Problem Set 1

Farris Sabir 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 29, 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.

```
meanIQ <- sum(y)/length(y)
demeanedSum <- NULL
for(i in 1:length(y)){
  demeanedSum[i] <- y[i] - meanIQ
}
```

```r
squaredError <- demeanedSum^2
variance <- sum(squaredError)/(length(y) - 1)
sdIQ <- sqrt(variance)
z90 <- qt((1-.9)/2, 24)
lower_90 <- meanIQ + z90*sdIQ/sqrt(length(y))
upper_90 <- meanIQ - z90*sdIQ/sqrt(length(y))
confint90 <- c(lower_90, upper_90)
confint90
t.test(y, conf.level = 0.9)
```

The z-scores of 1.71 and -1.71 on both sides of the t distribution corresponded to a 90% confidence interval. The standard error or standard deviation of the sampling distribution was determined by the sample's standard deviation (denoted by sdIQ) divided by the square root of the sample size. Hence, the product of the z-score and the standard error yielded the margin of error, which is 4.48. The margin of error was added to and subtracted from the mean, which was denoted by meanIQ, to obtain the confidence interval. This interval was verified by the R function t.test. Thus, the 90% confidence interval for the student IQ in the school is 94 to 103. If this was among 100 samples of students' IQ scores, confidence intervals calculated from 90 of them would have the true mean IQ score within them.

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:

The conditions meet the assumptions for a t-test because the data was collected from the students using a random sample. Also, because the sample is less than 30 individuals, the test cannot rely on central limit theorem to assume the sampling distribution is normally distributed. Rather, the IQ scores of the population are assumed to lie along a normal distribution since it is stated in the question.

The null hypothesis for this test is that the average IQ score of the students of this counselor's school is less than or equal to the IQ score of 100, the average for all schools in the country. The alternative hypothesis is that the average IQ score of the students of this counselor's school is greater than the IQ score of 100, the average for all schools in the country. In other

words, Ho: $\mu \leq 100$, and Ha: $\mu > 100$.

Because the population of IQ scores is normally distributed, the sampling distribution of sample mean ybar is normal about μ or the population mean or average IQ score of students in the school of the counselor. To assess how far this population mean deviates from the national mean IQ score of 100, the standard error denoted as SE must be calculated. Therefrom, a t-score corresponding to the sample mean or the test statistic can be calculated, denoted as TS.

```
demeanedSum <- NULL
for(i in 1:length(y)){
  demeanedSum[i] <- y[i] - meanIQ
}
squaredError <- demeanedSum^2
variance <- sum(squaredError)/(length(y) - 1)
sdIQ <- sqrt(variance)
SE = sdIQ/sqrt(length(y))
meanIQ <- sum(y)/length(y)
TS = (meanIQ - 100)/SE
TS
```

The t-score clculated using this method was -0.59574. The corresponding p-value was provided using this test statistic. The R function pt was able to do so, and because the p-value corresponds to the probability of receiving a test statistic as extreme as or more extreme, in this case positive, than the t-score calculated, the p-value was calculated for greater than the tail or calculating the lower tail was set to false. This p-value was confirmed by running the R function t.test as well.

```
pvalue <- pt(TS, df = 24, lower.tail=FALSE)
pvalue
t.test(y, mu = 100, alternative = "greater")
```

Because the p-value proved to be 0.7215, and this was greater than the predetermined significance level of 0.05, the t-test leads to the conclusion to fail to reject the null hypothesis. Therefore, it is 72.15% likely that this mean was obtained if the null hypothesis was true. Essentially we could not conclude it was unlikely if the true mean IQ score of the counselor's school's students was less than or equal to 100, the national average IQ score.

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:

```
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)
y <- as.character(y)
y = recode(y, "1" = "Freshman","2" = "Sophomore", "3" = "Junior", "4" =
   Senior")
y
```

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)
```

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

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

- C) 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.

A) The plots for the relationships between Y, X1, X2, and X3 are shown in Figures 1-6 at the end of this document. The code used to produce such plots and to conduct linear regressions on these plots are provided below.

