Problem Set 1

Zhuo Zhang

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

```r
IQ_scores <- 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)

# Perform a t-test
t_test_result <- t.test(IQ_scores, conf.level = 0.90)
lower <- t_test_result$conf.int[1]
upper <- t_test_result$conf.int[2]
# Print t-test results
cat("1. 90% Confidence Interval for School Students' Average IQ:\n")
cat( lower, "-", upper, "\n")
```

- **Result**:
  1. 90% Confidence Interval for School Students' Average IQ:
  93.95993 - 102.9201

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$\alpha = 0.05$.

```r
#Perform a t-test
t_test_result <- t.test(IQ_scores, mu = 100, alternative = "greater")

# Print t-test results
cat("2.Hypothesis Test Results:\n")
cat("Test Statistic (t):", t_test_result$statistic, "\n")
cat("p-value:", t_test_result$p.value, "\n")

# Determine whether to reject the null hypothesis based on the p-value
if (t_test_result$p.value < 0.05) {
  cat("Reject the null hypothesis: the average IQ of students in the
   school is above 100.\n")
} else {
  cat("Null hypothesis not rejected: There is insufficient evidence that
   the average IQ of students in the school is above 100.\n")
}
```

- **Result:**
  2.Hypothesis Test Results:
  Test Statistic (t): -0.5957439
  p-value: 0.7215383
  Null hypothesis not rejected: There is insufficient evidence that the average IQ of students in the school is above 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*

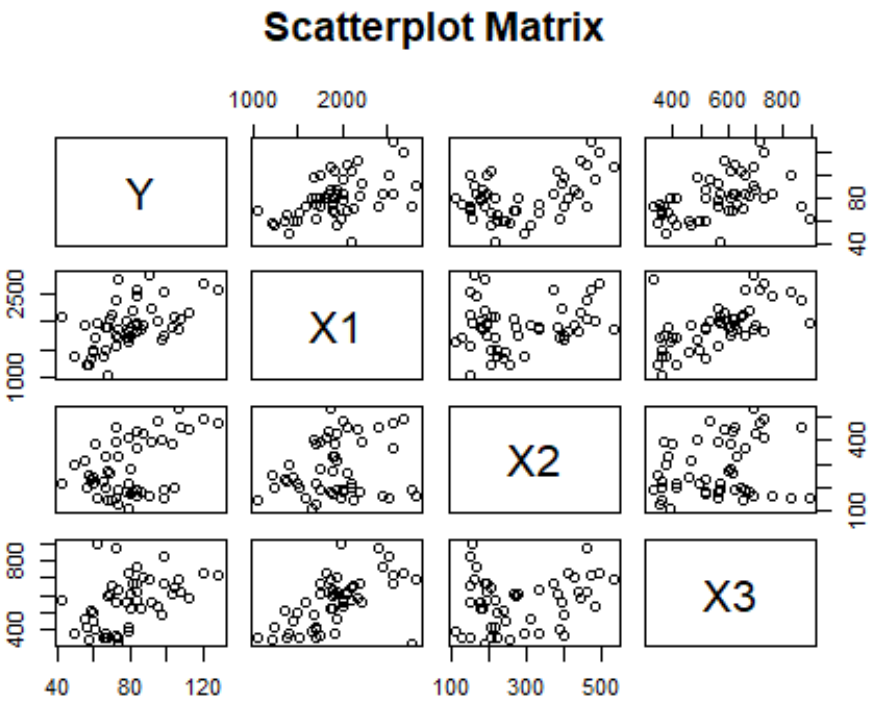
Explore the `expenditure` data set and import data into `R`.

```
df <- read.table("https://raw.githubusercontent.com/ASDS-TCD/StatsI_Fall2023/main/datasets/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)?

```
# Scatterplot matrix
pairs(df[c("Y", "X1", "X2", "X3")], main="Scatterplot Matrix")

