# 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 IQ of the students in her school and took a random sample of 25 students' IQ scores. The following is the data set:

```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)
```

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

```r
# First calculate the mean of IQ scores and the std. deviation
mean_y <- mean(y)
sd_y <- sd(y)

# And now the 90% CI (formula: mean(y) +/- z*(sd/sqrt(n)))
upper_90_y <- mean(y) + 1.645 * (sd_y/sqrt(length(y)))
lower_90_y <- mean(y) - 1.645 * (sd_y/sqrt(length(y)))
```

The 90% CI for the average student IQ score: 94.13244 (lower), 102.7476 (upper).

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
# Doing an one-sided t.test because we care about the IQ being higher than the average!
result <- t.test(y, mu = 100, alternative = "greater")

# Print the output
result
```

```
One Sample
 t-test data:  yt = -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
```

   The sample mean is 98.44 (lower than 100) and the p-value is greater than 0.05. We cannot reject the null hypothesis: There is not enough evidence that the average IQ is greater than 100.

# 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`.

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

```
head(expenditure)
```

```
  STATE  Y   X1  X2  X3 Region
1    ME 61 1704 388 399      1
2    NH 68 1885 272 598      1
3    VT 72 1745 397 370      1
4    MA 72 2394 458 868      1
5    RI 62 1966 157 899      1
6    CT 91 2817 162 690      1
```

```
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
```

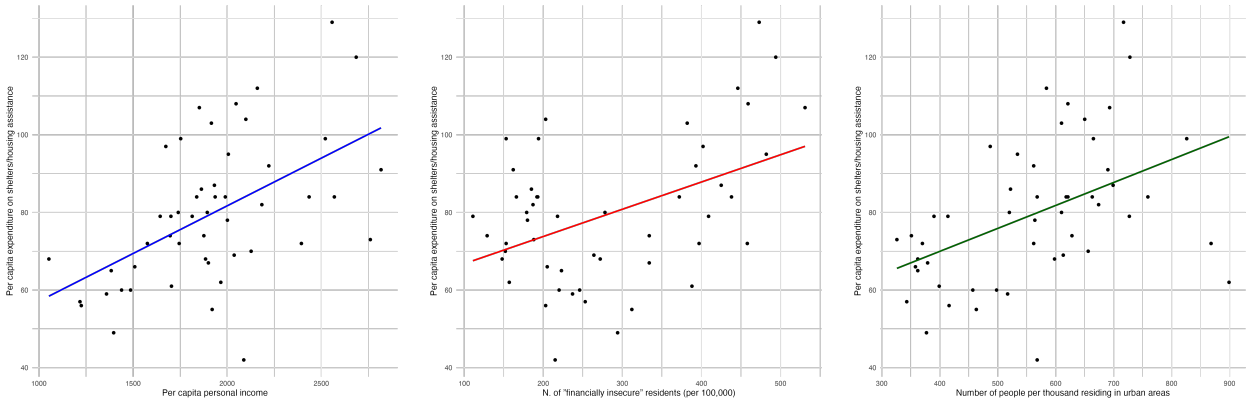
- Please plot the relationships among *Y*, *X1*, *X2*, and *X3*. What are the correlations among them (you just need to describe the graph and the relationships among them)?

```r
# X1
p1 <- ggplot(expenditure, aes(x = X1, y = Y)) +
  geom_point() +
  geom_smooth(method = "lm", se = F, color = "blue") +
  labs(x = "Per capita personal income", y = "Per capita expenditure on
    shelters/housing assistance") +
  theme_minimal()
ggsave("p1.png", p1)

# X2
p2 <- ggplot(expenditure, aes(x = X2, y = Y)) +
  geom_point() +
  geom_smooth(method = "lm", se = F, color = "red") +
  labs(x = "N. of financially insecure residents (per 100,000)",
       y = "Per capita expenditure on shelters/housing assistance") +
  theme_minimal()
ggsave("p2.png", p2)

