Statistische Modellbildung

Detlef Steuer steuer@hsu-hh.de

FT 2023

Struktur der Veranstaltung (11.4.2023)

Vorlesung und Übung

- Die Vorlesung ist dienstags, jeweils 14:00 Uhr, Raum 309.
- Es wird ein Skript geben. Das Skript wird jeweils (hoffentlich) spätestens Freitag Abend online auf der Seite der Veranstaltung hochgeladen. (und Illias natürlich.)
- Dort findet sich auch Material für die Übungen.
- Die Besprechung der Vorlesung und möglicher Übungsaufgaben findet in der vorgesehenen Zeit statt: Jeweils Montag 09:45-11:!5 Uhr. Tendentiell werden während der Veranstaltung ebenfalls Übungsaufgaben bearbeitet.

Prüfungsleistung

- Benotete Hausarbeit: Bearbeitung eines Datensatzes mit einer der vorgestellten Methoden.

Verwendete Werkzeuge

- Die Veranstaltung nutzt `R`, `RStudio` und `RMarkdown` bzw. `Quarto`.
- `RStudio` ist optional.

Warum RMarkdown bzw. Quarto?

- RMarkdown ist ein Dialekt von Markdown. Quarto ist ein Art Nachfolger von RMarkdown.
- In der Veranstaltung werden Dokumente in RMarkdown geschrieben und dann in PDF oder HTML umgewandelt.
- RMarkdown verbindet Markdown mit R, indem man R Programmcode und Formatierungsanweisungen mischen kann.
- Zur Veröffentlichung wird mittels pandoc der Quelltext z.B. in LaTeXübersetzt.
- Dadurch wird sogenanntes Literate Programming möglich!

Die Kombination von R und LaTeX ergibt Reproducible Research!

- Der Ursprung liegt im *literate programming* von Donald Knuth bereits um 1980.
  Programs are meant to be read by humans, and only incidentally for computers to execute. (Donald Knuth)

- Im Kern geht es bei *Reproducible Research* darum, Analysen in ihrer Gesamtheit einem zukünftigen Rezipienten an die Hand zu geben. Die Wiederholbarkeit wird erreicht indem sowohl die Daten *vollständig* vorhanden sind, als auch die Analyse *vollständig* und *vollständig dokumentiert* vorliegt. Die Analyse umfasst (idealerweise) auch die das *computational environments*, welches zur Analyse genutzt wird.

Der Mechanismus von Literate Programming

- Ausgangspunkt ist ein Dokument, welches Code und Kommentare enthält.
- Die einzelnen Codeteile werden *chunks* genannt.
- Zur Übersetzung des Dokuments wird es in seine Teile zerlegt. Die Codechunks werden in einer `R` Sitzung ausgeführt.
- Die einzelnen Teile werden in der richtigen Reihenfolge wieder zusammengesetzt, die Codechunks aber jeweils durch das Ergebnis ihrer Auswertung ersetzt.
- In `R` ist dafür das Paket `knitr` zuständig.
- Wir sehen das alles noch ausführlich in Aktion!

Warum Reproducible Research?

Non-reproducible single occurrences are of no significance to science. (Karl Popper)

- Im Jahr 2012 wurde eine Studie von Begley und Ellis in Nature veröffentlicht, die ein Jahrzehnt Forschung untersuchte. Diese Studie fand, dass 47 von 53 Papieren zu medizinischer Krebsforschung nicht reproduzierbar waren.
- In letzter Zeit wird das Problem auch bei vielen Studien zu AI festgestellt.
- Es gibt eine wirkliche Krise der wissenschaftlichen Methode.
- In der Pharmakologie ist die Wiederholbarkeit der Schlüssel zur Zulassung neuer Medikamente.

