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:

```
dataset_edu<-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 I calculated n using

   ```
   > length(dataset_edu)
   ```

Since n is smaller than 30 and I do not know the standard deviation of the whole sample, I decided to calculate the confidence interval using a t-distribution. To do so, I first calculated the t-score:

```
> t_score <- qt(0.95, df=25-1)
```

then I calculated the confidence interval

```
> conf_high <- mean(dataset_edu)+(t_score)*(sd(dataset_edu)/sqrt(25))
> conf_low <- mean(dataset_edu)-(t_score)*(sd(dataset_edu)/sqrt(25))
> conf_high
[1] 102.9201
> conf_low
[1] 93.95993
```

Thus, the 90% confidence interval is: [93.96; 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 $\alpha = 0.05$.

   H1: higher than average (meaning the H1 is one-sided)

   H0: lower than or equal to average

```
> t.test(dataset_edu, mu = 100, alternative = "greater")

One Sample t-test

data:  dataset_edu
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
```

The p-value = 0.7215

0.7215 is a higher value than 0.05, therefore we reject our alternative hypothesis (alternative hypothesis= the average student IQ in her school is higher than the average IQ score among all the schools in the country.)

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`.

```
plot(expenditure$X1, expenditure$X3, main = "Relationship between X1 and X3",
```

- 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)?

  First I created a 3x2 grid layout using

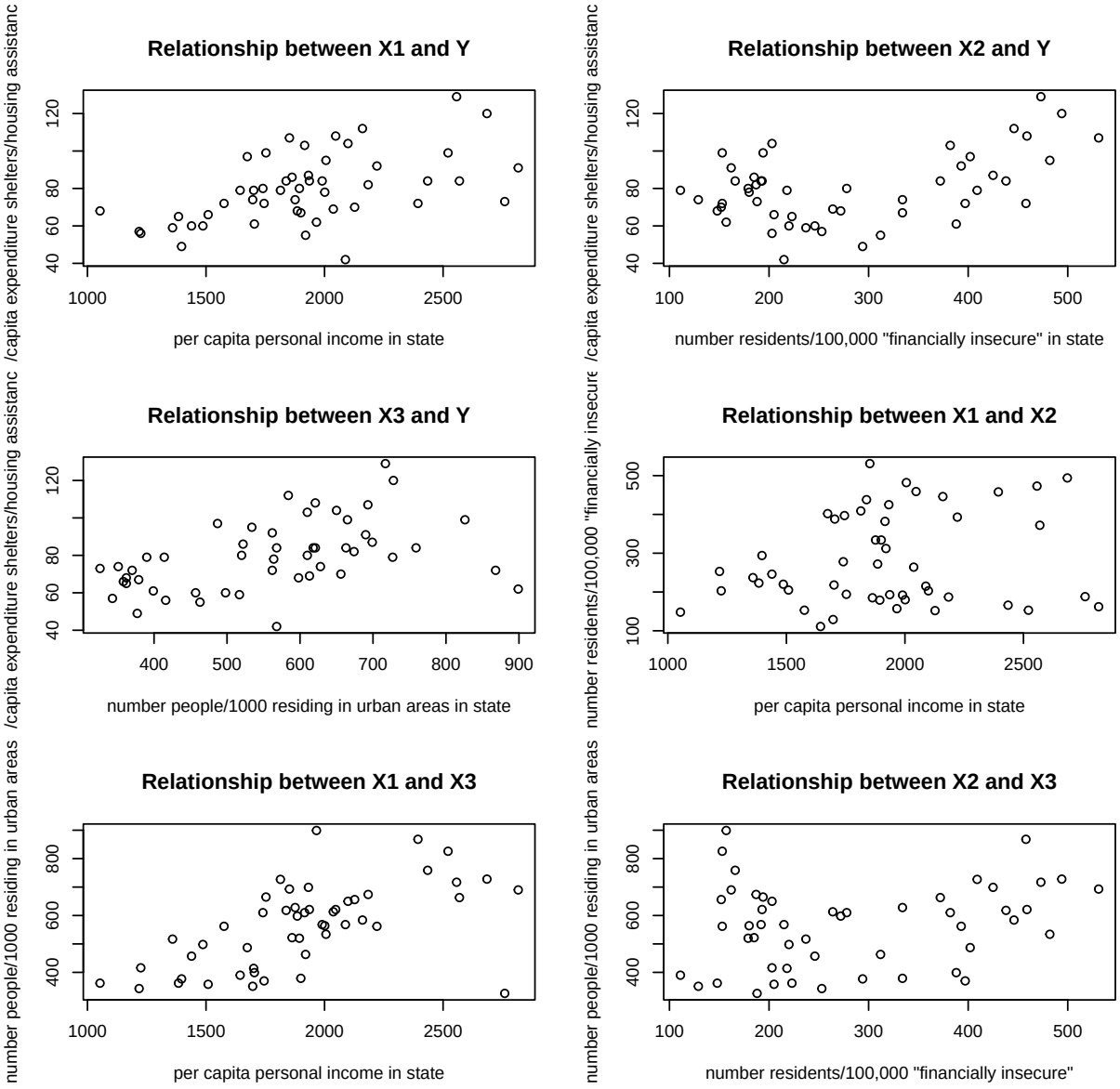
  ```
  > par(mfrow = c(3, 2))
  ```

  then I plotted all individual plots into that grid:

