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 ######
2 # here is answer number 1
3
4 # As the sample size is less than 30 (n=25), it's more appropriate to use
5 # the t-distribution rather than the z-distribution.
```

```
6  # The t-distribution better reflects the uncertainty in sampling
       distribution
7  # for small sample sizes.
8  # Use t.test function to directly calculate the confidence interval
9  CI_result <- t.test(y, conf.level = 0.90, alternative = "two.sided", mu =
       0)
10
11 # Print the results
12 cat("90% Confidence Interval:", round(CI_result$conf.int[1], 2), "to",
       round(CI_result$conf.int[2], 2), "\n")
13
14 # Print mean for comparison
15 cat("Sample Mean:", round(mean(y), 2), "\n")
16
17 #here are my conclusions
18 # We are 90% confident that the true population mean IQ for students in
       this school
19 # falls between 93.96 to 102.92. The sample mean IQ is 98.44.
```

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  ######
2  # here is answer number 2
3
4  # Perform one-tailed t-test
5  test_result <- t.test(y, mu = 100, alternative = "greater")
6
7  # Print results
8  cat("Hypothesis Test Results:\n")
9  cat("t-statistic:", round(test_result$statistic, 4), "\n")
10 cat("p-value:", round(test_result$p.value, 4), "\n")
11 cat("95% Confidence Interval:\n")
12 cat("   Lower bound:", round(test_result$conf.int[1], 2), "\n")
13 cat("   Upper bound:", round(test_result$conf.int[2], 2), "\n")
14 cat("Sample mean:", round(test_result$estimate, 2), "\n")
15
16 #here are my conclusions
17 # The calculated t-statistic is -0.5957 and the p-value is 0.7215.
18 # Since the p-value (0.7215) is greater than our significance level
       (0.05),
19 # we fail to reject the null hypothesis. There is not enough statistical
20 # evidence to conclude that the average IQ in this school is
       significantly
21 # higher than 100 (the national average).
22 # The school counselor should interpret this result as indicating that
23 # the school's average IQ is not significantly different from the
       national
24 # average of 100, based on this sample.
```

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`.

```
1 # read in expenditure data
2 expenditure <- read.table("https://raw.githubusercontent.com/ASDS-TCD/StatsI_
    Fall2024/main/datasets/expenditure.txt", header=T)
3 # explore the data
4 str(expenditure)  # Examine the structure of the dataframe
5 summary(expenditure)  # Get summary statistics for all variables
```

```
STATE      Y               X1             X2              X3             Region
   Min.   : 42.00   Min.   :1053   Min.   :111.0   Min.   :326.0   Min.   :1.00
   1st Qu.: 67.25   1st Qu.:1698   1st Qu.:187.2   1st Qu.:426.2   1st Qu.:2.00
   Median : 79.00   Median :1897   Median :241.5   Median :568.0   Median :3.00
   Mean   : 79.54   Mean   :1912   Mean   :281.8   Mean   :561.7   Mean   :2.66
   3rd Qu.: 90.00   3rd Qu.:2096   3rd Qu.:391.8   3rd Qu.:661.2   3rd Qu.:3.75
   Max.   :129.00   Max.   :2817   Max.   :531.0   Max.   :899.0   Max.   :4.00
```

