# Problem Set 1 - Solution

Dino Wildi

Due: October 1, 2021

## 1 Question 1: Education

### 1.1 Task 1

*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)
```

In the data provided by the school counselor, we have a sample size of n = 25; this is not enough to use a normal distribution. Hence, I am using a t-distribution to account for the small sample size. In R, I am using the qt() command to find the t-score for the given data and multiply it with the estimated standard deviation to get the appropriate margin of error. I am then subtracting and adding that margin to the mean of the sample to find the bounds of the confidence interval:

```
tscore <- qt(p=0.1/2, df = 24, lower.tail = F)
margin <- tscore * (sd(y)/sqrt(25))
lower_ci <- mean(y) - margin
upper_ci <- mean(y) + margin
```

According to this calculation, the confidence interval at the significance level alpha = 90% is **[93.960, 102.920]**, with the estimated mean being **98.44**.

## 1.2 Task 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* $\alpha = 0.05$.

The second question asks for a hypothesis test for the hypothesis that the mean level of intelligence at the school is higher than the average intelligence level in the country, which is stated to be 100. Formally speaking, the alternative hypothesis and the corresponding null hypothesis read:

$$HA : \mu > 100$$

$$H0 : \mu \leq 100$$

We try to reject the null hypothesis with a one-sided t-test, with the alternative hypothesis being specified as "greater" according to the direction of the sign in HA. The confidence level we are specifying is 95%, i.e. we seek a p-value below 0.05.

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

The results of the t-test are clearly not sufficient to reject the null hypothesis. The estimated mean is **98.44**, below the hypothesised mean of 100. The p-value for the null hypothesis is **0.72**, far above the specified confidence level of p = 0.05. We have to assume that the true mean intelligence of students at the counselor's school is not above 100.

# 2 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`****.*

```
expenditure <- read.table("https://raw.githubusercontent.com/
    ASDS-TCD/StatsI_Fall2021/main/datasets/expenditure.txt",
    header=T)
```

## 2.1 Task 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)?*

The first task asks for the associations between Y and the three X-variables given in the expenditure dataset. I have created the six plots for all respective associations using ggplot, both with a scatterplot and a line delineating the association of the two variables in each individual plot. The results are represented in Figure 1, with the top row representing the individual associations between Y and each X; the bottom row represents the correlations of the individual X variables.

```
y.x1 <- ggplot(data = expenditure, aes(x = X1, y = Y)) + geom_
    point() + geom_smooth(method = "lm") +
 xlab("Personal income per capita") + ylab("Expenditure on
    shelters/housing assistance per capita")
y.x2 <- ggplot(data = expenditure, aes(x = X2, y = Y)) + geom_
    point() + geom_smooth(method = "lm") +
  xlab("Financially insecure people per 100'000") + ylab("
    Expenditure on shelters/housing assistance per capita")
y.x3 <- ggplot(data = expenditure, aes(x = X3, y = Y)) + geom_
    point() + geom_smooth(method = "lm") +
```

```
  xlab("People residing in urban areas per 100'000") + ylab("
    Expenditure on shelters/housing assistance per capita")
x1.x2 <- ggplot(data = expenditure,aes(x = X1, y = X2)) + geom_
    point() + geom_smooth(method = "lm") +
  xlab("Personal income per capita") + ylab("Financially
    insecure people per 100'000")
x1.x3 <- ggplot(data = expenditure,aes(x = X1, y = X3)) + geom_
    point() + geom_smooth(method = "lm") +
  xlab("Personal income per capita") + ylab("People residing in
    urban areas per 100'000")
x2.x3 <- ggplot(data = expenditure,aes(x = X2, y = X3)) + geom_
    point() + geom_smooth(method = "lm") +
  xlab("Financially insecure people per 100'000") + ylab("People
    residing in urban areas per 100'000")


