Problem Set 1 Answers

Kaley Burg

Due: October 1, 2023

Instructions

This is my answer set

Question 1 (40 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.

   - **First I made variables for the mean, standard deviation, length, standard error, and degrees of freedom so that I could easily put them into the qt function**

     ```r
     #mean of sample
     y_mean <- mean(y)
     print(y_mean)

     #standard deviation of sample
     y_sd <- sd(y)
     print(y_sd)

     #length of sample
     n <- length(y)
     print(n)
     ```

```r

#standard error of sample
y_se <- y_sd/(sqrt(n))
print(y_se)
```

```r
#degrees of freedom for y
df_y <- n-1
print(df_y)
```

- **Mean = 98.44, Standard Deviation = 13.09, Length (n) = 25, Standard Error = 2.62**

  Note: I rounded these to put the results there but I used the non-rounded versions to compute everything

- **Next I used the qt function,using (1 -.90)/2 (or 0.05) because we want to divide it over both tails along with the degrees of freedom. We are looking in the upper tail because we are using the positive t score. The code using these values in the qt function looks like this. Note: I called the variable t90 but I understand that the exact value I found is actually the t score at 0.95**

```r
t90 <- qt((1 - .90)/2, df_y, lower.tail = FALSE)
```

- **I also showed how to do it this way too:**

```r
t90_alt <- qt(.95, df_y)
```

- **Now I just create the lower and upper bounds using the mean, t score, and standard error. Finally, I put these together to make a confidence interval**

```r
lower_90 <- y_mean - (t90 * (y_se))
upper_90 <- y_mean + (t90 * (y_se))
```

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

- **My confidence interval is [93.96, 102.92] (rounded to 2 decimal places)**
- **What this means is that 95% of the means of the sampling distribution of the school would fall within this range**

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 $\alpha = 0.05$.

   - **First, we look at our assumptions. We assume that the data is discrete because there are no non-integer scores. We also assume that the sample is random**

- Next, we set our hypotheses. Because it is asking if the average student IQ is *higher* than the average IQ score among all schools in the country, we are going to be using a one-tailed t test. Our hypotheses are as follows:
  Null Hypothesis: $\mu \leq 100$
  Alternative Hypothesis: $\mu > 100$
- For our third step, we calculate the test statistic. I used the formula from page 150 in the textbook:

$$t = \frac{\bar{y}-\mu_0}{se}$$

- Here I do this using R code: I find that the t statistic is -0.5957 (rounded to 4 decimal places)

```r
y_t <- ((y_mean - 100)/y_se)
```

- For the 4th step, we calculate the p value for our test statistic step 4 calculate p value for this t score, we are looking in the upper tail because our null hypothesis is that the mean IQ at the school is HIGHER than the population mean of 100. Here is my code for that: The p value I got was 0.7215 (rounded to 4 decimal places)

```r
p_y <- pt(y_t, df_y, lower.tail = F)
```

- For step 5, we draw a conclusion. In this case, we fail to reject the null hypothesis because our p value is 0.7215 (this means that the probability of this occurring by chance was 72.15%) while and our alpha value was 0.05. Therefore, we fail to reject the null hypothesis
- I also found the critical value of t needed to reject the null in this example just to check my work. I found that the critical t value is 1.71 here is the code to find this:

```r
t95 <- qt((1 - .95), df_y, lower.tail = F)
```

- I also checked my work with the t.test function in r and got the same results. My code for that is here:

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

Question 2 (40 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.

