AMS 315 First Computing Assignment

Formal Report for Part A

Name: Parv Joshi, ID: 112169570, net-id: pvjoshi Files Used: P1A_DV69570.csv, P1A_IV69570.csv

Introduction

This report describes the linear regression model of the dependent variable DV on the Independent Variable IV. I was given two data sets, one with IDs and IVs, other with IDs and DVs. The statistical programming language R was used in generating the results. The objective of this report is to perform analysis using software and find the fitted regression function, confidence and prediction intervals for the true slope/intercept, and the Anova Table.

Methodology

I was given two data sets, one with IDs and IVs, other with IDs and DVs. I used Excel to merge them and have a brief count of the data sets. Both had 648 IDs, with 599 IV observations, 495 DV observations, 456 IDs with both IV and DV observations, 143 IDs with IV observations but no DV observations, 39 IDs with DV observations but no IV observations, and 10 observations with no IV or DV observations. In total, there were 638 IDs with at least an IV or DV observation. I read both data files in R and merged them (to cross-check my excel work). I first removed the 10 IDs that had no observation. Hence 638 observations were used for the regression. I then used the MICE package to deal with missing data. I used the linear regression using bootstrap imputation (norm.boot) method for it. I then found the Anova table, plot, confidence interval, and the prediction interval.

Results

The model $y = \beta_0 + \beta_1 x$ is DV = (4.9679)IV + 17.8454. The R^2 value is 0.5802 (Variance explained is 58.02%) and the Adjusted R^2 is 0.5795. Both are greater than 0.5, which means this regression line is a fair estimate. The 95% confidence interval for the slope is (4.6389, 5.2970). The p-value of the test H_0: $\beta_1 = 0$ against H_1: $\beta_1 \neq 0$ is 2.2×10^{-16}. Therefore, we reject the null hypothesis. The F-statistic is 890.7798. The Analysis of Variance table (Anova Table) is shown below.

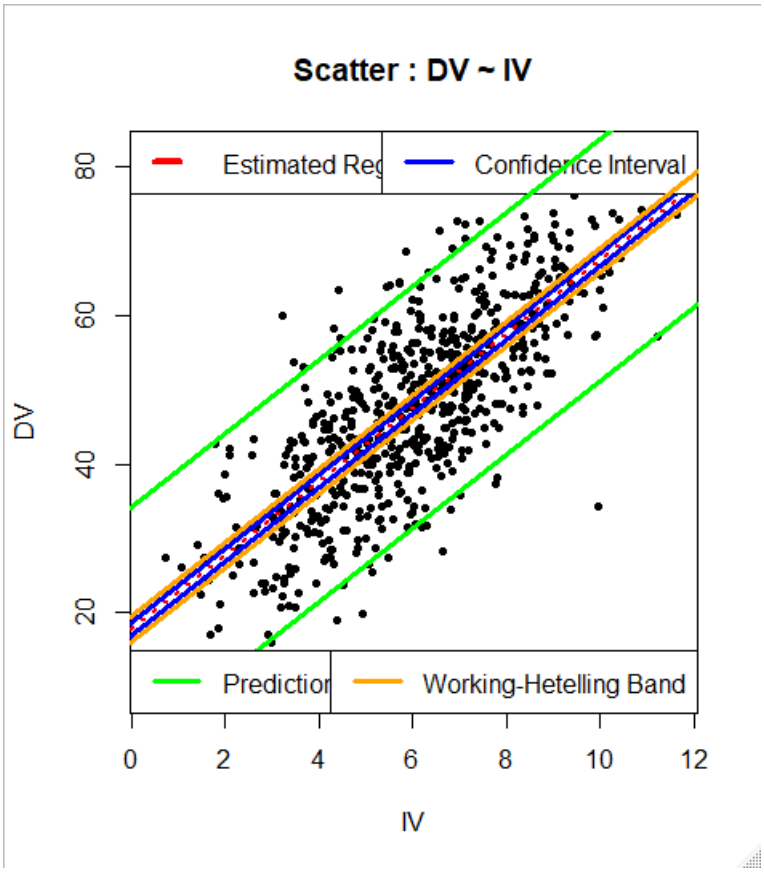
Conclusion

The p-value for part A is 2.2×10^{-16} which signifies that the linear association between the independent and dependent variable is highly significant with 58.02% variance explained. The correlation coefficient of IV and DV is 0.7617. The regression model DV = (4.9679)IV + 17.8454 accurately describes the relationship of IV and DV. See Anova Table and the Scatter Plot below.

