

Problem Set 1

Applied Stats/Quant Methods 1

Due: October 1, 2023

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.
- This problem set is due before 23:59 on Sunday October 1, 2023. No late assignments will be accepted.
- Total available points for this homework is 80.

Question 1 (40 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:

```
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)
```

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

```
1 #1.  
2 n <- length(y)  
3 sample_mean <- mean(y)  
4 sample_sd <- sd(y)
```

```

5 z90<-qnorm((1-.9)/2,25,24,lower.tail = FALSE)
6 z90<-qnorm((1-.9)/2,lower.tail = FALSE)
7 lower90<-sample_mean-(z90*(sample_sd/sqrt(n)))
8 upper_90<-sample_mean+(z90*(sample_sd/sqrt(n)))
9 confident90<-c(lower90, upper_90)
10 confident90

```

- a 90% confidence interval for the average student IQ is [94.13283,102.74717]

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 $\alpha = 0.05$.

```

1 df<-n-1
2 t<-(sample_mean-100)/(sample_sd/sqrt(n))
3 p_value<-2*pt(abs(t),df,lower.tail = FALSE)
4 p_value

```

- p-value= 0.5569233, which is bigger than 0.05,so we can not reject the null hypothesis, i.e. the average student IQ in her school is higher than the ave IQ score (100)

Question 2 (40 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

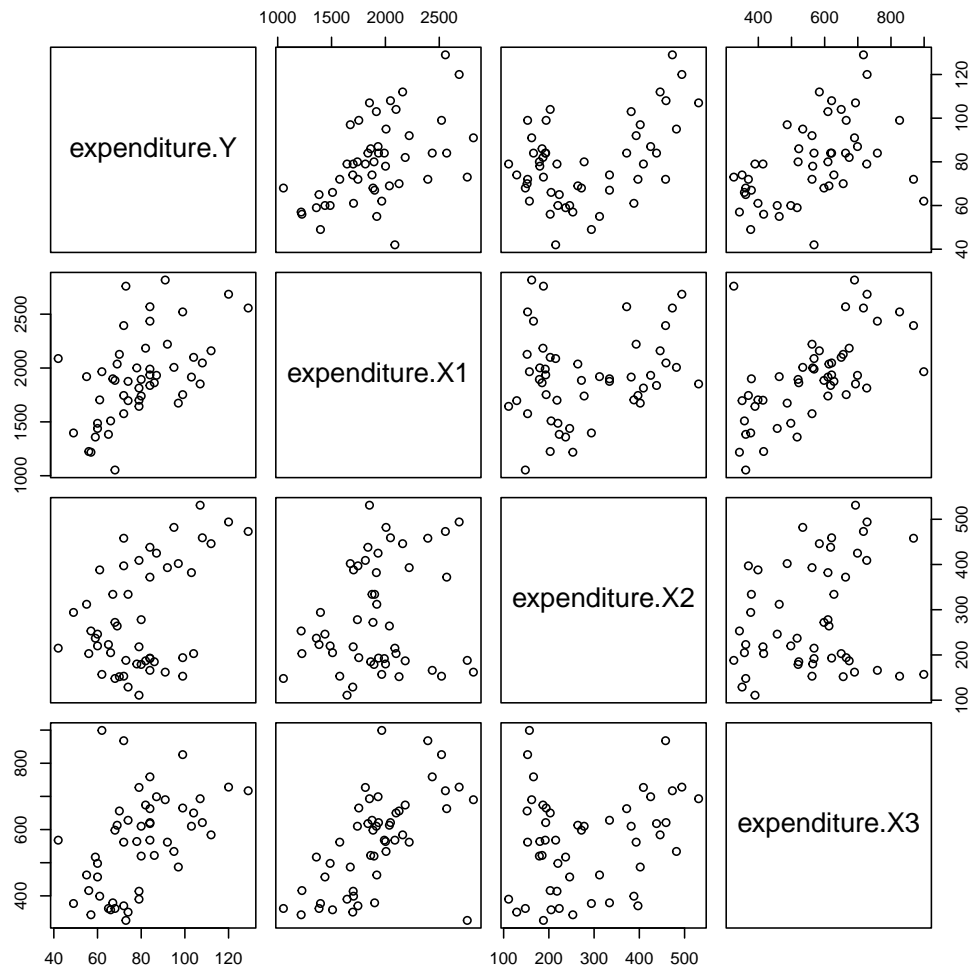
Explore the `expenditure` data set and import data into R.

