Data Screening

Pradeep Paladugula

2020-03-01

# Dataset:

600 employees participated in a company-wide experiment to test if an educational program would be effective at increasing employee satisfaction. Half of the employees were assigned to be in the control group, while the other half were assigned to be in the experimental group. The experimental group was the only group that received the educational intervention. All groups were given an employee satisfaction scale at time one to measure their initial levels of satisfaction. The same scale was then used half way through the program and at the end of the program. The goal of the experiment was to assess satisfaction to see if it increased across the measurements during the program as compared to a control group.

## Variables:

a) Gender (1 = male, 2 = female)  
b) Group (1 = control group, 2 = experimental group)  
c) 3 satisfaction scores, ranging from 2-100 points. Decimals are possible! The control group was measured at the same three time points, but did not take part in the educational program.   
 i) Before the program  
 ii) Half way through the program   
 iii) After the program

eduResearchData <- read.csv('06\_data.csv')  
summary(eduResearchData)

## Gender Group Begin Middle   
## Min. :1.000 Min. :1.000 Min. : 61.15 Min. :37.35   
## 1st Qu.:1.000 1st Qu.:1.000 1st Qu.: 94.72 1st Qu.:59.88   
## Median :2.000 Median :2.000 Median :102.26 Median :64.12   
## Mean :1.505 Mean :1.508 Mean :102.17 Mean :63.86   
## 3rd Qu.:2.000 3rd Qu.:2.000 3rd Qu.:110.04 3rd Qu.:68.18   
## Max. :2.000 Max. :2.000 Max. :148.25 Max. :83.79   
## NA's :8 NA's :8 NA's :8 NA's :8   
## After   
## Min. : 48.15   
## 1st Qu.: 89.99   
## Median : 97.42   
## Mean : 95.83   
## 3rd Qu.:103.73   
## Max. :120.41   
## NA's :8

# Data screening:

## Accuracy:

a) Include output and indicate how the data are not accurate.  
b) Include output to show how you fixed the accuracy errors, and describe what you did.

eduResearchData$Gender = factor(eduResearchData$Gender, levels = c(1,2), labels = c("male", "female"))  
  
eduResearchData$Group = factor(eduResearchData$Group, levels = c(1,2), labels = c("control", "experimental"))  
  
table(eduResearchData$Gender)

##   
## male female   
## 194 198

table((eduResearchData$Group))

##   
## control experimental   
## 193 199

## Missing data:

a) Include output that shows you have missing data.  
b) Include output and a description that shows what you did with the missing data.  
 i) Replace all participant data if they have less than or equal to 20% of missing data by row.   
 ii) You can leave out the other participants (i.e. you do not have to create allrows).

summary(eduResearchData)

## Gender Group Begin Middle   
## male :194 control :193 Min. : 61.15 Min. :37.35   
## female:198 experimental:199 1st Qu.: 94.72 1st Qu.:59.88   
## NA's : 8 NA's : 8 Median :102.26 Median :64.12   
## Mean :102.17 Mean :63.86   
## 3rd Qu.:110.04 3rd Qu.:68.18   
## Max. :148.25 Max. :83.79   
## NA's :8 NA's :8   
## After   
## Min. : 48.15   
## 1st Qu.: 89.99   
## Median : 97.42   
## Mean : 95.83   
## 3rd Qu.:103.73   
## Max. :120.41   
## NA's :8

percentageMissingData = function(x){sum(is.na(x))/length(x)\*100}  
missingData = apply(eduResearchData, 1, percentageMissingData)  
table(missingData)

## missingData  
## 0 20 40 60   
## 363 35 1 1

replcacedData = subset(eduResearchData, missingData <= 20)  
notReplcedData = subset(eduResearchData, missingData > 20)  
missingDataLessThanTwenty = apply(replcacedData, 1, percentageMissingData)  
table(missingDataLessThanTwenty)

## missingDataLessThanTwenty  
## 0 20   
## 363 35

library('mice')

