

# Problem Set 1 Applied Stats/Quant Methods 1

Zengyuan Zhao/zhaoze@tcd.ie

Due: September 30, 2024

## 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 GitHub.
- This problem set is due before 23:59 on Monday September 30, 2024. No late assignments will be accepted.

## Question 1: Education

A 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:

```
1 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)
```

1. Find a 90% confidence interval for the average student IQ in the school.

```
1 #import data
2 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
      )
3 #Calculating the confidence interval usually requires taking into account
   the mean, standard deviation, standard error, and critical value of
   the sample, so the above results must be obtained first
```

```

4 #Calculate sample mean
5 y_mean <- mean(y)
6 #Calculate standard deviation
7 y_sd <- sd(y)
8 #Calculate sample size
9 n <- length(y)
10 #The critical value can be found by looking up the statistical table or
    using code
11 a1 <- 0.1
12 t1_quantile <- qt(1-a1/2, df = n - 1)
13 #Calculate the error
14 error <- t1_quantile*(y_sd/sqrt(n))
15 #Construct the Confidence Interval
16 lower_bound <- y_mean - error
17 upper_bound <- y_mean + error
18 #the consequence is (lower_bound, upper_bound)
19 paste("The confidence interval is: (", lower_bound, ", ", upper_bound, ")
    ", sep="")

```

"The confidence interval is: (93.9599275120757, 102.920072487924)"

2. Next, the school counselor was curious whether the average student IQ in her school is higher than the average IQ score (100) among all the schools in the country. Using the same sample, conduct the appropriate hypothesis test with  $\alpha = 0.05$ .

```

1 #The mean and standard deviation have been calculated in the previous
    question
2 #national_data <- 100
3 national_data <- 100
4 #Calculate the t-statistic
5 t_statistic <- (y_mean - national_data) / (y_sd / sqrt(n))
6 t_statistic
7 #Because this test already has a clear hypothesis about one direction of
    the population parameter, and only has one tail to look for extreme
    values, it is therefore a one-tailed test.
8 #This is often used to test whether the mean is significantly greater
    than some hypothesized value
9 a2 <- 0.05
10 t2_quantile <- qt(1-a2, df=n-1, lower.tail = TRUE)
11 t2_quantile
12 if(t_statistic > t2_quantile){
13   print('The average IQ of students in this school is higher than the
    national average.')
14 }else{
15   print('The average IQ of students in this school is not higher than the
    national average.')
16 }

```

”The average IQ of students in this school is not higher than the national average.”

## Question 2: Political Economy

Researchers are curious about what affects the amount of money communities spend on addressing homelessness. The following variables constitute our data set about social welfare expenditures in the USA.

State	50 states in US
Y	per capita expenditure on shelters/housing assistance in state
X1	per capita personal income in state
X2	Number of residents per 100,000 that are "financially insecure" in state
X3	Number of people per thousand residing in urban areas in state
Region	1=Northeast, 2= North Central, 3= South, 4=West

Explore the `expenditure` data set and import data into R.

