MD Mahmudur Rahman

Data Analysis Project 1 Part A

Introduction:

To start off, we were given two data sets. One was labeled with ID and Independent Variable while the other was labeled with ID and Dependent Variable. We would need to first sort and merge the two datasets together. After that we would need to address as a statistician, the issues that our future supervisor would want to know about. We would need to use a statistical package to figure out the missing values. In this case I will be using R studio. We will be calculating the summary statistics on each variable, examine the scatterplot, calculate the bivariate statistics and calculate the Chapter 11 regression statistics. Our problem is to recover the function that was used to generate the dependent variable value based on the value of the independent variable.

Methods:

I first used R studio to sort the two files by subject ID and merge them. I used the code merged.data <- merge(AMS315_P1A_IV_949881,AMS315_P1A_DV_949881, by="ID"). It then gave me 464 observations of 3 variables. Afterwards, I installed the package known as mice. After that, I used the code md.pattern (merged.data_incomplete) and found out there are 419 complete data sets, 24 IV are missing, 22 DV are missing, and both are missing in 1 case. Then, we dropped the 1 case with missing IV and DV using the code merged_data_imp <- merged.data[!is.na(merged.data$IV)==TRUE |!is.na(merged.data$DV)==TRUE,]. Then, I used norm.boot to impute the missing values. I then used the summary function in R to get all of the summary statistics. After that, the knitr package was installed. This code line kable(anova(M), caption='ANOVA Table'), was used to create the ANOVA Table. The scatter plot was created using the code plot(merged.data_complete$DV ~ merged.data_complete$IV, main='Scatter : DV ~ IV', xlab='IV', ylab='DV', pch=20). After that, the estimated regression line was created using the code abline(M, col='red', lty=3, lwd=2). The confint function in R was used to calculate the 95% and the 99% confidence interval.

Results:

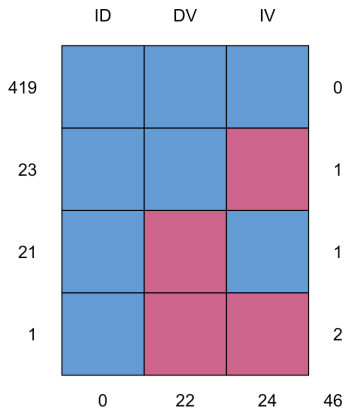
For the model $Y= B+B_1 X$ the fitted function was DV= 48.7215+3.0464IV. The 95% confidence interval for the intercept was [43.730212,53.712832] and for the slope was [2.716824, 3.375939]. The 99% confidence interval for the intercept was [42.151856, 55.291188] and for the slope was [2.612611, 3.480152]. The analysis of the variance table is shown below labeled ANOVA Table.

Conclusion:

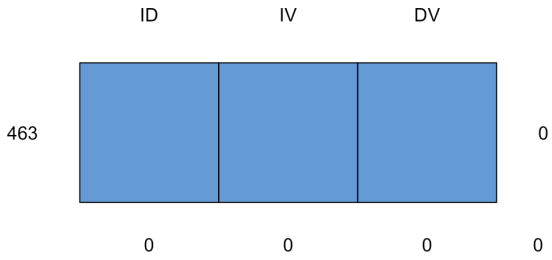
The p-value for our data was 2.2*10^-16 which tells us that there is strong evidence that we should reject the null hypothesis that the slope is 0. We concluded that the Residual standard error is 10.54, the multiple R-squared value 0.4172, and the adjusted R-squared value is 0.4159. The fitted function was DV= 48.7215+3.0464IV, which shows a strong relationship between the two variables.

```
library(readr)
AMS315_P1A_IV_949881 <- read_csv("~/Downloads/P1A_IV/AMS315_P1A_IV_949881.csv")
View(AMS315_P1A_IV_949881)
library(readr)
AMS315_P1A_DV_949881 <- read_csv("~/Downloads/P1A_DV/AMS315_P1A_DV_949881.csv")
View(AMS315_P1A_DV_949881)
merged.data <- merge(AMS315_P1A_IV_949881,AMS315_P1A_DV_949881, by="ID")
View(merged.data)
View(merged.data)
str(AMS315_P1A_IV_949881)
View(AMS315_P1A_IV_949881)
str(AMS315_P1A_DV_949881)
view(AMS315_P1A_DV_949881)
View(AMS315_P1A_DV_949881)
any(is.na(AMS315_P1A_IV_949881[,2]) == TRUE)
str(merged.data)
View(merged.data)
any(is.na(merged.data[,2]) == TRUE)
any(is.na(merged.data[,3]) == TRUE)
install.packages('mice')
merged.data_incomplete <- merged.data
md.pattern(merged.data_incomplete)
library(mice)
md.pattern(merged.data_incomplete)
merged.data_imp <- merged.data[!is.na(merged.data$IV)==TRUE|!is.na(merged.data$DV)==True,]
merged_data_imp <- merged.data[!is.na(merged.data$IV)==TRUE |!is.na(merged.data$DV)==TRUE,]
imp <- mice(merged_data_imp, method = "norm.boot", printFlag = FALSE)
merged.data_complete <- complete(imp)
md.pattern(merged.data_complete)
```

