Multivariate Data Analysis mit R

(und etwas SPSS)

Dr. Uwe Remer

WS 2020/2021

Contents

1 Syllabus

Campus-Nr.: 273081

Link zum Kurs auf ILIAS

1.1 Seminarbeschreibung

Die professionelle Analyse empirischer Daten ist in der sozialwissenschaftlichen Forschung und in vielen Berufsfeldern von Sozialwissenschaftlern wichtig. Meist erfordern entsprechende Analysen auch multivariate Methoden. Deshalb sind Kenntnisse multivariater Analyseverfahren - deren praktische Anwendung und die Interpretation der Ergebnisse - wichtige Voraussetzungen, um einerseits empirische Texte besser verstehen (und kritisieren) sowie andererseits im Studium und ggf. später im Beruf eigene empirische Analysen durchführen zu können. Im Seminar sollen die Studierenden mit Hilfe zahlreicher Übungsaufgaben die dafür notwendigen Kompetenzen entwickeln, indem sie die Verfahren zunächst vorgestellt bekommen und sich unter Anleitung und später eigenständig die entsprechenden Lösungswege erarbeiten. Bevor multivariate Verfahren durchgeführt werden können, sind in der Regel einige Vorarbeiten an den Daten zu leisten wie z.B. Rekodierungen, die Bildung von Indizes etc. In der ersten Sitzung werden deshalb kurz die Verfahren zur Datentransformation wiederholt. Im weiteren Verlauf werden dann zentrale Analyseverfahren wie Faktorenanalyse, t-Test, Varianzanalyse und die lineare Regression angewendet und die Interpretation der Ergebnisse an Beispielen eingeübt.

1.2 R - Freie Software und "State of the Art"

> "As is evident in the content of this journal from its inception, and in books on statistical computing published recently by and for statisticians, R has come to

> dominate the development of statistical software by statisticians. The use of R among political scientists and others in the social sciences is apparently also on an upwards trajectory" — Altman 2011[1]

Fast zehn Jahre nach dieser Aussage ist R mittlerweile das Statistikpaket der Wahl für Alle, die quantitative Analysen auf höchstem Niveau und in größter Flexibilität durchführen wollen und (neben Python) der Standard in Data Science, Digital Humanities und Computational Social Science.

R ist eine Programmiersprache zur statistischen Datenanalyse und als Statistikprogramm R unter der GNU General Public License frei verfügbar. In der aktuellen Version 4.0.2 läuft es auf Windows, MacOS und Linux. Das Grundgerüst von R lässt sich durch eine Vielzahl von Paketen erweitern, mit denen sich unterschiedliche statistische Problemstellungen bearbeiten lassen.

Ebenso ist es einfach möglich, eigene Erweiterungen zu programmieren. R ist kommandozeilenbasiert (vergleichbar mit der SPSS-Syntax). Dies hat den Vorteil, dass die Anwender*In genau wissen muss, was sie überhaupt vorhat und zu welchem Zweck sie welche statistischen Tests vom Programm anfordert. Eine besondere Stärke von R liegt in den umfangreichen Möglichkeiten der Datenvisualisierung.

Zwar ist der Einstieg in R etwas schwieriger, jedoch steigt die Lernkurve nach einigen Stunden intensiver Beschäftigung (wenn man die grundsätzliche Logik verstanden hat) stark an. Mit R gibt es dann keinerlei Beschränkungen mehr, was an Auswertung möglich ist (bis auf das, was statistisch möglich ist). Zu keinem anderen Statistikprogramm gibt es eine solche Bandbreite an Unterstützung im Internet (Video-Tutorials, Mailinglisten, Blogs und Diskussionsseiten). Darüber hinaus stehen zu R und allen Paketen Handbücher als pdf's frei zur Verfügung.

