

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:

```
1 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.

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

```

6 t <- qt(0.05, n-1, lower.tail = F)
7 # Step 2: Calculate lower and upper parts for the 90%
8 lower_CI <- mean(y)-(t*(sd(y)/sqrt(n)))
9 upper_CI <- mean(y)+(t*(sd(y)/sqrt(n)))
10 # print CIs with mean
11 c(lower_CI, mean(y), upper_CI) #Confidence interval (93.95993 102.92007)
    mean value(98.44000)
12 # double check our answer
13 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 $\alpha = 0.05$.

```

1 # (b) Step 1: Calculate the standard error
2 SE <- sd(y)/sqrt(n)
3 # Step 2: Calculate the test statistic for this hypothesis testing of
    mean
4 t <- (mean(y) - 100)/SE
5 # Get the p-value from t-distribution
6 pvalue <- pt(t, n-1, lower.tail = F)
7 # Or another way to do this hypothesis testing is to use the function t.
    test directly
8 t.test(y, mu = 100, conf.level = 0.95, alternative = "greater")
9 #           One Sample t-test
10 #data:  y
11 #t = -0.59574, df = 24, p-value = 0.7215
12 #(The t-value is close to 0, indicating that there is not much difference
    between the sample mean and the assumed mean (100))
13 #(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)
14 #alternative hypothesis: true mean is greater than 100
15 #(Indicating the hypothesis that the sample mean is greater than 100)
16 #95 percent confidence interval:
17 # 93.95993      Inf
18 #(The lower limit of the confidence interval is 93.95993.The upper limit
    of the confidence interval is infinite)
19 #sample estimates:
20 #mean of x
21 # 98.44
22 #(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)?

```
1 # read in expenditure data
2 # *Need to install and load the HTTPR package first
3 # if (!requireNamespace("httpr", quietly = TRUE)) install.packages("httpr")
4 # library(httpr)
5 # url <- "https://raw.githubusercontent.com/ASDS-TCD/StatsI_Fall2024/main
  /datasets/expenditure.txt"
6 # response <- GET(url)
7 # if (response$status_code == 200) {
8 #   content <- content(response, "text")
9 #   lines <- unlist(strsplit(content, "\n"))
10 #   expenditure <- read.table(text = paste0(lines, collapse =
    "\n"), header = TRUE)
11 # } else {
12 #   stop("Failed to download the file: HTTP status ", response$
    status_code)
13 # }
14 expenditure <- read.table("https://raw.githubusercontent.com/ASDS-TCD/
  StatsI_Fall2024/main/datasets/expenditure.txt", header=T)
15 pairs(expenditure[, c("Y", "X1", "X2", "X3")], main = "Scatterplot Matrix
  ", pch = 19) #Draw a scatter plot matrix of Y with X1, X2, X3
16 RJ.C <- cor(expenditure[, c("Y", "X1", "X2", "X3")]) #Calculate
  correlation
17 print(RJ.C)
18 summary(expenditure) #Output the statistical results as a text file
19 sink("summary.txt")
20 print(summary(expenditure))
21 sink() #complete the first question
```

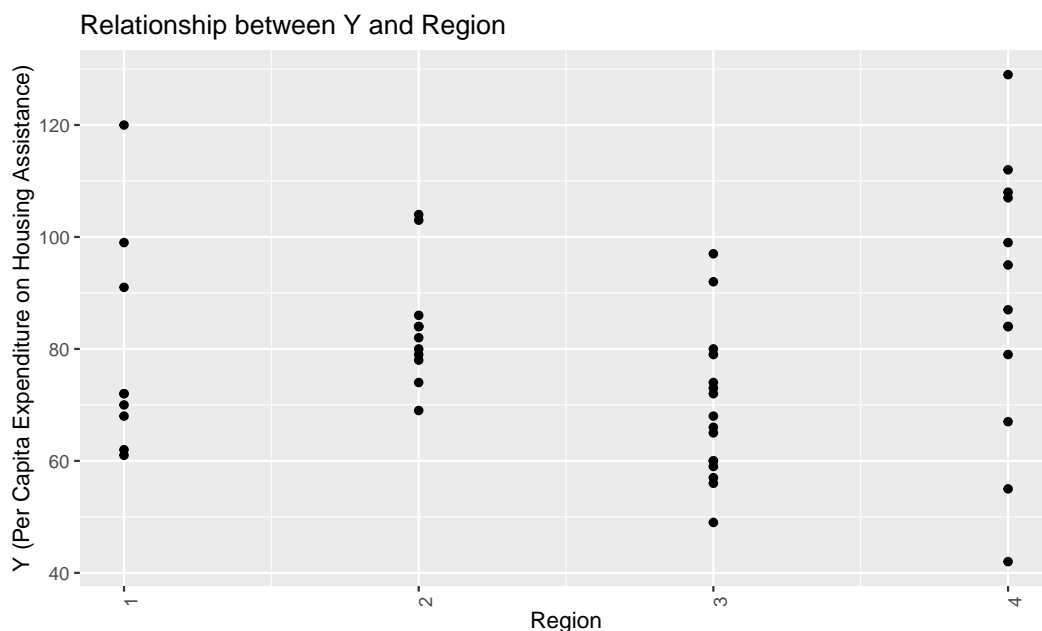
STATE	Y	X1	X2	X3
Length:50	Min. : 42.00	Min. :1053	Min. :111.0	Min. :326.0
Class :character	1st Qu.: 67.25	1st Qu.:1698	1st Qu.:187.2	1st Qu.:426.2
Mode :character	Median : 79.00	Median :1897	Median :241.5	Median :568.0
	Mean : 79.54	Mean :1912	Mean :281.8	Mean :561.7
	3rd Qu.: 90.00	3rd Qu.:2096	3rd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

