Statistische Analysen mit R

Liars, Outlier und selbstgemachte Statistik

Bruno Hopp, Linuxusergroup der Universität zu Köln

Software Freedom Day; 17. Sept. 2016

In the beginning . . .

Grundsatzfrage “Wat soll dat?”

“ R is a language and environment for statistical computing and graphics” (`www.r-project.org/about.html`)

In the beginning . . .

Grundsatzfrage “Wat soll dat?”

“ R is a language and environment for statistical computing and graphics” (`www.r-project.org/about.html`)

“is an integrated suite of software facilities for data manipulation, calculation and graphical display.” (dto).

In the beginning . . .

Grundsatzfrage “Wat soll dat?”

“ R is a language and environment for statistical computing and graphics” (`www.r-project.org/about.html`)

“is an integrated suite of software facilities for data manipulation, calculation and graphical display.” (dto).

Lizenz

Sowohl GNU/GPL 2 **und** GNU/GPL 3

In the beginning . . .

Grundsatzfrage “Wat soll dat?”

“ R is a language and environment for statistical computing and graphics” (`www.r-project.org/about.html`)

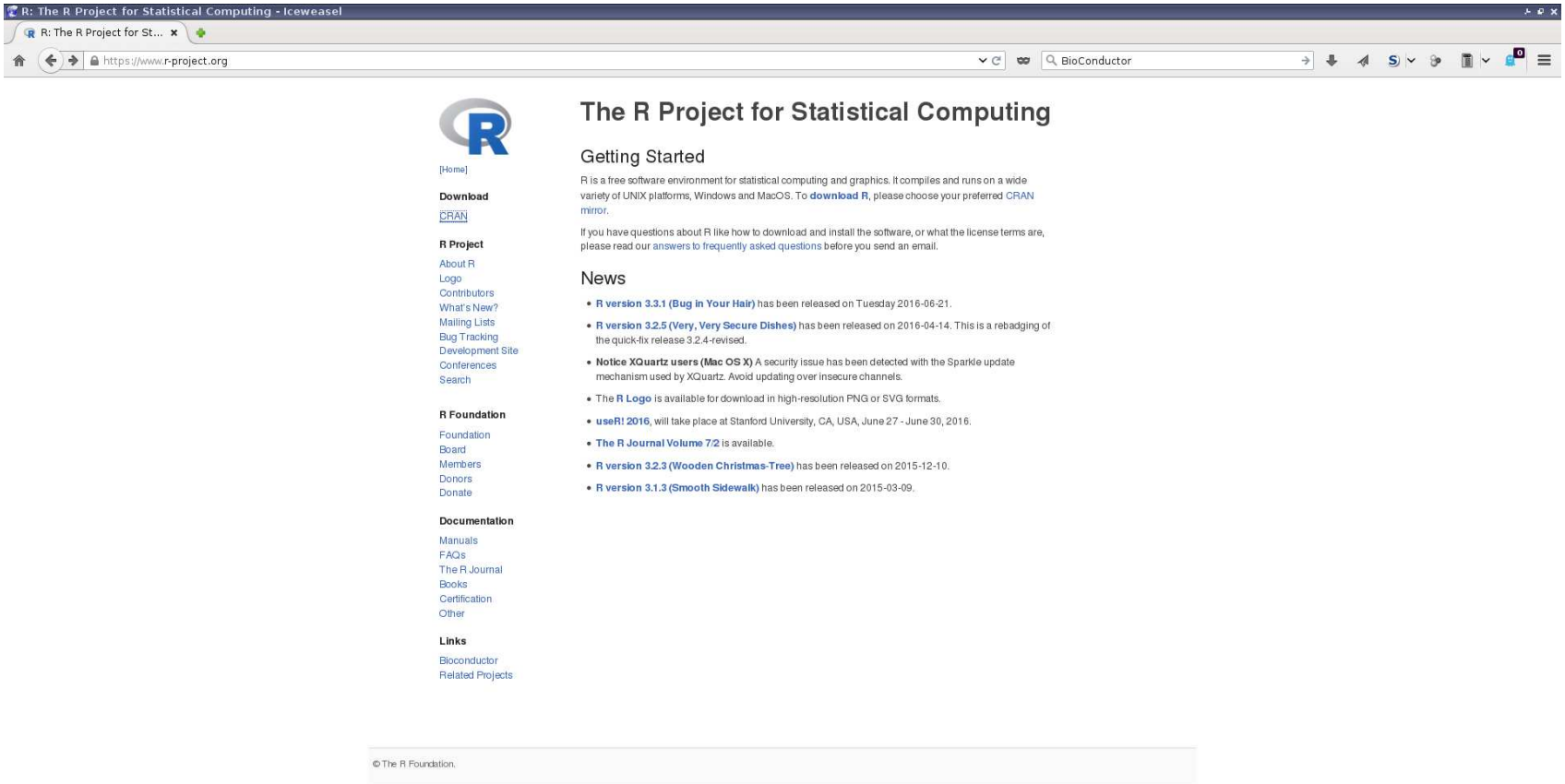
“is an integrated suite of software facilities for data manipulation, calculation and graphical display.” (dto).

