Problem Set 1

Applied Stats/Quant Methods 1

Due: October 1, 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:

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

```
# Calculation using standard normal distribution, but the sample size was small
# confidence_level <- 0.90
# z90 <- qnorm((1 - confidence_level) / 2, lower.tail = FALSE)
# sample_data <- y
# n <- length(sample_data)
# sample_mean <- mean(sample_data)
# sample_sd <- sd(sample_data)
# lower_90 <- sample_mean - (z90 * (sample_sd / sqrt(n)))
# upper_90 <- sample_mean + (z90 * (sample_sd / sqrt(n)))
# confint90 <- c(lower_90, upper_90)
# cat("90% Confidence Interval:", lower_90, "-", upper_90, "\n")
# Calculate it with the t distribution
sample_mean <- mean(y)
sample_sd <- sd(y)
sample_size <- length(y)
confidence_level <- 0.90
degrees_of_freedom <- sample_size - 1
t_critical <- qt(1 - (1 - confidence_level) / 2, df = degrees_of_freedom)
standard_error <- sample_sd / sqrt(sample_size)
confidence_interval <- c(sample_mean - t_critical * standard_error, sample_mean + t_critical * standard_error)
cat("90% Confidence Interval for Average IQ:", confidence_interval, "\n")
```

0.90 Confidence Interval for Average IQ: 93.95993 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.

```r
# Hypothesis testing:
# Null hypothesis (H0) : The average student IQ of the school is less than or equal to
    the national average IQ score.
# Alternative hypothesis (H1) : The average student IQ in the school is greater than
    the national average IQ score.
# The alternative hypothesis (H1) describes my research hypothesis,
# so it focuses on a 'greater than' scenario.
# Therefore, a right/upper-tailed hypothesis test should be used,
# with p-value calculated as p_value = 1 - pt(t_statistic, df = degrees_of_freedom)."
# Make a decision
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)
sample_mean <- mean(y)
sample_sd <- sd(y)
sample_size <- length(y)
confidence_level <- 0.05
degrees_of_freedom <- sample_size - 1
t_statistic <- (sample_mean - 100) / (sample_sd / sqrt(sample_size))
p_value <- 1 - pt(t_statistic, df = degrees_of_freedom)
if (p_value <= confidence_level) {
  cat("Reject the null hypothesis. The average student IQ is higher than the national
    average IQ score")
} else {
  cat("Fail to reject the null hypothesis. The average student IQ is less than or equal
     to the national average IQ score")
}
```

Fail to reject the null hypothesis. The average student IQ is less than or equal to the national average IQ score

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.

