Hands-On Data Analysis for ININ Using R

Prof. Dr. Cornelia Storz, M.Sc. Fei (Michael) Wang

Management and Microeconomics, Goethe-Universität Frankfurt

This document was prepared for students who are taking ININ course and planning to take the exam. It is a collection of notes and codes for the course. The notes are based on the tutorials we had in the course. I am trying to make it concise and easy to understand. I hope it can help you to review the course and prepare for the exam. *We are living in a very noisy world, therefore let's keep it simple and clear.* I setup a challenge for myself to deliver a clear and concise review notes within 15 pages. This brings the trade-off, which means some figures and tables are not included in the notes. Therefore, you have to run the codes to see the results.

I hope you enjoy reading it. I also hope you will have this notes with you whenever you want to do some data analysis. If one day, you still refer to this notes and find it still useful, I would be very happy to hear that.

Keywords: econometrics, data analysis, regression models, empirical research, innovation, management

Contents

1 Introduction

All statistical or econometric or machine learning models are based on the following assumptions:

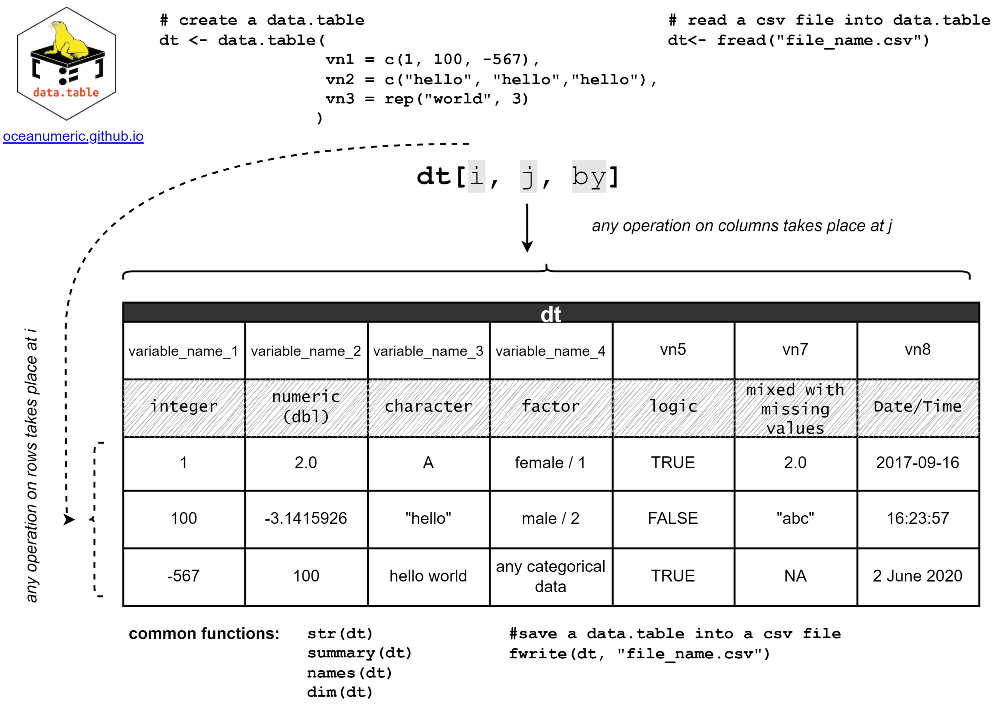
- there are something we know - **data**
- and something we don't know - **error** ε.

In summary, according to confucius, *to know what we know and what we do not know*, that is called **wisdom**. Or like Plato said, *I know that I know nothing*. To help you to review the course, the notes will be organized as follows:

1. **Data**: using `data.table` to get familiar with the data
2. **Simple linear regression**: how to estimate a simple linear regression model, how to interpret the results
3. **Multiple linear regression**: how to estimate a multiple linear regression model, how to interpret the results, how to test the model
4. **Introduction to logistic regression**: why do we need logistic regression
5. **Data manipulation**: will not be tested in the exam, but it is very useful for your future work or research

2 Introduction to Data and `data.table`

Broadly speaking, there are two kinds of data: **structured data** and **unstructured data**. Structured data is data that has a structure, such as a table, whereas unstructured data is data that does not have a structure, such as a text file. In this course, we focus on structured data. This means all the data we will use look like tables, such as the following one:


data.table
oceanumeric.github.io
dt[i, j, by]
any operation on columns takes place at j
any operation on rows takes place at i
common functions:

```
# create a data.table
dt <- data.table(
                 vn1 = c(1, 100, -567),
                 vn2 = c("hello", "hello","hello"),
                 vn3 = rep("world", 3)
                )
```

```
# read a csv file into data.table
dt<- fread("file_name.csv")
```

dt

variable_name_1	variable_name_2	variable_name_3	variable_name_4	vn5	vn7	vn8
integer	numeric (dbl)	character	factor	logic	mixed with missing values	Date/Time
1	2.0	A	female / 1	TRUE	2.0	2017-09-16
100	-3.1415926	"hello"	male / 2	FALSE	"abc"	16:23:57
-567	100	hello world	any categorical data	TRUE	NA	2 June 2020

```
str(dt)
summary(dt)
names(dt)
dim(dt)
```

```
#save a data.table into a csv file
fwrite(dt, "file_name.csv")
```