- 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)?

```
1 #import datasets
2 expenditure <- read.table('/Users/zach/Downloads/expenditure.txt', header
  = TRUE)
3 head(expenditure)
4 #Check and clean data
5 is.na(expenditure)
6 class(expenditure)
7 str(expenditure)
8 #Calculate correlation coefficient
9 expenditure_correlation <- cor(expenditure[,c("Y", "X1", "X2", "X3")])
10 expenditure_correlation
11 library(ggplot2)
12 #Draw scatter plots and regression lines
13 # Draw a scatter plot of Y versus X1
14 Y_vs_X1 <- ggplot(data = expenditure, aes(x = X1, y = Y)) +
15   geom_point() +
16   geom_smooth(method = "lm", se = TRUE, color = "blue") +
17   ggtitle("The scatter plot of Y versus X1") +
18   xlab('per capita personal income in state (X1)') +
19   ylab('per capita expenditure on shelters assistance in state (Y)')
```

```

20 ggsave("Y_vs_X1_scatterplot.pdf", plot = Y_vs_X1, width = 8, height = 6,
      units = "in")
21 #Draw a scatter plot of Y versus X2
22 Y_vs_X2 <- ggplot(data = expenditure, aes(x = X2, y = Y)) +
23   geom_point() +
24   geom_smooth(method = "lm", se = TRUE, color = "yellow") +
25   ggtitle("The scatter plot of Y versus X2") +
26   xlab('Number of residents per 100,000 that are financially insecure in
      state (X2)') +
27   ylab('per capita expenditure on shelters assistance in state (Y)')
28 ggsave("Y_vs_X2_scatterplot.pdf", plot = Y_vs_X2, width = 8, height = 6,
      units = "in")
29 #Draw a scatter plot of Y versus X3
30 Y_vs_X3 <- ggplot(data = expenditure, aes(x = X3, y = Y)) +
31   geom_point() +
32   geom_smooth(method = "lm", se = TRUE, color = "green") +
33   ggtitle("The scatter plot of Y versus X3") +
34   xlab('Number of people per thousand residing in urban areas in state (
      X3)') +
35   ylab('per capita expenditure on shelters assistance in state (Y)')
36 ggsave("Y_vs_X3_scatterplot.pdf", plot = Y_vs_X3, width = 8, height = 6,
      units = "in")
37 #Draw a scatter plot of X1 versus X2
38 X1_vs_X2 <- ggplot(data = expenditure, aes(x = X1, y = X2)) +
39   geom_point() +
40   geom_smooth(method = "lm", se = TRUE, color = "purple") +
41   ggtitle("The scatter plot of X2 versus X1") +
42   xlab('per capita personal income in state (X1)') +
43   ylab('Number of residents per 100,000 that are financially insecure in
      state (X2)')
44 ggsave("X1_vs_X2_scatterplot.pdf", plot = X1_vs_X2, width = 8, height =
      6, units = "in")
45 #Draw a scatter plot of X1 versus X3
46 X1_vs_X3 <- ggplot(data = expenditure, aes(x = X1, y = X3)) +
47   geom_point() +
48   geom_smooth(method = "lm", se = TRUE, color = "orange") +
49   ggtitle("The scatter plot of X3 versus X1") +
50   xlab('per capita personal income in state (X1)') +
51   ylab('Number of people per thousand residing in urban areas in state (
      X3)')
52 ggsave("X1_vs_X3_scatterplot.pdf", plot = X1_vs_X3, width = 8, height =
      6, units = "in")
53 #Draw a scatter plot of X2 versus X3
54 X2_vs_X3 <- ggplot(data = expenditure, aes(x = X2, y = X3)) +
55   geom_point() +
56   geom_smooth(method = "lm", se = TRUE, color = "red") +
57   ggtitle("The scatter plot of X3 versus X2") +
58   xlab('Number of residents per 100,000 that are financially insecure in
      state (X2)') +
59   ylab('Number of people per thousand residing in urban areas in state (
      X3)')

```

```

60 ggsave("X2_vs_X3_scatterplot.pdf", plot = X2_vs_X3, width = 8, height =
    6, units = "in")
61 par(mfrow = c(2, 2))
62 # launch pdf device
63 pdf(file = "scatterplot_matrix.pdf", width = 8, height = 6)
64 pairs(~ X1 + X2 + X3 + Y, data = expenditure, main = "X1,X2,X3,Y scatter
    plot matrix")
65 dev.off()

```

Figure 1: YvsX1 scatterplot.

