Applied Stats - Problem Set 1

Yana Konshyna

September 29, 2023

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:

```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.
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 $\alpha = 0.05$.

Answers to Question 1: Education

1. **Code in R:**

```r
# Calculating length, mean and standard deviation of our sample y
length(y)
mean(y)
sd(y)/sqrt(length(y))
# Finding confidence interval using t distribution, because n<30
# critical value
t_score <- qt(0.95, df=length(y)-1)
#margin of error
me<-(t_score)*(sd(y)/sqrt(length(y)))
# Lower bound, 90 confidence level
lower_90_t <- mean(y)-me
# Upper bound, 90 confidence level
upper_90_t <- mean(y)+me
#Print result
conf_int90 <- c(lower_90_t, upper_90_t)
conf_int90
```

Output:

```
> length(y)
[1] 25
> mean(y)
[1] 98.44
> sd(y)/sqrt(length(y))
[1] 2.618575
> # Finding confidence interval using t distribution, because n<30
> # critical value
> t_score <- qt(0.95, df=length(y)-1)
> #margin of error
> me<-(t_score)*(sd(y)/sqrt(length(y)))
> # Lower bound, 90 confidence level
> lower_90_t <- mean(y)-me
> # Upper bound, 90 confidence level
> upper_90_t <- mean(y)+me
> #Print result
> conf_int90 <- c(lower_90_t, upper_90_t)
> conf_int90
[1]  93.95993 102.92007
```

Conclusion: For finding a 90% confidence interval for the average student IQ in the school I used the t distribution, because n less than 30. As result, 90% confidence interval is [93.95993, 102.92007].

2. I set null hypothesis H0: the average student IQ in school is equal to the average IQ score (100) among all the schools in the country. Alternative hypothesis Ha: the average student IQ in school is greater than 100.

```r
# Setting null hypothesis that mean = 100
t.test(y, mu = 100, alternative = "greater")
```

Output:

```
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
```

Conclusion: P-value = 0.7215. It is greater than $\alpha = 0.05$, so I cannot reject null hypothesis. I don't have a significant evidence that the average student IQ in school is higher than the average IQ score (100) among all the schools in the country.

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`.

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

- 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.

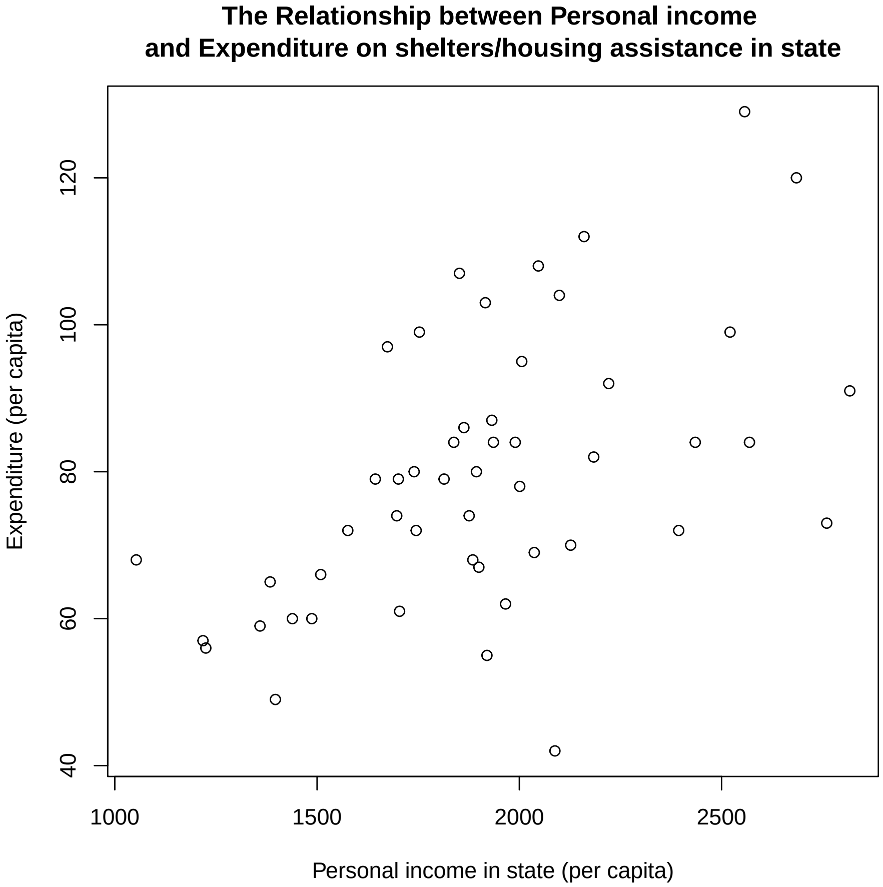
Answers to Question 2: Political Economy

1.1. The relationships between *Y* and *X1*:

```
pdf(file=”C:/Users/HOME/Documents/GitHub/StatsI_Fall2023/problemSets/PS01/my_answers/plot1_1.pdf”)
plot(expenditure$X1,
     expenditure$Y,
     xlab=”Personal income in state (per capita)”,
     ylab=”Expenditure (per capita)”,
     main=”The Relationship between Personal income \nand Expenditure on shelters/housing assistance in state”)
dev.off()
```

Scatterplot 1.

Figure 1:



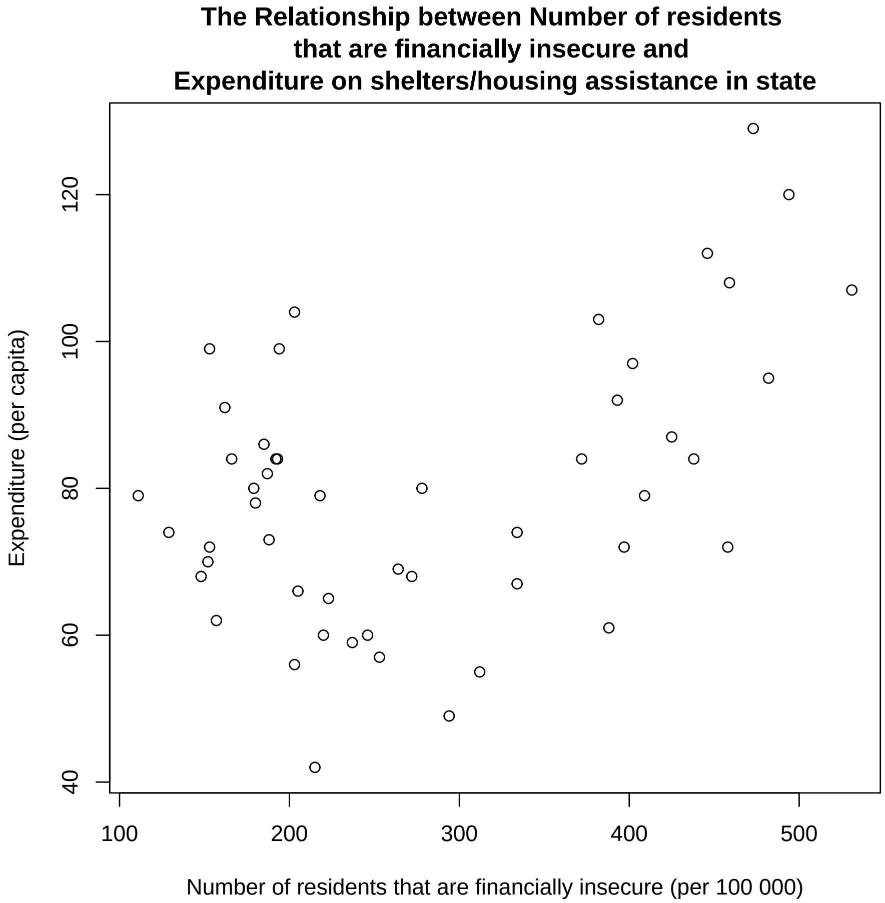
Conclusion: The scatterplot has a positive linear correlation. It shows a tendency for state with higher personal income to have higher levels of expenditure on shelters/housing assistance (per capita).

1.2. The relationships between Y and $X2$:

```
pdf(file=”C:/Users/HOME/Documents/GitHub/StatsI_Fall2023/problemSets/PS01/my_answers/plot1_2.pdf”)
plot(expenditure$X2,
     expenditure$Y,
     xlab=”Number of residents that are financially insecure (per 100 000)”,
     ylab=”Expenditure (per capita)”,
     main=”The Relationship between Number of residents \nthat are financially insecure and \nExpenditure on shelters/housing assistance in state”)
dev.off()
```

Scatterplot 2.

Figure 2:



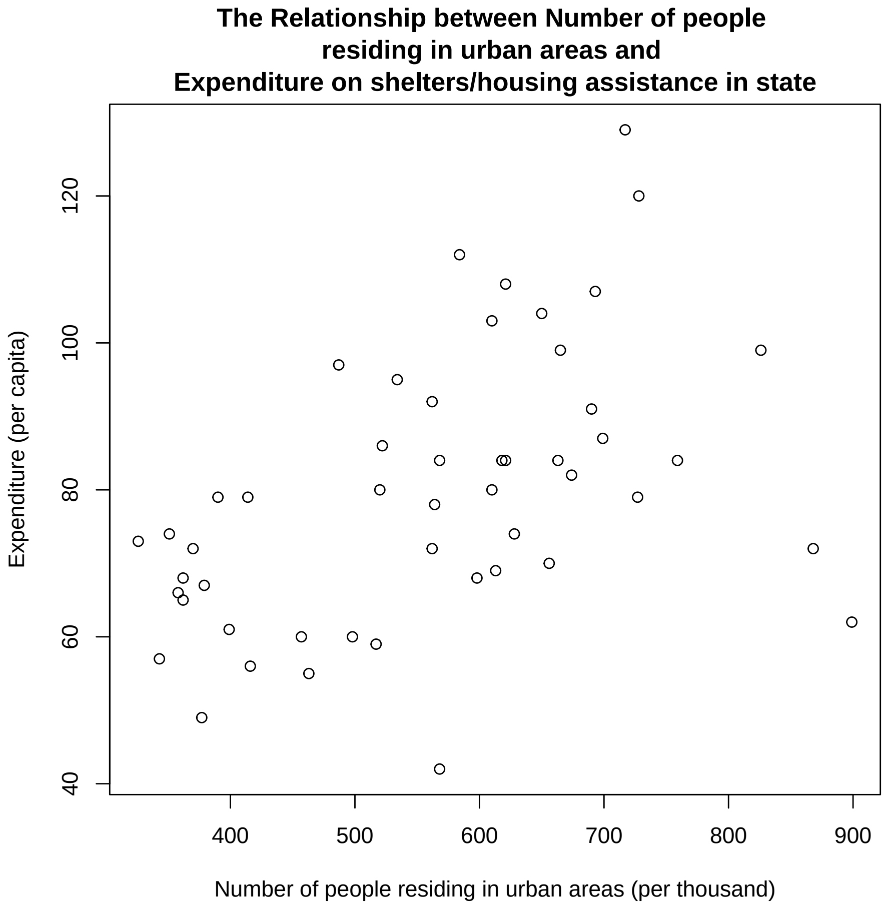
Conclusion: The scatterplot has a positive linear correlation. It shows a tendency for state with higher number of residents that are ”financially insecure” to have the higher levels of expenditure on shelters/housing assistance (per capita).

1.3. The relationships between Y and $X3$:

```
plot(expenditure$X3,
     expenditure$Y,
     xlab=”Number of people residing in urban areas (per thousand)”,
     ylab=”Expenditure (per capita)”,
     main=”The Relationship between Number of people \nresiding in urban areas
    and \nExpenditure on shelters/housing assistance in state”)
dev.off()
```

Scatterplot 3.

Figure 3:



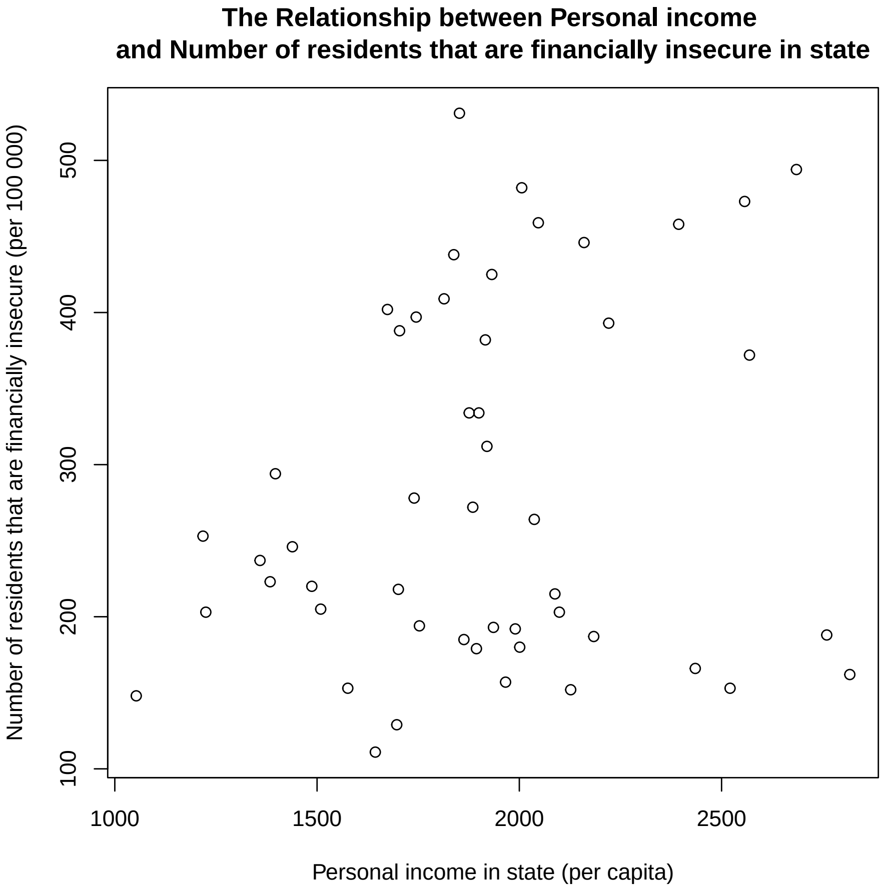
Conclusion: The scatterplot has a positive linear correlation. It shows a tendency for state with higher number of people residing in urban areas to have higher levels of expenditure on shelters/housing assistance (per capita).

1.4. The relationships between *X1* and *X2*:

```
plot(expenditure$X1,
     expenditure$X2,
     xlab=”Personal income in state (per capita)”,
     ylab=”Number of residents that are financially insecure (per 100 000)”,
     main=”The Relationship between Personal income \nand Number of residents that are financially insecure in state”)
dev.off()
```

Scatterplot 4.

Figure 4:



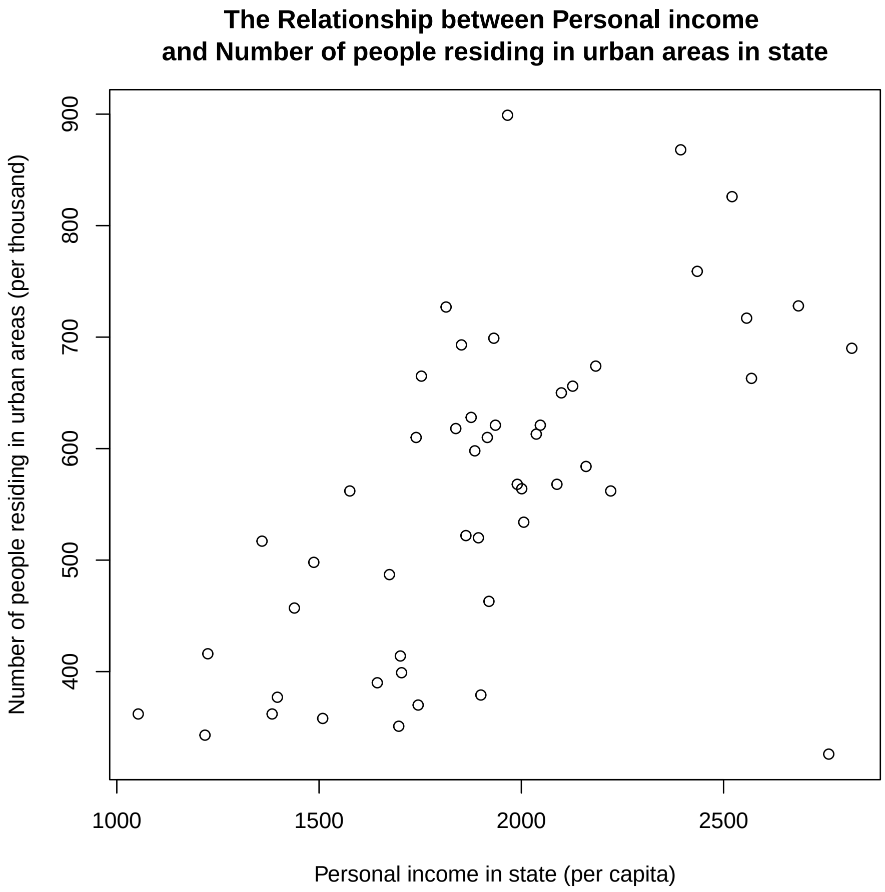
Conclusion: The scatterplot shows that a tendency for state with number of people who have middle and higher personal income to have higher levels of number of residents that are ”financially insecure”.

1.5. The relationships between *X1* and *X3*:

```
plot(expenditure$X1,
     expenditure$X3,
     xlab=”Personal income in state (per capita)”,
     ylab=”Number of people residing in urban areas (per thousand)”,
     main=”The Relationship between Personal income \nand Number of people
   residing in urban areas in state”)
dev.off()
```

Scatterplot 5.

Figure 5:



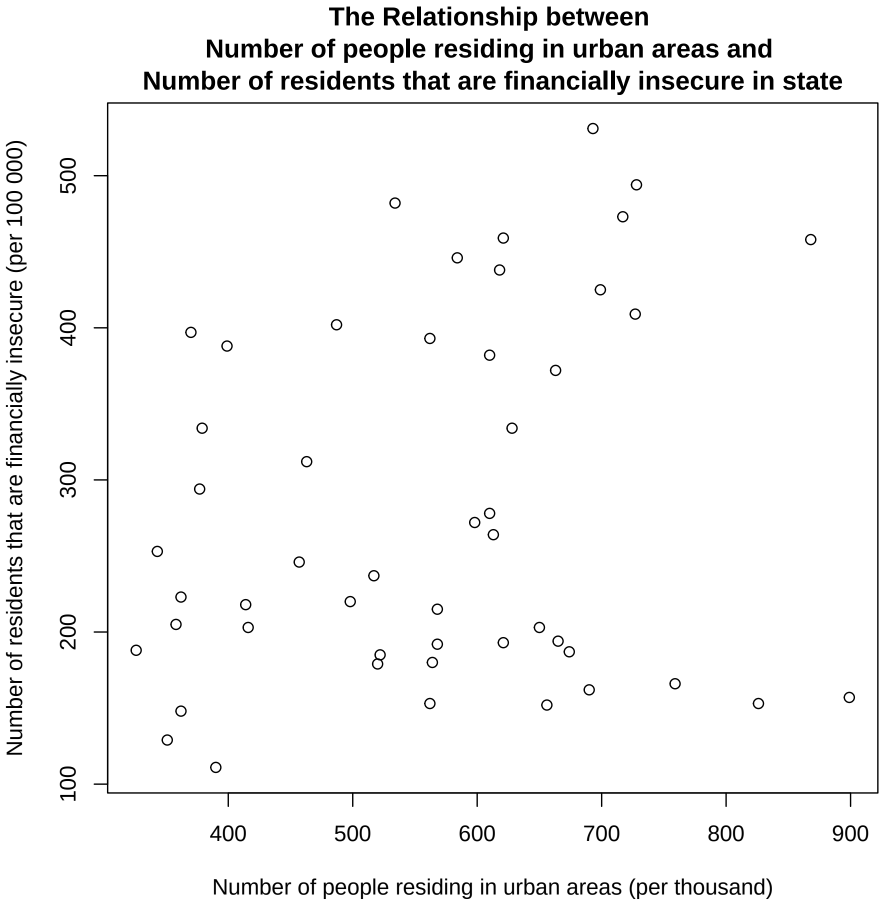
Conclusion: The scatterplot has a positive linear correlation. It shows a tendency for state with higher personal income to have higher number of people residing in urban areas.

1.6. The relationships between *X2* and *X3*:

```
plot(expenditure$X3,
     expenditure$X2,
     xlab="Number of people residing in urban areas (per thousand)",
     ylab="Number of residents that are financially insecure (per 100 000)",
     main="The Relationship between \nNumber of people residing in urban areas
    and \nNumber of residents that are financially insecure in state")
dev.off()
```

Scatterplot 6.

Figure 6:



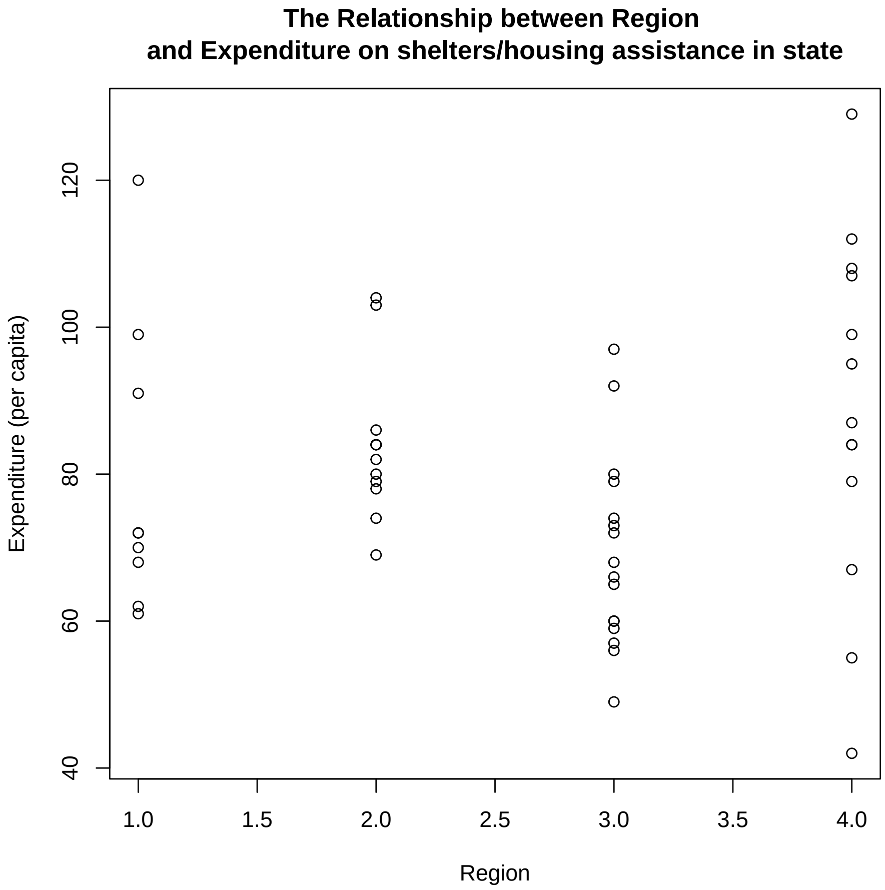
Conclusion: The scatterplot shows a tendency for state with low and middle number of people residing in urban area to have higher number of of residents that are ”financially insecure”.

2. The relationship between Y and *Region*:

```
plot(expenditure$Region,
     expenditure$Y,
     xlab="Region",
     ylab="Expenditure (per capita)",
     main="The Relationship between Region \nand Expenditure on shelters/
   housing assistance in state")
dev.off()
```

Scatterplot 7.

Figure 7:



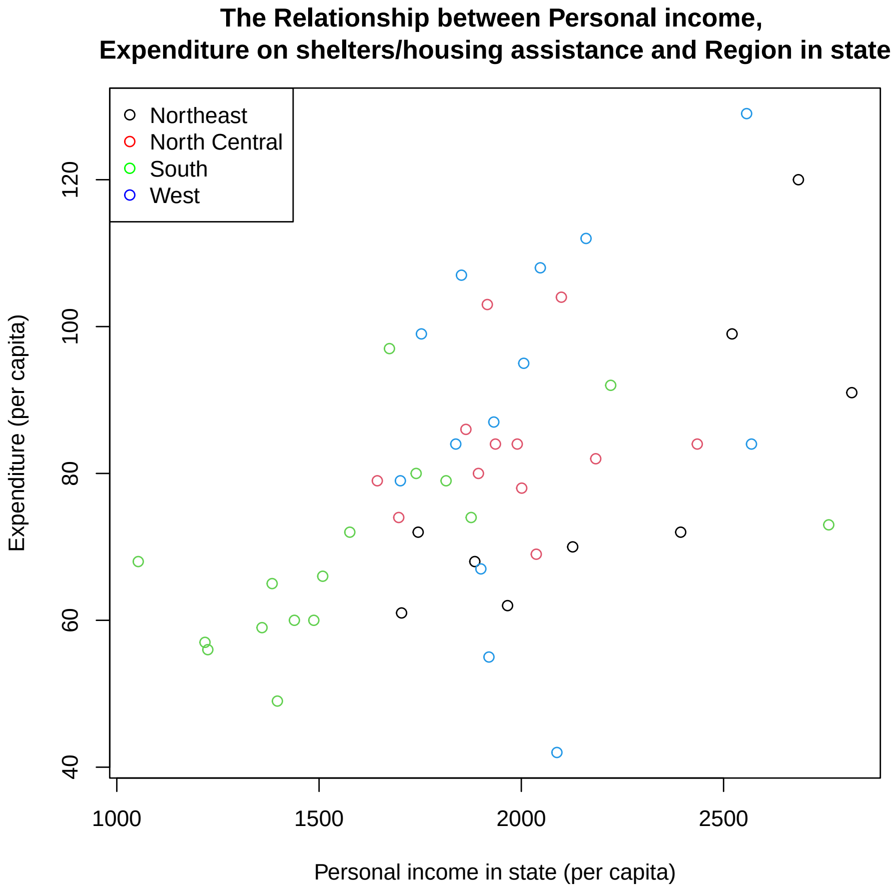
Conclusion: The scatterplot shows that on average regions 1 and 4 have the highest expenditure on shelters/housing assistance in state (per capita). Rigion 3 has the lowest expenditure on shelters/housing assistance in state (per capita)

3. The relationship between Y and $X1$ including one more variable *Region*:

```
pdf(file="C:/Users/HOME/Documents/GitHub/StatsI_Fall2023/problemSets/PS01/my_answers/plot3_1.pdf")
plot(expenditure$X1,
     expenditure$Y,
     col=expenditure$Region,
     xlab="Personal income in state (per capita)",
     ylab="Expenditure (per capita)",
     main="The Relationship between Personal income, \nExpenditure on shelters/housing assistance and Region in state")
legend("topleft",
        c("Northeast", "North Central", "South", "West"),
        col=c("black","red", "green", "blue"),
        pch=1) # Marker type (1 is default)
dev.off()
```

Scatterplot 8.

Figure 8:



Conclusion: The scatterplot has a positive linear correlation. It shows a tendency for state with higher personal income to have higher levels of expenditure on shelters/housing assistance (per capita). Especially in the two region: West and Northeast. South region has lowest level of expenditure on shelters/housing assistance (per capita).