4            0.04324    0.35040   0.123    0.902
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.22 on 2673 degrees of freedom
## Multiple R-squared:  5.698e-06,  Adjusted R-squared:  -0.0003684
## F-statistic: 0.01523 on 1 and 2673 DF,  p-value: 0.9018
```

```
#Regression model Y vs E5
mod.E5 <- lm(Y ~ E5, data = projectdata)
summary(mod.E5)
```

```
##
## Call:
## lm(formula = Y ~ E5, data = projectdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -58.756 -12.356   0.137  12.396  54.691
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 116.9512     1.7881  65.405   <2e-16 ***
## E5           -0.2355     0.3497  -0.673    0.501
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.21 on 2673 degrees of freedom
## Multiple R-squared:  0.0001696,  Adjusted R-squared:  -0.0002045
## F-statistic: 0.4534 on 1 and 2673 DF,  p-value: 0.5008
```

```
#Regression model Y vs E6
mod.E6 <- lm(Y ~ E6, data = projectdata)
summary(mod.E6)
```

```
##
## Call:
## lm(formula = Y ~ E6, data = projectdata)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -54.95 -11.09   0.00  11.12  54.04
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  79.6414     1.6652   47.83   <2e-16 ***
## E6            7.2487     0.3277   22.12   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.75 on 2673 degrees of freedom
## Multiple R-squared:  0.1547, Adjusted R-squared:  0.1544
## F-statistic: 489.2 on 1 and 2673 DF,  p-value: < 2.2e-16
```

```r
#Regression model Y vs E7
mod.E7 <- lm(Y ~ E7, data = projectdata)
summary(mod.E7)

##
## Call:
## lm(formula = Y ~ E7, data = projectdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -58.587 -12.377   0.156  12.459  54.616
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 113.7872     1.7345  65.604   <2e-16 ***
## E7            0.3981     0.3409   1.168    0.243
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.21 on 2673 degrees of freedom
## Multiple R-squared:  0.0005101,  Adjusted R-squared:  0.0001361
## F-statistic: 1.364 on 1 and 2673 DF,  p-value: 0.2429

#Regression model Y vs E8
mod.E8 <- lm(Y ~ E8, data = projectdata)
summary(mod.E8)

##
## Call:
## lm(formula = Y ~ E8, data = projectdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -53.677 -10.682  -0.373  10.737  51.031
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  70.7901     1.5718   45.04   <2e-16 ***
## E8            8.9898     0.3081   29.18   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.86 on 2673 degrees of freedom
## Multiple R-squared:  0.2416, Adjusted R-squared:  0.2413
## F-statistic: 851.4 on 1 and 2673 DF,  p-value: < 2.2e-16
```

A (4)

```r
for (i in 1:30) {
  print("T-test Results for indicator:")
  print(i)
```

```
  print(t.test(formula = Y ~ get(paste0("G",i))))
}
```

```
## [1] "T-test Results for indicator:"
## [1] 1
##
##  Welch Two Sample t-test
##
## data:  Y by get(paste0("G", i))
## t = -1.2292, df = 2256.1, p-value = 0.2191
## alternative hypothesis: true difference in means between group 0 and group
1 is not equal to 0
## 95 percent confidence interval:
##  -2.2883542  0.5249621
## sample estimates:
## mean in group 0 mean in group 1
##        115.4293        116.3110
##
## [1] "T-test Results for indicator:"
## [1] 2
##
##  Welch Two Sample t-test
##
## data:  Y by get(paste0("G", i))
## t = -0.60715, df = 1190.7, p-value = 0.5439
## alternative hypothesis: true difference in means between group 0 and group
1 is not equal to 0
## 95 percent confidence interval:
##  -2.056791  1.084637
## sample estimates:
## mean in group 0 mean in group 1
##        115.6478        116.1338
##
## [1] "T-test Results for indicator:"
## [1] 3
##
##  Welch Two Sample t-test
##
## data:  Y by get(paste0("G", i))
## t = 0.20904, df = 1505.2, p-value = 0.8344
## alternative hypothesis: true difference in means between group 0 and group
1 is not equal to 0
## 95 percent confidence interval:
##  -1.350279  1.672408
## sample estimates:
## mean in group 0 mean in group 1
##         115.819         115.658
##
## [1] "T-test Results for indicator:"
## [1] 4
```

```
##
##  Welch Two Sample t-test
##
## data:  Y by get(paste0("G", i))
## t = 1.0701, df = 1458.2, p-value = 0.2848
## alternative hypothesis: true difference in means between group 0 and group
1 is not equal to 0
## 95 percent confidence interval:
##  -0.6902616  2.3472640
## sample estimates:
## mean in group 0 mean in group 1
##        116.0133        115.1848
##
## [1] "T-test Results for indicator:"
## [1] 5
##
##  Welch Two Sample t-test
##
## data:  Y by get(paste0("G", i))
## t = -0.63581, df = 2213.2, p-value = 0.525
## alternative hypothesis: true difference in means between group 0 and group
1 is not equal to 0
## 95 percent confidence interval:
##  -1.8706998  0.9546575
## sample estimates:
## mean in group 0 mean in group 1
##        115.5949        116.0530
##
## [1] "T-test Results for indicator:"
## [1] 6
##
##  Welch Two Sample t-test
##
## data:  Y by get(paste0("G", i))
## t = -0.10192, df = 1263.6, p-value = 0.9188
## alternative hypothesis: true difference in means between group 0 and group
1 is not equal to 0
## 95 percent confidence interval:
##  -1.647966  1.485200
## sample estimates:
## mean in group 0 mean in group 1
##        115.7490        115.8304
##
## [1] "T-test Results for indicator:"
## [1] 7
##
##  Welch Two Sample t-test
##
## data:  Y by get(paste0("G", i))
## t = 1.1722, df = 1764.8, p-value = 0.2413
```