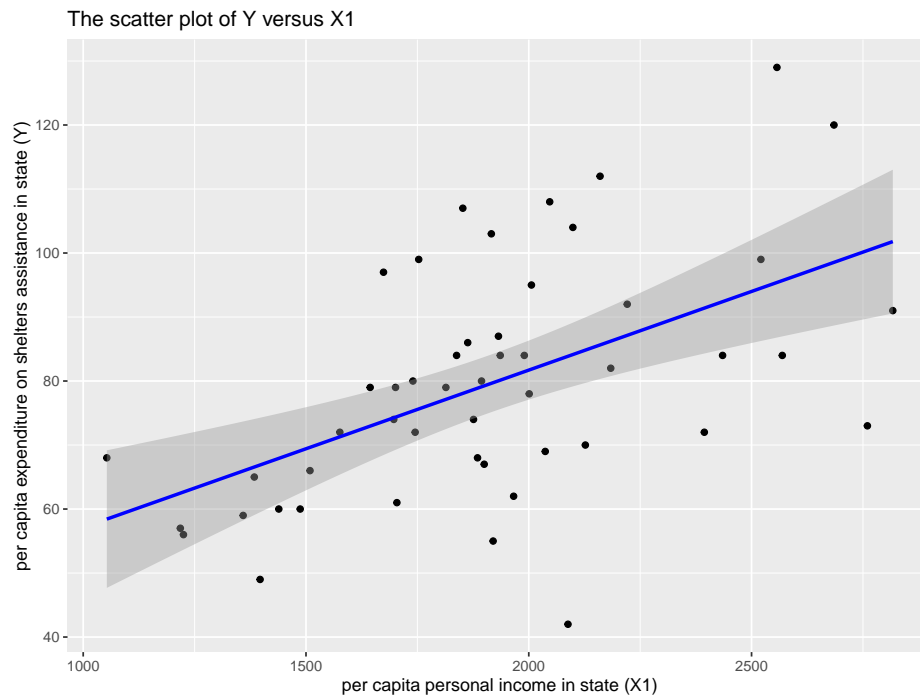


Figure 2: YvsX2 scatterplot.

The scatter plot of Y versus X2

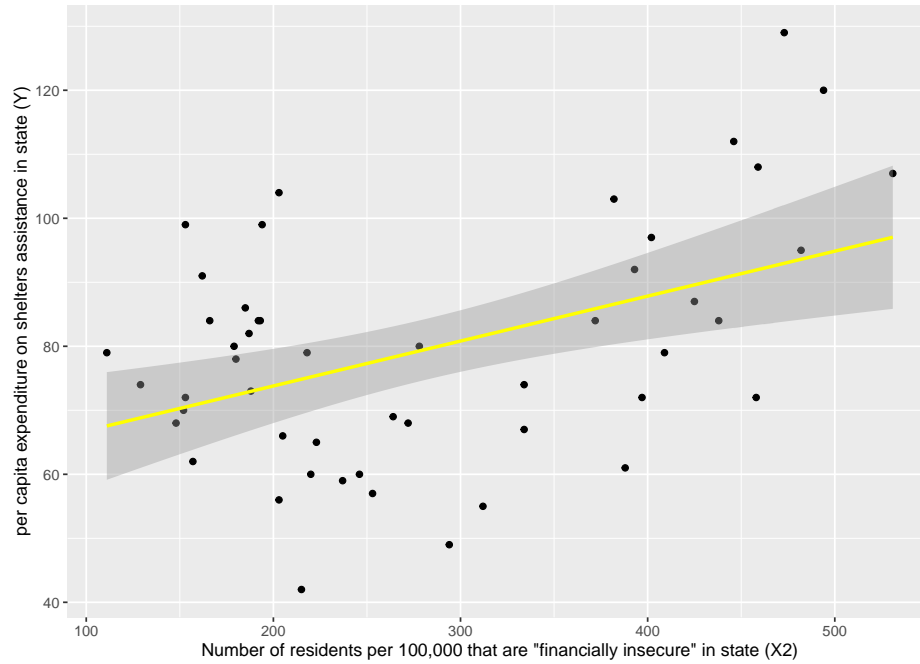


Figure 3: YvsX3 scatterplot.

