Angel Li 112784616 AMS 315 Project 1 Part A

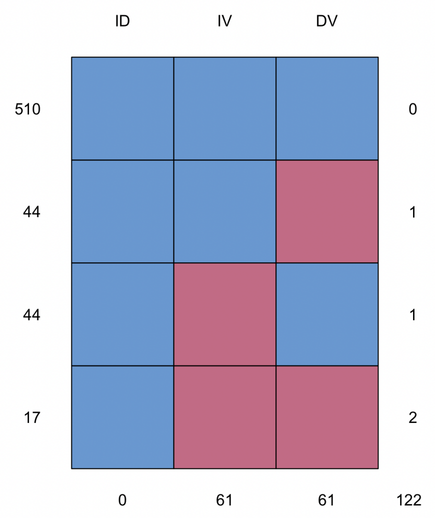
Introduction

In Part A, we looked at two files, one file containing a column for subject ID and an independent variable value, and another file containing a column for subject ID and a dependent variable value. The purpose of this report was to statistically process the data we were given. With the given data, we generate an ANOVA table to display the information.

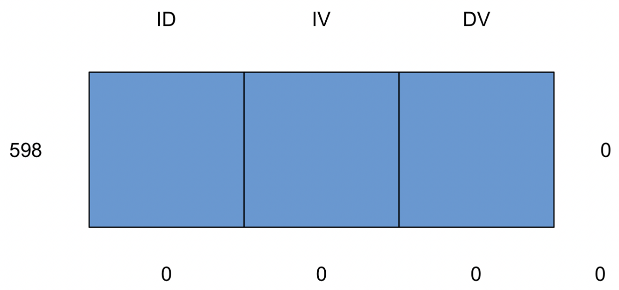
Methods

We used R to process the data. Within R, the file containing the independent variables and the file containing the dependent file were read. Then the R function, merge, was used to merge the two files into one file. There were 510 complete data sets. After that, we had to deal with missing data values. The statistical package, mice, was used for the imputation of the missing values. The md.pattern() function was used to inspect the pattern of missing data. It was shown that there were 61 cases of independent variable values missing, 61 cases of dependent variable values missing, and both independent and dependent variable variables missing in 17 cases. We deal with the missing data by first dropping the 17 observations that are missing in both variables. Once we saw the data set complete, there were 559 data sets.We used linear regression by bootstrap. We then fit a regression model to the data set and generated an anova table along with finding the 95% confidence interval of the slope and 99% confidence interval of the slope .

Pattern of missing data



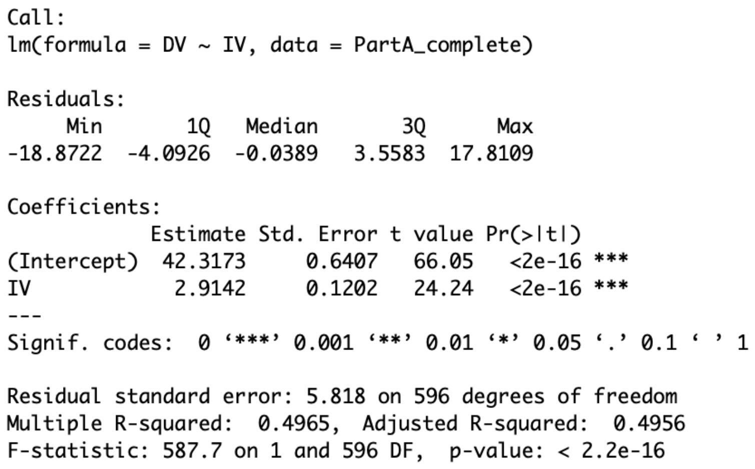
Complete data set



Results

The fitted function was DV = 2.9142IV + 42.3173. The ANOVA table computed for the data is shown below. The residual standard error was 5.818 on 596 degrees of freedom.. The multiple r-square was 0.4965 and the adjusted r squared was 0.4956. The F statistic was 587.7 on 1 and 596 degrees of freedom. The 95% confidence interval for the slope was [2.678118, 3.150295]. The 99% confidence interval for the slope was [2.603569, 3.224844]. The p value was < 2e -16. Null hypothesis is rejected that the slope was zero.