Vor allem durch die freie Zugänglichkeit und Unabhängigkeit von Lizenzen, ist R im universitären Kontext besonders attraktiv. Darüber hinaus erlernen die Studierenden ein Programm, auf das sie auch nach der Zeit an der Universität (wo es in der Regel die Lizenzen für die teuren Statistikprogramme gibt), nutzen können. Die wenigsten Studierende treffen auf einen Arbeitgeber, bei dem SPSS oder Stata vorhanden ist. Mit R sind sie dennoch in der Lage statistische Analysen auf höchstem Niveau durchzuführen, ohne skeptische Verantwortliche („das geht bestimmt auch mit Excel") vom Nutzen eines teuren Statistikprogramms überzeugen zu müssen.

Ein Hinweis zum Schluss: In der R Community gibt es zwei Strömungen: "*base R*" (die reine Lehre) und "*tidyverse*" (der Sündenfall, der R massentauglich macht). Sicherlich werden einige von Ihnen über kurz oder lang gefallen an tidyverse finden. Aus didaktischen Gründen lernen Sie im Seminar aber base R (trozdem nutzen wir natürlich eine VIelzahl an Paketen und Befehlen aus dem tidyverse).

1.3 Ressourcen

R ist Open-Source – also kostenlose, freie Software. Bitte installieren Sie auf Ihrem Computer/Laptop schon Mal R. Dazu benötigen Sie zwei Dinge (bitte auch in dieser Reihenfolge installieren):

1. R: https://cran.r-project.org/bin/windows/base/
2. R-Studio: https://rstudio.com/products/rstudio/download/#download

[1] Altman, Micah et al. (2011): An Introduction to the Special Volume on Political Methodology. In: Journal of Statistical Software, Vol. 41, Nr.1. DOI: http://dx.doi.org/10.18637/jss.v042.i01

1.3.1 Online Ressourcen

- R finden Sie im Internet unter https://www.r-project.org. Über den Link CRAN (Comprehensive R Archive Network) können Sie R herunterladen.
  - Außerdem gibt es ein ausführliches FAQ: https://cran.r-project.org/doc/FAQ/R-FAQ.html
  - Und Sie können herausfinden, welche Pakete es zu welchem Zweck gibt: https://cran.r-project.org/web/views/
- Um das Syntax-, Output- und Datenmanagement mit R zu erleichtern, gibt es eine Reihe von grafischen Oberflächen für R. Wir nutzen für das Seminar RStudio: https://rstudio.org
- Tipps und eine Übersicht über die Möglichkeiten mit R findet man bei Quick-R unter https://www.statmethods.net.
- Aktuelle Entwicklungen, Informationen über neue Pakete und besonders schöne Beispiele für Analysen gibt es bei den R-Bloggers http://www.r-bloggers.com.

1.3.2 Literaturempfehlungen

1.3.2.1 R

- Adler, Joseph (2012): R in a Nutshell. Sebastopol, Calif: O'Reilly Media. UB
- Field, Andy/Miles, Jeremy/Field, Zoë (2012): Discovering Statistics Using R. London: Sage. UB
- Fox, John/Weisberg, Sanford (2018): An R companion to applied regression. Thousand Oaks, Calif: SAGE Publications. UB
- Gelman, Andrew/Hill, Jennifer/Vehtari, Aki (2020): Regression and Other Stories. Cambridge: Cambridge University Press. doi: 10.1017/9781139161879
- Gelman, Andrew/Hill, Jennifer (2012): Data analysis using regression and multilevel, hierarchical models. Cambridge: Cambridge University Press. UB
- Kabacoff, Robert (2015): R in Action. Data Analysis and Graphics with R. Shelter Island, London: Manning. UB
- Verzani, John (2014): Using R for introductory statistics. Boca Raton: Chapman and Hall. UB

1.3.2.2 Wer etwas zu SPSS sucht...

- Urban, Dieter/Mayerl, Jochen 2018: Angewandte Regressionsanalyse: Theorie, Technik und Praxis. Wiesbaden: VS Verlag für Sozialwissenschaften. pdf von SpringerLink via VPN frei verfügbar
- Backhaus, Klaus/Erichson, Bernd/Plinke, Wulf/Weiber, Rolf 2011: Multivariate Analysemethoden. Berlin: Springer. UB
- Field, Andy 2018: Discovering Statistics Using SPSS. London: Sage. UB

