1 PS01 Answers

Aryan Goyal
Student number: 18306046

2 Question 1

2.1 Answer 1.1

In order to find the 90 percent confidence interval, I calculate some descriptive statistics:

```
mean(y) #central tendency, mean
sd(y)   #Standard deviation to help gauge the spread of the data
length(y) #to calculate the sample size
```

Using these functions, I find that the mean = 98.44, the standard deviation = 13.09 and the sample size (n) = 25
These values are used to calculate the standard error. The standard error describes how the mean varies from sample to sample

```
standard_error <- sd(y)/sqrt(25) #25 is the sample size
```

From the above values, we can calculate the 90 percent confidence interval
The confidence interval formula is: mean +- (t-score * (standard error))

```
t_score <- qt(0.95, df=length(y)-1)
lower_90_t <- mean(y)-(t_score)*(standard_error)
upper_90_t <- mean(y)+(t_score) * (standard_error)
```

From this, I got the following confidence intervals:

Lower interval:93.9599
Upper interval: 102.9201

Hence, the average student IQ was 98.44, 90 percent CI[93.9599,102.9201]

2.2 Answer 1.2

Based on the question, we have to decide the appropriate hypothesis test. Due to the small sample size, I have chosen the t-test.

My hypotheses:
Null hypothesis: The mean IQ in her school is less than/equal to 100
Alternative hypothesis: the mean IQ in her school is greater than 100

I conduct a one tailed t test as we are interested in finding out if our mean IQ is greater than 100.
To calculate this in R:

```
test_statistic <- (mean(y)-100)/(sd(y)/sqrt(length(y)))

P_value <- pt((test_statistic), df = 24, lower.tail = TRUE)
```

In order to double check, I use the t-test function in R:

```
t.test(y, mu = 100, alternative = 'less')
```

From both of these methods, I find that the p-value = 0.2785

This value is greater than our significance level = 0.05 and therefore, we do not have sufficient evidence to reject the null hypothesis.

We cannot reject the null hypothesis that the mean IQ in her school is less than/equal to 100

3 Question 2

3.1 Answer 2.1

I individually plot each relationship among Y, X1, X2 and X3 Next, I display the code from the first 3 graphs from R. The same code was used with the other variables for each graph:

```
plot(expenditure$Y,expenditure$X1,
     xlab="Per capita expenditure on shelters/housing assistance in state",
     ylab="Per capita personal income in state",
     main="The Relationship between expenditure on shelters and per capital
     personal income")
abline(lm(expenditure$X1 ~ expenditure$Y),col='blue',lty='dashed')
cor(expenditure$Y,expenditure$X1)
text(50, 2500,sprintf("Correlation=%s",
    round(cor(expenditure$X1,expenditure$Y),4)))

plot(expenditure$Y,expenditure$X2,
     xlab="Per capita expenditure on shelters/housing assistance in state",
     ylab=""Financially insecure" residents in state per 100,000",
     main="The Relationship between expenditure
     on shelters and No. of residents that are financially insecure")
abline(lm(expenditure$X2 ~ expenditure$Y),col='blue',lty='dashed')
cor(expenditure$Y,expenditure$X2)
text(50, 450,sprintf("Correlation=%s",
    round(cor(expenditure$Y,expenditure$X2),4)))

plot(expenditure$Y,expenditure$X3,
     xlab="Per capita expenditure on shelters/housing assistance in state",
     ylab="No. of people per 1000 residing in urban areas in state",
     main="Relationship between expenditure on shelters
    and no of people per 1000 residing in urban areas")
abline(lm(expenditure$X3 ~ expenditure$Y),col='blue',lty='dashed')
cor(expenditure$Y,expenditure$X3)
text(50, 800,sprintf("Correlation=%s",
    round(cor(expenditure$Y,expenditure$X3),4)))
```