Anova Table

Model	DF	Sum of Squares	Mean Square	F value	Pr(>F)
IV	1	60545.25	60545.25092	878.9009	0
Residuals	636	43812.43	68.88746	-	-

Scatter Plot



AMS 315 First Computing Assignment

Formal Report for Part B

Name: Parv Joshi, ID: 112169570, net-id: pvjoshi File Used: P1B69570.csv

Introduction

This report describes the best linear regression model of the dependent variable y on the Independent variable x using various transformations of x, y, or both. Given a data set with IDs, independent variable (x), and dependent variable (y), the best transformation was to be modelled between the two variables. The statistical programming language R was used in generating the results. The objective of this report is to perform an approximate lack of fit (LOF) test of cut and binned data using the best-found transformation that associates x and y.

Methodology

I was given a data set with 430 observations of x and y, with their IDs. I used all 430 observations for the regression as there was no missing data. I first plotted the x vs. y graph to look at the graph and try to have an intuitive guess on what function should I use for my transformation. The graph looked like it may have a curve of best fit like a log function. I then used the r-squared values of different transformations to compare them. The highest r-squared value I found from doing approximately 150 transformations (including box cox) was 0.6111 when I used lm(exp(y)~log(x)). Then I cut the data with a by-value of 0.01, which gave me exactly 130 groups. I then found the Pure Error Anova table after binning the data.

Results

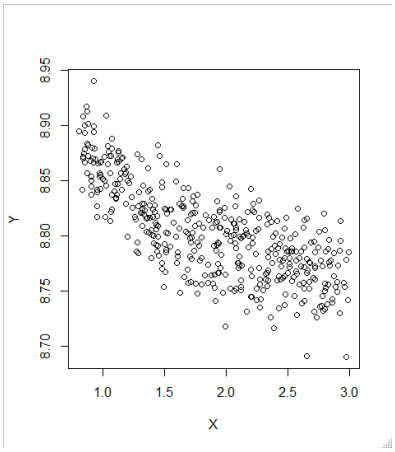
The simple linear regression is y = (5.1019) x +17.2235 with an R^2 of 0.5849. The best transformed model $Y = \beta_0 + \beta_1 X$ is exp(y) = (-616.10) log(x) + 7015.92. The R^2 value is 0.6115 (Variance explained is 61.15%) and the Adjusted R^2 is 0.6106. Both are greater than 0.5, which means this regression line is a fair estimate. The 95% confidence interval for the transformation slope is (-662.7628, -569.446). The p-value of the test H_0: *There is NO lack of fit*, against H_1: *There IS lack of fit* is 0.3954. This is greater than 0.01, 0.05, and 0.1. Therefore, we accept the null hypothesis proving the model is good. The F-statistic for lack of fit is 1.0376 confirming the model is good. The Analysis of Variance table (Anova Table) for pure error is shown below.

