My Answers Problem Set One Clémence Pellissier

2024-09-25

```
``` r
# I remove objects
rm(list=ls())

# I detach all libraries
detachAllPackages <- function() {
  basic.packages <- c("package:stats", "package:graphics",
                      "package:grDevices", "package:utils",
                      "package:datasets", "package:methods",
                      "package:base")
  package.list <- search()[ifelse(unlist(gregexpr("package:",
                                                  search()))==1, TRUE, FALSE)]
  package.list <- setdiff(package.list, basic.packages)
  if (length(package.list)>0)  for (package in package.list)
    detach(package,  character.only=TRUE)
}
detachAllPackages()

# I load libraries
pkgTest <- function(pkg){
  new.pkg <- pkg[!(pkg %in% installed.packages()[,  "Package"])]
  if (length(new.pkg))
    install.packages(new.pkg,  dependencies = TRUE)
  sapply(pkg,  require,  character.only = TRUE)
}
```

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 assign- ments 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:

```
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.

I Load my data as a vector :

```
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)
```

I calculate the mean of the sample :

```
mean_y <- mean(y)
mean(y)
```

```
## [1] 98.44
```

I calculate the standard deviation of the sample :

```
sd_y <- sd(y)
sd(y)
```

```
## [1] 13.09287
```

I determine the size of the sample :

```
n <- length(y)
n
```

```
## [1] 25
```

I find the t-value : Since our sample is smaller than 30, I use a t-distribution The confidence interval is 0.90

```
alpha <- 0.10
t_value <- qt(1 - alpha/2, df = n - 1)
qt(1 - alpha/2, df = n - 1)
```

```
## [1] 1.710882
```

I calculate the appropriate confidence interval for the mean level of students' IQ in the school. Since our sample is samller than 30, I conduct a t-test. To do so, I calculate the standard error of the mean

```
SE <- sd(y) / sqrt(n)
sd(y) / sqrt(n)
```

```
## [1] 2.618575
```

I calculate the confidence interval :

```
lower_bound90 <- mean(y) - t_value * SE
mean(y) - t_value * SE
```

```
## [1] 93.95993
```

```
upper_bound90 <- mean(y) + t_value * SE
mean(y) + t_value * SE
```

```
## [1] 102.9201
```

```
lower_bound90
```

```
## [1] 93.95993
```

```
upper_bound90
```

```
## [1] 102.9201
```

The results are : 93.95993 and 102.9201

Interpretation : At least 90/100 of the time, the IQ of the population would be between 93.95993 and 102.9201

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.

To do so, I run a 5 steps hypothesis testing test

Step One: Assomptions about my data : 1) The data is continuous 2) The sample is small, 25 3) The sampling method is randomnization

Step Two: Formulating the hypotheses H0 : The average IQ in her school is equal to the average 100 IQ score H1: The average IQ in her school is greater than the average 100 IQ score The set significant level is alpha = 0.05

Step Three : Choosing the testing methodology Since the sample is small (25), I will use t-statistic

Step Four : Calculating t-value and performing one-sample t-test :

I start by loading my data as vector :

```
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)
```

I calculate the t value, to do so I use 90 as my confidence interval :

```
t_value <- qt(1 - alpha/2, df = n - 1)
qt(1 - alpha/2, df = n - 1)
```

```
## [1] 1.710882
```

I calculate the degree of freedom :

```
df <- 25-1
25-1
```

```
## [1] 24
```

I start the hypothesis test by performing one-sample t-test :

```
t_test_result <- t.test(y,mu=100, alternative = "greater")
t.test(y, mu = 100, alternative = "greater")
```

```
##
##  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
```

The results are : t-value = -0.59574, p-value = 0.7215

Step Five : Drawing conclusion

Interpretation of the results : A negative value means that the mean of the sample, i.e. the IQ of the counselor's students, is less than 100. On top of this, the p-value is greater than the significant level.This means that the null hypotheses cannot be rejected. This means that there is not enough evidence to conclude that the average IQ in the school is greater than the national one, i.e.100.

Question 2 - 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.

