WZB

Wissenschaftszentrum Berlin
für Sozialforschung

Einführung in die Quantitative Datenanalyse

Sitzung 1: Datenanalyse im Forschungsprozess

Proseminar an der Freien Universität Berlin
24.04.2017 - Marcus Spittler

Einführung in die Quantitative Datenanalyse

Tab. 1: Regressionsanalysen des Einkommens

	Modell 1				Modell 2			
	$\hat{\beta}$	$s_{\hat{\beta}}$	B^*	t	$\hat{\beta}$	$s_{\hat{\beta}}$	B^*	t
Konstante	635	80		7,91	740	114		6,51
Westen	557	53	0,27	10,50	513	52	0,25	9,82
Männlich	199	66	0,11	3,02	215	64	0,12	3,36
verheiratet	−90	73	−0,05	−1,24	−84	71	−0,05	−1,18
Mann × verheiratet	330	90	0,19	3,66	349	87	0,21	4,00
Kinder	65	24	0,08	2,70	73	23	0,09	3,13
Bildung (Ref. HS, Lehre)								
MR, Lehre	339	52	0,20	6,58	210	53	0,13	3,98
Techn./Meister	415	90	0,12	4,60	228	91	0,07	2,51
FH	889	90	0,26	9,83	569	97	0,17	5,86
Uni	1362	70	0,54	19,53	895	91	0,36	9,82
Berufserfahrung	139	22	0,18	6,43	125	21	0,16	5,95
Berufsprestige					71	10	0,25	7,48
Deutsch					72	78	0,02	0,93
R^2	0,46				0,49			
R^2_{korr}	0,45				0,49			

Datenbasis: ALLBUS 2006; gewichtet mit Ost-West Transformationsgewicht (n=907).
Nur ganztags Erwerbstätige mit abhängiger Beschäftigung.

Wer ist der Typ da vorne?

- Marcus Spittler, M.A.
- Studium der Politikwissenschaft an der **Otto-Friedrich-Universität Bamberg** und an der **Freien Universität Berlin**.
- Gastaufenthalte an der **BGU in Minsk** und an der **Central European University** in Budapest.
- Gast im **Electoral Integrity Projekt** im Sydney.
- Seit 2015 Wissenschaftlicher Mitarbeiter am **Wissenschaftszentrum Berlin** in der Abteilung für **Demokratie und Demokratisierung**.
- Forschungsinteresse liegt v.a. in der **Wahl- und Einstellungsforschung**.

Kommunikation im Kurs

- Alle Präsentationen, Datensätze und Skripte die wir im Kurs verwenden stehen auf mspittler.gitlab.io zum Download.
- Die **Begleitlektüre** wird in der ersten Woche per email versendet und steht im Blackboard zur Verfügung.
- Ihr könnt gerne in die Sprechstunde kommen. Diese findet nach Vereinbarung statt.

Marcus Spittler
Wissenschaftszentrum Berlin (WZB)
Reichpietschufer 50
10785 Berlin
+49 30 25491 309
marcus.spittler@wzb.eu
www.wzb.eu/de/personen/marcus-spittler
mspittler.gitlab.io/

Teilnahmebedingungen, Hausaufgabe und Klausur

Hausaufgabe

- Kurze Bearbeitung einer gegebenen Aufgabenstellung.
- Bewertung möglich, kann zu 50% in die Endnote einfließen.
- Bearbeitung ist Bedingung für einen Teilnahmeschein

Die Klausur gliedert sich in:

- 1/3 Zentrale Konzepte und Begriffe (Wissen)
- 1/3 Interpretation statistischer Ergebnisse (Transfer)
- 1/3 Angewandte Datenanalyse (Software)

Kriterien:

- Vmtl. Open Book Exam
- 120 Minuten Bearbeitungszeit
- Termin: vmtl. letzte Seminar-Stunde (24.07.2017)

Geschichte der Statistik

Geschichte der Statistik

- **Ursprung der praktischen Statistik**
  - Rinderzensus in Ägypten (ca. 2500 BCE)
  - Erste Volkszählungen (Altes Testament / China (2300 BCE))
  - Ende des 18 Jhdt. - Gründung statistischer Zentralämter
  - Das Glücksspiel