- 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 ######
2 # here is answer number 1
3
4 #plot all of the bivariates relationships between the outcome and the
    predictors
5 #as well as the predictors themselves
6 #To visualize relationships between these variables
7 #we can create a scatter plot matrix using the pairs() function
8 pdf("plot_pairs.pdf")
9 pairs(expenditure[, c("Y", "X1", "X2", "X3")])
10 dev.off()
```

Figure 1: The Plot Relationships Among Y, X1, X2, X3
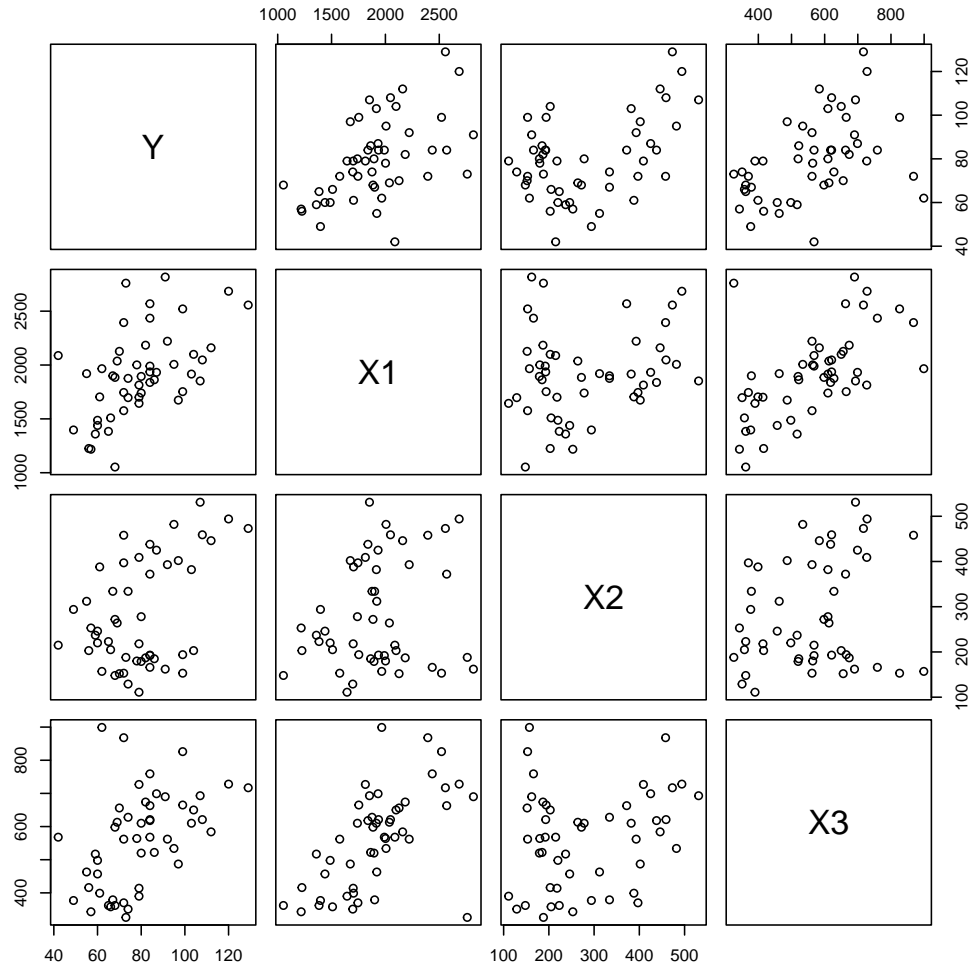
```
11 # Calculate correlation coefficients
12 cor(expenditure[,c("Y", "X1", "X2", "X3")])
13
14 #here are my conclusions
15 #Y and X1 show a strong positive correlation (r = 0.5317). The
       scatterplot clearly displays a positive linear relationship.
16 #Y and X2 have a moderate positive correlation (r = 0.4483). The
       scatterplot indicates a positive relationship, though not as strong as
       Y–X1.
17 #Y and X3 also exhibit a moderate positive correlation (r = 0.4637). The
       scatterplot shows a similar positive trend.
18 #X1 and X2 have a weak positive correlation (r = 0.2056). The scatterplot
       shows a more dispersed pattern with no clear linear trend.
19 #X1 and X3 demonstrate a strong positive correlation (r = 0.5925). This
```

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

Figure 2: The Plot Relationships Between Y and Region

Per Capita Expenditure on Housing Assistance by Region
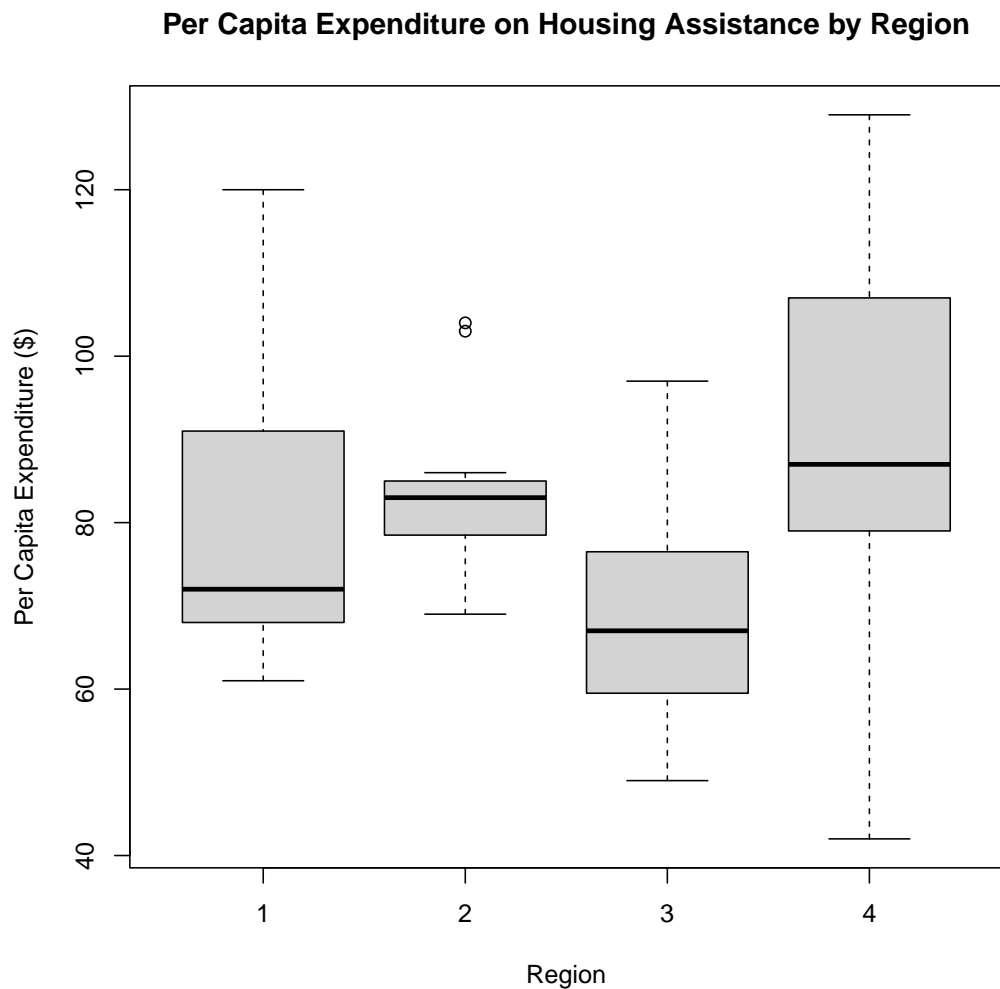
```
1 ######
2 # here is answer number 2
3
4 #Plot the relationship between Y and Region
5 #To visualize this relationship, we can use a boxplot
6 pdf("plot_boxplot.pdf")
```

