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:

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.

```
######################
# Problem 1
######################

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 the sum
sum(y)
#Find the mean
```

```r
sum(y)/length(y)
sample_mean=mean(y)
#Find the sum of errors
demeanedSumSimple = y - mean(y)
demeanedSumSimple
#Find the squared error
squaredError= demeanedSumSimple^2
squaredError
#Find the varience
varience=sum(squaredError)/length(y)
varience
#Find the standard deviation
Sd= sd(y,na.rm=FALSE)
sample_sd=Sd
#Given confidence coefficient = 0.9
z90=qt((1-0.9)/2,df=length(y)-1, lower.tail = FALSE)
n= length(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)
# [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

######################
# 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)

# set Ho:mu < 100
# set Ha:mu >= 100

mu = 100
#Find the sum
```

```r
sum(y)
#Find the mean
mean(y)
#Find the sum of errors
demeanedSumSimple = y - mean(y)
demeanedSumSimple
sum(demeanedSumSimple)
#Find the squared error
squaredError= demeanedSumSimple^2
#Find the varience
varience=sum(squaredError)/length(y)-1
varience

z=(mean(y)-mu)/varience

critical_value = qt(0.05, df =length(y)-1, lower.tail = T )

if(critical_value<z){
   print("there is not enough evidence to disprove Ho, the null hypothesis, which states that students in her school is lower than the average IQ score 100 among all the schools in the country")
}
#There is not enough evidence to disprove Ho, the null hypothesis, which states that students in her school is lower than the average IQ score 100 among all the schools in the country, thus the hypothesis testing suggest the counselor's hypothesis might not be right
```

Question 3 (50 points)

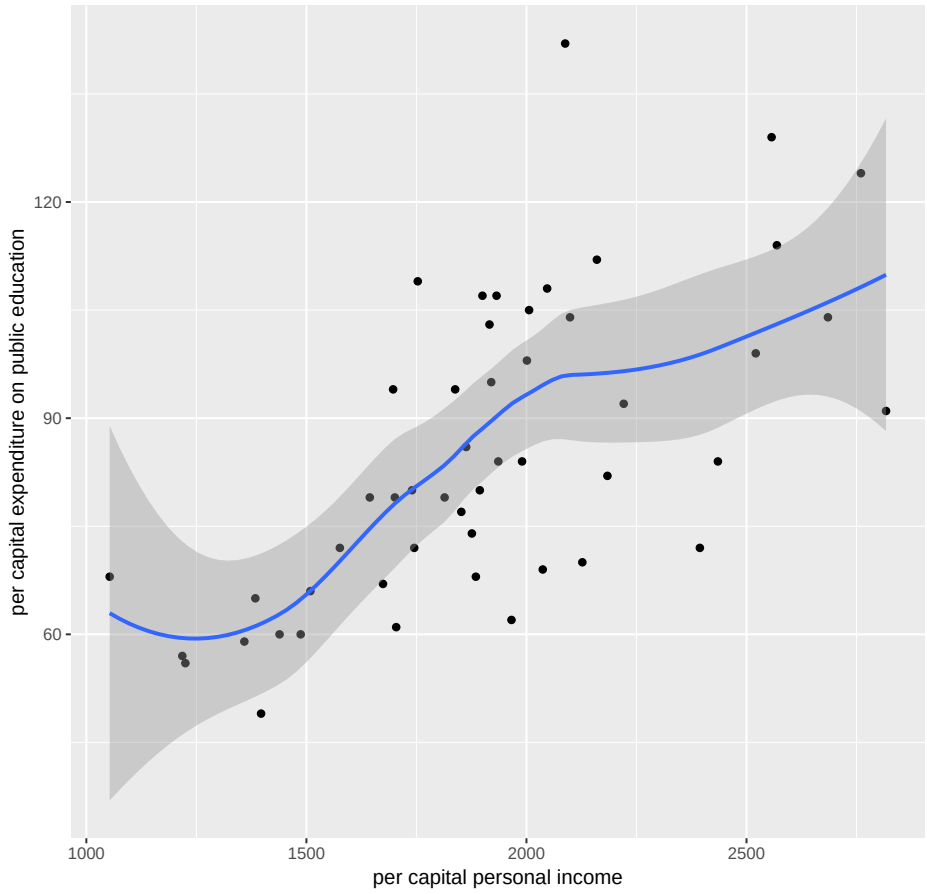
Researchers are curious about what affects the education expenditure on public education. The following is availabe variables in a data set about the education expenditure.

`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)
```

Figure 1: Relationship between per capital personal income and per capita expenditure on public education



- 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)?