```
library(readxl)
lmYX1 = lm(Y~X1, data = expenditure)
summary(lmYX1) #significant positive slope
lmYX2 = lm(Y~X2, data = expenditure)
```

```
summary(lmYX2)
lmYX3 = lm(Y~X3, data = expenditure)
summary(lmYX3)
lmX1X2 = lm(X1~X2, data = expenditure)
summary(lmX1X2) #significant negative slope
lmX1X3 = lm(X1~X3, data = expenditure)
summary(lmX1X3) #significant positive slope
lmX3X2 = lm(X3~X2, data = expenditure)
summary(lmX3X2) #significant negative slope

library(gridExtra)
library(ggpubr)
X1Y <- ggplot(expenditure, aes(x=X1, y=Y)) + geom_point() + geom_smooth(method
    = "lm") +
  ggtitle("Figure 1: Y by X1") +
  labs(x="Per Capita Personal Income", y="Per Capita Expenditure on Public
    Education")
X2Y <- ggplot(expenditure, aes(x=X2, y=Y)) + geom_point()+ geom_smooth(method
    = "lm") +
  ggtitle("Figure 2: Y by X2") +
  labs(x="Number of Residents per 1000 under 18 Years of Age", y="Per Capita
    Expenditure on Public Education")
X3Y <- ggplot(expenditure, aes(x=X3, y=Y)) + geom_point()+ geom_smooth(method
    = "lm") +
  ggtitle("Figure 3: Y by X3") +
  labs(x="Number of People per 1000 Residing in Urban Areas", y="Per Capita
    Expenditure on Public Education")
X1X2 <- ggplot(expenditure, aes(x=X1, y=X2)) + geom_point()+ geom_smooth(
    method = "lm") +
  ggtitle("Figure 4: X2 by X1") +
  labs(x="Per Capita Personal Income", y="Number of Residents per 1000 under
    18 Years of Age")
X1X3 <- ggplot(expenditure, aes(x=X1, y=X3)) + geom_point()+ geom_smooth(
    method = "lm") +
  ggtitle("Figure 5: X3 by X1") +
  labs(x="Per Capita Personal Income", y="Number of People per 1000 Residing
    in Urban Areas")
X2X3 <- ggplot(expenditure, aes(x=X2, y=X3)) + geom_point()+ geom_smooth(
    method = "lm") +
  ggtitle("Figure 6: X3 by X2") +
  labs(x="Number of Residents per 1000 under 18 Years of Age", y="Number of
    People per 1000 Residing in Urban Areas")
ggarrange(X1Y, X2Y, X3Y, X1X2, X1X3, X2X3,  ncol = 2, nrow = 3)
```

The results show that there is a moderately positive correlation between per capita personal income (X1) and per capita expenditure on public education (Y) due to an R squared value of 0.41, and the plot's slope of 0.034 is statistically significantly positive. In contrast, there is virtually no correlation between number of residents per thousand under eighteen years of age (X2) and per capita expenditure on public education (Y) with an R squared value of 0.02,

and the plot's slope of -0.08 is not statistically significantly different from zero. Similarly, there is virtually no correlation between number of people per thousand residing in urban areas (X3) and per capita expenditure on public education (Y) with an R squared value of 0.04, and the plot's slope of 0.04 is not statistically significantly different from zero.

Within the explanatory variables, there is some important correlations however. Between per capita personal income (X1) and number of residents per thousand under eighteen years of age (X2), there is a weakly negative correlation with an R squared value of 0.27, and the plot's slope of -3.9 is statistically significantly negative. In addition, between per capita personal income (X1) and number of people per thousand residing in urban areas (X3), there is a moderately positive correlation with an R squared value of 0.34, and the slope of 1.64 is statistically significantly positive. Furthermore, between number of residents per thousand under eighteen years of age (X2) and number of people per thousand residing in urban areas (X3), there is a weakly negative correlation with an R squared value of 0.12, and the slope of -.98 is statistically significantly negative.

B) The plot for the relationship between Y and Region is shown in Figure 7. The work to create this plot is shown in the code below.