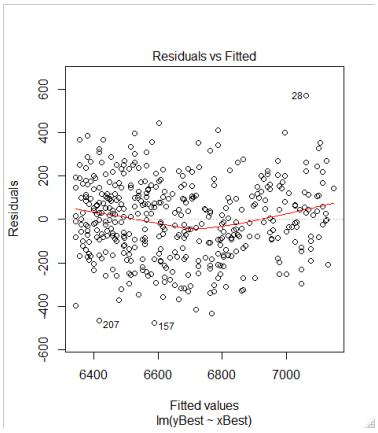
Conclusion

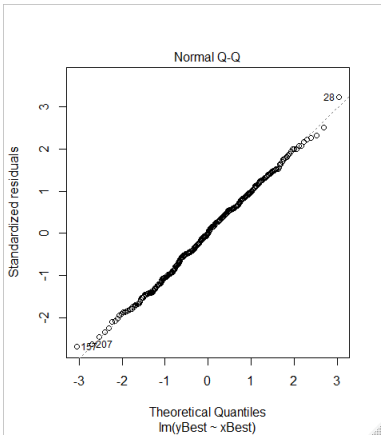
The p-value for the Lack of Fit (LOF) test in part B is 0.3954 which signifies that the linear association between the transformed independent and dependent variables is highly significant with 61.15% variance explained. The correlation coefficient for y and x is 0.7820. The *final regression model is* exp(y) = (-616.10) log(x) + 7015.92 accurately describes the relationship of y and x. See Pure Error Anova Table below. Also see attached plots of Leverage vs. Standardized residuals, fitted values vs. Square Root of Standardized residuals, Theoretical Quantities vs. Standardized residuals, fitted Values vs. residuals, x vs. y.

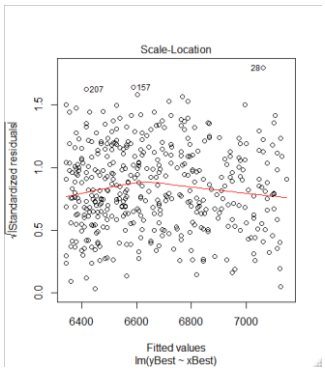
Pure Error Anova Table

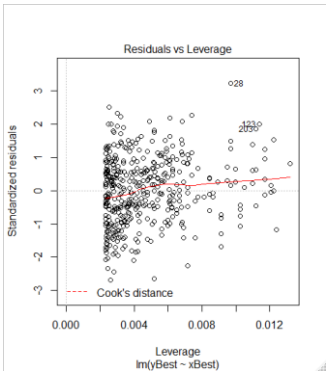
Model	DF	Sum of Squares	Mean Square	F value	Pr(>F)
IV	1	21365387	21365387	680.7564	2×10^{-16}
Residuals	428	13575290	31718	-	-
Lack of Fit	121	3940163	32563	1.0376	0.3954
Residuals	307	9635126	31385	-	-











Appendix

Code for Part A:

```
wdir -> "C:\Users\Parv\Documents\Spring 2020\AMS 315\Project 1"
setwd(wdir)
PartA_IV <- read.csv('P1A_IV69570.csv', header = TRUE)
PartA_DV <- read.csv('P1A_DV69570.csv', header = TRUE)
PartA <- merge(PartA_IV, PartA_DV, by = 'ID')
str(PartA)
View(PartA)
any(is.na(PartA[,2]) == TRUE)
any(is.nan(PartA[,2]) == TRUE)
PartA_incomplete <- PartA
install.packages('mice')
library(mice)
md.pattern(PartA_incomplete)
# There are 456 complete data sets
# IV is missing in 49, DV is missing in 153 and both are missing in 10 cases.
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)
install.packages('knitr')
library(knitr)
kable(anova(M), caption='ANOVA Table')
coef(M)
```

```
