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:

```
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.

```
#import data
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
    )
#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
```

```r
#Calculate sample mean
y_mean <- mean(y)
#Calculate standard deviation
y_sd <- sd(y)
#Calculate sample size
n <- length(y)
#The critical value can be found by looking up the statistical table or using code
a1 <- 0.1
t1_quantile <- qt(1-a1/2, df = n - 1)
#Calculate the error
error <- t1_quantile*(y_sd/sqrt(n))
#Construct the Confidence Interval
lower_bound <- y_mean - error
upper_bound <- y_mean + error
#the consequence is (lower_bound,upper_bound)
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 $\alpha = 0.05$.

```r
#The mean and standard deviation have been calculated in the previous question
#national_data <- 100
national_data <- 100
#Calculate the t-statistic
t_statistic <- (y_mean - national_data) / (y_sd / sqrt(n))
t_statistic
#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.
#This is often used to test whether the mean is significantly greater than some hypothesized value
a2 <- 0.05
t2_quantile <- qt(1-a2, df=n-1, lower.tail = TRUE)
t2_quantile
if(t_statistic > t2_quantile){
  print('The average IQ of students in this school is higher than the national average.')
}else{
  print('The average IQ of students in this school is not higher than the national average.')
}
```

"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)?

```r
#import datasets
expenditure <- read.table('/Users/zach/Downloads/expenditure.txt',header = TRUE)
head(expenditure)
#Check and clean data
is.na(expenditure)
class(expenditure)
str(expenditure)
#Calculate correlation coefficient
expenditure_correlation <- cor(expenditure[,c("Y", "X1", "X2", "X3")])
expenditure_correlation
library(ggplot2)
#Draw scatter plots and regression lines
# Draw a scatter plot of Y versus X1
Y_vs_X1 <- ggplot(data = expenditure, aes(x = X1, y = Y)) +
  geom_point() +
  geom_smooth(method = "lm", se = TRUE, color = "blue") +
  ggtitle("The scatter plot of Y versus X1") +
  xlab('per capita personal income in state (X1)') +
  ylab('per capita expenditure on shelters assistance in state (Y)')+
  theme(plot.title = element_text(hjust = 0.5))
```

```
ggsave("Y_vs_X1_scatterplot.pdf", plot = Y_vs_X1, width = 8, height = 6,
    units = "in")
#Draw a scatter plot of Y versus X2
Y_vs_X2 <- ggplot(data = expenditure, aes(x = X2, y = Y)) +
  geom_point() +
  geom_smooth(method = "lm", se = TRUE, color = "yellow") +
  ggtitle("The scatter plot of Y versus X2") +
  xlab('pNumber of residents per 100,000 that are financially insecure in
     state (X2)') +
  ylab('per capita expenditure on shelters assistance in state (Y)')+
  theme(plot.title = element_text(hjust = 0.5))
ggsave("Y_vs_X2_scatterplot.pdf", plot = Y_vs_X2, width = 8, height = 6,
    units = "in")
#Draw a scatter plot of Y versus X3
Y_vs_X3 <- ggplot(data = expenditure, aes(x = X3, y = Y)) +
  geom_point() +
  geom_smooth(method = "lm", se = TRUE, color = "green") +
  ggtitle("The scatter plot of Y versus X3") +
  xlab('Number of people per thousand residing in urban areas in state (
    X3)') +
  ylab('per capita expenditure on shelters assistance in state (Y)')+
  theme(plot.title = element_text(hjust = 0.5))
ggsave("Y_vs_X3_scatterplot.pdf", plot = Y_vs_X3, width = 8, height = 6,
    units = "in")
