Lesson-4.—09.05.2023-data-manipulation.R

bramu

2023-05-08

```
# Reference manual for the package
# https://cloud.r-project.org/web/packages/wooldridge/wooldridge.pdf
library(wooldridge)

rm(list = ls())

# Sctructure of the dataset ----

# We first take a look at the dataset
#View(wage1)

# How many rows in the dataset?
nrow(wage1)
```

```
## [1] 526
```

```
# How many columns?
ncol(wage1)
```

```
## [1] 24
```

```
# How many columns and rows?
dim(wage1)
```

```
## [1] 526  24
```

```
# Display the internal structure
str(wage1)
```

```
## ’data.frame':    526 obs. of  24 variables:
##  $ wage    : num  3.1 3.24 3 6 5.3 ...
##  $ educ    : int  11 12 11 8 12 16 18 12 12 17 ...
##  $ exper   : int  2 22 2 44 7 9 15 5 26 22 ...
##  $ tenure  : int  0 2 0 28 2 8 7 3 4 21 ...
##  $ nonwhite: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ female  : int  1 1 0 0 0 0 0 1 1 0 ...
##  $ married : int  0 1 0 1 1 1 0 0 0 1 ...
##  $ numdep  : int  2 3 2 0 1 0 0 0 2 0 ...
##  $ smsa    : int  1 1 0 1 0 1 1 1 1 1 ...
##  $ northcen: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ south   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ west    : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ construc: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ ndurman : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ trcommpu: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ trade   : int  0 0 1 0 0 0 1 0 1 0 ...
##  $ services: int  0 1 0 0 0 0 0 0 0 0 ...
##  $ profserv: int  0 0 0 0 0 1 0 0 0 0 ...
##  $ profocc : int  0 0 0 0 0 1 1 1 1 1 ...
##  $ clerocc : int  0 0 0 1 0 0 0 0 0 0 ...
##  $ servocc : int  0 1 0 0 0 0 0 0 0 0 ...
##  $ lwage   : num  1.13 1.18 1.1 1.79 1.67 ...
##  $ expersq : int  4 484 4 1936 49 81 225 25 676 484 ...
##  $ tenursq : int  0 4 0 784 4 64 49 9 16 441 ...
##  - attr(*, "time.stamp")= chr "25 Jun 2011 23:03"
```

```
# Names of the columns
colnames(wage1)
```

```
##  [1] "wage"     "educ"     "exper"    "tenure"   "nonwhite" "female"
##  [7] "married"  "numdep"   "smsa"     "northcen" "south"    "west"
## [13] "construc" "ndurman"  "trcommpu" "trade"    "services" "profserv"
## [19] "profocc"  "clerocc"  "servocc"  "lwage"    "expersq"  "tenursq"
```

```
# Indexing ----

# Useful when you need to address a particular element of a vector,
# for example the years of education of the fifth worker in the dataset
wage1$educ[5]
```

```
## [1] 12
```

```
# If we want data for more than one worker
wage1$educ[c(2,3,5)]
```

```
## [1] 12 11 12
```

```
# Data from worker one to worker five
wage1$educ[1:5]
```

```
## [1] 11 12 11  8 12
```

```
# If we want to modify one particular observation
wage1$educ[5] <- NA

# Negative indexing allows to show data except those specified in parenthesis
wage1$educ[-c(5:526)]
```

```
## [1] 11 12 11  8
```

```r
# Negative indexing allows to drop specific rows and columns.
# Here we drop the first row of the dataset
#wage1[-1,]

# Here we we drop the first column of the dataset
#wage1[,-1]

# We can also keep specific column selecting them by name
#wage1[, colnames(wage1) %in% c("wage", "educ", "exper")]

# We convert dollars to euros and add the new column to the dataset
wage1$wage_EUR <- wage1$wage * 0.86

# We can compare the variable wage, originally expressed in dollars,
# and the new variable `wage_EUR`. Here we take a look at the first five rows
wage1[1:5, c("wage", "wage_EUR")]
```

```
##   wage wage_EUR
## 1 3.10   2.6660
## 2 3.24   2.7864
## 3 3.00   2.5800
## 4 6.00   5.1600
## 5 5.30   4.5580
```

```r
# Conditional selection ----

# To be used when you need to extract data that satisfy certain criteria
# Workers that have more than 15 years of education
wage1$educ[wage1$educ > 15]
```

```
##   [1] NA 16 18 17 16 16 16 16 16 16 16 16 16 18 16 16 16 16 17 18 16 17 18 16 18
##  [26] 16 18 16 16 16 16 16 18 18 18 16 16 16 16 16 17 16 16 16 18 16 16 18 16 18
##  [51] 16 16 18 17 16 16 16 18 16 16 16 16 16 16 18 16 16 18 17 16 17 16 16 16 16
##  [76] 18 16 17 16 16 16 17 18 18 16 17 17 16 16 16 16 16 16 16 16 16 16 17 16 16
```

```r
# How many workers have more than 15 years of education?
length(wage1$educ[wage1$educ > 15])
```

```
## [1] 100
```

```r
# How many workers have education between 15 and 18 years of education?
length(wage1$educ[wage1$educ >= 15 & wage1$educ <= 18])
```

```
## [1] 121
```

