Workflow of statistical data analysis

Oliver Kirchkamp
Workflow of empirical work may seem obvious. It is not. Small initial mistakes can lead to a lot of hard work afterwards. In this course we discuss some techniques that hopefully facilitate the organisation of your empirical work.

This handout provides a summary of the slides from the lecture. It is not supposed to replace a book.

Many examples in the text are based on the statistical software R. I urge you to try these examples on your own computer.

As an attachment of this PDF you find a file \texttt{wf.zip} with some raw data. You also find a file \texttt{wf.Rdata} with some R functions and some data already in R's internal format.

The drawing on the previous page is Albercht Dürer's "Der Hafen von Antwerpen" — an example for workflow in a medieval city.

\section*{Contents}

\textbf{1 Introduction} \hfill 4
\subsection*{1.1 Motivation} \hfill 4
\subsection*{1.2 Structure of a paper} \hfill 6
\subsection*{1.3 Aims of statistical data analysis} \hfill 6
\subsection*{1.4 Creativity and chaos} \hfill 8
\subsection*{1.5 Making the analysis reproducible} \hfill 10
\subsection*{1.6 Preserve raw data} \hfill 11
\subsection*{1.7 Interaction with coauthors} \hfill 11

\section*{2 Digression: R} \hfill 11
\subsection*{2.1 Installation of R} \hfill 11
\subsection*{2.2 Types and assignments} \hfill 12
\subsection*{2.3 Functions} \hfill 15
\subsection*{2.4 Random numbers} \hfill 16
\subsection*{2.5 Example Datasets} \hfill 16
\subsection*{2.6 Graphs} \hfill 19
\subsection*{2.7 Basic Graphs} \hfill 19
\subsubsection*{2.7.1 Plotting functions} \hfill 20
\subsubsection*{2.7.2 Empty plots} \hfill 21
\subsubsection*{2.7.3 Line type} \hfill 21
\subsubsection*{2.7.4 Points} \hfill 22
\subsubsection*{2.7.5 Legends} \hfill 22
\subsubsection*{2.7.6 Auxiliary lines} \hfill 23
\subsubsection*{2.7.7 Axes} \hfill 24
\subsection*{2.8 Fancy math} \hfill 25
\subsubsection*{2.8.1 Several diagrams} \hfill 25
\subsection*{2.9 Tables} \hfill 27
# Organising work

## Scripts

- Robust scripts
- Robustness towards different computers
- Robustness towards changes in context
- Functions increase robustness

## Calculations that take a lot of time

## Nested functions

## Reproducible randomness

## Recap — writing scripts and using functions

## Human readable scripts

# Some programming techniques

## Debugging functions

## Lists of variables

## Return values of functions

## Repeating things

# Data manipulation

## Subsetting data

## Merging data

## Reshaping data

# Preparing Data

## Reading data

- Reading z-Tree Output
- Reading and writing R-Files
- Reading Stata Files
- Reading CSV Files
- Filesize

## Checking Values

- Range of values
- (Joint) distribution of values
- (Joint) distribution of missings
- Checking signatures

## Naming variables

## Labeling (describing) variables

## Labeling values

## Recoding data

- Replacing values by missings
- Replacing values by other values
1 Introduction

1.1 Motivation

**Literature:** Surprisingly, there is not much literature about workflow of statistical data analysis:

**General Literature**

Hadley Wickham; Tidy Data; *Journal of Statistical Software*, 2014.

**Literate Programming**

- Friedrich Leisch; *Sweave User Manual*.
- Nicola Sartori; *Sweave = \text{R} \cdot \LaTeX^2*.
- Max Kuhn; CRAN Task View: Reproducible Research.

**Version control**

- Scott Chacon, Ben Straub; *Pro Git*.
- Ben Collins-Sussman, Brian W. Fitzpatrick, C. Michael Pilato; *Version Control with Subversion*.

**What is workflow:**

- A sequence of operations.
- A pattern of actions that can be documented and learned.

- We spend a lot of time explaining statistical methods to students.
- We do not tell students how to apply these methods (how to integrate methods into a “workflow”)
- Why?
- Is “workflow” obvious? — I do not think so.
  - Is the wrong workflow not costly? — On the contrary.
    - Mistakes in the statistical method can always be cured.
    - Mistakes in the workflow can render the entire project invalid — no cure possible (e.g. loss of data, loss of understanding the data, loss of methods applied)
- Isn’t it sufficient to simply store and backup everything?
  - unfortunately not — statistical analysis tends to create a lot of data. → storing everythings means hiding everything very well from ourselves and from others.
1.2 Structure of a paper

- Describe the research question
  Which *economic model* do we use to structure this question?
  Which *statistical model* do we use for inference? (Estimation, hypothesis testing, classification . . .)

- Describe the economic method (experiment, field data, . . .)

- Describe the sample
  How many observations, means, distributions of main variables, key statistics
  Is there enough variance in the independent variables to test what we want to test?

- Statistical inference (estimate, test hypotheses, classify, . . .)
  possibly different variants of the model (increasing complexity)

- Discuss the model, robustness checks

1.3 Aims of statistical data analysis

- Limit work and time

- Get interesting results

- Replicability
  - for us, to understand our data and our methods after we get back to work after a short break
  - for our friends (coauthors), so that they can understand what we are doing
  - for our enemies — we should always (even years after) be able to *prove* our results *exactly*

- If statistical analysis was a straightforward procedure, then there would be no problem:
  - Store the raw data. All methods we apply are obvious and trivial.

- In the real world our methods are far from obvious:
  - We think quite a lot about details of our statistical analysis

- Assume we have another look at our paper (and our analysis) after a break of 6 month:
  - What does it mean if sex==1 ?
- For the variable meanContribution: was the mean taken with respect to all players and the same period, or with respect to the same player and all periods, or ...
- What is the difference between payoff and payoff2...
- Do the tables and figures in version 27 of the paper ...
  * ...refer to all periods of the experiment or only to the last 6 periods?
  * ...do they include data from the two pilot experiments we ran?
  * ...do they refer to the “cleaned” dataset, or to the “cleaned dataset in long form” (where we eliminated a few outliers)
  * Do all tables and figures and p-values and t-tests... actually refer to the same data? (or do some include outliers, some not, ...)

Assume we take only 10 not completely obvious decisions between two alternatives during our analysis (which perhaps took us 1 week), ...

(coding of data, data to include, treatments to compare, lags to include, outliers to remove, interaction terms to include, types of model comparison, dealing with non-linearities, correlation structure of error terms, ...)

... ⇒ we will have to explore $2^{10} = 1024$ variants of our analysis (= 1024 weeks) to recover what we actually did.

Often we take more than 10 not completely obvious decisions.
⇒ we should follow a workflow that facilitates replicability.
This is not obvious, since workflow is (unfortunately) not linear:

During this process we create a lot of intermediate results. How can we organise these results?
Solutions and restrictions:

- Store everything — not feasible
- We want to be creative, take shortcuts, we want to explore things, play with different representations of a solution...
- During this phase we can not document everything.

1.4 Creativity and chaos

Living two lives:

- creative (undocumented)
- permanent (documented)

Let our computer(s) reflect these two lives:

```
.../projectXYZ/
   /permanent/
     /rawData
     /cleanData
     /R
     /Paper
     /Slides
   /creative/
     /cleanData
     /R
     /Paper
     /Slides
```

You might need more directories for your work.

(In terms of version control, which we will cover later, “permanent” could be a trunk, while “creative” could be a branch)

Rules

1. Anything that we give to other people (collaborators, journals,...) must come entirely from permanent

2. Never delete anything from permanent

3. Never change anything in permanent

4. We must be able trace back everything in permanent clearly to our raw data.

Since we give things to other people more than once (first draft, second draft,..., first revision, ..., second revision,...), we must be able to replicate each of these instances.
Consequences — permanent data has versions  (Below we will discuss the advantages of a version control system (git, svn). Let us assume for a moment that we have to do everything manually.)

- We will accumulate versions in our permanent life (do not delete them, do not change them)
  
cleaned_data_150721.Rdata
cleaned_data_150722.Rdata
cleaned_data_150722b.Rdata
:
preparingData_150721.R
preparingData_150722.R
descriptives_150722.R
econometrics_150723.R
:
paper_150724.Rnw
paper_150725.Rnw
paper_150727.Rnw
:

What it the optimal workflow?  The optimal workflow is different for each of us

Aims

- Exactness (allow clear replication)
- Efficiency
- We must like it (otherwise we don’t do it)
- Whatever we do, we should do it in a systematic way
  - Follow a routine in our work (all projects should follow similar conventions)
  - Let the computer follow a routine (a mistake made in a routine will show up “routinely”, a hand coded mistake is harder to detect).
    Use functions, try to make them as general as possible.
  - Prepare for the unexpected! We should not assume that our data will always look the way it looks at the moment.

More on routines  Example:

- Probability to make a mistake: 0.1
- Probability to discover (and fix) a mistake: 0.8
Now you solve two related problems, A and B:

- Both problems are solved independently:
  - Probability of (undiscovered) mistake A: $0.1 \cdot 0.2$
  - Probability of (undiscovered) mistake B: $0.1 \cdot 0.2$
  - Probability of some undiscovered mistake: $1 - .98^2 \approx 0.04$

- Both problems are solved with the same routine (one function in your code):
  - Probability of some undiscovered mistake: $0.1 \cdot 0.2^2 = 0.004$

Producing your results with the help of identical (and computerised) routines makes it much easier to discover mistakes.

### 1.5 Making the analysis reproducible

Here are again the steps in writing a paper:

1. organise raw data
2. descriptive analysis (figures, descriptive tables…)
3. develop methods for analysis
4. get results (run program code)
5. write paper (mix results with text and explanations)
6. interact with collaborators

- All these tasks require decisions.
- All these decisions should be documented.
- When is our documentation sufficient? — If a third person, without our help, can find out what we were doing in all the above steps. If we want to have another look at our data in one year’s time we will be in the same position as an outsider today.
- We keep a log where we document the above steps for a given project on a daily basis (research log) (nobody wants to keep logs, so this must be easy)
1.6 Preserve raw data

- If our raw data comes from z-Tree experiments: We better keep all programs (the current version can always be found as @1.ztt,... in the working directory).

- If our raw data includes data from a questionnaire:
  - We need a codebook
    * variable name — question number — text of the questions
    * branching in the questionnaire
    * levels (value labels) used for factors
    * missing data, how was it coded?
    * cleaned data, how was it cleaned? (if we have no access to the raw data)

1.7 Interaction with coauthors

- Clear division of labour
  - the “experimenter” decides how the experiment is actually run
  - the “empiricist” decides what statistics and graphs are produced
  - the “writer” decides how to present the text
  - help, do not interfere

- In your communication: concentrate on the essentials:
  - exchange one file
  - make only essential changes to this file
  - clearly explain why these changes are necessary

2 Digression: R

For the purpose of the course we take R as an example for one statistical language. Even if you use other languages for your work, you will find that the concepts are similar.

2.1 Installation of R

On the Homepage of the R Projekt you find in the menu on the left a link Download / CRAN. This link leads to a choice of “mirrors”. If you are in Jena, the GWDG Mirror in Göttingen might be fast. There you also find instructions how to install R on your OS.
Installation of Libraries  If the command `library` complains about not being able to find the required library, then the library is most likely not installed. The command

```r
install.packages("Ecdat")
```

installs the library Ecdat. Some installations have a menu “Packages” that allows you to install missing libraries. Users of operating systems of Microsoft find support at the [FAQ for Packages](#).

### 2.2 Types and assignments

R knows about different *types* of data. We will meet some types in this chapter. To assign a number (or a value, or any object) to a variable, we use the operator `<-

```r
x <- 4
```

R stores the result of this assignment as *double*

```r
typeof(x)
[1] "double"
```

Now we can use `x` in our calculations:

```r
2 * x
[1] 8
```

```r
sqrt(x)
[1] 2
```

Often our calculations will not only involve a single number (a scalar) but several which are connected as a vector. Several numbers are connected with `c`

```r
x <- c(21, 22, 23, 24, 25, 16, 17, 18, 19, 20)
x
```

When we need a long list of subsequent numbers, we use the operator `:`

```r
21:30
```

```r
y <- 21:30
```

**Subsets**  We can access single elements of a variable with `[]`
When we want to access several elements at the same time, we simply use several indices (which are connected with `c`). We can use this to change the sequence of values (e.g. to sort).

```r
x[c(3,2,1)]
[1] 23 22 21

x[3:1]
[1] 23 22 21

x
```

(to sort a long vector we would use the function `order`).

```r
order(x)
[1] 6 7 8 9 10 1 2 3 4 5

x[order(x)]
[1] 16 17 18 19 20 21 22 23 24 25
```

(order determines an “ordering”, i.e. a sequence in which the elements of the dataset should be ordered. We use `x[...]` to see the ordered result.)

Negative indices drop elements:

```r
x[-1:-3]
[1] 24 25 16 17 18 19 20
```

**Logicals**  
Logicals can be either `TRUE` or `FALSE`. When we compare a vector with a number, then all the elements will be compared (this follows from the recycling rule, see below):

```r
x < 20
[1] FALSE FALSE FALSE FALSE FALSE TRUE TRUE TRUE TRUE FALSE
```

We can use logicals as indices, too:

```r
```
Characters  Not only numbers, also character strings can be assigned to a variable:

```r
x <- "Mary"
```

We can also work with vectors of character strings:

```r
x <- c("John","Mary","Jane")
x[2]
[1] "Mary"
x[3]<="Lucy"
x
[1] "John" "Mary" "Lucy"
```

Factors  Often it is clumsy to store a string of characters again and again if this string appears in the dataset several times. We might, e.g., want to store whether an observation belongs to a man or a woman. This can be done in an efficient way by storing 2 for "male", and 1 for "female".

```r
x <- as.factor(c("male","female","female","male"))
levels(x)
[1] "female" "male"
x[2]
[1] female
Levels: female male
as.numeric(x)
[1] 2 1 1 2
```

Usually the first level in a factor is the level that comes first on the alphabet. If we do not want this, we can `relevel` a factor:

```r
x<-relevel(x,"male")
x
[1] male female female male
Levels: male female
as.numeric(x)
[1] 1 2 2 1
```
Note that the meaning of the values remains unchanged. Sometimes, when we have more than only two levels, we want to order levels of a factor along a third variable. This is done by `reorder`.

```r
y <- c(12,7,8,11)
reorder(x,y)
```

```
[1] male female female male
attr(,"scores")
 male female
 11.5  7.5
Levels: female male
```

### 2.3 Functions

R knows many built-in functions:

```r
mean(x)
median(x)
max(x)
min(x)
length(x)
unique(c(1,2,3,4,1,1,1))
```

When we need more, we can write our own:

```r
square <- function(x) {
  x*x
}
```

The last expression in a function (here `x*x`) is the return value. Now we can use the function.

```r
square(7)
```

```
[1] 49
```

When we want to apply a function to many numbers, `sapply` helps:

```r
range <- 1:10
sapply(range,square)
```

```
[1]  1  4  9 16 25 36 49 64 81 100
```

With `sapply` we do not have to define a name for a function:

```r
sapply(range,function(x) x*x)
```

```
[1]  1  4  9 16 25 36 49 64 81 100
```
2.4 Random numbers

Random numbers can be generated for rather different distributions. R calculates pseudo-random numbers, i.e. R picks numbers from a very long list that appears random. Where we start in this long list is determined by `set.seed`:

```r
set.seed(123)

10 pseudo-random numbers from a normal distribution can be obtained with

```r
rnorm(10)
```

```
[1] -0.56047565 -0.23017749 1.55870831 0.07050839 0.12928774 1.71506499
[7] 0.46091621 -1.26506123 -0.68685285 -0.44566197
```

We get the same list when we initialise the list with the same starting value:

```r
set.seed(123)
rnorm(10)
```

```
[1] -0.56047565 -0.23017749 1.55870831 0.07050839 0.12928774 1.71506499
[7] 0.46091621 -1.26506123 -0.68685285 -0.44566197
```

This is very useful, when we want to replicate the same “random” results. 10 uniformly distributed random numbers from the interval \([100,200]\) can be obtained with

```r
runif(10,min=100,max=200)
```

```
[1] 188.9539 169.2803 164.0507 199.4270 165.5706 170.8530 154.4066 159.4142
[9] 128.9160 114.7114
```

Often we use random numbers when we simulate (stochastic) processes. To replicate a process we use the command `replicate`. E.g.