```
## alternative hypothesis: true difference in means between group 0 and group
1 is not equal to 0
## 95 percent confidence interval:
##  -0.5903121  2.3440043
## sample estimates:
## mean in group 0 mean in group 1
##        116.0612        115.1844
##
## [1] "T-test Results for indicator:"
## [1] 8
##
##  Welch Two Sample t-test
##
## data:  Y by get(paste0("G", i))
## t = 1.9958, df = 1650.4, p-value = 0.04612
## alternative hypothesis: true difference in means between group 0 and group
1 is not equal to 0
## 95 percent confidence interval:
##  0.02621244 3.01333306
## sample estimates:
## mean in group 0 mean in group 1
##        116.2622        114.7425
##
## [1] "T-test Results for indicator:"
## [1] 9
##
##  Welch Two Sample t-test
##
## data:  Y by get(paste0("G", i))
## t = -1.703, df = 1118.1, p-value = 0.08885
## alternative hypothesis: true difference in means between group 0 and group
1 is not equal to 0
## 95 percent confidence interval:
##  -2.9961840  0.2118415
## sample estimates:
## mean in group 0 mean in group 1
##        115.4278        116.8200
##
## [1] "T-test Results for indicator:"
## [1] 10
##
##  Welch Two Sample t-test
##
## data:  Y by get(paste0("G", i))
## t = -0.6563, df = 1145, p-value = 0.5118
## alternative hypothesis: true difference in means between group 0 and group
1 is not equal to 0
## 95 percent confidence interval:
##  -2.167531  1.080929
## sample estimates:
```

```
## mean in group 0 mean in group 1
##        115.6304        116.1737
##
## [1] "T-test Results for indicator:"
## [1] 11
##
##  Welch Two Sample t-test
##
## data:  Y by get(paste0("G", i))
## t = -1.6163, df = 1990.3, p-value = 0.1062
## alternative hypothesis: true difference in means between group 0 and group
1 is not equal to 0
## 95 percent confidence interval:
##  -2.6319157  0.2536847
## sample estimates:
## mean in group 0 mean in group 1
##        115.3383        116.5274
##
## [1] "T-test Results for indicator:"
## [1] 12
##
##  Welch Two Sample t-test
##
## data:  Y by get(paste0("G", i))
## t = 1.5671, df = 1442.8, p-value = 0.1173
## alternative hypothesis: true difference in means between group 0 and group
1 is not equal to 0
## 95 percent confidence interval:
##  -0.3029544  2.7100476
## sample estimates:
## mean in group 0 mean in group 1
##        116.1127        114.9092
##
## [1] "T-test Results for indicator:"
## [1] 13
##
##  Welch Two Sample t-test
##
## data:  Y by get(paste0("G", i))
## t = 0.24594, df = 1488.3, p-value = 0.8058
## alternative hypothesis: true difference in means between group 0 and group
1 is not equal to 0
## 95 percent confidence interval:
##  -1.311317  1.687287
## sample estimates:
## mean in group 0 mean in group 1
##        115.8253        115.6373
##
## [1] "T-test Results for indicator:"
## [1] 14
```

```
##
##  Welch Two Sample t-test
##
## data:  Y by get(paste0("G", i))
## t = -1.5145, df = 1719.4, p-value = 0.1301
## alternative hypothesis: true difference in means between group 0 and group
1 is not equal to 0
## 95 percent confidence interval:
##  -2.580418  0.331737
## sample estimates:
## mean in group 0 mean in group 1
##        115.4169        116.5412
##
## [1] "T-test Results for indicator:"
## [1] 15
##
##  Welch Two Sample t-test
##
## data:  Y by get(paste0("G", i))
## t = 0.319, df = 1964, p-value = 0.7498
## alternative hypothesis: true difference in means between group 0 and group
1 is not equal to 0
## 95 percent confidence interval:
##  -1.202121  1.669150
## sample estimates:
## mean in group 0 mean in group 1
##        115.8526        115.6191
##
## [1] "T-test Results for indicator:"
## [1] 16
##
##  Welch Two Sample t-test
##
## data:  Y by get(paste0("G", i))
## t = -0.49298, df = 1463, p-value = 0.6221
## alternative hypothesis: true difference in means between group 0 and group
1 is not equal to 0
## 95 percent confidence interval:
##  -1.918184  1.147675
## sample estimates:
## mean in group 0 mean in group 1
##        115.6550        116.0402
##
## [1] "T-test Results for indicator:"
## [1] 17
##
##  Welch Two Sample t-test
##
## data:  Y by get(paste0("G", i))
## t = 1.4334, df = 1239.6, p-value = 0.152
```

```
## alternative hypothesis: true difference in means between group 0 and group
1 is not equal to 0
## 95 percent confidence interval:
##  -0.418275  2.687291
## sample estimates:
## mean in group 0 mean in group 1
##        116.0639        114.9293
##
## [1] "T-test Results for indicator:"
## [1] 18
##
##  Welch Two Sample t-test
##
## data:  Y by get(paste0("G", i))
## t = 0.55268, df = 1244, p-value = 0.5806
## alternative hypothesis: true difference in means between group 0 and group
1 is not equal to 0
## 95 percent confidence interval:
##  -1.116103  1.991554
## sample estimates:
## mean in group 0 mean in group 1
##        115.8844        115.4466
##
## [1] "T-test Results for indicator:"
## [1] 19
##
##  Welch Two Sample t-test
##
## data:  Y by get(paste0("G", i))
## t = -0.73789, df = 2244.6, p-value = 0.4607
## alternative hypothesis: true difference in means between group 0 and group
1 is not equal to 0
## 95 percent confidence interval:
##  -1.9508301  0.8841051
## sample estimates:
## mean in group 0 mean in group 1
##        115.5578        116.0912
##
## [1] "T-test Results for indicator:"
## [1] 20
##
##  Welch Two Sample t-test
##
## data:  Y by get(paste0("G", i))
## t = 1.6195, df = 2199.7, p-value = 0.1055
## alternative hypothesis: true difference in means between group 0 and group
1 is not equal to 0
## 95 percent confidence interval:
##  -0.2476975  2.5967350
## sample estimates:
```

