ARBEIDSNOTAT 48/2006

Bjarne Strøm

Utvalgsseleksjon og manglende data: Noen metodemessige utfordringer

© NIFU STEP Studier av innovasjon, forskning og utdanning
Wergelandsveien 7, 0167 Oslo

Arbeidsnotat 48/2006
ISSN 1504-0887

For en presentasjon av NIFU STEPs øvrige utgivelser, se www.nifustep.no

Forord

Dette arbeidsnotatet inngår i prosjektet "Høyere utdanning og relevansen til arbeidsmarkedet", som er et Strategisk instituttprosjekt (SIP) ved NIFU STEP. Prosjektet er finansiert av Norges forskningsråd.

Arbeidsnotatet diskuterer noen metodemessige utfordringer ved håndtering av utvalgsseleksjon og manglende data, spesielt hvordan James Heckman's seleksjonsmodell kan anvendes for dette formålet. Arbeidsnotatet er skrevet på bakgrunn av et metodekurs som forfatteren holdt på NIFU STEP i 2005.

Forfatteren ønsker å takke kursdeltagerne for nyttige innspill og spesielt Jens B. Grøgaard for interessante diskusjoner og nyttige kommentarer til tidligere utkast.

Desember 2006

Petter Aasen
Direktør

Aris Kaloudis
Forskningsleder

Innhold

1 Innledning

Etterspørselen etter empiriske analyser av samfunnsvitenskapelige problemstillinger er stadig voksende. Både offentlige myndigheter, private bedrifter og andre har behov for å vite hvordan individer og institusjoner reagerer på endringer i økonomiske rammebetingelser, offentlige tiltak, og institusjonsendringer. Typiske spørsmål er: Hva skjer med gjennomstrømningen i høyere utdanning når studiefinansieringsordningen endres? Hvordan påvirkes lønnsnivået for folk med høyere utdanning av endringer i arbeidsledigheten? Presterer elever ved frittstående skoler bedre eller dårligere enn elever ved offentlige skoler? Hvor stort er lønnsgapet mellom offentlig og privat sektor?

Slik spørsmålene er formulert er det de kausale sammenhengene som etterlyses. For eksempel: Vi er interessert i finne ut om to identiske elever, der den ene tilfeldig er plassert i en friskole og den andre i en offentlig skole, presterer forskjellig. Vi ønsker altså å fjerne bidraget til prestasjonsforskjellen fra elevenes valg av skole som påvirkes av motivasjon, foreldrebakgrunn etc. Samtidig er tilgangen på data både fra spørreundersøkelser og offentlige registre økt sterkt. I dette perspektivet er det viktig å klargjøre i hvilken grad det empiriske materialet som anvendes gjør oss stand til å avdekke kausale sammenhenger og hvilke metoder som bør brukes for å avdekke slike.

Dette notatet systematiserer og gir en oversikt over noen av de metodemessige utfordringene som møter empirisk forskning i situasjoner der utvalget som observeres og analyseres er selektert på grunn av manglende respons i intervjuundersøkelser eller på grunn av individenes atferd forøvrig. Det legges særlig vekt på å presentere problemstillinger og metodeutfordringer som er relevant for NIFU STEP innenfor arbeidsmarkeds og utdanningsforskning.

Første del av notatet gir en drøfting av i hvilke situasjoner utvalgsseleksjon og manglende data er et problem og hvordan det tradisjonelt løses. Deretter behandles den grunnleggende modellen for utvalgsseleksjon (Heckman) og det diskuteres under hvilke forutsetninger korreksjon for utvalgsseleksjon i denne modellen kan gi mer pålitelige resultater. Det gis eksempler på anvendelse innenfor arbeidsmarkeds og utdanningsforskning. Spesielle problemstillinger rundt frafall i longitudinale intervjuundersøkelser behandles. Deretter

presenteres en variant av seleksjonsmodellen som er sentral innenfor effektstudier. Til slutt gis noen konkluderende bemerkninger.

2 Informasjonskilder, utvalg og seleksjon

I empirisk forskning vil problemet med manglende data og selekterte utvalg dukke opp i en rekke sammenhenger. I mange studier vil datagrunnlaget være en kombinasjon av registerbasert og intervjubasert informasjon. Den intervjubaserte informasjonen vil være et resultat av respondentenes valg av om de vil delta eller ikke, og i hvilken grad den informasjonen de leverer fra seg er pålitelig. Selv om den registerbaserte informasjonen gjerne oppfattes som objektiv og pålitelig, kan det også her være seleksjonsproblemer som gir metodemessige utfordringer når kausale sammenhenger skal avdekkes.

2.1 Informasjonskilder

Det kan være nyttig å sortere de ulike situasjonene som kan oppstå og vi starter med å skille mellom intervjubasert og registerbasert informasjon..

Intervjubasert informasjon

La oss anta at det er gjennomført en spørreundersøkelse ved at spørreskjema er sendt til et tilfeldig trukket utvalg i populasjonen som vi ønsker å studere. Et eksempel her kan være Kandidatundersøkelsen fra NIFU STEP, hvor det sendes ut spørreskjema til et tilfeldig (eller stratifisert) utvalg av de studentene som avsluttet høyere utdanning i en gitt periode. Følgende situasjoner kan da gjerne oppstå. Før det første vil noen ikke returnere skjema slik at vi mangler alle opplysninger om individet, bortsett fra de registerbaserte opplysninger som vi hadde i utgangspunktet. I engelskspråklig terminologi betegnes dette som ”unit nonresponse”. For det andre vil noen returnere skjema med noen opplysninger ubesvart (Engelsk terminologi: ”item nonresponse”). Endelig vil vi ha noen individer som fyller ut skjemaet feil.

Registerbasert informasjon

La oss dernest se på problemer som kan oppstå ved bruk av registerbasert informasjon. For det første kan vi ha registerbaserte data i utgangspunktet, men enhetene er uvillige til å delta eller unndrar seg registrering. Et eksempel på dette er gjennomføringen av de nasjonale prøvene i ungdomsskolen og videregående skole i 2005, der det kom rapporter om at flere skoler valgte å ikke gjennomføre prøvene, eller gjennomførte dem på en måte som avvek fra myndighetenes intensjon. Et annet eksempel er bruk av registrerte arbeidsledighetsdata. Registrert arbeidsledighet er basert på de individer som registrerer seg som ledige og mottar

dagpenger. Her vil opplagt individenes beslutning om å melde seg ledig eller ikke kunne variere over tid og mellom individer. Variasjoner i registrert arbeidsledighet vil derfor delvis oppstå som følge av variasjoner i meldetilbøyeligheten. Et annet eksempel er data fra skattestatistikken, der registerinformasjon fra selvangivelsesstatistikken om individenes inntekt kan gi et skjevt bilde av den reelle inntektssituasjon for noen på grunn av skatteunndragelse. Dersom omfanget av skatteunndragelse varierer over tid og mellom grupper, vil inntektsvariable basert på den registerbaserte selvangivelsesstatistikken kunne gi systematisk feilinformasjon. Vi kommer tilbake til problemstillinger knyttet til atferdsrelatert seleksjon senere i notatet.

