Målefeil i regresjonsanalyse

Magne Thoresen

Seksjon for medisinsk statistikk, Universitetet i Oslo
E-post: magne.thoresen@basalmed.uio.no

SAMMENDRAG

Vi vil i denne artikkelen se på effekter av målefeil i forskjellige regresjonsmodeller. Vi vil i hovedsak konsentrere oss om målefeil i forklaringsvariablene, men vil også kort nevne målefeil i responsvariabelen. Vi vil se på effekter både med hensyn på estimering og med hensyn på hypotesetesting. Vi vil plukke ut situasjoner hvor det er mulig å si noe sikkert om slike effekter, men vårt hovedbudskap er at effektene av målefeil generelt er vanskelig forutsigbare. Vi vil til slutt si noen ord om metoder for å korrigere for målefeil.

Thoresen M. **Measurement error in regression analysis**. *Nor J Epidemiol* 2003; **13** (2): 257-263.

ENGLISH SUMMARY

Measurement errors are inevitable in epidemiologic studies. In this paper we will summarize what is known about effects of measurement errors in different regression models. Most of the literature in this area has concentrated on measurement errors in the explanatory variables, but it is of course also possible to have errors in the response variables. We will also mention this very briefly. We will discuss effects both on parameter estimation and on hypothesis testing. We will describe situations where such effects are possible to foretell, but our main goal is to point to the fact that in general, effects of measurement errors are unpredictable. Finally, we will very briefly describe three methods for measurement error correction; regression calibration, simex, and likelihood methods.

INNLEDNING

Moderne epidemiologi gjør en utstrakt bruk av regresjonsanalyser, typisk lineær regresjon, logistisk regresjon, Cox regresjon og Poisson regresjon. Bakgrunnen for dette vil kunne være at man er interessert i å studere effekten av en eller flere eksponeringsvariable, på en responsvariabel, justert for effekten av en eller flere konfunderende variable. Felles for alle regresjonsmodeller er at de forutsetter at forklaringsvariablene er målt uten feil. I epidemiologisk sammenheng vil dette sjelden være oppfylt. Innenfor enkelte former for epidemiologi vil man tvert i mot operere med temmelig store målefeil. Eksempler på dette kan være studier av en eventuell sammenheng mellom luftforurensning og risiko for å utvikle astma, hvor man forsøker å måle individuell eksponering for luftforurensning. Et annet eksempel er ernæringsepidemiologi, hvor man studerer sammenhenger mellom inntak av forskjellige matvarer eller næringsstoffer og risiko for å utvikle for eksempel kreft. Målinger av enkeltpersoners eksponering for luftforurensning eller inntak av matvarer/næringsstoffer er forbundet med relativt stor usikkerhet.

Det er også verdt å legge merke til at vi kan ha målefeil i responsvariablene våre. Dersom man i eksemplet over registrerer om en person har astma eller ikke ved hjelp av spørreskjema, er disse dataene opplagt utsatt for en viss grad av målefeil.

Det har vært en vanlig oppfatning at målefeil vil medføre at man underestimerer sammenhengen mellom eksponering og utfall. Som vi skal se er dette opplagt kun i enkelte spesielle tilfeller.

Effekter av målefeil i lineær regresjon har vært et fokus av interesse i den statistiske faglitteraturen gjennom hele det forrige århundret. På dette temaet foreligger det dermed en omfattende mengde publikasjoner og teori. For de spesielt interesserte kan vi referere til Fuller (1987), men dette er en teknisk krevende bok og ikke spesielt godt egnet som introduksjon. Fra midten av 1980-tallet har det også kommet en mengde litteratur som omhandler målefeil i andre regresjonsmodeller enn lineær regresjon (ikke-lineær regresjon). I den epidemiologiske litteraturen begynte man å fokusere på målefeil, eller feilklassifisering, i enkle 2x2 tabeller allerede på 1950-tallet (se f.eks. Bross, 1954), og mye av den utviklingen som har funnet sted når det gjelder teori rundt målefeil i ikke-lineær regresjon er motivert i epidemiologiske problemstillinger. En god introduksjon til målefeilsproblemer i epidemiologiske studier kan man finne i *Statistics in Medicine* (1989), Vol. 8, som i sin helhet omhandler dette temaet.

I denne artikkelen skal vi forsøke å gi en oversikt over effekten av forskjellige former for målefeil i forskjellige regresjonssituasjoner. Vi vil begynne med å gi noen eksempler på epidemiologiske situasjoner hvor målefeil er viktig. Vi vil videre beskrive forskjellige typer målefeil, og deretter se på effekten av disse med hensyn på både estimering og testing i regresjonsmodeller. Til slutt vil vi også si litt om mulige korreksjoner for målefeil.