##   
## Attaching package: 'mice'

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

noMissingDataTemp = mice(replcacedData)

##   
## iter imp variable  
## 1 1 Gender Group Begin Middle After  
## 1 2 Gender Group Begin Middle After  
## 1 3 Gender Group Begin Middle After  
## 1 4 Gender Group Begin Middle After  
## 1 5 Gender Group Begin Middle After  
## 2 1 Gender Group Begin Middle After  
## 2 2 Gender Group Begin Middle After  
## 2 3 Gender Group Begin Middle After  
## 2 4 Gender Group Begin Middle After  
## 2 5 Gender Group Begin Middle After  
## 3 1 Gender Group Begin Middle After  
## 3 2 Gender Group Begin Middle After  
## 3 3 Gender Group Begin Middle After  
## 3 4 Gender Group Begin Middle After  
## 3 5 Gender Group Begin Middle After  
## 4 1 Gender Group Begin Middle After  
## 4 2 Gender Group Begin Middle After  
## 4 3 Gender Group Begin Middle After  
## 4 4 Gender Group Begin Middle After  
## 4 5 Gender Group Begin Middle After  
## 5 1 Gender Group Begin Middle After  
## 5 2 Gender Group Begin Middle After  
## 5 3 Gender Group Begin Middle After  
## 5 4 Gender Group Begin Middle After  
## 5 5 Gender Group Begin Middle After

noMissingData = complete(noMissingDataTemp, 1)  
summary(noMissingData)

## Gender Group Begin Middle   
## male :198 control :197 Min. : 61.15 Min. :37.35   
## female:200 experimental:201 1st Qu.: 94.65 1st Qu.:59.92   
## Median :102.26 Median :64.12   
## Mean :102.14 Mean :63.87   
## 3rd Qu.:110.10 3rd Qu.:68.14   
## Max. :148.25 Max. :83.79   
## After   
## Min. : 48.15   
## 1st Qu.: 90.08   
## Median : 97.42   
## Mean : 95.78   
## 3rd Qu.:103.64   
## Max. :120.41

allRow = rbind(notReplcedData, noMissingData)  
summary(allRow)

## Gender Group Begin Middle   
## male :198 control :197 Min. : 61.15 Min. :37.35   
## female:201 experimental:201 1st Qu.: 94.72 1st Qu.:59.87   
## NA's : 1 NA's : 2 Median :102.22 Median :64.12   
## Mean :102.12 Mean :63.85   
## 3rd Qu.:110.04 3rd Qu.:68.12   
## Max. :148.25 Max. :83.79   
## NA's :1   
## After   
## Min. : 48.15   
## 1st Qu.: 89.92   
## Median : 97.40   
## Mean : 95.75   
## 3rd Qu.:103.63   
## Max. :120.41   
## NA's :1

## Outliers:

a) Include a summary of your mahal scores that are greater than the cutoff.  
b) What are the df for your Mahalanobis cutoff?  
c) What is the cut off score for your Mahalanobis measure?  
  
```r  
cutoffScore = qchisq(0.999, ncol(allRow[,c(3,4,5)]))  
cutoffScore   
```  
  
```  
## [1] 16.26624  
```  
  
d) How many outliers did you have?  
 Solution: 1 outliers in Gender and 2 Outliers in Group  
   
e) Delete all outliers.

noMissData1 = noMissingData[,c(3,4,5)]  
cutoff = qchisq(1-.001,ncol(noMissData1))  
mahal = mahalanobis(noMissData1, colMeans(noMissData1, na.rm = TRUE), cov(noMissData1, use = "pairwise.complete.obs"))  
summary(mahal>cutoff)

## Mode FALSE TRUE   
## logical 396 2

output = subset(noMissingData, mahal<cutoff)  
str(output)

## 'data.frame': 396 obs. of 5 variables:  
## $ Gender: Factor w/ 2 levels "male","female": 2 2 2 2 2 2 2 2 2 2 ...  
## $ Group : Factor w/ 2 levels "control","experimental": 2 2 2 2 2 2 2 2 2 2 ...  
## $ Begin : num 104.3 69.8 87.3 94.8 98.6 ...  
## $ Middle: num 64 63.8 56.2 63.5 71 ...  
## $ After : num 105.3 98.4 93 106.9 112.7 ...

