Problem Set 1

Applied Stats/Quant Methods 1

Due: October 1, 2021

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 GitHub in `.pdf` form.
- This problem set is due before 8:00 on Friday October 1, 2021. No late assignments will be accepted.
- Total available points for this homework is 100.

Question 1 (50 points): Education

A 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:

```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)
```

1. Find a 90% confidence interval for the average student IQ in the school.

2. Next, the school counselor was curious whether the average student IQ in her school is higher than the average IQ score (100) among all the schools in the country.

   Using the same sample, conduct the appropriate hypothesis test with $\alpha = 0.05$.

Answer 1

1. I first loaded the variable y which was already provided. To estimate the confidence interval of 90% I transformed it into .10 coefficient used in the following formula:

```
z90 <- qnorm((1 - .10)/2, lower.tail = FALSE) ## (1-confidence
    coefficient)/2
```

This formula provided me with the value of Z score is 1.644. Further, I estimated the number of cases (25) and calculated the sample mean (98.44) and standard deviation (13.09) using available R functions:

```
n <- length(y)
sample_mean <- mean(y, na.rm = TRUE)
sample_sd <- sd(y, na.rm = TRUE)
```

Even though na.rm function is not necessary in this case, I left it because it does not create any problems in the calculation. I then proceeded with a standard formula for calculating the confidence interval. I used two tailed confidence intervals and calculated both upper and lower tail:

```
lower_90 <- sample_mean - (z90 * (sample_sd/sqrt(n)))
upper_90 <- sample_mean + (z90 * (sample_sd/sqrt(n)))
confint90 <- c(lower_90, upper_90)
```

Using this formula for variable y, I estimated that the confidence interval lies between 94.1 and 102.7.

2. To check if the average student IQ in the school is the same as average student IQ in the country (= 100) I used a t.test function in R. Before calculating the t test I set up null and alternative hypothesis. Having in mind I already had the sample mean from the previous task, I assumed that the mean of the students in school is not the same as the mean in the country (98.44:100). Therefore my null and alternative hypotheses were defined as following:

   (a) H0: Sample mean = Country mean;

   (b) Ha: Sample mean =/= Country mean;

   I then tested my hypothesis using the following t test features:

```
IQ_null <- t.test(y, mu = 100, conf.level = .05)
IQ_null
```

As per R help guidelines and the tutorial script I did this one-sample t test using y as my sample; mu took the value of 100 as a expected true value of the mean and as per task I took the confidence interval of .05. The results of my t test are the following: t = -0.59574, df = 24, p-value = 0.5569. The 5% confidence interval lies between 98.27407 and 98.60593 while mean is 98.44 (same as in the previous task). As per R output, the alternative hypothesis that sample mean and country mean are not the same is

corroborated. Yet the value of p (0.5569) is bigger than the confidence level of .05 which tells that the null hypothesis cannot be rejected.

Question 2 (50 points): Political Economy

Researchers are curious about what affects the amount of money communities spend on addressing homelessness. The following variables constitute our data set about social welfare expenditures in the USA.

`State`	*50 states in US*
`Y`	*per capita expenditure on shelters/housing assistance in state*
`X1`	*per capita personal income in state*
`X2`	*Number of residents per 100,000 that are "financially insecure" in state*
`X3`	*Number of people per thousand residing in urban areas in state*
`Region`	*1=Northeast, 2= North Central, 3= South, 4=West*

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

- 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)?

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

- 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.

Answer 2

1. As part of the Q2 1st task I loaded the expenditure table. I inspected the following data frame using the following functions:

```
class (expenditure)
typeof (expenditure)
objects (expenditure)
ls.str (expenditure)
attributes (expenditure)
summary (expenditure)
```

   I ran all of them to test the features of R and repeat ways of inspecting variables and datasets from the coding camp and the tutorial.

