

Markus Tausendpfund

Datenanalyse mit R. Eine Einführung

Fakultät für
**Kultur- und
Sozialwissen-
schaften**

Das Werk ist urheberrechtlich geschützt. Die dadurch begründeten Rechte, insbesondere das Recht der Vervielfältigung und Verbreitung sowie der Übersetzung und des Nachdrucks, bleiben, auch bei nur auszugsweiser Verwertung, vorbehalten. Kein Teil des Werkes darf in irgendeiner Form (Druck, Fotokopie, Mikrofilm oder ein anderes Verfahren) ohne schriftliche Genehmigung der FernUniversität reproduziert oder unter Verwendung elektronischer Systeme verarbeitet, vervielfältigt oder verbreitet werden. Wir weisen darauf hin, dass die vorgenannten Verwertungsalternativen je nach Ausgestaltung der Nutzungsbedingungen bereits durch Einstellen in Cloud-Systeme verwirklicht sein können. Die FernUniversität bedient sich im Falle der Kenntnis von Urheberrechtsverletzungen sowohl zivil- als auch strafrechtlicher Instrumente, um ihre Rechte geltend zu machen.

Der Inhalt dieses Studienbriefs wird gedruckt auf Recyclingpapier (80 g/m², weiß), hergestellt aus 100 % Altpapier.

Vorwort

Die quantitative Datenanalyse und die Arbeit mit klassischen Statistikprogrammen wie SPSS gehört zum Curriculum vieler Studiengänge. Seit einigen Jahren erleben die Sozialwissenschaften allerdings ein neues Zeitalter: Vielfalt und Umfang sozialwissenschaftlicher Daten nehmen rapide zu, unterschiedliche Datenbestände werden systematisch verknüpft und immer leistungsfähigere Hardware erlaubt die Analyse immer größerer Datenbestände. Diese Datenbestände sowie neuere Analysetechniken erfordern allerdings neue Kompetenzen, die im Rahmen der Methodenausbildung vermittelt werden müssen. Dabei ist auch die Softwareausbildung in den Blick zu nehmen, die für die Arbeit mit den alten und neuen Datenbeständen erforderlich ist. Dabei sprechen mehrere Gründe für die Programmierumgebung R.

Erstens ist R ein Open-Source-Programm und steht für mehrere Plattformen (Windows, Mac, Linux) kostenfrei zur Verfügung. Über frei verfügbare Erweiterungen (Packages) kann der Funktionsumfang von R beträchtlich erweitert werden. Mit Blick auf Aufbereitung, Visualisierung, Analyse von Daten und Ankopplung an Datenbanksysteme fungiert R damit als Programmierumgebung, die für unterschiedlichste Aufgaben genutzt werden kann. *Zweitens* ist R methodenagnostisch und kann sowohl in der quantitativen *und* qualitativen Sozialforschung eingesetzt werden. Es existieren Erweiterungspakete sowohl für die quantitative (z.B. Regression, Faktorenanalyse) als auch für die qualitative Sozialforschung (z.B. Qualitative Comparative Analysis). *Drittens* gilt R als zukunftssicher. Als Open-Source-Programm wird R ständig weiterentwickelt. R überwindet zudem die Ein-Datensatzlogik und verfügt über Schnittstellen zu webbasierten Datensätzen. R kann auch genutzt werden, um unstrukturierte oder strukturierte Daten zu sammeln und weiterzuarbeiten.

Im Vergleich zu klassischer (kostenpflichtiger) Software wie Stata und SPSS wird R ausschließlich über die Befehlssprache gesteuert und gilt als ein Programm mit einer eher steilen Lernkurve. Zwar existieren mittlerweile einige (ebenfalls) kostenfreie Ergänzungsprogramme, die die Arbeit mit R erleichtern (z.B. RStudio), aber aufgrund der enormen Flexibilität ist der Einstieg in die Datenanalyse sicherlich anspruchsvoller als mit SPSS oder Stata.

Diese Lerneinheit bietet einen ersten Einstieg in das Statistikprogramm R und die Entwicklungsumgebung RStudio. In den nächsten Monaten wird dieser Kurs erweitert. Für hilfreiche Verbesserungsvorschläge bin ich Juliane Döschner, Patrick Heiser, Dorothee Köstlin, Ilka Schadow und Simon Stocker zu Dank verpflichtet. Für die Weiterentwicklung und Verbesserung bin ich auch auf Ihre Unterstützung angewiesen. Deshalb freue ich mich sehr über Hinweise und Anregungen zur weiteren Verbesserung dieses Skripts (gerne auch Tippfehler). Sie können Hinweise und Anregungen gerne in der Moodle-Lernumgebung posten oder mir via E-Mail (Markus.Tausendpfund@fernuni-hagen.de) mitteilen. Vielen Dank.