EKSEMPLER

Vi skal her kort gi noen eksempler på studier hvor målefeil på forklaringsvariable har vært ansett for å være viktig, og hvor det har vært gjort forsøk på å anslå noe om effekten av disse.

Eksempel 1

Det første arbeidet som omhandlet korreksjon for målefeil i ikke-lineær regresjon er vanligvis ansett for å være Prentice (1982). Hans arbeid var motivert av en studie av overlevende etter atombombeeksplosjonene over Hiroshima og Nagasaki. Få år etter eksplosjonene ble det gjort et forsøk på å etablere en kohorte av overlevende. Hovedinteressen var grad av eksponering for radioaktiv stråling (dosering) og potensiell sammenheng med død av forskjellige typer kreft. Strålingsnivået ble estimert på bakgrunn av lokalisering innen byen på tidspunktet for bomben, og på bakgrunn av individuelle opplysninger om vern. Det ble antatt at selv om antagelsene om strålenivå som en funksjon av avstand fra episenteret var korrekt, ville de individuelle strålingsdoseestimatene avvike fra de sanne verdiene med så mye som 30%. Målsettingen i dette arbeidet (Prentice, 1982) var å bruke informasjon om (eller antagelser om) målefeilene til å korrigere risikoestimatene.

Eksempel 2

MacMahon et al. (1990) ville estimere sammenhengen mellom blodtrykk og hjerte-karsykdom, basert på en metaanalyse av ni store prospektive observasjonsstudier. Totalt var 420 000 individer inkludert i disse studiene. MacMahon og medarbeidere innså at tidligere estimater var skjeve fordi de vanligvis baserte seg på enkeltmålinger av blodtrykk, gjort ved starten av studien. Slike målinger vil være utsatt for tilfeldige svingninger som igjen vil skape en skjevhet i de estimerte sammenhengene med sykdom. Ved å bruke data fra Framinghamstudien, hvor det ble gjort gjentatte målinger av blodtrykk, estimerte de de overnevnte sammenhengene til å være 60% sterkere enn tidligere antatt (i ukorrigerte analyser). Disse metodene og resultatene ble senere diskutert av Carroll og Stefanski (1994).

Eksempel 3

Mye av litteraturen på målefeil i ikke-lineær regresjon er motivert av arbeider innen ernæringsepidemiologi, hvor målefeilene er virkelig store. Et eksempel er "Nurses' Health Study" hvor over 80 000 kvinner er fulgt over tid, og et av målene har vært å studere sammenhengen mellom inntak av mettet fett og risiko for brystkreft (Willett et al. 1992). Man brukte her et matvarefrekvensskjema (spørreskjema) til å måle inntak av mettet fett. Data fra slike skjema inneholder store målefeil, både tilfeldige og systematiske. Skjemaet var validert i en egen valideringsstudie hvor man brukte et gjennomsnitt av fire en-ukes kostregistreringer gjort hver tredje måned i løpet av et år som et uttrykk for sant inntak. Rosner (1996) estimerte sammenhengen mellom inntak av mettet fett og risiko for brystkreft til å være 20% sterkere enn i tidligere ukorrigerte analyser. Denne studien har motivert viktige arbeider av Rosner et al. (1989, 1990).

TYPER MÅLEFEIL

Som nevnt i innledningen kan man ha målefeil både på forklaringsvariable og på responsvariable. Fokuset i forskningen på effekter av målefeil har i hovedsak vært på effekter av målefeil på forklaringsvariable. Dette vil også være vårt hovedfokus. Vi vil imidlertid også inkludere et avsnitt om effekter av målefeil på responsvariable.

Videre er det vanlig å skille mellom såkalte differensielle og ikke-differensielle målefeil. Snakker man om feil på forklaringsvariablene så er ikke-differensielle målefeil feil som er uavhengige av responsvariabelen (Y), mens differensielle målefeil er målefeil som er avhengige av responsvariabelen. Omvendt, dersom man snakker om feil på responsvariablene er ikke-differensielle målefeil feil som er uavhengige av forklaringsvariablene (X), mens differensielle målefeil er målefeil som er avhengige av forklaringsvariablene. Differensielle målefeil vil typisk kunne oppstå i case-control studier, hvor man henter informasjon om eksponering tilbake i tid, etter at responsvariabelen er registrert (etter diagnose). Det er lett å tenke seg at den informasjonen man får tak i om eksponering tilbake i tid kan være påvirket av diagnosen. I tilfeller med differensielle målefeil er det vanskelig å si noe generelt om effekten av feilen. I kohortestudier vil dette problemet være betraktelig mindre, da man vil samle inn data om eksponering før responsvariabelen registreres. Man skal imidlertid være klar over at man kan innføre differensiell målefeil ved å kategorisere en kontinuerlig forklaringsvariabel målt med ikke-differensiell feil (Flegal et al. 1991). Vi vil i denne artikkelen konsentrere oss om situasjoner med ikke-differensielle målefeil.