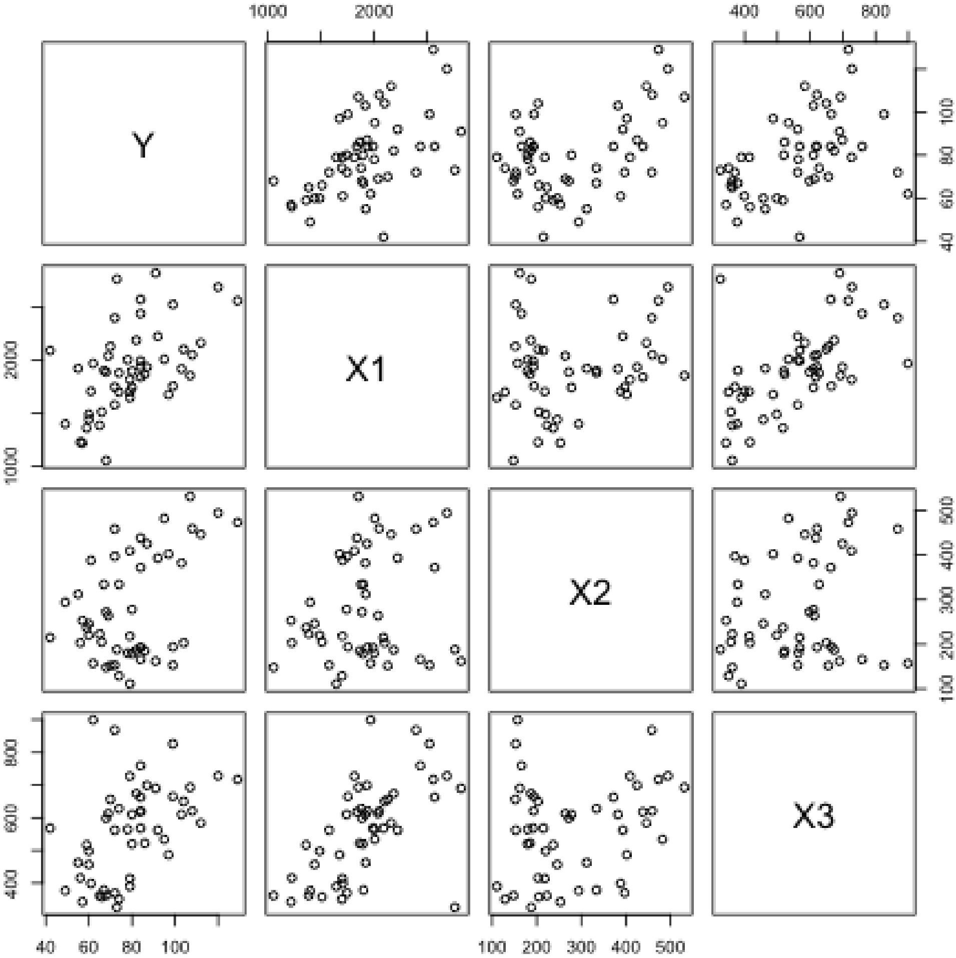
`State`	*50 states in US*
`Y`	*per capita expenditure on shelters/housing assistance in state*
`X1`	*per capita personal income in state*
`X2`	*Number of residents per 100,000 that are "financially insecure" in state*
`X3`	*Number of people per thousand residing in urban areas in state*
`Region`	*1=Northeast, 2= North Central, 3= South, 4=West*

Explore the `expenditure` data set and import data into `R`.

1. 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
# read in expenditure data
expenditure <- read.table("https://raw.githubusercontent.com/ASDS-TCD/StatsI_Fall2023/main/datasets/expenditure.txt", header=T)

# 1
data <- read.table(text = "
STATE  Y   X1    X2   X3   Region
ME    61   1704      388  399  1
NH    68   1885      272  598  1
VT    72   1745      397  370  1
MA    72   2394      458  868  1
RI    62   1966      157  899  1
CT    91   2817      162  690  1
NY    120  2685      494  728  1
NJ    99   2521      153  826  1
PA    70   2127      152  656  1
OH    82   2184      187  674  2
IN    84   1990      192  568  2
IL    84   2435      166  759  2
MI    104  2099      203  650  2
WI    84   1936      193  621  2
MN    103  1916      382  610  2
IA    86   1863      185  522  2
MO    69   2037      264  613  2
ND    74   1697      129  351  2
SD    79   1644      111  390  2
NB    80   1894      179  520  2
KS    78   2001      180  564  2
DE    73   2760      188  326  3
MD    92   2221      393  562  3
VA    97   1674      402  487  3
WV    66   1509      205  358  3
NC    65   1384      223  362  3
SC    57   1218      253  343  3
GA    60   1487      220  498  3
FL    74   1876      334  628  3
KY    49   1397      294  377  3
TN    60   1439      246  457  3
AL    59   1359      237  517  3
MS    68   1053      148  362  3
AR    56   1225      203  416  3
```