plot_grid(y.x1, y.x2, y.x3, x1.x2, x1.x3, x2.x3, nrow = 2)
```

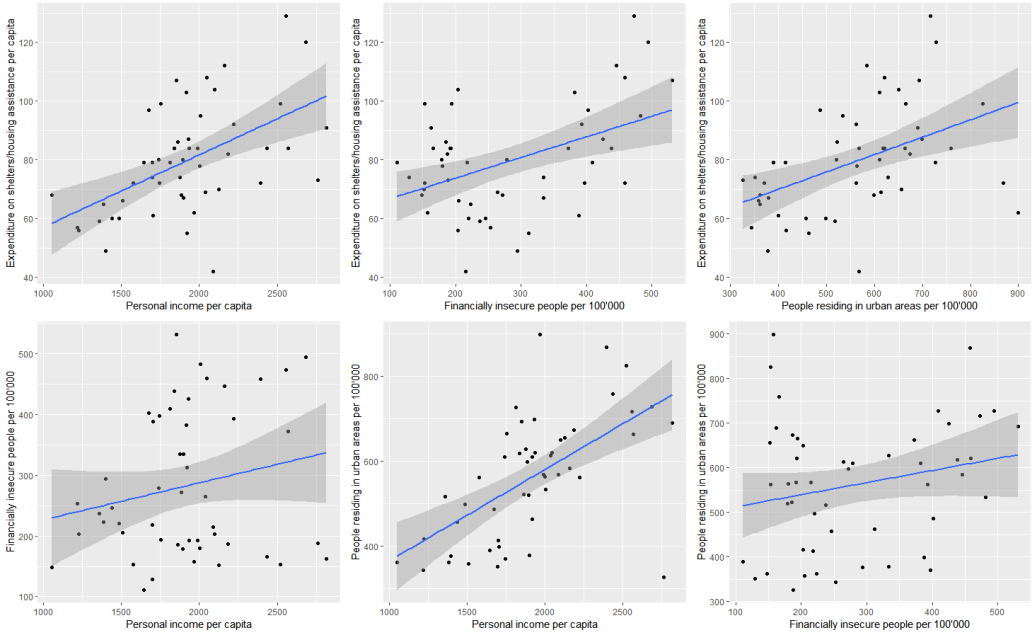


Figure 1: Relations between Y, X1, X2, and X3

In the top row, we can observe a positive correlation between the expenditure per capita on shelters and housing assistance (i.e. Y) and each individual X. The effect is largest for X1, personal income per capita. The confidence intervals plotted alongside the regression line also indicate that the effect is unlikely to be purely random, as they are relatively narrow alongside the regression line.

In the bottom row, we see three plots outlining the correlations between individual explanatory variables. Here, we would ideally observe no correlations whatsoever. However, we see a strong association between personal income per capita (X1) and people residing in urban areas per 100'000 (X3). This suggests that the two variables are not independent of each other, and including both of them in a regression model might confound results. Both plots involving the amount of financially insecure individuals per 100'000 (X2) show only a small positive correlation; in both cases, the upper bound of the confidence interval at the lower end of the curve overlaps with the lower bound of the confidence interval at the upper end of the curve. This indicates that any correlation between those variables is small, and unlikely to lead to large issues.

## 2.2 Task 2

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

The second task asks to plot expenditures for shelters and housing assistance, i.e. Y, across the regions of the United States. For this purpose, I have chosen a box plot to show both the means and the spread of expenditures in the four different regions.

```
plot2 <- ggplot(data = expenditure, aes(x = as.factor(Region), y
    = Y)) + geom_boxplot() + xlab("Region") +
  ylab("Expenditure on shelters/housing assistance per capita")
   + scale_x_discrete(labels = c("Northeast","North Central", "
   South", "West"))
```

Figure 2 shows the highest mean expenditure in Region 4, the West of the USA:

## 2.3 Task 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.*

For the third task, I plotted the correlation between expenditures for shelters and housing assistance (Y), and personal income per capita (X1), separated by region. I have added individual regression lines for each region, in the same colour as the region's observations.

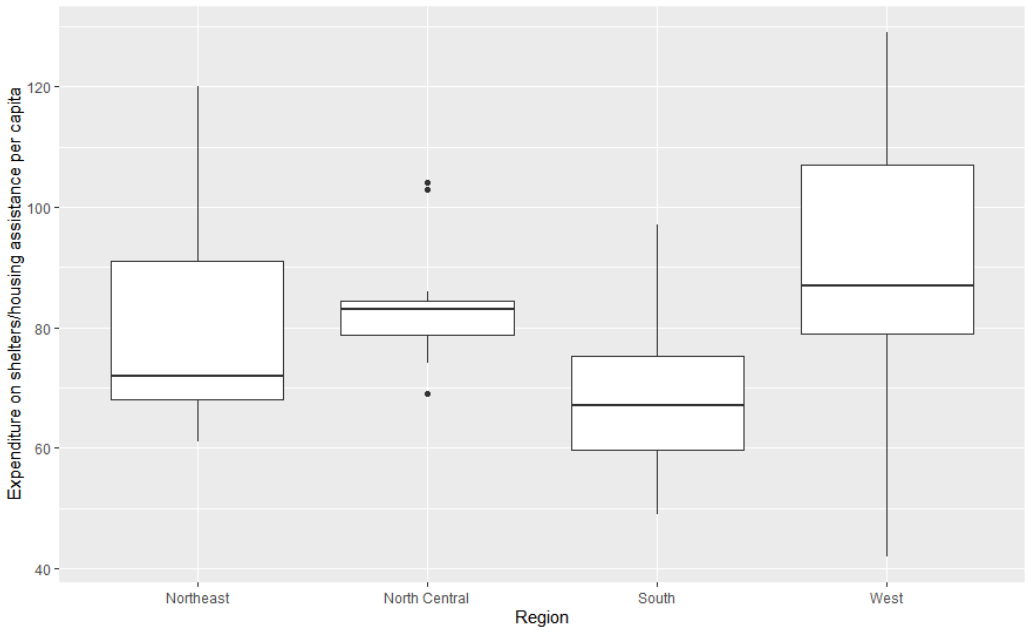


Figure 2: Per capita expenditure per region