The scatter plot of Y versus X3

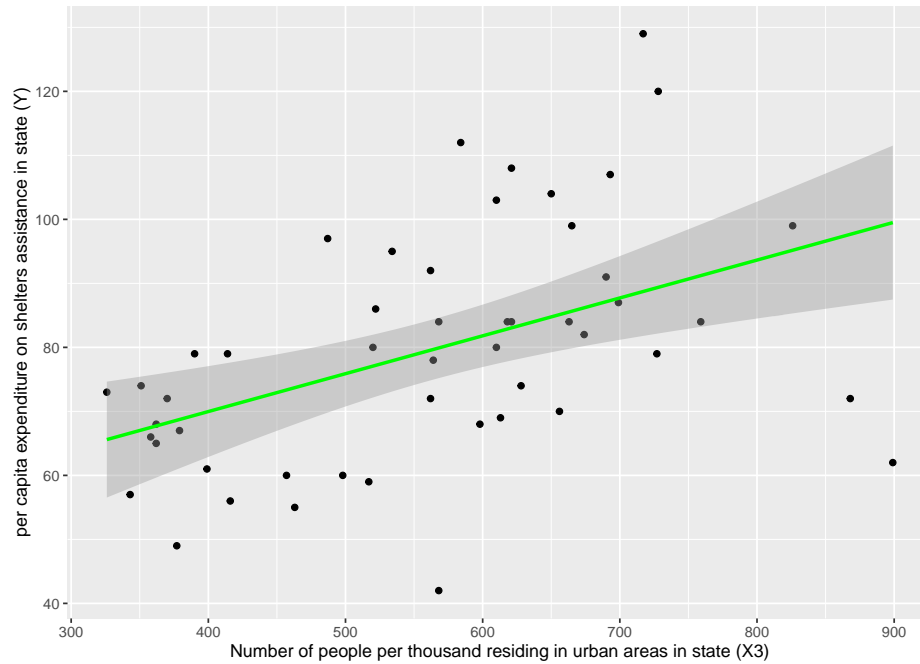


Figure 4: X1vsX2 scatterplot.

The scatter plot of X2 versus X1

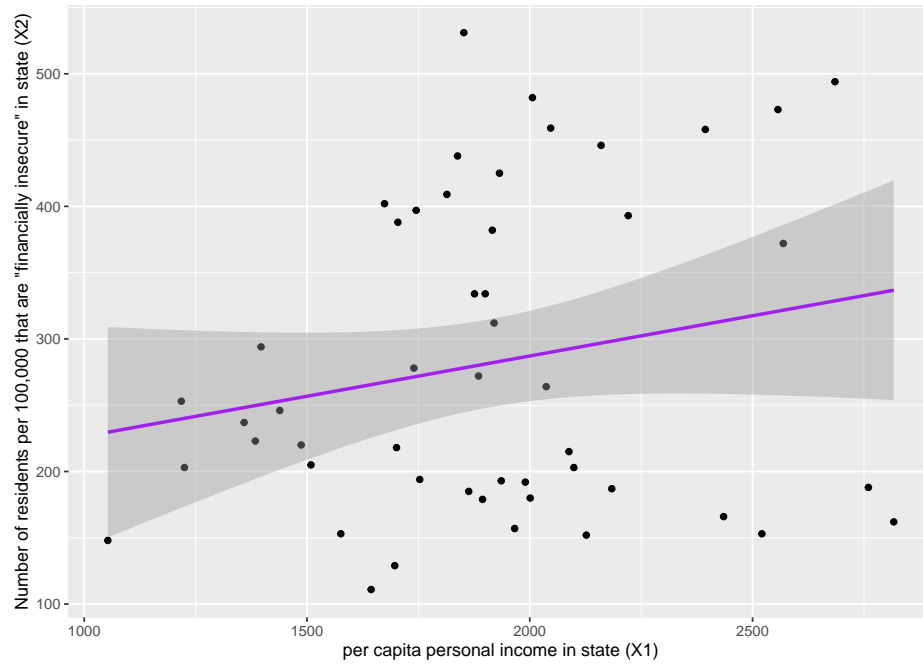


Figure 5: X1vsX3 scatterplot.

