Problem Set 1

Applied Stats/Quant Methods 1
Clare Zureich

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.

```
90% Confidence Interval = [93.96, 102.92]

Step #1: Calculate y_bar
```

```r
mean_scores <-mean(scores)
```

```
Step #2: Calculate S and sigma_hat_y
```

```r
sd_scores <- sd(scores)
se_scores <- sd_scores/sqrt(length(scores))
```

```
Step #3: Calculate the curve area to the left and right
Area to the right: (1-.9)/2 = .05
Area to the left: (.9)/2 = .45
```

```
Step #4: Find the T-score associated with confidence level and DF
T-score = 1.711
```

```r
t90 <- qt((.9+(1-.9)/2), df=length(scores)-1)
```

```
Step #5: Calculate the confidence interval
```

```r
lower_90 <- mean_scores - t90*se_scores
upper_90 <- mean_scores + t90*se_scores
```

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 $\alpha = 0.05$.

```
Step #1: Assumptions of random, normal, and continuous quantitative data with a
sample size of 25
```

```r
sample_size = length(scores)
class(scores)
str(scores)
```

Step #2: Hypotheses
Null hypothesis: The average IQ scores of the counselor's students is less than or equal to 100, the national IQ score
Alternative hypothesis: The average IQ scores of the counselor's students is greater than 100, the national IQ score

Step #3: Calculate a test statistic
Test statistic = -.596

```
national_average = 100
test_statistic <- (mean_scores-national_average)/se_scores
```

Step #4: Calculate a p-value
p-value = .722

```
p_value <- (1 - pt(test_statistic , length(scores)-1))
```

Step #5: Conclusion
Based on the p-value of .72 and a .05 significance level, we fail to reject the null hypothesis that the average IQ score is less than or equal to 100.
There is insufficient evidence to conclude that the average IQ of the counselor's students is higher than the average of 100.

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`.

```r
expenditure <- read.table(”https://raw.githubusercontent.com/ASDS-TCD/StatsI_Fall2024/main/datasets/expenditure.txt”, header=T)
summary(expenditure)
names(expenditure)
str(expenditure)
```

- 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)?

  Graph 1 (Y vs X1): Housing expenditure (per capita) vs personal income (per capita) shows a moderate, positive, linear correlation. As personal income per capita increases, housing expenditure tends to increase.

  Graph 2 (Y vs X2): Housing expenditure (per capita) vs financially insecure residents (per 100,000) appears to have a nonlinear relationship (the shape resembles a parabola). There is weak to moderate correlation.

  Graph 3 (Y vs X3): Housing expenditure (per capita) vs urban area residents (per 1,000) shows a moderate, positive, linear correlation. As the number of urban area residents increases, housing expenditure tends to increase.

