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
######
# here is answer number 1

# As the sample size is less than 30 (n=25), it's more appropriate to use
# the t-distribution rather than the z-distribution.
```

```r
# The t-distribution better reflects the uncertainty in sampling distribution
# for small sample sizes.
# Use t.test function to directly calculate the confidence interval
CI_result <- t.test(y, conf.level = 0.90, alternative = "two.sided", mu = 0)

# Print the results
cat("90% Confidence Interval:", round(CI_result$conf.int[1], 2), "to", round(CI_result$conf.int[2], 2), "\n")

# Print mean for comparison
cat("Sample Mean:", round(mean(y), 2), "\n")

#here are my conclusions
# We are 90% confident that the true population mean IQ for students in this school
# 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 $\alpha = 0.05$.

```r
######
# here is answer number 2

# Perform one-tailed t-test
test_result <- t.test(y, mu = 100, alternative = "greater")

# Print results
cat("Hypothesis Test Results:\n")
cat("t-statistic:", round(test_result$statistic, 4), "\n")
cat("p-value:", round(test_result$p.value, 4), "\n")
cat("95% Confidence Interval:\n")
cat("  Lower bound:", round(test_result$conf.int[1], 2), "\n")
cat("  Upper bound:", round(test_result$conf.int[2], 2), "\n")
cat("Sample mean:", round(test_result$estimate, 2), "\n")

#here are my conclusions
# The calculated t-statistic is -0.5957 and the p-value is 0.7215.
# Since the p-value (0.7215) is greater than our significance level (0.05),
# we fail to reject the null hypothesis. There is not enough statistical
# evidence to conclude that the average IQ in this school is significantly
# higher than 100 (the national average).
# The school counselor should interpret this result as indicating that
# the school's average IQ is not significantly different from the national
# 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`.

```
# read in expenditure data
expenditure <- read.table("https://raw.githubusercontent.com/ASDS-TCD/StatsI_Fall2024/main/datasets/expenditure.txt", header=T)
# explore the data
str(expenditure)  # Examine the structure of the dataframe
summary(expenditure)  # Get summary statistics for all variables
```

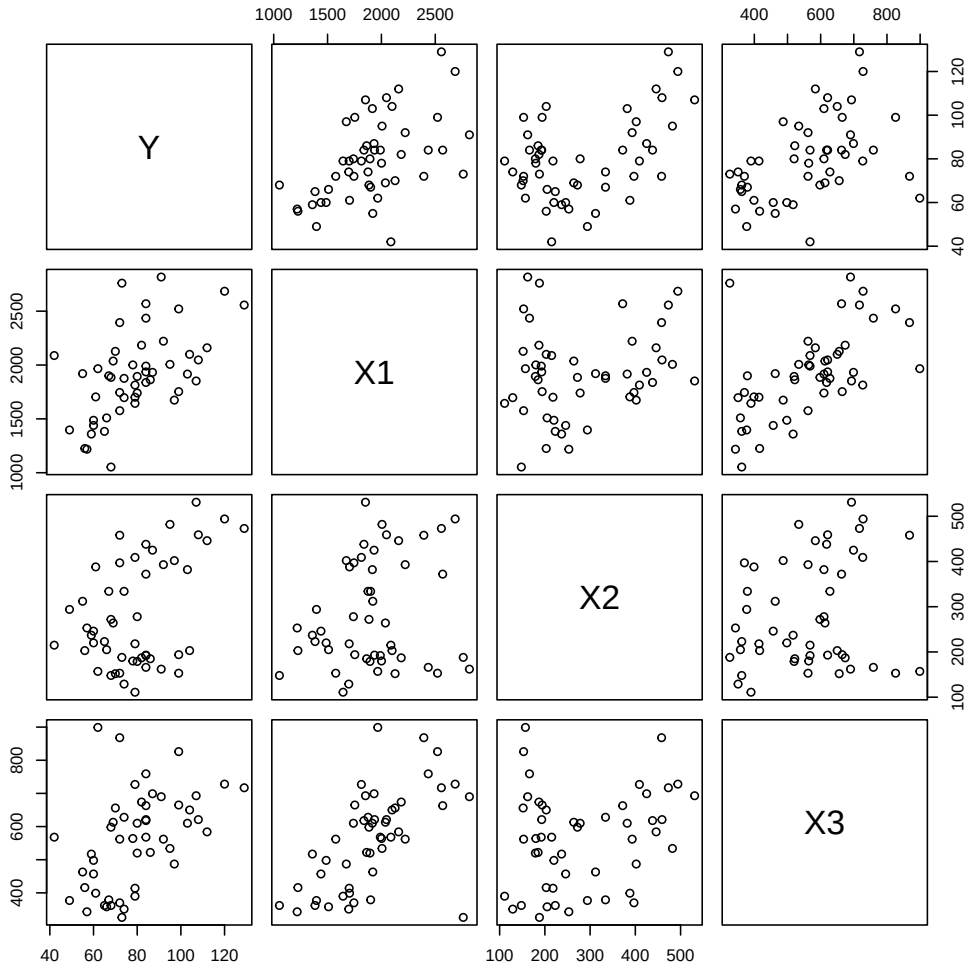
```
STATE    Y               X1             X2             X3             Region
    Min.   : 42.00   Min.   :1053   Min.   :111.0   Min.   :326.0   Min.   :1.00
    1st Qu.: 67.25   1st Qu.:1698   1st Qu.:187.2   1st Qu.:426.2   1st Qu.:2.00
    Median : 79.00   Median :1897   Median :241.5   Median :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)?

```
######
# here is answer number 1

#plot all of the bivariates relationships between the outcome and the predictors
#as well as the predictors themselves
#To visualize relationships between these variables
#we can create a scatter plot matrix using the pairs() function
pdf("plot_pairs.pdf")
pairs(expenditure[, c("Y", "X1", "X2", "X3")])
dev.off()
```

