Treatment-Evaluationsproblematik

Seminarvortrag
Seminar: "Statistische Herausforderungen im Umgang mit fehlenden bzw. fehlerbehafteten Daten"

Micha Fischer

23.01.2015

Betreuer: Prof. Dr. Thomas Augustin

Institut für Statistik
Ludwig-Maximilians-Universität München

Fundamentalprolem kausaler Inferenz

$$D \longrightarrow Y$$

- Treatmentindikator: $D \in \{0, 1\}$
- Outcomevariable: Y
- Treatmenteffekt für das i-te Subjekt: $\delta_i = Y_i^1 - Y_i^0$

$$S_i \begin{cases} \nearrow D_i = 1 \longrightarrow Y_i^1 \\ \searrow D_i = 0 \longrightarrow Y_i^0 \end{cases}$$

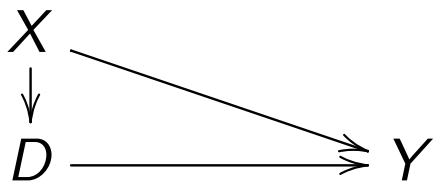
Fundamentalproblem kausaler Inferenz

Subjekt	D	Y^0	Y^1
1	1	NA	0.4
2	1	NA	0.5
3	0	0.94	NA
4	1	NA	0.45
5	0	0.78	NA
6	0	0.94	NA
7	1	NA	1.94
8	1	NA	0.5
9	0	0.66	NA
10	1	NA	0.40

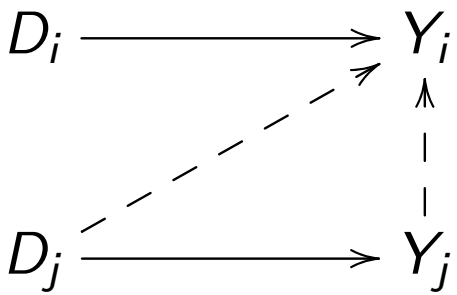
- Average Treatment Effect (ATE):
  $\mathbb{E}(Y^1 - Y^0) = \mathbb{E}(Y^1) - \mathbb{E}(Y^0)$
- Average Treatment Effect on the Treated (ATT):
  $\mathbb{E}(Y^1 - Y^0|D = 1) =$
  $\mathbb{E}(Y^1|D = 1) - \mathbb{E}(Y^0|D = 1)$

Annahmen und Probleme

- Selection Bias: $\mathbb{E}(Y^0|D=1) \neq \mathbb{E}(Y^0|D=0)$



- Stable Unit Treatment Value Assumption (SUTVA):



- Zufallsstichprobe für Inferenz nötig

Methodenüberblick

Experimentelles Design

Nicht-experimentelles Design:

- Matching-Verfahren
- Regressionsansatz
- Instrumentalvariablen
- Methoden mit Paneldaten

Experimentelles Design

Randomisierung:

- Bildung von vergleichbaren Gruppen
- Zuteilung zum Treatment erfolgt zufällig
- $f(X|D=1) = f(X|D=0) \Rightarrow \mathbb{E}(Y^0|D=1) = \mathbb{E}(Y^0|D=0)$

Vorteile:

- Selection Bias wird eliminiert, da MCAR-Situation vorliegt
- Einfache Analysemethoden mit wenigen Annahmen:

$$\widehat{ATE} = \overline{Y}^1 - \overline{Y}^0$$

Nachteil:

- Randomisierung nicht in allen Situationen möglich (z.B. Beobachtungsdaten, ethische Schwierigkeiten)
- Oft keine Zufallsstichprobe durchführbar

Nicht-Experimentelles Design

- Keine zufällige Zuweisung zum Treatment
- Es liegt MAR oder NMAR vor
- $\widehat{ATE} = \overline{Y}^1 - \overline{Y}^0$ nicht mehr erwartungstreu

Matching-Verfahren

Vorgehen:

- Nachträgliche Bildung von vergleichbaren Gruppen
- Jedem Subjekt mit Treatment wird ein Subjekt ohne Treatment zugeordnet

