Problem Set 1

Applied Stats/Quant Methods 1

Due: October 1, 2023

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 Sunday October 1, 2023. No late assignments will be accepted.
- Total available points for this homework is 80.

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:

```
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.

```
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)

confidence_level <- 0.90
```

```r

y_size <- length(y)
y_mean <- mean(y)
y_sd <- sd(y)

degrees_of_freedom <- length(y) - 1
t_critical <- qt((1 - confidence_level) / 2, df = degrees_of_freedom)
lower_Confidence <- y_mean - (t_critical * (y_sd / sqrt(y_size)) )
upper_Confidence <- y_mean + (t_critical * (y_sd / sqrt(y_size)) )

cat("Confidence Interval : [", lower_Confidence, ", " , upper_Confidence,
    "]\n")
```

Confidence Interval (90%): [102.9201 , 93.95993]

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 $\alpha = 0.05$.

```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)
alpha <- 0.05
mu<-100
t_statistic <- (mean(y) - mu) / (sd(y) / sqrt(length(y)))
p_value <- 2 * pt(abs(t_statistic), degrees_of_freedom, lower.tail = FALSE)
t_test_IQ <- t.test(y, mu = 100)
```

p value: 0.557, which is greater than alpha= 0.05; so no enough evidence to reject the null hypothesis; In other words no sufficient evidence to prove the IQ score of random 25 students in the school was higher than the average IQ score among all the schools in the country.

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*

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)?

```
install.packages("tidyverse")
library(tidyverse)
expenditure <- read.table("https://raw.githubusercontent.com/ASDS-TCD/StatsI_Fall2023/main/datasets/expenditure.txt", header=T)
relationship_expenditure <- expenditure[, c("Y", "X1", "X2", "X3")]
par(mfrow = c(1, 3))
plot(relationship_expenditure$X1, relationship_expenditure$Y, main = "Y vs. X1", xlab = "X1", ylab = "Y", col = "yellow")
lm_model <- lm(Y ~ X1, data = relationship_expenditure)
abline(lm_model, col = "blue")
```

plot1 : scatterplot with linner regression *Y,X1*

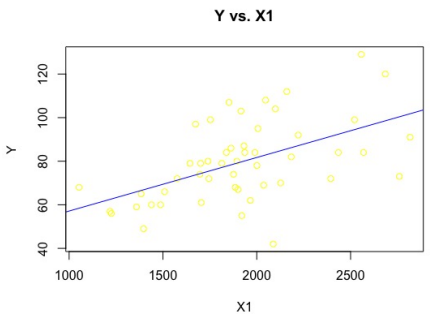

Figure 1: Y vs X1

```r
plot(relationship_expenditure$X2, relationship_expenditure$Y, main = "Y vs. X2", xlab = "X2", ylab = "Y", col = "pink")
lm_model<- lm(Y~X2, data = relationship_expenditure)
abline(lm_model,col="black")
```

plot2 : scatterplot with linner regression *Y,X2*

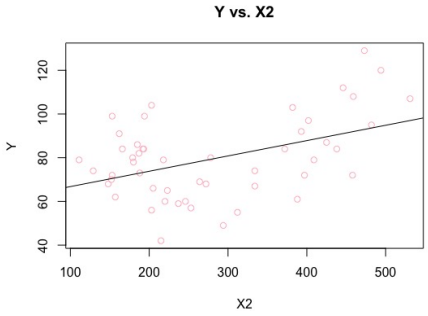

Figure 2: Y vs X2

```r
plot(relationship_expenditure$X3, relationship_expenditure$Y, main = "Y vs. X3", xlab = "X3", ylab = "Y", col = "blue")
lm_model<- lm(Y~X3, data = relationship_expenditure)
abline(lm_model,col="black")
```

plot3 : scatterplot with linner regression *Y,X3*

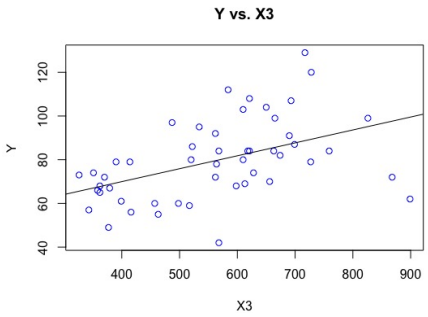

Figure 3: Y vs X3

