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:

```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.

```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)
# capture the number of observations
n <- length(y)
# (a) Calculate the 90% confidence interval for the student IQ
# Step 1: get t-score
```

```r
t <- qt(0.05, n-1, lower.tail = F)
# Step 2: Calculate lower and upper parts for the 90%
lower_CI <- mean(y)-(t*(sd(y)/sqrt(n)))
upper_CI <- mean(y)+(t*(sd(y)/sqrt(n)))
# print CIs with mean
c(lower_CI, mean(y), upper_CI) #Confidence interval (93.95993 102.92007)
    mean value(98.44000)
# double check our answer
t.test(y, conf.level = 0.9)$"conf.int" #Use the t.test() function to
    directly calculate the 90% confidence interval and extract the
    confidence interval
```

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
# (b) Step 1: Calculate the standard error
SE <- sd(y)/sqrt(n)
# Step 2: Calculate the test statistic for this hypothesis testing of
    mean
t <- (mean(y) - 100)/SE
# Get the p-value from t-distribution
pvalue <- pt(t, n-1, lower.tail = F)
# Or another way to do this hypothesis testing is to use the function t.
    test directly
t.test(y, mu = 100, conf.level = 0.95, alternative = "greater")
#                 One Sample t-test
#data:  y
#t = -0.59574, df = 24, p-value = 0.7215
#(The t-value is close to 0, indicating that there is not much difference
     between the sample mean and the assumed mean (100))
#(The p-value is much greater than 0.05, which means there is not enough
    evidence to reject the null hypothesis, i.e. there is no evidence to
    suggest that the sample mean is significantly greater than 100)
#alternative hypothesis: true mean is greater than 100
#(Indicating the hypothesis that the sample mean is greater than 100)
#95 percent confidence interval:
#  93.95993      Inf
#(The lower limit of the confidence interval is 93.95993.The upper limit
    of the confidence interval is infinite)
#sample estimates:
#mean of x
#   98.44
#(The sample mean is 98.44)
```

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`.

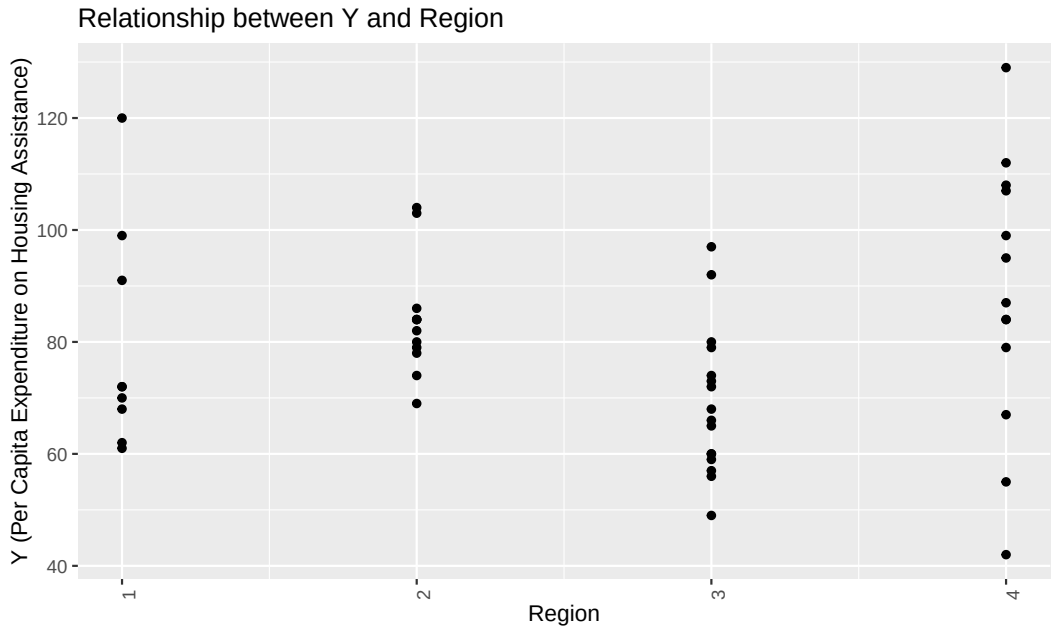
- Please plot the relationships among *Y*, *X1*, *X2*, and *X3*? What are the correlations among them (you just need to describe the graph and the relationships among them)?

```r
# read in expenditure data
# *Need to install and load the HTTR package first
# if (!requireNamespace("httr", quietly = TRUE)) install.packages("httr")
# library(httr)
# url <- "https://raw.githubusercontent.com/ASDS-TCD/StatsI_Fall2024/main/datasets/expenditure.txt"
# response <- GET(url)
# if (response$status_code == 200) {
#             content <- content(response, "text")
#             lines <- unlist(strsplit(content, "\n"))
#             expenditure <- read.table(text = paste0(lines, collapse = "\n"), header = TRUE)
# } else {
#         stop("Failed to download the file: HTTP status ", response$status_code)
# }
expenditure <- read.table("https://raw.githubusercontent.com/ASDS-TCD/StatsI_Fall2024/main/datasets/expenditure.txt", header=T)
pairs(expenditure[, c("Y", "X1", "X2", "X3")], main = "Scatterplot Matrix", pch = 19) #Draw a scatter plot matrix of Y with X1, X2, X3
RJ.C <- cor(expenditure[, c("Y", "X1", "X2", "X3")]) #Calculate correlation
print(RJ.C)
summary(expenditure) #Output the statistical results as a text file
sink("summary.txt")
print(summary(expenditure))
sink() #First question completed
```

```
   STATE                Y                X1              X2              X3
 Length:50          Min.   : 42.00   Min.   :1053   Min.   :111.0   Min.   :326.0
 Class :character   1st Qu.: 67.25   1st Qu.:1698   1st Qu.:187.2   1st Qu.:426.2
 Mode  :character   Median : 79.00   Median :1897   Median :241.5   Median :568.0
                    Mean   : 79.54   Mean   :1912   Mean   :281.8   Mean   :561.7
                    3rd Qu.: 90.00   3rd Qu.:2096   3rd Qu.:391.8   3rd Qu.:661.2
                    Max.   :129.00   Max.   :2817   Max.   :531.0   Max.   :899.0
```

