Problem Set 1

QTM 200: Applied Regression Analysis

Due: January 27, 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:

```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)
```

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.

```r
#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)
#setting n as sample size
n <- 25
#finding z-score
```

```
z90 <- qnorm((1 - .90)/2, lower.tail = FALSE)
#finding the mean and standard deviation
sample_mean <- mean(y, na.rm = TRUE)
sample_sd <- sd(y, na.rm = TRUE)
#finding the lower and upper bounds of the interval
lower_90 <- sample_mean - (z90 * (sample_sd/sqrt(n)))
upper_90 <- sample_mean + (z90 * (sample_sd/sqrt(n)))
#calculating a 90% confidence interval
confint90 <- c(lower_90, upper_90)
confint90
#The confidence interval is (94.13283, 102.74717).
#This means that if we repeated this process 100 times, we can expect 90% of
    confidence intervals to contain the true population mean.
```

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:

```
#####################
```

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

```
#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)
#set sample size as 25
n2 <- 25
#find the mean of the sample
xbar <- mean(y)
#the mean score is 98.44
s <- sd(y)
#the standard deviation of the sample is 13.09287
#one-sample, one-sided t-test:
tscore <- (xbar - 100)/(sqrt((s^2)/n2))
tscore
#The t-score is -0.5957439. Now find the p-value:
pt(-abs(tscore), lower.tail = FALSE, df = length(y) - 1)
#The p-value is 0.72153.
#Check using the t.test function:
t.test(y, mu=100, alternative="greater", conf.level = 0.95)
#If the average IQ of the students was 100, the probability of selecting a
    sample with a mean less than or equal to 98.44 is about 72%.
```

Question 3 (50 points)

Assume y is variable with values 1,2,3,4 standing for “Freshman”, “Sophomore”, “Junior”, and “Senior”, convert y from numbers to characters in R:

```
######################
```

```
#data set:
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)
#converting numerical values into character variables:
y2 <- as.character(y)
y2[y=="1"] <- "freshman"
y2[y=="2"] <- "sophomore"
y2[y=="3"] <- "junior"
y2[y=="4"] <- "senior"
y2
```

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.

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*

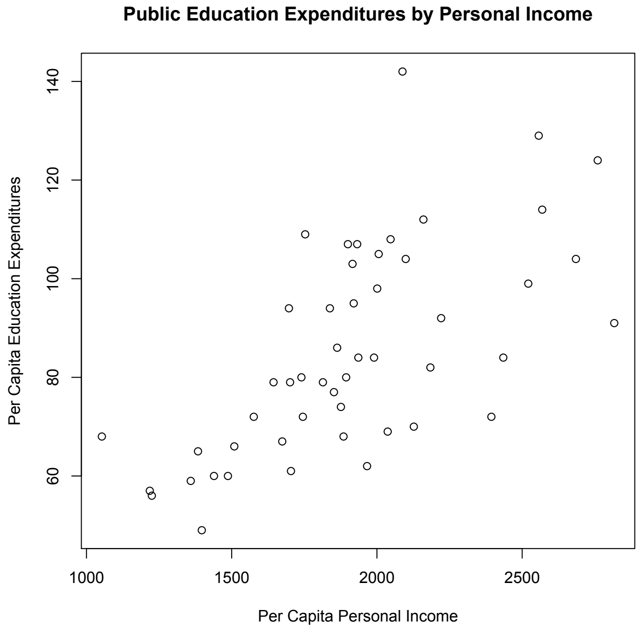
Explore the `expenditure` data set and import data into R.

```
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)
```

```
#import expenditure data set
expenditure <- read.table("expenditure.txt", header=T)
```

- Please plot the relationships among *Y*, *X1*, *X2*, and *X3*? What are the correlations among them? Describe the graph and the relationships among them.

```
#Plot the relationship between y and x1.
plot(x = expenditure$X1, y = expenditure$Y, xlab = "Per Capita Personal
   Income", ylab = "Per Capita Education Expenditures", main = "Public
   Education Expenditures by Personal Income")
```