```r
replicate(10,mean(rnorm(100)))
```

```
[1] 0.016749257 -0.024755975 0.061320514 -0.028205903 0.087712299
[6] -0.025113287 -0.141043824 0.123989920 0.109293109 -0.002743263
```

takes 10 times the mean of each 100 pseudo-normally distributed random numbers.

2.5 Example Datasets

We just saw that the command `c` allows us to describe the elements of a vector. For long datasets this is not very convenient. R contains already a lot of example datasets. These datasets are, similar to statistical functions, organised in libraries. To save space
and time R does not load all libraries initially. The command `library` allows us to load a library with a dataset at any time.

The library `Ecdat` provides a lot of interesting economic datasets. The library `memisc` gives access to some interesting functions that help us organising our data.

When we need a specific function and we do not know in which library to look for this function we can use the command `RSiteSearch` or the R Site Search Extension for Firefox.

The dataset `BudgetFood` is, e.g., contained in the library `Ecdat`.

```r
data(BudgetFood, package="Ecdat")
```

To really see the numbers, we can use the command `fix`:

```r
fix(BudgetFood)
```

Usually we do not want to see many numbers. Instead we want to derive (in a structured way) a few numbers (parameters, confidence intervals, p-values,...)

The command `help` aids us in finding out the meaning of the numbers of the different columns of a dataset.

```r
help(BudgetFood)
```

An important command to get a summary is `summary`

```r
summary(BudgetFood)
```

How can we access specific columns from our dataset? Since R may have several datasets at the same time in its memory, there are several possibilities. One possibility is to append the name of the dataset `BudgetFood` with a `$` and then the name of the column.

```r
BudgetFood$age
```

This is helpful when we work with several different datasets at the same time.

The example also shows that R does not flood our screen with long lists of numbers. Instead we only see the first few numbers, and then the text “omitted ... entries”.

When we want to use only one dataset, then the command `attach` is helpful.

```r
attach(BudgetFood)
```

```r
age
```

```r
[1] 43 40 28 60 37 35 40 68 43 51 48 51 58 61 53 58 64 50 50 47 76 49 44 49
[26] 51 56 63 30 70 29 60 50 56 36 46 43 32 45 34
[ reached getOption("max.print") -- omitted 23932 entries ]
```
From now on, all variables will first be searched in the dataset BudgetFood. When we no longer want this, then we say

```r
detach(BudgetFood)
```

A third possibility is the command `with`:

```r
with(BudgetFood, age)
```

```
[1] 43 40 28 60 37 35 40 68 43 48 51 43 48 51 58 61 53 58 64 50 50 47 76 49 44 49
[26] 51 56 63 30 70 29 60 50 56 36 46 43 32 45 34
[ reached getOption("max.print") -- omitted 23932 entries ]
```

We often use `with` when we use a function and want to refer to a specific dataset in this function. E.g. `hist` shows a histogram:

```r
with(BudgetFood, hist(age))
```

![Histogram of age](image)

Most commands have several options which allow you to fine-tune the result. Have a look at the help-page for `hist` (you can do this with `help(hist)`). Perhaps you prefer the following graph:

```r
with(BudgetFood, hist(age, breaks=40, xlab="Age [years]", col=gray(.7), main="Spain"))
```
2.6 Graphs

There is more than one way to represent numbers as graphs.

2.7 Basic Graphs

Here are three basic graphs:

```r
with(BudgetFood, {
  hist(age)
  plot(density(age))
  boxplot(age ~ sex, main="Boxplot")
})
```

Two further helpful plots are `ecdf` and `qqnorm:`
Sometimes it is obvious how to prepare our data for these functions. Sometimes it is more complicated. Then other commands help and calculate an object that can be plotted (with `plot`):

- `density`, `ecdf`, `xyplot`...

Some commands then plot whatever we have prepared:

- `plot`, `hist`, `boxplot`, `barplot`, `curve`, `mosaicplot`...

Yet other commands add something to an existing plot:

- `points`, `text`, `lines`, `abline`, `qqline`...

### 2.7.1 Plotting functions

We can plot functions of `x` with `curve`.

```r
curve(dchisq(x, 3), from = 0, to = 10)
```
2.7.2 Empty plots

Sometimes it is helpful to start with an empty plot. Then we have to help `plot` a little bit. Usually, `plot` can guess from the data the limits and labels of the axes. With an empty plot we have to specify them explicitly.

```
plot(NULL, xlim=c(0,10), ylim=c(-3,6), xlab="x", ylab="y", main="empty plot")
```

2.7.3 Line type

Almost all commands that draw lines follow the following conventions:

- `lty` linetype ("dashed", "dotted", or simply a number)
2.7.4 Points

The character used to draw points is determined with pch.

```r
range=1:20
plot(range,range/range,pch=range,frame=FALSE)
text(range,range/range+.2,range)
```

2.7.5 Legends

When we use more than one line or more than one symbol in our plot we have to explain their meaning. This is done in a legend.

Usually legend gets as an option a vector of linetypes lty and symbols pch. They will be used to construct example lines and symbols next to the actual text of the legend. If the lty or pch is NA, then no line or point is drawn.

```r
plot(NULL,xlim=c(0,10),ylim=c(-3,6),xlab="x",ylab="y",main="empty plot")
legend("topleft",c("Text 1","more Text","even more"),lty=1:3,pch=1:3)
legend("bottomright",c("no line no symbol","line only","line and symbol","symbol only"),lty=c(NA,1:3),pch=c(NA,1:4))
```
2.7.6 Auxiliary lines

The command `abline` allows us to add auxiliary lines to a plot.

```r
plot(NULL, xlim=c(0,10), ylim=c(-3,6), xlab="x", ylab="y", main="main title")
abline(h=2:6, lty="dotted")
abline(v=5, lty="dashed")
abline(a=-1, b=1, lwd=5, col=grey(.7))
legend("bottomright", c("h", "v", "a/b"), lty=c("dotted", "dashed", "solid"), col=c("black", "black", grey(.7)), lwd=c(2,2,5))
```

`abline` knows the following important parameters:

- **h** = for horizontal lines
- **v** = for vertical lines
• $a=\ldots$, $b=\ldots$ for lines with intercept $a$ and slope $b$

Note, that these arguments can be vectors if we want to draw several lines at the same time.

### 2.7.7 Axes

The options `log='x'`, `log='y'` or `log='xy'` determine whether which axis is shown in a logarithmic style.

```r
data(PE, package="Ecdat")
x<-as.data.frame(PE)
attach(x)

plot(price, earnings)  # Normal plot
plot(price, earnings, log="x")  # Logarithmic x-axis
plot(price, earnings, log="xy")  # Logarithmic x and y-axes

To gain more flexibility `axis` can draw a wide range of axes. Before using `axis` the previous axes can be removed entirely (`axes=FALSE`) or suppressed selectively (`xaxt="n"` or `yaxt="n"`).

```r
plot(price, earnings, log="xy", axes=FALSE)
plot(price, earnings, log="xy", xaxt="n")
plot(price, earnings, log="xy", xaxt="n")
axis(1, at=c(5,10,20,40,80,160,320,640,1280))
```
If we specify a lot of axes labels, as in the example above, R does not print them all if they overlap.

### 2.8 Fancy math

R can also render more than only textual labels with `plotmath`:

```r
plot(price, earnings, xlab='\(\pi_1\)', ylab='\(\hat{\gamma}_0\)',
     main="the \(\int_\theta^{\bar{\theta}} \sqrt{\xi} d\phi\)"
abline(lm(earnings~price))
legend("bottomright",c("legend","\(\xi^2\","line \(\phi\)"),pch=c(NA,1,NA),lty=c(NA,NA,1))
```

2.8.1 Several diagrams

#### Diagrams side by side

To put several diagrams on one plot side by side we can call `par(mfrow=c(...))` or `layout` or `split.screen`.

```r
par(mfrow=c(1,2))
with(BudgetFood, {
  hist(age)
  plot(density(age))
})
```
Superimposed graphs

- Anything that can create lines or points (like density or ecdf) can immediately be added to an existing plot.

- Plot-objects that would otherwise create a new figure (like plot, hist, or curve) can be added to an existing plot with the optional parameter add=TRUE.

```r
with(BudgetFood, {
  plot(density(age), lwd=2)
  lines(density(age[sex=="man"], na.rm=TRUE), lty=3, lwd=2)
  hist(age, freq=FALSE, add=TRUE)
  curve(dnorm(x, mean(age), sd(age)), add = TRUE, lty=2)
})
```

Coplots We will discuss coplots in section ??.
2.9 Tables

Tables of frequencies The command `table` calculates a table of frequencies. Here we show only the first 16 columns:

```r
with(BudgetFood, table(sex, age))[,1:16]
```

<table>
<thead>
<tr>
<th>age</th>
</tr>
</thead>
<tbody>
<tr>
<td>sex</td>
</tr>
<tr>
<td>16</td>
</tr>
<tr>
<td>17</td>
</tr>
<tr>
<td>18</td>
</tr>
<tr>
<td>19</td>
</tr>
<tr>
<td>20</td>
</tr>
<tr>
<td>21</td>
</tr>
<tr>
<td>22</td>
</tr>
<tr>
<td>23</td>
</tr>
<tr>
<td>24</td>
</tr>
<tr>
<td>25</td>
</tr>
<tr>
<td>26</td>
</tr>
<tr>
<td>27</td>
</tr>
<tr>
<td>28</td>
</tr>
<tr>
<td>29</td>
</tr>
<tr>
<td>30</td>
</tr>
<tr>
<td>31</td>
</tr>
</tbody>
</table>

Other statistics The command `aggregate` groups our data by levels of one or several factors and applies a function to each group. In the following example the factor is `sex`, the function is the mean which is applied to the variable `age`.

```r
with(BudgetFood, aggregate(age ~ sex, FUN=mean))
```

<table>
<thead>
<tr>
<th>sex</th>
<th>age</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 man</td>
<td>49.08985</td>
</tr>
<tr>
<td>2 woman</td>
<td>59.47445</td>
</tr>
</tbody>
</table>

2.10 Regressions

Simple regressions can be estimated with `lm`. The operator `~` allows us to describe the regression equation. The dependent variable is written on the left side of `~`, the independent variables are written on the right side of `~`.

```r
lm(wfood ~ totexp, data=BudgetFood)
```

Call:
`lm(formula = wfood ~ totexp, data = BudgetFood)`

Coefficients:
```
(Intercept) totexp
 0.4950397225 -0.0000001348
```

The result is a bit terse. More details are shown with the command `summary`.

```r
summary(lm(wfood ~ totexp, data=BudgetFood))
```

Call:
`lm(formula = wfood ~ totexp, data = BudgetFood)`

Residuals:
```
  Min  1Q Median  3Q    Max
-0.49307 -0.09374 -0.01002 0.08617 1.06182
```
2.11 Starting and stopping R

Whenever we start R, the program attempts to find a file .Rprofile, first in the current working directory, then in the home directory. If the file is found, it is “sourced”, i.e. all R commands in this file are executed. This is useful when we want to run the same commands whenever we start R. The following line

```r
options(browser = "/usr/bin/iceweasel")
```

in .Rprofile makes sure that the help system of R always uses iceweasel.

Also when we quit R with the command q(), the application tries to make our life easier.

```r
q()
```

R first asks us

Save workspace image? [y/n/c]:

Here we have the possibility to save all the data that we currently use (and that are in our workspace) in a file .Rdata in the current working directory. When we start R for the next time (from this directory) R automatically reads this file and we can continue our work.

3 Organising work

3.1 Scripts

Most of the practical work in data analysis and statistics can be see as a sequence of commands to a statistical software.

How can we run these commands?

- Execute a command in the command window (or with mouse and dialog boxes)
  - clumsy
– hard to repeat actions
– hard to replicate what we did and why we did it (logs don’t really help).
– hard to find mistakes (structure of the mistake is easy to overlook).

• Write a file (.R or .do) and execute single lines (or small regions) from the file while editing the file.
  – great way to creatively develop code line by line. Not reproducible since the file changes permanently.
  – one window with the file another window with mainly the R output

• Write a source file (.R or .do), open it in an editor and then always execute the entire file (while editing the file).
  – great way to creatively develop larger parts of code

• Source “public” files (.R or .do) from a “master file”

```r
source("read_data_160715.R")
source("clean_data_160715.R")
source("create_figures_160715.R")
```

This is the first step to reproducible research. When our script seems to do what it is supposed to do, we make it “public”, give it a unique name, and never change it again.

• From a master file, first source a file which defines functions. Then call these functions.

```r
source("functions_XYZ_160715.R")
read_data()
clean_data()
create_figures()
```

Advantages of using functions:
– functions can take parameters.
– several functions go in one file (still do not harm each other). Systematic changes are easier with only one file.
– Regardless whether we divide our work into source files or into functions: This division allows us to save time. Some of these steps take a lot of time. Once they work, we do not have to do them over and over again.

Advantages of using source files (with or without functions):
• We keep a record of our work.
• We can work incrementally, fix mistakes and introduce small changes (if we refer to a public file, we should work on a copy of this file with a new name).
• We can use the editor of our choice (Emacs is a nice editor)
3.1.1 Robust scripts

How can we make our scripts “robust”? Remember:

- The structure of the data may change over time.
  - New variables might come with new treatments of our experiment.
  - New treatments might require that we code variables differently.

- The scripts may not only run on our computer.

- The scripts are not always sourced in the same context.

- Our random number generator may start from different seeds.