Figure 1: The Plot Relationships Among Y, X1, X2, X3



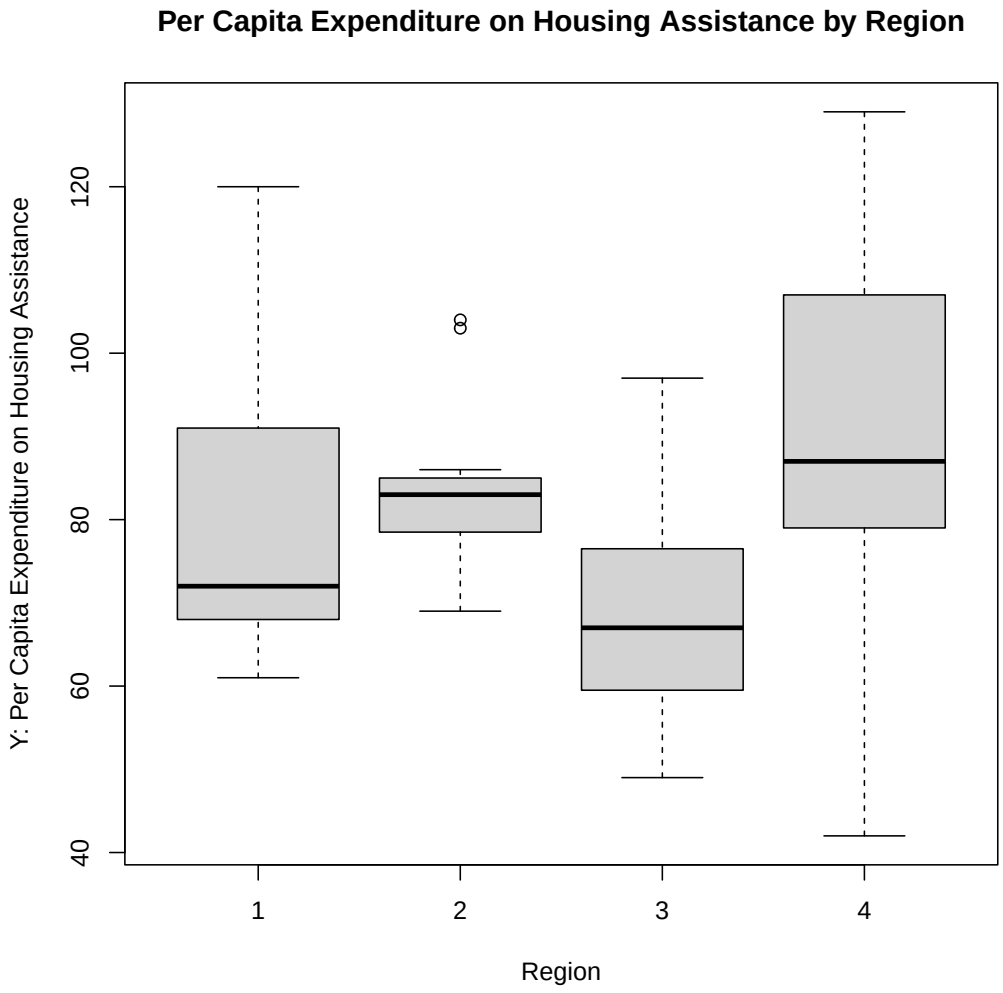

```
# Calculate correlation coefficients
cor(expenditure[,c("Y", "X1", "X2", "X3")])

#here are my conclusions
#Y and X1 show a strong positive correlation (r = 0.5317). The
    scatterplot clearly displays a positive linear relationship.
#Y and X2 have a moderate positive correlation (r = 0.4483). The
    scatterplot indicates a positive relationship, though not as strong as
     Y–X1.
#Y and X3 also exhibit a moderate positive correlation (r = 0.4637). The
    scatterplot shows a similar positive trend.
#X1 and X2 have a weak positive correlation (r = 0.2056). The scatterplot
     shows a more dispersed pattern with no clear linear trend.
#X1 and X3 demonstrate a strong positive correlation (r = 0.5925). This
```

```
   positive relationship is clearly visible in the scatterplot.
#X2 and X3 show a weak positive correlation (r = 0.2210). The scatterplot
    displays a somewhat random distribution of points.
```

