Problem Set 1

Eleni Karagianni

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:

```
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.

```
# 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
# Setting up the hypotheses:
# H0 (null): The average IQ score of the counselor's students is not greater than 100.
# H1 (altenative): The average IQ score of the counselor's students is greater than 100.

# How many observations do we have?
n <- length(y) # 25 -> t-statistic because n < 30

# Calculate the standard error
se_y <- sd_y / sqrt(n)

# Calculate the t-statistic
t_stat <- (mean_y - 100) / se_y

# ...and the p-value
p_value <- (1 - pt(t_stat, df = n - 1)) # 0.7215383

# To confirm:
# 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`.

With the command 'summary' I explored the main measures of central tendency to get an image about the range of the values included in the dataset. The results can be seen below.

```
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)?

```
ggpairs(expenditure, columns = c("Y", "X1", "X2", "X3"),
        title = "Pairwise Correlation Matrix",
```

```
        upper = list(continuous = wrap("cor", size = 4, color = "darkblue
    "), combo = wrap("cor", size = 3)),
        lower = list(continuous = "smooth", combo = "smooth",
                     continuous_params = lisIf t(color = "darkblue", size
    = 0.5)),
        diag = list(continuous = wrap("barDiag", color = "darkblue"),
                    discrete = wrap("barDiag", color = "darkblue"))) +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"),
        axis.text = element_text(size = 10),
        panel.grid.major = element_blank())
dev.off()
```

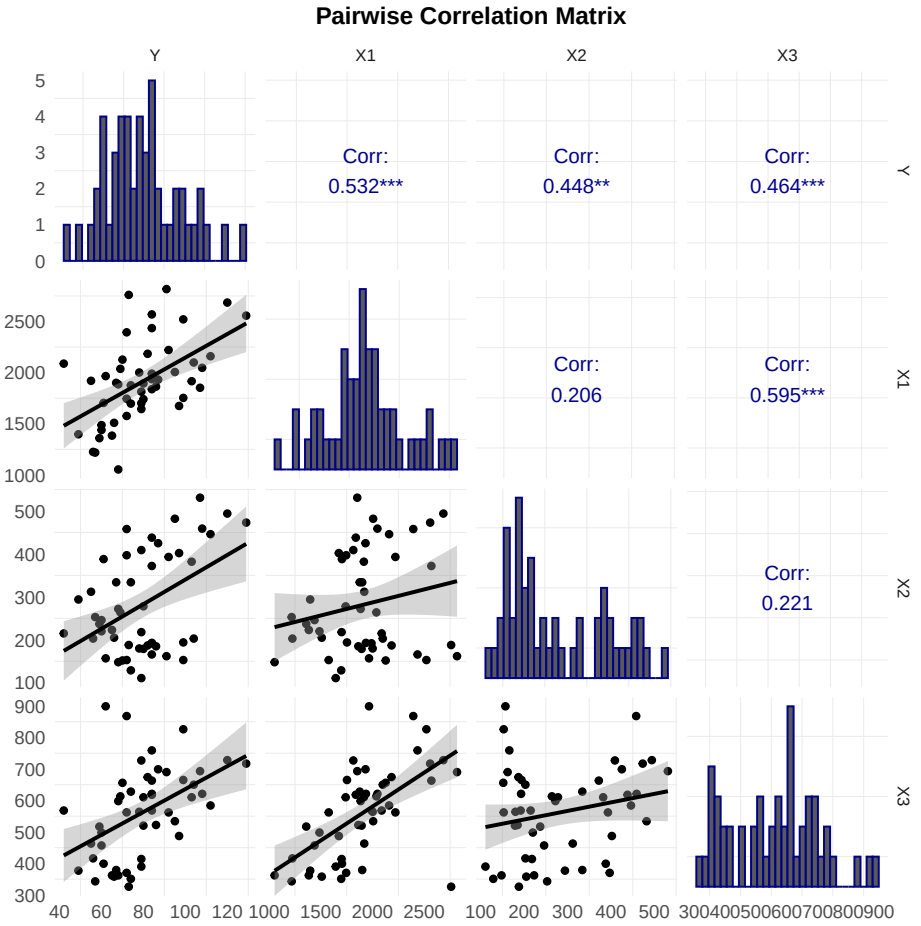

Figure 1: Correlation matrix plot between the output and explanatory variables.

Figure 1 depicts the correlations between the outcome variable (Y), namely the per capite expenditure on shelters/housing assistance in state and the explanatory variables ($X1$, $X2$, $X3$). A relatively strong positive correlation is seen between Y and personal income (PCI) in state (0.532). This might indicate that wealthier states are more likely to allocate resources towards housing assistance. The scatterplot between

Y and $X2$ also shows a positive but weaker relationship. States with a higher number of 'financially insecure' residents tend to spend slightly more on shelters/housing assistance, though the relationship is not as strong as with personal income. For Y and $X3$, the scatterplot reveals another positive relationship with a correlation coefficient of 0.464. States with higher urban populations tend to allocate more resources towards housing assistance.

Focusing on the relationship between the input variables, the association between PCI and number of 'financially insecure' residents is not so strong, with a positive correlation of just 0.206. On the contrary, a strong association exists between higher PCI and residents of urban areas (0.595). Finally, the number of citizens that identify as financial insecure is weakly associated with urban area citizens, suggesting that urbanization alone does not necessarily increase the rate of financial insecurity.

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