2.2 Når er missing eller manglende respons på spørreundersøkelser et problem?

Før vi går videre er det viktig å drøfte mer spesifikt i hvilke tilfeller manglende data (missing) og seleksjon er et problem eller ikke. For det første avhenger det av om missing er systematisk knyttet til kjennetegn ved utvalgsenheten eller ikke. For det andre avhenger det av hva vi ønsker å beregne. For å illustrere poengene kan vi se på tilfellet hvor man ønsker å beregne for eksempel populasjonsandeler. Ta som eksempel at vi ønsker å beregne andelen i populasjonen av kandidater med høyere utdanning som er arbeidsledig et år etter eksamen basert på spørreundersøkelse ala NIFU STEP's kandidatundersøkelse. En relevant situasjon er at noen utdanningskategorier har mer missing på spørreskjemaene enn andre. De tradisjonelle metodene for å korrigere for dette er

i) Vekte observasjonene med inverse av frafallsandelen

ii) Imputere verdier hvis vi har informasjon om noen kjennetegn ved de som ikke har svart.

La oss se nærmere på ii). Vi kan estimere en regresjonsmodell for interessevariabelen (arbeidsmarkedsstatus) mot kjennetegn ved respondentene.

La

$y =$ arbeidsmarkedsstatus et år etter eksamen

$$y = \begin{cases} 0 \text{ hvis ikke jobb} \\ 1 \text{ hvis jobb} \end{cases}$$

Sett at vi har følgende kjennetegn på alle i utvalget (både respondenter og ikke respondenter)

$x_1 = alder$

$x_2 = kjønn$

$x_3 = utdanningstype$

Vi er altså interessert i å beregne andelen i populasjonen av uteksaminerte kandidater som er ledig et år etter eksamen.

For å illustrere poenget velger vi den enkleste varianten en kan tenke seg og estimerer en lineær sannsynlighetsmodell for y med x_1, x_2, x_3 som forklaringsvariable basert på de N respondentene. Det vil si at vi estimerer følgende relasjon med OLS:

(1) $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + restledd \quad i = 1,....N$

Vi betegner de estimerte koeffisienter fra denne relasjonen med $\widehat{\beta_0}, \widehat{\beta_1}, \widehat{\beta_2}, \widehat{\beta_3}$ og bruker simpelthen disse estimatene sammen med observerte verdier på x_1, x_2, x_3 til å beregne verdien på y for de i=N+1,……,M ikke-respondentene[1]. En implisitt antakelse bak denne prosedyren er at den samme prosessen genererer ledighetshistorien for ikke-respondentene som for respondentene. Hvis utvalget i utgangspunktet var representativt, vil da andelen ledige i utvalget (inklusive ikke-respondentene som har fått en beregnet verdi på y) være et konsistent anslag på ledigheten i populasjonen[2].

Men hva hvis responsen på spørreskjemaet er systematisk knyttet til både observerte og uobserverte kjennetegn ved individet som også påvirker ledighetshistorien? Da vil generelt imputeringene bli systematisk feil, og følgelig også de konstruerte ledighetsrater. Denne og lignende situasjoner er utgangspunkt for den mer generelle litteraturen om seleksjonsmodeller som er systematisk behandlet i den klassiske artikkelen av Heckman (1979) som vi kommer grundig tilbake til nedenfor.

[1] Selv om den lineære sannsynlighetsmodellen under standardforutsetningen om at forklaringsvariablene er ukorrelert med restleddet gir konsistente estimatorer for koeffisientene i (1) er det velkjent at OLS/lineær sannsynlighetsmodell har betydelige svakheter, for eksempel heteroskedastiske restledd og prediksjoner utenfor 0-1 intervallet. Ikke-lineære alternativer som logit og probit kan derfor være å foretrekke, men vi går ikke inn på dette i denne omgang.

[2] Imputering blir også brukt når man mangler data for enkeltobservasjoner i regresjonsmodeller. Imputering vil påvirke standardavvikene til de estimerte parametrene og inferens fra standard lineære regresjonsmodeller med imputerte data kan dermed bli feil. Cameron og Trivedi (2005), kapittel 27 gir en diskusjon av moderne metoder for regresjon med imputerte data som håndterer slike problemer.

2.2 Utvalgsseleksjon i økonometriske modeller.

Ovenfor behandlet vi tilfellet der en populasjonsandel skulle beregnes på basis av ufullstendige opplysninger gitt av en del av respondentene. I mange tilfeller vil en empirisk undersøkelse innebære at vi estimerer en multippel regresjonsmodell med en avhengig variabel og flere forklaringsvariable. Lærebokssituasjonen er at utvalget vi benytter er tilfeldig trukket fra den underliggende populasjonen. Standardprosedyren hvis noen av enhetene i utvalget mangler opplysninger om en eller flere av de relevante variable i modellen er å ekskludere observasjonene for disse enhetene ved estimeringen. Dette reduserer altså det antall observasjoner vi har til rådighet. Spørsmålet er om det er andre statistiske konsekvenser knyttet til denne datareduksjonen. Dersom frafallet (missing) er generert rent tilfeldig er det eneste problemet at estimatorene blir mindre presise fordi informasjonen i datamaterialet blir mindre.

Den mest interessante situasjonen for en økonometriker er imidlertid når frafallet (missing) ikke er tilfeldig. Vi skal nå se på hvilke problemer slik systematisk utvalgsseleksjon gir oss når vi skal estimere økonometriske modeller[3]. Her er det nyttig å skille mellom eksogen og endogen seleksjon. Eksogen seleksjon innebærer at seleksjonen skjer på basis av verdien på en eksogen variabel (forklaringvariabel), mens endogen seleksjon innebærer at seleksjonen skjer på basis av den endogene (avhengige) variabel. For å illustrere dette vil vi betrakte et eksempel der vi ønsker å estimere sammenhengen mellom logaritmen til individuell lønn, w, alder og antall år utdanning og formulerer følgende enkle ligning

$$(2)\quad \ln w_i = \beta_0 + \beta_1 Utdanning_i + \beta_2 Alder_i + \beta_3 Alder^2{}_i + u_i$$

der u er et stokastisk restledd som oppfyller standardforutsetningene: uavhengig og identisk fordelt, samt ukorrelert med høyresidevariablene i ligningen.

