Betsy Fridman - Project 1 Part A

Introduction

In this project, I was given two lists of independent and dependent variables, containing 696 observations total. Each list had a variable that was assigned to the ID numbers of the individuals in the sample. Some of the data provided was missing. If the observations have neither an independent or dependent variable, they must be removed. However, if the observations are only missing one of the variables, I have to impute the values, and then further find the linear regression line. Based on the independent variables and dependent variables provided and using the R Studio program, my objective is to generate a line of best fit for the dataset, and see if there is an association between the two variables. My main research question is: Is there a strong association between the independent and dependent variables?

Methods

To perform this statistical analysis, I imported my datasets into R Studio. As shown in Appendix 1, I imported the two separate datasets, one with ID and independent variables and the other with ID and dependent variables. Using the merge() function, as seen in Appendix 1, my next step was to merge the datasets and sort them by ID number. Using R, I inspected the pattern of missing data and created a plot that showcased how many variables were missing, shown in Appendix 2. Then, as seen in Appendix 3, I used the logical "or" operator to locate which observations had either an independent or dependent variable. I used that method to remove the 5 observations that were missing both independent and dependent variables, as those observations give us no information. Using the bootstrap method for linear regression and imputation, MICE in R bootstrapped the incomplete dataset and then imputed each bootstrap a number of times. My end result was a linear regression of the dependent variables over the independent variables, having the previously missing values be imputed. Further in Appendix 3, it's clear that the bootstrap method was used; mean imputation, median imputation, or listwise deletion were not used. Now, 691 observations are completed with both independent and dependent variables since I imputed the values. The next step was to fit a regression model to the data set and summarize it in an ANOVA table. In Appendix 4, I can see how I attained the lines of best fit and the information about the explained variance. Using the code shown in Appendix 5, I was able to create a scatter plot of all the points, which included the estimated linear regression line. Afterwards, I found the confidence intervals for the intercept and the independent variable coefficient with 95% and 99% confidence, which can be found in Appendix 6.

Results

Out of the total 696 observations, 612 had an independent variable and 637 had a dependent variable. Thus, the fraction of missing data in the independent variable was 0.1207 and the fraction of missing data in the dependent variable was 0.0848. There were 691 observations that had at least one independent variable value or dependent variable value and there were 558 observations with both an independent and dependent variable. The fitted function was $y = 49.1337 + 6.6963x$, where 49.1337 is the estimate for the intercept and 6.6963 is the estimate for the coefficient of x. Here, x represents the independent variables and y represents the dependent variables. The value for r-squared was 0.8477, and this shows the fraction of the variation of the dependent variable that was explained. The correlation coefficient r was calculated to be 0.9207. According to the ANOVA table, the sum of squares explained by the regression is 31130.884, and the total sum of squares is 36722.093. If I divide the explained by the total, I get the same value as our r-squared, and this shows that 84.77% of our data is explained by the dependent variables. The 95% confidence interval for the intercept was (48.0559, 50.2114) and the 95% confidence interval for the independent variable coefficient was (6.4840, 6.9086). The 99% confidence interval for the intercept is (47.7159, 50.5515) and the 99% confidence interval for the independent variable coefficient was (6.4170, 6.9755). The null hypothesis of the slope being 0 would be rejected because 0 is not in the 95% or 99% confidence interval that I found for the coefficient of the independent variable.

Conclusion

For this project, I tested 696 observations to see if there was association between the independent and dependent variables. After removing the observations missing both variables and imputing variables for the observations missing one, I created a linear regression line. The fitted function was $y = 49.1337 + 6.6963x$, with x representing the independent variables and y representing the dependent variables. The r-squared value was 0.8477 and the correlation coefficient was 0.9207, meaning that there was a very strong association between the independent and dependent variables. Thus, a linear regression line is the best regression line for this dataset, and our fitted function was very strongly associated with our dataset.

