

.title-wrap[Intro to Programming with R for Political Scientists]

Trivia

Workflow

CalculatoR

CalculatoR

CalculatoR

A Primer on OOP (“Object Oriented Programming”)

Object Oriented Programming

Functions

Functions

Making Objects: Assignment (Operators)

Making Objects: Assignment (Operators)

Naming Conventions

Workspace/Environment

Vectors

Lists

Lists

Factors

Data Frames

Data Frames

Type Conversion

Loading/Installing Packages

Loading/Installing Packages

The Tidyverse

Next up: Data Wrangling

Lys

goes Xaringan

Markus Freitag

2021-06-08



class: inverse, center, title-slide, middle

.title-wrap[Intro to Programming with R for Political Scientists]

.handw[]

Markus Freitag

Geschwister Scholl Institute of Political Science, LMU

🐦 (<https://twitter.com/MarkusGFreitag>) 🌐
(<https://markusfreitag.netlify.app/>)

2021-06-08

(<https://github.com/m-freitag>)

Trivia

- R was designed in 1993 by Ross Ihaka and Robert Gentleman
- Builds upon the S programming language by John Chambers
 - Named R as a play on S and bc of the first names of the authors
- There are 17661 packages available on CRAN (<https://cran.r-project.org/>) as of 2021-06-08.
- R-Studio (<https://www.rstudio.com/about/>) \neq R-Core Team (<https://www.r-project.org/contributors.html>); the former is a mix of a for-profit and a non-profit company; highly committed to produce free & open-source products; has some business solutions



.center[.font50[Image source and more R-History trivia

(<https://rss.onlinelibrary.wiley.com/doi/epdf/10.1111/j.1740-9713.2018.01169.x>)]]

Workflow

- You forked and cloned the course repo.

- Navigate to `Session Scripts > Session 2` and open `Session_2_script.R`.
- You will see a pre-formatted Script containing some useful information on comments and formatting.
- The script is otherwise empty. Fill it with the stuff I will discuss on the slides.
- Make comments for yourself. Learn as you write. Explore. Stage, commit and push.



- If you have a second monitor, great! If not, split your screen.

class: inverse, center, middle name: intro

CalcuatoR

CalcuatoR

```
7 + 5 # [n] stands for the nth element printed to the console.
```

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

```
4 * 5 + 2 / 3^3 # Multiplication and division first, then addition and subtraction
```

```
## [1] 20.07407
```

```
# Modulo Operators:
```

```
10 %/% 3 # Integer division
```

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

```
10 %% 3 # Remainder ("Rest")
```

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

CalculatoR

```
# Relational and logical operators
```

```
3 < 4
```

```
## [1] TRUE
```

```
2 == 1 & 4 > 2 # == "equal to"; & "element wise logical AND"
```

```
## [1] FALSE
```

```
2 == 1 | 4 > 2 # | "element wise logical or"
```

```
## [1] TRUE
```

```
3 != 4 # != "not equal"
```

```
## [1] TRUE
```

```
7 %in% 300:500 # %in% can be used to evaluate matches in objects
```

```
## [1] FALSE
```

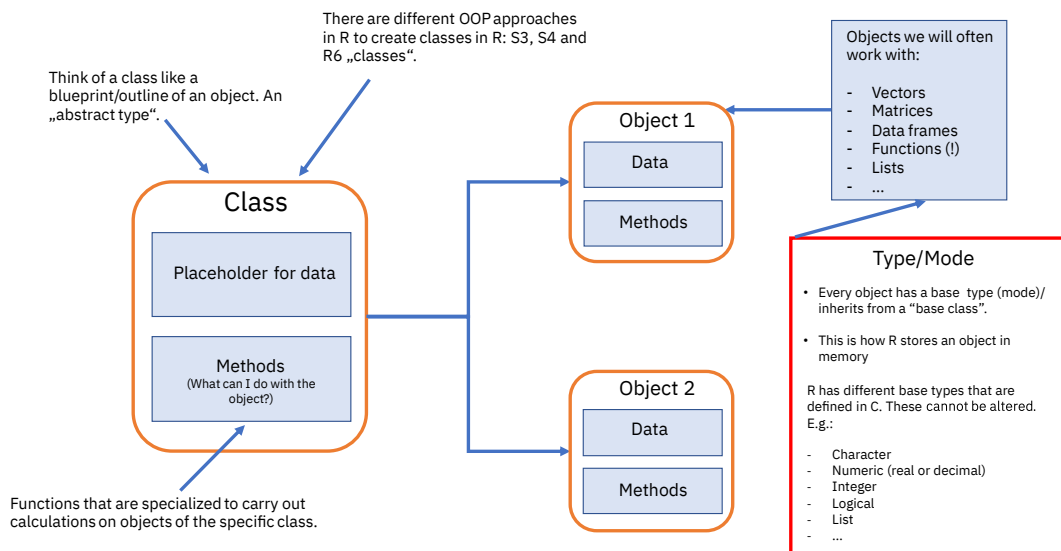
```
# Take care about the order of precedence...
```

class: inverse, center, middle name: intro

A Primer on OOP (“Object Oriented Programming”)

Object Oriented Programming

Everything is an object and everything has a name.



Functions

- Functions (<https://cran.r-project.org/doc/manuals/r-release/R-lang.html#Function-objects>) are objects; we will discuss them in more detail in session x
- For now, just think of them in the usual mathematical sense, where we pass some argument and get back a value. E.g. $f(x) = \frac{2x+3}{\sqrt{3}}$.
- In R, one or multiple arguments get passed to the function body (where the function is defined) and you get back some results.

- For instance, to define the above function and call it, we specify the following:

```
f <- function(x) {  
  (2 * x + 3) / sqrt(3)  
}
```

```
f(3)
```

```
## [1] 5.196152
```

Functions

- There are many functions in R, some are written by users and scientists and put into some package, some are built-in functions of “base” R.
- A really helpful built-in function - in the literal sense - is the... `help()` function.
- Gives you more information about the usage and arguments of a built-in/user-written function.

We can also call `help()` to get help about help:

```
# help(help)
# Or for short:
?help
```

- But let's not get ahead of ourselves... what did this arrow (`<-`) thing do?

[Fine Point:] The arithmetic, logical and relational operators we met are actually also function calls.]

Making Objects: Assignment (Operators)

- You can use `<-` or `=` for assignment

- For instance,

```
a <- 3 # Or a = 3; under the hood, assignment works more like a -> 3
```

assigns the name `x` to an object of type/mode numeric holding the value 3. I.e. binds an object to a name.

Simplification: `>` creates an object named `a`, containing the value 3.]

-

-Using

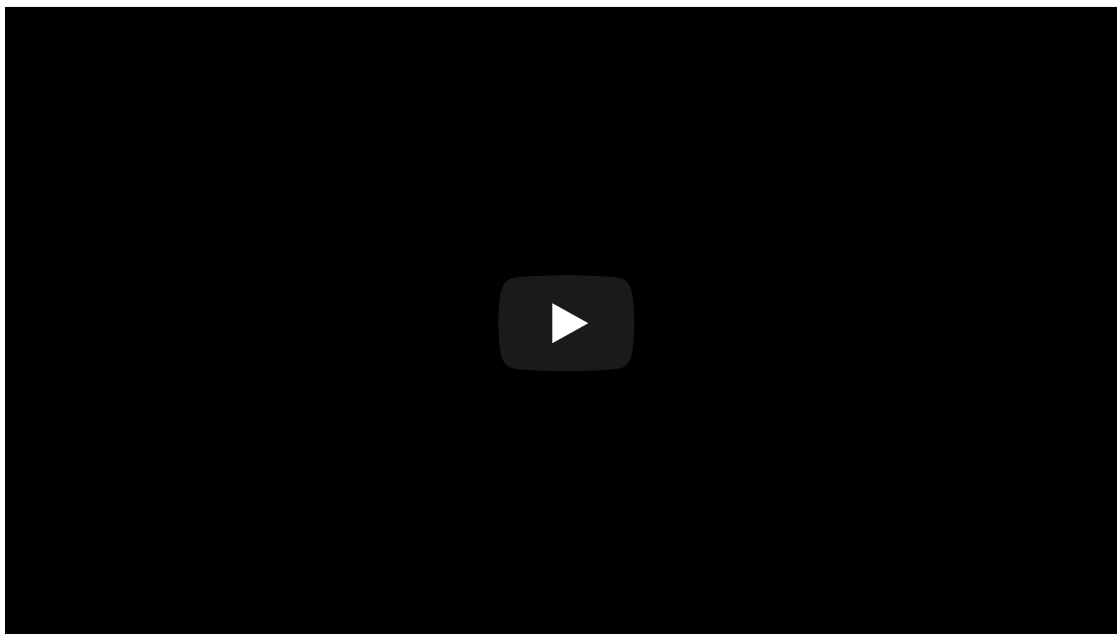
```
class(a)
mode(a)
```

gives you information about the class/type of the object. `class()` gives the class of the object from an OOP POV, `typeof()` the base type.

- In this case, this is not very interesting as we created an **atomic** numeric vector.
- What is the class/type of `b <- 3 > 2` and `c <- "string"`?

Making Objects: Assignment (Operators)

- Using `=` is legal as per the man, the myth, the legend Ross Ihaka himself:



- But there are also some sensible arguments to stick mostly to `->` in casual settings (e.g. readability: easier to discriminate from arguments)
- Bottom line: **be consistent**.

Naming Conventions

- For readability, we want names of/bound to objects to be meaningful.
- Pretty easy to do when working with “real” data; laziness pretty much the only obstacle ;)
- There are several naming conventions. I like `snake_case` and `camelCase` the most. We will use snake cases in this course.
- Be consistent.
- There are some “forbidden”/“reserved” words (<https://stat.ethz.ch/R-manual/R-devel/library/base/html/Reserved.html>) that cannot be assigned as names. E.g. `NA` (logical constant indicating missing values), `if`, `else`, `break`.

Workspace/Environment

- In contrast to Stata, R can hold multiple objects at a time.

- This is very convenient; you can copy, modify etc. R also copies **ALOT** internally (one reason why its sort of a slow language).
- The global environment is the interactive workspace you usually work with.
- In R-Studio you can inspect some objects by clicking on them (equivalent to calling `View()`)

—

- We won't go into the details of environments but see here (<https://adv-r.hadley.nz/environments.html>) for an advanced treatment.
- This will get more intuitive soon (e.g. when we take on different data structures on the next few slides)

Vectors

- Vectors are the most fundamental data structure in R.
- As vectors in R have to be of the same type (e.g. numeric), they are often called **atomic** vectors
- *Technical Fine Point:* (<https://adv-r.hadley.nz/vectors-chap.html>) Vectors have attributes, importantly dimension and class. With the dimension attribute, vectors can become arrays and matrices. With the class attribute, we can build an S3 object on top of the base type (see e.g. factor type).

—

We can build longer vectors by concatenating shorter ones using the `c()` function:

```
chr_var <- c("a", "b", "c", "d") # A 4-element atomic vector of type character
```

Indexing:

```
chr_var[3] # Returns the third element of object "chr_var".  
chr_var[1:3] # Returns the 1st three.
```

Lists

- Objects of type list are highly versatile. They are vectors but more “generic”.
- I.e. elements of a list can have different types.^{font70[[2]]}

For example:


```
list_a <- list(
  5:7,
  "string",
  c(TRUE, TRUE),
  c(1.23, 4.20)
)
```

constructs a list using the function `list()` .

.font70[^[2].hl[Fine Point:] A list does not store objects of different types, it references to them.]

Lists

- Indexing:

.pull-left[

```
list_a[3] # This returns a list with element four.
```

```
## [[1]]
## [1] TRUE TRUE
```

```
list_a[[4]] # We need double brackets to index a list element.
```

```
## [1] 1.23 4.20
```

```
# Index within a list element:
list_a[[4]][2]
```

```
## [1] 4.2
```

]

.pull-right[

```
# For named lists, we can index with $:
names(list_a) <- c("a", "b", "c", "d") # We can also name elements directly:
  list(x = 1, y = 2)
list_a$b
```

```
## [1] "string"
```

```
list_a$d[1]
```

```
## [1] 1.23
```

```
]
```

Factors

- To represent categorical variables, we often use factor objects in R (e.g. makes it easier to get counts of all categories).
- They are build on top of integer vectors and come with a levels attribute.

```
.code70[
```

```
backgrounds_char <- c("none", "Stata", "Stata", "Stata", "R") # Student's pro  
  g. backgrounds  
backgrounds_fac <- factor(backgrounds_char, levels = c("none", "Stata", "R"))  
  # Or:  
# backgrounds_fac <- factor(c("none", "Stata", "Stata", "Stata", "R"))
```

```
]
```

```
.code70[.pull-left[
```

```
class(backgrounds_fac)
```

```
## [1] "factor"
```

```
]]
```

```
.code70[.pull-right[
```

```
typeof(backgrounds_fac)
```

```
## [1] "integer"
```

```
]] .code70[.pull-left[
```

```
table(backgrounds_fac)
```

```
]] .code70[.pull-right[
```

```
## backgrounds_fac
##   none Stata      R
##     1     3     1
```

```
]]
```

Data Frames

- Data frames are special lists: lists of evenly sized vectors.
- You likely already have a grasp for their structure from Stata or other software; they are crucial for data analysis → the rectangle we love.

.code60[.pull-left[

```
set.seed(666)
df_langs <- data.frame(
  background = factor(c(
    "Python",
    "Stata", "Stata", "Stata",
    "R"
  )),
  skill = rnorm(5)
)
# View(df_langs)
head(df_langs) # print first few lines of an object (6 by default)
# Indexing
df_langs[2, 2] # Second row of second column
df_langs$background[1]
```

```
]]
```

.code60[.pull-right[

```
##   background      skill
## 1    Python  0.7533110
## 2     Stata  2.0143547
## 3     Stata -0.3551345
## 4     Stata  2.0281678
## 5        R -2.2168745
```

```
## [1] 2.014355
```

```
## [1] Python
## Levels: Python R Stata
```

]]

Data Frames

- When passing columns of a data frame we need need to take care that the named references to them “live” inside the data frame and not in the global environment.

.code60[

```
# Throws an error:
lm(skill ~ background) # [3]; the lm object is of type list
```

]

.code60[

```
lm(df_langs$skill ~ df_langs$background) # Or: lm(skill ~ background, df_langs)
```

```
##
## Call:
## lm(formula = df_langs$skill ~ df_langs$background)
##
## Coefficients:
##              (Intercept)          df_langs$backgroundR  df_langs$backgroundSta
ta
##              0.7533              -2.9702              0.47
58
```

] .font70[^[3].hl[NOTE:] The formula object/class is built on the base type “language” and important for many stats applications.]

Type Conversion

In most cases, we can also convert an object to another type. Sometimes this has side-effects.

.pull-left[.code70[

```
df_langs_m <- as.matrix(df_langs) # convert data.frame to matrix
print(df_langs_m)
```

```
##      background skill
## [1,] "Python"      " 0.7533110"
## [2,] "Stata"       " 2.0143547"
## [3,] "Stata"       "-0.3551345"
## [4,] "Stata"       " 2.0281678"
## [5,] "R"           "-2.2168745"
```

```
]]
```

```
.pull-right[.code70[
```

```
typeof(df_langs_m[2, 1])
```

```
## [1] "character"
```

```
]]
```

- As matrices are atomic, an implicit coercion happened.^{.font70[[4]]}
- If stuff cannot get converted, you get NAs.

.font70^[4].hl[Fine Point:] Coercion order is character > double > integer > logical.]

Loading/Installing Packages

Classic:

```
install.packages("tidyverse")
library(tidyverse)
```

If you are lazy (...for CRAN packages; recommended only for smaller projects/problem sets):

```
install.packages("pacman")

pacman::p_load(
  tidyverse,
  patchwork,
  rio,
  data.table
)
```

Loading/Installing Packages

- As mentioned, R comes with 17661 packages on CRAN up to date.

- This is both a blessing and a curse (https://en.wikipedia.org/wiki/Dependency_hell).
- Many useful packages also live only on Github
- Escaping package/library dependency hell is a bit of a pain in R (much easier in Python and Julia).^[5]
- To take it to another level, Dockers (<https://environments.rstudio.com/docker>) expand reproducibility by providing a full runtime environment (and they work well with R).

.font70^[5].hl[NOTE:] The packrat (<https://github.com/rstudio/packrat>) package is one option. renv (<https://rstudio.github.io/renv/articles/renv.html>) is much better. You should definitely use it if you work on a larger project]

The Tidyverse

“The tidyverse is an opinionated collection of R packages designed for data science. All packages share an underlying design philosophy, grammar, and data structures.” tidyverse website (<https://www.tidyverse.org/>)

- Developed by Hadley Wickham (https://en.wikipedia.org/wiki/Hadley_Wickham), supported by RStudio
 - Packages that are easy to learn, that have a similar syntax, and are convenient
- There are alternatives... base R is still stable and useful; data.table is fast
- In my opinion, the tidyverse is a bit verbose/wordy. A lot of functions doing one particular thing.
- You will always end up googling a lot. But that's something you will do a lot anyways (I do too!)

class: inverse, center, middle name: intro

Next up: Data Wrangling