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.

```
#find t statistics for a 90% confidence level
t90 <- qt((1-0.9)/2, df= 24, lower.tail = FALSE)
n <- length(y) #sample size
ymean <- mean(y) # sample mean
ysd <- sd(y) # sample standard deviation
lower <- ymean - t90*(ysd/sqrt(n))
```

```r
upper <- ymean + t90*(ysd/sqrt(n))
#construct the confidence interval
confint90 <- c(lower, upper)
confint90 #[1] 93.95993 102.92007
```

We are 90% confident that the population IQ lies within the range of 93.95993 to 102.92007

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.

```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 assuming that the school average IQ is higher than the country average
    (100)
t.test(y, mu=100, alternative="greater", conf.level=0.95)
#t = -0.59574, df = 24, p-value = 0.7215
```

Since p-value = 0.7215, which is greater than significance level 0.05, This does not reject the null hypothesis and thus we cannot conclude that the school mean IQ is higher than the country's mean IQ

Question 3 (50 points)

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

`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`.

```
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)
```

- 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 we can see in *Figure 1* and *Figure 3*, public education expenditure is positively correlated with both income and urban population; As seen in *Figure 2*, education expenditure is weakly negatively correlated with number of residents at school age.

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

  As seen in *Figure 4*, All four regions have very different expenditures on public education. The West has by far the highest per capita expenditure on public education and the three other regions are all comparatively lower.

- 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.

  As seen in *Figure 5*, the public educational expenditure is positively correlated to personal income per capita. Generally, there are no evident clusters for any of the regions, but differences in regional distribution are still distinct. We can also see from the distribution of the green plus sign that the South contributed to this correlation the most, with other regions spreading more to the upper right.

```
#Relationship between Y,X1,X2 and X3
plot(expenditure$X1, expenditure$Y, main = "Public education Expenditure
    vs. personal income",
                                    xlab = "personal income per capita",
                                    ylab = "expenditure per capita")
#The best fit line
abline(lm(expenditure$Y~expenditure$X1))

#X2
plot(expenditure$X2, expenditure$Y, main = "Public education Expenditure
    vs. underaged population",
                                    xlab = "number of resident per
    thousands under 18",
                                    ylab = "expenditure per capita")
#The best fit line
```

```r
abline(lm(expenditure$Y~expenditure$X2))
#X3
plot(expenditure$X3, expenditure$Y, main = "Public education Expenditure vs. urban population",
                                    xlab = "number of urban resident per thousands",
                                    ylab = "expenditure per capita")
#The best fit line
abline(lm(expenditure$Y~expenditure$X3))

#Relationship between Y and Regions
library(Rmisc)
summarySE(data = expenditure, measurevar = "Y", groupvars = "Region")
#Create a new factor variable YR
expenditure$YR <- expenditure$Y
expenditure$YR <- factor(NA, levels = c("Northeast", "North Central", "South", "West"))
expenditure$YR[expenditure$Region == 1] <- "Northeast"
expenditure$YR[expenditure$Region == 2] <- "North Central"
expenditure$YR[expenditure$Region == 3] <- "South"
expenditure$YR[expenditure$Region == 4] <- "West"
summary(expenditure$YR)
#A side by side boxplot
boxplot(expenditure$Y~expenditure$YR, main = "Public education Expenditure vs. Regions",
                                      xlab = "Regions",
                                      ylab = "expenditure per capita")
#Y vs X1 plus regional information
plot(expenditure$X1, expenditure$Y, main = "Public education Expenditure vs. personal income",
     xlab = "personal income per capita",
     ylab = "expenditure per capita", col = expenditure$Region, pch = expenditure$Region)
legend(x="topright", legend = levels(expenditure$YR), col = c(1,2,3,4), pch= c(1,2,3,4))
```