I. Eksogen seleksjon

Sett at datamaterialet er et tilfeldig trukket utvalg av personer over 35 år. Seleksjonen er med andre ord basert på nivået på den eksogene variabelen Alder. Så lenge modellen er den samme for alle delutvalg (her aldersgrupper) i populasjonen og vi har tilstrekkelig variasjon i

[3] Framstillingen her er basert på Woolridge (2003), kap.9.4.

den avhengige variabelen i delutvalget så vil seleksjon på basis av alder ikke gi skjevhet i estimatorene for $\beta_0, \beta_1, \beta_2, \beta_3$.

II. Endogen seleksjon

Sett nå i stedet at utvalget som kan brukes i estimeringen er bestemt av nivået på den avhengige variabelen:

Trunkering

Sett at bare individer med lønnsnivå w<NOK1000 000 er inkludert i utvalget.

Utvalget er altså ikke tilfeldig, men basert på verdien på den avhengige variabel. Dette vil gi skjeve OLS-estimatorer for parametrene i (1). Årsaken er, løst formulert, at forventningen til den avhengige variabel, betinget på forklaringsvariablene i populasjonsmodellen (1) ikke er den samme som forventet verdi betinget på w<NOK1000 000. Dette kan håndteres ved å estimere en sensurert regresjonsmodell som kan betraktes som et spesialtilfelle av mer generelle modeller for utvalgsseleksjon som behandles nedenfor[4].

[4] Se Woolridge (2003), kap.17.4 for en enkel innføring i temaet, mens Cameron og Trivedi (2005), kap.16.2, Green (2003) kap.22 og Woolridge (2002), kap 16 inneholder mer avanserte framstillinger.

3 Seleksjon på grunn av individenes atferd.

3.1 Heckman's seleksjonsmodell

Økonomer er ofte interessert i effekten av for eksempel utdanning på lønnstilbudet som et individ får. Dette danner utgangspunkt for det klassiske eksemplet for denne type utvalgsseleksjon. La oss for å illustrere anta at vi anvender en variant av lønnsmodellen ovenfor der vi ønsker å estimere en modell for lønnsnivået for populasjonen av kandidater som har avsluttet høyere utdanning et år etter endt utdanning. Datagrunnlaget kan vi tenke på som NIFU STEP's kandidatundersøkelse.
La den underliggende lønnsmodellen være:

(3) $\ln w_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + b_3 x_{3i} + u_i$

La oss foreløpig se bort fra problemet med de som ikke har besvart spørreskjema (anta at frafallet er rent tilfeldig). Vi ønsker å finne anslag på sammenhengen mellom lønnstilbudet, w og forklaringsvariablene x_1, x_2, x_3 som kan representere kjønn, utdanning etc. for populasjonen. Noen individer i populasjonen arbeider ikke. I standardtilfellet antas utvalget å være tilfeldig trukket fra den bakenforliggende populasjonen: Men nå er spørsmålet: Hvilke faktorer påvirker hvilke individer som er med i utvalget vi kan benytte? Vi kan bare benytte observasjonene for de individer som har rapportert lønnsnivå (altså de som jobber) i regresjonsmodellen. Det kritisk spørsmålet vi må stille nå er: Hvilke faktorer bestemmer om personen jobber eller ikke? Påvirker disse også lønnsnivået? For å få tak på de problemer dette medfører, er det nødvendig med litt formalisering.

Vi må modellere den prosessen som genererer valget mellom å delta eller ikke i arbeidslivet, og denne beslutningen kan enkelt oppsummeres i **seleksjonsligningen**.
La z_i^* være en latent variabel som indikerer nettogevinsten ved å jobbe: Kandidaten jobber dersom nettogevinsten er positiv. I økonomspråk betyr det at lønnstilbudet er høyere enn individets reservasjonslønn.

La videre nettogevinsten z_i^* være en lineær funksjon av et sett av observerbare variable pluss et stokastisk restledd som representerer uobserverbare variable:

(4) $z_i^* = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \alpha_4 x_4 + v_i$

der det stokastiske restleddet v_i er normalfordelt.

I tillegg til kjønn, alder og utdanningstype har vi antatt at variabelen

x_4 = Antall barn

påvirker nettogevinsten ved jobb

Det vi observerer er imidlertid ikke z_i^* men en indikator z_i som tar verdien 1 hvis individet jobber, og 0 ellers. Hvis vi som ovenfor antar at individet jobber dersom nettogevinsten ved å jobbe er positiv, har vi at:

(5) $$z_i = \begin{cases} 1 \text{ hvis } z_i^* > 0 \quad \text{dvs hvis } v_i > -(\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \alpha_4 x_4) \\ 0 \text{ hvis } z_i^* < 0 \quad \text{dvs hvis } v_i < -(\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \alpha_4 x_4) \end{cases}$$

Vi er altså interessert i å estimere parametrene i (3), men lønna $\ln w_i$ observeres bare når

$z_i = 1$

Ligningene (4) og (5) karakteriserer nå hvordan utvalget er trukket fra populasjonen

Ytterligere forutsetninger i modellen er:

a) $x_{1i}, x_{2i}, x_{3i}, x_{4i}$ observeres alltid, y_i observeres bare når $z_i = 1$

b) u_i og v_i er uavhengig av $x_{1i}, x_{2i}, x_{3i}, x_{4i}$ og har forventning 0.

c) u_i og v_i følger en bivariat normalfordeling $(u_i, v_i) \sim N(0, \sigma)$

σ er varians-kovariansmatrisa for u og v og er gitt ved

$$\sigma = \begin{pmatrix} \sigma_u & \sigma_{uv} \\ \sigma_{uv} & \sigma_v \end{pmatrix}$$

Det er verdt å merke seg betydningen av de to første forutsetningene. Forutsetning a) innebærer at de relevante variablene som inngår i så vel lønnsmodellen som i seleksjonsligninga (altså de variablene som påvirker om du jobber eller ikke) er observerbare. Forutsetning b) innebærer at både modellen for lønnstilbudet og seleksjonsmodellen er velspesifiserte, i den forstand at restleddene i de to ligningene er ukorrelert med forklaringsvariablene.

Imidlertid følger det fra forutsetningene over at

d) $E(u_i \mid v_i) = \frac{\sigma_{uv}}{\sigma_v} v_i = \delta v_i$

Det betyr at restleddet i lønnsligningen er korrelert med restleddet i seleksjonsligningen og tolkingen er at de uobserverbare variablene som påvirker lønna er korrelert med de uobserverbare variablene som påvirker sannsynligheten for at individet jobber.

