Markus Tausendpfund

Datenanalyse mit R. Eine Einführung

Fakultät für
Kultur- und Sozialwissen-schaften

FernUniversität in Hagen

Der Inhalt dieses Studienbriefs wird gedruckt auf Recyclingpapier (80 g/m^2, 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

Tabellenverzeichnis

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.

Entwicklung von R

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.

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.

Inhalte dieser Lerneinheit

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.

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 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.

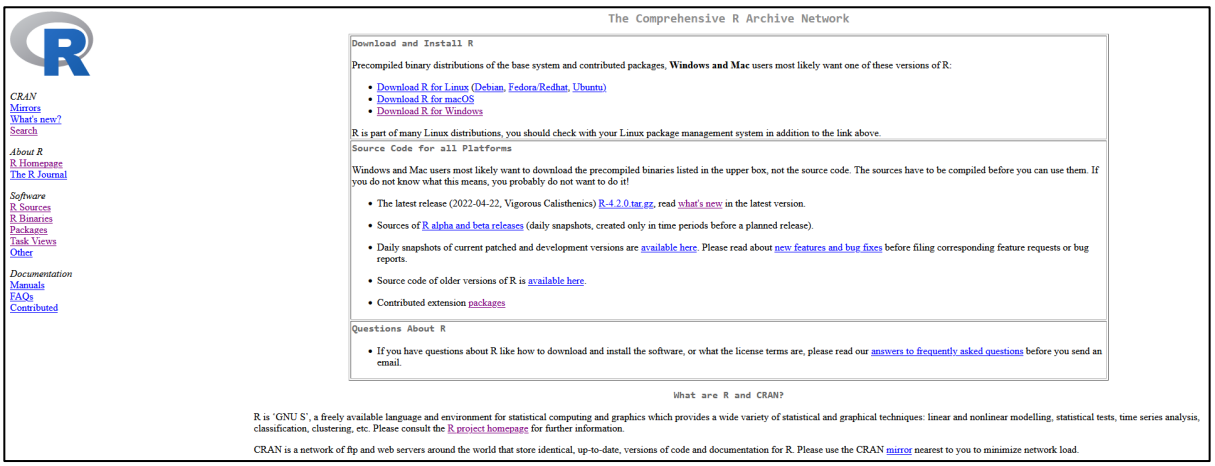
2.1 R

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.

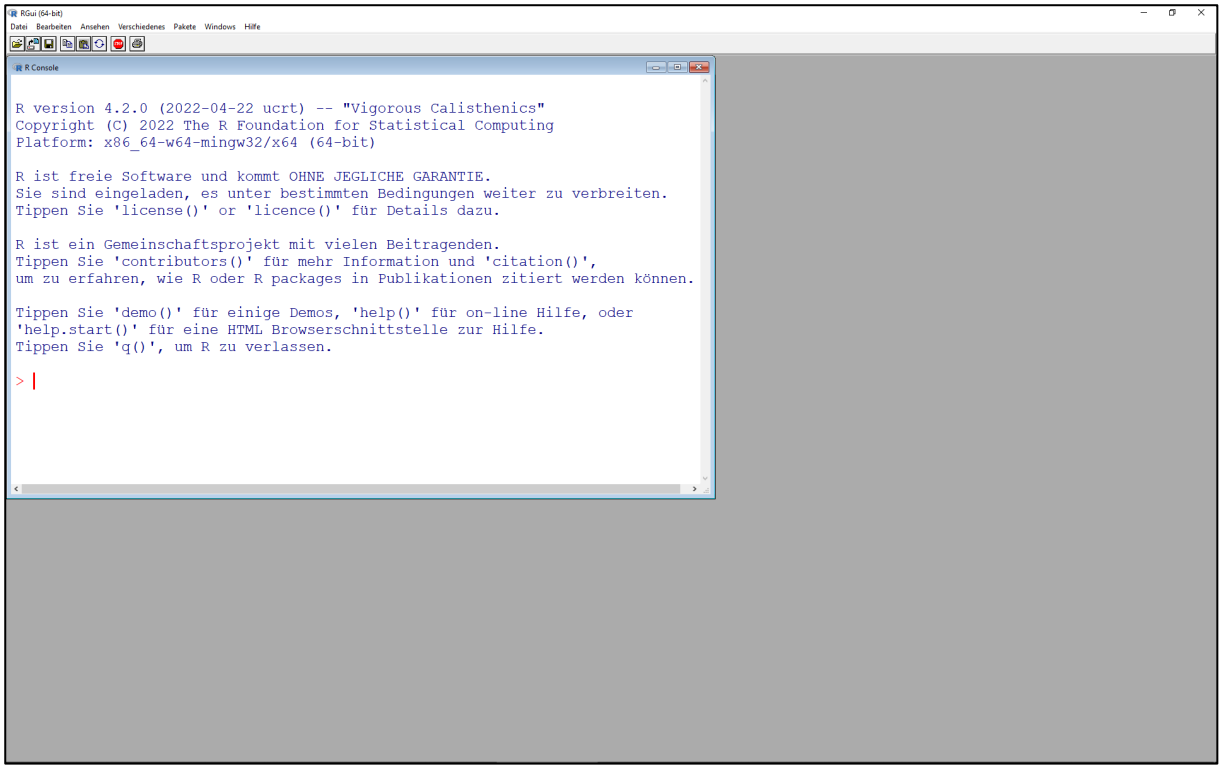
Abbildung 1: The Comprehensive R Archive Network (CRAN)



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.

2.2 RStudio

Links

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.