# X3
p3 <- ggplot(expenditure, aes(x = X3, y = Y)) +
  geom_point() +
  geom_smooth(method = "lm", se = F, color = "darkgreen") +
  labs(x = "Number of people per thousand residing in urban areas",
       y = "Per capita expenditure on shelters/housing assistance") +
  theme_minimal()
ggsave("p3.png", p3)
```

The correlations depicted by each of the three plots differ. Plot (a) shows a positive correlation between per capita personal income ($X1$) and state expenditure on housing/shelters (Y), though many values deviate significantly from the line, indicating substantial variability. Plot (b) illustrates the relationship between the number of 'financially insecure' residents ($X$) and expenditure. Here, the relationship is less straightforward, forming a parabola-like pattern. Most states have relatively few financially insecure residents, and their expenditure levels vary widely. A small number of states have both a large financially insecure population and high housing assistance expenditure, while the middle ground features fewer cases. Finally, plot (c) depicts the relationship between the number of people residing in urban areas ($X3$) and state housing expenditure. Similar to plot (a), this relationship appears positive and roughly




(a) X1 (b) X2 (c) X3

Figure 1: Relationship between Y and X1, X2, X3 variables.

linear, suggesting that states with larger urban populations tend to spend more on shelters and housing assistance.

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

```
p4 <- ggplot(expenditure, aes(x = factor(Region), y = Y)) +
  geom_boxplot(fill = "lightblue", color = "darkblue") +
  labs(x = "Region", y = "Per capita expenditure on shelters/housing
    assistance",
       title = "Expenditure by region") +
  scale_x_discrete(labels = c("1" = "Northeast", "2" = "North Central", "
    3" = "South", "4" = "West")) +
  theme_minimal()
```

On average, out of the four regions, Western states have the highest expenditure per capita for shelters/housing assistance. However, the CIs are larger in comparison to the rest of the regions. Therefore, more analysis is needed to see why there is so much variation within Western states.

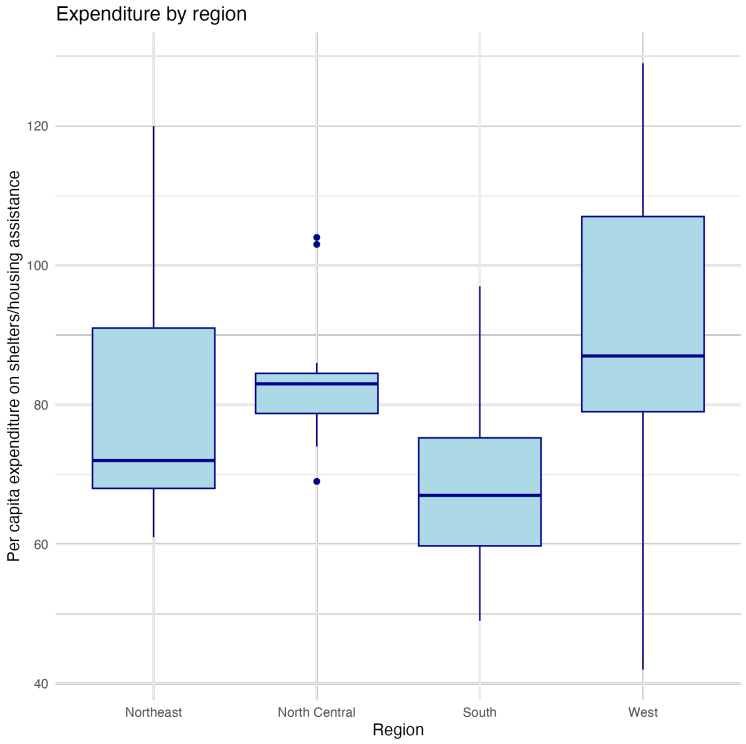



Figure 2: Boxplots of expenditure by region.

- 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.

