ERLING BERGE

SOS316 ”REGESJONSANALYSE”

Kausalanalyse og seleksjonsproblem

Institutt for sosiologi og statsvitenskap,
NTNU, Trondheim

- **Litteratur**
- Breen, Richard 1996 “Regression Models. Censored, Sample Selected, or Truncated Data”, Sage University Paper: QASS 111, London, Sage
- Winship, Chrisopher, and Stephen L. Morgan 1999 “The Estimation of Causal Effects from Observational Data”, Annual Review of Sociology Vol 25: 659-707

KAUSALANALYSE

- Eksperiment
  - randomisering av påverknad (“behandling”) gir presise kausale konklusjonar om verknader (“respons”) ved signifikant skilnad i gjennomsnitt
  - kan vere umogeleg på grunn av
    - praktisk tilhøve
    - økonomiske skrankar
    - etiske vurderingar
- Kvasi-eksperiment der eksperiment er umogeleg
  - t.d. regresjonsanalyse

Eksperimetet plasserer ”case” tilfeldig i ei av to grupper:

- **BEHANDLING (T)** med observasjon
  - FØR behandling
  - ETTER behandling
- **KONTROLL (C)** med observasjon
  - FØR ikkje-behandling
  - ETTER ikkje-behandling

Modell av kausaleffektar

- Studiar av observasjonsdata brukar omgrep fra eksperimentell design
- "Påverknad/ Behandling", "Stimulus" (Treatment/ Stimulus)
- "Effekt", "Utfall" (Effect/ Outcome)

Modell av kausaleffektar:

Den "Kontrafaktiske" hypotesa for studiet av kausalitet

- Individet "i" kan i utgangspunktet tenkjast "selektert" til ei av to grupper
  - behandlingsgruppa, T, eller kontrollgruppa, C.
- Behandlinga, t, så vel som ikkje-behandling, c, kan i utgangspunktet tenkjast gitt til individ både i T- og C-gruppa
- Faktisk vil vi kunne observere t i T og c i C

Modell av kausaleffektar:
Den "Kontrafaktiske" hypotesa

- For kvart individ "i" kan ein tenkje seg fire mogelege utfall
  - **$Y_i(c,C)$** eller $Y_i(t,C)$; ved plassering i kontrollgruppe
  - $Y_i(c,T)$ eller **$Y_i(t,T)$** ; ved plassering i behandlingsgruppa
- Berre Y_i(c,gitt "i" med i C) eller Y_i(t,gitt "i" med i T) kan observerast for eit gitt individ

Modell av kausaleffektar:
Den "Kontrafaktiske" hypotesa

Mogelege utfall for person i

	Behandling: t	Ikkje beh.: c
T-gruppa	$\mathbf{Y^t_i}$ Î **T**	$Y^c_i \in T$
C-gruppa	$Y^t_i \in C$	$\mathbf{Y^c_i}$ Î **C**

Modell av kausaleffektar:
Den "Kontrafaktiske" hypotesa

- Kausaleffekten for individ i er da
- $\delta_i = Y_i(t) - Y_i(c)$
- Berre ein av desse to storleikane kan observerast for eit gitt individ

Modell av kausaleffektar:
Den "Kontrafaktiske" hypotesa

- Vi kan til dømes observere
  Y_i (c; gitt i med i C),
  men ikkje Y_i (t; gitt i med i C)
- Problemet kan seiast å vere manglande data
- I staden for individeffektar vil ein estimere gjennomsnittseffektar i heile populasjonen

Modell av kausaleffektar:

- Gjennomsnittseffektar lar seg estimere, men som regel berre med store vanskar
- Ein føresetnad er at effekten av påverknad vil vere den samme for eit gitt individ uansett kva gruppe individet er plassert i
- Dette er imidlertid ikkje sjølvsagt

Modell av kausaleffektar:

Den "Kontrafaktiske" hypotesa antar

- at endring av behandlingsgruppe for eitt individ ikkje verkar inn på utfallet for andre individ (fravær av interaksjon)
- at behandlinga, "påverknaden", faktisk er manipulerbar (t.d. kjønn er ikkje manipulerbar)

Modell av kausaleffektar:

- Ein av vanskane er at i eit utval vil den prosessen som plasserer personen ”i” i kontroll- eller behandlings-gruppa kunne verke inn på det estimerte gjennomsnittsutfallet (seleksjonsproblemet)
- I ein del tilhøve er imidlertid den interessante størrelsen gjennomsnitts-effekten for dei som får påverknaden

Modell av kausaleffektar:

- Det kan visast at det er to kjelder til feil (bias) i estimata av gjennomnsitseffekten
  - ein eksisterande skilnad mellom C- og T-gruppene
  - behandlinga verkar i prinsippet ulikt for dei som er i T-gruppa samanlikna med dei som er i C -gruppa
- For å handtere dette må vi utvikle modellar for korleis folk hamnar i C- og T-gruppene

Modell av kausaleffektar:

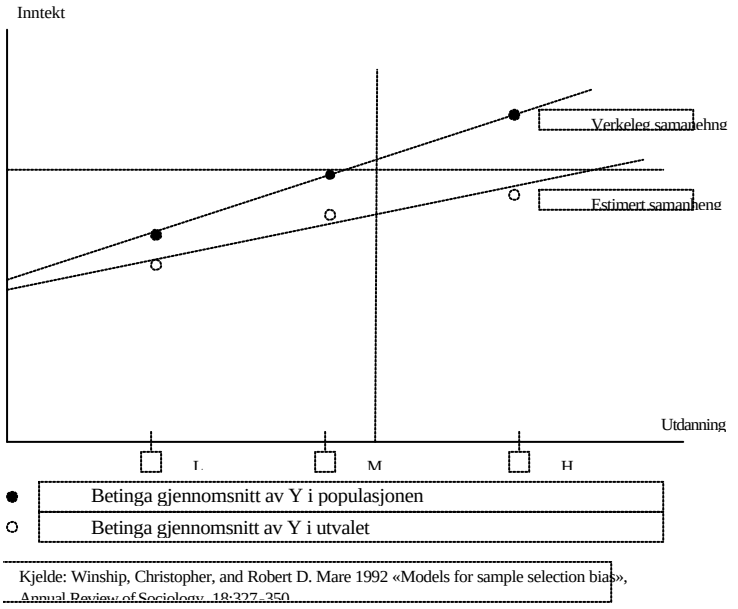
- Ein generell klasse metodar som kan nyttast til å estimere kausaleffektar er regresjonsmodellane
- Dei vil kunne "kontrollere" for observerbare skilnader mellom T- og C-gruppene, men ikkje for ulik respons på behandling

KAUSALANALYSE

- Kvasi-eksperiment (Regresjonsanalyse)
  - vil (som regel) ha store problem ved skeive utval (sensurerte, selekterte eller trunkerte utval)
  - slike utval oppstår både fordi samfunnet fungerer "selektivt" og fordi vi ikkje får fullstendige svar på spørsmål vi stiller

derfor

- seleksjonsprosessen må inkluderast i modellen eller i analysen



Kjelde: Winship, Christopher, and Robert D. Mare 1992 «Models for sample selection bias», Annual Review of Sociology, 18:327-350

Trunkering

- Ein variabel, X, vert kalla trunkert dersom vi for X<c eller for X>c ikkje veit meir enn at X<c eller X>c
- Dette kan omtalast som venstre eller høgretrunkering
- Vi kan også ha multippel trunkering, t.d. samtidig venstre og høgretrunkering

SKEIVE UTVAL OG MANGLANDE DATA I

- Sensurerte utval (eksplisitt seleksjon på Y)
  - Y er ukjent når Y har verdiar over eller under c
  - x er kjent for heile utvalet
- Selekterte utval (usystematisk seleksjon)
  - Y er ukjent dersom t.d. Z=1 og kjent når Z=0
  - x er kjent for heile utvalet