#Draw a scatter plot of X1 versus X2
X1_vs_X2 <- ggplot(data = expenditure, aes(x = X1, y = X2)) +
geom_point() +
  geom_smooth(method = "lm", se = TRUE, color = "purple") +
  ggtitle("The scatter plot of X2 versus X1") +
  xlab('per capita personal income in state (X1)') +
  ylab('Number of residents per 100,000 that are financially insecure in
    state (X2)')+
  theme(plot.title = element_text(hjust = 0.5))
ggsave("X1_vs_X2_scatterplot.pdf", plot = X1_vs_X2, width = 8, height =
    6, units = "in")
#Draw a scatter plot of X1 versus X3
X1_vs_X3 <- ggplot(data = expenditure, aes(x = X1, y = X3)) +
  geom_point() +
  geom_smooth(method = "lm", se = TRUE, color = "orange") +
  ggtitle("The scatter plot of X3 versus X1") +
  xlab('per capita personal income in state (X1)') +
  ylab('Number of people per thousand residing in urban areas in state (
    X3)')+
  theme(plot.title = element_text(hjust = 0.5))
ggsave("X1_vs_X3_scatterplot.pdf", plot = X1_vs_X3, width = 8, height =
    6, units = "in")
#Draw a scatter plot of X2 versus X3
X2_vs_X3 <- ggplot(data = expenditure, aes(x = X2, y = X3)) +
  geom_point() +
  geom_smooth(method = "lm", se = TRUE, color = "red") +
  ggtitle("The scatter plot of X3 versus X2") +
```

```
  xlab('Number of residents per 100,000 that are financially insecure in state (X2)') +
  ylab('Number of people per thousand residing in urban areas in state (X3)')+
  theme(plot.title = element_text(hjust = 0.5))
ggsave("X2_vs_X3_scatterplot.pdf", plot = X2_vs_X3, width = 8, height = 6, units = "in")
par(mfrow = c(2, 2))
# launch pdf device
pdf(file = "scatterplot_matrix.pdf", width = 8, height = 6)
pairs(~ X1 + X2 + X3 + Y, data = expenditure, main = "X1,X2,X3,Y scatter plot matrix")
dev.off()
#In order to obtain a more accurate relationship, regression analysis can be performed and the linear regression formula can be obtained
#Regression of X1 and X2
model_X1_X2 <- lm(X1 ~ X2, data = expenditure)
summary(model_X1_X2)
paste0("X1 = ", round(model_X1_X2$coefficients[1],3), " + ", round(model_X1_X2$coefficients["X2"],3), "*X2")
#Regression of X1 and X3
model_X1_X3 <- lm(X1 ~ X3, data = expenditure)
summary(model_X1_X3)
paste0("X1 = ", round(model_X1_X3$coefficients[1],3), " + ", round(model_X1_X3$coefficients["X3"],3), "*X3")
#Regression of X2 and X3
model_X2_X3 <- lm(X2 ~ X3, data = expenditure)
summary(model_X2_X3)
paste0("X2 = ", round(model_X2_X3$coefficients[1],3), " + ", round(model_X2_X3$coefficients["X3"],3), "*X3")
#Regression of Y and X1
model_Y_X1 <- lm(Y ~ X1, data = expenditure)
summary(model_Y_X1)
paste0("Y = ", round(model_Y_X1$coefficients[1],3), " + ", round(model_Y_X1$coefficients["X1"],3), "*X1")
#Regression of Y and X2
model_Y_X2 <- lm(Y ~ X2, data = expenditure)
summary(model_Y_X2)
paste0("Y = ", round(model_Y_X2$coefficients[1],3), " + ", round(model_Y_X2$coefficients["X2"],3), "*X2")
#Regression of Y and X3
model_Y_X3 <- lm(Y ~ X3, data = expenditure)
summary(model_Y_X3)
paste0("Y = ", round(model_Y_X3$coefficients[1],3), " + ", round(model_Y_X3$coefficients["X3"],3), "*X3")
#Regression of Y and X1,X2,X3
model_Y_multiple <- lm(Y ~ X1+X2+X3, data = expenditure)
summary(model_Y_multiple)
paste0("Y = ", round(model_Y_multiple$coefficients[1],3), " + ", round(model_Y_multiple$coefficients["X1"],3), "*X1"," + ", round(model_Y_multiple$coefficients["X2"],3), "*X2"," + ", round(model_Y_multiple$
```

```
    coefficients["X3"],3), "*X3")
```

Figure 1: Y vs X1 scatter plot.

