

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 Dezember 2023

Markus Tausendpfund

Inhaltsverzeichnis

Abbildungsverzeichnis	VIII
Tabellenverzeichnis	X
1 Einführung	11
2 R und RStudio kennenlernen	12
2.1 Installation	12
2.1.1 R.....	12
2.1.2 RStudio	13
2.1.3 Pakete.....	17
2.1.4 Aktualisierungen	17
2.1.5 posit Cloud: eine Alternative zur lokalen R-Installation	18
2.2 Ein erster Überblick.....	18
2.2.1 Einfache Rechenoperationen.....	18
2.2.2 Objekt erstellen	20
2.2.3 Variable erstellen	20
2.2.4 Variable mit fehlenden Werten erstellen	21
2.2.5 Grafiken mit R.....	23
2.2.6 Datensatz erstellen	25
2.2.7 Hilfe in R	26
3 Erste Analysen	28
3.1 Daten laden	28
3.2 Pakete.....	29
3.3 Daten kennenlernen	31
3.4 Datenaufbereitung	33
3.5 Univariate Datenanalyse.....	34
3.5.1 Häufigkeitstabelle.....	35
3.5.2 Lagemaße	37
3.5.3 Streuungsmaße	37
3.5.4 Formmaße.....	38
3.5.5 Erweiterte Möglichkeiten	38
3.6 Bivariate Datenanalyse	40
3.6.1 Kreuztabellen	40
3.6.2 Zusammenhangsmaße.....	43
3.7 Multivariate Datenanalyse.....	46

3.7.1	Lineare Regression	46
3.7.2	Logistische Regression	51
3.8	Grafiken	54
3.8.1	Säulen- und Balkendiagramm	54
3.8.2	Histogramm	56
3.8.3	Boxplots	57
4	Arbeiten mit R	59
4.1	Daten laden	59
4.2	Objekttypen	60
4.3	Vektoren, Datensätze und Listen	63
4.4	Objekttypen testen und konvertieren	63
4.5	Daten importieren	64
4.5.1	Excel	65
4.5.2	SPSS	66
4.5.3	CSV	67
4.5.4	Alternative Pakete	68
4.6	Saubere Skripte erstellen	69
4.7	Projekte in R	70
5	Datenaufbereitung	72
5.1	Erforderliche Pakete	72
5.2	Daten importieren	72
5.3	Pipe-Operator	74
5.4	Datenmanagement mit dplyr	76
5.4.1	count	76
5.4.2	select	76
5.4.3	rename	77
5.4.4	filter	78
5.4.5	summarise und group_by	79
5.4.6	mutate	79
5.4.7	Weitere Optionen	80
5.5	Datenaufbereitung mit sjmisc	80
5.5.1	Variablen erkunden	80
5.5.2	Variablen kodieren	82
5.5.3	Weitere Optionen	84

5.6	Labels konvertieren mit sjlabelled	85
5.7	Weitere Pakete der Datenmodifikation	86
6	Multivariate Datenanalyse	87
6.1	Lineare Regression	87
6.1.1	Fragestellung, Pakete und Daten	87
6.1.2	Datenaufbereitung	88
6.1.3	Regressionsmodelle	90
6.1.4	Regressionsergebnisse präsentieren	95
6.2	Logistische Regression	100
6.2.1	Fragestellung, Pakete und Daten	100
6.2.2	Datenaufbereitung	101
6.2.3	Regressionsmodell	103
6.2.4	Regressionsmodelle präsentieren	104
6.3	Regressionsdiagnostik	107
6.3.1	Das Anscombe-Quartett	107
6.3.2	Grafiken zur Regressionsdiagnostik	109
6.3.3	Weitere Möglichkeiten der Regressionsdiagnostik	113
7	Grafiken mit ggplot2	114
7.1	Pakete und Daten	114
7.2	Grundlagen	115
7.3	Diagrammtypen	118
7.3.1	Liniendiagramm	118
7.3.2	Teilgrafiken mit Facetten	119
7.3.3	Säulen- und Balkendiagramme	121
7.3.4	Boxplot	123
7.3.5	Histogramm	124
7.3.6	Streudiagramm	125
7.3.7	Streudiagramm mit Regressionsgerade	126
7.4	Themen	128
7.5	Erweiterungen	129
7.5.1	patchwork	130
7.5.2	ggthemes	130
8	Explorative Faktorenanalyse	132
8.1	Pakete und Daten	132

8.2	Prüfung der Items	134
8.3	Anzahl der Faktoren	137
8.4	Faktorenanalyse und Rotation der Faktorenmatrix.....	140
8.5	Gütekriterien und Skalenkonstruktion	141
9	Literatur	144