On the basis of the correlation value and line of best fit, I can say that all variables have a positive weak to moderate correlation.
The strongest positive moderate correlation (0.59) is between X1 (Per capital personal income in state) and X3(Number of people per 1000 residing in urban areas in state)
Whereas, the weakest positive correlation (0.21) is between X1 (Per capita personal income in state) and X2("Financially insecure" residents in state per 1000)

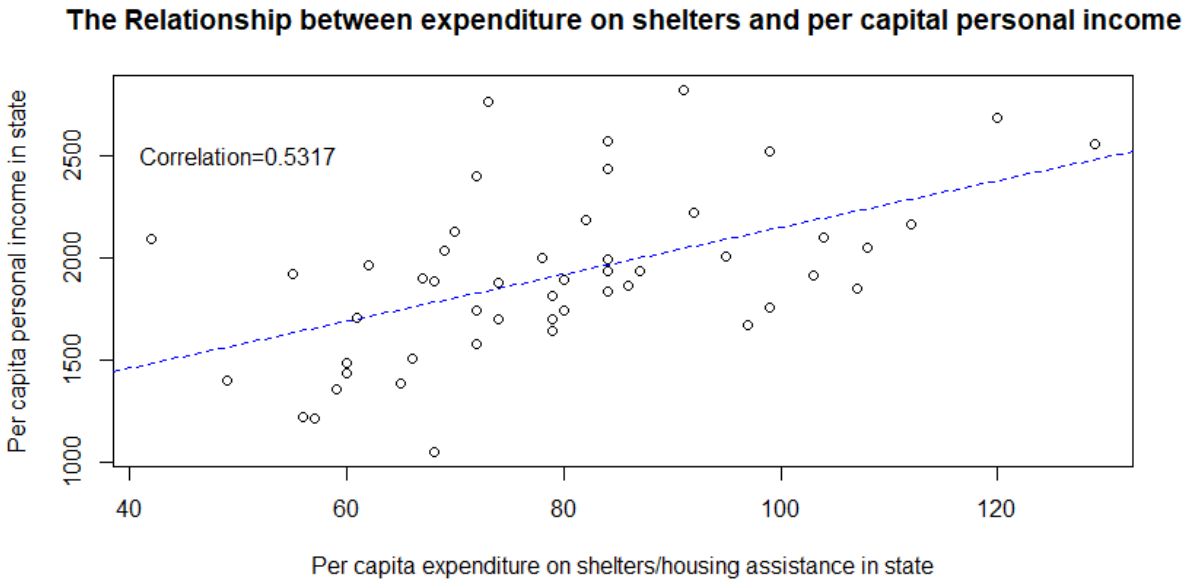

Figure 1: Y and X1

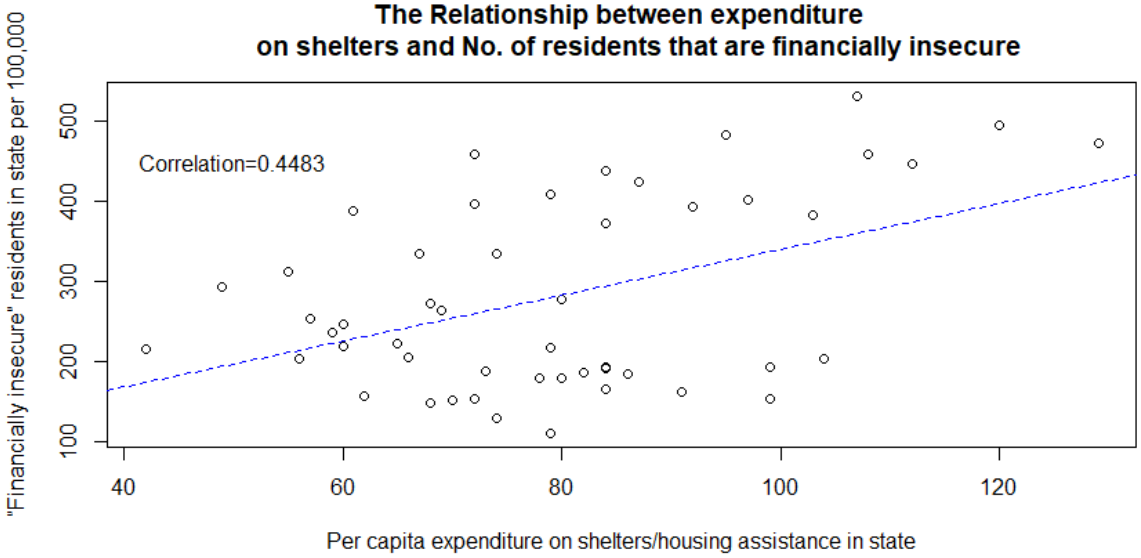

Figure 2: Y and X2

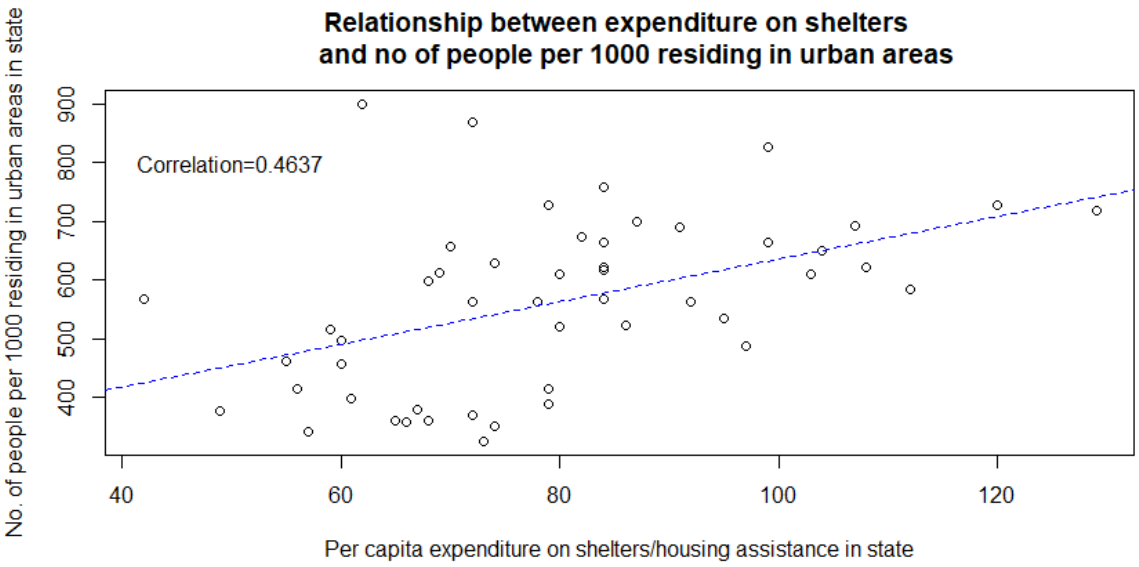

Figure 3: Y and X3

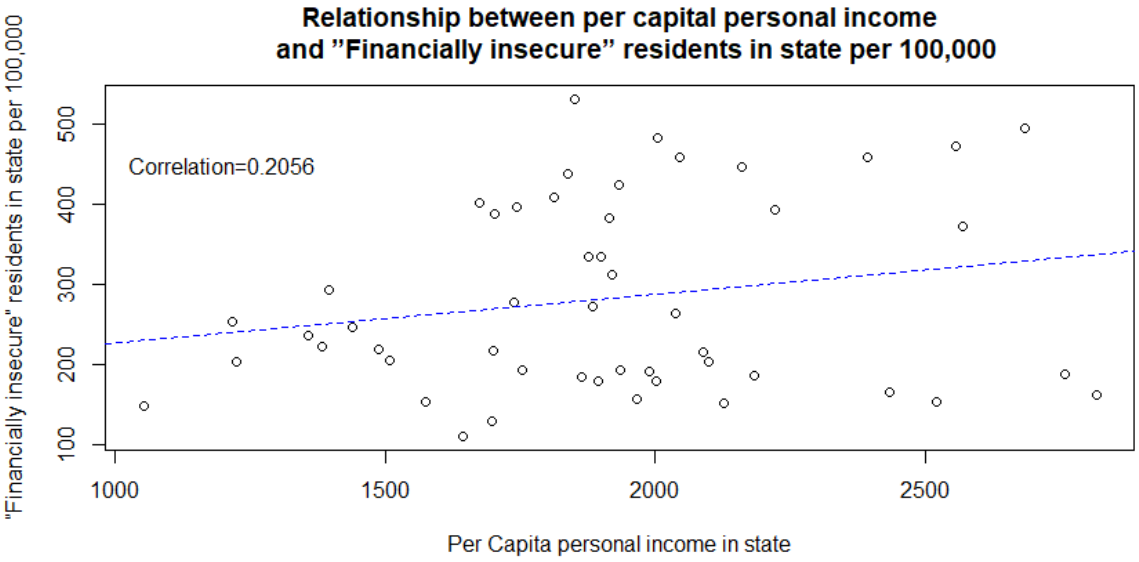

Figure 4: X1 and X2

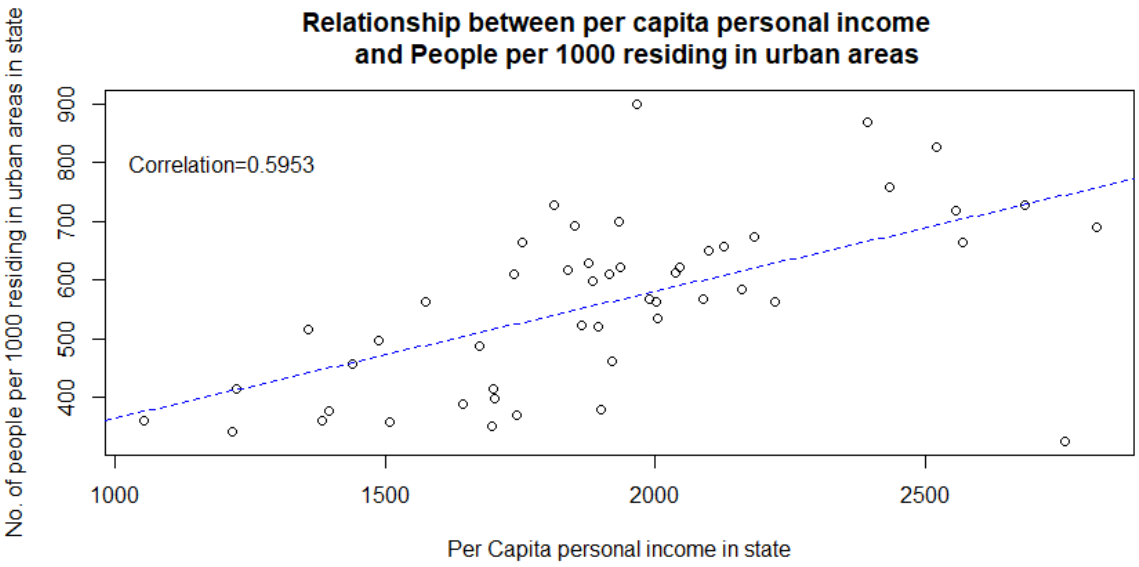

Figure 5: X1 and X3

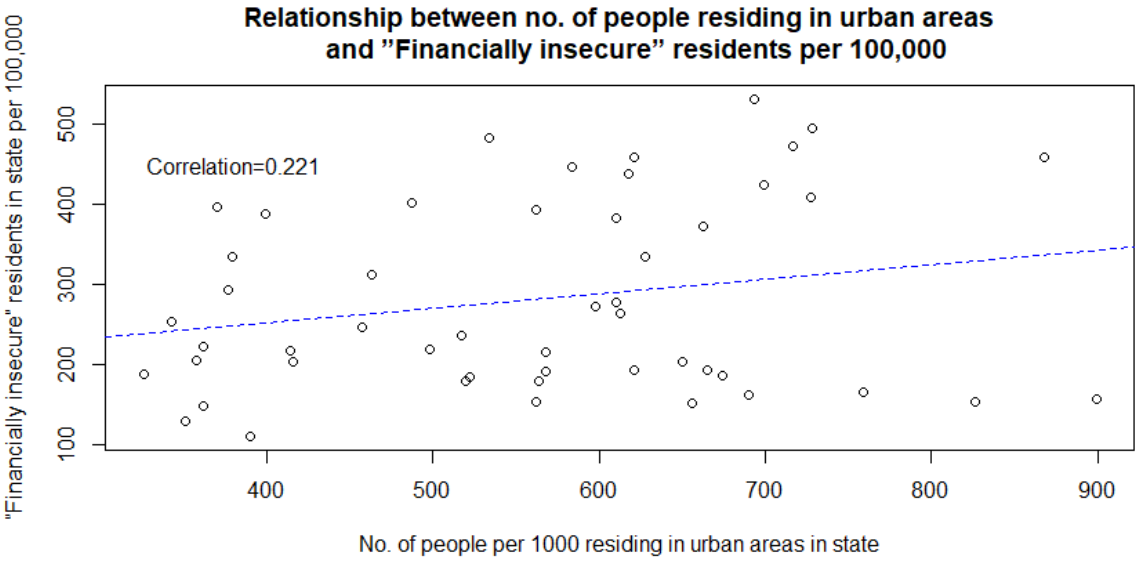

Figure 6: X2 and X3

3.2 Answer 2.2

Next, I plot the relationship between Y(per capita expenditure on shelters/housing assistance in state) and Region. For this, I make a boxplot using this code:

```
    boxplot(expenditure$Y ~ expenditure$Region,
        main="Boxplot of per capita expenditure on shelters
        by region",
        ylab="Per capita expenditure on shelters",
        xlab="Region",
        names=c("Northeast","North Central","South","West"))
means <- tapply(expenditure$Y, expenditure$Region, mean)
points(means, pch=20)
```

The boxplot (Figure 7) is displayed below.
The black dots in each boxplot indicate the mean. On the basis of this, I can say that the West (Region 4) has the highest per capita expenditure on housing assistance on average

I also used this code to calculate the mean of the four regions:

```
    aggregate(expenditure$Y, list(expenditure$Region), FUN=mean)
```

This reiterated the same finding that the West had the highest per capita expenditure on housing assistance on average at 88.3.

