AMS 315 - Project 01

TUAN BUI

SBU ID: 113141951

11/04/2021

Report for Part A

Introduction:

The objective is to find the model describing the dataset in Problem A. In this report, I am using the R statistical package to conduct the necessary functions to draw my conclusions. I will generate the ANOVA table to check the precision of the linear model, and use a least-squares fit to generate a linear model for the data.

Methods:

In problem A, the original datasets were supplied with two data sheets in .csv file. One dataset had the ID of an observation and its associated independent variable value, and the other had the ID and associated dependent variable value. Both datasets had total of 540 observations with ID# ranging from 1 to 540. First, I merged the two files into one new dataset. There were 540 observations and three columns in the new datasets, called 'ID', 'IV', and 'DV'. Each is numerical. After inspecting the dataset, I found that there were 451 complete data sets, IV was missing in 47 cases, DV was missing in 58 cases, and both IV and DV were missing in 16 cases. Since the fraction of missing both IV and DV data was about 3%, my choice was using the imputation method to impute the missing data. After the imputation, 16 observations with both IV and DV missing values were removed. The final dataset had 524 complete data sets. Next, I fitted a regression model to the final dataset. Based on the regression model, I generated the ANOVA table for the final dataset, and calculated the 95% and 99% confidence interval of slope and intercept.

Results:

The fitted function for the model Y = B + B1*X was DV = 47.9576 + 4.6419*IV. The coefficient of determination (R-square) is 0.4754. The 95% confidence interval for the slope was [4.22265, 5.061174]. The 95% confidence interval for the intercept was [45.70434, 50.210775]. The 99% confidence interval for the slope was [4.090169, 5.193655]. The 99% confidence interval for the intercept was [44.992349, 50.922762]. The analysis of variance table is shown below and the association between the independent variable and dependent variable was highly significant ($p < 2.2e-16$).

ANOVA Table (Dependent Variable regressed on Independent Variable)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Regression	1	42702.51	42702.51055	473.0815	0
Residuals	522	47118.12	90.26459	NA	NA
Total	523	89820.63			

Conclusions and Discussions:

For part A, the association between the independent and dependent variables is evidently significant ($p < 2e-16$), with 99.9% of the dependent variable variation explained. The linear model, DV = 47.9576 + 4.6419*IV, accurately describes the relationship between the two variables. The plot of residual versus predicted value confirmed the validity of this model. Suppose the null hypothesis that the slope of the linear regression is 0, and the alternative hypothesis that the slope of the linear regression is not 0, then the null hypothesis is rejected because the analysis of variance reveals the F statistic of 473.0815, which is greater than 10.9512, the critical value of F at $\alpha = 0.001$, df1 = 1, and df2 = 523.

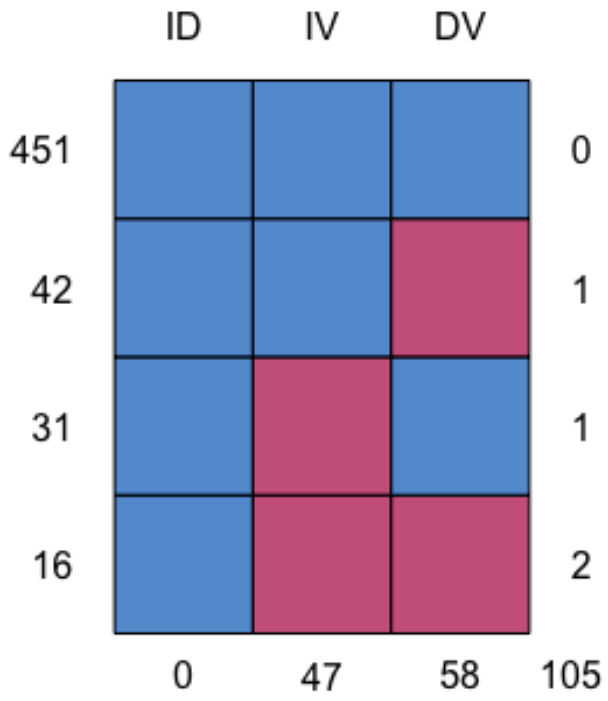
R Codes for Part A:

```
library(mice)

##
## Attaching package: 'mice'

## The following object is masked from 'package:stats':
##
##     filter

## The following objects are masked from 'package:base':
##
##     cbind, rbind

library(knitr)
wdir <- "~/OneDrive - Stony Brook University/SBU/MAT + AMS/Fall 2021/AMS 315/
project/01/Data"
setwd(wdir)

PartA_IV <- read.csv("141951_IV.csv", header = T)
PartA_DV <- read.csv("141951_DV.csv", header = T)
PartB <- read.csv("141951_PartB.csv", header = T)

PartA <- merge(PartA_IV, PartA_DV, by = 'ID')

str(PartA)

## 'data.frame':    540 obs. of  3 variables:
##  $ ID: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ IV: num  3.1 8.21 6.65 4.49 8.36 ...
##  $ DV: num  60.7 90 81.1 NA 109.9 ...
```