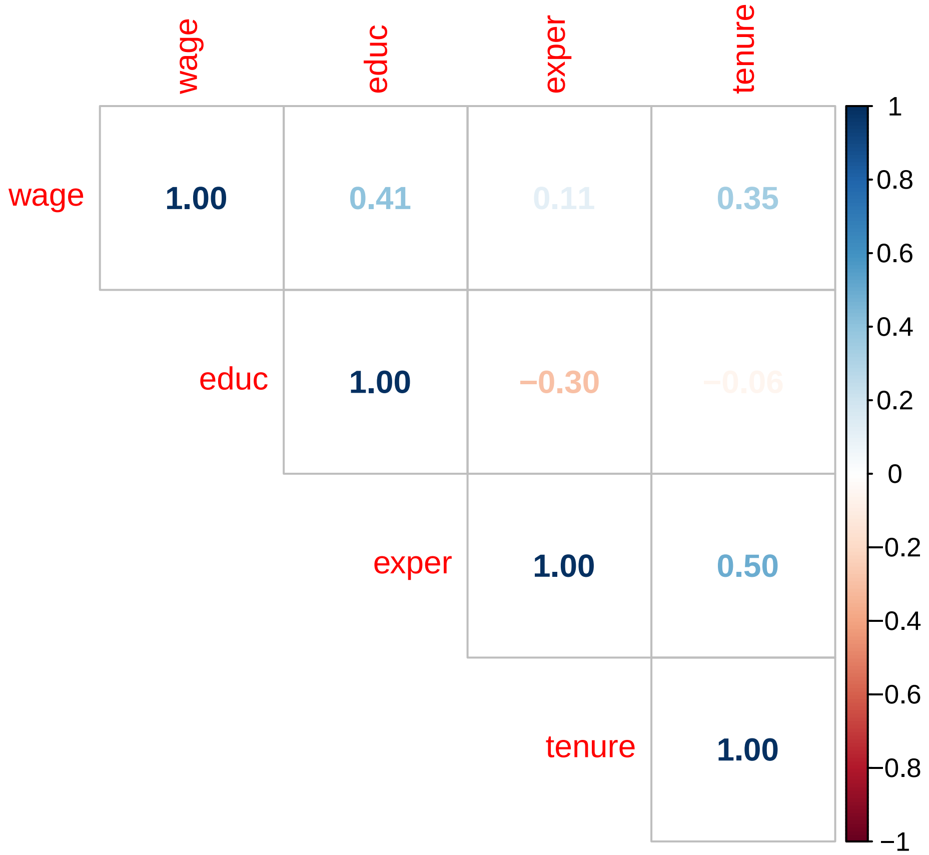
```r
# What is the mean wage of workers that have between 15 and 18
# years of education?
mean(wage1$educ[wage1$educ >= 15 & wage1$educ <= 18])
```

```
## [1] NA
```

```
# What is the percentage of workers who have more than 15 years of experience?

# Correlation matrix ----
rm(list = ls())
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
data_W <- wage1[, colnames(wage1) %in% c("wage", "educ", "exper", "tenure")]

data_W_corr <- cor(data_W)

corrplot(data_W_corr,
         method = "number",
         type = "upper")
```


wage
educ
exper
tenure
wage
educ
exper
tenure