```
plot(relationship_expenditure$X1, relationship_expenditure$X2, main = "X1
    vs. X2", xlab = "X1", ylab = "X2", col = "blue")
lm_model<- lm(X1~X2, data = relationship_expenditure)
abline(lm_model,col="black")
```

plot4 : scatterplot with linner regression *X1*,*X2*

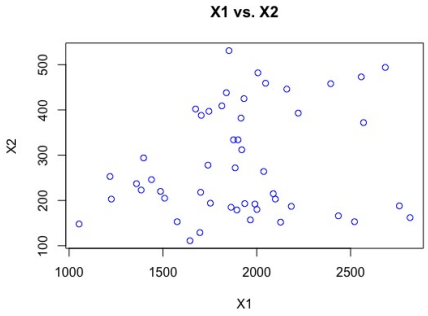

Figure 4: X1 vs X2

```
plot(relationship_expenditure$X1, relationship_expenditure$X3, main = "X1
    vs. X3", xlab = "X1", ylab = "X3", col = "green")
lm_model<- lm(X1~X3, data = relationship_expenditure)
abline(lm_model,col="black")
```

plot5 : scatterplot with linner regression *X1*,*X3*

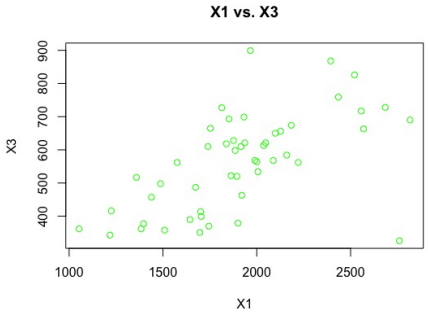

Figure 5: X1 vs X3

```r
plot(relationship_expenditure$X2, relationship_expenditure$X3, main = "X2 vs. X3", xlab = "X2", ylab = "X3", col = "orange")
lm_model<- lm(X2~X3, data = relationship_expenditure)
abline(lm_model,col="black")
```

plot6 : scatterplot with linner regression *X2,X3*

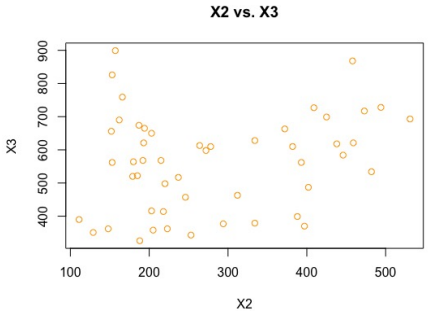

Figure 6: X2 vs X3

```r
pairs(relationship_expenditure[, c("Y", "X1", "X2", "X3")], col = c("blue", "red", "green", "purple"))
```

plot7 : Matrix scatterplots *Y,X1* to*X3*

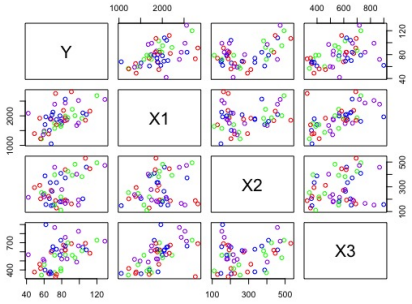

Figure 7: Y vs X1-X3

```r
ggplot(expenditure, aes(x = X1)) +
  geom_line(aes(y = Y, color = "Y vs. X1")) +
  geom_line(aes(x = X2, y = Y, color = "Y vs. X2")) +
  geom_line(aes(x = X3, y = Y, color = "Y vs. X3")) +
  labs(title = "plot of Y, X1, X2, and X3", x = "X1_X2_X3", y = "Y_
    expenditures") +
  theme_minimal()
```

plot8 : line plot *Y*,*X1* to*X3*

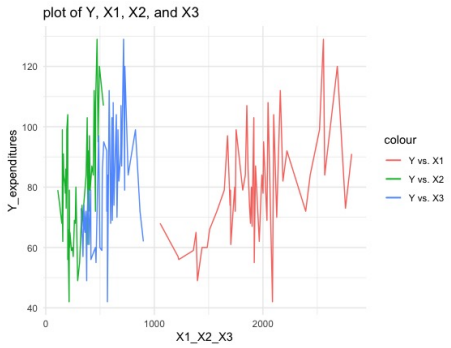

Figure 8: Y vs X1-X3

Description:

we can see
plot1: Y and X1 are two variables with positive regression line
plot2: Y and X2 are two variables with positive regression line
plot3: Y and X3 are two variables with positive regression line
plot4: X1 and X2 are two variables, with Positive coefficients
plot5: X1 and X3 are two variables, with Positive coefficients
plot6: X2 and X3 are two variables , with Positive coefficients

From plot7 and plot 8: independent variable X1: per capital personal income make more obvious positive effect on dependant variable Y: per capital expenditure on shelters.

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