3.1.2 Robustness towards different computers

We better use relative pathnames.

Assume that on my computer the script is stored in

/home/oliver/projectXYX/R

next to it we have

/home/oliver/projectXYX/data/munich/1998/test.Rdata

From the script I might call either (absolute path)

```r
load(file="/home/oliver/projectXYX/data/munich/1998/test.Rdata")
```

or (relative path)

```r
load(file="../data/munich/1998/test.Rdata")
```

The latter assumes that there is a file

../data/munich/1998/test.Rdata

to the script. But it does not assume that everything is in

/home/oliver/projectXYZ

Hence, the latter works even if my coauthor has stored everything as

C:/users/eva/PhD/projectXYX/R
C:/users/eva/PhD/projectXYX/data/munich/1998/test.Rdata

If a lot happens in ../data/munich/1998/ anyway, use the `setwd` command

```r
setwd("../data/munich/1998/")
...
load(file="test.Rdata")
```

(and remember to make the `setwd` relative, i.e. avoid the following:...
3.1.3 Robustness towards changes in context

Assume we have the following two files

```r
# script1.R
load("someData.Rdata")
# now two variables, x and y are defined
source("script2.R")
```

The content of `script2.R` might be this:

```r
# script2.R
est <- lm ( y ~ x)
```

In this example `script2.R` *assumes* that variables `y` and `x` are defined. As long as `script2.R` is called in this context, everything is fine.

Changing `script1.R` might have unexpected side effects since we transport variables from one script to the other. The call

```r
source("script2.R")
```

does not reveal how `y` and `x` are used by the script.

3.1.4 Functions increase robustness

```r
# script1.R
source("script2.R")
load("someData.Rdata")
myFunction(y,x)
```

```r
# script2.R
# defines myFunction
myFunction <- function(y,x) {
  est <<- lm ( y ~ x)
}
```

Now `script2.R` only defines a function. The function has arguments, hence, when we use it in `script1.R` we realise which variable goes where.

Note that the function takes *arguments*. This is more elegant (and less risky) than to write functions like this one:
and then say

```r
# script1.R
source("script2.R")
load("someData.Rdata")
x <- ...
y <- ...
myFunction()
```

It will still work, but later it will be less clear to us that the assignments before the function call are essential for the function.

```r
myFunction <- function(y,x) {
est <<- lm ( y ~ x)
}
```

This function has a *side effect*. It changes a variable `est` outside the function. Often it is less confusing to define functions with *return values* and no side effects.

```r
myFunction <- function(y,x) {
lm ( y ~ x)
}
```

When we call this function later as

```r
est <- myFunction(y,x)
```

it is clear where the result of the function goes.

**Recap**

- Functions which use global variables: risky
- Functions with side effects: risky
- Functions which only use arguments and return values: often better

**Note:** If we use *scripts* instead of *functions*:
- Scripts must use *global variables* and can only produce *side effects*.
- Scripts are more likely to lead to *mistakes* than functions.
- Replace scripts by *functions* (with arguments) whenever possible.
3.2 Calculations that take a lot of time

If a sequence of functions takes a lot of time to run, let it generate intermediate data. Our master-R-file could look like this:

```r
set.seed(123)
...
source("projectXYZ_init_160715.R")
getAndCleanData() # takes a lot of time
save(cleanedData, file="cleanedData160715.Rdata")
load("cleanedData160715.Rdata")
doBootstrap() # takes a lot of time
save(bsData, file="bsData160715.Rdata")
load("cleanedData160715.Rdata")
load("bsData160715.Rdata")
doFigures()
...
```

3.3 Nested functions

If our functions become long and complicated, we can divide them into small chunks.

```r
doAnalysis <- function () {
  firstStepAnalysis()
  secondStepAnalysis()
  thirdStepAnalysis()
  ...
}

firstStepAnalysis <- function () {
  ...
}

secondStepAnalysis <- function () {
  ...
}

Actually, if we need some functions only within a specific other function then we can define them within this function:

```r
doAnalysis <- function () {
  firstStep <- function () {
    ...
  }
  secondStep <- function () {
    ...
  }
  ...
}
• Advantage: These functions are only visible from within doAnalysis and can do no harm elsewhere (where we, perhaps, defined functions with the same name that do different things).

Nesting of functions has three advantages:

• It structures our work.
• It facilitates debugging.
• It facilitates communication with coauthors.

"...there is a problem in thirdStep in doAnalysis..."

### 3.4 Reproducible randomness

```r
set.seed(123)
```

Random numbers affect our results:

• Simulation
• MCMC samples
• Bootstrapping
• Approximate permutation tests
• Selection of training and confirmation samples
• ...

### 3.5 Recap — writing scripts and using functions

• If there is a systematic structure in our problem, then we can exploit it
• If we make mistakes, we make them systematically!
Compare

```r
t.test(profit88, data=myData)$p.value
t.test(profit89, data=myData)$p.value
t.test(profit98, data=myData)$p.value
```

with

```r
sapply(grep("profit", colnames(myData), value=TRUE), function(x) t.test(myData[,x])$p.value)
```

The first looks simpler.
The second is more robust against

- a change in the dataset (instead of myData we now use myDataClean)
- a change in the names of the variables (profit becomes Profit_)
- adding another profit-variable (profit2016...)
- typos (use profit88 twice, instead of profit88 and profit89 once each.

### 3.6 Human readable scripts

- Weaving and knitting → we do this later
- Comments at the beginning of each file

```r
# scriptExample160715.R
#
# the purpose of this script is to illustrate the use of
# comments
#
# first version: 160715
# this version: 160715
# last change by: Oliver
# requires: test160715.Rdata, someFunctions160715.R
# provides: ...
#
set.seed(123)
```

- Comments at the beginning of each function

```r
# # exampleFun transforms two vectors into an example
# # side effects: ...
# # returns: ...
# # exampleFun <- function(x,y) {
# ...
# }
```
Comment non-obvious steps

```r
# to detect outliers we use lrt-method.
# We have tried depth.trim and depth.pond
# but they produce implausible results...
outl <- foutliers(data, method="lrt")
```

Document your thoughts in your comments

```r
# 16/07/21: although I thought that age should not affect
# profits, it does here! I also checked
# xyz-specification and it still does.
# Perhaps age is a proxy for income.
# Unfortunately we do not have data on
# income here.
```

Formatting

Compare

```r
lm ( s1 ~ trust + ineq + sex + age + latitude )
lm ( otherinvestment ~ trust + ineq + sex + age + latitude )
```

with

```r
lm ( s1 ~ trust + ineq + sex + age + latitude )
lm ( otherinvestment ~ trust + ineq + sex + age + latitude )
```

Insert linebreaks

Compare

```r
lm ( otherinvestment ~ trust + ineq + sex + age + latitude, data=trustExp, subset=sex="female"
```

with

```r
lm ( otherinvestment ~ trust + ineq + sex + age + latitude, data=trustExp,
     subset=sex="female" )
```

Variables names

short but not too short

```r
lm ( otherinvestment ~ trustworthiness + inequalityaversion + sexOfProposer + ageOfProposer
lm ( otherinvestment ~ trust + ineq + sex + age + latitude)
lm ( o1 ~ t + i + s + a + 11 + 12)
lm ( R100234 ~ R100412 + R100017 + R100178 + R100671 + R100229 )
```
We will say more about variable names in section 6.3.

- Abbreviations in scripts
  
  R (and other languages too) allows you to refer to parameters in functions with names:

  ```r
  qnorm(p=.01, lower.tail=FALSE)
  [1] 2.326348
  ```

  To save space, you can abbreviate these names:

  ```r
  qnorm(p=.01, low=FALSE)
  [1] 2.326348
  ```

4 Some programming techniques

4.1 Debugging functions

```r
library(Ecdat)
data(Kakadu)
head(Kakadu)

      lower upper answer recparks jobs lowrisk wildlife future aboriginal finben
1 0 2 nn 3 1 5 5 1 1 1
mineparks moreparks gov envcon vparks tvenv conservation sex age schooling
1 4 5 1 yes yes 1 no male 27 3
income major
1 25 no
[ reached getOption("max.print") -- omitted 5 rows ]
```

general strategies: debug the function with a simple example

```r
sqMean <- function (x) {
  z <- mean(x)
  z^2
}
sqMean(Kakadu$lower)

[1] 2361.471
```

Is this correct? Take a (simpler) subsample of the data:
Assume that we still do not trust the function. `debug` allows us to debug a function. `ls` allows us to list the variables in the current environment.

```r
debug(sqMean)
```

```
sqMean(xx)
```

```
debugging in: sqMean(xx)
debug at <text>#1: {
  z <- mean(x)
  z^2
}
debug at <text>#2: z <- mean(x)
debug at <text>#3: z^2
exiting from: sqMean(xx)
[1] 1600
```

```r
undebug(sqMean)
```

If the function returns with an error, it helps to set

```
options(error=recover)
```

In the following function we refer to the variable `xxx` which is not defined. The function will, hence, fail. With `options(error=recover)` we can inspect the function at the time of the failure.

```r
sqMean <- function (x) {
  z <- mean(xxx)
  z^2
}
sqMean(xx)
```

```
Error in mean(xxx) (from #2) : object 'xxx' not found
Enter a frame number, or 0 to exit
1: sqMean(xx)
2: #2: mean(xxx)
```

Selection: 1
Called from: top level
4.2 Lists of variables

To make the analysis more consistent.

Whenever things repeat, we define them in variables at the top of the paper:

```r
models <- list(a="income",
               b="income + age + sex",
               c="income + age + sex + conservation + vparks")
```

(We use here character strings to represent parts of formulas. Alternatively, we could also store objects of class `formula`. However, manipulating these objects is not always to obvious. To keep things simple, we will use character strings here.) Later in the paper we compare the different models:

```r
mylm <- function (m) lm(paste("as.integer(answer) ~ ",m),data=Kakadu)
lmList<-lapply(models,mylm)
class(lmList)<-c("list","by")
mtable(lmList)
```
Now we use the same explanatory variables to explain a different dependent variable:

```r
mylogit <- function(m) glm(paste("answer=='yy' ~ ",m),
                         data=Kakadu,family=binomial(link=logit))
logitList <- lapply(models,mylogit)
class(logitList)<-c("list","by")
mtable(logitList)
```
Similarly, we might define at the beginning of the paper...

- lists of random effects
- lists of variables to group by
- palettes for plots

### 4.3 Return values of functions

Most functions do not only return a number (or a vector) but rather complex objects. In R `str()` helps us to learn more about the structure of these objects. (In Stata similar return values are provided by `return`, `ereturn`, and `sreturn`)

```r
lm1 <- mylm (models[[1]])
str(lm1)
```

List of 12

$ coefficients : Named num [1:2] 2.12202 0.00278
  ..- attr(*, "names")= chr [1:2] "(Intercept)" "income"
$ residuals : Named num [1:1827] -1.19 -1.15 -1.19 -1.19 -1.22 ...
   ..- attr(*, "names")= chr [1:1827] "1" "2" "3" "4" ...
$ effects : Named num [1:1827] -93.28 1.95 -1.17 -1.17 -1.21 ...
   ..- attr(*, "names")= chr [1:1827] "(Intercept)" "income" "" "" ...
$ rank : int 2
$ fitted.values: Named num [1:1827] 2.19 2.15 2.19 2.19 2.22 ...
   ..- attr(*, "names")= chr [1:1827] "1" "2" "3" "4" ...
$ assign : int [1:2] 0 1
$ qr :List of 5
   ..$ qr : num [1:1827, 1:2] -42.7434 0.0234 0.0234 0.0234 0.0234 ...
   ..   ..- attr(*, "dimnames")=List of 2
   ..   ..$ : chr [1:1827] "1" "2" "3" "4" ...
   ..   ..$ : chr [1:2] "(Intercept)" "income"
   ..   ..- attr(*, "assign")= int [1:2] 0 1
   ..$ qr: List of 2
   ..$ tol : num 0.0000001
   ..$ rank : int 2
   ..- attr(*, "class")= chr "qr"
$ df.residual : int 1825
$ xlevels : Named list()
$ call : language lm(formula = paste("as.integer(answer) ~ ", m), data = Kakadu)
$ terms :Classes 'terms', 'formula' language as.integer(answer) ~ income
   ..- attr(*, "variables")= language list(as.integer(answer), income)
   ..- attr(*, "factors")= int [1:2, 1] 0 1
   ..- attr(*, "dimnames")=List of 2
   ..$ : chr [1:2] "as.integer(answer)" "income"
   ..$ : chr [1:2] "(Intercept)"
   ..- attr(*, "term.labels")= chr "income"
   ..- attr(*, "order")= int 1
   ..- attr(*, "intercept")= int 1
   ..- attr(*, "response")= int 1
   ..- attr(*, "Environment")=<environment: 0x6c0fc80>
   ..- attr(*, "predvars")= language list(as.integer(answer), income)
   ..- attr(*, "dataClasses")= Named chr [1:2] "numeric" "numeric"
   ..- attr(*, "names")= chr [1:2] "as.integer(answer)" "income"
$ model : 'data.frame': 1827 obs. of 2 variables:
   ..$ as.integer(answer): int [1:1827] 1 1 1 1 1 1 1 1 1 1 ...
   ..- attr(*, "terms")=Classes 'terms', 'formula' language as.integer(answer) ~ income
   ..- attr(*, "variables")= language list(as.integer(answer), income)
   ..- attr(*, "factors")= int [1:2, 1] 0 1
   ..- attr(*, "dimnames")=List of 2
   ..$ : chr [1:2] "as.integer(answer)" "income"
   ..$ : chr "income"
   ..- attr(*, "term.labels")= chr "income"
   ..- attr(*, "order")= int 1
   ..- attr(*, "intercept")= int 1
   ..- attr(*, "response")= int 1
   ..- attr(*, "Environment")=<environment: 0x6c0fc80>
   ..- attr(*, "predvars")= language list(as.integer(answer), income)
   ..- attr(*, "dataClasses")= Named chr [1:2] "numeric" "numeric"
There are at least two ways to extract data from these objects:

- Extractor functions

```r
coef(lm1)
```

```
(Intercept)    income
2.122018102  0.002781938
```

```r
cov(lm1)
```

```
  (Intercept)     income
(Intercept) 0.00121806075 -0.000035685812
income -0.00003568581  0.000001647787
```

```r
hccm(lm1)
```

```
  (Intercept)     income
(Intercept)  0.00123366056 -0.000036812592
income -0.00003681259  0.000001719666
```

```r
logLik(lm1)
```

```
'log Lik.' -2402.751 (df=3)
```

```r
effects(lm1)
fitted.values(lm1)
residuals(lm1)
```

(the equivalent in Stata are postestimation commands)

- Whatever is a list item can also be accessed directly:

```r
lm1$coefficients
lm1$residuals
lm1$fitted.values
```

Note: Some interesting values are not provided by the `lm`-object itself. These can often be accessed as part of the `summary`-object.

```r
slm1 <- summary(lm1)
slm1$r.squared
slm1$adj.r.squared
slm1$fstatistic
```
4.4 Repeating things

Looping  The simplest way to repeat a command is a loop:

```r
for (i in 1:10) print(i)
```

```
[1] 1
[1] 2
[1] 3
[1] 4
[1] 5
[1] 6
[1] 7
[1] 8
[1] 9
[1] 10
```

If the command is a sequence of expressions, we have to enclose it in braces.

```r
for (i in 1:10) {
  x <- runif(i)
  print(mean(x))
}
```

```
[1] 0.3565607
[1] 0.9663778
[1] 0.5063639
[1] 0.4378409
[1] 0.487012
[1] 0.5853594
[1] 0.3502112
[1] 0.499148
[1] 0.5078825
[1] 0.4557163
```

Avoiding loops  In R loops should be avoided. It is more efficient (faster) to apply a function to a vector.

```r
sapply(1:10,print)
```

```
[1] 1 2 3 4 5 6 7 8 9 10
```
Or, the more complex example:

```r
sapply(1:10, function(i) {
  x <- runif(i)
  mean(x)
})
```

```
[1] 0.6538133 0.4623162 0.8092458 0.4935831 0.6997635 0.4856793 0.6413399
[8] 0.5610393 0.5781580 0.4712342
```

Note that `sapply` already returns a vector which is in many cases what we want anyway.

