Problem Set 1

Junjie LIU[1,*]

2023-09-28

[1] Department of Political Science, Trinity College Dublin, 2 Clare, Street, Dublin 2, Ireland

[*] Correspondence: Junjie LIU <liuj13@tcd.ie>

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:

```
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 of Question 1

For the first sub-task in Question 1:

```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)

### For the Q1.1
q11.n <- length(y)
q11.y_bar <- mean(y)
q11.y_std <- sqrt(var(y))

q11.margin_left <- q11.y_bar - qt(0.05, q11.n, lower.tail = F) *
    q11.y_std/sqrt(q11.n)
q11.margin_right <- q11.y_bar + qt(0.05, q11.n, lower.tail = F) *
    q11.y_std/sqrt(q11.n)
sprintf("The 90%% CI of the Avg IQ is (%1$.3f, %2$.3f)", q11.margin_left,
    q11.margin_right)
```

```
## [1] "The 90% CI of the Avg IQ is (93.967, 102.913)"
```

For the second sub-task of Question 1:

- $\mathcal{H}_0$: The average student IQ in her school is not higher than the average IQ score (100) among all the scools in the country, i.e. $\text{IQ}_{\text{school}} \leq \text{IQ})_{\text{country}}$
- $\mathcal{H}_1$: The average student IQ in her school is higher than the average IQ score (100) among all the schools in the country, i.e. $\text{IQ}_{\text{school}} > \text{IQ})_{\text{country}}$

```r
### For the Q1.2
alpha <- 0.05
q12.t_test <- t.test(y, mu = 100, conf.level = (1 - alpha), alternative = "greater")
print(q12.t_test)
```

```
##
##  One Sample t-test
##
## data:  y
## t = -0.59574, df = 24, p-value = 0.7215
## alternative hypothesis: true mean is greater than 100
## 95 percent confidence interval:
##  93.95993      Inf
## sample estimates:
## mean of x
##     98.44
```

Since the p-value is 0.7215, we could not reject the null hypothesis.

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)?

- 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.

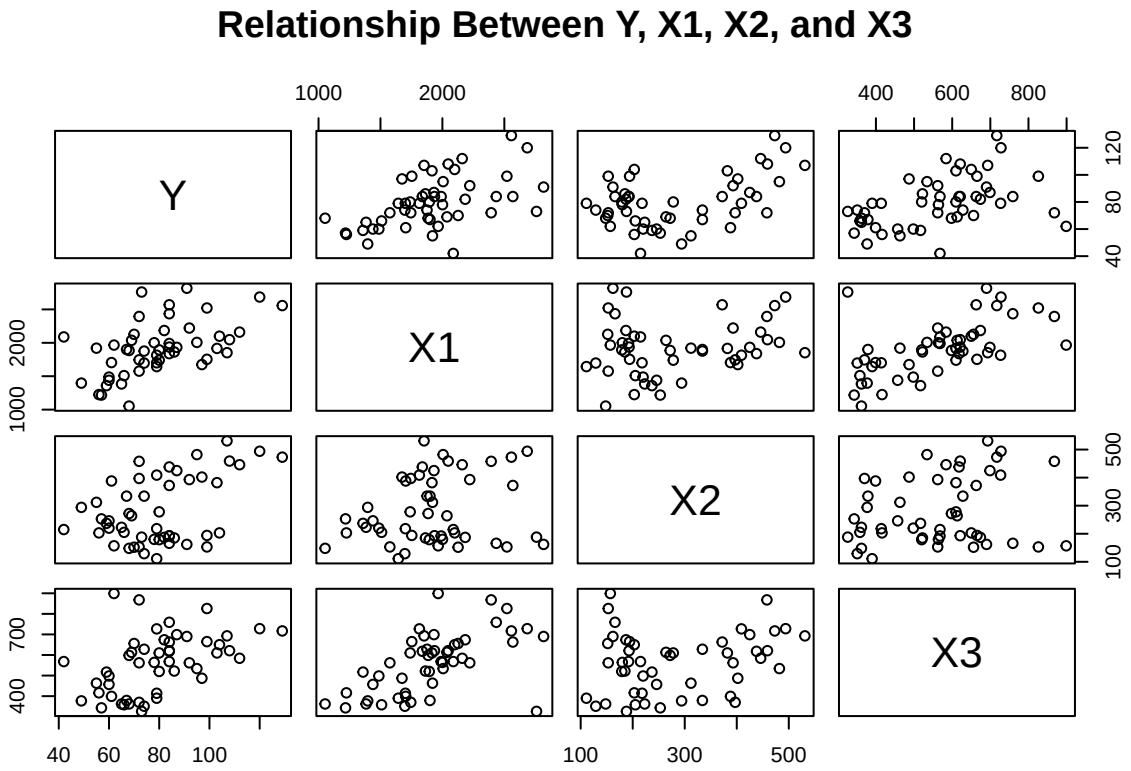
Answer of Question 2

For the first sub-task in Question 2:

```
expenditure <- read.table("https://raw.githubusercontent.com/ASDS-TCD/StatsI_Fall2023/main/datasets/exp
    header = T)
summary(expenditure)
```

```
##     STATE                  Y               X1             X2
##  Length:50          Min.   : 42.00   Min.   :1053   Min.   :111.0
##  Class :character   1st Qu.: 67.25   1st Qu.:1698   1st Qu.:187.2
##  Mode  :character   Median : 79.00   Median :1897   Median :241.5
##                     Mean   : 79.54   Mean   :1912   Mean   :281.8
##                     3rd Qu.: 90.00   3rd Qu.:2096   3rd Qu.:391.8
##                     Max.   :129.00   Max.   :2817   Max.   :531.0
##        X3            Region
##  Min.   :326.0   Min.   :1.00
##  1st Qu.:426.2   1st Qu.:2.00
##  Median :568.0   Median :3.00
##  Mean   :561.7   Mean   :2.66
##  3rd Qu.:661.2   3rd Qu.:3.75
##  Max.   :899.0   Max.   :4.00
```

```r
### For the Q2.1
pairs(expenditure[, 2:5], main = "Relationship Between Y, X1, X2, and X3")
```

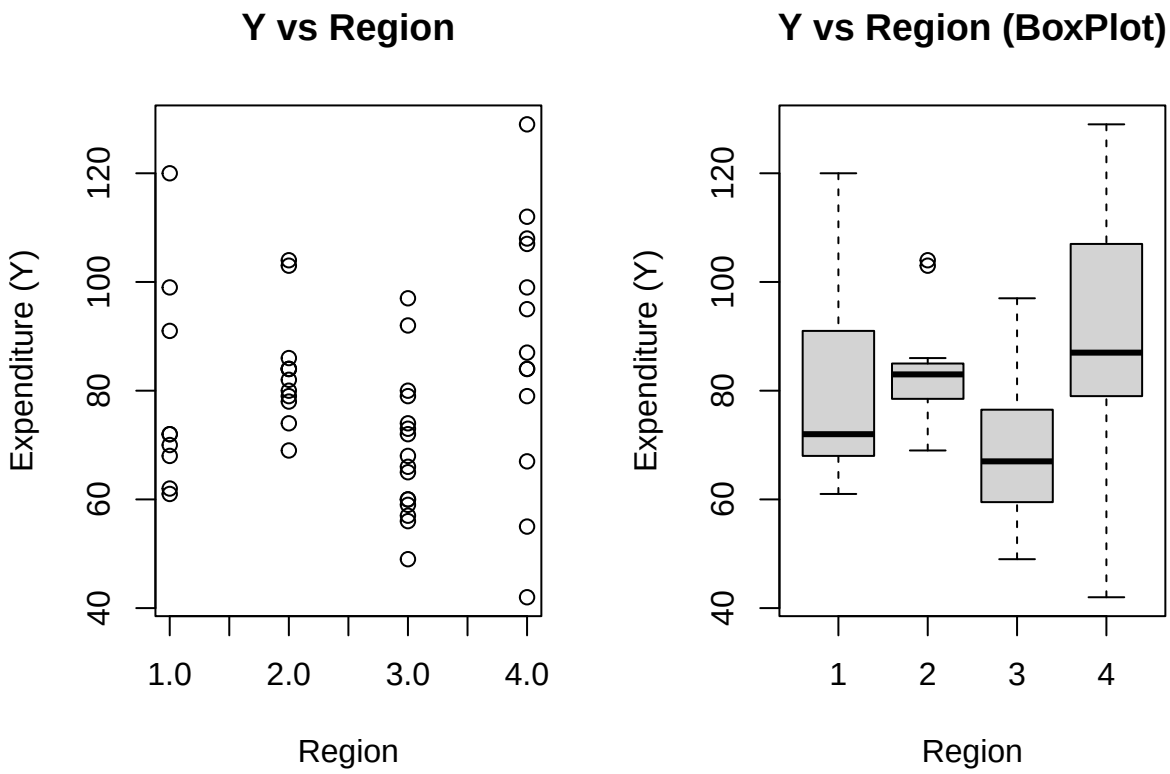



```r
correlation <- cor(expenditure[, c("Y", "X1", "X2",
    "X3")])
print(correlation)
```