Regression model

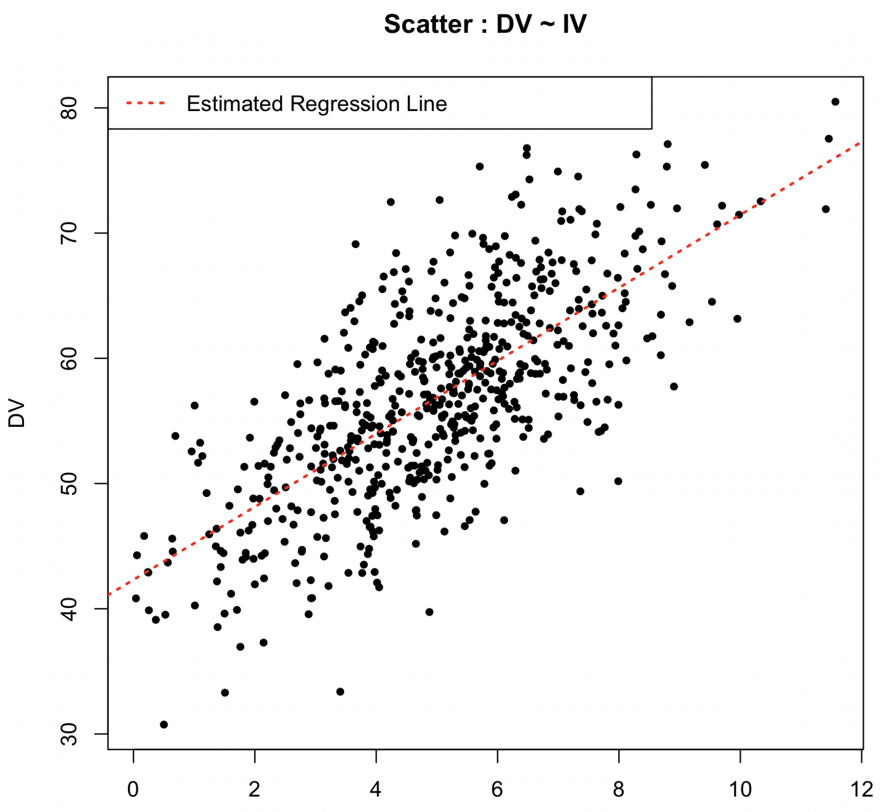


```
Call:
lm(formula = DV ~ IV, data = PartA_complete)

Residuals:
     Min       1Q   Median       3Q      Max 
-18.8722  -4.0926  -0.0389   3.5583  17.8109 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  42.3173     0.6407   66.05   <2e-16 ***
IV            2.9142     0.1202   24.24   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.818 on 596 degrees of freedom
Multiple R-squared:  0.4965,	Adjusted R-squared:  0.4956 
F-statistic: 587.7 on 1 and 596 DF,  p-value: < 2.2e-16
```

Conclusions and Discussion

The R squared value being 0.4956 and the p value being < 2e -16 showed that moderate effect size on the dependent variables from the independent variables. There is clear indication of an association between the variables. The fitted function of DV = 2.9142IV + 42.3173 describes the relationship between the independent variable and the dependent variable. The table below shows the scatter plot with an estimated regression line for the data.

Dataset plot



Angel Li 112784616 AMS 315 Project 1 Part B

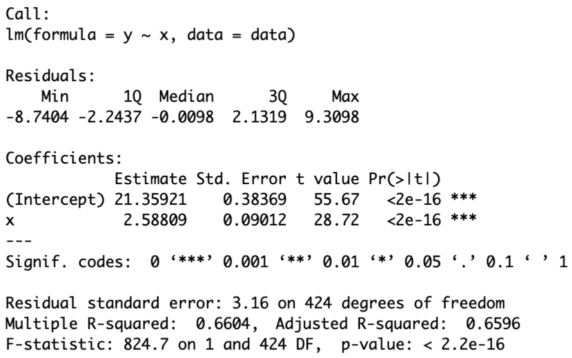
Introduction

In Part B, we were given a subject ID, an x value, and a y value. The data values may have had a transformation applied to them so it was necessary to look into if there actually was one. From there, there was a fitted transformed regression model and data was binned and the data was applied the lack of fit test.

