Problem Set 1

QTM 200: Applied Regression Analysis

Due: January 27, 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.

```
##finding the confidnce interval
#first the mean was found
m <- mean(y)
#then the standard deviation
s <- sd(y)
#n is the number of samples
```

```r
n <- 25
#finding the error using t distribution as n is less than 30
error <- qt(0.95, df=n-1)*s/sqrt(n)
#left confidence interval
left <- m-error
#right confidence interval
right <- m+error
left
right
```

In order to find the confidence interval, the mean and standard deviation of the data was determined and stated as "m" and "s" respectively. Then I assigned "n" as the total number of sample. With these variables, the error could be determined by the use of a t-distribution as the number of samples is less than 30, thus illustrating a small sample. With the confidence interval being 90%, we determined the 5% error on each side of the t-distribution. In order to determine the lower confidence interval, the difference between mean and error was determined. To determine the upper confidence interval, the sum of mean and error was determined.

It was found that the 90% confidence interval for the student IQ in the school is 93.96 and 102.92

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
#null hypothesis Ho: pi < pio
pio <- 100
pi <- mean(y)
pi
ts <- qt(.95, df=n-1)*s/sqrt(n)
ts
p = pt(abs(ts), df=n-1, lower.tail=F)
p
```

To conduct a test with 0.05 significance level, first the assumptions need to be stated. The data is continuous, random sampling and the sample is distributed normally.

The null hypothesis is $\nu < 100$ and the alternative hypothesis is $\nu \geq 100$.

The population is mean is 100, while the sample mean is 98.44. The test statistic is determined by using a t-distribution with the confidence interval being 0.95 as this is a one-sided test. The t-distribution is used as this is a small sample (less than 30). With the test statistic, the p-value is determined to be 7.79×10^{-5}, which is smaller than 0.05, and thus leading to rejecting the null hypothesis. We can state that the average IQ of the students in her school is equal or higher to the average IQ score among the population.

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)
```

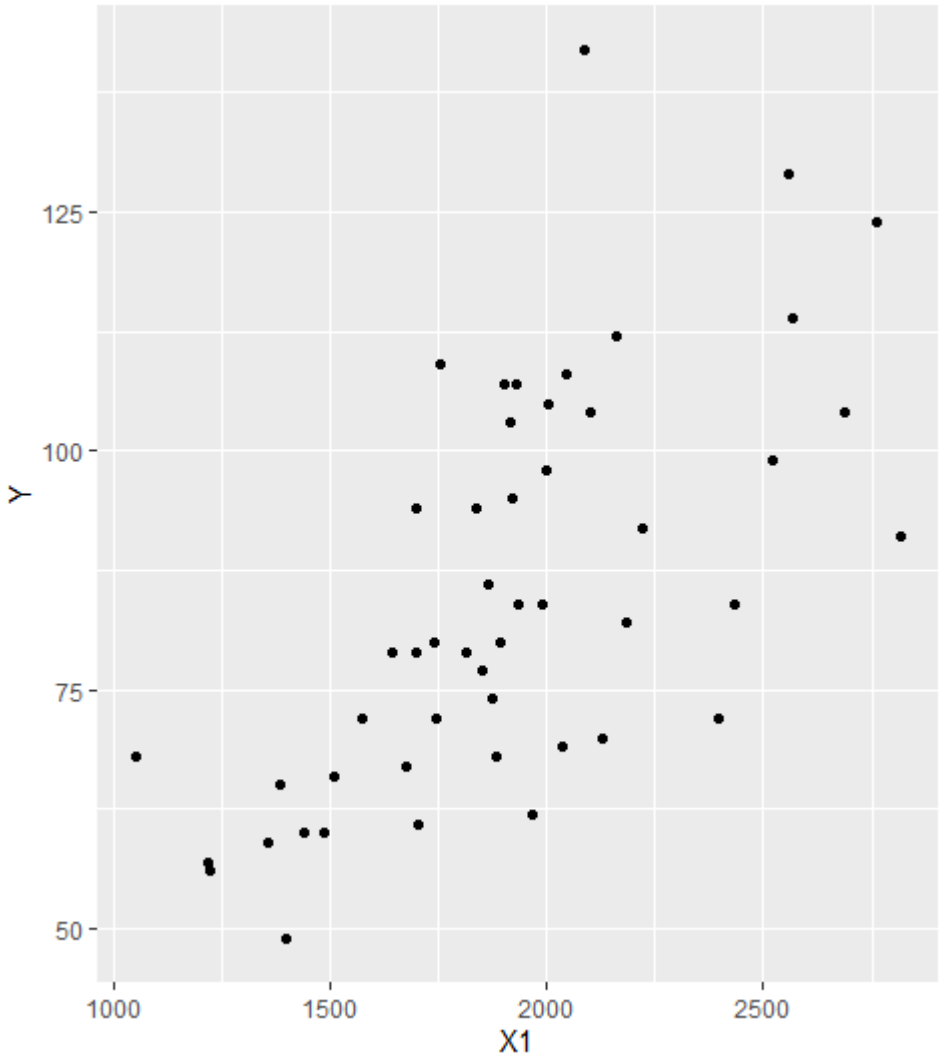
Researchers are curious about what affects the education expenditure on public education. The following is availabe variables in a data set about the education expenditure.

