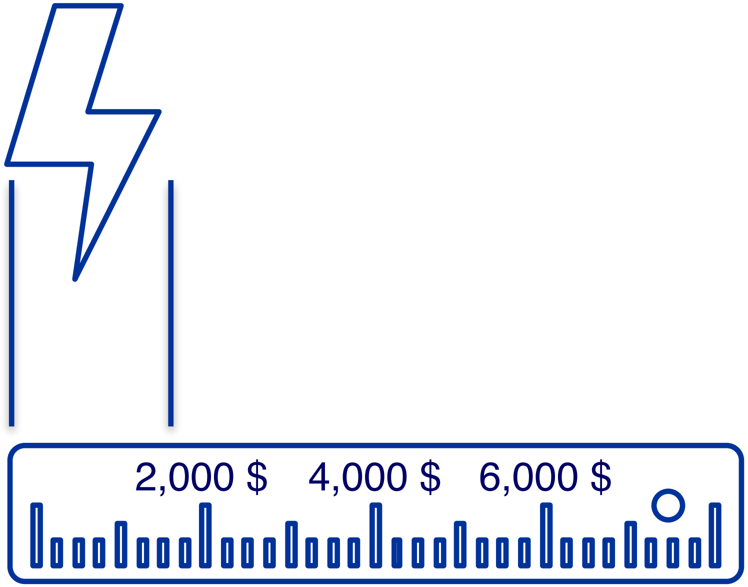
Feature Engineering in GLM

Titanic Exercise

Peter Cvacho, Data Scientist
EUBA - Bratislava
2018-03-26

Risk in our everyday life

How to measure and predict?



Solution

Generalized Linear Models

George E. P. Box:

*"Essentially, all models are wrong,
but some of them are useful"*

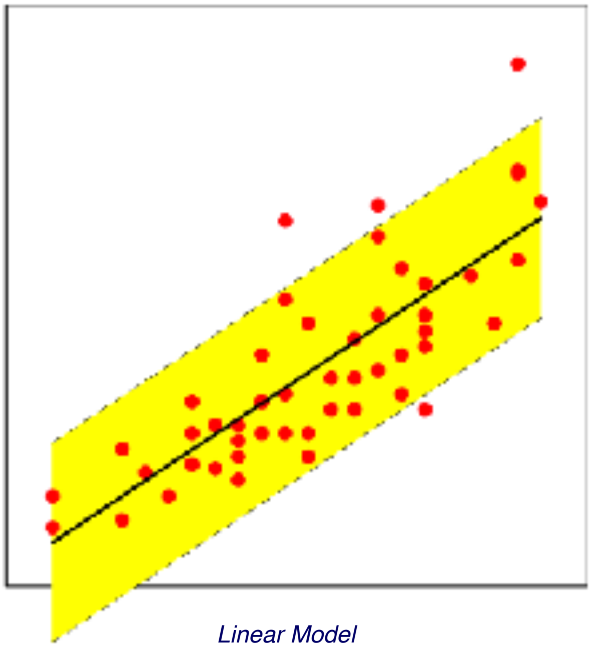
So, Why GLM?