```
LA   72   1576     153  562  3
OK   80   1740     278  610  3
TX   79   1814     409  727  3
MT   55   1920     312  463  4
ID   79   1701     218  414  4
WY   42   2088     215  568  4
CO   108  2047     459  621  4
NM   84   1838     438  618  4
AZ   87   1932     425  699  4
UT   99   1753     194  665  4
NV   84   2569     372  663  4
WA   112  2160     446  584  4
OR   95   2006     482  534  4
CA   129  2557     473  717  4
AK   67   1900     334  379  4
HI   107  1852     531  693  4
", header = TRUE)
install.packages("ggplot2")
plot(expenditure[c("Y", "X1", "X2", "X3")])

# 0.5317212  Y   and X1: Positively correlated, strong relationship
cor(expenditure["Y"],expenditure["X1"])

# 0.4482876  Y   and X2: Positively correlated, moderate strength of relationship
cor(expenditure["Y"],expenditure["X2"])

# 0.4636787  Y   and X3: Positively correlated, moderate strength of relationship.
cor(expenditure["Y"],expenditure["X3"])

# 0.2056101  X1   and X2: Positively correlated, weak relationship.
cor(expenditure["X1"],expenditure["X2"])

# 0.5952504  X1   and X3: Positively correlated, strong relationship.
cor(expenditure["X1"],expenditure["X3"])

# 0.2210149  X2   and X3: Positively correlated, weak relationship.
cor(expenditure["X2"],expenditure["X3"])
```

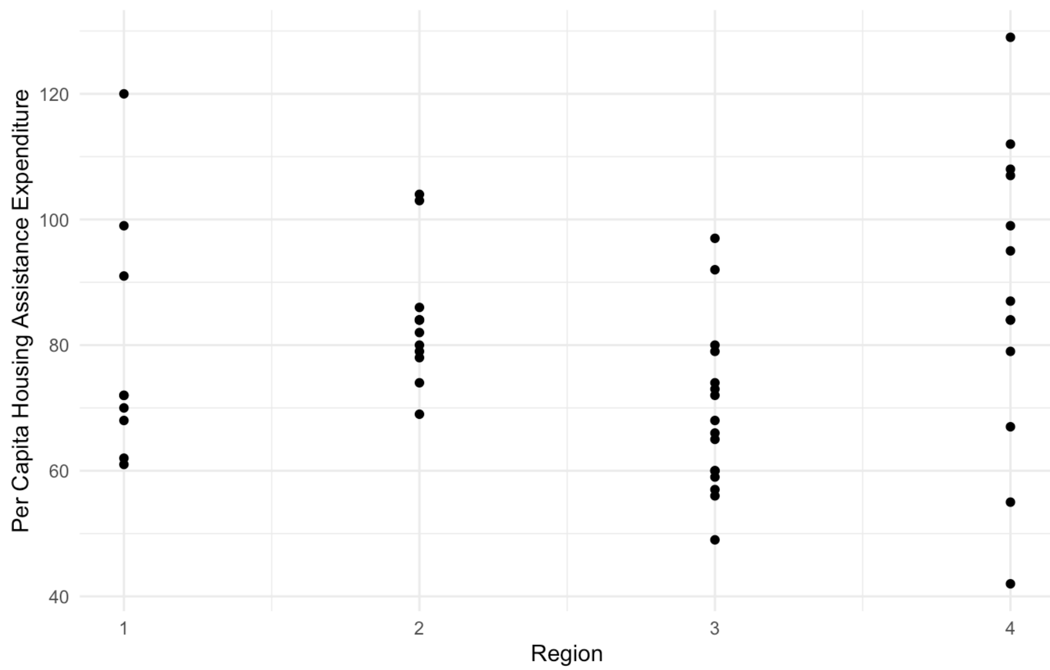


Y and X1: Positively correlated, strong relationship.
Y and X2: Positively correlated, moderate strength of relationship.
Y and X3: Positively correlated, moderate strength of relationship.
X1 and X2: Positively correlated, weak relationship.
X1 and X3: Positively correlated, strong relationship.
X2 and X3: Positively correlated, weak relationship.

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

