Data Analysis Project 1 Part B

Introduction

In Part B of this assignment, we are given a merged set of data. The task asks us to bin the data and apply the LOF test. We need to find the repeated or nearly repeated independent variables to apply the process of binning. The purpose is to get us familiar with data binning and materials about LOF test.

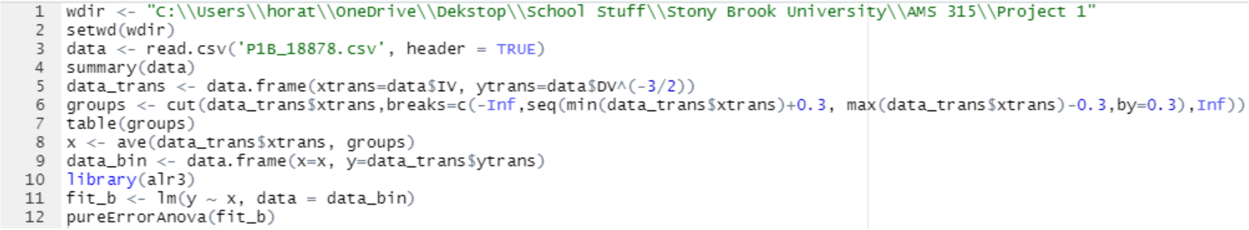
Methodology

Since the task for us is to bin the data and then apply the LOF test. Even there are not exactly repeated IV values, we could still bin these points into one group of nearly repeated points. We can use the cut() function in R. After binning, we will compute the average IV value. We can use the function ave() to compute the average IV value. We do not need to bin dependent variable values. After binning data, the next step is to apply the Lack of Fit test. We can do so with the help of pureErrorAnova() function. It generates an ANOVA table that is used to determine the F statistic for the test of lack of fit. Through a series of calculations, we can determine if the F test for lack of fit is or is not significant.

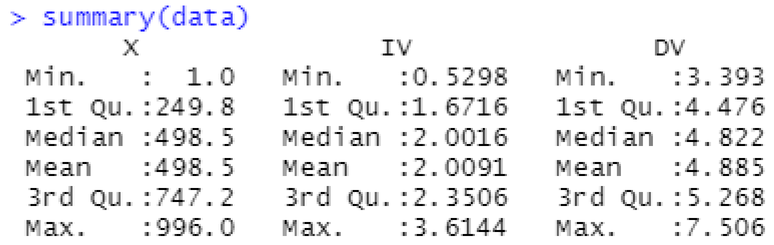
Result

As you can see from the appendix, SSLF=0.003141 and SSPE=0.154816 so the sum SSE=0.154816. From Df column we can see that n=996 and c=10 because that c-2=8, n-c=986, n-2=994. This is the basic analysis of the variance table. The above information is used to calculate mean squares. MSLF=SSLF/(c-2)=0.003141/8 and MSPE=SSPE/(n-c)=0.154816/986. We are going to make hypotheses H_0: there is no lack of fit; and H_a: there is lack of fit. To conduct the lack of fit test, we calculate the value of the F-statistic, which is F=MSLF/MSPE=2.503. We then follow standard hypothesis test procedures. H_0 means the model has no lack of fit, while H_1 means there is a lack of fit. We can make a decision that since the F*-statistic is 2.5005 and the P-value is 0.01086, and the P-value is smaller than the significance level α=0.05, so we reject the null hypothesis. Therefore, there is a lack of fit in the model.

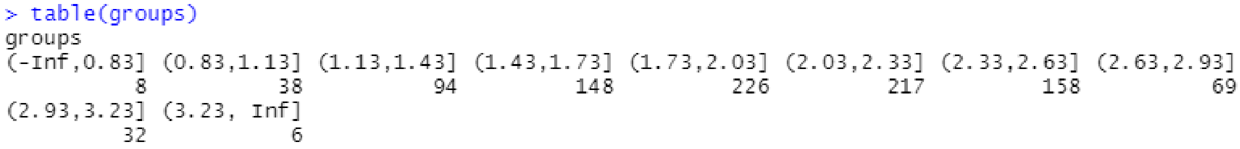
Appendix



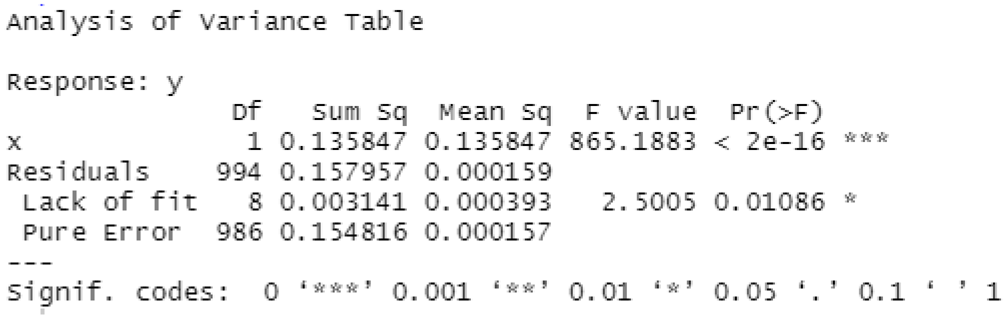
```
wdir <- "C:\\Users\\horat\\OneDrive\\Dekstop\\School Stuff\\Stony Brook University\\AMS 315\\Project 1"
setwd(wdir)
data <- read.csv('P1B_18878.csv', header = TRUE)
summary(data)
data_trans <- data.frame(xtrans=data$IV, ytrans=data$DV^(-3/2))
groups <- cut(data_trans$xtrans,breaks=c(-Inf,seq(min(data_trans$xtrans)+0.3, max(data_trans$xtrans)-0.3,by=0.3),Inf))
table(groups)
x <- ave(data_trans$xtrans, groups)
data_bin <- data.frame(x=x, y=data_trans$ytrans)
library(alr3)
fit_b <- lm(y ~ x, data = data_bin)
pureErrorAnova(fit_b)
```



```
> summary(data)
       X               IV               DV       
 Min.   :  1.0   Min.   :0.5298   Min.   :3.393  
 1st Qu.:249.8   1st Qu.:1.6716   1st Qu.:4.476  
 Median :498.5   Median :2.0016   Median :4.822  
 Mean   :498.5   Mean   :2.0091   Mean   :4.885  
 3rd Qu.:747.2   3rd Qu.:2.3506   3rd Qu.:5.268  
 Max.   :996.0   Max.   :3.6144   Max.   :7.506  
```



```
> table(groups)
groups
(-Inf,0.83] (0.83,1.13] (1.13,1.43] (1.43,1.73] (1.73,2.03] (2.03,2.33] (2.33,2.63] (2.63,2.93] 
          8          38          94         148         226         217         158          69 
(2.93,3.23]  (3.23, Inf] 
         32           6 
```



```
Analysis of Variance Table

Response: y
              Df   Sum Sq  Mean Sq  F value  Pr(>F)    
x              1 0.135847 0.135847 865.1883 < 2e-16 ***
Residuals    994 0.157957 0.000159                     
 Lack of fit   8 0.003141 0.000393   2.5005 0.01086 *  
 Pure Error  986 0.154816 0.000157                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```