Income graph.pdf

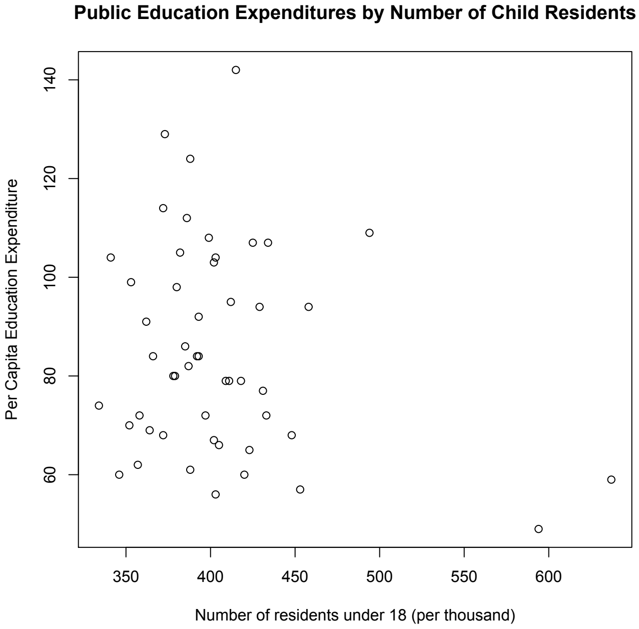
```
#This scatterplot shows a moderately strong positive linear correlation between the per capita expenditure on public education and per capita personal income within a state.
```

```
#Plot the relationship between y and x2.
plot(expenditure$X2, expenditure$Y, xlab = ”Number of residents under 18 (per thousand)”, ylab = ”Per Capita Education Expenditure”, main = ”Public Education Expenditures by Number of Child Residents”)
#This scatterplot indicates that there is no correlation between per capita public education expenditures and the number of residents under 18 years old in a state.
```

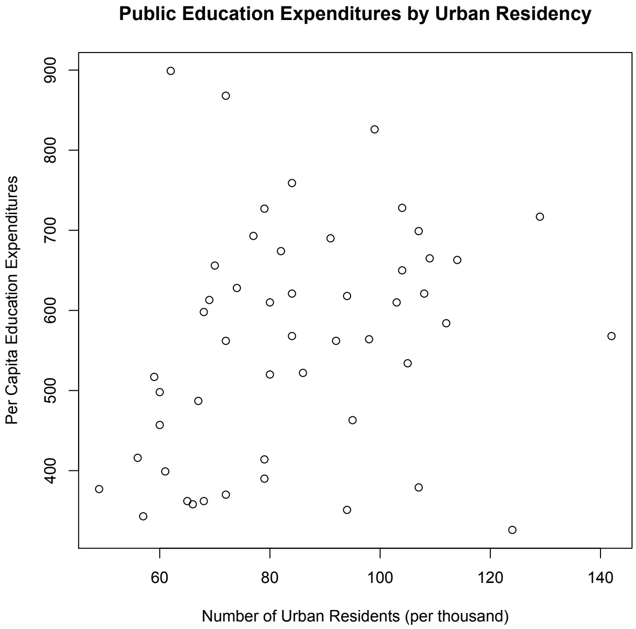
```
#Plot the relationship between y and x3.
plot(expenditure$Y, expenditure$X3, xlab = ”Number of Urban Residents (per thousand)”, ylab = ”Per Capita Education Expenditures”, main = ”Public Education Expenditures by Urban Residency”)
#This scatterplot may show a very weak positive (possibly linear?) association between per capita public education expenditures and the number of urban residents in a state.
```

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

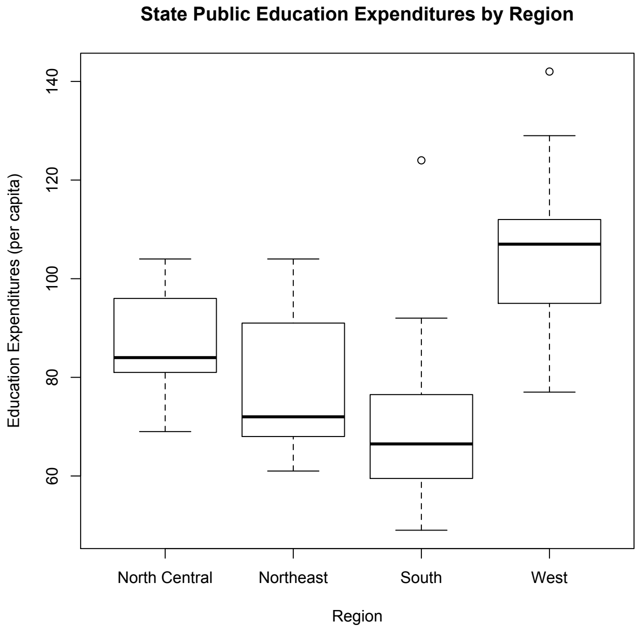
```
#Plot the relationship between Y and Region.
class(expenditure$Region)
#Change ”Region”'s variable type from integer to characters
expenditure$RegionNames <- as.character(expenditure$Region)
```



