Lab 3

Aniket

2018-12-16

# Dataset:

350 students were given a short questionnaire about their perception of parking space available on campus. The participants were first asked demographic information:

a) Gender   
b) Class level (freshman, sophomore, junior, senior)

Next, participants were asked to indicate their opinion on three questions on parking lot space a 1-7 likert-type scale (1 = strongly disagree, 7 = strongly agree).

a) I always drive to campus.  
b) I spend several minutes each day looking for parking.  
c) I think MSU should build additional parking.

#set working directory  
setwd("C:/Users/scyth/OneDrive/Documents/HarrisburgU/AN\_500\_52\_MK/Lab")  
#load the csv file  
raw = read.csv("Lab3.csv")  
library(mice)

## Warning: package 'mice' was built under R version 3.4.4

## Loading required package: lattice

## Warning: package 'lattice' was built under R version 3.4.4

##   
## Attaching package: 'mice'

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

# 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.

#check if there are datapoints which are inaccurate as per the data definition  
summary(raw)

## Gender Class Q1 Q2   
## : 17 Min. : 1.000 Min. :0.500 Min. :1.000   
## 0 : 2 1st Qu.: 2.000 1st Qu.:2.000 1st Qu.:5.000   
## 3 : 3 Median : 3.000 Median :4.000 Median :6.000   
## Female :153 Mean : 3.016 Mean :4.032 Mean :5.931   
## FemaleMale: 2 3rd Qu.: 4.000 3rd Qu.:6.000 3rd Qu.:7.000   
## Male :171 Max. :44.000 Max. :8.000 Max. :8.000   
## MaleFemale: 2 NA's :17 NA's :8 NA's :8   
## Q3   
## Min. :1.000   
## 1st Qu.:4.000   
## Median :5.000   
## Mean :5.359   
## 3rd Qu.:6.000   
## Max. :8.000   
## NA's :6

#gender cannot be malefemale or femalemale  
table(raw$Gender)

##   
## 0 3 Female FemaleMale Male   
## 17 2 3 153 2 171   
## MaleFemale   
## 2

#the class level cannot be decimal  
table(raw$Class)

##   
## 1 1.3 1.4 2 3 4 4.5 13 22 24 43 44   
## 67 1 1 89 88 80 1 1 2 1 1 1

#use new dataframe name  
notypos = raw  
  
#set correct factors for Gender  
notypos$Gender = factor(notypos$Gender,   
 levels = c("Female", "Male"),   
 labels = c("Female", "Male"))  
  
#set correct factors for grade and change the number to actual class levels  
notypos$Class = factor(notypos$Class,   
 levels = c(1,2,3,4),  
 labels = c("freshman", "sophomore", "junior", "senior"))  
  
#output of the fixed categorical data  
table(notypos$Class)

##   
## freshman sophomore junior senior   
## 67 89 88 80

table(notypos$Gender)

##   
## Female Male   
## 153 171

#fixing other numeric colums - Highest value can be 7  
#anything more than 7 is invalid datapoint  
notypos[ , 3:5][ notypos[ , 3:5] > 7 ] = NA  
summary(notypos)

## Gender Class Q1 Q2   
## Female:153 freshman :67 Min. :0.500 Min. :1.000   
## Male :171 sophomore:89 1st Qu.:2.000 1st Qu.:5.000   
## NA's : 26 junior :88 Median :4.000 Median :6.000   
## senior :80 Mean :4.021 Mean :5.921   
## NA's :26 3rd Qu.:6.000 3rd Qu.:7.000   
## Max. :7.000 Max. :7.000   
## NA's :9 NA's :10   
## Q3   
## Min. :1.000   
## 1st Qu.:4.000   
## Median :5.000   
## Mean :5.351   
## 3rd Qu.:6.000   
## Max. :7.000   
## NA's :7

## 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).

#function to calculate how much data is missing on each row  
percentmiss = function(x){ sum(is.na(x))/length(x) \*100 }  
  
#storing the result in a different vector  
missing = apply(notypos, 1, percentmiss)  
table(missing)

## missing  
## 0 20 40 60 80   
## 293 47 4 1 5

#separating the data based on how much data is missing  
#removepeople dataframe will not be used in analysis  
removepeople = subset(notypos, missing>20)  
replacepeople = subset(notypos, missing<40)  
  
#we want to fix the missing values in the replacepeople dataframe  
#columns where missing data is to be filled  
replacecolumn = replacepeople[ , -c(1,2)]  
  
#columns which should not be fixed  
dontcolumn = replacepeople[ , c(1,2)]  
  
#using mice lib to generate values  
#using complete to return a dataframe with the missing values filled  
tempnomiss = mice(replacecolumn)

##   
## iter imp variable  
## 1 1 Q1 Q2 Q3  
## 1 2 Q1 Q2 Q3  
## 1 3 Q1 Q2 Q3  
## 1 4 Q1 Q2 Q3  
## 1 5 Q1 Q2 Q3  
## 2 1 Q1 Q2 Q3  
## 2 2 Q1 Q2 Q3  
## 2 3 Q1 Q2 Q3  
## 2 4 Q1 Q2 Q3  
## 2 5 Q1 Q2 Q3  
## 3 1 Q1 Q2 Q3  
## 3 2 Q1 Q2 Q3  
## 3 3 Q1 Q2 Q3  
## 3 4 Q1 Q2 Q3  
## 3 5 Q1 Q2 Q3  
## 4 1 Q1 Q2 Q3  
## 4 2 Q1 Q2 Q3  
## 4 3 Q1 Q2 Q3  
## 4 4 Q1 Q2 Q3  
## 4 5 Q1 Q2 Q3  
## 5 1 Q1 Q2 Q3  
## 5 2 Q1 Q2 Q3  
## 5 3 Q1 Q2 Q3  
## 5 4 Q1 Q2 Q3  
## 5 5 Q1 Q2 Q3

nomiss = complete(tempnomiss, 1)  
  