```

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

```



```

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

```

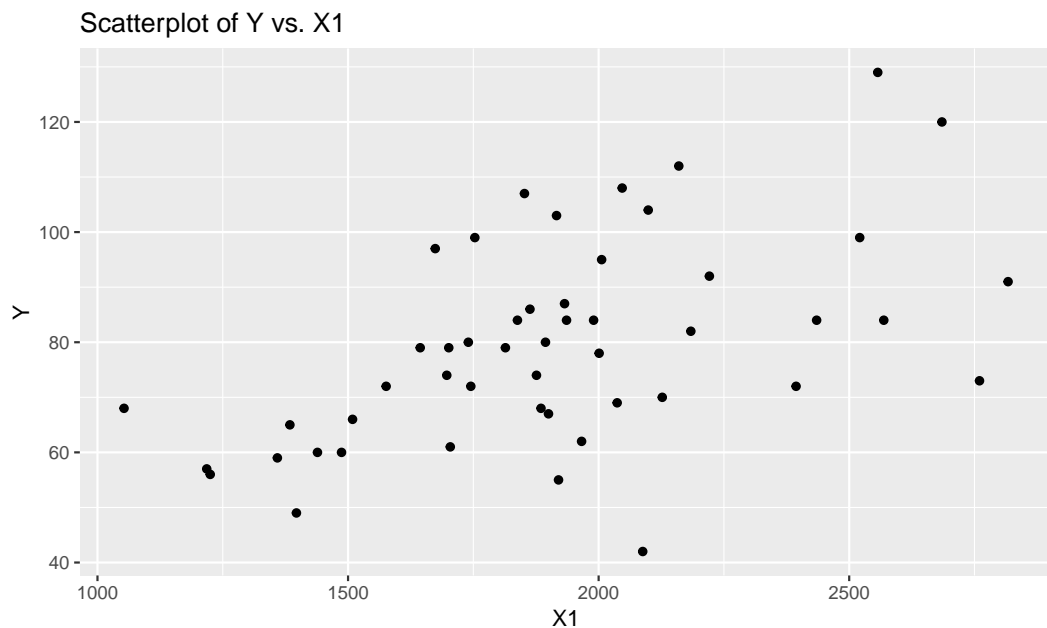
Region	Y
4	4 88.30769

- 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.

```

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

```



```

1 expenditure <- read.table("https://raw.githubusercontent.com/ASDS-TCD/
  StatsI_Fall2024/main/datasets/expenditure.txt", header=T)
2 regression_model <- lm(Y ~ X1, data=expenditure)
3 summary(regression_model)
4 output_stargazer <- function(outputFile, model) {
5   output <- capture.output(stargazer(model, type = "text"))
6   writeLines(output, con = outputFile)
7 }
8 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(Y/X1) of the third question

```

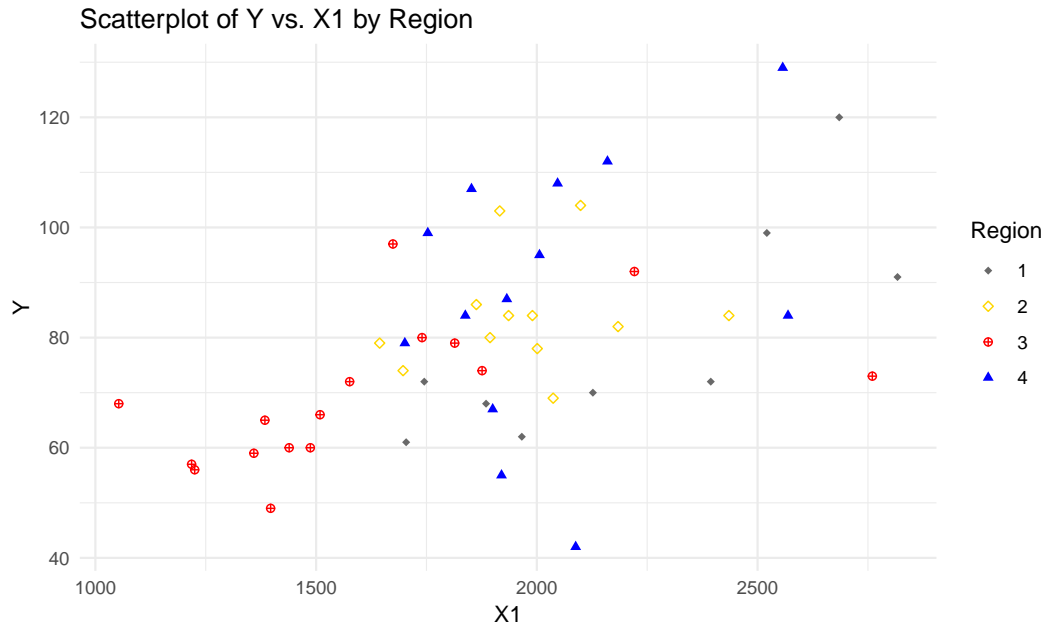
Table 1:

<i>Dependent variable:</i>	
	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)
<i>Note:</i> *p<0.1; **p<0.05; ***p<0.01	

