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:

```
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.

Answer:

First, I calculated the t-score for a t-distribution with 24 degrees of freedom (degrees of freedom = sample size - 1; 25 - 1 = 24) using the qt function. I used the t-distribution to answer this question because the sample size is less than 30. The area of each tail is (1 - 0.9) / 2 using the formula (1 - confidence coefficient) / 2. Using the "lower.tail = FALSE" argument, I calculated the positive t-score for the distribution.

```
t90 <- qt((1 - 0.9) / 2, lower.tail = FALSE, df = 24) # T-score for 90%
    confidence interval with 24 degrees of freedom
```

Then, I stored the sample size, sample mean, and the standard deviation into variables using R functions.

```
n <- length(y) # Sample size
sample_mean <- mean(y) # Sample mean
sample_sd <- sd(y) # Sample standard deviation
```

Then, I calculated the standard error and the upper/lower bounds of the 90% confidence interval using these two formulas:

$$se_{\bar{x}} = \frac{s}{\sqrt{n}}$$

Figure 1: Formula for standard error of sampling distribution

$$\bar{x} \pm t_{\alpha} \times se_{\bar{x}}$$

Figure 2: Formula for confidence interval about a mean

```
sample_se <- sample_sd / sqrt(n) # Standard error
lower_90 <- sample_mean - (t90 * sample_se) # Lower bound of 90% confidence
    interval
upper_90 <- sample_mean + (t90 * sample_se) # Upper bound of 90% confidence
    interval
confint <- c(lower_90, upper_90)
confint # 90% confidence interval (93.96, 102.92)

t.test(y, conf.level = 0.9) # checking answer
```

The 90% confidence interval for student IQ in the school is (93.96, 102.92).

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:

```
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.

Answer:

Based on the question, some assumptions that are satisfied are that the sample is normally distributed and that the data is quantitative (I'm still using the t-distribution because of the small sample size). The null hypothesis is that the true mean IQ of students is less than or equal to 100, while the alternative hypothesis is that the true mean is greater than 100.

First, I calculated the t-statistic using the sample mean, the hypothesized mean, and the standard error (same variables and values from Question 1). Then, I calculated the p-value, or the probability that the t-statistic takes a value this extreme or more, given that the null hypothesis is true. I conducted a one-tailed test to see if the true mean IQ is higher than 100.

$$t = \frac{\bar{x} - \mu_0}{se_{\bar{x}}}$$

Figure 3: Formula for t-statistic

```
tstat <- (sample_mean - 100) / sample_se # Calculating test statistic
pvalue <- pt(tstat, df = 24, lower.tail = F) # Calculating the p-value for a
    one-tailed test
pvalue # 0.7215; greater than the alpha value of 0.05

t.test(y, mu = 100, alternative = "greater") # Checking answer
```

Since the p-value of 0.7215 is greater than the alpha value of 0.05, we fail to reject the Ho. Therefore, there is not sufficient evidence to conclude that the true national mean IQ in schools is significantly higher than 100.

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`:

I created a variable yNew for y and assigned student classifications based on the corresponding numerical values.

```
yNew <- y # Creating new variable for class year

# Converting the numbers to class year
yNew[y==1] <-"Freshman"
yNew[y==2] <-"Sophomore"
yNew[y==3] <-"Junior"
yNew[y==4] <-"Senior"
table(yNew) # Checking recoding
```