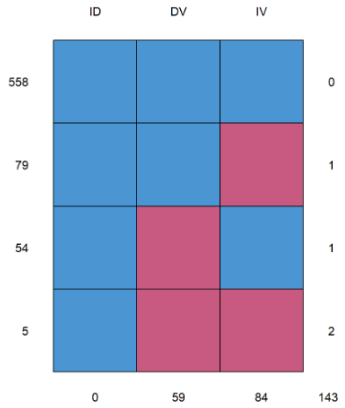
Appendix A

Appendix 1 - Importing and Merging the Datasets

```
> P1A_IV_sample <-
read.csv("C:/Users/Angel/OneDrive/Desktop/P1A_IV_sample.csv")
> P1A_DV_sample <-
read.csv("C:/Users/Angel/OneDrive/Desktop/P1A_DV_sample.csv")
> PartA <- merge(P1A_IV_sample,P1A_DV_sample,by = 'ID')
```

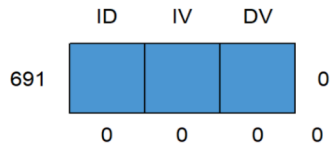
Appendix 2 - Pattern of Missing Variables in the Merged Dataset

```
> PartA_incomplete <- PartA
> install.packages('mice')
> library(mice)
> md.pattern(PartA)
```



Appendix 3 - Pattern of Complete Merged Dataset

```
> PartA_imp <- PartA[!is.na(PartA$IV)==TRUE|!is.na(PartA$DV)==TRUE,]
> imp <- mice(PartA_imp,method="norm.boot",printFlag=FALSE)
> PartA_complete <- complete(imp)
> md.pattern(PartA_complete)
```



Appendix 4 - Summary of the Variables and ANOVA table

```
> M <- lm(DV ~ IV, data=PartA_complete)
> summary(M)
> install.packages('knitr')
> library(knitr)
> kable(anova(M),caption='ANOVA Table')
```

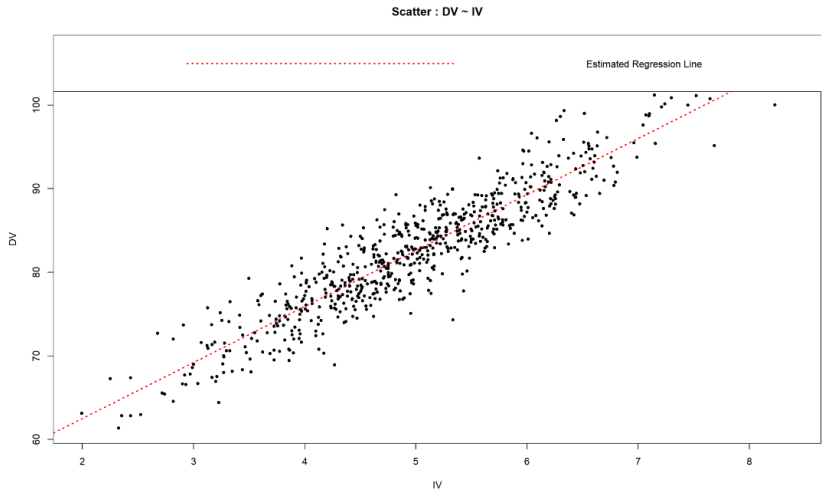
```
Residuals:
     Min       1Q   Median       3Q      Max
-10.5239  -2.0239  -0.0515   1.9608   7.9332

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  49.1337     0.5489   89.51   <2e-16 ***
IV            6.6963     0.1081   61.94   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.849 on 689 degrees of freedom
Multiple R-squared:  0.8477,    Adjusted R-squared:  0.8475
F-statistic:  3836 on 1 and 689 DF,  p-value: < 2.2e-16
```

Table: ANOVA Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
IV	1	31130.884	31130.884325	3836.233	0
Residuals	689	5591.209	8.114962	NA	NA

Appendix 5 - Scatter Plot of Complete Dataset with Estimated Regression Line

```
> plot(PartA_complete$DV ~ PartA_complete$IV, main='Scatter : DV ~ IV',
xlab='IV', ylab='DV', pch=20)
> abline(M, col='red', lty=3, lwd=2)
> legend('topleft', legend='Estimated Regression Line', lty=3, lwd=2,
col='red')
```