Because we can.*

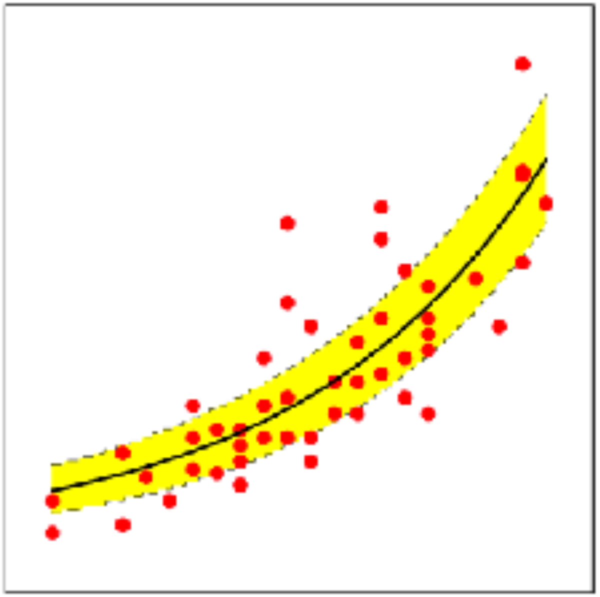
**if assumptions of GLM are satisfied*

© Zurich

LM vs. GLM



Linear Model



Generalized Linear Model (Poisson)

Credit: https://freakonometrics.hypotheses.org/9593

**simplified visualisation*

Always care about assumptions!

I assume they will catch me

Case Intro

UW Department: Our pricing is not accurate

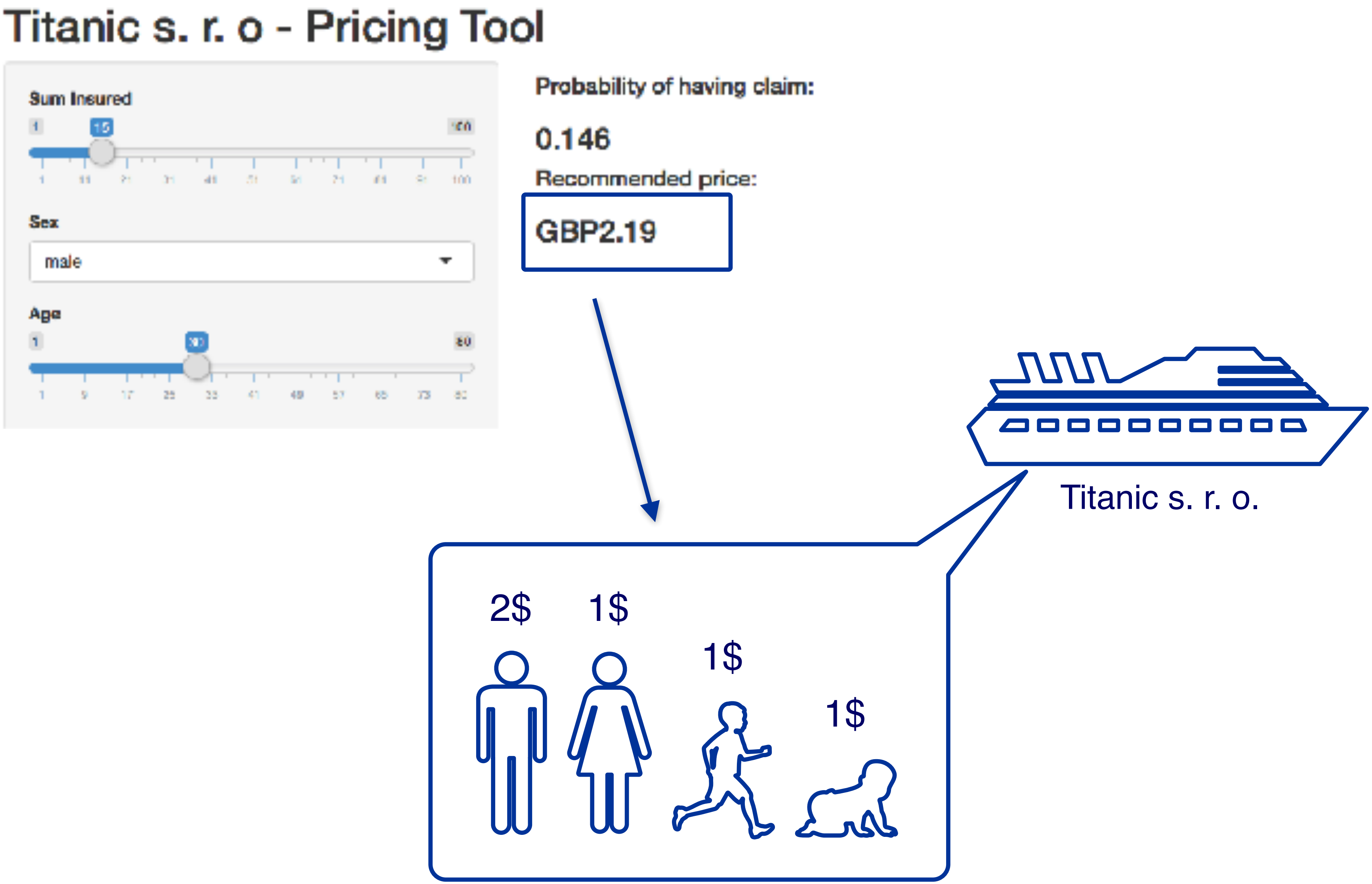
Can you help to price better?

