Problem Set 1

Applied Statistical Analysis 1

Maiia Skrypnyk 23371609

Question 1 (40 points): 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.

```
#1) Calculating the mean of y — our point estimate.
mean(y)
mean <- mean(y)

#2) Calculating standard deviation with the help of R function (the
    simple way)...
sd(y)
s <- sd(y)
#... or calculating standard deviation by ourselves as s=sqrt(sum((y -
    mean(y))^2)/(n - 1))
n <- length(y)
s1 <- sqrt(sum((y-mean(y))^2)/(n-1))
#We can see that the results are identical:
s
s1

#3) Calculating the standard error using the formula se=s/sqrt(n)
sd(y)/sqrt(n) #n being length(y)
se <- sd(y)/sqrt(n)

#4) Calculating how much area do we need under the curve (if CI=90%)...
#4.1) ... to the right? (1-Confidence Coefficient)/2
(1-.90)/2
right <- (1-.90)/2
right
#4.2) ... to the left? (1+Confidence Coefficient)/2
```

```
(1+.90)/2
left <- (1+.90)/2
left

#5) Calculating Z-score.
#I used the Z-score table, and estimated that Z-score lies close to 1.65.
#But I also googled Z-score for 90% CI, and the result was 1.645, which is close, but still not the same.

#I found an explanation in this resource: https://www.vedantu.com/question-answer/z-value-for-a-90-95-and-99-percent-confidence-class-11-maths-cbse-606c515d034c9021d4c5b4f0
#"Looking for this value in the normal distribution table given below,
#we can see that this value [area under the curve to the left = 0.95 — MS]
#lies close to the row containing 1.6 and column containing 0.05.
#It also lies close to the row containing 1.6 and column containing 0.04.
#So, we take a mean of these values to obtain the z value at this point.

(1.64+1.65)/2
z <- (1.64+1.65)/2
z

#6) Calculating Confidence Interval via formulas 1.1) mean(y)+z(y)+se(=s/sqrt(n)), 1.2) mean(y)-z(y)+se(=s/sqrt(n))

mean+(z*s/sqrt(n))
mean-(z*s/sqrt(n))

upper_90 <- mean+(z*s/sqrt(n))
lower_90 <- mean-(z*s/sqrt(n))

ConfidenceInterval90 <- c(lower_90, upper_90)
ConfidenceInterval90

#RESULTS: Confidence Interval [94.13244, 102.74756] =
#With repeated sampling, 90% of CIs will fall between these bounds: [94.13244, 102.74756]
```

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 $\alpha = 0.05$.

```
length(y)
n <- length(y)

#1) Assumptions: continuous data; random sample; small sample (as n<30).
#As n<30, we should use a t-statistic.

```

```r
#2) Null hypothesis: the average IQ score of the School A is lower than
# the average IQ score (100) among all the schools in the country.
#H0: mean(y) < 100
#Ha: mean(y) > 100

#3) Test statistic: t = (mean(y)-100)/se; df=(n-1), se = sd(y)/sqrt(n)
t <- (mean(y)-100)/se
df <- (n-1) #24
df

#4) P-value: as Ha: mean(y) > 100,
#P = probability to the right of observed t-value (Agresti 2018, 155)
#As I am performing a one-tailed test (right-tailed), I am not
    multiplying pt by 2 (2*pt),
#but use the next formula:

p <- pt((t), df = n-1, lower.tail=FALSE) #if left-tailed, then lower.tail
    =TRUE

#I have been checking my results in multiple places, so
#https://www.statology.org/p-value-of-t-score-r/ suggested the next
    formula (quite the same in meaning)

p1 <- pt(q=t, df=24, lower.tail=FALSE)

#(q represents the quantile (or value) for which we want to calculate the
     cumulative probability)
#and I also checked the calculator here (got the same result): https://
    www.omnicalculator.com/statistics/p-value
#Also, not using pnorm() because we are using a t-statistic, not a z-
    statistic.

#Let's also run a t-test using t-test() formula to check our results once
     again:

t_test <- t.test(y, mu=100, alternative="greater")
t_test

p>.05
p1>.05

#As our P-value is larger than Alpha level (.05), our result is NOT
    statistically significant.
#Therefore, we CANNOT REJECT the null hypothesis/
#there is NOT enough evidence to reject the null hypothesis.
```

Question 2 (40 points): 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`.