Appendix 6 - 95% and 99% Confidence Intervals for the Coefficient of the Intercept and the Independent Variable

```
> confint(M, level = 0.95)
                2.5 %    97.5 %
(Intercept) 48.055938 50.211408
IV           6.484012  6.908556
> confint(M, level = 0.99)
                0.5 %   99.5 %
(Intercept) 47.715850 50.55150
IV           6.417028  6.97554
```

Betsy Fridman - Project 1 Part B

Introduction

In this project, I was given a list containing 569 observations. Each observation had an independent and dependent variable. The data points formed a pattern that resembled a line that was decreasing at a decreasing rate. Rather than completing a linear regression, which would very poorly represent the data, my goal was to create a fitted nonlinear model by transforming the data. My objective is to choose a transformation, of the independent or dependent variable, that most accurately fits my data. By using the R Studio program, I will be able to recover the best function that was used to generate the dependent variable value based on the independent variable value. My main research question is: Which transformation of the independent or dependent variable best fits the given dataset?

Methods

To begin my statistical analysis, I imported my dataset into R Studio. As shown in Appendix 1, after importing my dataset, I created a scatter plot and a linear model for my dataset, along with a summary table. My first step is to determine which transformation to apply to the data to find the nonlinear line of best fit. According to the textbook provided, since my scatter plot looked as though it was decreasing at a decreasing rate, it is highly recommended to test the transformation of x to ln(x), 1/x, and sqrt(x), and y to y^2. Since these transformation plots may not be as strongly associated with my scatter plot as desired, I will additionally test the transformation of x to 1/sqrt(x). For each transformation, I have to follow a specific process, starting with the transformation of the dataset itself. Using the data.frame() function, I am able to transform the variables of my data in any way. Afterwards, I can use the cut() function to generate my groups, then average the x values, as a way to bin my data points together. By binning my data together, I simplify my dataset and may reduce the possibility of error. Each of my groups will have close to 50 groups, to attain an accurate r-squared. Once I create the groups and average the x values, I will have officially binned my data, and then I can begin to apply the Lack of Fit test. Using pureErrorAnova() and looking for the higher p-values for the Lack of Fit component in the ANOVA table will help us identify the best fit. Finally, when I create the summary table for the transformation of x or y, the r-squared will be given, and this will be the best indicator of a good fit. All of the previously mentioned steps will be applied to each transformation, in hopes of finding the best fit. In the Appendices 2,5,8,11, and 14, the steps of this statistical analysis are shown for each chosen transformation.

Results

The linear model I created for my data set had an r-squared of 0.5259, which was moderately low. When transforming x to sqrt(x), I chose to bin the data into intervals of 0.04 which created 52 groups, as seen in Appendices 2 and 3. From the ANOVA table and Summary table in Appendix 4, the p-value is extremely low and the r-squared value is 0.5877. I transformed x to ln(x) next, and grouped the data into intervals of 0.05 which created 55 groups, as seen in Appendices 5 and 6. In Appendix 7, the ANOVA and Summary tables show us that the p-value is 0.1532 and the r-squared value is 0.6356. The third transformation is x to 1/x, and the data is grouped into 46 groups with an interval of 0.04, shown in Appendices 8 and 9. Appendix 10 shows us the ANOVA and Summary tables, where we can see that the p-value is extremely low and the r-squared value is 0.6166. Next, as seen in Appendices 11 and 12, the first transformation of y was done. The transformation of y to y^2 was cut into intervals of 0.15, amounting to 49 groups. This transformation had an extremely low p-value and the r-squared value was 0.5249, as written in the ANOVA and Summary tables in Appendix 13. Finally, the last transformation was from x to 1/sqrt(x). As shown in Appendices 14 and 15, using intervals of 0.02, the data was cut and binned into 53 groups. This transformation, as displayed in Appendix 16, produced a p-value of 0.636 and an r-squared value of 0.6503, which shows the fraction of the variation of the dependent variable that was explained. In Appendix 17, we can finally see how the transformed fitted line looked with our dataset, and we can similarly see the residual plot. Found in the Summary table of Appendix 16, the fitted line for the final transformation was $y = 7.897540 + 0.400509x$, with 7.897540 being the estimate for the intercept and

0.400509 being the estimate for the coefficient of the independent variable. As r-squared was found to be 0.6503, r is equal to 0.8064, which is the correlation coefficient.