The scatter plot of X3 versus X1

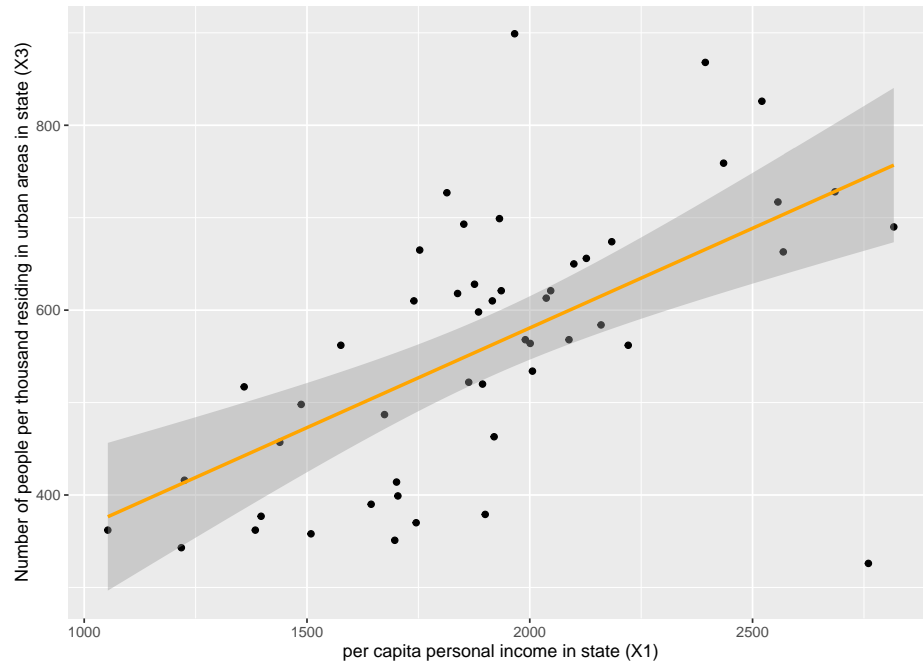


Figure 6: X2vsX3 scatterplot.

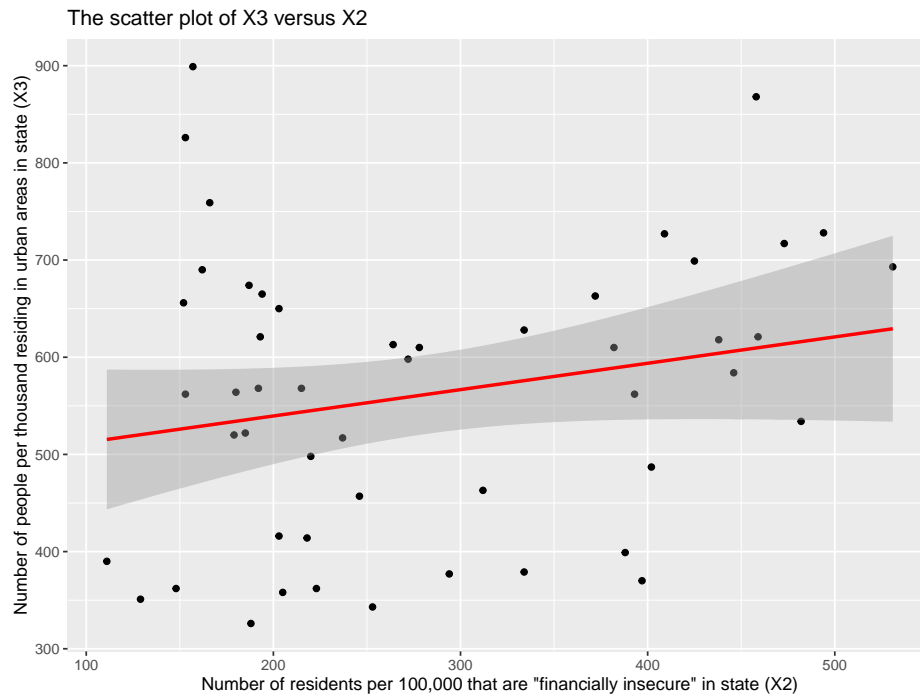
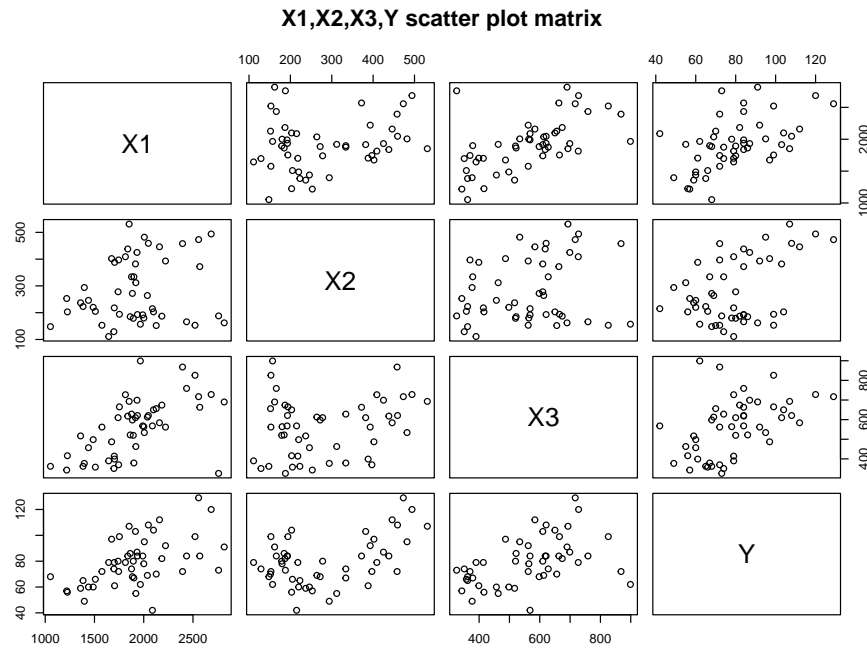