1.3.2.3 Forschungsdesign, Methoden und Statistik

- Agresti, Alan/Finlay, Barbara (2009): Statistical Methods for the Social Sciences. Prentice Hall. UB
- Kellstedt, Paul M./Whitten, Guy D. (2013): The Fundamentals of Political Science Research. New York: Cambridge University Press. UB
- Powner, Leanne C. (2015): Empirical research and writing. A political science student's practical guide. Los Angeles: Sage CQ Press. UB

- Wolf, Christof/Best, Henning (Hrsg.): Handbuch der sozialwissenschaftlichen Datenanalyse. Wiesbaden: VS Verlag für Sozialwissenschaften. pdf von SpringerLink via VPN frei verfügbar

1.4 Formalia

Das Seminar "Multivariate Datenanalyse mit R (und etwas SPSS)" ist Teil 2 des Moduls 282502 "Statistik-Software für Sozialwissenschaftler" im Studiengang BA Sozialwissenschaften (PO 2012 und PO 2018). Für das Modul erhalten Sie 6 ECTS.

Die Prüfungsleistung für das Modul besteht in einer lehrveranstaltungsbegleitenden Prüfung (Nr. 28252), die Sie als kurze Hausarbeit im Umfang von sechs bis acht Seiten in diesem Seminar erbringen. Abgabefrist für die Hausarbeit ist (Stand 2.11.2020) der 31.03.2021. Die Abgabe erfolgt ausschließlich digital als pdf und R Code über einen Dateibriefkasten in ILIAS.

Das Seminar hat mit 3 ECTS einen Workload von 90 Stunden, davon 21 Stunden in Präsenz. Präsenz heißt in diesem Semester: ILIAS-Lernmodule, Lernvideos, Webex-Sitzungen. Darüber hinaus sollten Sie ca. ein bis zwei Stunden fürs Lesen und eigenständiges Bearbeiten von Übungsaufgaben einplanen.

Falls Sie einen Termin in der Sprechstunde möchten, vereinbaren wir einen Termin per Mail. Die Sprechstunde findet dann über Webex statt: https://unistuttgart.webex.com/meet/uwe.remer

1.5 Sitzungen

1.5.1 02.11.2020 Sitzung 1 - Konstituierende Sitzung

1.5.2 09.11.2020 Sitzung 2 - Erste Schritte mit R

1.5.3 16.11.2020 Sitzung 3 - R Workflow

- Daten einlesen
- Daten und Code handling (R Projekte, Workingdirectory, Ordnerstruktur)
- Code schreiben (Variablenanme, Einrückungen, Kommentare etc.)
- R Markdown

1.5.4 23.11.2020 Sitzung 4 - Wiederholung: recodieren, transformieren, Indices

1.5.5 30.11.2020 Sitzung 5 - Wiederholung: uni- und bivariate Statistik (aufbereiten und ausgeben)

1.5.6 07.12.2020 Sitzung 6 - Mittelwertvergleich (t-Test)

1.5.7 14.12.2020 Sitzung 7 - Varianzanalyse (ANOVA)

1.5.8 21.12.2020 Sitzung 8 - Explorative Faktorenanalyse und Reliabilität

1.5.9 11.01.2021 Sitzung 9 - Regressionsanalyse I, Einführung (bivariater Fall)

1.5.10 18.01.2021 Sitzung 10 - Regressionsanalyse II, multivariate Regression

1.5.11 25.01.2021 Sitzung 11 - Regressionsanalyse III, Kausalität und Drittvariablenkontrolle

1.5.12 01.02.2021 Sitzung 12 - Regressionsanalyse IV, Regressionsannahmen und -diagnostik

1.5.13 08.02.2021 Sitzung 13 - Regressionsanalyse V, Moderator- und Mediator-Effekte