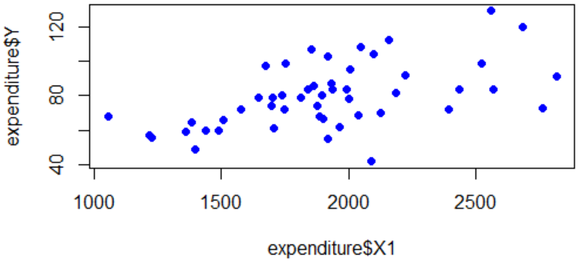
`State`	*50 states in US*
`Y`	*per capita expenditure on shelters/housing assistance in state*
`X1`	*per capita personal income in state*
`X2`	*Number of residents per 100,000 that are "financially insecure" in state*
`X3`	*Number of people per thousand residing in urban areas in state*
`Region`	*1=Northeast, 2= North Central, 3= South, 4=West*

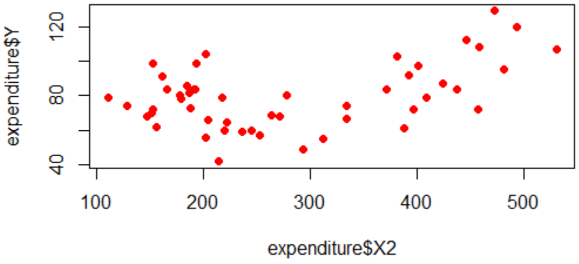
- Here is my code for exploring the data set and importing the data into r

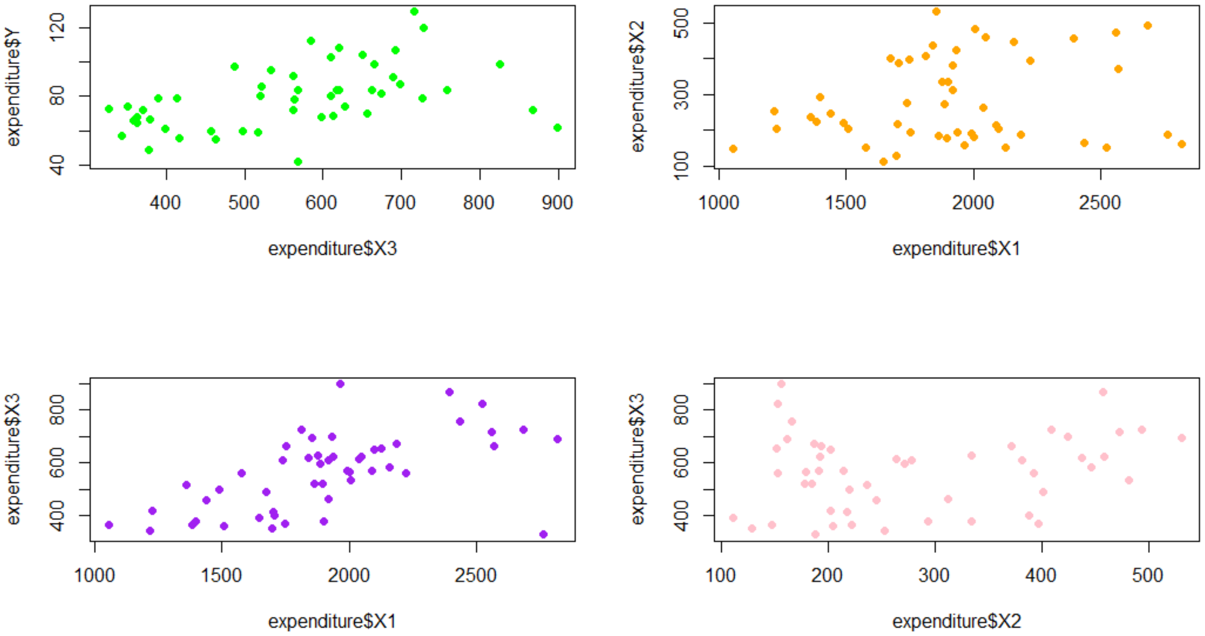
```r
expenditure <- read.table("https://raw.githubusercontent.com/ASDS-TCD/StatsI_Fall2023/main/datasets/expenditure.txt", header=T)
str(expenditure)
head(expenditure)
```

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)?
  - **Here are my 6 plots for each of the two variables:**