Hagen, im Juni 2022

Markus Tausendpfund

Inhaltsverzeichnis

Abbildungsverzeichnis	VII
Tabellenverzeichnis	VIII
1 Einführung	9
2 Installation	10
2.1 R.....	10
2.2 RStudio	11
2.3 Pakete.....	15
2.4 Aktualisierungen	15
3 R und RStudio kennenlernen.....	16
3.1 Einfache Rechenoperationen.....	16
3.2 Objekt erstellen	17
3.3 Variable erstellen.....	18
3.4 Variable mit fehlenden Werten erstellen.....	19
3.5 Grafiken mit R.....	20
3.6 Datensatz erstellen	22
3.7 Hilfe in R	24
4 Erste Analysen	25
4.1 Daten laden	25
4.2 Pakete.....	26
4.3 Daten kennenlernen	28
4.4 Datenaufbereitung	30
4.5 Univariate Datenanalyse.....	31
4.5.1 Häufigkeitstabelle.....	32
4.5.2 Lagemaße	34
4.5.3 Streuungsmaße	34
4.5.4 Formmaße.....	35
4.5.5 Erweiterte Möglichkeiten	36
4.6 Bivariate Datenanalyse	37
4.6.1 Kreuztabellen	37
4.6.2 Zusammenhangsmaße.....	40
4.7 Multivariate Datenanalyse.....	44
4.7.1 Lineare Regression.....	44
4.7.2 Logistische Regression	49

4.8	Grafiken.....	51
4.8.1	Säulen- und Balkendiagramm	51
4.8.2	Histogramm	53
4.8.3	Boxplots	54
5	Arbeiten mit R	56
5.1	Daten laden	56
5.2	Objekttypen	57
5.3	Vektoren, Datensätze und Listen	60
5.4	Objekttypen testen und konvertieren	60
5.5	Daten importieren	61
5.5.1	Excel	62
5.5.2	SPSS.....	63
5.5.3	Alternative Pakete für den Import von Datensätzen	64
6	Datenaufbereitung	65
6.1	Erforderliche Pakete.....	65
6.2	Daten importieren	65
6.3	Pipe-Operator.....	67
6.4	dplyr	68
6.4.1	count	68
6.4.2	select	68
6.4.3	rename	69
6.4.4	filter	70
6.4.5	summarise und group_by.....	71
6.4.6	mutate.....	71
6.4.7	Weitere Optionen	72
6.5	sjmisc.....	72
6.5.1	Variablen erkunden	72
6.5.2	Variablen kodieren.....	74
6.5.3	Weitere Optionen	76
6.6	sjlabelled	77
6.7	Weitere Pakete der Datenmodifikation	78
7	Multivariate Datenanalyse	79
7.1	Lineare Regression	79
7.1.1	Fragestellung, Pakete und Daten.....	79

7.1.2	Datenaufbereitung	80
7.1.3	Regressionsmodelle	82
7.1.4	Regressionsergebnisse präsentieren.....	87
7.2	Logistische Regression	92
7.2.1	Fragestellung, Pakete und Daten.....	92
7.2.2	Datenaufbereitung	93
7.2.3	Regressionsmodell.....	94
7.2.4	Regressionsmodelle präsentieren.....	96
8	Grafiken mit ggplot2	99
8.1	Pakete und Daten.....	99
8.2	Grundlagen.....	100
8.3	Diagrammtypen.....	103
8.3.1	Liniendiagramm.....	103
8.3.2	Teilgrafiken mit Facetten.....	104
8.3.3	Säulen- und Balkendiagramme.....	106
8.3.4	Boxplot	108
8.3.5	Histogramm	109
8.3.6	Streudiagramm	110
8.3.7	Streudiagramm mit Regressionsgerade.....	111
8.4	Themen	113
8.5	Erweiterungen.....	114
8.5.1	patchwork.....	115
8.5.2	ggthemes.....	115
9	Explorative Faktorenanalyse	117
9.1	Pakete und Daten.....	117
9.2	Prüfung der Items.....	119
9.3	Anzahl der Faktoren	123
9.4	Faktorenanalyse und Rotation der Faktorenmatrix.....	125
9.5	Gütekriterien und Skalenkonstruktion	126
10	Literatur	129

