GEOG 714 - Assignment 3

Haoran Xu

Sep 29, 2024

```
# hi!  don't know why this can't be detected!

a <- 5
change <- function(x) {
    return(x + a)
}
change(5)
# print(x) # this will generate an error
```

```
a <- 5
x <- 6
change <- function(x) {
    y <- 5
    return(x + a + y)
}
change(5)
print(x)
# print(y) # this will generate an error
```

```
stuff <- 1
blah <- function() {
    stuff <- 5
}
print(stuff)
blah()
print(stuff)

# compare below:
stuff <- 1
blah <- function() {
    stuff <<- 5  # '<<-' assignment operator can be used to change global variables from within functio
}
print(stuff)
blah()
print(stuff)
```

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

```
## here() starts at /Users/horanxu/Desktop/Assignment/Assignment3
```

```r
df <- read.csv(paste0(here(), "/data/Canada2006_WVS - Sheet1.csv"), stringsAsFactors = FALSE)

names(df) <- c("respondent", "happiness", "healthg", "friends", "satisfaction", "membership",
    "science", "age", "size", "weight")

m <- matrix(c(1, 4, 3, 2, -1, 5, 3, 4, -1), ncol = 3, nrow = 3)
apply(m, 1, function(x) x < 0)  # return TRUE/FALSE
m[apply(m, 1, function(x) x < 0)] <- 0
```

```r
df[df < 0] <- NA  # no need add quotation marks
length(df$healthg[is.na(df$healthg) == TRUE])
```

Q1. Write a two to three sentence explanation about what is happening in this code. You may need to read up on the length() and is.na() functions.

The code uses `is.na()` to identify missing values in the `healthg` column and applies `length()` to count the total number of `NA` entries in that column. So it can calcutes the number of missing values of `df$healthg`.

```r
df <- df[apply(df, 1, function(x) !any(is.na(x))), ]  # return TRUE/FALSE
# The innermost function is.na() returns TRUE if a value is NA.  The any()
# function returns TRUE if any value inside the row is TRUE.  The '!' operator
# reverses any() so that it is FALSE if any value in the row is TRUE.  If any
# value in any row is NA, then that row is not included in the output file.

df$happiness <- as.factor(df$happiness)
```

Q2. Write code to convert all of these variables (healthg, friends, satisfaction, membership, science,size) into factor variables

```r
df$healthg <- as.factor(df$healthg)
df$friends <- as.factor(df$friends)
df$satisfaction <- as.factor(df$satisfaction)
df$membership <- as.factor(df$membership)
df$science <- as.factor(df$science)
df$size <- as.factor(df$size)
```

```r
unique(df$happiness)  # used in factors
unique(df$healthg)

df$happiness <- as.character(df$happiness)
df$healthg <- as.character(df$healthg)

df$happiness[df$happiness == "1"] <- "1. Very happy"
df$happiness[df$happiness == "2"] <- "2. Quite happy"
df$happiness[df$happiness == "3"] <- "3. Not very happy"
df$happiness[df$happiness == "4"] <- "4. Not at all happy"

df$healthg[df$healthg == "1"] <- "1. Very good"
df$healthg[df$healthg == "2"] <- "2. Good"
```

```r
df$healthg[df$healthg == "3"] <- "3. Fair"
df$healthg[df$healthg == "4"] <- "4. Poor"

df$happiness <- as.factor(df$happiness)
df$healthg <- as.factor(df$healthg)

table(df$happiness, df$healthg)
```

Q3. Write code to make this change.

```r
df$happiness <- as.character(df$happiness)
df$happiness[df$happiness == "3. Not very happy" | df$happiness == "4. Not at all happy"] <- "3. Not ve
df$happiness <- as.factor(df$happiness)
table(df$happiness, df$healthg)
```

```r
chisq.test(df$happiness, df$healthg, simulate.p.value = TRUE)$expected
# install.packages('weights')
library(weights)
```

```
## Loading required package: Hmisc

##
## Attaching package: 'Hmisc'

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

```r
wtd.chi.sq(df$happiness, df$healthg, weight = df$weight)
```

Q4. Provide a short written interpretation of this result (no more than two sentences).

The results showed p-value is less than 0.0001. This assumes that the null hypothesis is rejected, which means the two categorical variables are not independent.