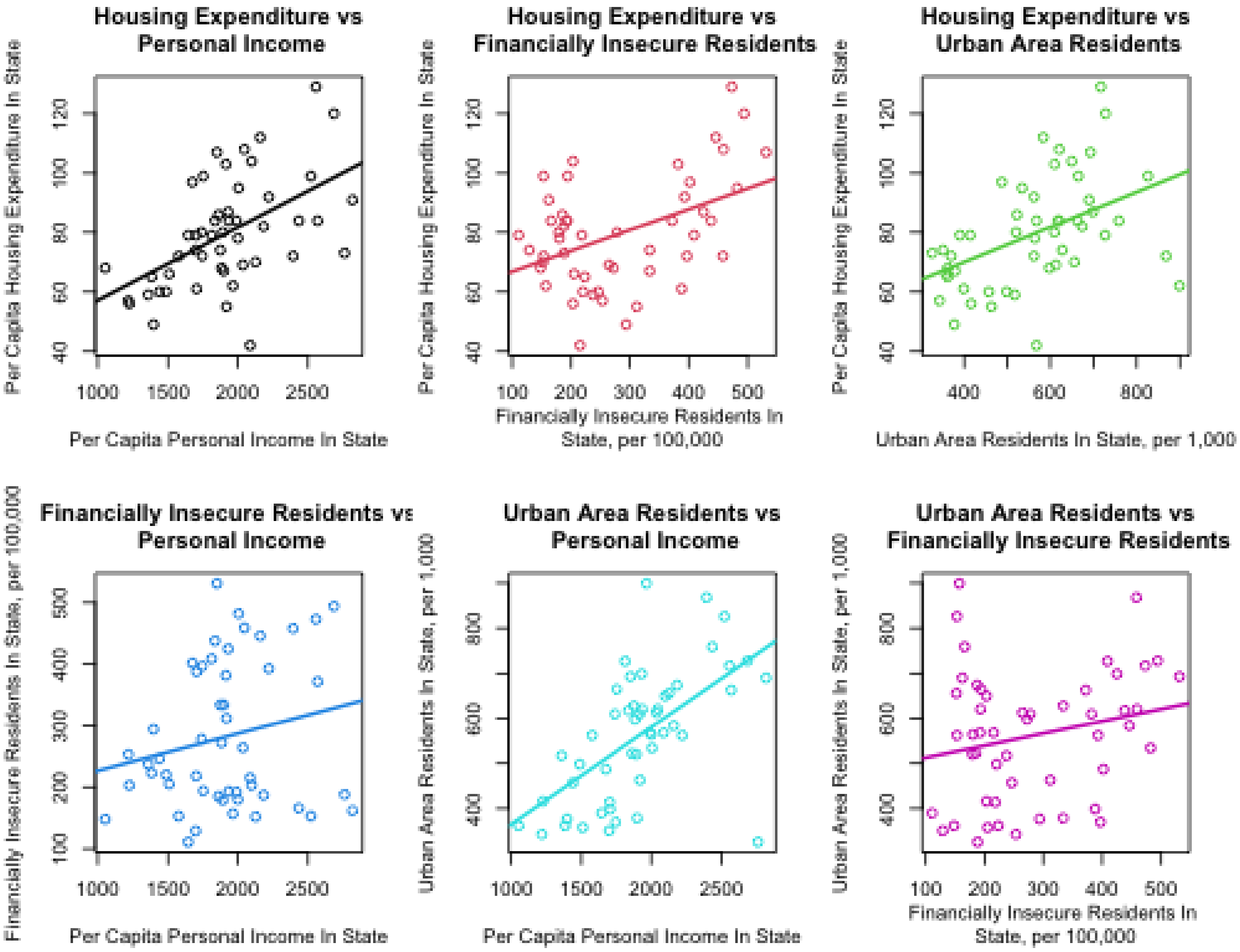
  Graph 4 (X2 vs X1): Financially insecure residents (per 100,000) vs personal income (per capita) appears to have a nonlinear relationship. There is weak, negligible correlation between the covariates.

  Graph 5 (X3 vs X1): Urban area residents (per 1,000) vs personal income (per capita) shows a moderate, positive, linear positive between the covariates.

As the per capita personal income increases, the number of urban area esidents tends to increase.

Graph 6 (X3 vs X2): Urban area residents (per 1,000) vs financially insecure residents appears to have a nonlinear relationship. There is weak, negligible correlation between the covariates.

Figure 1: Variable Relationships.