Number of data sets complete, Missing IV, Missing DV, Missing Both IV and DV



After imputing missing values and removing the data set where both IV and DV was missing



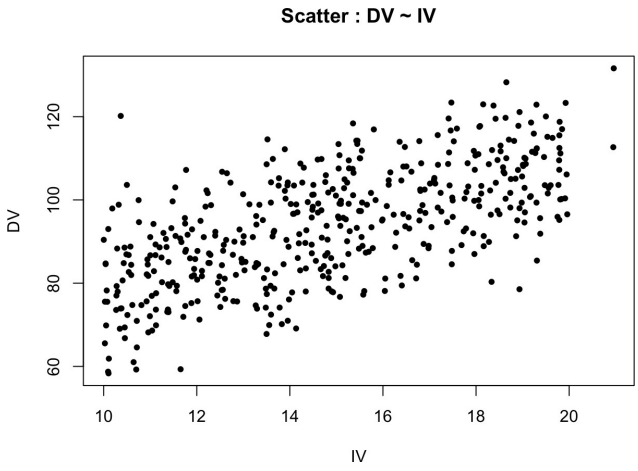
Steps for finding linear regression of the independent variable and dependent variable

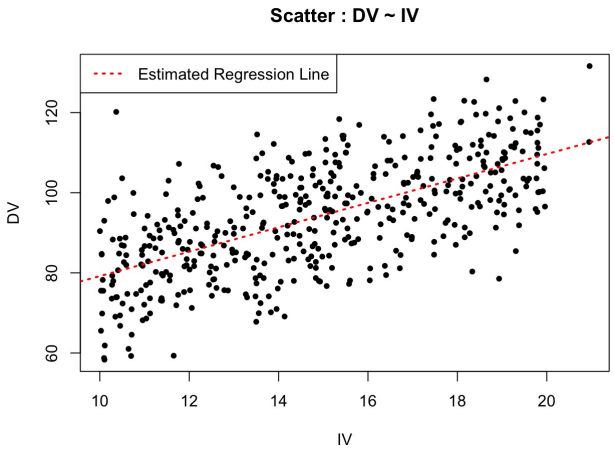
```
M <- lm(DV ~ IV, data=merged.data_complete)
summary(M)
install.packages('knitr')
library(knitr)
kable(anova(M), caption='ANOVA Table')
plot(merged.data_complete$DV ~ merged.data_complete$IV, main='Scatter : DV ~ IV', xlab='IV', ylab='DV', pch=20)
abline(M, col='red', lty=3, lwd=2)
legend('topleft', legend='Estimated Regression Line', lty=3, lwd=2, col='red')
confint(M, level = 0.95)
confint(M, level = 0.99)
View(merged.data_complete)
load("~/AMS 315 Project/.RData")
```

```
> M <- lm(DV ~ IV, data=merged.data_complete)
> summary(M)

Call:
lm(formula = DV ~ IV, data = merged.data_complete)

Residuals:
    Min      1Q  Median      3Q     Max
-27.854  -7.670   0.236   7.322  39.874

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  48.7215     2.5399   19.18   <2e-16 ***
IV            3.0464     0.1677   18.16   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.54 on 461 degrees of freedom
Multiple R-squared:  0.4172,    Adjusted R-squared:  0.4159
F-statistic:   330 on 1 and 461 DF,  p-value: < 2.2e-16
```

Table: ANOVA Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
IV	1	36646.60	36646.5995	329.9787	0
Residuals	461	51197.49	111.0575	NA	NA






```
> confint(M, level = 0.95)
                2.5 %    97.5 %
(Intercept) 43.730212 53.712832
IV           2.716824  3.375939
> confint(M, level = 0.99)
                0.5 %    99.5 %
(Intercept) 42.151856 55.291188
IV           2.612611  3.480152
> 
```

MD Mahmudur Rahman

Data Analysis Project 1 Part B

Introduction:

For this part of the project, we are already given the merged data that included the Subject ID, Independent Variable and Dependent Variable. Our goal is to first figure out the transformation of the data. Then we have to do an analysis of the data and have very few exact repeats of an independent variable value. This means, we have to bin the data and apply the lack of fit test. We have to bin near repeated data into one level.