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`.

```
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.

- Please plot the relationship between *Y* and *Region*? On average, which region does have 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.

Answer:

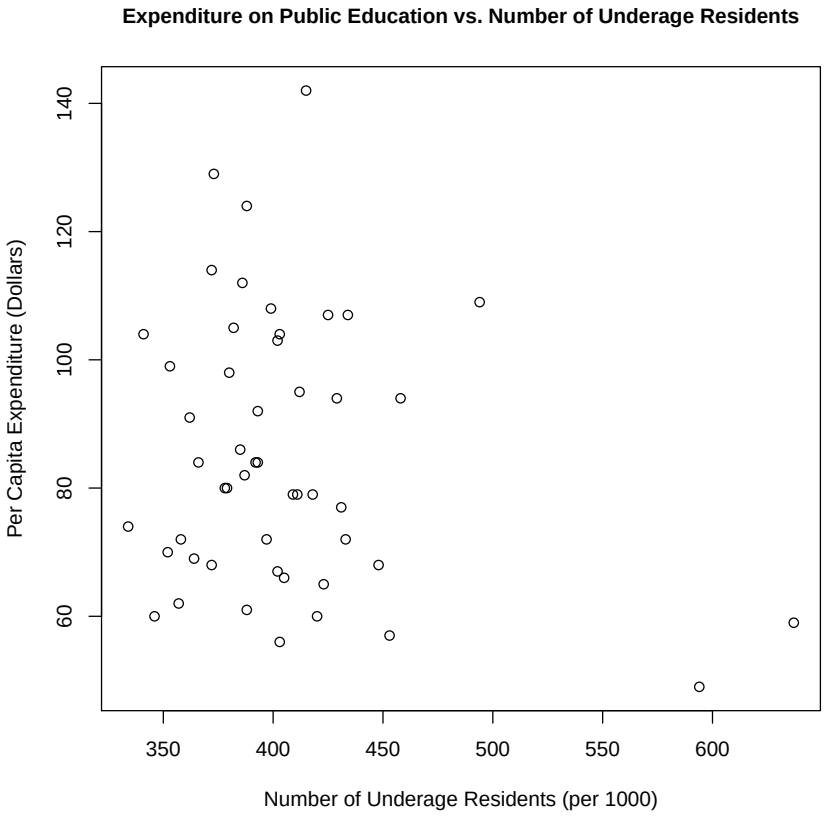

Figure 4: Relationship between Y and X2

```
# Expenditure vs. Residents under 18 years old
pdf("Graph1.pdf")
plot(expenditure$X2, expenditure$Y, cex.main = 1, main = "Expenditure on
    Public Education vs. Number of Underage Residents", xlab = "Number of
    Underage Residents (per 1000)", ylab = "Per Capita Expenditure (
    Dollars)")
dev.off()
```

The graph shows a weak to no correlation between the number of residents under 18 years old and the per capita expenditure on public education. This is interesting since I expected an increased demand and expenditure for public education with more children/teenagers.

Also, there are a few outliers with unusually large numbers of underage residents and corresponding low per capita expenditure on public education.

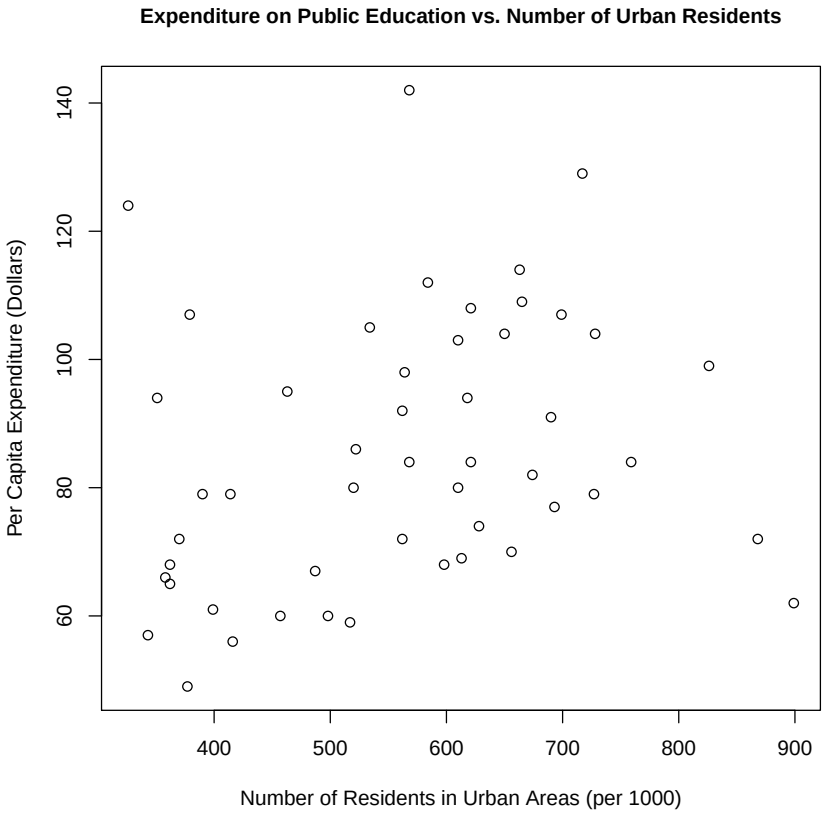

Figure 5: Relationship between Y and X3

