

Markus Tausendpfund

Datenanalyse mit R. Weiterführende Verfahren

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 methoden-agnostisch 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.

Diese Lerneinheit verfolgt zwei Ziele: Zum einen sollen die Grundlagen der Arbeit mit R bzw. RStudio vorgestellt und zum anderen die Durchführung von fortgeschrittenen Analyseverfahren illustriert werden. Im Mittelpunkt steht dabei die multivariate Datenanalyse.

Die vorliegende Lerneinheit ist kein „Endprodukt“. Die regelmäßige Aktualisierung stellt eine Daueraufgabe dar. Deshalb freue ich mich sehr über alle Hinweise und Anregungen zur weiteren Verbesserung der Lerneinheit. 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 2023

Markus Tausendpfund

Inhaltsverzeichnis

1	Einführung	9
2	R und RStudio kennenlernen	10
2.1	Installation	10
2.1.1	R.....	10
2.1.2	RStudio	11
2.1.3	Pakete.....	15
2.1.4	Aktualisierungen	15
2.2	Ein erster Überblick.....	16
2.2.1	Einfache Rechenoperationen.....	16
2.2.2	Objekt erstellen	17
2.2.3	Variable erstellen	18
2.2.4	Variable mit fehlenden Werten erstellen	19
2.2.5	Grafiken mit R.....	20
2.2.6	Datensatz erstellen	22
2.2.7	Hilfe in R	23
3	Erste Analysen	25
3.1	Daten laden	25
3.2	Pakete.....	26
3.3	Daten kennenlernen	28
3.4	Datenaufbereitung	30
3.5	Univariate Datenanalyse.....	31
3.5.1	Häufigkeitstabelle.....	32
3.5.2	Lagemaße	34
3.5.3	Streuungsmaße	34
3.5.4	Formmaße.....	35
3.5.5	Erweiterte Möglichkeiten	35
3.6	Bivariate Datenanalyse.....	37
3.6.1	Kreuztabellen	37
3.6.2	Zusammenhangsmaße.....	40
3.7	Multivariate Datenanalyse.....	43
3.7.1	Lineare Regression.....	43
3.7.2	Logistische Regression	48
3.8	Grafiken.....	51

3.8.1	Säulen- und Balkendiagramm	51
3.8.2	Histogramm	53
3.8.3	Boxplots	54
4	Arbeiten mit R	56
4.1	Daten laden	56
4.2	Objekttypen	57
4.3	Vektoren, Datensätze und Listen	60
4.4	Objekttypen testen und konvertieren	60
4.5	Daten importieren	61
4.5.1	Excel	62
4.5.2	SPSS	63
4.5.3	Alternative Pakete für den Import von Datensätzen	64
5	Datenaufbereitung	65
5.1	Erforderliche Pakete	65
5.2	Daten importieren	65
5.3	Pipe-Operator	67
5.4	dplyr	69
5.4.1	count	69
5.4.2	select	69
5.4.3	rename	70
5.4.4	filter	71
5.4.5	summarise und group_by	72
5.4.6	mutate	72
5.4.7	Weitere Optionen	73
5.5	sjmisc	73
5.5.1	Variablen erkunden	73
5.5.2	Variablen kodieren	75
5.5.3	Weitere Optionen	77
5.6	sjlabelled	78
5.7	Weitere Pakete der Datenmodifikation	79
6	Multivariate Datenanalyse	80
6.1	Lineare Regression	80
6.1.1	Fragestellung, Pakete und Daten	80
6.1.2	Datenaufbereitung	81

6.1.3	Regressionsmodelle	83
6.1.4	Regressionsergebnisse präsentieren.....	88
6.2	Logistische Regression	93
6.2.1	Fragestellung, Pakete und Daten.....	93
6.2.2	Datenaufbereitung	94
6.2.3	Regressionsmodell.....	96
6.2.4	Regressionsmodelle präsentieren.....	97
6.3	Regressionsdiagnostik.....	100
6.3.1	Das Anscombe-Quartett	100
6.3.2	Grafiken zur Regressionsdiagnostik	102
6.3.3	Weitere Möglichkeiten der Regressionsdiagnostik.....	106
7	Grafiken mit ggplot2	107
7.1	Pakete und Daten.....	107
7.2	Grundlagen.....	108
7.3	Diagrammtypen.....	111
7.3.1	Liniendiagramm.....	111
7.3.2	Teilgrafiken mit Facetten.....	112
7.3.3	Säulen- und Balkendiagramme.....	114
7.3.4	Boxplot	116
7.3.5	Histogramm	117
7.3.6	Streudiagramm	118
7.3.7	Streudiagramm mit Regressionsgerade	119
7.4	Themen	121
7.5	Erweiterungen.....	122
7.5.1	patchwork.....	123
7.5.2	ggthemes.....	123
8	Explorative Faktorenanalyse	125
8.1	Pakete und Daten.....	125
8.2	Prüfung der Items.....	127
8.3	Anzahl der Faktoren	130
8.4	Faktorenanalyse und Rotation der Faktorenmatrix.....	133
8.5	Gütekriterien und Skalenkonstruktion	134
9	Literatur	137

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	51
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	92
Abbildung 21: Koeffizientenplot einer logistischen Regression mit sjPlot	99
Abbildung 22: Streudiagramme des Anscombe-Quartetts.....	101
Abbildung 23: Grafiken zur Prüfung der Modellannahmen einer linearen Regression I	103
Abbildung 24: Grafiken zur Prüfung der Modellannahmen einer linearen Regression II	104
Abbildung 25: Grafiken zur Prüfung der Modellannahmen einer linearen Regression III	105
Abbildung 26: Lebenserwartung in Deutschland	109
Abbildung 27: Lebenszufriedenheit und Bruttonsozialprodukt in Deutschland	111
Abbildung 28: Liniendiagramm.....	112
Abbildung 29: Entwicklung der Lebenserwartung in Europa	113
Abbildung 30: Bruttonsozialprodukt nach Land (Säulendiagramm).....	114
Abbildung 31: Bruttonsozialprodukt nach Land (Balkendiagramm).....	115
Abbildung 32: Lebenserwartung nach Kontinent (Boxplot)	116
Abbildung 33: Durchschnittliche Lebenserwartung in Jahren (Histogramm)	117
Abbildung 34: Lebenserwartung und Bruttonsozialprodukt (Streudiagramm I).....	118
Abbildung 35: Lebenserwartung und Bruttonsozialprodukt (Streudiagramm II).....	119
Abbildung 36: Streudiagramm mit Regressionsgerade	120
Abbildung 37: Streudiagramm mit Regressionsgerade	121
Abbildung 38: Verschiedene Themen in ggplot	122
Abbildung 39: Verschiedene Themen in ggthemes	124
Abbildung 40: Korrelationsmatrix der Vertrauensitems	128
Abbildung 41: Histogramme der Vertrauensitems.....	130
Abbildung 42: Scree plot	131
Abbildung 43: Parallelanalyse	132

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: Varianten des SPSS-Imports in R (Variable: health).....	67
Tabelle 10: Auswahl von Filter-Möglichkeiten.....	71
Tabelle 11: Auswahl von rec-Elementen.....	77
Tabelle 12: Weitere R-Pakete zur Datenaufbereitung	79
Tabelle 12: Determinanten der Demokratiezufriedenheit (Modell m1a)	89
Tabelle 13: Determinanten der Demokratiezufriedenheit (Modell m1a)	90
Tabelle 14: Determinanten der Demokratiezufriedenheit (Modell m5a und m5b).....	91
Tabelle 15: Weitere R-Pakete zur Präsentation von Regressionsergebnissen	92
Tabelle 16: Determinanten der Wahlbeteiligung.....	98
Tabelle 17: Das Anscombe-Quartett.....	100
Tabelle 18: Weitere R-Pakete zur Regressionsdiagnostik	106
Tabelle 19: Zuordnung visueller Eigenschaften (aesthetics – aes)	109
Tabelle 20: Zuordnung geometrischer Objekte (geometric object).....	110
Tabelle 21: Items und Fragetext	126
Tabelle 22: Deskriptive Statistiken der Vertrauensitems.....	127
Tabelle 23: Korrelationsmatrix der Vertrauensitems	128
Tabelle 24: Faktorenladungen und Kommunalitäten.....	134
Tabelle 25: Deskriptive Informationen der Vertrauensskalen.....	136

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 noch 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 (<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 soll Ihnen einen Einstieg in wichtige Themenfelder der fortgeschrittenen Datenanalyse bieten. In der Lerneinheit werden zum einen die Grundlagen der Arbeit mit R bzw. RStudio vorgestellt und zum anderen wird die Durchführung von fortgeschrittenen Analyseverfahren illustriert. Im Mittelpunkt steht dabei die multivariate Datenanalyse.