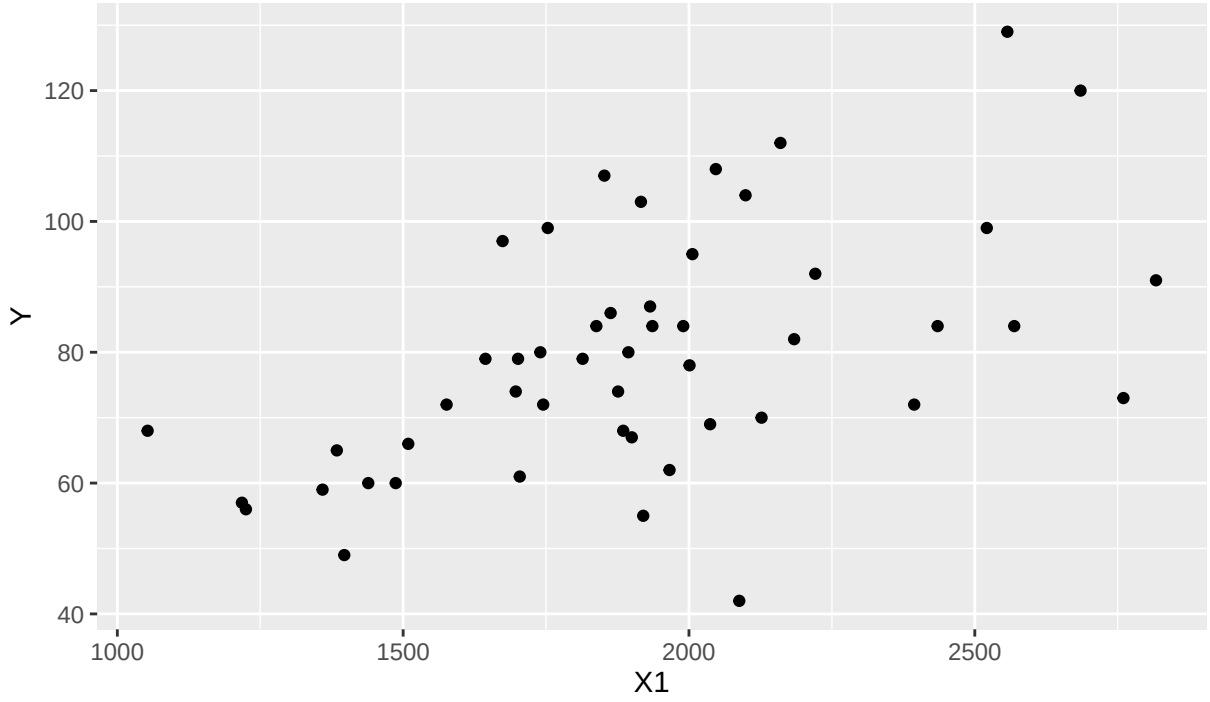
Doing the first task I was not sure how to get the best graphics with barplot function in R, so I loaded ggplot2:

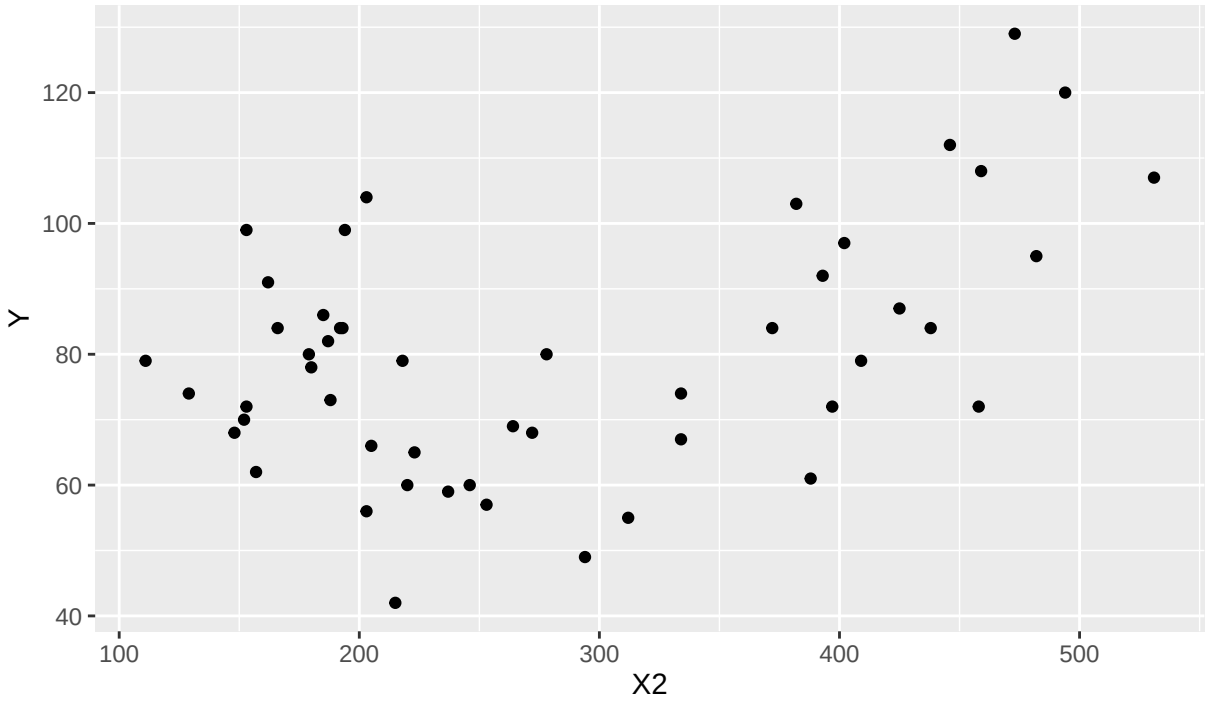
```
library(ggplot2)
```

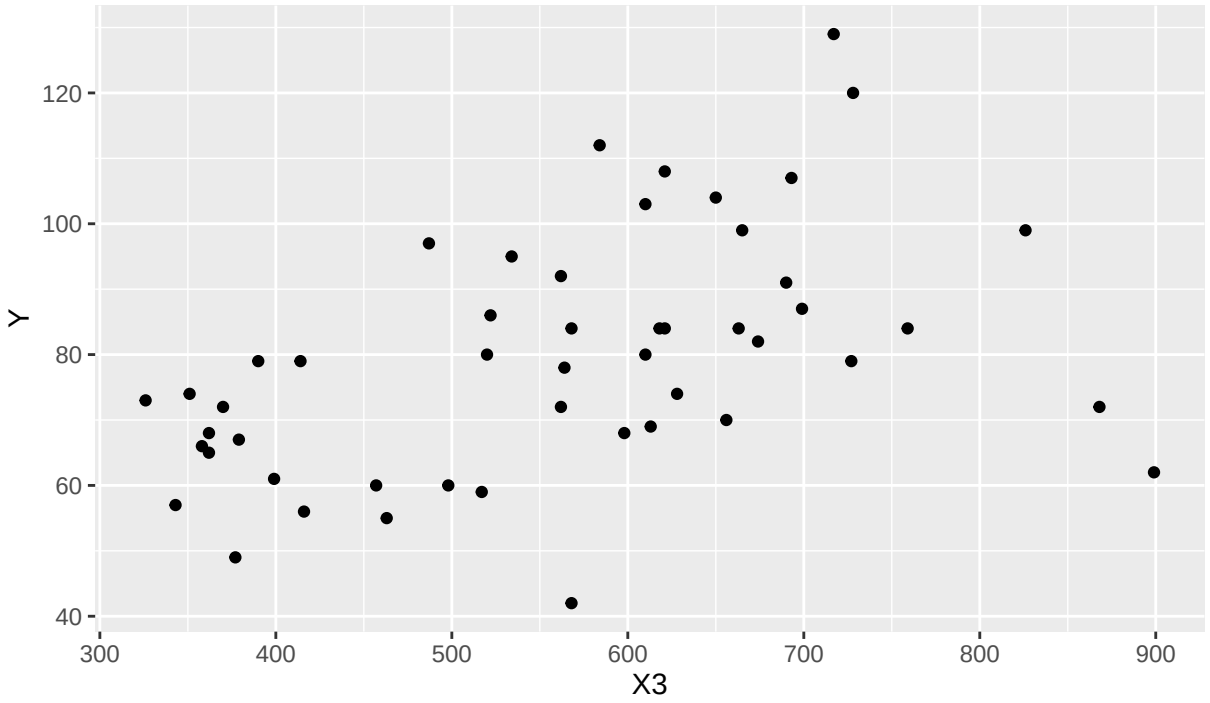
I used option qplot to render graphics of variable y (per capita expenditure on shelters/-housing assistance in state) with X1 (per capita personal income in state), X2 (Number of residents per 100,000 that are "financially insecure" in state) and X3 (Number of people per thousand residing in urban areas in state) separately:

```
qplot (Y,X1, data=expenditure)
qplot (Y,X2, data=expenditure)
qplot (Y,X3, data=expenditure)
```