- 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)?

```
1 #1.
2 expenditure <- read.table("https://raw.githubusercontent.com/ASDS-TCD/
  StatsI_Fall2023/main/datasets/expenditure.txt", header=T)
3
4 YX123<-data.frame(expenditure$X1, expenditure$Y,expenditure$X2,
  expenditure$X3)
5 pdf("plot_YX123.pdf")
6 plot(YX123)
7 dev.off()
```

- According to the graphs, there is a positive correlation between $X1$ and Y , $X2$ and Y , $X3$ and Y , $X1$ and $X2$, $X1$ and $X3$; Besides, there is no correlation between $X2$ and $X3$.

Figure 1: Relationship among Y, X1, X2, X3



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

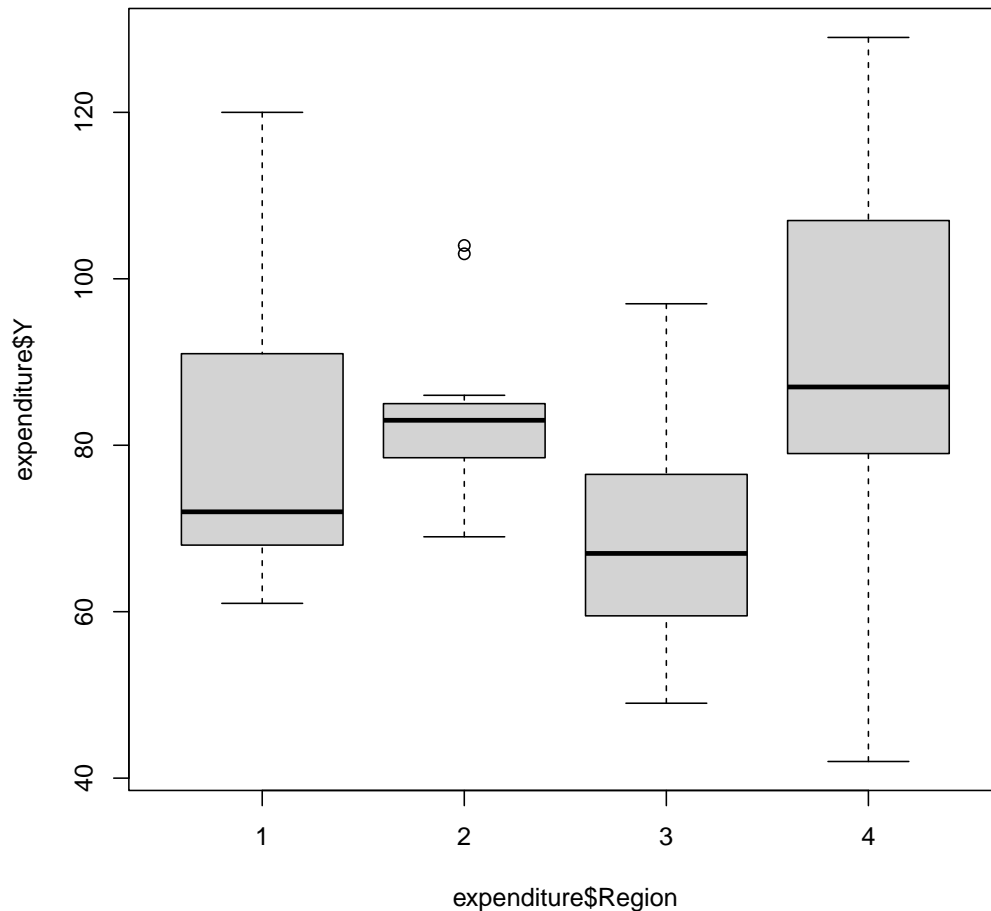
```

1 #2.
2 # create scatterplot of region and Y
3 pdf("plot_regionandY.pdf")
4 plot(expenditure$Region, expenditure$Y)
5 dev.off()

```

- According to the graphs, the second region has the highest per capita expenditure on housing assistance on average.

