Problem Set 1

QTM 200: Applied Regression Analysis

January 29, 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.

```r
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)
# 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:

```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)
```

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

ANSWER: p-value = 0.278 greater than 0.05, therefore sample mean is not significantly greater than 100

```r
#null hypothesis: xbar = 100
#alternative hypothesis: xbar > 100
#calculate test statistic and p-value
mean(y)
#mean of sample is 98.44
sd(y)
#sample size
n <- length(y)
#standard deviation of sample is 13.09
teststata <- (98.44-100)
teststatb <- 13.09/(sqrt(n))
teststatistic <- ((teststata)/(teststatb))
#test statistic = -0.596
# n=25 therefore df = 24
#test statistic = -0.119
1-pt(abs(-0.596), df=24)
# p-value = 0.278
# ANSWER/INTERPRETATION: p-value = 0.278 > 0.05 therefore sample mean is not
    significantly greater than 100
```

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)
```

ANSWER:

```
highschool[y==1] <- "Freshman"
highschool[y==2] <- "Sophomore"
highschool[y==3] <- "Junior"
highschool[y==4] <- "Senior"
str(highschool)
highschool[1:10]
highschool
#Recode numerical vector y so that values are assigned to corresponding
    categories

#####
```

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)
```

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

```

Figure 1: Y and X1: Moderate, linear, positive correlation

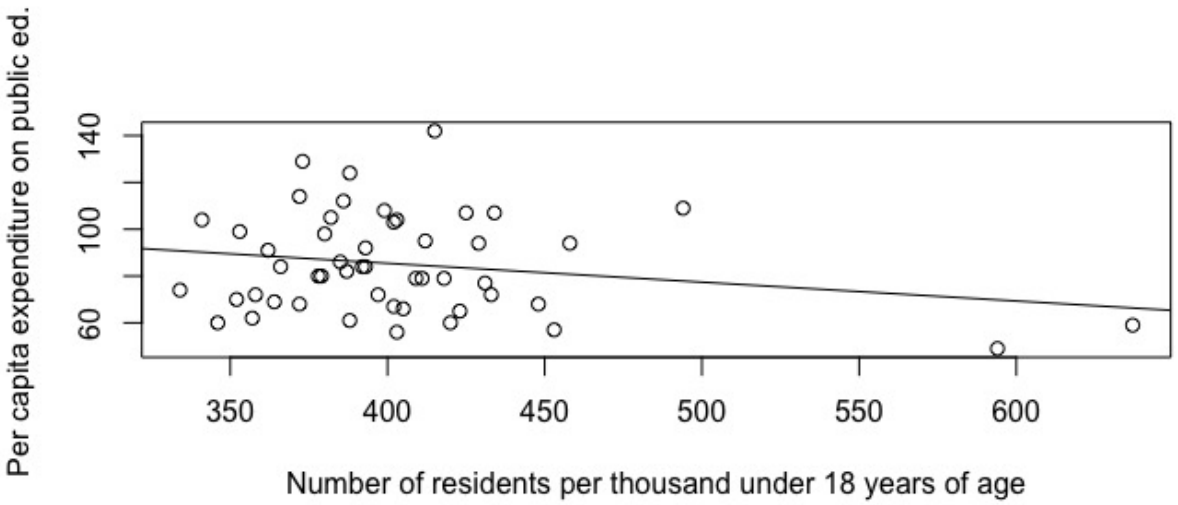

Figure 2: Y and X2: Weak, negative, non-linear correlation

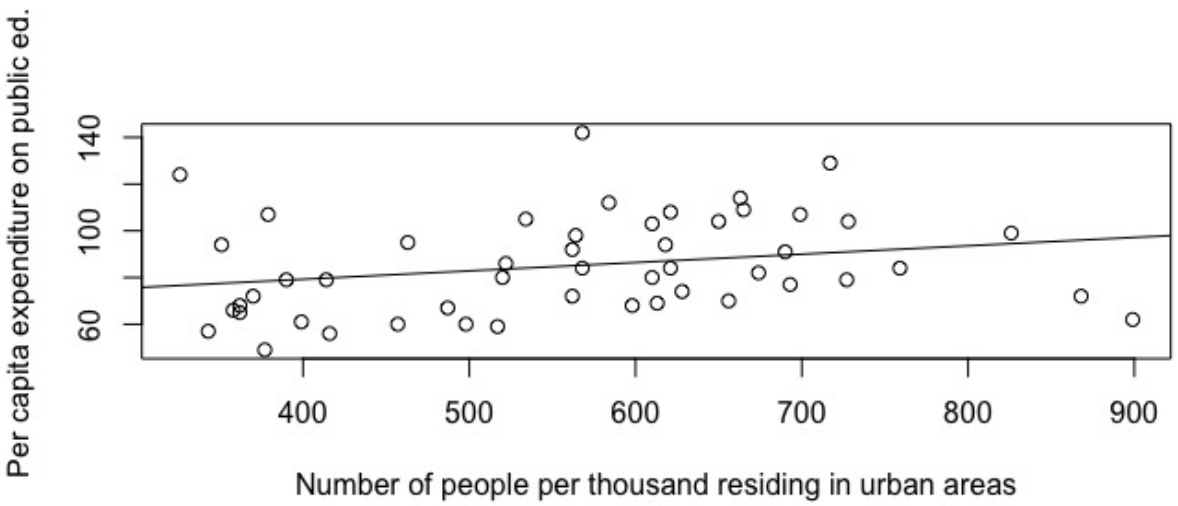

Figure 3: Y and X3: Weak positive correlation

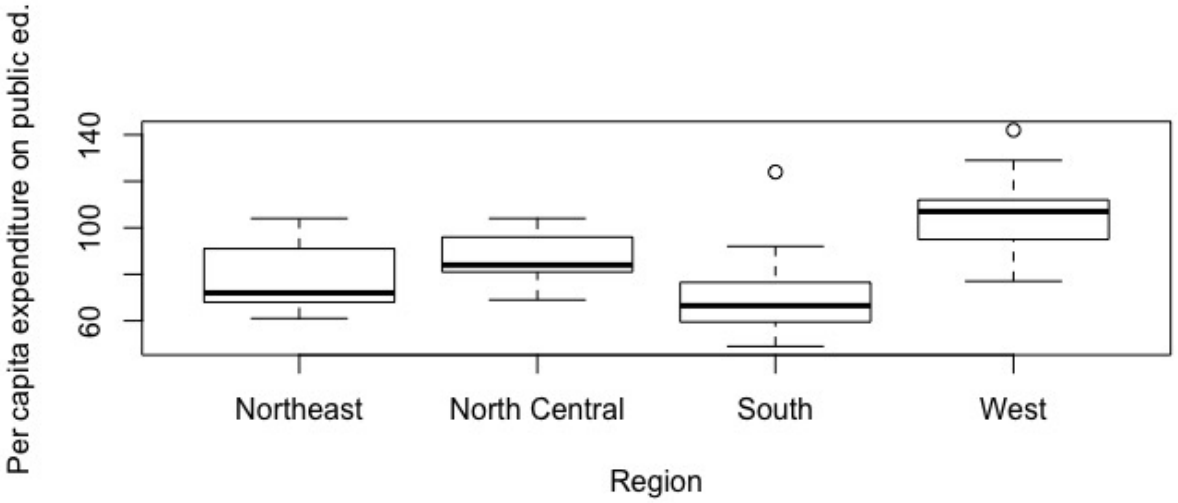

Figure 4: On average, the Western region has the highest per capita expenditure on public education

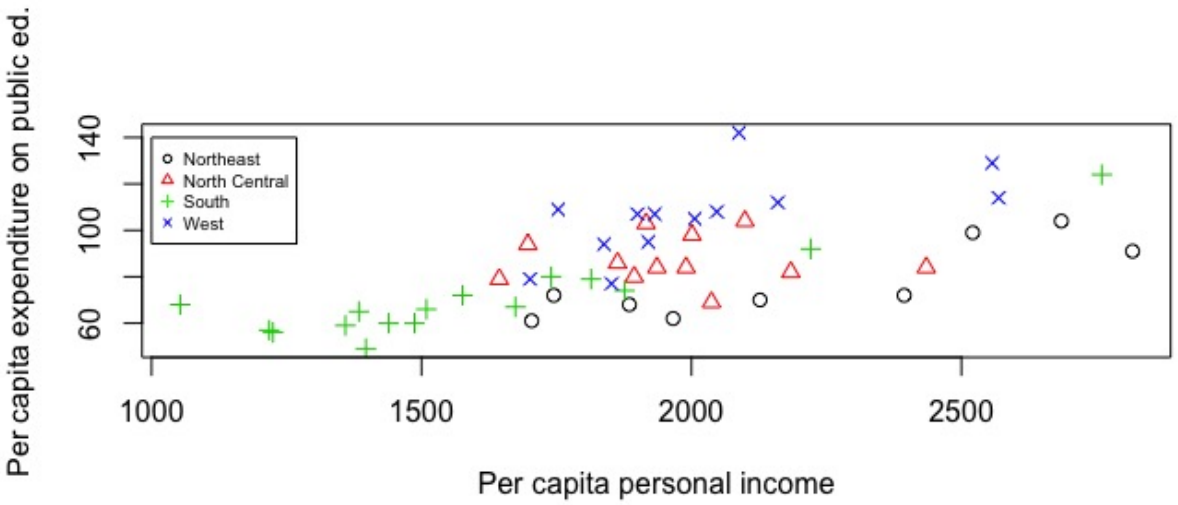

Figure 5: Moderate, linear, positive correlation between per capita expenditure on public education and per capita personal income

```r
# PLOT 2
plot(expenditure$X2, expenditure$Y,
     xlab="Number of residents per thousand under 18 years of age", ylab="Per capita expenditure on public ed.")
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 ed.")
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
```

```
32
33 # PLOT 5
34 plot(expenditure$X1, expenditure$Y,
35      xlab="Per capita personal income", ylab="Per capita expenditure on
       public education")
36 abline(lm(expenditure$Y ~ expenditure$X1))
37
38 # PLOT 6
39 dev.off()
40 plot(expenditure$X1, expenditure$Y,
41              xlab="Per capita personal income", ylab="Per capita
       expenditure on public education", col=expenditure$Region, pch=
       expenditure$Region)
42 legend(x=1000, y=140, c("Northeast","North Central","South","West"), cex
       =0.6, col=c("black","red","green","blue"),pch=c(1:4))
```