La X være vår uobserverbare forklaringsvariabel, og la W være det vi observerer. W er relatert til X gjennom en feilmodell. Vi kan spesifisere en relativt enkel feilmodell ved

$$W = \beta_0 + \beta_1 X + U, \qquad (1)$$

hvor U har forventning null. Dette gir at vår observerbare forklaringsvariabel W er en lineær funksjon av vår uobserverbare variabel X, pluss tilfeldig støy. Hvis vi nå antar $\beta_0 = 0$ og $\beta_1 = 1$ har vi $W = X + U$. Her observerer vi altså X kun tillagt tilfeldig støy. Dette er en modell som vil være relevant i forhold til f.eks. målinger av blodtrykk, kolesterol og lignende, hvor de målte verdiene fluktuerer rundt et sant, underliggende gjennomsnitt. Man vil vanligvis anta at U har konstant varians, altså at støyen er like stor for alle personer i

studien. Veldig mye av teorien som er utviklet rundt målefeil baserer seg på denne modellen. Det er imidlertid klart at man kan oppleve situasjoner hvor dette ikke holder. Vi kan anta at variansen vil variere med person, og vi kan anta at variansen vil variere med verdien av X, f.eks. slik at det er større tilfeldig støy rundt høye verdier av X enn rundt lave.

En mer avansert modell for feilene vil være å anta at hver enkelt person har en egen konstant skjevhet som kommer i tillegg til den tilfeldige variasjonen, eller å anta at de observerte verdiene W er en funksjon også av andre forklaringsvariable enn X.

EFFEKTER AV MÅLEFEIL

Når man snakker om effekter av målefeil, konsentrerer man seg som oftest om parameterestimeringen. Man er altså interessert i om man estimerer effektene av forklaringsvariablene korrekt. I tillegg til dette vil man naturlig også være interessert i effekten av målefeil på de statistiske testene. Vi skal se på begge disse aspektene. Vi skal innlede med å se nærmere på de absolutt enkleste tilfellene.

Hvis vi går tilbake til likning (1), så vil det å sette $\alpha_0 \neq 0$, $\alpha_1 = 1$ og $U \equiv 0$ bety at vi har en systematisk feil = α_0, hvor α_0 er konstant ($W = \alpha_0 + X$). Dette kan f.eks. forekomme hvis vi kalibrerer et måleinstrument feil. Denne typen feil vil ikke ha betydning i regresjonsmodellene, så sant vi kun er interessert i å estimere effektene av våre forklaringsvariable. Vi vil få en skjevhet i det estimerte konstantleddet (β_0), men dette har jo ingen betydning for effekten av forklaringsvariablene våre.

Hvis vi derimot setter $\alpha_0 = 0$, $\alpha_1 \neq 1$ og $U \equiv 0$, hvor α_1 er konstant, har vi en situasjon med et konstant avvik som er avhengig av nivået av X, altså et prosentvis konstant avvik ($W = \alpha_1 X$). Nå er det selvfølgelig enkelt å vise at det å bruke W i regresjonen istedenfor X vil føre til at vi overestimerer effekten av X i tilfellet med $\alpha_1 < 1$ og underestimerer effekten av X i tilfellet med $\alpha_1 > 1$.

Dette er altså de helt enkle situasjonene. I det som følger vil vi ta utgangspunkt i modellen med tilfeldig feil, altså $W = X + U$ der U har konstant varians. I mer kompliserte feilmodeller er det vanskelig å si noe generelt om effekten av feilene.

Estimering

Når det gjelder estimering vil vi se på de forskjellige regresjonsmodellene hver for seg. Spesielt er det naturlig å se på lineær regresjon isolert fra andre regresjonsmodeller, fordi man i lineær regresjon kan uttrykke estimeringsskjevheten eksplisitt.

Lineær regresjon

I den enkle lineære regresjonsmodellen $Y = \beta_0 + \beta_x X +$ ε hvor vi gjennomfører regresjonen på W istedenfor på den uobserverbare X kan man vise at den estimerte koeffisienten β_w vil underestimere β_x med en faktor $\lambda = \sigma_x^2 / \sigma_w^2$ (forholdet mellom variansen til X og variansen til W). Vi estimerer altså $\lambda\beta_x$ istedenfor β_x. Dette betyr at hvis man for eksempel vil se på sammenhengen mellom systolisk blodtrykk og grad av depresjon (målt på en gitt skala), gjennom en enkel lineær regresjonsanalyse, vil man underestimere denne sammenhengen med en faktor som tilsvarer forholdet mellom variasjonen til det sanne, underliggende blodtrykket (X) og variasjonen til det observerte blodtrykket (W). Det finnes data som kan tyde på at dette forholdet er omtrent 0,8.

