

Teil IV

Methoden zur Behandlung defizitärer Daten

6 Ausblick: Einige weitere Aspekte zur Behandlung defizitärer Daten

6.1 Messfehlerproblematik in komplexen Modellen

6.2 Fehlklassifikation

6.2.1 Univariater Fall

Ausgangssituation im binären Fall:

- ξ wahre binäre Variable
- x beobachtete binäre Variable
- Defizitätsmodell

$$P(x = 1|\xi = 1) =: \pi_{11} \quad \text{Sensitivität}$$

$$P(x = 0|\xi = 0) =: \pi_{00} \quad \text{Spezifität}$$

$$P(x = 0|\xi = 1) = 1 - \pi_{11} =: \pi_{01}$$

$$P(x = 1|\xi = 0) = 1 - \pi_{00} =: \pi_{10}$$

- Fehlklassifikationsmatrix

$$\pi = \begin{pmatrix} \pi_{00} & \pi_{01} \\ \pi_{10} & \pi_{11} \end{pmatrix}$$

Auf der Diagonale: Wahrscheinlichkeiten konkreter Klassifikationen.

Auswirkungen der Fehlklassifikation

- gesucht $P(\xi = 1) =: p$

- naiver Schätzer:

$$\hat{p}^- = \frac{1}{n} \sum_{i=1}^n x_i$$

-

$$\begin{aligned} P(x = 1) &= P(x = 1 | \xi = 1) \cdot P(\xi = 1) \\ &\quad + P(x = 1 | \xi = 0) \cdot P(\xi = 0) \\ &= \pi_{11} \cdot P(\xi = 1) + \pi_{10} \cdot P(\xi = 0) \end{aligned}$$

- Ausmaß des Bias

$$\begin{aligned}P(x = 1) - P(\xi = 1) &= \pi_{11} \cdot P(\xi = 1) + \pi_{10} \cdot P(\xi = 0) \\ &\quad - (\pi_{11} + \pi_{01}) \cdot P(\xi = -1) \\ &= \pi_{10} \cdot P(\xi = 0) - \pi_{01} \cdot P(\xi = 1)\end{aligned}$$

- $Bias = 0$, falls $P(\xi = 1)$ und $\pi_{00} = \pi_{11}$
- positiver wie negativer Bias möglich

Korrektur

- $P(x = 1)$ konsistent schätzbar
- Löse Gleichung

$$\begin{aligned}P(x = 1) &= \pi_{11} \cdot P(\xi = 1) + \pi_{10} \cdot P(\xi = 0) \\ &= \pi_{11} \cdot P(\xi = 1) + \underbrace{\pi_{10}}_{1 - \pi_{00}} \cdot (1 - P(\xi = 1))\end{aligned}$$

nach $P(\xi = 1)$ auf

- $P(\xi = 1) = \frac{P(x=1) - \pi_{10}}{\pi_{11} + \pi_{00} - 1}$
- Voraussetzungen
 - π_{11}, π_{00} bekannt
 - $\pi_{11} + \pi_{00} > 1$

Multinomialer Fall

- Fehlklassifikationswahrscheinlichkeiten

$$\pi_{ij} = P(x = i | \xi = j)$$

- Fehlklassifikationsmatrix Π (bekannt oder konsistent geschätzt)

$$\bullet \begin{pmatrix} P(x = 1) \\ P(x = 2) \\ \vdots \\ P(x = k) \end{pmatrix} = \Pi \cdot \begin{pmatrix} P(\xi = 1) \\ P(\xi = 2) \\ \vdots \\ P(\xi = k) \end{pmatrix}$$

- „Matrixmethode“

$$\begin{pmatrix} P(\xi = 1) \\ P(\xi = 2) \\ \vdots \\ P(\xi = k) \end{pmatrix} = \Pi^{-1} \cdot \begin{pmatrix} P(x = 1) \\ P(x = 2) \\ \vdots \\ P(x = k) \end{pmatrix}$$

- nicht immer zielführend, die Schätzungen mancher Komponenten müssen nicht im Intervall $[0; 1]$ liegen.