Inhalte dieser Lerneinheit

Das zweite Kapitel bietet ein erstes Kennenlernen von R und RStudio. Nach Hinweisen zur Installation werden erste Analysen mit R präsentiert. Im dritten Kapitel werden univariate, bivariate und multivariate Analysen auf Basis eines Beispieldatensatzes durchgeführt. Das vierte Kapitel stellt wichtige Objekttypen in R vor und das fünfte Kapitel bietet eine Einführung in die Datenmodifikation mit R. Das sechste Kapitel behandelt die Schätzung einer linearen und logistischen Regression sowie die Darstellung von Regressionstabellen und Koeffizientenplots mit dem Paket sjPlot. Im siebten Kapitel wird mit ggplot2 ein Paket zur Datenvisualisierung vorgestellt, Kapitel 8 behandelt die Durchführung einer explorativen Faktorenanalyse mit R.

Die Arbeit mit R bzw. RStudio muss trainiert werden. Deshalb werden in der Moodle-Lernumgebung Quizze 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 R und RStudio kennenlernen

Vorschau



Dieses Kapitel bietet ein erstes Kennenlernen der Programme R und RStudio. Nach der Installation der beiden Programme bietet der zweite Abschnitt einen ersten Überblick über das Arbeiten mit R bzw. RStudio.

2.1 Installation

In diesem Abschnitt 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 Firma Posit (früher: 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.

2.1.1 R

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.

Abbildung 1: The Comprehensive R Archive Network (CRAN)

The Comprehensive R Archive Network

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 (2023-04-21, Already Tomorrow) [R-4.3.0 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](#)

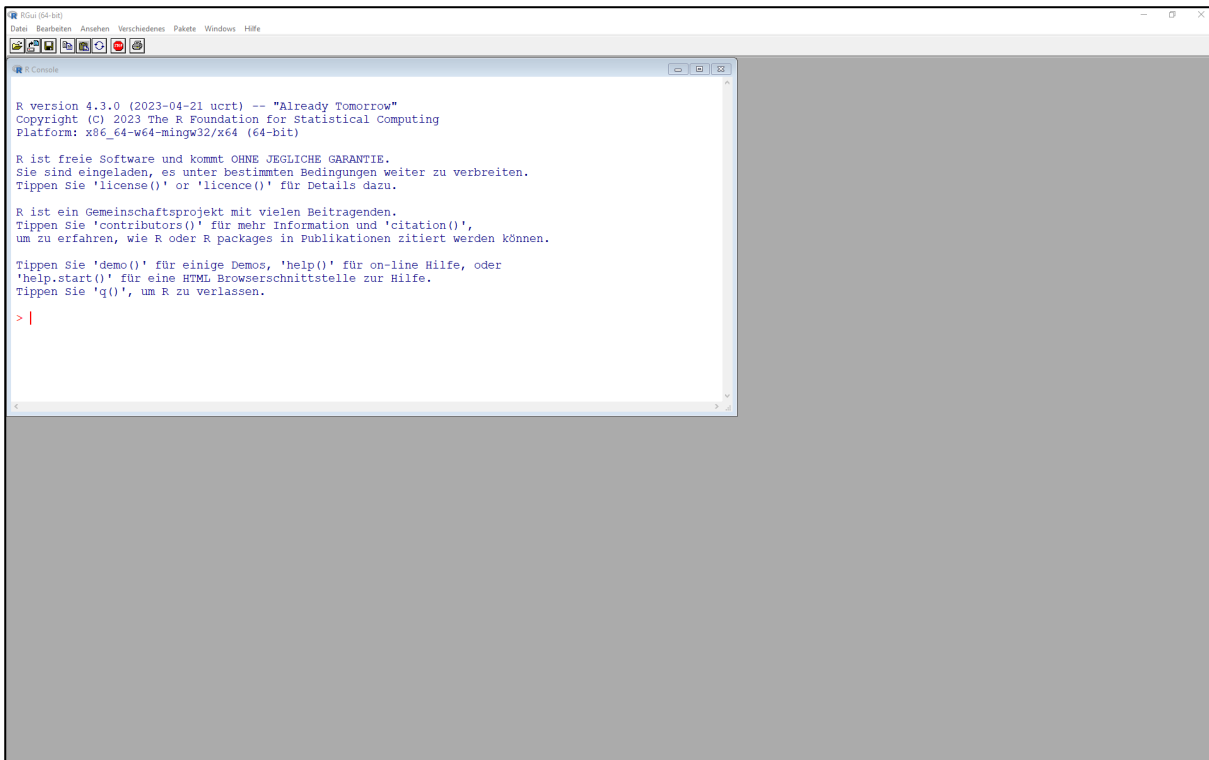
Questions About R

- 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.

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 version 4.3.0 (2023-04-21 ucrt) -- "Already Tomorrow"
Copyright (C) 2023 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.

2.1.2 RStudio

Posit (früher: 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://posit.co> zu erreichen, die Installationsdatei für die Benutzeroberfläche RStudio findet sich unter

<https://posit.co/downloads>

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