Residents.pdf



Residency graph.pdf

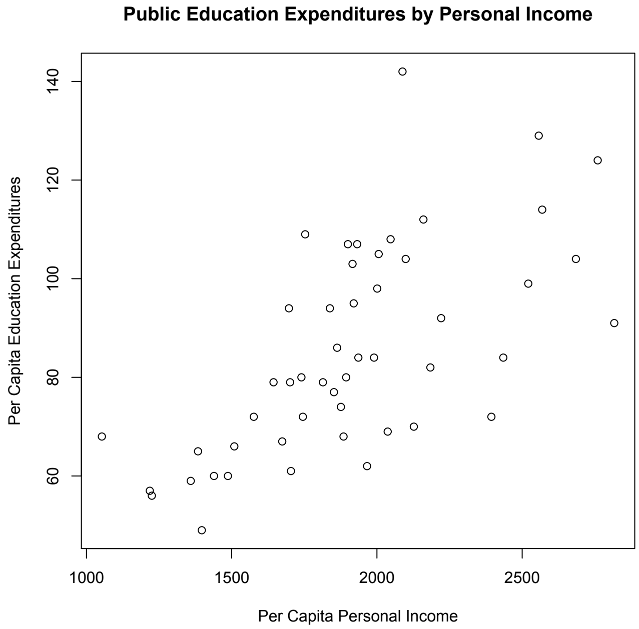


boxplot.pdf

```
expenditure$RegionNames[expenditure$Region=="1"] <- "Northeast"
expenditure$RegionNames[expenditure$Region=="2"] <- "North Central"
expenditure$RegionNames[expenditure$Region=="3"] <- "South"
expenditure$RegionNames[expenditure$Region=="4"] <- "West"
#Since we have a continuous variable (education expenditures), and a categorical variable (region), we will use a boxplot.
boxplot1 <- boxplot(expenditure$Y~expenditure$RegionNames, xlab = "Region", ylab = "Education Expenditures (per capita)", main = "State Public Education Expenditures by Region")
#The boxplot displays the average public education expenditures by region. It appears that the West has the highest average expenditures out of all the regions, with an average expenditure around 110.
```

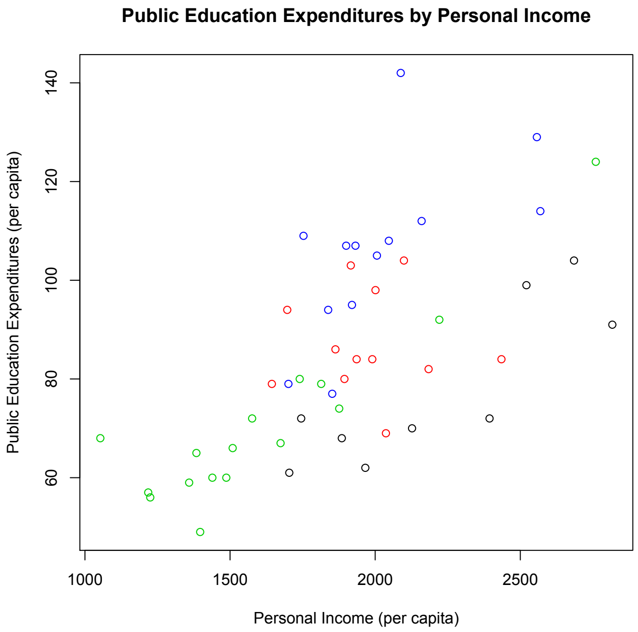
- Please plot the relationship between *Y* and *X1*? Describe this graph and the relationship. Reproduce the above graph including one more variable *Region* and display different regions with different types of symbols and colors.

```
#Plot the relationship between Y and X1.
plot(expenditure$X1, expenditure$Y, xlab = "Personal Income (per capita)", ylab = "Public Education Expenditures (per capita)", main = "Public Education Expenditures by Personal Income")
#Again, this scatterplot seems to display a moderately strong, positive, linear association between personal income per capita and the state's public education expenditures per capita.
#Add "Region" by adding colors.
plot(expenditure$X1, expenditure$Y, col = expenditure$Region, xlab = "Personal Income (per capita)", ylab = "Public Education Expenditures (per capita)", main = "Public Education Expenditures by Personal Income
```