I det følgende normaliserer vi variansen til restleddet i seleksjonsligningen til 1, dvs $\sigma_v = 1$

Dersom vi nå tar betinga forventning i lønnsligninga (forventet lønn, gitt størrelsen på forklaringsvariablene i lønnsligningen og gitt at individet jobber) får vi:

$$E(\ln w_i \mid x_{1i}, x_{2i}, x_{3i}; z_i = 1) = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + E[u_i \mid \mathrm{x}_{1\mathrm{i}}, x_{2i}, x_{3i}; z_i = 1]$$

Benytter vi forutsetning (d), kan denne skrives:

(6) $$E(\ln w_i \mid x_{1i}, x_{2i}, x_{3i}; z_i = 1) = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + \delta E[v_i \mid \mathrm{z_i} = 1]$$

Her kan vi skille mellom to mulige tilfeller:

1) $\delta = 0$. I dette tilfellet er de utelatte variablene i lønnsligningen ukorrelert med de utelatte variablene i seleksjonsligningen, $\sigma_{uv} = 0$ og vi er tilbake i standardsituasjonen. Siste ledd i (6) forsvinner og vi kan estimere parametrene i lønnsligninga konsistent ved OLS på observerte lønnsnivåer

2) $\delta \neq 0$

Dette tilfellet er utgangspunkt for James Heckman's artikkel i Econometrica fra 1979, Heckman (1979), hvor bidraget nettopp var formuleringen av seleksjonsproblemet som et problem med utelatte variable.

La oss se nærmere på situasjonen med $\delta \neq 0$ og vi må da studere egenskapene til leddet $E[v_i \mid \mathrm{z_i} = 1]$ i (6).

Vi har at

$$E[v_i \mid \mathrm{z_i} = 1] = \mathrm{E}[v_i \mid \mathrm{z_i}^* > 0] = E[v_i \mid \underbrace{v_i > -(\alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \alpha_3 x_{3i} + \alpha_4 x_{4i})}_{z_i^* > 0}]$$

Denne forventningen er den inverse Mills-ratioen eller "Heckman's lambda". Vi kan nemlig vise at når v_i er en standard normalfordelt variabel så er:

$$E[v_i \mid \underbrace{v_i > -(\alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \alpha_3 x_{3i} + \alpha_4 x_{4i})}_{z_i^* > 0}] =$$

$$= \frac{\phi(\alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \alpha_3 x_{3i} + \alpha_4 x_{4i})}{\Phi(\alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \alpha_3 x_{3i} + \alpha_4 x_{4i})} = \lambda_i$$

der ϕ er tetthetsfunksjonen til en standard normalfordelt variabel og Φ er den kumulative tetthetsfunksjonen til standard normalfordelt variabel.

Det viktige for oss er nå at både ϕ og Φ er kjente funksjoner og (6) kan nå skrives:

(7) $E(\ln w_i \,|\, x_{1i}, x_{2i}, x_{3i}; z_i = 1) = b_0 + b_1 x_{1i} + b_2 x_{2i} + b_3 x_{3i} + \delta\lambda_i$

Ut fra (7) ser vi nå at vi har et utelatt variabelproblem dersom vi estimerer lønnsligningen (3) basert på de observerte lønnsnivåene uten å korrigere for λ_i. Effekten av variablene x_1, x_2, x_3 på lønna vil da representere de kausale effektene pluss effekten av seleksjonen inn i jobb. Korreksjonsfaktoren λ_i er i utgangspunktet ukjent, men ideen er nå å lage en konsistent estimator for denne i et første steg. Andre steget består da simpelthen i å estimere lønnsrelasjonen med korreksjonsfaktoren λ_i ,representert ved estimatet, inkludert.

Framgangsmåten er altså enkel:

1.steg: Estimer seleksjonsligningen

(8)
$$P(z_i = 1) = P(z_i^* > 0) = P(\alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \alpha_3 x_{3i} + \alpha_4 x_{4i} + v_i > 0) =$$
$$= P(v_i > -(\alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \alpha_3 x_{3i} + \alpha_4 x_{4i}))$$

Når v_i er en standard normalfordelt variabel [forutsatt over] så er dette en standard Probit-modell. Probitmodellen kan estimeres på vanlig måte ved Maximum-Likelihood-metoden (ML) og vi får estimater på parametrene i seleksjonsligningen som: $\widehat{\alpha_0}, \widehat{\alpha_1}, \widehat{\alpha_2}, \widehat{\alpha_3}, \widehat{\alpha_4}$.

Vi kan da beregne et estimat på λ_i som:

$$\widehat{\lambda_i} = \frac{\phi(\widehat{\alpha_0} + \widehat{\alpha_1} x_{1i} + \widehat{\alpha_2} x_{2i} + \widehat{\alpha_3} x_{3i} + \widehat{\alpha_4} x_{4i})}{\Phi(\widehat{\alpha_0} + \widehat{\alpha_1} x_{1i} + \widehat{\alpha_2} x_{2i} + \widehat{\alpha_3} x_{3i} + \widehat{\alpha_4} x_{4i})}$$

Siden dette er forholdet mellom tettheten og den kumulative fordelingsfunksjonen til en standard normalfordelt variabel, som er kjente funksjoner, så kan denne størrelsen enkelt beregnes.

2.steg:Estimer lønnsmodellen:

$$\ln w_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + b_3 x_{3i} + \delta \widehat{\lambda}_i + \varepsilon_i$$

med OLS. Gitt våre forutsetninger vil dette gi konsistente estimatorer for b-ene.

Det er for øvrig grunn til å merke seg at de vanlige standardavvikene til OLS-estimatorene i 2. steget er inkonsistente. Heckman (1979) foreslo en konsistent varians-kovariansestimator som også er beskrevet i Greene (2003) s. 785. Økonometriprogrammer som STATA genererer de korrekte standardavvikene automatisk når Heckman-opsjonen benyttes.
Dessuten kan vi også teste med standard tester om koeffisienten foran λ er signifikant ulik null. Dersom koeffisienten er ikke-signifikant tyder det på at seleksjonsproblemet er av liten betydning. Det er også mulig å estimere modellen med maximum likelihood-metoden (ML) direkte, se Cameron og Trivedi (2005) s. 548. I STATA er både ML-estimering og to-stegsprosedyren beskrevet over tilgjengelig.