# Assumptions:

## Additivity:

a) Include the symnum bivariate correlation table of your continuous measures.  
b) Do you meet the assumption for additivity?  
  
 Solution: Yes, I have met the assumption of aditivity.

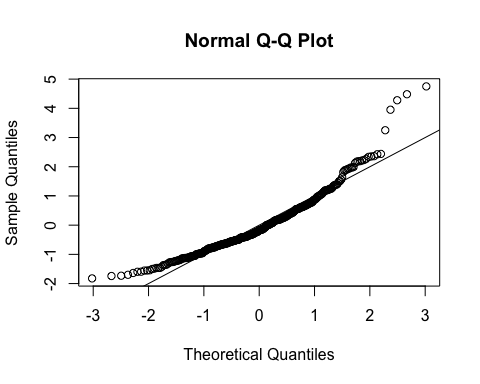
symnum(cor(output[, c(3,4,5)]))

## B M A  
## Begin 1   
## Middle 1   
## After 1  
## attr(,"legend")  
## [1] 0 ' ' 0.3 '.' 0.6 ',' 0.8 '+' 0.9 '\*' 0.95 'B' 1

## Linearity:

a) Include a picture that shows how you might assess multivariate linearity.  
b) Do you think you've met the assumption for linearity?  
  
 Solution: Yes, it seems I have reached the linearity.

linearityOutput = rchisq(nrow(output), 7)  
plot1 = lm(linearityOutput~., data = output)  
standardizedPlot = rstudent(plot1)  
as.numeric(unlist(qqnorm(standardizedPlot))) + abline(0,1)

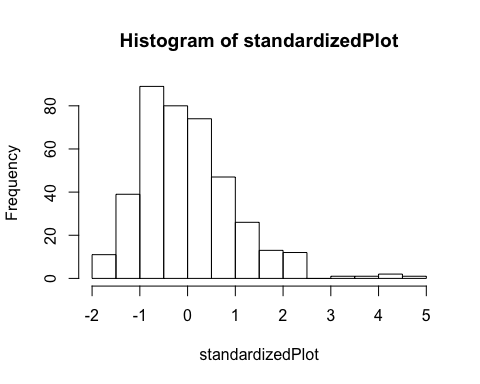


## numeric(0)

## Normality:

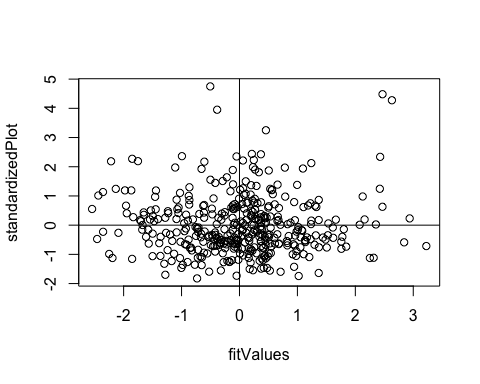
a) Include a picture that shows how you might assess multivariate normality.  
b) Do you think you've met the assumption for normality?   
  
 Solution: From the above picture it is clearly seen that data is skewed to the left

hist(standardizedPlot, breaks = 15)

 ## Homogeneity/Homoscedasticity:

a) Include a picture that shows how you might assess multivariate homogeneity.  
b) Do you think you've met the assumption for homogeneity?  
 Solution: No, I dont think I met the homogeneity. left to right is met, but no the top to bottom.  
c) Do you think you've met the assumption for homoscedasticity?  
 Solution: Yes, suumption of homoscedasticity is met.

fitValues = scale(plot1$fitted.values)  
{plot(fitValues, standardizedPlot)+abline(0,0)+abline(v =0)}



## integer(0)