Mer interessant blir dette straks man i tillegg også har forklaringsvariable som er målt uten feil. Det er nemlig slik at estimatet av effekten av slike variable også vil være påvirket av målefeilene i andre forklaringsvariable. Dette på en slik måte at dersom det er en positiv korrelasjon mellom en forklaringsvariabel målt uten feil og en forklaringsvariabel målt med tilfeldig feil, vil effekten av variabelen målt uten feil, overestimeres. Tilsvarende vil den underestimeres dersom det er en negativ korrelasjon. Anta at vi har en situasjon med en forklaringsvariabel X som måles med feil (gjennom W) og en forklaringsvariabel Z som måles uten feil. Det kan da vises (Carroll et al. 1995) at den estimerte effekten av Z kan uttrykkes som

$$\beta_z^* = \beta_z + \beta_x(1 - \lambda)\gamma_z.$$

Her er $\lambda = \sigma_{x|z}^2 / \sigma_{w|z}^2$ hvor $\sigma_{x|z}^2$ betegner variansen til X gitt Z, $\sigma_{w|z}^2$ betegner tilsvarende variansen til W gitt Z, og γ_z er koeffisienten i regresjonen av X på Z. Dette vil si at γ_z er positiv dersom det er en positiv korrelasjon mellom X og Z, og negativ dersom det er en negativ korrelasjon mellom X og Z.

Spesielt kan vi se på en situasjon hvor Z er binær, altså en kategorisk variabel som kun tar to verdier. Et eksempel kan være kjønn. Da kan man vise (Carroll, 1989) at den estimerte effekten av Z kan uttrykkes som

$$\beta_z^* = \beta_z + \frac{\beta_x \sigma_u^2 (\mu_1 - \mu_2)}{2(\sigma_x^2 + \sigma_u^2 - (\mu_1 - \mu_2)^2/4)}.$$

Her betegner μ_1 og μ_2 gjennomsnittsverdiene av X i de to gruppene. Ser man nærmere på dette uttrykket finner man at dette betyr at effekten av Z både kan overestimeres og underestimeres, og sammenhengen kan også i enkelte tilfeller snues, alt avhengig av størrelsen på målefeilen og sammenhengen mellom X og Z. Vi ser også at i tilfellet hvor $\mu_1 = \mu_2$ vil $\beta_z^* = \beta_z$. Vi estimerer altså effekten av Z korrekt i tilfellet hvor det ikke er noen korrelasjon mellom X og Z. Dette vil typisk være tilfelle i randomiserte studier, hvor den binære variabelen Z vil angi behandlingsgruppe (intervensjonsgruppe). Det bør nevnes at dersom man fortsatt bare har én variabel som måles med feil, vil effekten av denne fortsatt underestimeres.

Dersom man har en situasjon med to eller flere forklaringsvariable målt med tilfeldig feil, vil man overestimere eller underestimere effekten av disse, avhengig av størrelsen på målefeilen og korrelasjonen

mellom variablene. Man kan tenke seg en enkel situasjon med to forklaringsvariable, begge målt med tilfeldig feil, og hvor disse to variablene er positivt korrelert. Imidlertid er det slik at den ene variabelen måles med stor feil, og den andre med liten feil. Da vil man kunne oppleve at den store feilen fullstendig dominerer den mindre feilen, slik at variabelen med den lille feilen inntar samme rolle som Z over, og effekten av denne vil altså i en slik situasjon kunne overestimeres.

I en generell regresjonsmodell med en rekke variable målt både med og uten feil, kan man ikke forutsi noe om i hvilken grad effekten av de forskjellige variablene over- eller underestimeres.

Logistisk regresjon

Generelt kan man si at man vil oppleve de samme effektene av tilfeldig målefeil i logistisk regresjon som i lineær regresjon, selv om man ikke kan uttrykke skjevhetene eksplisitt på samme måte. Et unntak finnes dog: Man kan vise (Stefanski & Carroll, 1985) at i den enkle logistiske regresjonsmodellen

$$\ln(\frac{p}{1 - p}) = \beta_0 + \beta_x X$$

finnes det tilfeller hvor det å sette inn W og gjennomføre regresjonen på denne vil føre til overestimering av β_x. Imidlertid er det snakk om så vidt sære tilfeller at man sjelden vil oppleve det i praksis. Utover dette vil effektene i logistisk regresjon tilsvare effektene i lineær regresjon.

