GEOG 714 - Assignment 4

Haoran Xu

Oct 13, 2024

```
m <- matrix(runif(20), nrow = 5)
row_means <- apply(m, 1, mean)
col_means <- apply(m, 2, mean)

L <- list(a = c(1, 2, 3), b = c(4, 5), c = c(6, 5, 5, 5))
L_mean <- lapply(L, mean)
first_element <- lapply(L, function(x) x[1])
weights <- lapply(L, function(x) x[1]/mean(x))

library(here)
```

```
## here() starts at /Users/horanxu/Desktop/GEOG714_Applied_Data_Analysis_for_Geographers_and_Earth_Scie
```

```
data <- data.frame(read.csv(paste0(here(), "/Assignments/Assignment4/Assignment 3 data.csv")))
summary(data$CO2.emissions..metric.tons.per.capita.)
summary(data$Fossil.fuel.energy.consumption....of.total.)
summary(data$Energy.use..metric.tonne.of.oil.equivalent.per.capita.)
hist(data$CO2.emissions..metric.tons.per.capita.)
```

```
hist(data$Fossil.fuel.energy.consumption....of.total.)
```

```
hist(data$Energy.use..metric.tonne.of.oil.equivalent.per.capita.)
```

```
## Normal Distribution tesing Shapiro-Wilk Test
shapiro_result <- shapiro.test(data$CO2.emissions..metric.tons.per.capita.)
print(shapiro_result)  # p-value  0.05: Reject the null hypothesis (data is not normally distributed).
```

Q1. Briefly reflect on whether or not any of these three variables have a 'Normal' distribution. Use both the summary statistics and the histograms to make your case.

The `CO2.emissions..metric.tons.per.capita.` variable does not conform to normal distribution. The data skews right as the mean (5.79) is greater than median (3.13), and the maximum value (61.99) is much larger than the 3rd quartile (7.87), which suggests extreme outliers or a long right tail.

The `Fossil.fuel.energy.consumption....of.total.` variable does not conform to normal distribution. From the histogram, the data has two peaks on both sides. And the mean (59.26) is lower than the median (72.37), suggesting a left skew. Moreover, the 1st Quantile is much lower than the median and third quartile, indicating a possible concentration of values either in the lower or higher range.

The `Energy.use..metric.tonne.of.oil.equivalent.per.capita.` variable does not conform to normal distribution. It has a similar histogram as the `CO2.emissions..metric.tons.per.capita.` variable. The data skews right as the mean (2.46) is greater than median (1.23), and the maximum value (19.91) is much larger than the 3rd quartile (3.15), which suggests extreme outliers or a long right tail.

```
log_co2_emissions <- log(data$CO2.emissions..metric.tons.per.capita.)
log_ffuel <- log(data$Fossil.fuel.energy.consumption....of.total.)
log_energy <- log(data$Energy.use..metric.tonne.of.oil.equivalent.per.capita.)
hist(log_co2_emissions)
```

```
hist(log_ffuel)
```

```
hist(log_energy)
```

```
## Normal Distribution tesing Shapiro-Wilk Test
shapiro_result1 <- shapiro.test(log_co2_emissions)
print(shapiro_result1)  # p-value  0.05: Reject the null hypothesis (data is not normally distributed).
log_ffuel_clean <- log_ffuel[!is.na(log_ffuel) & !is.infinite(log_ffuel)]  # As `data$Fossil.fuel.energ
shapiro_result2 <- shapiro.test(log_ffuel_clean)
print(shapiro_result2)
shapiro_result3 <- shapiro.test(log_energy)
print(shapiro_result3)
```

Q2. Write code to delete the record corresponding to 'Small states'. Note: you must write this code in base R without using any external libraries or functions.

```
data <- data[data$Country_Name != "Small states", ]
```

```
cor.test(log_co2_emissions, log_energy, method = "pearson")  # Measures linear correlation
cor.test(log_co2_emissions, log_energy, method = "spearman")  # Measures monotonic correlation
```

Q3. Do a little independent searching on the web, and explain (in no more than three sentences, and in simple language) the difference between the Pearson and Spearman correlation coefficients.