```
## mean in group 0 mean in group 1
##        116.2340        115.0595
##
## [1] "T-test Results for indicator:"
## [1] 21
##
##  Welch Two Sample t-test
##
## data:  Y by get(paste0("G", i))
## t = -9.3997, df = 2138.9, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group 0 and group
1 is not equal to 0
## 95 percent confidence interval:
##  -8.107571 -5.308535
## sample estimates:
## mean in group 0 mean in group 1
##        113.2405        119.9486
##
## [1] "T-test Results for indicator:"
## [1] 22
##
##  Welch Two Sample t-test
##
## data:  Y by get(paste0("G", i))
## t = 1.3486, df = 1979, p-value = 0.1776
## alternative hypothesis: true difference in means between group 0 and group
1 is not equal to 0
## 95 percent confidence interval:
##  -0.449843  2.430486
## sample estimates:
## mean in group 0 mean in group 1
##        116.1255        115.1351
##
## [1] "T-test Results for indicator:"
## [1] 23
##
##  Welch Two Sample t-test
##
## data:  Y by get(paste0("G", i))
## t = 0.88159, df = 1636, p-value = 0.3781
## alternative hypothesis: true difference in means between group 0 and group
1 is not equal to 0
## 95 percent confidence interval:
##  -0.8138266  2.1426751
## sample estimates:
## mean in group 0 mean in group 1
##        115.9769        115.3125
##
## [1] "T-test Results for indicator:"
## [1] 24
```

```
##
##  Welch Two Sample t-test
##
## data:  Y by get(paste0("G", i))
## t = -1.1803, df = 1695.8, p-value = 0.238
## alternative hypothesis: true difference in means between group 0 and group
1 is not equal to 0
## 95 percent confidence interval:
##  -2.3697957  0.5891631
## sample estimates:
## mean in group 0 mean in group 1
##        115.4822        116.3725
##
## [1] "T-test Results for indicator:"
## [1] 25
##
##  Welch Two Sample t-test
##
## data:  Y by get(paste0("G", i))
## t = -0.67413, df = 2275.5, p-value = 0.5003
## alternative hypothesis: true difference in means between group 0 and group
1 is not equal to 0
## 95 percent confidence interval:
##  -1.8828724  0.9195104
## sample estimates:
## mean in group 0 mean in group 1
##        115.5853        116.0670
##
## [1] "T-test Results for indicator:"
## [1] 26
##
##  Welch Two Sample t-test
##
## data:  Y by get(paste0("G", i))
## t = -0.38489, df = 2120, p-value = 0.7004
## alternative hypothesis: true difference in means between group 0 and group
1 is not equal to 0
## 95 percent confidence interval:
##  -1.710830  1.149463
## sample estimates:
## mean in group 0 mean in group 1
##        115.6636        115.9442
##
## [1] "T-test Results for indicator:"
## [1] 27
##
##  Welch Two Sample t-test
##
## data:  Y by get(paste0("G", i))
## t = -1.0278, df = 1407.3, p-value = 0.3042
```

```
## alternative hypothesis: true difference in means between group 0 and group
1 is not equal to 0
## 95 percent confidence interval:
##  -2.3234144  0.7258293
## sample estimates:
## mean in group 0 mean in group 1
##        115.5438        116.3426
##
## [1] "T-test Results for indicator:"
## [1] 28
##
##  Welch Two Sample t-test
##
## data:  Y by get(paste0("G", i))
## t = -1.5495, df = 1647, p-value = 0.1214
## alternative hypothesis: true difference in means between group 0 and group
1 is not equal to 0
## 95 percent confidence interval:
##  -2.6553250  0.3115127
## sample estimates:
## mean in group 0 mean in group 1
##        115.4006        116.5725
##
## [1] "T-test Results for indicator:"
## [1] 29
##
##  Welch Two Sample t-test
##
## data:  Y by get(paste0("G", i))
## t = -0.37882, df = 1718.3, p-value = 0.7049
## alternative hypothesis: true difference in means between group 0 and group
1 is not equal to 0
## 95 percent confidence interval:
##  -1.759890  1.190118
## sample estimates:
## mean in group 0 mean in group 1
##        115.6779        115.9628
##
## [1] "T-test Results for indicator:"
## [1] 30
##
##  Welch Two Sample t-test
##
## data:  Y by get(paste0("G", i))
## t = 0.74741, df = 2168.5, p-value = 0.4549
## alternative hypothesis: true difference in means between group 0 and group
1 is not equal to 0
## 95 percent confidence interval:
##  -0.8771532  1.9575187
## sample estimates:
```

```
## mean in group 0 mean in group 1
##        115.9753        115.4352
```

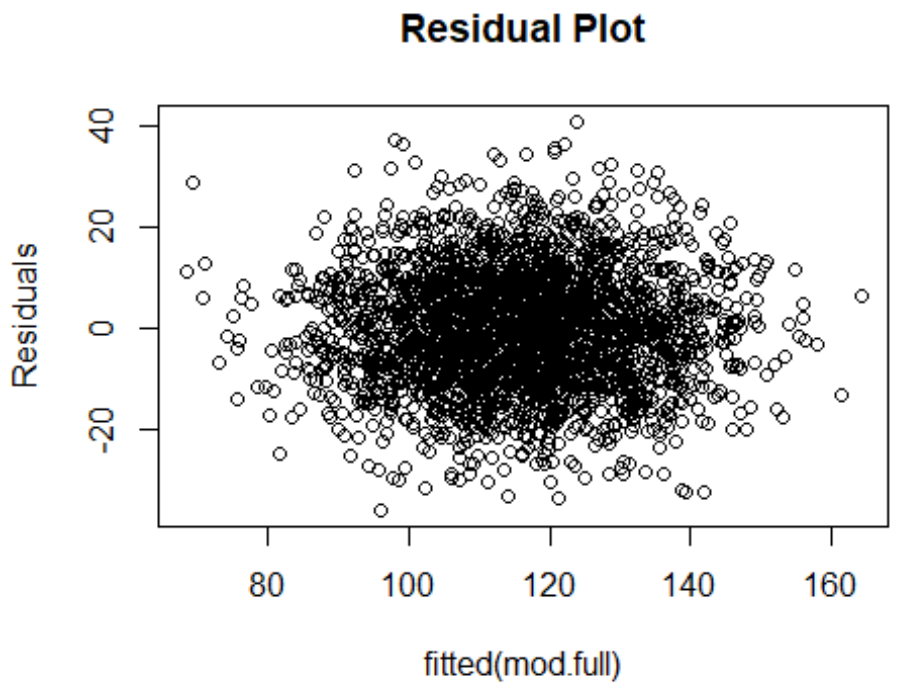
A (5)

