

# Introduction to R

Ista Zahn



**The Institute**  
*for* Quantitative Social Science  
at Harvard University

# Outline

- 1 Workshop Materials and Introduction
- 2 Graphical User Interfaces
- 3 Data and Functions
- 4 Help and package management
- 5 Getting data into R
- 6 Data Manipulation
- 7 Basic Statistics and Graphs
- 8 Wrap-up

# Topic

- 1 Workshop Materials and Introduction
- 2 Graphical User Interfaces
- 3 Data and Functions
- 4 Help and package management
- 5 Getting data into R
- 6 Data Manipulation
- 7 Basic Statistics and Graphs
- 8 Wrap-up

# Materials and setup

Everyone should have R installed –if not:

- Open a web browser and go to <http://cran.r-project.org> and download and install it
- Also helpful to install RStudio (download from <http://rstudio.com>)

Materials for this workshop include slides, example data sets, and example code.

- Download materials from <http://j.mp/intro-r>
- Extract the zip file containing the materials to your desktop

Workshop notes are available in .html and .pdf format. Navigate to your desktop and open either Rintro.pdf or Rintro.html.

# What is R?

R is a programming language designed for statistical computing. Notable characteristics include:

- Vast capabilities, wide range of statistical and graphical techniques
- Very popular in academia, growing popularity in business:  
<http://r4stats.com/articles/popularity/>
- Written primarily by statisticians
- FREE (no cost, open source)
- Excellent community support: mailing list, blogs, tutorials
- Easy to extend by writing new functions

# Coming to R

Coming from...

**Stata** <http://www.princeton.edu/~otorres/RStata.pdf>

**SAS/SPSS** <http://www.et.bs.ehu.es/~etptupaf/pub/R/RforSAS&SPSSusers.pdf>

**matlab** <http://www.math.umaine.edu/~hiebler/comp/matlabR.pdf>

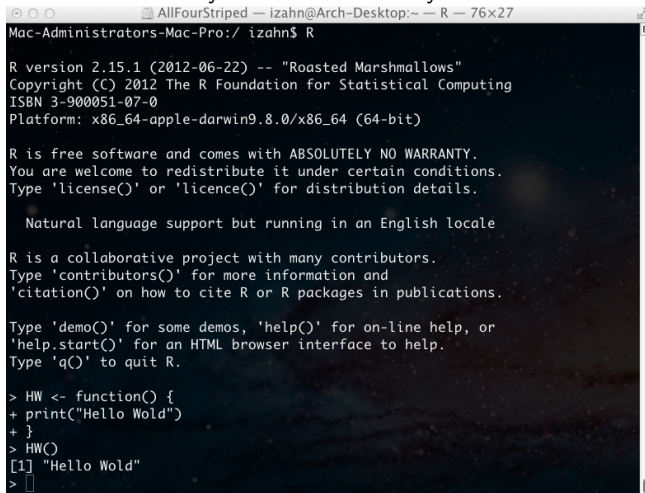
**Python** <http://mathesaurus.sourceforge.net/matlab-python-xref.pdf>

# Topic

- 1 Workshop Materials and Introduction
- 2 Graphical User Interfaces
- 3 Data and Functions
- 4 Help and package management
- 5 Getting data into R
- 6 Data Manipulation
- 7 Basic Statistics and Graphs
- 8 Wrap-up