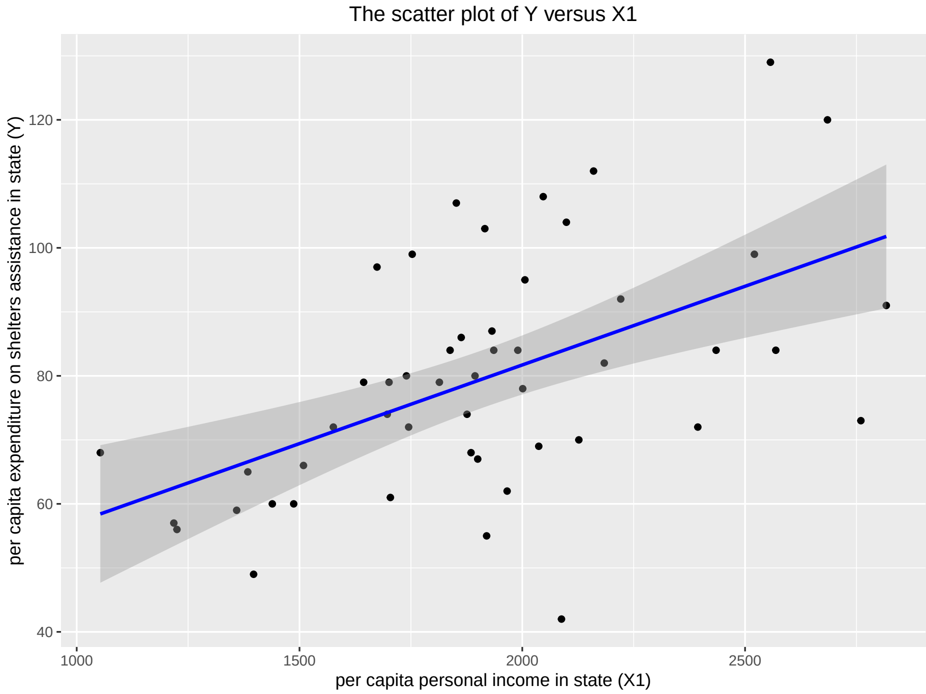


Figure 2: Y vs X2 scatter plot.

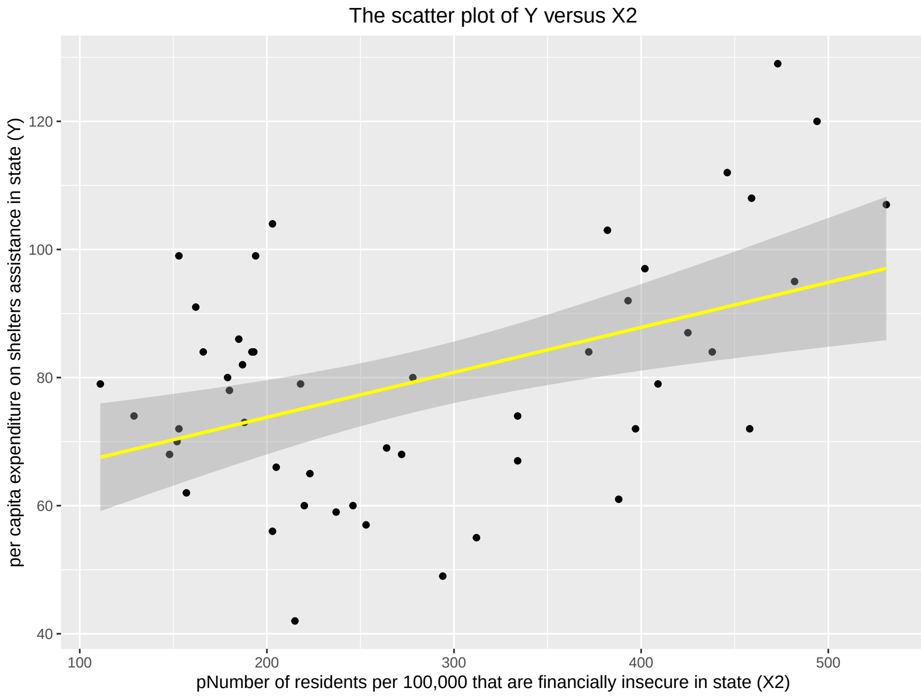


Figure 3: Y vs X3 scatter plot.