```r
# Expenditure vs. Number of urban residents
pdf("Graph2.pdf")
plot(expenditure$X3, expenditure$Y, cex.main = 1, main = "Expenditure on
    Public Education vs. Number of Urban Residents", xlab = "Number of
    Residents in Urban Areas (per 1000)", ylab = "Per Capita Expenditure (
    Dollars)")
dev.off()
```

The graph shows a weak to moderate, positive, linear correlation between the number of residents in urban area and the per capita expenditure on public education. Since population density in urban areas are greater than that of rural areas, the expenditure on public education may be greater as a result.

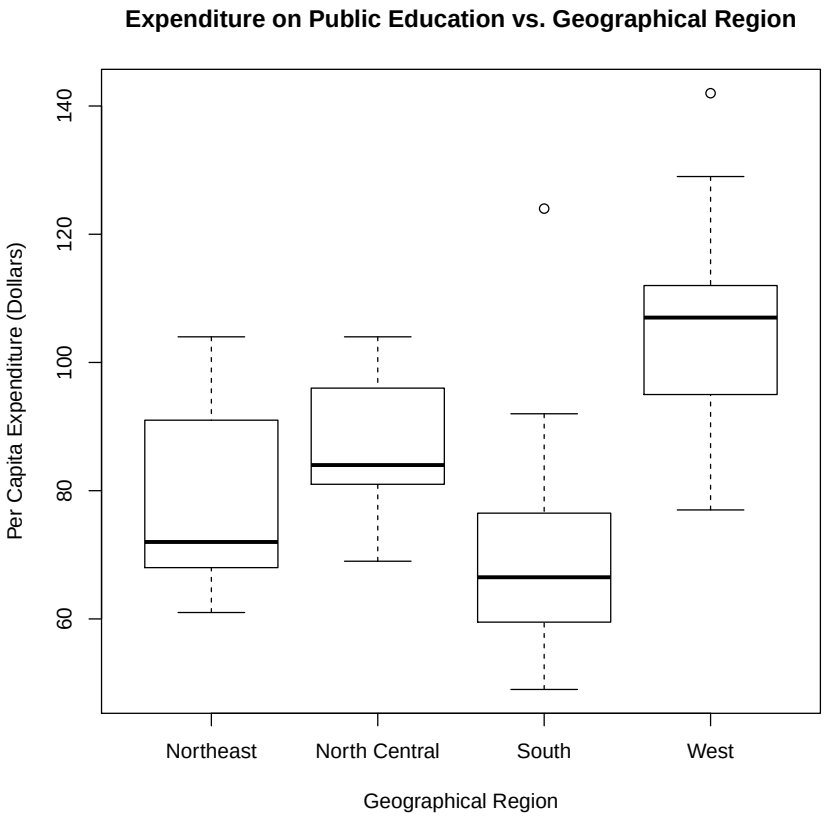

Figure 6: Relationship between Y and Region

```
# Expenditure vs. Region
pdf("Graph3.pdf")
boxplot(expenditure$Y ~ expenditure$Region, names = c("Northeast", "North
    Central", "South", "West"), main = "Expenditure on Public Education
   vs. Geographical Region", xlab = "Geographical Region", ylab = "Per
   Capita Expenditure (Dollars)")
dev.off()
```

First, I recoded the variable Region to show names of the regions instead of numbers.