Abbildungsverzeichnis

Abbildung 1: The Comprehensive R Archive Network (CRAN)	10
Abbildung 2: Startbildschirm von R	11
Abbildung 3: RStudio mit drei Fenstern	12
Abbildung 4: RStudio mit vier Fenstern	13
Abbildung 5: Global Options bei RStudio	14
Abbildung 6: Streudiagramm der Variablen Alter und Einkommen	21
Abbildung 7: Plotsymbole bei R	22
Abbildung 8: Laden eines R-Datensatzes	25
Abbildung 9: Geladener Datensatz	26
Abbildung 10: Installierte Packages	27
Abbildung 11: Säulendiagramme der Variable Bildung	52
Abbildung 12: Balkendiagramm der Variable Bildung	52
Abbildung 13: Säulendiagramme der Variable gesund2 (links) und gesund (rechts).....	53
Abbildung 14: Histogramm des Alters.....	54
Abbildung 15: Boxplot der Lebenszufriedenheit in Abhängigkeit der Gesundheit	55
Abbildung 16: „Import Dataset“-Funktion in RStudio	61
Abbildung 17: Excel-Datensatz importieren	62
Abbildung 18: SPSS-Datensatz importieren	63
Abbildung 19: Datenimport mit RStudio	66
Abbildung 20: Koeffizientenplot einer linearen Regression mit sjPlot	91
Abbildung 21: Koeffizientenplot einer logistischen Regression mit sjPlot	98
Abbildung 22: Lebenserwartung in Deutschland	101
Abbildung 23: Lebenszufriedenheit und Bruttonsozialprodukt in Deutschland.....	103
Abbildung 24: Liniendiagramm.....	104
Abbildung 25: Entwicklung der Lebenserwartung in Europa	105
Abbildung 26: Bruttonsozialprodukt nach Land (Säulendiagramm).....	106
Abbildung 27: Bruttonsozialprodukt nach Land (Balkendiagramm).....	107
Abbildung 28: Lebenserwartung nach Kontinent (Boxplot)	108
Abbildung 29: Durchschnittliche Lebenserwartung in Jahren (Histogramm)	109
Abbildung 30: Lebenserwartung und Bruttonsozialprodukt (Streudiagramm I).....	110
Abbildung 31: Lebenserwartung und Bruttonsozialprodukt (Streudiagramm II).....	111
Abbildung 32: Streudiagramm mit Regressionsgerade	112
Abbildung 33: Streudiagramm mit Regressionsgerade	113
Abbildung 34: Verschiedene Themen in ggplot	114
Abbildung 35: Verschiedene Themen in ggthemes	116
Abbildung 36: Korrelationsmatrix der Vertrauensitems	120
Abbildung 37: Histogramme der Vertrauensitems.....	122
Abbildung 38: Scree plot.....	123
Abbildung 39: Parallelanalyse	124

Tabellenverzeichnis

Tabelle 1: Beispieldaten mit Alter und Einkommen	19
Tabelle 2: Beschreibung des Datensatzes	28
Tabelle 3: Wichtige Zusammenhangsmaße bei der bivariaten Datenanalyse.....	40
Tabelle 4: Informationen zu einer linearen Regression.....	45
Tabelle 5: Ausgewählte Parameter der glm-Funktion	49
Tabelle 6: Beispieldatensatz (peanuts_r)	56
Tabelle 7: Ausgewählte Funktionen zum Testen und Konvertieren von Objekten.....	61
Tabelle 8: Alternative Pakete für den Import von Datensätzen.....	64
Tabelle 9: Auswahl von Filter-Möglichkeiten.....	70
Tabelle 10: Auswahl von rec-Elementen.....	76
Tabelle 11: Weitere R-Pakete zur Datenaufbereitung	78
Tabelle 12: Determinanten der Demokratiezufriedenheit (Modell m1a)	88
Tabelle 13: Determinanten der Demokratiezufriedenheit (Modell m1a)	89
Tabelle 14: Determinanten der Demokratiezufriedenheit (Modell m5a und m5b).....	90
Tabelle 15: Weitere R-Pakete zur Präsentation von Regressionsergebnissen	91
Tabelle 16: Determinanten der Wahlbeteiligung.....	97
Tabelle 17: Zuordnung visueller Eigenschaften (aesthetics – aes)	101
Tabelle 18: Zuordnung geometrischer Objekte (geometric object).....	102
Tabelle 19: Items und Fragetext	118
Tabelle 20: Deskriptive Statistiken der Vertrauensitems.....	119
Tabelle 21: Korrelationsmatrix der Vertrauensitems	120
Tabelle 22: Faktorenladungen und Kommunalitäten.....	126
Tabelle 23: Deskriptive Informationen der Vertrauensskalen.....	128

