Problem Set 1

QTM 200: Applied Regression Analysis

Due: 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.

```
#Q1
#calculate the mean, sd and upper lower boundary
z90 <- qt((1-.90)/2, df=24, lower.tail = FALSE)
n<-length(na.omit(y))
```

```r
sample_mean<-mean(y , na.rm = TRUE)
sample_sd<-sd(y , na.rm = TRUE)
lower_90<-sample_mean-(z90 * (sample_sd/sqrt(n)))
upper_90<-sample_mean+(z90 * (sample_sd/sqrt(n)))
confint90<-c(lower_90, upper_90)
print(confint90)
#we get the result of a 90% confidence interval[1]  93.95993 102.92007
```

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.

```r
#Q2
#conduct one-sample t-test to see if true mean is greater than 100
t.test(y, mu = 100, alternative = "greater")
#One Sample t-test
#data:  y
#t = -0.59574, df = 24, p-value = 0.7215
#alternative hypothesis: true mean is greater than 100
#95 percent confidence interval:
#  93.95993      Inf
#sample estimates:
#  mean of x
#     98.44
#since p-value > 0.05 we fail to reject the null hypothesis thus we cannot
   conclude that her school mean IQ is higher than the country's.
```

Question 3 (50 points)

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 (you just need to describe the graph and the relationships among them)?

```
#Q3
#import data
expenditure<-read.table("expenditure.txt" , header=T)
#1)
plot(expenditure$X1,expenditure$Y, main="Personal income vs expenditure
    on public education",xlab="per capita personal income",ylab="per
    capita expenditure on public education")
#Figure 1 shows a positive linear relationship
plot(expenditure$X2,expenditure$Y, main="Residents under 18 vs
    expenditure on public education",xlab="Number of residents per
    thousand under 18 years of age",ylab="per capita expenditure on public
     education")
#Figure 2 shows a negative relationship