Annahmen:

- Beruht nur auf gemessenen Kovariablen X (MAR-Situation)
- $P(D = 1|X) \in (0, 1)$

Ziel:

- Vergleichsgruppen für die gilt: $f(X|D = 1) = f(X|D = 0)$

Matching-Verfahren

Exaktes Matching:

- Paarung von Subjekten mit gleichen Kovariablen-Ausprägungen
- Problem: Mit p dichotomen Kovariablen sind 2^p Kombinationsmöglichkeiten vorhanden

Propensity-Score Matching:

- Aggregation der Kovariablen zu einem Score
- $e(x_i) = P(D_i = 1 | X_i = x_i)$
- Paarung von Subjekten mit ähnlichem Score-Wert

Regressionsansatz

Vorgehen:

- Schätzung von getrennten Modellen:
  $\mu_1(x) = \mathbb{E}(Y_i|D = 1, X_i = x_i)$, $\mu_0(x) = \mathbb{E}(Y_i|D = 0, X_i = x_i)$
- Schätzung des kausalen Effekts durch:

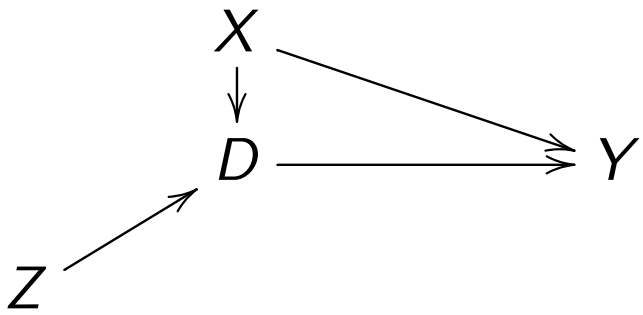
$$\widehat{ATE} = \frac{1}{N}\sum_{i=1}^{N}(\hat{\mu}_1(x_i) - \hat{\mu}_0(x_i))$$

Annahmen:

- Beruht nur auf gemessenen Kovariablen X (MAR-Situation)
- $\epsilon_i \perp D_i|X_i$

Instrumentalvariablen

- Auffinden einer Instrumentalvariable Z mit der folgenden Eigenschaft:



Vorteil:

- Auch unbeobachtete Confounder können berücksichtigt werden (NMAR-Situation)

Nachteil:

- Instrumentalvariable oft schwer zu finden

Kausalanalyse mit Paneldaten

Paneldaten:

- Mehrere Messungen (zu verschiedenen Zeitpunkten) am selben Subjekt
- Individuelle Verläufe beobachtbar

Differences-in-Differences-Schätzer:

- Vergleich von arithmetischen Mitteln vor und nach dem Treatment für beide Vergleichsgruppen

$$DID = (\bar{Y}_t^1 - \bar{Y}_{t'}^0 | D = 1) - (\bar{Y}_t^0 - \bar{Y}_{t'}^0 | D = 0)$$

- Annahme: $\mathbb{E}(Y_t^0 - Y_{t'}^0 | D = 1) = \mathbb{E}(Y_t^0 - Y_{t'}^0 | D = 0)$
- Vorteil: Ungemessene zeitkonstante Confounder werden implizit kontrolliert

Fixed-Effects-Regression

Idee:

- Betrachtung individueller Differenzen zwischen den Messzeitpunkten
- Durch personenspezifischen Intercept α_i werden zeitkonstante ungemessene Confounder kontrolliert

$$y_{it} = \alpha_i + \mathbf{x}_{it}\beta + \epsilon_{it}$$

Annahme:

- Alle Beobachtungen werden als unabhängig angesehen

Problem:

- $N - 1$ zusätzliche Parameter im Modell

Fixed-Effects-Regression

Within-Transformation:

$$y_{it} = \alpha_i + \mathbf{x}_{it}\beta + \epsilon_{it} \tag{1}$$

$$\overline{y}_i = \alpha_i + \overline{\mathbf{x}}_i\beta + \overline{\epsilon}_i \tag{2}$$

$$\overline{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{it}, \quad \overline{\mathbf{x}}_i = \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_{it}, \quad \overline{\epsilon}_i = \frac{1}{T}\sum_{t=1}^{T} \epsilon_{it}$$