```
1
2 #1
3
4 #Y & X1
5 pdf("plot1.pdf")
6 ggplot(expenditure, aes(x=X1, y=Y)) + geom_point() + labs(x="per capital
    personal income", y="per capital expenditure on public education",
    title="")+geom_smooth()
7 dev.off()
8 #Observation: The relationship between X1 and Y is almost linear and the
    line is upwards slowping. With per capital personal income increases,
```

Figure 2: Relationship between number of residents per thousand under 18 years of age and per capital expenditure on public education

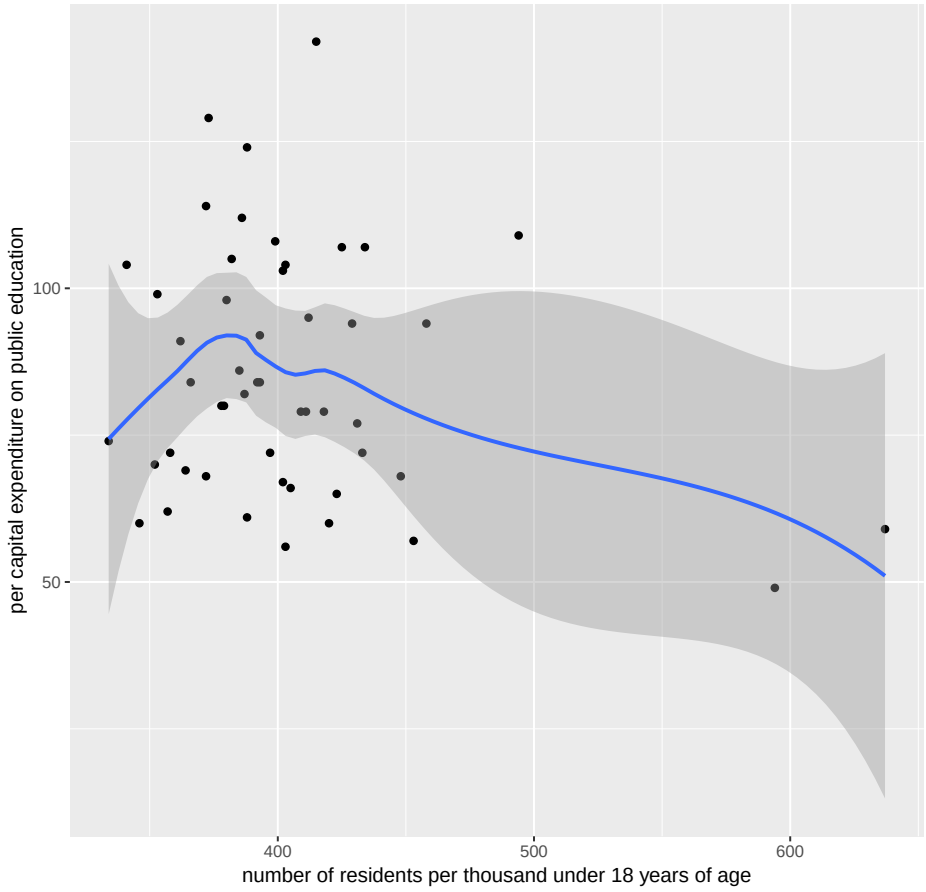
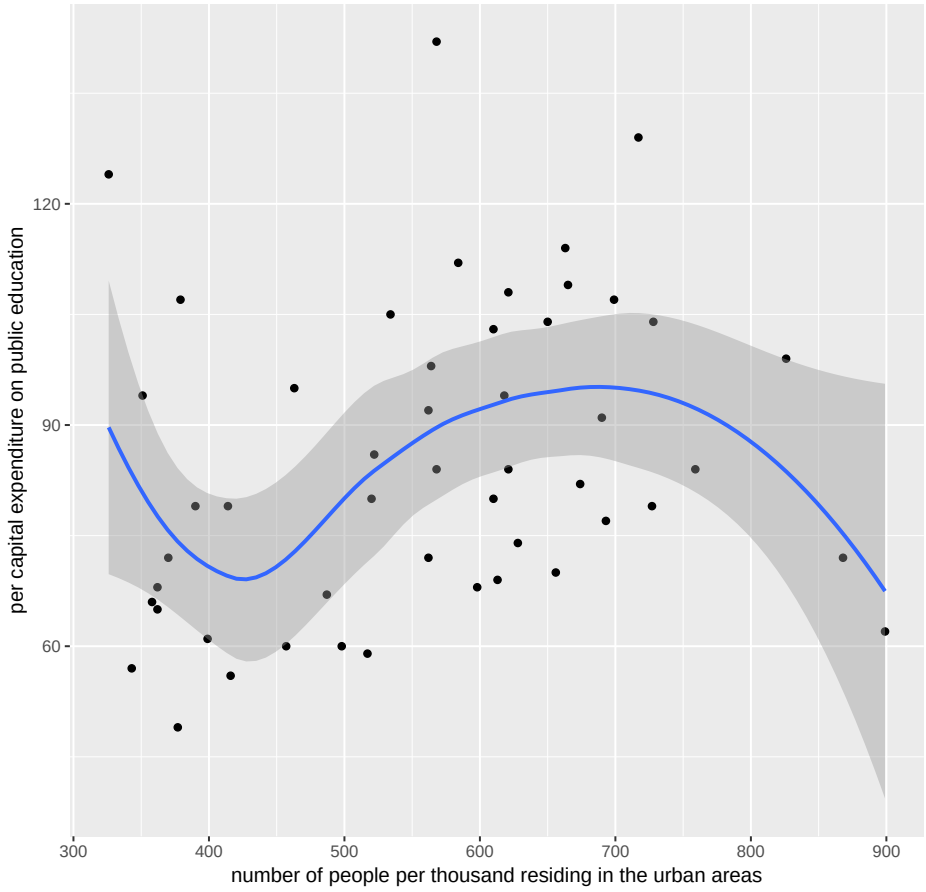


Figure 3: Relationship between number of people per thousand residing in the urban areas and per capital expenditure on public education



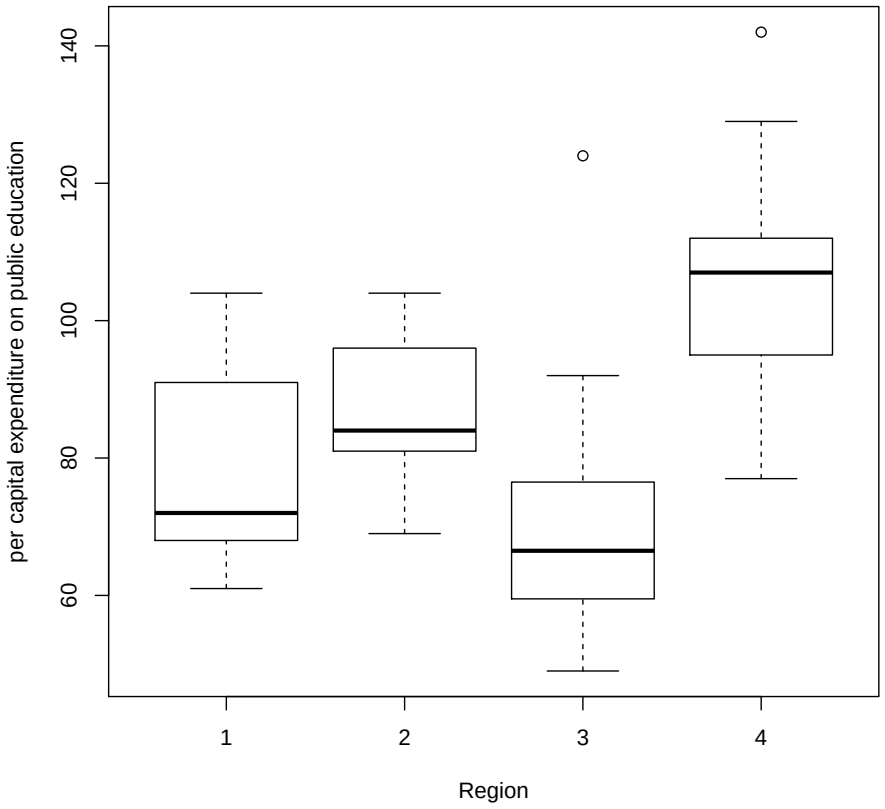
```
    per capital expenditure on public education also increases.