```

1 library(ggplot2)
2 expenditure <- read.table("https://raw.githubusercontent.com/ASDS-TCD/
  StatsI_Fall2024/main/datasets/expenditure.txt", header=T)
3 expenditure$Region <- as.factor(expenditure$Region)
4 ggplot(expenditure, aes(x = X1, y = Y, color = Region, shape = Region)) +
5   geom_point() +
6   labs(title = "Scatterplot of Y vs. X1 by Region", x = "X1", y = "Y") +
7   theme_minimal() +
8   scale_color_manual(values = c("dimgray", "gold", "red", "blue", "coral"
9   )) +
10  scale_shape_manual(values = c(18, 23, 10, 17, 2))
11 pdf("plot_symbols_colors_RJ.C.pdf")
12 plot(expenditure$X1, expenditure$Y)
  dev.off() #Complete the third question(Y/X1) of the third question

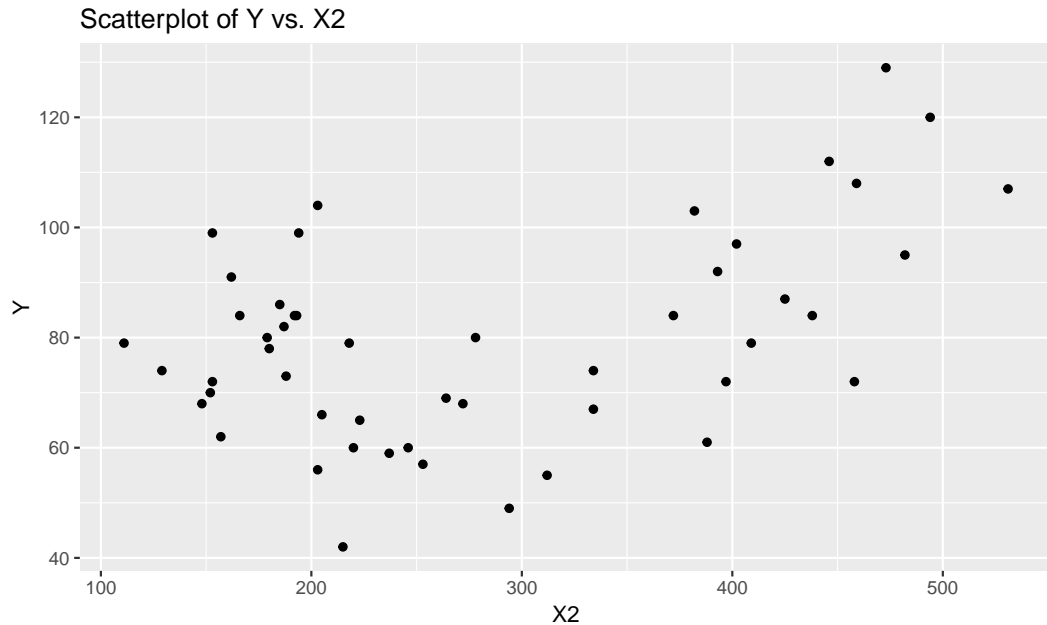
```



```

1 library(ggplot2)
2 expenditure <- read.table("https://raw.githubusercontent.com/ASDS-TCD/
  StatsI_Fall2024/main/datasets/expenditure.txt", header=T)
3 ggplot(expenditure, aes(x = X2, y = Y)) +
4   geom_point() +
5   labs(title = "Scatterplot of Y vs. X2", x = "X2", y = "Y") #Draw a
  point of Y and X2
6 pdf("plot.Y.X2.RJ.C.pdf")
7 plot(expenditure$X2, expenditure$Y)
8 dev.off() #Complete the first question(Y/X2) of the third question

```



```

1 expenditure <- read.table("https://raw.githubusercontent.com/ASDS-TCD/
  StatsI_Fall2024/main/datasets/expenditure.txt", header=T)
2 regression_model <- lm(Y ~ X2, data=expenditure)
3 summary(regression_model)
4 output_stargazer <- function(outputFile, model) {
5   output <- capture.output(stargazer(model, type = "text"))
6   writeLines(output, con = outputFile)
7 }
8 library(stargazer)
9 output_stargazer("regression_output2_RJ.C.tex", regression_model) #
  Complete the second question(Y/X2) of the third question

```

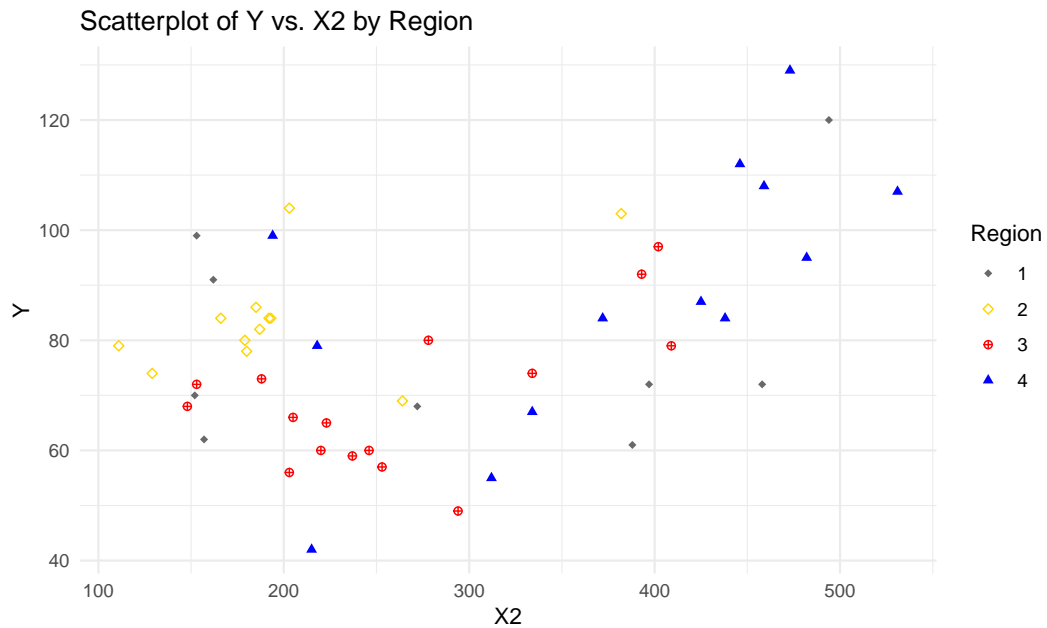
```