Using qplot, I rendered the following graphs:







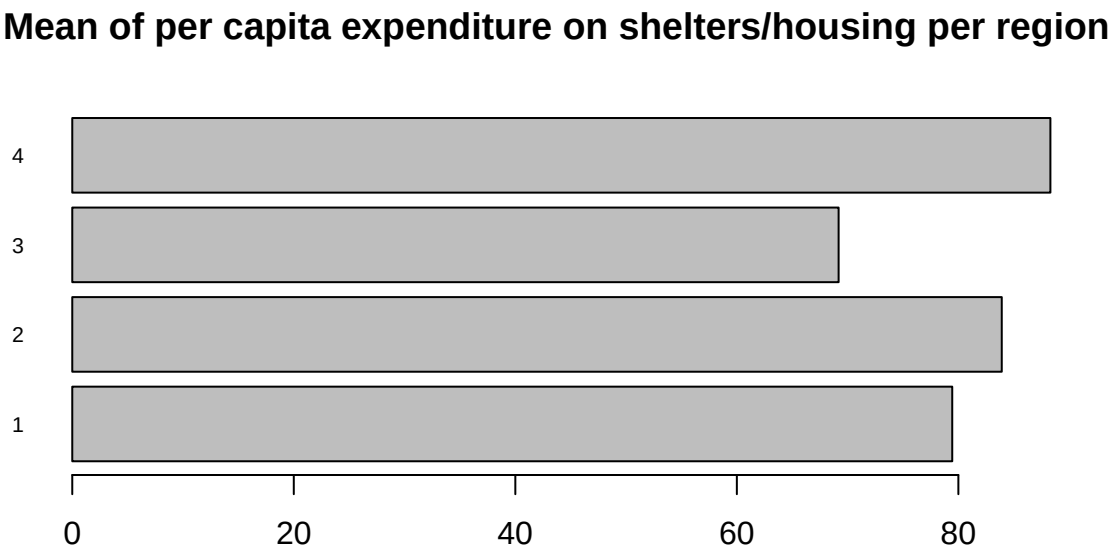
On each graph I put Y (per capita expenditure on shelters/housing assistance in state) on the y axis, while the X1, X2 and X3 were put on the X axis. In the first plot there seem to be a slight correlation between Y and X1 (per capita personal income in state). This correlation seem to exist until the income reaches a point of 1750$

after which the data become more disperse. On the second graph, the relationship per capita expenditure on shelters/housing assistance in state and the number of residents per 100,000 that are "financially insecure" in state does not show a visible trend due to dispersed data until the number of financially insecure residents reaches 300 after which the increasing trend becomes visible. On the third graph, that depicts the relationship between per capita expenditure on shelters/housing assistance in state and a number of people per thousand residing in urban areas in state no clear trend could be observed due to highly dispersed data.