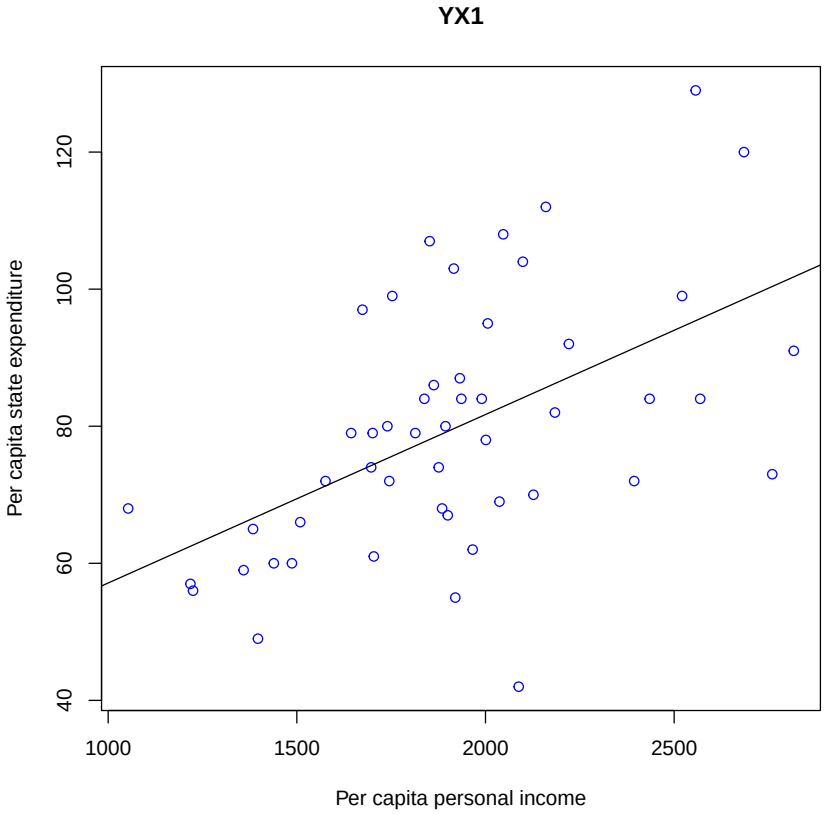
```
expenditure <- read.table("https://raw.githubusercontent.com/ASDS-TCD/StatsI_Fall2023/main/datasets/expenditure.txt", header=T)
```

- **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)?**

2.1.1:YX1

(Please see the plot on the next page)

Figure 1: Relationship between Per capita state expenditure and Per capita personal income