Lizenz

Sowohl GNU/GPL 2 **und** GNU/GPL 3

Homepage: `www.r-project.org`

Homepage www.r-project.org



weitere Eigenschaften (Juni 2016 . . .)

- Relevante Architekturen: Linux auf i386+AMD64+ARM, Microsoft Windows, SPARC, BSD unices: OpenBSD, FreeBSD . . .
- per RDBMS (SQlite, MySQL, MariaDB, PostgreSQL)
- I/O in Fremdformaten wie JSON oder XML fürs Web,
- API/Zugang für: C, C++, Python, Fortran, Java.

In the beginning . . .

es war im letzten Jahrtausend . . .

"R was created by Ross IHAKA and Robert GENTLEMAN at the University of Auckland, New Zealand. . ."
`http://en.wikipedia.org/wiki/R_(programming_language)`
Start 1992, seit 1993 öffentlich, seit 1995 unter GNU/GPL.

In the beginning . . .

es war im letzten Jahrtausend . . .

"R was created by Ross IHAKA and Robert GENTLEMAN at the University of Auckland, New Zealand. . ."
`http://en.wikipedia.org/wiki/R_(programming_language)`
Start 1992, seit 1993 öffentlich, seit 1995 unter GNU/GPL.

Release policy

Major Releases zweimal im Jahr: Ende April und Ende Oktober.
Zur Zeit: Version 3.3.1, aka **Bug in your hair**, bis max. drei sub-Releases (3.3.3 oder 3.3.4).

CRAN - Mirrors weltweit

Das “Comprehensive R Archive Network” dient effizienter Verbreitung und leichtem Zugriff:
`http://cran.r-project.org/mirrors.html`
CRAN bietet 8729 Zusatzpakete für **R** (Stand 8.Juli 2016)
`http://cran.at.r-project.org/web/packages`

Daneben gibt es spezialisierte Sammlungen von Zusatzpaketen, etwa Bioconductor 3.3: `https://bioconductor.org` für die Biologischen Wissenschaften mit momentan 1211 software packages, 293 experimental data packages und 916 up-to-date annotation packages.

TaskViews als Installationshilfe

http://cran.at.r-project.org/web/views

hilft bei der Bewältigung komplexer Aufgaben: z.B. Lineare Modellierung, Bayesian Statistik verlangen *mehrere* Pakete gleichzeitig. Um schnell ans Ziel zu kommen, installiere eine der **TaskViews** (N=33)

```
install.packages(''ctv'')
install.views(''WebTechnologies'')
install.views(''SocialSciences'')
update.views(''SocialSciences'')
```

Haben will — Debian Linux und Gentoo Linux

`http://packages.debian.org/stable`

liefert eine veraltete Version (sources.list).
`http://cran.at.r-project.org/bin/linux/debian` in sources.list, falls du das aktuelle release möchtest.

Haben will — Debian Linux und Gentoo Linux

`http://packages.debian.org/stable`

liefert eine veraltete Version (sources.list).
`http://cran.at.r-project.org/bin/linux/debian` in sources.list, falls du das aktuelle release möchtest.

GENTOO am Puls der Zeit

emerge dev-lang/R

`http://cran.at.r-project.org/sources.html`

für die Sourcen.

Basispakete und Dokumentation

r-base r-recommended lmtest Rcmdr

Basispakete und Dokumentation

r-base r-recommended Imtest Rcmdr

Dokumentation im LaTeX 2ε Format

das ist für **R** die Regel - für jedes Paket!

