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.

```
# Calculate the mean of the sample, y, and save it as sample.mean
mean(y)
sample.mean<-mean(y)
#Calculate the standard deviation of the sample, y, and save it as sample.sd
sd(y)
sample.sd<-sd(y)
```

```r
# Calculate the standard error of the mean by dividing the standard deviation by the square root of the sample size. Save as sample.stderror
sample.sd/(sqrt(25))
sample.stderror<-sample.sd/(sqrt(25))
#Calculate the t statistic for a 90% confidence interval using 25 df. Save as sample.t
qt(1-.1/2,df=25)
sample.t<-qt(1-.1/2,df=25)
# Calculate the confidence interval using the formula mean +/- tvalue * standard error
sample.mean+(sample.t*sample.stderror)
sample.mean-(sample.t*sample.stderror)
# SOLUTION: (93.96711,102.9129). With 90% certainty, the true mean of the students' IQ falls between 93.96711 and 102.9129
```

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
# Ho: there is no difference between the average IQ of students in the counselor's school and the average IQ of all students in the country (mean = 100)
# Ha: The average IQ of students in the counselor's school is higher than the average IQ of all students in the country (mean > 100) *this is a one-sided test
# The significance level is 0.05 (alpha=0.05)
# Calculate the mean of the sample, y, and save it as sample.mean
mean(y)
sample.mean<-mean(y)
#Calculate the standard deviation of the sample, y, and save it as sample.sd
sd(y)
sample.sd<-sd(y)
# Calculate the standard error of the mean by dividing the standard deviation by the square root of the sample size. Save as sample.stderror
sample.sd/(sqrt(25))
sample.stderror<-sample.sd/(sqrt(25))
# Calculate the test statistic by subtracting the population mean from the sample mean and then dividing by the standard error.
(sample.mean-100)/sample.stderror
```