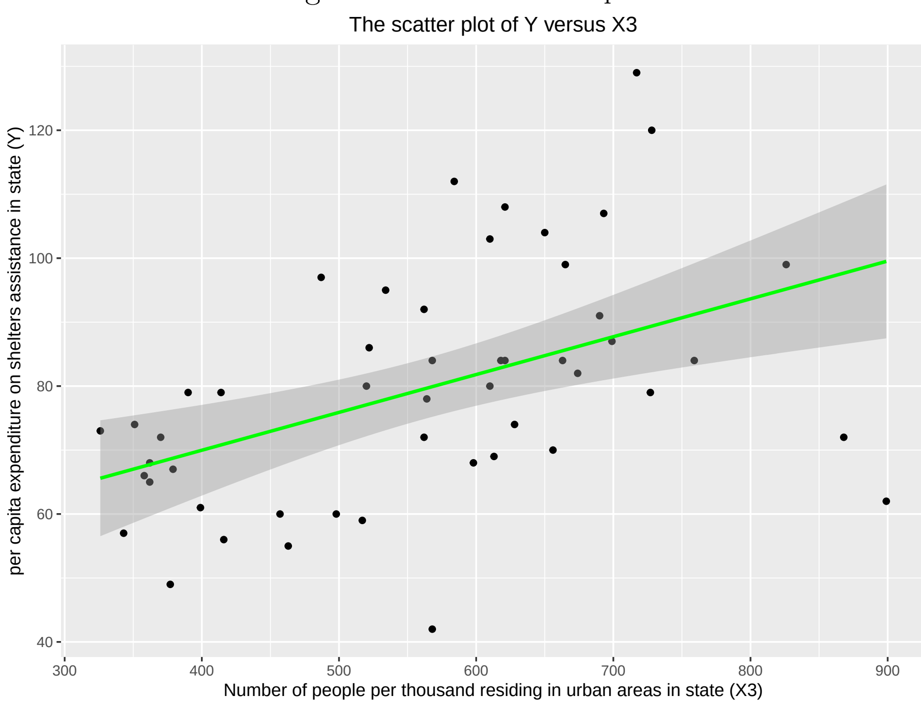


Figure 4: X1 vs X2 scatter plot.

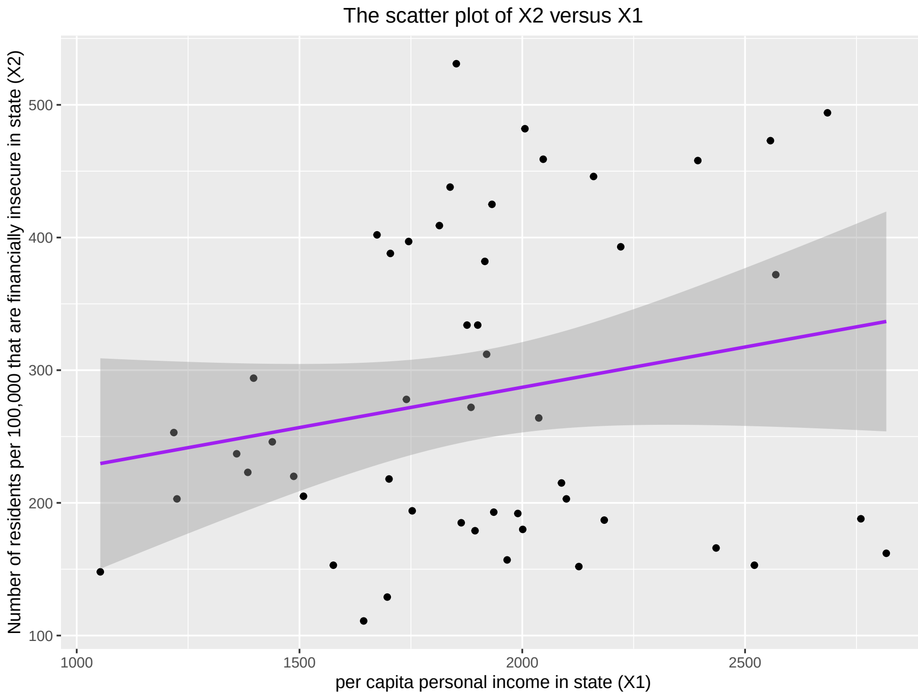


Figure 5: X1 vs X3 scatter plot.