Underwriter

Titanic s. r. o.

Case Exploration

Statistics might help!



Titanic s. r. o - Pricing Tool

Sum Insured

1 — 15 — 100

Sex

male

Age

1 — 30 — 80

Probability of having claim:

0.146

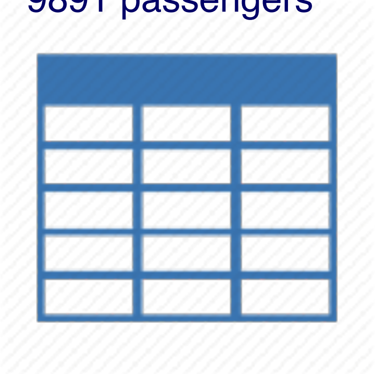
Recommended price:

GBP2.19

Solution

Generalized Linear Models

9891 passengers



Having (1) / Not having (0) Claim

Features: Sex, Age, …

Binomial Distribution *in Regression*

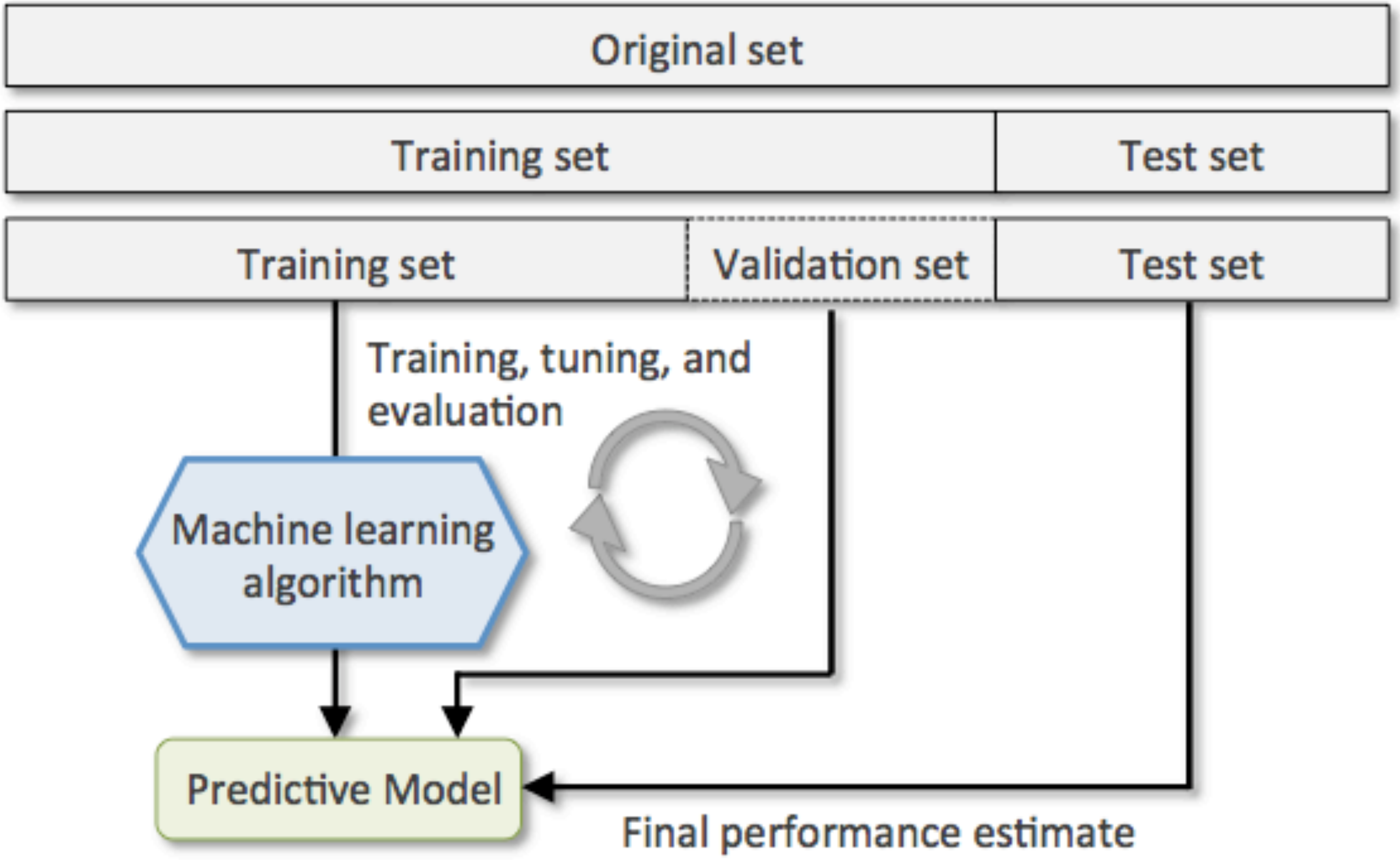
Def.: $\mathbf{E}(\mathbf{Y}) = \mu = g^{-1}(\mathbf{X}\beta)$