```r
# Calculate the p value given the test statistic of -0.5957439 and 24 degrees of freedom
qt(abs(-0.5957439), df=24)
# SOLUTION: p= 0.2450352, so fail to reject the null hypothesis. There is not enough evidence to reject the claim that the student's at the counselor's school have the same average IQ as the average IQ of all students.
```

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*

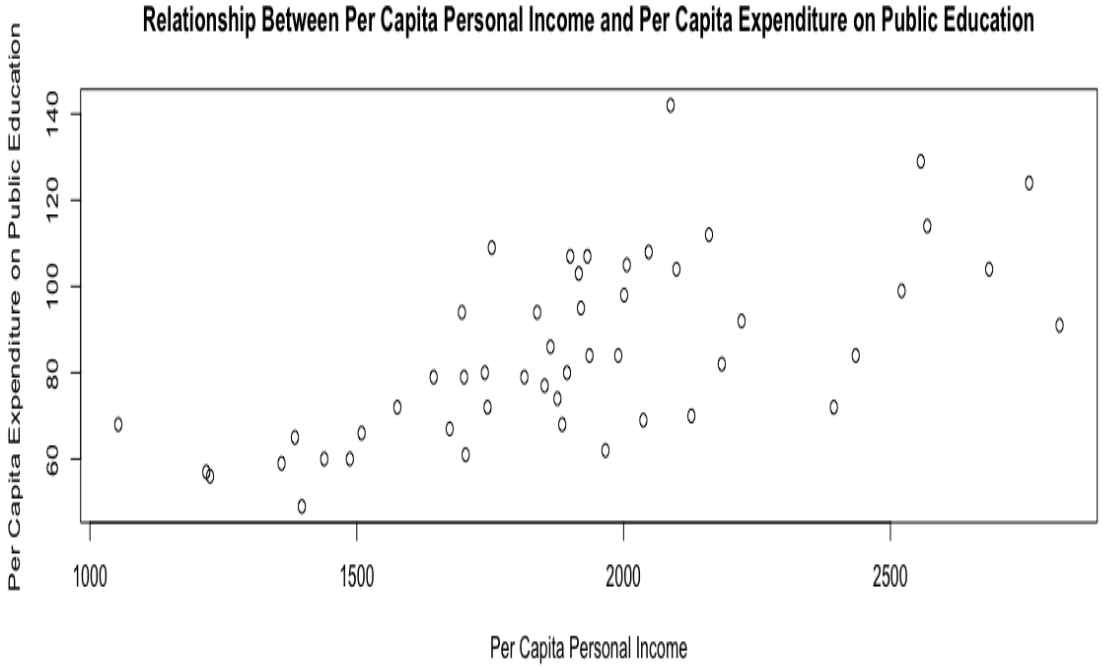
Explore the `expenditure` data set and import data into `R`.

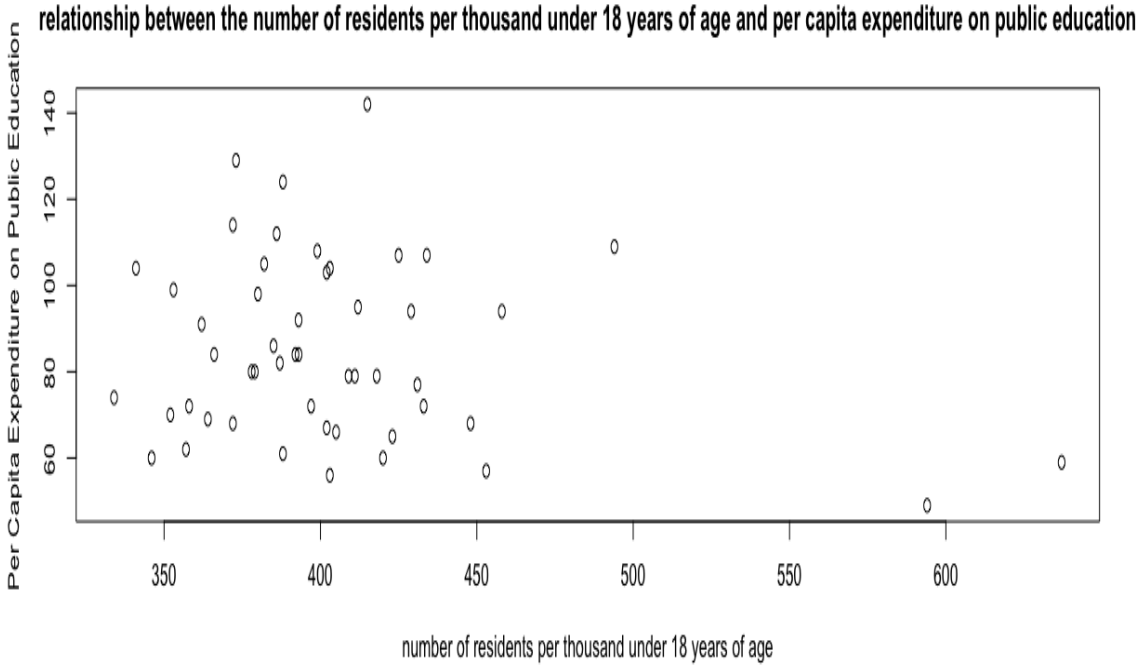
```r
expenditure <- read.table("expenditure.txt", header=T)
```

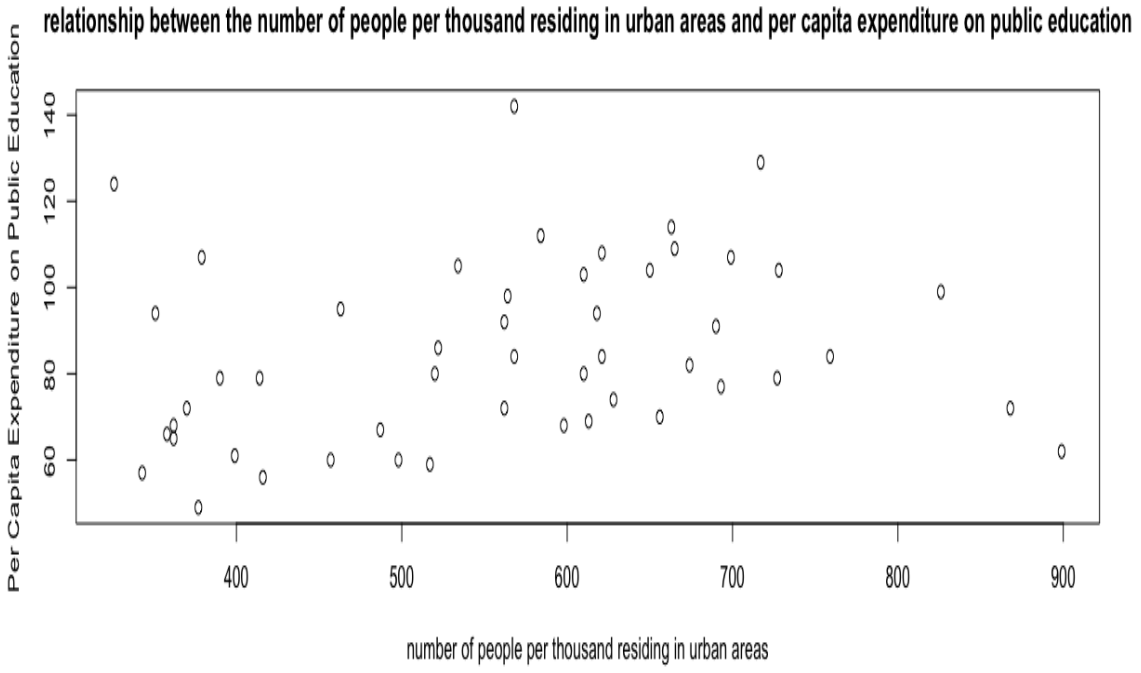
- Please plot the relationships among *Y*, *X1*, *X2*, and *X3*? What are the correlations among them (you just need to describe the graph and the relationships among them)?