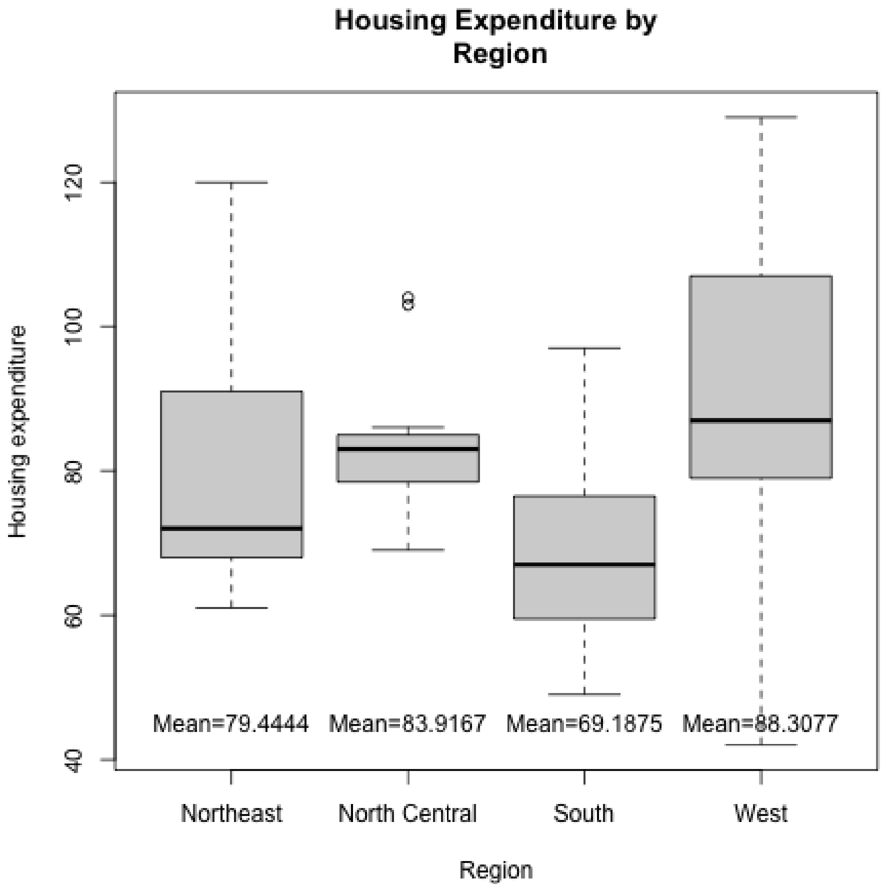
```
fit1 <- lm(Y ~ X1, data = expenditure)
fit2 <- lm(Y ~ X2, data = expenditure)
fit3 <- lm(Y ~ X3, data = expenditure)
fit4 <- lm(X2 ~ X1, data = expenditure)
fit5 <- lm(X3 ~ X1, data = expenditure)
fit6 <- lm(X3 ~ X2, data = expenditure)

#Correlations for all relationships
```

```r
png(file="political_economy_scatter_plot.png",
    width = 500,
    height = 400)
par(mfrow = c(2,3))
plot(expenditure$X1, expenditure$Y, col=1,
     ylab = "Per Capita Housing Expenditure In State",
     xlab = "Per Capita Personal Income In State",
     main = "Housing Expenditure vs\n Personal Income")
abline(fit1 , col = 1, lwd = 2)
plot(expenditure$X2, expenditure$Y, col=2,
     ylab = "Per Capita Housing Expenditure In State",
     xlab = "Financially Insecure Residents In \nState, per 100,000",
     main = "Housing Expenditure vs\n Financially Insecure Residents")
abline(fit2 , col = 2, lwd = 2)
plot(expenditure$X3, expenditure$Y, col=3,
     ylab = "Per Capita Housing Expenditure In State",
     xlab = "Urban Area Residents In State, per 1,000",
     main = "Housing Expenditure vs\n Urban Area Residents")
abline(fit3 , col = 3, lwd = 2)
plot(expenditure$X1, expenditure$X2, col=4,
     xlab = "Per Capita Personal Income In State",
     ylab = "Financially Insecure Residents In State, per 100,000",
     main = "Financially Insecure Residents vs\n Personal Income")
abline(fit4 , col = 4, lwd = 2)
plot(expenditure$X1, expenditure$X3, col=5,
     xlab = "Per Capita Personal Income In State",
     ylab = "Urban Area Residents In State, per 1,000",
     main = "Urban Area Residents vs\n Personal Income")
abline(fit5 , col = 5, lwd = 2)
plot(expenditure$X2, expenditure$X3, col=6,
     xlab = "Financially Insecure Residents In \nState, per 100,000",
     ylab = "Urban Area Residents In State, per 1,000",
     main = "Urban Area Residents vs\n Financially Insecure Residents")
abline(fit6 , col = 6, lwd = 2)
```

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

  `Region 4, the West, has the highest average of per capita expenditure on housing assistance`