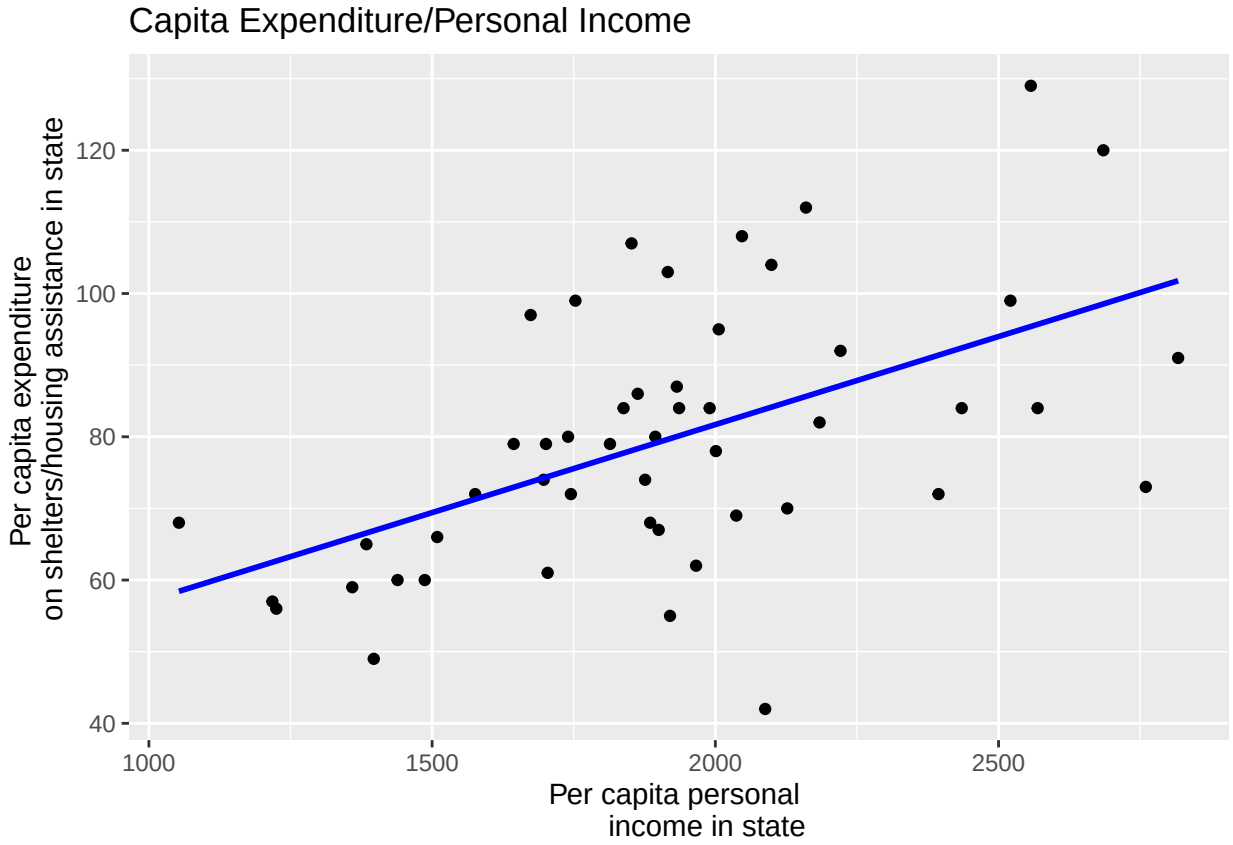
• Please plot the relationships among Y, X1, X2, and X3. What are the correlations among them (you just need to describe the graph and the relationships among them)?

I load my data :

```
expenditure <- read.table("https://raw.githubusercontent.com/ASDS-TCD/StatsI_Fall2024/main/datasets/exp
                          , header=T)
data <- expenditure[, c("Y", "X1", "X2", "X3")]
```

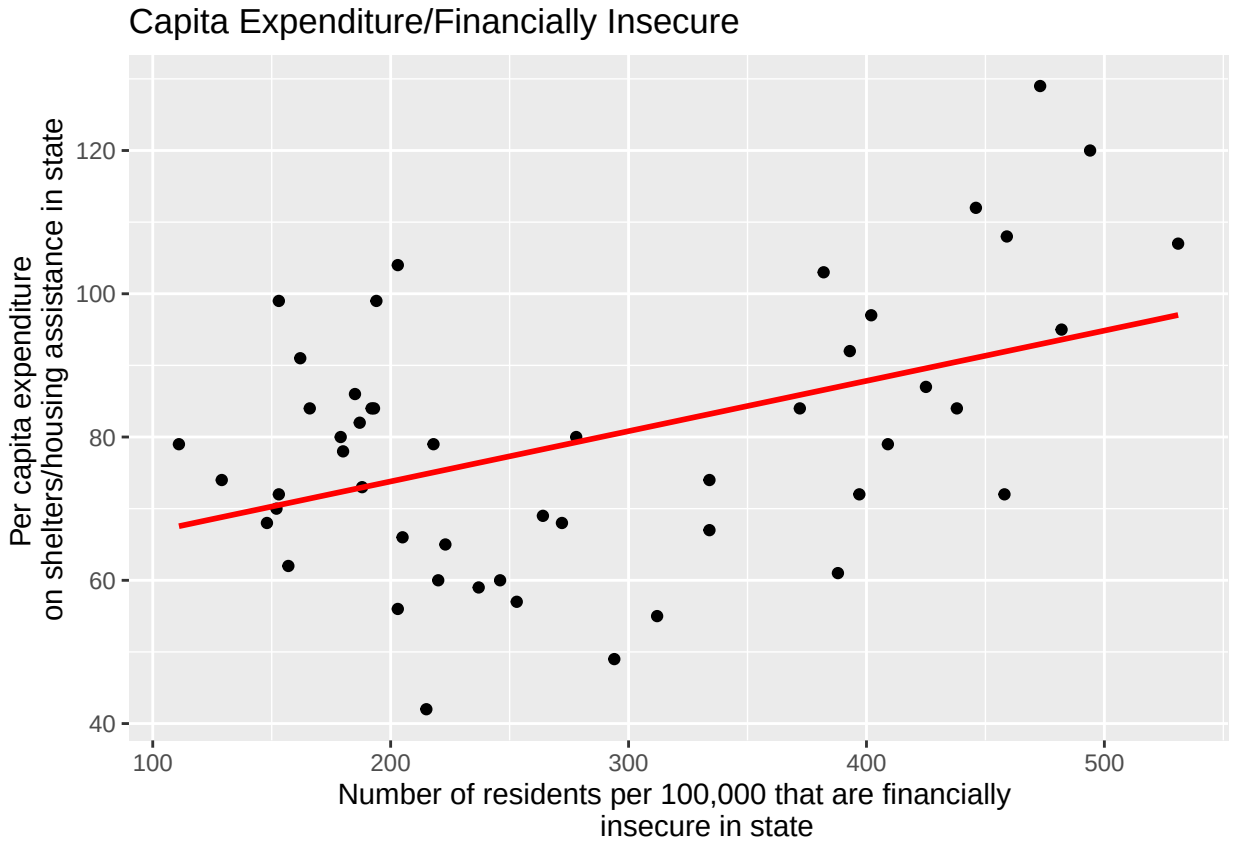
I start analyzing the six different relationships :