  ```
  > plot(expenditure$X1, expenditure$Y,  main = "Relationship between X1 and Y",
  +      xlab = "per capita personal income in state",
  +      ylab = "/capita expenditure shelters/housing assistance")
  > plot(expenditure$X2, expenditure$Y,  main = "Relationship between X2 and Y",
  +      xlab = "number residents/100,000 "financially insecure" in state",
  +      ylab = "/capita expenditure shelters/housing assistance")
  > plot(expenditure$X3, expenditure$Y,  main = "Relationship between X3 and Y",
  +      xlab = "number people/1000 residing in urban areas in state",
  +      ylab = "/capita expenditure shelters/housing assistance")
  > plot(expenditure$X1, expenditure$X2, main = "Relationship between X1 and X2",
  +      xlab = "per capita personal income in state",
  +      ylab = "number residents/100,000 "financially insecure"")
  > plot(expenditure$X1, expenditure$X3, main = "Relationship between X1 and X3",
  +      xlab = "per capita personal income in state",
  +      ylab = "number people/1000 residing in urban areas")
  ```

```
> plot(expenditure$X2, expenditure$X3, main = "Relationship between X2 and X3",
+      xlab = "number residents/100,000 "financially insecure"",
+      ylab = "number people/1000 residing in urban areas")
```