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=2, lwd=4, col='red')

confint(M, level=0.95)

# Confidence Interval

obs <- nrow(PartA_complete)

CI_L <- fitted(M) - qt(0.975, df=obs-2)*summary(M)$sigma*sqrt(1/obs + (PartA_complete$IV-mean(PartA_complete$IV))^2/(var(PartA_complete$IV)*(obs-1)))

CI_U <- fitted(M) + qt(0.975, df=obs-2)*summary(M)$sigma*sqrt(1/obs + (PartA_complete$IV-mean(PartA_complete$IV))^2/(var(PartA_complete$IV)*(obs-1)))

M1 <- lm(CI_L ~ IV, data=PartA_complete)

abline(M1, col='blue', 3, lwd=3)

M2 <- lm(CI_U ~ IV, data=PartA_complete)

abline(M2, col='blue', 3, lwd=3)

legend('topright', legend='Confidence Interval', lty=1, lwd=3, col='blue')

# Prediction Interval

PI_L <- fitted(M) - qt(0.975, df=obs-2)*summary(M)$sigma*sqrt(1+1/obs + (PartA_complete$IV-mean(PartA_complete$IV))^2/(var(PartA_complete$IV)*(obs-1)))

PI_U <- fitted(M) + qt(0.975, df=obs-2)*summary(M)$sigma*sqrt(1+1/obs + (PartA_complete$IV-mean(PartA_complete$IV))^2/(var(PartA_complete$IV)*(obs-1)))

M3 <- lm(PI_L ~ IV, data=PartA_complete)

abline(M3, col='green', 3, lwd=3)

M4 <- lm(PI_U ~ IV, data=PartA_complete)

abline(M4, col='green', 3, lwd=3)

legend('bottomleft', legend='Prediction Interval', lty=1, lwd=3, col='green')

# Working-Hetelling Band

WH_L <- fitted(M) - qf(0.975, 2, obs-2)*summary(M)$sigma*sqrt(1/obs + (PartA_complete$IV-mean(PartA_complete$IV))^2/(var(PartA_complete$IV)*(obs-1)))#

WH_U <- fitted(M) + qf(0.975, 2, obs-2)*summary(M)$sigma*sqrt(1/obs + (PartA_complete$IV-mean(PartA_complete$IV))^2/(var(PartA_complete$IV)*(obs-1)))

M5 <- lm(WH_L ~ IV, data=PartA_complete)
```

```
abline(M5, col='orange', 3, lwd=3)
M6 <- lm(WH_U ~ IV, data=PartA_complete)
abline(M6, col='orange', 3, lwd=3)
legend('bottomright', legend='Working-Hetelling Band', lty=1, lwd=3, col='orange')
```