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

```r
install.packages("ggplot2") #Install ggplot2 package to draw charts
library(ggplot2) #Load ggplot2 package
expenditure <- read.table("https://raw.githubusercontent.com/ASDS-TCD/StatsI_Fall2024/main/datasets/expenditure.txt", header=T)
head(expenditure) #Check the first six items of the read webpage text
ggplot(expenditure, aes(x = Region, y = Y)) +
  geom_point() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  labs(title = "Relationship between Y and Region",
       x = "Region",
       y = "Y (Per Capita Expenditure on Housing Assistance)")  #Draw a point of Y and Region
pdf("plot.Y.Region_RJ.C.pdf")
plot(expenditure$Region,expenditure$Y)
dev.off() #Complete the first question of the second question
```

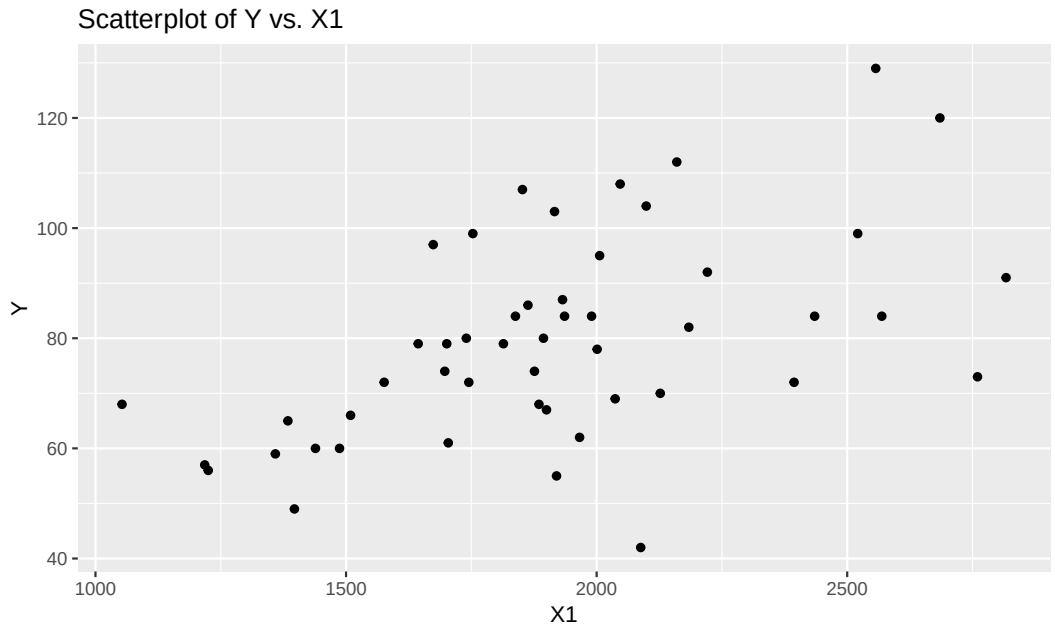


```r
expenditure <- read.table("https://raw.githubusercontent.com/ASDS-TCD/StatsI_Fall2024/main/datasets/expenditure.txt", header=T)
average_expenditure <- aggregate(Y ~ Region, data=expenditure, FUN=mean)
highest_region <- average_expenditure[which.max(average_expenditure$Y),]
print(highest_region)#Complete the second question of the second question
```

```
  Region        Y
4      4 88.30769
```