Methods

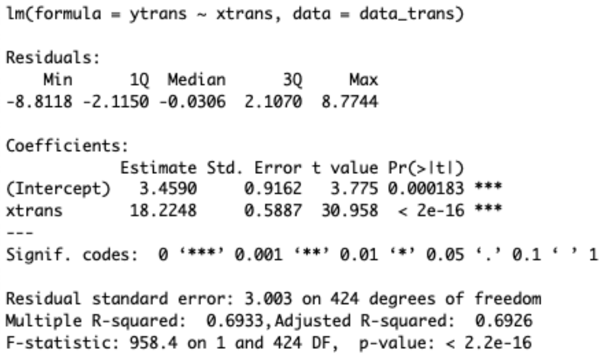
I used R to process the data. First, I needed to see whether a transformation was done on the x value, y value, or even both. I first found the regression model for the given data and looked at the R-squared value. The R-squared value was 0.6604. From there, I did the same procedure but with transformations such as square root, cube root, power, and logarithmic transformations. I compared the R-squared values by performing the transformations and found the transformation with the largest R-squared value. Regression models of the original data and the transformed data are shown below to the left and right respectively.From there, I found that a transformation was indeed done and it happened to be a transformation of x^(1/3). After discovering the transformation, the transformed data was cut to generate groups for binned data. From there, the pureErrorAnova() function was used to apply the Lack of Fit Test.

Regression Model y~x



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

Residuals:
    Min      1Q  Median      3Q     Max 
-8.7404 -2.2437 -0.0098  2.1319  9.3098 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 21.35921    0.38369   55.67   <2e-16 ***
x            2.58809    0.09012   28.72   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.16 on 424 degrees of freedom
Multiple R-squared:  0.6604,	Adjusted R-squared:  0.6596 
F-statistic: 824.7 on 1 and 424 DF,  p-value: < 2.2e-16
```

Regression Model y~x^(⅓)

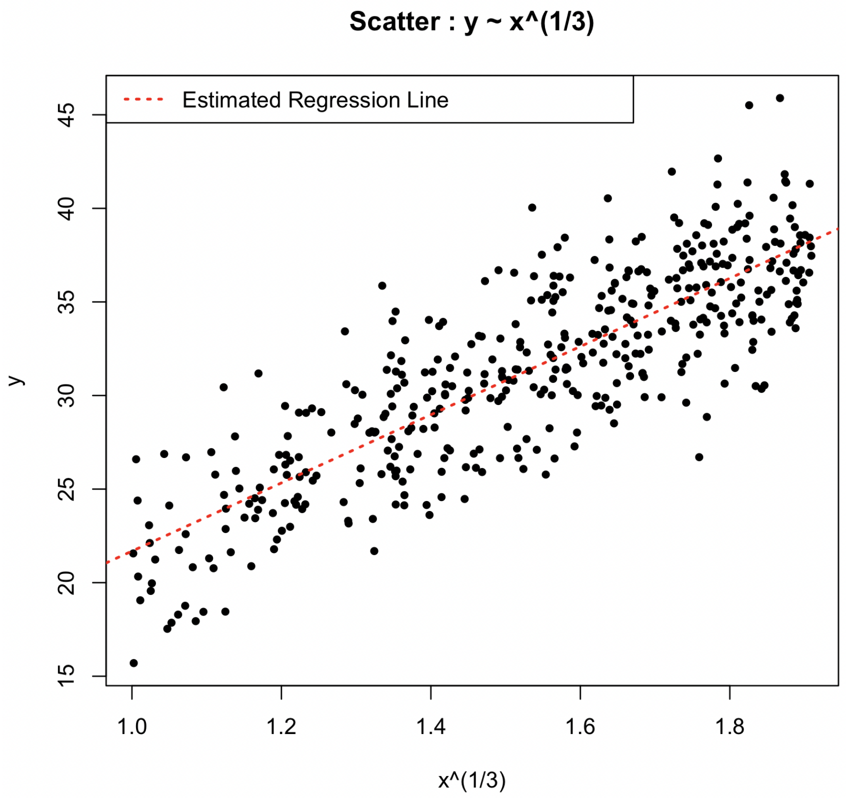


