Problem Set 1

Applied Stats/Quant Methods 1

Due: October 1, 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 1, 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:

```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.**

   The confidence interval is a range that a parameter is believed to fall within. In this question we are using a 90% confidence coefficient, which is how confident we are that the parameter will fall within the range.

   As the sample has less than 30 people we need to use a t distribution.

1. First we calculate the mean. The mean is the sum of the numbers divided by the number of numbers.

```
sample_mean <- mean(y , na.rm = TRUE)
```

2. We calculate the standard deviation of the sample. The standard deviation shows the variation in the data. This is calculated manually by dividing the sum of the squared deviations by the sample size minus one. R calculates this using the following code:

```
sample_sd <- sd(y, na.rm = TRUE)
```

3. Next we specify how much area under the curve we are looking at. At it it a 90 percent confidence interval we use:

```
(1-.90)/2)
```

4. We find the t value associated with this number.

```
t90 <- qt((1-.90)/2,df=24)
```

This gives us a t-statistic of -1.71

5. We then use the values above to calculate the range using the code:

```
n <- length(na.omit(y))
lower_90 <- sample_mean - (t90 * (sample_sd/sqrt(n)))
upper_90 <- sample_mean + (t90 * (sample_sd/sqrt(n)))

confint90 <- c(lower_90, upper_90)
print(confint90)
```

This prints out that the confidence interval is [93.96,102.92]

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 $\alpha = 0.05$.

   To conduct a hypothesis test we need to follow 5 steps:

1. First we need to acknowledge the assumptions about the data. The data is *quantitative*, *random sampling* and *normally distributed.* Because the number of people in this sample is 25, we need to use a t-distribution.

2. We set up the null and alternative hypotheses to enable proof by contradiction. The null hypothesis is that there is going to be no difference between our sample and the population mean. The alternative hypothesis is that the sample mean is greater than the population mean.

H0 : sample mean = the population mean

H1 : sample mean is higher than the population mean.

This is a one sided test as we are only interested in whether the school's mean IQ is higher than the population mean.

3. We then calculate a test statistic. This summarises how much the sample differs from what we expected if the null is true. As the sample group has n less than 30 we can use a t test to calculate the p-value.

```
t.test(y, mu = 100,
        alternative = "greater")
```

This gives the following output:

```
One Sample t-test

data:  y
t = -0.59574, df = 24, p-value = 0.7215
alternative hypothesis: true mean is greater than 100
95 percent confidence interval:
 93.95993      Inf
sample estimates:
mean of x
     98.44
```

The t-statistic is: *-.059574.* Degrees of Freedom are: *24.* Mean: *98.44.*

4. Next we calculate a p-value. The t.test shows that the p-value *0.7215* which is greater than our alpha value of *0.05.*

5. Draw a conclusion. This p-value suggests we fail to reject the null hypothesis. Which suggests the average IQ within this sample is not higher than the population mean.

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)?**

  As each of the X variables are measuring different values it is not possible to plot all three X variables against the Y. I have therefore plotted each of the variables against each other in scatterplots.

  This has shown that most of the relationships are weak, positive and linear. The associations below do not signify that one causes the other, but just describes the relationship between them.

  1. *Y* and *X1*

  ```
  plot(expenditure$Y, expenditure$X1)
  cor.test(expenditure$Y, expenditure$X1)
  ```

  This shows that the the relationship among per capita expenditure and per capita personal income is positive and linear with a correlation coefficient of .53. This suggests that an association between increased shelter spending and higher personal income in the state.

  2. *Y* and *X2*

  ```
  plot(expenditure$Y, expenditure$X2)
  cor.test(expenditure$Y, expenditure$X2)
  ```

This shows that the relationship among per capita expenditure and the number of residents that are financially insecure is positive and linear with a correlation coefficient of .45. This suggests that there is a weak positive association between higher spending in states with increased number of people that are financially insecure.

3. *Y* and *X3*

```
plot(expenditure$Y, expenditure$X3)
cor.test(expenditure$Y, expenditure$X3)
```

This shows that the the relationship among per capita expenditure and the number of people per thousand residing in urban areas is weak, positive and linear with a correlation coefficient of .46.

4. *X1* and *X2*

