Einführung in R

Vom ÖGD für den ÖGD

Florian Beese, Jacob Schumacher

24. Oktober 2023

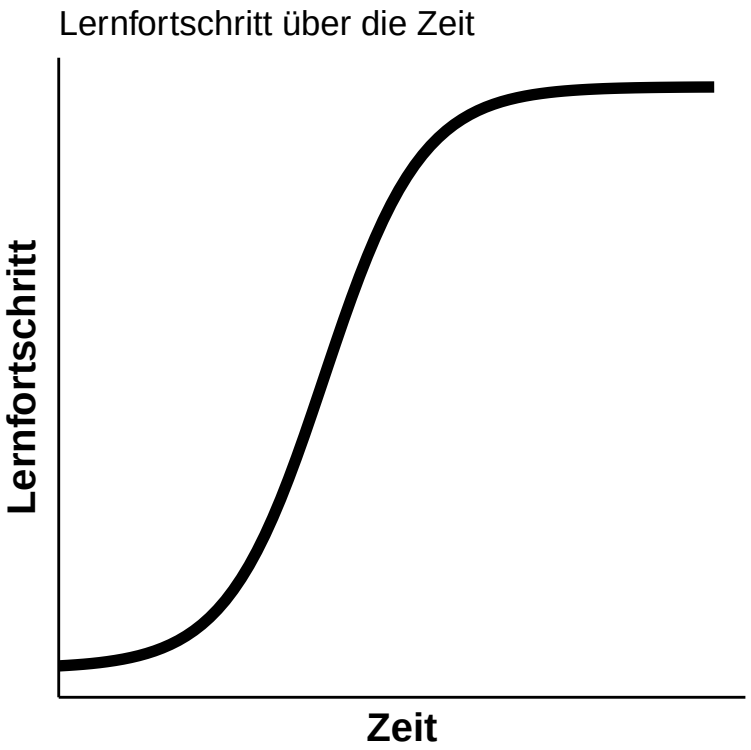
Contents

Liebe Teilnehmende,

vielen Dank für Ihre Teilnahme am Einführungskurs in die statistische Programmierung mit R der Akademie für öffentliches Gesundheitswesen.

In der Fortbildung haben wir Ihnen einen ersten Einblick in die Programmiersprache R und die Entwicklungsumgebung RStudio vorgestellt. Sie haben die ersten Schritte in der Programmiersprache `base R` und in der modernen Datenanalyse mit dem `tidyverse` gemacht. Dabei haben Sie erste Schritte im Datenimport mit `readr` oder `rio`, in der Datenmanipulation mit `dplyr`, in der Umstrukturierung von Daten mit `tidyr` oder in der Visualisierung mit `ggplot2` kennengelernt. Zudem haben Sie einen ersten Eindruck in die Erstellung von dynamischen und reproduzierbaren Berichten mit `rmarkdown` bekommen.

Der Start mit R kann mitunter sehr holprig wahrgenommen werden und wirkt im ersten Moment oft überfordernd, was für einige frustrierend sein kann. R zu nutzen bedeutet stetiges Lernen und Sie werden schnell merken, dass Sie nach dem Erlernen der Basics einen steilen Anstieg in der Lernkurve erfahren werden.



Um Sie nach diesem Einführungskurs trotzdem nicht alleine zu lassen, möchten wir Ihnen neben den Inhalten aus den 2 Kurstagen eine Auswahl weiterführender Ressourcen zur Verfügung stellen, die Ihnen bei Ihrer Reise in die Welt von R weiterhelfen können.

Gesamtüberblick

Einen guten Überblick in die (epidemiologische) Datenanalyse mit R bietet The Epidemiologist R Handbook. Diese Ressource beginnt bei den Grundlagen der Programmierung mit `base R` und führt Sie durch die klassischen Schritte der Datenanalyse.

? in R

Das Fragezeichen oder auch `help()` ist ein nützliches Tool, das bereits in R integriert ist und ermöglicht es innerhalb von RStudio Hilfeseiten zu bestimmten Funktionen aufzurufen. Hierfür können Sie einfach `?funktion` oder `help(funktion)`

Base R

Auch wenn es mittlerweile modernere Tools der Datenanalyse (bspw. `tidyverse`) gibt, ist ein grundlegendes Verständnis der Programmiersprache seiner Komponenten (Objekte, Funktionen, Packages, Datenklassen, Operatoren) essentiell, um darauf aufbauen zu können. Eine ausführliche (aber auch sehr technische) Einführung findet sich beim Comprehensive R Archive Network. Eine "quick and easy" Einführung mit weiterführenden Links findet sich bei Statistical tools for high-throughput data analysis

Data Science mit Tidyverse

Das `tidyverse` haben Sie im Kurs bereits kennengelernt. Es wurde von seinem "Urvater" Hadley Wickham entwickelt, um die Arbeit mit Daten vom Datenimport über die Analyse bis zur Visualisierung intuitiver zu gestalten. Das `tidyverse` ist ein package, das wiederum zahlreiche andere packages enthält. Jedes dieser

packages hat einen bestimmten Fokus (bspw. Datenmanipulation) und zusammen bilden diese packages im `tidyverse` einen umfassenden Werkzeugkasten, um den Umgang mit Daten zu erleichtern.