First (outcome/predictor one): plot Y/X1 (with a regression line):



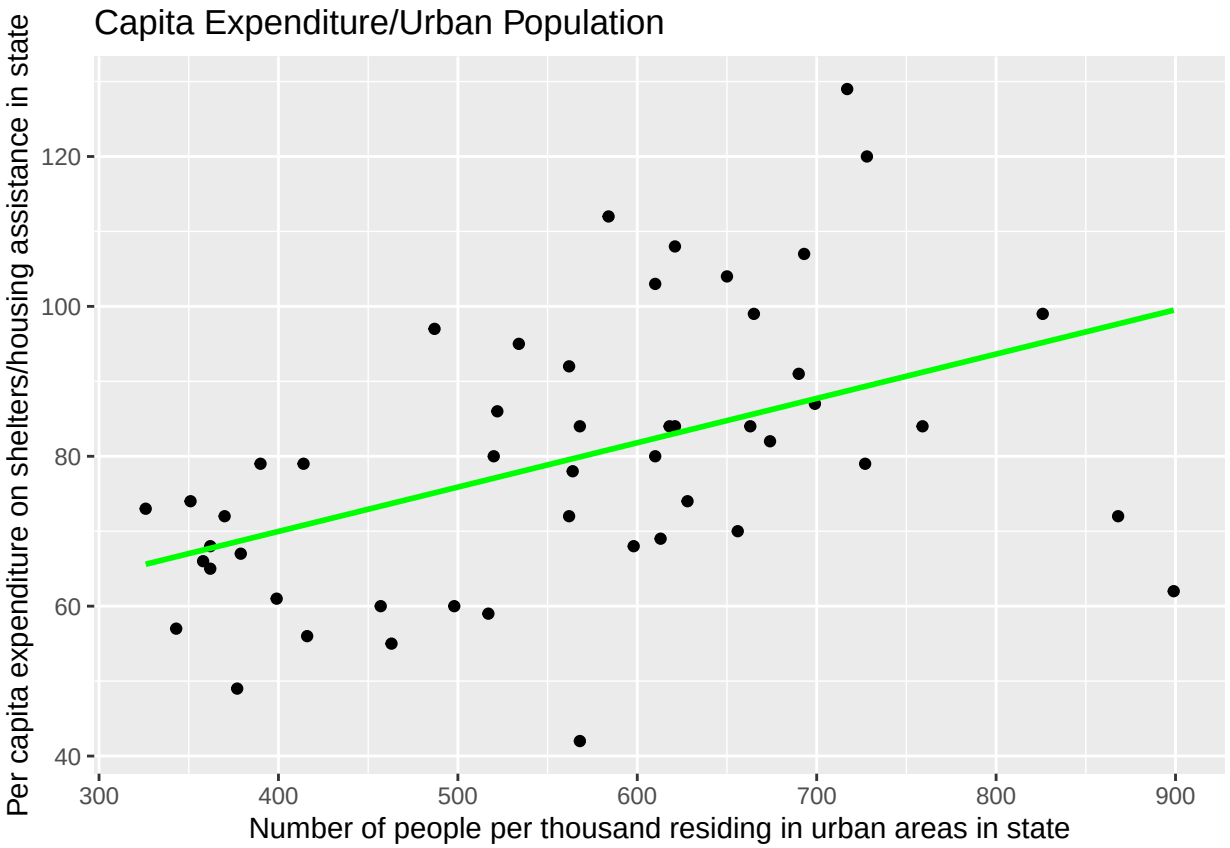
Interpretation : I can see a strong positive linear correlation between capital personal income (X) and capital expenditure (Y). In other words, when there is a increase of personal income in the sate, there is also an increase of capital expenditure on shelters/housing assistance.