1 Einführung

Die Datenanalyse ist die Phase in einem wissenschaftlichen Forschungsprojekt, in der die verwendeten Daten beschrieben und die Hypothesen empirisch geprüft werden. Für die Datenanalyse stehen mittlerweile zahlreiche Computerprogramme zur Verfügung, die komplexe statistische Verfahren sehr schnell und zuverlässig durchführen können. In der sozialwissenschaftlichen Methodenausbildung dominieren aktuell SPSS und Stata (Munzert 2018, S. 391), aber seit einigen Jahren gewinnt das Statistikprogramm R zunehmend an Bedeutung.

Das Statistikprogramm R wurde in den 1990er Jahren von Ross Ihaka und Robert Gentleman entwickelt (Ihaka und Gentleman 1996) und orientiert sich an der Programmiersprache S und an Scheme. R ist unter der General Public Licence (GNU) veröffentlicht und damit frei zugänglich. Ein R Core Team verantwortet die Weiterentwicklung von R (Fox 2009). Bereits die Basisversion von R enthält zahlreiche statistische Analyseverfahren und Möglichkeiten der grafischen Darstellung. Bei R handelt es sich aber nicht um eine geschlossene Statistikumgebung, sondern es kann durch Pakete (sogenannte Packages) erweitert werden. Dadurch kann R für die unterschiedlichsten Aufgaben verwendet werden, nicht nur für die statistische Datenanalyse.

Entwicklung von R

Auf der R-Homepage unter <https://www.r-project.org> finden sich Informationen zu R und zum Download der aktuellen Version. R wird praktisch ausschließlich über die Befehlssprache gesteuert und gilt als ein Programm mit einer eher steilen Lernkurve (Kabacoff 2015, S. xvii; Sauer 2019, S. 17). Allerdings existieren mittlerweile einige (ebenfalls) kostenfreie Ergänzungsprogramme, die die Arbeit mit R erleichtern (z.B. RStudio). Dennoch ist eine gewisse Frustrationstoleranz erforderlich, um mit gelegentlichen Fehlermeldungen umzugehen.



Diese Lerneinheit bietet ein erstes Kennenlernen von R und RStudio. Das zweite Kapitel informiert über die Installation von R und der grafischen Benutzeroberfläche RStudio. Außerdem finden sich Informationen zu Paketen, die in R installiert werden können. Im dritten Kapitel werden erste Analysen mit R präsentiert. Im vierten Kapitel werden univariate, bivariate und multivariate Analysen auf Basis eines Beispieldatensatzes durchgeführt. Das fünfte Kapitel stellt wichtige Objekttypen in R vor und das sechste Kapitel bietet eine Einführung in die Datenmodifikation mit R. Das siebte Kapitel behandelt die Schätzung einer linearen und logistischen Regression sowie die Darstellung von Regressionstabellen und Koeffizientenplots mit dem Paket sjPlot. Im achten Kapitel wird mit ggplot2 ein Paket zur Datenvisualisierung vorgestellt. Kapitel 9 behandelt die Durchführung einer explorativen Faktorenanalyse mit R.

Inhalte dieser Lerneinheit

Die Arbeit mit R bzw. RStudio muss trainiert werden. Deshalb werden in der Moodle-Lernumgebung Quizzes und Aufgabenblätter bereitgestellt, die die Auseinandersetzung mit R bzw. RStudio fördern sollen. Zudem illustrieren Vodcasts typische Arbeitsschritte der Datenanalyse. In der Moodle-Lernumgebung finden Sie auch weitere Literaturhinweise zur Arbeit mit R.