In the above examples we applied a function to a vector. Sometimes we want to apply functions to a matrix.

### Applying a function along one dimension of a matrix

In the following example we apply a function along the second dimension of the dataset `Kakadu`.

```r
apply(Kakadu, 2, function(x) mean(as.integer(x)))
```

<table>
<thead>
<tr>
<th>lower</th>
<th>upper</th>
<th>answer</th>
<th>recparks</th>
<th>jobs</th>
<th>lowrisk</th>
</tr>
</thead>
<tbody>
<tr>
<td>48.594964</td>
<td>536.714286</td>
<td>NA</td>
<td>3.688560</td>
<td>2.592228</td>
<td>2.790367</td>
</tr>
<tr>
<td>wildlife</td>
<td>future</td>
<td>aboriginal</td>
<td>finben</td>
<td>mineparks</td>
<td>moreparks</td>
</tr>
<tr>
<td>4.739464</td>
<td>4.466886</td>
<td>3.569787</td>
<td>2.915709</td>
<td>3.643678</td>
<td>3.864806</td>
</tr>
<tr>
<td>gov</td>
<td>envcon</td>
<td>vparks</td>
<td>tvenv</td>
<td>conservation</td>
<td>sex</td>
</tr>
<tr>
<td>1.083196</td>
<td>NA</td>
<td>NA</td>
<td>1.785441</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>age</td>
<td>schooling</td>
<td>income</td>
<td>major</td>
<td></td>
<td></td>
</tr>
<tr>
<td>42.968254</td>
<td>3.683634</td>
<td>21.656814</td>
<td>NA</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

```r
cbind(apply(Kakadu, 2, function(x) mean(as.integer(x))))
```

```
[,1]
lower 48.594964
upper 536.714286
answer NA
repcarks 3.688560
jobs 2.592228
lowrisk 2.790367
wildlife 4.739464
future 4.466886
aboriginal 3.569787
finben 2.915709
mineparks 3.643678
moreparks 3.864806
gov 1.083196
eenvcon NA
varks NA
tvenv 1.785441
conservation NA
sex NA
```
Rectangular and ragged arrays

Rectangular array:

<table>
<thead>
<tr>
<th>wide</th>
<th>long</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>hor</td>
</tr>
<tr>
<td></td>
<td>a</td>
</tr>
<tr>
<td></td>
<td>b</td>
</tr>
<tr>
<td></td>
<td>c</td>
</tr>
<tr>
<td>A</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>4</td>
</tr>
</tbody>
</table>

Ragged array:

<table>
<thead>
<tr>
<th>wide</th>
<th>long</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>hor</td>
</tr>
<tr>
<td></td>
<td>b</td>
</tr>
<tr>
<td></td>
<td>c</td>
</tr>
<tr>
<td></td>
<td>a</td>
</tr>
<tr>
<td></td>
<td>b</td>
</tr>
<tr>
<td>A</td>
<td>2</td>
</tr>
<tr>
<td>B</td>
<td>4</td>
</tr>
</tbody>
</table>

Applying a function to each element of a ragged array

In R ragged arrays can be represented as datasets grouped by one or more factors. These variables describe which records belong together (e.g. to the same person, year, firm, …)

In the following example we use the dataset Fatality. This dataset contains for each state of the United States and for each year in 1982 to 1988 in `mrall` the traffic fatality rate (deaths per 10000).

```r
data(Fatality)
head(Fatality)
```

```r
state year mrall beertax mlda jaild comserd vmiles unrate perinc
1 1 1982 2.12836 1.539379 19.00 no no 7.233887 14.4 10544.15
2 1 1983 2.34848 1.788991 19.00 no no 7.836348 13.7 10732.80
3 1 1984 2.33643 1.714286 19.00 no no 8.262990 11.1 11108.79
4 1 1985 2.19348 1.652542 19.67 no no 8.726917 8.9 11332.63
```

[ reached getOption("max.print") -- omitted 2 rows ]

```r
by(Fatality,list(Fatality$year),function(x) mean(x$mrall))
```

: 1982

[1] 2.089106
by(Fatality,list(Fatality$state),function(x) mean(x$mrall))

: 1
[1] 2.412627

: 4
[1] 2.7059

: 5
[1] 2.435336

: 6
[1] 1.904977

: 8
[1] 1.866981

: 9
[1] 1.463509

: 10
[1] 2.068231

: 12
[1] 2.477799

: 13
[1] 2.401569

: 16
[1] 2.571667
<table>
<thead>
<tr>
<th></th>
<th>1.319227</th>
</tr>
</thead>
<tbody>
<tr>
<td>35</td>
<td>[1] 3.653197</td>
</tr>
<tr>
<td></td>
<td>[1] 1.207581</td>
</tr>
<tr>
<td>37</td>
<td>[1] 2.34471</td>
</tr>
<tr>
<td></td>
<td>[1] 1.601454</td>
</tr>
<tr>
<td>39</td>
<td>[1] 1.550474</td>
</tr>
<tr>
<td></td>
<td>[1] 2.177147</td>
</tr>
<tr>
<td>41</td>
<td>[1] 1.541673</td>
</tr>
<tr>
<td></td>
<td>[1] 1.110077</td>
</tr>
<tr>
<td>45</td>
<td>[1] 2.821669</td>
</tr>
<tr>
<td></td>
<td>[1] 2.04929</td>
</tr>
<tr>
<td>47</td>
<td>[1] 2.403066</td>
</tr>
<tr>
<td></td>
<td>[1] 2.27587</td>
</tr>
<tr>
<td>49</td>
<td>[1] 1.835836</td>
</tr>
<tr>
<td></td>
<td>[1] 2.092991</td>
</tr>
<tr>
<td>51</td>
<td>[1] 1.740946</td>
</tr>
<tr>
<td></td>
<td>[1] 1.677211</td>
</tr>
</tbody>
</table>
by does not return a vector but an object of class by. If we actually need a vector we have to use `c` and `sapply`.

In the following example we let by actually return two values.

```r
byObj <- by(Fatality, list(Fatality$year),
           function(x) c(year=median(x$year),
                         fatality=mean(x$mrall),
                         meanbeertax=mean(x$beertax)))
sapply(byObj, c)
```

```
fatality 2.0891059 2.007846 2.0171225 1.9736708 2.0650710
meanbeertax 0.5302734 0.532393 0.5295902 0.5169272 0.5086639

1987 1988
year 1987.0000000 1988.0000000
fatality 2.0606956 2.0695941
meanbeertax 0.4951288 0.4798154
```

```r
xx <- data.frame(t(sapply(byObj, c)))
xyplot(fatality ~ meanbeertax, type="l", data=xx) +
       layer(with(xx, panel.text(label=year, y=fatality, x=meanbeertax, adj=c(1,1))))
```

We can do more complicated things in by. In the following example we calculate
a regression. To get only the coefficients from the regression (and not fitted values, residuals, etc.) we use the extractor function `coef`.

```r
byObj <- by(Fatality,list(Fatality$year),function(x)
    lm(mrall ~ beertax + jaild, data=x))
sapply(byObj,coef)
```

```
(Intercept) 1.9079924 1.7503870 1.6768093 1.6567128 1.7108657 1.7188081
beertax 0.1824028 0.2991742 0.4066922 0.4057889 0.4944595 0.4920275
jaildyes 0.4500807 0.3625151 0.4283417 0.3430232 0.3286131 0.3369277
```

```
1988
(Intercept) 1.7411593
beertax 0.4509099
jaildyes 0.3842788
```

```r
xx<-data.frame(t(sapply(byObj,coef)))
xyplot(beertax ~ jaildyes,type="l",data=xx)+
   layer(with(xx,panel.text(label=rownames(xx),y=beertax,x=jaildyes,adj=c(1,1))))
```

```
by is very complex. It offers the entire subset of the dataframe, as defined by the index variable, to the function.
Sometimes we want simply to apply a function of only a vector along a ragged array.
```

```r
with(Fatality,aggregate(mrall~year,FUN=mean))
```

```
    year  mra ll
   1 1982 2.089106
   2 1983 2.007846
   3 1984 2.017122
   4 1985 1.973671
   5 1986 2.065071
```
Again, the function (which was \texttt{mean} in the previous example) can be defined by us:

```r
with(Fatality, aggregate(mrall~year, FUN=function(x) sd(x)/mean(x)))
```

<table>
<thead>
<tr>
<th>year</th>
<th>mrall</th>
</tr>
</thead>
<tbody>
<tr>
<td>1982</td>
<td>0.3196449</td>
</tr>
<tr>
<td>1983</td>
<td>0.3017002</td>
</tr>
<tr>
<td>1984</td>
<td>0.2721300</td>
</tr>
<tr>
<td>1985</td>
<td>0.2726437</td>
</tr>
<tr>
<td>1986</td>
<td>0.2709500</td>
</tr>
<tr>
<td>1987</td>
<td>0.2738153</td>
</tr>
<tr>
<td>1988</td>
<td>0.2518286</td>
</tr>
</tbody>
</table>

## 5 Data manipulation

### 5.1 Subsetting data

There are several ways to access only a part of a dataset:

- Many functions take an option `...`, \texttt{subset=}...

```r
lm(mrall ~ beertax + jaild, data=Fatality, subset = year == 1982)
```

Call:

```r
lm(formula = mrall ~ beertax + jaild, data = Fatality, subset = year == 1982)
```

Coefficients:

<table>
<thead>
<tr>
<th>(Intercept)</th>
<th>beertax</th>
<th>jaildyes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.9080</td>
<td>0.1824</td>
<td>0.4501</td>
</tr>
</tbody>
</table>

- The \texttt{subset()} function

```r
subset(Fatality, year == 1982 )
```

<table>
<thead>
<tr>
<th>state</th>
<th>year</th>
<th>mrall</th>
<th>beertax</th>
<th>mlda</th>
<th>jaild</th>
<th>comserd</th>
<th>vmiles</th>
<th>unrate</th>
<th>perinc</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1982</td>
<td>2.12836</td>
<td>1.539379</td>
<td>48</td>
<td>no</td>
<td>no</td>
<td>4.8</td>
<td>14.4</td>
<td>10544.152</td>
</tr>
<tr>
<td>8</td>
<td>1982</td>
<td>2.49914</td>
<td>0.214797</td>
<td>14.0</td>
<td>yes</td>
<td>yes</td>
<td>6.8</td>
<td>9.9</td>
<td>12309.069</td>
</tr>
<tr>
<td>15</td>
<td>1982</td>
<td>2.38405</td>
<td>0.650358</td>
<td>21.0</td>
<td>no</td>
<td>no</td>
<td>7.2</td>
<td>9.8</td>
<td>10267.303</td>
</tr>
<tr>
<td>22</td>
<td>1982</td>
<td>1.86194</td>
<td>0.107398</td>
<td>57</td>
<td>no</td>
<td>no</td>
<td>6.8</td>
<td>9.9</td>
<td>15797.136</td>
</tr>
</tbody>
</table>

[ reached getOption("max.print") -- omitted 44 rows ]

```r
with(subset(Fatality, year == 1982 ), lm(mrall ~ beertax + jaild))
```
Call:
`lm(formula = mrall ~ beertax + jaild)`

Coefficients:

<table>
<thead>
<tr>
<th></th>
<th>(Intercept)</th>
<th>beertax</th>
<th>jaildyes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimate</td>
<td>1.9080</td>
<td>0.1824</td>
<td>0.4501</td>
</tr>
</tbody>
</table>

- The first index of the dataset

```R
Fatality[ Fatality$year==1982 , ]
```

```R
<table>
<thead>
<tr>
<th>state</th>
<th>year</th>
<th>mrall</th>
<th>beertax</th>
<th>mlda</th>
<th>jaild</th>
<th>comserd</th>
<th>vmiles</th>
<th>unrate</th>
<th>perinc</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1982</td>
<td>2.12836</td>
<td>1.53937948</td>
<td>19.0</td>
<td>no</td>
<td>no</td>
<td>7.233887</td>
<td>14.4</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
<td>1982</td>
<td>2.49914</td>
<td>0.21479714</td>
<td>19.0</td>
<td>yes</td>
<td>yes</td>
<td>6.810157</td>
<td>9.9</td>
</tr>
<tr>
<td>15</td>
<td>5</td>
<td>1982</td>
<td>2.38405</td>
<td>0.65035802</td>
<td>21.0</td>
<td>no</td>
<td>no</td>
<td>7.208500</td>
<td>9.8</td>
</tr>
<tr>
<td>22</td>
<td>6</td>
<td>1982</td>
<td>1.86194</td>
<td>0.10739857</td>
<td>21.0</td>
<td>no</td>
<td>no</td>
<td>6.858677</td>
<td>9.9</td>
</tr>
</tbody>
</table>
```

[reached getOption("max.print") -- omitted 44 rows]

```R
with(Fatality[ Fatality$year==1982 , ],lm(mrall ~ beertax + jaild))
```

Call:
`lm(formula = mrall ~ beertax + jaild)`

Coefficients:

<table>
<thead>
<tr>
<th></th>
<th>(Intercept)</th>
<th>beertax</th>
<th>jaildyes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimate</td>
<td>1.9080</td>
<td>0.1824</td>
<td>0.4501</td>
</tr>
</tbody>
</table>

### 5.2 Merging data

- Appending two datasets

```R
library(plyr)
rbind.fill(x,y)
```

(In Stata this is done by `append`)

- Matching two datasets (inner join)

```R
merge(x,y)
```

(In Stata this is done by `merge`)

- Joining two datasets (left join)

```R
merge(x,y,all.x=TRUE)
```

(In Stata this is done by `joinby`)
Dataset A
Name  Grade
Eva   2.0  
Mary  1.0  
Mike  3.0  

Dataset B
Name    eMail
Eva    eva@... 
Eva    eva2@... 
Susan  susan@... 
Mike   mike@... 