```
expenditure$newRegion <- factor(NA, c("Northeast", "North Central", "
   South", "West"))
expenditure$newRegion[expenditure$Region == 1] <- "Northeast"
expenditure$newRegion[expenditure$Region == 2] <- "North Central"
expenditure$newRegion[expenditure$Region == 3] <- "South"
expenditure$newRegion[expenditure$Region == 4] <- "West"
```

On average, states in the West have the highest per capita expenditure on public education compared to other states.

Using the summarySE function also shows that the average expenditure on public

education is greatest in the West.

	newRegion	N	Y	sd	se	ci
1	Northeast	9	77.66667	16.07016	5.356720	12.352617
2	North Central	12	87.25000	10.46314	3.020448	6.647961
3	South	16	70.50000	17.84750	4.461876	9.510263
4	West	13	106.00000	17.73885	4.919871	10.719478

Figure 7: Result from summarySE

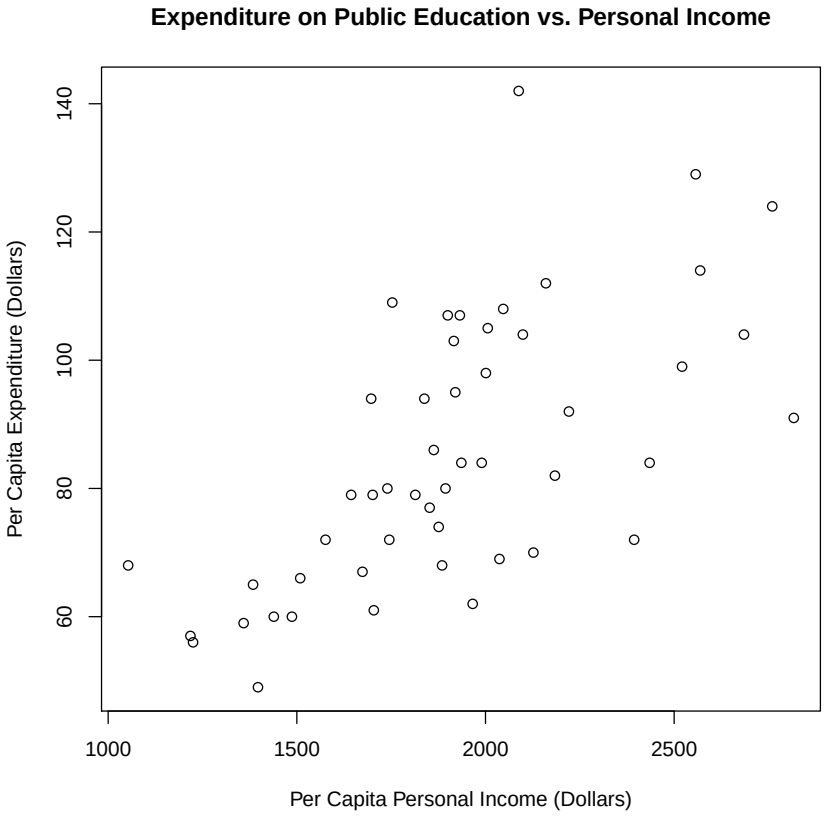

Figure 8: Relationship between Y and X1