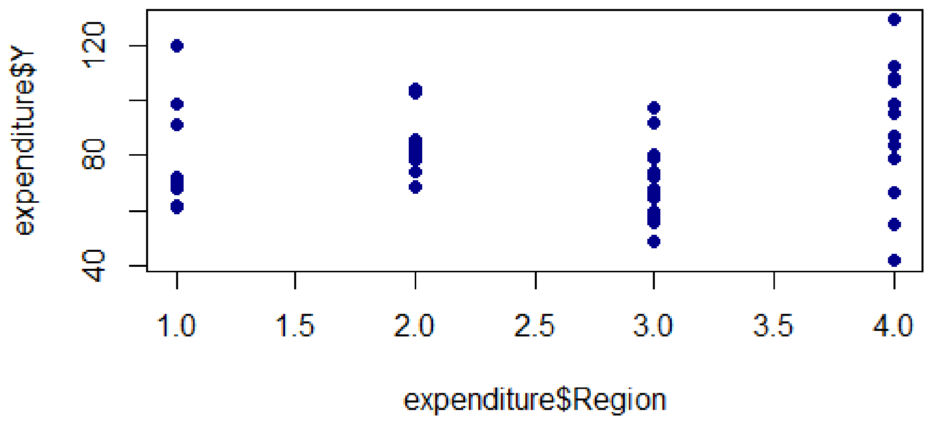




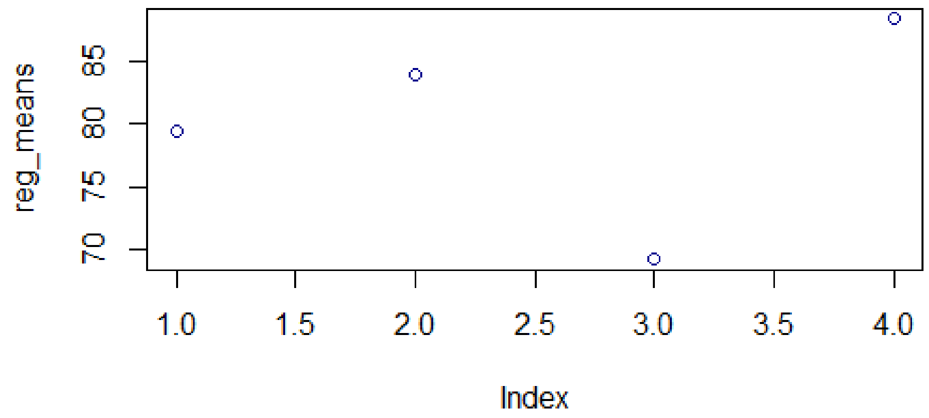


  - **For the plot of X1 and Y we see a positive correlation. It looks somewhere in the middle of strong and weak**
  - **For the plot of X2 and Y, the graph is somewhat u-shaped, so neither positive or negative**
  - **For the plot of X3 and Y, the graph is positive and looks weak**
  - **For the plot of X1 and X2, the graph is positive and looks weak**
  - **For the plot of X1 and X3, the graph is positive and looks somewhere between weak and strong**
  - **For the plot of X1 and X2, the graph is positive and looks weak**

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



- **I found which region has the highest capita expenditure on housing assistance by making 4 new variables, one for each region. Then I made 4 more variables computing the mean of each region. From this I could see that Region 4 (West) had the highest per capita expenditure on housing assistance of 88.31 (rounded to 2 decimal places)**
- **Then I put these means into one variable and I plotted it. Results below:**