Durch (1)-(2) folgt (3):

$$y_{it} - \overline{y}_i = (\mathbf{x}_{it} - \overline{\mathbf{x}}_i)\beta + (\epsilon_{it} - \overline{\epsilon}_i) \tag{3}$$

Fixed-Effects-Regression

Vorteile:

- Ungemessene zeitkonstante Confounder werden kontrolliert
- Erweiterung mit subjektspezifischen Steigungen möglich

Nachteile:

- Schätzung der Standardfehler problematisch, da $Cov(\epsilon_{it}, \epsilon_{is}) \neq 0$
- Im Allgemeinen größere Standardfehler als in gemischten Modellen
- Keine Schätzung der Parameter zeitkonstanter Kovariablen möglich

Partielle Identifikation

Gesetz der abnehmenden Glaubwürdigkeit:

- Die Glaubwürdigkeit der Inferenz verringert sich je stärker die benötigten Annahmen sind

Idee:

- Beginn mit schwachen/keinen Annahmen
- Eingrenzen des Ergebnisraumes durch sukzessive Hinzunahme von Annahmen

Partielle Identifikation

Beispiel: Fehlende Daten

Person	Y (Körpergröße)
1	1,77m
2	1,75m
3	1,65m
4	1,79m
5	2,00m
6	1,53m
7	1,71m
8	1,90m
9	1,60m
10	NA

- $\overline{Y} = \frac{1}{10}\sum_{i=1}^{9} Y_i + \frac{1}{10} * NA$

$$= 1,57 + \frac{1}{10} * NA$$

Annahmen über NA:

- NA ist MCAR
- $NA \in \mathbb{R}_+$
- $NA \in [1m; 3m]$
- $NA \in [1,53m; 2,00m]$
- $NA = 1,65m$

Partielle Identifikation bei fehlenden Daten

Formaler und abstrakter:

- Ziel: Verteilung $f(Y)$ durch Zufallsstichprobe schätzen
- $D = 0$ ist Indikator für fehlenden Wert
- Satz der totalen Wahrscheinlichkeit liefert:

$$f(Y) = f(Y|D = 1)f(D = 1) + f(Y|D = 0)f(D = 0)$$

- Die Identifikationsregion $H[f(Y)]$ wird wiefolgt definiert:

$$H[f(Y)] = [f(Y|D = 1)f(D = 1) + \gamma f(D = 0), \gamma \in \Gamma_Y]$$

- Beispielannahme: $f(Y|D = 1) = f(Y|D = 0)$ (MCAR)

Partielle Identifikation bei Treatmentevaluation

- Ziel: Menge der Verteilungen $\{f(Y^d|X=x), d \in D\}$ durch Zufallsstichprobe schätzen
- Satz der totalen Wahrscheinlichkeit liefert für $f(Y^1|X=x)$ wieder:

$$f(Y^1|X=x) = f(Y^1|X=x, D=1)f(D=1|X=x)$$
$$+f(Y^1|X=x, D=0)f(D=0)$$

- Die Identifikationsregion $H[f(Y^1|X=x)]$ wird analog definiert:

$$H[f(Y^1|X=x)] = [f(Y^1|X=x, D=1)f(D=1|X=x)$$
$$+\gamma f(D=0|X=x), \gamma \in \Gamma_Y]$$

Verbindung zur vorgestellten Methodik:

- Randomisierung liefert: $f(Y^1|X=x) = f(Y^1|X=x, D=1)$

Fazit

- Missing-Mechanismus beeinflusst die Methodenwahl stark
- Plausibilität der Modellannahmen ist nach Situation verschieden
- Nur durch die Kombinationen verschiedener Methoden werden Ergebnisse glaubwürdiger

Vielen Dank für die Aufmerksamkeit