Second (outcome/predictor two): plot Y/X2 (with a regression line) :



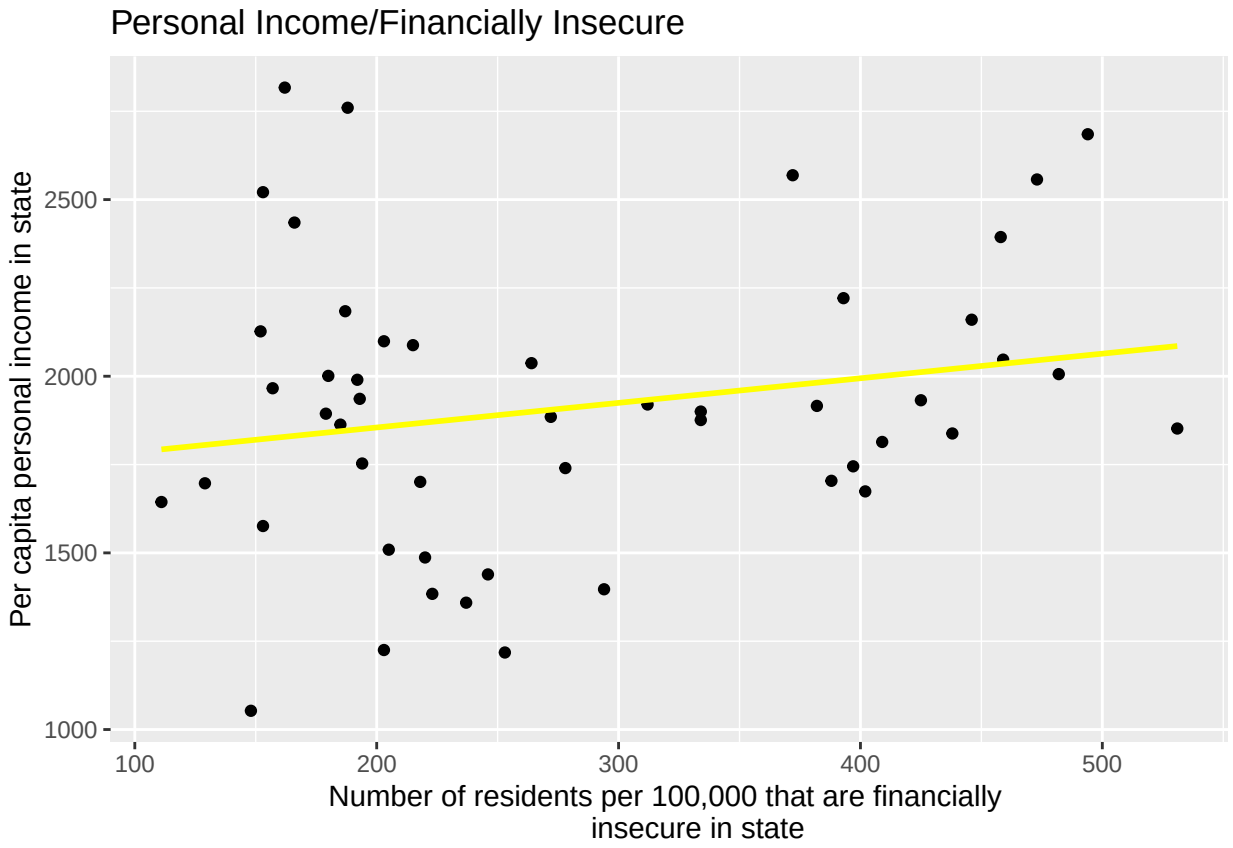
Interpretation : I can see no clear linear positive or negative correlation between the number of residents financially unsecured (X) and capital expenditure (Y), the dots forming a parabola. In other words,I can see that if there is an increase from 100 to 300 of the number of residents financially unsecured, the capital expenditure on shelters/housing assistance decreases. However, above 300 of residents financially unsecured, there is a linear positive correlation between this number of residents financially unsecured and the capital expenditure.