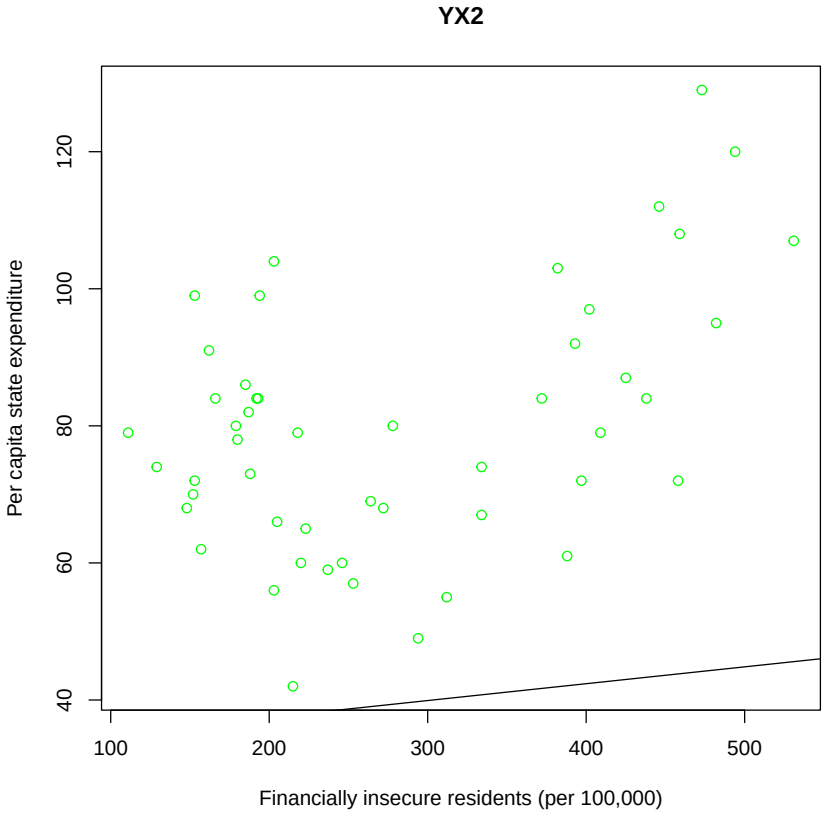


cor = 0.5, positive moderate

- The correlation between Per capita expenditure on shelters/housing assistance in state and Per capita personal income in state is LINEAR, POSITIVE, MODERATE (with a few outliers).
- As Personal income increases, so does Expenditure on shelters/HA, though this relationship is not strong (+there are exceptions).

2.1.2:YX2

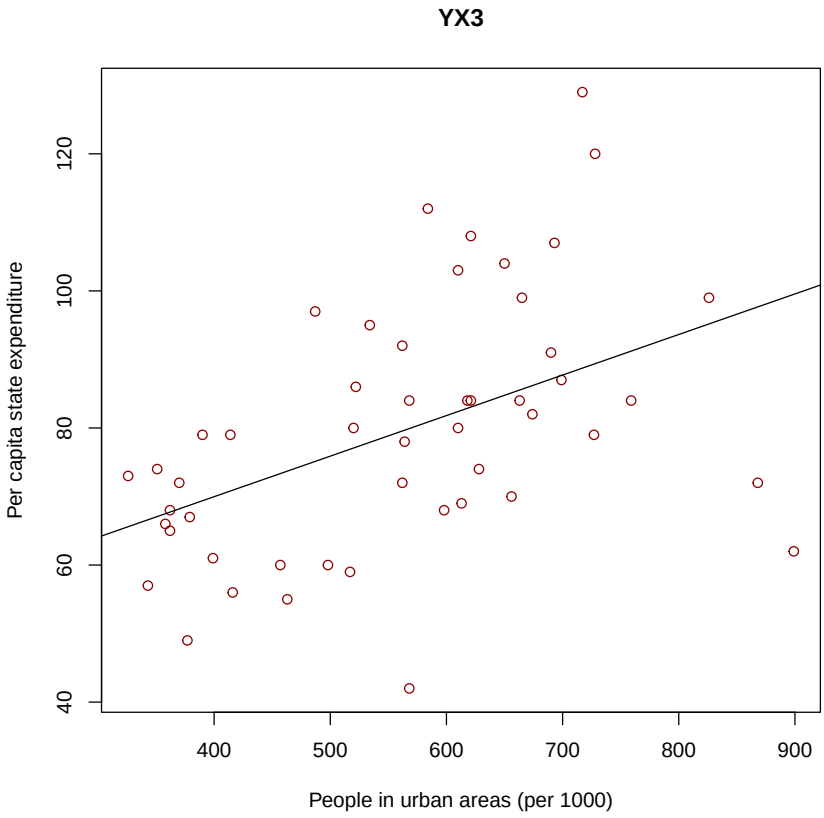
Figure 2: Relationship between Per capita state expenditure and № of Financially insecure residents in state (per 100,000)



- The correlation between Per capita expenditure on shelters/housing assistance in state and Number of residents per 100,000 that are "financially insecure" in state follows a (NON-LINEAR) U-SHAPED pattern.
- *Being honest, this is my first time hearing about a U-shaped relationship, so I searched for the answers on the Internet to understand it:* "U-shaped relationship... usually means that the relationship is first decreasing and then increasing, or vice versa. In other words, it means that the relationship is not monotonic (non-monotonic), but instead has exactly one extremum (maximum or minimum)."
- From my understanding, there is an initial range where changes in Number of "financially insecure" residents do not have a significant effect on Per capita state expenditure, but starts to increase its effect after a 'turning point' in the bottom of the U-shape.

