September 26, 2020

knitr::opts\_chunk$set(echo = TRUE)

Please answer the following questions. These assume you have performed the readings and watched each of the lectures up to this point. Where necessary, please print your R output as your answer.

1: Please find the examScores.xlsx file (Canvas > Files > homeworks > homework\_03) and determine if exam1 scores are related to exam2 scores. (5 points)

# Libraries needed   
library(ggplot2)  
library(tidyverse)

## -- Attaching packages ----------------------------------------------------------------------------------------------------------------- tidyverse 1.3.0 --

## v tibble 3.0.3 v dplyr 1.0.2  
## v tidyr 1.1.2 v stringr 1.4.0  
## v readr 1.3.1 v forcats 0.5.0  
## v purrr 0.3.4

## -- Conflicts -------------------------------------------------------------------------------------------------------------------- tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

#library(car)  
library(readxl)  
library(dplyr)  
library(car)

## Loading required package: carData

##   
## Attaching package: 'car'

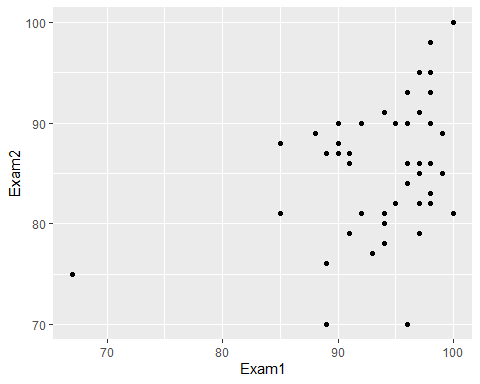
## The following object is masked from 'package:dplyr':  
##   
## recode

## The following object is masked from 'package:purrr':  
##   
## some

# reading and exploring Excel file  
exam\_scores <- read\_excel("examScores.xlsx")

## New names:  
## \* `` -> ...1

#glimpse(exam\_scores)  
exam\_scores <-data.frame(exam\_scores) # creating data frame  
#working on variables required  
Exam1<-exam\_scores$Exam1  
Exam2<-exam\_scores$Exam2  
## make graph object  
graph <- ggplot(exam\_scores,aes(Exam1,Exam2))  
## creating a scatter plot  
graph+geom\_point()



# Pearsons Correlation  
cor.test(Exam1,Exam2,alternative="two.sided",method="pearson", conf.level=0.95)

##   
## Pearson's product-moment correlation  
##   
## data: Exam1 and Exam2  
## t = 2.657, df = 46, p-value = 0.0108  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.08994502 0.58795647  
## sample estimates:  
## cor   
## 0.3647578

Since it has been clear here that correlation which is not equal to 0 so relation do exist between scores of Exam1 and Exam2 and there are positive values more than zero to 1, there is positive relation between scores of two exams with small effect.

2: Can exam1 scores be used as an accurate predictor of exam2 scores? Please provide the model that captures the relationship between these two variables. Please plot your model as well. (5 points)

### performing simple least squares regression  
lm\_scores=lm(Exam2~Exam1,data=exam\_scores)  
print(lm\_scores)

##   
## Call:  
## lm(formula = Exam2 ~ Exam1, data = exam\_scores)  
##   
## Coefficients:  
## (Intercept) Exam1   
## 44.2707 0.4406

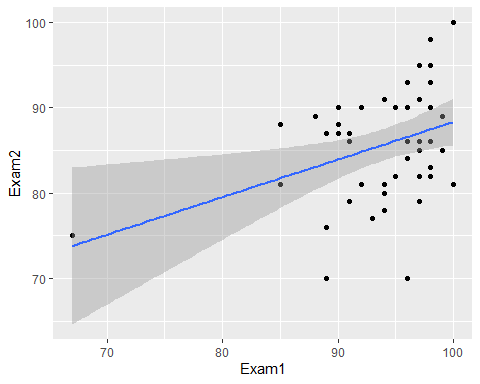
# display statistical summary of the regression  
summary(lm\_scores)

##   
## Call:  
## lm(formula = Exam2 ~ Exam1, data = exam\_scores)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -16.5636 -4.5042 0.2755 4.3594 11.6741   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 44.2707 15.5993 2.838 0.00673 \*\*  
## Exam1 0.4406 0.1658 2.657 0.01080 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.352 on 46 degrees of freedom  
## Multiple R-squared: 0.133, Adjusted R-squared: 0.1142   
## F-statistic: 7.059 on 1 and 46 DF, p-value: 0.0108

Since, p-value is less than 0.05 so there is accuracy in the predictor but Its not good idea to take p-value as a single factor to decide accuracy. According to ASA (American statistical association), lots of other factors such as previously done analysis on similar data, how properly the analysis is done and in case of lack of evidences the reproducibility should be checked well before stepping to the concrete decision.