```
7  boxplot(Y ~ Region, data = expenditure,
8          main = "Per Capita Expenditure on Housing Assistance by Region",
9          xlab = "Region",
10         ylab = "Per Capita Expenditure ($)")
11 dev.off()
12
13 #here are my conclusions
14 #On average, Region 4 (West) has the highest per capita expenditure on
      housing assistance.
15 #This can be seen from the box plot, where Region 4
16 #has the highest median (represented by the thick black line in the
      middle of the box)
17 #and the largest interquartile range.
```

- 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  ######
2  # here is answer number 3
3
4  #Plot the relationship between Y and X1
5  pdf("plot_X1&Y.pdf")
6  plot(expenditure$X1, expenditure$Y,
7       xlab="X1: Per Capita Personal Income",
8       ylab="Y: Per Capita Expenditure on Housing Assistance",
9       main="Relationship between Income and Housing Assistance Expenditure
      ")
10 # Add regression line
11 abline(lm(Y ~ X1, data=expenditure), col="red")
12 dev.off()
13
14 # Calculate correlation coefficient
15 cor(expenditure$Y, expenditure$X1)
16
17 #install stargazer packages
18 install.packages("stargazer")
19 library(stargazer)
20
21 # Run and show regression analysis
22 regression1 <- lm(Y ~ X1, data=expenditure)
23
24 output_stargazer <- function(outputFile, ...) {
25    output <- capture.output(stargazer(...))
26    cat(paste(output, collapse = "\n"), "\n", file=outputFile, append=TRUE)
27 }
28 output_stargazer("regression_output1.tex", regression1)
29
30 #Plot the relationship between Y and X1 and region
31 #display different regions with different types of symbols and colors.
```

```r
32 pdf("plot_X1&Y&region.pdf")
33 plot(expenditure$X1, expenditure$Y,
34     col = expenditure$Region,
35     pch = as.numeric(expenditure$Region),  # Different symbol for each
    region
36     xlab = "X1: Per Capita Personal Income",
37     ylab = "Y: Per Capita Expenditure on Housing Assistance",
38     main = "Relationship between Income and Housing Assistance
    Expenditure")
39 # Add legend
40 legend("topright",
41         legend = c("Northeast", "North Central", "South", "West"),
42         col = 1:4,
43         pch = 1:4)
44 dev.off()
45
46 #here are my conclusions
47 #The relationship between Y (Per Capita Expenditure on Housing Assistance
    )
48 #and X1 (Per Capita Personal Income) shows a positive correlation.
49
50 #The scatter plot displays a clear upward trend, with the
51 #red regression line indicating a positive relationship. As personal
    income
52 #increases, there's a general tendency for housing assistance expenditure
     to increase as well.
53
54 #The correlation coefficient between X1 and Y is 0.5317212, indicating a
    moderate positive correlation.
55
56 #The regression coefficient for X1 is 0.025, which is statistically
    significant
57 #at the 1% level (p<0.01). This means that for every $1 increase in per
    capita
58 #personal income, we expect an average increase of $0.025 in per capita
    housing assistance expenditure.
59
60 #The R  value is 0.283, indicating that about 28.3% of the variation in
61 #housing assistance expenditure is explained by personal income.
62
63 #The F-statistic (18.920, p<0.01) suggests that the overall model is
    statistically significant.
```