Conclusion

After finishing the statistical analysis, I have come to the conclusion that the best transformation for our dataset is transforming x into 1/sqrt(x). This provided the highest p-value and r-squared value of all the transformation trials, being 0.636 and 0.6503 respectively. The r-squared value of the transformation was greater than the r-squared value of the linear regression, which was 0.5259. The correlation coefficient of the transformation was 0.8064, meaning that there was a very strong association between our transformed dataset and the transformed fitted line, $y = 7.897540 + 0.400509x$, as seen in the graphs of Appendix 17. Additionally, the fairly high p-value from the Lack of Fit test helps show that there is not a significant Lack of Fit in this regression model. Thus, transforming x into 1/sqrt(x) is the best transformation for this given dataset.

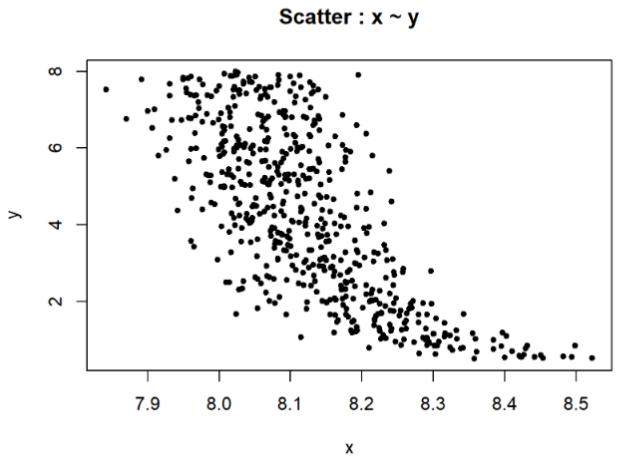
Appendix B

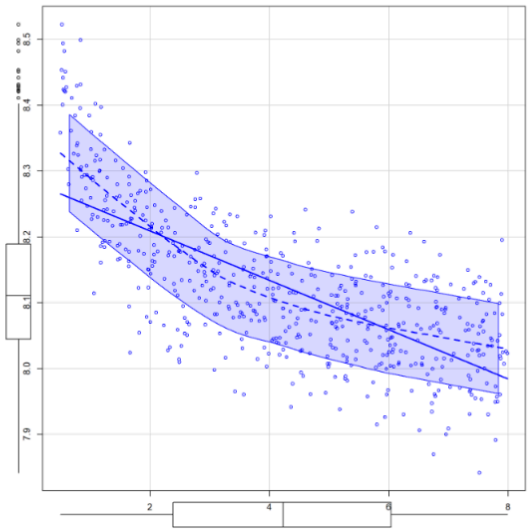
Appendix 1 - Importing the Dataset and Creating a Scatter Plot of the Dataset

```
> data <- read_csv("C:/Users/Angel/OneDrive/Desktop/P1B_sample.csv")
> plot(data$x ~ data$y, main='Scatter : x ~ y',xlab = 'x',ylab = 'y', pch=20)
> scatterplot(data$x,data$y)
```

```
> originalData <- lm(y ~ x, data=data)
```

```
> summary(originalData)
```





```
Call:
lm(formula = y ~ x, data = data)

Residuals:
      Min        1Q    Median        3Q       Max
-0.197188 -0.052660 -0.002599  0.050383  0.258308

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.284054   0.007131 1161.70   <2e-16 ***
x           -0.037500   0.001495  -25.08   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07753 on 567 degrees of freedom
Multiple R-squared:  0.5259,    Adjusted R-squared:  0.525
F-statistic: 628.8 on 1 and 567 DF,  p-value: < 2.2e-16
```

Appendix 2 - Transformation from x to sqrt(x): Process

```
> data_trans <- data.frame(xtrans=sqrt(data$x), ytrans=data$y)
> groups <-
cut(data_trans$xtrans,breaks=c(-Inf,seq(min(data_trans$xtrans)+0.04,
max(data_trans$xtrans)-0.04,by=0.04),Inf))
> table(groups)
> x <- ave(data_trans$xtrans, groups)
> data_bin <- data.frame(x=x, y=data_trans$ytrans)
> library(remotes)
> library(alr3)
> fit_b <- lm(y ~ x, data = data_bin)
> pureErrorAnova(fit_b)
> summary(fit_b)
```

