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.

```
1
2 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)
3
4 # Section1:Calculate the 90% Confidence Interval
5
6 # Given the sample size is less than 30(n=25),so we can use t-test to
    calculate the confidence interval
7 # Sample size
8 n <- length(y)
9
10 # Set the significance level and get the t-critical value
11 alpha <- 0.10
12 t_critical <- qt(0.05, df=n-1)
```

```r

# Calculate the standard deviation of the sample
sample_sd <- sd(y)

# This value reflects the uncertainty around the sample mean
margin_of_error <- t_critical * (sd(y) / sqrt(n))

#Calculate the confidence interval limits
CI_lower <- mean(y) - margin_of_error
CI_upper <- mean(y) + margin_of_error

#print confidence interval
print(c(CI_lower, mean(y), CI_upper))

# The calculated 90% confidence interval is approximately [93.95, 102.92]
```

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
# Section2:Conduct the Hypothesis Test

#Perform a one-sample t-test to test whether the hypothesized mean is significantly greater than 100
# The null hypothesis (H0) is that the true mean IQ is equal to 100.
# The alternative hypothesis (H1) is that the true mean IQ is greater than 100.

#Perform the t-test
t_test_result <- t.test(y, mu = 100, alternative = "greater", conf.level = 0.95)

#Print the result of the t-test
print(t_test_result)

# The output includes:
# t-value: -0.59574
# p-value: 0.7215, which is quite large and suggests we fail to reject the null hypothesis.
# Alternative hypothesis: true mean is greater than 100.
# 95% Confidence Interval: [93.96, Inf] - The lower bound is 93.96, with no upper bound (as it's a one-tailed test).
# Sample mean: 98.44, which is less than 100
# aligning with the test's result that the average IQ is not significantly greater than 100.
```

The average IQ of students in her school is higher than the average IQ

Part I
title

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*

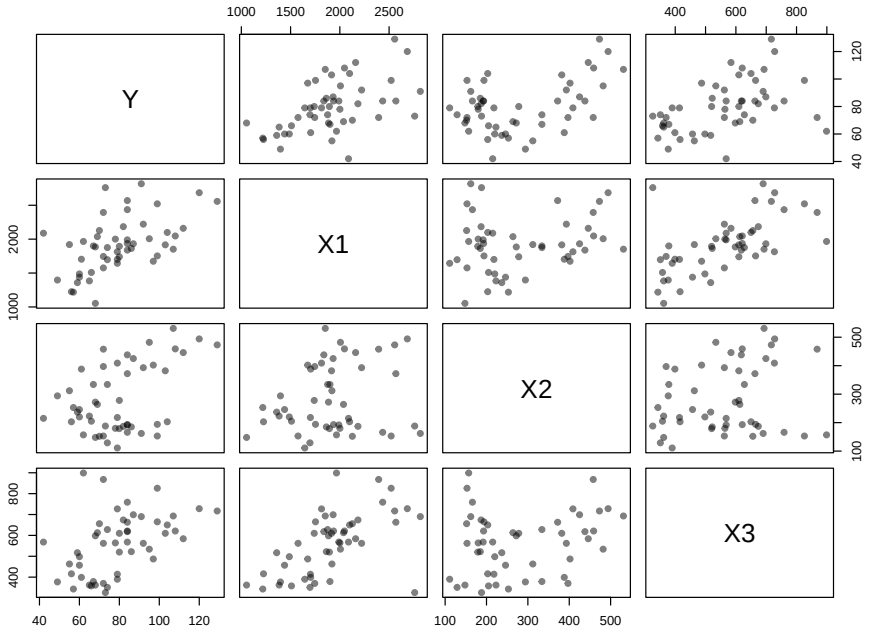
Explore the `expenditure` data set and import data into `R`.

- 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)?