6.2.2 Erweiterungen

- Bivariater Fall, insbesondere 2×2 Tafeln
- Regressionsmodelle
 - Fehlklassifikation in binärer abhängiger Variable:
Hausmann et al (Journal of Econometrics, 1998), Neuhaus (Biometrika, 1999)
Likelihood bei „bekannter“ Fehlklassifikationsmatrix
 - Erweiterung SIMEX-Verfahren: Küchenhoff & Lesaffre (2006, Biometrics)

6.3 Fehlende Daten

6.3.1 Grundlegende Begriffe

- unit-nonresponse:
 - keine Information über Fall,
 - „non coverage“,
 - Ausschöpfungsquote
- item-nonresponse: nur Werte bestimmter Variablen fehlen, hier im folgenden
- *Complete Case Analysis*:
 - Analysiere nur die Fälle, bei denen zu allen Variablen die Ausprägungen vorliegen
 - wenn Fehlen nicht zufällig \Rightarrow Verzerrung, z.B. Zensierung

- auf alle Fälle Effizienzverlust
- eventuell bleiben nur wenige Fälle übrig
- Available Case Analysis
benutze jeweils alle Fälle, die bezüglich der jeweils betrachteten Frage vollständig sind
- Imputation
„Fülle Datenmatrix auf“
verschiedene Verfahren
- komplexe Schätzverfahren:
 - EM-Algorithmus
 - Bayesianische Methoden
 - Gewichtung (Horvitz-Thompson)

- Fehlendmechanismen bei partiell fehlendem Response Y und Kovariable X

- betrachte Indikatorvariable $\Delta_i = \begin{cases} 1 & Y_i \text{ beobachtet} \\ 0 & Y_i \text{ nicht beobachtet} \end{cases}$

- und die Wahrscheinlichkeit $P(\Delta_i = 1)$

- unterscheide die folgenden Fälle: Δ_i hängt

- α) weder von X noch von Y

- β) von X , aber nicht von Y

- γ) von Y , aber nicht von X

- δ) von Y und von X

- ab.

- bei α) spricht man von Missing completely at random MCAR
(Beobachtete Daten sind eine echte Zufallsstichprobe aller Daten)
- bei β) spricht man von Missing at random MAR
Beobachtete Daten sind in jeder bezüglich x gebildeten Klasse eine Zufallsstichprobe
- γ) und δ) Hier fehlen die Daten nicht zufällig Missing not at random (MNAR)

- Viele Verfahren benötigen MAR

- Was tun bei MCAR
 - Defizitätsprozess modellieren
 - MAR ist nicht testbar

- „Vorsichtige Analyse fehlender Daten“, Sensitivitätsanalyse
 - Ökonometrie: insbesondere Manski (2002, Partial Identification of Probability Distributions, Springer)
 - Biometrie: systematische Sensitivitätsanalyse (Vansteelandt, Goetghebeur, Kenward, Molenberghs. Statistica Sinica, 2006.)
 - Manski's „Law of Decreasing Credibility“ (insbesondere Manski (Ökonometrie): Die Glaubwürdigkeit statistischer Aussagen sinkt mit der Stärke der Annahmen, auf denen sie beruhen.
 - Gib den Anspruch auf, jede statistische Analyse müsse eine eindeutige Antwort geben
 - Betrachte die Menge *aller* mit den Daten verträglichen Modelle
 - Partielle Identifikation statt Punktidentifikation z.B. „Intervallwertige Punktschätzer“ für Parameter oft für substanzwissenschaftlich relevante Aussagen ausreichend.

- unschärfere, aber dafür glaubwürdigere Aussagen
- insbesondere wird deutlich, wie stark gewisse Annahmen die substanzwissenschaftlichen Folgerungen prägen.