Inner join: merge(A,B)
Name  Grade  eMail
Eva   2.0    eva@... 
Eva   2.0    eva2@... 
Mike  3.0    mike@... 

Left join: merge(A,B,all.x=TRUE)
Name  Grade  eMail
Eva   2.0    eva@... 
Eva   2.0    eva2@... 
Mary  1.0    NA 
Mike  3.0    mike@... 

Appending
In the following example we first split the data from an experiment into two parts. Merge helps us to append them to each other.

load("data/160716_060x.Rdata")
experiment1 <- subset(trustGS$subjects,Date="160716_0601")
experiment2 <- subset(trustGS$subjects,Date="160716_0602")
dim(experiment1)
[1] 108 14
dim(experiment2)
[1] 108 14
library(plyr)
dim(rbind.fill(experiment1,experiment2))
[1] 216 14

Joining
A frequent application for a join are tables in z-Tree that have something to do with each other. E.g. the globals and the subjects tables both provide information about each period. Common variables in these tables are Date, Treatment, and Period.

By merging globals with subjects, merge looks up for each record in the subjects table the matching record in the globals table and adds the variables which are not already present in subjects.

head(trustGS$global)

<table>
<thead>
<tr>
<th>Date</th>
<th>Treatment</th>
<th>Period</th>
<th>NumPeriods</th>
<th>RepeatTreatment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 160716_0601</td>
<td>1</td>
<td>1</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>2 160716_0601</td>
<td>1</td>
<td>2</td>
<td>6</td>
<td>0</td>
</tr>
</tbody>
</table>
In the following example we simply get two more variables in the dataset (NumPeriods and RepeatTreatment). With more variables in globals we would, of course, also get more variables in the merged dataset.

**Joining aggregates** A common application for a join is a comparison of our individual data with aggregated data. Let us come back to the Fatalities example. We want to compare the traffic fatality rate \( mrall \) for each state with the average values for each year.

```r
head(Fatality)
```

<table>
<thead>
<tr>
<th>state</th>
<th>year</th>
<th>mrall</th>
<th>beertax</th>
<th>mlda</th>
<th>jaild</th>
<th>comserd</th>
<th>vmiles</th>
<th>unrate</th>
<th>perinc</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1982</td>
<td>2.12836</td>
<td>1.539379</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1983</td>
<td>2.34848</td>
<td>1.788991</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1984</td>
<td>2.33643</td>
<td>1.714286</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1985</td>
<td>2.19348</td>
<td>1.652542</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

[ reached getOption("max.print") -- omitted 2 rows ]

```r
Aggregate(c(avgMrall=mean(mrall)) ~ year,data=Fatality)
```

```r
year avgMrall
1 1982 2.089106
```
merge(Fatality, Aggregate(c(avgMrall=mean(mrall)) ~ year, data=Fatality))

merge has joined the two datasets, the large Fatality one, and the small aggregated one, on the variable `year`.

### 5.3 Reshaping data

Sometimes we have different observations of the same (or similar) variable in the same row (e.g. `profit.1` and `profit.2`), sometimes we have them stacked in one column (e.g. as `profit`). We call the first format `wide`, the second `long`.

For the `long` case we need a variable that distinguishes the different instances of this variable (`profit.1` and `profit.2`) from each other. Such a variable is called `timevar` (Stata call them `j`).

We also need one or more variables that tells us, which observations actually belonged to one row in the `wide` format. We call these variables `idvar` (Stata call this variable `i`).

Let us look at a part of our trust dataset

```r
trustLong <- trustGS$subjects[,c("Date","Period","Subject","Pos","Group","Offer")]
trustLong[1:4,]
Date Period Subject Pos Group Offer
1 160716_0601 1 1 2 1 0.000
2 160716_0601 1 2 2 4 0.000
3 160716_0601 1 3 1 5 0.495
4 160716_0601 1 4 2 2 0.000
```

```r
trustWide <- reshape(trustLong, v.names=c("Offer","Subject"), idvar=c("Date","Period","Group"), timevar="Pos", direction="wide")
trustWide[1:4,]
```

```r
merge(Fatality,Aggregate(c(avgMrall=mean(mrall)) ~ year, data=Fatality))
```
Reshaping back returns more or less the original data. The ordering has changed and rows have got names now.

Reshaping with `reshape2` `recast` does not give us Subject, though.

### 6 Preparing Data

- read data
- check structure (names, dimension, labels)
- check values
- create new data:
  - recode variables
  - rename variables
  - label variables
– eliminate outliers
– reshape data

6.1 Reading data
6.1.1 Reading z-Tree Output

The function `zTreeTables(...vector of filenames...[,vector of tables])` reads zTree .xls files and returns a list of tables. Here we use `list.files` to find all files that match the typical z-Tree pattern. If we ever get more experiments our command will find them and use them.

```r
setwd("rawdata/Trust/")
files <- list.files(pattern = "\[0-9\]{6}_\[0-9\]{4}.xls\$",recursive=TRUE)

[1] "160716_0601.xls"   "160716_0602.xls"
[3] "160716_0603.xls"   "160716_0604.xls"
[5] "rawdata/PublicGood/090622_0601.xls" "rawdata/PublicGood/090622_0602.xls"
[7] "rawdata/PublicGood/090622_0603.xls" "rawdata/PublicGood/090622_0604.xls"
[9] "rawdata/PublicGood/130616_0601.xls" "rawdata/PublicGood/130616_0602.xls"
[11] "rawdata/PublicGood/130616_0603.xls" "rawdata/PublicGood/130616_0604.xls"
[13] "rawdata/Trust/130716_0601.xls"  "rawdata/Trust/130716_0602.xls"
[15] "rawdata/Trust/130716_0603.xls"  "rawdata/Trust/130716_0604.xls"

trustGS <- zTreeTables(files)
reading 160716_0601.xls ...
Skipping:
Doing: globals
Doing: subjects
*** 160716_0602.xls is file 2 / 16 ***
reading 160716_0602.xls ...
Skipping:
Doing: globals
Doing: subjects
*** 160716_0603.xls is file 3 / 16 ***
reading 160716_0603.xls ...
Skipping:
Doing: globals
Doing: subjects
*** 160716_0604.xls is file 4 / 16 ***
reading 160716_0604.xls ...
Skipping:
Doing: globals
Doing: subjects
*** rawdata/PublicGood/090622_0601.xls is file 5 / 16 ***
reading rawdata/PublicGood/090622_0601.xls ...
Skipping:
Doing: globals
Doing: subjects
*** rawdata/PublicGood/090622_0602.xls is file 6 / 16 ***
reading rawdata/PublicGood/090622_0602.xls ...
Skipping:
Doing: globals
Doing: subjects
*** rawdata/PublicGood/090622_0603.xls is file 7 / 16 ***
reading rawdata/PublicGood/090622_0603.xls ...
Skipping:
Doing: globals
Doing: subjects
*** rawdata/PublicGood/090622_0604.xls is file 8 / 16 ***
reading rawdata/PublicGood/090622_0604.xls ...
Skipping:
Doing: globals
Doing: subjects
*** rawdata/PublicGood/130616_0601.xls is file 9 / 16 ***
reading rawdata/PublicGood/130616_0601.xls ...
Skipping:
Doing: globals
Doing: subjects
*** rawdata/PublicGood/130616_0602.xls is file 10 / 16 ***
reading rawdata/PublicGood/130616_0602.xls ...
Skipping:
Doing: globals
Doing: subjects
*** rawdata/PublicGood/130616_0603.xls is file 11 / 16 ***
reading rawdata/PublicGood/130616_0603.xls ...
Skipping:
Doing: globals
Doing: subjects
*** rawdata/PublicGood/130616_0604.xls is file 12 / 16 ***
reading rawdata/PublicGood/130616_0604.xls ...
Skipping:
Doing: globals
Doing: subjects
*** rawdata/Trust/130716_0601.xls is file 13 / 16 ***
reading rawdata/Trust/130716_0601.xls ...
Skipping:
Doing: globals
Doing: subjects
*** rawdata/Trust/130716_0602.xls is file 14 / 16 ***
reading rawdata/Trust/130716_0602.xls ...
Skipping:
Doing: globals
Doing: subjects
*** rawdata/Trust/130716_0603.xls is file 15 / 16 ***
reading rawdata/Trust/130716_0603.xls ...
Skipping:
Doing: globals
Doing: subjects
6.1.2 Reading and writing R-Files

If we want to save one or more R objects in a file, we use `save`:

```r
save(trustGS,zTreeTables,file="data/160716_060x.Rdata")
```

To retrieve them, we use `load`.
Advantages:

- Rdata is very compact, files are small
- All attributes are saved together with the data
- We can save functions together with data

6.1.3 Reading Stata Files

<table>
<thead>
<tr>
<th>package</th>
<th>command</th>
<th>limitation</th>
</tr>
</thead>
<tbody>
<tr>
<td>foreign</td>
<td>read.dta</td>
<td>Stata version 5-12</td>
</tr>
<tr>
<td></td>
<td>write.dta</td>
<td>Stata version 5-12</td>
</tr>
<tr>
<td>memisc</td>
<td>Stata.file</td>
<td>Stata version 5-12</td>
</tr>
<tr>
<td></td>
<td>readstata13</td>
<td>read.dta13</td>
</tr>
</tbody>
</table>

library(foreign)
sta <- read.dta("data/160716_060x.dta")

sta2 <- Stata.file("data/160716_060x.dta")

foreign (read.dta stores internal Stata information as attributes of the data frame. memisc (Stata.file) stores internal Stata information as attributes of the variable.

str(sta)

'data.frame': 3936 obs. of 19 variables:
$ Date     : chr  "090622_0601"  "090622_0601"  "090622_0601"  "090622_0601"  ...
$ Treatment: num  1  1  1  1  1  1  1  1  1  1  ...
$ Period   : num  1  1  1  1  1  1  1  1  1  1  ...
$ NumPeriods: num  12 12 12 12 12 12 12 12 12 12  ...
$ RepeatTreatment: num  0  0  0  0  0  0  0  0  0  0  ...
$ Subject  : num  1  2  3  4  5  6  7  8  9  10  ...
$ Pos      : num  3  1  3  3  2  3  4  2  4  1  ...
$ Group    : num  4  4  2  1  3  3  4  1  3  1  ...
$ Offer    : num  NA NA NA NA NA NA NA NA NA NA  ...
$ Receive  : num  NA NA NA NA NA NA NA NA NA NA  ...
$ Return   : num  NA NA NA NA NA NA NA NA NA NA  ...
$ GetBack  : num  NA NA NA NA NA NA NA NA NA NA  ...
$ country  : num  NA NA NA NA NA NA NA NA NA NA  ...
$ siblings : num  NA NA NA NA NA NA NA NA NA NA  ...
$ sex      : num  1  2  1  2  1  1  2  1  1  2  ...
$ age       : num  24 21 27 19 29 33 25 98 20 99  ...
$ Contrib1 : num  0.528 0.721 0.621 0.6 0.48 0.611 0.513 0.336 0.594 0.854  ...
$ Contrib2 : num  0.691 0.542 0.5 0.465 0.694 0.481 0.572 0.619 1 0.54  ...
The data frame created by Stata file looks different:

```r
str(sta2)
```

Formal class 'Stata.importer' [package "memisc"] with 5 slots
  .Data :List of 19
    ..$ : Nmnl. item chr(0)
    ..$ : Itvl. item num(0)
    ..$ : Itvl. item num(0)
    ..$ : Itvl. item num(0)
    ..$ : Itvl. item num(0)
    ..$ : Itvl. item num(0)
    ..$ : Itvl. item num(0)
    ..$ : Itvl. item num(0)
    ..$ : Itvl. item num(0)
    ..$ : Itvl. item num(0)
    ..$ : Itvl. item num(0)
    ..$ : Itvl. item num(0)
    ..$ : Itvl. item num(0)
    ..$ : Itvl. item num(0)
    ..$ : Itvl. item num(0)
    ..$ : Itvl. item num(0)
  data.spec:List of 8
    ..$ names : chr [1:19] "Date" "Treatment" "Period" "NumPeriods" ...
    ..$ types : raw [1:19] 0b ff ff ff ...
    ..$ nobs : int 3936
    ..$ nvar : int 19
    ..$ varlabs : Named chr [1:19] "Date" "Treatment" "Period" "NumPeriods" ...
    ..$ value.labels : Named chr(0)
    ..$ missing.values: NULL
    ..$ version.string: chr "Stata 7"
  @ ptr :<externalptr>
  @ document : chr("data/160716_060x.dta")
  @ names : chr [1:19] "Date" "Treatment" "Period" "NumPeriods" ...
```

Also the attributes are different:
attributes(sta)

$datalabel
[1] "Written by R."