```r
# Load required libraries
library(ggplot2)
library(dplyr)
library(corrplot)

# Import the data
expenditure <- read.table("https://raw.githubusercontent.com/ASDS-TCD/StatsI_Fall2024/main/datasets/expenditure.txt", header = TRUE)
head(expenditure)

# Rename columns for clarity
colnames(expenditure) <- c("State", "Y", "X1", "X2", "X3", "Region")

# Convert Region to factor
expenditure$Region <- factor(expenditure$Region,levels = 1:4,labels = c("Northeast", "North Central", "South", "West"))

# Create correlation matrix
cor_matrix <- cor(expenditure[, c("Y", "X1", "X2", "X3")])
print(cor_matrix)

# section 1:Plot relationships among Y, X1, X2, and X3
# Create scatter plots for all bivariate relationships
pdf(file="scatter_Matrix.pdf",width=8,height=6)

```

```r
# Plot the pairwise scatterplots between Y, X1, X2, X3
pairs(expenditure[, c("Y", "X1", "X2", "X3")], pch = 19, col = alpha("black", 0.5))
dev.off()

# The correlation among Y, X1, X2, and X3
# Y and X1 show a moderate positive correlation (~0.53), suggesting that states with higher personal income tend to spend more on housing assistance.
# Y and X2 have a moderate positive correlation (~0.45), indicating a link between the number of financially insecure residents and housing expenditure.
# Y and X3 also show a moderate positive correlation (~0.46), implying that urbanization (X3) has some relationship with higher housing assistance spending.
# The relationships between X1 and X2, X1 and X3, and X2 and X3 show varying degrees of correlation.
```

Figure 1: scatter Matrix