Basispakete und Dokumentation

r-base r-recommended lmtest Rcmdr

Dokumentation im LaTeX 2ε Format

das ist für **R** die Regel - für jedes Paket!

`https://cran.r-project.org/other-docs.html`

liefert kurze, lange, deutsche, englische Dokumentationen!

Programmstart

Shell

R (case-sensitiv!). Leichte Frustration ...

R -g TK bietet nur wenig mehr an Luxus.

Programmstart

Shell

R (case-sensitiv!). Leichte Frustration ...
R -g TK bietet nur wenig mehr an Luxus.

Rcmdr

library(Rcmdr) startet nach einem Moment die grafische Oberfläche des **R** Commander. Es gibt weitere GUIs (ca. zehn oder mehr).

Programmstart

Shell

R (case-sensitiv!). Leichte Frustration ...
R -g TK bietet nur wenig mehr an Luxus.

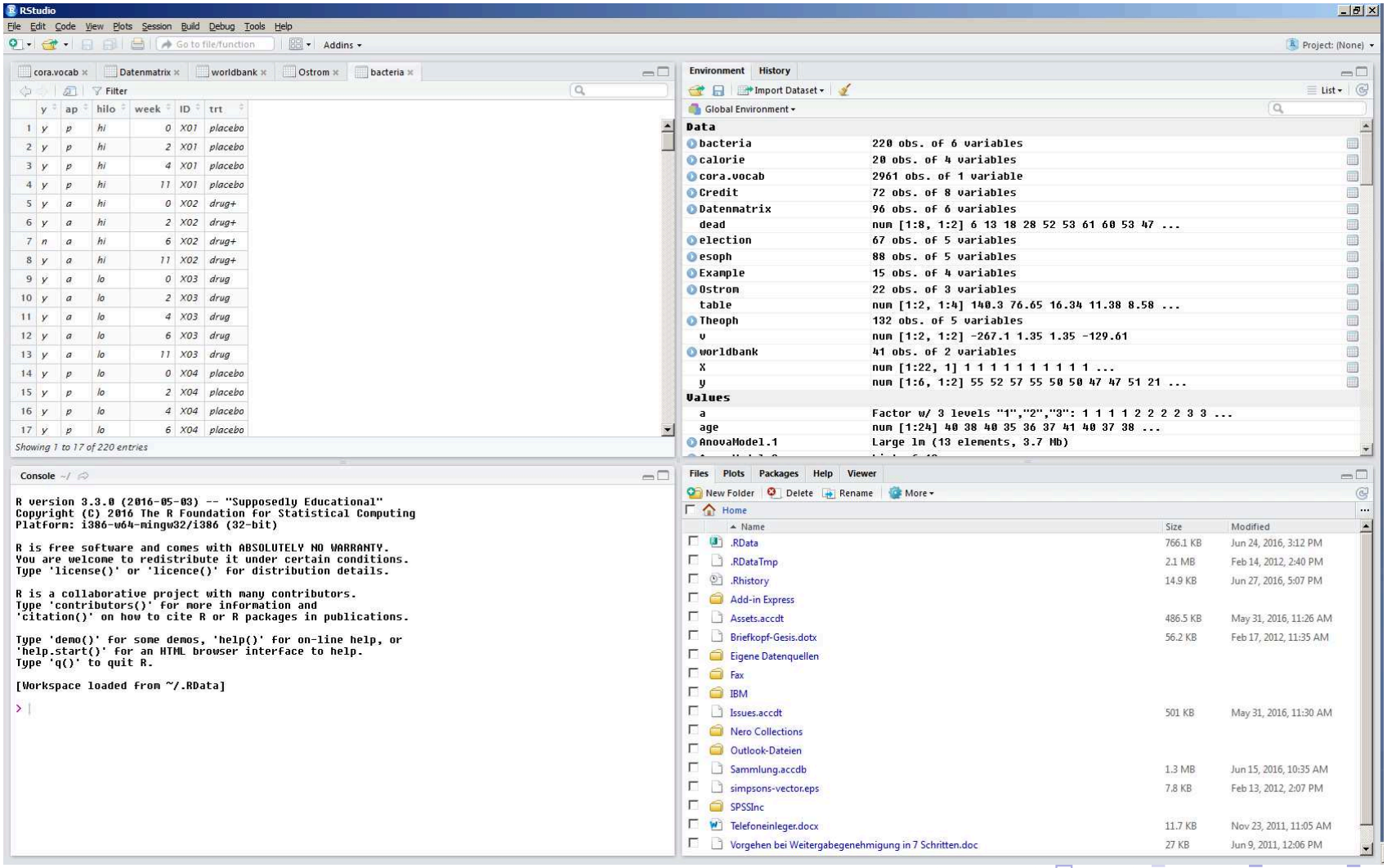
Rcmdr

library(Rcmdr) startet nach einem Moment die grafische Oberfläche des **R** Commander. Es gibt weitere GUIs (ca. zehn oder mehr).