#combine all columns  
allcolumns = cbind(dontcolumn, nomiss)  
summary(allcolumns)

## Gender Class Q1 Q2   
## Female:151 freshman :67 Min. :0.500 Min. :1.000   
## Male :170 sophomore:89 1st Qu.:2.000 1st Qu.:5.000   
## NA's : 19 junior :86 Median :4.000 Median :6.000   
## senior :80 Mean :4.021 Mean :5.912   
## NA's :18 3rd Qu.:6.000 3rd Qu.:7.000   
## Max. :7.000 Max. :7.000   
## Q3   
## Min. :1.000   
## 1st Qu.:4.000   
## Median :5.000   
## Mean :5.366   
## 3rd Qu.:6.000   
## Max. :7.000

#save it as dataframe with more recognizable name  
nomissing = allcolumns

## 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?  
d) How many outliers did you have?  
e) Delete all outliers.

#calculate mahal score  
mahal = mahalanobis(nomissing[ , -c(1,2)],   
 colMeans(nomissing[ , -c(1,2)], na.rm = TRUE),  
 cov(nomissing[ , -c(1,2)], use="pairwise.complete.obs"))  
  
#calculate cutoff score  
cutoff = qchisq(1 - .001,ncol(nomissing[ , -c(1,2)]))   
  
#DF of the mahalonbis cutoff  
ncol(nomissing[ , -c(1,2)])

## [1] 3

#summary of mahal scores greater than cutoff - There are 6 outliers  
#False is bad  
summary(mahal < cutoff)

## Mode FALSE TRUE   
## logical 6 334

#saving the data in a new dataframe with outliers removed  
noout = subset(nomissing, mahal < cutoff)

# Assumptions:

## Additivity:

a) Include the symnum bivariate correlation table of your continuous measures.  
b) Do you meet the assumption for additivity?

correlations = cor(noout[,-c(1,2)], use="pairwise.complete.obs")  
#symnum bivariate correlation table  
symnum(correlations)

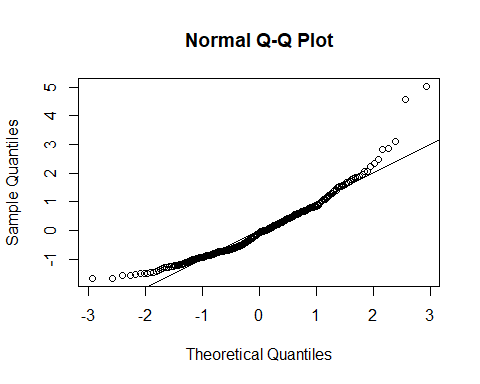
## Q1 Q2 Q3  
## Q1 1   
## Q2 1   
## Q3 1   
## attr(,"legend")  
## [1] 0 ' ' 0.3 '.' 0.6 ',' 0.8 '+' 0.9 '\*' 0.95 'B' 1

#correlation table does not have any +,\* or B hence we meet assumption

## Linearity:

a) Include a picture that shows how you might assess multivariate linearity.  
b) Do you think you've met the assumption for linearity?

random = rchisq(nrow(noout), 7)  
fake = lm(random~., data=noout)  
  
standardized = rstudent(fake)  
qqnorm(standardized)  
abline(0,1)



## Normality:

a) Include a picture that shows how you might assess multivariate normality.  
b) Do you think you've met the assumption for normality?

library(moments)

## Warning: package 'moments' was built under R version 3.4.4

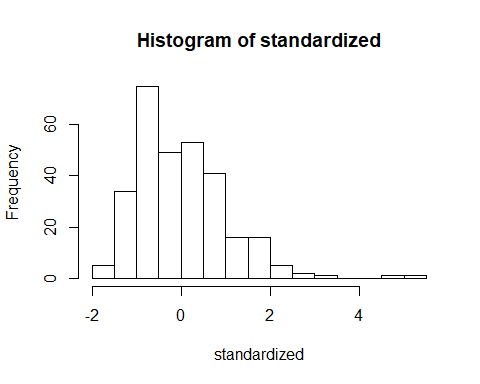
##get the values  
skewness(noout[ , -c(1,2)], na.rm=TRUE)

## Q1 Q2 Q3   
## -0.10731014 -0.11725051 -0.09573304

kurtosis(noout[ , -c(1,2)], na.rm=TRUE)

## Q1 Q2 Q3   
## 1.868345 2.011863 2.182097

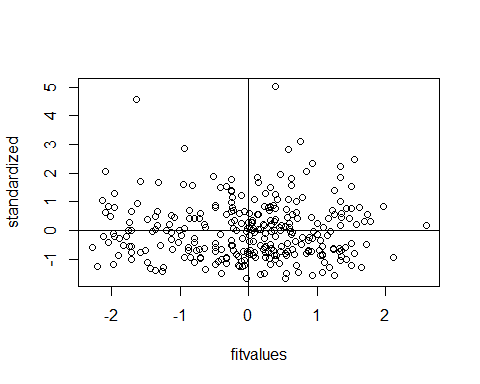
hist(standardized, breaks=15)



## 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?  
c) Do you think you've met the assumption for homoscedasticity?

fitvalues = scale(fake$fitted.values)  
plot(fitvalues, standardized)   
abline(0,0)  
abline(v = 0)



#the datapoints are distributed uniformly and do not form any shapes or biases