Abbildungsverzeichnis

Abbildung 1: The Comprehensive R Archive Network (CRAN)	12
Abbildung 2: Startbildschirm von R	13
Abbildung 3: RStudio mit drei Fenstern	14
Abbildung 4: RStudio mit vier Fenstern	15
Abbildung 5: Global Options bei RStudio	16
Abbildung 6: posit Cloud	18
Abbildung 7: Streudiagramm der Variablen Alter und Einkommen	24
Abbildung 8: Plotsymbole bei R	25
Abbildung 9: Laden eines R-Datensatzes	28
Abbildung 10: Geladener Datensatz	29
Abbildung 11: Installierte Packages	30
Abbildung 12: Säulendiagramme der Variable Bildung	54
Abbildung 13: Balkendiagramm der Variable Bildung	55
Abbildung 14: Säulendiagramme der Variable gesund2 (links) und gesund (rechts)	56
Abbildung 15: Histogramm des Alters	57
Abbildung 16: Boxplot der Lebenszufriedenheit in Abhängigkeit der Gesundheit	58
Abbildung 17: „Import Dataset“-Funktion in RStudio	64
Abbildung 18: Excel-Datensatz importieren	65
Abbildung 19: SPSS-Datensatz importieren	66
Abbildung 20: Datenimport mit RStudio	73
Abbildung 21: Koeffizientenplot einer linearen Regression mit sjPlot	99
Abbildung 22: Koeffizientenplot einer logistischen Regression mit sjPlot	106
Abbildung 23: Streudiagramme des Anscombe-Quartetts	108
Abbildung 24: Grafiken zur Prüfung der Modellannahmen einer linearen Regression I	110
Abbildung 25: Grafiken zur Prüfung der Modellannahmen einer linearen Regression II	111
Abbildung 26: Grafiken zur Prüfung der Modellannahmen einer linearen Regression III	112
Abbildung 27: Lebenserwartung in Deutschland	116
Abbildung 28: Lebenszufriedenheit und Bruttosozialprodukt in Deutschland	118
Abbildung 29: Liniendiagramm	119
Abbildung 30: Entwicklung der Lebenserwartung in Europa	120
Abbildung 31: Bruttosozialprodukt nach Land (Säulendiagramm)	121
Abbildung 32: Bruttosozialprodukt nach Land (Balkendiagramm)	122
Abbildung 33: Lebenserwartung nach Kontinent (Boxplot)	123
Abbildung 34: Durchschnittliche Lebenserwartung in Jahren (Histogramm)	124
Abbildung 35: Lebenserwartung und Bruttosozialprodukt (Streudiagramm I)	125
Abbildung 36: Lebenserwartung und Bruttosozialprodukt (Streudiagramm II)	126
Abbildung 37: Streudiagramm mit Regressionsgerade	127
Abbildung 38: Streudiagramm mit Regressionsgerade	128
Abbildung 39: Verschiedene Themen in ggplot	129
Abbildung 40: Verschiedene Themen in ggthemes	131
Abbildung 41: Korrelationsmatrix der Vertrauensitems	135
Abbildung 42: Histogramme der Vertrauensitems	137
Abbildung 43: Scree plot	138

Abbildung 44: Parallelanalyse 139

Tabellenverzeichnis

Tabelle 1: Beispieldaten mit Alter und Einkommen	22
Tabelle 2: Beschreibung des Datensatzes	31
Tabelle 3: Wichtige Zusammenhangsmaße bei der bivariaten Datenanalyse.....	43
Tabelle 4: Informationen zu einer linearen Regression.....	48
Tabelle 5: Ausgewählte Parameter der glm-Funktion	52
Tabelle 6: Beispieldatensatz (peanuts_r)	59
Tabelle 7: Ausgewählte Funktionen zum Testen und Konvertieren von Objekten.....	64
Tabelle 8: Alternative Pakete für den Import von Datensätzen.....	69
Tabelle 9: Varianten des SPSS-Imports in R (Variable: health).....	74
Tabelle 10: Auswahl von Filter-Möglichkeiten.....	78
Tabelle 11: Auswahl von rec-Elementen.....	84
Tabelle 12: Weitere R-Pakete zur Datenaufbereitung	86
Tabelle 13: Determinanten der Demokratiezufriedenheit (Modell m1a)	96
Tabelle 14: Determinanten der Demokratiezufriedenheit (Modell m1a)	97
Tabelle 15: Determinanten der Demokratiezufriedenheit (Modell m5a und m5b).....	98
Tabelle 16: Weitere R-Pakete zur Präsentation von Regressionsergebnissen	99
Tabelle 17: Determinanten der Wahlbeteiligung.....	105
Tabelle 18: Das Anscombe-Quartett.....	107
Tabelle 19: Weitere R-Pakete zur Regressionsdiagnostik	113
Tabelle 20: Zuordnung visueller Eigenschaften (aesthetics – aes)	116
Tabelle 21: Zuordnung geometrischer Objekte (geometric object).....	117
Tabelle 22: Items und Fragetext	133
Tabelle 23: Deskriptive Statistiken der Vertrauensitems.....	134
Tabelle 24: Korrelationsmatrix der Vertrauensitems	135
Tabelle 25: Faktorenladungen und Kommunalitäten.....	141
Tabelle 26: Deskriptive Informationen der Vertrauensskalen.....	143