$time.stamp
[1] ""

$names
[1] "Date" "Treatment" "Period" "NumPeriods"
[5] "RepeatTreatment" "Subject" "Pos" "Group"
[9] "Offer" "Receive" "Return" "GetBack"
[13] "country" "siblings" "sex" "age"
[17] "Contrib1" "Contrib2" "Contrib3"

$formats
[1] "%11s" "%9.0g" "%9.0g" "%9.0g" "%9.0g" "%9.0g" "%9.0g" "%9.0g" "%9.0g"
[10] "%9.0g" "%9.0g" "%9.0g" "%9.0g" "%9.0g" "%9.0g" "%9.0g" "%9.0g" "%9.0g"
[19] "%9.0g"

$types
[1] 138 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100

$val.labels
[1] "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" ""

$var.labels
[1] "Date" "Treatment" "Period" "NumPeriods"
[5] "RepeatTreatment" "Subject" "Pos" "Group"
[9] "Offer" "Receive" "Return" "GetBack"
[13] "country" "siblings" "sex" "age"
[17] "Contrib1" "Contrib2" "Contrib3"

$row.names
[1] "1" "2" "3" "4" "5" "6" "7" "8" "9" "10" "11" "12" "13" "14" "15"
[16] "16" "17" "18" "19" "20" "21" "22" "23" "24" "25" "26" "27" "28" "29" "30"
[31] "31" "32" "33" "34" "35" "36" "37" "38" "39" "40"

[ reached getOption("max.print") -- omitted 3896 entries ]

$version
[1] 7

$class
[1] "data.frame"

Stata.file stores variable labels as attributes of the variables:

attributes(sta2)

$ptr
<pointer: 0x9c1e090>
attr(,"file.name")
$document
class character(0)

$names
[1] "Date"   "Treatment"   "Period"   "NumPeriods"
[5] "RepeatTreatment" "Subject"   "Pos"     "Group"
[9] "Offer"    "Receive"   "Return"   "GetBack"
[13] "country"  "siblings"  "sex"     "age"
[17] "Contrib1" "Contrib2"  "Contrib3"

$data.spec
$data.spec$names
[1] "Date"   "Treatment"   "Period"   "NumPeriods"
[5] "RepeatTreatment" "Subject"   "Pos"     "Group"
[9] "Offer"    "Receive"   "Return"   "GetBack"
[13] "country"  "siblings"  "sex"     "age"
[17] "Contrib1" "Contrib2"  "Contrib3"

$data.spec$types
[1] 0b ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff ff

$data.spec$nobs
[1] 3936

$data.spec$nvar
[1] 19

$data.spec$varlabs

<table>
<thead>
<tr>
<th></th>
<th>Treatment</th>
<th>Period</th>
<th>NumPeriods</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Date</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>RepeatTreatment</td>
<td>Subject</td>
<td>Pos</td>
</tr>
<tr>
<td></td>
<td>Offer</td>
<td>Receive</td>
<td>Return</td>
</tr>
<tr>
<td></td>
<td>country</td>
<td>siblings</td>
<td>sex</td>
</tr>
<tr>
<td></td>
<td>Contrib1</td>
<td>Contrib2</td>
<td>Contrib3</td>
</tr>
</tbody>
</table>

$data.spec$value.labels
named character(0)

$data.spec$missing.values
NULL

$data.spec$version.string
[1] "Stata 7"

$class
Within the `memisc` world you can obtain more information with `codebook`.

```r
codebook(sta2)
```

<table>
<thead>
<tr>
<th>Variable</th>
<th>Type</th>
<th>Measurement</th>
<th>Mean</th>
<th>Variance</th>
<th>Skewness</th>
<th>Kurtosis</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date</td>
<td>character</td>
<td>nominal</td>
<td>1.000</td>
<td>0.000</td>
<td>NaN</td>
<td>NaN</td>
<td>090622_0601</td>
<td>160716_0604</td>
</tr>
<tr>
<td>Treatment</td>
<td>double</td>
<td>interval</td>
<td>1.000</td>
<td>0.000</td>
<td>NaN</td>
<td>NaN</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>Period</td>
<td>double</td>
<td>interval</td>
<td>5.841</td>
<td>11.483</td>
<td>0.290</td>
<td>-1.082</td>
<td>1.000</td>
<td>12.000</td>
</tr>
</tbody>
</table>
### NumPeriods 'NumPeriods'

<table>
<thead>
<tr>
<th>Storage mode: double</th>
<th>Measurement: interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean: 10.683</td>
<td></td>
</tr>
<tr>
<td>Variance: 6.168</td>
<td></td>
</tr>
<tr>
<td>Skewness: -1.355</td>
<td></td>
</tr>
<tr>
<td>Kurtosis: -0.163</td>
<td></td>
</tr>
<tr>
<td>Min: 6.000</td>
<td>Max: 12.000</td>
</tr>
</tbody>
</table>

### RepeatTreatment 'RepeatTreatment'

<table>
<thead>
<tr>
<th>Storage mode: double</th>
<th>Measurement: interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean: 0.000</td>
<td></td>
</tr>
<tr>
<td>Variance: 0.000</td>
<td></td>
</tr>
<tr>
<td>Skewness: NaN</td>
<td></td>
</tr>
<tr>
<td>Kurtosis: NaN</td>
<td></td>
</tr>
<tr>
<td>Min: 0.000</td>
<td>Max: 0.000</td>
</tr>
</tbody>
</table>

### Subject 'Subject'

<table>
<thead>
<tr>
<th>Storage mode: double</th>
<th>Measurement: interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean: 8.720</td>
<td></td>
</tr>
<tr>
<td>Variance: 22.665</td>
<td></td>
</tr>
<tr>
<td>Skewness: 0.028</td>
<td></td>
</tr>
<tr>
<td>Kurtosis: -1.165</td>
<td></td>
</tr>
<tr>
<td>Min: 1.000</td>
<td>Max: 18.000</td>
</tr>
</tbody>
</table>

### Pos 'Pos'
Storage mode: double
Measurement: interval

Mean: 2.280
Variance: 1.202
Skewness: 0.317
Kurtosis: -1.214
Min: 1.000
Max: 4.000

Group 'Group'

Storage mode: double
Measurement: interval

Mean: 3.049
Variance: 3.510
Skewness: 1.287
Kurtosis: 1.685
Min: 1.000
Max: 9.000

Offer 'Offer'

Storage mode: double
Measurement: interval

Mean: 0.327
Variance: 0.137
Skewness: 0.481
Kurtosis: -1.424
Min: 0.000
Max: 1.000
Miss.: 3072.000
NAs: 3072.000

Receive 'Receive'
Storage mode: double
Measurement: interval

Mean: 0.981
Variance: 1.230
Skewness: 0.481
Kurtosis: -1.424
Min: 0.000
Max: 3.000
Miss.: 3072.000
NAs: 3072.000

================================================================================
Return 'Return'
================================================================================
Storage mode: double
Measurement: interval

Mean: 0.409
Variance: 0.381
Skewness: 1.502
Kurtosis: 1.437
Min: 0.000
Max: 2.763
Miss.: 3072.000
NAs: 3072.000

================================================================================
GetBack 'GetBack'
================================================================================
Storage mode: double
Measurement: interval

Mean: 0.409
Variance: 0.381
Skewness: 1.502
Kurtosis: 1.437
Min: 0.000
Max: 2.763
Miss.: 3072.000
NAs: 3072.000

================================================================================
country 'country'
siblings 'siblings'

sex 'sex'

age 'age'
Storage mode: double
Measurement: interval

Mean: 31.421
Variance: 511.610
Skewness: 2.478
Kurtosis: 4.619
Min: 16.000
Max: 99.000

Contrib1 'Contrib1'

Storage mode: double
Measurement: interval

Mean: 0.504
Variance: 0.047
Skewness: -0.047
Kurtosis: -0.269
Min: 0.000
Max: 1.000
Miss.: 864.000
NAs: 864.000

Contrib2 'Contrib2'

Storage mode: double
Measurement: interval

Mean: 0.507
Variance: 0.048
Skewness: 0.006
Kurtosis: -0.429
Min: 0.000
Max: 1.000
Miss.: 864.000
NAs: 864.000

Contrib3 'Contrib3'
The `memisc` approach preserves more information. Often this is more intuitive. Some packages are, however, confused by these attributes.

**Stata 13** Every now and then stata changes their file format:

```r
library(readstata13)
sta13<-read.dta13("data/160716_060x.dta13")
```

### 6.1.4 Reading CSV Files

CSV-Files (Comma-Separated-Value) Files are in no way always comma separated. The term is rather used to denote any table with a constant separator. Some of the parameters that always change are:

- Separators: ,; TAB
- Quoting of strings: " ’ —
- Headers: with / without

As a result, the `read.table` has many parameters.

```r
csv <- read.csv("data/160716_060x.csv",sep="\t")
str(csv)
```

The advantage of CSV as a medium to exchange data is: CSV can be read by any software.

The disadvantage is: No extra information (variable labels, levels of factors, . . .) can be stored.
6.1.5 Filesize

For our example we obtain the following sizes:

<table>
<thead>
<tr>
<th>Format</th>
<th>Size / Bytes</th>
</tr>
</thead>
<tbody>
<tr>
<td>xlsx</td>
<td>96048</td>
</tr>
<tr>
<td>xls</td>
<td>30856</td>
</tr>
<tr>
<td>dta</td>
<td>613274</td>
</tr>
<tr>
<td>dta13</td>
<td>618378</td>
</tr>
<tr>
<td>csv</td>
<td>301048</td>
</tr>
<tr>
<td>Rdata</td>
<td>24508</td>
</tr>
</tbody>
</table>

6.2 Checking Values

```r
load("data/160716_060x_C.Rdata")
```

6.2.1 Range of values

```r
codebook(data.set(trustC))
```

```
--------------------------------------------------------------------------------
trustC.Offer 'trustor’s offer'

Storage mode: double
Measurement: interval

Min:  0.000
Max:  1.000
Mean: 0.654
Std.Dev.: 0.244
Skewness: -0.684
Kurtosis: 0.034
Miss.:  216.000
NAs:   216.000

--------------------------------------------------------------------------------
trustC.country 'country of origin'

Storage mode: double
Measurement: nominal
```
6.2.2 (Joint) distribution of values

Basic plots

```r
with(trustC, hist(GetBack/Offer))
boxplot(GetBack/Offer ~ sub("_","",Date), data=trustC, main="Boxplot")
with(trustC, plot(ecdf(GetBack/Offer)))
abline(v=1)
```
Joint distributions  First pool all data:

```r
plot(GetBack ~ Offer, data=trustC)
abline(a=0, b=3)
```

If something is suspicious (which does not seem to be the case here) plot the data for subgroups:

```r
coplot(GetBack ~ Offer | Period + Date, data=trustC, show.given=FALSE)
```

The Kakadu data contains variables `lower` and `upper`.

```r
data(Kakadu)
nrow(Kakadu)
```

```
[1] 1827
```

- `lower`: lowerbound of willingness to pay, 0 if observation is left censored
• upper upper bound of willingness to pay, 999 if observation is right censored

When our data falls into a small number of categories a simple scatterplot is not too informative. The right graph shows a scatterplot with some jitter added.

```
plot(lower ~ upper, data=Kakadu)
abline(a=0, b=1)
plot(jitter(lower, factor=50) ~ jitter(upper, factor=50), cex=.1, data=Kakadu)
```

With such a large number of observations, and so few categories, a table might be more informative

```
with(Kakadu, table(lower, upper))
```

```
upper

<table>
<thead>
<tr>
<th>lower</th>
<th>2</th>
<th>5</th>
<th>20</th>
<th>50</th>
<th>100</th>
<th>250</th>
<th>999</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>129</td>
<td>147</td>
<td>156</td>
<td>176</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>63</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>20</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>69</td>
<td>0</td>
<td>0</td>
<td>321</td>
</tr>
<tr>
<td>50</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>76</td>
<td>0</td>
<td>281</td>
</tr>
<tr>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>61</td>
<td>187</td>
</tr>
<tr>
<td>250</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>152</td>
</tr>
</tbody>
</table>
```

6.2.3 (Joint) distribution of missings

• Do we expect any missings at all?

• Are missings where they should be?
  - e.g. number of siblings=0, age of oldest sibling=NA
– e.g. number of siblings=NA, age of oldest sibling=25

In our dataset we do not have the age of the oldest sibling, but let us just pretend:

```r
with(trustGS$subjects, table(siblings, age, useNA='always'))
```

<table>
<thead>
<tr>
<th>age</th>
<th>siblings</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>16 17 18 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 38 &lt;NA&gt;</td>
</tr>
<tr>
<td>0</td>
<td>12 24 0 0 12 24 36 0 0 12 24 0 0 12 24 12 0</td>
</tr>
</tbody>
</table>

[reached getOption("max.print") -- omitted 5 rows]

```r
with(trustGS$subjects, table(siblings, is.na(age)))
```

<table>
<thead>
<tr>
<th>siblings</th>
<th>FALSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>228</td>
</tr>
<tr>
<td>1</td>
<td>180</td>
</tr>
<tr>
<td>2</td>
<td>192</td>
</tr>
<tr>
<td>3</td>
<td>252</td>
</tr>
<tr>
<td>99</td>
<td>12</td>
</tr>
</tbody>
</table>

The discussion of value labels in section 6.5 contains more details on missings.

### 6.2.4 Checking signatures

How can we make sure that we are working on the “correct dataset”? Assume you and your coauthors work with what you think is the same dataset, but you get different results.

Solution: compare checksums.

```r
library(tools)
md5sum("data/160716_060x.Rdata")
data/160716_060x.Rdata
"9551b43d01a79e6659c15d86b2cae879"
```

It might be worthwhile to include in the draft version of your paper the checksum of your datasets.

### 6.3 Naming variables

We already mentioned variable names in section 83.

• short but not too short

```r
lm ( otherinvestment ~ trust + ineq + sex + age + latitude + longitude )
lm ( R100234 ~ R100412 + R100017 + R100178 + R100671 + R100229 + R100228 )
lm ( otherinvestment ~ trustworthiness + inequalityaversion + sexOfProposer + ageOfProposer )
lm ( oi ~ t + i + s + a + l1 + l2 )
```
changing existing variables creates confusion, better create new ones

Keep related variables alphabetically together.

... ProfitA ProfitB ProfitC ...

and not

... AProfit BProfit CProfit ...

How do we order variable names anyway?

trustC[,sort(names(trustC))]

6.4 Labeling (describing) variables

Variable names should be short...

but after a while we forget the exact meaning of a variable

- What was the difference between Receive and GetBack?
- Did we code male=1 and female=2 or the opposite?

Labels provide additional information.

Either...

- use a small number of source files, and keep the information somewhere in the file

... or ...

- use many source files and few data files, and keep the information with the data.

```r
load("data/160716_0601.Rdata")
trust <- within(within(trustGS, merge(globals, subjects)), {
  description(Pos)<- "(1=trustor, 2=trustee)"
  description(Offer)<- "trustor's offer"
  description(Receive)<- "amount received by trustee"
  description(Return)<- "amount trustee sends back to trustor"
  description(GetBack)<- "amount trustor receives back from trustee"
  description(country)<- "country of origin"
  description(sex)<- "participant's sex (1=male, 2=female)"
  description(siblings)<- "number of siblings"
  description(age)<- "true age"
})
codebook(data.set(trust))
attr(trust,"annotation")<-"Note: 160716_0601 was a pilot,..."
annotation(trust)["note"]="Note: This is not a real dataset..."
```
labels can be long, but they should be meaningful even if they are truncated.
The following is not a label but a wording:

```r
description(uncondSend) <- "how much would you send to the
other player if no binding contract was possible"
description(condSend) <- "how much would you send to the
other player if you had the possibility of a binding contract"
```

Better:

```r
description(uncondSend) <- "how much to send without binding
contract"
description(condSend) <- "how much to send with binding
contract"
wording(uncondSend) <- "how much would you send to the other
player if no possibility of a binding contract was possible"
wording(condSend) <- "how much would you send to the other
player if you had the possibility of a binding contract"
```

**General attributes**
- `description()`: short description of the variable — always
- `wording()`: wording of a question — if necessary
- `annotation() ["..."]`: e.g. specific property of dataset — if necessary
  - how a variable was created — if necessary

### 6.5 Labeling values

Let us again list some interesting datatypes:

- **numbers**: 1, 2, 3
- **characters**: “male”, “female”, ...
- **factors**: “male”=1, “female”=2, ...
  - factors are integers + levels, often treated as characters.
  - factors have only one type of missing (this is not a restriction, since the type of missingness could be stored in another variable)

The **memisc**-package provides another type: *item*

- **item**: “male”=1, “female”=2, ...
  - items are numbers + levels, often treated as numbers.
  - items can have several types of missing. Useful for questionnaire (or from z-Tree).
```
codebook(trustC$sex)

================================================================================
trustC$sex 'participant's sex (1=male, 2=female)'
================================================================================

Storage mode: double
Measurement: nominal
Missing values: 98, 99

<table>
<thead>
<tr>
<th>Values and labels</th>
<th>N</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 'male'</td>
<td>174</td>
<td>44.6</td>
</tr>
<tr>
<td>2 'female'</td>
<td>216</td>
<td>55.4</td>
</tr>
<tr>
<td>98 M 'refused'</td>
<td>18</td>
<td>4.2</td>
</tr>
<tr>
<td>99 M 'missing'</td>
<td>24</td>
<td>5.6</td>
</tr>
</tbody>
</table>

`table(as.factor(trustC$sex),useNA="always")`

```
<table>
<thead>
<tr>
<th>male</th>
<th>female</th>
<th>&lt;NA&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>174</td>
<td>216</td>
<td>42</td>
</tr>
</tbody>
</table>
```

`table(as.numeric(trustC$sex),useNA="always")`

```
<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>&lt;NA&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>174</td>
<td>216</td>
<td>42</td>
</tr>
</tbody>
</table>
```

`table(as.character(trustC$sex),useNA="always")`

```
<table>
<thead>
<tr>
<th>female</th>
<th>male</th>
<th>missing</th>
<th>refused</th>
<th>&lt;NA&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>216</td>
<td>174</td>
<td>24</td>
<td>18</td>
<td>0</td>
</tr>
</tbody>
</table>
```

`useNA="always"` allows us to count missings. `mean(is.na())` allows us to calculate the fraction of missings. The result depends on the representation.