Figure 2: Relationship Between Region and Y



- 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.

```

1 #3.
2 # create scatterplot of X1 and Y and region
3 pdf("plot_X1andYandregion.pdf")
4 plot(expenditure$Region, expenditure$Y, col=expenditure$Region)
5 dev.off()

```

- According to the graphs, there is positive correlation between Y and $X1$.

Figure 3: Relationship Between Y, X1

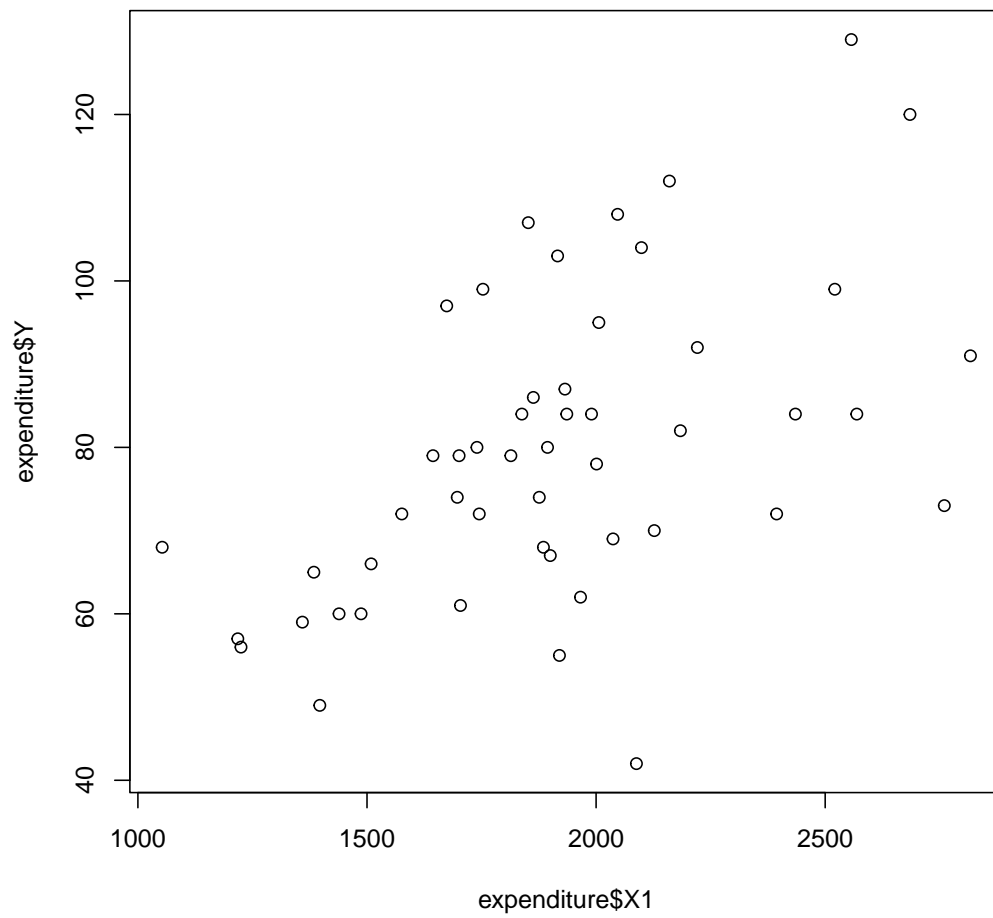


Figure 4: Relationship Between Y, X1 and Region