Figure 7: scatterplot matrix





From the above six scatter plots between X1, X2, X3 and Y and the linear model smooth line, we can know that there is a positive correlation between X1, X2, X3 and Y. Among them, there is a strong positive correlation between Y and X1, which shows that the more personal income in each region, the more investment in assisted housing, and vice versa. X1 and X3 also have a strong positive correlation, which shows that the more individuals in each region earn more, the greater the proportion of people living in urban areas. The correlation between X2 and X3 is weak, indicating that the number of "financially insecure" residents per 100,000 people in the state has little to do with the number of people living in urban areas.

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

```

1 #Extract the region column and Y column, and form a new dataset.
2 Y_region <- expenditure[, c("Y", "Region")]
3 Y_region
4 Region1_vs_Y1 <- ggplot(Y_region, aes(x = factor(Region), y = Y, fill =
   factor(Region))) +
5   geom_bar(stat = "identity", position = "dodge") +
6   labs(title = "Per capita housing assistance expenditure by region",
   x = "Region", y = "per capita expenditure on shelters assistance in
   state")
7 ggsave("Region1_vs_Y1_barchart.pdf", plot = Region1_vs_Y1, width = 8,
   height = 6, units = "in")
8 #Regional variables are discrete data. To analyze the relationship
   between discrete data and continuous data, regression analysis is
   usually used.
9 Y_region_correlation <- cor(Y_region[, c('Y', 'Region')])
10 Y_region$Region <- as.factor(Y_region$Region)
11 # use linear regression
12 model <- lm(Y ~ Region, data = Y_region)
13 # check the model
14 summary(model)
15 #The scatter plot of Region and Y
16 Region2_vs_Y2 <- ggplot(data = expenditure, aes(x = Region, y = Y)) +
17   geom_point() +
18   ggtitle("The scatter plot of Region versus Y") +
19   xlab('Region') +
20   ylab('per capita expenditure on shelters assistance in state (Y)')
21 ggsave("Region2_vs_Y2_scatterplot.pdf", plot = Region_vs_Y, width = 8,
   height = 6, units = "in")
22 #Calculate housing support expenditure per capita in each region
23 # Group by Region
24 region_split <- split(Y_region, Y_region$Region)
25 # Calculate the mean of Y in each group
26 Y_averages <- lapply(region_split, function(x) mean(x$Y, na.rm = TRUE))