```
plot(expenditure$X1, expenditure$X2)
cor.test(expenditure$X1, expenditure$X2)
```

This shows that the the relationship among per capita income and residents in a state who are financially insecure is very weak, positive and linear with a correlation coefficient of .21.

5. *X1* and *X3*

```
plot(expenditure$X1, expenditure$X3)
cor.test(expenditure$X1, expenditure$X3)
```

This shows that the the relationship among per capita income and the number of people living in urban areas is positive and linear with a correlation coefficient of .59.

6. *X2* and *X3*

```
plot(expenditure$X2 expenditure$X3)
cor.test(expenditure$X2, expenditure$X3)
```

This shows that the the relationship among the number of people who are financially insecure and number of residents living in urban areas is very weak, positive and linear with a correlation coefficient of .22.

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

  I have plotted the relationship between *Y* and *Region* using the code below:

```
ggplot(expenditure, aes(x=Region, y=Y)) +
  geom_point(size=2, shape=23)
```

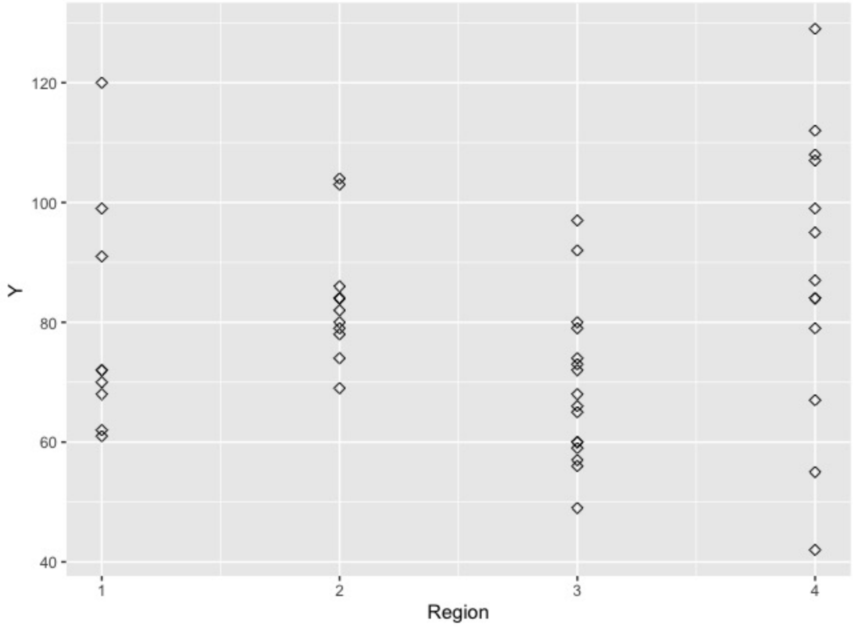

Figure 1: Plotting Y and Region

This shows visually that on average 4 (the west) has the highest per capita spending on housing assistance. To help backup this statement I used the following code to assess which *Region* has the highest mean spending per capita:

```
region_means <- aggregate(expenditure$Y, by = list(expenditure$Region), FUN = mean)

print(region_means)
```

- **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.**

Part One:

Below I have plotted the relationship between *Y* and *X1* using the code:

```
ggplot(expenditure, aes(x=X1, y=Y)) +
  geom_point(size=2, shape=23)
```

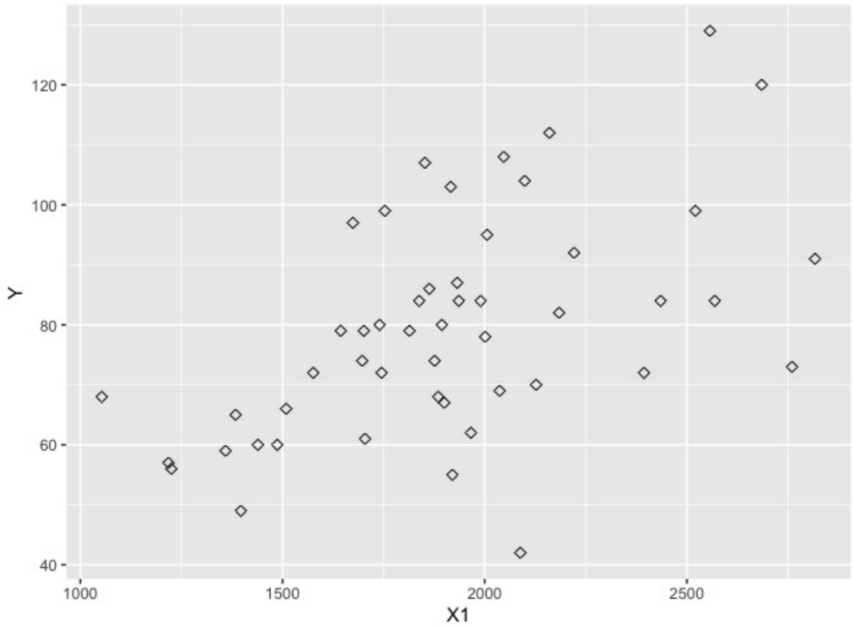

Figure 2: Plotting Y and X1

The below code adds in the line of best fit which helps describe the relationship:

```
ggplot(expenditure, aes(x=X1, y=Y)) +
  geom_point(size=2, shape=23)+
  geom_smooth(method=lm, se=FALSE, fullrange=TRUE)
```

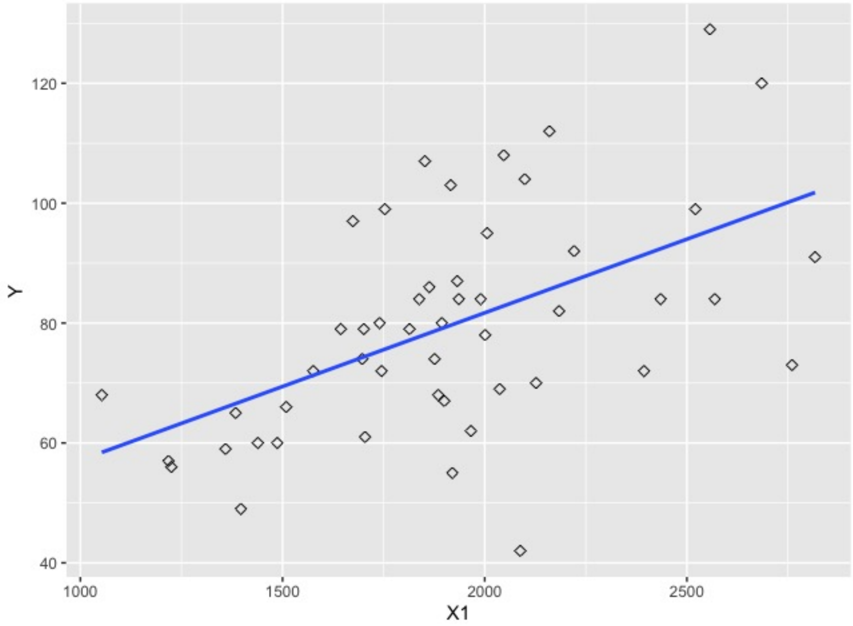

Figure 3: Plotting Y and X1, with line of best fit

We can see that it is a linear positive relationship. This describes that as *X1* (the per capita personal income) increases so does *Y* (the expenditure on shelters/housing assistance). This is just an association and not suggesting causation.

Part Two:

Below I have reproduced the graph to include *Region* using the code below:

```
ggplot(expenditure, aes(x=X1, y=Y, color=as.factor(Region), shape=as.factor(Region)
  geom_point() +
  scale_shape_manual(values=c(1, 2, 3, 4))+
  scale_color_manual(values=c(’red’,’orange’, ’seagreen1’, ’royalblue4’))+
  theme(legend.position="bottom")
```

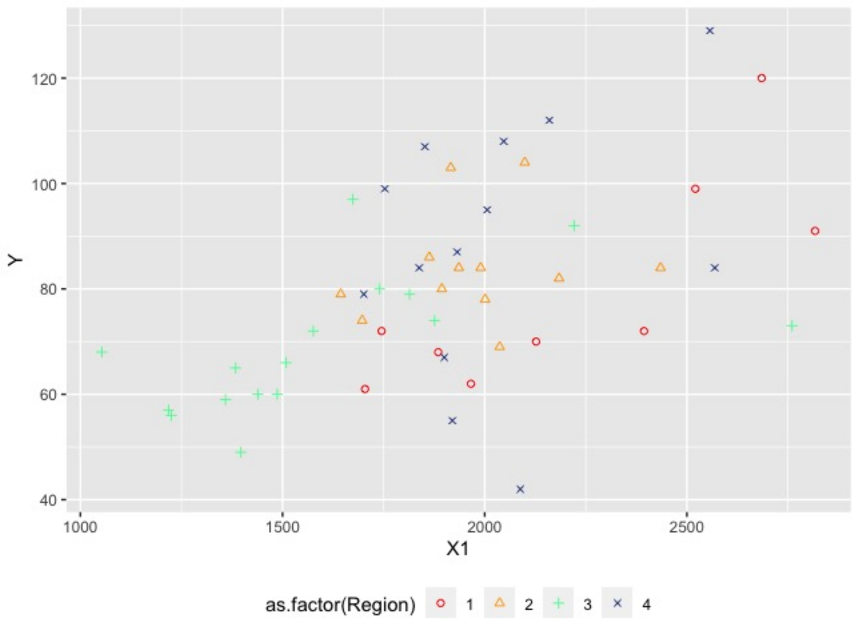

Figure 4: Plotting Y and X1, with variable Region

This shows that when including the variable *Region* the relationship in each region is still positive and linear. By adding the line of best fit code for each *Region* into the graph it makes reading the relationships more visually accessible.

```
ggplot(expenditure, aes(x=X1, y=Y, color=as.factor(Region), shape=as.factor(Region)
  geom_point() +
  scale_shape_manual(values=c(1, 2, 3, 4))+
  scale_color_manual(values=c('red','orange', 'seagreen1', 'royalblue4'))+
  theme(legend.position="bottom")+
  geom_smooth(method=lm, se=FALSE, fullrange=TRUE)
```

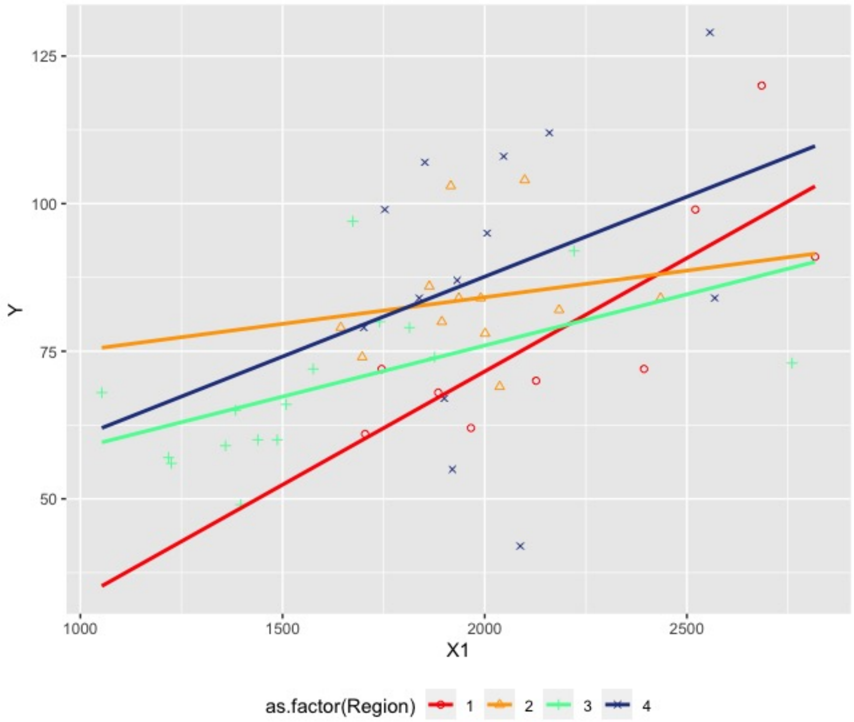

Figure 5: Plotting Y and X1, incl variable Region and line of best fit