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. Find a 90% confidence interval for the average student IQ in the school.

   (a) Find the Mean

   ```r
   y_mean <- mean(y) # Point estimate
   ```

   (b) Find the Standard Error

   ```r
   y_err <- sd(y)/sqrt(length((y)))
   ```

   (c) Calculate the associated t-statistic using the degrees of freedom

```r
# calculate the degrees of freedom
deg_free <- length(y) - 1

# as n<30 the clt cannot be used
t_stat_1 <- qt(1 - 0.05, deg_free)
```

(d) Obtain the Upper and Lower Bounds

```r
# upper bound
upper_90 <- y_mean + t_stat_1*y_err
```

```r
# lower bound
lower_90 <- y_mean - t_stat_1*y_err
```

(e) This provides us with a confidence interval of:
[93.96 , 102.92] (3d.p)

2. Next, the school counsellor 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 $\alpha = 0.05$.

The 5 steps of Hypothesis Testing

(a) Assumptions
We have assumed that the data is approximately normally distributed.
We have also assumed that the IQ scores are independent of each other, as well as the sample being representative of the school.

(b) State H_0 and H_a

Our H_0 is that the average IQ in her school is $\leq$100.

Our H_a is that the average IQ in her school is > 100

(c) Calculate the T-Statistic

```r
# calculate the t-statistic, mu0 = 100
t_stat_2 <- (y_mean - 100) / y_err
# t = -0.5957439
```

(d) Calculate the P-Value

```r
p_val <- pt(t_stat_2, deg_free, lower.tail = FALSE)
# p = 0.7215383
```

(e) Conclusion

Given that the associated p-value is much larger than our significance level:

$$0.721 > 0.05$$

We can say that there is **not** sufficient evidence for us to reject the null hypothesis. This suggests that there is no evidence to support the students having a higher IQ than the average (100) among all the schools in the country.

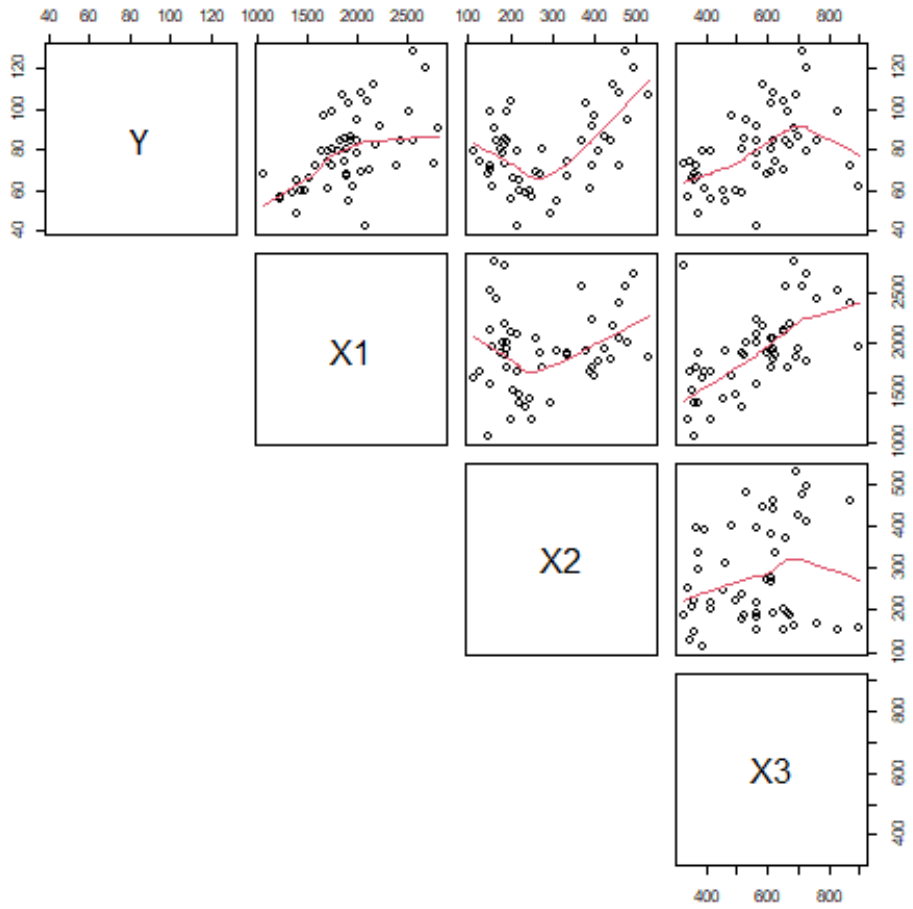
Question 2: Political Economy

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)?



X1,Y - We see a slight positive correlation.
This suggest there could be a positive relationship between state personal income per capita and state expenditure on housing assistance.

X2,Y - We see a non linear correlation.
This implies the existence of some relationship between the number of financially insecure residence and state expenditure on housing assistance.

