GEOG 714 - Assignment 1

Haoran Xu

Sep 22, 2024

```
m <- matrix(1:9, nrow = 3)
elements <- m[2, 3]
m[1, ] <- c(10, 20, 30)
# m <- matrix(1:9, nrow = 3, byrow = TRUE) # if you want to do row-major order.
```

Q1. Describe what is happening in the code to the right.

Create a matrix of three rows using integer 1 to 9 in the default column-major order. Index the 2nd row and 3rd column number in the m matrix and assign the value to elements. Re-assign the values for the first row of m matrix with 10, 20, 30.

```
# Create a list called L (three vectors)
L <- list(number = 1:10, characters = letters[1:5], logical = c(TRUE, FALSE)  # TRUE/FALSE should be in
)
L
L[1]
L[[1]][1]  # first 1 means the number of vector, then the second means element
L[[1]]
L[[1]][which(L[[3]][1] == TRUE)]  # which() function returns the index rather than the value, here it r
print(which(L[[3]][2] == TRUE))  # return 'integer(0)'
```

```
# install.packages('curl')
library(curl)
```

```
## Using libcurl 8.7.1 with LibreSSL/3.3.6
```

```
url <- "https://datazone.healthgeomatics.com//Hawaii_CO2.csv"
data <- read.csv(url)

pre <- data[data$Year < 1980, ]
post <- data[data$Year >= 1980, ]
data$CO2
summary(data$CO2)
avg_m <- summary(data$CO2)[4]
```

Q2. Do some research online and find another function that can be used to calculate some simple statistics for the CO2 variable. Don't install any new packages to accomplish this. Write the code to use that function.

```r
cor(data)
apply(data, 2, median)
table(data$CO2)  # probability
prop.table(table(data$CO2))  # relative pro
aggregate(pre$CO2, by = list(pre$Year), FUN = mean)  # calculateing mena
```

```r
hist(data$CO2)
```

```r
data$diff <- data$CO2 - avg_m
plot(seq(1:161), data$CO2, main = "Monthly CO2 concentrations May 1974 - September 1987, Hawaii",
    xlab = "Time", ylab = expression(CO[2] * " (ppm)"))  # use 'expression to make the '2' subscripted'
```

Q3. Write your own code to create a plot in which the Y axis is the 'diff' variable you created a few steps ago. Ensure that the y axis is labeled "CO2 (ppm difference from series average)"

```r
plot(seq(1:161), data$diff, main = "Monthly CO2 concentrations May 1974 - September 1987, Hawaii",
    xlab = "Time", ylab = expression(CO[2] * " (ppm difference from series average)"))
```

```r
data <- data[data$Year >= 1975 & data$Year <= 1986, ]
agg <- aggregate(data$CO2, by = list(data$Year), FUN = mean)
names(agg) <- c("Year", "CO2")
plot(agg$Year, agg$CO2, main = "Annual CO2 concentrations May 1974 - September 1987, Hawaii",
    xlab = "Year", ylab = expression(CO[2] * " (ppm)"))
```

```r
agg2 <- aggregate(data$CO2, by = list(data$Month), FUN = mean)
names(agg2) <- c("Month", "CO2")
barplot(agg2$CO2, main = "Monthly CO2 concentrations", xlab = "Month (1975 to 1986 aggregated)")  #   pl
```

```r
barplot(agg2$CO2, main = "Monthly CO2 concentrations", xlab = "Month (1975 to 1986 aggregated)",
    ylab = "CO2 (ppm)", ylim = c(300, 350), xpd = FALSE, names.arg = c("J", "F",
        "M", "A", "M", "J", "J", "A", "S", "O", "N", "D"))
```

```r
# ylim sets the limit, xpd means the bars stop at the bottom of the y-axis
```

```r
data2 <- data[data$Month == 4 | data$Month == 10, ]
barplot(data2$CO2, main = "Monthly CO2 concentrations", xlab = "April and October (1975 to 1986 aggrega
    ylab = "CO2 (ppm)", ylim = c(300, 350), xpd = FALSE, col = rep(c("black", "white"),
        12))
legend("topleft", legend = c("April", "October"), fill = c("black", "white"))
```

Q4. Do you think a bar plot of CO2 measurements is the best way of representing this information? What do you think is a better way of visually representing mean monthly CO2 concentrations? Explain your answer (two to three sentences)

A bar plot may not effectively show continuous trends in CO2 levels. A smooth curve plot may better highlight seasonal fluctuations and long-term patterns, which would show better how CO2 levels change over time.

Q5. The data in this series goes until September 1987. If you had to make predictions about the CO2 concentrations in October, November and December, what would you do? Come up with some code that makes a prediction of the remainder of 1987 CO2 concentrations. The predictions do not have to be perfect, and you don't need to do any fancy modeling at this stage—keep the procedure as simple as you want provided that they generate a prediction better than random chance. Explain your reasoning. Importantly, do not import or use any new libraries for this analysis; only use base R functions.