```r
boxplot(df$age ~ df$healthg)
```

```r
summary(aov(df$age ~ df$healthg))
```

Q5. Provide a brief interpretation of this result. One of the assumptions of ANOVA is that the dependent variable (age in this case) is normally distributed. Is this a reasonable assumption? Justify your answer in no more than 2 sentences total.

The ANOVA results indicate that there is a significant difference in the mean `ages` among the different health levels (`healthg`), as the p-value is less than 0.001. This suggests that at least one `healthg` group has a mean age that is significantly different from the others.

Regarding normal distribution, it is reasonable to assume that the age variable is normally distributed. However, in large samples (e.g., the residual degrees of freedom here is 1990), ANOVA is still robust due to the Central Limit Theorem even if the variable might not be normally distributed.

```
shapiro.test(df$age)  # test if df$age is normally distributed --> 'p < 0.001' means it's not
```

Q6. Find some data on your own. Ensure that the data has one categorical variable with at least three levels, and one continuous numeric variable. Your null hypothesis is that the numeric variable does not vary across the groups. Use one-way ANOVA or Kruskal-Wallis to analyse these data. In your answer, provide 1) a link to the data 2) a description of the data, 3) the results of your analysis and 4) an interpretation of your results. Make it look nice and pretty, and ensure it is no more than 1 page in length.

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

boxplot(pitchfork_albums_updated$score ~ pitchfork_albums_updated$year)
```

```
agg_albums1 <- aggregate(pitchfork_albums_updated$cnt, by = list(pitchfork_albums_updated$year),
    FUN = sum)

pitchfork_albums_updated$year[pitchfork_albums_updated$year <= 1995] <- 1
pitchfork_albums_updated$year[pitchfork_albums_updated$year >= 1996 & pitchfork_albums_updated$year <=
    2003] <- 2
pitchfork_albums_updated$year[pitchfork_albums_updated$year > 2003 & pitchfork_albums_updated$year <=
    2008] <- 3
pitchfork_albums_updated$year[pitchfork_albums_updated$year > 2008 & pitchfork_albums_updated$year <=
    2013] <- 4
pitchfork_albums_updated$year[pitchfork_albums_updated$year > 2013 & pitchfork_albums_updated$year <=
    2018] <- 5
pitchfork_albums_updated$year[pitchfork_albums_updated$year > 2018] <- 6

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)
```

```
agg_albums2 <- aggregate(pitchfork_albums_updated$cnt, by = list(pitchfork_albums_updated$year),
    FUN = sum)

boxplot(pitchfork_albums_updated$score ~ pitchfork_albums_updated$year)
```

```r
summary(aov(pitchfork_albums_updated$score ~ pitchfork_albums_updated$year))
```

Introduction

Pitchfork is an American online music publication founded in 1996 by Ryan Schreiber in Minneapolis. Since then, it began to gain popularity among indie music fans and till now it has become a professionalmusic publication loved by a lot of people.

In this project I used "Pitchfork Reviews: Music Critiques Over the Years" Kaggle dataset uploaded by Tim Stafford.

Data Preparation and Descriptive Analysics

This dataset has 25708 rated album records that were released from 1952 to 2023. First, I screened out 24699 album records that do not have missing information about scores or years. Then I aggregated the album numbers by their released years and found that before 2002, the number of albums rated each year increased with years. And after 2002, the albums rated each year range stably from 850 to 1250 (which means averagely they publish three to four reviews every day). I chose 1996 and 2002 as two important years and since recategorize the continuous `year` variable into a 1-6 categorical variable, which are <= 1995, 1996-2003, 2004-2008, 2009-2013, 2014-2018, 2019-2023).

ANOVA Aanalysis

I did a ANOVA test with the continuous `score` variable and the categorical `year` variable. The result shows that the `p-value` is less than 0.001, which means there exist significant score differences between groups. According to the box plot, it is obviously that the scores in the "before 1996" group are significantly higher than the other gourps, with an average score of 8.7/10.0. This is partly because all the albums released before 1996 that Pitchfork rated were not contemporary albums. They either use their "Sunday Review" section to review the albums from the past once a week, or they would review those legacy albums after one artist make anniversity reissues (e.g., Aphex Twin, Joni Mitchell) or deceased (e.g., David Bowie, Prince). These "old" albums are typically rated higher.

The other interesting phenomenon is the scores given were gradually increasing since 1996. And along with it, the variances between album scores in a single year are decreasing. This would mean that Pitchfork does not give that many incredibly "lower" scores ($< 4.0/10.0$), neither does it give those "super high" scores that often ($> 9.5/10.0$). This is how Pitchfork has been evolving, that it is not that bold and brashy as it used me, instead, it became more "conservative" and "safe" in recent years.