2.1.3:YX3

Figure 3: Relationship between Per capita state expenditure and № of Urban residents (per 1000)

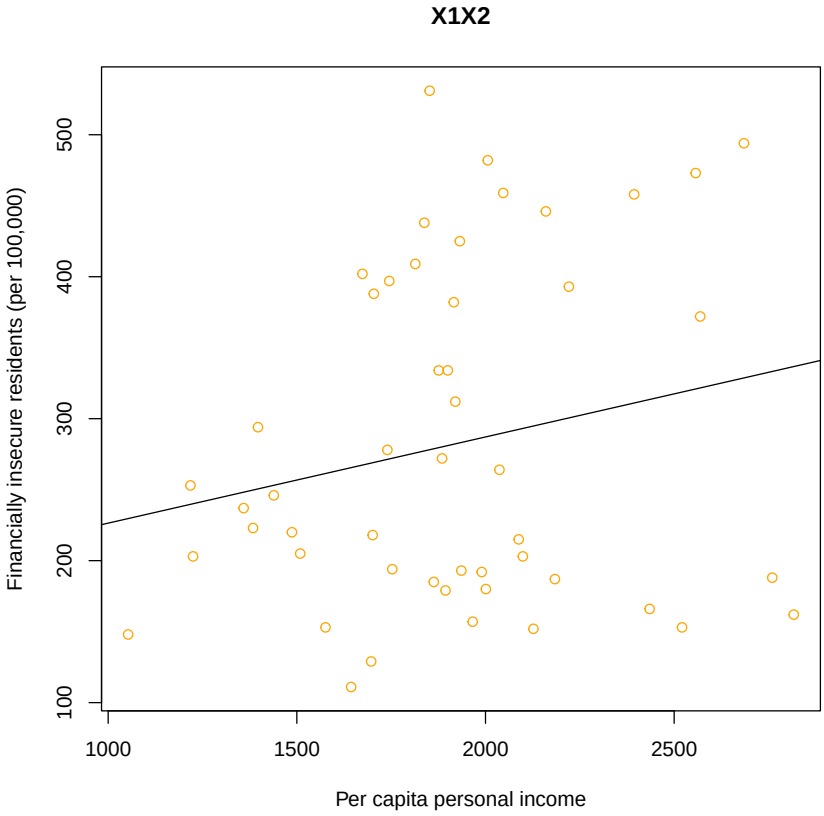


cor = 0.46, positive moderate

- The correlation between Per capita expenditure on shelters/housing assistance in state and Number of people per thousand residing in urban areas in state is LINEAR, POSITIVE, MODERATE (with a few outliers).
- As the Number of urban residents increases, Per capita state expenditure on shelters/HA tends to increase as well, but this correlation is not strong.

2.1.4:X1X2

Figure 4: Relationship between № of Financially insecure residents in state (per 100,000) and Per capita personal income in state

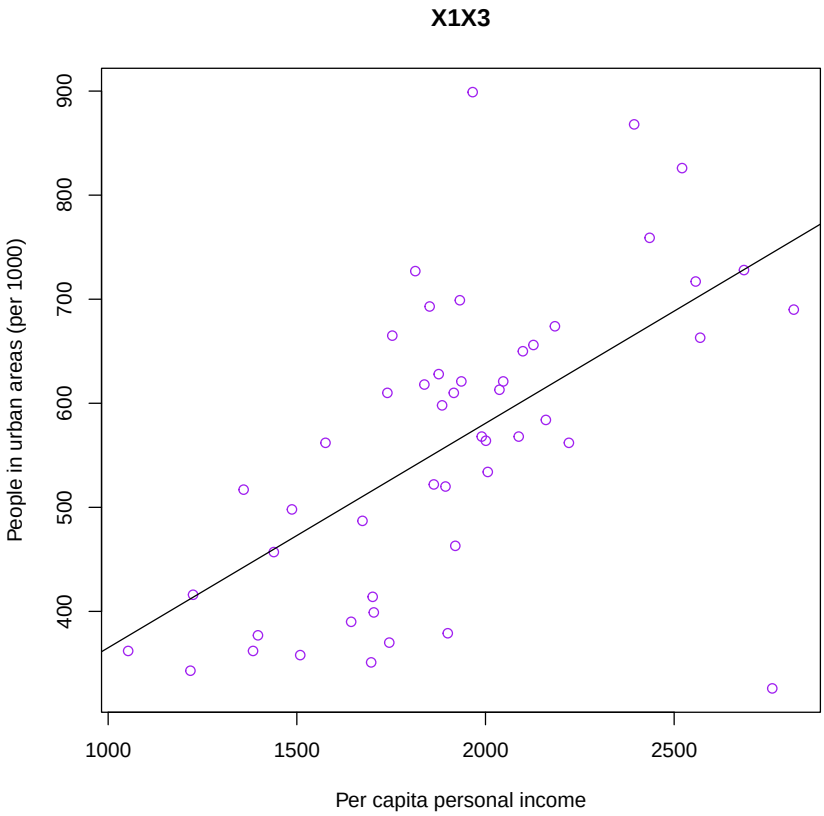


cor = 0.2, positive weak

- The correlation between Per capita personal income in state and Number of residents per 100,000 that are "financially insecure" in state is LINEAR, POSITIVE, WEAK.
- As Per capita personal income in state increases, Number of Financially insecure residents in that state tends to decrease, but this tendency is quite weak, and there are probably other factors that influence this variable more strongly.

2.1.5:X1X3

Figure 5: Relationship between Per capita personal income in state and № of Urban residents (per 1000).

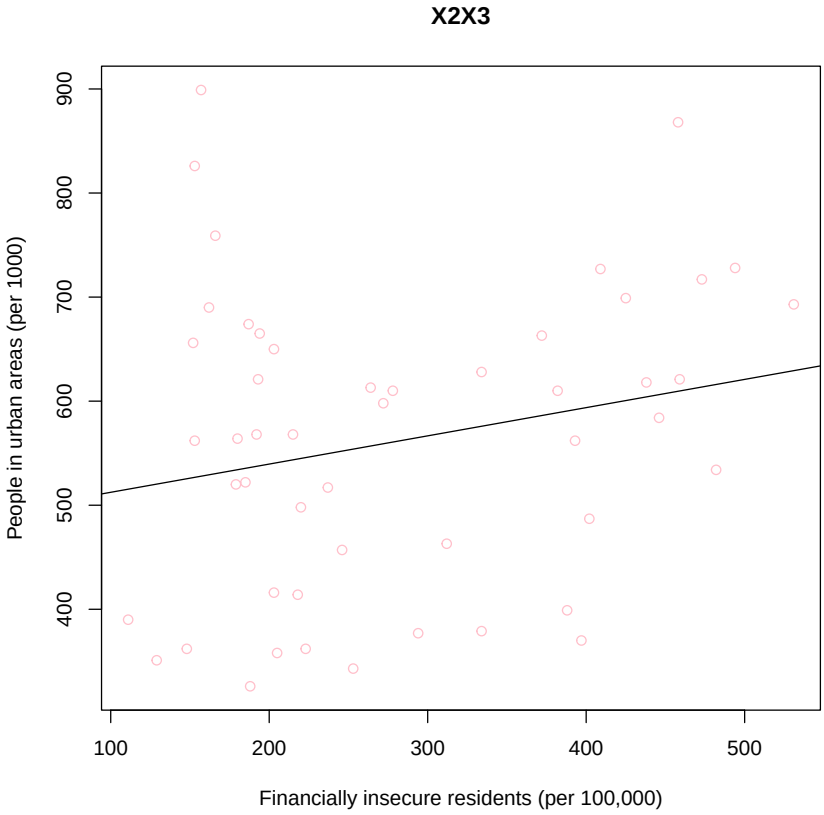


cor = 0.6, positive moderate

- The correlation between Per capita personal income in state and Number of people per thousand residing in urban areas in state is LINEAR, POSITIVE, MODERATE.
- As Per capita personal income in state increases, Number of urban dwellers tends to increase as well. However, the correlation is not very strong(though the strongest of all our cases), and other factors' influence may play its role.

2.1.6:X2X3

Figure 6: Relationship between № of Financially insecure residents in state (per 100,000) and № of Urban residents (per 1000).

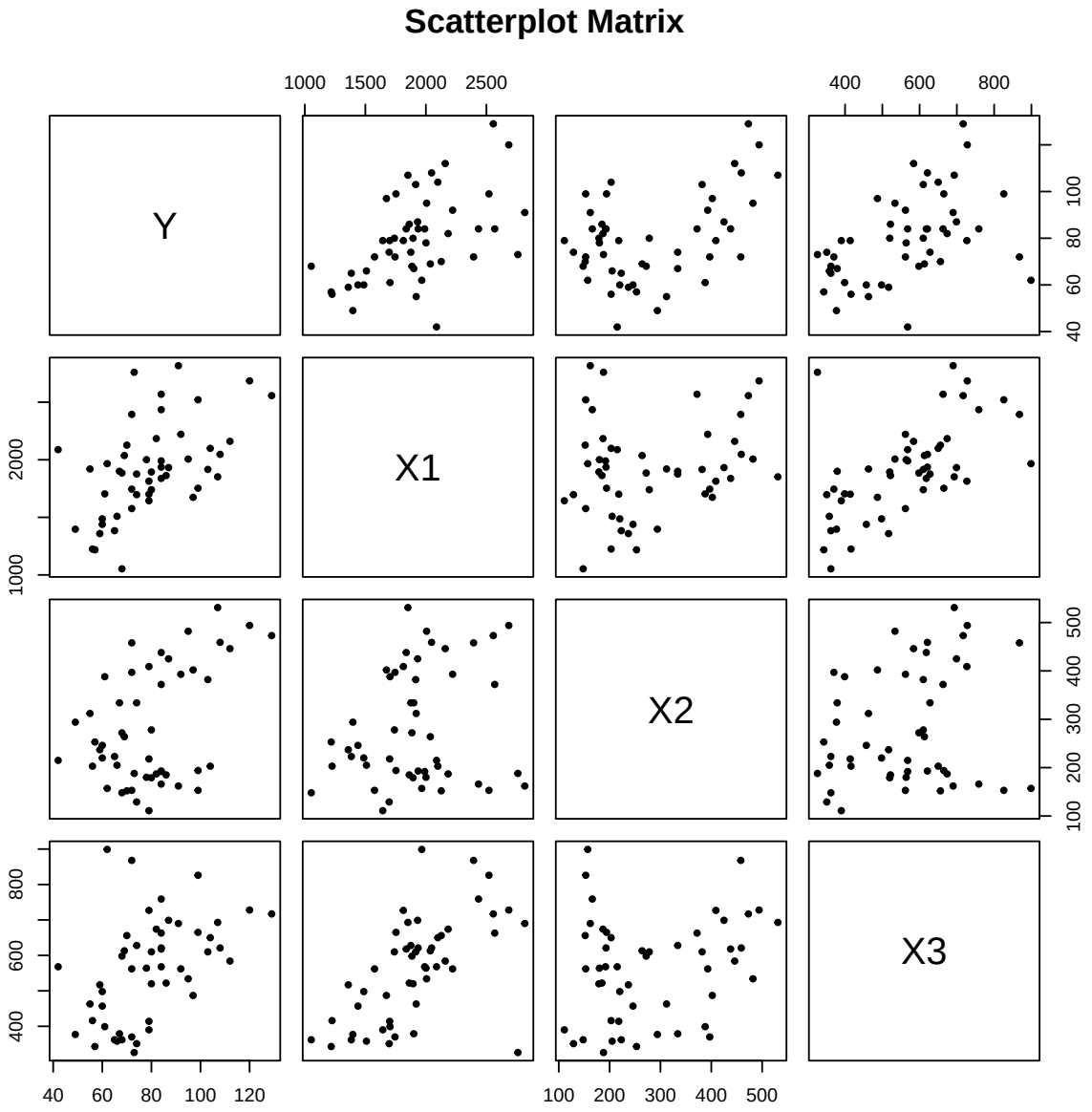


cor = 0.2, positive weak

- The correlation between Number of residents per 100,000 that are "financially insecure" in state and Number of people per thousand residing in urban areas in state is LINEAR, POSITIVE, WEAK.
- As Financial Insecurity increases, Number of Urban dwellers also tends to increase, but this tendency is weak, and there are probably other factors that influence this variable more strongly.