Report:

There are 540 observations in this file and three columns in the data set, called 'ID', 'IV' and 'DV'. Each is numerical.

```
md.pattern(PartA)
```




```
##     ID IV DV
## 451  1  1  1   0
## 42   1  1  0   1
## 31   1  0  1   1
## 16   1  0  0   2
##      0 47 58 105
```

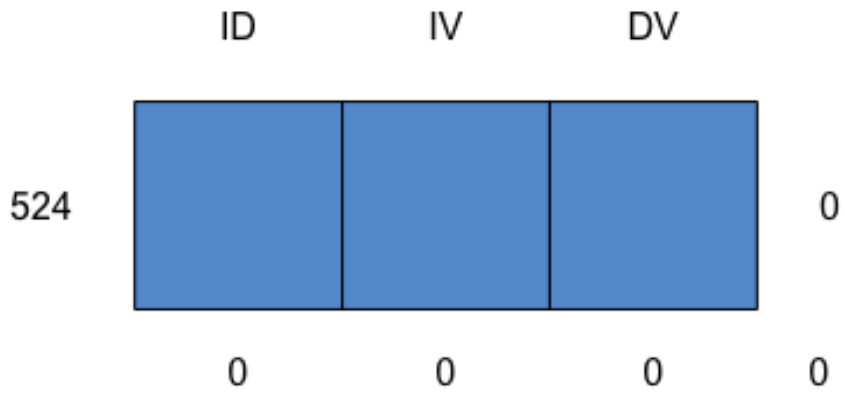
Report:

There are 451 complete data sets.

IV is missing in 47 cases, DV is missing in 58 cases, and both IV and DV are missing in 16 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)

##  /\     /\
## {  `---'  }
## {  O   O  }
## ==>  V <==  No need for mice. This data set is completely observed.
```

```
##  \  \|/  /
##   `-----'
```




```
##     ID IV DV
## 524  1  1  1 0
##      0  0  0 0
```

Report:

There are 524 complete data sets after imputing missing data. There are no data for 16 observations

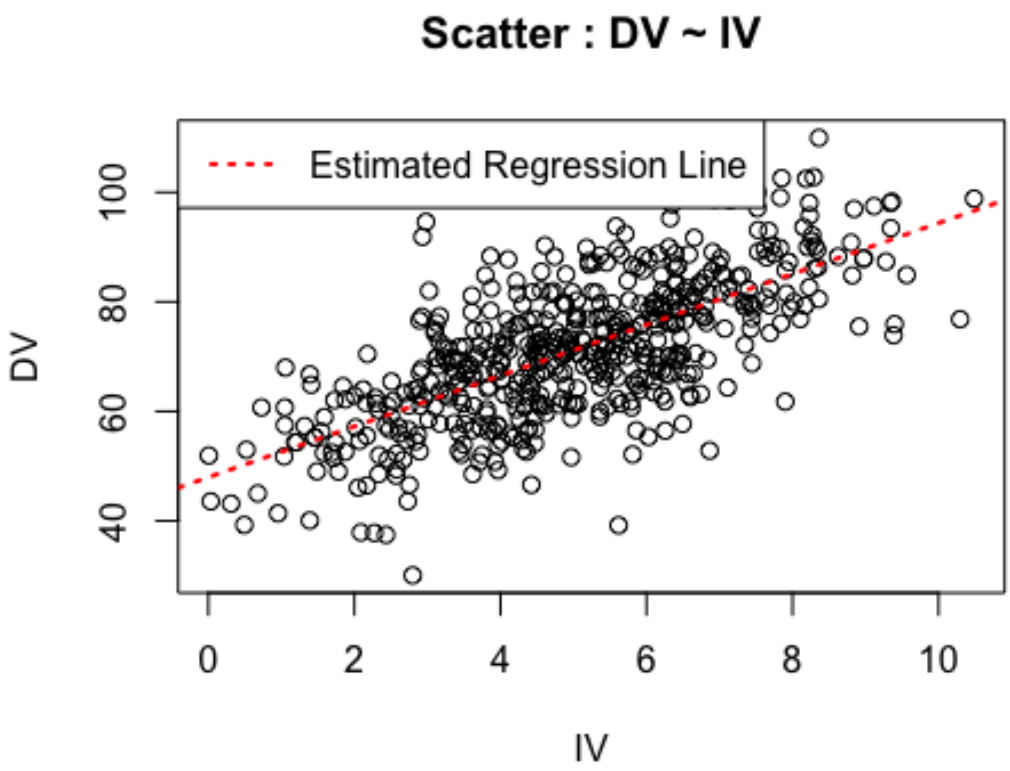
```
fit <- lm(DV ~ IV, data = PartA_complete)
summary(fit)