Identifikasjonsspørsmålet

I eksemplet foran hadde vi en variabel, x_4 (antall barn) som vi antok påvirket sannsynligheten for å jobbe, men ikke lønnsnivået direkte. I prinsippet kan vi estimere seleksjonsmodellen også når det samme sett av variable inngår både i seleksjonsligningen og i lønnsligningen. Men da er det bare ikke-lineariteten i λ_i som bidrar til å identifisere koeffisientene. Jo mer lineær den er, jo vanskeligere blir det å identifisere b-ene i lønnsligningen, fordi λ_i da vil være høyt korrelert med de øvrige variablene i lønnsligningen. Vi vil forvente høye standardavvik på de estimerte parametrene i dette tilfellet. Det er en stor litteratur som har undersøkt estimatorens egenskaper i dette tilfellet med Monte-Carlo-simuleringer, se Nawata og Nagase (1995) og Vella (1998) for en oppsummering.[5] Resultatene tyder på at modellen fungerer dårlig i endelige utvalg uten ekskluderingsrestriksjoner. I dag vil det i praksis være vanskelig å få internasjonal publisering av artikler som benytter Heckmans seleksjonsmodell uten en overbevisende ekskluderingsrestriksjon.

[5] Grasdal (2001) er en norsk studie som sammenligner hvordan ulike estimatorer fungerer på et eksperiment i helsesektoren (Bergenseksperimentet) der noen av individene i eksperimentet faller fra underveis.

3.2 Anvendelse for ikke-respons i intervju-data

Til nå har vi sett på hvordan seleksjon inn i jobb medfører et problem i estimering av lønnsmodeller. Et lignende tilfelle vil oppstå dersom det utvalget som besvarer spørreskjema er en selektert gruppe. Dersom vi har noe informasjon om alle enhetene (både de som responderer og de som ikke responderer) har vi en klar parallell til seleksjonsmodellen diskutert over. La oss derfor se nærmere på dette. Ta som eksempel at vi har sendt spørreskjema til N personer men bare N_1 personer besvarer.

La

y = avhengig variabel (utfall)

x_1 og x_2 er forklaringsvariable

Vi formuler følgende populasjonsmodell for den avhengige variabelen:

(9) $y_i = \alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + u_i$ der indeks i angir individet.

Vi er altså i utgangspunktet interessert i å finne konsistente estimatorer for parametrene i (9).

Vi formulerer i tillegg en modell for hvorvidt individet besvarer spørreskjema eller ikke:

La den latente variabel z_i^* representere den (uobserverbare) tilbøyeligheten til å besvare spørreskjema.

(10) $z_i^* = b_0 + b_1 x_{1i} + b_2 x_{2i} + b_3 x_{3i} + v_i$

Her har vi altså antatt at den latente variabelen z_i^* påvirkes av en ekstra variabel x_3 i tillegg til de variablene som inngår i (9). Vi har en indikator z som tar verdien 1 hvis individet responderer, og 0 ellers.

Individet responderer (og vi observerer y) dersom $z_i^* > 0$, altså når

$$v_i < -(b_0 + b_1 x_{1i} + b_2 x_{2i} + b_3 x_{3i})$$

Dette er altså helt tilsvarende lønnsmodellen i det foregående avsnittet og vi kan dermed bruke en tilsvarende estimeringsprosdyre som korrigerer for seleksjon på grunn av responstilbøyeligheten.

Det er verd å merke seg at denne prosedyren stiller betydelige krav til data for å kunne gjennomføres. Den krever at vi har informasjon for alle i utvalget (både de som responderer

og de som ikke responderer) om de andre relevante variablene som påvirker y og svartilbøyeligheten. Disse kravene kan være oppfylt dersom vi har registerinformasjon om alle individene i utvalget, men bruker informasjon fra spørreundersøkelsen for å lage den avhengige variabelen y. I tillegg vil også her identifikasjonsproblemet stå sentralt. Med andre ord, bør vi ha en ekstra variabel (x_{3i}) som påvirker sannsynligheten for å svare, men som ikke påvirker utfallet i (1). Dette er ofte kjerneproblemet her. Mulige kandidater til dette er:
1) Egenskaper ved intervjueren (ved personlig intervju)
2) Variable som karakteriserer forholdet mellom den institusjon som gjennomfører surveyen og respondenten

I det følgende vil vi gjengi et eksempel på en identifikasjonsstrategi av type 2). Hamermesh og Donald (2006) studerer en problemstilling som er relevant for arbeidet ved NIFU-STEP, nemlig avkastningsforskjeller målt ved lønnsforskjeller på ulike typer høyere utdanning (majors). De har data fra et stort universitet i USA med mange utdanningstyper (majors): Arkitechture&fine arts, Business-soft, Business-hard, Communications, Engineering, Humanities, Social sciences, Natural Sciences&Pharmacology, Nursing&Social work. De kombinerer registerdata fra universitetet med spørreundersøkelse til kandidater fra kullene 79/80, 84/85, 89/90, 94/95 og 1999/2000 om arbeidsmarkedssituasjonen deres i 2001-2002. De sendte spørreskjema til 7970 tidligere studenter og bare 2015 svarte, altså en svarprosent på bare 25. De estimerte først en probit-modell for sannsynligheten for respons, og fant at dummyer for utdanningstype (majors) hadde signifikant effekt på responsen. Dette tyder på at ikke-respons er et betydelig problem ved estimering av avkastningsforskjeller mellom majors.

For å håndtere dette setter Hamermesh og Donald opp en generell modell som både innebærer responsseleksjon og seleksjon mellom jobb og ikke jobb, altså en dobbel seleksjonsmodell.

(11) $y_1 = 1$ hvis $(x_1'\delta_1 + \varepsilon_1 > 0)$ Respons

(12) $y_2 = 1$ hvis $(x_2'\delta_2 + \varepsilon_2 > 0)$ Sysselsatt hvis svart på spørreskjema

(13) $\ln y_3 = x_3'\delta_3 + \varepsilon_3$ hvis $y_2 = 1$ Lønn hvis sysselsatt

Identifikasjonsproblemet her består i å finne variable som
a) påvirker svartilbøyeligheten, men
b) ikke påvirker sysselsetting og lønn direkte (ekskluderingsrestriksjonen)

Hamermesh og Donald's forslag går ut på å bruke en indikator for om individet var medlem av universitetets alumniorganisasjon på intervjutidspunktet som identifiserende variabel. De viser empirisk at responstilbøyeligheten er klart positivt korrelert med indikatoren, altså tilfredsstiller den krav a). Deres problem er imidlertid at krav b) ikke er testbart. Spørsmålet er altså om vi tror alumnimedlemskap er ukorrelert med lønn og sysselsetting. Hamermesh og Donald er selvsagt oppmerksom på dette, og argumenterer for at denne eksklusjonsrestriksjonen er mer troverdig enn ekskluderingsrestriksjonen som brukes i tradisjonelle lønns-sysselsettingsmodeller, hvor antall barn forutsettes bare å påvirke beslutningen om yrkesdeltakelse og ikke lønn. Det får bli opp til leseren å vurdere om dette er et godt argument eller ikke. Hamermesh og Donald utleder ML-estimatoren for denne doble seleksjonsmodellen. Resultatene deres viser for det første at forskjellen i avkastning mellom ulike utdanningstyper (majors) målt ved standardavviket i lønnsforskjellen, reduseres kraftig (halveres) når det kontrolleres for kjennetegn ved individene. Korreksjonen for seleksjon reduserer standardavviket i lønnsforskjellene ytterligere med 10 prosent.