```
#For length reasons I have chosen to hide these results.
E1.squared <- (projectdata$E1)^2
E2.squared <- (projectdata$E2)^2
E3.squared <- (projectdata$E3)^2
E4.squared <- (projectdata$E4)^2
E5.squared <- (projectdata$E5)^2
E6.squared <- (projectdata$E6)^2
E7.squared <- (projectdata$E7)^2
E8.squared <- (projectdata$E8)^2

mod.E1.squared <- lm(Y ~ E1.squared, data = projectdata)
summary(mod.E1.squared)
mod.E2.squared <- lm(Y ~ E2.squared, data = projectdata)
summary(mod.E2.squared)
mod.E3.squared <- lm(Y ~ E3.squared, data = projectdata)
summary(mod.E3.squared)
mod.E4.squared <- lm(Y ~ E4.squared, data = projectdata)
summary(mod.E4.squared)
mod.E5.squared <- lm(Y ~ E5.squared, data = projectdata)
summary(mod.E5.squared)
mod.E6.squared <- lm(Y ~ E6.squared, data = projectdata)
summary(mod.E6.squared)
mod.E7.squared <- lm(Y ~ E7.squared, data = projectdata)
summary(mod.E7.squared)
mod.E8.squared <- lm(Y ~ E8.squared, data = projectdata)
summary(mod.E8.squared)

E1.sqrt <- (projectdata$E1)^.5
E2.sqrt <- (projectdata$E2)^.5
E3.sqrt <- (projectdata$E3)^.5
E4.sqrt <- (projectdata$E4)^.5
E5.sqrt <- (projectdata$E5)^.5
E6.sqrt <- (projectdata$E6)^.5
E7.sqrt <- (projectdata$E7)^.5
E8.sqrt <- (projectdata$E8)^.5

mod.E1.sqrt <- lm(Y ~ E1.sqrt, data = projectdata)
summary(mod.E1.sqrt)
mod.E2.sqrt <- lm(Y ~ E2.sqrt, data = projectdata)
summary(mod.E2.sqrt)
mod.E3.sqrt <- lm(Y ~ E3.sqrt, data = projectdata)
summary(mod.E3.sqrt)
mod.E4.sqrt <- lm(Y ~ E4.sqrt, data = projectdata)
summary(mod.E4.sqrt)
mod.E5.sqrt <- lm(Y ~ E5.sqrt, data = projectdata)
summary(mod.E5.sqrt)
```

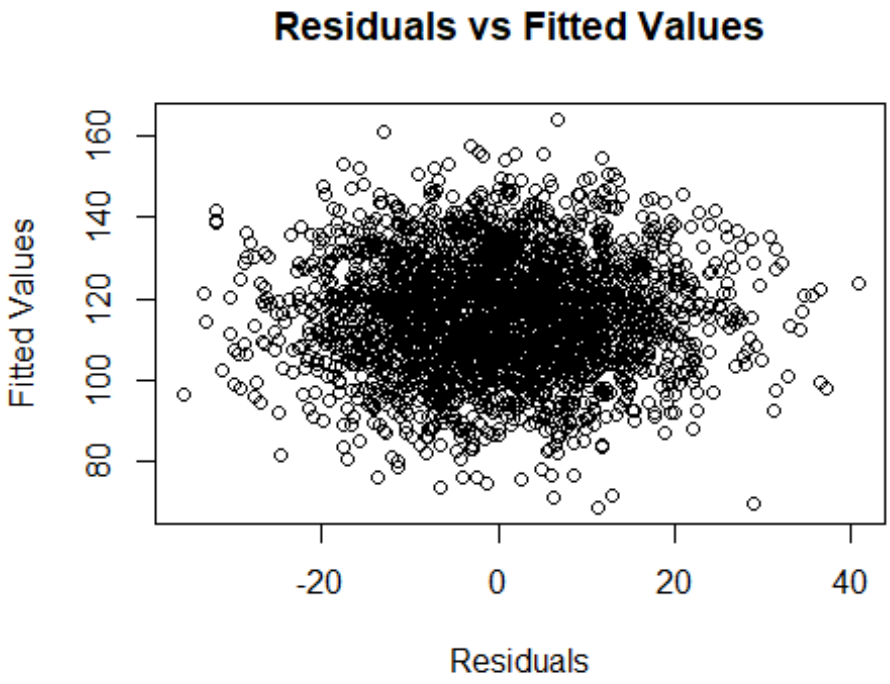
```
mod.E6.sqrt <- lm(Y ~ E6.sqrt, data = projectdata)
summary(mod.E6.sqrt)
mod.E7.sqrt <- lm(Y ~ E7.sqrt, data = projectdata)
summary(mod.E7.sqrt)
mod.E8.sqrt <- lm(Y ~ E8.sqrt, data = projectdata)
summary(mod.E8.sqrt)
```

A (6)

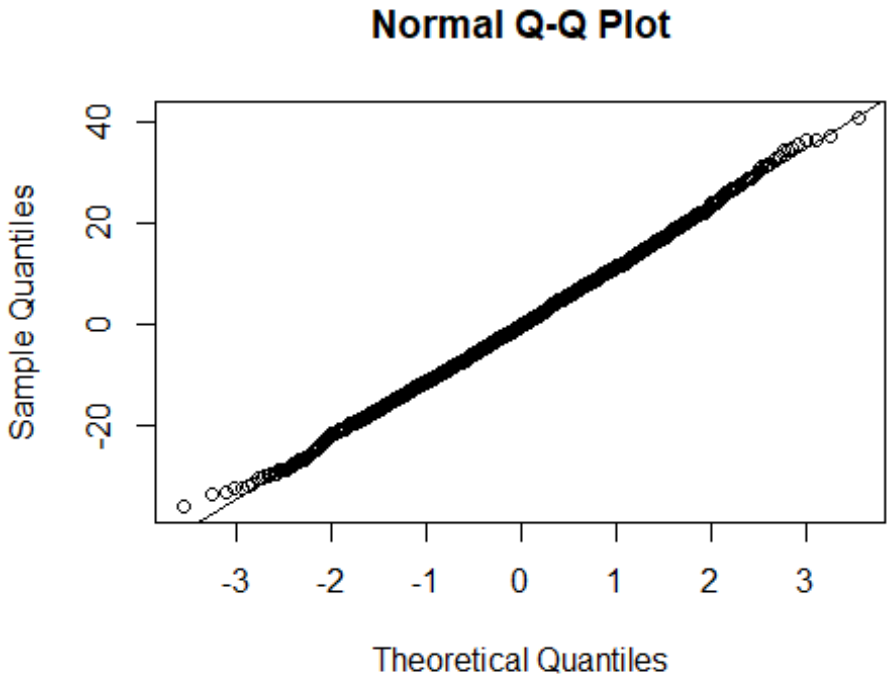
```
mod.full <- lm(Y ~
(E1+E2+E3+E4+E5+E6+E7+E8+G1+G2+G3+G4+G5+G6+G7+G8+G9+G10+G11+G12+G13+G14+G15+G16+G17+G18+G19+G20+G21+G22+G23+G24+G25+G26+G27+G28+G29+G30)^2, data =
projectdata)
plot(resid(mod.full) ~ fitted(mod.full), main='Residual
Plot',ylab='Residuals')
```




