MODULE 2: REGRESSION AND PREDICTION

CASE STUDY INSTRUCTIONS

R START-UP INSTRUCTIONS

CASE STUDY INSTRUCTIONS

R START-UP INSTRUCTIONS

Installing R and RStudio

- Install R, a free software environment for statistical computing and graphics from CRAN, the Compre-hensive R Archive Network. It is recommended to install a precompiled binary distribution for your operating system.
- Install RStudio's IDE (Integrated Development Environment), a powerful user interface for R. RStudio Desktop is freely available.

Installing R packages

- A strength of R is that many add on packages are available which extend the capability of R. The "official" packages are hosted at CRAN and can be installed in R/RStudio with command install.packages(). The package hdm which we will use later in the course can be installed by

```
install.packages("hdm")
```

- Many packages are collected in so-called task views. For Machine Learning a good starting point is here.
- Often the most current version of packages, but also some packages which are not hosted at CRAN, are hosted at file repositories like R Forge and Github. E.g., the most current version of hdm can be installed from R-Forge by specifying the corresponding repository:

```
install.packages("hdm", repos="http://R-Forge.R-project.org")
# binary, if your system is up to date

install.packages("hdm", repos="http://R-Forge.R-project.org", type="source")
# source code, if your system is not up to date
```

- After installing a package it can be made available in the current R session with the command

```
library(hdm)
```

Packages for Machine Learning

One of the strength of R is that many useful packages for Machine Learning are available. Some of the most important ones which will also useful during the course of this course are given in the table.

Package	Description
rpart, rpart.plot, tree, party	tree-structured models for regression and classification
randomForest	random forests
nnet	single-hidden-layer neural network

mboost, gbm	boosting methods
hdm, glmnet	lasso implementations

Loading Data Sets

Using data sets is core for statistical analysis and R and its packages are shipped with many data sets for demonstration purpose. E.g. the package hdm contains some data sets. Here is an example:

```
library(hdm)

data(pension)

help(pension)
```

Finding help in R and on the web

- R has a comprehensive built-in help system. E.g. to get help for the function lm which conducts linear regression, you can use any of the following at the program's command prompt :

```
help.start()            # general help
help(lm)                # help about function lm
?lm                     # same result
apropos("lm")           # list all functions containing string lm
??lm                    # extensive search on all documents containing the string "lm"
example(lm)             # show an example of function lm
RSiteSearch("lm")       # search for lm in help manuals and archived mailing lists
```

Moreover, many packages contain introductions called "vignettes".

```
# get vignettes on using installed packages
vignette()                  # show available vingettes
vignette(package="hdm")     # show the names of vignettes contained in the package hdm
vignette("hdm_introduction") # show the vignette
```

For information on help search in are can be found on this stackoverflow question

- R is shipped with different manuals where "An Introduction to R" is a good starting point to learn more about R. Moreover, good sources for help are stackoverflow and the archive of the R-help list where solutions to many problems can be found.