Appendix 3 - Transformation from x to sqrt(x): Groups

```
groups
 (-Inf,0.747] (0.747,0.787] (0.787,0.827] (0.827,0.867]
            7             5             3             3
(0.867,0.907] (0.907,0.947] (0.947,0.987]  (0.987,1.03]
            5             7             5             9
  (1.03,1.07]   (1.07,1.11]   (1.11,1.15]   (1.15,1.19]
            5             8            10             3
  (1.19,1.23]   (1.23,1.27]   (1.27,1.31]   (1.31,1.35]
            8             4            14             6
  (1.35,1.39]   (1.39,1.43]   (1.43,1.47]   (1.47,1.51]
            5            16             8             6
  (1.51,1.55]   (1.55,1.59]   (1.59,1.63]   (1.63,1.67]
            6            12             9            10
  (1.67,1.71]   (1.71,1.75]   (1.75,1.79]   (1.79,1.83]
            9             9            15            10
  (1.83,1.87]   (1.87,1.91]   (1.91,1.95]   (1.95,1.99]
            8            12            12             9
  (1.99,2.03]   (2.03,2.07]   (2.07,2.11]   (2.11,2.15]
           16            13            14            13
  (2.15,2.19]   (2.19,2.23]   (2.23,2.27]   (2.27,2.31]
           10             6            18            22
  (2.31,2.35]   (2.35,2.39]   (2.39,2.43]   (2.43,2.47]
           13            10            12            27
  (2.47,2.51]   (2.51,2.55]   (2.55,2.59]   (2.59,2.63]
           10            14            11            25
  (2.63,2.67]   (2.67,2.71]   (2.71,2.75]   (2.75, Inf]
           10             7            24            36
```

Appendix 4 - Transformation from x to sqrt(x): ANOVA Table and Summary

```
Analysis of Variance Table

Response: y
             Df Sum Sq Mean Sq  F value    Pr(>F)
x             1 4.2240  4.2240 937.4580 < 2.2e-16 ***
Residuals   567 2.9638  0.0052
 Lack of fit  50 0.6343  0.0127   2.8153  4.64e-09 ***
 Pure Error  517 2.3295  0.0045
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Call:
lm(formula = y ~ x, data = data_bin)

Residuals:
      Min        1Q    Median        3Q       Max
-0.203633 -0.050494 -0.000634  0.050622  0.215928

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.419730   0.010806  779.20   <2e-16 ***
x           -0.149102   0.005245  -28.43   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0723 on 567 degrees of freedom
Multiple R-squared:  0.5877,    Adjusted R-squared:  0.5869
F-statistic: 808.1 on 1 and 567 DF,  p-value: < 2.2e-16
```

Appendix 5 - Transformation from x to ln(x): Process

```
> data_trans <- data.frame(xtrans=sqrt(data$x), ytrans=data$y)
> groups <-
cut(data_trans$xtrans,breaks=c(-Inf,seq(min(data_trans$xtrans)+0.05,
max(data_trans$xtrans)-0.05,by=0.05),Inf))
> table(groups)
> x <- ave(data_trans$xtrans, groups)
> data_bin <- data.frame(x=x, y=data_trans$ytrans)
> library(remotes)
> library(alr3)
> fit_b <- lm(y ~ x, data = data_bin)
> pureErrorAnova(fit_b)
> summary(fit_b)
```

