Seminarplan - Programmieren in den Sozialwissenschaften

Henrik-Alexander Schubert

2024-05-08

Dozent

Mein Name ist Henrik-Alexander Schubert, nennt mich bitte stets Henrik, und ich bin Doktorand am Max-Planck-institut für demografische Forschung und an der Universität Oxford. Ich habe einen Masterabschluss in Demografie und Abschlüsse in Soziologie und Politikwissenschaft. Meine Interessen sind quantiative und komputergestützte Methoden, Programmieren, Sport (vor Allem Rudern und Hockey),

Voraussetzungen

Es sind keine Programmiererfahrung vorausgesetzt, aber Erfahrungen in R, Python oder anderen Programmiersprachen sind hilfreich. Eine Bereitschaft für mindestens 4 Stunden Arbeit in der Woche (2 Stunden Seminar + 2 Stunden Vor- und Nachbereitung) ist essentiell. Programmieren kann Spaß machen, aber aller Anfang ist müßig. Desweiteren ist ein Interesse an Datenanalysen, Software, Mathematik, Logik oder Automatisierungen von Prozessen hilfreich. Technische Voraussetzungen sind der Besitz und Zugang zu einem Computer und/oder Zugang zum Datenlabor. Darüber hinaus sind zwei Fertigkeiten hilfreich und entscheidend für den Arbeitsaufwand. Erstens, gute Englischkenntnisse sind hilfreich, weil die Dokumentation, die meisten und besten Quellen und die Diskussionsforen auf Englisch sind. Außerdem sind die englischen Resourcen meistens zugänglicher. Zweitens, das Behrrschen der 10-Finger Schreibtechnik am Computer ist hilfreich. Man kann erstens zeitgleich Lesen und Schreiben, schneller und mit weniger Fehlern schreiben. Programmieren heißt vor Allem fehlerfrei den Computer Anweisungen in Schrift zu geben, wenn das Schreiben bereits viel Energie und Aufmerksamkeit auf sich zieht, dann bleibt wenig für die eigentliche Arbeit übrig.

Lernziele

Der Kurs verfolg eine Reihe von Lernzielen, die sowohl im Bereich der Kenntnisse, der Fertigkeiten als auch der Characterbildung liegen. Erstens und Vordergründig soll die Programmiersprache Python oder das Statistikprogramm **R** gelernt werden. Jedoch sind die **grundlegenden Konzepte beim Programmieren**, wie *Loops*, *Funktionen* und *Datenstrukturen*, auf viele andere Programmiersprachen übertragbar. Somit werden die Konzepte stets im Vordergrund stehen. Desweiteren wird die **Organisation von einem quantitativen Forschungsprojekt** vermittelt. In den meisten Kursen wird der Fokus entweder auf das Schreiben oder die Analyse gelegt, aber die Qualität und auch die Dauer eines Projekts ist unmittelbar abhängig von strukturellen Entscheidungen bei der Ordnerstruktur, Datenaufbereitung, dem Forschungsdesign und der Verschriftlichung. Darüber hinaus sollen auch noch einige *Soft-skills* vermitelt werden. Erstens, eine generelle **Problemlösungskompetenz** wird erlernt. Programmieren zwingt einen zu einem strukturierten und logischenDenken. Somit verbessern sich Fähigkeiten bei der Identifikation eines Problems, Entwickeln von verschiedenen Lösungen, Abwegen von Optionen. Desweiteren wird das **strategisches Denken** geschult. Projekte, Texte und Analysen werden auf ein Ziel ausgerichtet und im Voraus geplant.

Prüfungsleistungen

- Ein kleines Forschungsprojekt in **R** oder [**Python**].

- wöchtenliche Hausaufgaben (ca. 2 Stunden Bearbeitungszeit, inklusive Wiederholung der Inhalte)

Kommunikation

- inhaltliche Fragen: Werden im Forum auf Stud.IP gestellt und kollektiv beantwortet.
- organisatorische Fragen:
  - Kursrelevant: an mich
  - Prüfungsrelevant: an das Prüfungsamt
- Beschwerden:
  - falls es strukturelle Probleme des Kurses sind oder Konflikte mit Kommilitonen: an mich
  - Konflikt mit dem Dozierenden: an die Lehrstuhlinhaberin Prof. Doblhammer oder den StuRa

Hilfsmittel

- für konkrete Befehle die Dokumentation: help([Befehlsname])
- das Internet: stackoverflow, cran-r.org, python.org, chatGPT
- textbücher:
- Kommilitonen

Ablauf

- Das Seminar ist inhaltlich aufgeteilt in 10 Themenblöcke
- Die Veranstaltungen bauen aufeinander auf und sätzen die Bearbeitung der Hausaufgaben sowie das Verständnis aus den vorherigen Wochen voraus
- Hausaufgaben sind in RMarkdown zu bearbeiten und einzureichen (Samt Codesegmenten)

1. Veranstaltung: Einleitung (Literature: Wickham: Grammar of graphics, Wickham: R for data science, Intro)

Vorbereitung: 1. Lesen der Texte. 2. Lernen der 10-Finger Schreibtechnik

- Installation von R und RStudio
- Wie sieht ein Projektordner aus
- Aufbau von Rstudio (*Console, R-script, Environments*)
- Was ist ein Programm (*Beispiel: "Hello, World", "Hello, Student"*)
- RMarkdown (*Open new file*, *Formatting*, *Write code*)
- Hans-Rosling Gedenkplot

Hausaufgabe: Erstellen Sie eine Grafik samt Code basierend auf den Daten von der Gapminder-Foundation in einem RMarkdown-Dokument. Die Grafik soll kurz interpretiert werden. Hilfsmittel: Ko-operation mit Kommilitionen, Internet, und das Video .

