Statistische Verfahren SS 2021

Projekt Kriminalitätsstatistik

Problemstellung:

Identifizierung von Faktoren, die die Kriminalität beeinflussen

Datensatz: (NCCrime.csv, Baltagi 2006)

Der Datensatz enthält Daten zur Kriminalitätsstatistik in 90 Counties von North Carolina im Jahr 1981. Im Unterschied zum Originaldatensatz ist die Zielgröße nicht die Kriminalitätsrate, sondern die absolute Zahl von Verbrechen (*crimes*). Folgende potentiell erklärende Variablen sind im Datensatz enthalten:

- prbarr – Probability of arrest: Anteil der Straftäter, die anschließend arrestiert werden
- prbpris – Probability of prison: Anteil der Straftäter, die zu einer Gefängnisstrafe verurteilt werden
- polpc – Anzahl der Polizisten pro Kopf der Bevölkerung
- density – Populationsdichte (Einwohner pro sq. Mile)
- area – Fläche des Counties
- taxpc – Pro-Kopf-Steueraufkommen
- region – Einteilung in die Regionen „west“, „central“ und „other“
- pctmin – Anteil von Minderheiten an der Gesamtbevölkerung
- pctymale – Anteil der jungen männlichen Bevölkerung (15-24 Jahre)
- wcon- wöchentlicher Lohn im Baugewerbe
- wsta – wöchentlicher Lohn der Staatsangestellten
- wser – wöchentlicher Lohn im Dienstleistungssektor
- wtrd – wöchentlicher Lohn im Handel
- wfir – wöchentlicher Lohn in Finanz, Versicherung und Immobilien

Aufgaben:

- Entwickeln Sie ein geeignetes statistisches Modell für die Zahl der Verbrechen. Berücksichtigen Sie dabei insbesondere die qualitative Einflussgröße *region* und deren mögliche Wechselwirkungen mit anderen Prädiktoren.

Simulationsaufgabe:

- Vergleichen Sie die statistischen Eigenschaften (Einhaltung der angestrebten Überdeckungswahrscheinlichkeit $1-\alpha$) der in der Vorlesung vorgestellten approximativen Konfidenzintervalle für die Koeffizienten mit denen der in R verwendeten Profile-Likelihood-Konfidenzintervallen.
- Wählen Sie sich dazu in Anlehnung an den ersten Teil ein konkretes (einfaches) Modell als wahres Modell. Wählen Sie dann zu jedem untersuchten Stichprobenumfang eine feste Designmatrix, die aus zufällig ausgewählten Zeilen der ursprünglichen Design-Matrix besteht (evtl. mit Wiederholung) und simulieren mehrfach Pseudobeobachtungen der Zielgröße basierend auf dieser Design-Matrix.