Third (outcome/predictor three): plot Y/X3 (with a regression line) :



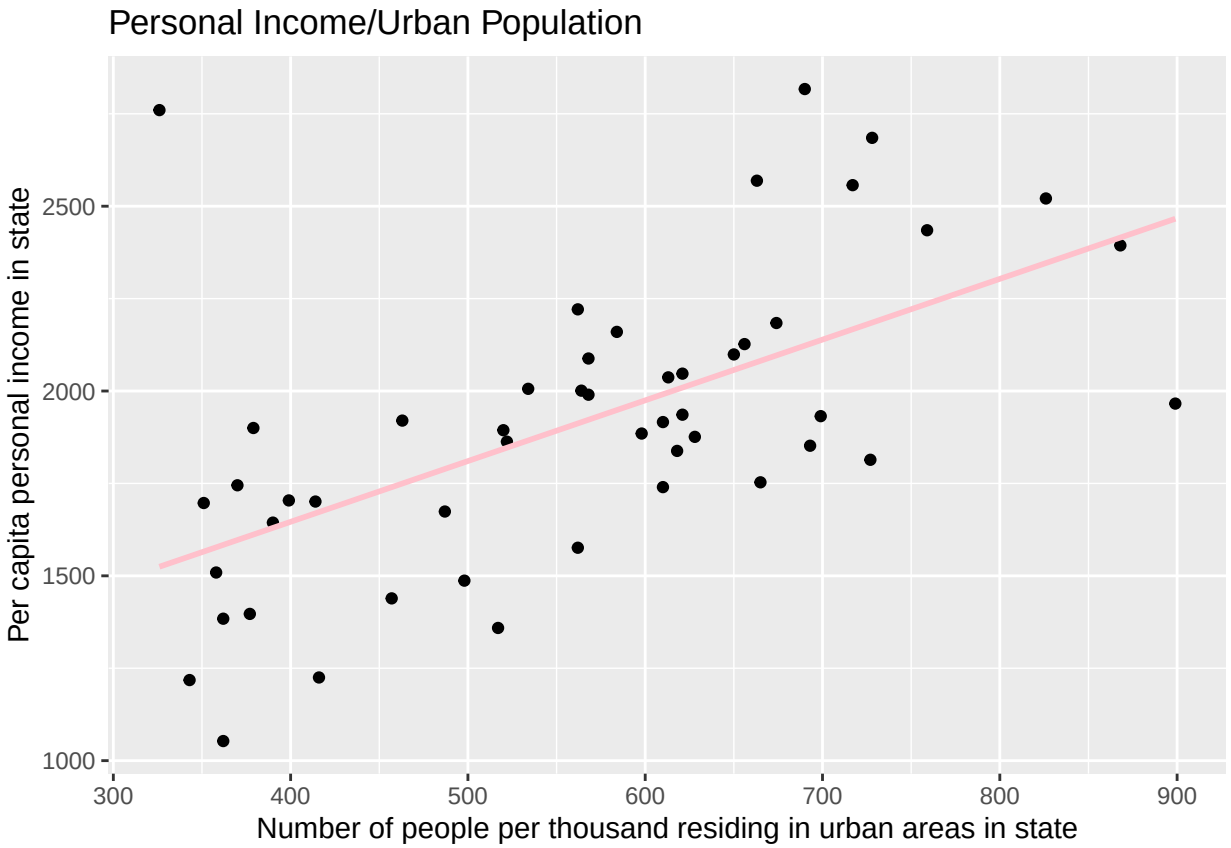
Interpretation : I can see a weak linear correlation between the number of people in urban areas (X) and capital expenditure (Y). In other words, when there is a increase of the number of people in urban areas in the sate, there is also a small tendancy of an increase of capital expenditure on shelters/housing assistance.

Fourth (predictor one/predictor two): plot X1 vs X2(with a regression line):