Die Werkzeuge (standing on the shoulders of giants)

- TEX bzw. LATEX (Donald Knuth)
- pandoc (John MacFarlane)
- `R` (The R-core team)
- RStudio (Hadley Wickham, ggplot, dplyr, devtools etc.)
- `knitr` (Yihui Xie)
- SWeave (Fritz Leisch)
- git (Linus Torvalds)
- Jabref (Team aus D: Oliver Kopp (ab 2011), Stefan Kolb (ab 2015), Matthias Geiger (ab 2015), Tobias Diez (ab 2015), Christoph Schwentker (ab 2016), Linus Dietz (ab 2017), Carl Christian Snethlage (ab 2020), Dominik Voigt (ab 2020)) oder
- zotero (Ray Rosenzweig Center for History and New Media, George Mason University)

RMarkdown (bzw. Quarto), ein erstes Dokument

- RMarkdown hat ebenfalls einen Header. Dort finden sich z.B. das Zielformat, der Autor etc.
- Das Format ist das sog. YAML (yet another markup language).

```
---
title: "RMarkdown and Quarto Intro"
author: "Detlef Steuer"
date: "20 1 2023"
output: html_document
---
```

- Danach folgt in RMarkdown formatierter Klartext.
- Beschreibung der Möglichkeiten unter http://rmarkdown.rstudio.com.
- Dateiendung: .Rmd
- Oder Quarto mit Dateiendung .qmd

Der knitr-Zauber

- knitr kann nicht nur `R`, sondern auch python, Julia, sql, bash oder C Programmblöcke
- knitr verwaltet die Grafiken
- knitr cached die Ergebnisse von Rechnungen
- knitr erlaubt Verweise auf Codeblocks an anderer Stelle im Dokument
- knitr controlliert den Ablauf all der beteiligten externen Programme!

Übung 11.4.2023

Lesen Sie
https://rmarkdown.rstudio.com/articles_intro.html und
spielen Sie das dortige Beispiel durch.

Ziel der Veranstaltung

- Herausarbeiten der speziellen statistischen Herangehensweise an die Modellbildung.
- Einbeziehung der Prinzipien des Reproducible Research.
- Bewertungskriterien von konkurrierenden Modellen.
- Schwerpunkte: Multiple Regression, Nichtlineare Regression, Varianzanalyse, Diskriminanzanalyse, Saisonale Modelle, Modellauswahl, Informationsmaße

Literatur

- Hain, Statistik mit R, RRZN Hannover 2011 (über die Uni erhältlich)
- Dalgaard, Introductory statistics with R, Springer (elektronisch über die Bibliothek verfügbar)
- Faraway, Linear Models in R, Chapmann and Hall
- Ligges, Programmieren in R, Springer (elektronisch über die Bibliothek verfügbar)
- Literatur für den ersten Teil der Vorlesung, Beispiele sind dort zum Teil entnommen
- Yihui Xie (2012). knitr: A general-purpose package for dynamic report generation in R. R package version 0.6. http://CRAN.R-project.org/package=knitr
- Statistische Modellbildung, Walter Gruber, 2019
- Reichhaltige Informationen im Netz!

Was kennzeichnet Statistische Modellbildung?

- Donald Knuth hat die Datenanalyse als Kunst bezeichnet, um diese von der reinen Wissenschaft zu unterscheiden. (Programming as an art, Knuth 2008)
- Die statistische Modellbildung strebt nicht hauptsächlich danach, wissenschaftliche Wahrheiten zu entdecken, sondern danach, die vorhandenen Daten zu strukturieren und durch einfache Formel(n) die Zusammenhänge zwischen den beobachteten Variablen zu beschreiben.
- Durch diese Beschreibung wird eine Abstraktion weg vom Einzelfall geschaffen, was ermöglicht, gewinnbringend Folgerungen aus Daten zu ziehen.
- Diese Abstraktion rechtfertigt den Begriff des Modells!