1 library(ggplot2)
2 expenditure <- read.table("https://raw.githubusercontent.com/ASDS-TCD/
  StatsI_Fall2024/main/datasets/expenditure.txt", header=T)
3 expenditure$Region <- as.factor(expenditure$Region)
4 ggplot(expenditure, aes(x = X2, y = Y, color = Region, shape = Region)) +
5   geom_point() +
6   labs(title = "Scatterplot of Y vs. X2 by Region", x = "X2", y = "Y") +
7   theme_minimal() +
8   scale_color_manual(values = c("dimgray", "gold", "red", "blue", "coral"
9   )) +
10  scale_shape_manual(values = c(18, 23, 10, 17, 2))
11 pdf("plot.symbols.colors2_RJ.C.pdf")
12 plot(expenditure$X2, expenditure$Y)
13 dev.off() #Complete the third question(Y/X2) of the third question

```


Table 2:

	<i>Dependent variable:</i>
	Y
X2	0.070*** (0.020)
Constant	57.761*** (6.164)
Observations	50
R ²	0.201
Adjusted R ²	0.184
Residual Std. Error	16.714 (df = 48)
F Statistic	12.072*** (df = 1; 48)
<i>Note:</i>	*p<0.1; **p<0.05; ***p<0.01



```

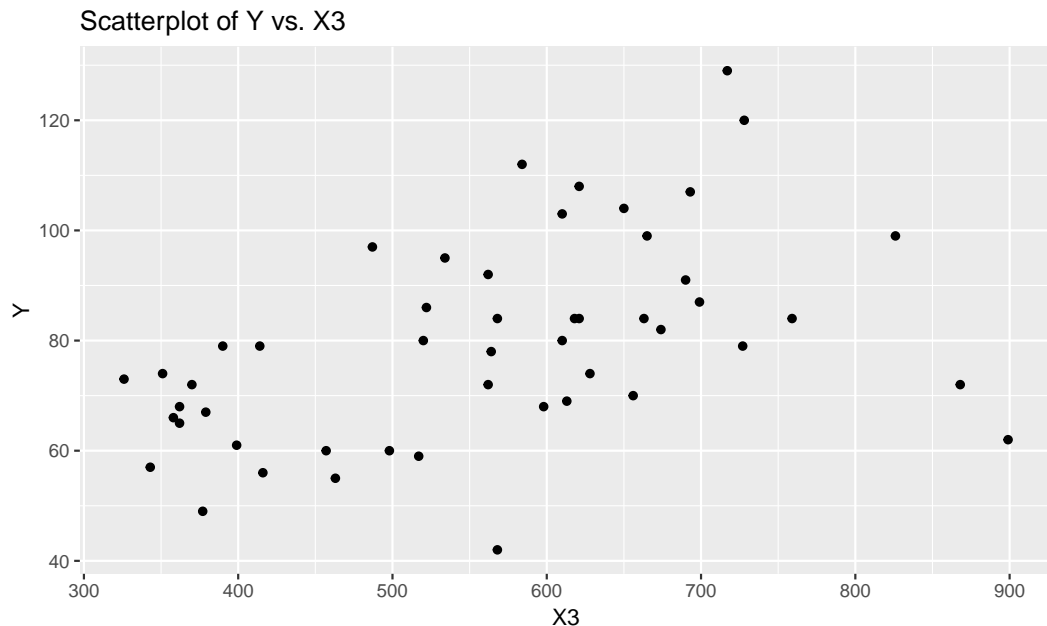
1 library(ggplot2)
2 expenditure <- read.table("https://raw.githubusercontent.com/ASDS-TCD/
  StatsI_Fall2024/main/datasets/expenditure.txt", header=T)
3 ggplot(expenditure, aes(x = X3, y = Y)) +
4   geom_point() +
5   labs(title = "Scatterplot of Y vs. X3", x = "X3", y = "Y") #Draw a
  point of Y and X3

```

```

6 pdf("plot.Y.X3-RJ.C.pdf")
7 plot(expenditure$X3, expenditure$Y)
8 dev.off() #Complete the first question(Y/X3) of the third question

```