3.3 Frafall i longitudinale intervjuundersøkelser.

Foreløpig har vi behandlet situasjoner der datamaterialet er et rent tverrsnittsmateriale. Seleksjonsproblemet kan imidlertid også oppstå i arbeidet med paneldata (longitudinale data). Dette kan også være en problemstilling i noen av de undersøkelsene som NIFU STEP arbeider med. Et eksempel kan være utvalget i kandidatundersøkelsen 2000 der noen spørres igjen i 2004 og 2008[6]. I utgangspunktet kan paneldata være en fordel ved identifikasjon av kausale effekter. Grunnen er at man da har mulighet for å kontrollere for alle individvariable som er konstante over tid ved å inkludere såkalte faste individeffekter i regresjonsmodellene eller transformere modellene til førstedifferanser. Så lenge de variable som er av interesse for undersøkelsen varierer tilstrekkelig over tid innen hver enhet (individ) kan dermed empiriske analyser basert på paneldata gi mer pålitelige resultater enn tilsvarende analyser basert på rene tverrsnitt.

[6] Ifølge beskrivelsen av Kandidatundersøkelsen 2004 ble de som deltok i 2000-undersøkelsen spurt om de var villige til å delta i nye undersøkelser i 2004 og 2008. Halvparten av de uteksaminerte ble spurt om å delta i nye spørreskjemaundersøkelser mens den andre halvparten ble bedt om å gi tillatelse til at det nyttes opplysninger fra Statistisk sentralbyrås registre til å følge deres videre yrkeskarriere. Mellom halvparten og to tredjedeler av kullet sa seg villige til videre deltakelse i 2004 og 2008. Se http://www.nifustep.no/norsk/innhold/prosjekter/kandidatunders_kelsene__1/spesialunders_kelser/kandidatunders_kelsen_2004

Et problem som ofte kan dukke opp i paneldata (longitudinale data) er imidlertid at noen av enhetene faller fra i løpet av observasjonsperioden.

Eksemplet med NIFU STEP sin kandidatundersøkelse fra 2002 kan belyse dette. Problemet består i at noen av de som svarte i 2000-undersøkelsen faller fra i 2004 og eventuelt. i 2008. I den engelskspråklige litteraturen betegnes dette som "the attrition problem".[7]

For å gå videre, la oss formalisere litt. Longitudinale design på spørreundersøkelsen gir oss et panel. Med paneldata kan vi eliminere individspesifikke variable som er konstante over tid ved å ta første-differanser.
La modellen være

(14) $y_{it} = a_1 x_i + a_2 z_{it} + u_{it} + \eta_i$ t=1,……,T og i=1,……,N

η_i er den individspesifikke restleddskomponenten, mens u_{it} er et idiosynkratisk restledd (både variasjon over tid og mellom individer). x betegner en observerbar variabel som bare varierer mellom individer, mens z betegner en observerbar variabel som varierer både mellom og innen individer. Hvis den avhengige variabelen er lønnsnivå kan z være nivået på arbeidsledighetsraten i det området individet jobber. Vi er bekymret for korrelasjon mellom det individspesifikke restleddet og den observerbare personkarakteristikken x.

Ved å ta første differanser av modellen fjernes den individspesifikke komponenten, se s. 585-586 i Woolridge (2002).

(15) $\triangle y_{it} = a_2 \triangle z_{ti} + \triangle u_{it}$ t=2,…. T

Problemet som nå gjenstår er at noen enheter faller fra. La oss si at vi starter på tidspunkt 1 med N individer. Fra og med tidspunkt 2 og framover, vil noen av individene falle fra, og vi antar at de som faller fra forblir ute av utvalget i de resterende tidsperiodene. Hvis disse frafallene ikke er tilfeldig, har vi et seleksjonsproblem. En måte å håndtere dette på er å modellere sannsynligheten for at individet faller fra. Det betyr at vi introduserer en seleksjonsligning på samme måte som i de foregående modellene. La s_{it}^* være den latente

[7] Framstillingen her bygger i stor grad på Woolridge (2002).

tilbøyeligheten til å falle fra fra et tidspunkt til et annet, mens s_{it} er en indikator for om individet faktisk falt fra.

Seleksjonsligningen for individ i på tidspunkt t kan da skrives:

(16) $s_{it} = 1$ dersom $s^* = w_{it}\beta + v_{it} > 0$

w_{it} kan være laggede verdi på Z, Z_{it-1} eller variable som er mulig å beregne for alle, så lenge de er observert på det initiale tidspunktet (Eksempel: alder)

Vi antar nå at z er en eksogen variabel og at seleksjonen ikke er korrelert med $\triangle z_{it}$, når w er kontrollert for.

Videre antar vi at forventingen til restleddet $\triangle u_{it}$ betinget på $\triangle z_{it}$ og w_{it} kan skrives[8]:

$$E(\triangle u_{it} \mid \triangle z_{it}, w_{it}, s_{it} = 1) = E(\triangle u_{it} \mid v_{it}) = \delta_t v_{it}$$

Da blir den betingede forventning for Δy:

(17) $E(\triangle y_{it} \mid \triangle z_{it}, w_{it}, s_{it} = 1) = a_2 \triangle z_{it} + \delta_t \lambda(w_{it}\beta)$

der λ tilsvarer Heckman's lambda i de forrige anvendelsene

Igjen kan vi følge en tostegsprosedyre:

Steg 1:

Estimer for hvert tidspunkt t=2,….T en probitligning for om individet i som responderte på tidspunkt t-1 også er med i utvalget på tidspunkt t. Gir oss T-1 probit-ligninger og et sett av estimerte "Lambdaer", $\widehat{\lambda_{it}}$ for hver observasjon inkludert i år t.

