MATH2349 Semester 2, 2018

Assignment 1

Amal Joy (s3644794)

10 August 2018

Contents

1 Introduction

The aim of this report is to locate an open data from the web, import it into R, and reflect upon the data types, formats and structures in the selected data set. The tool used for the analysis is R.

2 Setup

The following packages are used in this report for data preparation and data exploration.

```
library(knitr) # Used for knittng the document
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.5.1
```

3 Data Description

The open source data is collected from **UCI Machine Learning Repository**. The link for downloading the data is https://archive.ics.uci.edu/ml/datasets/Contraceptive+Method+Choice. This dataset is a subset of the 1987 **National Indonesia Contraceptive Prevalence Survey**. This dataset is created by Tjen-Sien Lim. The samples are married women who were either not pregnant or do not know if they were at the time of interview. This dataset was collected to predict choice of contraceptive method of a women based on her demographic and socio-economic characteristics. According to the source of the dataset, the attribute information is given as below:

- V1 = **Wife's age** (numerical)
- V2 = **Wife's education** (categorical) 1=low, 2, 3, 4=high
- V3 = **Husband's education** (categorical) 1=low, 2, 3, 4=high
- V4 = **Number of children ever born** (numerical)
- V5 = **Wife's religion** (binary) 0=Non-Islam, 1=Islam
- V6 = **Wife's now working?** (binary) 0=Yes, 1=No
- V7 = **Husband's occupation** (categorical) 1, 2, 3, 4
- V8 = **Standard-of-living index** (categorical) 1=low, 2, 3, 4=high
- V9 = **Media exposure** (binary) 0=Good, 1=Not good
- V10 = **Contraceptive method used** (class attribute) 1=No-use, 2=Long-term, 3=Short-term

4 Read/Import Data

The data was read from the URL into a data file named 'contraceptive'. The base R function 'read.csv' is used to read the data. The first 5 elements of the data set are given below.

```
link = "https://archive.ics.uci.edu/ml/machine-learning-databases/cmc/cmc.data"
contraceptive <- read.csv(url(link), header = FALSE) # reading the data
head(contraceptive,5) # head of the dataset
```

```
##   V1 V2 V3 V4 V5 V6 V7 V8 V9 V10
## 1 24  2  3  3  1  1  2  3  0   1
## 2 45  1  3 10  1  1  3  4  0   1
## 3 43  2  3  7  1  1  3  4  0   1
## 4 42  3  2  9  1  1  3  3  0   1
## 5 36  3  3  8  1  1  3  2  0   1
```

```r
class(contraceptive) # class of the dataset
```

```
## [1] "data.frame"
```

The class of the data is a data frame. As seen from the data frame head, variable names are not properly marked in the data frame.