```
mean(is.na(trustC$sex))
[1] 0
mean(is.na(as.factor(trustC$sex)))
[1] 0.09722222
mean(is.na(as.numeric(trustC$sex)))
[1] 0.09722222
mean(is.na(as.character(trustC$sex)))
[1] 0
```
How do we add labels to values? (requires memisc)

```r
trust <- within(trust,
  labels(sex)<-c("male"=1,"female"=2,"refused"=98,"missing"=99)
  labels(siblings)<-c("refused"=98,"missing"=99)
  labels(age)<-c("refused"=98,"missing"=99)
  labels(country)<-c("a"=1, "b"=2, "c"=3, "d"=4, "e"=5, "f"=6, "g"=7, "h"=8, "i"=9, "j"=10, "k"=11, "l"=12, "m"=13, "n"=14, "o"=15, "p"=16,"refused"=98,"missing"=99)
  missing.values(sex)<-c(98,99)
  missing.values(siblings)<-c(98,99)
  missing.values(age)<-c(98,99)
  missing.values(country)<-c(98,99)
)
```

### 6.6 Recoding data

#### 6.6.1 Replacing meaningless values by missings

In our trust game not all players have made all decisions. z-Tree coded these “decisions” as zero. This can be misleading. Better code them as missing.

```r
trustC <- within(trust, {
  Offer [Pos==2 & Offer==0] <-NA
  GetBack[Pos==2 & GetBack==0]<-NA
  Receive[Pos==1 & Receive==0]<-NA
  Return [Pos==1 & Return==0] <-NA
})

save(trustC,file="data/160716_060x_C.Rdata")
```

Introducing missings makes a difference. The left graph shows the plot where missings were coded (wrongly) as zeroes, the right graph shows the plot with missings coded as missings.

```r
c(ecdfplot(~Offer,data=trust),ecdfplot(~Offer,data=trustC))
```
mean(trust$Offer)

[1] NA

mean(trustC$Offer)

[1] NA

mean(trustC$Offer, na.rm=TRUE)

[1] 0.6536776

### 6.6.2 Replacing values by other values

Sometimes we want to simplify our data. E.g. the siblings variable might be too detailed.

```r
trustC <- within(trustC, altSiblings<-recode(siblings,
    "single child" = 0 <- 0,
    "siblings" = 1 <- range(1,50),
    "refused" = 98 <- 98,
    "missing" = 99 <- 99))
```

### 6.6.3 Comparison of missings

We can not compare `NA`s. The following will fail in R:

```r
if(NA == NA) print("ok")
```

```r
Error in if (NA == NA) print("ok"): missing value where TRUE/FALSE needed
```

```r
if(7 < NA) print("ok")
```

```r
Error in if (7 < NA) print("ok"): missing value where TRUE/FALSE needed
```

(Note that the equivalent in Stata, `. == .` and `7 < .`, do not fail but returns TRUE.)

The following works:

```r
x<-NA
if(is.na(x)) print("x is na")
```

[1] "x is na"

### 6.7 Creating new variables

- give them new names (overwriting “forgets” previous information)
• give them labels
• keep the old variables

6.8 Select subsets
(See the remarks on subsetting in section 5.1)

• delete records you will never ever use (in the cleaned data, not in the raw data)

```r
trust <- subset(trust, Pos != 2)
```

• generate indicator variables for records you will use in a specific context

```r
trust <- within(trust, youngSingle <- age < 25 & siblings == 0)
with(subset(trust, youngSingle), ...)
```

7 Weaving and tangling

• Describe the research question
  Which model do we use to structure this question?
  Which hypotheses do we want to test?

• Describe the method

• Describe the sample
  How many observations, means, distributions of main variables, key statistics
  Is there enough variance in the independent variables to test what you want to test?

• Test hypotheses based on the model.
  possibly different variants of the model (increasing complexity)

• Discuss model, robustness checks

7.1 How can we link paper and results?

Lots of notes in the paper, e.g. the following:
In your \LaTeX-file...:

```latex
\% the following table was created by tableAvgProfits()
\% from projectXYZ_160621.R
\% \begin{table}
\% ...
```

Better: Weave (Sweave, knitr)
7.2 A history of literate programming

Donald Knuth: The CWEB System of Structured Documentation (1993)

What is “literate programming”:

- meaningful and readable high-quality documentation
- details are usually not included in comments
- supposed to be read
- facilitates feedback and reuse of code
- reduces the amount of text one must read to understand the code

Literate programming for empiricists:

- tangle (Stangle, knit(..., tangle=TRUE)): foo.Rnw $\rightarrow$ foo.R
- weave (Sweave, knit): foo.Rnw $\rightarrow$ foo.tex (may contain parts of foo.R)

What does Rnw mean:

- R for the R project
- nw for noweb (web for no particular language, or Norman Ramsey’s Web)
Nonliterate versus literate work

**Nonliterate:**

\[
\text{raw data} \rightarrow \text{statistical methods} \rightarrow \text{paper} \rightarrow \text{workflow}
\]

Remember: it is easy to confuse the different version of the analysis and their relation to the versions of the paper.

**Literate:**

\[
\text{raw data} \rightarrow \text{statistical methods} \rightarrow \text{workflow} \rightarrow \text{paper}
\]

With literate programming in the analysis we avoid one relevant source of errors: Confusion about which parts of our work do belong together and which do not.

**Advantages of literate programming**

- *Connection* of methods to paper (no more: ‘which version of the methods were used for which figure, which table’)

- The paper is *dynamic*
  - More raw data arrives: the new version of the paper writes itself
  - You organise and clean the data differently: the new version of the paper writes itself
  - You change a detail of the method which has implications for the rest of the paper: the new version of the paper writes itself

Don’t write:

We ran 12 sessions with 120 participants.

instead:

\[
\text{numSession} <- \text{length(unique(sessionID))}
\]

\[
\text{numPart} <- \text{length(unique(partID))}
\]

\[
\ldots
\]

We ran \Sexpr{numSession} sessions with \Sexpr{numPart} participants.
### 7.3 An example

Here is a brief Rnw-document:

```latex
\documentclass{article}
\begin{document}

\text{text that explains what you are doing and why it is interesting ...}

<<someCalculations,results='asis',echo=FALSE>>=
library(Ecdat)
library(xtable)
library(lattice)
data(Caschool)
attach(Caschool)
est <- lm(testscr ~ avginc)
xtable(anova(est))
@

<<aFigure,echo=FALSE,fig.width=4,fig.height=3>>=
plot(xyplot(testscr ~ avginc,xlab="average income",ylab="testscore",
            type=c("p","r","smooth")))
@

the correlation between average income and testscore is
\Sexpr{round(cor(testscr,avginc),4)}
more text ...
\end{document}
```

To compile this Rnw-file, we can do the following:

```r
l library(knitr)
k knit("<filename.Rnw>")
s system("pdflatex <filename.tex>")
```

...or use a front end like RStudio.
The result, after knitting:

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>avginc</td>
<td>1</td>
<td>77204.39</td>
<td>77204.39</td>
<td>430.83</td>
<td>0.0000</td>
</tr>
<tr>
<td>Residuals</td>
<td>418</td>
<td>74905.20</td>
<td>179.20</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

the correlation between average income and testscores is 0.7124.

more text ...

7.4 Text chunks

What we saw:

- The usual LaTeX-text
- "chunks" like this

```
<<>>=
lm(testscr ~ avginc)
```

or "chunks" with parameters:

```
<<fig.height=2.5>>=
plot(est,which=1)
```

more generally

```
<<...parameters...>>=
...R-commands...
```

What are these parameters:
not necessary, but identifies the chunk. Also helps recycling chunks, e.g. a figure.

this chunk will not be evaluated

the code of this chunk will not be shown

All figures produced in this chunk will have this width and height.

The chunk produces \LaTeX\-output which should be inserted here

Furthermore you can include small parts of output in the text:
\Sexpr{...}

Elements of a knitr-document

\documentclass{article}
\begin{document}
\begin{document}

\Sexpr{...}

\begin{itemize}
\item dev='tikz', external=FALSE sets the format for plots
\end{itemize}

Since tikz is at the moment not part of the standard R packages, you have to install with install.packages("tikzDevice", repos="http://R-Forge.R-project.org")
This works well on Unix based systems. On a Microsoft Windows system you may need the Windows toolset for R which is not part of the standard distribution.
• fig.width=4.5, fig.height=3 the the size for plots
• echo=TRUE, warning=TRUE, error=TRUE, message=TRUE what kind of output is shown
• cache=TRUE, autodep=TRUE do calculate chunks only when they have changed
• size="footnotesize" size of the output

All these values can be overridden for specific knitr chunks.

**Words of caution** There is still something that might break:
In case something in R changes in the future, better put somewhere in your document:

```r
This document has been generated on \today, with \Sexpr{version$version.string}, on \Sexpr{version$platform}.
```

This document has been generated on July 29, 2016, with R version 3.3.1 (2016-06-21), on x86_64-pc-linux-gnu.

### 7.5 Advantages
- Accuracy (no more mistakes from copying and pasting)
- Reproducability (even years later, it is always clear how results were generated)
- Dynamic document (changes are immediately reflected everywhere, this speeds up the writing process)

### 7.6 Practical issues

**What if some calculations take too much time** Usually you will not be able (or willing) to *always* do the entire journey from your raw data to the paper in one single step.

```r
<<fastOrSlow>>=
FAST=FALSE
@

<<eval=!FAST>>=
read.csv('rawData.csv')
expensiveData<-thisTakesALongTime()
save(expensiveData,file='expensive.Rdata')
@

<<<>>>=
load('expensive.Rdata')
...@
```
Alternatively: caching intermediate results  knitr can also cache intermediate results:

```r
<<expensiveStep,cache=TRUE>>=
intermediateResults <- ....
@
```

The above chunk is executed only once (unless it changes), results are stored on disk and can be used lateron.

### 7.7 When R produces tables

#### 7.7.1 Tables

You can save a lot of work if you harness R to create and format your tables for you. A versatile function is `xtable`:

```r
(x <- rbind(c(1,2,3),c(4,5,6)))

[,1] [,2] [,3]
[1,] 1 2 3
[2,] 4 5 6

<<results='asis'>>
```

```r
library(xtable)
xtable(x)
```

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.00</td>
<td>2.00</td>
<td>3.00</td>
</tr>
<tr>
<td>2</td>
<td>4.00</td>
<td>5.00</td>
<td>6.00</td>
</tr>
</tbody>
</table>

#### 7.7.2 Estimation results

Estimation results in tabular form from `mtable` are typeset by `toLatex`:

```r
library(Ecdat)
data(Caschool)
est1 <- lm(testscr ~ str,data=Caschool)
est2 <- lm(testscr ~ str + elpct,data=Caschool)
est3 <- lm(testscr ~ str + elpct + avginc,data=Caschool)
toLatex(mtable(est1,est2,est3))
```
Nicer names for variables and equations

toLatex(relabel(mtable("simple"=est1,"intermediate"=est2, "final"=est3),c(str="student/teacher", elpct="English learners", avginc="average income", 
"(Intercept)"="Constant")))
<table>
<thead>
<tr>
<th></th>
<th>simple</th>
<th>intermediate</th>
<th>final</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>698.933***</td>
<td>686.032***</td>
<td>640.315***</td>
</tr>
<tr>
<td></td>
<td>(9.467)</td>
<td>(7.411)</td>
<td>(5.775)</td>
</tr>
<tr>
<td>student/teacher</td>
<td>−2.280***</td>
<td>−1.101**</td>
<td>−0.069</td>
</tr>
<tr>
<td></td>
<td>(0.480)</td>
<td>(0.380)</td>
<td>(0.277)</td>
</tr>
<tr>
<td>English learners</td>
<td>−0.650***</td>
<td>−0.488***</td>
<td>−0.069</td>
</tr>
<tr>
<td></td>
<td>(0.039)</td>
<td>(0.029)</td>
<td>(0.075)</td>
</tr>
<tr>
<td>average income</td>
<td></td>
<td></td>
<td>1.495***</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.075)</td>
</tr>
</tbody>
</table>

R-squared: 0.1
adj. R-squared: 0.0
sigma: 18.6
F: 22.6
p: 0.0
Log-likelihood: −1822.2
Deviance: 144315.5
AIC: 3650.5
BIC: 3662.6
N: 420

Requirements The default of toLatex assumes the dcolumn package, i.e. in the preamble of the document we have to say something like:

\usepackage{dcolumn}
\let\toprule\hline
\let\midrule\hline
\let\bottomrule\hline

7.7.3 Mixed effects

If we use lmer to estimate models with mixed effects, toLatex needs a summary.mer method. The following is one example:

```r
bootSize <- 1000
getSummary.mer <- function(mer) {
  msd <- sqrt(diag(vcov(mer)))
  coefs <- fixef(mer)
  mz <- mcmcsamp(mer, bootSize)
  mf <- mz@fixef
  mzp <- 2*pnorm(-abs(mz) - (mzcoef <- apply(mf, 1, mean))/
                  (mzsd <- apply(mf, 1, sd)))
  mzci <- cbind(coefs) %*% c(1, 1) + cbind(mzsd) %*% rbind(qnorm(c(.025, .975)))
  coef <- cbind(coefs, mzsd, mzp, mzci)
  colnames(coef) <- c("est", "se", "stat", "p", "lwr", "upr")
  smer <- summary(mer)
  return(coef)
}
```
We should note that our definition of *indep.obs.* and *participants* as the smallest and largest number of groups, respectively, is often reasonable if we have, indeed, two random effects, one for independent observations, the other for participants. This is frequently the case for experiments but need not always be the case for other mixed effects models.

We should also note that there are several ways to bootstrap *p*-values. In the example we use *mcmcsamp* and we assume that the distribution of coefficients follows a normal distribution.

### 7.8 The magic of make

In the same directory where I have my Rnw file, I also have a file that is called *Makefile*. Let us assume that the current version of my Rnw file is called *myProject_160601.Rnw*. Then here is my *Makefile*:

```makefile
PROJECT = myProject_160601

pdf: $(PROJECT).pdf

%.pdf: %.tex
  pdflatex $<
```
Let us go through the individual lines of this Makefile.

