Data Analysis Project 1 Part A

Introduction

As stated in the assignment description, the first part of the project is the first task that a newly hired statistician might be given usually. Part A of the project is to describe a linear regression model for the dataset provided. We were given two datasets, one for IV and another one for DV, to conduct numerous of statistician functions and draw a conclusion for my fictional manager. The purpose of this assignment is to practice how to better use the statistical computing resources and the information that is given to us to achieve the objective of the task.

Methodology

Since the data is in .csv format, there is no problem in reading the file in R. My files do not contain a header, so it is important to create headers that correspond to the data in the files. After we do that, we then start merging the two files by ID as the key. I will write a new .csv file that contains ID, IV, and DV that is sorted automatically by ID. After examining the new data file, I got 661 subject IDs that had at least one IV and DV value, 606 that had an IV value, 567 that had a DV value, 13 that are missing both ID and DV, and 512 that had both ID and DV value. We should not use mean imputation, listwise deletion methods or median imputation, so I used function “norm.boot”, which is linear regression using bootstrap. Now we have a dataset with total 661 IDs. The data was then used so that we find a line of best fit using a linear regression model.

Result

The linear fit model I have successfully computed is DV = 75.6936 - 2.8597*IV. SSR/SST=0.42279326. Therefore, it means that 42.28% of the variation in DV is explained by variation in IV. We can set a null hypothesis that the slope of the linear regression line is 0 and an alternative hypothesis that the slope is not 0. To compute the critical value, we set alpha = 0.001, df1=1 and df2=659, and compute the critical value in R. We will reject the null hypothesis because the ANOVA shows F statistics is 482.4164 which is greater than the critical value of 10.92535. The 99% confidence interval puts the value of the slope between -3.196047 and -2.523356. We reject the null hypothesis that the slope is zero at the 0.001 significance level. I can safely say that slope I computed was -2.8597.

Appendix

ANOVA[a]					
Model	Sum of Squares	DF	Mean Square	F	Sig.
Regression	1223.510	1	1223.509978	482.4164	0[b]
Residual	1671.363	659	2.536211		
Total	2893.873	660			

a: Dependent Variable: DV

b: Predictors: (Constant), IV

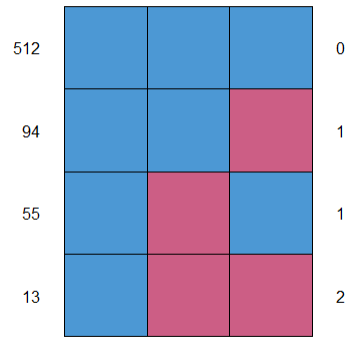
code

```
wdir <- "C:\\Users\\horat\\OneDrive\\Dekstop\\School Stuff\\Stony Brook University\\AMS 315\\Project 1"
setwd(wdir)
PartA_IV <- read.csv('P1A_IV_18878.csv', header = FALSE)
PartA_DV <- read.csv('P1A_DV_18878.csv', header = FALSE)
colnames(PartA_IV) <- c("ID","IV")
colnames(PartA_DV) <- c("ID","DV")
PartA <- merge(PartA_IV, PartA_DV, by = 'ID')
str(PartA)
any(is.na(PartA[,2]) == TRUE)
any(is.nan(PartA[,2]) == TRUE)
any(is.null(PartA[,2]) == TRUE)
library(mice)
PartA_incomplete <- PartA
md.pattern(PartA_incomplete)
PartA_incomplete <- PartA[(!is.na(PartA_IV[,2])==TRUE)|(!is.na(PartA_DV[,2])==TRUE),]
imp <- mice(PartA_incomplete, method = "norm.boot", printFlag = FALSE)
PartA_complete <- complete(imp)
write.csv(PartA_complete, "PartA_complete.csv", row.names=FALSE)
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 = 0.99)
qf(0.999, df1=1, df2=659)
```