Pearson correlation measures if there exists a linear relationship between two variables, assuming the data is normally distributed. In contrast, the Spearman correlation measures if there exists a monotonic relationship (increasing or decreasing trend) without requiring the data to be linear or normally distributed. Pearson focuses on linear trends while Spearman is more flexible.

Q4. Make a scatter plot (using plot()) of the log energy and log CO2 emissions variables. Be sure to label the plot.

```
plot(log_energy, log_co2_emissions, main = "relationships betwenen logged energy use per capita and log
    xlab = "Energy use equivalent per capita (tonne of oil)", ylab = "CO2 emissions per capita (tons)")
```

```r
lm(log_co2_emissions ~ log_energy)
summary(lm(log_co2_emissions ~ log_energy))
# Multiple R-squared doesn't adjust for the number of predictors, so it may
# overestimate the model's explanatory power when there are many variables.
# Adjusted R-squared adjusts for the number of predictors, providing a more
# reliable measure, especially in models with many variables.
```

Q5. Use the lm() function to predict CO2 (dependent variable) with energy (independent variable) in their natural (non-logged) form

```r
plot(data$CO2.emissions..metric.tons.per.capita. ~ data$Energy.use..metric.tonne.of.oil.equivalent.per.
```

```r
summary(lm(data$CO2.emissions..metric.tons.per.capita. ~ data$Energy.use..metric.tonne.of.oil.equivalen
```

```r
data[data$Country_Name == "Canada", 2]
data[data$Country_Name == "Canada", 5]
```

Q6. Write code to predict the CO2 emissions for all of Canada if there was a two unit increase in energy use per capita. Assume that Canada's population is 40 million people.

```r
B0 <- 0.36686
Bx1 <- 2.20205
x1 <- 8.365201 + 2
prediction <- (B0 + Bx1 * x1) * 4e+07
print(prediction)  #
```

Q7. Find some data (a dependent and independent variable) and use lm() to analyse their relationship. Briefly describe the data, put the regression results in a table, and offer a brief interpretation.

```r
library(here)
pitchfork_albums <- data.frame(read.csv(paste0(here(), "/Assignments/Assignment4/pitchfork_reviews.csv"

pitchfork_albums_updated <- pitchfork_albums[pitchfork_albums$score != "Not Available" &
    !(pitchfork_albums$year == "Not Available"), ]
pitchfork_albums_updated$cnt <- 1
pitchfork_albums_updated$score <- as.numeric(pitchfork_albums_updated$score)
pitchfork_albums_updated$year <- as.numeric(pitchfork_albums_updated$year)

summary(lm(pitchfork_albums_updated$score ~ pitchfork_albums_updated$year))
```

In the last assignment I did a ANOVA with the Pitchfork datasets, which are 25708 album records rated by the American online music publication Pitchfork. The albums in the datasets were released from 1952 to 2023 and given a score between 0.0 to 10.0.

In this assignment, I did a simple linear regression with the `year` of the albums and the `score` of the albums. The intercept (10.9879) suggests the predicted album score would be around 10.99 if given the year 0. The slope (-0.0019) suggests a very slight decrease in album score over time. Specifically, for each additional year, the album's score decreases by about 0.0019 points. The p-value for the slope is 0.0625, which is slightly above the 0.05. This means the relationship between year and score is not statistically significant at the 95% confidence level, though it is significant at the 90% confidence level (p-value < 0.10). Comparing the multiple R-squared (0.00014) and the adjusted R-squared (9.996e-05), they are both very low and close to 0, which means that the year of release explains only 0.014% of the variation in album scores.

The results suggest that the year of release has no strong influence on the score of an album, as shown with very low R-squared value and non-significant p-value. If given better time-frame accuracy (which means the if use "month" or "date" as the time variable of a year), or given more sample data (especially those before the year 1998), the model regression would be better.