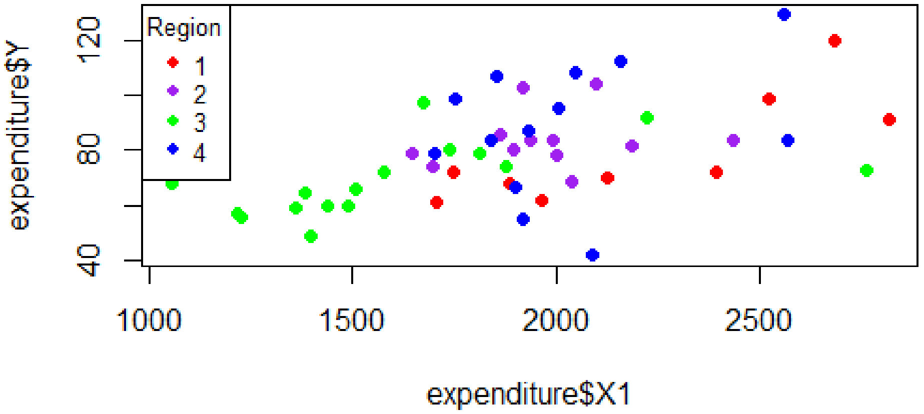


- Please plot the relationship between *Y* and *X1*? Describe this graph and the relationship. Reproduce the above graph including one more variable *Region* and display different regions with different types of symbols and colors.

- **First I used base r plot and I used code from** this website **and my code is as follows:**

```r
{plot(expenditure$X1, expenditure$Y, pch=16,
      col=c("red", "purple", "green", "blue")[expenditure$Region])
legend("topleft", pch=16, col=c("red", "purple", "green", "blue"),
        c("1", "2", "3", "4"), cex=0.8,
        title="Region")}
```

- **My plot for this is below**



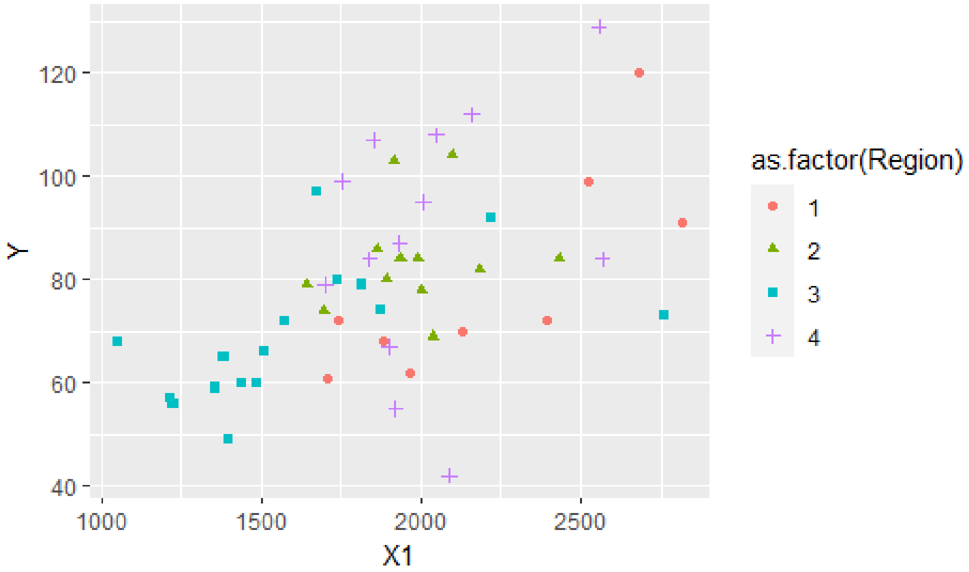
- **But, this only allowed me to change the colors for each region when I wanted to change the shape also. I tried to find a response for this on google and had no luck, so instead I used ggplot. To do so, I first loaded tidyverse:**

```r
library(tidyverse)
```

- **Then, I used code from coding camp but also used some code from** here **to get this code below:**

```r
expenditure %>%
  filter(Region %in% c("1", "2", "3", "4")) %>%
  group_by(Region) %>%
  ggplot(aes(X1, Y, color = as.factor(Region), shape = as.factor(
    Region))) +
  geom_point()
  theme_classic() +
  labs(title = "test")
```

- **The result is here:**


Y
X1
as.factor(Region)