Cox regresjon, Poisson regresjon

Igjen er det slik at man vil oppleve de samme effektene i disse regresjonsmodellene som i lineær regresjon. I den enkle modellen med én forklaringsvariabel målt med tilfeldig feil, vil vi oppleve en underestimering av effekt. I Cox-modellen vil graden av underestimering være avhengig av sensureringen (Hughes, 1993), mens i Poisson-modellen vil man i tillegg også kunne oppleve såkalt overspredning på grunn av målefeilen. Dette kan være problematisk fordi det kan lede en vekk fra Poisson-modellen også i tilfeller hvor denne er korrekt (Guo & Li, 2002). I mer komplekse modeller er det igjen slik at man ikke kan forutsi om effekter vil bli over- eller underestimert.

Hypotesetesting

Det meste av den litteraturen som finnes på målefeilsområdet, både innenfor epidemiologi og statistikk, fokuserer på estimering. Nå er det et faktum at man i anvendt epidemiologi er vel så opptatt av testing, altså om man har en signifikant effekt av en gitt eksponeringsvariabel eller ikke. Det er da opplagt viktig å vite noe om i hvilken grad målefeilsproblemet påvirker hypotesetestene. Vi skal si litt kort om dette. Alle resultatene i dette avsnittet er beskrevet i Carroll et al. (1995), og gjelder generelt for regresjonsmodeller.

Vi må først definere hva det vil si at en test er valid. Med dette mener vi at testen faktisk holder det signifikansnivå vi sier at den skal holde. Det at en test er valid kan sees på som et minstekrav til en statistisk test.

Vi tar fortsatt utgangspunkt i den enkle feilmodellen $W = X + U$ hvor U har forventning null og konstant varians, altså en situasjon med ren tilfeldig støy.

Dersom vi har en regresjonsmodell med kun én forklaringsvariabel målt med feil, kan man vise at testen for 'ingen effekt' av denne variabelen er valid. Imidlertid er det opplagt slik at testen vil ha lavere styrke enn om vi hadde vært i stand til å observere den sanne underliggende X. Dette kommer både av at vi underestimerer effekten av X og av at usikkerheten i estimatet er større når vi baserer oss på den observerte variabelen W.

Dersom vi i regresjonsmodellen vår også inkluderer variable målt uten feil, er det slik at testen for 'ingen effekt' av disse generelt ikke er valid. Det kan vises at den vil være valid i enkelte spesialsituasjoner, men altså ikke generelt. I lineær regresjon vil testen være valid dersom variablene målt uten feil er uavhengige av variabelen(e) målt med feil. I andre regresjonsmodeller holder ikke dette. Testen på effekt av variabelen målt med feil er fortsatt valid.

Til slutt; dersom vi inkluderer flere variable med feil kan man heller ikke stole på at testen på de enkelte av disse variablene er valide.

Totalt sett betyr altså dette at dersom man er i en situasjon hvor man har én eksponeringsvariabel som er målt med tilfeldig støy, alle kovariater er målt uten feil, og man er kun interessert i effekten av eksponeringsvariabelen, så er man på trygg grunn dersom testen av denne viser en signifikant effekt.

MÅLEFEIL I KATEGORISKE FORKLARINGSVARIABLE

Vi har til nå kun behandlet kontinuerlig fordelte forklaringsvariable. I epidemiologi er det vel så ofte snakk om kategoriske forklaringsvariable. Når vi snakker om målefeil på kategoriske variable blir dette som regel omtalt som feilklassifisering heller enn målefeil. Effekten av feilklassifiserte forklaringsvariable er naturlig nok i hovedsak den samme som effekten av målefeil på kontinuerlige forklaringsvariable, men enkelte detaljer gjør at denne situasjonen fortjener en egen omtale.

La oss begynne med en situasjon med en enkel binær forklaringsvariabel, for eksempel røyker/ikke-røyker. Graden av feilklassifisering på en slik variabel kan beskrives ved hjelp av begrepene sensitivitet og spesifisitet i forhold til den underliggende, sanne situasjonen. Da er det slik at dersom summen av sensitiviteten og spesifisiteten til den observerte variabelen er større enn én vil effekten av forklaringsvariabelen underestimeres i en regresjonsanalyse (se f.eks. Reade-Christopher and Kupper, 1995).