Methods:

The first step that I did was to find the linear regression of the independent variable and dependent variable without transformation and saw it was not a good model. I used the code reg1 <- lm(DV ~ IV, data=AMS315_P1B_949881), summary(reg1). I did the scatter plot and drew in the estimated regression line. Afterwards, I had to figure out the best transformation that would make the data closest to normal distribution. I used automatic transformations to do that, instead of trial and error. I implemented Tukey's Ladder of Powers. I used the code transform2 <- data.frame(xtrans=transformTukey(AMS315_P1B_949881$IV),ytrans=AMS315_P1B_949881$ DV), to find at which lambda value the IV data is closest to normality. My transformation was -1*IV^-0.825. Then, I generated my groups using cut(x, breaks) and the table was shown. Next, I binned the data using the code x <- ave(transform2$xtrans, groups5), data_bin3 <- data.frame(x=x, y=transform2$ytrans). I performed the lack of fit test by downloading the alr 3 package. Then, using the code, fit_b4 <- lm(y ~ x, data = data_bin3)> pureErrorAnova(fit_b4), I was able to create the ANOVA table to verify the fit of this data. Lastly, I did the linear regression of the transformed independent variable and dependent variable. I did the scatter plot and drew in the estimated regression line and found 95% CI and 99%CI.

Results

For the model $Y= B+B_1 X$ the fitted function was DV= 53.9501-24.3751IV. The 95% confidence interval for the intercept was [53.21213,54.68812] and for the slope was [-26.05549, -22.69465]. The 99% confidence interval for the intercept was [52.97927, 54.92097] and for the slope was [-26.58569, -22.16445]. The analysis of the variance table is shown below labeled ANOVA Table. After transforming IV using Tukey's Ladder of Powers, the model for the transformation was DVtrans= 28.8123-6.3187IVtrans using -1*IV^-0.825. The 95% confidence interval for the intercept was [27.89365,29.731045] and for the slope was [-6.67757, -5.959877]. The 99% confidence interval for the intercept was [27.60379, 30.020909] and for the slope was [-6.790792, -5.846655]. The ANOVA table is shown below labeled ANOVATableTransformation

Conclusion

There is a significant linear correlation between the transformation of IV and DV, which is (-1*IV^-0.825, DV), data sets based on Tukey's Ladder of Powers. We can confirm visually that this transformation is a good fit since our p value is 0.7901 and F value is 0.5582 or 55.82% for the Lack of Fit. We concluded that the Residual standard error is 3.768, the multiple R-squared value is 0.627, and the adjusted R-squared value is 0.6265. The linear regression between them was DVtrans= 28.8123 - 6.3187 IVtrans.

MD Mahmudur Rahman

Steps for finding linear regression of the independent variable and dependent variable without transformation