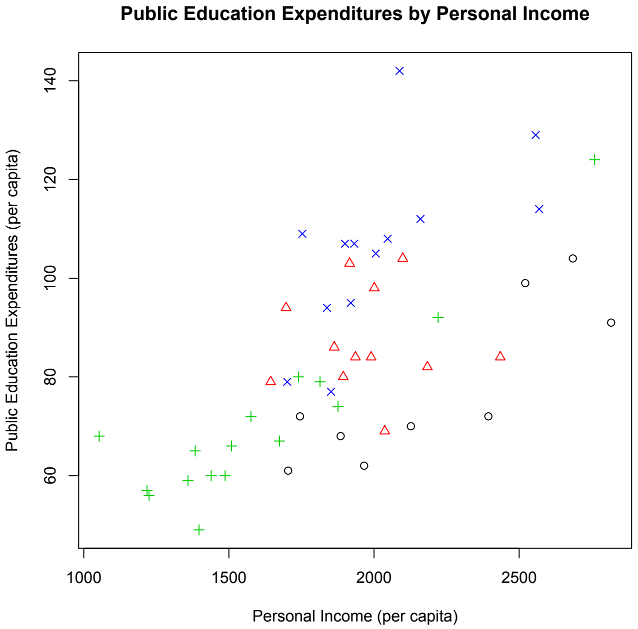


Income graph.pdf

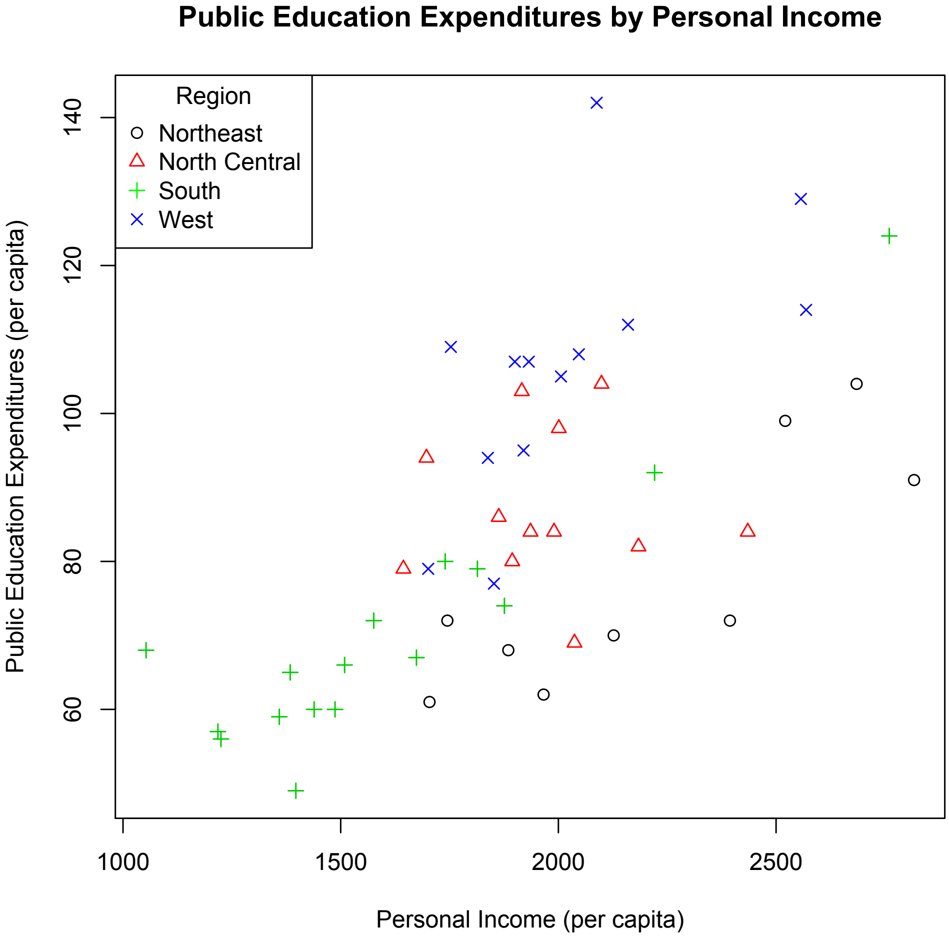
```
    ")
#Change plotting symbols according to Region.
plot(expenditure$X1, expenditure$Y,  col = expenditure$Region, pch =
    expenditure$Region, xlab = "Personal Income (per capita)", ylab = "
    Public Education Expenditures (per capita)", main = "Public Education
    Expenditures by Personal Income")
#Add legend explaining the colors and symbols.
legend("topleft", title = "Region", legend = c("Northeast", "North
    Central", "South", "West"), col = c("black", "red", "green", "blue"),
    pch = c(1, 2, 3, 4))
```



Income + color.pdf



Income + color + symbols.pdf



Income (region) with legend.pdf