Dersom vi har en kategorisk forklaringsvariabel med flere enn to kategorier, vil disse typisk analyseres i form av såkalte dummy-variable, og vi er over i situasjonen med flere forklaringsvariable målt med feil. Her er det, som i situasjonen med kontinuerlige forklaringsvariable, umulig å si noe generelt om effekten av feilklassifisering. Det eneste man kan si er at man vil underestimere effekten av den kategorien som har den sterkeste sammenhengen med responsen. En annen måte å se dette på er at man vil underestimere risikoen til den gruppen som har den høyeste eksponeringen hvis man legger til grunn en dose-respons tankegang.

Når vi i tillegg inkluderer variable målt uten feil er det også umulig å si noe generelt om effekten av feilklassifisering. Se ellers Reade-Christopher and Kupper (1991, 1995) og Veierød and Laake (2001) for detaljerte beskrivelser av effekten av feilklassifisering i forskjellige regresjonsmodeller.

MÅLEFEIL I RESPONSVARIABELEN

Så langt har vi behandlet målefeil i form av feil på forklaringsvariablene. Det er klart at man også kan oppleve feil i responsvariablene. Dette er imidlertid et felt som er mindre studert. Et par ting er verdt å merke seg.

Ser vi først på lineær regresjon og enkel tilfeldig feil, så ser vi fort at dette ikke har noen betydning i forhold til korrekt estimering av regresjonskoeffisientene. Den tilfeldige feilen i Y vil bare inngå i residualen (ε) i regresjonslikningen, og dermed medføre en større usikkerhet i estimatene. Man vil altså fortsatt ha et såkalt *forventningsrett* estimat, men standardfeilen til estimatet øker.

Når det gjelder logistisk regresjon, kan man derimot vise (se for eksempel Neuhaus, 1999) at i tilfellet med en feilklassifisert respons hvor feilen er uavhengig av forklaringsvariablene (altså ikke-differensiell feil), vil effekten av forklaringsvariablene systematisk underestimeres. Kjenner man feilklassifiseringssannsynlighetene kan man enkelt korrigere for denne feilen slik at man også her står med et forventningsrett estimat. Imidlertid kan man også i denne modellen vise at standardfeilen til estimatet øker i forhold til situasjonen uten feil i responsen (Neuhaus, 1999).

Det er verdt å bemerke at det at effekten av forklaringsvariablene vil underestimeres i logistisk regresjon gjelder i tilfellet med klart definerte utvalg eller kohorter, hvor feilklassifiseringen foregår innen utvalget. I forbindelse med pasient-kontroll studier kan man tenke seg at pasienter blir feilklassifisert som "ikke-pasienter", og dermed ikke i det hele tatt blir inkludert i studien. Da gjelder ikke lenger dette resultatet. Man kan imidlertid diskutere om dette er et feilklassifiseringsproblem eller om det er et seleksjonsproblem.

Til slutt vil vi her nevne at det selvfølgelig også er mulig å ha feil på både responsvariable og forklaringsvariable samtidig. Spesielt kan disse feilene være korrelerte. Dette tilfellet er behandlet inngående av Schaalje og Butts (1993) i lineær regresjon. Kristensen (1992) behandler et tilfelle som er relevant for logistisk regresjon ut fra en tabell tankegang. I begge situasjoner er det slik at skjevheten i effektestimatene kan gå i både positiv og negativ retning.

KORREKSJON

Når man ser hvilke problemer målefeil fører med seg i forbindelse med regresjonsanalyse vil man ønske seg metoder for å komme unna disse problemene. Den mest opplagte løsningen er selvfølgelig å forsøke å unngå feilene. Mye feil kan man unngå ved å bruke validerte måleinstrumenter. Imidlertid er det opplagt slik at man uansett vil sitte igjen med en mengde målefeil det ikke er mulig å gjøre noe med. Neste skritt vil da være å forsøke å korrigere for disse feilene. I lineær regresjon kan man i mange tilfeller gjøre dette relativt enkelt. Igjen kan vi referere til Fuller (1987) for en oversikt. I andre regresjonsmodeller er det vanskeligere, men det har etter hvert blitt utarbeidet en del forslag til slike korreksjonsmetoder. Til tross for dette finnes det meget få eksempler på at slike korreksjoner er gjennomført i praktiske epidemiologiske arbeider. Det er antagelig i hovedsak to grunner til dette; metodene er for det meste publisert i statistiske tidsskrifter og beskrivelsen av dem er ofte relativt teknisk, og metodene har ikke vært implementert i standard programvare. Situasjonen er nå noe endret, i det pakken Stata har implementert to forskjellige metoder i spesielt tilpasset software, samt at noe også er mulig å gjøre i modulen 'gllamm' (se annet sted i dette nummer). Vi vil derfor bruke noe plass på å kort beskrive de mest populære metodene. Se ellers Carroll et al. (1995) for en innføring i korreksjonsmetoder generelt.