```

Figure 8: Region2vsY2 scatterplot

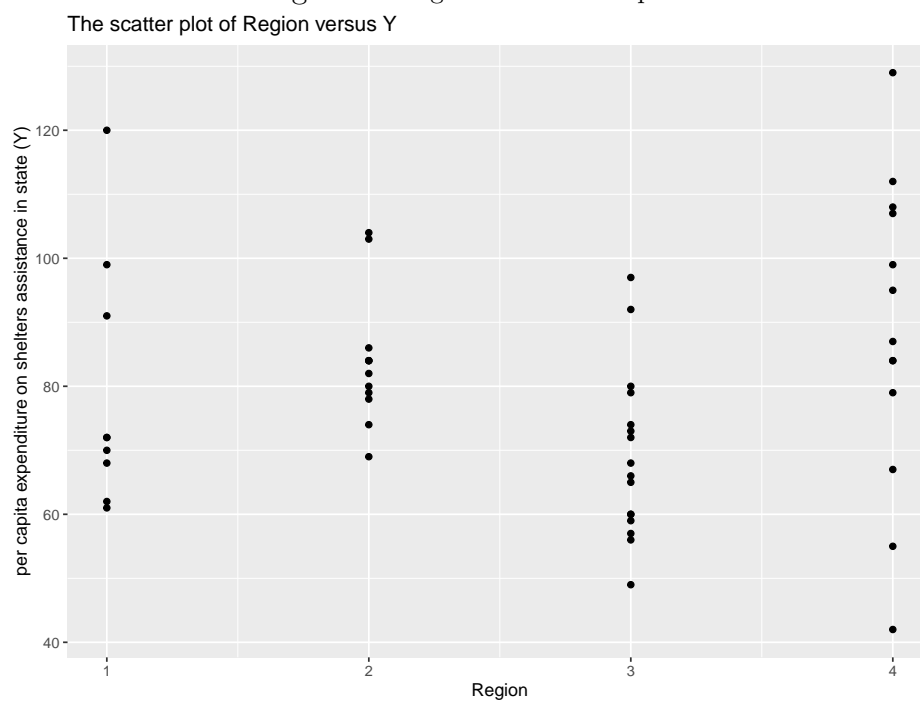


Figure 9: Region2vsY2 scatterplot

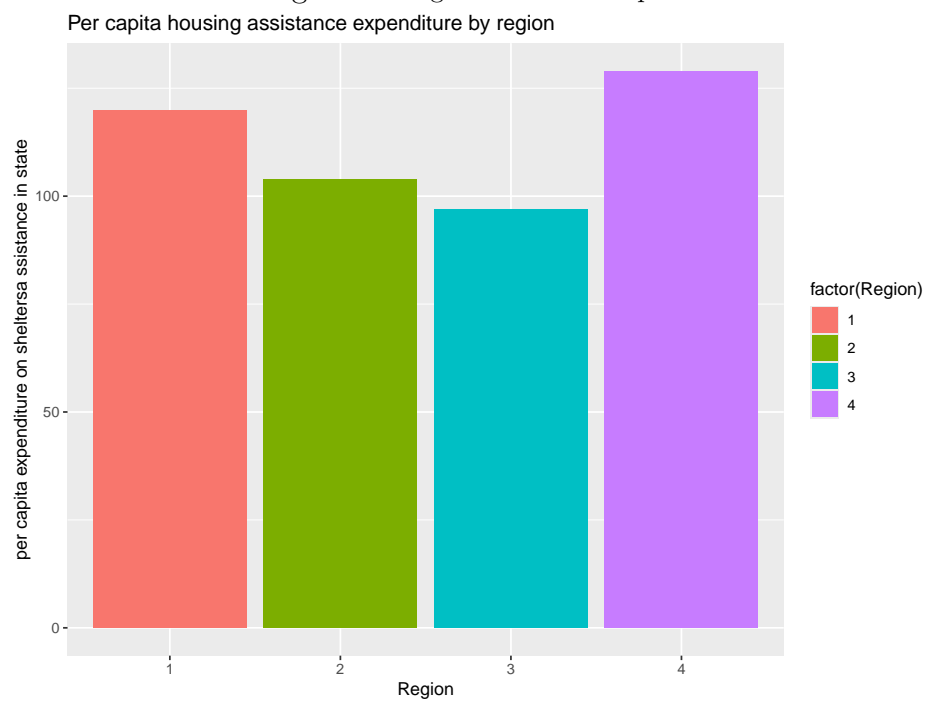


Figure 10: model summary

```
Call:
lm(formula = Y ~ Region, data = Y_region)

Residuals:
    Min       1Q   Median       3Q      Max
-46.308  -9.410  -2.552   10.472   40.692

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  79.444      5.782   13.741 <2e-16 ***
Region2       4.472      7.648    0.585  0.562
Region3     -10.257      7.227   -1.419  0.163
Region4       8.863      7.521    1.178  0.245
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17.34 on 46 degrees of freedom
Multiple R-squared:  0.1754, Adjusted R-squared:  0.1216
F-statistic: 3.262 on 3 and 46 DF, p-value: 0.02973
```

The p-value of the F statistic is 0.02973, which is less than the commonly used significance level of 0.05, which indicates that the Region variable has a significant impact on Y as a whole. From the scatter plot, the data points in the North Central region are concentrated and may have a strong correlation. On average, the western region has the highest spending on housing support, at about 88.30769.

