TØI notat 1161/2000
Forfattere: Jens Rekdal og Tom N Hamre
Oslo 2000, 108 sider

Sammendrag:

Noen problemstillinger knyttet til estimering av transportmodeller

I dette notatet tar vi opp noen sentrale problemstillinger knyttet til estimering av transportmodeller med logit-metodikk. Først i notatet ser vi på konsekvenser for modellene når de estimeres på **data som inneholder målefeil**. Et datasett etablert på tradisjonelle RVU-er vil alltid inneholde ulike former for målefeil, spesielt når man har benyttet nettverksmodeller til å beregne såkalte "transportstandard-variable" som skal inngå i estimeringen. Deretter henledes oppmerksomheten mot **konsekvenser av varierende restleddsvarians**. I standard økonometri kalles dette fenomenet for *heteroskedastisitet*. Når man estimerer transportmodeller ved hjelp av logit-metodikk, forutsettes det at restleddene i de såkalte nyttefunksjonene er identisk og uavhengig fordelt. Dette er en meget streng forutsetning spesielt når det er store variasjoner i reiseavstandene. Til sist i notatet ser vi på mulige **alternative formuleringer og strukturer** i estimering av transportmodeller for valg av transportmiddel for lange reiser. Dette arbeidet er knyttet opp mot mulige forbedringer av de modeller som i dag er implementert i Den nasjonale persontransportmodellen.

Målefeil i data

Problemer med målefeil i variable er et område som har vært systematisk undersøkt når det gjelder *standard* økonometriske problemstillinger, og et generelt resultat er at hvis man har en variabel med *store tilfeldige* målefeil, vil man, selv om variabelen i gjennomsnitt er målt/registrert riktig, få forventningsskjeve resultater for variabelens koeffisient *og også for andre koeffisienter i modellen.* Problemet med tilfeldige målefeil er etter vår oppfatning ikke i tilsvarende grad tatt alvorlig i modeller for diskrete valg som for eksempel i regresjonsmodeller. Vi har åpenbart betydelige måleproblemer når det gjelder en rekke av de viktigste variablene i transport-modellene. Helt markert blir dette blant annet for gangavstander til holdeplasser når vi i stedet for den korrekte avstanden bruker et sonegjennomsnitt beregnet ved hjelp av et kodet transportnett og en rutevalgsmodell.

I dette arbeidet har vi konsentrert oss om en enkel binær reisemiddelvalgmodell. Modellen estimeres på et datasett som er syntetisk i den forstand at det er manuelt etablert med utgangspunkt i statistikk for reisevaner og nettverksdata. Vi forutsetter at respondentene i datamaterialet har homogene preferanser og at alle forutsetninger for estimering av logitmodeller er oppfylt. Vi forutsetter videre at datasettet inneholder fullstendig korrekte opplysninger om de reisene respondentene har gjennomført og om mulige alternative reisemåter.

Notatet kan bestilles fra:
Transportøkonomisk institutt, Postboks 6110 Etterstad, 0602 Oslo
Telefon: 22 57 38 00 Telefax: 22 57 02 90

Datasettet representerer individer bosatt Oslo-området og omfatter bare 200 observasjoner. På disse dataene er det estimert en modell som vi betrakter som "sann", i den forstand at alle data og observasjoner er korrekte og alle forutsetninger for modellen er oppfylt. Denne modellen benyttes så til å simulere effektene av målefeil i gangtidsvariabelen og effektene av ulik spesifisering av kostnadsvariabelen for individer med månedskort. Simuleringene er gjennomført i programmeringsspråket GAUSS som såkalte Monte Carlo-simuleringer. Teorien for logitmodeller forutsetter at nyttefunksjonene for hvert enkelt individs alternativer (bil og kollektivt i vårt tilfelle) har et Gumbelfordelt restledd. Dette restleddet fanger i teorien opp uobserverte og utelatte variable, målefeil, preferanseforskjeller m m. Den sanne modellen representerer de valgene individene har foretatt og som er beskrevet i datasettet. Simuleringene innebærer at vi for hver observasjon og hvert transportmiddel trekker et uniformt fordelt tall i intervallet 0 til 1 og transformerer dette til Gumbelfordelingen. Sammen med den sanne modellen og dataverdiene for hvert individ benyttes de Gumbelfordelte tallene til beregning av et nytt simulert transpormiddelvalg for hvert individ. Så estimeres en ny modell på dette transportmiddelvalget.