Steg2:

Estimer

(18) $\triangle y_{it} = a_2 \triangle z_{it} + \delta_2 d2_t \hat{\lambda}_{it} + \delta_3 d3_t \hat{\lambda}_{it} + ...\delta_T d3_T \hat{\lambda}_{it} + restledd \qquad t = 2,....,T$

på panelet med OLS.

der $d2_t, d3_t, ...dT_T$ er tidsdummier

[8] Denne vil gjelde dersom Δu_{it} og v_{it} er simultant normalfordelt.

Under våre forutsetninger vil dette gi konsistente anslag på interessevariabelen a_2. En test på om seleksjonsskjevhet er et problem kan enkelt gjennomføres ved å teste den simultane hypotesen om $\delta_2 = \delta_3 = ... = \delta_T = 0$ i (18). Woolridge (2002) s. 586 viser også hvordan ligningen kan estimeres ved en instrumentvariabel-metode (IV-metode) når en eller flere av forklaringsvariablene i modellen er endogene. Vi vil imidlertid ikke gå nærmere inn på dette her.

Dette avsnittet har bare gitt en smakebit på de metodeutfordringer som spesielt reiser seg ved frafall i paneldata. Problemene og mulige estimeringsmetoder er inngående behandlet i Nijman og Verbeek (1992), Woolridge (1995) og Vella og Verbeek (1999).

4 Heckman's seleksjonsmodell anvendt i effektstudier.

Dette er en relativt rett fram applisering av metodikken foran. Situasjonen som skal håndteres er imidlertid noe annerledes. I motsetning til det foregående har vi observasjoner om den avhengige variabelen for alle i utvalget, men forklaringsvariabelen (dummyvariabel for "behandling" (treatment)) er endogen. Den klassiske litteraturreferansen her er Heckman (1978). Anvendelsen av metodikken er svært omfattende. Her er noen anvendelser på arbeidsmarkeds og utdanningsområdet som kan være relevant for NIFU-STEP:

- Studier av lønnsforskjeller mellom privat og offentlig sektor
- Studier av lønnsforskjeller mellom fagorganiserte og ikke fagorganiserte
- Studier av lønnsforskjell mellom individer med og uten høyere utdanning ("college wage premium)

For å komme videre må vi formalisere litt. Vi betrakter en enkel lineær modell for den avhengige variabel y:

(19) $y_i = \beta_0 + \alpha D_i + X_i \beta_x + u_i$

α er vår interesseparameter. Den angir gjennomsnittseffekten av "behandling"("treatment"), der "behandlingen", altså tolkingen av variabelen D, kan være ansettelse i privat sektor, fagorganisering eller høyere utdanning.

D er altså en dummyvariabel som angir om individet er "behandlet" eller "ikke-behandlet" For eksempel

$D_i = 1$ hvis individet arbeider i privat sektor.

$D_i = 0$ ellers

X er et sett av andre variable som påvirker den avhengige variabel y.

Det sentrale nå er at hvorvidt individet er behandlet eller ikke, ikke beror på rene tilfeldigheter men er et resultatet av egne valg. Vi formulerer derfor en deltakermodell som er helt parallell til seleksjonsmodellen foran.

(20) $D_i^* = Z_i \gamma + v_i$ der D* er en latent variabel for individets nettogevinst ved å delta.

Z er et sett av variable som påvirker deltakelsen og Z kan i prinsippet være identisk med X. Vi observerer

(21) $D_i = 1$ hvis $D^*_i = Z_i\gamma + v_i > 0 \Leftrightarrow v_i > -Z_i\gamma$

Her er det verdt å gjenta forskjellen fra den forrige seleksjonsmodellen: Her observerer vi altså den avhengige variabelen for alle enheter, men vi har en behandlingsindikator, D som er resultat av beslutninger. En annen måte å se det på er at vi ikke kan observere det kontrafaktiske for samme person.

Vi har en seleksjonsskjevhet fordi det er en systematisk sammenheng mellom restleddet u, og behandlingsvariabelen D, betinget på X. Her kan vi skille mellom ulike situasjoner

1) Seleksjon på observerbare variable
Dette er situasjonen hvis seleksjonsskjevheten skyldes korrelasjon mellom u og Z

2) Seleksjon på uobserverbare variable.
Dette er situasjonen hvis seleksjonsskjevheten skyldes korrelasjon mellom u og v:

Vi skal i dette notatet rendyrke situasjon 2):
Her er altså situasjonen at u og v er ukorrelert med X og Z, men innbyrdes korrelert, altså samme som i Heckman-modellen foran.

Dersom vi tar betinga forventning til (19) får vi

(22) $E(y_i \mid X_i, Z_i, D_i) = \beta_0 + \beta_x X_i + \alpha D_i + E(u_i \mid X_i, Z_i, D_i)$

Vårt problem er at leddet $E(u_i \mid X_i, Z_i, D_i)$ ikke er lik null pga korrelasjonen mellom restleddene i de to ligningene.

Intuisjonen bak løsning på problemet er at vi ønsker å finne et uttrykk for dette leddet og så estimere modellen med leddet (eller et estimat på det) inkludert i modellen.
Under normalitetsforutsetningene om restleddene, dvs at u og v er bivariat normalfordelt med forventning null, samt at vi har normalisert variansen til v til 1, får vi

$$E(u_i \mid X_i, Z_i, D_i = 1) = \delta\lambda_{1i} = \delta \frac{\varphi(Z_i\gamma)}{\Phi(Z_i\gamma)}$$

$$E(u_i \mid X_i, Z_i, D_i = 0) = \delta\lambda_{0i} = \delta \frac{-\varphi(Z_i\gamma)}{1-\Phi(Z_i\gamma)}$$

Dette er formelt vist hos Heckman (1978), side 937-938, ved bruk av relativt standard resultater fra multivariat fordelingsteori.

Setter vi dette resultatet inn i (4) kan vi skrive modellen (for eksempel lønnsmodellen) som:

(23) $$y_i = \beta_0 + \beta_x X_i + \alpha D_i + \delta \underbrace{\left[D_i \lambda_{1i} + (1 - D_i)\lambda_{0i} \right]}_{\text{"LAMBDA"}} + u_i^*$$

der u* er et restledd med "klassiske" egenskaper."LAMBDA" i (23) fungerer altså på samme måte som korreksjonsleddet i utvalgsseleksjonsmodellen foran. Poenget er nå at "LAMBDA" kan beregnes siden den er en kjent funksjon av observerbare variable.