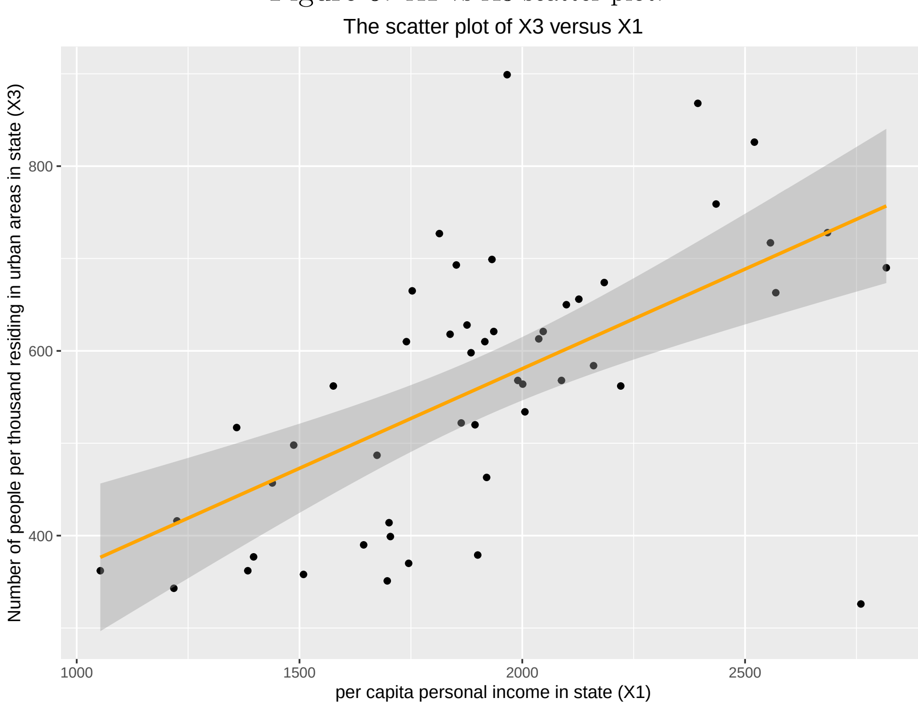


Figure 6: X2 vs X3 scatter plot.

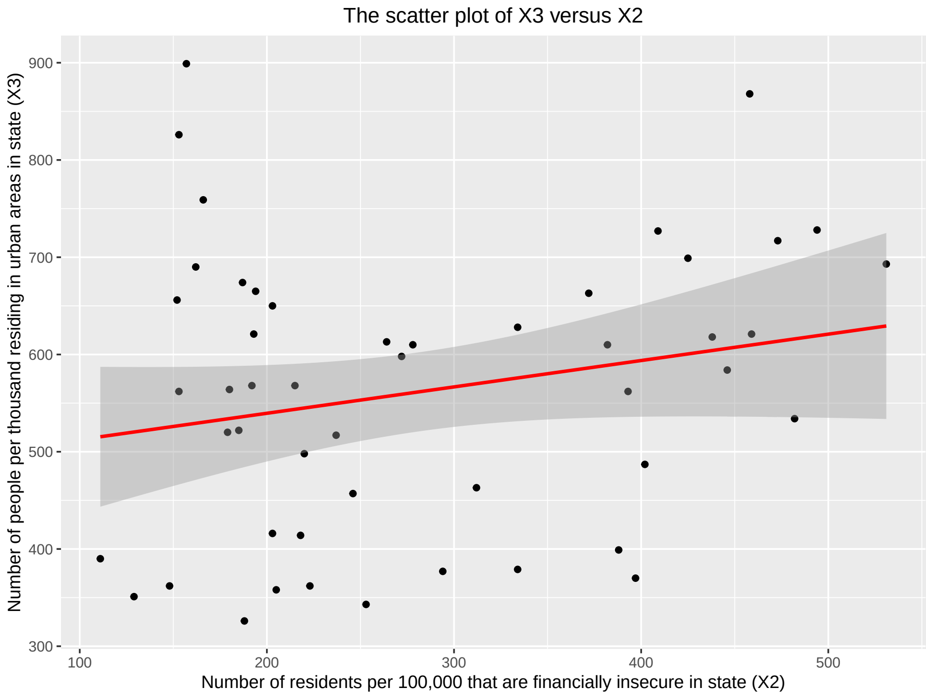
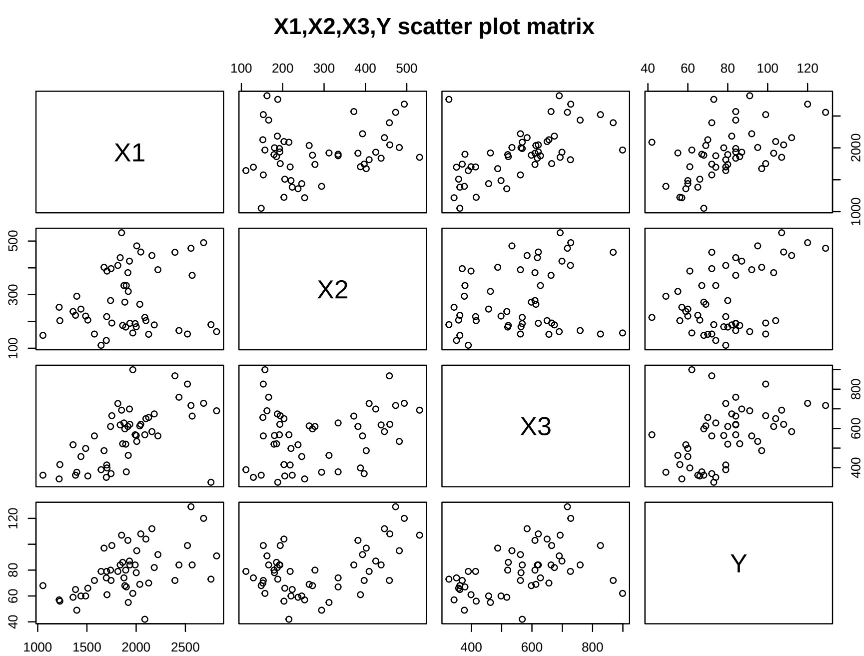