Regresjonskalibrering

Av de mange mer 'ad-hoc' metoder som er publisert, er regresjonskalibrering den metoden som har fått mest gjennomslag. Dette er også en av de metodene som nå er implementert i Stata. Tanken bak denne metoden er at om man ikke kan observere forklaringsvariabelen X direkte, må man kunne estimere (eller predikere) X på bakgrunn av det man faktisk klarer å måle (W og andre forklaringsvariable). Når man har gjort dette kan man gjennomføre den analysen man opprinnelig hadde tenkt seg, med den estimerte verdien av X. Denne metoden synes å virke rimelig bra i veldig mange situasjoner. Unntaket er situasjoner med veldig skjevfordelte forklaringsvariable. Det må også bemerkes at dette er en metode for å behandle målefeil på kontinuerlige forklaringsvariable. Sentrale referanser er Rosner et al. (1989, 1990).

Simex

Den andre korreksjonsmetoden som er implementert i Stata er den såkalte Simex (Simulation-Extrapolation) metoden. Dette er en simuleringsbasert metode, som

kun tar hånd om situasjoner med enkel, tilfeldig feil ($W = X + U$). Metoden er besnærende enkel. Den forutsetter at man har informasjon om variasjonen i U. Basert på denne informasjonen kan man simulere ytterligere målefeil og etablere en trend i de estimerte regresjonskoeffisientene. Basert på denne trenden kan man så ekstrapolere seg tilbake til situasjonen uten målefeil. Den sentrale referansen her er Cook & Stefanski (1994).

Maksimum likelihood

Såkalt likelihood tankegang er statistikkens svar på mange estimeringsproblemer. Når man estimerer effekten av en gitt eksponeringsvariabel i f.eks. logistisk regresjon, så gjøres dette gjennom å maksimere en såkalt likelihood funksjon. Imidlertid synes ikke dette for den vanlige bruker, da alle slike tekniske detaljer ligger innbakt i programvaren. Å bruke likelihood tankegangen i målefeilssituasjonen innebærer at man må ta stilling til hvordan forholdet er mellom X og W. Man må altså spesifisere en modell for feilen. I tillegg må man spesifisere en fordeling for den sanne underliggende variabelen X, gitt variablene målt uten feil. Dette siste kan være vanskelig. Det vil ofte være naturlig å påstå at denne er normalfordelt, men dette vil i stor grad være gjetning.

Det er også slik at estimering i disse modellene er forbundet med relativt store numeriske problemer, og dette er fortsatt et aktivt forskningsfelt.

Likelihood estimering i målefeilstilfellet kan i mange situasjoner gjøres i 'gllamm' modulen i Stata (Rabe-Hesketh et al. 2003), som også har en mulighet for å la fordelingen til X være uspesifisert.

En viktig ting å legge merke til i forbindelse med disse korreksjonsmetodene er at man vil trenge ekstra informasjon for å kunne gjennomføre dette. I situasjonen med den helt enkle tilfeldige feilen ($W = X + U$) holder det med å ha gjentatte målinger av W for et utvalg personer. På bakgrunn av disse gjentakene kan man estimere variansen til U. I mer kompliserte situasjoner er man avhengig av å kunne måle den sanne X (eller i alle fall X målt med bare tilfeldig feil) på et utvalg, slik at man kan si noe om forholdet mellom den sanne X og målingen W. Dette betyr at hvis man skal forsøke å korrigere for målefeil, bør man ta hensyn til dette allerede i planleggingsfasen av en studie.

OPPSUMMERING

Hensikten med denne artikkelen har vært å gjøre et forsøk på å oppsummere effekten av målefeil/feilklassifisering i forbindelse med de regresjonsmodeller som er mest aktuelle i epidemiologiske studier.

Først og fremst har vi forsøkt å peke på at den såkalte attenueringseffekten som man i epidemiologisk sammenheng har vært opptatt av, altså det at man underestimerer effekter på grunn av målefeil, ikke er et så generelt fenomen som man ofte får inntrykk av. Det er tvert i mot slik at man kun i spesielt enkle modeller kan stole på at man har denne effekten.

Videre har vi forsøkt å påpeke at målefeil også påvirker både estimering og testing av effekt av forklaringsvariable målt uten feil. Dette er det dessverre veldig lett å overse.

Til slutt har vi også beskrevet metoder for å korrigere for målefeil, og vi har påpekt at for at dette skal kunne la seg gjøre er man avhengig av noe informasjon om forholdet mellom den sanne variabelen og den observerte variabelen. Dette betyr at man er nødt til å ta hensyn til målefeil allerede i planleggingsfasen av en studie.

