Problem Set 1

QTM 200: Applied Regression Analysis

Due: January 29, 2020

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 the course GitHub page in `.pdf` form.
- This problem set is due at the beginning of class on Wednesday, January 22, 2020. No late assignments will be accepted.
- Total available points for this homework is 100.

Question 1 (25 points)

A private 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)
```

Find a 90% confidence interval for the student IQ in the school assuming the population of IQ from which our random sample has been selected is normally distributed.

Solution

`R` code:

```
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)
z90 <- qnorm((1 - .90)/2, lower.tail = FALSE)
```

```r
n <- length(y)
sample_mean <- mean(y)
sample_sd <- sd(y)
lower_90 <- sample_mean - (z90 * (sample_sd/sqrt(n)))
upper_90 <- sample_mean + (z90 * (sample_sd/sqrt(n)))
confint90 <- c(lower_90, upper_90)
confint90
```

Results: [94.13283, 102.74717]

Question 2 (25 points)

A private school counselor was curious whether the average of IQ of the students in her school is higher than the average IQ score 100 among all the schools in the country. She 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)
```

Conduct a test with 0.05 significance level assuming the population of IQ from which our random sample has been selected is normally distributed.

Solution

R code:

```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)
t.test(y, mu = 100,
        alternative = "greater")
# OR
mu <- 100
n <- length(y)
sample_mean <- mean(y)
sample_sd <- sd(y)
teststatistic <- (sample_mean-mu)/(sample_sd/sqrt(n))
teststatistic
pt(abs(teststatistic), df = n - 1)
```

Results

Due to the p-value of 0.7215 being larger than the required significance level of 0.05, we fail to reject the null hypothesis and cannot accept the alternative hypothesis. The evidence does not supports the hypothesis that the average IQ students in the private school is higher than the average IQ score 100 among all the schools in the country.

Question 3 (50 points)

Researchers are curious about what affects the education expenditure on public education. The following is available variables in a data set about the education expenditure.

Solution to converting y from numbers to characters

R code:

```
y <- c(1, 2, 1, 3, 4, 1, 1, 4, 2, 1, 3, 4, 3, 2, 1, 3, 4, 1, 2, 3, 1, 1, 2, 1,
    1, 3, 4)
y.characters <- c("Freshman", "Sophomore", "Junior", "Senior")[y]
y.characters
```

State	*50 states in US*
Y	*per capita expenditure on public education*
X1	*per capita personal income*
X2	*Number of residents per thousand under 18 years of age*
X3	*Number of people per thousand residing in urban areas*
Region	*1=Northeast, 2= North Central, 3= South, 4=West*

Explore the `expenditure` data set and import data into R.

```
expenditure <- read.table("expenditure.txt", header=T)
```

- 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)?

R code:

```
expenditure <- read.table("expenditure.txt", header=T)
head(expenditure,6)
cor(expenditure$Y, expenditure$X1)
exp_data <- expenditure[,2:length(expenditure)]
round(cor(exp_data),2)

plot(expenditure$X1, expenditure$Y)
plot(expenditure$X2, expenditure$Y)
plot(expenditure$X3, expenditure$Y)

plot(expenditure$X1, expenditure$X2)
plot(expenditure$X1, expenditure$X3)

plot(expenditure$X2, expenditure$X3)
```

After plotting the relationships and building a correlation matrix among *Y*, *X1*, *X2*, and *X3*. I've come up with these observations.

- Between Y and $X1$:

From	*linear association*
Direction	*positive association*
Strength	*moderately strong*
Outliers	*present*
Correlation	*0.65*

- Between Y and $X2$:

From	*linear association*
Direction	*negative association*
Strength	*weak*
Outliers	*present*
Correlation	*-0.21*

- Between Y and $X3$:

From	*linear association*
Direction	*positive association*
Strength	*weak*
Outliers	*present*
Correlation	*0.25*

- Between $X1$ and $X2$:

From	*linear association*
Direction	*negative association*
Strength	*moderately strong*
Outliers	*present*
Correlation	*-0.53*

- Between $X1$ and $X3$:

From	*linear association*
Direction	*positive association*
Strength	*moderately strong*
Outliers	*present*
Correlation	*0.60*

- Between $X2$ and $X3$:

From	*linear association*
Direction	*negative association*
Strength	*weak*
Outliers	*present*
Correlation	*-0.37*

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

```r
plot(expenditure$Region, expenditure$Y)

region1 <- filter(expenditure, expenditure$Region == 1)
mean(region1$Y)

region2 <- filter(expenditure, expenditure$Region == 2)
```

```r
mean(region2$Y)

region3 <- filter(expenditure, expenditure$Region == 3)
mean(region3$Y)

region4 <- filter(expenditure, expenditure$Region == 4)
mean(region4$Y)
```