Link Function: $\mathbf{X}\beta = \ln\left(\frac{\mu}{1-\mu}\right)$

Mean Function: $\mu = \frac{\exp(\mathbf{X}\beta)}{1+\exp(\mathbf{X}\beta)} = \frac{1}{1+\exp(-\mathbf{X}\beta)}$

Modeling

Training vs. Validation


Original set
Training set
Test set
Training set
Validation set
Test set
Training, tuning, and evaluation
Machine learning algorithm
Predictive Model
Final performance estimate

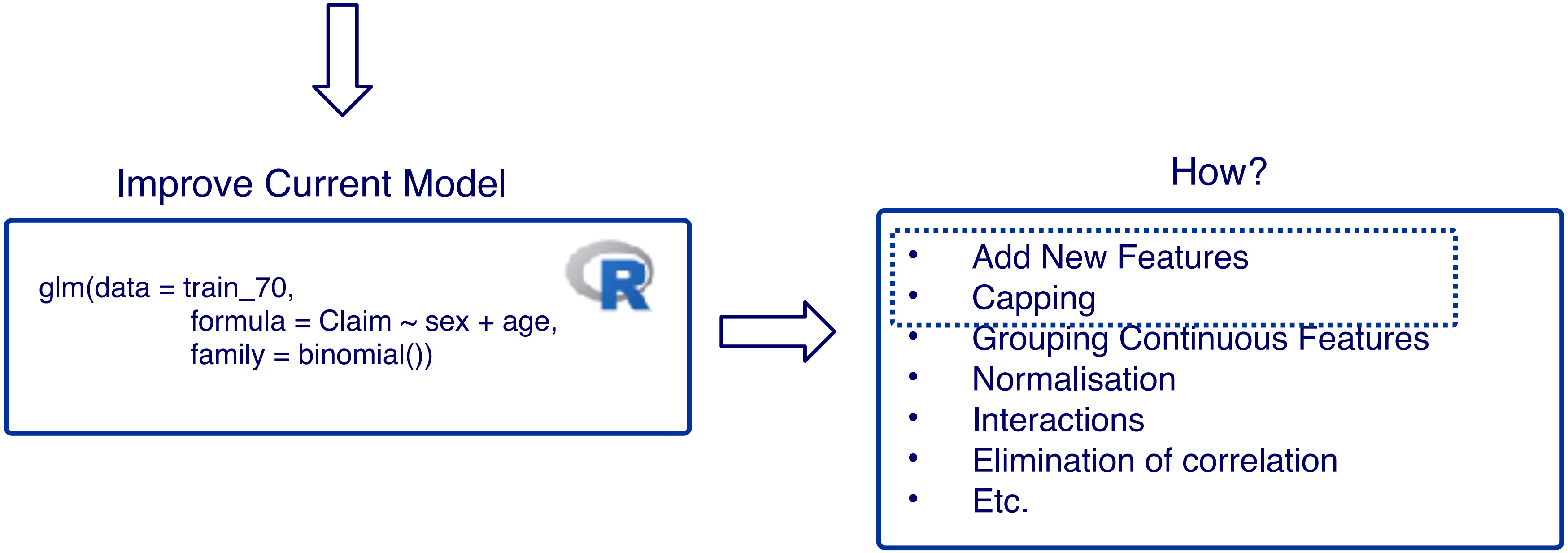
We will use rule of 70:20:10