#Y & X2
pdf("plot2.pdf")
ggplot(expenditure, aes(x=X2, y=Y)) + geom_point() + labs(x="number of
    residents per thousand under 18 years of age", y="per capital
    expenditure on public education", title="")+geom_smooth()
dev.off()
#Observation: The relationship between X2 and Y is almost linear and the
    line is downwards slowping. With number of residents per thousand
    under 18 years of age increases, per capital expenditure on public
    education also decreases.

#Y & X3
pdf("plot3.pdf")
ggplot(expenditure, aes(x=X3, y=Y)) + geom_point() + labs(x="number of
    people per thousand residing in the urban areas", y="per capital
    expenditure on public education", title="")+geom_smooth()
dev.off()
#Observation: The relationship between X3 and Y is not linear because the
     curve has 3 inflection points. When X3 approaches to 400, Y value is
    decreasing; when X3 approaches to 700, Y value is increasing; when X3
    increases after 700, Y value decreases.
```

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

```

#2(graphics for this question is under #3 for no reason: I can't fix it
    :(  )
pdf("plot4.pdf")

input=expenditure[,c('Region','Y')]
print(head(input))
boxplot(Y~Region, data = expenditure, xlab = "Region",
ylab = "per capital expenditure on public education", main = "")

dev.off()

#Observation: I use the box plot to compare across regions in terms of
    per capital expenditure on public education. Accrording to the chart
    we find on average  West(4) region has the largest per capital
    expenditure on public education whereas  South(3) region has the
    lowest
```

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

Figure 4: Relationship between Regions and per capital expenditure on public education



different regions with different types of symbols and colors.

```
1
2 #3
3
4 pdf("plot5.pdf")
5
6 ggplot(expenditure, aes(x=X1, y=Y))+ geom_point()+labs(x="per capital
    personal income", y="per capital expenditure on public education",
    title="")+ geom_smooth()
7 #Observation: The relationship between X1 and Y is almost linear and the
    line is upwards slowping. With per capital personal income increases,
    per capital expenditure on public education also increases.
8 dev.off()
9
10 pdf("plot6.pdf")
11 ggplot(expenditure, aes(x=X1, y=Y))+geom_point(aes(color=Region))+ labs(x
    ="per capital personal income", y="per capital expenditure on public
    education", title="")
12 dev.off()
```

Figure 5: Relationship between per capital personal income and per capital expenditure on public education

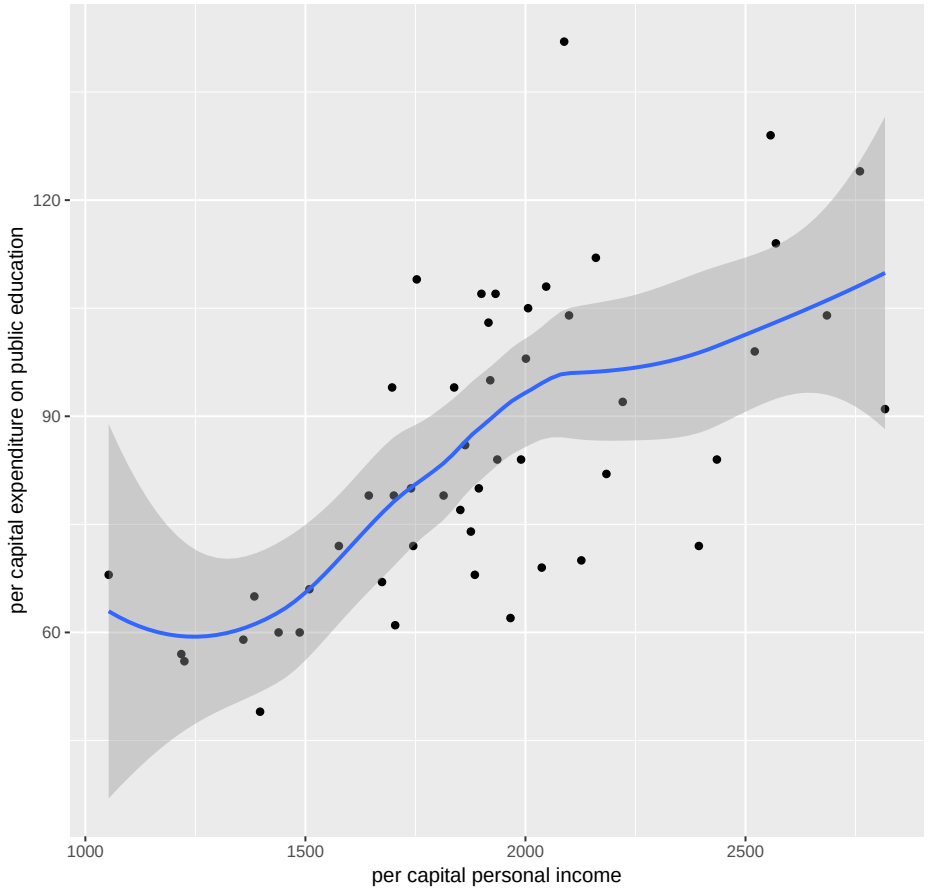


Figure 6: Relationship between per capital personal income and per capital expenditure on public education with Region on the side