Code for Part B:

```
wdir -> "C:\Users\Parv\Documents\Spring 2020\AMS 315\Project 1"
setwd(wdir)
Data <- read.csv('P1B69570.csv', header = TRUE)
nrow(Data)
X=Data$x
Y=Data$y
plot(Y~X)
install.packages('knitr')
library(knitr)
M <- lm(Y ~ X, data=Data)
kable(anova(M), caption='ANOVA Table')
summary(M)
summary(lm(Y~X))$r.squared
summary(lm(Y~X^2))$r.squared
summary(lm(Y~X^3))$r.squared
summary(lm(Y~abs(X)))$r.squared
summary(lm(Y~(1/X)))$r.squared
summary(lm(Y~sqrt(X)))$r.squared
summary(lm(Y~exp(X)))$r.squared
summary(lm(Y~log(X)))$r.squared
summary(lm(Y~log10(X)))$r.squared
summary(lm(Y~log2(X)))$r.squared
summary(lm(Y~cos(X)))$r.squared
```

```
summary(lm(Y~sin(X)))$r.squared
summary(lm(Y~tan(X)))$r.squared
summary(lm(Y~poly(X)))$r.squared
summary(lm(Y~X))$r.squared
summary(lm(Y^2~X))$r.squared
summary(lm(Y^3~X))$r.squared
summary(lm(abs(Y)~X))$r.squared
summary(lm((1/Y)~X))$r.squared
summary(lm(sqrt(Y)~X))$r.squared
summary(lm(exp(Y)~X))$r.squared
summary(lm(log(Y)~X))$r.squared
summary(lm(log10(Y)~X))$r.squared
summary(lm(log2(Y)~X))$r.squared
summary(lm(cos(Y)~X))$r.squared
summary(lm(sin(Y)~X))$r.squared
summary(lm(tan(Y)~X))$r.squared
summary(lm(poly(Y)~X))$r.squared
summary(lm(Y^2~X))$r.squared
summary(lm(Y^2~X))$r.squared
summary(lm(Y^2~X^2))$r.squared
summary(lm(Y^2~X^3))$r.squared
summary(lm(Y^2~abs(X)))$r.squared
summary(lm(Y^2~(1/X)))$r.squared
summary(lm(Y^2~sqrt(X)))$r.squared
summary(lm(Y^2~exp(X)))$r.squared
summary(lm(Y^2~log(X)))$r.squared
summary(lm(Y^2~log10(X)))$r.squared
summary(lm(Y^2~log2(X)))$r.squared
summary(lm(Y^2~cos(X)))$r.squared
```

```
summary(lm(Y^2~sin(X)))$r.squared
summary(lm(Y^2~tan(X)))$r.squared
summary(lm(Y^2~poly(X)))$r.squared
summary(lm(Y^3~X))$r.squared
summary(lm(Y^3~X))$r.squared
summary(lm(Y^3~X^2))$r.squared
summary(lm(Y^3~X^3))$r.squared
summary(lm(Y^3~abs(X)))$r.squared
summary(lm(Y^3~(1/X)))$r.squared
summary(lm(Y^3~sqrt(X)))$r.squared
summary(lm(Y^3~exp(X)))$r.squared
summary(lm(Y^3~log(X)))$r.squared
summary(lm(Y^3~log10(X)))$r.squared
summary(lm(Y^3~log2(X)))$r.squared
summary(lm(Y^3~cos(X)))$r.squared
summary(lm(Y^3~sin(X)))$r.squared
summary(lm(Y^3~tan(X)))$r.squared
summary(lm(Y^3~poly(X)))$r.squared
summary(lm(abs(Y)~X))$r.squared
summary(lm(abs(Y)~X^2))$r.squared
summary(lm(abs(Y)~X^3))$r.squared
summary(lm(abs(Y)~abs(X)))$r.squared
summary(lm(abs(Y)~(1/X)))$r.squared
summary(lm(abs(Y)~sqrt(X)))$r.squared
summary(lm(abs(Y)~exp(X)))$r.squared
summary(lm(abs(Y)~log(X)))$r.squared
summary(lm(abs(Y)~log10(X)))$r.squared
summary(lm(abs(Y)~log2(X)))$r.squared
summary(lm(abs(Y)~cos(X)))$r.squared
```

```
summary(lm(abs(Y)~sin(X)))$r.squared
summary(lm(abs(Y)~tan(X)))$r.squared
summary(lm(abs(Y)~poly(X)))$r.squared
summary(lm((1/Y)~X))$r.squared
summary(lm((1/Y)~X^2))$r.squared
summary(lm((1/Y)~X^3))$r.squared
summary(lm((1/Y)~abs(X)))$r.squared
summary(lm((1/Y)~(1/X)))$r.squared
summary(lm((1/Y)~sqrt(X)))$r.squared
summary(lm((1/Y)~exp(X)))$r.squared