```
# Expenditure vs. Income
pdf("Graph4.pdf")
plot(expenditure$X1, expenditure$Y, main = "Expenditure on Public
    Education vs. Personal Income", xlab = "Per Capita Personal Income (
    Dollars)", ylab = "Per Capita Expenditure (Dollars)")
dev.off()
```

There is a moderate to strong, positive, linear correlation between personal income and per capita expenditure on public education. Since communities with higher income

have higher spending power, expenditure on public education may be greater as a result.

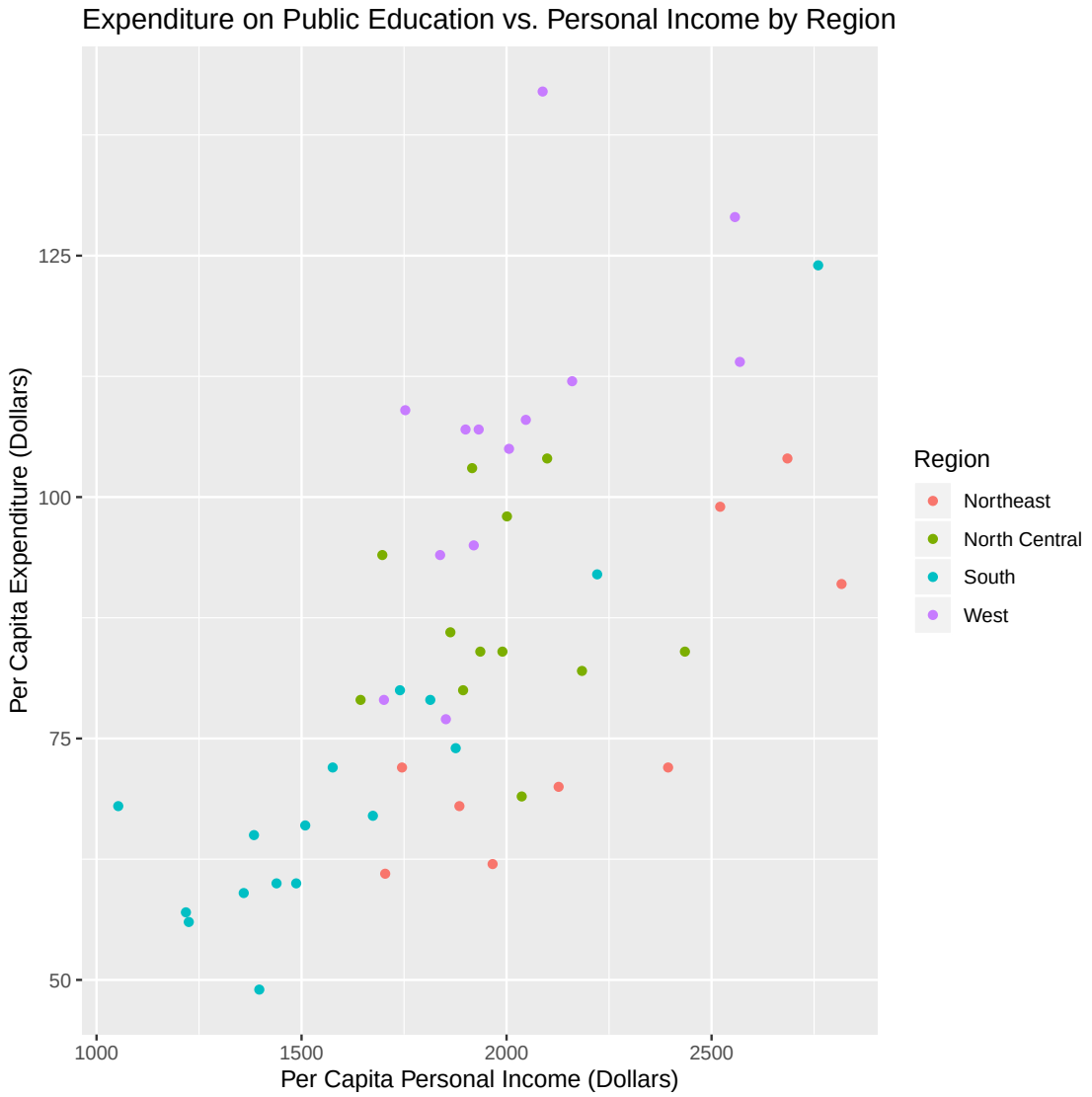

Figure 9: Relationship between Y and X1 illustrated with Respective Regions Colored

```
# Replotting Expenditure vs. Income color-coded by region
pdf("Graph5.pdf")
Graph5 <- ggplot(expenditure, aes(X1, Y, colour = newRegion)) + geom_point
    ()
print(Graph5 + labs(colour = "Region", title = "Expenditure on Public
    Education vs. Personal Income by Region", x = "Per Capita Personal
    Income (Dollars)", y = "Per Capita Expenditure (Dollars)"))
dev.off()
```

Based on this graph, we can see that the Western region tend to have higher expenditure compared to other regions and has moderate to high per capita income. In contrast, the South tends to have lower per capita income and lower per capita expenditure. It would be interesting to explore the infrastructure and the public policy of different regions to see how they interact with variables like individual income and expenditure on public education.