```

1 expenditure <- read.table("https://raw.githubusercontent.com/ASDS-TCD/
  StatsI_Fall2024/main/datasets/expenditure.txt", header=T)
2 regression_model <- lm(Y ~ X2, data=expenditure)
3 summary(regression_model)
4 output_stargazer <- function(outputFile, model) {
5   output <- capture.output(stargazer(model, type = "text"))
6   writeLines(output, con = outputFile)
7 }
8 library(stargazer)
9 output_stargazer("regression_output2-RJ.C.tex", regression_model) #
  Complete the second question(Y/X2) of the third question

```

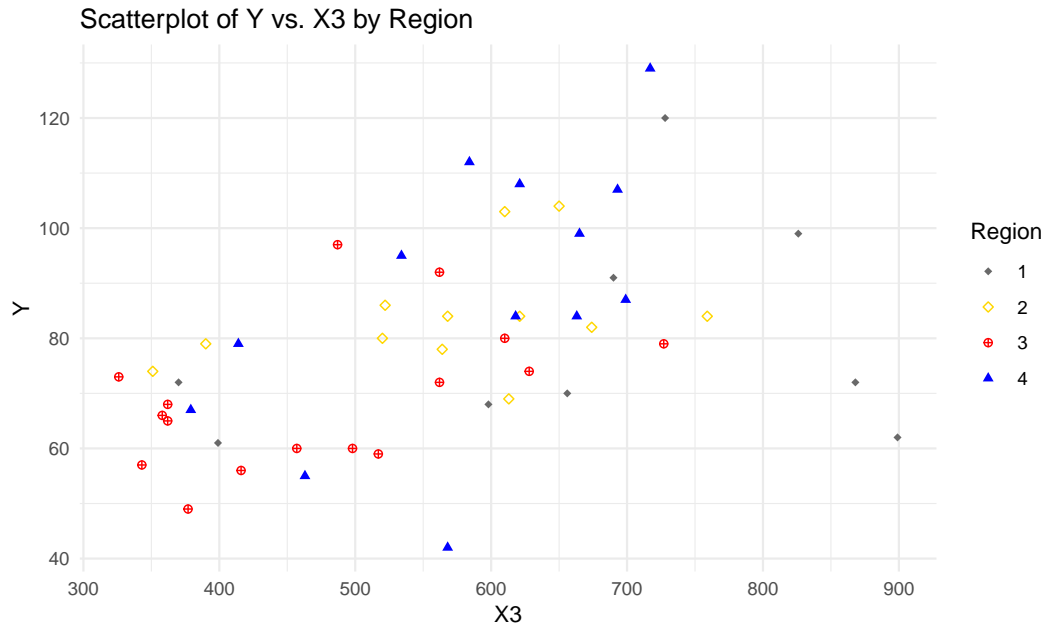
```

1 library(ggplot2)
2 expenditure <- read.table("https://raw.githubusercontent.com/ASDS-TCD/
  StatsI_Fall2024/main/datasets/expenditure.txt", header=T)
3 expenditure$Region <- as.factor(expenditure$Region)
4 ggplot(expenditure, aes(x = X3, y = Y, color = Region, shape = Region)) +
5   geom_point() +
6   labs(title = "Scatterplot of Y vs. X3 by Region", x = "X3", y = "Y") +
7   theme_minimal() +
8   scale_color_manual(values = c("dimgray", "gold", "red", "blue", "coral")
9   ) +
10  scale_shape_manual(values = c(18, 23, 10, 17, 2))
11 pdf("plot.symbols.colors3-RJ.C.pdf")
12 plot(expenditure$X3, expenditure$Y)
13 dev.off() #Complete the third question(Y/X3) of the third question

```

Table 3:

	<i>Dependent variable:</i>
	Y
X3	0.059*** (0.016)
Constant	43.306*** (9.461)
Observations	50
R ²	0.215
Adjusted R ²	0.199
Residual Std. Error	16.567 (df = 48)
F Statistic	13.1146*** (df = 1; 48)
<i>Note:</i>	*p<0.1; **p<0.05; ***p<0.01



```

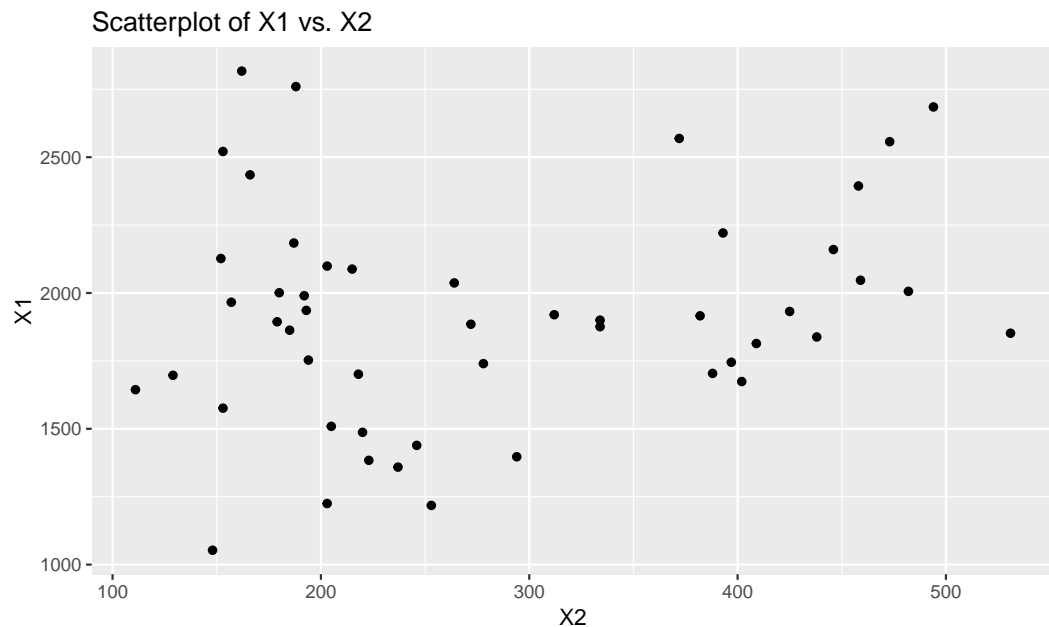
1 library(ggplot2)
2 expenditure <- read.table("https://raw.githubusercontent.com/ASDS-TCD/
  StatsI_Fall2024/main/datasets/expenditure.txt", header=T)
3 ggplot(expenditure, aes(x = X2, y = X1)) +
4   geom_point() +
5   labs(title = "Scatterplot of X1 vs. X2", x = "X2", y = "X1") #Draw a
  point of X1 and X2

```