# R GUI alternatives (no GUI)

The old-school way is to run R directly in a terminal



```
Mac-Administrators-Mac-Pro:/ izahn$ R

R version 2.15.1 (2012-06-22) -- "Roasted Marshmallows"
Copyright (C) 2012 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
Platform: x86_64-apple-darwin9.8.0/x86_64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

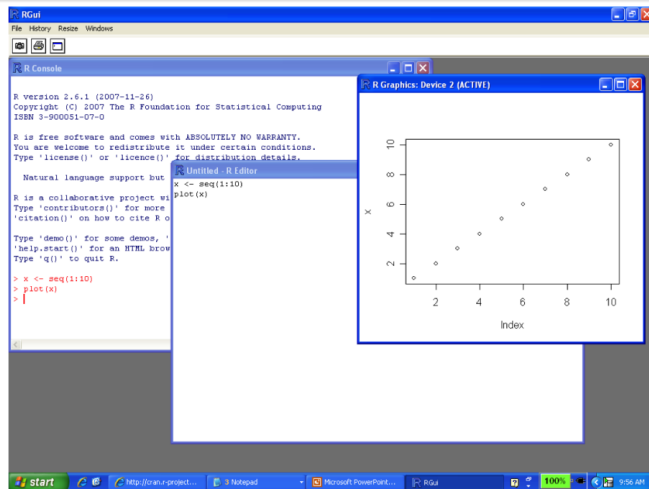
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> HW <- function() {
+ print("Hello Wold")
+ }
> HW()
[1] "Hello Wold"
>
```

But hardly anybody does it that way anymore!



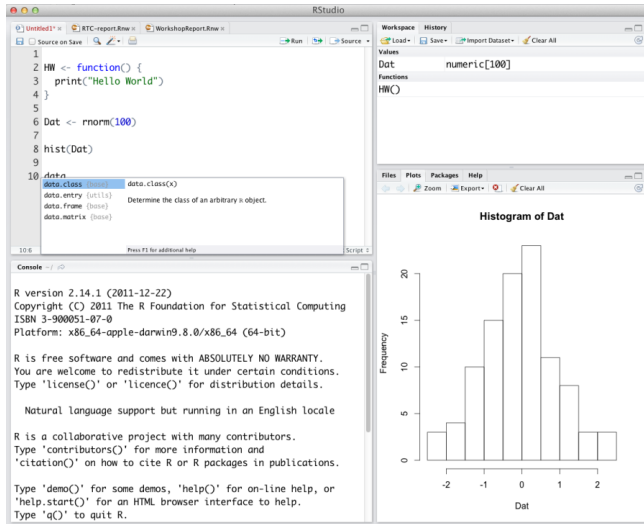
# R GUI alternatives (Windows default)



The default windows GUI is not very good

- No parentheses matching or syntax highlighting
- No work-space browser

# R GUI Alternatives (Rstudio on Mac)



Rstudio has many useful features, including parentheses matching and auto-completion

# R GUI Alternatives (Emacs with ESS)

The screenshot shows the Emacs editor interface with the ESS package running R. The main window displays the following R code:

```
HW <- function() {
  print("Hello World")
}

Dat <- rnorm(100)

png(file = "myHist.png")
hist(Dat)
dev.off()

data.frame
```

The sidebar on the right shows the results of the code execution:

```
mode length
Dat numeric 100
HW function 1
```

Below the code, a table lists the contents of the `data.frame` object:

data.frame	data.frame	package:base
data.class	R Documentation	
data.entry		
data.matrix	Data Frames	

Description:

This function creates data frames.

The bottom right of the sidebar shows a histogram titled "Histogram of Dat" with the following data:

Dat	Frequency
-3	4
-2	9
-1	32
0	35
1	13
2	4
3	1

Emacs + ESS is a very powerful combination, but can be difficult to set up

# Components of R GUIs

- The R console
  - Displays command history and results
  - Commands can be typed directly in the console
  - R Console work disappears once session is closed
- A text editor
  - A plain text editor for writing R code
  - Good ones will have syntax highlighting, parentheses matching etc.
  - Anything that modifies your data should be done in a text editor
- Graphics windows
  - View, re-size, and save graphics
  - A good GUI will allow you to cycle through graph history
- Work-space viewer
  - Some GUIs have work-space browsers that allow you to see stored objects
  - Very helpful if you are absentminded like me and frequently forget what names you gave your data!

