Problem Set 1

Applied Stats/Quant Methods 1

Due: September 30, 2024

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 Monday September 30, 2024. No late assignments will be accepted.

Question 1: 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
mean_y <- mean(y)

se_y <- sd(y) / sqrt(length(y))

critical_value <- 1.645

upper_90 <- (mean(y)) + (critical_value) * (sd(y) / sqrt(length(y)))

```

```
lower_90 <- (mean(y)) - (critical_value) * (sd(y) / sqrt(length(y)))

lower_90
mean_y
upper_90
```

The school counselor conducted a study by taking a random sample of 25 students' IQ scores to estimate the average IQ of all students in the school. After analyzing the data, the sample mean IQ was found to be 98.44.

A 90% confidence interval was calculated to estimate the range within which the true average IQ of all students in the school is likely to fall.

The 90% confidence interval for the mean IQ ranges from 94.13 to 102.75. This means that we are 90% confident that the true average IQ of the students in the school lies between 94.13 and 102.75.

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 $\alpha = 0.05$.

```

t_test_result <- t.test(y, mu = 100, alternative = ”greater”, conf.level
    = 0.95)

t_test_result
```

The one-sample t-test was conducted to determine if the average student IQ at the counselor's school is higher than the national average of 100. With a t-value of -0.596 and a p-value of 0.7215, which is greater than the significance level of 0.05, there is insufficient evidence to reject the null hypothesis. Therefore, we conclude that the average IQ of students in this school is not significantly higher than the national average. In fact, the sample mean of 98.44 suggests the average IQ might even be slightly lower, but this difference is not statistically significant.

Question 2: 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`.

```
upper_90
```

- 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)?

```

expenditure <- read.table("https://raw.githubusercontent.com/ASDS-TCD/StatsI_Fall2024/main/datasets/expenditure.txt", header=T)

pkgTest <- function(x) {
  if (!require(x, character.only = TRUE)) {
    install.packages(x, dep = TRUE)
    if (!require(x, character.only = TRUE)) stop("Package not found")
  }
}

required_packages <- c("ggplot2", "corrplot", "GGally")

lapply(required_packages, pkgTest)

library(GGally)
ggpairs(expenditure, columns = c("Y", "X1", "X2", "X3"))

library(corrplot)
cor_matrix <- cor(expenditure[, c("Y", "X1", "X2", "X3")], use = "complete.obs")