```
expenditure$REGION <- as.character(expenditure$Region)
expenditure$REGION <- recode(expenditure$REGION, "1" = "Northeast", "2" = "
    North Central", "3" = "South", "4" = "West")
ggplot(expenditure, aes(x=REGION, y=Y)) +
  geom_bar(stat="identity") +
  ggtitle("Figure 7: Y by Region") +
  labs(x="Region", y="Per Capita Expenditure on Public Education")
```

Figure 7 shows that the West has the highest per capita expenditure on public education overall, adding each state's per capita expenditure on public education within each region. Similarly, after averaging the per capita expenditure on public education for each state within each region, this group-by/summarize function also confirmed that the West on average have the highest per capita expenditure on public education.

```
group_by(expenditure, REGION) %>%
  summarize(meanY = mean(as.numeric(Y))) %>%
  arrange(desc(meanY))
```

C) The plot for the relationship between Y and X1 is shown in Figure 1. This graph between per capita personal income and per capita expenditure on public education shows a moderately positive correlation with an R squared value of 0.41 and statistically significantly positive slope of 0.034.

```
ggplot(expenditure, aes(x=X1, y=Y)) +
```

```
  geom_point(aes(shape=Region, color=Region)) +
  ggtitle("Figure 8: Y by X1 and by Region") +
  labs(x="Per Capita Personal Income", y="Per Capita Expenditure on Public
    Education")
```

This graph can be reproduced to include the variable Region, which is displayed in Figure 8. Region is shown by color and symbol.