I would use a simple linear regression model to predict CO2 in future years.

```
model <- lm(CO2 ~ Year + Month, data = data)  # create a linear regression model (Year represents yearl
future_data <- data.frame(Year = rep(1987, 3), Month = c(10, 11, 12))  # predict three months in 1987
predictions <- predict(model, newdata = future_data)  # 'model' is the model we use, this returns the p
predictions
```

```
url2 <- "https://datazone.healthgeomatics.com//JanMayen.csv"
data_JMIsland <- read.csv(url2)
```

Q6. Keep data only from 1940 to 2001

```
data_JMIsland <- data_JMIsland[data_JMIsland$Year >= 1940 & data_JMIsland$Year <=
    2001, ]
```

Q7. Calculate summary statistics for the Temperature variable using the summary() function

```
summary(data_JMIsland$Temperature)
```

Q8. Create a histogram of Temperature

```
hist(data_JMIsland$Temperature)
```

Q9. Plot the December average temperatures from 1940 to 2001, and ensure that the plot is properly labeled. Ensure that temperature is on the y axis and year is on the x axis. Don't worry about the tick mark labels for the years on the x-axis

```
december_data <- data_JMIsland[data_JMIsland$Month == 12, ]
plot(december_data$Year, december_data$Temperature, main = "Annual December average temperature 1940-20
    xlab = "Year", ylab = "Temperature (celsius)")
```

Q10. Calculate the annual average temperatures (by aggregating on year) from 1940 to 2001 and put the result in a dataframe called agg

```
agg <- aggregate(data_JMIsland$Temperature, by = list(data_JMIsland$Year), FUN = mean)
names(agg) <- c("Year", "Ave_t")
```

Q11. Plot the average temperatures by year on a plot.

```
plot(agg$Year, agg$Ave_t, main = "Average temperatures by year", xlab = "Year", ylab = "Temperature (ce
```

Q12. Write code to delete any records with age equal to 99 or internet use values equal to 8 or 9.

```
df <- data.frame(read.csv("/Users/horanxu/Library/CloudStorage/OneDrive-Personal/GEOG 714 - Applied Dat
df <- df[df$Age != 99 & !(df$Internet_use %in% c(8, 9)), ]  # %in% ('in') to check if the elements
```

```
t.test(df$Age ~ df$Internet_use)
```

```
# install.packages('weights')
library(weights)
```

```
## Loading required package: Hmisc
```

```
##
## Attaching package: 'Hmisc'
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
df1 <- df[df$Internet_use == 1, ]
df2 <- df[df$Internet_use == 2, ]
wtd.t.test(x = df1$Age, y = df2$Age, weight = df1$Weight, weighty = df2$Weight, samedata = FALSE)
```

Q13. Find some data online and use the t-test() function in R. Provide a brief description of your reasoning and interpretation of the results of your analysis.

Pitchfork is an American online music publication founded in 1996 by Ryan Schreiber in Minneapolis. I have been keeping reading the album reviews on its webstie for six more years. For this project I used "Pitchfork Reviews: Music Critiques Over the Years" Kaggle dataset uploaded by Tim Stafford.

cleaning pitchfork-rated album dataset

```
pitchfork_albums <- data.frame(read.csv("/Users/horanxu/Library/CloudStorage/OneDrive-Personal/GEOG 714
pitchfork_albums_updated <- pitchfork_albums[pitchfork_albums$score != "Not Available" &
    !(pitchfork_albums$year == "Not Available"), ]
pitchfork_albums_updated$score <- as.numeric(pitchfork_albums_updated$score)
agg_albums <- aggregate(pitchfork_albums_updated$score, by = list(pitchfork_albums_updated$year),
    FUN = , mean)
names(agg_albums) <- c("Year", "ave_scores")
plot(agg_albums$Year, agg_albums$ave_scores)
```

t-test

My null hypothesis is: the scored of albums in 2020 and 2021 rated by Pitchfork do not have a significant difference.

```
pitchfork_albums_2020_2021 <- pitchfork_albums[pitchfork_albums$year == "2020" |
    pitchfork_albums$year == "2021", ]
pitchfork_albums_2020_2021$score <- as.numeric(pitchfork_albums_2020_2021$score)
t.test(pitchfork_albums_2020_2021$score ~ pitchfork_albums_2020_2021$year)
```

From the t-test results, the p-value is 0.9566, the mean scores of albums in 2020 and 2021 are respectively 7.250 and 7.248, this means there's a high probability that the null hypothesis is true. So it means the scored of albums in 2020 and 2021 rated by Pitchfork do not have a significant difference.