Region	Mean
Northeast	79.44
North-Central	83.92
South	69.19
West	88.31

Table 1: Mean of each region

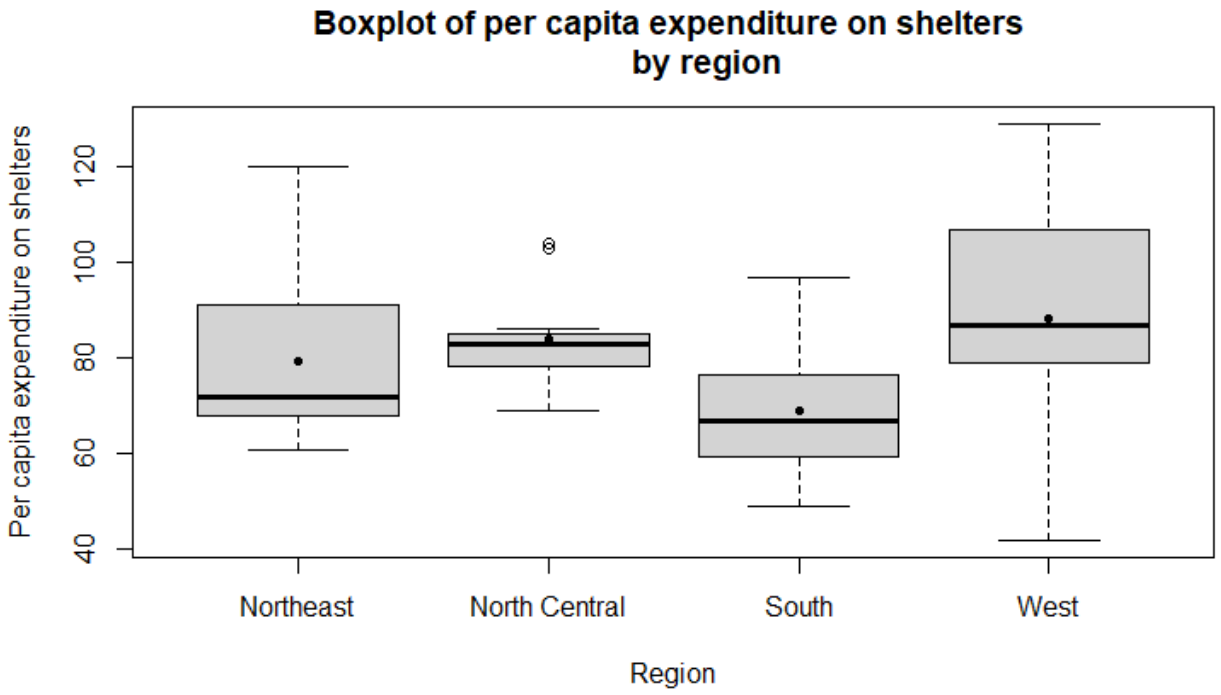

Figure 7: Y and Region

3.3 Answer 2.3

First, I plot the relationship between Y and X1.
From the graph below, it is possible to say that there is a moderate positive correlation (0.53) between per capita expenditure on shelters and per capital personal income in the state.

Using this code in R, I added region to the above graph:

```
 library(car)
?scatterplot
scatterplot(expenditure$X1 ~ expenditure$Y|
    expenditure$Region,
    regLine=TRUE,smooth=FALSE, grid=FALSE,
    legend=c(title="Region",coords="topleft"),
    main="The relationship between per capita personal income
    and per capita expenditure on shelters by region",
    xlab="Per capita personal income in state",
    ylab="Per capita expenditure on shelters in state")
```