5 Inspect and Understand

Dimensions of the data frame are checked using the 'dim' function. Structure of the dataset is also inspected using 'str'. Relevant column names are provided as given in the source of the dataset.

```r
dim(contraceptive) # Dimensions of the dataset
```

```
## [1] 1473   10
```

```r
colnames(contraceptive) # column names
```

```
##  [1] "V1"  "V2"  "V3"  "V4"  "V5"  "V6"  "V7"  "V8"  "V9"  "V10"
```

```r
colnames(contraceptive) <-
  c("Wife_age","Wife_edu",
    "Hus_edu","Num_Children",
    "Wife_rel","Wife_work",
    "Hus_occup","std_liv",
    "media","contraceptive")

str(contraceptive) # structure of the dataset
```

```
## 'data.frame':    1473 obs. of  10 variables:
##  $ Wife_age     : int  24 45 43 42 36 19 38 21 27 45 ...
##  $ Wife_edu     : int  2 1 2 3 3 4 2 3 2 1 ...
##  $ Hus_edu      : int  3 3 3 2 3 4 3 3 3 1 ...
##  $ Num_Children : int  3 10 7 9 8 0 6 1 3 8 ...
##  $ Wife_rel     : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Wife_work    : int  1 1 1 1 1 1 1 0 1 1 ...
##  $ Hus_occup    : int  2 3 3 3 3 3 3 3 3 2 ...
##  $ std_liv      : int  3 4 4 3 2 3 2 2 4 2 ...
##  $ media        : int  0 0 0 0 0 0 0 0 0 1 ...
##  $ contraceptive: int  1 1 1 1 1 1 1 1 1 1 ...
```

```r
contraceptive$Wife_edu <- as.factor(contraceptive$Wife_edu)
contraceptive$Hus_edu <- as.factor(contraceptive$Hus_edu)
contraceptive$Wife_rel <- as.factor(contraceptive$Wife_rel)
contraceptive$Wife_work <- as.logical(contraceptive$Wife_work)
contraceptive$Hus_occup <- as.factor(contraceptive$Hus_occup)
contraceptive$std_liv <- as.factor(contraceptive$std_liv)
contraceptive$media <- as.factor(contraceptive$media)
contraceptive$contraceptive <- as.factor(contraceptive$contraceptive)

sapply(contraceptive, levels) # Checking the levels of each variable
```

```
## $Wife_age
## NULL
##
## $Wife_edu
```

```
## [1] "1" "2" "3" "4"
##
## $Hus_edu
## [1] "1" "2" "3" "4"
##
## $Num_Children
## NULL
##
## $Wife_rel
## [1] "0" "1"
##
## $Wife_work
## NULL
##
## $Hus_occup
## [1] "1" "2" "3" "4"
##
## $std_liv
## [1] "1" "2" "3" "4"
##
## $media
## [1] "0" "1"
##
## $contraceptive
## [1] "1" "2" "3"
```

```
levels(contraceptive$contraceptive) <- c("no use","long term","short term")
levels(contraceptive$Wife_rel) <- c("Non-Islam","Islam")
levels(contraceptive$media) <- c("Good","Not-Good")

str(contraceptive) # structure of the dataset
```

```
## 'data.frame':    1473 obs. of  10 variables:
##  $ Wife_age     : int  24 45 43 42 36 19 38 21 27 45 ...
##  $ Wife_edu     : Factor w/ 4 levels "1","2","3","4": 2 1 2 3 3 4 2 3 2 1 ...
##  $ Hus_edu      : Factor w/ 4 levels "1","2","3","4": 3 3 3 2 3 4 3 3 3 1 ...
##  $ Num_Children : int  3 10 7 9 8 0 6 1 3 8 ...
##  $ Wife_rel     : Factor w/ 2 levels "Non-Islam","Islam": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Wife_work    : logi  TRUE TRUE TRUE TRUE TRUE TRUE ...
##  $ Hus_occup    : Factor w/ 4 levels "1","2","3","4": 2 3 3 3 3 3 3 3 3 2 ...
##  $ std_liv      : Factor w/ 4 levels "1","2","3","4": 3 4 4 3 2 3 2 2 4 2 ...
##  $ media        : Factor w/ 2 levels "Good","Not-Good": 1 1 1 1 1 1 1 1 1 2 ...
##  $ contraceptive: Factor w/ 3 levels "no use","long term",..: 1 1 1 1 1 1 1 1 1 1 ...
```

The dataset consists of 10 columns and 1473 variables. By inspecting the structure of the dataset, it is found that all the attributes are treated as integers which are not true. Some of them are factors and some of them are logical variables. From the attribute information given in the data description, we change those variables which are not actually integers. We use 'as.factor' and 'as.logical' functions to do this. 'sapply' function is used to check the levels of each variable in the dataset. It is found that 'Wife's religion', 'Media exposure' and 'Contraceptive method used' are denoted by numbers. The levels for these variables are renamed to respective levels. The variables like 'Wife's education', 'Husband's education', 'Husband's occupation', and 'Standard-of-living index' are ordinal variables and are ranked from 1 to 4, with '1' being low and '4' being high. Structure of the data frame is examined to check the changes to the data frame.

6 Subsetting I

The data frame is a subset using the first 10 observations including all the variables. Then it is converted into a matrix. Structure of that matrix is inspected using 'str' function.

```
contra_sub <- head(contraceptive,10) # Subsetting the data frame
contra_sub_matrix <- as.matrix(contra_sub) # Converting in to a matrix
contra_sub_matrix
```

```
##    Wife_age Wife_edu Hus_edu Num_Children Wife_rel Wife_work Hus_occup
## 1  "24"     "2"      "3"     " 3"         "Islam"  " TRUE"   "2"
## 2  "45"     "1"      "3"     "10"         "Islam"  " TRUE"   "3"
## 3  "43"     "2"      "3"     " 7"         "Islam"  " TRUE"   "3"
## 4  "42"     "3"      "2"     " 9"         "Islam"  " TRUE"   "3"
## 5  "36"     "3"      "3"     " 8"         "Islam"  " TRUE"   "3"
## 6  "19"     "4"      "4"     " 0"         "Islam"  " TRUE"   "3"
## 7  "38"     "2"      "3"     " 6"         "Islam"  " TRUE"   "3"
## 8  "21"     "3"      "3"     " 1"         "Islam"  "FALSE"   "3"
## 9  "27"     "2"      "3"     " 3"         "Islam"  " TRUE"   "3"
## 10 "45"     "1"      "1"     " 8"         "Islam"  " TRUE"   "2"
##    std_liv media      contraceptive
## 1  "3"     "Good"     "no use"
## 2  "4"     "Good"     "no use"
## 3  "4"     "Good"     "no use"
## 4  "3"     "Good"     "no use"
## 5  "2"     "Good"     "no use"
## 6  "3"     "Good"     "no use"
## 7  "2"     "Good"     "no use"
## 8  "2"     "Good"     "no use"
## 9  "4"     "Good"     "no use"
## 10 "2"     "Not-Good" "no use"
```

```
str(contra_sub_matrix) # Structure of the matrix
```

```
##  chr [1:10, 1:10] "24" "45" "43" "42" "36" "19" "38" "21" "27" "45" ...
##  - attr(*, "dimnames")=List of 2
##   ..$ : chr [1:10] "1" "2" "3" "4" ...
##   ..$ : chr [1:10] "Wife_age" "Wife_edu" "Hus_edu" "Num_Children" ...
```

```
attributes(contra_sub_matrix) # Attributes of the matrix
```

```
## $dim
## [1] 10 10
##
## $dimnames
## $dimnames[[1]]
##  [1] "1"  "2"  "3"  "4"  "5"  "6"  "7"  "8"  "9"  "10"
##
## $dimnames[[2]]
##  [1] "Wife_age"      "Wife_edu"      "Hus_edu"       "Num_Children"
##  [5] "Wife_rel"      "Wife_work"     "Hus_occup"     "std_liv"
##  [9] "media"         "contraceptive"
```