Appendix 6 - Transformation from x to ln(x): Groups

```
groups
    (-Inf,-0.643]   (-0.643,-0.593]   (-0.593,-0.543]
                2                 4                 3
  (-0.543,-0.493]   (-0.493,-0.443]   (-0.443,-0.393]
                3                 2                 0
  (-0.393,-0.343]   (-0.343,-0.293]   (-0.293,-0.243]
                2                 1                 4
  (-0.243,-0.193]   (-0.193,-0.143]  (-0.143,-0.0926]
                2                 7                 1
(-0.0926,-0.0426] (-0.0426,0.00743]  (0.00743,0.0574]
                4                 1                 8
   (0.0574,0.107]     (0.107,0.157]     (0.157,0.207]
                4                 5                 5
    (0.207,0.257]     (0.257,0.307]     (0.307,0.357]
                8                 2                 3
    (0.357,0.407]     (0.407,0.457]     (0.457,0.507]
                7                 3                 9
    (0.507,0.557]     (0.557,0.607]     (0.607,0.657]
                7                 6                 6
    (0.657,0.707]     (0.707,0.757]     (0.757,0.807]
               13                 9                 4
    (0.807,0.857]     (0.857,0.907]     (0.907,0.957]
                7                 8                11
     (0.957,1.01]       (1.01,1.06]       (1.06,1.11]
                8                11                11
      (1.11,1.16]       (1.16,1.21]       (1.21,1.26]
               16                12                 7
      (1.26,1.31]       (1.31,1.36]       (1.36,1.41]
               13                13                21
      (1.41,1.46]       (1.46,1.51]       (1.51,1.56]
               17                16                16
      (1.56,1.61]       (1.61,1.66]       (1.66,1.71]
               10                27                24
      (1.71,1.76]       (1.76,1.81]       (1.81,1.86]
               16                34                19
      (1.86,1.91]       (1.91,1.96]       (1.96,2.01]
               19                30                21
     (2.01, Inf]
               47
```

Appendix 7 - Transformation from x to ln(x): ANOVA Table and Summary

```
Analysis of Variance Table

Response: y
              Df Sum Sq Mean Sq   F value Pr(>F)
x              1 4.5688  4.5688 1008.5984 <2e-16 ***
Residuals    567 2.6189  0.0046
 Lack of fit  52 0.2860  0.0055    1.2142 0.1532
 Pure Error  515 2.3329  0.0045
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Call:
lm(formula = y ~ x, data = data_bin)

Residuals:
      Min        1Q    Median        3Q       Max
-0.197821 -0.045879  0.000365  0.049071  0.185359

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.290986   0.006001 1381.67   <2e-16 ***
x           -0.131826   0.004191  -31.45   <2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06796 on 567 degrees of freedom
Multiple R-squared:  0.6356,    Adjusted R-squared:  0.635
F-statistic: 989.2 on 1 and 567 DF,  p-value: < 2.2e-16
```

Appendix 8 - Transformation from x to 1/x: Process

```
> data_trans <- data.frame(xtrans=(1/(data$x)), ytrans=data$y)
> groups <-
cut(data_trans$xtrans,breaks=c(-Inf,seq(min(data_trans$xtrans)+0.04,
max(data_trans$xtrans)-0.04,by=0.04),Inf))
> table(groups)
> x <- ave(data_trans$xtrans, groups)
> data_bin <- data.frame(x=x, y=data_trans$ytrans)
> library(remotes)
> library(alr3)
> fit_b <- lm(y ~ x, data = data_bin)
> pureErrorAnova(fit_b)
> summary(fit_b)
```

Appendix 9 - Transformation from x to 1/x: Groups

```
groups
 (-Inf,0.165] (0.165,0.205] (0.205,0.245] (0.245,0.285] (0.285,0.325]
          142            98            56            47            30
(0.325,0.365] (0.365,0.405] (0.405,0.445] (0.445,0.485] (0.485,0.525]
           27            20            14            10            18
(0.525,0.565] (0.565,0.605] (0.605,0.645] (0.645,0.685] (0.685,0.725]
            8             9             9             8             4
(0.725,0.765] (0.765,0.805] (0.805,0.845] (0.845,0.885] (0.885,0.925]
            2             7             4             7             3
(0.925,0.965]  (0.965,1.01]   (1.01,1.05]   (1.05,1.09]   (1.09,1.13]
            5             5             1             4             1
  (1.13,1.17]   (1.17,1.21]   (1.21,1.25]   (1.25,1.29]   (1.29,1.33]
            1             6             0             3             2
  (1.33,1.37]   (1.37,1.41]   (1.41,1.45]   (1.45,1.49]   (1.49,1.53]
            1             1             1             1             0
  (1.53,1.57]   (1.57,1.61]   (1.61,1.65]   (1.65,1.69]   (1.69,1.73]
            1             1             0             1             2
  (1.73,1.77]   (1.77,1.81]   (1.81,1.85]   (1.85,1.89]   (1.89,1.93]
            1             1             3             1             2
  (1.93, Inf]
            1
```

