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. The area of each tail is (1 - 0.9) / 2, and 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)

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

n <− length(y)
sample_mean <− mean(y)
sample_sd <− sd(y)

Then, I calculated the standard error and the upper/lower bounds of the 90% confidence interval.

sample_se <- sample_sd / sqrt(n)
lower_90 <- sample_mean - (t90 * sample_se)
upper_90 <- sample_mean + (t90 * sample_se)
confint <- c(lower_90, upper_90)
confint = (93.95993, 102.92007)

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:

First, I conducted a one sample t-test:

```
t.test(y, mu = 100)
```

Output:

```
        One Sample t-test

data:  y
t = -0.59574, df = 24, p-value = 0.5569
alternative hypothesis: true mean is not equal to 100
95 percent confidence interval:
  93.03553 103.84447
sample estimates:
mean of x
    98.44
```

Since the p-value of 0.5569 is greater than the alpha of 0.05, we fail to reject the Ho. Therefore, we conclude that the true national mean IQ in schools is not significantly different from 100. Moreover, we are 95% confident that the true national mean IQ in schools is between 93.04 and 103.84. There is not sufficient evidence to conclude that the true mean is significantly different from 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.

```
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)
yNew <- y
yNew[y==1] <- "Freshman"
yNew[y==2] <- "Sophomore"
yNew[y==3] <- "Junior"
yNew[y==4] <- "Senior"
table(yNew)
```

```
Freshman    Junior    Senior Sophomore
      11         6         5         5
```

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: