PS1 Response 19336078

Owen Eglinton

26/06/2024

My first step was to set the working directory:

```
setwd(”C:\\Users\\Owen Eglinton\\Documents\\GitHub\\StatsI_Fall2024\\
    problemSets\\PS01\\template\\myanswers”)
```

Question 1

I began by loading in the 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)
```

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

I coded the process of finding the necessary descriptive statistics and using them to calculate the upper and lower bounds of the CI as follows:

```
confidence <- 0.9
n <- length(y)
t <- qt((confidence + 1)/2, df = n-1)
ci_u <- round(mean(y) + t * sd(y)/sqrt(n), 2)
ci_l <- round(mean(y) - t * sd(y)/sqrt(n), 2)
paste(”The 90% CI for average student IQ is”, ci_l, ”-”, ci_u)
```

After running this code, I got the following output:

```
[1] "The 90% CI for average student IQ is 93.96 - 102.92"
```

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 $\alpha = 0.05$.

I set up a one-tailed t-test as follows:

```r
t.test(y, mu = 100, alternative = "greater")
```

and got the following output:

```
One Sample t-test

data:  y
t = -0.59574, df = 24, p-value = 0.7215
alternative hypothesis: true mean is greater than 100
95 percent confidence interval:
93.95993      Inf
sample estimates:
mean of x
98.44
```

Since the `p-value` is greater than α, we fail to reject the null hypothesis, and thus must conclude that the average student IQ is not greater than 100.

Question 2

I began by loading in the data:

```r
expend <- read.table("https://raw.githubusercontent.com/ASDS-TCD/StatsI_
   Fall2024/main/datasets/expenditure.txt", header=TRUE)
```

1. 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)?

First, I imported the data descriptions as a label vector:

```r
labels_exp = c("State", "Housing Expenditure p.c.", "Income p.c.", "#
   Financially Insecure per 100,000", "# Urban Residents per 1000", "Region")
```

Then, for ease of reference, I built a function to plot specific variables with labels:

```r
plot_expend <- function(x, y) {
    plot(expend[[y]],
        expend[[x]],
        xlab = labels_exp[[y]],
        ylab = labels_exp[[x]],
    )
}
```

With these prepared, I wrote the following code to generate and save the necessary graphs, while describing their correlation coefficient:

```
for (i in 2:4) {
    for (j in (i+1):5) {
        png(file = paste(colnames(expend)[[i]], "_",
                         colnames(expend)[[j]], ".png",
                         sep = ""))
        plot_expend(i, j)
        dev.off()
        print(paste("The correlation between",
                    colnames(expend)[[i]],
                    "and",
                    colnames(expend)[[j]],
                    "is",
                    round(cor(expend[[i]], expend[[j]]), 2))
        )
    }
}
```

This generated *Figures 1-6*, and the following output:

```
[1] "The correlation between Y and X1 is 0.53"
[1] "The correlation between Y and X2 is 0.45"
[1] "The correlation between Y and X3 is 0.46"
[1] "The correlation between X1 and X2 is 0.21"
[1] "The correlation between X1 and X3 is 0.6"
[1] "The correlation between X2 and X3 is 0.22"
```

- Observing *Figure 1*, Y (per-capita expenditure on housing) is positively correlated with X1 (per-capita income). However, the relationship seems heteroscedastic, with the variance of Y seemingly increasing in X1.

- Observing *Figure 2*, Y (per-capita expenditure on housing) is positively correlated with X2 (the number of financially insecure persons per 100,000), if a linear relationship is assumed. However, the shape of the scatterplot would seem to suggest a non-linear relationship, with Y decreasing in X2 to a ”trough” at c.(X2=275, Y=60), before Y increases in X2 from that point.

- Observing *Figure 3*, Y (per-capita expenditure on housing) is positively correlated with X3 (the number of urban residents per 1000). However, the relationship again seems heteroscedastic, with the variance of Y seemingly increasing in X3.

- Observing *Figure 4*, X1 (per-capita income) is relatively uncorrelated with X2 (the number of financially insecure persons per 100,000), if a linear relationship is assumed. However, the shape of the scatterplot would again seem to suggest a non-linear relationship, with X1 decreasing in X2 to a ”trough” at c.(X2=275, X1=1700), before X1 increases in X2 from that point.

- Observing *Figure 5*, X1 (per-capita income) is positively correlated with X3 (the number of urban residents per 1000). The relationship seems linear and homoscedastic.

- Observing *Figure 6*, X2 (the number of financially insecure persons per 100,000) is relatively uncorrelated with X3 (the number of urban residents per 1000), if a linear relationship is assumed. However, the shape of the scatterplot would yet again seem to suggest a non-linear relationship, with X3 decreasing in X2 to a ”trough” at c.(X2=275, X3=475), before X3 increases in X2 from that point.

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

I began by coding *Region* as a factor variable, with the respective region names as factors, before calling a plot of expenditure against region:

```r
region_labels = c("Northeast", "North Central", "South", "West")
expend$Region <- factor(expend$Region, labels = region_labels)

png(file = "Y_Region.png")
plot_expend(2, 6)
dev.off()
```

This generated *Figure 7*.

Then, for ease of reference, I built a function to return specific region means:

```r
region_factor <- function(x) {
    as.numeric(expend$Region) == x
}

region_mean <- function(x) {
    sum(expend$Y*(region_factor(x)))/sum(region_factor(x))
}
```

With this prepared, I wrote the following code to output the region with the highest average expenditure:

```r
mean_high <- region_mean(1)
highest <- region_labels[[1]]
for (i in 2:length(levels(expend$Region))) {
    if (mean_high < region_mean(i)) {
        mean_high <- region_mean(i)
        highest <- region_labels[[i]]
    }
}

paste("The region with the highest per capita expenditure is the ",
        highest,
        ", with an average of $",
        round(mean_high, 2),
        sep = ""
)
```

This gave the following output:

```
[1] "The region with the highest per capita expenditure is the West, with
an average of $88.31"
```

3. 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.

Recalling the description of this relationship given in the answer to Part 1 of this question:

```
[1] "The correlation between Y and X1 is 0.53"
```

- Observing *Figure 1*, Y (per-capita expenditure on housing) is positively correlated with X1 (per-capita income). However, the relationship seems heteroscedastic, with the variance of Y seemingly increasing in X1.

I proceeded to write the following code to re-generate the graph to include the variable *Region*, represented via the shapes and colours of the symbols:

```r
png(file = "Y_X1_Region.png")

plot(expend[[3]],
     expend[[2]],
     col = expend[[6]],
     pch = as.numeric(expend$Region),
     xlab = labels_exp[[3]],
     ylab = labels_exp[[2]],
)

legend(x = "topleft",
    legend=region_labels,
    col=c(1:4),
    pch=c(1:4)
)

dev.off()
```

This generated *Figure 8*.

Figures

Figure 1: The relation between Y and X1.

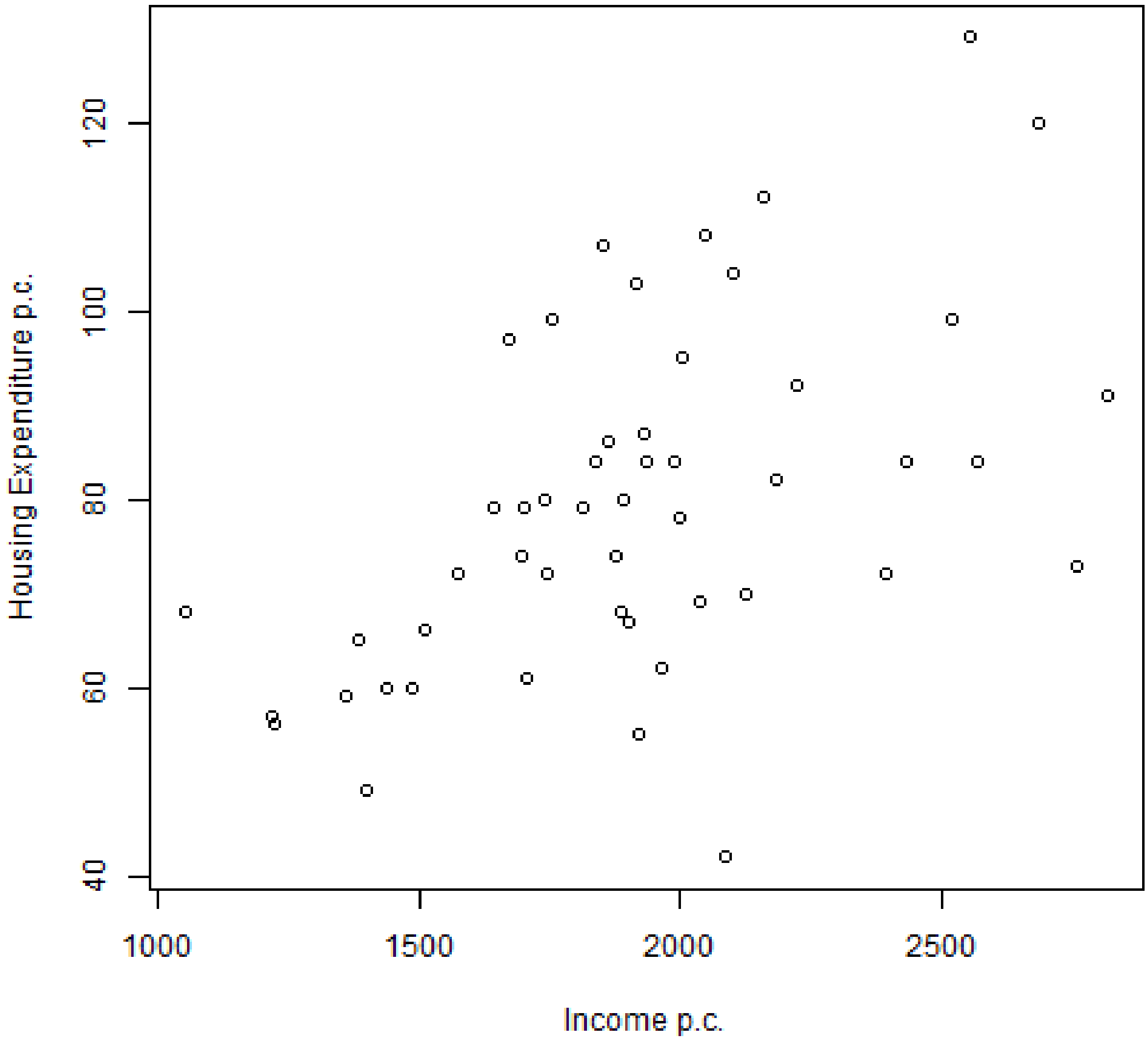


Figure 2: The relation between Y and X2.

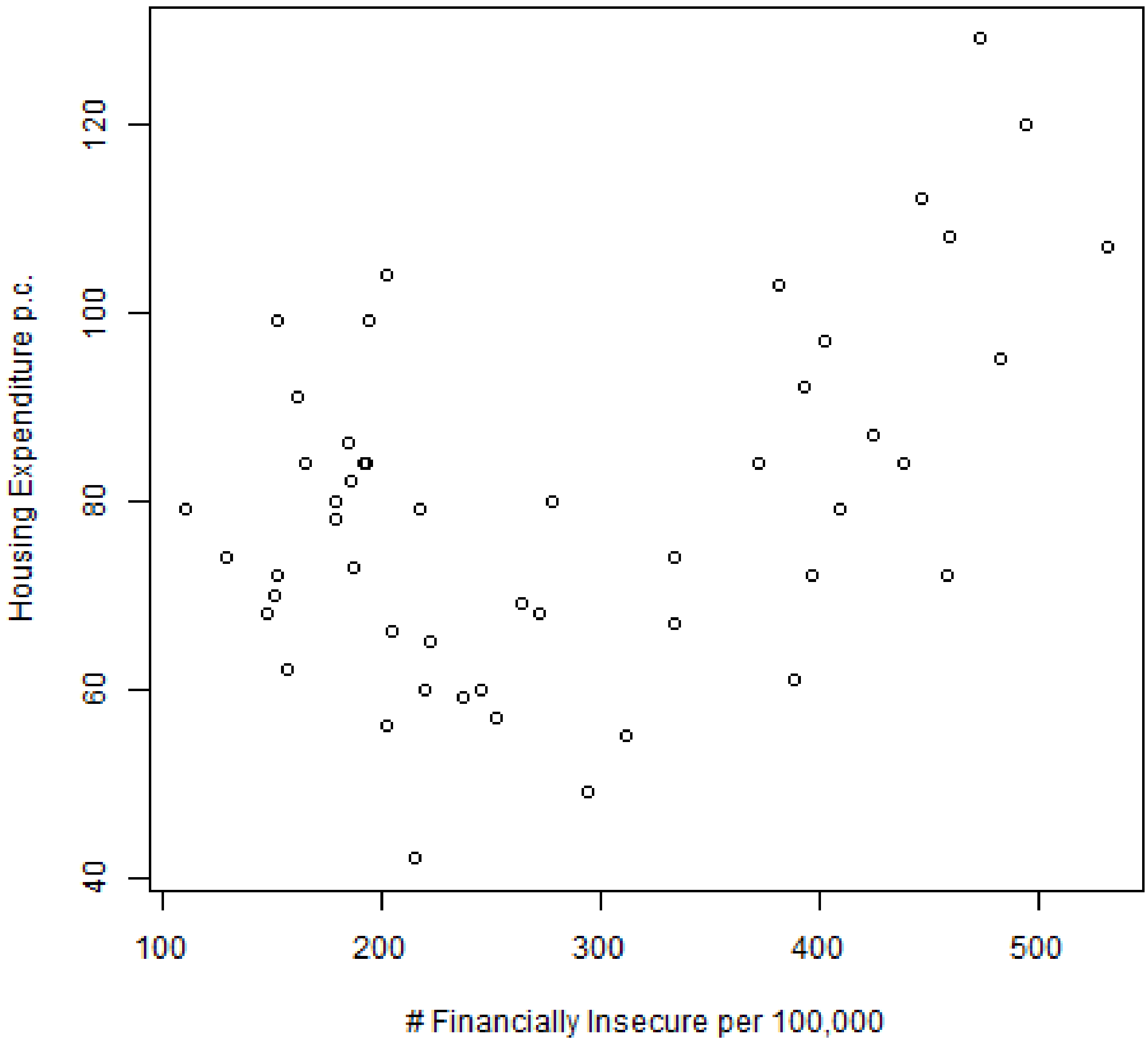


Figure 3: The relation between Y and X3.

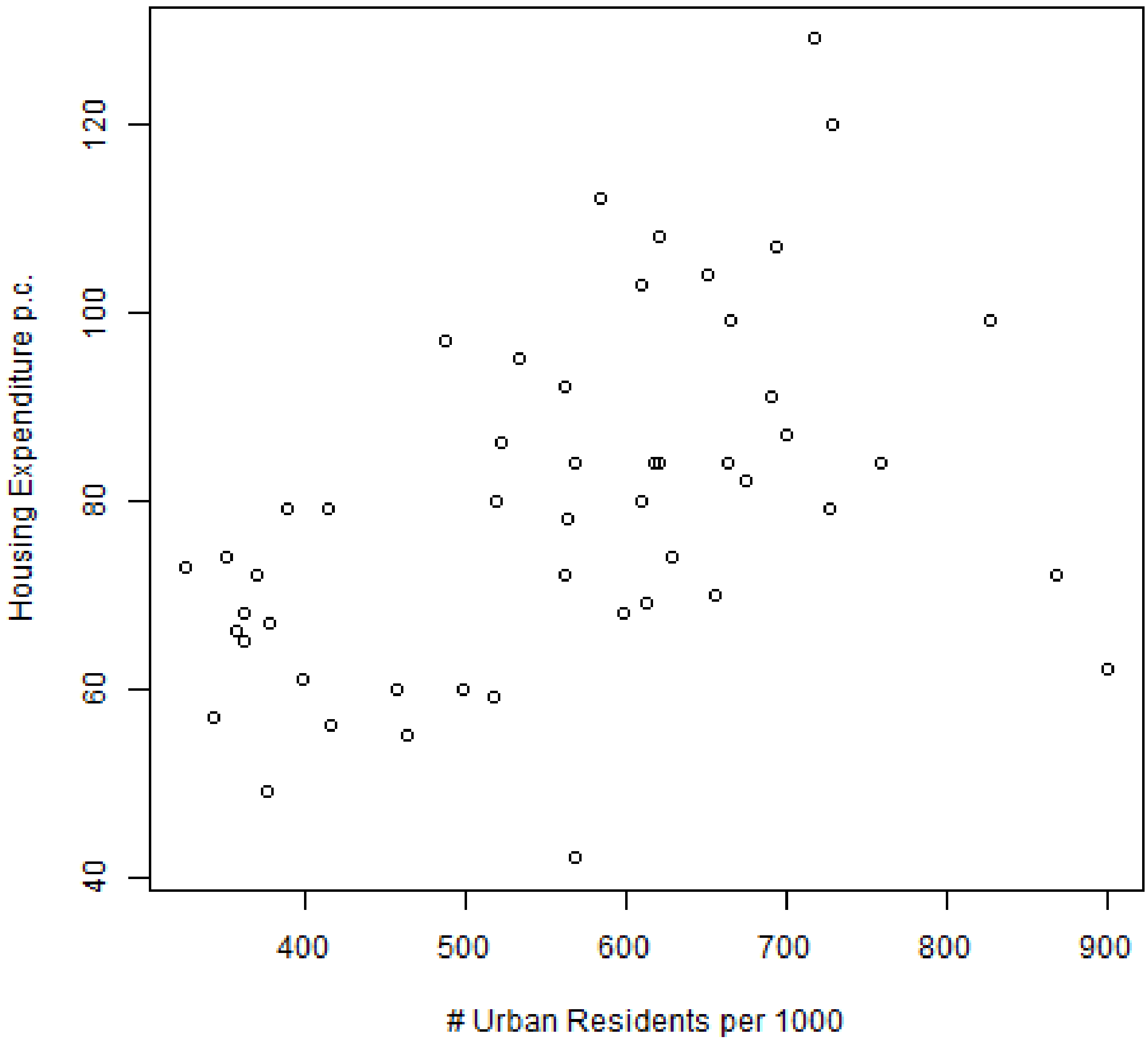


Figure 4: The relation between X1 and X2.

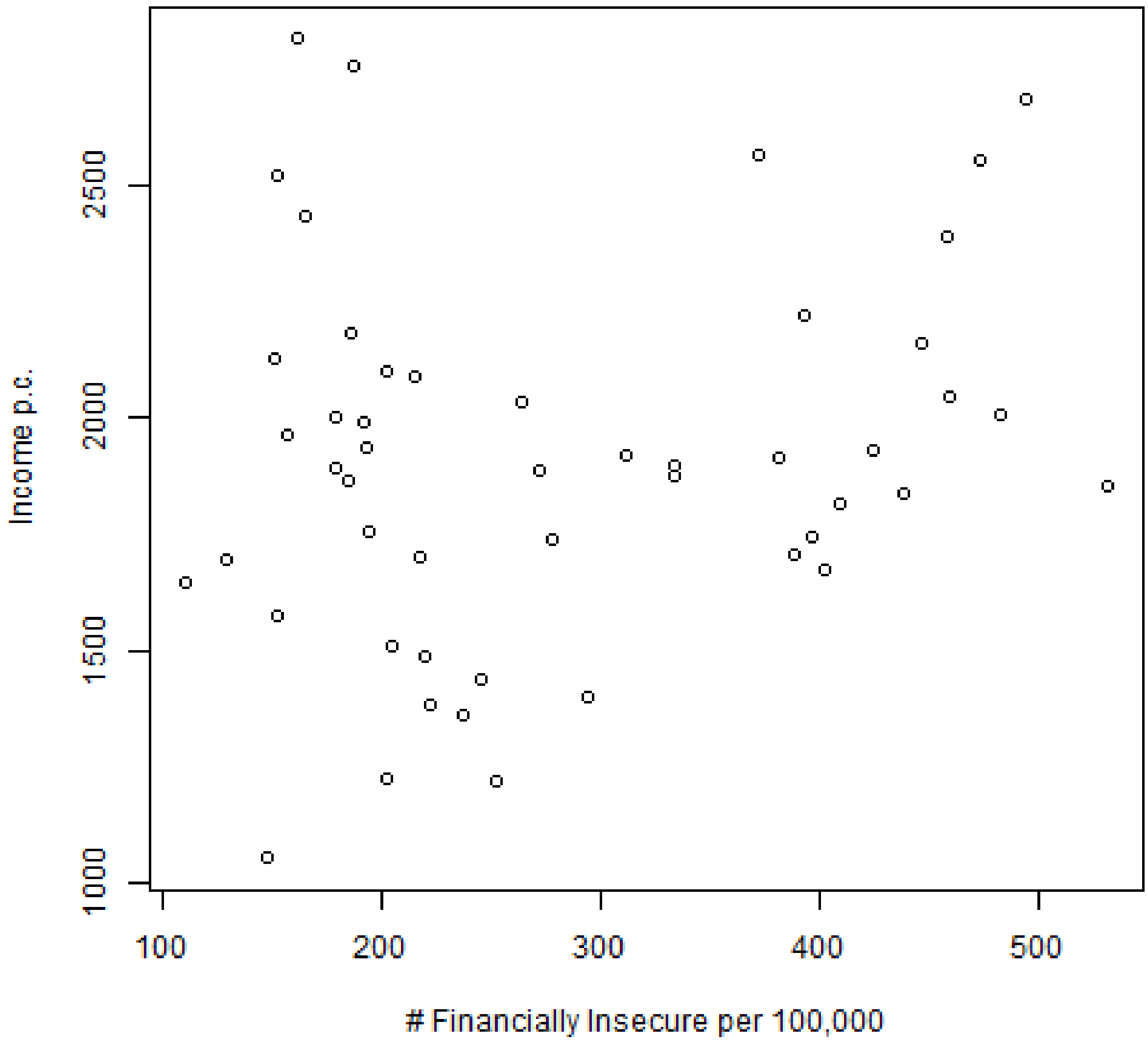


Figure 5: The relation between X1 and X3.

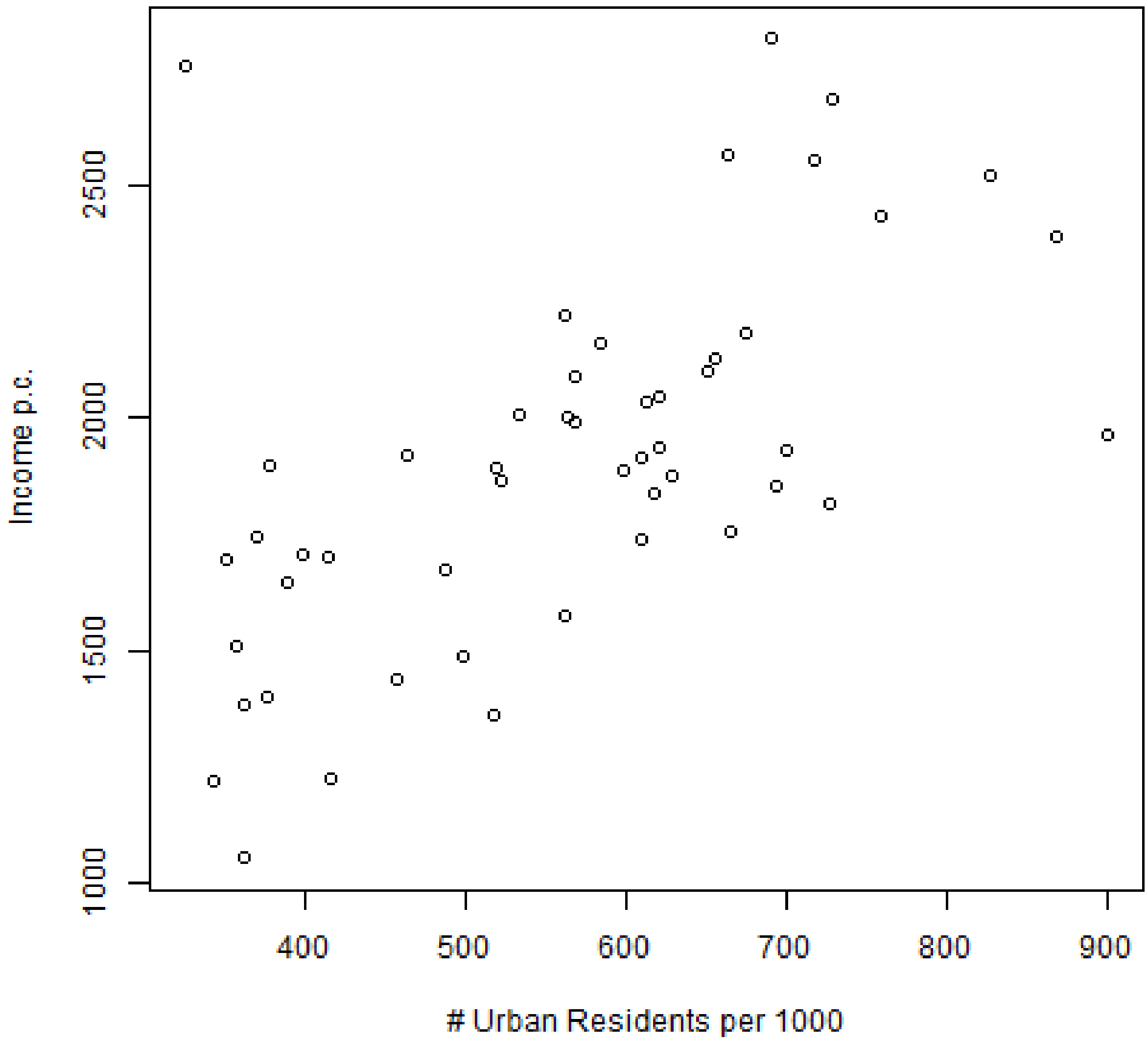


Figure 6: The relation between X2 and X3.

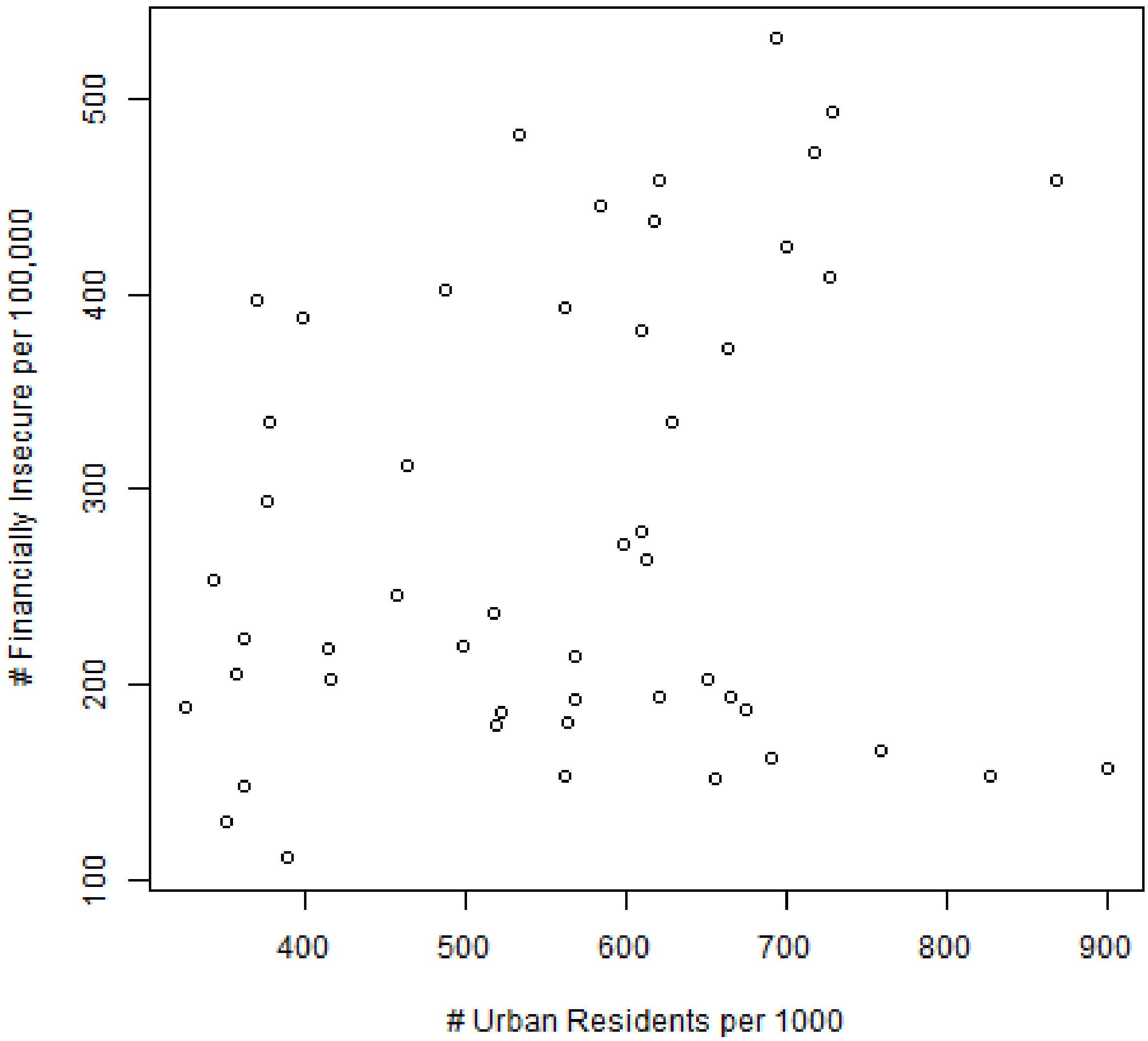


Figure 7: Expenditure by Region

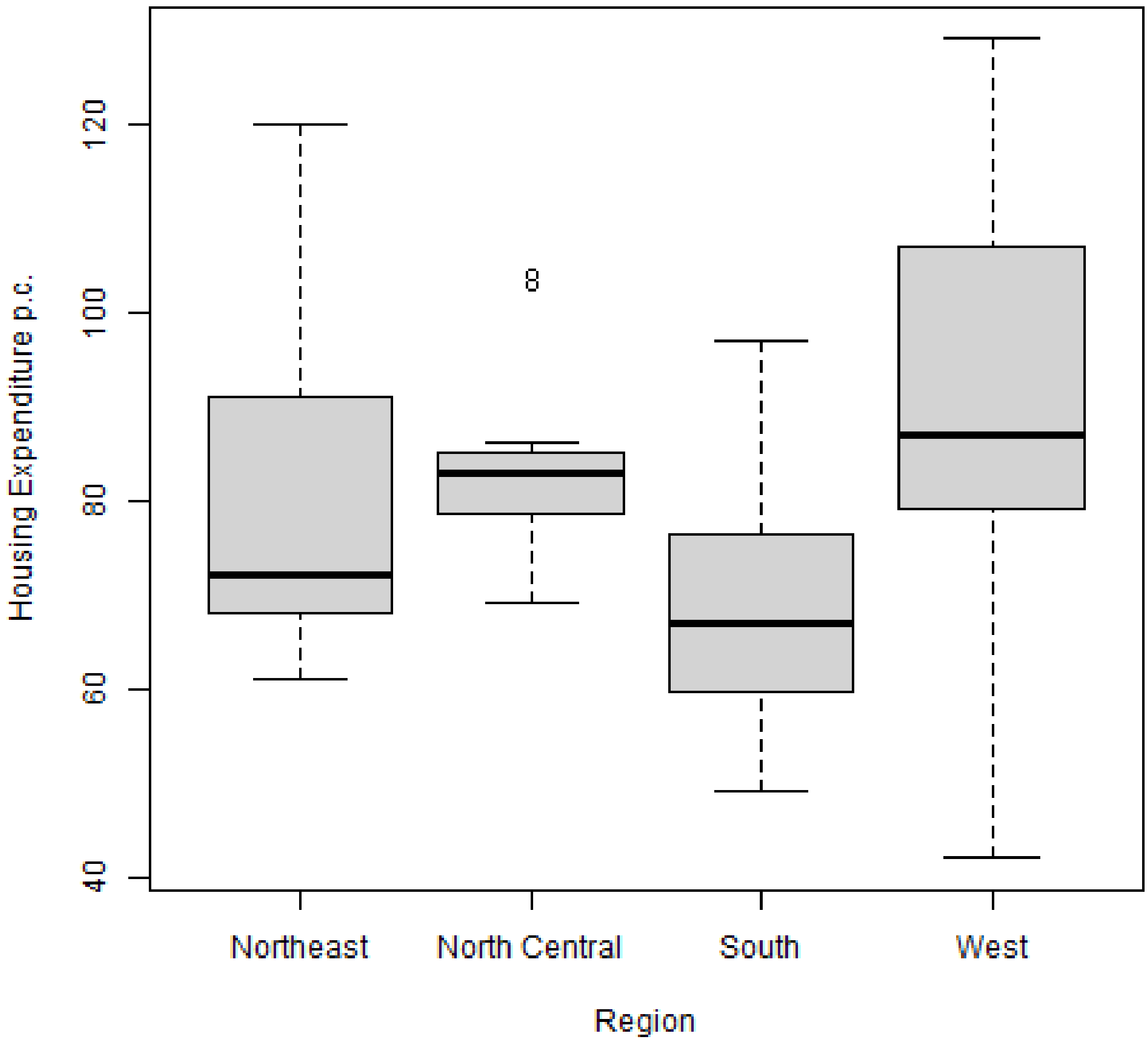
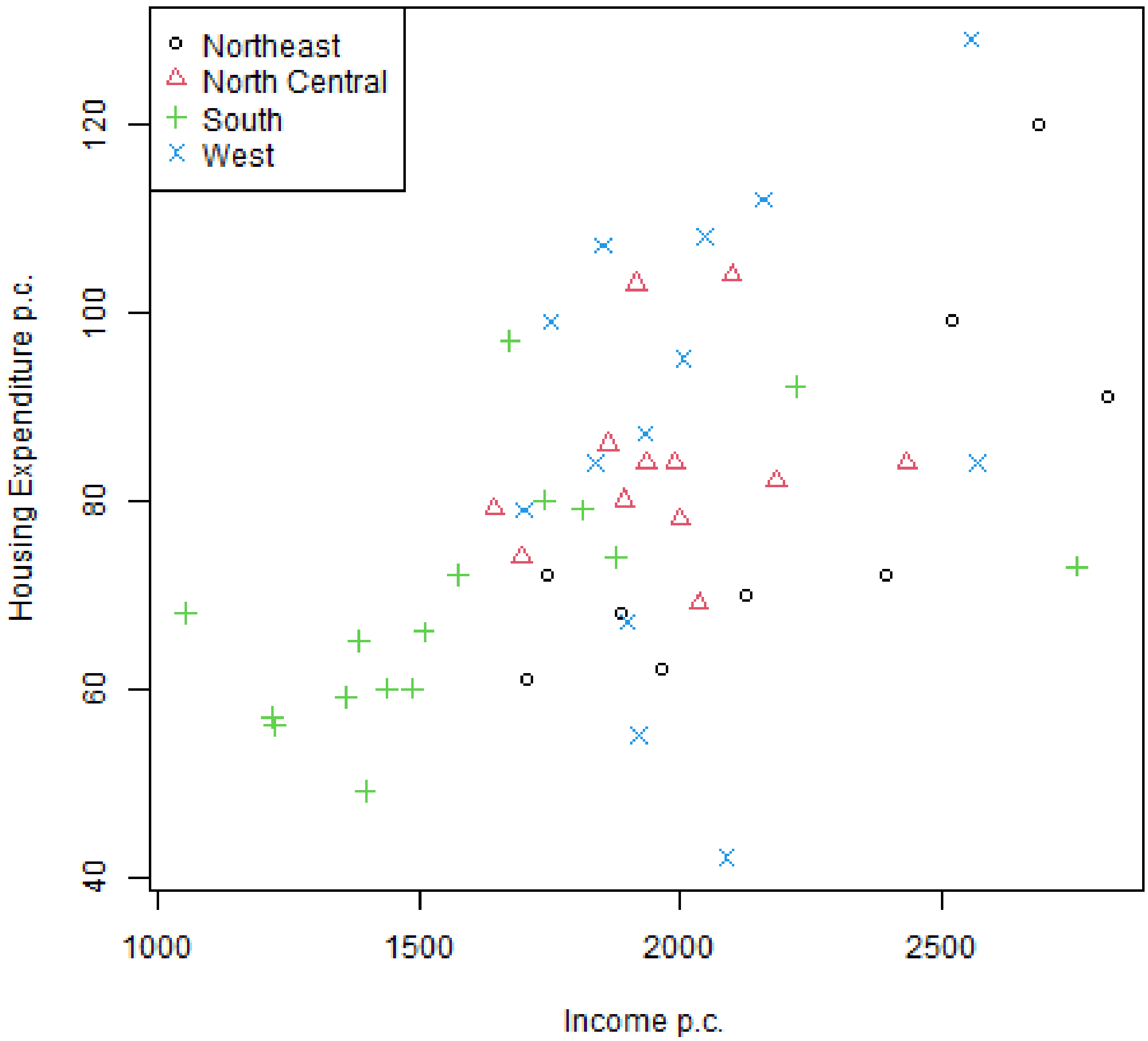


Figure 8: The relation between Y and X1, inclusive of Region.


○ Northeast
△ North Central
+ South
× West
Housing Expenditure p.c.
Income p.c.