```
lm(formula = ytrans ~ xtrans, data = data_trans)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.8118 -2.1150 -0.0306  2.1070  8.7744 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   3.4590     0.9162   3.775 0.000183 ***
xtrans       18.2248     0.5887  30.958  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.003 on 424 degrees of freedom
Multiple R-squared:  0.6933,Adjusted R-squared:  0.6926 
F-statistic: 958.4 on 1 and 424 DF,  p-value: < 2.2e-16
```

Results

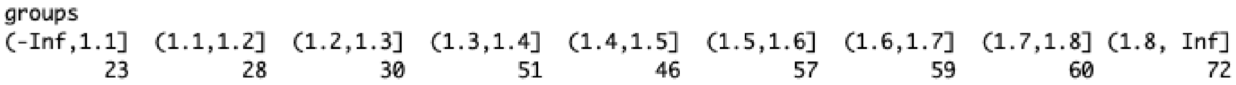
The given data had a transformation of x^(1/3) applied. The data set is plotted and shown below with the estimated regression line. The R-squared value of the original data was 0.6604. After trying different transformations, the biggest R-squared value was associated with a transformation of x^(1/3) with 0.6933. Then the data was cut to generate groups of binned data. The binned data is shown below the graph of plotted values.Using the pureErrorAnova() function, we got an analysis of variance table, which is shown at the bottom of the next page.The lack of fit P-value was 0.269. The high P-value shows that the model does not have much significant lack of fit. We accept the null hypothesis that the linear model is adequate.

Dataset plot

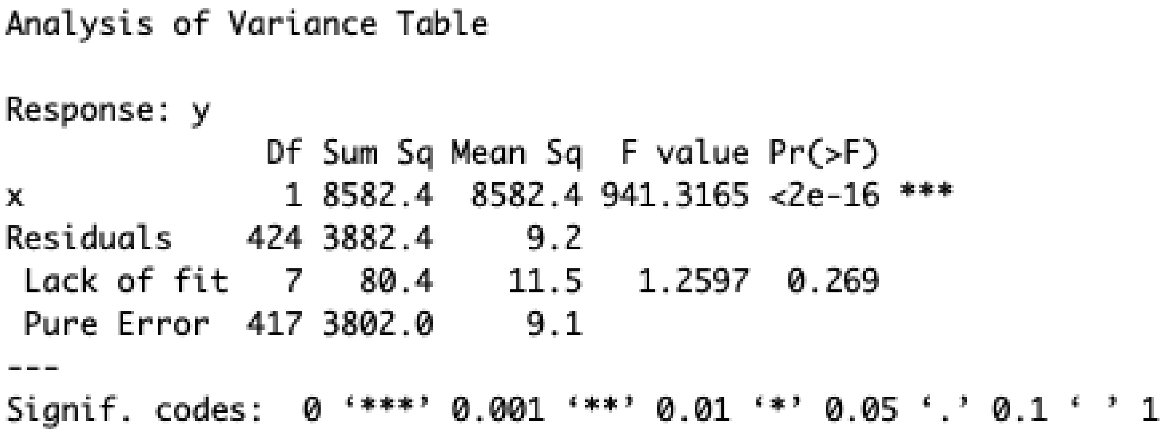


Scatter : y ~ x^(1/3)

Grouped data



```
groups
(-Inf,1.1]  (1.1,1.2]  (1.2,1.3]  (1.3,1.4]  (1.4,1.5]  (1.5,1.6]  (1.6,1.7]  (1.7,1.8] (1.8, Inf]
        23         28         30         51         46         57         59         60         72
```



```
Analysis of Variance Table

Response: y
             Df Sum Sq Mean Sq  F value Pr(>F)    
x             1 8582.4  8582.4 941.3165 <2e-16 ***
Residuals   424 3882.4     9.2                    
 Lack of fit   7   80.4    11.5   1.2597  0.269    
 Pure Error  417 3802.0     9.1                    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusions and Discussion