```
plot(mod.full$residuals,mod.full$fitted.values,xlab =
'Residuals',ylab='Fitted Values',main = 'Residuals vs Fitted Values')
```

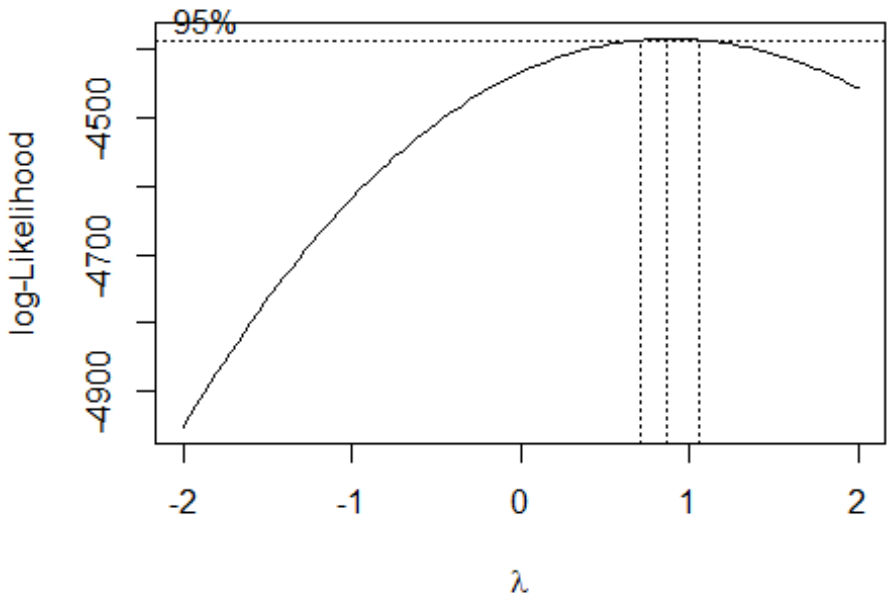



```
qqnorm(residuals(mod.full))
qqline(mod.full$residuals)
```



A (7)

```
bc <- boxcox(mod.full)
```



Appendix B

B (1)

```
M <- regsubsets(model.matrix(mod.full)[,-1], Y,nbest = 1 , nvmax=5, method =
'forward', intercept = TRUE )
temp <- summary(M)
Var <- colnames(model.matrix(mod.full))
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')
```

Model Summary

model	adjR2	BIC
(Intercept)+E6:E8	0.390520768895639	-1309.74448764976
(Intercept)+E1:E8+E6:E8	0.434550704908514	-1503.43657478448
(Intercept)+E1:E8+E1:G21+E6:E8	0.46302075211848	-1634.73983264956
(Intercept)+E6+E1:E8+E1:G21+E6:E8	0.46707538219761	-1648.12488214673
(Intercept)+E6+E8+E1:E8+E1:G21+E6:E8	0.473433902021459	-1673.34357909049

B (2)

```
M_1st <- lm(Y ~
E1+E2+E3+E4+E5+E6+E7+E8+G1+G2+G3+G4+G5+G6+G7+G8+G9+G10+G11+G12+G13+G14+G15+G1
6+G17+G18+G19+G20+G21+G22+G23+G24+G25+G26+G27+G28+G29+G30, data =
projectdata)
temp <- summary(M_1st)
kable(temp$coefficients[abs(temp$coefficients[,4]) <= 0.01,], caption='Sig
Coefficients')
```

Sig Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	11.579037	3.7965584	3.049877	0.0023121
E1	3.745493	0.2576601	14.536564	0.0000000
E6	7.425736	0.2599990	28.560634	0.0000000
E8	9.082536	0.2581060	35.189177	0.0000000
G21	6.257588	0.5299745	11.807337	0.0000000

```
M_1stage <- lm(Y ~
E1+E2+E3+E4+E5+E6+E7+E8+G1+G2+G3+G4+G5+G6+G7+G8+G9+G10+G11+G12+G13+G14+G15+G1
6+G17+G18+G19+G20+G21+G22+G23+G24+G25+G26+G27+G28+G29+G30, data=projectdata)
temp <- summary(M_1stage)
temp$coefficients[abs(temp$coefficients[,3]) >= 3,]

##              Estimate Std. Error   t value      Pr(>|t|)
## (Intercept) 11.579037  3.7965584  3.049877  2.312124e-03
## E1           3.745493  0.2576601 14.536564  4.119570e-46
## E6           7.425736  0.2599990 28.560634 1.517103e-156
## E8           9.082536  0.2581060 35.189177 1.021722e-222
## G21          6.257588  0.5299745 11.807337  2.178123e-31
```

B (3)

```
M_2stage <- lm(Y ~ (E1+E6+E8+G21)^2, data=projectdata)
temp <- summary(M_2stage)
temp$coefficients[abs(temp$coefficients[,3]) >= 1,]

##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 12.3115240 11.8382813  1.039976 2.984456e-01
## E6           9.1802987  1.8451357  4.975406 6.925547e-07
## E8           9.4410644  1.8887080  4.998689 6.146400e-07
## E1:E8        0.2684001  0.2512993  1.068050 2.855949e-01
## E6:E8       -0.3951692  0.2640841 -1.496376 1.346741e-01
## E6:G21      -0.6415722  0.5330916 -1.203493 2.288924e-01
## E8:G21       0.9022165  0.5345240  1.687888 9.154985e-02
```

B (4)