2. To plot the relationship between Y and Region (1=Northeast, 2= North Central, 3= South, 4=West) I used barplot function in R. In the first line I aggregated the data for per capita expenditure on shelters/housing assistance and listed them per regions:

```
Region_mode <- aggregate(expenditure$Y, by = list(expenditure$Region),
    FUN = mean)
#b. Plot our grouped means
bp <- barplot(Region_mode[,2], #use square bracket subsetting to select
    the second col
              names.arg = Region_mode[,1],
              horiz = TRUE, #Flip our axes
              las = 1, #rotate our text to fit it in,
              cex.names = 0.7, #make our axis text a bit smaller to fit
              main = "Mean of per capita expenditure on shelters/housing
    per region")
```

This code rendered the following bar chart:

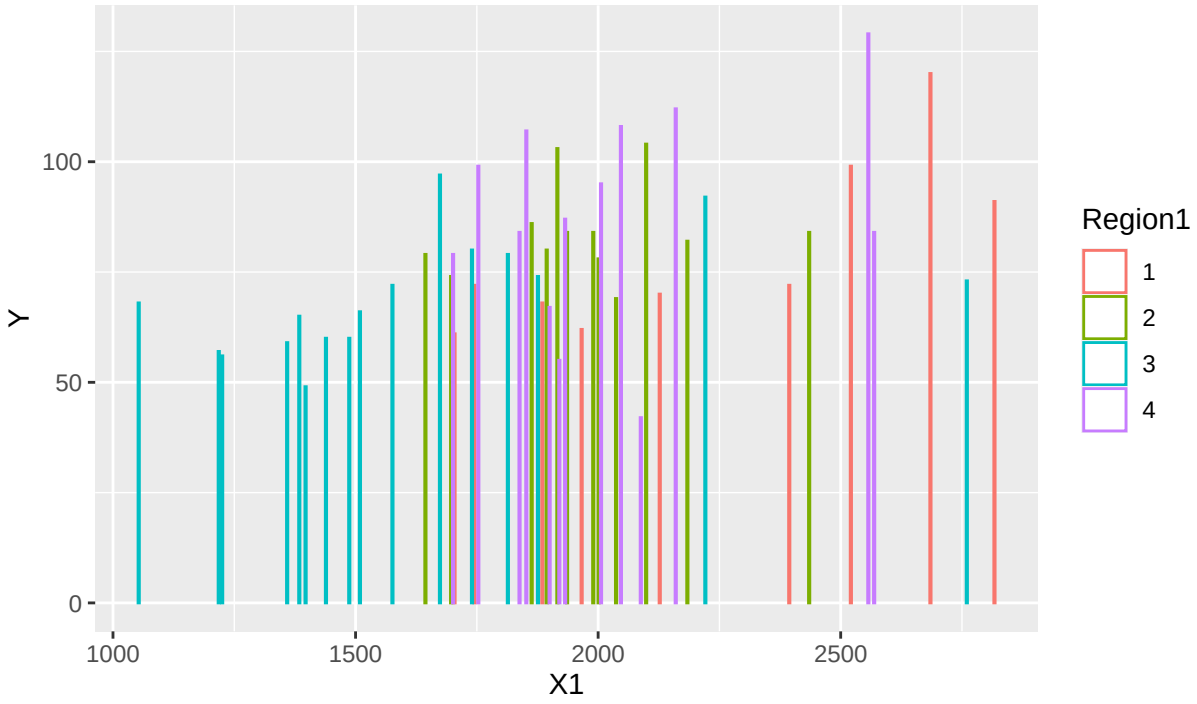


From this graph it is visible that the highest amount of per capita expenses is in the 4th region (West), while the lowest level is in the 3rd region (South).

3. To resolve the third task, I used ggplot2 again. Inside ggplot I used colour option to indicate the differences among regions while plotting the relationship between per capita expenditure and per capita personal income. Before rendering the graph I created and new variable Region1, overwriting Region as factor variable:

```
expenditure$Region1 <- as.factor(expenditure$Region)
```

I put per capita expenditure on y axis and per capita personal income in state on x axis (in accordance with previous graphs). The code rendered the following graph:



From this graph it is visible that the lowest level of per capita income correlates the lowest levels of expenditure which are most common in the southern states. On average the highest amount of expenditure is most densely concentrated in the region around $2000 income level.