```

6 pdf("plot.X1.X2.RJ.C.pdf")
7 plot(expenditure$X2, expenditure$X1)
8 dev.off()#Complete the first question(X1/X2) of the third question

```



```

1 expenditure <- read.table("https://raw.githubusercontent.com/ASDS-TCD/
  StatsI_Fall2024/main/datasets/expenditure.txt", header=T)
2 regression_model <- lm(X1 ~ X2, data=expenditure)
3 summary(regression_model)
4 output_stargazer <- function(outputFile, model) {
5   output <- capture.output(stargazer(model, type = "text"))
6   writeLines(output, con = outputFile)
7 }
8 library(stargazer)
9 output_stargazer("regression_output4_RJ.C.tex", regression_model) #
  Complete the second question(X1/X2) of the third question

```

```

1 library(ggplot2)
2 expenditure <- read.table("https://raw.githubusercontent.com/ASDS-TCD/
  StatsI_Fall2024/main/datasets/expenditure.txt", header=T)
3 expenditure$Region <- as.factor(expenditure$Region)
4 ggplot(expenditure, aes(x = X2, y = X1, color = Region, shape = Region))
5   +
6   geom_point() +
7   labs(title = "Scatterplot of X1 vs. X2 by Region", x = "X2", y = "X1")
8   +
9   theme_minimal() +
10  scale_color_manual(values = c("dimgray", "gold", "red", "blue", "coral")
11  ) +
12  scale_shape_manual(values = c(18, 23, 10, 17, 2))

```

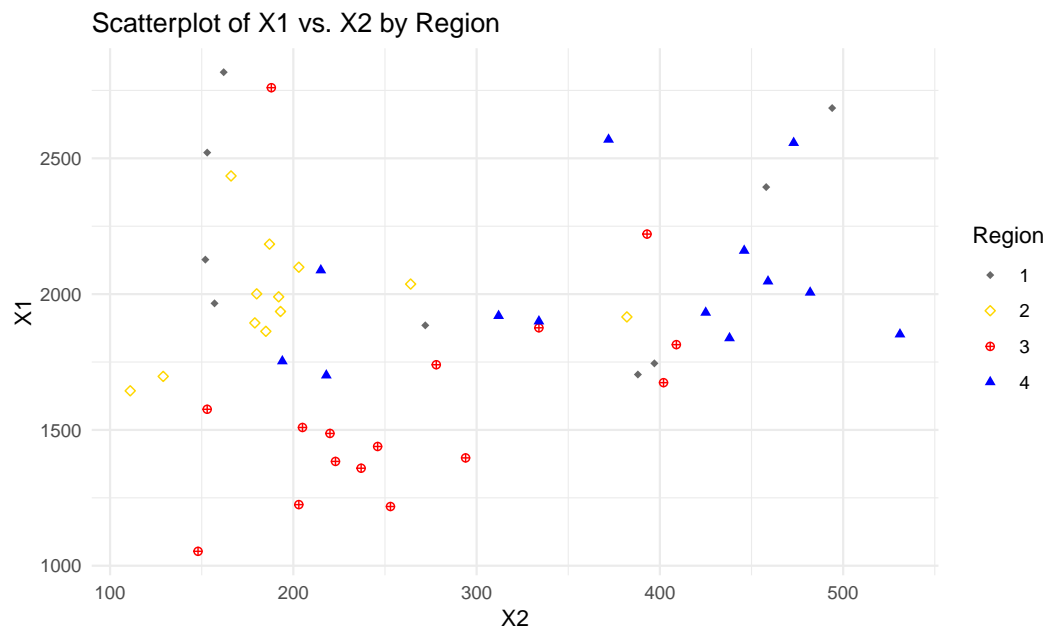
Table 4:

	<i>Dependent variable:</i>
	X1
X2	0.696*** (0.478)
Constant	1715.655*** (145.981)
Observations	50
R ²	0.042
Adjusted R ²	0.022
Residual Std. Error	395.854 (df = 48)
F Statistic	2.119*** (df = 1; 48)
<i>Note:</i> *p<0.1; **p<0.05; ***p<0.01	

```

10 pdf("plot.symbols.colors4_RJ.C.pdf")
11 plot(expenditure$X2, expenditure$X1)
12 dev.off() #Complete the third question(X1/X2) of the third question

```