Figure 7: scatter plot matrix



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.

In order to get a more accurate relationship, we get the regression formula between Y, X1, X2, and X3 based on the linear regression model of Y, X1, X2, and X3:

1. The regression relationship between X1 and X2 is: "X1 = 1715.655 + 0.696*X2"
The p value is 0.152, which is greater than 0.05. There is no significant relationship between X1 and X2

2. The regression relationship between X1 and X3 is: "X1 = 988.947 + 1.643*X3"
The p value is 0.00000513, which is less than 0.05. There is significant positive relationship between X1 and X3

3. The regression relationship between X2 and X3 is: "X2 = 180.609 + 0.18*X3"
The p value is 0.123, which is greater than 0.05. There is no significant relationship between X2 and X3

4. The regression relationship between Y and X1 is: "Y = 32.546 + 0.025*X1"
The p value is 0.00007079, which is less than 0.05. There is significant positive relationship between Y and X1

5. The regression relationship between Y and X2 is: "Y = 59.761 + 0.07*X2"
The p value is 0.001095, which is less than 0.05. There is significant positive relationship between Y and X2

6. The regression relationship between Y and X3 is: "Y = 46.306 + 0.059*X3"
The p value is 0.0006955, which is less than 0.05. There is significant positive relationship between Y and X3

7. The regression relationship between Y and X1,X2,X3 is: "Y = 20.466 + 0.017*X1 + 0.053*X2 + 0.023*X3"

The p value is 0.00001203, which is less than 0.05. There is significant positive relationship between Y and X1,X2,X3

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

```r
#Extract the region column and Y column, and form a new dataset.
Y_region <- expenditure[, c("Y", "Region")]
Y_region
Region1_vs_Y1 <- ggplot(Y_region, aes(x = factor(Region), y = Y, fill = factor(Region))) +
    geom_bar(stat = "identity", position = "dodge") +
    labs(title = "Per capita housing assistance expenditure by region", x = "Region", y = "per capita expenditure on sheltersa ssistance in state")+
  theme(plot.title = element_text(hjust = 0.5))
ggsave("Region1_vs_Y1_barchart.pdf", plot = Region1_vs_Y1, width = 8, height = 6, units = "in")
#Regional variables are discrete data. To analyze the relationship between discrete data and continuous data, regression analysis is usually used.
Y_region_correlation <- cor(Y_region[,c("Y","Region")])
Y_region_correlation
#Region is a categorical variable, and regression analysis can better determine the correlation.
Y_region$Region <- as.factor(Y_region$Region)
# use linear  regression
model <- lm(Y ~ Region, data = Y_region)
# check the model
summary(model)
#The scatter plot of Region and Y
Region2_vs_Y2 <- ggplot(data = expenditure, aes(x = Region, y = Y)) +
    geom_point() +
    ggtitle("The scatter plot of Region versus Y") +
    xlab('Region') +
    ylab('per capita expenditure on shelters assistance in state (Y)')+
    theme(plot.title = element_text(hjust = 0.5))
ggsave("Region2_vs_Y2_scatterplot.pdf", plot = Region2_vs_Y2, width = 8, height = 6, units = "in")
#Calculate housing support expenditure per capita in each region
# Group by Region
region_split <- split(Y_region, Y_region$Region)
# Calculate the mean of Y in each group
Y_averages <- lapply(region_split, function(x) mean(x$Y, na.rm = TRUE))
```