- **Woher kommt der Begriff?**
  - Ursprung im lateinischen *Status* (Zustand, Staat) und im italienischen *statista* (Staatsmann)
  - Vorlesung M. Schmeitzel in Halle (1679-1747) mit dem Namen "collegium politico-statisticum"

- **Einige Vertreter**
  - Euler, Gauss, *Condorcet*, Bernoulli, **Pearson**, **Fisher**
  - Cox, Nightingale

Statistik-Software

- Wir werden mit den frei zugänglichen Open-Source Statistik-Umgebungen **R** und **R Studio** arbeiten.
- In den Sozialwissenschaften werden auch Alternativen eingesetzt. Bekannte Beispiele sind **STATA**, **SPSS**, **SAS**, aber auch **Excel.**

Was ist R ?

- Ursprünglich als Statistikumgebung **S** in den Bell Labs entwickelt
- 1992 veröffentlichten die Statistiker Ross Ihaka und Robert Gentleman **R**
- **R** besteht aus einem Kern (R Core/Base R) plus Zusatzpaketen.
- Die wichtigsten Zusatzpakete die wir nutzen werden sind das **tidyverse** und **ggplot2**
- Wo bekomme ich Hilfe?
  - CRAN / eingebaute R-Hilfe
  - stackoverflow.com
  - http://www.cookbook-r.com

R Vor- und Nachteile

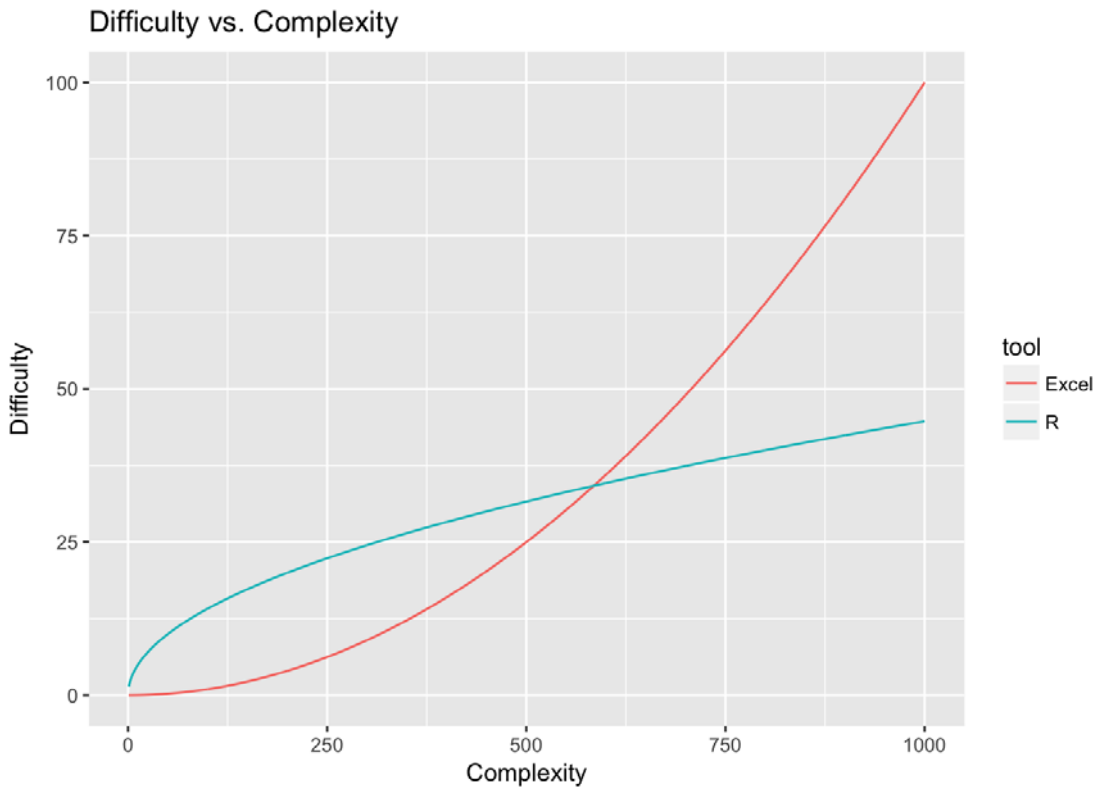
Vorteile

- Open Source
- Kostenfreie Nutzung
- Akzeptiert alle Datenformate
- Sehr viele und unterschiedliche Zusatzpakete
- Große Community