```

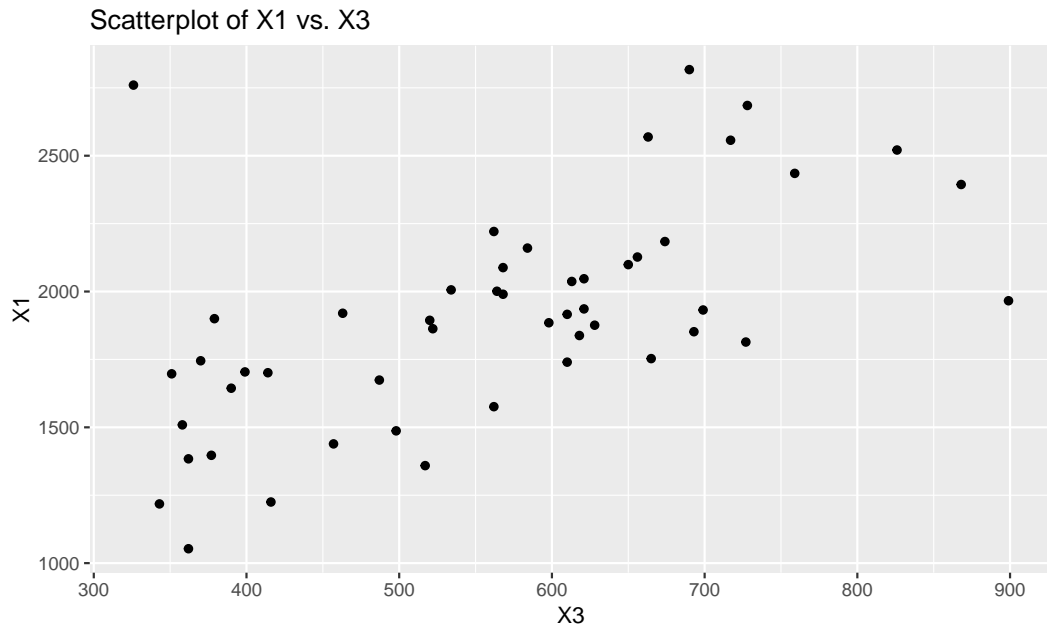
1 library(ggplot2)
2 expenditure <- read.table("https://raw.githubusercontent.com/ASDS-TCD/
  StatsI_Fall2024/main/datasets/expenditure.txt", header=T)

```

```

3 ggplot(expenditure, aes(x = X3, y = X1)) +
4   geom_point() +
5   labs(title = "Scatterplot of X1 vs. X3", x = "X3", y = "X1") #Draw a
6   point of X1 and X3
7 pdf("plot.X1.X3.RJ.C.pdf")
8 plot(expenditure$X3, expenditure$X1)
9 dev.off()#Complete the first question(X1/X3) of the third question

```



```

1 expenditure <- read.table("https://raw.githubusercontent.com/ASDS-TCD/
2   StatsI_Fall2024/main/datasets/expenditure.txt", header=T)
3 regression_model <- lm(X1 ~ X3, data=expenditure)
4 summary(regression_model)
5 output_stargazer <- function(outputFile, model) {
6   output <- capture.output(stargazer(model, type = "text"))
7   writeLines(output, con = outputFile)
8 }
9 library(stargazer)
10 output_stargazer("regression_output5.RJ.C.tex", regression_model) #
11   Complete the second question(X1/X3) of the third question

```

```

1 library(ggplot2)
2 expenditure <- read.table("https://raw.githubusercontent.com/ASDS-TCD/
3   StatsI_Fall2024/main/datasets/expenditure.txt", header=T)
4 expenditure$Region <- as.factor(expenditure$Region)
5 ggplot(expenditure, aes(x = X3, y = X1, color = Region, shape = Region))
6   +
7   geom_point() +
8   labs(title = "Scatterplot of X1 vs. X3 by Region", x = "X3", y = "X1")
9   +

```

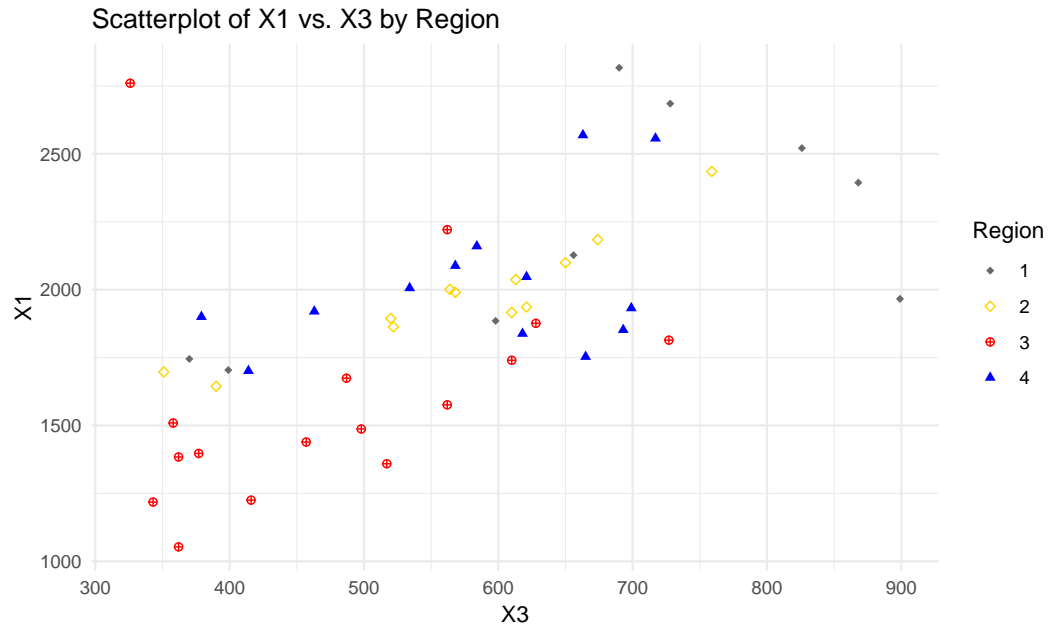
Table 5:

	<i>Dependent variable:</i>
	X1
X3	1.643*** (0.320)
Constant	988.947*** (185.614)
Observations	50
R ²	0.354
Adjusted R ²	0.341
Residual Std. Error	325.029 (df = 48)
F Statistic	26.341*** (df = 1; 48)
<i>Note:</i>	*p<0.1; **p<0.05; ***p<0.01