I had to shorten some of the names of the scales to fit in the graph. Please assume every scale to refer to the situation "in state".



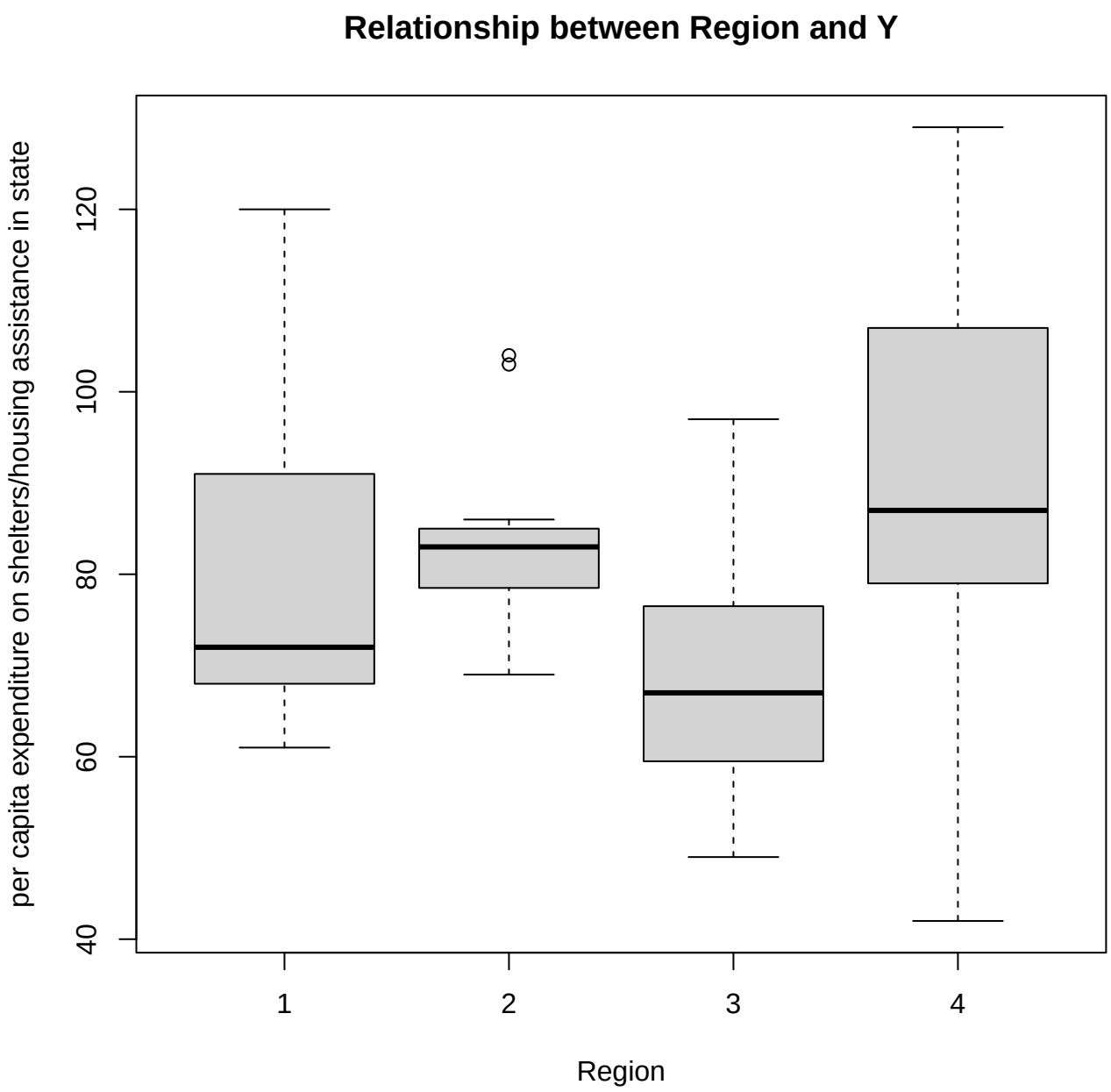
I don't find any of the correlations to be very clea-cut linear correlations. However, the first graph (Relationship between X1 and Y) shows what I would consider still a linear correlation throughout the whole graph, especially a the most frequent values for X1 (X1=1500-2000). The relationship between X2 and Y seems to be linear starting at

about X2=200. Before that the distribution seems almost U-shaped. The relatinship between X3 and Y appear linear to me as well, however the highest X3-values (X3=800-900) seem to have lower Y-values again. Since there are not many observation like this, I would still consider the overall distribution linear. The distribution between X1 and X3 also appears to be linear, while the relationship between X1 and X2 as well as between X2 and X3 seems to have, in my opinion, no discernible distribution, linear or otherwise. In general, I believe more observations would be helpful in determining linearity, especially in the case of X3 and Y.

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

  To do so I created a boxplot, since it is easiest to see relevant values like median or maximum value in that kind of graph.