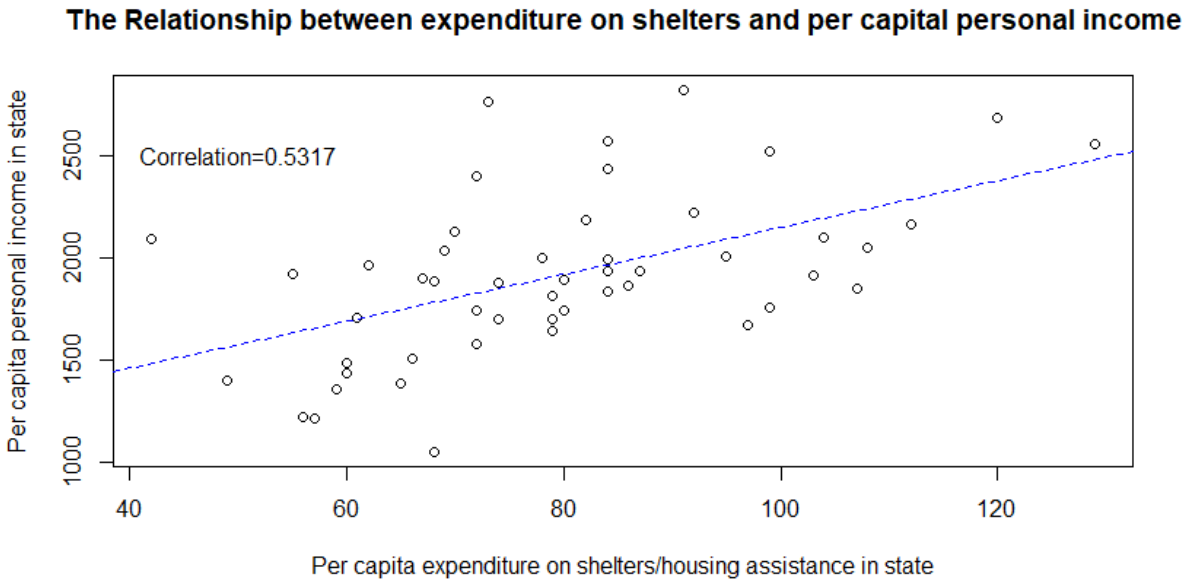

Figure 8: Y and X1

On the basis of the graph below, it is possible to make some inferences.

First, Region 1 (Northeast) has the highest per capita expenditure on shelters. Whereas, Region 3 (South) has the lowest per capita expenditure on shelters on average.

On the basis of the line of best fit, we can say that there is a moderate positive correlation between per capita personal income and per capita expenditure on shelters in Region 1(Northeast) and 3(South). Whereas, there is a weak positive correlation for Region 2(North Central) and 4(West). The per capital personal income has a large spread for Region 4 (West) which could be a possible explanation for why there is a weak correlation to expenditure on shelters.

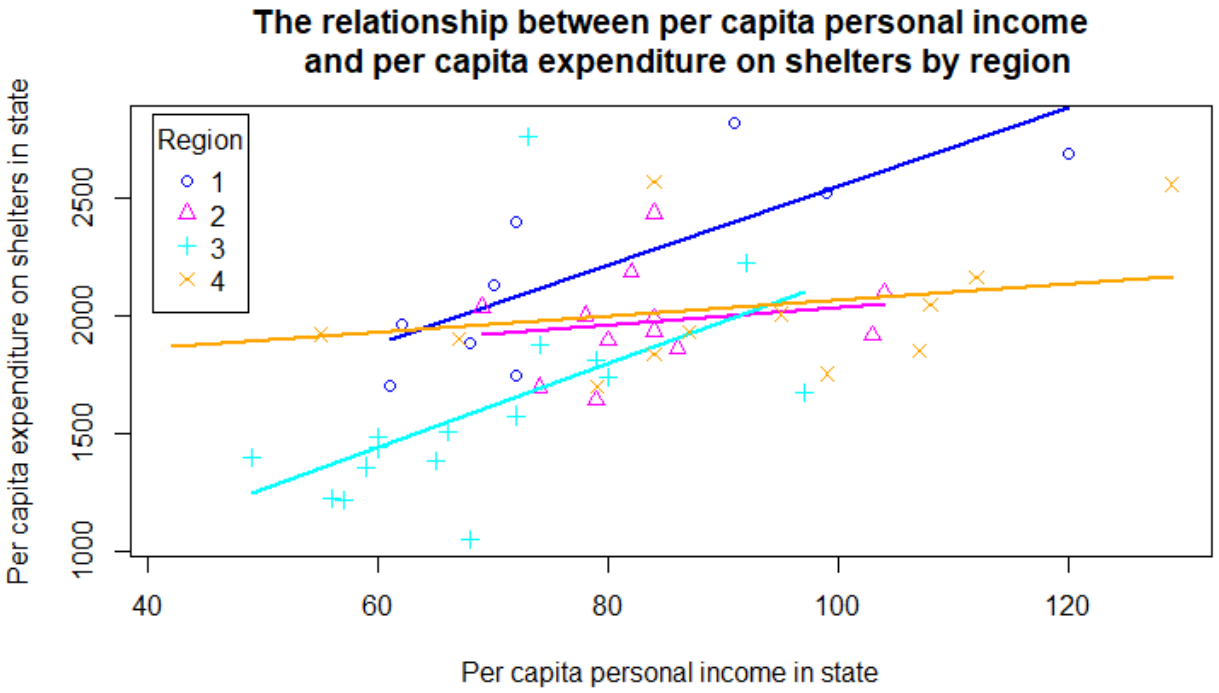

Figure 9: Y and X1 by region