The scatter plots show a moderate positive relationship between per capita expenditure on housing assistance (Y) and each of the predictors: personal income (X1), financially insecure residents (X2), and urban population percentage (X3). This suggests that states with higher incomes, larger urban populations, and more financially vulnerable residents tend to spend more on addressing homelessness.

Figure 2: Y X1

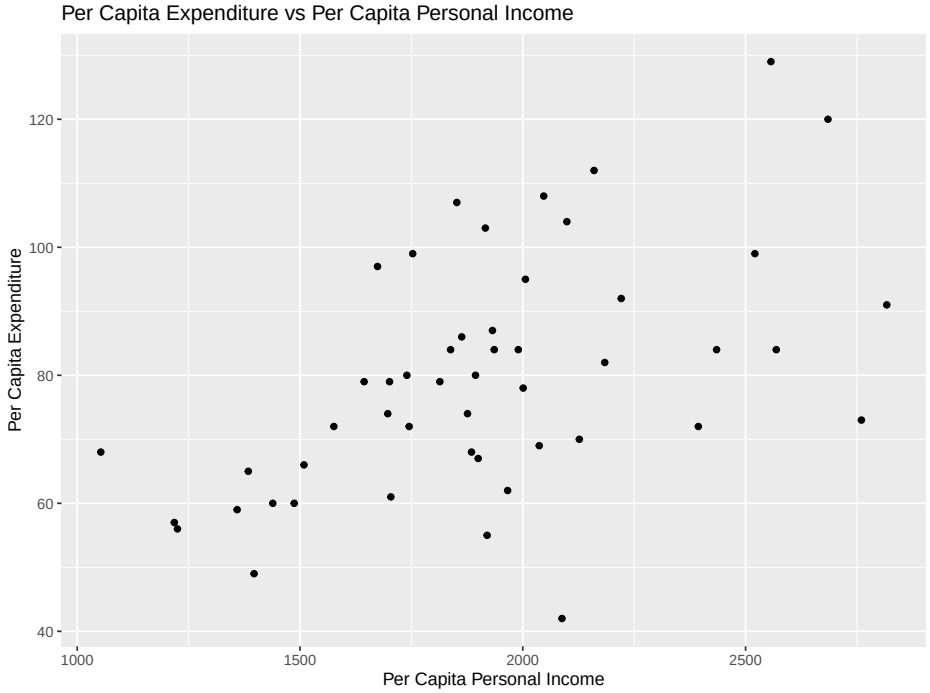


Figure 3: X2 Y

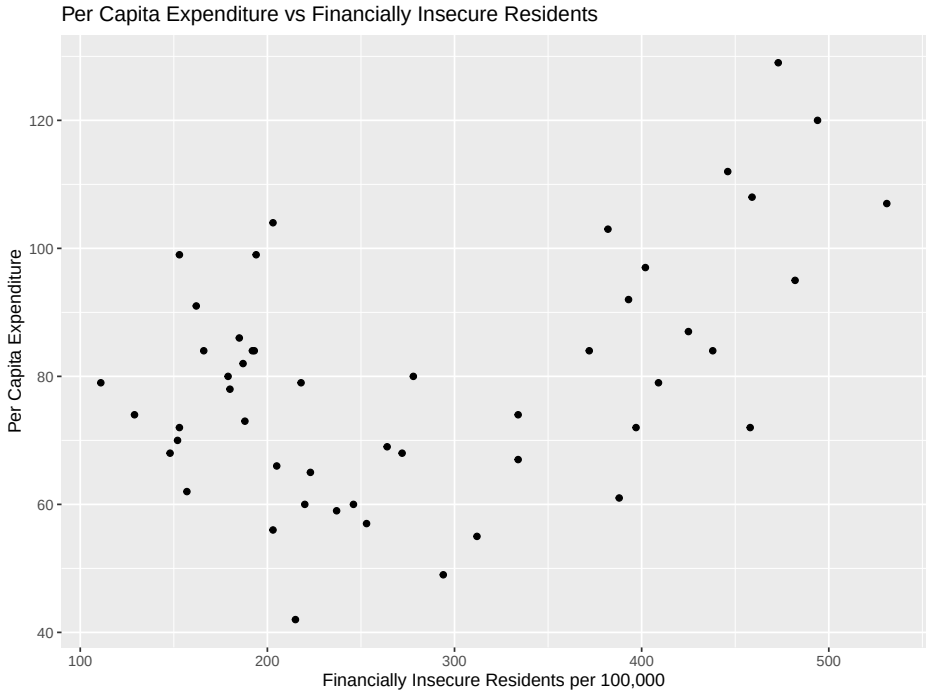


Figure 4: X3 Y

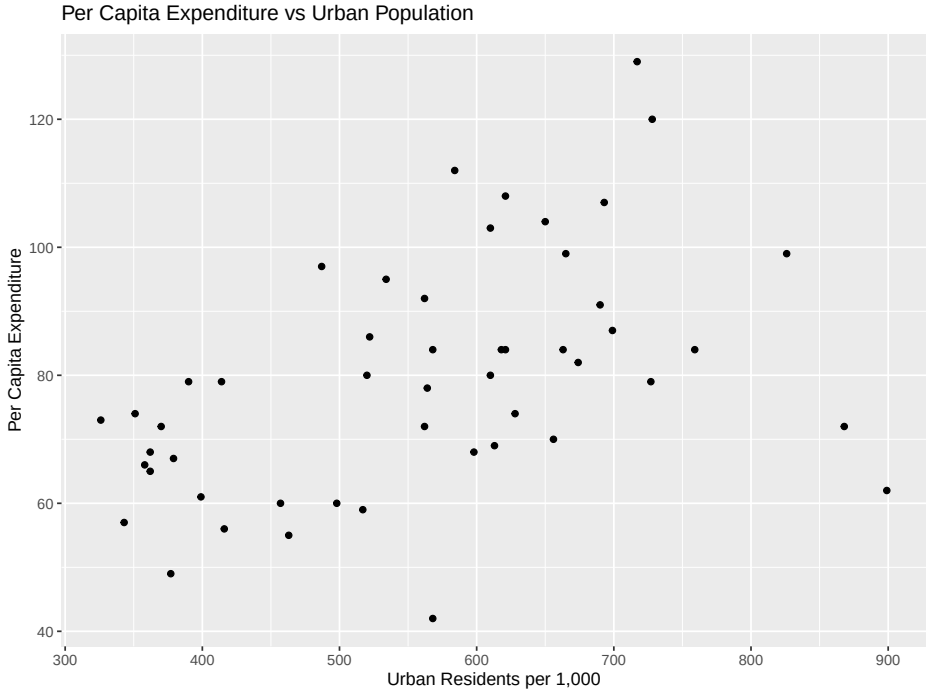


Figure 5: X1 X2

Figure 6: X1 X3

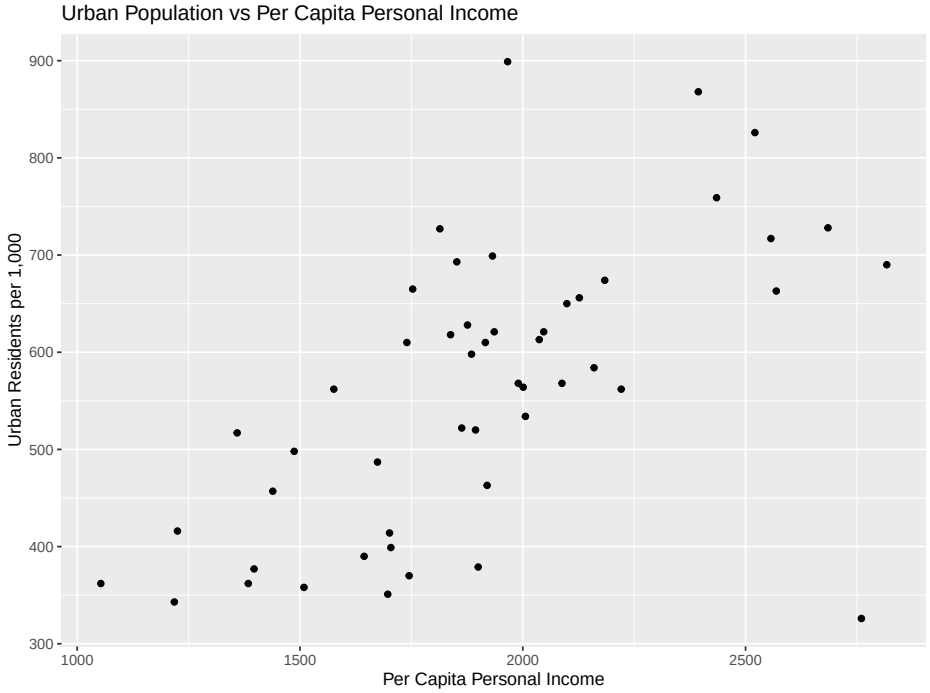
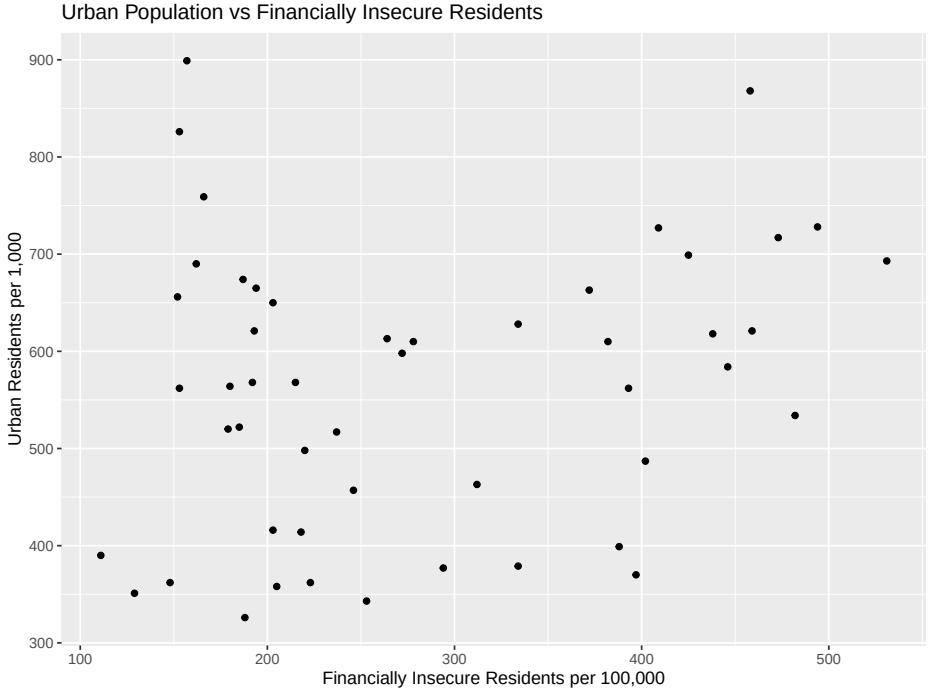


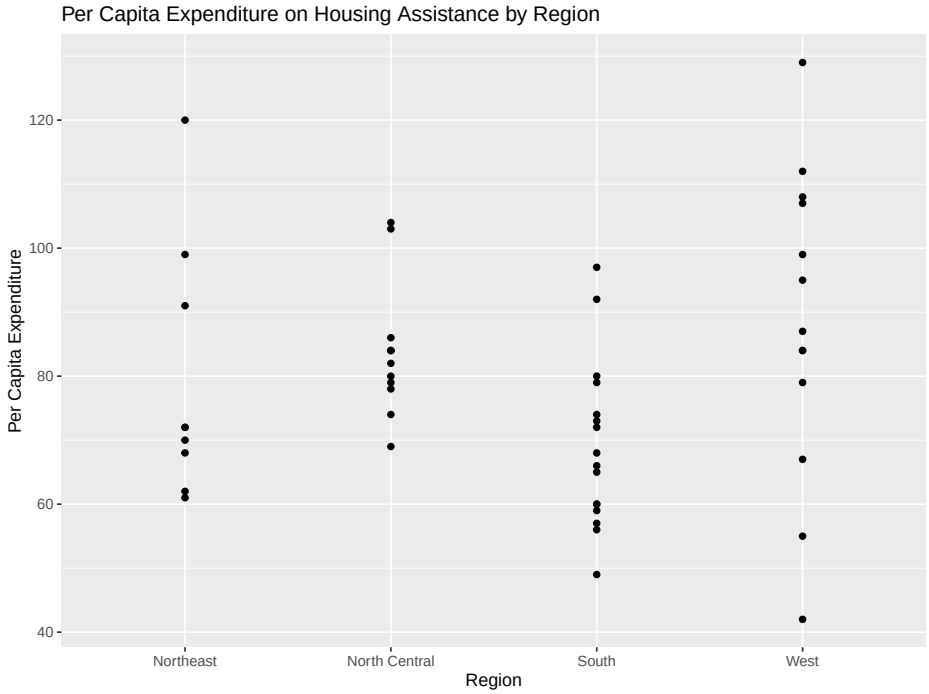
Figure 7: X2 X3



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