`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*

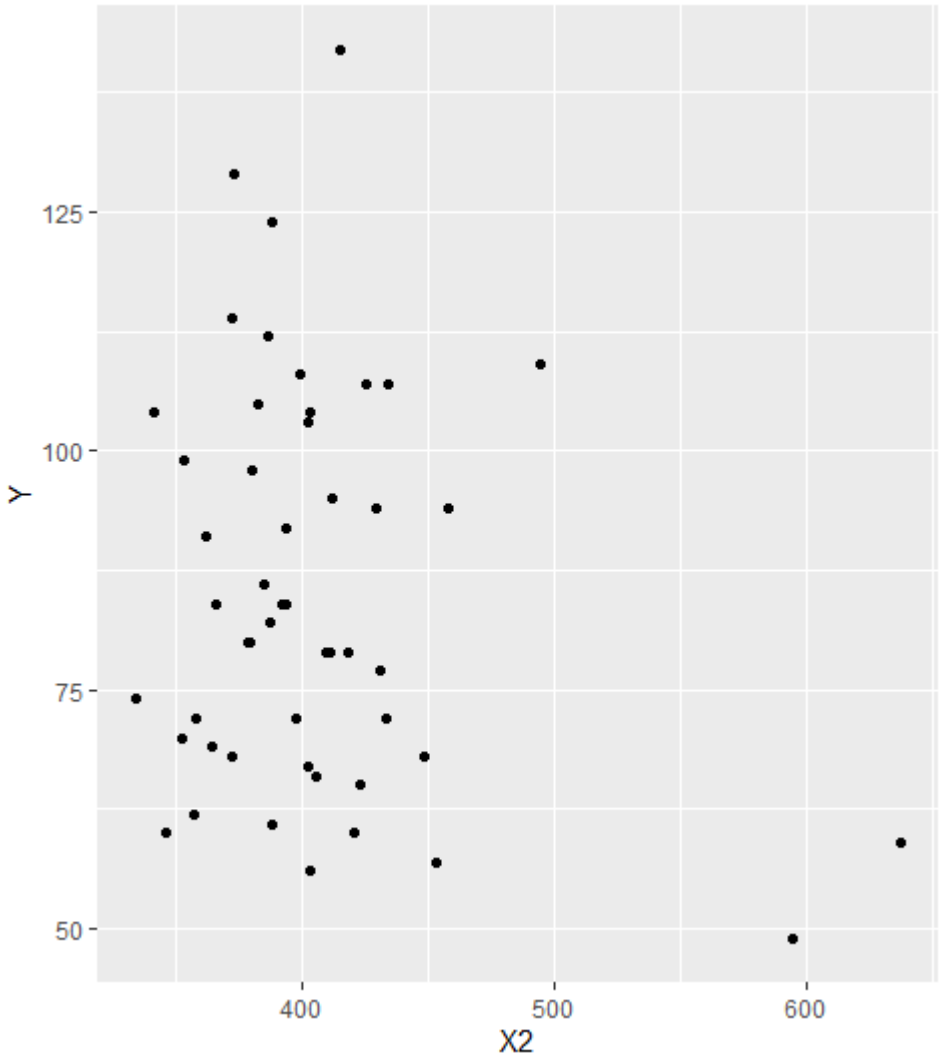
Explore the `expenditure` data set and import data into `R`.