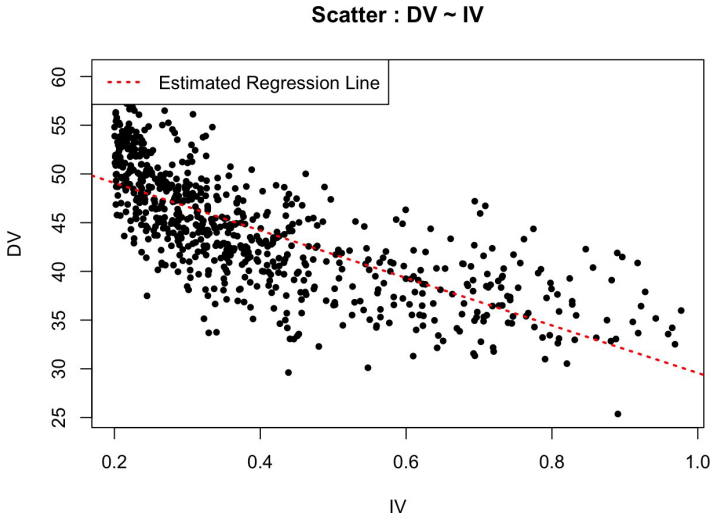
```
reg1 <- lm(DV ~ IV, data=AMS315_P1B_949881)
summary(reg1)
install.packages('knitr')
library(knitr)
kable(anova(reg1), caption= 'ANOVA Table')
plot(AMS315_P1B_949881$DV ~ AMS315_P1B_949881$IV, main='Scatter : DV ~ IV', xlab='IV', ylab='DV', pch=20)
abline(reg1, col='red', lty=3, lwd=2)
legend('topleft', legend='Estimated Regression Line', lty=3, lwd=2, col='red')
confint(reg1, level = 0.95)
confint(reg1, level = 0.99)
```

```
> reg1 <- lm(DV ~ IV, data=AMS315_P1B_949881)
> summary(reg1)

Call:
lm(formula = DV ~ IV, data = AMS315_P1B_949881)

Residuals:
     Min       1Q   Median       3Q      Max
-13.6417  -2.9061   0.0155   2.6770  11.2569

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  53.9501     0.3759  143.52   <2e-16 ***
IV          -24.3751     0.8559  -28.48   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.217 on 711 degrees of freedom
Multiple R-squared:  0.5329,    Adjusted R-squared:  0.5322
F-statistic:   811 on 1 and 711 DF,  p-value: < 2.2e-16
```

Table: ANOVA Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
IV	1	14423.37	14423.37348	811.0213	0
Residuals	711	12644.57	17.78421	NA	NA

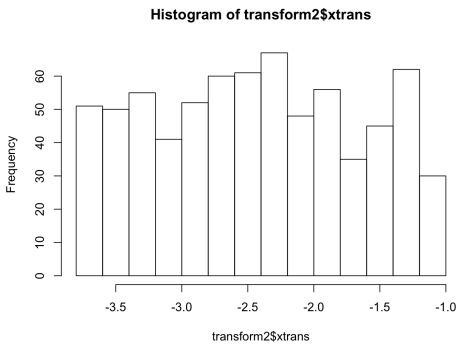
```
> library(knitr)
> kable(anova(reg1), caption= 'ANOVA Table')


Table: ANOVA Table

|          |  Df|   Sum Sq|     Mean Sq|  F value| Pr(>F)|
|:---------|---:|--------:|-----------:|--------:|------:|
|IV        |   1| 14423.37| 14423.37348| 811.0213|      0|
|Residuals | 711| 12644.57|    17.78421|       NA|     NA|
> plot(AMS315_P1B_949881$DV ~ AMS315_P1B_949881$IV, main='Scatter : DV ~ IV', xlab='IV', ylab='DV', pch=2
0)
> abline(reg1, col='red', lty=3, lwd=2)
> legend('topleft', legend='Estimated Regression Line', lty=3, lwd=2, col='red')
> confint(reg1, level = 0.95)
                2.5 %    97.5 %
(Intercept)  53.21213  54.68812
IV          -26.05549 -22.69465
> confint(reg1, level = 0.99)
                0.5 %    99.5 %
(Intercept)  52.97927  54.92097
IV          -26.58569 -22.16445
>
```