- 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




```
######
# here is answer number 2

#Plot the relationship between Y and Region
#To visualize this relationship, we can use a boxplot
pdf("plot_boxplot.pdf")
```

```
boxplot(Y ~ Region, data = expenditure,
        main = "Per Capita Expenditure on Housing Assistance by Region",
        xlab = "Region",
        ylab = "Y: Per Capita Expenditure on Housing Assistance")
dev.off()

#here are my conclusions
#On average, Region 4 (West) has the highest per capita expenditure on
    housing assistance.
#This can be seen from the box plot, where Region 4
#has the highest median (represented by the thick black line in the
    middle of the box)
#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.

```
######
# here is answer number 3

#Plot the relationship between Y and X1
pdf("plot_X1&Y.pdf")
plot(expenditure$X1, expenditure$Y,
     xlab="X1: Per Capita Personal Income",
     ylab="Y: Per Capita Expenditure on Housing Assistance",
     main="Relationship between Income and Housing Assistance Expenditure
    ")
# Add regression line
abline(lm(Y ~ X1, data=expenditure), col="red")
dev.off()

# Calculate correlation coefficient
cor(expenditure$Y, expenditure$X1)

#install stargazer packages
install.packages("stargazer")
library(stargazer)

# Run and show regression analysis
regression1 <- lm(Y ~ X1, data=expenditure)

output_stargazer <- function(outputFile, ...) {
  output <- capture.output(stargazer(...))
  cat(paste(output, collapse = "\n"), "\n", file=outputFile, append=TRUE)
}
output_stargazer("regression_output1.tex", regression1)

#Plot the relationship between Y and X1 and region
#display different regions with different types of symbols and colors.
```

```r
pdf("plot_X1&Y&region.pdf")
plot(expenditure$X1, expenditure$Y,
     col = expenditure$Region,
     pch = as.numeric(expenditure$Region),  # Different symbol for each region
     xlab = "X1: Per Capita Personal Income",
     ylab = "Y: Per Capita Expenditure on Housing Assistance",
     main = "Relationship between Income and Housing Assistance Expenditure")
# Add legend
legend("topright",
       legend = c("Northeast", "North Central", "South", "West"),
       col = 1:4,
       pch = 1:4)
dev.off()

#here are my conclusions
#The relationship between Y (Per Capita Expenditure on Housing Assistance)
#and X1 (Per Capita Personal Income) shows a positive correlation.