print(cor_matrix)
pdf()
corrplot(cor_matrix, method = "circle", type = "upper", tl.col = "black", tl.srt = 45)
dev.off()
```

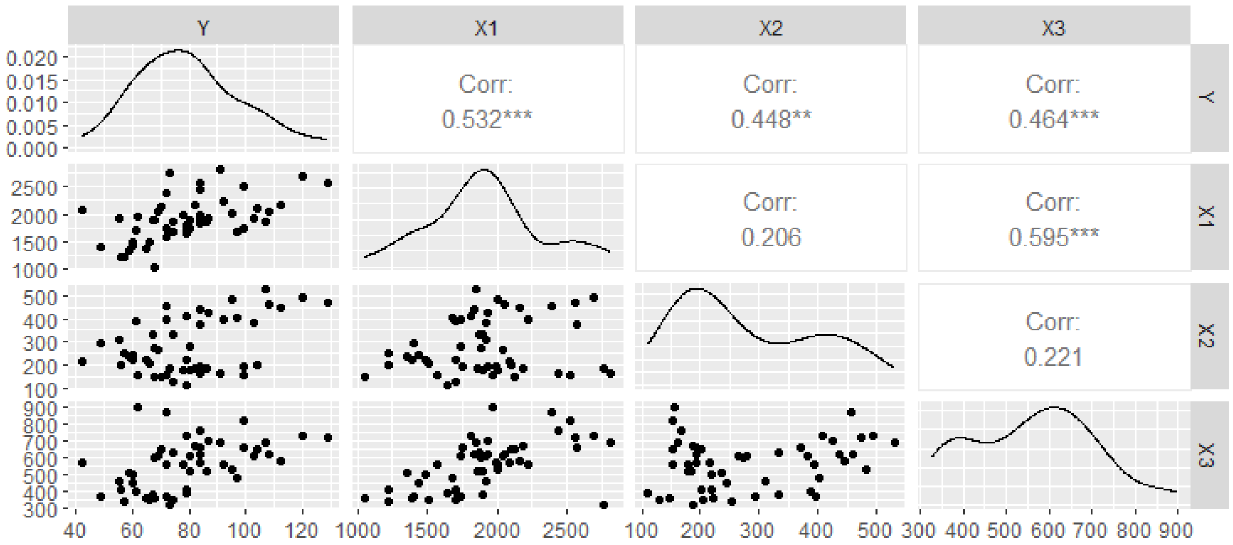

Figure 1: Relationship and Correlation of Variables

The pair plot shows the relationships among the variables Y (per capita expenditure on housing assistance), X1 (per capita personal income), X2 (number of financially insecure residents), and X3 (urban population). There is a moderate positive correlation between Y and X1 (0.532), indicating that states with higher personal income tend to spend more on housing assistance. Similarly, Y shows a positive correlation with both X2 (0.448) and X3 (0.464), suggesting that states with more financially insecure and urban residents also tend to allocate more towards homelessness expenditures. The strongest relationship is between X1 and X3 (0.595), reflecting that wealthier states generally have larger urban populations. Overall, the data indicate that personal income, urbanization, and financial insecurity are moderately associated with increased housing assistance expenditure.

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

```r
library(ggplot2)

expenditure$Region <- factor(expenditure$Region,
                              levels = c(1, 2, 3, 4),
                              labels = c("Northeast", "North Central", "South", "West"))

ggplot(expenditure, aes(x = Region, y = Y)) +
  geom_boxplot(fill = "lightblue", color = "black") +
  labs(title = "Per Capita Expenditure on Housing Assistance by Region",
       x = "Region",
       y = "Per Capita Expenditure (Y)") +
  theme_minimal()
```

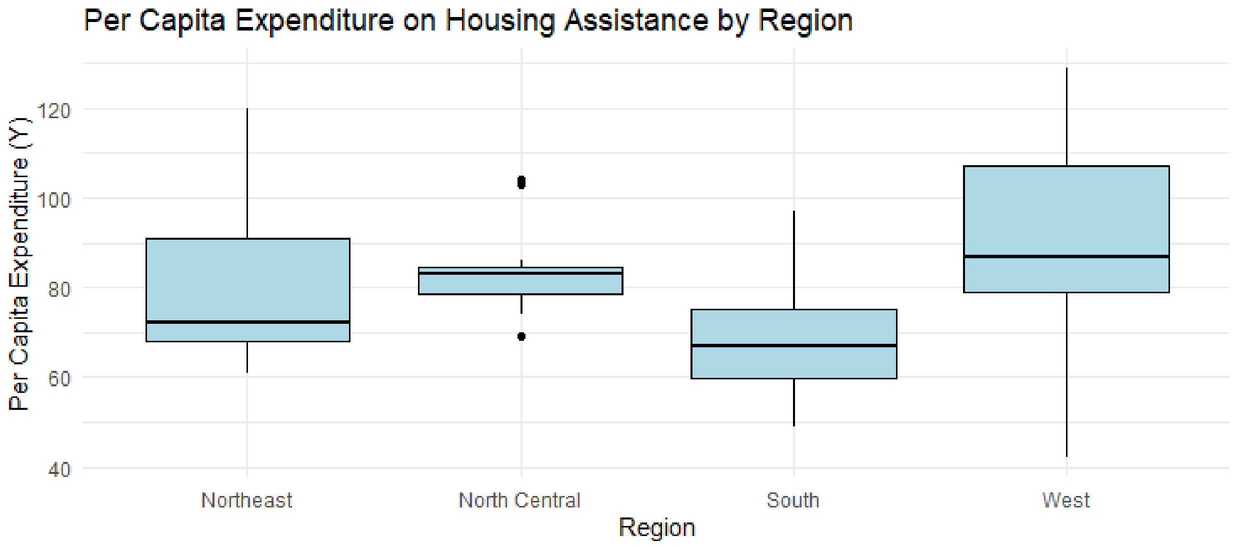

Figure 2: Housing Assistance Expenditure by Region

The boxplot illustrates the relationship between per capita expenditure on housing assistance (Y) and different regions. The black line inside each box represents the median per capita expenditure for that region, which is the midpoint of the data. The West region has the highest median per capita expenditure, around 90, followed closely by the Northeast, with a median just below 80. The North Central region has a median expenditure of around 78, while the South has the lowest median, around 65. The West also displays the widest range of expenditures, with values ranging from approximately 60 to over 120, indicating significant variability in spending across states in this region. In contrast, the North Central and South regions show much narrower ranges of spending, with the North Central spanning from about 75 to 85. Overall, the West not only has the highest median expenditure but also the greatest variability in per capita spending.

- 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.