Simuleringene gjennomføres parallelt på det sanne datasettet og på fem andre datasett som er identiske med det sanne med unntak av den variabelen som inneholder målefeil. I fire av disse datasettene inneholder gangtidsvariabelen ulike former for feil, i ett av dem er gangtid utelatt og i det siste datasettet inneholder kostnadsvariabelen for kollektivtransport gjennomsnittlig pris pr reise for de individer som har månedskort, i motsetning til i den sanne modellen hvor kostnaden for disse trafikantene er satt lik 0. Vi har gjennomført 1000 simuleringer parallelt (det vil si med samme Gumbelfordelte restledd) på hvert av de fem datasettene. Resultatet fra simuleringene blir dermed parameterverdier for 1000 modeller for hvert datasett.

Vi har sett på følgende tre typer målefeil i gangtidsvariabelen:

"p" $T_P = T + e_P$,
"1" $T_1 = aT + e_1$
"2" $T_2 = T + k + e_2$

Feiltype "p" innebærer at vi erstatter de sanne gangtidene, *T*, med gangtider hentet fra et transportnett med standard soneinndeling i transportnettene for Oslo-området (PROSAM-soner). Leddet e_p kan vi betrakte som en stokastisk variabel med ukjent fordeling, bortsett fra at min $e_p > -T$ og max $e_p < T^*$ hvor T^* vil avhenge av sonestørrelsen. Feiltype "1" og "2" er generert med a = 0,9, k=-5, e_1 et tilfeldig trukket tall uniformt fordelt mellom 0 og 1, og e_2 et tilfeldig trukket uniformt fordelt tall mellom 0 og 5.

Resultatene av simuleringene viser at målefeilene i variabelen for gangtid ikke bare slår ut på parameterverdien for gangtid, men også på parametre for de andre variablene i modellen. Dette gjør at forholdet mellom parameterne (tidsverdier, gangtidsvekt, ventetidsvekt og vektfaktor for omstigninger) varierer betydelig ved ulike typer målefeil. I feiltype "1" er den tilfeldige feilen liten (under 1 minutt), mens den systematiske feilen er 10 % av den sanne gangtiden. Denne type feil har størst effekt på gangtidsparameteren. I feiltype "2" er den tilfeldige feilen opp til 5 minutter. Denne type feil slår ut ganske kraftig på alle parameterverdier i modellen. Det samme gjør feiltype "p". Når man har variable man vet inneholder målefeil, ser

det heller ikke ut til å være noen heldig løsning å droppe variabelen i modellen. Dette påvirker parameterverdiene for de andre variable i relativt stor grad.

Varierende restleddsvarians (heteroskedastisitet)

Gjennom arbeidet med Den nasjonale persontransportmodellen (NTM), har vi flere ganger fått erfare at det er problematisk å estimere modeller og matriser for persontrafikk mellom kommunene i Norge. Ett av problemene synes å være knyttet til de store spenn i avstand (og dermed reisetid og reisekostnad) som kan hevdes å være spesielt for det langstrakte Norge. Eksempelvis kan det være grunnlag for å hevde at karakteristika ved en tromsøværings reisemiddelvalg for en reise til Bergen, Stavanger eller Oslo vil være av en helt annen art enn det som karakteriserer et tilsvarende valg for en reise til Bodø, Harstad eller Alta.