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

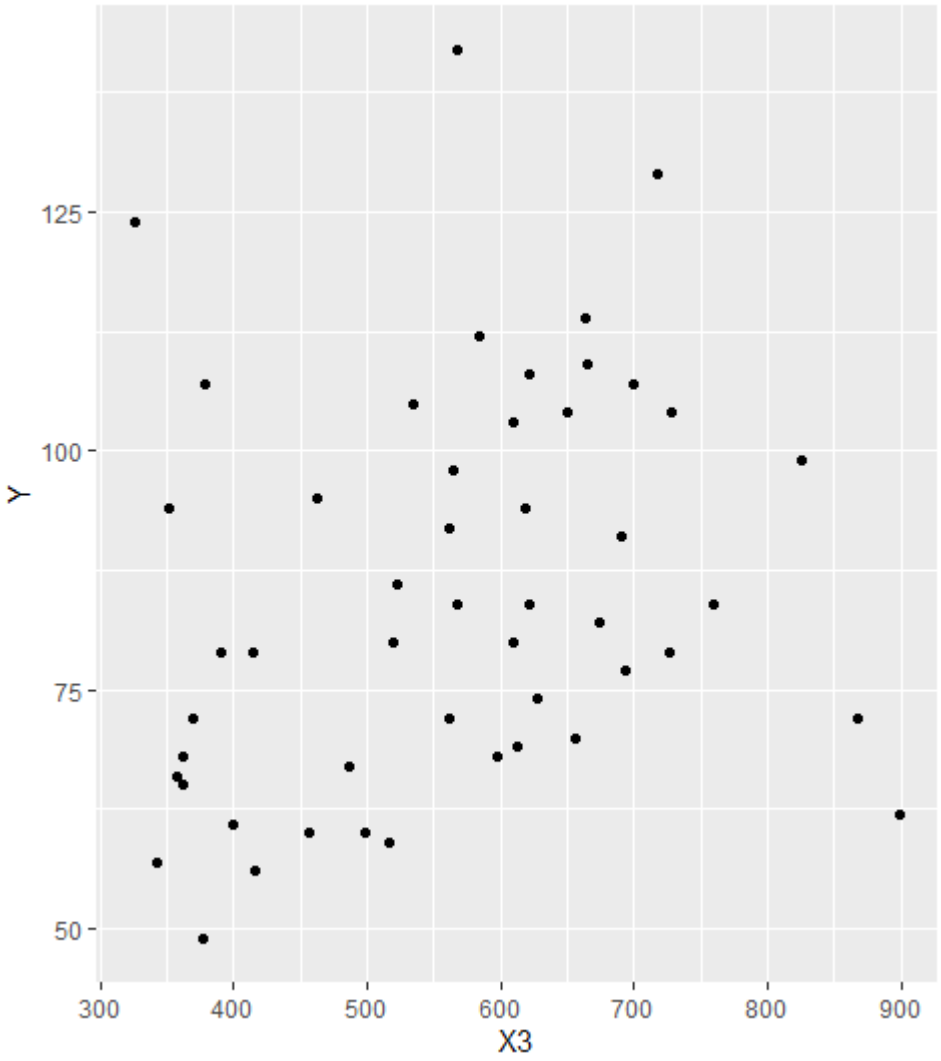
- Please plot the relationships among *Y*, *X1*, *X2*, and *X3*? What are the correlations among them? Describe the graph and the relationships among them.