Figure 8: Region2 vs Y2 scatterplot

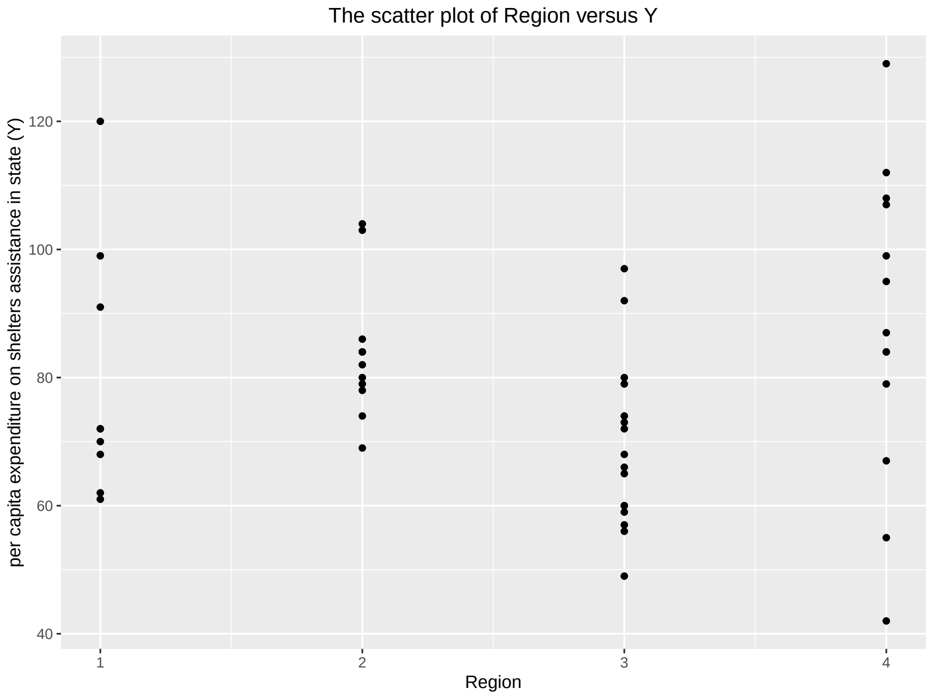


Figure 9: Region2 vs Y2 scatterplot

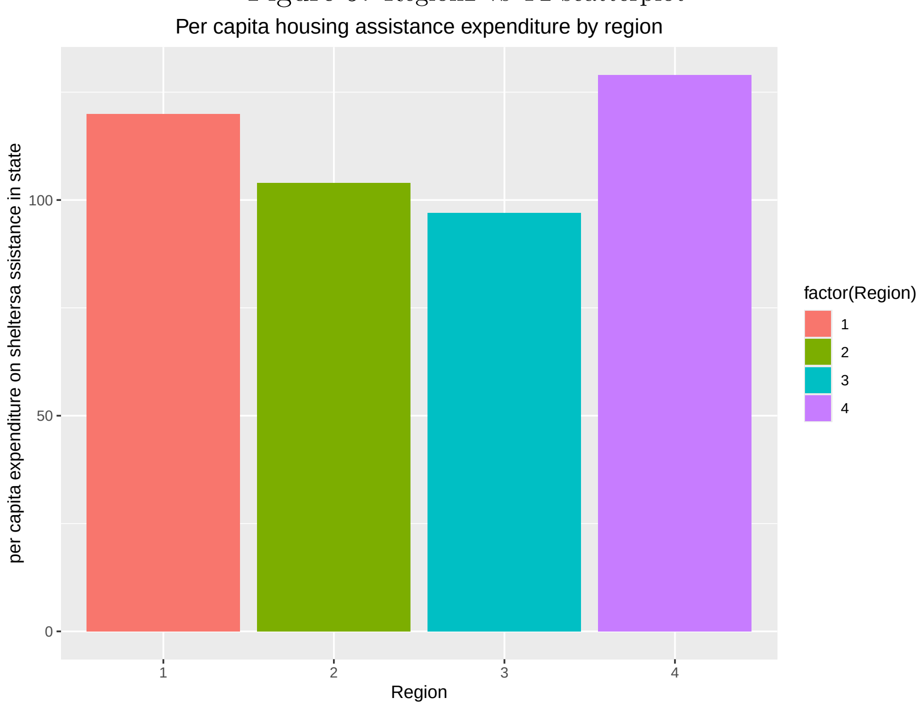


Figure 10: model summary

```
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.