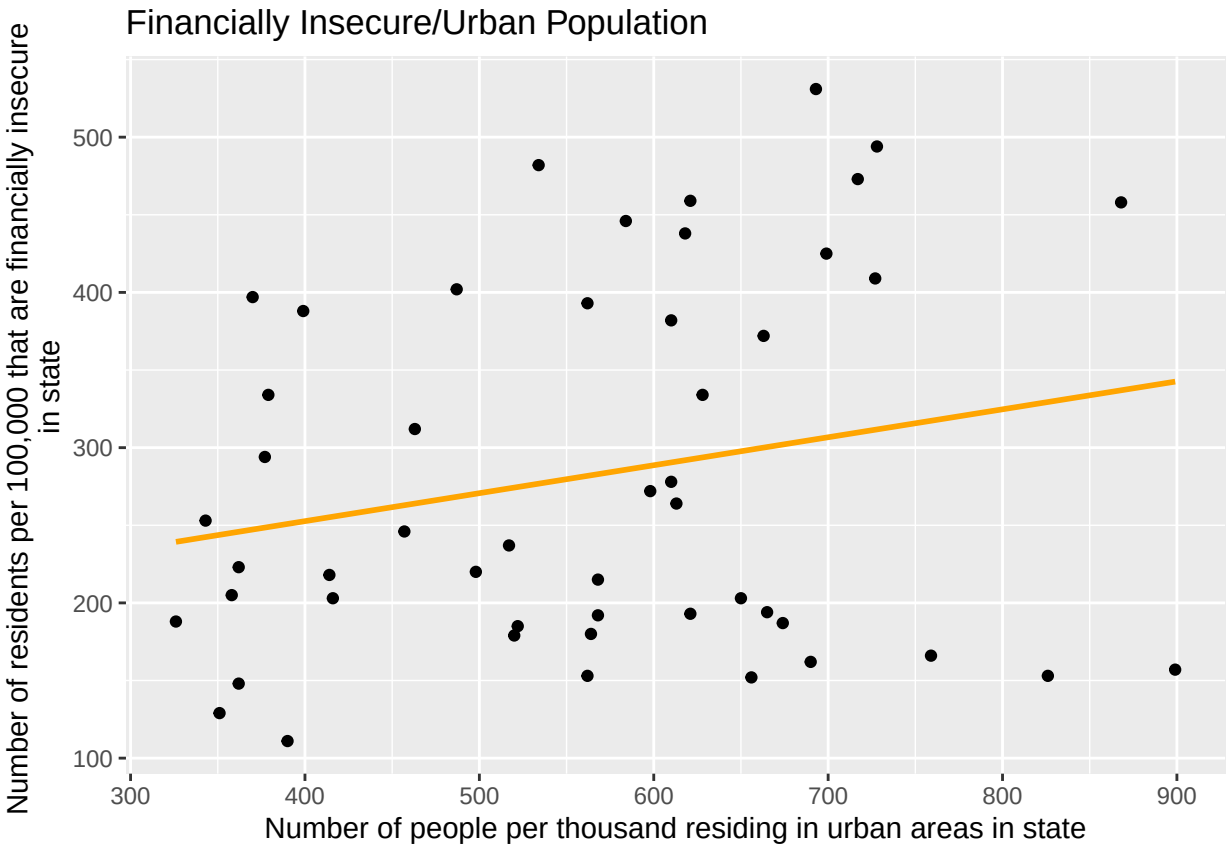
Interpretation : Here there is no clear linear correlation between the number of residents financially unsecured (X) and capital personal income (Y), the dots forming a parabola. In other words, if there is an increase from 100 to 250 of the number of residents financially unsecured, the capital personal income decreases. However, above 300 of residents financially unsecured, there is an increasing correlation between the number of residents financially unsecured and the capital personal income.

Fifth (predictor one/predictor three): plot X1 vs X3(with a regression line) :



Interpretation : I can see a strong linear correlation between the number of people in urban areas (X) and capital personal income (Y). In other words, when there is an increase of the number of people in urban areas in the sate, there is also an increase of capital personal income.

Sixth (predictor two/predictor three): plot X2 vs X3 (with a regression line):



Interpretation : I can see a weak linear correlation between the number of people in urban areas (X) and the number of residents financially unsecured (Y). In other words, when there is an increase of the number of people in urban areas in the sate, there is also a small tendancy of an increase of the number of residents financially unsecured.

I complete this analyse by calculating correlations :

```
correlation_matrix <- cor(data[, c("Y", "X1", "X2", "X3")])
```

Interpretation : I know that : => A value close to 1 indicates a strong positive correlation. => A value close to -1 indicates a strong negative correlation. => A value close to 0 means no correlation.

Here we have : 1) Y/X1 = 0.5317212, it is a moderate positive correlation. This means that there is a not particular strong correlation between the increase of personal income in the sate and the increase of capital expenditure on shelters/housing assistance.

2) Y/X2 = 0.4482876, it is a moderate positive correlation.This means that there is a not particular strong correlation between the increase of the number of residents financially unsecured and the increase of capital expenditure on shelters/housing assistance decreases.

3) Y/X3 = 0.4636787, it is a moderate positive correlation. This means tha there is a not particular strong correlation between the increase of the number of people in urban areas and the increase of capital expenditure on shelters/housing assistance decreases.