The given data has a total of 426 observations. There is enough data to show significant linear correlation between x and y. The two variables x and y have a fit with a regression equation of y = 3.459 + 18.2248x^(1/3)

Code for Part A

```
>wdir <-”/Users/angel/Documents/Project 1 Part A Data Files”
>setwd(wdir)
>PartA_IV<-read.csv(“784616_IV.csv”, header=TRUE)
>PartA_DV<-read.csv(“784616_DV.csv”, header=TRUE)
>PartA<-merge(PartA_IV, PartA_DV, by = ‘ID’)
>str(PartA)
>any(is.na(PartA[,2]) == TRUE)
>any(is.nan(PartA[,2]) == TRUE)
>any(is.na(PartA[,3]) == TRUE)
>any(is.nan(PartA[,3]) == TRUE)
>PartA_incomplete <- PartA
>library(mice)
>md.pattern(PartA_incomplete)
>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)
>M <- lm(DV ~ IV, data=PartA_complete)
>summary(M)
>library(knitr)
>kable(anova(M), caption='ANOVA Table')
>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')
>confint(M, level = 0.95)
>confint(M, level = 0.99)
```

Code for Part B

```
> wdir <- ‘/Users/angel/Documents/Project 1 Part B Data Files’
>setwd(wdir)
>data<-read.csv(‘784616_PartB.csv’, header=TRUE)
>M<- lm(y~x, data=data)
>summary(M)
>library(knitr)
>kable(anova(M), caption = ‘ANOVA Table’)
>data_trans <- data.frame(xtrans=data$x^(1/2), ytrans=data$y)
> Model <- lm (ytrans ~ xtrans, data = data_trans)
> summary(Model)
>data_trans <- data.frame(xtrans=data$x^(-1), ytrans=data$y)
> Model <- lm (ytrans ~ xtrans, data = data_trans)
> summary(Model)
>data_trans <- data.frame(xtrans=data$x^(2), ytrans=data$y)
> Model <- lm (ytrans ~ xtrans, data = data_trans)
> summary(Model)
>data_trans <- data.frame(xtrans=data$x^(3), ytrans=data$y)
> Model <- lm (ytrans ~ xtrans, data = data_trans)
> summary(Model)
>data_trans <- data.frame(xtrans=data$x, ytrans=data$y^(-1))
> Model <- lm (ytrans ~ xtrans, data = data_trans)
> summary(Model)
>data_trans <- data.frame(xtrans=data$x, ytrans=data$y^(-2))
> Model <- lm (ytrans ~ xtrans, data = data_trans)
> summary(Model)
>data_trans <- data.frame(xtrans=data$x, ytrans=data$y^(2))
> Model <- lm (ytrans ~ xtrans, data = data_trans)
> summary(Model)
>data_trans <- data.frame(xtrans=data$x, ytrans=data$y^(3))
> Model <- lm (ytrans ~ xtrans, data = data_trans)
> summary(Model)
>data_trans <- data.frame(xtrans=data$x, ytrans=data$y^(1/2))
> Model <- lm (ytrans ~ xtrans, data = data_trans)
> summary(Model)
>data_trans <- data.frame(xtrans=data$x, ytrans=data$y^(1/3))
> Model <- lm (ytrans ~ xtrans, data = data_trans)
> summary(Model)
>data_trans <- data.frame(xtrans=data$x^(1/3), ytrans=data$y)
```

```
> Model <- lm (ytrans ~ xtrans, data = data_trans)
> summary(Model)
>plot(data_trans$ytrans~data_trans$xtrans, main= ‘Scatter : y~x^(⅓)’, xlab =’x^(⅓)’, ylab=’y’,
pch=20)
>abline(Model, col='red', lty=3,Iwd=2)
>legend(‘topleft’, legend='Estimated Regression Line', tv=3. lwd=2. col='red')
>data_trans <-data. frame(xtrans=data$x^(⅓),ytrans=data$y)
>groups <-cut(data_trans$xtrans,breaks=c(-Inf,seq(min(data_trans$xtrans)+0.1,
max(data_trans$xtrans)-0.1,by=0.1),Inf))
>table(groups)
>x <- ave(data_trans$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)
```