plot(expenditure$X3,expenditure$Y, main="People in urban areas vs
    expenditure on public education",xlab="Number of people per thousand
    residing in urban areas",ylab="per capita expenditure on public
    education")
#Figure 3 shows a almost flat positive linear relationship
```

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

```
#2) for relationship between Y and Region
expenditure$R<-expenditure$Y
expenditure$R<-factor(NA, level=c("Northeast", "North Central", "South",
    "West"))
#rename region number to actual type
expenditure$R[expenditure$Region==1]<-"Northeast"
expenditure$R[expenditure$Region==2]<-"North Central"
```

```
expenditure$R[expenditure$Region==3]<-"South"
expenditure$R[expenditure$Region==4]<-"West"
#do a boxplot
boxplot(expenditure$Y~expenditure$R, main="Expenditure on public education vs Region",xlab="Regions",ylab="per capita expenditure on public education")
#we see on Figure 4 on average West has 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.

```
#3)for relationship between Y and X1
plot(expenditure$X1,expenditure$Y, main="personal income vs expenditure on public education",xlab="per capita personal income",ylab="per capita expenditure on public education", col=expenditure$Region, pch=expenditure$Region)
legend(x="topleft", legend = levels(expenditure$R), col = c(1,2,3,4), pch = c(1,2,3,4))
#from Figure 5 we can tell that the 4 regions have a positive linear relationship with West and Central in the middle, Northest with the most income and South the least
```

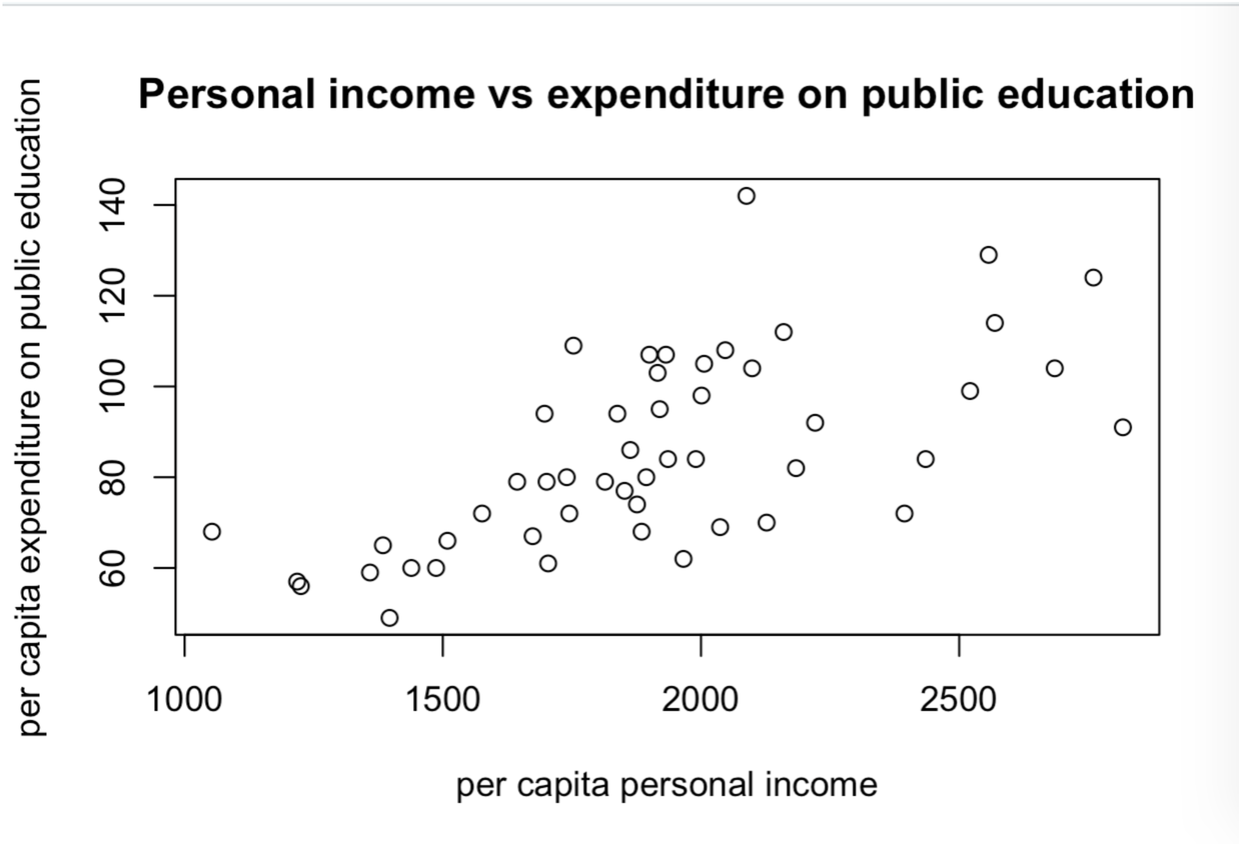

Figure 1: Y and X1.

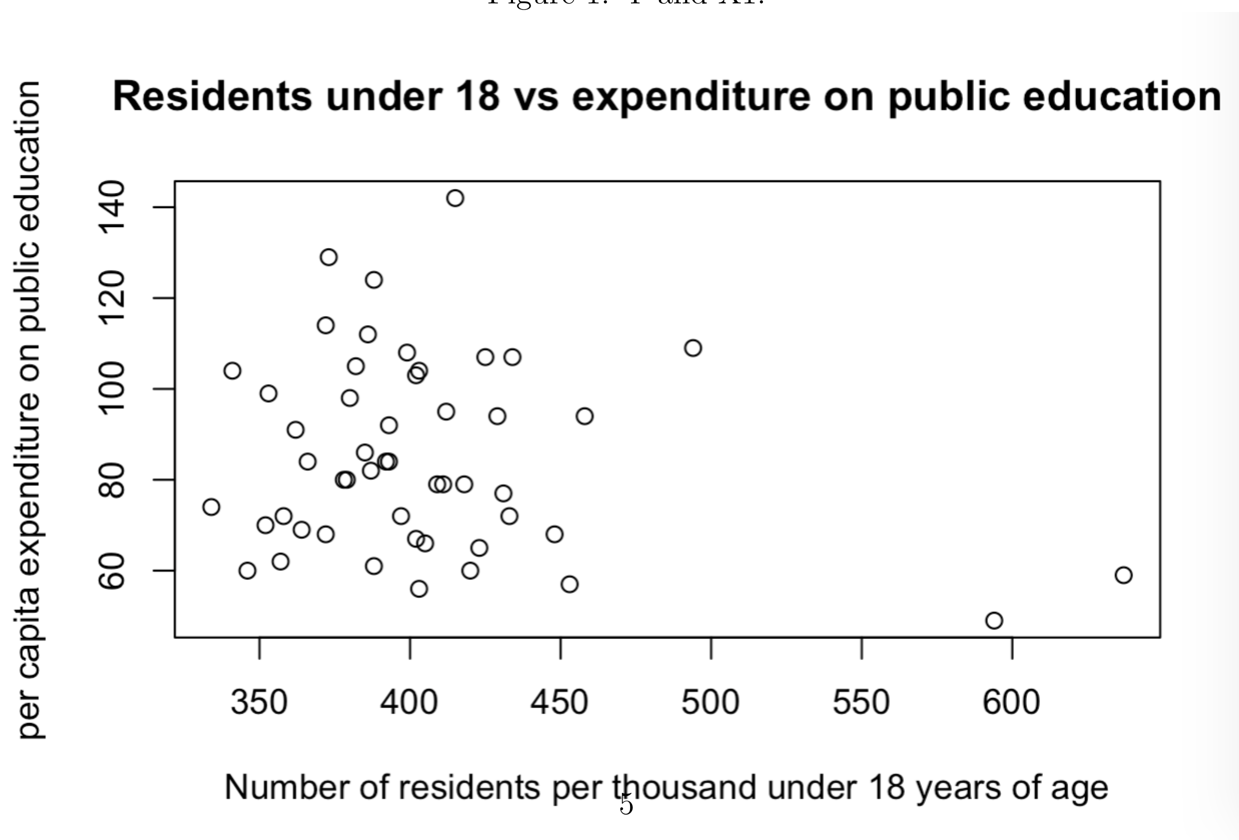

Figure 2: Y and X2.

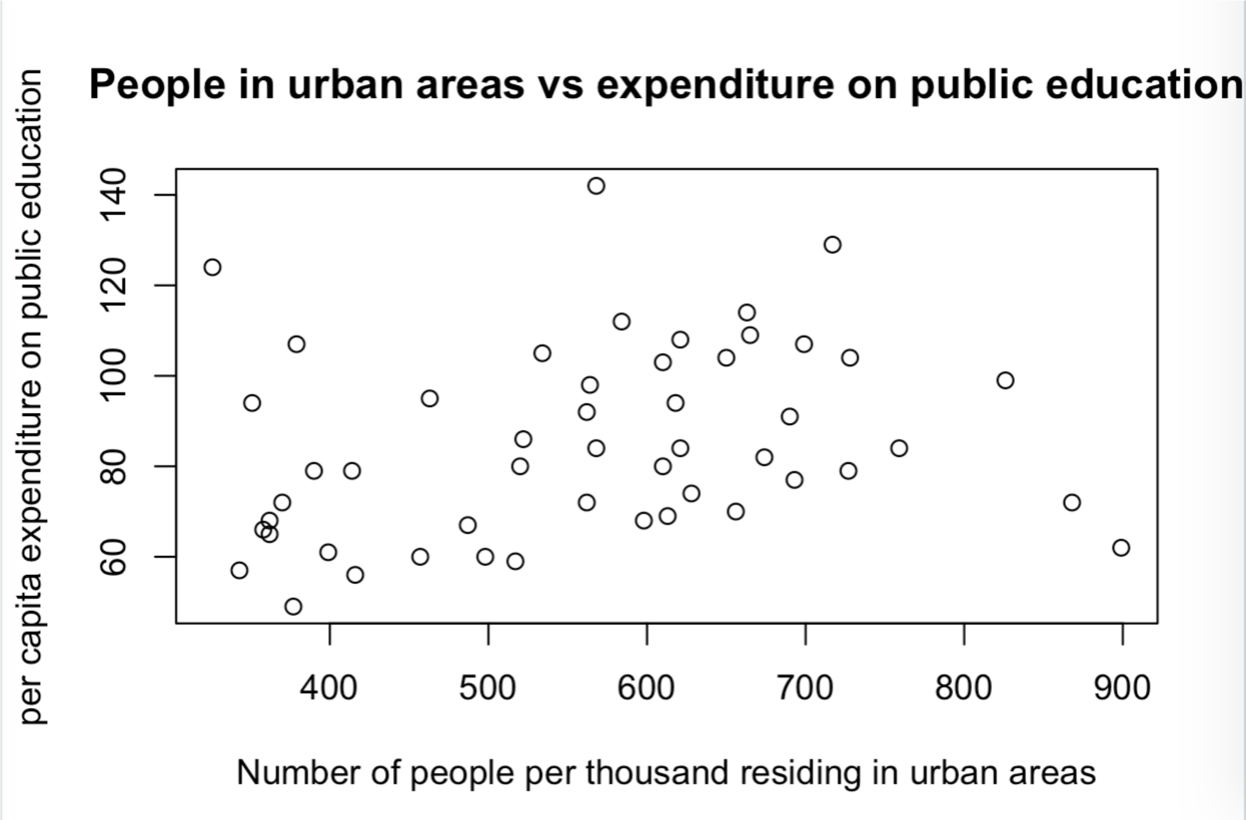

Figure 3: Y and X3.

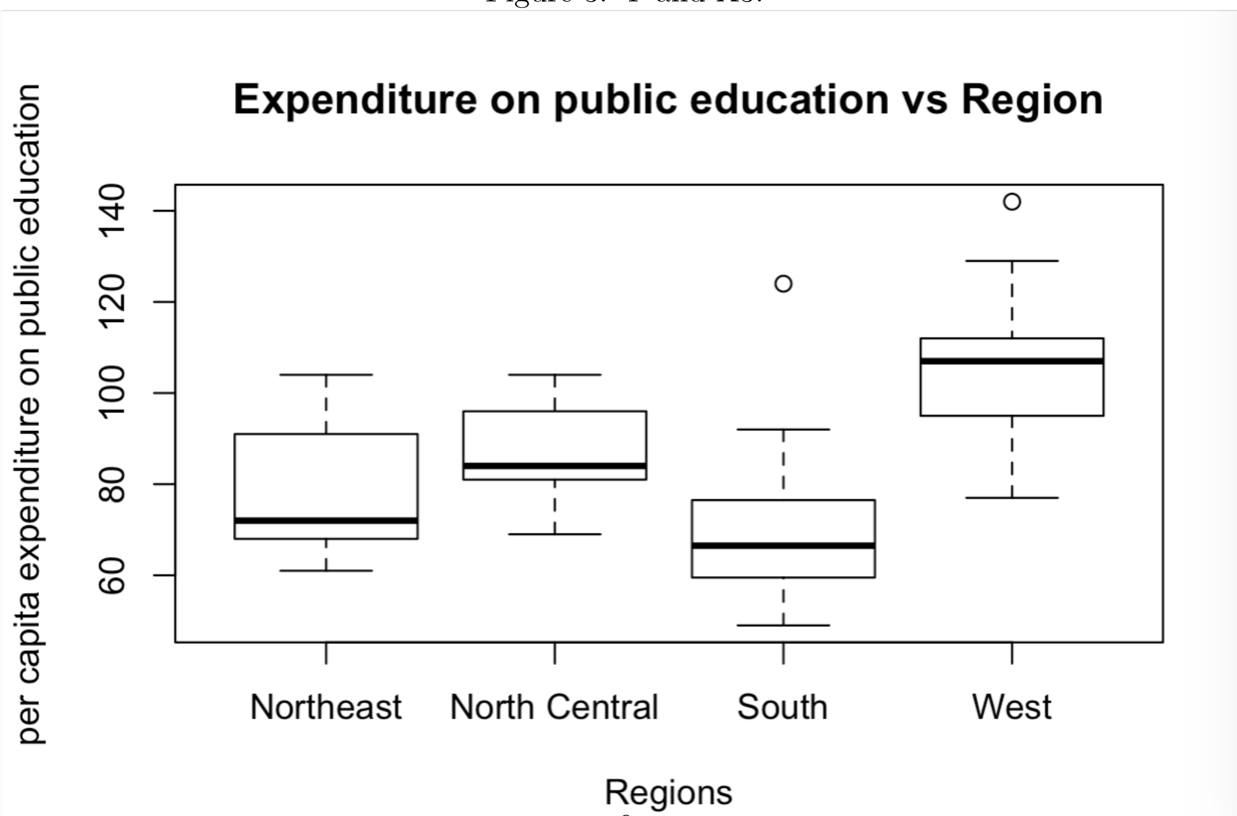

Figure 4: Y and Region.