Appendix 10 - Transformation from x to 1/x: ANOVA Table and Summary

```
Analysis of Variance Table

Response: y
             Df Sum Sq Mean Sq  F value    Pr(>F)
x             1 4.4323  4.4323 996.5212 < 2.2e-16 ***
Residuals   567 2.7555  0.0049
 Lack of fit  41 0.4159  0.0101   2.2809 1.983e-05 ***
 Pure Error  526 2.3395  0.0044
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Call:
lm(formula = y ~ x, data = data_bin)

Residuals:
      Min        1Q    Median        3Q       Max
-0.223730 -0.047656  0.000586  0.051484  0.179484

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 8.029282   0.004308  1863.6   <2e-16 ***
x           0.256000   0.008477    30.2   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06971 on 567 degrees of freedom
Multiple R-squared:  0.6166,    Adjusted R-squared:  0.616
F-statistic:   912 on 1 and 567 DF,  p-value: < 2.2e-16
```

Appendix 11 - Transformation from y to y^2: Process

```
> data_trans <- data.frame(xtrans=data$x, ytrans=data$y^(2))
> groups <-
cut(data_trans$xtrans,breaks=c(-Inf,seq(min(data_trans$xtrans)+0.15,
max(data_trans$xtrans)-0.15,by=0.15),Inf))
> table(groups)
> x <- ave(data_trans$xtrans, groups)
> data_bin <- data.frame(x=x, y=data_trans$ytrans)
> install.packages('remotes')
> library(remotes)
> install_github("cran/alr3")
> library(alr3)
> fit_b <- lm(y ~ x, data = data_bin)
> pureErrorAnova(fit_b)
> summary(fit_b)
```

Appendix 12 - Transformation from y to y^2: Groups

```
groups
(-Inf,0.65] (0.65,0.8] (0.8,0.95] (0.95,1.1] (1.1,1.25] (1.25,1.4] (1.4,1.55] (1.55,1.7]
         14          9         12         12         15          8         11         15
 (1.7,1.85]   (1.85,2]   (2,2.15] (2.15,2.3] (2.3,2.45] (2.45,2.6] (2.6,2.75] (2.75,2.9]
          9         12         14          6         11         13          9         11
 (2.9,3.05] (3.05,3.2] (3.2,3.35] (3.35,3.5] (3.5,3.65] (3.65,3.8] (3.8,3.95] (3.95,4.1]
         11         15         12          7         11         12          9         16
 (4.1,4.25] (4.25,4.4] (4.4,4.55] (4.55,4.7] (4.7,4.85]   (4.85,5]   (5,5.15] (5.15,5.3]
         13         10         13          8          9          6         16         15
 (5.3,5.45] (5.45,5.6] (5.6,5.75] (5.75,5.9] (5.9,6.05] (6.05,6.2] (6.2,6.35] (6.35,6.5]
         15          5         14          7         22         12         10          7
 (6.5,6.65] (6.65,6.8] (6.8,6.95] (6.95,7.1] (7.1,7.25] (7.25,7.4] (7.4,7.55] (7.55,7.7]
          9         16         13          8          3         13         15         13
 (7.7, Inf]
         23
```

Appendix 13 - Transformation from y to y^2: ANOVA Table and Summary

```
Analysis of Variance Table

Response: y
             Df  Sum Sq Mean Sq  F value    Pr(>F)
x             1 1005.89 1005.89 846.2748 < 2.2e-16 ***
Residuals   567  910.54    1.61
 Lack of fit  47  292.46    6.22   5.2352 < 2.2e-16 ***
 Pure Error  520  618.08    1.19
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Call:
lm(formula = y ~ x, data = data_bin)

Residuals:
    Min      1Q  Median      3Q     Max
-3.2246 -0.8676 -0.0508  0.8296  4.3572

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 68.62351    0.11658  588.65   <2e-16 ***
x           -0.61189    0.02445  -25.03   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.267 on 567 degrees of freedom
Multiple R-squared:  0.5249,    Adjusted R-squared:  0.524
F-statistic: 626.4 on 1 and 567 DF,  p-value: < 2.2e-16
```