```
##            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
```

For the second sub-task in Question 2:

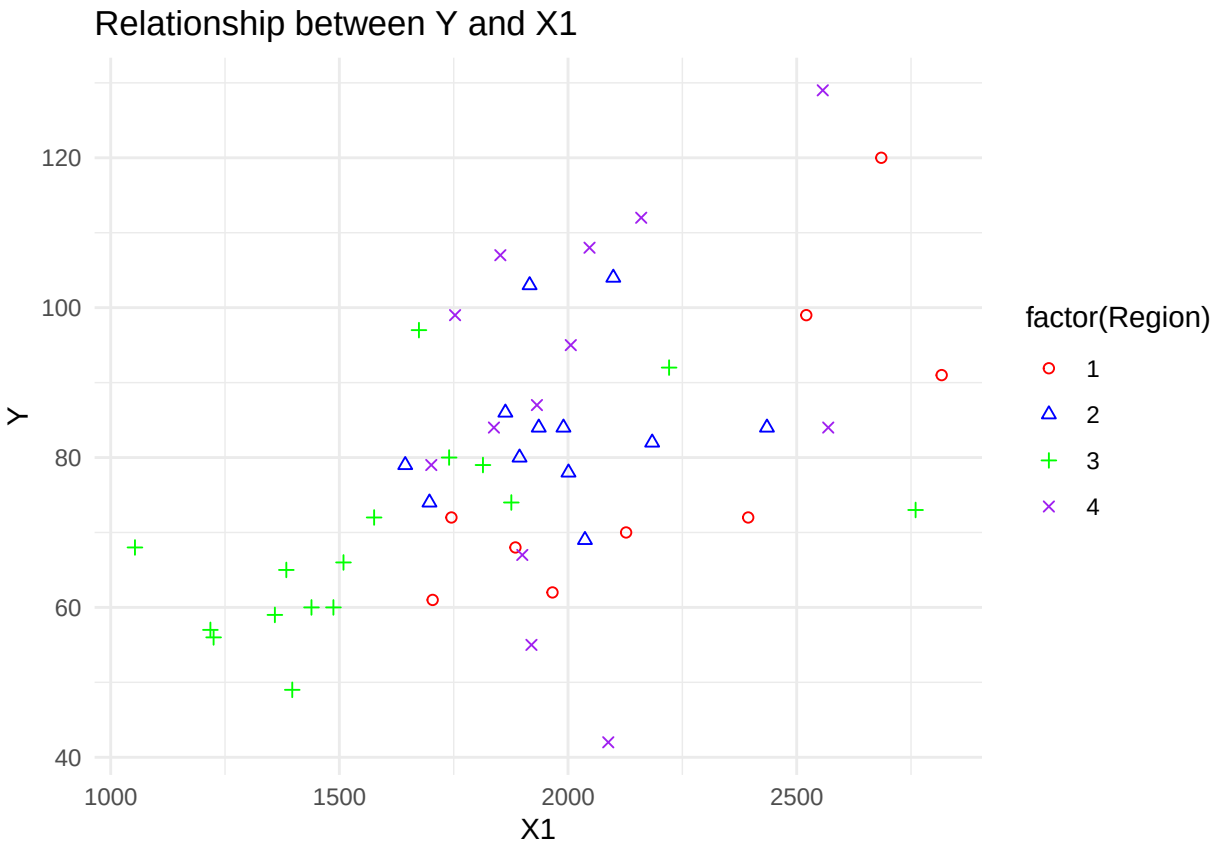
```r
par(mfrow = c(1, 2))
plot(expenditure$Region, expenditure$Y, xlab = "Region",
    ylab = "Expenditure (Y)", main = "Y vs Region")
boxplot(expenditure$Y ~ expenditure$Region, xlab = "Region",
    ylab = "Expenditure (Y)", main = "Y vs Region (BoxPlot)")
```



The figure on the left is the scattor plot and the figure on the right is the box plot of Y and *Region*, where both of them with the same x-axis and y-axis, i.e, the x-axis represents the Region and the y-axis represents the Y (expenditure) values.

For the third sub-task of Question 2:

```
library(ggplot2)
# Create the scatter plot with Y and X1
ggplot(expenditure, aes(x = X1, y = Y)) + geom_point(aes(shape = factor(Region),
    color = factor(Region))) + labs(x = "X1", y = "Y",
    title = "Relationship between Y and X1") + scale_shape_manual(values = c(1,
    2, 3, 4)) + scale_color_manual(values = c("red",
    "blue", "green", "purple")) + theme_minimal()
```


Relationship between Y and X1
120
100
80
60
40
Y
1000
1500
2000
2500
X1
factor(Region)
1
2
3
4