From the structure of the matrix, it is clear that the matrix former is a matrix of character values. Since matrix can accept only one type of data and the variables in the data frame had integers factors and logical variables, the matrix could only be formed as a character matrix by converting all elements into character.

7 Subsetting II

The data frame was subset to contain only first and the last variable. The 'ncol' function is used to get the number of columns in the data frame. This is then used to specify the last variable in the data frame. By omitting the row specifications and specifying only the first and the last variable in the dataset a subset is created. This subset of the data frame was then saved as a **.RData** file in the working directory to be transferred or stored for later use.

```
ncol(contraceptive) # total number of columns
```

```
## [1] 10
```

```
contra_sub_2 <- contraceptive[,c(1,ncol(contraceptive))] # Subsetting
head(contra_sub_2) # head of new subset
```

```
##   Wife_age contraceptive
## 1       24        no use
## 2       45        no use
## 3       43        no use
## 4       42        no use
## 5       36        no use
## 6       19        no use
```

```
save(contra_sub_2, file = "contra_sub_2.RData") # saving the output as .RData
```

8 Create a new Data Frame

A data frame named 'df' is created using the function 'data.frame'. The data frame consists of two variables named 'Education' and 'Age', each containing 4 observations each. The variable 'Education' is an ordinal variable with ordered levels from **"elementary", "high school", "Bachelor's"**, and **"Master's"**. The variable 'Age' is a numeric variable. Structure of the variables is shown below using 'str' function. Levels of the variable 'Education' is displayed using the 'levels' function. A numeric vector called salary is created and is combined the to the previous data frame using 'cbind' function. Attributes, dimensions, and structure of the new data frame are inspected using the 'attributes', 'dim' and 'str' functions respectively.

```
df <- data.frame(Education =
                   c("elementary","high school",
                     "Bachelor's","Master's"),
                 Age =
                   c(12,25,54,29),stringsAsFactors=TRUE) # Creating dataframe
df
```

```
##     Education Age
## 1  elementary  12
## 2 high school  25
## 3  Bachelor's  54
## 4    Master's  29
```

```
str(df) # structure of the data
```

```
## 'data.frame':    4 obs. of  2 variables:
##  $ Education: Factor w/ 4 levels "Bachelor's","elementary",..: 2 3 1 4
##  $ Age      : num  12 25 54 29
```

```
levels(df$Education) # Levels of education
```

```
## [1] "Bachelor's"  "elementary"  "high school" "Master's"
```

```
salary <- c(8000,7000,12000,15000) # creating new vector
newDF <- cbind(df,salary) # binding the vector and Dataframe
newDF
```

```
##     Education Age salary
## 1  elementary  12   8000
## 2 high school  25   7000
## 3  Bachelor's  54  12000
## 4    Master's  29  15000
```

```
attributes(newDF) # attributes of new Dataframe
```

```
## $names
## [1] "Education" "Age"       "salary"
##
## $class
## [1] "data.frame"
##
## $row.names
## [1] 1 2 3 4
```

```
dim(newDF) # dimensions of new Dataframe
```

```
## [1] 4 3
```

```
str(newDF) # structure of new Dataframe
```

```
## 'data.frame':    4 obs. of  3 variables:
##  $ Education: Factor w/ 4 levels "Bachelor's","elementary",..: 2 3 1 4
##  $ Age      : num  12 25 54 29
##  $ salary   : num  8000 7000 12000 15000
```