#The scatter plot displays a clear upward trend, with the
#red regression line indicating a positive relationship. As personal income
#increases, there's a general tendency for housing assistance expenditure to increase as well.

#The correlation coefficient between X1 and Y is 0.5317212, indicating a moderate positive correlation.

#The regression coefficient for X1 is 0.025, which is statistically significant
#at the 1% level (p<0.01). This means that for every $1 increase in per capita
#personal income, we expect an average increase of $0.025 in per capita housing assistance expenditure.

#The R  value is 0.283, indicating that about 28.3% of the variation in
#housing assistance expenditure is explained by personal income.

#The F-statistic (18.920, p<0.01) suggests that the overall model is statistically significant.
```

Figure 3: The Plot Relationships Between Y and X1

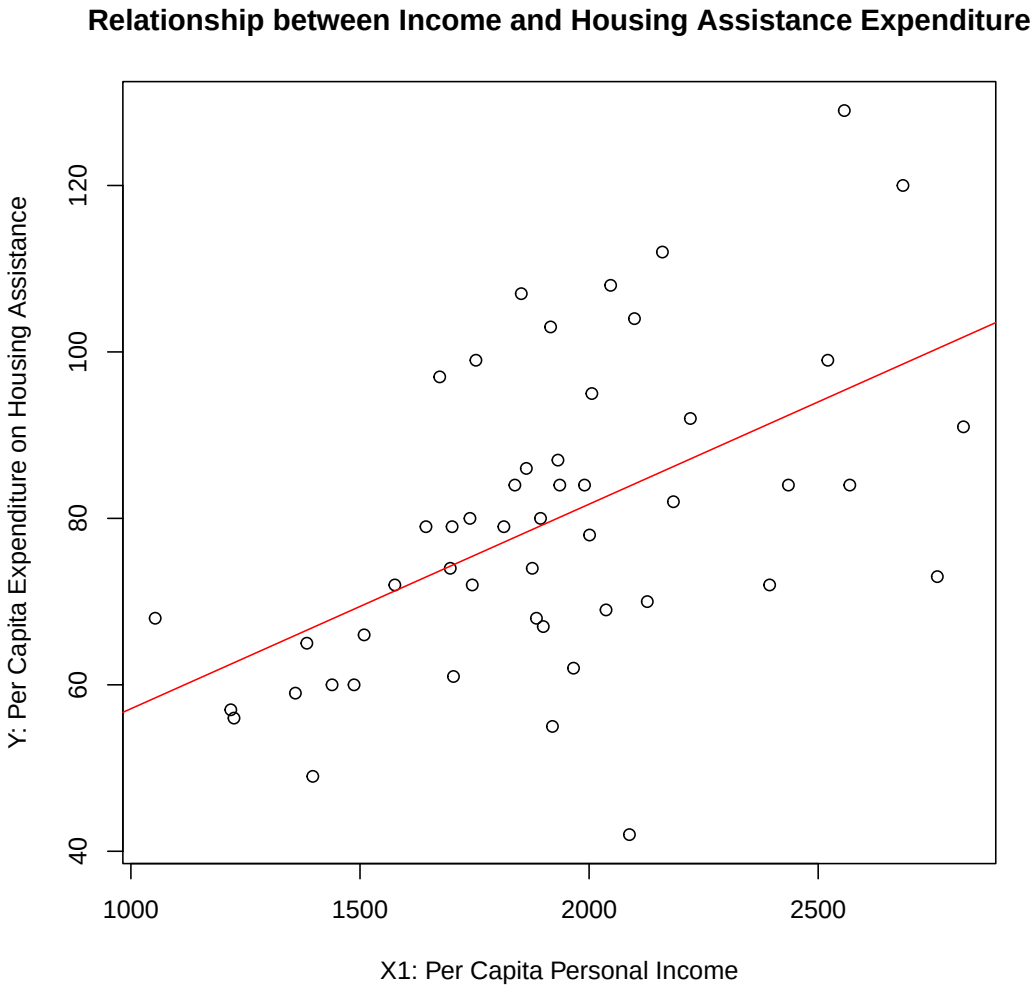


Table 1

	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

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