```
#section 2: Plot the relationship between Y and Region
Y_vs_Region <- ggplot(expenditure, aes(x = Region, y = Y)) +
  geom_point() +
  labs(title = "Per Capita Expenditure on Housing Assistance by Region",
       x = "Region", y = "Per Capita Expenditure")
ggsave("Y_VS_Region.pdf",plot=Y_vs_Region,width=8,height=6,units="in")

# Based on the plot, Region 1 (Northeast) has the highest average per
    capita expenditure on housing assistance.
```

Figure 8: Y Region

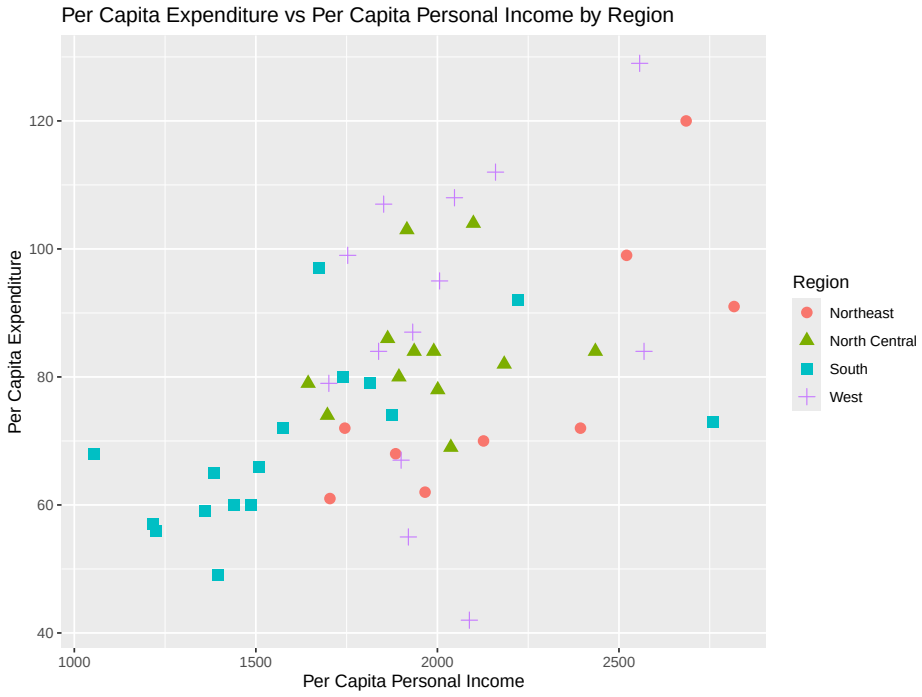


Please plot the relationship between Y and *X1*? Describe this graph and the relationship. Reproduce the above graph including one more variable *Region* and display different regions with different types of symbols and colors.