SKEIVE UTVAL OG MANGLANDE DATA II

- Trunkert utval (eksplisitt seleksjon på Y)
  - Y er ukjent når Y har verdiar over eller under c
  - x er kjent når Y er kjent
- Seleksjon på uavhengig variabel
  - Y er kjent år x har verdiar over eller under c
  - x er kjent når Y er kjent

SKEIVE UTVAL OG MANGLANDE DATA III

- Seleksjon på uavhengig variabel er uproblematisk
- Trunkerte, selekterte og sensurerte utval fører til at restleddet er korrelert med dei uavhengige variablane. Både ekstern og intern validitet er kompromitert.

SKEIVE UTVAL OG MANGLANDE DATA IV

- Datainnsamlingsprosedyrer og manglande svar kan gi opphav til trunkerte, selekterte eller sensurerte utval
  - eks: "missing" på avhengig variabel gir eit selektert utval basert på Z: gir svar eller ikkje
- I alle ikkje-tilfeldige utval er det eit potensiale for feil i konklusjonane på grunn av skeive utval

SKEIVE UTVAL OG MANGLANDE DATA V

- Ein bør i analysen ta utgangspunkt i dette og bruke modellar som korrigerer for biasen dersom ein ikkje kan argumentere for at problemet er lite.
- Løysinga er
  - 1) lage ein modell som predikerer seleksjonen
  - 2) bruke dette i ein modell som predikerer y gitt at personen er selektert

Basismodellen for sensurerte utval

E[Y| X] = Pr[Y>c | X]*E[Y | Y>c & X] +

Pr[Y<=c | X]*E[Y | Y<=c & X]

Venstretrunkering av Y ved c gir

E[Y | Y<=c & X]=c

Kan alltid transformer Y slik at c=0, dermed er den verkelege regresjonen, E[Y| X] :

- **E[Y| X] = Pr[Y>c | X]*E[Y | Y>c & X]**

Modellen i trunkerte utval

- $Y_i = E[Y_i \mid Y_i<a \ \& \ X_i] + e_i$

Det kan visast at dette er det samme som

- $Y_i = E[Y_i \mid X_i] - \sigma\lambda'_i(m) + e_i$

der $\lambda'_i(m)$ er estimert hasardraten for punktet

$m= (a - E[Y_i \mid X_i])/\sigma$

☑ parametrane i $E[Y_i \mid X_i]$ vert overestimert

Modellen kan estimerast med ML-metoden

Tostegsestimering i sensurerte utval

- Seleksjonsmodellen, Pr[Y>c | X], kan vi modellere ved probit regresjon på det sensurerte utvalet
- Utfallsmodellen, E[Y | Y>c & X], kan vi estimere på det sensurerte utvalet

- Resultata blir truverdige berre i STORE utval

Problem i tostegsmodellen

- Resultata er sensistive i høve til føresetnader om fordelinga på restledda
  - Homoskedastisitet: brot er meir alvorleg enn i vanleg OLS sidan estimata i sensurert modell da verken er konsistente eller effisiente
  - Normalfordeling

  Begge føresetnadene må testast grundig
- Problem med identifikasjon av parametrane (multikollinearitet mellom hasardraten og forklaringsvariablane, sjå t.d. Breen 1996, avsn. 2.2, s.16, likning 2.7)

Tostegsestimering med OLS

er sensitiv for

- korrelasjon mellom feil i seleksjonslikning (u_i) og feil i utfallslikning (e_i)
- korrelasjon mellom variablane i seleksjons og utfallslikninga
- graden av sensurering i utvalet (kor stor del av observasjonane av y manglar)

Konklusjon: bruk ML-estimering

Selektert eller sensurert utval ?

- Generelt er dette eit spørsmål om tolking og teoretisk fornuft
  - Når manglande observasjon av Y skuldast målemetode eller data innsamling er utvalet sensurert
  - Når manglande observasjon av Y skuldast atferd hos individa er utvalet selektert