- 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
library(ggplot2)
expenditure <- read.table("https://raw.githubusercontent.com/ASDS-TCD/StatsI_Fall2024/main/datasets/expenditure.txt", header=T)
ggplot(expenditure, aes(x = X1, y = Y)) +
  geom_point() +
  labs(title = "Scatterplot of Y vs. X1", x = "X1", y = "Y")  #Draw a point of Y and X1
pdf("plot.Y.X1_RJ.C.pdf")
plot(expenditure$X1,expenditure$Y)
dev.off()  #Complete the first question of the third question
```

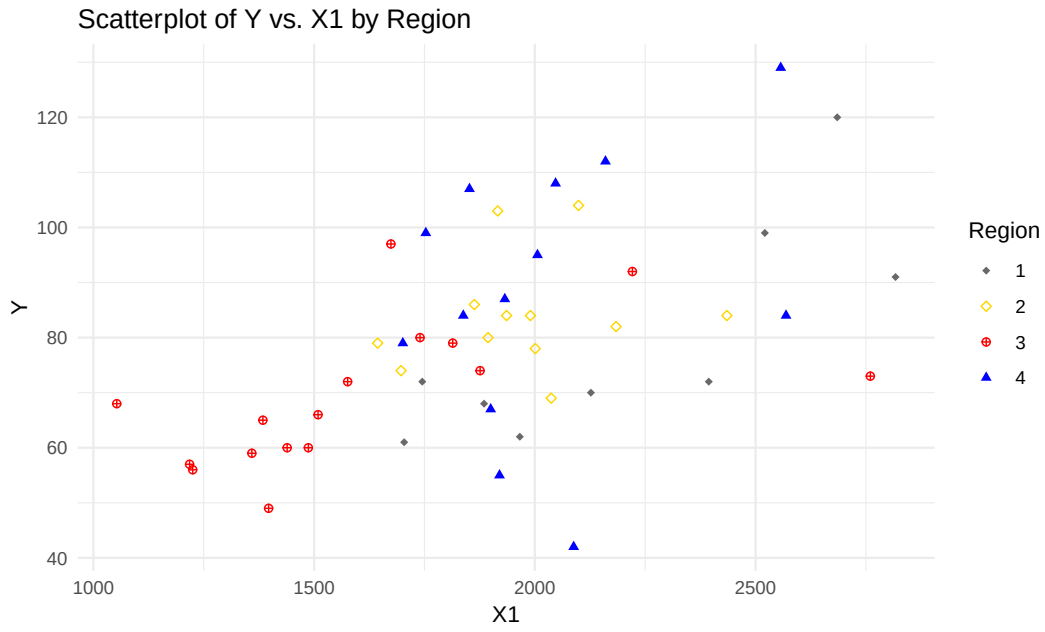


```r
expenditure <- read.table("https://raw.githubusercontent.com/ASDS-TCD/StatsI_Fall2024/main/datasets/expenditure.txt", header=T)
regression_model <- lm(Y ~ X1, data=expenditure)
summary(regression_model)
output_stargazer <- function(outputFile, model) {
  output <- capture.output(stargazer(model, type = "text"))
  writeLines(output, con = outputFile)
}
output_stargazer("regression_output_RJ.C.tex", regression_model) #This will write the output of stargazer to the 'regression_output_RJ.C.tex' file #Complete the second question of the third question
```

Table 1:

	Dependent variable:
	Y
X1	0.025***
	(0.006)
Constant	32.546***
	(11.034)
Observations	50
R^2	0.283
Adjusted R^2	0.268
Residual Std. Error	15.836 (df = 48)
F Statistic	18.920*** (df = 1; 48)
Note:	*p<0.1; **p<0.05; ***p<0.01

```r
library(ggplot2)
expenditure <- read.table("https://raw.githubusercontent.com/ASDS-TCD/StatsI_Fall2024/main/datasets/expenditure.txt", header=T)
expenditure$Region <- as.factor(expenditure$Region)
ggplot(expenditure, aes(x = X1, y = Y, color = Region, shape = Region)) +
  geom_point() +
  labs(title = "Scatterplot of Y vs. X1 by Region", x = "X1", y = "Y") +
  theme_minimal() +
  scale_color_manual(values = c("dimgray", "gold", "red", "blue", "coral")) +
  scale_shape_manual(values = c(18, 23, 10, 17, 2))
pdf("plot.symbols.colors_RJ.C.pdf")
plot(expenditure$X1,expenditure$Y)
dev.off()   #Complete the third question of the third question
```


Scatterplot of Y vs. X1 by Region
120
100
80
60
40
Y
1000
1500
2000
2500
X1
Region
1
2
3
4