# Things to keep in mind

- Case sensitive, like Stata (unlike SAS)
- Comments can be put almost anywhere, starting with a hash mark ('#'); everything to the end of the line is a comment
- The command prompt ">" indicates that R is ready to receive commands
- If a command is not complete at the end of a line, R will give a different prompt, '+' by default
- Parentheses must always match (first thing to check if you get an error)
- R Does not care about spaces between commands or arguments
- Names should start with a letter and should not contain spaces
- Can use "." in object names (e.g., "my.data")
- Use forward slash ("/") instead of backslash in path names, even on Windows

# Exercise 0

- 1 Try to get R to add 2 plus 2
- 2 See if you can find the help page for the "mean" topic
- 3 Using any means available, try to figure out how to run a linear regression model in R
- 4 Go to <http://cran.r-project.org/web/views/> and skim the topic closest to your field/interests

# Topic

- 1 Workshop Materials and Introduction
- 2 Graphical User Interfaces
- 3 Data and Functions**
- 4 Help and package management
- 5 Getting data into R
- 6 Data Manipulation
- 7 Basic Statistics and Graphs
- 8 Wrap-up

# Assignment

Values can be assigned names and used in subsequent operations

- The `<-` operator (less than followed by a dash) is used to save values
- The name on the left gets the value on the right.

```
x <- 11 # Assign the value 10 to a variable named x
x + 1 # Add 1 to x
y <- x + 1 # Assign y the value x + 1
y
```

Saved variables can be listed, overwritten and deleted

```
ls() # List variables in workspace
x # Print the value of x
x <- 100 # Overwrite x. Note that no warning is given!
x
rm(x) # Delete x
ls()
```



# Functions

Using R is mostly about applying **functions** to **variables**. Functions

- take **variable(s)** as input **argument(s)**
- perform operations
- **return** values which can be **assigned**
- optionally perform side-effects such as writing a file to disk or opening a graphics window

The general form for calling R functions is

FunctionName(arg.1, arg.2, ... arg.n)

Arguments can be matched by position or name

Examples:

```
#?sqrt
a <- sqrt(y) # Call the sqrt function with argument x=y
round(a, digits = 2) # Call round() with arguments x=x and digits=2
# Functions can be nested so an alternative is
round(sqrt(y), digits = 5) # Take sqrt of a and round
```

# Topic

- 1 Workshop Materials and Introduction
- 2 Graphical User Interfaces
- 3 Data and Functions
- 4 Help and package management**
- 5 Getting data into R
- 6 Data Manipulation
- 7 Basic Statistics and Graphs
- 8 Wrap-up

# Asking R for help

- Start html help, search/browse using web browser

- at the R console:

```
help.start()
```

- or use the help menu from you GUI

- Look up the documentation for a function

```
help(topicName)
```

```
?topicName
```

- Look up documentation for a package

```
help(package="packageName")
```

- Search documentation from R (not always the best way... google often works better)

```
help.search("topicName")
```

# R packages and libraries

There are thousands of R packages that extend R's capabilities.

- To view available packages:

```
library()
```

- To see what packages are loaded:

```
search()
```

- To load a package:

```
library("packageName")
```

- Install new package:

```
install.packages("packageName")
```

# Topic

- 1 Workshop Materials and Introduction
- 2 Graphical User Interfaces
- 3 Data and Functions
- 4 Help and package management
- 5 Getting data into R**
- 6 Data Manipulation
- 7 Basic Statistics and Graphs
- 8 Wrap-up

# The gss dataset

The next few examples use a subset of the General Social Survey data set.  
The variables in this subset include

```
head(read.csv("dataSets/gssInfo.csv"))  
#see gssInfo.csv for rest of the variable descriptions
```

# The "working directory" and listing files

R knows the directory it was started in, and refers to this as the "working directory". Since our workshop examples are in the Rintro folder on the desktop, we should all take a moment to set that as our working directory:

```
setwd("~/Desktop/Rintro")
```