2 Installation

Vorschau



In diesem Kapitel wird die Installation des Statistikprogramms R und der sogenannten Entwicklungsumgebung RStudio erläutert. Es handelt sich um zwei verschiedene Programme. R ist das Statistikprogramm, RStudio eine grafische Benutzeroberfläche zu R. Die gleichnamige Firma RStudio hat diese Oberfläche entwickelt, die die Arbeit mit R deutlich erleichtert. RStudio ohne das Statistikprogramm R funktioniert aber nicht. Wenn Sie mit RStudio arbeiten möchten, dann muss auch das Statistikprogramm R installiert sein. Beide Programme sind kostenlos. Das Statistikprogramm R kann durch Zusatzpakete – sogenannte Packages – erweitert werden. Deshalb beschäftigt sich der letzte Abschnitt mit diesen Zusatzpaketen.

2.1 R

Links



R ist ein Open-Source-Programm und daher auch frei (kostenlos) verfügbar. Informationen zu R finden sich auf der R-Homepage unter <http://www.r-project.org>. Das Statistikprogramm R steht für Linux, Mac und Windows zur Verfügung. Sie finden die jeweiligen Installationsdateien unter <http://www.cran.r-project.org>.

Abbildung 1 zeigt die Startseite des Comprehensive R Archive Network (CRAN). Mit CRAN wird das Server-Netzwerk bezeichnet, das in verschiedenen Ländern eine Kopie des Statistikprogramms bereithält. In der Navigation auf der linken Seite findet sich der Eintrag Manuals. Dort finden sich Einführungen in das Statistikprogramm R und ausführliche Informationen zur Installation. In den meisten Fällen sollte der Download der entsprechenden Installationsdatei (je nach Betriebssystem) und die anschließende Installation allerdings selbsterklärend sein. Bei Manderscheid (2017, S. 8-11) finden sich einige Hinweise zur R-Installation unter Windows, Linux und Mac.

Abbildung 1: The Comprehensive R Archive Network (CRAN)

The screenshot shows the CRAN website with a navigation menu on the left and a main content area. The main content area is titled 'Download and Install R' and contains the following text:

Download and Install R

Precompiled binary distributions of the base system and contributed packages. **Windows and Mac** users most likely want one of these versions of R:

- [Download R for Linux \(Debian, Fedora/Redhat, Ubuntu\)](#)
- [Download R for macOS](#)
- [Download R for Windows](#)

R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

Source Code for all Platforms

Windows and Mac users most likely want to download the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it!

- The latest release (2022-04-22, Vigorous Callisthenics) [R 4.2.0 tar.gz](#), read [what's new](#) in the latest version.
- Sources of [R alpha and beta releases](#) (daily snapshots, created only in time periods before a planned release).
- Daily snapshots of current patched and development versions are [available here](#). Please read about [new features and bug fixes](#) before filing corresponding feature requests or bug reports.
- Source code of older versions of R is [available here](#).
- Contributed extension [packages](#)

Questions About R

- If you have questions about R like how to download and install the software, or what the license terms are, please read our [answers to frequently asked questions](#) before you send an email.

What are R and CRAN?

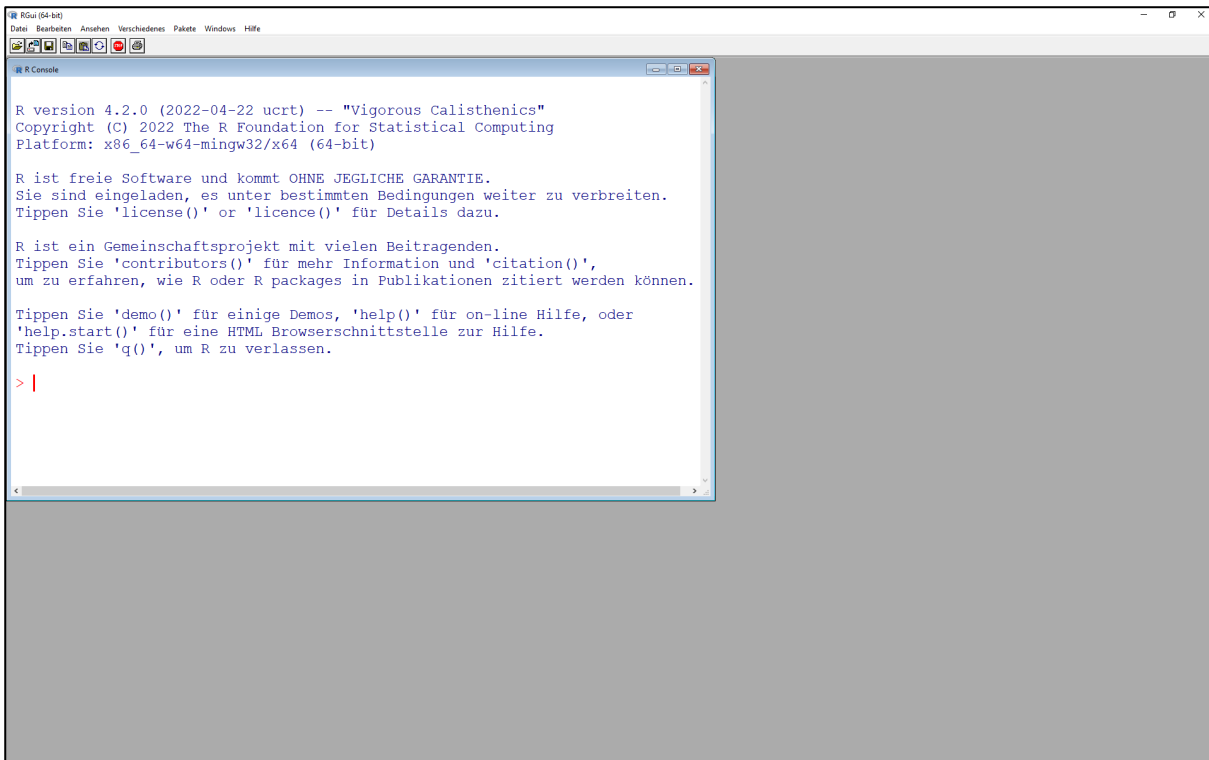
R is 'GNU S', a freely available language and environment for statistical computing and graphics which provides a wide variety of statistical and graphical techniques: linear and nonlinear modelling, statistical tests, time series analysis, classification, clustering, etc. Please consult the [R project homepage](#) for further information.