Modeling

Exercise

Our Goal:

We want to know, what is the risk, passengers are facing during the trip and what should be sustainable price in case of accident.



Improve Current Model

```
glm(data = train_70,
        formula = Claim ~ sex + age,
        family = binomial())
```

How?

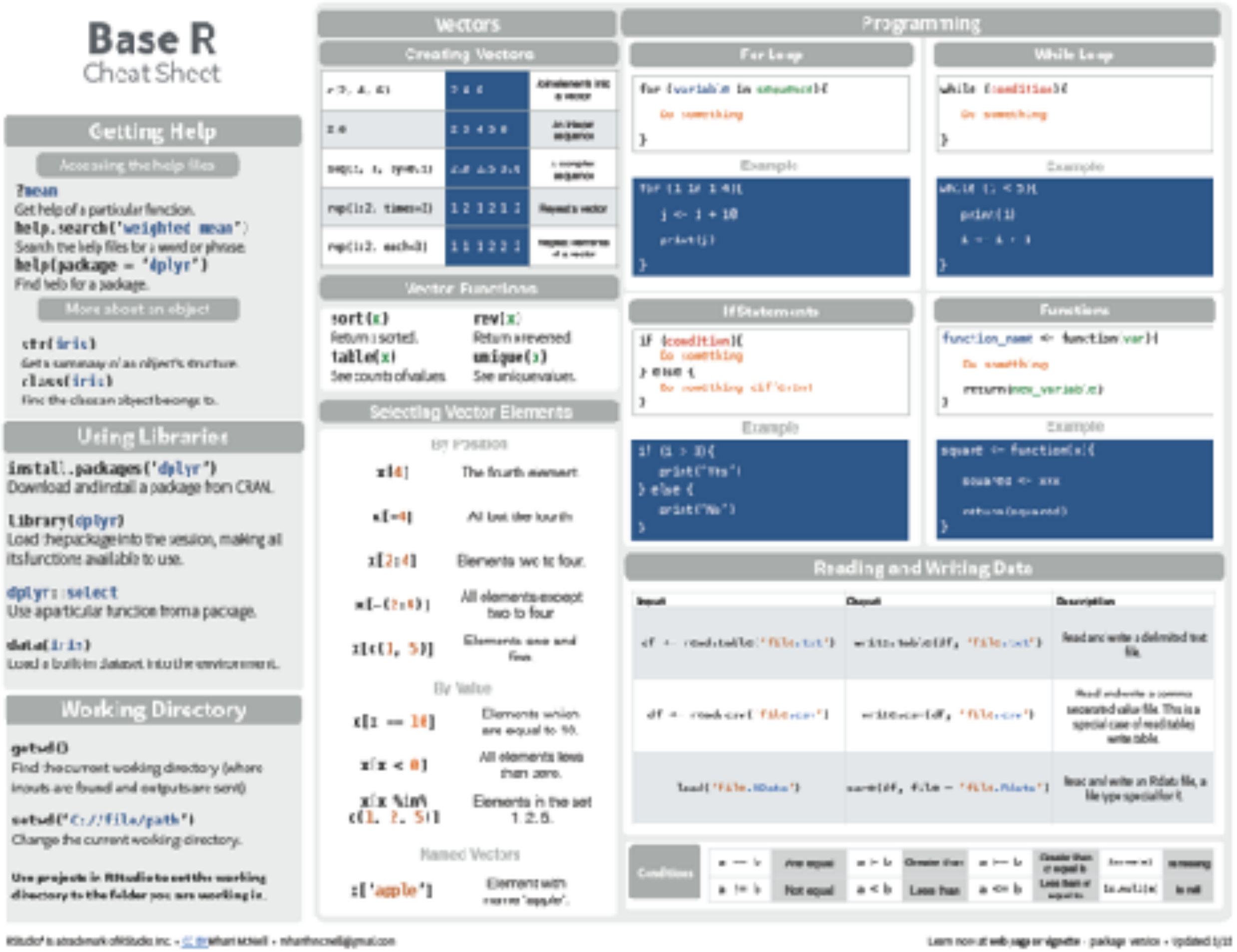
- Add New Features
- Capping
- Grouping Continuous Features
- Normalisation
- Interactions
- Elimination of correlation
- Etc.