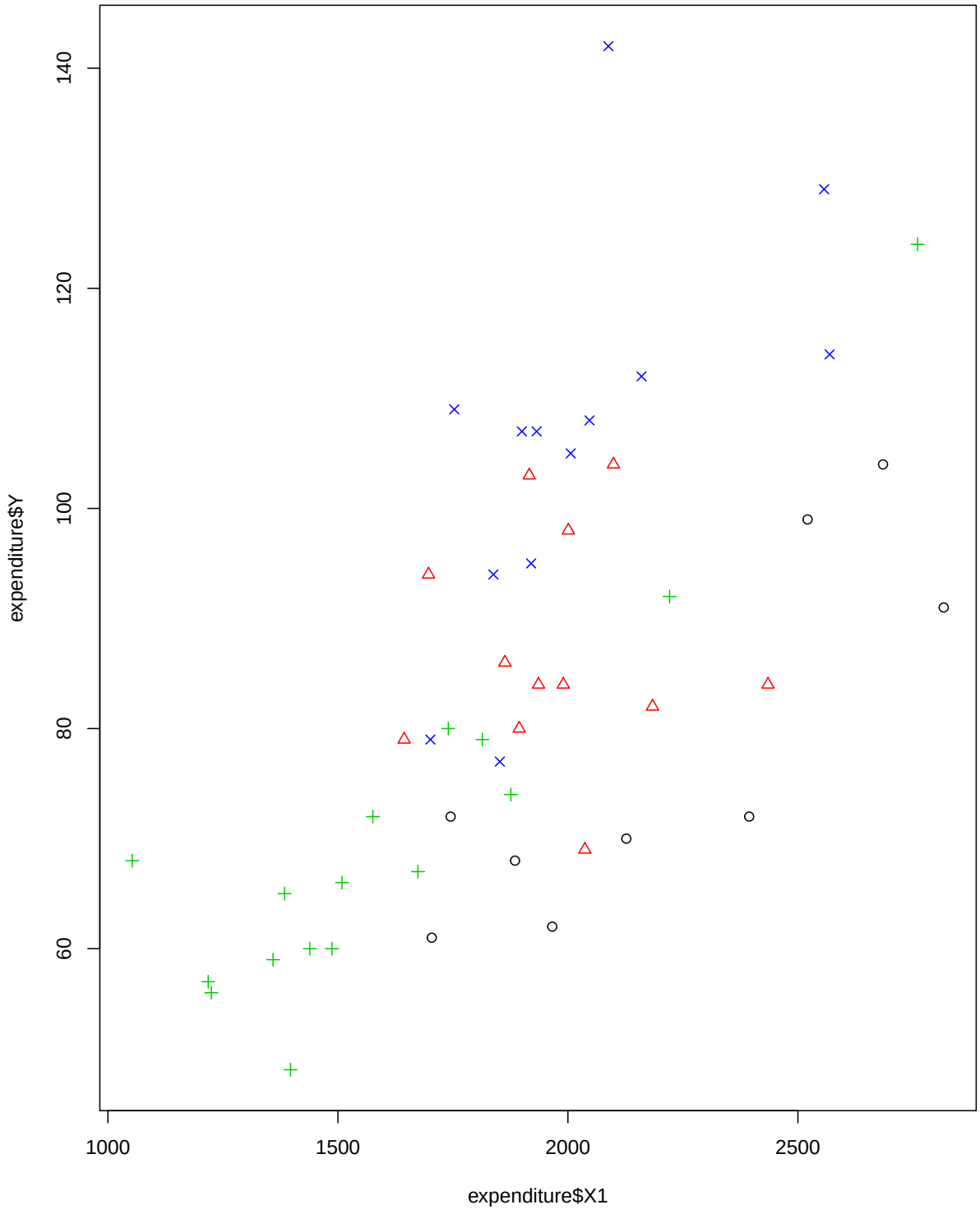
On average, Region 4 has the highest per capita expenditure on public education.

- 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.

```r
plot(expenditure$Region, expenditure$Y)

region1 <- filter(expenditure, expenditure$Region == 1)
mean(region1$Y)

region2 <- filter(expenditure, expenditure$Region == 2)
mean(region2$Y)

region3 <- filter(expenditure, expenditure$Region == 3)
mean(region3$Y)

region4 <- filter(expenditure, expenditure$Region == 4)
mean(region4$Y)
```

Between Y and $X1$:

`From`	*linear association*
`Direction`	*positive association*
`Strength`	*moderately strong*
`Outliers`	*present*
`Correlation`	*0.65*


expenditure$Y
140
120
100
80
60
1000
1500
2000
2500
expenditure$X1