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 Tests 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 ein weiterer 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 screenshot shows the CRAN website with the following content:

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

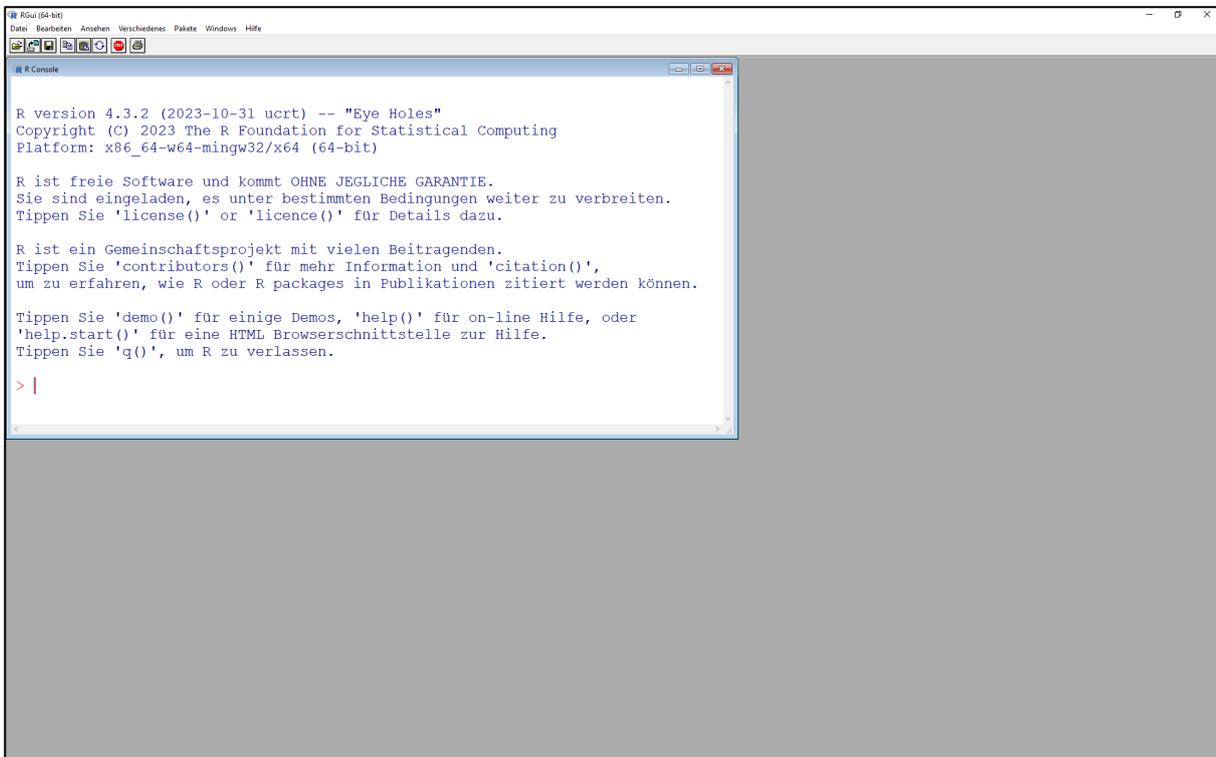
- If you have questions about R like how to download and install the software, or what the license terms are, please read our [answers to frequently asked questions](#) before you send an email.

Navigation links on the left side of the screenshot include: CRAN, Mirrors, What's new?, Search, CRAN Team, About R, R Homepage, The R Journal, Software, R Sources, R Binaries, Packages, Task Views, Other, Documentation, Manuals, FAQs, and Contributed.

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



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 (<https://posit.co>). 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.