```makefile
PROJECT = myProject_160601
```

Here we define a variable. This is useful, since this most of the time the only line of the Makefile I ever have to change (instead of changing every occurrence of the filename)

```makefile
pdf: $(PROJECT).pdf
```

The part pdf before the colon is a target. Since it is the first target in the file it is also the default target. I.e. make will try to make it whenever I just say

```
make
```

Make will do the same when I call it explicitly

```
make pdf
```

The part after the colon tells make on which file(s) the target actually depends (the prerequisites). Here it is only one but there could be several. If all prerequisites exists, and if they are up-to-date (newer than all files they depends on), make will apply the rule. Otherwise, make will try to create the prerequisites (the pdf file in this case, with the help of other rules) and then apply this rule.

```makefile
%.tex: %.Rnw
    echo "library(knitr);knit("$<");");" | R --vanilla
```

This is a rule that make can use to create tex files. So above we requested the pdf file myProject_160601.pdf, and now make knows that we require a file myProject_160601.tex. If this already exists and is up-to-date (i.e. newer than all files it depends on), make will apply this rule. Otherwise, make will first try to create the prerequisite (the single tex file in this case would be created with the help of other rules) and then apply this rule.

To create our pdf it is now sufficient to say (from the command line, not from R)

```
make
```

and make will do everything that is needed.

Note: In this context a simple shell script would work almost as well. However, make is very helpful when your pdf file depends on more than one tex or Rnw file.
A Makefile for a larger project  When I wrote this handout I split it into several Rnw files. This saves time. When I make changes to one part, only this part has to be compiled again. The files were all in the same directory. The directory also contained a “master”-tex file that would assemble the tex-files for each Rnw-file.

The following example shows how we assemble the output of several files to make one document:

```
PROJECT = myProject_160601
RPARTS   = $(wildcard $(PROJECT)_[1-9].Rnw)
TEXPARTS = $(RPARTS:.Rnw=.tex)

pdf: $(PROJECT).pdf
    $(PROJECT).pdf: $(TEXPARTS) $(PROJECT).tex
        pdflatex $(PROJECT)

# only the tex files who belong to Rnw files
# should be knitted:
$(TEXPARTS) : %.tex : %.Rnw ;
    echo "library(knitr);knit("$<");" | R --vanilla
```

8 Version control

8.1 Problem I – concurrent edits

What happens if two authors, Anna and Bob, simultaneously want to work on the same file. Chances are that one is deleting the changes of the other. (This problem is similar to one author working on two different machines)

- Anna’s work is lost — very inefficient (50% of the contribution is lost)

8.2 A “simple” solution: locking

Serialising the workflow might help. Anna could put a “lock” on a file while she wants to edit this file. Only when she is finished, the “unlocks” the file and Bob can continue.
Bob can only work with Anna’s permission — very inefficient (50% of the time Anna and Bob are forced to wait)

8.3 Problem II – nonlinear work

Even when Anna works on a problem on her own she can be in conflict with herself. Imagine the following: Anna successfully completed the steps A, B, and C on a paper and has now something readable that she could send around. Perhaps she actually has sent it around. Now she continues to work on some technical details D and E, but so far her work in incomplete – D and E are not ready for the public. Suddenly the need arises to go back to the last public version (C) and to add some work there (e.g. Anna decides to submit the paper to a conference, but wants to rewrite the introduction and the conclusion. It will take too much time to first finish the work on D and E, so she has to go back to C. Rewriting the introduction and conclusion are steps F and G. Once the paper (G) has been submitted, Anna wants to return to the technical bits D and E and merge them with F and G.

8.4 Version control

(revision control, source control) Traditional:

- Editions of a book
- Revisions of a specification

Software:

- Concurrent Versions System (CVS)
- Subversion (SVN)
- Git
• Mercurial
• Bazaar
• ...

In this course we will use Git.

• Free
• Distributed repository
• Supports many platforms, formats
• ...

8.5 Solution to problem II: nonlinear work

Before we create our first git-repository, we have to provide some basic information about ourselves:

```sh
git config --global user.name "Your Name Comes Here"
git config --global user.email you@yourdomain.example.com
```

Now we can create our first repository:

```sh
git init
```

We can check the current “status” as follows:

```sh
git status
```

```sh
# On branch master
# Initial commit
# nothing to commit (create/copy files and use "git add" to track)
```

Now we create a file `test.Rnw`

```sh
git status
```

```sh
# On branch master
# Initial commit
# Untracked files:
#   (use "git add <file>..." to include in what will be committed)
#   test.Rnw
nothing added to commit but untracked files present (use "git add" to track)
```

```sh
git add test.Rnw
```
git status

# On branch master
# Initial commit
# Changes to be committed:
# (use “git rm --cached <file>...” to unstage)
# new file:  test.Rnw

git commit -a -m "first version of test.Rnw"

git status

# On branch master
nothing to commit, working directory clean

git log --oneline

3ea6194 first version of test.Rnw

Note that git denotes versions with identifiers like “3ea6194” (and not A, B, C).
After some changes to test.Rnw...

git status

# On branch master
# Changes not staged for commit:
# (use “git add <file>...” to update what will be committed)
# (use “git checkout -- <file>...” to discard changes in working directory)
# modified:  test.Rnw
# no changes added to commit (use "git add" and/or "git commit -a")

git commit -a -m "introduction and first results"

git status

nothing to commit, working directory clean

git log -oneline

74fd521 introduction and first results
3ea6194 first version of test.Rnw

More changes and...

git commit -a -m "draft conclusion"

more changes and...

git commit -a -m "improved regression results (do not fully work)"

more changes and...

git commit -a -m "added funny model (does not fully work yet)"
Assume we want to go back to 1d05e8f but not forget what we did between 1d05e8f and f965066.

Remember current state:

```bash
git branch funny
```

Now that we have given the current branch a name we can revert to the old state:

```bash
git reset 1d05e8f
```

Unstaged changes after reset:
```
M test.Rnw
```

```bash
git checkout test.Rnw
```

# On branch master
nothing to commit, working directory clean

```
3ea6194 → 74fd521 → 1d05e8f → 9100277 → f965066
```

do more work...

```bash
git commit -a -m "rewrote introduction"
```

do even more work...

```bash
git commit -a -m "rewrote conclusion, added literature"
```
eventually we want to join the two branches:

```
git merge funny
```

now two things can happen: Either this...

```
Merge made by recursive.
test.Rnw |   1 +
1 files changed, 1 insertions(+), 0 deletions(-)
```

...or that...

```
Auto-merging test.Rnw
CONFLICT (content): Merge conflict in test.Rnw
Automatic merge failed; fix conflicts and then commit the result
```

We can fix this with `git mergetool`:

```
git mergetool
```

Now we can make detailed merge decisions in an editor.

```
git commit -m "merged funny"
```
### 8.6 Solution to problem 1: concurrent edits

Version control allows all authors to work on the file(s) simultaneously.

In this example we start with an empty repository. In a first step both Anna and Bob “checkout” the repository, i.e. they create a local copy of the repository on their computer.

Anna creates a file, adds it to version control and commits it to the repository. Bob then updates his copy and, thus, obtains Anna’s changes.

- First step: create a “bare” repository on a “server”
  ```
git --bare init
  ```

- This repository can now be accessed from “clients”, either on the same machine...
  ```
git clone /path/to/repository/
  ```

  ...or on a different machine via ssh (where user has access rights):
  ```
git clone ssh://user@my.server.org/path/to/repository
  ```

<table>
<thead>
<tr>
<th>Anna</th>
<th>Repository</th>
<th>Bob</th>
</tr>
</thead>
<tbody>
<tr>
<td>empty</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Anna

```r
A=...  
B=...
```

`git add test.Rnw`

`git commit`

uploads the file:

```r
A=...  
B=...
```

`git push`

Bob

```r
A=...  
B=...
```

`git pull`

---

### 8.7 Edits without conflicts:

To make this more interesting we now assume that both work on the file. Anna works on the upper part (A), Bob works on the lower part (B). Both update and commit their
changes. Since they both edit different parts of the file, the version control system can silently merge their changes.

<table>
<thead>
<tr>
<th>Anna</th>
<th>Repository</th>
<th>Bob</th>
</tr>
</thead>
<tbody>
<tr>
<td>A=1</td>
<td>A=...</td>
<td>A=...</td>
</tr>
<tr>
<td>B=...</td>
<td>B=...</td>
<td>B=2</td>
</tr>
</tbody>
</table>

Both commit their work to their own local repos:

```
git commit -a -m "..."
git commit -a -m "..."
```

Anna pulls, but there is no conflict

```
git pull
```

Anna pushes her changes

```
git push
```

Bob pulls, and finds a merge conflict

```
git pull
git mergetool
```

Bob commits his merge

```
git commit -a -m "..."
```

Bob pushes his merge

```
git push
```

Anna pulls to get the current version

```
git pull
```

8.8 Going back in time

Version control is not only helpful to avoid conflicts between several people, it also helps when we change our mind and want to have a look into the past. `git log` provides a list of the different revision of a file:

```
git log --oneline
```

f965066 added funny model (does not fully work yet)
8.9 git and subversion

- **git-Server**: requires ssh access to the server machine
- **subversion-Server**: provided by the URZ at the FSU Jena
  
  git can use subversion as a remote repository:
  
  - `git clone`  
  - `git pull`  
  - `git push`
  - `git svn clone`  
  - `git svn rebase`  
  - `git svn dcommit`

- Conceptual differences:
  
  - subversion has only one repository (on the server), git has one or more local repositories plus one or more on different servers.
  
  - inconsistent uploads to a server:
    
    subversion will not complain if after a push/commit the state on the server is different from the state on any of the clients. git will not allow this (git forces you to pull first, merge, commit, and push then)

8.10 Steps to set up a subversion repository at the URZ at the FSU Jena

If you need to set up a subversion repository here at the FSU, tell me about it and tell me the \langle urz-login\rangle s of the people who plan to use it. Technically, setting up a new repository means the following:

- ssh to subversion.rz.uni-jena.de
- `svnadmin create /data/svn/ewf/\langle repository\rangle`
- `chmod -R g+w /data/svn/ewf/\langle repository\rangle`
• set access rights for all involved \( \langle \text{urz-login} \rangle \)s in \( /\text{svn/\text{access-ewf}} \)

• then, at the local machine in a directory that actually contains only the files you want to add: \( \text{svn -username } \langle \text{urz-login} \rangle \text{ import . } \text{https://subversion.rz.uni-jena.de/svn/\text{ewf}} \) "Initial import"
  (this “imports” data into the repository)

• then, at all client machines,
  \( \text{svn -username } \langle \text{urz-login} \rangle \text{ checkout https://subversion.rz.uni-jena.de/svn/\text{ewf/}} \langle \text{repository} \rangle \)

8.11 Setting up a subversion repository on your own computer

• On your own computer: \( \text{svnadmin create } \langle \text{path} \rangle/\langle \text{repository} \rangle \) 
  \( \langle \text{path} \rangle \) is a complete path, e.g. \( /\text{home/user/Documents/} \) or \( /\text{C:MyDocuments/} \)

• then, in a directory that actually contains only the files you want to add:
  \( \text{svn import . file://}\langle \text{path} \rangle/\langle \text{repository} \rangle -m "Initial import" \)

• then, wherever you actually want to work on your own computer:
  \( \text{svn checkout file://}\langle \text{path} \rangle/\langle \text{repository} \rangle \)

• if you have \text{ssh} access to your computer you can also say from other machines:
  \( \text{svn checkout svn+ssh://}\langle \text{yourComputer} \rangle/\langle \text{path} \rangle/\langle \text{repository} \rangle \)

8.12 Usual workflow with git

While setting up a repository looks a bit complicated, using it is quite simple:

• \texttt{git pull} check whether the others did something

• editing
  - \texttt{git add} add a file to version control
  - \texttt{git mv} move a file under version control
  - \texttt{git rm} delete a file under version control

• \texttt{git commit} commit own work to local repository

• \texttt{git pull} check whether the others did something

• \texttt{git mergetool} merge their changes

• \texttt{git commit} commit merge

• \texttt{git push} upload everything to the server
8.13 Exercise

Create (in \langle path \rangle) four directories \texttt{A}, \texttt{B}, and \texttt{C}.

From \texttt{A} create a repository: \texttt{svadmin create ../R}

In \texttt{A} create a file \texttt{test.txt} with some text:

\begin{verbatim}
A=...  
B=...
\end{verbatim}

Initial import. In \texttt{A} say:

\texttt{svn import . file://\langle path \rangle/R -m "My first initial import"}

in \texttt{B}:

\texttt{svn checkout file://\langle path \rangle/R}

in \texttt{C}:

\texttt{svn checkout file://\langle path \rangle/R}

Simultaneous changes to \texttt{test.txt}

\begin{verbatim}
A=1  
B=...  
A=...  
B=2
\end{verbatim}

Commit changes

\texttt{svn commit}

Update

\texttt{svn update}

9 Exercises

Exercise 1
Have a look at the dataset \texttt{Workinghours} from the library \texttt{Ecdat}. Compare the distribution of “other household income” for whites and non-whites. Do the same for the different types of occupation of the husband.

Exercise 2
Read the data from a hypothetical experiment from \texttt{rawdata/Coordination}. Does the \texttt{Effort} change over time?

Exercise 3-a
Read the data from a hypothetical z-Tree experiment from \texttt{rawdata/Trust}. Do you find any relation between the number of siblings and trust?

Exercise 3-b
For the same dataset: Attach a label (description) to \texttt{siblings}. Attach value labels to this variable.

Exercise 3-c
Make the above a function.
Also write a function that compares the offers of all participants with \( n \) siblings with the other offers. This function should (at least) return a p-value of a two-sample Wilcoxon test (`wilcox.test`). The number \( n \) should be a parameter of the function.

**Exercise 4**
Read the data from a hypothetical z-Tree experiment from `rawdata/PublicGood`. The three variables `Contrib1`, `Contrib2`, and `Contrib3` are contributions of the participants to the other three players in their group (in groups of four).

1. Check that, indeed, in each period, players are equally distributed into four groups.
2. Produce for each period a boxplot with the contribution (i.e. 16 boxplots in one graph).
3. Add a regression line to the graph.
4. Produce for each contribution partner a boxplot with the contribution (i.e. 3 boxplots in one graph).
5. Produce an `Sweave` file that generates the two graphs. In this file also write when you estimate the average contribution reaches zero.