Statistics with R - Exercise 2

Philipp Satlawa - h0640348

04/12/2020

This document contains the answered questions of exercise 2 for the course "Statistics with R".

Task 1 - Robustness of mean/median

Task 2 - Estimation

Task 3 - Working with a real data set

```
#load library data.table
library("data.table")
```

1. Import data from different sources

import from file .Rdata (R specific file type)

```
# load data from file
load("~/Documents/boku/statistics_with_R/ex_02/CO2.Rdata")
# convert data.frame to data.table
datCO2 <- setDT(dat)
# check if data was properly loaded
str(dat)
```

```
## Classes 'data.table' and 'data.frame':   1619494 obs. of  6 variables:
##  $ unit     : Factor w/ 2 levels "MIO_T","THS_T": 1 1 1 1 1 1 1 1 1 1 ...
##  $ airpol   : Factor w/ 11 levels "CH4","CH4_CO2E",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ airemsect: Factor w/ 172 levels "CRF1","CRF1-6X4_MEMO",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ geo      : Factor w/ 35 levels "AT","BE","BG",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ time     : num  2017 2017 2017 2017 2017 ...
##  $ values   : num  0.02497 0.04194 0.05682 0.01095 0.00059 ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

import from file .csv (comma separated values)

import from file .txt (tabulator separated values)

import from file .xlsx (MS Excel specific file type)

2. Set seed to the student id

```r
# prepare seed
set.seed(as.numeric(format(Sys.time(), "%H%M%S")))
# set student id
id <- 640348
# set seed with student id
set.seed(id)
```

3. Data exploration

```r
# show the data structure
str(dat)
```

```
## Classes 'data.table' and 'data.frame':   1619494 obs. of  6 variables:
##  $ unit     : Factor w/ 2 levels "MIO_T","THS_T": 1 1 1 1 1 1 1 1 1 1 ...
##  $ airpol   : Factor w/ 11 levels "CH4","CH4_CO2E",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ airemsect: Factor w/ 172 levels "CRF1","CRF1-6X4_MEMO",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ geo      : Factor w/ 35 levels "AT","BE","BG",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ time     : num  2017 2017 2017 2017 2017 ...
##  $ values   : num  0.02497 0.04194 0.05682 0.01095 0.00059 ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

```r
# show the data summary
summary(dat)
```

```
##     unit              airpol            airemsect              geo         
##  MIO_T:809747   GHG     :249878   CRF1-6XMEMO:  21590   HU     :  57198  
##  THS_T:809747   CH4     :247102   CRF2       :  21374   SI     :  52588  
##                 CH4_CO2E:247102   CRF2B      :  21298   PL     :  49938  
##                 CO2     :232890   CRF2C      :  20040   RO     :  49414  
##                 N2O     :229428   CRF2B10    :  19148   ES     :  49118  
##                 N2O_CO2E:229428   CRF6       :  18722   EU28   :  48834  
##                 (Other) :183666   (Other)    :1497322   (Other):1312404  
##       time          values        
##  Min.   :1985   Min.   :-458348  
##  1st Qu.:1996   1st Qu.:      0  
##  Median :2003   Median :      0  
##  Mean   :2003   Mean   :   5235  
##  3rd Qu.:2010   3rd Qu.:      5  
##  Max.   :2017   Max.   :5729428  
##                 NA's   :5284
```

The provided dataset contains 6 variables (columns) and 1619494 records (rows). The variables have the following types: * unit : categorical Factor with 2 levels (unit abriviation) * airpol : categorical Factor with 11 levels (chemical compound) * airemsect: categorical Factor with 172 levels (code for the sector) * geo : categorical Factor with 35 levels (country abrivation) * time : discrete numerical (year) * values : continuous numerical (the value)

The variable values has 5284 NA values and is the only variable that contains NA.

4. Select randomly two countries

```
# create vector with all countries
geo_col <- datCO2[, unique(geo)]
# take 2 random samples
geo_c <- sample(geo_col, 2)
```

5. Filter data

```
# filter data
datFilter <- datCO2[unit == "THS_T" &
                      airpol == "GHG" &
                      airemsect %in% c("CRF3", "CRF31", "CRF1A3") &
                      geo %in% geo_c]
```

6. Remove columns/variables unit and airpol

```
# remove variables
datFilter <- datFilter[, -c("unit", "airpol")]
```

7. Show records per country

```
#
datFilter[, .N, by = geo]
```

```
##    geo  N
## 1:  IS 84
## 2:  LU 84
```