QUESTIONS?

Cheatsheet

Base R

http://github.com/rstudio/cheatsheets/raw/master/base-r.pdf



Base R
Cheat Sheet

Getting Help

Accessing the help files

```
?mean
```
Get help of a particular function.

```
help.search('weighted mean')
```
Search the help files for a word or phrase.

```
help(package = 'dplyr')
```
Find help for a package.

More about an object

```
str(iris)
```
Get a summary of an object's structure.

```
class(iris)
```
Find the class an object belongs to.

Using Libraries

```
install.packages('dplyr')
```
Download and install a package from CRAN.

```
library(dplyr)
```
Load the package into the session, making all its functions available to use.

```
dplyr::select
```
Use a particular function from a package.

```
data(iris)
```
Load a built-in dataset into the environment.

Working Directory

```
getwd()
```
Find the current working directory (where inputs are found and outputs are sent).

```
setwd('C://file/path')
```
Change the current working directory.

Use projects in RStudio to set the working directory to the folder you are working in.

Vectors

Creating Vectors

`c(2, 4, 6)`	2 4 6	Join elements into a vector
`2:6`	2 3 4 5 6	An integer sequence
`seq(2, 3, by=0.5)`	2.0 2.5 3.0	A complex sequence
`rep(1:2, times=3)`	1 2 1 2 1 2	Repeat a vector
`rep(1:2, each=3)`	1 1 1 2 2 2	Repeat elements of a vector

Vector Functions

```
sort(x)
```
Return x sorted.

```
rev(x)
```
Return x reversed.

```
table(x)
```
See counts of values.

```
unique(x)
```
See unique values.

Selecting Vector Elements

By Position

`x[4]`	The fourth element.
`x[-4]`	All but the fourth.
`x[2:4]`	Elements two to four.
`x[-(2:4)]`	All elements except two to four.
`x[c(1, 5)]`	Elements one and five.

By Value

`x[x == 10]`	Elements which are equal to 10.
`x[x < 0]`	All elements less than zero.
`x[x %in% c(1, 2, 5)]`	Elements in the set 1, 2, 5.

Named Vectors

`x['apple']`	Element with name 'apple'.