2.1.7: Scatterplot Matrix

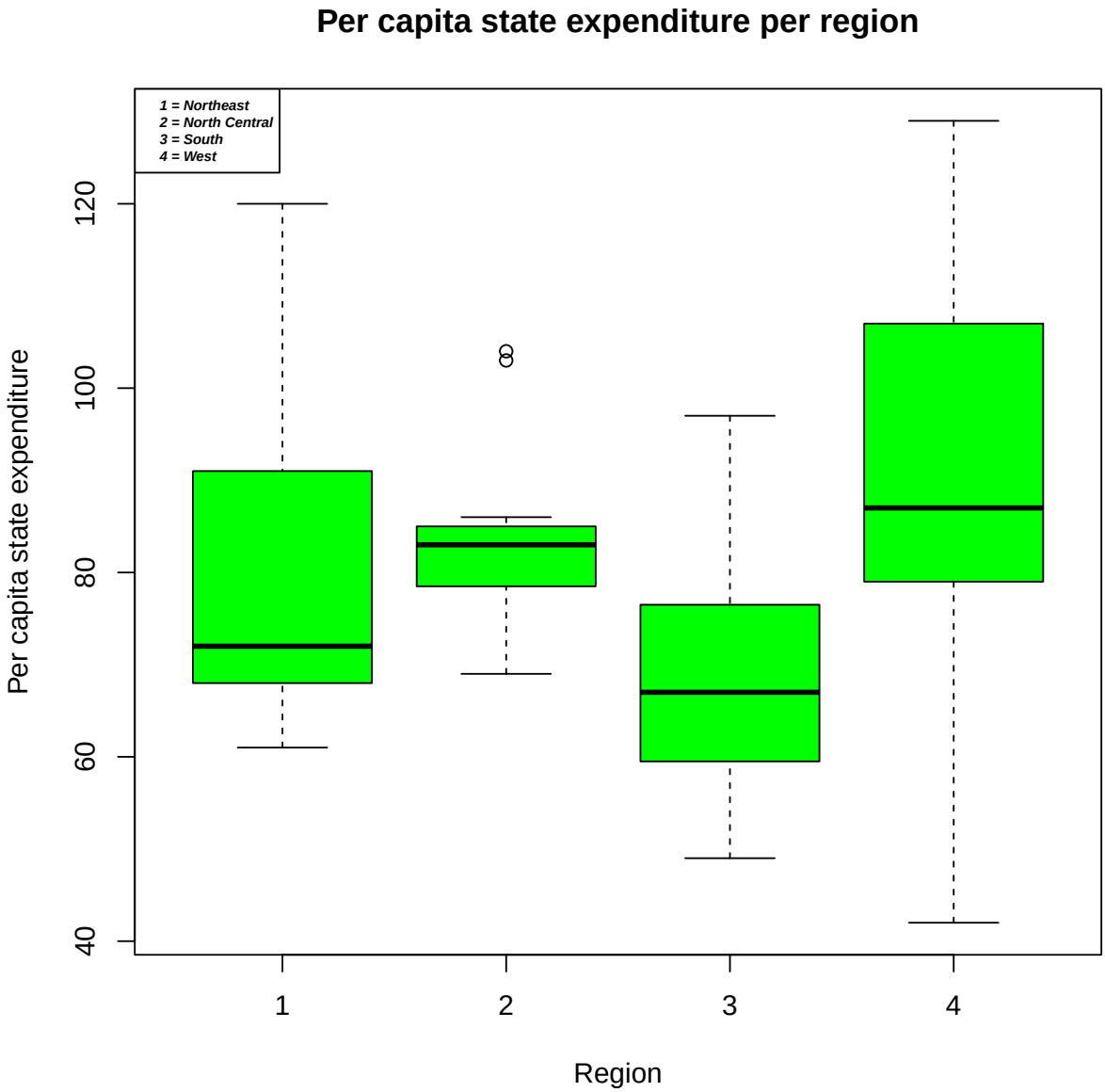
Figure 7: Scatterplot Matrix (Y, X1, X2, X3)