Beispiel für Statistische Modellbildung

- Beispiel Maskenpflicht in der Pandemie. Der genaue physikalische Zusammenhang zwischen Virusgröße, Viruslast, Filtermaß, Passform, korrekter Sitz, etc. ist extrem schwer zu modellieren, wenn man die "Wahrheit" sucht. Zählt man jedoch einfach die Infektionen in Gruppen mit und ohne Masken, so ist eine Wirkung klar.
- Natürlich ist man hocherfreut, wenn man klar interpretierbare Modelle findet.
- Es reicht in der Statistik ein Modell zu finden, dass für eine gegebene Fragestellung gut genug ist!
- Für die kinetische Energie beim Zusammenprall zweier Fahrzeuge ist es völlig ok, keine relativistischen Effekte mit einzurechnen, auch wenn die "ganze Wahrheit" dies erfordern würde.

Kontrast zum Theoriegetriebenen Modell

- Gerade in der VWL sind Modelle, die aus einer theoretischen Begründung folgen sehr verbreitet.
- Das Modell wird in diesen Fällen ohne Daten postuliert.
- Aus Statistikersicht ist das völlig ok, solange die Empirie nicht gegen das Modell spricht. Dann hat die Theorie einen Fehler ("Schwachpunkt")!
- Hier geht es nicht zuerst um Nützlichkeit, sondern um wissenschaftliche Erkenntnis, also einen Wahrheitsanspruch.
- Wenn die Theorie der Empirie zu stark entgegensteht, verlangt der westliche Wissenschaftsbegriff (Popper) eine Korrektur der Theorie! (Beispiel Mindestlohn!)

Der Begriff des Statistischen Modells

- Herkunft aus der Mathematischen Statistik: Es ist ein Tripel aus einer Menge von Elementarereignissen, eine σ-Algebra dieser Ereignisse und einer Menge von möglichen Wahrscheinlichkeitsmaßen über diesem Mengensystem.
- Die Modellbildung versucht in gewisser Weise ein bestes Wahrscheinlichkeismaß aus der Menge der möglichen Maße zu identifizieren.
- Im Modell stecken Annahmen über den Daten generierenden Prozess.
- Sind die Daten stetig, diskret? Kennt man eine parametriesierte Verteilung oder muss man nichtparametrisch arbeiten? Habe ich Vermutungen über die Struktur des Zusammenhangs oder nicht?

Mögliche Strukturannahmen

- Das lineare Modell unterstellt metrische Ziel- und Einflussgrößen und normalverteilte, unabhängige Fehler
- Ein Modell für einen Würfelwurf die Gleichverteilung über die Würfelseiten
- Ein neuronales Netz verzichtet auf ein interpretierbares Modell zugunsten eines riesigen Parameterraumes, um eine möglichst gute Prognose zu erreichen.
- Nichtparametrische Glätter verzichten auf eine interpretierbare funktionale Beschreibung, überwinden dafür aber die Nutzung desselben Zusammenhanges über den ganzen Beobachtungsraumes.
- Von black-box Modellen spricht man, wenn auf eine Erklärung des Ergebnisses durch die Modellannahmen verzichtet wird.
- Großes Problem für ADM!

Der Prozess der Modellbildung

- “Bildung” ist nichts fertiges, sondern eine aktive Tätigkeit
- Zu Beginn der Modellbildung steht die Fragestellung, die mit dem Modell beantwortet werden soll. Im Optimalfall sogar vor der Datensammlung.
- Die Datensammlung! Wo schaue ich nach welchen Daten? Ist die Qualität der Daten gut genug für den geplanten Zweck? (Luca-App fail!)
- Die deskriptive und explorative Datenanaylse, als das Anschauen der Daten! Wichtig: Die Hypothese, welches Modell zu den Daten passen könnte, sollte nicht hier entwickelt werden, da ansonsten die p-Werte nicht zuverlässig sind. Passiert aber trotzdem, da keine anderen Daten vorhanden! Eigentliches Ziel ist ein Desastercheck.
- Statistische Inferenz

Worum es in der Vorlesung nicht geht

- ML machine learning (manche packen auch LDA da hin)
- AI artificial intelligence (auch eigentlich nur Statistik)
- Prof Gertheiss hält dieses Semester die entsprechende Vorlesung!