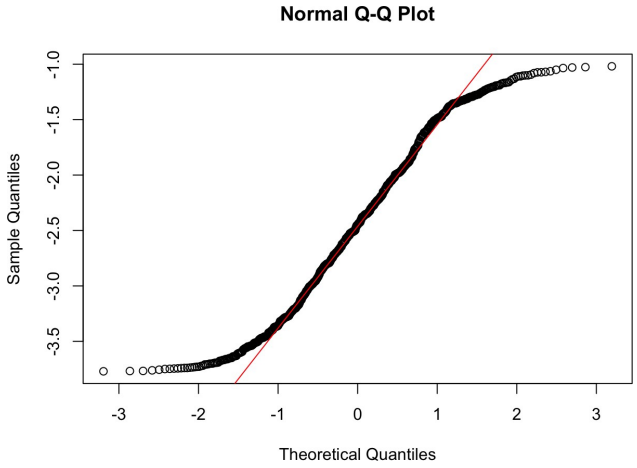


Steps for finding the Transformation of IV

```
install.packages("rcompanion")
library(rcompanion)
transform2 <- data.frame(xtrans=transformTukey(AMS315_P1B_949881$IV), ytrans=AMS315_P1B_949881$DV)
View(transform2)
hist(transform2$xtrans)
```

```
> library(rcompanion)
> transform2 <- data.frame(xtrans=transformTukey(AMS315_P1B_949881$IV), ytrans=AMS315_P1B_949881$DV)

    lambda     W Shapiro.p.value
368 -0.825 0.961       7.727e-13

if (lambda >  0){TRANS = x ^ lambda}
if (lambda == 0){TRANS = log(x)}
if (lambda <  0){TRANS = -1 * x ^ lambda}
```






```
groups5 <- cut(transform2$xtrans,breaks = c(-Inf,seq(min(transform2$xtrans)+0.3, max(transform2$xtrans)-0.3,by=0.3),Inf))
table(groups5)
x <- ave(transform2$xtrans, groups5)
data_bin3 <- data.frame(x=x, y=transform2$ytrans)
library(carData)
library(car)
library(alr3)
fit_b4 <- lm(y ~ x, data = data_bin3)
pureErrorAnova(fit_b4)
```

```
> groups5 <- cut(transform2$xtrans,breaks = c(-Inf,seq(min(transform2$xtrans)+0.3, max(transform2$xtrans)
-0.3,by=0.3),Inf))
> table(groups5)
groups5
 (-Inf,-3.47] (-3.47,-3.17] (-3.17,-2.87] (-2.87,-2.57] (-2.57,-2.27] (-2.27,-1.97] (-1.97,-1.67]
           81            83            64            87           100            87            57
(-1.67,-1.37]  (-1.37, Inf]
           68            86
```

```
> x <- ave(transform2$xtrans, groups5)
> data_bin3 <- data.frame(x=x, y=transform2$ytrans)
> library(carData)
> library(car)
> library(alr3)
> fit_b4 <- lm(y ~ x, data = data_bin3)
> pureErrorAnova(fit_b4)
Analysis of Variance Table

Response: y
             Df  Sum Sq Mean Sq   F value Pr(>F)
x             1 16771.9 16771.9 1153.1541 <2e-16 ***
Residuals   711 10296.1    14.5
 Lack of fit   7    56.8     8.1    0.5582 0.7901
 Pure Error  704 10239.2    14.5
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Steps for finding linear regression of the transformed independent variable and dependent variable

```
reg2transform <- lm(ytrans ~ xtrans, data=transform2)
summary(reg2transform)
plot(transform2$ytrans ~ transform2$xtrans, main='Scatter : ytrans ~ xtrans', xlab='IV', ylab='DV', pch=20)
abline(reg2transform, col='red', lty=3, lwd=2)
legend('topleft', legend='Estimated Regression Line', lty=3, lwd=2, col='red')
confint(reg2transform, level = 0.95)
confint(reg2transform, level = 0.99)
```

```
> reg2transform <- lm(ytrans ~ xtrans, data=transform2)
> summary(reg2transform)

Call:
lm(formula = ytrans ~ xtrans, data = transform2)

Residuals:
     Min       1Q   Median       3Q      Max
-11.6665  -2.4197  -0.0899   2.3231  10.6076

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  28.8123     0.4679   61.57   <2e-16 ***
xtrans       -6.3187     0.1828  -34.57   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.768 on 711 degrees of freedom
Multiple R-squared:  0.627,     Adjusted R-squared:  0.6265
F-statistic:  1195 on 1 and 711 DF,  p-value: < 2.2e-16

>
```

```
> library(knitr)
> kable(anova(reg2transform), caption = 'ANOVA Table for Transformation')


Table: ANOVA Table for Transformation

|          |  Df|   Sum Sq|     Mean Sq|  F value| Pr(>F)|
|:---------|---:|--------:|-----------:|--------:|------:|
|xtrans    |   1| 16971.45| 16971.44726| 1195.137|      0|
|Residuals | 711| 10096.50|    14.20042|       NA|     NA|
>
```

```
> plot(transform2$ytrans ~ transform2$xtrans, main='Scatter : ytrans ~ xtrans', xlab='IV', ylab='DV', pch
=20)
> abline(reg2transform, col='red', lty=3, lwd=2)
> legend('topleft', legend='Estimated Regression Line', lty=3, lwd=2, col='red')
> confint(reg2transform, level = 0.95)
               2.5 %    97.5 %
(Intercept) 27.89365 29.731045
xtrans      -6.67757 -5.959877
> confint(reg2transform, level = 0.99)
                0.5 %    99.5 %
(Intercept) 27.603790 30.020909
xtrans      -6.790792 -5.846655
>
```