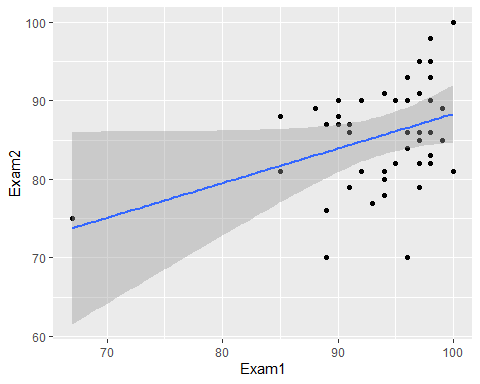
# plotting the model  
## graph scatter plot with linear model (lm)  
graph <- ggplot(exam\_scores,aes(Exam1,Exam2))  
## create a scatter plot + regression line  
graph+geom\_point()+geom\_smooth(method="lm",se=T)

## `geom\_smooth()` using formula 'y ~ x'

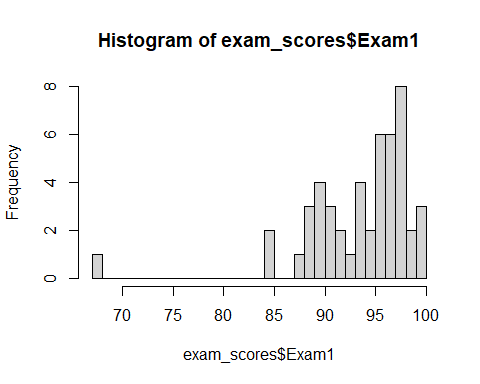


## create a scatter plot + regression line and 99% CI  
graph+geom\_point()+geom\_smooth(method="lm", level=0.99)

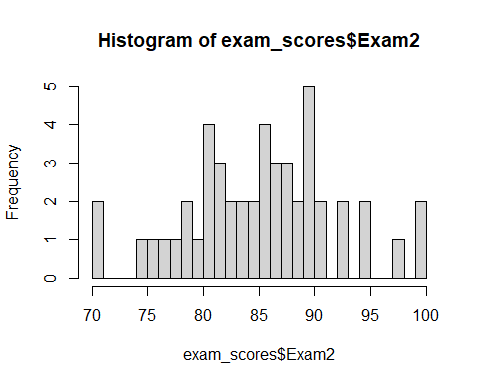
## `geom\_smooth()` using formula 'y ~ x'



## Normality check  
### Qualitative , plotting histograms with bin 30  
hist(exam\_scores$Exam1,30)



hist(exam\_scores$Exam2,30)



# Spearman's Correlation (there is outlier present in Exam1)  
cor.test(Exam1,Exam2,alternative="two.sided",method="spearman", conf.level=0.95)

## Warning in cor.test.default(Exam1, Exam2, alternative = "two.sided", method =  
## "spearman", : Cannot compute exact p-value with ties

##   
## Spearman's rank correlation rho  
##   
## data: Exam1 and Exam2  
## S = 12663, p-value = 0.03048  
## alternative hypothesis: true rho is not equal to 0  
## sample estimates:  
## rho   
## 0.3126864

# Getting rid of the warning message   
cor.test(Exam1,Exam2,alternative="two.sided",method="spearman", conf.level=0.95, exact=FALSE)

##   
## Spearman's rank correlation rho  
##   
## data: Exam1 and Exam2  
## S = 12663, p-value = 0.03048  
## alternative hypothesis: true rho is not equal to 0  
## sample estimates:  
## rho   
## 0.3126864

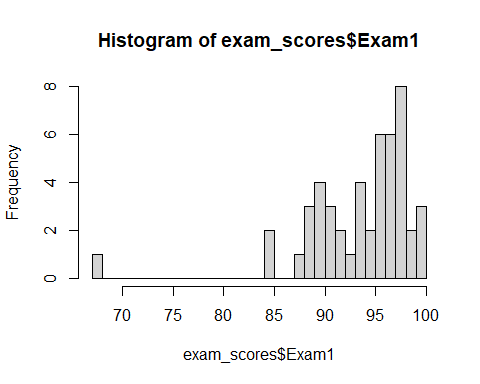
### Quantitative  
shapiro.test(exam\_scores$Exam1)