```r
# Use scatter plots to determine the relationship between 2 quantitative variables
# Use a scatter plot to determine the relationship between per capita personal income and per capita expenditure on public education
plot(expenditure$X1,expenditure$Y, main = "Relationship Between Per Capita Personal Income and Per Capita Expenditure on Public Education", xlab = "Per Capita Personal Income", ylab = "Per Capita Expenditure on Public Education")
# SOLUTION: There is a moderate, positive, linear correlation between these two variables (X1 and Y)

# Use a scatter plot to determine the relationship between the number of residents per thousand under 18 years of age and per capita expenditure on public education
plot(expenditure$X2,expenditure$Y, main = "relationship between the number of residents per thousand under 18 years of age and per capita expenditure on public education", xlab = "number of residents per
```

```
    thousand under 18 years of age", ylab = "Per Capita Expenditure on
    Public Education")
# SOLUTION: There is no correlation between these two variables (X2 and Y
    )

# Use a scatter plot to determine the relationship between the number of
    people per thousand residing in urban areas and per capita expenditure
     on public education
plot(expenditure$X3, expenditure$Y, main = "relationship between the
    number of people per thousand residing in urban areas and per capita
    expenditure on public education", xlab = "number of people per
    thousand residing in urban areas", ylab = "Per Capita Expenditure on
    Public Education")
#SOLUTION: There is a weak, positive correlation between these two
    variables (X3 and Y)
```



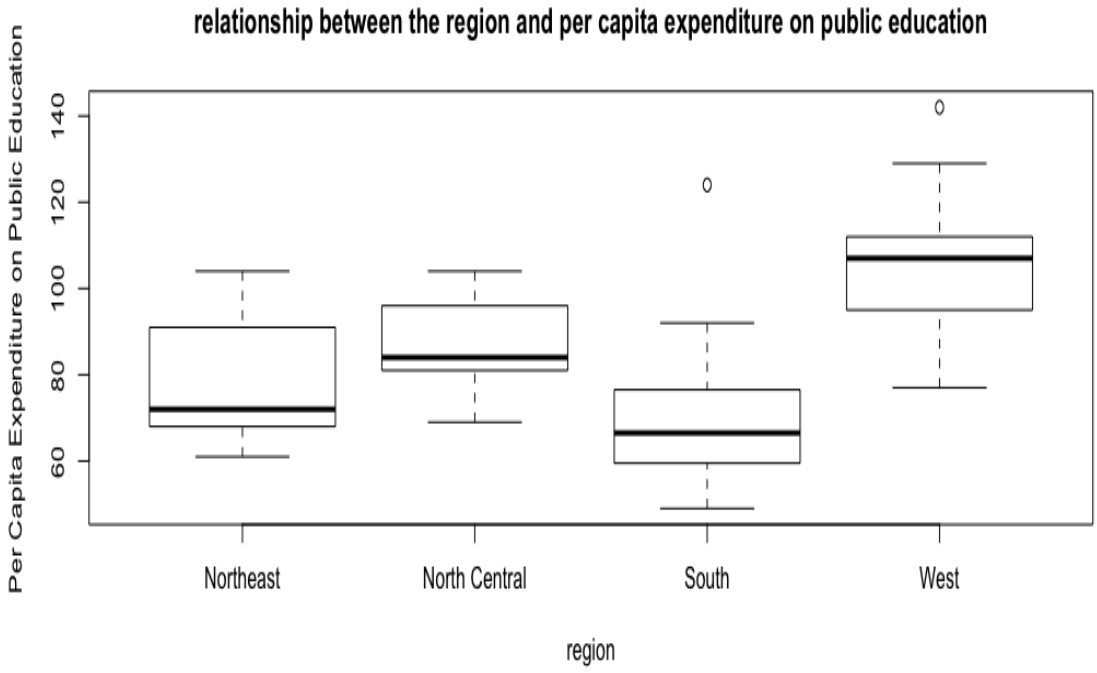




- Please plot the relationship between Y and *Region*? On average, which region has the

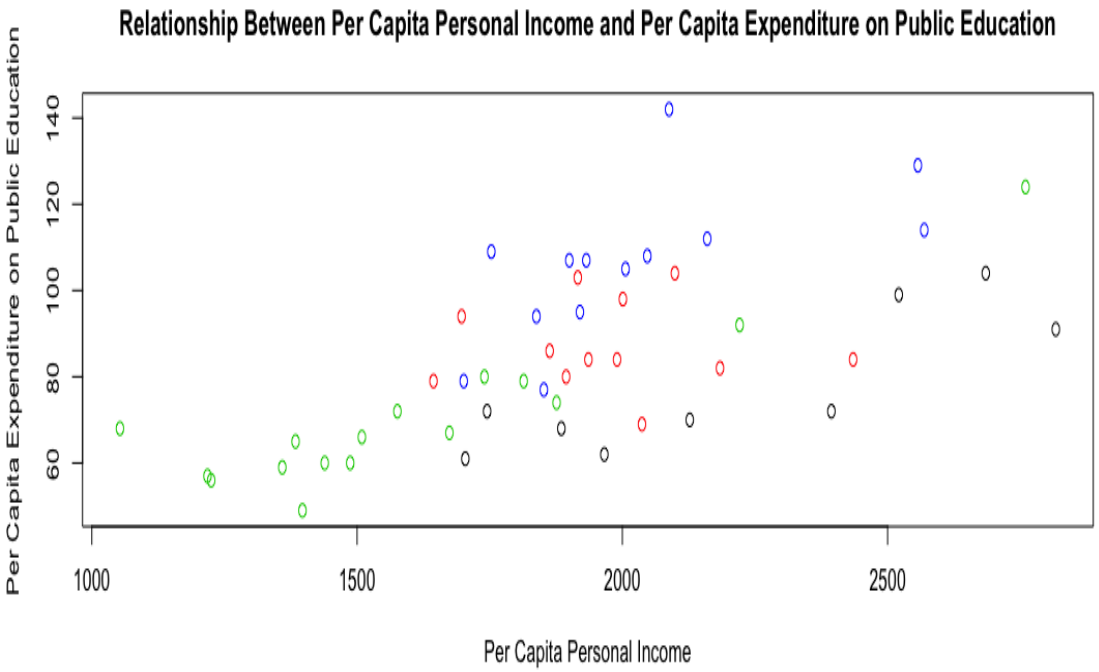
highest per capita expenditure on public education?