summary(lm((1/Y)~log(X)))$r.squared
summary(lm((1/Y)~log10(X)))$r.squared
summary(lm((1/Y)~log2(X)))$r.squared
summary(lm((1/Y)~cos(X)))$r.squared
summary(lm((1/Y)~sin(X)))$r.squared
summary(lm((1/Y)~tan(X)))$r.squared
summary(lm((1/Y)~poly(X)))$r.squared
summary(lm(sqrt(Y)~X))$r.squared
summary(lm(sqrt(Y)~X^2))$r.squared
summary(lm(sqrt(Y)~X^3))$r.squared
summary(lm(sqrt(Y)~abs(X)))$r.squared
summary(lm(sqrt(Y)~(1/X)))$r.squared
summary(lm(sqrt(Y)~sqrt(X)))$r.squared
summary(lm(sqrt(Y)~exp(X)))$r.squared
summary(lm(sqrt(Y)~log(X)))$r.squared
summary(lm(sqrt(Y)~log10(X)))$r.squared
summary(lm(sqrt(Y)~log2(X)))$r.squared
summary(lm(sqrt(Y)~cos(X)))$r.squared
summary(lm(sqrt(Y)~sin(X)))$r.squared
```

```
summary(lm(sqrt(Y)~tan(X)))$r.squared
summary(lm(sqrt(Y)~poly(X)))$r.squared
summary(lm(exp(Y)~X))$r.squared
summary(lm(exp(Y)~X^2))$r.squared
summary(lm(exp(Y)~X^3))$r.squared
summary(lm(exp(Y)~abs(X)))$r.squared
summary(lm(exp(Y)~(1/X)))$r.squared
summary(lm(exp(Y)~sqrt(X)))$r.squared
summary(lm(exp(Y)~exp(X)))$r.squared
summary(lm(exp(Y)~log(X)))$r.squared
summary(lm(exp(Y)~log10(X)))$r.squared
summary(lm(exp(Y)~log2(X)))$r.squared
summary(lm(exp(Y)~cos(X)))$r.squared
summary(lm(exp(Y)~sin(X)))$r.squared
summary(lm(exp(Y)~tan(X)))$r.squared
summary(lm(exp(Y)~poly(X)))$r.squared
summary(lm(log(Y)~X))$r.squared
summary(lm(log(Y)~X^2))$r.squared
summary(lm(log(Y)~X^3))$r.squared
summary(lm(log(Y)~abs(X)))$r.squared
summary(lm(log(Y)~(1/X)))$r.squared
summary(lm(log(Y)~sqrt(X)))$r.squared
summary(lm(log(Y)~exp(X)))$r.squared
summary(lm(log(Y)~log(X)))$r.squared
summary(lm(log(Y)~log10(X)))$r.squared
summary(lm(log(Y)~log2(X)))$r.squared
summary(lm(log(Y)~cos(X)))$r.squared
summary(lm(log(Y)~sin(X)))$r.squared
summary(lm(log(Y)~tan(X)))$r.squared
```

```
summary(lm(log(Y)~poly(X)))$r.squared
summary(lm(log10(Y)~X))$r.squared
summary(lm(log10(Y)~X^2))$r.squared
summary(lm(log10(Y)~X^3))$r.squared
summary(lm(log10(Y)~abs(X)))$r.squared
summary(lm(log10(Y)~(1/X)))$r.squared
summary(lm(log10(Y)~sqrt(X)))$r.squared
summary(lm(log10(Y)~exp(X)))$r.squared
summary(lm(log10(Y)~log(X)))$r.squared
summary(lm(log10(Y)~log10(X)))$r.squared
summary(lm(log10(Y)~log2(X)))$r.squared
summary(lm(log10(Y)~cos(X)))$r.squared
summary(lm(log10(Y)~sin(X)))$r.squared
summary(lm(log10(Y)~tan(X)))$r.squared
summary(lm(log10(Y)~poly(X)))$r.squared
summary(lm(log2(Y)~X))$r.squared
summary(lm(log2(Y)~X^2))$r.squared
summary(lm(log2(Y)~X^3))$r.squared
summary(lm(log2(Y)~abs(X)))$r.squared
summary(lm(log2(Y)~(1/X)))$r.squared
summary(lm(log2(Y)~sqrt(X)))$r.squared
summary(lm(log2(Y)~exp(X)))$r.squared
summary(lm(log2(Y)~log(X)))$r.squared
summary(lm(log2(Y)~log10(X)))$r.squared
summary(lm(log2(Y)~log2(X)))$r.squared
summary(lm(log2(Y)~cos(X)))$r.squared
summary(lm(log2(Y)~sin(X)))$r.squared
summary(lm(log2(Y)~tan(X)))$r.squared
summary(lm(log2(Y)~poly(X)))$r.squared
```

```
summary(lm(cos(Y)~X))$r.squared
summary(lm(cos(Y)~X^2))$r.squared