CRAN is a network of ftp and web servers around the world that store identical, up-to-date, versions of code and documentation for R. Please use the CRAN [mirror](#) nearest to you to minimize network load.

Quelle: Eigene Darstellung

Starten Sie nach der Installation das Statistikprogramm R. Je nach Installationsart können Sie zum Öffnen einfach das Programmsymbol auf dem Desktop verwenden oder das Programm über das Startmenü öffnen. Abbildung 2 zeigt den Startbildschirm von R.

Abbildung 2: Startbildschirm von R



```
R GUI (64-bit)
Datei Bearbeiten Ansehen Verschiedenes Pakete Windows Hilfe

R Console

R version 4.2.0 (2022-04-22 ucrt) -- "Vigorous Calisthenics"
Copyright (C) 2022 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R ist freie Software und kommt OHNE JEGLICHE GARANTIE.
Sie sind eingeladen, es unter bestimmten Bedingungen weiter zu verbreiten.
Tippen Sie 'license()' or 'licence()' für Details dazu.

R ist ein Gemeinschaftsprojekt mit vielen Beitragenden.
Tippen Sie 'contributors()' für mehr Information und 'citation()',
um zu erfahren, wie R oder R packages in Publikationen zitiert werden können.

Tippen Sie 'demo()' für einige Demos, 'help()' für on-line Hilfe, oder
'help.start()' für eine HTML Browserschnittstelle zur Hilfe.
Tippen Sie 'q()', um R zu verlassen.

> |
```

Quelle: Eigene Darstellung

In der Basisversion von R wird mit der sogenannten R Console gearbeitet. In diese Console werden Befehle getippt, die dann anschließend von R bearbeitet werden. Die Befehle werden nach dem Prompt-Symbol (>) eingegeben und mit der Eingabetaste (↵) abgeschlossen. Tippen Sie doch einmal $2+2$. Als Ergebnis wird 4 ausgegeben.

Für die Arbeit mit R ist die Benutzeroberfläche RStudio allerdings deutlich angenehmer. Schließen Sie R bitte vor der Installation von RStudio. Klicken Sie mit der Maus auf das rechte, obere Kreuz oder wählen Sie in der oberen Menüzeile Datei und dann Beenden.

2.2 RStudio

RStudio ist eine amerikanische Firma, die die gleichnamige grafische Benutzeroberfläche für das Statistikprogramm R entwickelt hat. Die Benutzeroberfläche RStudio steht ebenfalls als Open Source Edition zur Verfügung. Die Homepage der Firma RStudio ist unter <https://rstudio.com> zu erreichen, die Installationsdatei für die Benutzeroberfläche RStudio findet sich unter

<https://rstudio.com/products/rstudio/download/>

Wählen Sie die Open Source Version von RStudio, die für verschiedene Betriebssysteme zur Verfügung steht.