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

2.2: YRegion

Figure 8: Relationship between Per capita expenditure on shelters/housing assistance in state and Region



```r
#We can calculate our mean per region manually, by subsetting the dataframe
    first...:

north_east <- expenditure[expenditure$Region == 1,]
north_central <- expenditure[expenditure$Region == 2,]
south <- expenditure[expenditure$Region == 3,]
west <- expenditure[expenditure$Region == 4,]

#... and using mean() function to calculate average Per capita state
    expenditure
#for each of the four regions.

mean1 <- mean(north_east$Y)
mean2 <- mean(north_central$Y)
mean3 <- mean(south$Y)
mean4 <- mean(west$Y)

#The Internet suggested I use the aggregate() function
#('...allows you to specify a dataframe, a condition, and a function to apply
#to each group of rows that meet the condition' — https://saturncloud.io/blog
    /aggregate-dataframe-by-condition-in-r-a-comprehensive-guide)
#which is a much simpler way and returns the same results.

mean_y_by_region <- aggregate(Y~Region, data = expenditure, FUN = base::mean)
mean_y_by_region

max(mean_y_by_region)
```

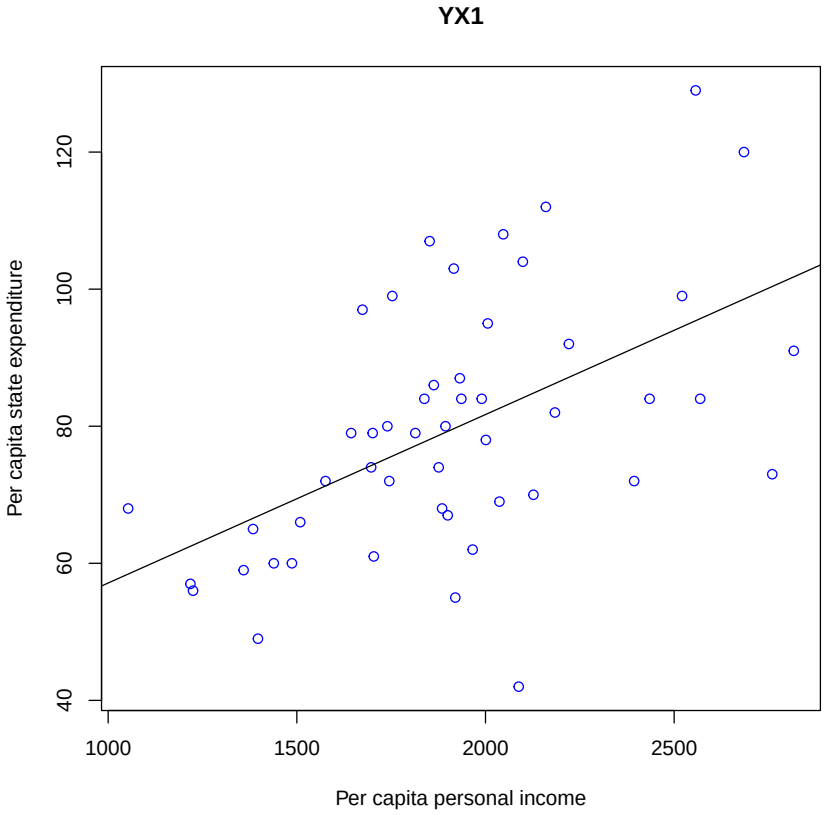
Therefore, REGION 4=WEST, on average, has the highest per capita expenditure on housing assistance (88.30769).

- **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.**

2.3.1: YX1

(Please see the plot on the next page)

Figure 9: Relationship between Per capita state expenditure and Per capita personal income



cor = 0.5, positive moderate

- The correlation between Per capita expenditure on shelters/housing assistance in state and Per capita personal income in state is LINEAR, POSITIVE, MODERATE (with a few outliers).
- As Personal income increases, so does Expenditure on shelters/HA, though this relationship is not strong (+there are exceptions).

2.3.2: YX1 by Region

Figure 10: Relationship between Per capita state expenditure and Per capita personal income by Region