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.

```
#Find 90% confidence interval assuming normality
#Find z score
z90 <- qnorm((1-.9)/2, lower.tail = FALSE)

```

```r
#Find sample mean
sample_mean<-mean(y)

#Find sample standard deviation
sample_sd<-sd(y)

#Find upper and lower bounds of confidence interval
upper<-sample_mean + (z90*(sample_sd/sqrt(25)))
lower<-sample_mean - (z90*(sample_sd/sqrt(25)))

#Final confidence interval
CI90<-c(upper, lower)
CI90

#Confidence Interval: (94.133, 102.747)
#Interpretation: when taking random samples of the population, the mean IQ scores of these samples will fall between 94.133 and 102.747 90% of the time.
```

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
#Assumptions met: continuous scale, random sampling (albeit with small issues), population normally distributed, n not quite over 30 but close

#H0: mu<=100, Ha: mu>100

# one sided t test

t.test(y, mu=100, alternative="greater",conf.level=.95)

#With a p value of .7215 and an alpha value of .05, we fail to reject H0. It is entirely possible that we could have obtained this sample mean given a random sample of the entire population, indicating that this school's students likely do not have higher IQs on average.
```

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.

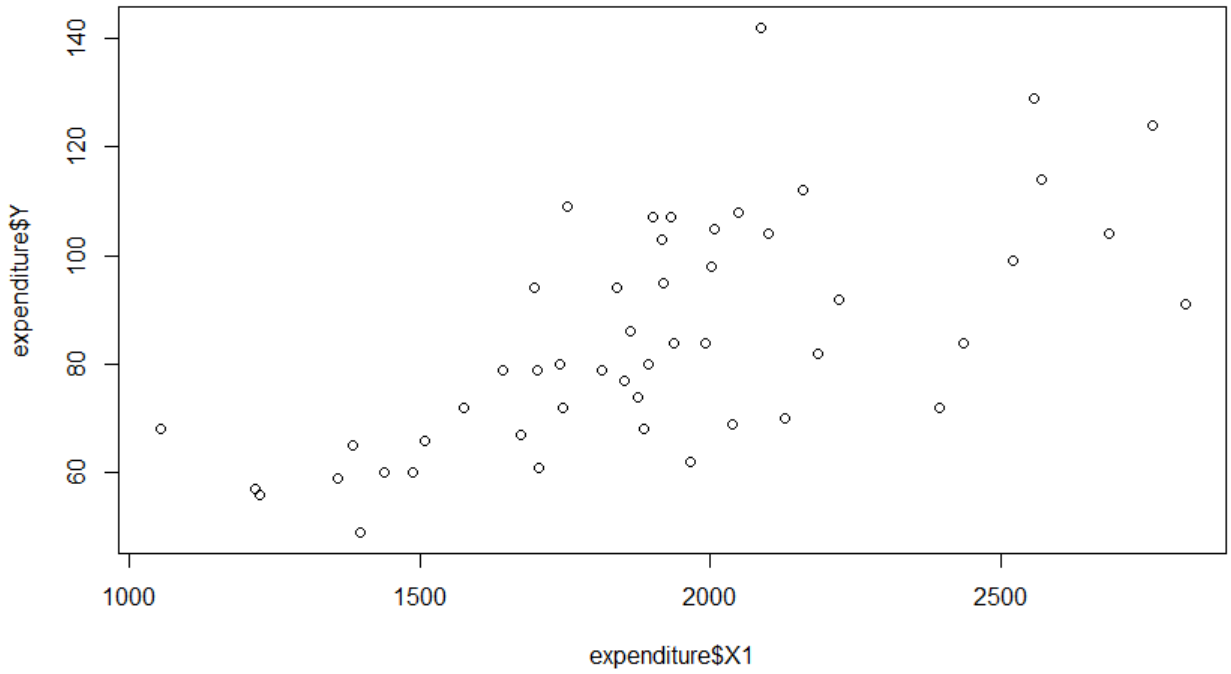
`State`	*50 states in US*
`Y`	*per capita expenditure on public education*
`X1`	*per capita personal income*
`X2`	*Number of residents per thousand under 18 years of age*
`X3`	*Number of people per thousand residing in urban areas*
`Region`	*1=Northeast, 2= North Central, 3= South, 4=West*

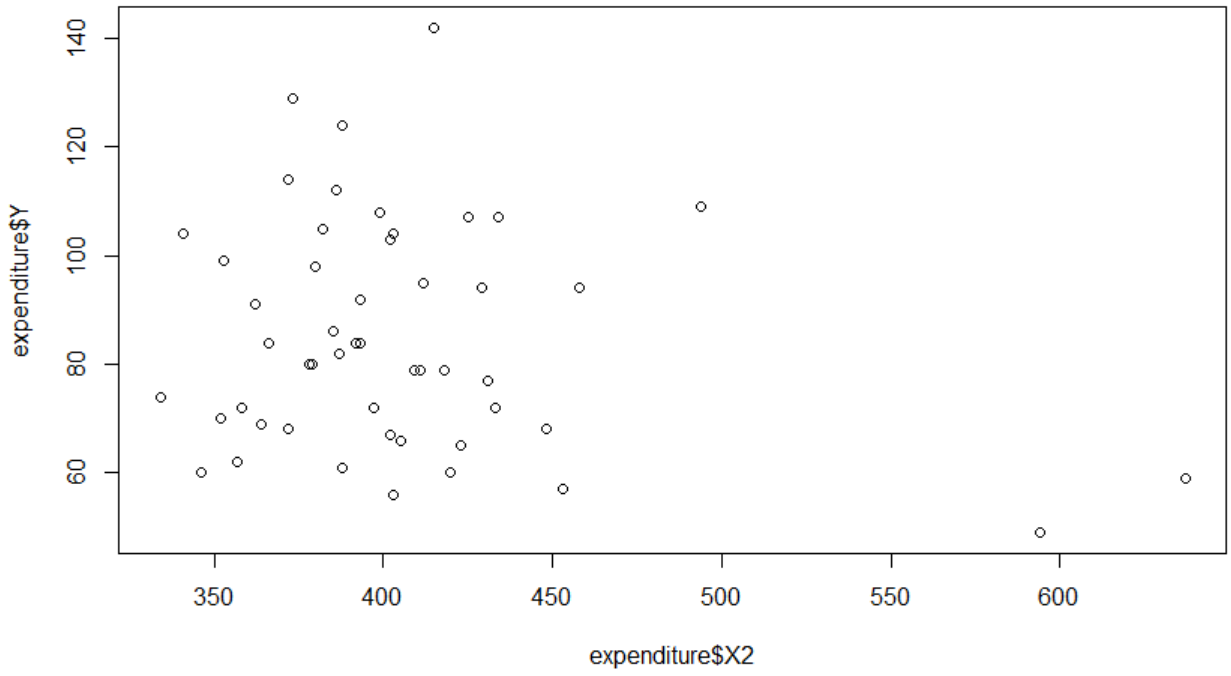
Explore the `expenditure` data set and import data into `R`.

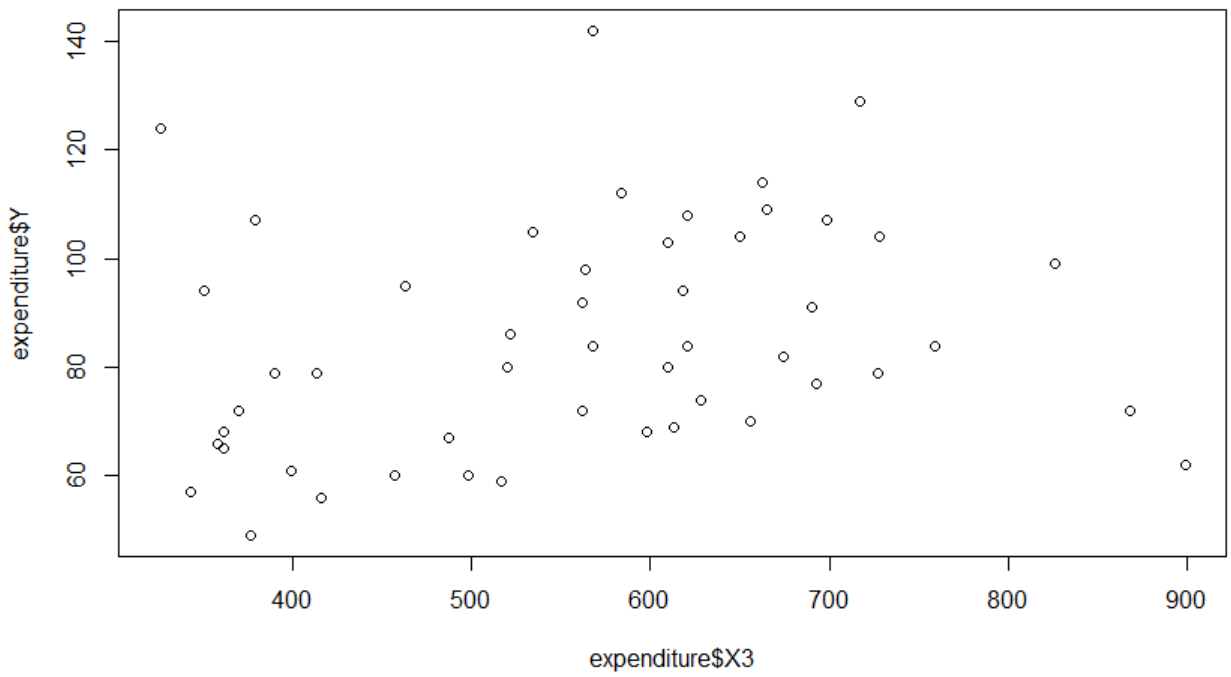
```
expenditure<-read.table("C:/Users/camer/OneDrive/Documents/Emory/Emory Classes
  /QTM 200/expenditure.txt", header=T)
```

- 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)?