```

7  theme_minimal() +
8  scale_color_manual(values = c("dimgray", "gold", "red", "blue", "coral"
9  )) +
9  scale_shape_manual(values = c(18, 23, 10, 17, 2))
10 pdf("plot.symbols.colors5_RJ.C.pdf")
11 plot(expenditure$X3, expenditure$X1)
12 dev.off()    #Complete the third question(X1/X3) of the third question

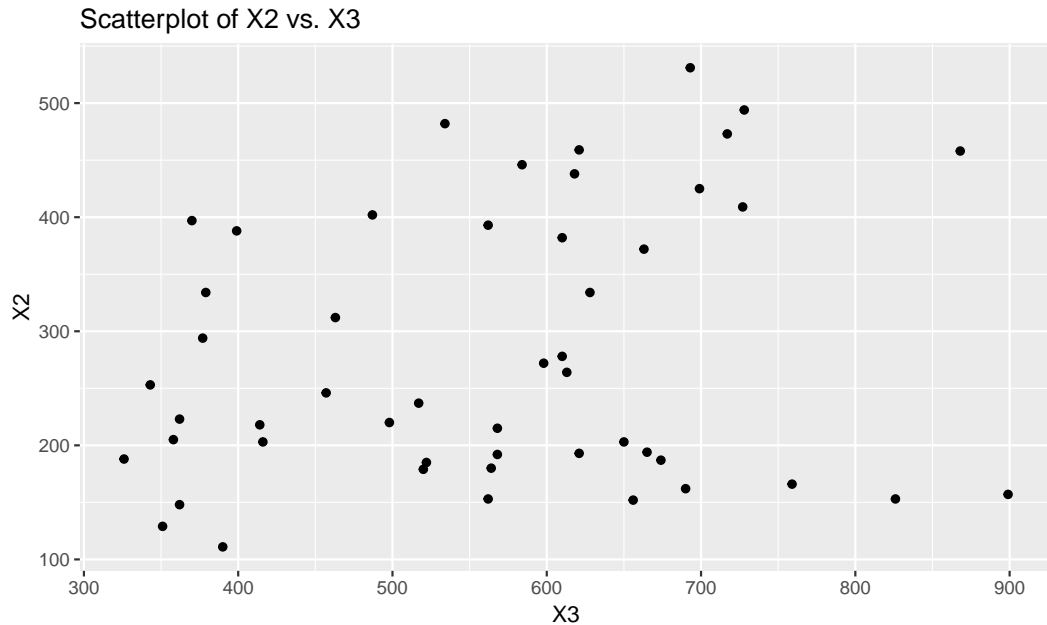
```



```

1 library(ggplot2)
2 expenditure <- read.table("https://raw.githubusercontent.com/ASDS-TCD/
  StatsI_Fall2024/main/datasets/expenditure.txt", header=T)
3 ggplot(expenditure, aes(x = X3, y = X2)) +
4   geom_point() +
5   labs(title = "Scatterplot of X2 vs. X3", x = "X3", y = "X2") #Draw a
  point of X2 and X3
6 pdf("plot.X2.X3.RJ.C.pdf")
7 plot(expenditure$X3, expenditure$X2)
8 dev.off()#Complete the first question(X2/X3) of the third question

```

```

1 expenditure <- read.table("https://raw.githubusercontent.com/ASDS-TCD/
  StatsI_Fall2024/main/datasets/expenditure.txt", header=T)
2 regression_model <- lm(X2 ~ X3, data=expenditure)
3 summary(regression_model)
4 output_stargazer <- function(outputFile, model) {
5   output <- capture.output(stargazer(model, type = "text"))
6   writeLines(output, con = outputFile)
7 }
8 library(stargazer)
9 output_stargazer("regression_output6_RJ.C.tex", regression_model) #
  Complete the second question(X2/X3) of the third question

```

```

1 library(ggplot2)
2 expenditure <- read.table("https://raw.githubusercontent.com/ASDS-TCD/
  StatsI_Fall2024/main/datasets/expenditure.txt", header=T)
3 expenditure$Region <- as.factor(expenditure$Region)
4 ggplot(expenditure, aes(x = X3, y = X2, color = Region, shape = Region))
5   +
6   geom_point() +
7   labs(title = "Scatterplot of X2 vs. X3 by Region", x = "X3", y = "X2")
8   +
9   theme_minimal() +
10  scale_color_manual(values = c("dimgray", "gold", "red", "blue", "coral"
11    )) +
12  scale_shape_manual(values = c(18, 23, 10, 17, 2))
13 pdf("plot.symbols.colors6_RJ.C.pdf")
14 plot(expenditure$X3, expenditure$X1)
15 dev.off() #Complete the third question(X2/X3) of the third question

```

Table 6:

<i>Dependent variable:</i>	
X2	
X3	0.180*** (0.115)
Constant	180.609*** (66.509)
Observations	50
R ²	0.049
Adjusted R ²	0.029
Residual Std. Error	116.465 (df = 48)
F Statistic	2.465*** (df = 1; 48)
<i>Note:</i> *p<0.1; **p<0.05; ***p<0.01	