Appendix 14 - Transformation from x to 1/sqrt(x): Process

```
> data_trans <- data.frame(xtrans=(1/sqrt(data$x)), ytrans=data$y)
> groups <-
cut(data_trans$xtrans,breaks=c(-Inf,seq(min(data_trans$xtrans)+0.02,
max(data_trans$xtrans)-0.02,by=0.02),Inf))
> table(groups)
> x <- ave(data_trans$xtrans, groups)
> data_bin <- data.frame(x=x, y=data_trans$ytrans)
> library(remotes)
> library(alr3)
> fit_b <- lm(y ~ x, data = data_bin)
> pureErrorAnova(fit_b)
> summary(fit_b)
```

Appendix 15 - Transformation from x to 1/sqrt(x): Groups

```
groups
 (-Inf,0.374] (0.374,0.394] (0.394,0.414] (0.414,0.434]
           67            50            51            32
(0.434,0.454] (0.454,0.474] (0.474,0.494] (0.494,0.514]
           40            27            28            26
(0.514,0.534] (0.534,0.554] (0.554,0.574] (0.574,0.594]
           22            17            17            14
(0.594,0.614] (0.614,0.634] (0.634,0.654] (0.654,0.674]
           14            13             9             9
(0.674,0.694] (0.694,0.714] (0.714,0.734] (0.734,0.754]
            6            14             8             6
(0.754,0.774] (0.774,0.794] (0.794,0.814] (0.814,0.834]
            8            10             3             6
(0.834,0.854] (0.854,0.874] (0.874,0.894] (0.894,0.914]
            4             1             4             7
(0.914,0.934] (0.934,0.954] (0.954,0.974] (0.974,0.994]
            7             2             3             7
 (0.994,1.01]   (1.01,1.03]   (1.03,1.05]   (1.05,1.07]
            2             3             2             0
  (1.07,1.09]   (1.09,1.11]   (1.11,1.13]   (1.13,1.15]
            6             1             3             2
  (1.15,1.17]   (1.17,1.19]   (1.19,1.21]   (1.21,1.23]
            1             1             2             0
  (1.23,1.25]   (1.25,1.27]   (1.27,1.29]   (1.29,1.31]
            1             1             1             2
  (1.31,1.33]   (1.33,1.35]   (1.35,1.37]   (1.37,1.39]
            2             2             2             2
  (1.39, Inf]
            1
```

Appendix 16 - Transformation from x to 1/sqrt(x): ANOVA Table and Summary

```
Analysis of Variance Table

Response: y
              Df Sum Sq Mean Sq   F value Pr(>F)
x              1 4.6739  4.6739 1046.6362 <2e-16 ***
Residuals    567 2.5138  0.0044
 Lack of fit  49 0.2006  0.0041    0.9167  0.636
 Pure Error  518 2.3132  0.0045
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Call:
lm(formula = y ~ x, data = data_bin)

Residuals:
      Min        1Q    Median        3Q       Max
-0.200896 -0.043425  0.000003  0.046500  0.170903

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.897540   0.007538 1047.71   <2e-16 ***
x           0.400509   0.012335   32.47   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06658 on 567 degrees of freedom
Multiple R-squared:  0.6503,    Adjusted R-squared:  0.6496
F-statistic:  1054 on 1 and 567 DF,  p-value: < 2.2e-16
```

Appendix 17 - Transformation from x to 1/sqrt(x): Transformed Fitted Scatter Plot and Residual Plot

```
> scatterplot(data_trans$x,data_trans$y)
> plot(data_trans$x ~ data_trans$y, main='Scatter : x ~ y',xlab = 'x',ylab =
'y', pch=20)
> plot(fitted(fit_b),resid(fit_b))
> abline(0,0)
```