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.

```
#Since n<30, I use t-score
t90 <- qt ((1-0.90)/2, df= (25-1), lower.tail = FALSE)
sample_mean <- mean (y)
sample_sd <- sd (y)
```

```r
lower_90 <- sample_mean-(t90 * (sample_sd/sqrt(25)))
upper_90 <- sample_mean+(t90 * (sample_sd/sqrt(25)))
CI90 <- c(lower_90, upper_90)
CI90 #The confidence interval is (93.96, 102.92).
```

The confidence interval is (93.96, 102.92). Therefore, we're 90 percent confident that the true mean will fall within the interval of (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:

```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
#Ho: u=u0;    Ha: u>u0.
t.test(y,mu=100,alterntive="greater",conf.level=0.95)
```

I fail to reject the null hypothesis because p-value = 0.557, which is greater than 0.05. The average IQ of the school does not significantly differ from 100 on average.

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`.

```
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)?

```
ggplot(expenditure, aes(expenditure$X1, expenditure$Y))+
  geom_point()+
  labs(x="Personal Income",y="Expenditure",title="Figure 1:Personal
    Income and Educational Expenditure")
```

The graph shows that as the per capita personal income increases, per capita expenditure on public education increases.

```
ggplot(expenditure, aes(expenditure$X2, expenditure$Y))+
  geom_point()+
  labs(x="People under 18 yrs old",y="Expenditure",title="Figure 2:People
     under 18 and Educational Expenditure")
```

The graph does not show a clear relationship between the number of residents under 18 years of age and the per capita expenditure on public education.

```
ggplot(expenditure, aes(expenditure$X3, expenditure$Y))+
  geom_point()+
  labs(x="People in urban area",y="Expenditure",title="Figure 3:People in
     urban area and Educational Expenditure")
```

The graph shows that as the number of people residing in urban areas increases, per capita expenditure on public education increases.

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