Figure 1: Y by X1

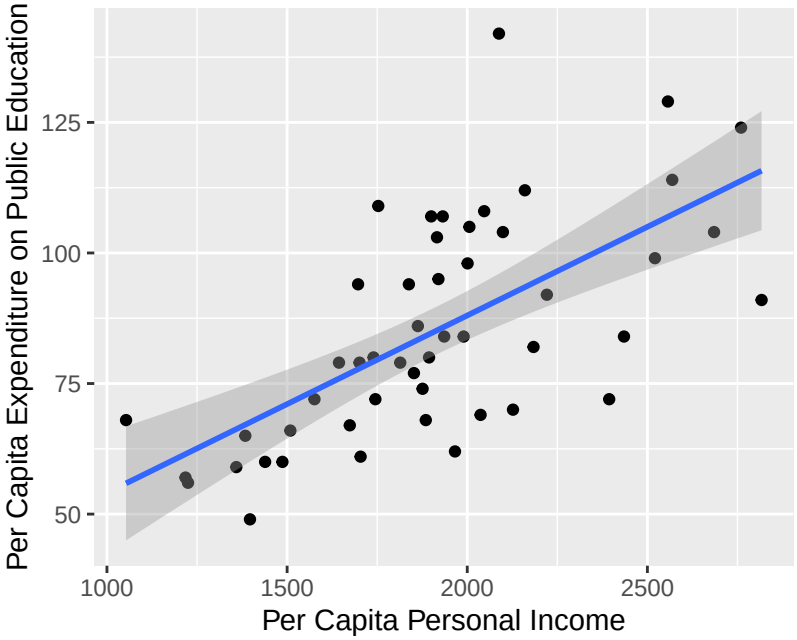

Figure 2: Y by X2

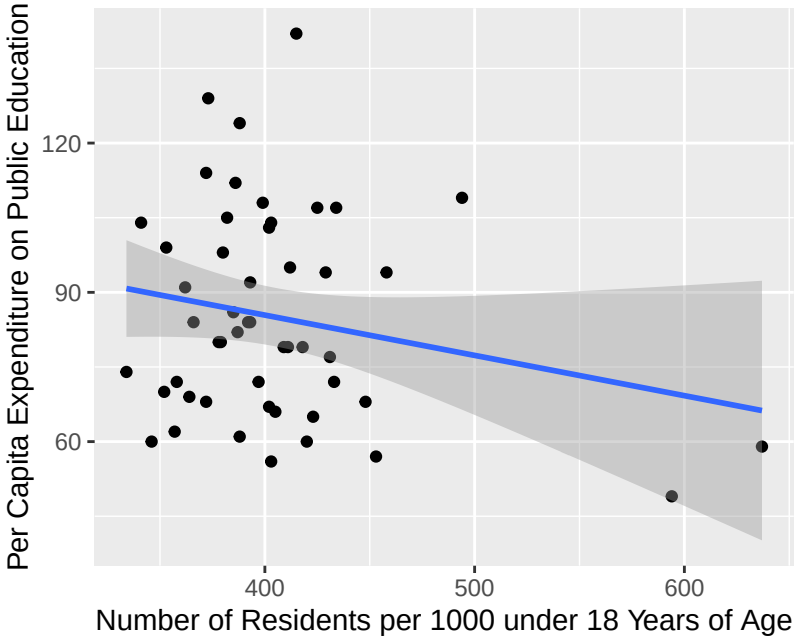

Figure 3: Y by X3

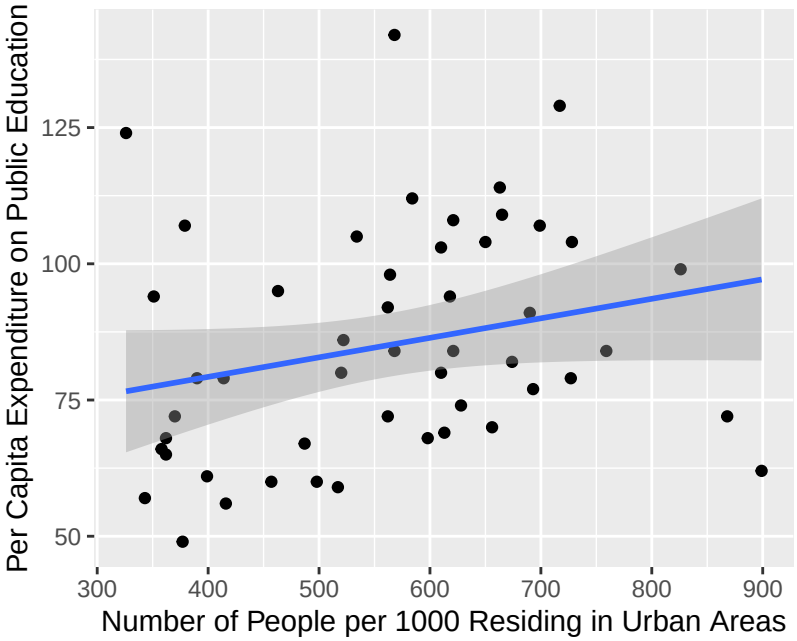

Figure 4: X2 by X1

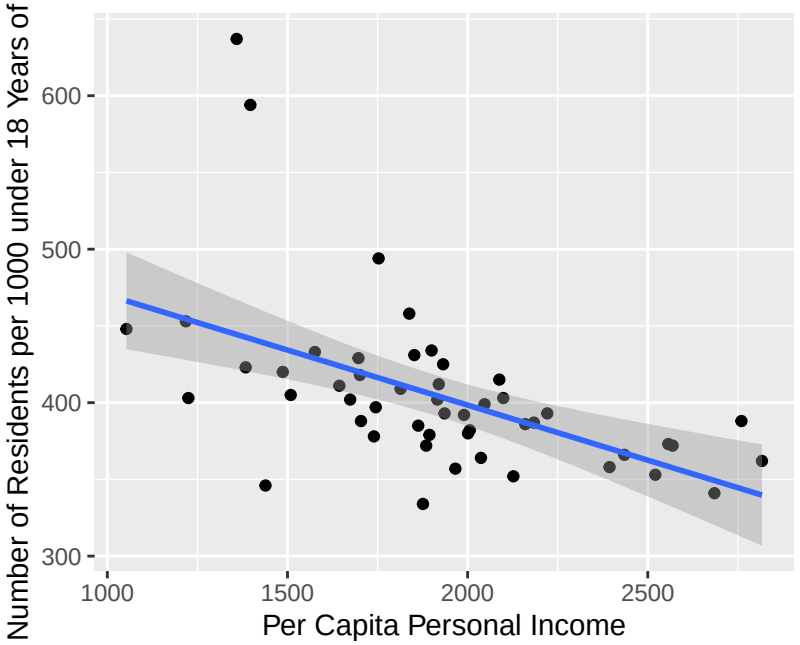

Figure 5: X3 by X1

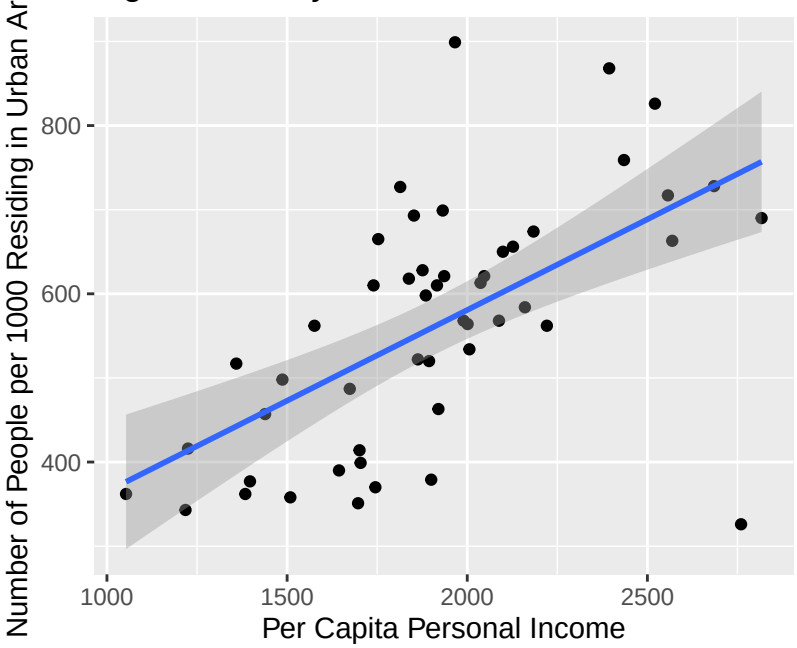

Figure 6: X3 by X2

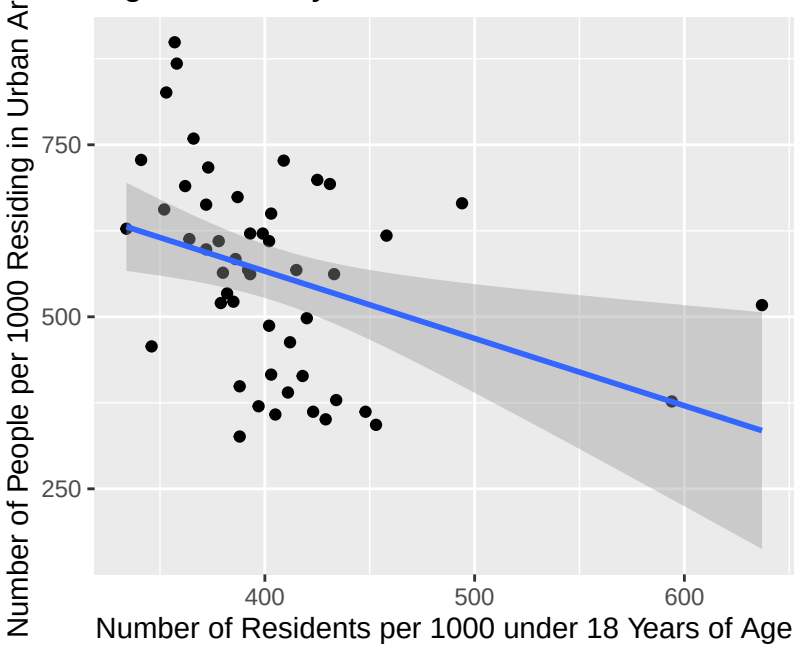

Figure 7: Y by Region

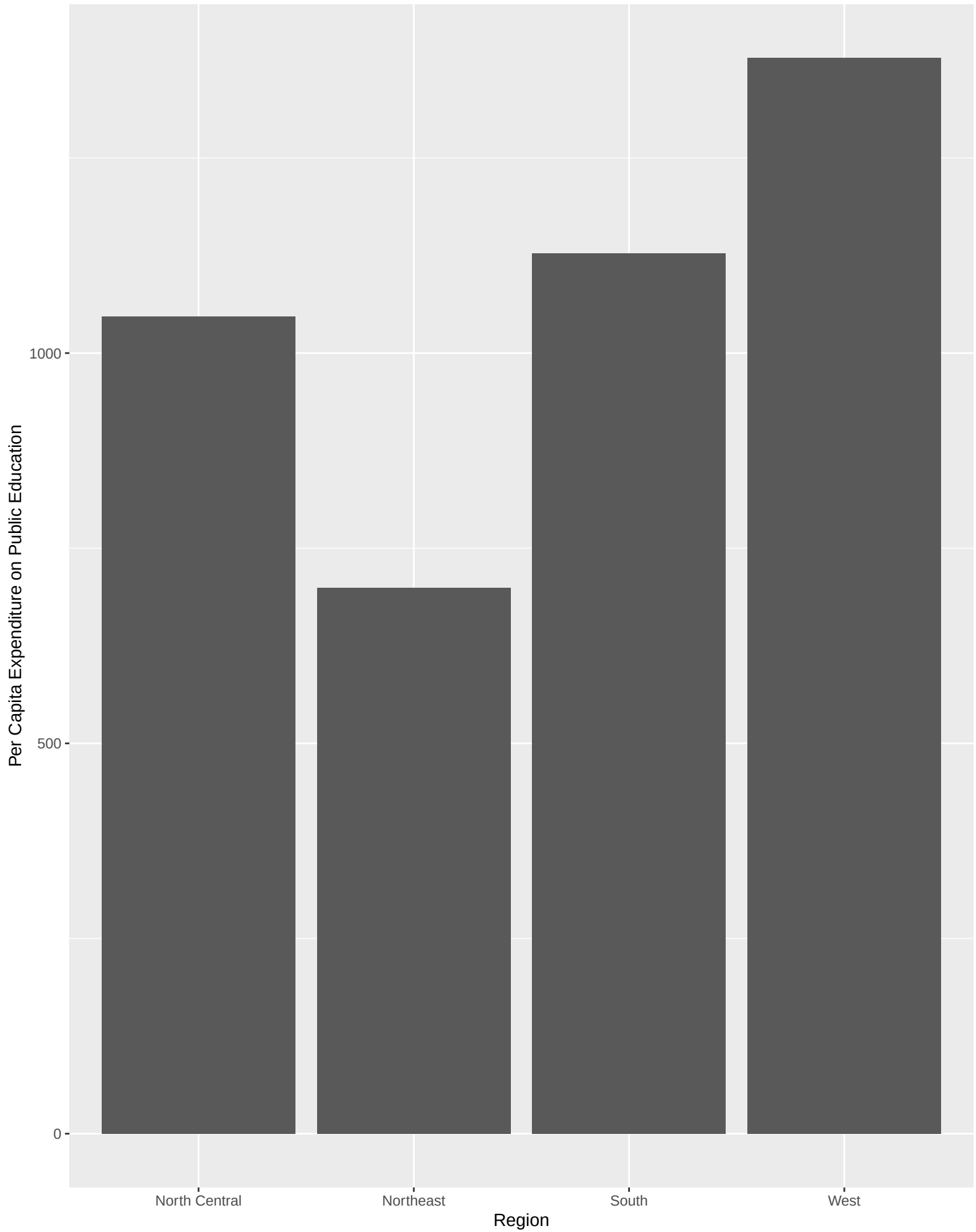

Figure 8: Y by X1 and by Region