```
# Use a boxplot plot to determine the relationship between the region and
    per capita expenditure on public education
# First, recode the region as a factor variable
expenditure$Region_Names<- factor(expenditure$Region, levels = c("
   Northeast","North Central","South","West"))
expenditure$Region_Names[expenditure$Region==1]<- "Northeast"
expenditure$Region_Names[expenditure$Region==2]<- "North Central"
expenditure$Region_Names[expenditure$Region==3]<- "South"
expenditure$Region_Names[expenditure$Region==4]<- "West"
# Next, plot the relationship between Region_Names and Y
plot(expenditure$Region_Names,expenditure$Y, main = "relationship between
    the region and per capita expenditure on public education", xlab = "
   region", ylab = "Per Capita Expenditure on Public Education")
# SOLUTION: On average, the West has the highest per capita expenditure
   on public education

# Use a scatter plot to determine the relationship between per capita
   personal income and per capita expenditure on public education. Color
   code based on region.
plot(expenditure$X1,expenditure$Y, main = "Relationship Between Per
   Capita Personal Income and Per Capita Expenditure on Public Education"
   , xlab = "Per Capita Personal Income", ylab = "Per Capita Expenditure
   on Public Education", col= c(expenditure$Region, pch=c(expenditure$
   Region)))
# Legend: Northeast = black, North Central = red, South = green, West =
   blue
# SOLUTION: The correlation between per capita personal income and per
   capita expenditure on public education is strongest for the South and
   the West. There is weak positive correlation for Northeast, and no
   correlation for North Central.
```



- 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.

```
# Use a scatter plot to determine the relationship between per capita
    personal income and per capita expenditure on public education. Color
    code based on region.
plot(expenditure$X1,expenditure$Y, main = "Relationship Between Per
    Capita Personal Income and Per Capita Expenditure on Public Education"
    , xlab = "Per Capita Personal Income", ylab = "Per Capita Expenditure
    on Public Education", col= c(expenditure$Region, pch=c(expenditure$
    Region)))
# Legend: Northeast = black, North Central = red, South = green, West =
    blue
# SOLUTION: The correlation between per capita personal income and per
    capita expenditure on public education is strongest for the South and
    the West. There is weak positive correlation for Northeast, and no
    correlation for North Central.
```


Relationship Between Per Capita Personal Income and Per Capita Expenditure on Public Education
Per Capita Expenditure on Public Education
60
80
100
120
140
1000
1500
2000
2500
Per Capita Personal Income