```
M_3stage <- lm(Y ~ (E1+E6+E8+G21)^3, data=projectdata)
temp <- summary(M_3stage)
temp$coefficients[abs(temp$coefficients[,3]) >= 1,]
```

```
##             Estimate Std. Error   t value   Pr(>|t|)
## E6         12.452285  7.0853442  1.757471 0.07895264
## E8         10.715965  7.1596186  1.496723 0.13458403
## G21        29.535227 24.1979401  1.220568 0.22235788
## E1:G21     -4.487552  3.8985747 -1.151075 0.24980481
## E6:G21     -6.086420  3.8362733 -1.586545 0.11273462
## E1:E6:G21   1.003047  0.5335623  1.879906 0.06023003
```

B (5)

```
M_4stage <- lm(Y ~ (E1+E6+E8+G21)^4, data=projectdata)
temp <- summary(M_4stage)
temp$coefficients[abs(temp$coefficients[,3]) >= 1,]
```

```
##                  Estimate Std. Error   t value   Pr(>|t|)
## E1             10.1542847  8.8523427  1.147073 0.25145459
## E6             17.6258857  8.6733059  2.032199 0.04223270
## E8             15.9325850  8.7579765  1.819208 0.06899205
## G21           104.3444592 76.2757476  1.367990 0.17143080
## E1:G21        -19.5805131 15.1056753 -1.296236 0.19500684
## E6:G21        -20.6895278 14.6321176 -1.413980 0.15748472
## E1:E6:G21       3.9533176  2.9021916  1.362184 0.17325536
## E6:E8:G21       2.9632931  2.8318061  1.046432 0.29545666
## E1:E6:E8:G21   -0.5799187  0.5607446 -1.034194 0.30113949
```

Appendix C

C (1)

```
mod.full.all <- lm(Y ~
(E1+E2+E3+E4+E5+E6+E7+E8+E1.sqrt+E2.sqrt+E3.sqrt+E4.sqrt+E5.sqrt+E6.sqrt+E7.s
qrt+E8.sqrt+E1.squared+E2.squared+E3.squared+E4.squared+E5.squared+E6.squared
+E7.squared+E8.squared+G1+G2+G3+G4+G5+G6+G7+G8+G9+G10+G11+G12+G13+G14+G15+G16
+G17+G18+G19+G20+G21+G22+G23+G24+G25+G26+G27+G28+G29+G30), data =
projectdata)
M <- regsubsets(model.matrix(mod.full.all)[,-1], Y,nbest = 1 , nvmax=5,
method = 'forward', intercept = TRUE )
temp <- summary(M)
Var <- colnames(model.matrix(mod.full.all))
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')
```

Model Summary

model	adjR2	BIC
(Intercept)+E8	0.241282750345908	-723.854472179308
(Intercept)+E8+E6.sqrt	0.406235775034723	-1372.73156683606
(Intercept)+E1+E8+E6.sqrt	0.446603613563784	-1554.18197128642
(Intercept)+E1+E8+E6.sqrt+G21	0.474319204220323	-1684.7344178803
(Intercept)+E1+E8+E6.sqrt+G19+G21	0.475077748289892	-1681.70751836151

C (2)

```
M_mainstage <- lm(Y ~ E1+E6.sqrt+E8+G21, data=projectdata)
temp <- summary(M_mainstage)
temp$coefficients[abs(temp$coefficients[,3]) >= 3,]
```

```
##               Estimate Std. Error   t value      Pr(>|t|)
## (Intercept) -23.271956  3.1355392 -7.421995  1.542452e-13
## E1            3.673954  0.2560639 14.347801  4.976049e-45
## E6.sqrt      32.742288  1.1314770 28.937652 2.179461e-160
## E8            9.124207  0.2566393 35.552653 5.306626e-227
## G21           6.275220  0.5269318 11.908978  6.724842e-32
```

C (3)

```
M_2stage <- lm(Y ~ (E1+E6.sqrt+E8+G21)^2, data=projectdata)
temp <- summary(M_2stage)
temp$coefficients[abs(temp$coefficients[,3]) >= 1,]
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -30.2569218 19.7372735 -1.532984 1.253986e-01
## E6.sqrt      39.9790299  8.0838437  4.945547 8.064950e-07
## E8           11.6331447  2.9279802  3.973095 7.282681e-05
## G21           7.0785979  6.4738916  1.093407 2.743140e-01
## E1:E8         0.2869065  0.2512628  1.141858 2.536156e-01
## E6.sqrt:E8   -1.9086821  1.1564586 -1.650454 9.896793e-02
## E6.sqrt:G21  -2.9975370  2.3310857 -1.285897 1.985906e-01
## E8:G21        0.9006185  0.5343390  1.685482 9.201280e-02
```

C (4)

```
M_3stage <- lm(Y ~ (E1+E6.sqrt+E8+G21)^3, data=projectdata)
temp <- summary(M_3stage)
temp$coefficients[abs(temp$coefficients[,3]) >= 1,]
```

```
##                  Estimate Std. Error   t value   Pr(>|t|)
## E6.sqrt         52.937322  31.221659  1.695532 0.09009158
## G21             54.176397  40.506961  1.337459 0.18118722
## E1:G21          -8.852034   6.035964 -1.466549 0.14261711
## E6.sqrt:G21    -24.833111  16.729274 -1.484411 0.13781852
## E1:E6.sqrt:G21   4.231989   2.344597  1.804996 0.07118832
```

C (5)

```
M_4stage <- lm(Y ~ (E1+E6.sqrt+E8+G21)^4, data=projectdata)
temp <- summary(M_4stage)
temp$coefficients[abs(temp$coefficients[,3]) >= 1,]
```

```
##                      Estimate Std. Error   t value   Pr(>|t|)
## (Intercept)       -118.087285  86.981980 -1.357606 0.17470384
## E1                  18.223399  17.187181  1.060290 0.28910890
## E6.sqrt             79.382322  38.560825  2.058626 0.03962728
## E8                  25.422962  16.974060  1.497754 0.13431593
## G21                217.044380 145.159896  1.495209 0.13497858
## E1:G21             -41.780367  28.821535 -1.449623 0.14728154
```