```
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() +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"),
        axis.test = element_text(size = 10))
```

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.

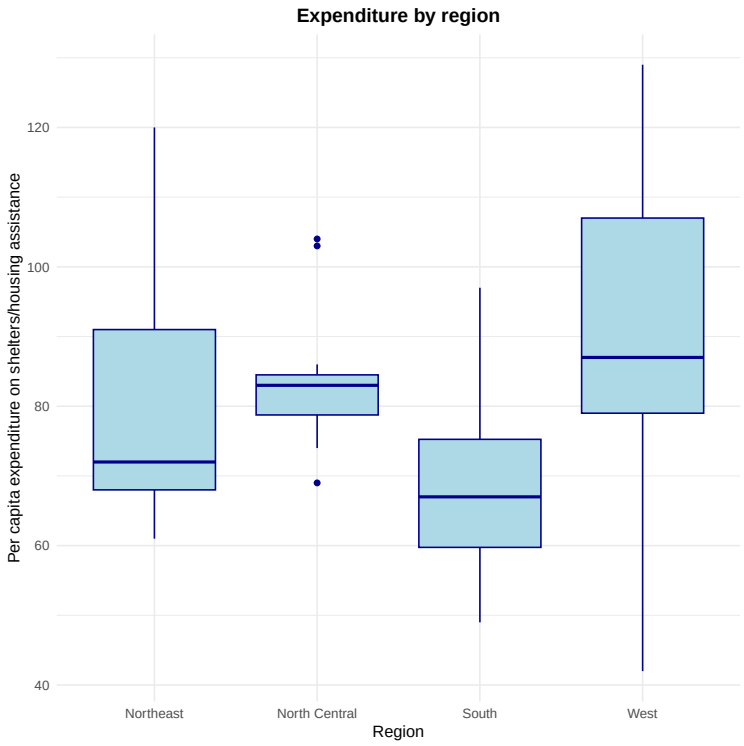

Figure 2: Boxplots of expenditure by region.

- 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.

```
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",
       title = "Expenditure on Shelters/Housing by PCI in state") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"),
        axis.test = element_text(size = 10))
```

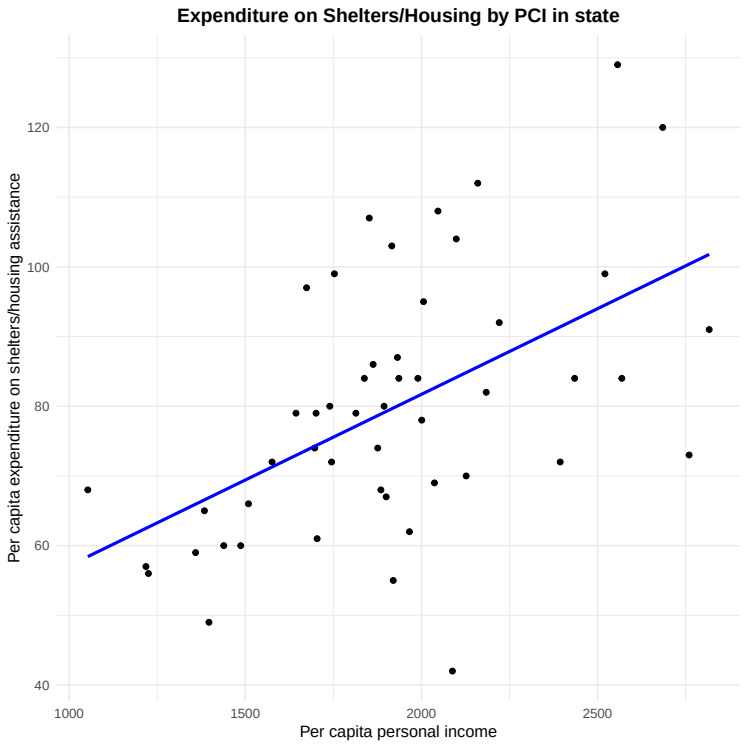

Figure 3: Relationship between expenditure on shelters/housing assistance and per capita personal income in state.

```
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()
```

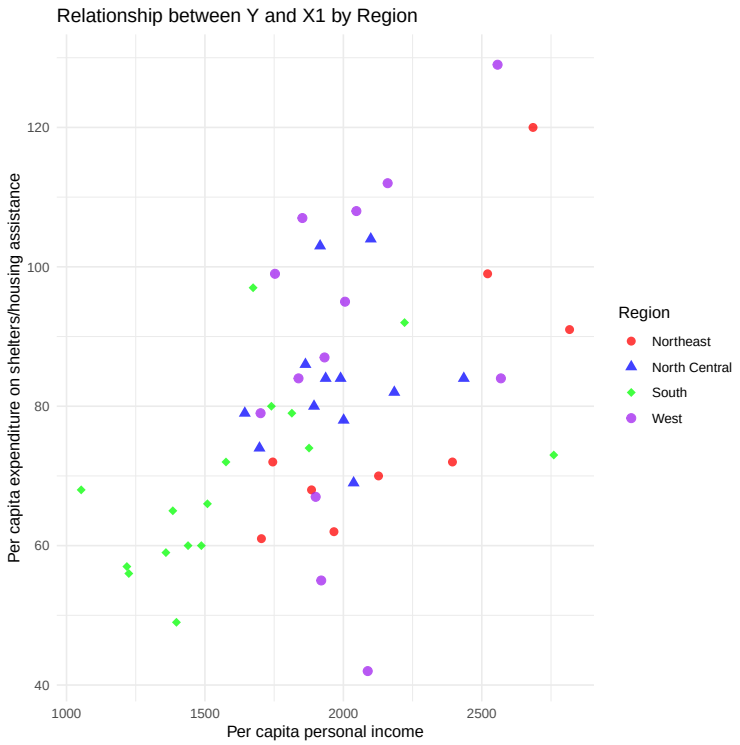

Figure 4: Relationship between state expenditure on shelters/housing assistance and per capita personal income by region.

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).