Nachteile

- Teilweise sehr steile Lernkurve

R Vor- und Nachteile



R Installation

CRAN
Mirrors
What's new?
Task Views
Search

About R
R Homepage
The R Journal

Software
R Sources
R Binaries
Packages
Other

Documentation
Manuals
FAQs
Contributed

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
- Download R for (Mac) OS X
- 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 (2016-04-14, Very, Very Secure Dishes) R-3.2.5.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.

What are R and CRAN?

R is 'GNU S', a freely available language and environment for statistical computing and graphics which provides a wide variety of statistical and graphical techniques: linear and nonlinear modelling, statistical tests, time series analysis, classification, clustering, etc. Please consult the R project homepage for further information.

CRAN is a network of ftp and web servers around the world that store identical, up-to-date, versions of code and documentation for R. Please use the CRAN mirror nearest to you to minimize network load.

Submitting to CRAN

To "submit" a package to CRAN, check that your submission meets the CRAN Repository Policy and then use the web form.

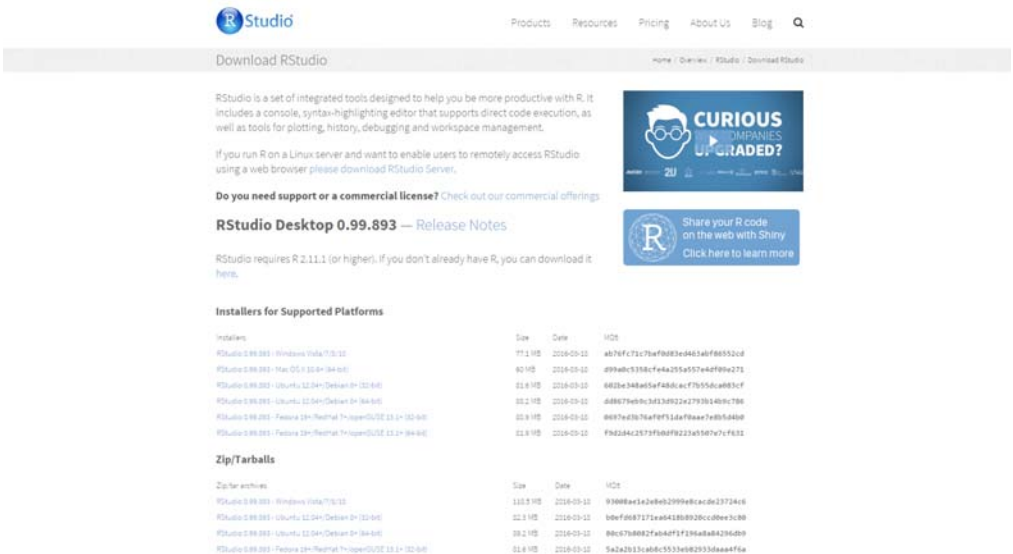
If this fails, upload to ftp://CRAN.R-project.org/incoming/ and send an email to CRAN@R-project.org following the policy. Please do not attach submissions to emails, because this will clutter up the mailboxes of half a dozen people.

Note that we generally do not accept submissions of precompiled binaries due to security reasons. All binary distribution listed above are compiled by selected maintainers, who are in charge for all

Zuerst muss R installiert werden, z.B. aus dieser Quelle:

https://cran.r-project.org/

RStudio Installation



RStudio ist die grafische Oberfläche die "über" R liegt. Mit ihr werden wir R bedienen werden.

https://www.rstudio.com/products/rstudio/download/

Ausblick

- In der Begleitlektüre finden Sie Hinweise zur Installation von R. Bitte versuchen Sie R bis zum nächsten mal zuhause zu installieren.

- Bearbeiten Sie bitte den ersten Kurs auf **DataCamp**
  https://www.datacamp.com/courses/free-introduction-to-r

- Zum Üben steht ihnen ein RStudio Server für die Zeit des Kurses zur Verfügung.