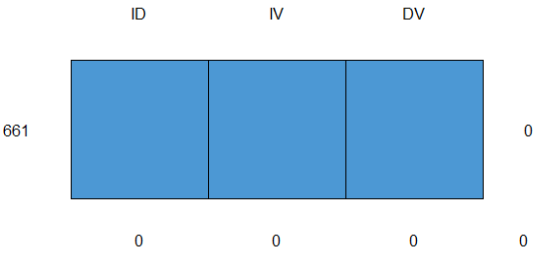
Some result

```
> str(PartA)
'data.frame':   674 obs. of  3 variables:
 $ ID: int  1 2 3 4 5 6 7 8 9 10 ...
 $ IV: num  1.53 1.94 2.46 NA 1.79 ...
 $ DV: num  72.8 69.7 70.2 75.6 NA ...
```

```
> md.pattern(PartA_incomplete)
    ID IV  DV
512  1  1   1   0
94   1  1   0   1
55   1  0   1   1
13   1  0   0   2
     0 68 107 175
```




```
> md.pattern(PartA_complete)
 /\     /\
{  `---'  }
{  O   O  }
==>  V <==  No need for mice. This data set is completely observed.
 \  \|/  /
  `-----'

    ID IV DV
661  1  1  1 0
     0  0  0 0
```

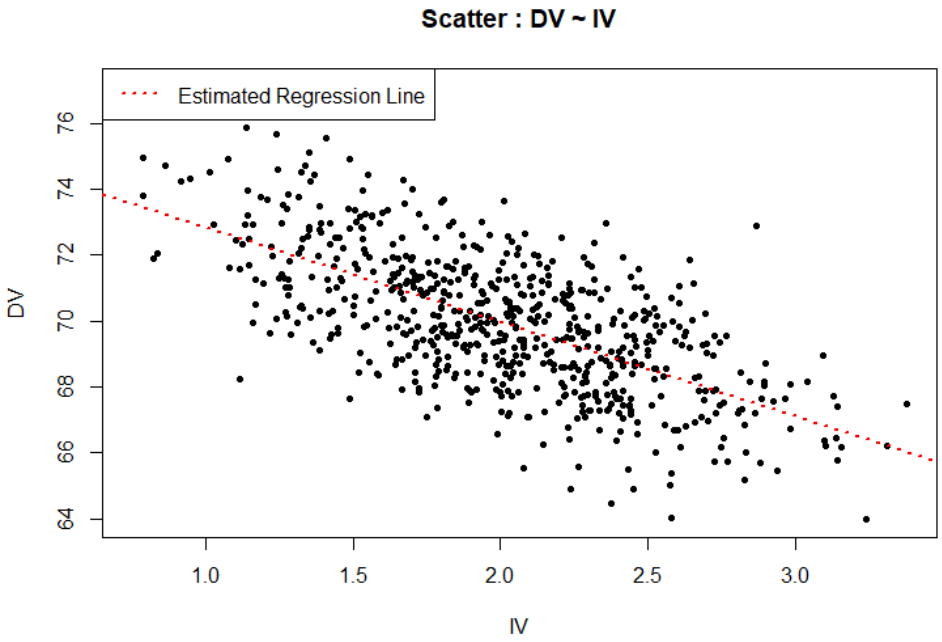



```
> summary(M)

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

Residuals:
    Min      1Q  Median      3Q     Max
-4.4319 -1.1382  0.0316  1.1511  5.3794

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  75.6939     0.2684  282.03   <2e-16 ***
IV           -2.8597     0.1302  -21.96   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.593 on 659 degrees of freedom
Multiple R-squared:  0.4226,    Adjusted R-squared:  0.4218
F-statistic: 482.4 on 1 and 659 DF,  p-value: < 2.2e-16
```

```
> kable(anova(M), caption='ANOVA Table')
```

Table: ANOVA Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
IV	1	1223.510	1223.509978	482.4164	0
Residuals	659	1671.363	2.536211	NA	NA




```
> confint(M, level = 0.99)
                0.5 %    99.5 %
(Intercept) 75.000608 76.387290
IV          -3.196047 -2.523356
> qf(0.999, df1=1, df2=659)
[1] 10.92535
```