Hvis vi tar forventningen til (23) betinget på de observerbare X, D og Z får vi

$$E(y \mid X, Z, D = 1) = \beta_0 + \beta_x X_i + \alpha + \delta\lambda_1 \qquad \text{(D=1)}$$

$$E(y \mid X, Z, D = 0) = \beta_0 + \beta_x X_i + \delta\lambda_0 \qquad \text{(D=0)}$$

Differansen i y mellom de behandlede og de ikke-behandlede blir derfor:

(24) $$E(y \mid X, Z, D = 1) - E(y \mid X, Z, D = 0) = \alpha + \delta(\lambda_1 - \lambda_0) = \alpha + \delta \frac{\varphi(Z\gamma)}{\Phi(Z\gamma)(1 - \Phi(Z\gamma))}$$

Seleksjonsskjevheten innebærer altså at dersom vi ikke kontrollerer for seleksjonen når vi estimerer ligningen med OLS vil vi få overvurdert behandlingseffekten dersom δ >0, dvs. dersom restleddene er positivt korrelerte og undervurdert den i det motsatte tilfellet.

For å illustrere dette kan det være nyttig å tenke på lønnseksemplet: Dersom uobserverbar dyktighet både gir høyere lønn uavhengig av sektor og samtidig øker sannsynligheten for å velge privat sektor, så vil OLS uten korreksjon overvurdere lønnsforskjellen mellom privat og offentlig sektor.

Ved estimeringen kan vi også her enten bruke en tostegs-prosedyre eller estimere begge ligningene i modellen simultant med maximum-likelihoodmetoden. Jeg skal her se nærmere på to-stegsmetoden.

STEG 1: Finn estimater for γ , $\hat{\gamma}$ ved å estimere en Probit-modell for D=1 versus D=0. Beregn elementene i "LAMBDA":

$$\hat{\lambda}_1 = \frac{\varphi(Z\hat{\gamma})}{\Phi(Z\hat{\gamma})} \qquad \text{og } \hat{\lambda}_0 = \frac{-\varphi(Z\hat{\gamma})}{1 - \Phi(Z\hat{\gamma})}$$

<u>Steg 2:</u>

Estimer motstykket til (23) med OLS:

(25) $$y = \beta_0 + \beta_x X + \alpha D + \delta \underbrace{\left[D\hat{\lambda}_1 + (1-D)\hat{\lambda}_0 \right]}_{\text{"LAMBDA"}} + u*$$

Under våre forutsetninger gir det oss konsistente anslag på parametrene i modellen.
Som i den ”vanlige” seleksjonsmodellen er ikke den estimerte varians-kovariansmatrisa som kommer ut av OLS-estimering av (25) konsistent. Heckman (1978) foreslo en konsistent estimator for denne, som er implementert i ”treatreg”-kommandoen i STATA. Treatreg gir også mulighet for å estimere modellen med ML.

Som i de andre seleksjonsmodellene er identifikasjonsproblemet også her avgjørende. Selv om modellen under de forutsetninger som er beskrevet over kan estimeres også når det samme settet av variable inngår både i strukturligningen og i deltakerligningen (X og Z er identiske) vil tilliten til resultatene øke sterkt dersom det kan pålegges troverdige eksklusjonsrestriksjoner. Det vil si at det finnes en eller flere variable som påvirker deltakerbeslutningen (for eksempel beslutningen om å jobbe i privat sektor), men ikke lønnsnivået direkte. Igjen er det en stor utfordring å finne slike variable.

5 Oppsummering og konkluderende merknader

Dette notatet har gitt en innføring i hvordan problemer med selvseleksjon og manglende respons i intervjudata kan behandles i økonometriske modeller. Litteraturen på området er meget stor, og det har derfor bare vært mulig å gi en liten smakebit på problemstillinger og metoder. Utgangspunktet er den klassiske modellen for korrigering av seleksjon introdusert av Heckman (1979). Det er grunn til å understreke at mange av de metodene som er diskutert har både sterke og svake sider. For eksempel legges ofte sterke fordelingsforutsetninger til grunn og det stilles også betydelige krav til datamaterialet. Anvendelse på utvalgsskjevhet på grunn av manglende respons på spørreundersøkelser krever blant annet at det finnes data for relevante variable som påvirker tilbøyeligheten for respons. I de tilfeller dette datakravet ikke er oppfylt (for eksempel hvis det ikke finnes noen opplysninger om de som ikke har besvart spørreskjema) er man gjerne henvist til å bruke mer skjønnsmessig vurdering av representativiteten av utvalget. Videre er det et felles trekk ved de metoder som er presentert at de fungerer godt bare dersom det kan pålegges troverdige identifiserende eksklusjonsrestriksjoner, dvs. at det finnes variable som påvirker deltakingsbeslutningen, men ikke den avhengige variabelen direkte. Det å finne slike eksklusjonsrestriksjoner er ofte en stor utfordring i praktiske anvendelser og krever ofte inngående kjennskap til sektoren som analyseres og til tidligere studier på det aktuelle området. En pragmatisk tilnærming er derfor å la estimering av modeller som korrigerer for seleksjonsskjevhet inngå som en del av en generell robusthetssjekk og sensitivitetsanalyse av økonometriske modeller.

Referanser:

Cameron, A. C. og P. K. Trivedi (2005): Microeconometrics. Methods and applications. Cambridge University Press.

Grasdal, A. (2001): The performance of sample selection estimators to control for attrition bias. *Health Economics 10*, 385-398.

Green, W. H. (2003): Econometric Analysis. Fifth Edition. Prentice Hall.

Hamermesh, D. S. og S. Donald (2006): The effect of college curriculum on earnings: An affinity identifier for non-ignorable non-response bias. Revidert versjon av NBER Working paper no. 10809, 2004.

Heckman, J. J. (1978): Dummy endogenous variables in a simultaneous equations system. *Econometrica 46*, 931-960.

Heckman, J. J. (1979: Sample selection bias as a specification error. *Econometrica 47*, 153-161.

Nawata, K. og N. Nagase (1996): Estimation of sample selection bias models. *Econometric Reviews 15*, 387-400.

Vella, F. (1998): Estimating Models with Sample Selection Bias: A Survey. *Journal of Human Resources* 33, 127-172.

Vella, F. og M. Verbeek (1999): Two-step estimation of panel data models with censored endogenous variables and selection bias. *Journal of Econometrics 90*, 239-263.

Woolridge, J. (1995): Selection corrections for panel data models under conditional mean independence assumptions. *Journal of Econometrics 68*, 115-132.

Woolridge, J. (2002): Econometric Analysis of Cross Section and Panel Data. MIT Press.

Woolridge, J. (2003): Introductory Econometrics. 2.edition. Thomson South-Western.