```
#Plot X1, X2, and X3 against Y
plot(expenditure$X1, expenditure$Y)
#Correlation: strong positive almost linear (.9)
plot(expenditure$X2,expenditure$Y)
#Correlation: none (0)
plot(expenditure$X3, expenditure$Y)
#Correlation: very weak positive linear (.1)
```

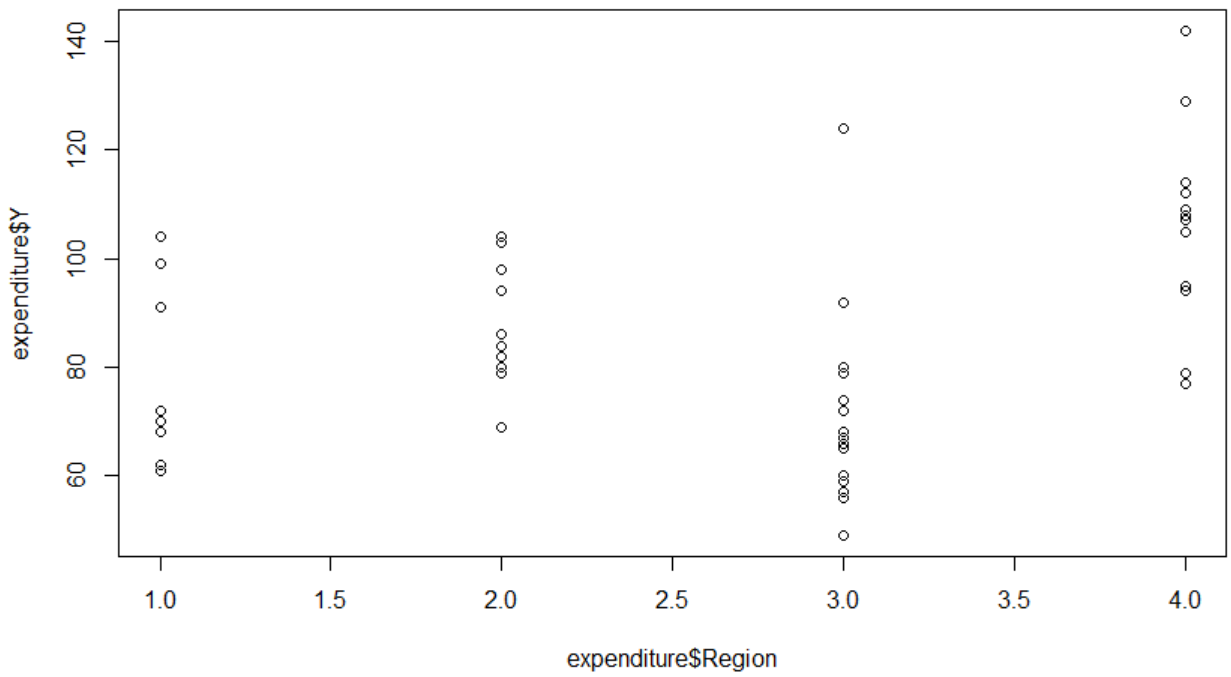

expenditure$Y
60
80
100
120
140
1000
1500
2000
2500
expenditure$X1


expenditure$Y
60
80
100
120
140
350
400
450
500
550
600
expenditure$X2



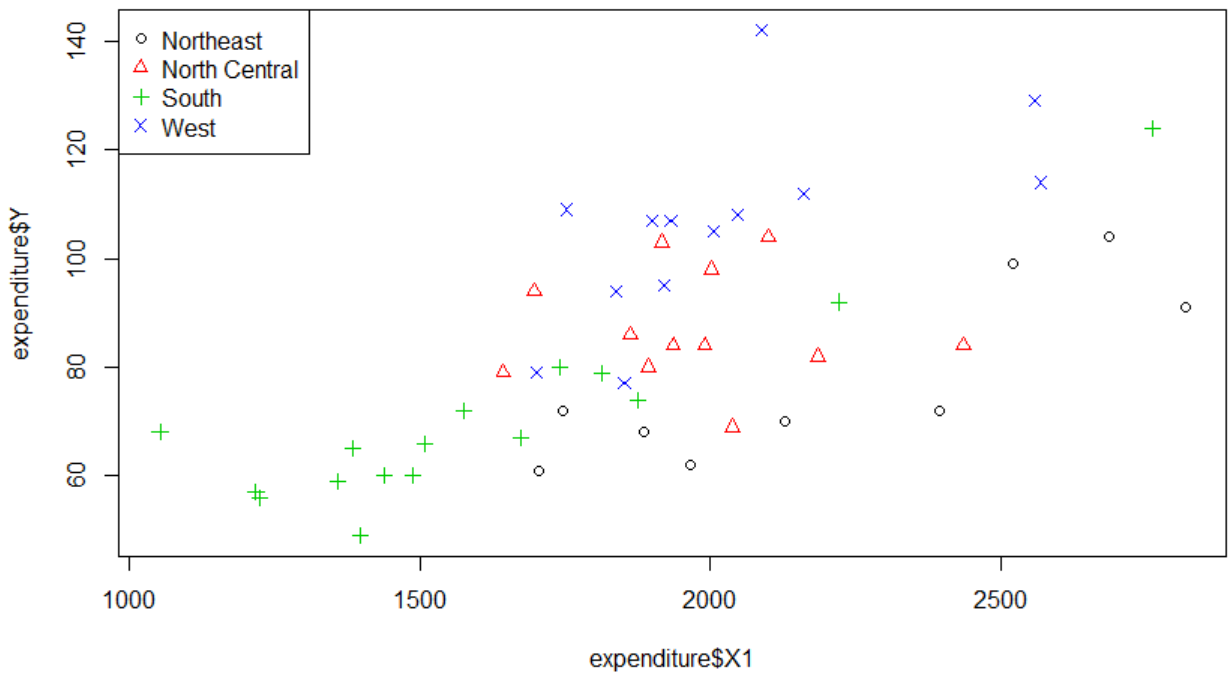
- Please plot the relationship between Y and *Region*? On average, which region has the highest per capita expenditure on public education?

```
#Plot relationship between Y and region
plot(expenditure$Region, expenditure$Y)
#The West has the highest per capita expenditure on education on average
```



- 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.

```
#Plot relationship between X1 and Y with different colors and shapes for
    different regions
plot(expenditure$X1, expenditure$Y, col=expenditure$Region, pch=expenditure$
    Region)
legend("topleft",legend=paste(c("Northeast","North Central","South","West
    ")),col=1:4,pch=1:4)
#Overall, the relationship has a strong, positive, linear correlation,
    but on the regional level this is really only true for the South and
    West. In the Northeast the relationship is much weaker and in the
    North Central region there is no correlation at all.
```


Northeast
North Central
South
West
expenditure$Y
60
80
100
120
140
1000
1500
2000
2500
expenditure$X1