```
# Plot
p5 <- ggplot(expenditure,
             aes(x = X1,
                 y = Y,
                 color = factor(Region),
                 shape = factor(Region))) +
  geom_point(size = 2.5, alpha = 0.75) +
  labs(title = "Relationship between Y and X1 by Region",
       x = "Per capita personal income",
       y = "Per capita expenditure on shelters/housing assistance",
       color = "Region",
       shape = "Region") +
  scale_color_manual(values = c("1" = "red", "2" = "blue", "3" = "green",
    "4" = "purple"),
                     labels = c("1" = "Northeast", "2" = "North Central",
    "3" = "South", "4" = "West")) +
  scale_shape_manual(values = c(16, 17, 18, 19),
                     labels = c("1" = "Northeast", "2" = "North Central",
    "3" = "South", "4" = "West")) +
  theme_minimal()
```

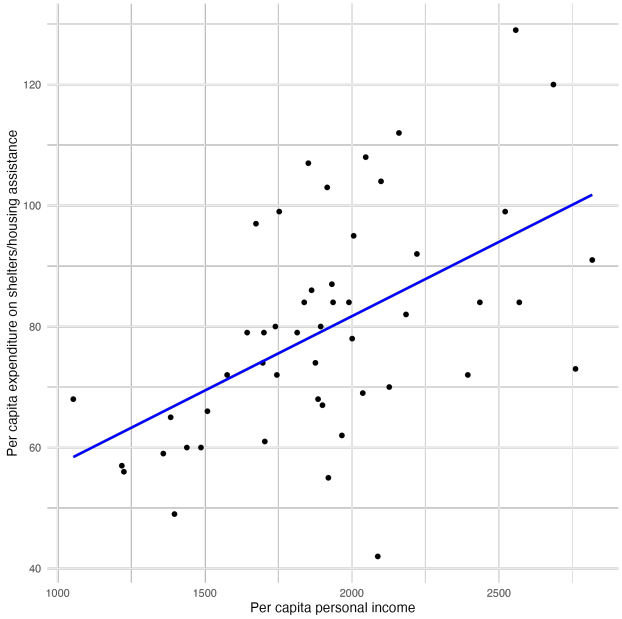



Figure 3: Relationship between expenditure on shelters/housing assistance and per capita personal income in state.

In Figure 3, we can see is a positive relationship between the state's expenditure on shelters and housing assictance ($Y$) and the personal income per capita in state ($X1$). The higher the personal income, the more money the state spends on shelters/housing. However, the line does not perfectly fit the data, indicating that there are other confounders that influence state's expenditue.

When we differentiate by 'Region' (see Figure 4), the correlation looks quite different and a linear relationship is harder to distinguish. For example, Southern states create a cluster on the bottom right of the plot, implying lower per capita expenditure on housing while also having lower personal income. Meanwhile, Western states have an average PCI but vary a lot in their expenditure levels (which also explains the large CIs in Firure 2).

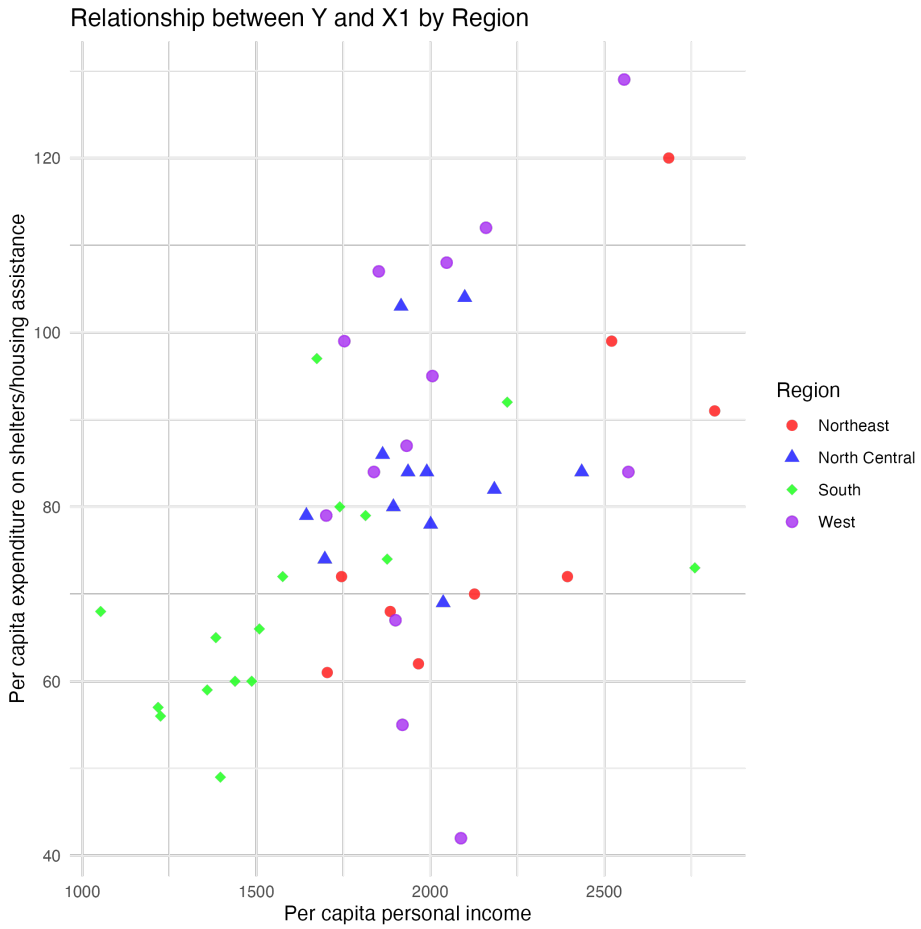



Figure 4: Relationship between state expenditure on shelters/housing assistance and per capita personal income by region.