```
# Section3:Plot the relationship between Y and X1
Y_vs_X1 <- ggplot(expenditure, aes(x = X1, y = Y)) +
  geom_point() +
  labs(title = "Per Capita Expenditure vs Per Capita Personal Income",
       x = "Per Capita Personal Income", y = "Per Capita Expenditure")
ggsave("Y_vs_X1.pdf",plot=Y_vs_X1,width=8,height=6,units="in")

# The scatterplot shows a positive linear relationship between Y and X1 (
    per capita income).
```

```r
# Higher personal income in a state tends to be associated with higher spending on housing assistance.

# Reproduce the Y vs X1 scatterplot, adding Region as a distinguishing factor
Y_X1_Region <- ggplot(expenditure, aes(x = X1, y = Y, color = Region, shape = Region)) +
  geom_point(size = 3) +
  labs(title = "Per Capita Expenditure vs Per Capita Personal Income by Region",
       x = "Per Capita Personal Income", y = "Per Capita Expenditure")
ggsave("Y_X1_Region.pdf",plot=Y_X1_Region,width=8,height=6,units="in")

# The plot shows that different regions have different expenditure trends with respect to personal income.
# Region 1 (Northeast) tends to have higher spending across the income spectrum, while Region 3 (South) generally shows lower expenditure.
```

Figure 9: Y X1 Region



The scatter plot of Y (per capita expenditure on housing assistance), X1 (personal income), and Region shows a positive relationship between Y and X1, with higher-income states spending more on housing assistance. However, regional differences are clear, with the Northeast and West regions generally spending more at similar income levels compared to the South and North Central. This suggests that both personal income and regional factors play significant roles in determining housing assistance expenditures.