This graph illustrates a slight positive correlation

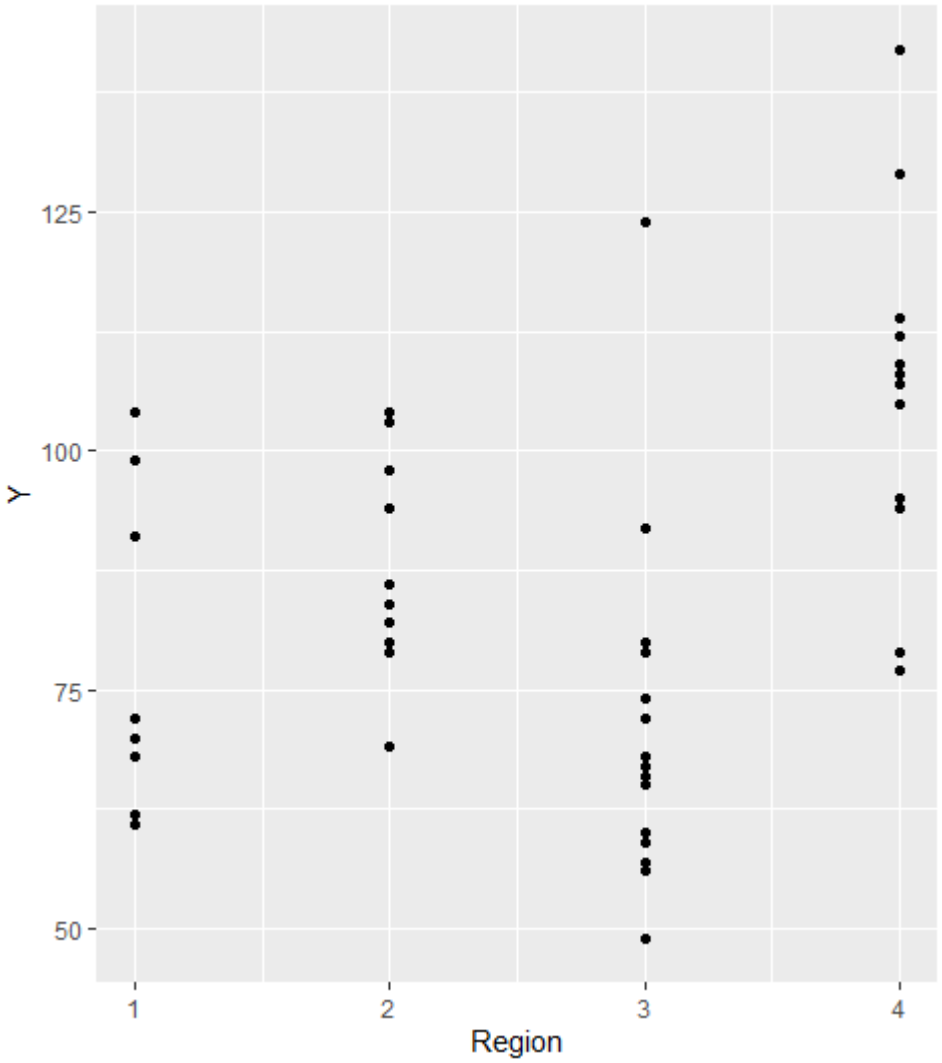


This graph illustrates no correlation, as most of the data is on the lower end of the x axis.



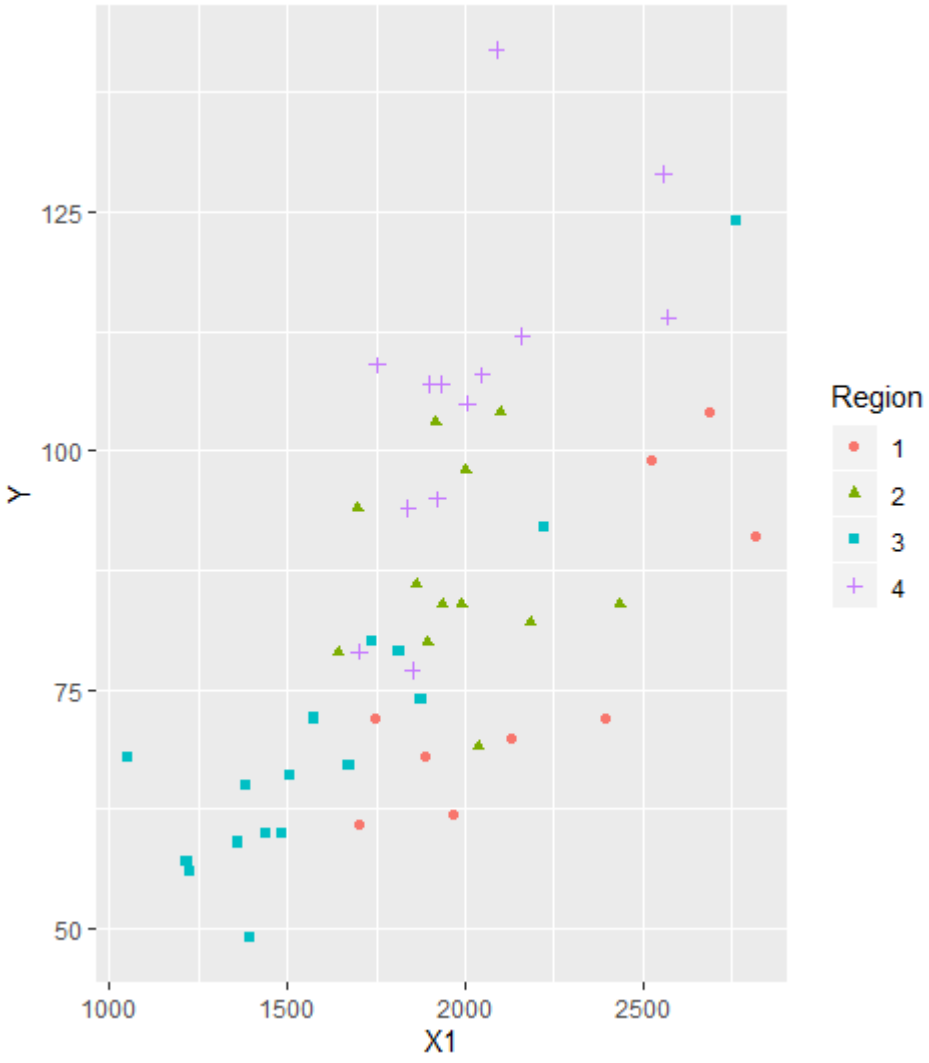
This graph is shown to also have no correlation, however the data is more spread out throughout the y and x axis

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



Region 4 on average has the highest per capital expenditure on public education compared to the other regions. Not only is the min and max the largest in Region 4, but on average, Region 4 is a lot higher per capital expenditure on public education compared to other regions.

- 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.


125
100
75
50
Y
1000
1500
2000
2500
X1
Region
1
2
3
4