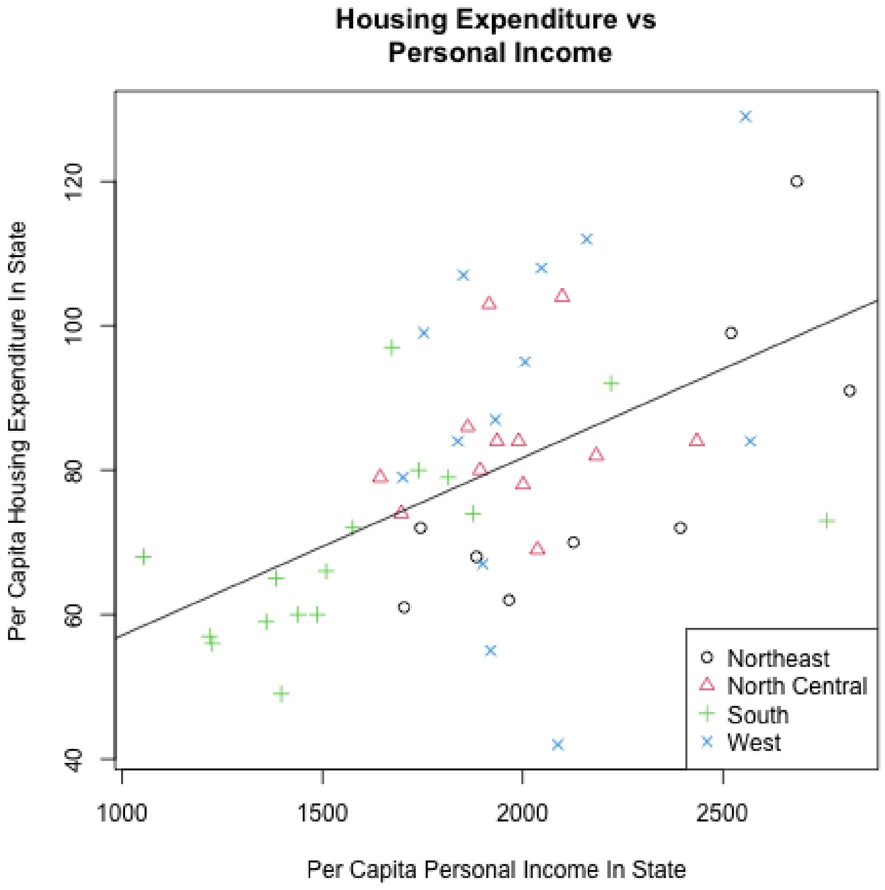
Figure 2: Housing Expenditure (Per Capita) by Region



```r
means1 <- mean(expenditure$Y[expenditure$Region == 1])
means2 <- mean(expenditure$Y[expenditure$Region == 2])
means3 <- mean(expenditure$Y[expenditure$Region == 3])
means4 <- mean(expenditure$Y[expenditure$Region == 4])

#Graph between Y and Region
png(file="housing_expenditure_region__scatter_plot.png")
par(mfrow = c(1,1))
region <- factor(expenditure$Region,
                    levels = c(1, 2, 3, 4),
                    labels = c("Northeast", "North Central", "South", "West"
   ))
expenditure$name <- region
plot(region, expenditure$Y,
      ylab = "Housing expenditure",
      xlab = "Region",
      main = "Housing Expenditure by\n Region")
text(1, 45, sprintf("Mean=%s", round(means1, 4)))
text(2, 45, sprintf("Mean=%s", round(means2, 4)))
text(3, 45, sprintf("Mean=%s", round(means3, 4)))
text(4, 45, sprintf("Mean=%s", round(means4, 4)))
```

- 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.

Figure 3: Figure 3.



The scatterplot of Y (Housing expenditure, per capita in state) vs X1 (Personal income, per capita in state), shows a moderate positive linear relationship. As per capita personal income (in state) increases, per capita housing expenditure (in state) tends to increase. Region 4, the West, has the most spread, and region 2 (North Central) has the least spread in terms of housing expenditure. The North Central region is clustered in the middle of the data, the South towards the lower end. The South (Region 3) region appears to have a strong positive relationship between housing expenditure and personal income (per capita). The Northeast region (region 1) also appears to have a stroing positive relationship between Y and X1. The North Central region (Region 2) and the West (Region 4) appear to have nonlinear relationships between per capita housing expenditure and per capita personal income, with weak to negligible correlation.

```r
png(file="housing_expenditure_personal_income_scatter_plot.png")
plot(expenditure$X1, expenditure$Y, col=expenditure$Region,
     ylab = "Per Capita Housing Expenditure In State",
     pch = expenditure$Region,
     xlab = "Per Capita Personal Income In State",
     main = "Housing Expenditure vs\n Personal Income")
legend("bottomright", legend=unique(expenditure$name),
       col=unique(expenditure$Region),
       pch=unique(expenditure$Region))
abline(fit1 , col = 1)
```