Rstudio - eine beliebte und brauchbare IDE

emerge sci-mathematics/rstudio

Arbeitsumgebung Rstudio



Verschlungene Wege zur richtigen Analyse

welche Analysemethode?

Regressionen, Faktoranalysen, Zeitreihen, Bayesian Statistik, Netzwerkanalysen, GLM, SEM und zahlreiche andere werden von **R** unterstützt. Klassische Test-Statistik, computergestützte Klassifikationsverfahren (kNN u.a.).
Matrixalgebra mit Einbinden von ATLAS bzw. LAPACK, ähnlich wie APL oder MatLab.
Für Mathematiker: Optimierung hat ein eigenes TaskView!

Verschlungene Wege zur richtigen Analyse

welche Analysemethode?

Regressionen, Faktoranalysen, Zeitreihen, Bayesian Statistik, Netzwerkanalysen, GLM, SEM und zahlreiche andere werden von **R** unterstützt. Klassische Test-Statistik, computergestützte Klassifikationsverfahren (kNN u.a.).
Matrixalgebra mit Einbinden von ATLAS bzw. LAPACK, ähnlich wie APL oder MatLab.
Für Mathematiker: Optimierung hat ein eigenes TaskView!

Missing Values sind oft wichtig

(NA oder NaN) werden gesondert behandelt falls gewünscht.
Spezialpakete: ACD, Amelia, cat, ForImp und weitere.

Grafik und Visualisierungen

Abbildungen jeder Art

sind eine DER wichtigen Stärken von **R**. Pakete: ggplot, Hmisc u.a. erzeugen Scatterplots, Kreisdiagramme, Balken, Linien, Heatmaps, Lattices. Gern eingebaut in Webseiten, die dynamisch generiert werden.

Grafik und Visualisierungen

Abbildungen jeder Art

sind eine DER wichtigen Stärken von **R**. Pakete: ggplot, Hmisc u.a. erzeugen Scatterplots, Kreisdiagramme, Balken, Linien, Heatmaps, Lattices. Gern eingebaut in Webseiten, die dynamisch generiert werden.

Aber Vorsicht

Die Gefahr von Fehlinterpretationen ist bei Grafiken groß . . . Sehgewohnheiten des Zielpublikums beachten! Manche Akademiker hassen 3D-Plots - weiß jemand, warum?

Datenquellen in externen Formaten: **foreign**-Paket

Import/Export nach SPSS, STATA und SAS funktioniert.
Import/Export nach MS Excel funktioniert (eigenes Paket!)
Import/Export nach OpenOffice/LibreOffice/Gnumeric funktioniert.
Import/Export als CSV/clear text funktioniert.
Import/Export als JSON funktioniert (**Rjson**)
Import/Export zu div. RDBMS funktioniert (mehrere Pkg).
Elementar: JSON via Rjson/jsonlite, Rmonetdb.

Lesestoff und menschlicher Austausch

Lesen

- Joseph ADLER (2010): R in a nutshell. Köln: O'Reilly.
- Peter DALGAARD (2002): Introductory Statistics with R. New York: Springer.
- W.M. VENABLES/D. Smith/R Core Project (2016): An Introduction to R.
  `http://cran.at.r-project.org/doc/manuals/R-intro.pdf`
  Juni 2016: R-version 3.3.1
- `http://www.r-project.org/doc/bib/R-books.html`

R-bloggers `http://www.r-bloggers.com/`

R Usergroup `www.meetup.com/de-DE/KoelnRUG`

Das Ende ist nah —

Fragen, Anmerkungen? Danke!

Gleich geht es weiter . . .

Technisches: diese Folien wurden natürlich erstellt mit LaTeX 2_ε, der documentclass “Beamer” und Linux. Was denn sonst?