##   
## Shapiro-Wilk normality test  
##   
## data: exam\_scores$Exam1  
## W = 0.77712, p-value = 4.099e-07

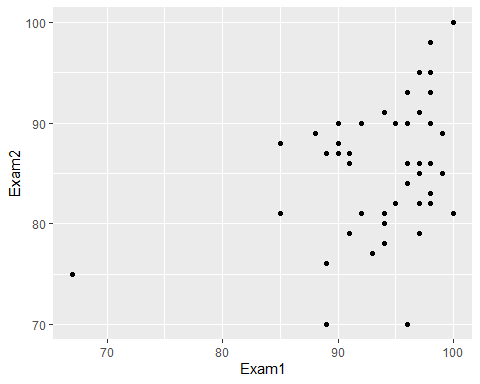
shapiro.test(exam\_scores$Exam2)

##   
## Shapiro-Wilk normality test  
##   
## data: exam\_scores$Exam2  
## W = 0.98473, p-value = 0.7803

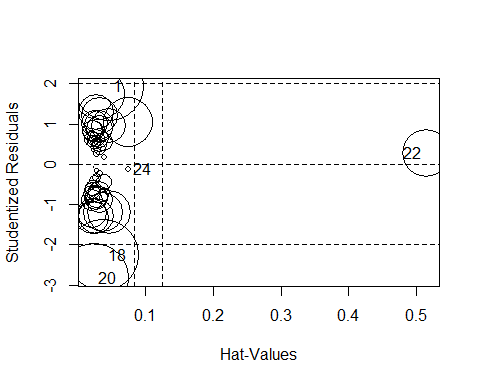
## As outlier is present, possibly influencing the regression  
hist(exam\_scores$Exam1, 30)



graph <- ggplot(exam\_scores,aes(Exam1,Exam2))  
graph+geom\_point()



### rerun model to   
lm\_scores=lm(exam\_scores$Exam2~exam\_scores$Exam1,data=exam\_scores)  
influencePlot(lm\_scores)

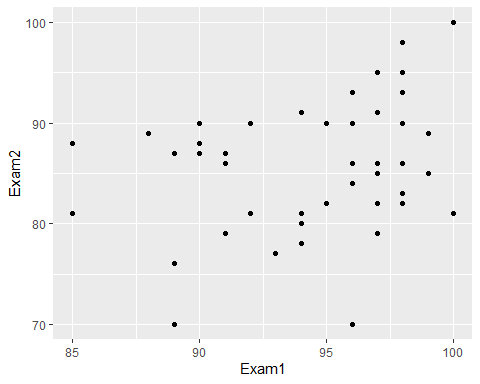


## StudRes Hat CookD  
## 1 1.9371186 0.04604815 0.0854534735  
## 18 -2.2569785 0.03730411 0.0906285328  
## 20 -2.8336017 0.02379060 0.0848695531  
## 22 0.2711333 0.51447876 0.0397494593  
## 24 -0.1161898 0.07500568 0.0005593399

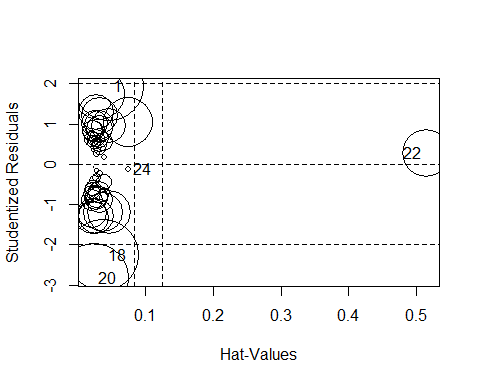
### Quantitative--> nonconstant variance test  
ncvTest(lm\_scores)

## Non-constant Variance Score Test   
## Variance formula: ~ fitted.values   
## Chisquare = 0.9637907, Df = 1, p = 0.32623

###filtering extreme value  
k=which(exam\_scores$Exam1!=min(exam\_scores$Exam1))  
cleanData=exam\_scores[k,]  
graph <- ggplot(cleanData,aes(Exam1,Exam2))  
graph+geom\_point()



lm\_scoresClean=lm(exam\_scores$Exam2~exam\_scores$Exam1,data=cleanData)  
influencePlot(lm\_scoresClean)



## StudRes Hat CookD  
## 1 1.9371186 0.04604815 0.0854534735  
## 18 -2.2569785 0.03730411 0.0906285328  
## 20 -2.8336017 0.02379060 0.0848695531  
## 22 0.2711333 0.51447876 0.0397494593  
## 24 -0.1161898 0.07500568 0.0005593399

Since, p-value is less than 0.05 so there is accuracy in the predictor but Its not good idea to take p-value as a single factor to decide accuracy. According to ASA (American statistical association), lots of other factors such as previously done analysis on similar data, how properly the analysis is done and in case of lack of evidences the reproducibility should be checked well before stepping to the concrete decision.cookD is below 1 which indicates there is no extreme scores. Overall the model is accurate.