```
boxplot(expenditure$Y~expenditure$Region)
```

The west has the highest per cpita expenditure on public education.

Figure 1: Figure 1:Personal Income and Educational Expenditure

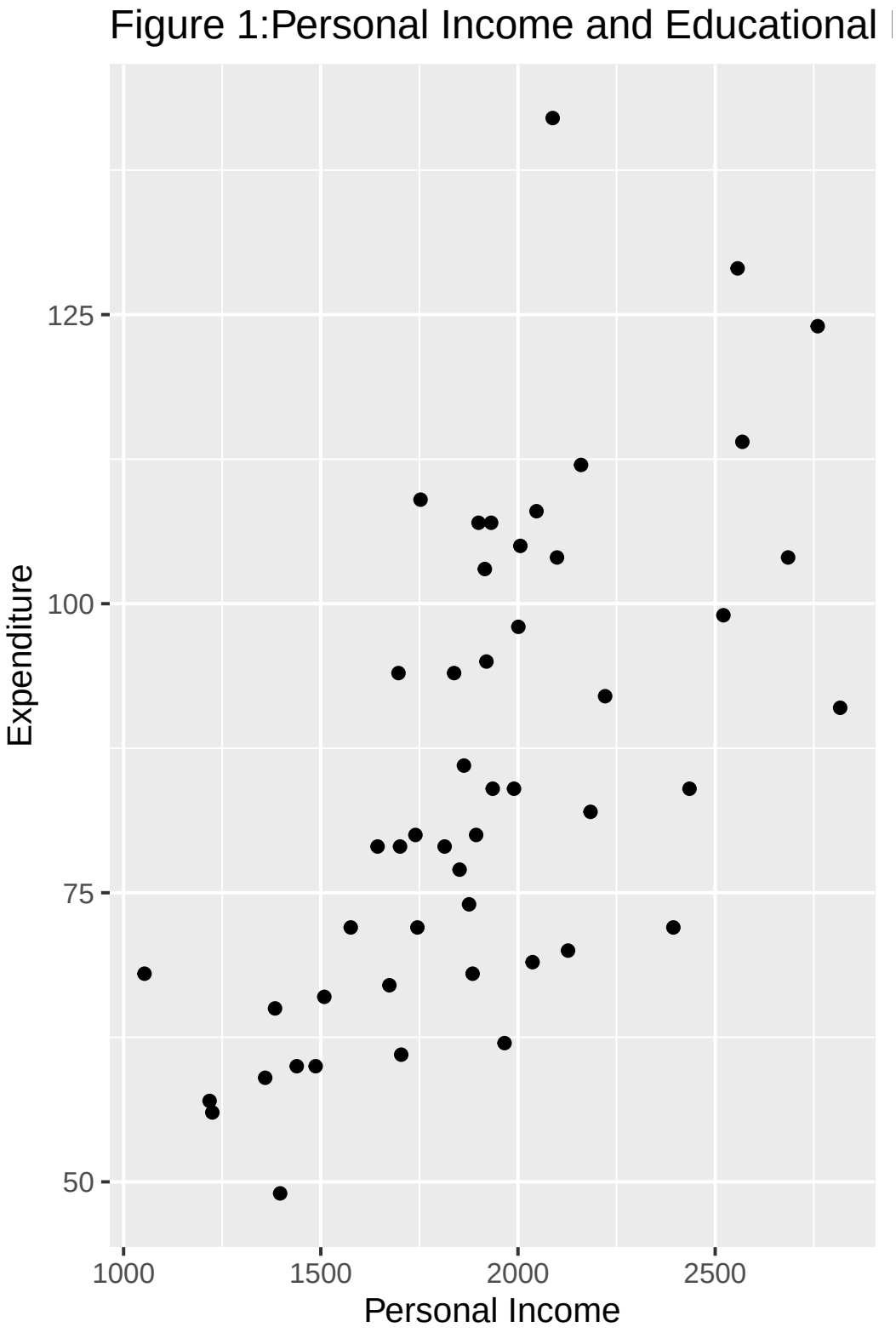


Figure 2: Figure 2:People under 18 and Educational Expenditure

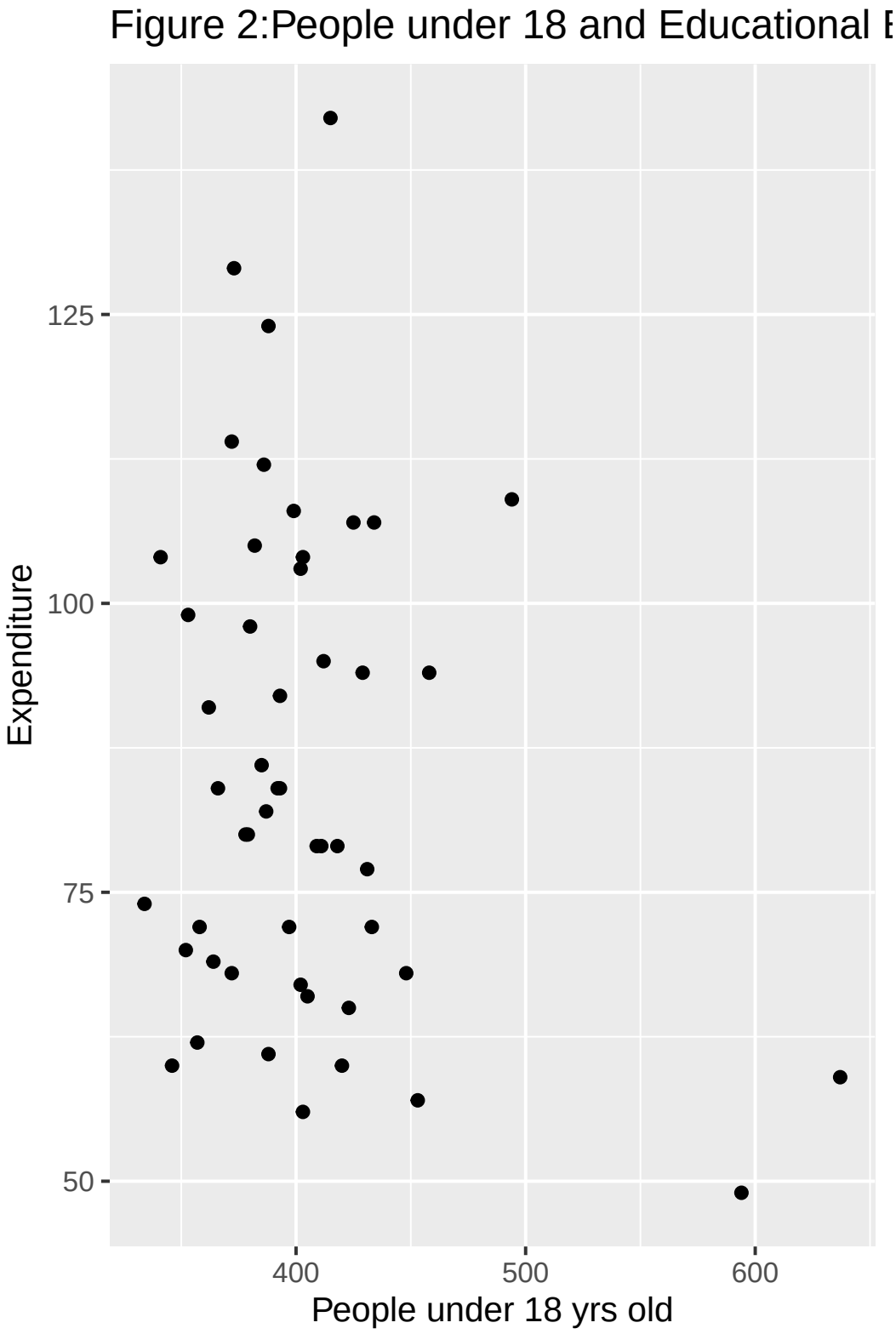


Figure 3: Figure 3:People in urban area and Educational Expenditure