Vi har kun behandlet såkalte ikke-differensielle feil. Man skal være klar over at differensielle feil kan oppstå, spesielt i forbindelse med case-control studier, og at det i slike tilfeller er betraktelig vanskeligere både å forutsi effekten av feilene og å korrigere for dem.

Man kan selvfølgelig spørre seg om målefeil egentlig har noen vesentlig praktisk betydning. Det er opplagt slik at man i epidemiologi har store nok problemer som man har, med å få kontroll over alle mulige konfunderende variable osv. Det er imidlertid også opplagt at man i de fleste epidemiologiske studier opererer med til dels meget store målefeil, og man skal ikke se bort fra at mange av de noe motstridende funn man opererer med kan skyldes nettopp dette. Se f.eks. Michels (2001) for slike betraktninger.

REFERANSER

Bross I, 1954. Misclassification in 2 x 2 tables. *Biometrics* **10**, 478-486.

Carroll RJ, 1989. Covariance analysis in generalized linear measurement error models. *Statistics in Medicine* **8**, 1075-1093.

Carroll RJ, Stefanski LA, 1994. Measurement error, instrumental variables and corrections for attenuation with applications to meta-analyses. *Statistics in Medicine* **13**, 1265-1282.

Carroll RJ, Ruppert D, Stefanski LA, 1995. *Measurement error in nonlinear models*. London: Chapman and Hall.

Cook JR, Stefanski LA, 1994. Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association* **89**, 1314-1328.

Flegal KM, Keyl PM, Nieto FJ, 1991. Differential misclassification arising from nondifferential errors in exposure measurement. *American Journal of Epidemiology* **134**, 1233-1244.

Fuller WA, 1987. *Measurement error models*. New York: Wiley.

Guo JQ, Li T, 2002. Poisson regression models with errors-in-variables: implication and treatment. *Journal of Statistical Planning and Inference* **104**, 391-401.

Hughes MD, 1993. Regression dilution in the proportional hazards model. *Biometrics* **49**, 1056-1066.

Kristensen P, 1992. Bias from nondifferential but dependent misclassification of exposure and outcome. *Epidemiology* **3**, 210-215.

MacMahon S, Peto R, Cutler J, Collins R, Sorlie P, Neaton J, Abbott R, Godwin J, Dyer A, Stamler J, 1990. Blood pressure, stroke, and coronary heart disease. Part 1, prolonged differences in blood pressure: prospective observational studies corrected for the regression dilution bias. *Lancet* **335**, 765-774.

Michels KB, 2001. A renaissance for measurement error. *International Journal of Epidemiology* **30**, 421-422.

Neuhaus JM, 1999. Bias and efficiency loss due to misclassified responses in binary regression. *Biometrika* **86**, 843-855.

Prentice RL, 1982. Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika* **69**, 331-342.

Rabe-Hesketh S, Skrondal A, Pickles A, 2003. Maximum likelihood estimation of generalized linear models with covariate measurement error. *Stata Journal*, In press.

Reade-Cristopher SJ, Kupper LL, 1991. Effects of exposure misclassification on regression analyses of epidemiologic follow-up study data. *Biometrics* **47**, 535-548.

Reade-Cristopher SJ, Kupper LL, 1995. On the effects of predictor misclassification in multiple linear regression analysis. *Communications in Statistics – Theory and Methods* **24**, 13-37.

Rosner B, Willett WC, Spiegelman D, 1989. Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. *Statistics in Medicine* **8**, 1051-1069.

Rosner B, Spiegelman D, Willett WC, 1990. Correction of logistic regression relative risk estimates and confidence intervals for measurement error: The case of multiple covariates measured with error. *American Journal of Epidemiology* **132**, 734-745.

Rosner B, 1996. Measurement error models for ordinal exposure variables measured with error. *Statistics in Medicine* **15**, 293-303.

Schaalje GB, Butts RA, 1993. Some effects of ignoring correlated measurement errors in straight line regression and prediction. *Biometrics* **49**, 1262-1267.

Stefanski LA, Carroll RJ, 1985. Covariate measurement error in logistic regression. *Annals of Statistics* **13**, 1335-1351.

Veierød MB, Laake P, 2001. Exposure misclassification: bias in category specific Poisson regression coefficients. *Statistics in Medicine* **20**, 771-784.

Willett WC, Hunter DJ, Stampfer MJ, Colditz G, Manson JE, Spiegelman D, Rosner B, Hennekens CH, Speizer FE, 1992. Dietary fat and fiber in relation to risk of breast cancer. *Journal of the American Medical Association* **268**, 2037-2044.