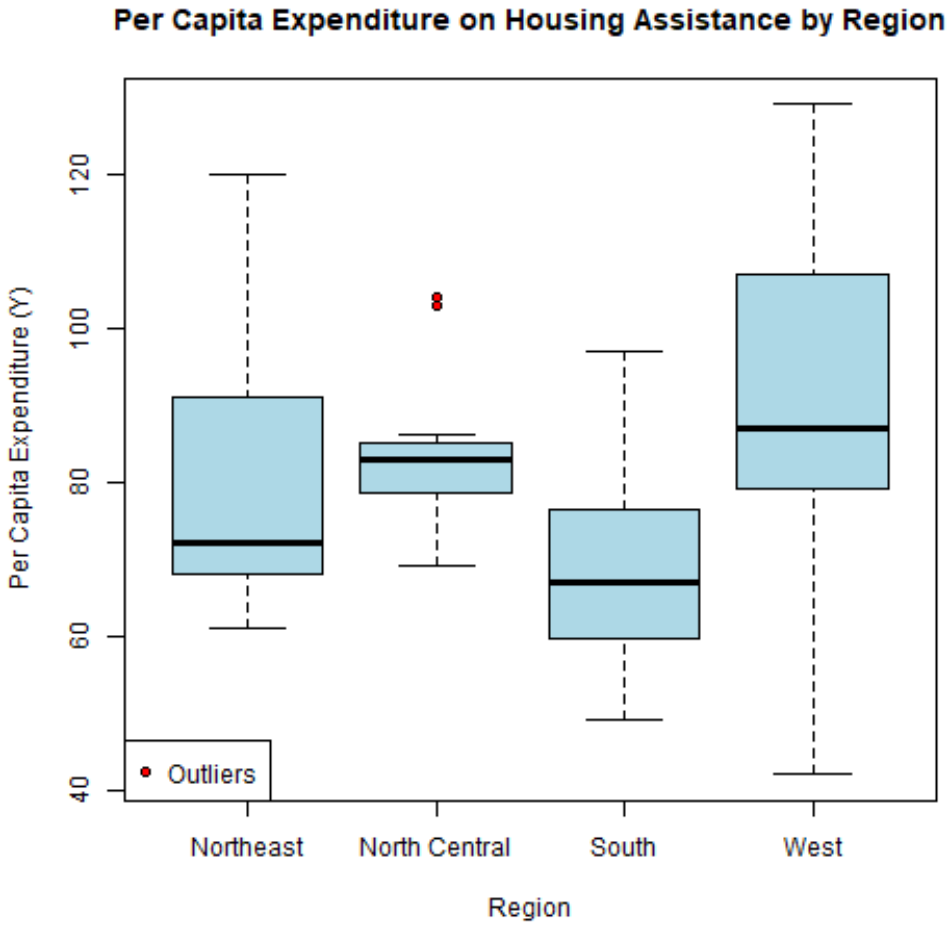
X3,Y - We observe no clear correlation.
This suggests no relationship between the number of residents in urban areas and state expenditure on housing assistance.

X2,X1 - We see a slight non linear correlation.
This implies the existence of some relationship between the number of financially insecure residence and state personal income per capita.

X3,X1 - We see a positive correlation.
This suggest there could be a positive relationship between the number of residents in urban areas and state personal income per capita.

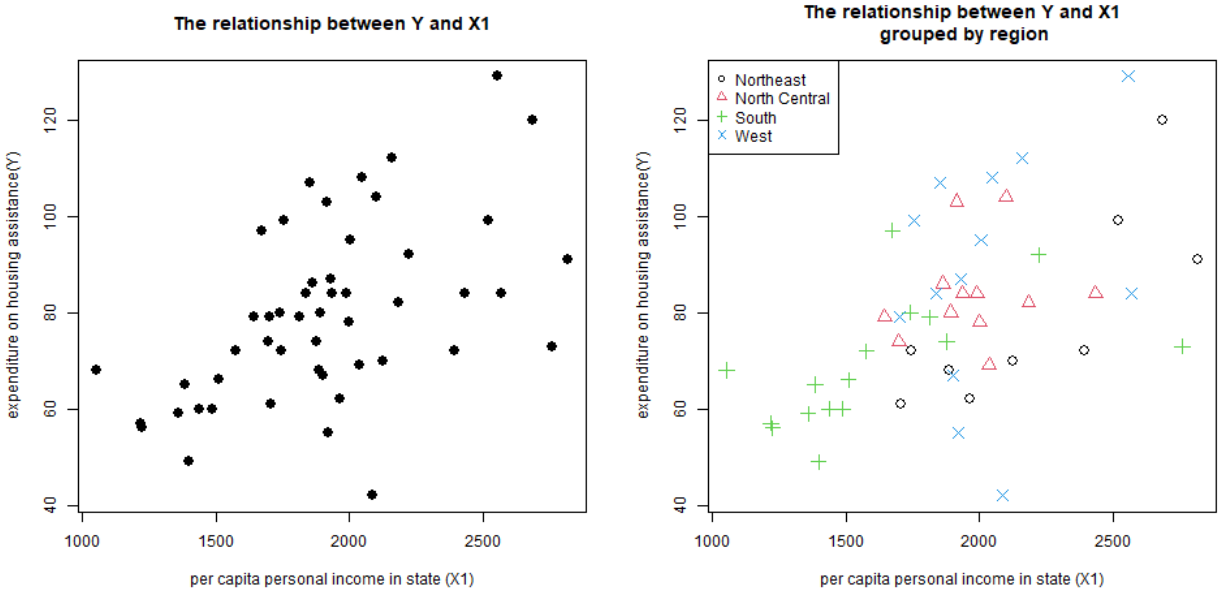
X3,X2 - We observe no clear correlation.
This suggests no relationship between the number of residents in urban areas and the number of financially insecure residence.

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



Looking at the box plot, we see on average (mean) the 'West' region has the highest expenditure on housing assistance per capita, the box plot shows also that it has the largest 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 colours.



Looking at the left hand graph we can see a general linear positive correlation between the two variables *(X1 and Y)*.

However, when we split up the variables by region we see some different patterns emerge. For example looking at the 'South' region we are able to identify the same positive correlation as before. But, when we observe the 'West' and 'North Central' regions this correlation is much less obvious.

Finally, upon observing the 'Northeast' region, the two variables appear to have a non-linear relationship. With Y climbing exponentially with minor increases *X1*