```r
library(ggplot2)

ggplot(expenditure, aes(x = X1, y = Y)) +
  geom_point(color = "blue", size = 3) +  # Scatterplot with blue points
  labs(title = "Relationship Between Per Capita Expenditure (Y) and Income (X1)",
       x = "Per Capita Personal Income (X1)",
       y = "Per Capita Expenditure on Housing Assistance (Y)") +
  theme_minimal()
```

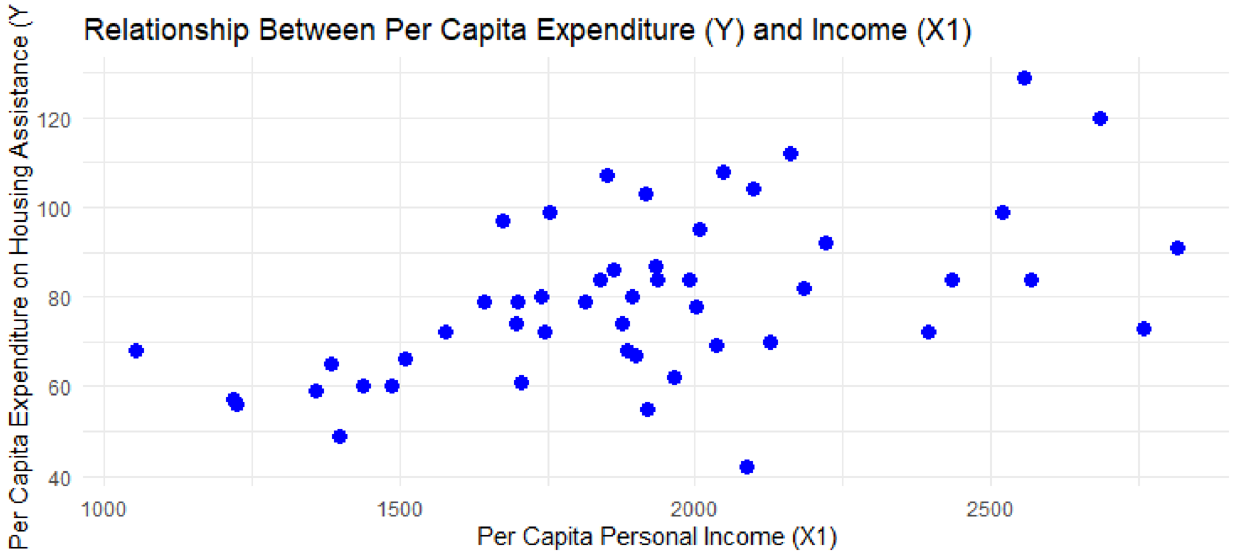

Figure 3: Relationship of Per Capita Expenditure of Housing Assistance and Per Capita Personal Income

The scatterplot shows the relationship between per capita personal income (X1) and per capita expenditure on housing assistance (Y). There appears to be a moderate positive correlation between the two variables, as higher personal income is generally associated with higher spending on housing assistance. Most data points are clustered between 1,500USD and 2,000USD for income, with expenditures ranging from 60USD to 100USD. A few outliers exist, with some states showing higher expenditures (up to 120USD) despite varying income levels. Overall, as income increases, there is a tendency for higher housing expenditure, though with some variability.