##
## Call:
## lm(formula = DV ~ IV, data = PartA_complete)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.848  -6.295   0.463   5.647  32.815
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  47.9576     1.1470   41.81   <2e-16 ***
## IV            4.6419     0.2134   21.75   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.501 on 522 degrees of freedom
## Multiple R-squared:  0.4754, Adjusted R-squared:  0.4744
## F-statistic: 473.1 on 1 and 522 DF,  p-value: < 2.2e-16

kable(anova(fit), caption='ANOVA Table')
```

ANOVA Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
IV	1	42702.51	42702.51055	473.0815	0
Residuals	522	47118.12	90.26459	NA	NA

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



95% confidence interval of the slope:

```
confint(fit, level = 0.95)
```

```
##                 2.5 %    97.5 %
## (Intercept) 45.70434 50.210775
## IV           4.22265  5.061174
```

99% confidence interval of the slope

```
confint(fit, level = 0.99)
```

```
##                  0.5 %    99.5 %
## (Intercept) 44.992349 50.922762
## IV           4.090169  5.193655
```

Report for Part B

Introduction:

The objective is to find the transformation of Independent and Dependent Variables, bin data, and apply the lack of fit test. In this report, I am using the R statistical package to conduct the necessary functions to draw my conclusions. My goal is to find a linear equation through the transformation of either independent or dependent variable, or both.

Methods:

In problem B, the dataset was supplied with one data sheet in .csv file. There are 444 observations in the dataset. There is no missing data field. The dataset had the ID of each observation and its associated independent and dependent variable value. I found that the best linear fit using linear regression model with a transformation of the dependent variable of $\sqrt{y}$. This was done after trying several various transformations such as $y \sim x^2$, $\log(y) \sim \log(x)$, $y \sim \log(x)$, $\frac{1}{y} \sim x$.

Results:

The fitted function for the model y = $(2.67823 + 0.03919x)^2$ with a transformation of the dependent variable of $\sqrt{y}$. The coefficient of determination (R-square) is 0.4069. The correlation value is 0.6379092. The 95% confidence interval for the slope was [0.03476374, 0.04360847]. The 95% confidence interval for the intercept was

[2.61767086, 2.73878486]. The 99% confidence interval for the slope was [0.03336492, 0.0450073]. The 99% confidence interval for the intercept was [2.59851626, 2.7579395].

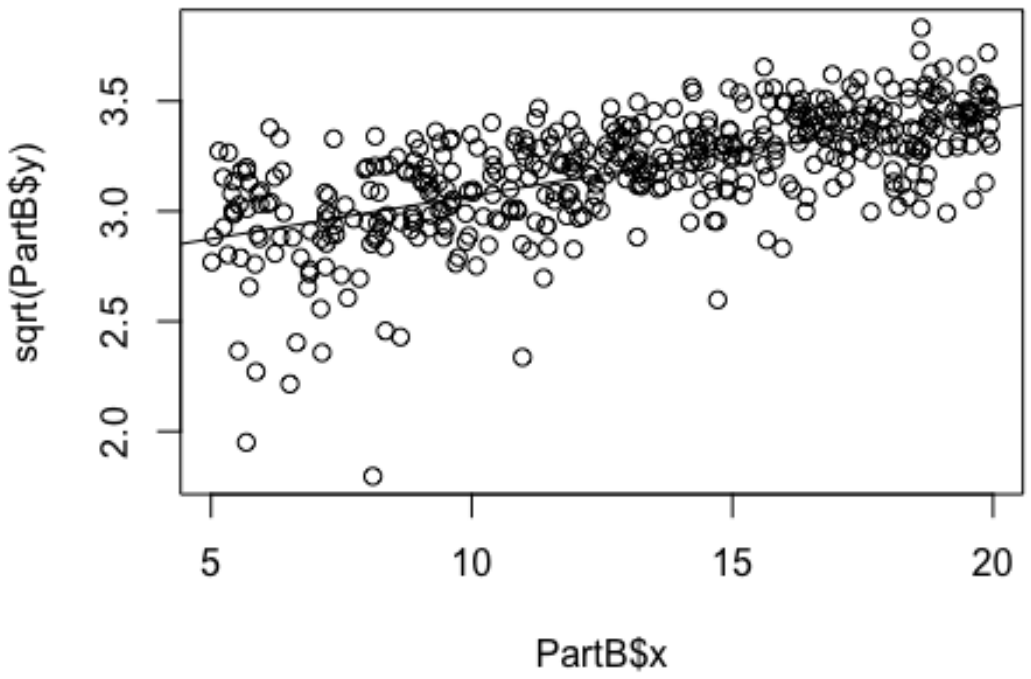
Soure	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	12.5890	12.5890	309.9736	< 2e-16
Residual Error	442	18.3961	0.0416		
Lack of Fit	47	2.3539	0.0501	1.2332	0.1486
Pure Error	395	16.0422	0.0406		
Total	443	30.9851			

Conclusions and Discussions:

Suppose the null hypothesis that the relationship assumed in the model is reasonable, and the alternative hypothesis that the relationship assumed in the model is not reasonable. The null hypothesis is rejected, i.e., there is lack of fit in the model. It is because the analysis of variance reveals the F statistic of 309.9736, which is greater than 10.9738, the critical value of F at α = 0.001, df1 = 1, and df2 = 442.

R Codes for Part B