Relationship between Income and Housing Assistance Expenditure
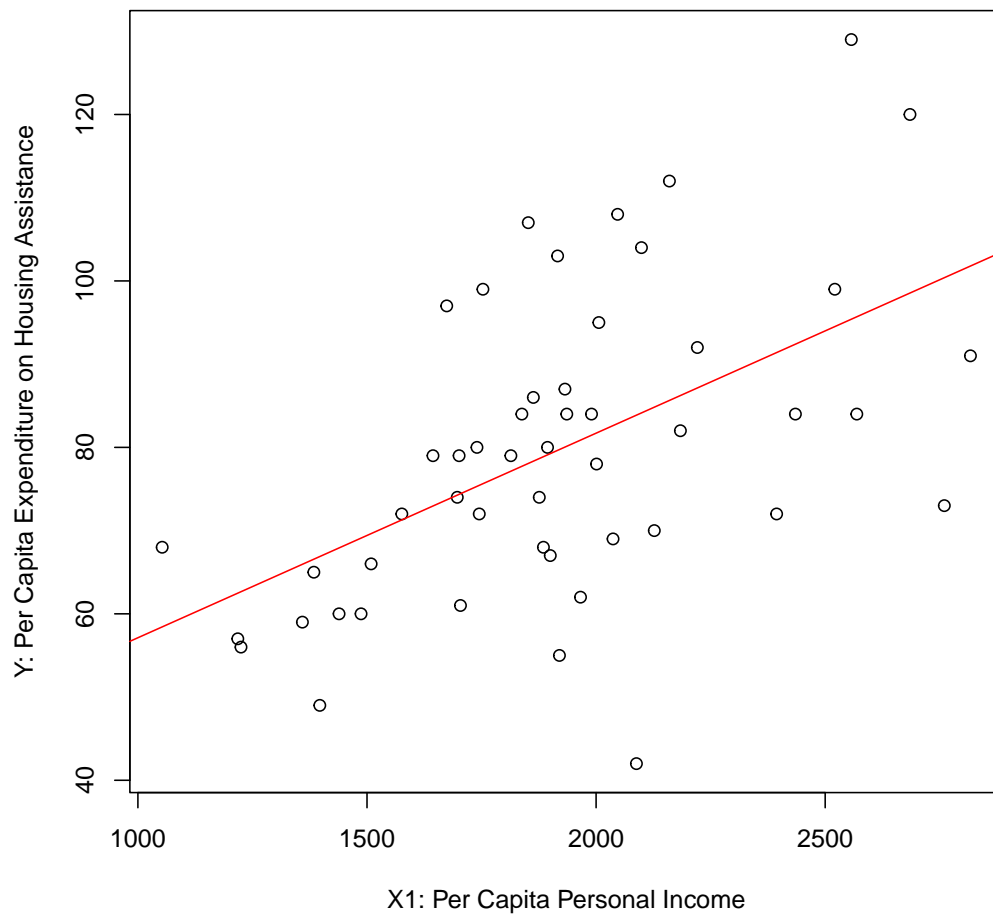
Table 1

	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

1

Figure 4: The Plot Relationships Between Y and X1 and Region

Relationship between Income and Housing Assistance Expenditure