```r
# 2
library(ggplot2)
# Create a scatter plot to show the relationship between Y and Region
# Calculate the average Y value by Region
```

```r
# Find the region with the highest average Y value
ggplot(data, aes(x = Region, y = Y)) +
  geom_point()  +
  labs(x = "Region", y = "Average Housing Assistance Expenditure (Y)") +
  theme_minimal()

average_by_region <- aggregate(data$Y, by = list(data$Region), FUN = mean)
max_region <- average_by_region[which.max(average_by_region$x), ]
cat("On average, the areas with the highest per capita expenditure on housing assistance are
    :", max_region$Group.1, "\n")
```

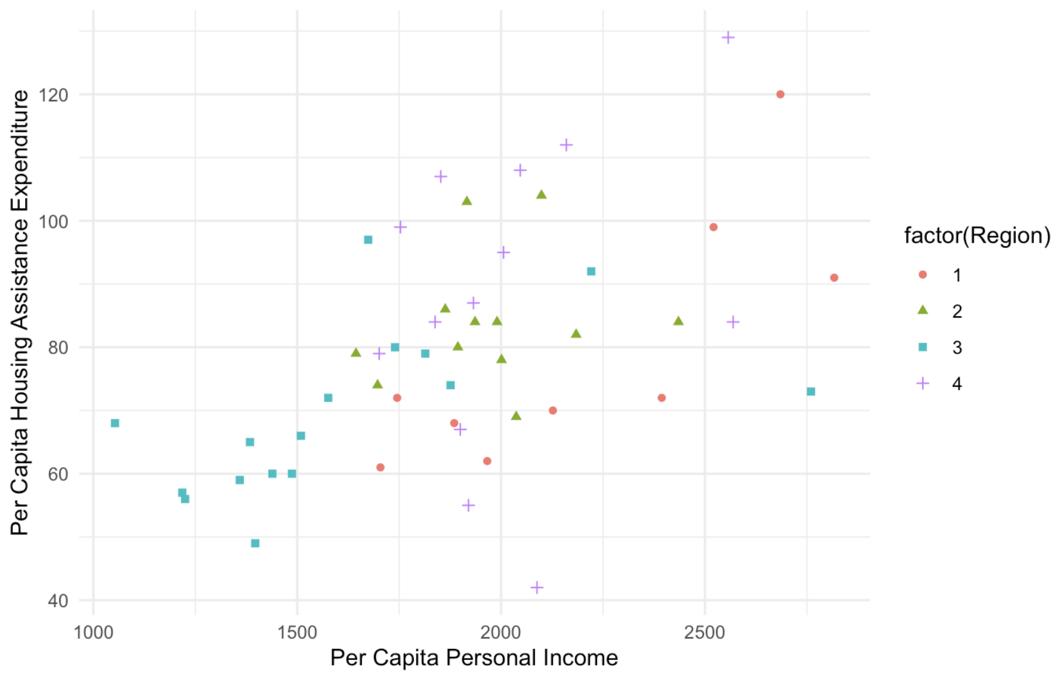


On average, the areas with the highest per capita expenditure on housing assistance are: 4

3. 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.

```r
# 3
# Plot a scatter plot of the relationship between Y and X1,
# using different types of symbols and colors for different regions
ggplot(data, aes(x = X1, y = Y, shape = as.factor(Region), color = as.factor(Region))) +
  geom_point() +
  labs(x = "per capita personal income in state", y = "housing assistance in state") +
  theme_minimal()
# 0.5317212 Y  and X1: Positively correlated, strong relationship
cor(expenditure["Y"],expenditure["X1"])
```

Y and X1: Positively correlated, strong relationshi



Per Capita Housing Assistance Expenditure
120
100
80
60
40
1000
1500
2000
2500
Per Capita Personal Income
factor(Region)
1
2
3
4