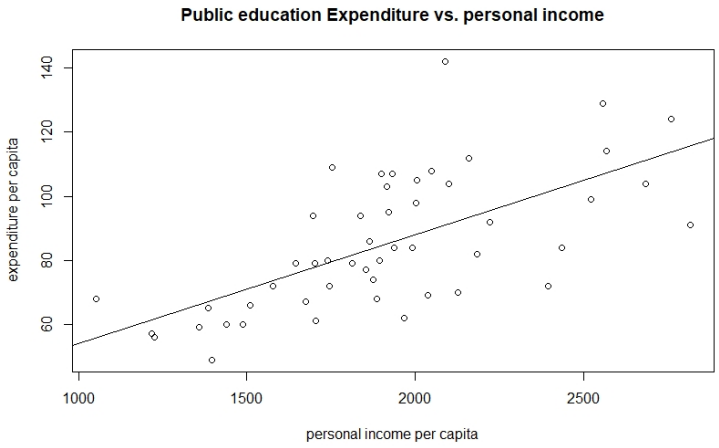

Figure 1: Y and X1

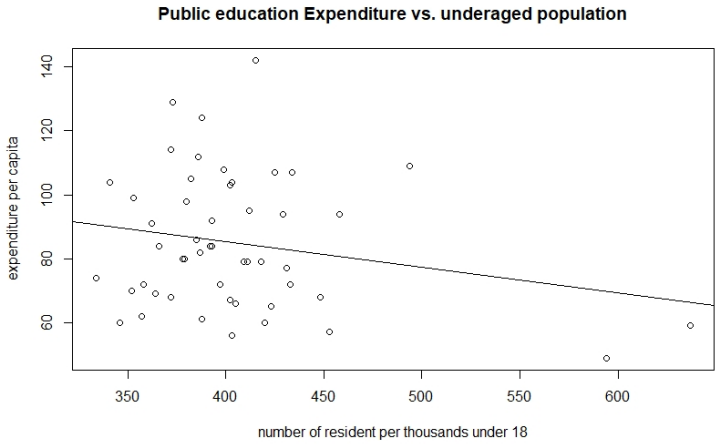

Figure 2: Y and X2

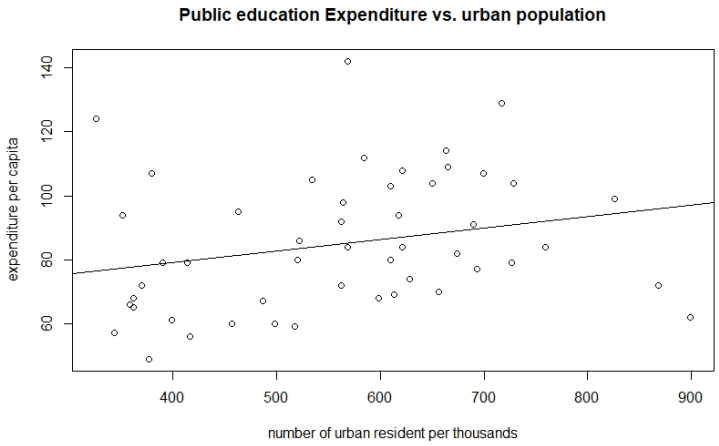

Figure 3: Y and X3

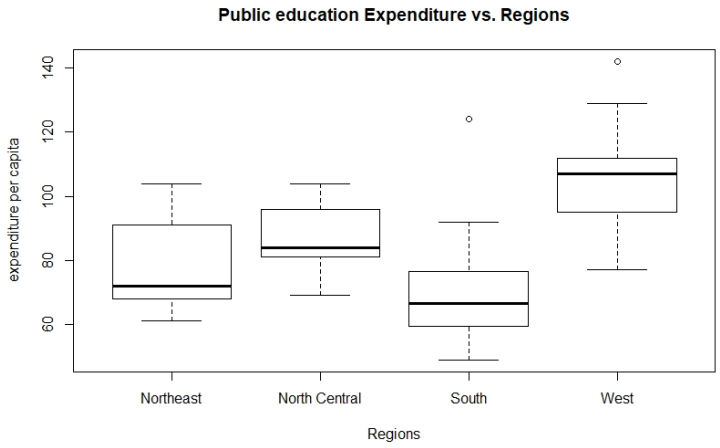

Figure 4: Y and Region

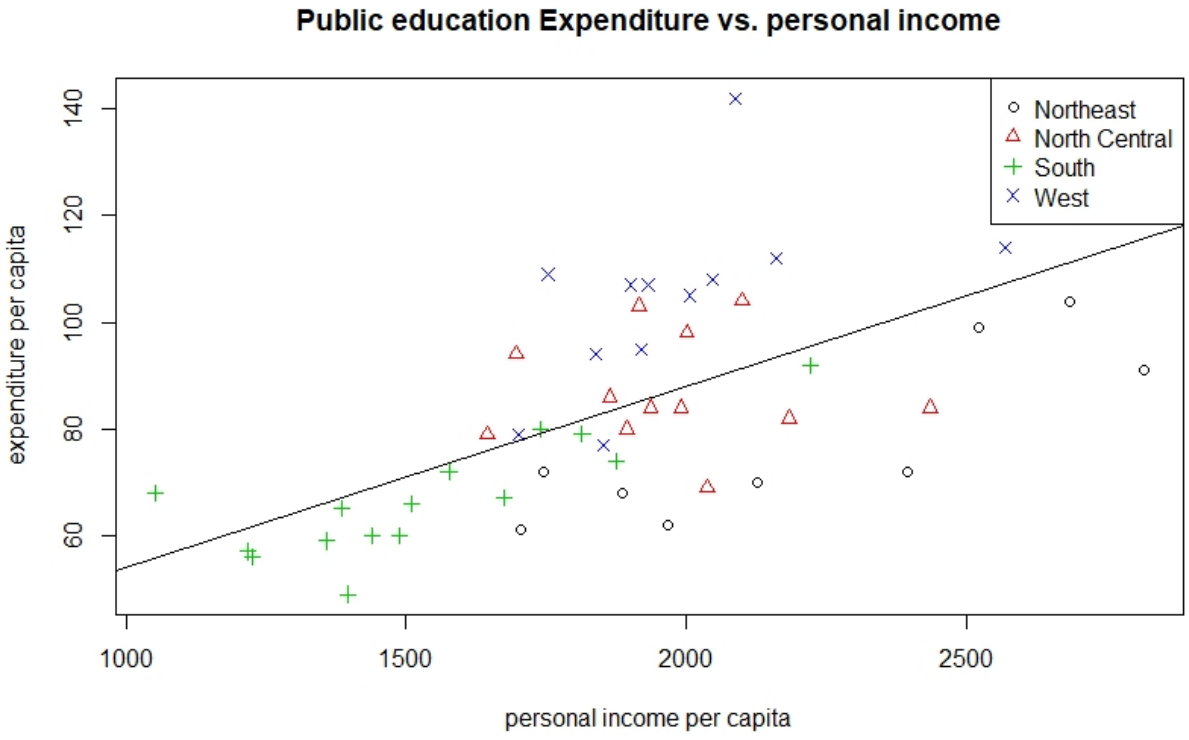

Figure 5: Y vs. X1 with Regions labeled