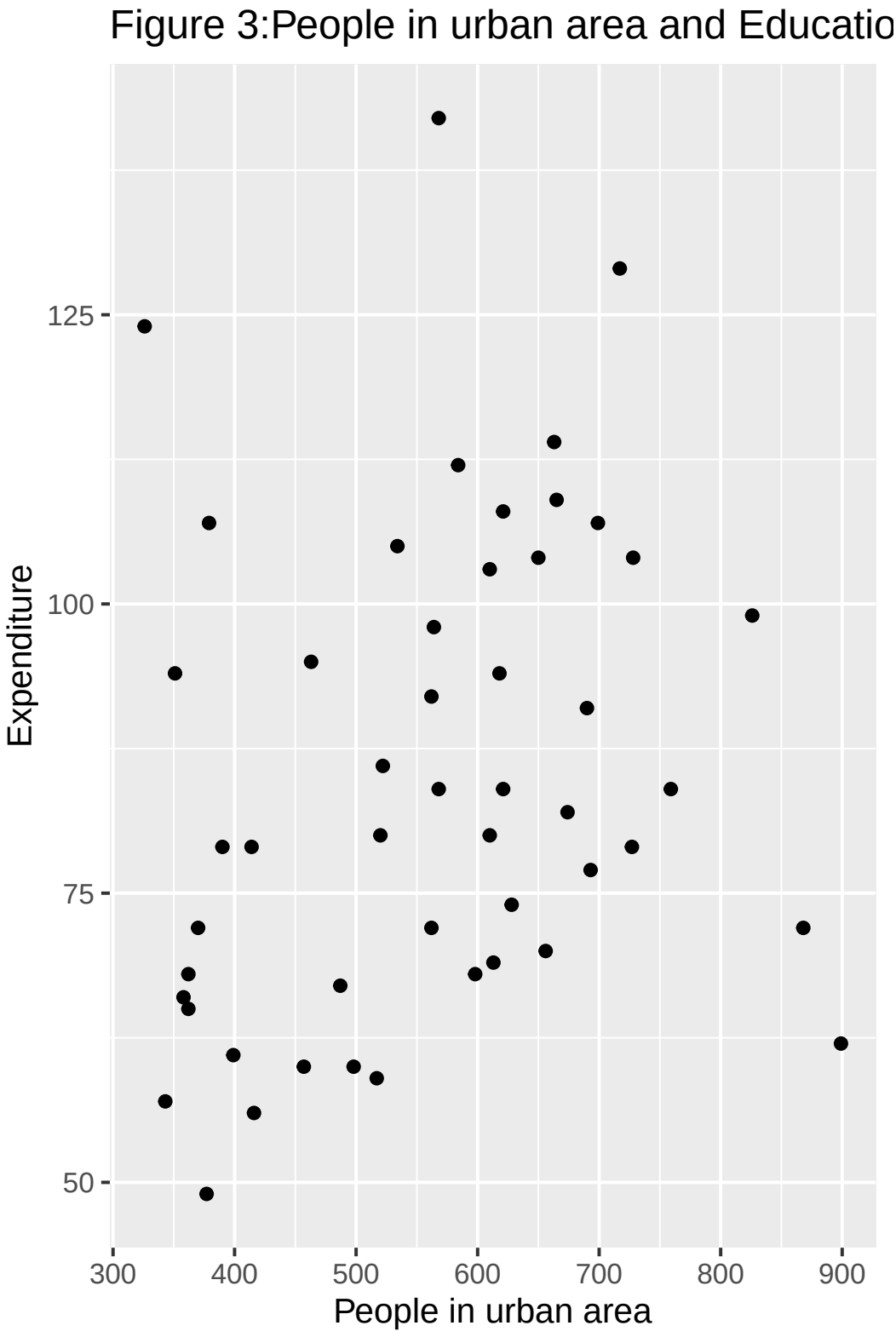
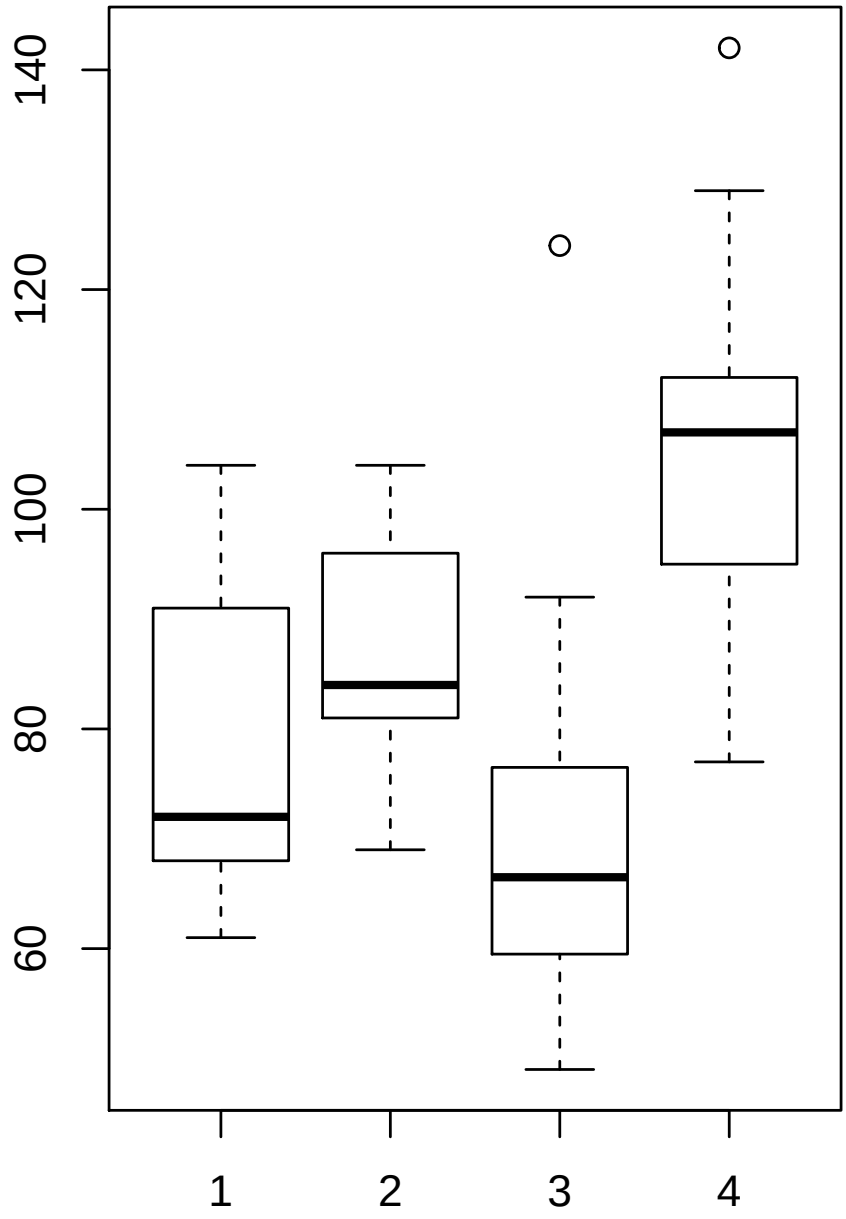


Figure 4: Figure 4:Region and Educational Expenditure



- 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.

```
ggplot(expenditure, aes(x=expenditure$X1,y=expenditure$Y))+
  geom_point(aes(color=expenditure$Region))+
  scale_color_gradient(name="Region", low="blue", high="orange")+
  labs(x="Personal Income",y="Expenditure")
```

People with higher income in the Northeast and West are more likely to spend more on public education. Whilst people in the North Central and South do not spend more on public education if they earn higher income.

Figure 5: Figure 5:Personal Income by Region and Educational Expenditure