```r
# scatter plot between X1 and Y
X1_Y_Region <- expenditure[, c("X1","Y", "Region")]
X1vsY_Region_scatterplot <- ggplot(X1_Y_Region, aes(x = X1, y = Y, color
    = factor(Region), shape = factor(Region))) +
  geom_point(size = 3) +
  geom_smooth(method = "lm", se = TRUE, color = "orange") +
  scale_color_manual(values = c("1" = "red", "2" = "blue", "3" = "green",
    "4" = "purple")) + # set color
  scale_shape_manual(values = c("1" = 16, "2" = 17, "3" = 18, "4" = 19))
   + # set shape
  labs(title = "Y vs X1 Relationship by Region",
       x = "X1",
       y = "Y",
       color = "Region",
```

```
        shape = "Region")+
  theme(plot.title = element_text(hjust = 0.5))
ggsave("X1vsY_Region_scatterplot.pdf", plot = X1vsY_Region_scatterplot,
   width = 8, height = 6, units = "in")
X1_Y_Region_correlation <- cor(X1_Y_Region[,c("Y", "X1", "Region")])
X1_Y_Region_correlation
```

Figure 11: X1 vs Y vs Region scatter plot

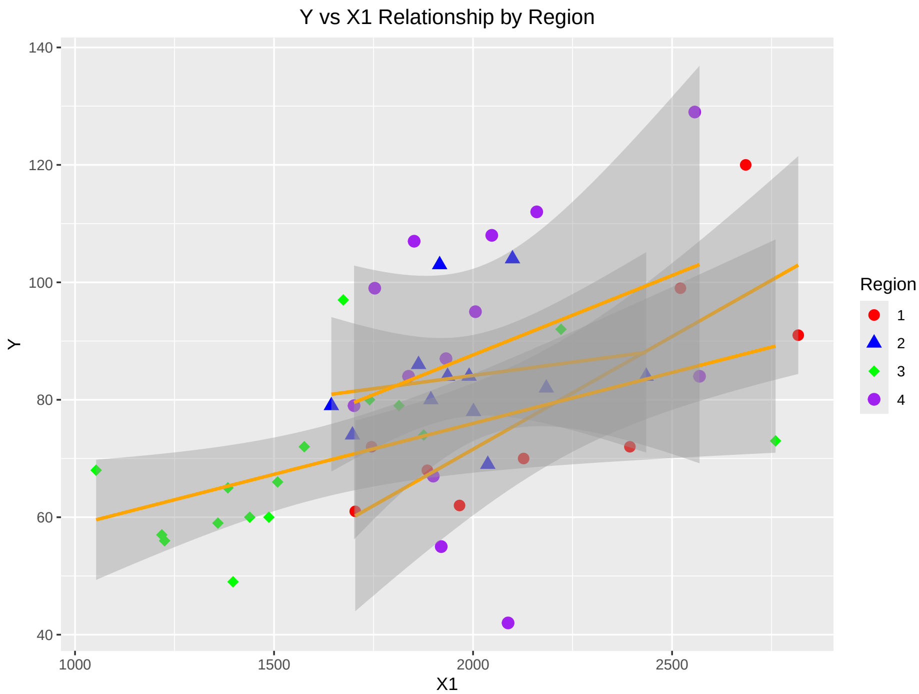


Figure 12: X1 vs Y by Region correlation

	Y	X1	Region
Y	1.00000000	0.5317212	0.06563026
X1	0.53172116	1.0000000	-0.21890587
Region	0.06563026	-0.2189059	1.00000000

From this scatter plot, we can see that the per capita income of each region is positively correlated with housing security expenditure, with a correlation coefficient of 0.5317212, which is a strong positive correlation. Therefore, the higher the per capita income of each region, the higher the housing security expenditure. In addition, the correlation between personal income and housing security expenditure in the Northeast region (region 1) is stronger than in other regions. After calculation, the correlation coefficient is about 0.802, which is a strong positive correlation.