```
plot3 <- ggplot(data = expenditure,aes(x = X1, y = Y)) + geom_
    point(aes(col = as.factor(Region), shape = as.factor(Region))
    ) +
  geom_smooth(method = "lm", se = F, aes(col = as.factor(Region)
    )) + xlab("Personal income per capita") +
  ylab("Expenditure on shelters/housing assistance per capita")
    + labs(col = "Region", shape = "Region") +
  scale_color_discrete(breaks = c(1,2,3,4), labels = c("
    Northeast","North Central", "South", "West")) +
  scale_shape_discrete(breaks = c(1,2,3,4), labels = c("
    Northeast","North Central", "South", "West"))
```

In Figure 3, we can again see the higher level of expenditure in the West (purple). Furthermore, we see that the correlation of Y and X1 is largest in the Northeastern region (red), and smallest in the North-Central region (green). We also can see that all of the states with the lowest personal income per capita; as well as most of the states with low expenditures (Y) are located in the South. Expenditures in this region are overall low, leading to a weaker effect compared to the Northeast or the West. Finally, we can see that there are a few observations in the Western region with uncharacteristically low values for Y; likely, these observations diminish the size of the effect in the West somewhat.

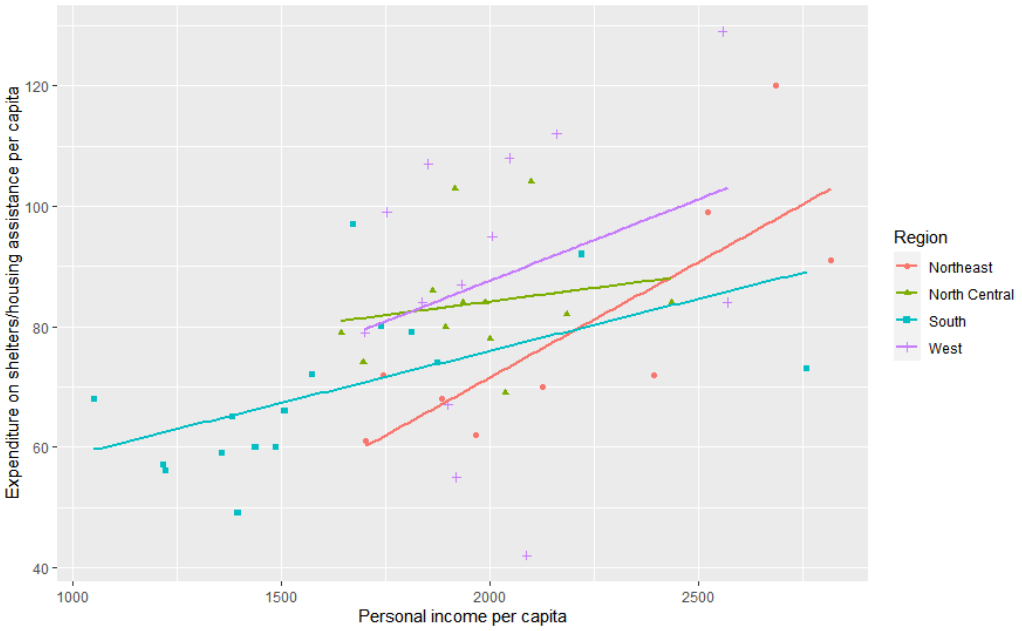


Figure 3: Per capita expenditure per region