```
install.packages("ggplot2")
library("ggplot2")

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

expenditure$Region <- as.factor(expenditure$Region)
library("ggplot2")

ggplot(expenditure, aes(x=Y, fill=Region))+ geom_histogram(binwidth = 6, position = "dodge")+labs(title = "histogram Y by region", x="expenditures on shelter", y="regions") + theme_minimal()

ggplot(expenditure,aes(x=factor(Region), y=Y, fill=factor(Region)))+ geom_bar(stat = "identity")+labs(title = "bar chart Y by region", X="region", y="expenditures")+ theme_minimal()
```

bar chart : bar chart Y,*Region* histogram: histogram Y,*Region*

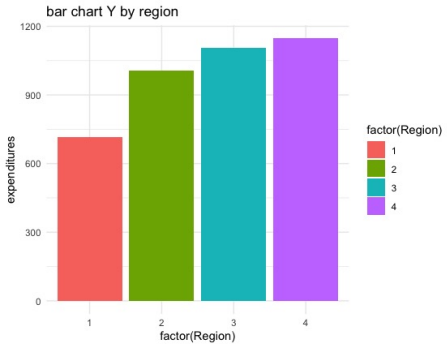

Figure 9: Y vs region

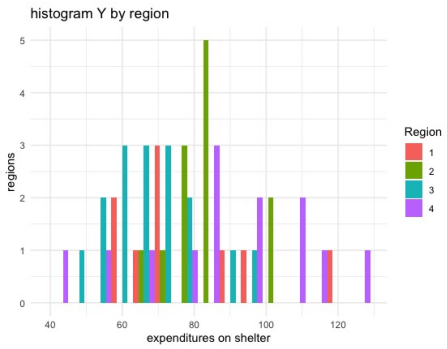

Figure 10: Y vs region

Description:

although region 4(west) has the highest per capita expenditures on shelters on the bar chart, region 2(North central) has the most times of expenditure between 75-95

- 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.

```
ggplot(expenditure, aes(x=X1, y=Y))+ geom_line()+labs(title = "line Y by X1",
    x="person income", y="expenditures on shelter")+theme_minimal()

ggplot(expenditure,aes(x=X1, y=Y, color= Region))+geom_line()+geom_point(aes(
    shape= Region), size=6) + labs(title = "line Y by X1&Region", x="person
    income", y="expenditures on shelter") + theme_minimal()+scale_shape_manual
    (values = c("1"=0, "2"=1,"3"=2,"4"=5))
```

line chart : line chart *Y*,*X1*

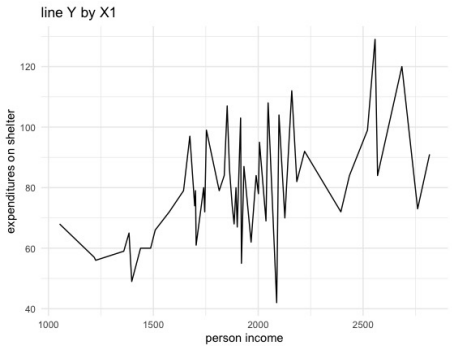

Figure 11: Y vs X1

line chart : line chart *Y,X1 with Region*

Figure 12: Y vs X1 with Region

Table 1:

	Dependent variable:
	Y
X1	0.025***
	(0.006)
Constant	32.546***
	(11.034)
Observations	50
R^2	0.283
Adjusted R^2	0.268
Residual Std. Error	15.836 (df = 48)
F Statistic	18.920*** (df = 1; 48)
Note:	*p<0.1; **p<0.05; ***p<0.01

Description:

Previously, we found that there is positive relationship between Y and X1; From the second line chart, X1(personal income) with different regions make diverse impact on expenditures on shelters; people live in region 4 who make income around 2500 willing to make the highest contribution on house assistance.