Soyoung Chung

Computer Assignment Report

Part A

Introduction

The objective of this report was to explain the process that we came across to our solution. There are two tasks given. The first task is to sort two files by subject ID and merge them. Second task is impute the missing values. We were given two different files in PartA. One has a list of ID's with corresponding independent variables and the other has a list of ID's with corresponding dependent variables. Afterwards, using the exist data to perform linear regression and analyze the result.

Methodology

I used the statistic packages, Excel and R studio. I used R studio to sort and merge two files into one file and used Excel to count the number of ID that satisfies conditions.

First, in order to merge two given files, read those two files in R, merge two files with sorted by ID. Then, I can get beautiful tables that have been sorted and merged by ID numbers ranged from 1 to 780. Using Excel count function, I can get the count of the number of subject IDs that had at least one independent variable value or dependent variable value or the count of the number of subject that have IV value. I used R studio to impute the missing values. I used a package called MICE. I chose to use bootstrap linear regression. Using R commands, drop cases(for my file, it is 6) that both miss IV and DV since they do not contain any information and choose MICE method.

Results

In order to check if two files are sorted and merged correctly, you may go to AMS315Proejct1-P1A_46738.cvs or Project1 Code.r under submission

Count of the number of subject IDs that have the value of		Excel Function
Dependent Variable(DV)	616	"=COUNT(B2:B781)"
Independent Variable(IV)	733	"=COUNT(C2:C781)"
At least one IV or DV	774	"780-COUNTIFS(B1:B781, {"NA"},C1:C781,{"NA"})"
Both IV and DV	575	"780-[COUNTIF(B2:B781,"NA") +COUNTIF(C2:C781,"NA")- COUNTIFS(B1:B781,{"NA"},C1:C781, {"NA"})]"

(Total number of data = 780, There are 575 complete data sets, IV is missing 47datas, DV is missing in 164 datas, both are missing 6 cases.)

ANOVA Table					
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
IV	1	45135.74	45135.73854	833.2836	0
Residuals	772	41816.24	54.16612	NA	NA

Conclusion

To conclude results of Part A, The 95% confidence interval for the slope is [3.649503, 4.182081]. The 99% confidence interval for the slope is [3.565512, 4.266072]. I provided an analysis of variance table above and my results found the association between the independent and dependent variables to be highly significant with a p-value of 2.2e-16, F value of 833.2836, Multiple R-squared of 0.5191 and Adjusted R-squared is 0.5185 .

Part B

Introduction

The objective of Part B to find repeated independent variable values and perform the lack of fit(LOF) test on the data set. The task Part B was to bin data first if there are few exact repeats of an independent variable value, then apply of the lack of fit(LOF) test. There is one file given, which contains on line for each subject ID. The list will contain the subject ID with the value of the Independent Variable(IV) and Dependent Variable(DV).

Methodology

I used the statistic package, R studio only. First, I had to find few exact repeats of an independent variable value. Since the values of the data are small and the number of significant number is around 15, it is hard to find the exact repeats of an independent variable value. Before binning the points into one group of "nearly" repeated points, I have to find my transformation first. The code is

```
>data_trans <- data.frame(xtrans=data$x, ytrans=data$y^(-3/2)),
```

which is IV and DV^(-3/2)

In order to bin appopriately, I had to go through my data throughly to check the bin interval. Some of data's difference is less than 0.01, So, I used 0.05 as bin interval.

Results

```
> table(groups)
groups
 (-Inf,1.107] (1.107,1.112] (1.112,1.117] (1.117,1.122] (1.122,1.127] (1.127,1.132] (1.132,1.137]
            6            3            6            8            5            5            8
(1.137,1.142] (1.142,1.147] (1.147,1.152] (1.152,1.157] (1.157,1.162] (1.162,1.167] (1.167,1.172]
            9            5            6            4            7            6            3
(1.172,1.177] (1.177,1.182] (1.182,1.187] (1.187,1.192] (1.192,1.197] (1.197,1.202] (1.202,1.207]
            5           10            7            2            1            1            5
(1.207,1.212] (1.212,1.217] (1.217,1.222] (1.222,1.227] (1.227,1.232] (1.232,1.237] (1.237,1.242]
            3            2            5            4            7            6            7
(1.242,1.247] (1.247,1.252] (1.252,1.257] (1.257,1.262] (1.262,1.267] (1.267,1.272] (1.272,1.277]
            5            3            6            5            7            5            4
(1.277,1.282] (1.282,1.287] (1.287,1.292] (1.292,1.297] (1.297,1.302] (1.302,1.307] (1.307,1.312]
            6            9            3           10            3            5            3
(1.312,1.317] (1.317,1.322] (1.322,1.327] (1.327,1.332] (1.332,1.337] (1.337,1.342] (1.342,1.347]
            3            7            3            1            3            6            4
(1.347,1.352] (1.352,1.357] (1.357,1.362] (1.362,1.367] (1.367,1.372] (1.372,1.377] (1.377,1.382]
            4            4           11            3            6            4            4
(1.382,1.387] (1.387,1.392] (1.392,1.397] (1.397,1.402] (1.402,1.407] (1.407,1.412] (1.412,1.417]
            5            5            6            6            7            7            5
(1.417,1.422] (1.422,1.427] (1.427,1.432] (1.432,1.437] (1.437,1.442] (1.442,1.447] (1.447,1.452]
            2            6            4            5            4            6            3
(1.452,1.457] (1.457,1.462] (1.462,1.467] (1.467,1.472] (1.472,1.477] (1.477,1.482] (1.482,1.487]
            5            7            5            3            9            3            4
(1.487,1.492] (1.492,1.497] (1.497,1.502] (1.502,1.507] (1.507,1.512] (1.512,1.517] (1.517,1.522]
            1            4            2            3            7            3            5
(1.522,1.527] (1.527,1.532] (1.532,1.537] (1.537,1.542] (1.542,1.547] (1.547,1.552] (1.552,1.557]
            7            3            4            8            8            4            2
(1.557,1.562] (1.562,1.567] (1.567,1.572] (1.572,1.577] (1.577,1.582] (1.582,1.587] (1.587,1.592]
            1            3            8            5            4            6            3
 (1.592, Inf]
            9
Analysis of Variance Table

Response: y
              Df Sum Sq Mean Sq  F value Pr(>F)
x              1 520.97  520.97 797.9299 <2e-16 ***
Residuals    488 302.16    0.62
 Lack of fit  97  46.88    0.48   0.7402 0.9628
 Pure Error  391 255.29    0.65
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

As you can see the square in the attached table, the group of nearly repeated point is (1.357,1.362], which the number of repeated is 11. We provided an analysis of variance table above and my results found the association between

the independent to be highly significant with the LOF F value is 0.7402, showing that there is no significant lack of fit when applying Lack of Fit(LOF) test.

Conclusion

To conclude our results of Part B, I found the association between the independent variables to be highly significant due to the LOF F value is 0.7402 and Sum Sq of 46.88 and Mean Sq of 0.48