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 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, 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 2021

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 Versionen 16
3 R und RStudio kennenlernen 17
3.1 Einfache Rechenoperationen 17
3.2 Objekt erstellen 18
3.3 Variable erstellen 19
3.4 Variable mit fehlenden Werten erstellen 20
3.5 Grafiken mit R 21
3.6 Datensatz erstellen 23
3.7 Hilfe in R 25
4 Erste Analysen 26
4.1 Daten laden 26
4.2 Pakete installieren und laden 27
4.3 Daten kennenlernen 29
4.4 Datenaufbereitung 31
4.5 Univariate Datenanalyse 32
4.5.1 Häufigkeitstabelle 33
4.5.2 Lagemaße 35
4.5.3 Streuungsmaße 35
4.5.4 Formmaße 36
4.5.5 Erweiterte Möglichkeiten 37
4.6 Bivariate Datenanalyse 38
4.6.1 Kreuztabellen 38
4.6.2 Zusammenhangsmaße 41
4.7 Multivariate Datenanalyse 45
4.7.1 Lineare Regression 45
4.7.2 Logistische Regression 50

4.8 Grafiken ... 52
4.8.1 Säulen- und Balkendiagramm ... 52
4.8.2 Histogramm ... 54
4.8.3 Boxplots ... 55
5 Datenaufbereitung ... 57
5.1 Erforderliche Pakete ... 57
5.2 Daten importieren ... 57
5.3 Pipe-Operator ... 59
5.4 dplyr ... 60
5.4.1 count ... 60
5.4.2 select ... 60
5.4.3 rename ... 61
5.4.4 filter ... 62
5.4.5 summarise und group_by ... 63
5.4.6 mutate ... 63
5.4.7 Weitere Optionen ... 64
5.5 sjmisc ... 64
5.5.1 Variablen erkunden ... 64
5.5.2 Variablen kodieren ... 66
5.5.3 Weitere Optionen ... 68
5.6 sjlabelled ... 69
5.7 Weitere Pakete der Datenmodifikation ... 70
6 Multivariate Datenanalyse ... 71
6.1 Lineare Regression ... 71
6.1.1 Fragestellung, Pakete und Daten ... 71
6.1.2 Datenaufbereitung ... 72
6.1.3 Regressionsmodelle ... 74
6.1.4 Regressionsergebnisse präsentieren ... 79
6.2 Logistische Regression ... 84
6.2.1 Fragestellung, Pakete und Daten ... 84
6.2.2 Datenaufbereitung ... 85
6.2.3 Regressionsmodell ... 86
6.2.4 Regressionsmodelle präsentieren ... 88

7 Grafiken mit ggplot2 91
7.1 Pakete und Daten 91
7.2 Grundlagen 92
7.3 Diagrammtypen 95
7.3.1 Liniendiagramm 95
7.3.2 Teilgrafiken mit Facetten 96
7.3.3 Säulen- und Balkendiagramme 98
7.3.4 Boxplot 100
7.3.5 Histogramm 101
7.3.6 Streudiagramm 102
7.3.7 Streudiagramm mit Regressionsgerade 103
7.4 Themen 105
7.5 Erweiterungen 106
7.5.1 patchwork 107
7.5.2 ggthemes 107
8 Literatur 109

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: CRAN Task Views 15
Abbildung 7: Streudiagramm der Variablen Alter und Einkommen 22
Abbildung 8: Plotsymbole bei R 23
Abbildung 9: Laden eines R-Datensatzes 26
Abbildung 10: Geladener Datensatz 27
Abbildung 11: Installierte Packages 28
Abbildung 12: Säulendiagramme der Variable Bildung 53
Abbildung 13: Balkendiagramm der Variable Bildung 53
Abbildung 14: Säulendiagramme der Variable gesund2 (links) und gesund (rechts) 54
Abbildung 15: Histogramm des Alters 55
Abbildung 16: Boxplot der Lebenszufriedenheit in Abhängigkeit der Gesundheit 56
Abbildung 17: Datenimport mit RStudio 58
Abbildung 18: Koeffizientenplot einer linearen Regression mit sjPlot 83
Abbildung 19: Koeffizientenplot einer logistischen Regression mit sjPlot 90
Abbildung 20: Lebenserwartung in Deutschland 93
Abbildung 21: Lebenszufriedenheit und Bruttosozialprodukt in Deutschland 95
Abbildung 22: Liniendiagramm 96
Abbildung 23: Entwicklung der Lebenserwartung in Europa 97
Abbildung 24: Bruttosozialprodukt nach Land (Säulendiagramm) 98
Abbildung 25: Bruttosozialprodukt nach Land (Balkendiagramm) 99
Abbildung 26: Lebenserwartung nach Kontinent (Boxplot) 100
Abbildung 27: Durchschnittliche Lebenserwartung in Jahren (Histogramm) 101
Abbildung 28: Lebenserwartung und Bruttosozialprodukt (Streudiagramm I) 102
Abbildung 29: Lebenserwartung und Bruttosozialprodukt (Streudiagramm II) 103
Abbildung 30: Streudiagramm mit Regressionsgerade 104
Abbildung 31: Streudiagramm mit Regressionsgerade 105
Abbildung 32: Verschiedene Themen in ggplot 106
Abbildung 33: Verschiedene Themen in ggthemes 108

Tabellenverzeichnis

Tabelle 1: Beispieldaten mit Alter und Einkommen 20
Tabelle 2: Beschreibung des Datensatzes 29
Tabelle 3: Wichtige Zusammenhangsmaße bei der bivariaten Datenanalyse 41
Tabelle 4: Informationen zu einer linearen Regression 46
Tabelle 5: Ausgewählte Parameter der glm-Funktion 50
Tabelle 6: Auswahl von Filter-Möglichkeiten 62
Tabelle 7: Auswahl von rec-Elementen 68
Tabelle 8: Weitere R-Pakete zur Datenaufbereitung 70
Tabelle 9: Determinanten der Demokratiezufriedenheit (Modell m1a) 80
Tabelle 10: Determinanten der Demokratiezufriedenheit (Modell m1a) 81
Tabelle 11: Determinanten der Demokratiezufriedenheit (Modell m5a und m5b) 82
Tabelle 12: Weitere R-Pakete zur Präsentation von Regressionsergebnissen 83
Tabelle 13: Determinanten der Wahlbeteiligung 89
Tabelle 14: Zuordnung visueller Eigenschaften (aestetics – aes) 93
Tabelle 15: Zuordnung geometrischer Objekte (geometric object) 94

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 des Programms R. 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 durchgeführt. Im vierten Kapitel werden univariate, bivariate und multivariate Analysen auf Basis eines Beispieldatensatzes durchgeführt 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 R-Paket sjPlot. Im siebten Kapitel wird mit ggplot2 ein mächtiges Paket zur Datenvisualisierung vorgestellt.

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.