Programming

For Loop

```
for (variable in sequence){
   Do something
}
```

Example

```
for (i in 1:4){
   j <- i + 10
   print(j)
}
```

While Loop

```
while (condition){
   Do something
}
```

Example

```
while (i < 5){
   print(i)
   i <- i + 1
}
```

If Statements

```
if (condition){
   Do something
} else {
   Do something different
}
```

Example

```
if (i > 3){
   print('Yes')
} else {
   print('No')
}
```

Functions

```
function_name <- function(var){
   Do something
   return(new_variable)
}
```

Example

```
square <- function(x){
   squared <- x*x
   return(squared)
}
```

Reading and Writing Data

Input	Output	Description
`df <- read.table('file.txt')`	`write.table(df, 'file.txt')`	Read and write a delimited text file.
`df <- read.csv('file.csv')`	`write.csv(df, 'file.csv')`	Read and write a comma separated value file. This is a special case of read.table/write.table.
`load('file.RData')`	`save(df, file = 'file.Rdata')`	Read and write an R data file, a file type special for R.

Conditions								
	`a == b`	Are equal	`a > b`	Greater than	`a >= b`	Greater than or equal to	`is.na(a)`	Is missing
	`a != b`	Not equal	`a < b`	Less than	`a <= b`	Less than or equal to	`is.null(a)`	Is null

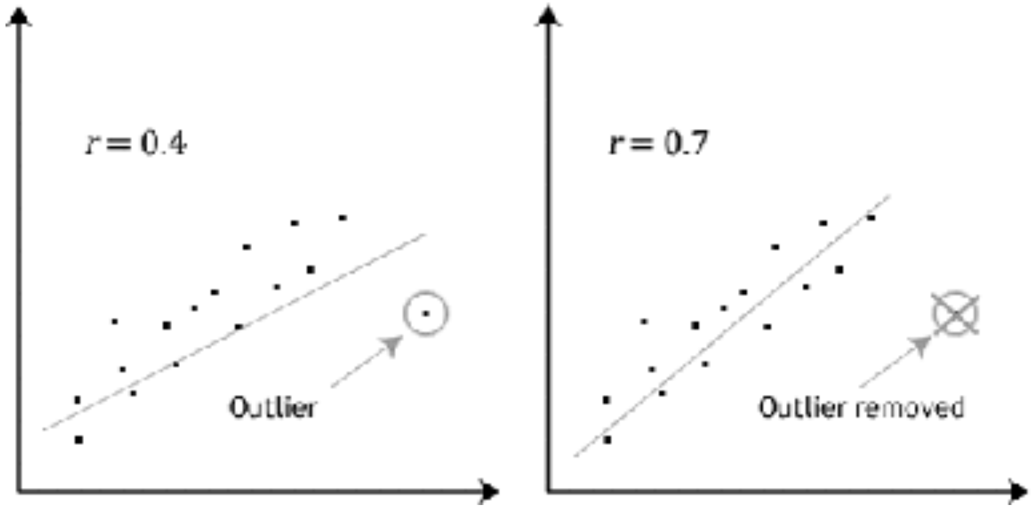
RStudio® is a trademark of RStudio, Inc. • CC BY Mhairi McNeill • mhairihmcneill@gmail.com

Learn more at web page or vignette • package version • Updated [illegible]

© Zurich

Modeling

Outliers



Summary

What have you learnt today?

- Training vs. Validation
- Checking of missing values and its imputation
- Identifying outliers and capping them

Materials

- Many Cheatsheets
  https://www.rstudio.com/resources/cheatsheets/
- More about Shiny (gallery, tutorials, articles, …)
  https://shiny.rstudio.com/
- R Programming
  http://www.cookbook-r.com/
- GLM Paper
  https://www.jstor.org/stable/2344614?seq=1#page_scan_tab_contents

https://www.surveymonkey.com/r/X6CTS3D