We can also set the working directory using paths relative to the current working directory:

```
getwd() # get the current working directory
setwd("dataSets") # set wd to the dataSets folder
getwd()
setwd("..") # set wd to enclosing folder ("up")
```

It can be convenient to list files in a directory without leaving R

```
list.files("dataSets") # list files in the dataSets folder
# list.files("dataSets", pattern = ".csv") # restrict to .csv files
```

# Importing data from files

In order to read data from a file, you have to know what kind of file it is. The table below lists the functions needed to import data from common file formats.

data type	function	package
comma separated (.csv)	read.csv()	utils (default)
other delimited formats	read.table()	utils (default)
Stata (.dta)	read.dta()	foreign
SPSS (.sav)	read.spss()	foreign
SAS (.sas7bdat)	read.sas7bdat()	sas7bdat
Excel (.xls, .xlsx)	readWorksheetFromFile()	XLConnect

## Examples:

```
# read gss data from the gss.rds R file
datGSS <- readRDS("dataSets/gss.rds")
# read gss data from the gss.csv comma separated file
gss.data <- read.csv("dataSets/gss.csv") # read gss data
# read a Stata dataset from gss.dta
library(foreign) # load foreign data functions
datGSS <- read.dta(file="dataSets/gss.dta")
```



# Checking imported data

Always a good idea to examine the imported data set—usually we want the results to be a `data.frame`

```
class(datGSS) # check to see that test is what we expect it to be
dim(datGSS) # how many rows and columns?
names(datGSS)[1:10] # first 10 column names
str(datGSS[1:5]) # more details about the first 5 columns
```

# Saving and loading R workspaces

In addition to importing individual datasets, R can save and load entire workspaces

- Save our entire workspace

```
ls() # list objects in our workspace
save.image(file="myWorkspace.RData") # save workspace
rm(list=ls()) # remove all objects from our workspace
ls() # list stored objects to make sure they are deleted
```

- Load the "myWorkspace.RData" file and check that it is restored

```
load("myWorkspace.RData") # load myWorkspace.RData
ls() # list objects
```

When you close R you will be asked if you want to save your workspace – if you choose yes then your workspace will be restored next time you start R

# Exercise 1

- ➊ Load the foreign package if you haven't already done so (`library(foreign)`)
- ➋ Look at the help page for the `read.spss` function
- ➌ Read the SPSS data set in `dataSets/gss.sav` and assign the result to an R data object named `GSS.sav`
- ➍ Make sure that the data loaded in step 2 is a `data.frame` (hint: check the arguments documented in the help page)
- ➎ Display the dimensions of the `GSS.sav`.
- ➏ BONUS: figure out how to read the Excel file "`gss.xlsx`" into R

# Topic

- 1 Workshop Materials and Introduction
- 2 Graphical User Interfaces
- 3 Data and Functions
- 4 Help and package management
- 5 Getting data into R
- 6 Data Manipulation**
- 7 Basic Statistics and Graphs
- 8 Wrap-up

# data.frame objects

- Usually data read into R will be stored as a **data.frame**
- A data.frame is a list of vectors of equal length
  - Each vector in the list forms a column
  - Each column can be a different type of vector
  - Often the columns are variables and the rows are observations
- A data.frame has two dimensions corresponding the number of rows and the number of columns (in that order)

# data.frame meta-data

A number of functions are available for inspecting data.frame objects:

```
# row and column names
head(names(datGSS)) # variable names in datGSS
head(rownames(datGSS)) # first few rownames of datGSS
# dimensions
dim(datGSS)
# structure
#str(datGSS) # get structure
```

# Logical operators

It is often useful to select just those rows of your data where some condition holds—for example select only rows where sex is 1 (male). The following operators allow you to do this:

`==` equal to

`!=` not equal to

`>` greater than

`<` less than

`>=` greater than or equal to

`<=` less than or equal to

`&` and

`|` or

Note the double equals signs for testing equality. The following example show how to use some of these operators to extract and replace elements matching specific conditions.