# Calculate correlations
correlations <- cor(df[c("Y", "X1", "X2", "X3")])
print(correlations)
```



– **Result:**

– As can be seen in the scatterplot matrix, the relationship between Y and X1 shows a positive correlation. This means that as personal income per person (X1) increases, there is a tendency for expenditures per person on housing assistance (Y) to increase as well.

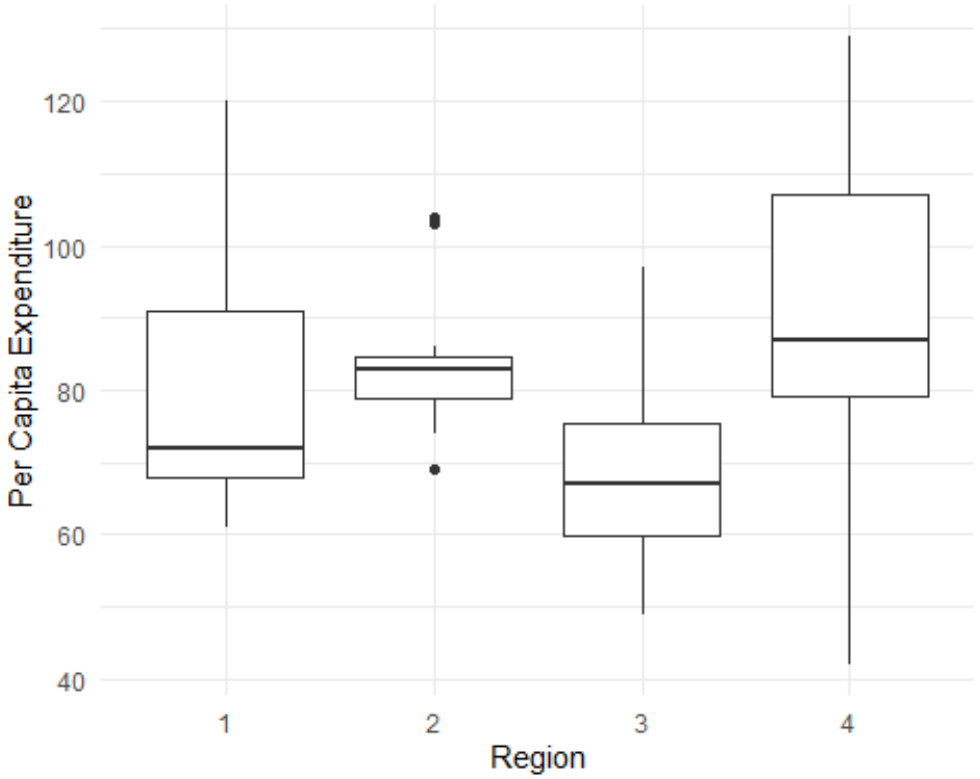
– The relationship between Y and X2 is less obvious, and the scatterplot distribution is relatively scattered, making it difficult to determine if there is a clear linear relationship.

– The relationship between Y and X3 appears to be negative. This implies that as the proportion of the population living in urban areas (X3) increases, spending per person on housing assistance (Y) may decrease.

– The relationship between X1 and X2 and X3 is difficult to determine intuitively.

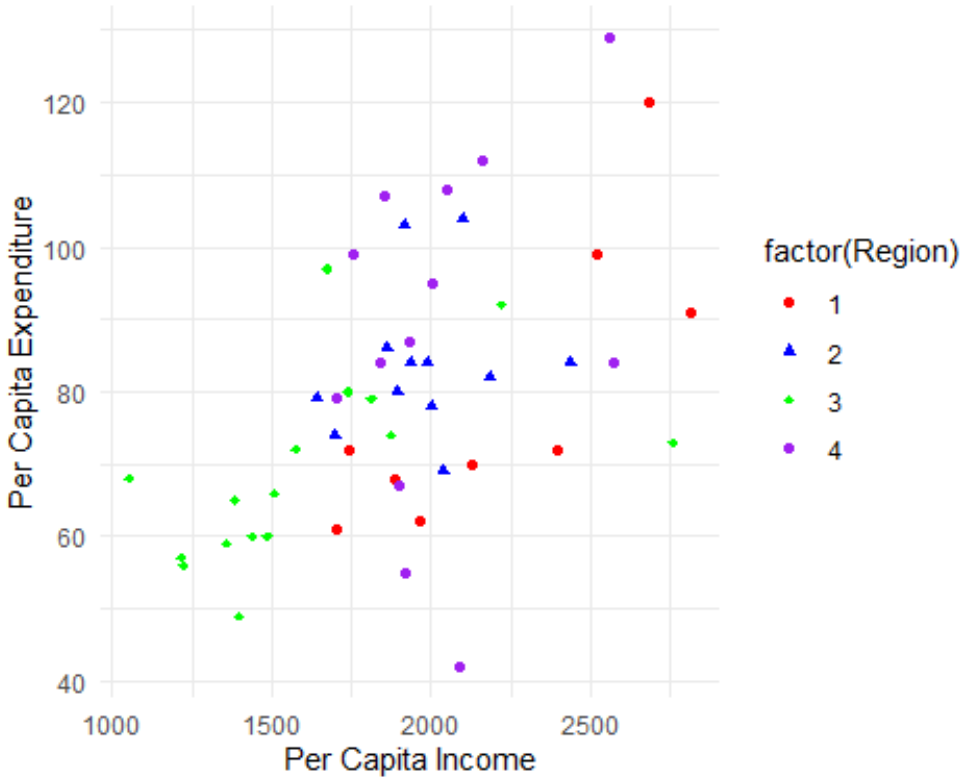
	Y	X1	X2	X3
Y	1.0000000	0.5317212	0.4482876	0.4636787
X1	0.5317212	1.0000000	0.2056101	0.5952504
X2	0.4482876	0.2056101	1.0000000	0.2210149
X3	0.4636787	0.5952504	0.2210149	1.0000000

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



  - **Result:**
  - On average, region 4 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.



- The x-axis represents per capita personal income (X1) in the state.
- The y-axis represents per capita expenditure on shelters/housing assistance (Y) in the state.
- Different colors and shapes are used to differentiate between regions (1 = Northeast, 2 = North Central, 3 = South, 4 = West). In the context of Region 3 (South), with its lower per capita income.
  - **Result:**
  - As we can see from the chart above, there is a positive relationship between per capita personal income (X1) and per capita housing/housing assistance expenditures (Y), with the higher the per capita personal income (X1), the higher the per capita housing/housing assistance expenditures (Y). The South region has the lowest per capita personal income (X1) and the lowest per capita housing/housing assistance expenditures (Y).