- 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.

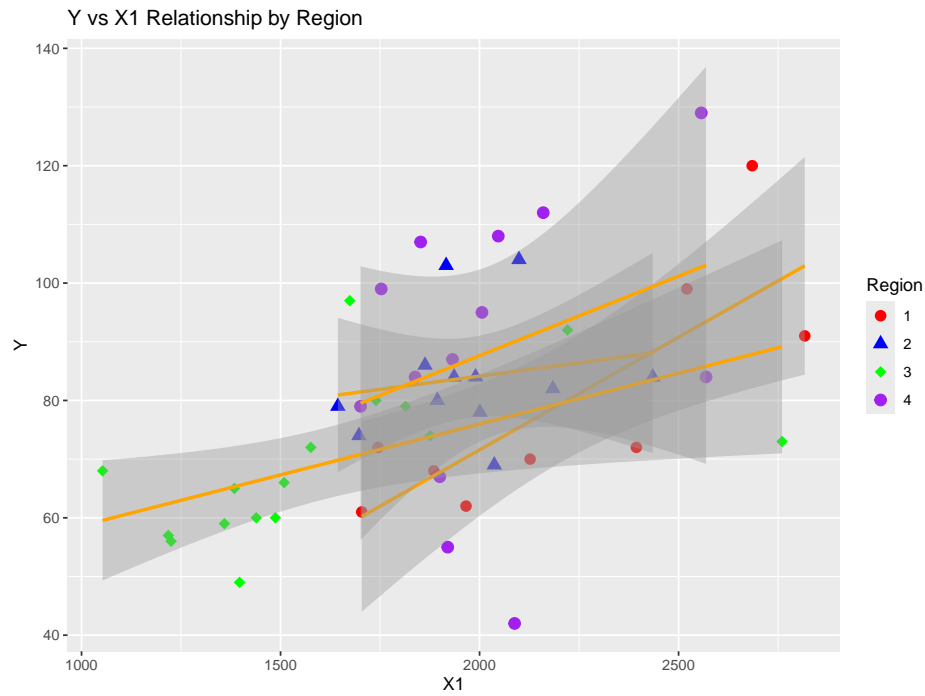
```
1 X1_Y_Region <- expenditure[, c("X1", "Y", "Region")]
2 X1vsY_Region_scatterplot <- ggplot(X1_Y_Region, aes(x = X1, y = Y, color
3   = factor(Region), shape = factor(Region))) +
4   geom_point(size = 3) +
5   geom_smooth(method = "lm", se = TRUE, color = "orange") +
6   scale_color_manual(values = c("1" = "red", "2" = "blue", "3" = "
7   green", "4" = "purple")) + # set color
8   scale_shape_manual(values = c("1" = 16, "2" = 17, "3" = 18, "4" =
9   19)) + # set shape
10  labs(title = "Y vs X1 Relationship by Region",
11       x = "X1",
12       y = "Y",
13       color = "Region",
14       shape = "Region")
```

```

12 ggsave("X1vsY_Region_scatterplot.pdf", plot = X1vsY_Region_scatterplot,
    width = 8, height = 6, units = "in")
13 X1_Y_Region_correlation <- cor(X1_Y_Region[,c("Y", "X1", "Region")])

```

Figure 11: X1vsYvsRegion scatterplot



From this scatter plot, we can see that the per capita income of each region and the expenditure on housing support are positively correlated, and the correlation coefficient is 0.5317212, which is a strong positive correlation. In addition, the correlation between personal income and expenditure on housing support in the Northeast region (Region 1) is stronger than in other regions.