# Extracting subsets of data.frames

You can extract subsets of data.frames using the `subset()` function.

```
# extracting subsets
subset(datGSS,
  # rows 1 through 3
  subset = rownames(datGSS) %in% 1:3,
  # column 1 to 5
  select = 1:4)

subset(datGSS,
  # rows where age > 90
  subset = age > 90,
  ## sex and age columns
  select = c("sex", "age"))

## the $ operator can be used to extract a single column
str(datGSS$age)
```

Note that `subset()` is a convenience function; see `?Extract` for a more powerful (and complicated) way to subset data.



# Transforming data.frames

You can modify data.frames using the `transform()` function.

```
# creating new variable mean centered age
datGSS <- transform(datGSS,
                    ageC = age - mean(age))

#education difference between wifes and husbands
datGSS <- transform(datGSS,
                    educ.diff = wifeduc - husbeduc)

## ifelse() is also useful; note that the $ operator can
## also be used to create new variables.
datGSS$young <- ifelse(datGSS$age < 30, "yes", "no")

## examine our newly created variables
head(subset(datGSS,
            select = c("age", "ageC", "young", "wifeduc",
                      "husbeduc", "educ.diff")),
      n = 8)
```

Note that `transform` is a convenience function; see `?Extract` for a more powerful way to modify data.frames.

# Exporting Data

Now that we have made some changes to our GSS data set, we might want to save those changes to a file. Everything we have done so far has only modified the data in R; the files have remained unchanged.

```
# write data to a .csv file
write.csv(datGSS, file = "gss.csv")
# write data to a Stata file
write.dta(datGSS, file = "gss.dta")
# write data to an R file
saveRDS(datGSS, file = "gss.rds")
```

## Exercise 2: Data manipulation

Use the gss.rds data set

- ➊ Generate the following variables:
  - "rich" equal to 0 if rincdol is less than 100000, and 1 otherwise
  - "sinc" equal to incomdol - rincdol
- ➋ Create a subset of the data containing only rows where "usecomp" = "Yes"
- ➌ Examine the data.frame created in step 2, and answer the following questions:
  - How many rows does it have?
  - How many columns does it have?
  - What is the class of the "satjob" variable?
- ➍ BONUS (hard): Generate a variable named "dual.earn" equal to 1 if both wkftwife = 1 and wkfthusb = 1, and zero otherwise

# Topic

- 1 Workshop Materials and Introduction
- 2 Graphical User Interfaces
- 3 Data and Functions
- 4 Help and package management
- 5 Getting data into R
- 6 Data Manipulation
- 7 Basic Statistics and Graphs**
- 8 Wrap-up

# Basic statistics

Descriptive statistics of single variables are straightforward:

```
mean(datGSS$educ) # calculate mean value of education
sd(datGSS$educ) # calculate standard deviation of x
# calculate min, max, quantiles, mean of educ, age, and ageC
summary(subset(datGSS, select = c("educ", "age", "ageC")))
```

Some of these functions (e.g., `summary`) will also work with `data.frames` and other types of objects, others (such as `sd`) will not.

# Counts and proportions

Start by using the `table()` function to tabulate counts, then perform additional computations if needed

```
sex.counts <- table(datGSS$sex) # tabulate sex categories
sex.counts
prop.table(sex.counts) # convert to proportions
```

Add variables for crosstabs

```
table(subset(datGSS, select = c("sex", "happy"))) # crosstab marital X happy
```

# Statistics by classification factors

The `by()` function can be used to perform a calculation separately for each level of a classifying variable

```
by(subset(datGSS, select = c("income", "educ")),  
   INDICES=datGSS["sex"],  
   FUN=summary)
```

# Correlations

Let's look at correlations among between age, income, and education

```
cor(subset(datGSS, select = c("age", "incomdol", "educ")))
```

For significance tests, use `cor.test()`

```
with(datGSS,  
      cor.test(age, educ))
```



# Multiple regression

Modeling functions generally use the *formula* interface with DV on left followed by "~" followed by predictors—for details see

```
help("formula")
```

- Predict the number of hours individuals spend on email (emailhrs)

```
m1 <- lm(educ ~ sex + age, data = datGSS)  
summary(m1)
```

# Save R output to a file

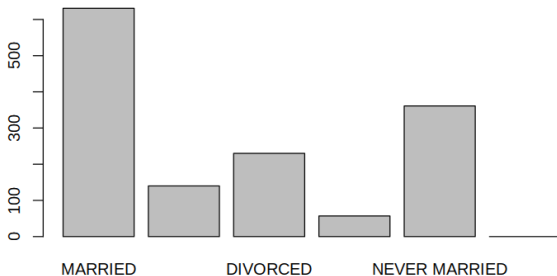
Earlier we learned how to write a data set to a file. But what if we want to write something that isn't in a nice rectangular format, like the results of our regression model? For that we can use the `sink()` function:

```
sink(file="output.txt", split=TRUE) # start logging
print("This is the result from model 1\n")
print(summary(m1))
sink() ## sink with no arguments turns logging off
```

# Basic graphics: Frequency bars

Thanks to classes and methods, you can `plot()` many kinds of objects:

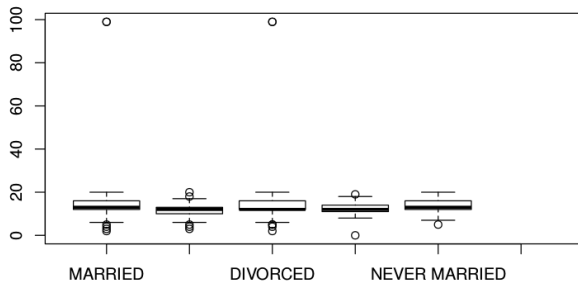
```
plot(datGSS$marital) # Plot a factor
```



# Basic graphics: Boxplots by group

Thanks to classes and methods, you can `plot()` many kinds of objects:

```
with(datGSS,  
  plot(marital, educ)) # Plot ordinal by numeric
```



# Basic graphics: Mosaic chart

Thanks to classes and methods, you can `plot()` many kinds of objects:

```
with(datGSS, # Plot factor X factor  
  plot(marital, happy))
```



# Exercise 3

Using the `datGSS` data.frame

- 1 Cross-tabulate sex and emailhrs
- 2 Calculate the mean and standard deviation of `incomdol` by sex
- 3 Save the results of the previous two calculations to a file
- 4 Create a scatter plot with `educ` on the x-axis and `incomdol` on the y-axis

# Topic

- 1 Workshop Materials and Introduction
- 2 Graphical User Interfaces
- 3 Data and Functions
- 4 Help and package management
- 5 Getting data into R
- 6 Data Manipulation
- 7 Basic Statistics and Graphs
- 8 Wrap-up**

# Help us make this workshop better!

- Please take a moment to fill out a very short feedback form
- These workshops exist for you – tell us what you need!
- <http://tinyurl.com/R-intro-feedback>



# Additional resources

- IQSS workshops:  
[http://projects.iq.harvard.edu/rtc/filter\\_by/workshops](http://projects.iq.harvard.edu/rtc/filter_by/workshops)
- IQSS statistical consulting: <http://rtc.iq.harvard.edu>
- Software (all free!):
  - R and R package download: <http://cran.r-project.org>
  - Rstudio download: <http://rstudio.org>
  - ESS (emacs R package): <http://ess.r-project.org/>
- Online tutorials
  - <http://www.codeschool.com/courses/try-r>
  - <http://www.datamind.org>
- Getting help:
  - Documentation and tutorials:  
<http://cran.r-project.org/other-docs.html>
  - Recommended R packages by topic:  
<http://cran.r-project.org/web/views/>
  - Mailing list: <https://stat.ethz.ch/mailman/listinfo/r-help>
  - StackOverflow: <http://stackoverflow.com/questions/tagged/r>