```
boxplot(Y ~ Region, data = expenditure,
xlab="Region",
ylab="per capita expenditure on shelters/housing assistance in state",
main="Relationship between Region and Y")
```



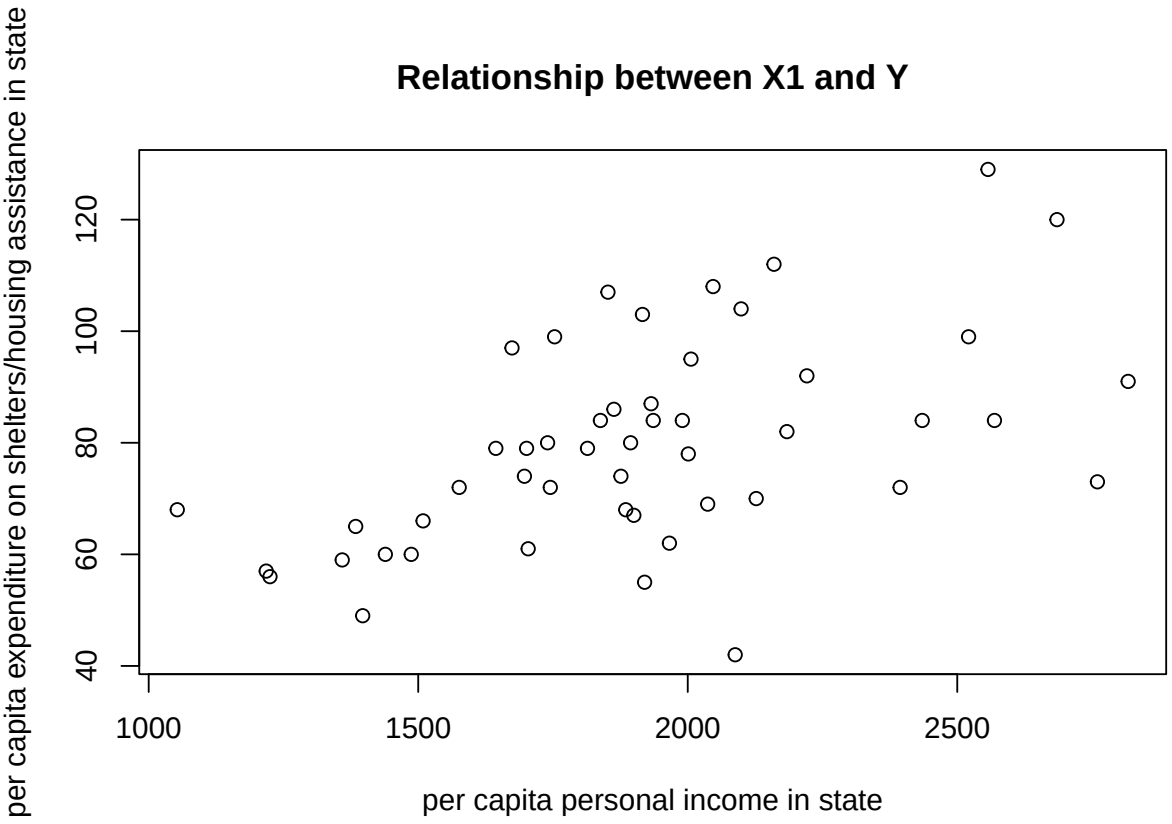
Region 4 (West) seems to have the highest median (shown by the black line), followed by Region 2 (North Central). However the maximum values of Y can be found in Region 4 and Region 1 (Northeast), not Region 2. So overall, I think the highest per capita expenditure on housing assistance can be found in the West.

- 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.

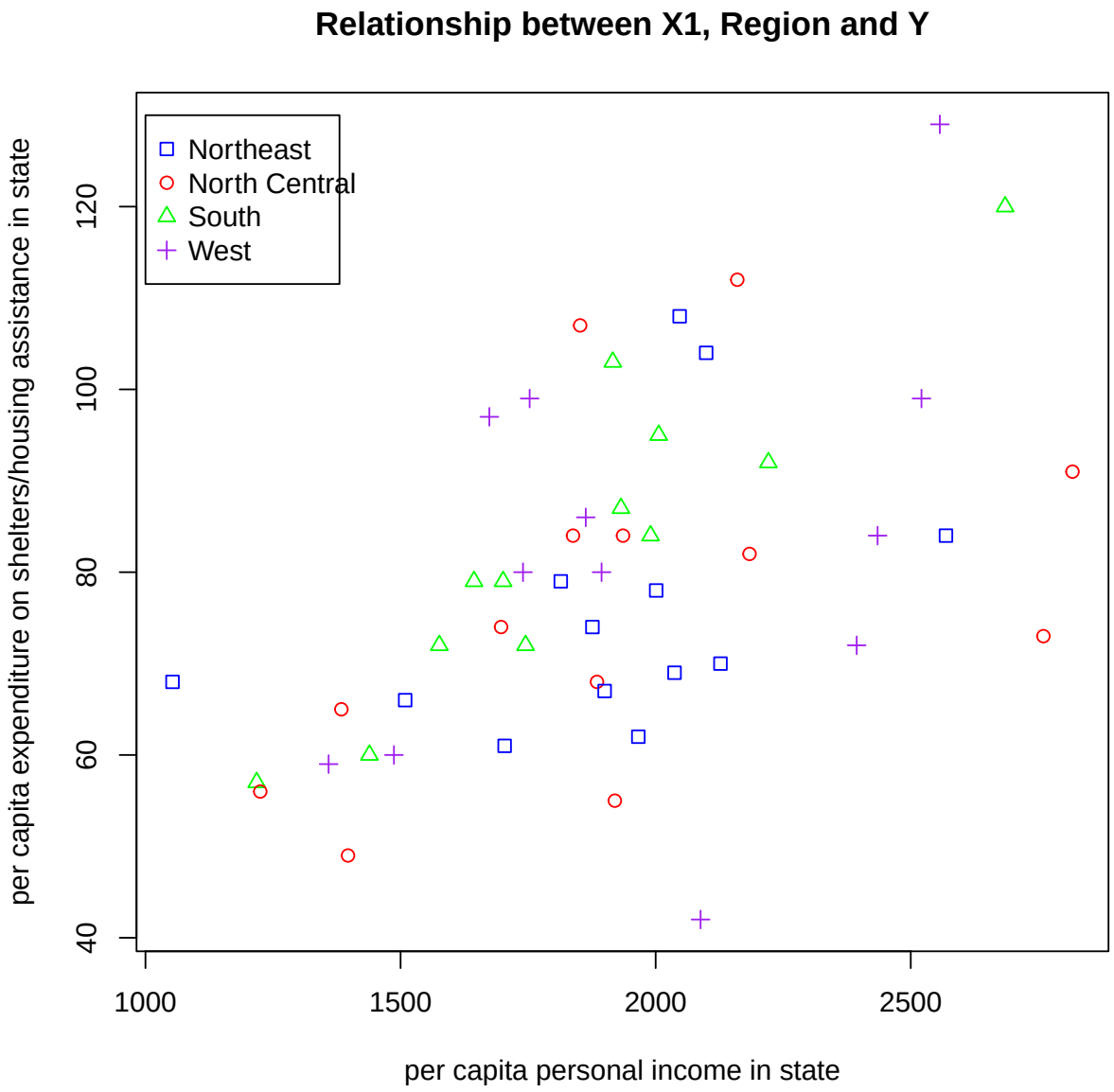
First, plotting the relationship between Y and X1:

```
> plot(expenditure$X1,expenditure$Y,
+      xlab = "per capita personal income in state",
+      ylab = "per capita expenditure on shelters/housing assistance in state",
+      main = "Relationship between X1 and Y")
```



As I answered before, the relationship between X1 and Y according to this scatterplot seems positively linear to me. The higher the per capita personal income in a state, the higher the per capita expenditure on shelters/housing assistance in that state. This can be seen especially where there are a lot of observations (around X1=1500-2000). However, there are not a lot of observations so it is difficult to tell, especially when considering the states with a higher per capita personal income (X1=2500 or more).

After that, I add the variable "Region"


Relationship between X1, Region and Y
Northeast
North Central
South
West
per capita expenditure on shelters/housing assistance in state
40
60
80
100
120
1000
1500
2000
2500
per capita personal income in state