Problem Set 1

Applied Stats/Quant Methods 1

Due: October 3, 2021

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 in `.pdf` form.
- This problem set is due before 8:00 on Friday October 3, 2021. No late assignments will be accepted.
- Total available points for this homework is 100.

Question 1 (50 points): 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.

   First I calculate the average IQ for the sample

   ```
   AverageIQ <- mean(y)
   ```

   Then I calculate the standard deviation in the sample

```
StandardDeviation <- sd(y)
```

Next I calculate the margin of error using the standard deviation and sample size.

```
margin_error <- qt(0.95,df=24)*StandardDeviation/sqrt(25)
```

Finally, I calculate the upper and lower intervals

```
UpperInterval <- mean(y) + margin_error
LowerInterval <- mean(y) - margin_error
UpperInterval
LowerInterval
# Intervals [93.95993, 102.9201]
```

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 $\alpha = 0.05$.

I conduct a one sample t-test

```
t.test(y, mu = 100, alternative = "greater")
# p-value = 0.7215
```

Because p>0.05, we do not have evidence to reject the null-hypothesis, that the average student IQ in the school is the same or lower than the average IQ score (100) among all the schools in the country.

Alternatively, we can calculate this manually:

```
AverageIQ_t <- ((AverageIQ - 100)/(StandardDeviation/(sqrt(25))))
AverageIQ_p <- pt(AverageIQ_t, 24, lower.tail = FALSE)
AverageIQ_p
# AverageIQ_p = 0.7215383
```

Question 2 (50 points): Political Economy

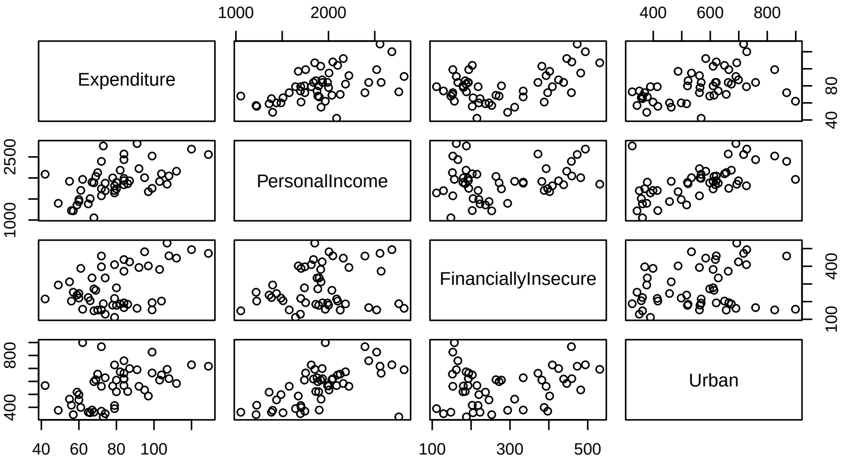
Researchers are curious about what affects the amount of money communities spend on addressing homelessness. The following variables constitute our data set about social welfare expenditures in the USA.

`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)?

-
  ```
  pairs(~ Expenditure + PersonalIncome + FinanciallyInsecure + Urban, data
      = Expenditure)
  ```



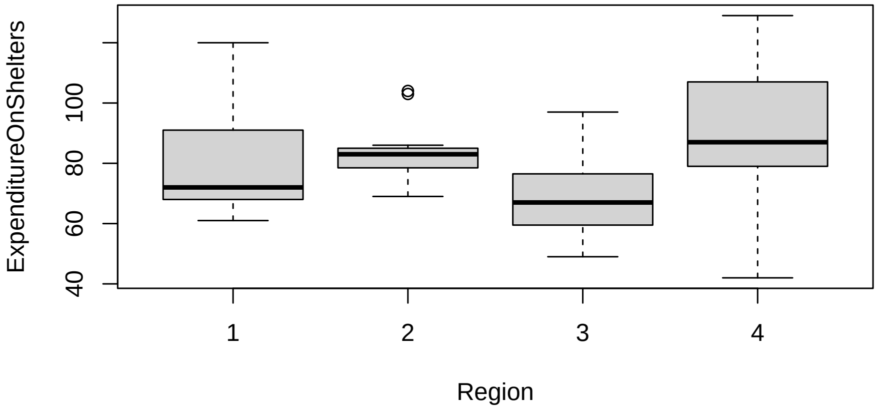
Each of the three independent variables is positively associated with Expenditure on Shelters/Housing Assistance. Personal income and

urban population appear to have a strong positive correlation.
On the other hand, Financial Insecurity doesn't appear to have a strong association with either Urban population or Personal Income.