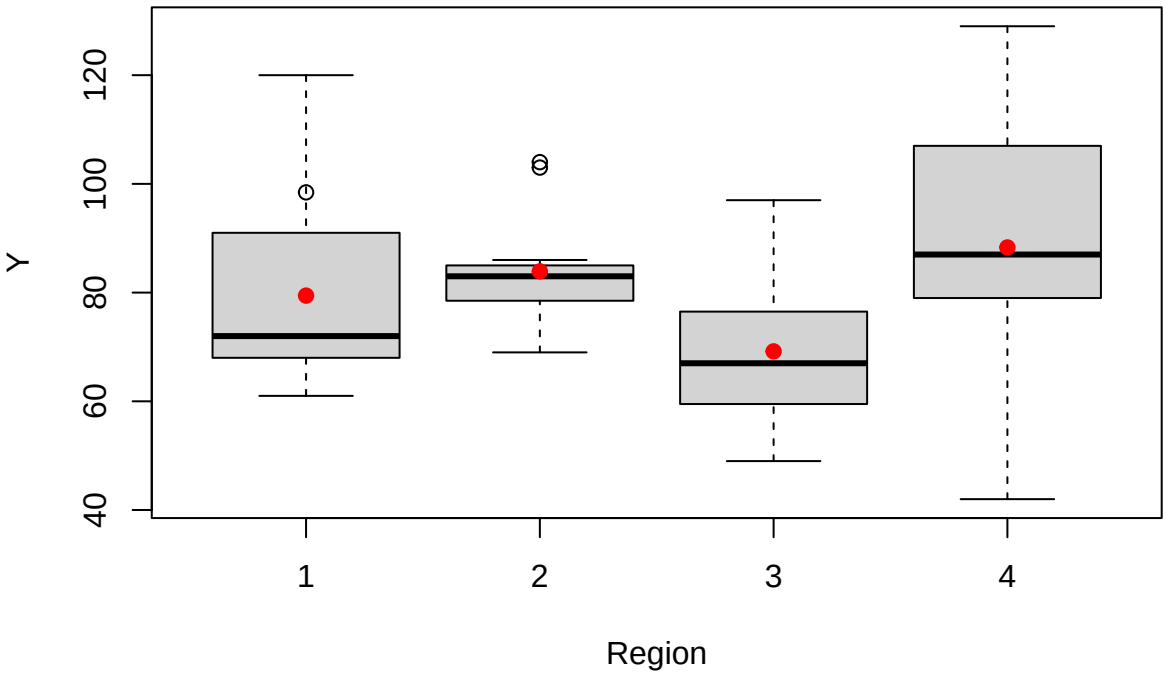
4) X1/X2 = 0.2056101, it is a weak positive correlation. This means there is a low correlation between the increase of personal income in the sate and the increase of the number of residents financially unsecured.

5) X1/X3 = 0.5952504, it is a moderate positive correlation.This means that there is a not particular strong correlation between the increase of personal income in the sate and the increase of the number of people in urban areas.

6) X2/X3 = 0.2210149, t is a weak positive correlation. This means there is a low correlation between the increase of the number of residents financially unsecured and the increase of the number of people in urban areas.

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

I plot the relationship between Y and Region, using a boxplot :



Interpretation : I can see which region (X) has the highest capita expenditure on housing assistance (Y). To do so, I inspect the average red dotes in the boxplot : generally I can say that the highest one represents the region with the highest per capita expenditure. Here are (in descending order): Region 4 (West) ; Region 2 (Centre-Nord); Region 1 (Nord-Est); and Region 3 (Sud). In this case, the same applies to the median line.

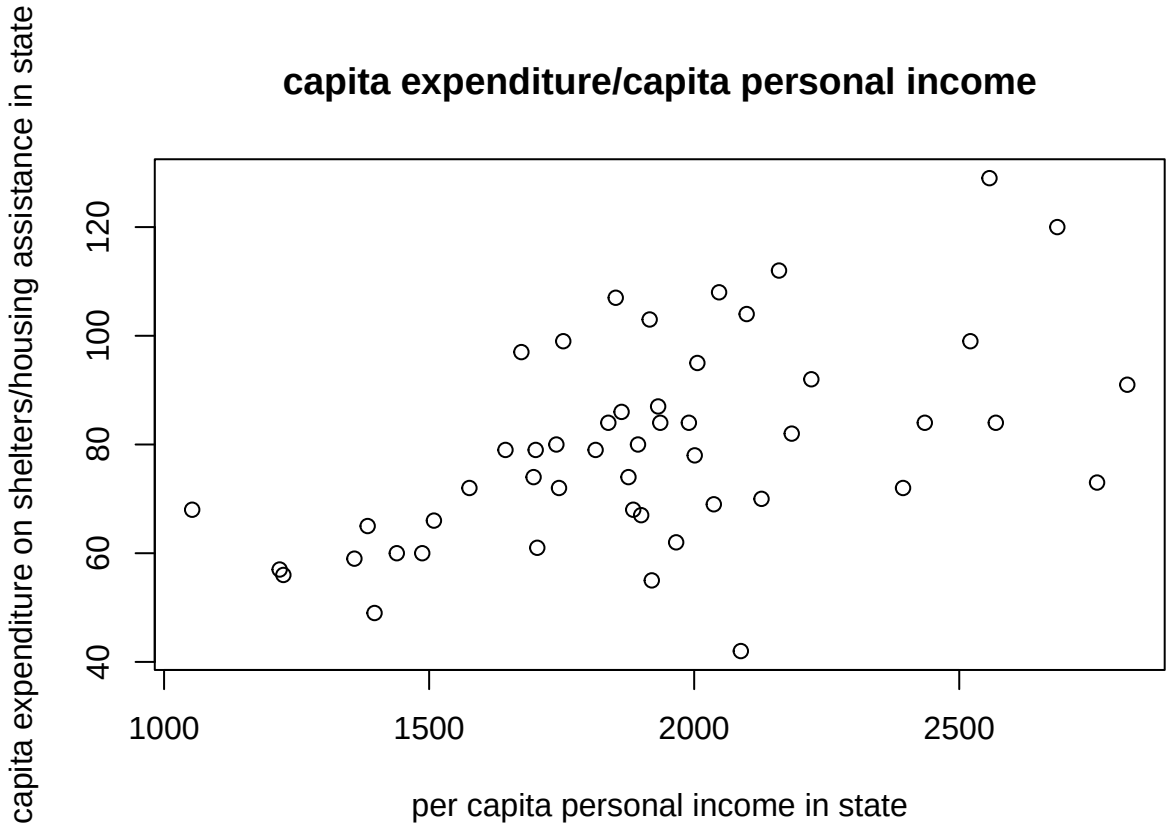
• Please plot the relationship between Y and X1? Describe this graph and the relationship. Reproduce the above graph including one more variable Region and display different regions with different types of symbols and colors.