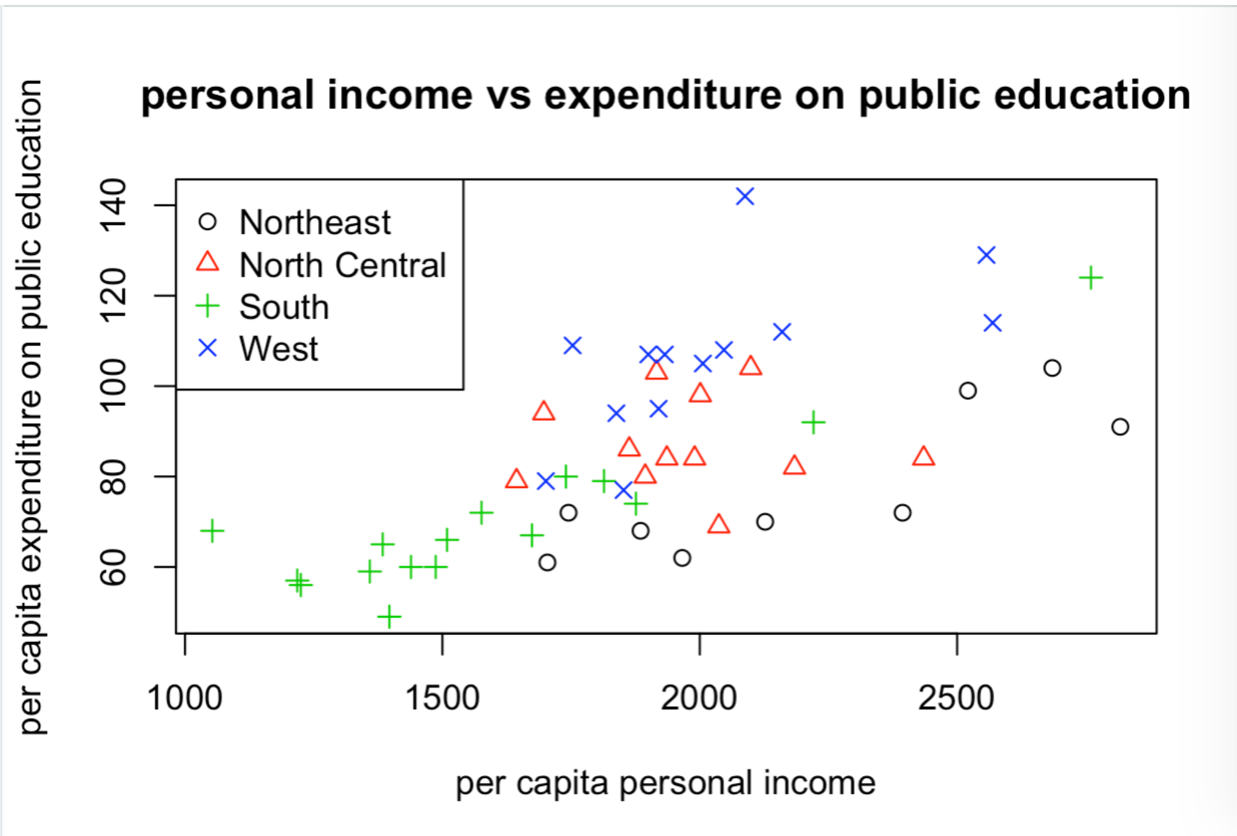

Figure 5: Y and RX1.