```
## E6.sqrt:E8          -8.111176   7.548550 -1.074534 0.28268077
## E6.sqrt:G21        -97.097608  64.071556 -1.515456 0.12977589
## E8:G21             -30.940779  27.888635 -1.109440 0.26734056
## E1:E6.sqrt:G21      18.852569  12.731097  1.480828 0.13877080
## E1:E8:G21            6.359301   5.523745  1.151266 0.24972618
## E6.sqrt:E8:G21      14.297458  12.333654  1.159223 0.24646930
## E1:E6.sqrt:E8:G21   -2.856856   2.445108 -1.168397 0.24275155
```

Appendix D

D (1)

```
mod.1 <- lm(Y ~ E6+E8+E1:E8+E1:G21+E6:E8,data = projectdata)
summary(mod.1)

##
## Call:
## lm(formula = Y ~ E6 + E8 + E1:E8 + E1:G21 + E6:E8, data = projectdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -43.931  -8.871  -0.219   8.509  49.251
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 20.88535    6.92019   3.018  0.00257 **
## E6           9.38028    1.35152   6.941 4.88e-12 ***
## E8           7.90876    1.37173   5.766 9.07e-09 ***
## E8:E1        0.63024    0.05097  12.364  < 2e-16 ***
## E1:G21       1.23033    0.10372  11.862  < 2e-16 ***
## E6:E8       -0.37951    0.26381  -1.439  0.15039
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.22 on 2669 degrees of freedom
## Multiple R-squared:  0.4744, Adjusted R-squared:  0.4734
## F-statistic: 481.8 on 5 and 2669 DF,  p-value: < 2.2e-16

summary(mod.1)$adj.r.squared

## [1] 0.4734339

vif(mod.1)

##      E6      E8   E8:E1  E1:G21   E6:E8
## 27.3120 28.5620  1.9827  1.0222 52.5720

press(mod.1)

## [1] 468223
```

D (2)

```
mod.2 <- lm(Y ~ E1+E6+E8+G21,data = projectdata)
summary(mod.2)
```

```
##
## Call:
## lm(formula = Y ~ E1 + E6 + E8 + G21, data = projectdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -43.694  -8.944  -0.304   8.612  49.164
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  12.2087     2.2605   5.401 7.22e-08 ***
## E1            3.6822     0.2561  14.378  < 2e-16 ***
## E6            7.4754     0.2585  28.915  < 2e-16 ***
## E8            9.1130     0.2567  35.504  < 2e-16 ***
## G21           6.2720     0.5270  11.901  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.21 on 2670 degrees of freedom
## Multiple R-squared:  0.4749, Adjusted R-squared:  0.4741
## F-statistic: 603.7 on 4 and 2670 DF,  p-value: < 2.2e-16
```

```
summary(mod.2)$adj.r.squared
```

```
## [1] 0.4741244
```

```
vif(mod.2)
```

```
##     E1     E6     E8    G21
## 1.0000 1.0007 1.0014 1.0007
```

```
press(mod.2)
```

```
## [1] 467425.1
```

D (3)

```
mod.3 <- lm(Y ~ E1+E6.sqrt+E8+G21,data = projectdata)
summary(mod.3)
```

```
##
## Call:
## lm(formula = Y ~ E1 + E6.sqrt + E8 + G21, data = projectdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -43.527  -8.964  -0.255   8.605  48.930
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -23.2720     3.1355  -7.422 1.54e-13 ***
## E1            3.6740     0.2561  14.348  < 2e-16 ***
## E6.sqrt      32.7423     1.1315  28.938  < 2e-16 ***
## E8            9.1242     0.2566  35.553  < 2e-16 ***
## G21           6.2752     0.5269  11.909  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.2 on 2670 degrees of freedom
## Multiple R-squared:  0.4751, Adjusted R-squared:  0.4743
## F-statistic: 604.2 on 4 and 2670 DF,  p-value: < 2.2e-16
```

```
summary(mod.3)$adj.r.squared
```

```
## [1] 0.4743192
```

```
vif(mod.3)
```

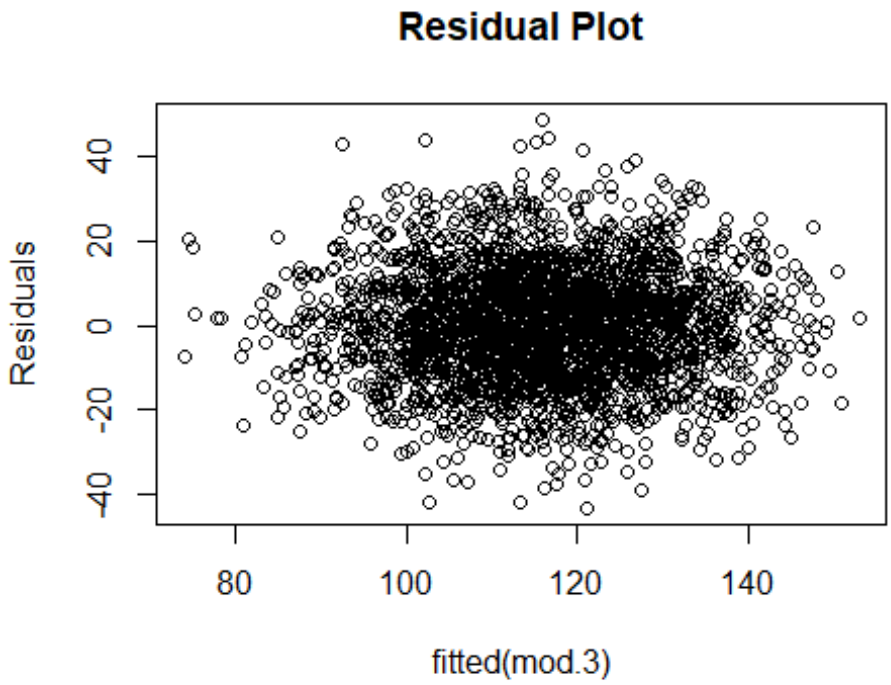
```
##      E1 E6.sqrt      E8     G21
##  1.0001  1.0008  1.0014  1.0007
```