```
fit_trial <- lm(sqrt(PartB$y) ~ PartB$x)
plot(sqrt(PartB$y) ~ PartB$x)
abline(fit_trial)
```




```
summary(fit_trial)

##
## Call:
## lm(formula = sqrt(PartB$y) ~ PartB$x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.19948 -0.10184  0.01495  0.13031  0.46013
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.67823    0.03081   86.92   <2e-16 ***
## PartB$x      0.03919    0.00225   17.41   <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2039 on 442 degrees of freedom
## Multiple R-squared:  0.4069, Adjusted R-squared:  0.4056
## F-statistic: 303.3 on 1 and 442 DF,  p-value: < 2.2e-16

PartB_trans <- data.frame(xtrans = PartB$x, ytrans = sqrt(PartB$y))

cor(PartB_trans$xtrans, PartB_trans$ytrans)

## [1] 0.6379092

confint(fit_trial, level = 0.95)

##                   2.5 %     97.5 %
## (Intercept) 2.61767086 2.73878486
## PartB$x     0.03476374 0.04360847

confint(fit_trial, level = 0.99)

##                   0.5 %    99.5 %
## (Intercept) 2.59851626 2.7579395
## PartB$x     0.03336492 0.0450073

groups <- cut(PartB_trans$xtrans, breaks = c(-Inf, seq(min(PartB_trans$xtrans
) + 0.3, max(PartB_trans$xtrans) - 0.3, by = 0.3), Inf))
table(groups)

## groups
## (-Inf,5.32] (5.32,5.62] (5.62,5.92] (5.92,6.22] (6.22,6.52] (6.52,6.82]
##           5          11          10           6           7           3
## (6.82,7.12] (7.12,7.42] (7.42,7.72] (7.72,8.02] (8.02,8.32] (8.32,8.62]
##           6          12           3           5          10           7
## (8.62,8.92] (8.92,9.22] (9.22,9.52] (9.52,9.82] (9.82,10.1] (10.1,10.4]
##          11           9          12           8           9           7
## (10.4,10.7]   (10.7,11]   (11,11.3] (11.3,11.6] (11.6,11.9] (11.9,12.2]
##           3          12           9          10          10          10
## (12.2,12.5] (12.5,12.8] (12.8,13.1] (13.1,13.4] (13.4,13.7]   (13.7,14]
##           8           8          11          13           8           4
##   (14,14.3] (14.3,14.6] (14.6,14.9] (14.9,15.2] (15.2,15.5] (15.5,15.8]
##           8          12           9          11           5          10
## (15.8,16.1] (16.1,16.4] (16.4,16.7]   (16.7,17]   (17,17.3] (17.3,17.6]
##           7          11          10          13           8           8
## (17.6,17.9] (17.9,18.2] (18.2,18.5] (18.5,18.8] (18.8,19.1] (19.1,19.4]
##          10          11           9          16           9           6
## (19.4, Inf]
##          24

x_b <- ave(PartB_trans$xtrans, groups)
data_bin <- data.frame(x = x_b, y = PartB_trans$ytrans)
```

```
library(remotes)
library(alr3)

## Loading required package: car

## Loading required package: carData

fit_b <- lm(y ~ x, data = data_bin)
pureErrorAnova(fit_b)

## Analysis of Variance Table
##
## Response: y
##               Df  Sum Sq Mean Sq  F value Pr(>F)
## x              1 12.5890 12.5890 309.9736 <2e-16 ***
## Residuals    442 18.3961  0.0416
##  Lack of fit  47  2.3539  0.0501   1.2332 0.1486
##  Pure Error  395 16.0422  0.0406
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```