Eine Einführung in das `tidyverse` finden Sie unter https://www.tidyverse.org/ und eine praxisorientierte Einführung finden Sie kostenfreien Onlinebuch R for Data Science, das von Hadley Wickham mitverfasst wurde.

Für die einzelnen Toolkits des `tidyverse` finden Sie weitere Onlineressourcen:

- Datenimport mit readr
- Datenmanipulation mit dplyr
- Umstrukturierung von Daten mit tidyr
- Visulisierung mit ggplot2
  - für die Datenvisualisierung gibt es außerdem ein eigenes kostenfreies Buch namens ggplot 2 - Elegant Graphics for Data Analysis von Hadley Wickham

Neben den genannten `tidyverse` packages gibt es noch zahlreiche weitere. Für viele Zwecke reichen jedoch die hier genannten.

Berichterstattung mit RMarkdown

Im Kurs haben Sie eine kurze Einführung in die Erstellung dynamischer und reproduzierbarer Berichte mit `rmarkdown` erhalten. Dieses Tool ist sehr hilfreich, wenn Sie wiederkehrende Berichte standardisiert immer und immer wieder generieren möchten. Die Nutzung basiert auf der `markdown` Syntax und ermöglicht somit die Verknüpfung zwischen ausführbarem Code und einfachen Textelementen. Für die Nutzung sollten Sie jedoch die grundlegenden Prinzipien der statistischen Programmierung mit R unter Nutzung von `base R` und dem `tidyverse` verstanden haben, da sich ein `rmarkdown`-Skript sich nochmal ein wenig von einem Standard-Skript in R unterscheidet. Die Mühe lohnt sich jedoch, da Sie so Prozesse in der Berichterstattung standardisieren können. EIne ausführliche Einführung in `rmarkdown` finden Sie in dem kostenfreien Onlinebuch R Markdown: The Definitive Guide.

Cheatsheets

Cheatsheets sind kurze übersichtliche Darstellungen von packages und deren grundlegenden Funktionen. Sie geben einen kurzen Einblick in die Funktionalität eines packages und zeigen, wie die darin befindlichen Funktionen anzuwenden sind. Cheatsheets gibt es bereits für zahlreiche packages unter https://posit.co/resources/cheatsheets/. Relevante Cheatsheets, für die Kursinhalte sind folgende:

- Grundlegende Funktionen von base R
- Arbeiten mit RStudio
- Datenimport mit readr
- Datenmanipulation mit dplyr
- Umstrukturierung von Daten mit tidyr
- Visulisierung mit ggplot2
- Arbeiten mit Datums- und Zeitangaben mit lubridate
- Berichterstattung mit rmarkdown

Weitere Ressourcen

Google

R ist eine Open-source software. Zusammen mit der vielfältigen Funktionalität von R hat das zur Folge, dass es eine große Community gibt, die sich mit R und dessen problemorientierten Weiterentwicklung beschäftigt. Die Community ist mittlerweile so groß, dass es fast zu allen Problemstellungen Lösungsansätze im Internet gibt. Wenn Sie nicht weiterkommen ist daher der erste Schritt oft das einfach "googeln". Dabei finden Sie in der Regel häufig Forendiskussionen oder Dokumentationen, die mit Ihrem Problem zutun haben und die Ihnen dabei helfen Ihr Problem zu lösen. Sie haben bereits im Kurs die Erfahrung gemacht häufig auf Fehlermeldungen zu stoßen. Können Sie mit den teils kryptischen Meldungen nichts anfangen, zögern Sie nicht und kopieren Sie diese zu Google und Sie werden Ihr Problem häufig lösen können.

Stackoverflow

Wenn nicht schon geschehen, werden Sie beim Googeln häufig auf die Seite Stackoverflow stoßen. Stackoverflow ist eine Plattform, in der sich zahlreiche Nutzerinnen und Nutzer zu verschiedensten Problemen zu allen möglichen Programmiersprachen (darunter auch R) austauschen und gegenseitig helfen. Häufig finden Sie zu Ihrer Frage bereits eine Antwort. Sollte das nicht der Fall sein, können Sie selbst Fragen im Forum stellen. Aber Vorsicht: Die Community hier ist (leider) teilweise sehr kritisch und setzt voraus, dass Sie sich schon mit dem Problem auseinandergesetzt haben und selbst Versuche aufzeigen, die Sie bei der Lösungsfindung unternommen haben aber nicht zielführend waren. Eine Frage bei Stackoverflow zu stellen, bedarf mitunter einer Vorbereitung - so zum Beispiel der Erstellung eines Beispieldatensatzes, um Ihr individuelles Problem zu reproduzieren (hier ein Leitfaden für die Erstellung eines reproduzierbaren Beispiels). Nichtsdestotrotz kann es sich lohnen bei Stackoverflow vorbeizuschauen (Sie werden früher oder später höchstwahrscheinlich sowieso hiermit in Kontakt kommen).

Youtube

Wie gesagt, die Community ist groß und dementsprechend gibt es zahlreiche Nutzerinnen und Nutzer von R, die ihre Kenntnisse in Tutorials und kostenlosen Einführungskursen auf Youtube stellen. Das hat den Vorteil, dass Sie aktiv folgen und mitarbeiten können.

ChatGPT

Mit ChatGPT chat.openai.com können einfache Codeschnipsel erstellt werden. Auch für die Analyse von Fehlermeldungen ist es sehr hilfreich.