```
press(mod.3)
```

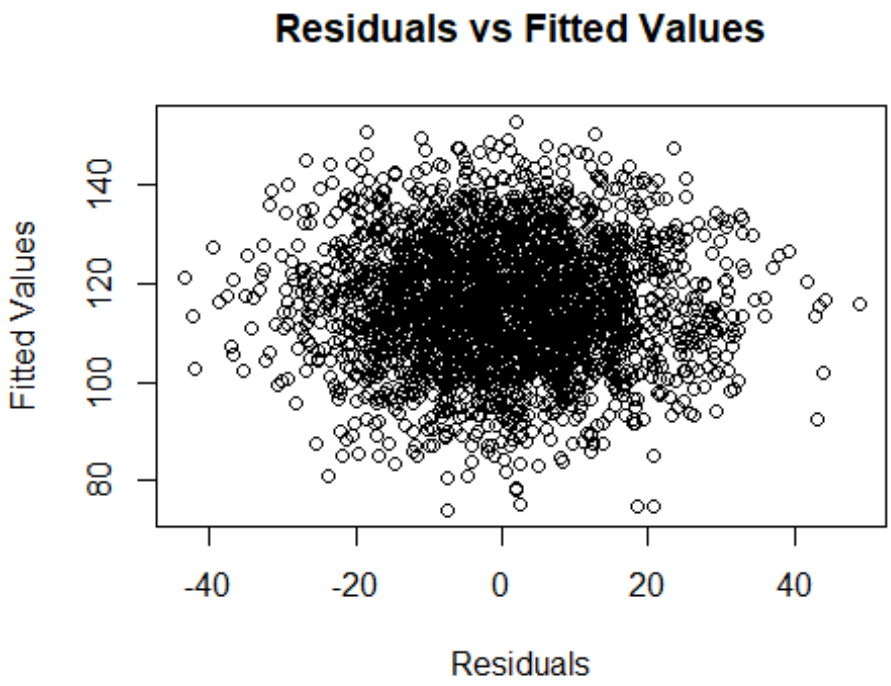
```
## [1] 467249.9
```

D (4)

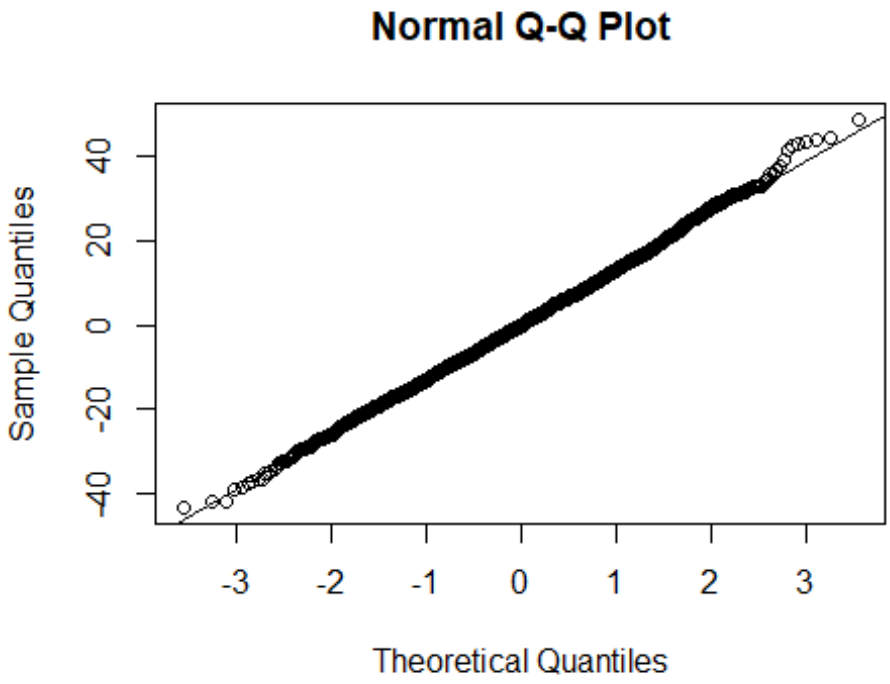
```
plot(resid(mod.3) ~ fitted(mod.3), main='Residual Plot',ylab='Residuals')
```



```
plot(mod.3$residuals,mod.3$fitted.values,xlab = 'Residuals',ylab='Fitted
Values',main = 'Residuals vs Fitted Values')
```




```
qqnorm(residuals(mod.3))
qqline(mod.3$residuals)
```




```
mean(mod.3$residuals)

## [1] 1.665335e-16

anova(mod.3)

## Analysis of Variance Table
##
## Response: Y
##             Df Sum Sq Mean Sq F value    Pr(>F)
## E1           1  36088   36088  206.98 < 2.2e-16 ***
## E6.sqrt      1 136079  136079  780.47 < 2.2e-16 ***
## E8           1 224476  224476 1287.46 < 2.2e-16 ***
## G21          1  24728   24728  141.82 < 2.2e-16 ***
## Residuals 2670 465528     174
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

D (5)

```
mod.onlyE <- lm(Y ~ E1+E6.sqrt+E8,data = projectdata)
summary(mod.onlyE)

##
## Call:
## lm(formula = Y ~ E1 + E6.sqrt + E8, data = projectdata)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -44.518  -9.414  -0.180   9.175  46.673
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -21.2038     3.2122  -6.601 4.91e-11 ***
## E1            3.6773     0.2627  13.997  < 2e-16 ***
## E6.sqrt      32.6865     1.1609  28.156  < 2e-16 ***
## E8            9.2053     0.2632  34.971  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.55 on 2671 degrees of freedom
## Multiple R-squared:  0.4472, Adjusted R-squared:  0.4466
## F-statistic: 720.3 on 3 and 2671 DF,  p-value: < 2.2e-16
```

```
mod.optimal <- lm(Y ~ E1+E6.sqrt+E8+G21,data = projectdata)
summary(mod.optimal)
```

```
##
## Call:
## lm(formula = Y ~ E1 + E6.sqrt + E8 + G21, data = projectdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -43.527  -8.964  -0.255   8.605  48.930
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -23.2720     3.1355  -7.422 1.54e-13 ***
## E1            3.6740     0.2561  14.348  < 2e-16 ***
## E6.sqrt      32.7423     1.1315  28.938  < 2e-16 ***
## E8            9.1242     0.2566  35.553  < 2e-16 ***
## G21           6.2752     0.5269  11.909  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.2 on 2670 degrees of freedom
## Multiple R-squared:  0.4751, Adjusted R-squared:  0.4743
## F-statistic: 604.2 on 4 and 2670 DF,  p-value: < 2.2e-16
```