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

- I produce a boxplot in order to compare the average expenditure on shelters/housing assistance across the 4 regions.

```r
boxplot(ExpenditureOnShelters ~ Region)
```

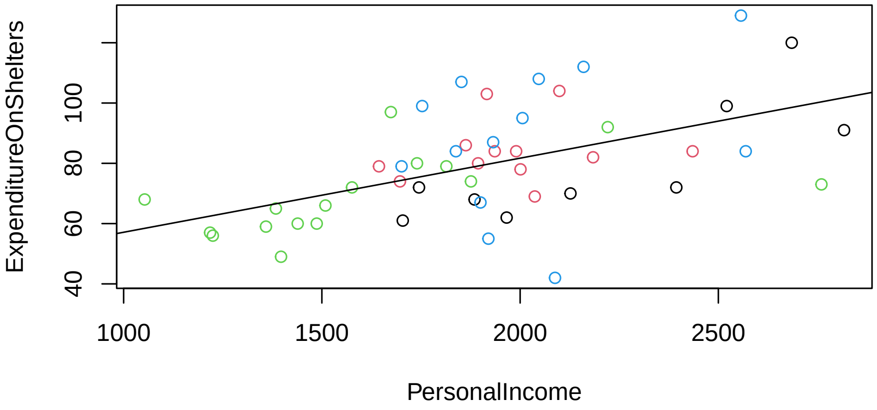


This box plot shows that expenditure on shelters/housing assistance is highest in the Western region of the US.

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.

I produce a scatterplot of Expenditure on Shelters vs Personal Income, with different regions in different colours:

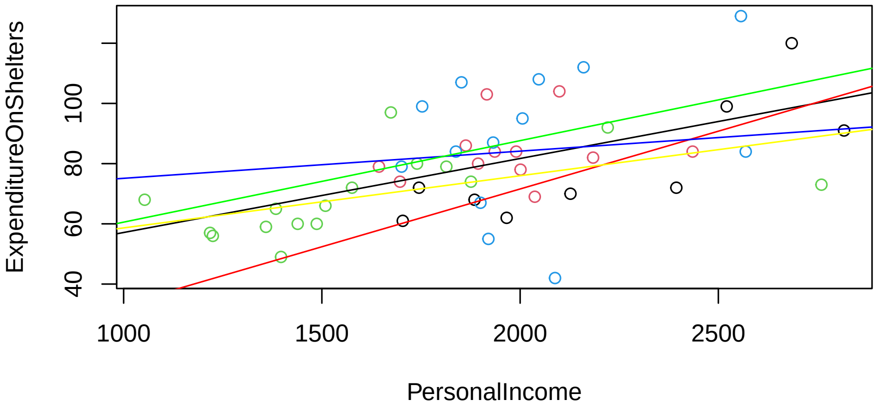
```r
plot(PersonalIncome, ExpenditureOnShelters, col=Region)
abline(lm(ExpenditureOnShelters ~ PersonalIncome))
```



to better understand the data and see how states in different regions differ, I add 4 lines of best fit, one for each of the 4 regions:

```
Northeast <- subset(Expenditure, Region == "1")
Northcentral <-subset(Expenditure, Region == "2")
South <-subset(Expenditure, Region == "3")
West <-subset(Expenditure, Region == "4")

abline(lm(Northeast$Expenditure ~ Northeast$PersonalIncome), col = "red")
abline(lm(Northcentral$Expenditure ~ Northcentral$PersonalIncome), col =
    "blue")
abline(lm(South$Expenditure ~ South$PersonalIncome), col = "yellow")
abline(lm(West$Expenditure ~ West$PersonalIncome), col = "green")
```



It appears from the data that the positive association between Personal Income and higher rates of expenditure on shelters and housing assistance is most significant in the Northeast, and is least significant in the Northcentral region, with the other two regions falling in betweeen. Northeast states also appear to have the highest Personal Incomes overall, whilst southern states have the lowest incomes.