summary(lm(cos(Y)~X^3))$r.squared
summary(lm(cos(Y)~abs(X)))$r.squared
summary(lm(cos(Y)~(1/X)))$r.squared
summary(lm(cos(Y)~sqrt(X)))$r.squared
summary(lm(cos(Y)~exp(X)))$r.squared
summary(lm(cos(Y)~log(X)))$r.squared
summary(lm(cos(Y)~log10(X)))$r.squared
summary(lm(cos(Y)~log2(X)))$r.squared
summary(lm(cos(Y)~cos(X)))$r.squared
summary(lm(cos(Y)~sin(X)))$r.squared
summary(lm(cos(Y)~tan(X)))$r.squared
summary(lm(cos(Y)~poly(X)))$r.squared
summary(lm(sin(Y)~X))$r.squared
summary(lm(sin(Y)~X^2))$r.squared
summary(lm(sin(Y)~X^3))$r.squared
summary(lm(sin(Y)~abs(X)))$r.squared
summary(lm(sin(Y)~(1/X)))$r.squared
summary(lm(sin(Y)~sqrt(X)))$r.squared
summary(lm(sin(Y)~exp(X)))$r.squared
summary(lm(sin(Y)~log(X)))$r.squared
summary(lm(sin(Y)~log10(X)))$r.squared
summary(lm(sin(Y)~log2(X)))$r.squared
summary(lm(sin(Y)~cos(X)))$r.squared
summary(lm(sin(Y)~sin(X)))$r.squared
summary(lm(sin(Y)~tan(X)))$r.squared
summary(lm(sin(Y)~poly(X)))$r.squared
summary(lm(tan(Y)~X))$r.squared
```

```
summary(lm(tan(Y)~X^2))$r.squared
summary(lm(tan(Y)~X^3))$r.squared
summary(lm(tan(Y)~abs(X)))$r.squared
summary(lm(tan(Y)~(1/X)))$r.squared
summary(lm(tan(Y)~sqrt(X)))$r.squared
summary(lm(tan(Y)~exp(X)))$r.squared
summary(lm(tan(Y)~log(X)))$r.squared
summary(lm(tan(Y)~log10(X)))$r.squared
summary(lm(tan(Y)~log2(X)))$r.squared
summary(lm(tan(Y)~cos(X)))$r.squared
summary(lm(tan(Y)~sin(X)))$r.squared
summary(lm(tan(Y)~tan(X)))$r.squared
summary(lm(tan(Y)~poly(X)))$r.squared
summary(lm(exp(Y)~log(X)))
#Highest R^2 Values:
#> summary(lm(exp(Y)~log(X)))$r.squared
#[1] 0.6110614
#> summary(lm(exp(Y)~log10(X)))$r.squared
#[1] 0.6110614
#> summary(lm(exp(Y)~log2(X)))$r.squared
#[1] 0.6110614
library(MASS)
Model=lm(exp(y)~log(x), data = Data)
bc=boxcox(Model, lambda = seq(-3,3))
best.lam = bc$x[which(bc$y==max(bc$y))]
best.lam
Model.inv = lm(y^-1 ~ x, data=Data)
plot(Model.inv)
#Boxcox R^2:
```

```
summary(lm((Y)^(-0.3030)~X))$r.squared

data_trans <- data.frame(xtrans=log(X), ytrans=exp(Y))

install.packages('alr3')

library(alr3)

fit_a <- lm(exp(Y)~log(X), data = data_trans)

pureErrorAnova(fit_a)

groupsx <- cut(data_trans$xtrans,breaks=c(-Inf,seq(min(data_trans$xtrans)+0.01,
max(data_trans$xtrans)-0.01,by=0.01),Inf))

table(groups)

groupsy <- cut(data_trans$ytrans,breaks=c(-Inf,seq(min(data_trans$ytrans)+0.01,
max(data_trans$ytrans)-0.01,by=0.01),Inf))

table(groups)

xBest <- ave(data_trans$xtrans, groupsx)

yBest <- ave(data_trans$ytrans, groupsy)

data_bin <- data.frame(x=x, y=data_trans$ytrans)

fit_b <- lm(yBest ~ xBest, data = data_bin)

pureErrorAnova(fit_b)

summary(fit_b)

coef(fit_b)

confint(fit_b, level=0.95)

plot(fit_b)
```

Citation: Help was taken for coding in R from the handout provided by professor. The link to that handout is:

https://blackboard.stonybrook.edu/bbcswebdav/pid-5297218-dt-content-rid-41158583_1/courses/1204-AMS-315-SEC01-49021/One%20Predictor%20Linear%20Regression%20Handout%20Spring%202020.html