```r
expenditure <- read.table("https://raw.githubusercontent.com/ASDS-TCD/StatsI_Fall2024/main/datasets/expenditure.txt", header=T)

library(ggplot2)

expenditure$Region <- factor(expenditure$Region,
                             levels = c(1, 2, 3, 4),
                             labels = c("Northeast", "North Central", "South", "West"))

ggplot(expenditure, aes(x = X1, y = Y, color = Region, shape = Region)) +
  geom_point(size = 3) +
  labs(title = "Relationship Between Per Capita Expenditure (Y), Income (X1) by Region",
       x = "Per Capita Personal Income (X1)",
       y = "Per Capita Expenditure on Housing Assistance (Y)") +
  theme_minimal() +
  theme(legend.title = element_text(face = "bold"),
        legend.position = "right")
```

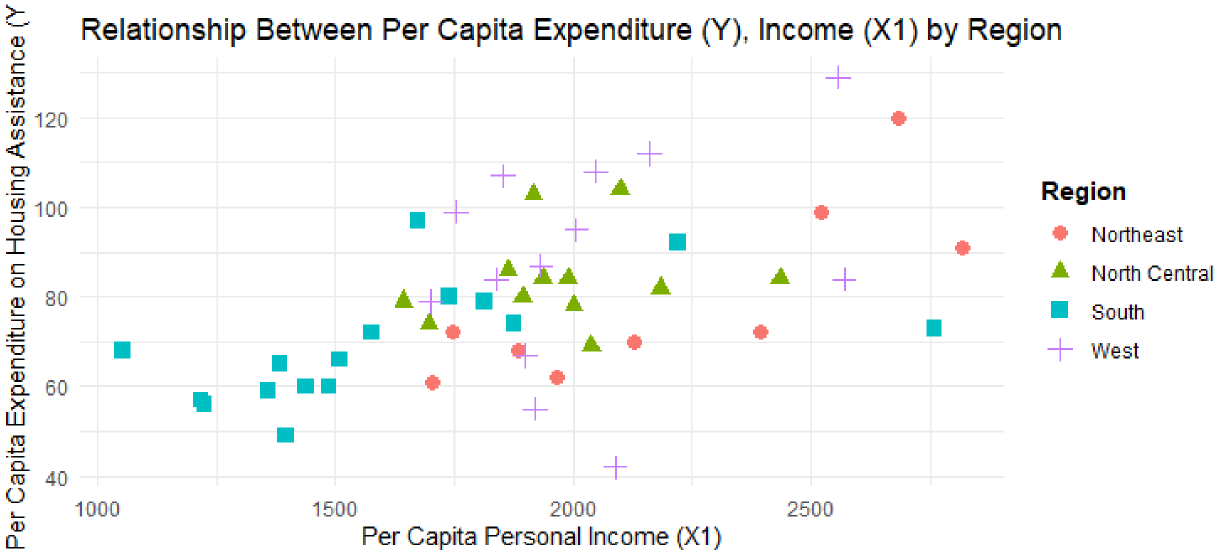

Figure 4: Relationship of Per Capita Expenditure on Housing Assistance and Per Capita Personal Income by Region

The scatterplot shows a positive relationship between per capita income (X1) and housing expenditure (Y), with regions distinguished by color and shape. The West and Northeast show higher and more variable expenditures, while the South and North Central have lower, more consistent expenditures. Overall, the graph shows again, that higher income is generally associated with greater housing assistance spending, with notable regional differences in variability.

Submitted by Juliane Wesselmann