I load my data:

```
expenditure <- read.table("https://raw.githubusercontent.com/ASDS-TCD/StatsI_Fall2024/main/datasets/expe
                          , header=T)
data <- expenditure[, c("Y", "X1", "X2", "X3", "Region")]

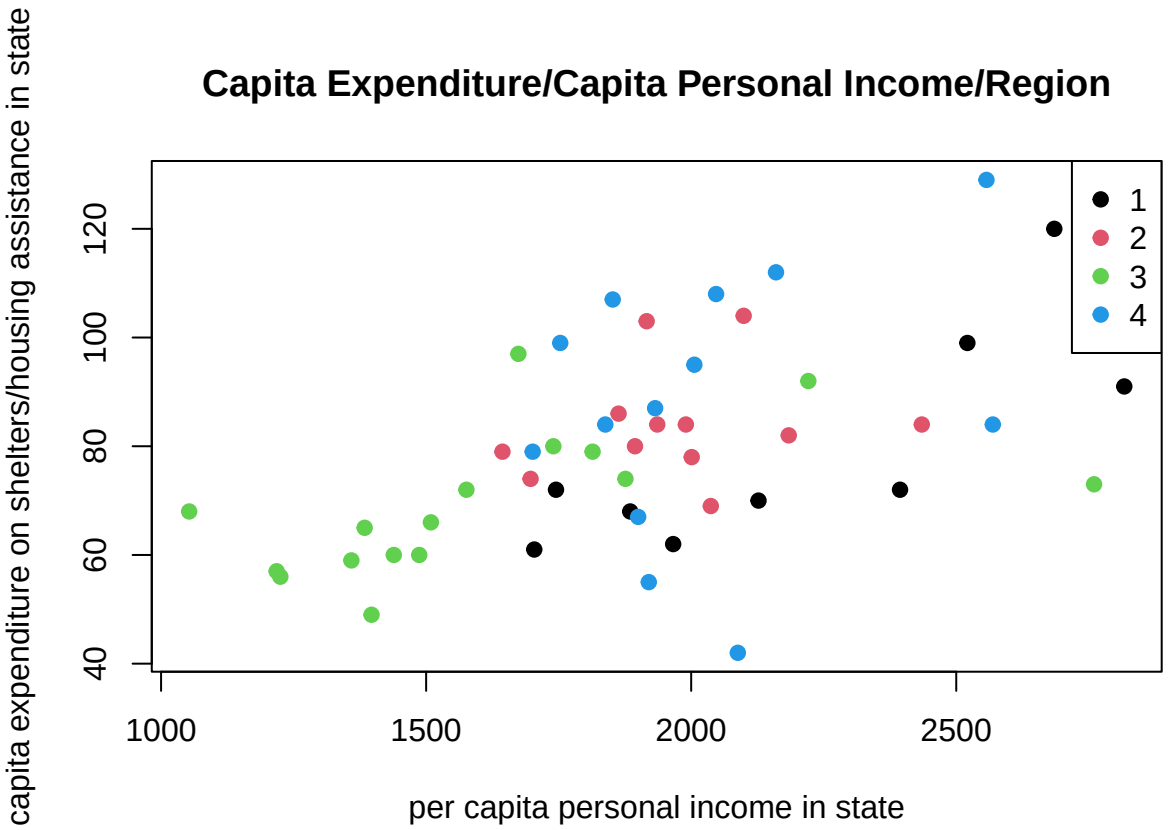
colors <- as.factor(data$Region)
```

I analyze the relationships (outcome/predictor one), plotting Y vs X1 :



Interpretation : I can see a strong positive linear correlation between capital personal income (X) and capital expenditure (Y). In other words, when there is a increase of personal income in the sate, there is also an increase of capital expenditure on shelters/housing assistance.

I add the extra variable (Region) :



Interpretation :I can see which region (Z) has the highest capita expenditure on housing assistance (Y), regarding its capita personal income (X). Inspecting the graph, I can say that region 3 (South) is far below, region 2 (North Central) is in the middle, and regions 4 (West) and 1 (Northeast) are above.