2. Veranstaltung: Datentypen, Objekte, Subsetting und Visualisierungen

Vorbereitung: 1. Lesen der Texte, 2. Watch the video on ...

- Assignment operator: ´<-´ (~/code/12/wer.R)
- Vektoren, Matrizen, Listen (~/code/w2/types.R, ~/code/w2/containers.R)
- Subsetting von Vektoren, Matrizen und Listen
- base R vs. tidyverse -> "Viele Wege führen nach Rom"
- Laden von Daten (*Beispiel csv. Rostock Temperaturdaten von [url]*)

Hausaufgabe: Laden sie die Wetterdaten für Rostock vom deutschen Wetterdienst. 1. Erstellen sie eine Grafik vom zeitlichen Verlauf der tagesdurchschnittstemperatur.(Liniendiagramm bzw. Zeitreihendiagram) 2. Filtern Sie die Daten von den

Sommermonaten (Juni, Juli, August) und Wintermonaten (Dezember, Januar, Februar). 3. Erstellen sie jeweils ein Histogram der Tagesdurchschnittstemperatur für den Sommer und den Winter.

3. Veranstaltung: Loops, Conditionals und Rechnungen

Vorbereitung:

- mathematische Operatoren: +, -, /, *, ^
- boolean operators: &, |, ==, !=, >=, <=, <, >
- if, if_else, if ... else
- Aufbau von Loops (Input, Body, Output)
- Verwendungsmöglichkeiten von Loops (Wiederholung, reduziert Code)
- Progressbar
- Vor- und Nachteile von Loops ()

Hausaufgabe: Schreib eine Loop welche eine Faktorisierung der Zahl 5 berechnet (!5). Teste im Anschluss ob die Zahl 10 ist.

4. Veranstaltung: Funktionen und Packages (Sources: 1. Video-CS50)

Vorbereitung: CS50 von 1:36 Stunde

- Aufbau einer Funktion: input -> body -> output (~/code/w4/make_mail.R, ~/code/w4/quadrieren.R)
- initial-values von Funktionen
- Vor- und Nachteile von Funktionen

Hausaufgabe: Schreib eine Funktion namens gerade, die testet, ob eine Zahl gerade oder ungerade ist. Wenn die Zahl ungerade ist, dann soll die Funktion FALSE ansonsten TRUE anzeigen.

5. Veranstaltung: Kausalität und Simulationen (Literature: Telling stories with data, Kapitel 2, Kosuke Imai (2021): Introduction to quantitative methods, Kapitel 1)

Vorbereitung: Read the literature

- potential outcome framework
- individueller und durchschnittlicher kausaler Effekt
- directed acyclic graphs (DAGs)
- randomized-control trials

Hausaufgabe: Zeichne einen DAG und simuliere die Daten für den Einfluss von Rauchen auf die Sterblichkeit.

6. Veranstaltung: Regression - Regression (Literatur: Travor, Hastie, : Introduction to statistical learning, Kapitel 1)

Vorbereitung:

- Geschichte der linearen Regression (Beispiel: Größe über Generationen hinweg)
- Regressionsgleichung: $y = \alpha + \beta + X + \epsilon$
- Was ist α und β
- Berechnung: Methode der kleinsten Quadrate und likelihood-Methode
- statistische Annahmen: *iid.* = independent and identically distributed observations
- Vorhersage

Hausaufgabe: Gompertz hat die Alterung in der Bevölkerung mit einer log-linearen Funktion beschrieben. Die Aussage des Gesetzes ist, dass die altersspezifischen Sterblichkeitsraten nach dem Alter 30 exponentiell steigen. Nutze die Daten von der Human Mortality Database für Schweden und berechne eine Funktion für das Jahr 1850 und 2022. Interpretiere die y-Achsenabschnitte und die Steigung. Wie lässt sich die Steigung interpretieren und wie lässt sich der y-Achsenabschnitss interpretieren. Was besagt die Änderung über die Zeit?

7. Veranstaltung: Karten und Wahrscheinlichkeit (Literatur: Kosuke Imai: Quantitative Social Science, Kapitel 6 & 7)

Vorbereitung:

- Bootstrap
- Illustration vom "Gesetz der großen Zahlen", "Asymptotische Theories" (Würfel)

Hausaufgabe: Erstellen sie eine Deutschlandkarte für die Lebenserwartung auf Kreisebene aus dem Zeitschriftenaufsatz von Rau und Schmertmann (2021).

8. Veranstaltung: Web-scraping (Literatur: Matthew Salganik: Bit-by-Bit.)

Vorbereitung:

- was ist web-scraping, API
- Beispiele für Webscraping (Facebook, Wohnungsmarkt)
- Download einer Tabelle von Wikipedia -> packages: httr2, rvest
- Analyse der Tabelle

Hausaufgabe: Erstelle eine Datenbank mit allen Angestellten des Lehrstuhls von Gabrielle Doblhammer samt Informationen zu Adresse, Sprechzeit.