Problem Set 1

QTM 200: Applied Regression Analysis

January 28, 2020

Instructions

- Please show your work! You may lose points by simply writing in the answer. If the problem requires you to execute commands in `R`, please include the code you used to get your answers. Please also include the `.R` file that contains your code. If you are not sure if work needs to be shown for a particular problem, please ask.
- Your homework should be submitted electronically on the course GitHub page in `.pdf` form.
- This problem set is due at the beginning of class on Wednesday, January 22, 2020. No late assignments will be accepted.
- Total available points for this homework is 100.

Question 1 (25 points)

A private school counselor was curious about the average of IQ of the students in her school and took a random sample of 25 students' IQ scores. The following is the data set:

```
y <- c(105, 69, 86, 100, 82, 111, 104, 110, 87, 108, 87, 90, 94, 113, 112, 98,
    80, 97, 95, 111, 114, 89, 95, 126, 98)
```

Find a 90% confidence interval for the student IQ in the school assuming the population of IQ from which our random sample has been selected is normally distributed.

- ANSWER: The 90% confidence interval is (94.1, 102.7).
- INTERPRETATION: We are 90% confident that that true average IQ of students at her school lies between 94.1 and 102.7.

Question 2 (25 points)

A private school counselor was curious whether the average of IQ of the students in her school is higher than the average IQ score 100 among all the schools in the country. She took a random sample of 25 students' IQ scores. The following is the data set:

```
y <- c(105, 69, 86, 100, 82, 111, 104, 110, 87, 108, 87, 90, 94, 113, 112, 98,
    80, 97, 95, 111, 114, 89, 95, 126, 98)
```

Conduct a test with 0.05 significance level assuming the population of IQ from which our random sample has been selected is normally distributed.

Question 3 (50 points)

Assume y is variable with values 1,2,3,4 standing for "Freshman", "Sophomore", "Junior", and "Senior", convert y from numbers to characters in `R`:

```
y <- c(1, 2, 1, 3, 4, 1, 1, 4, 2, 1, 3, 4, 3, 2, 1, 3, 4, 1, 2, 3, 1, 1, 2, 1,
    1, 3, 4)
```

Researchers are curious about what affects the education expenditure on public education. The following is availabe variables in a data set about the education expenditure.

`State`	*50 states in US*
`Y`	*per capita expenditure on public education*
`X1`	*per capita personal income*
`X2`	*Number of residents per thousand under 18 years of age*
`X3`	*Number of people per thousand residing in urban areas*
`Region`	*1=Northeast, 2= North Central, 3= South, 4=West*

Explore the `expenditure` data set and import data into `R`.

```
expenditure <- read.table("expenditure.txt", header=T)
```

- Please plot the relationships among *Y*, *X1*, *X2*, and *X3*? What are the correlations among them? Describe the graph and the relationships among them.

- Please plot the relationship between *Y* and *Region*? On average, which region does have the highest per capita expenditure on public education?

- Please plot the relationship between Y and $X1$? Describe this graph and the relationship. Reproduce the above graph including one more variable *Region* and display different regions with different types of symbols and colors.

- R code:

```
#####################
# load libraries
# set wd
# clear global .envir
#####################

# remove objects
rm(list=ls())
# detach all libraries
detachAllPackages <- function() {
  basic.packages <- c("package:stats", "package:graphics", "package:
    grDevices", "package:utils", "package:datasets", "package:methods", "
    package:base")
  package.list <- search()[ifelse(unlist(gregexpr("package:", search()))
    ==1, TRUE, FALSE)]
  package.list <- setdiff(package.list , basic.packages)
  if (length(package.list)>0)  for (package in package.list) detach(
    package,  character.only=TRUE)
}
detachAllPackages()

# load libraries
pkgTest <- function(pkg){
  new.pkg <- pkg[!(pkg %in% installed.packages()[,  "Package"])]
  if (length(new.pkg))
    install.packages(new.pkg,  dependencies = TRUE)
  sapply(pkg,  require,  character.only = TRUE)
}

# here is where you load any necessary packages
# ex: stringr
# lapply(c("stringr"),  pkgTest)

lapply(c(),  pkgTest)

# set working directory
setwd("~/GitHub/QTM200Spring2020/problem_sets/PS1")


#####################
# Problem 1
#####################

```

```r
y <- c(105, 69, 86, 100, 82, 111, 104, 110, 87, 108, 87, 90, 94, 113, 112, 98, 80, 97, 95, 111, 114, 89, 95, 126, 98)
z90 <- qnorm((1-.90)/2, lower.tail = FALSE)
n <- length(y)
mean(y)
sample_mean <- mean(y)
sample_sd <- sd(y)
lower_90 <- sample_mean - (z90 * (sample_sd/sqrt(n)))
upper_90 <- sample_mean + (z90 * (sample_sd/sqrt(n)))
confint90  <- c(lower_90, upper_90)
# ANSWER: 90% confidence interval is (94.1, 102.7)

######################
# Problem 2
######################

y <- c(105, 69, 86, 100, 82, 111, 104, 110, 87, 108, 87, 90, 94, 113, 112, 98, 80, 97, 95, 111, 114, 89, 95, 126, 98)
#null hypothesis: xbar = 100
#alternative hypothesis: xbar > 100
#calculate test statistic and p-value
mean(y)
#mean of sample is 98.44
sd(y)
#standard deviation of sample is 13.09
ts <- ((100-98.44)/(13.09))
#sample size
n <- length(y)
# n=25 therefore df = 24
#test statistic = -0.119
pt(abs(0.119), df=24, lower.tail=F)
# ANSWER: p-value = 0.453 > 0.05 therefore sample mean is not significantly greater than 100

######################
# Problem 3
######################

y <- c(1, 2, 1, 3, 4, 1, 1, 4, 2, 1, 3, 4, 3, 2, 1, 3, 4, 1, 2, 3, 1, 1, 2, 1, 1, 3, 4)

expenditure <- read.table("expenditure.txt", header=T)

# PLOT 1
plot(expenditure$X1, expenditure$Y,
     xlab="Per capita personal income", ylab="Per capita expenditure on public education")
abline(lm(expenditure$Y ~ expenditure$X1))
#Moderate positive correlation

# PLOT 2
```

```r
plot(expenditure$X2, expenditure$Y,
     xlab="Number of residents per thousand under 18 years of age", ylab="Per capita expenditure on public education")
abline(lm(expenditure$Y ~ expenditure$X2))
#Weak negative correlation

# PLOT 3
plot(expenditure$X3, expenditure$Y,
     xlab="Number of people per thousand residing in urban areas", ylab="Per capita expenditure on public education")
abline(lm(expenditure$Y ~ expenditure$X3))
#Weak positive correlation


# PLOT 4

expenditure$fourregions <- NA
expenditure$fourregions <- factor(NA, levels=c("Northeast", "North Central", "South", "West"))
expenditure$fourregions[expenditure$Region==1] <- "Northeast"
expenditure$fourregions[expenditure$Region==2] <- "North Central"
expenditure$fourregions[expenditure$Region==3] <- "South"
expenditure$fourregions[expenditure$Region==4] <- "West"
boxplot(expenditure$Y ~ expenditure$fourregions,
        xlab= "Region",
        ylab= "Per capita expenditure on public education")

# On average, the Western region has the highest per capita expenditure on public education

# PLOT 5

plot(expenditure$X1, expenditure$Y,
     xlab="Per capita personal income", ylab="Per capita expenditure on public education")
abline(lm(expenditure$Y ~ expenditure$X1))
```