Estimering av logit-modeller skjer under en implisitt forutsetning om at fordelingen av restleddene i nyttefunksjonene er uavhengig av reisemåte og reiselengde. Dette er en spesielt dristig forutsetning i den nasjonale langdistansemodellen, som spenner over et distanseintervall på 100-3000 km, men det kan også være en kritisk forutsetning for modellen for reiser i intervallet 0-100 km. Hvis det er en systematisk sammenheng mellom restleddsvarians og for eksempel reiselengde, bør man ta hensyn til dette på samme måte som ved estimeringsopplegg hvor man bruker vanlig regresjonsanalyse. En forholdsvis enkel måte å undersøke betydningen av dette problemet på er å sammenlikne "standard" estimeringsprosedyre med alternativer hvor man løsner på forutsetningen om konstant restleddsvarians.

I dette notatet har vi studert denne problemstillingen først på et aggregert datasett etablert med utgangspunkt i OD-matriser og nettverksdata fra Den nasjonale persontransportmodellen, og siden på et datasett bestående av individdata fra RVU 91/92.

I det første datasettet ble det kun estimert modeller for valg av transportmiddel. Arbeidet viste at modeller med lik modellformulering, men med ulike transformasjoner på variablene er sammenliknet med tilsvarende modeller uten transformasjoner på datamaterialet. Det er estimert modeller med følgende tre transformasjoner av data:

B) $f_B(ldist) = (ldist)^{-0.5}$
C) $f_C(ldist) = (ldist)^{-0.2}$
D) $f_D(ldist) = (ldist)^{-0.7}$

Her betegner (ldist) luftlinjeavstanden mellom sonecentroidene for de aktuelle kommunene. B-, C- og D-modellene er sammenliknet med en tilsvarende A-modell som er estimert uten transformasjon av variablene.

Det fremgår av arbeidet at det er vesentlig lettere å tilpasse modeller som er estimert på data hvor alle variabler er dividert på kvadratroten av luftlinjeavstand mellom soner[1], enn modeller som er estimert på data uten en slik transformasjon. Hvis transformasjonen er mindre enn variabelverdiene dividert på kvadratroten av luftlinjeavstand mellom sonene, for eksempel $(ldist)^{-0.2}$, er effekten av trans-

[1] Denne forutsetningen er ekvivalent med en forutsetning om at variansen på restleddet er proporsjonal med luftlinjeavstanden mellom sonecentroidene.

formasjonen mindre. Hvis transformasjonen er større enn variabelverdiene dividert på kvadratroten av luftlinjeavstand mellom sonene, for eksempel $(ldist)^{-0.7}$, er effektene noe mer uviss. Høyere (absolutte) log-likelihood-verdier på den sist omtalte transformasjonen, tyder imidlertid på at transformasjonen $(ldist)^{-0.5}$, gjennomgående ser ut til å gi best resultater. I og med at vi opererer på et svært høyt aggregeringsnivå, både når det gjelder venstresidevariable (totalt antall reiser mellom kommuner) og selve variablene (hentet fra grovt kodede nasjonale nettverksrepresentasjoner), vil vi imidlertid være forsiktige med å trekke for sterke konklusjoner basert bare på t-verdier og log-likelihood-verdier.

Arbeidet ble derfor videreført, men denne gang på individdata. Vi benyttet her en datafil for lange private reiser, som også er benyttet til å estimere dagens versjon av denne delmodellen i Den nasjonale persontransportmodellen (NTM4c). En vesentlig del av ressursene til denne oppgaven var nødvendig å bruke til finne ut av innholdet i datafiler og programoppsett og til å konvertere ulike datafiler mellom binært format og vanlig ascii-format.

På samme måte som for de aggregerte data ble det estimert modeller med og uten transformasjon av alle data som inngår i de ulike nyttefunksjonene. Variablene ble imidlertid bare transformert med kvadratroten av luftlinjedistansen, fordi det var denne transfomasjonen som var mest lovende i arbeidet med de aggregerte data. Transformeringen viste også her mer signifikante parameterestimater, i tallverdi høyere parameterestimater, og lavere likelihood-verdier for de modeller som er estimert på transformerte data. Med de programpakkene som er tilgjengelige i dag for estimering av logitmodeller, er det av tekniske årsaker ikke mulig å simultant estimere modeller for transportmiddelvalg og destinasjonsvalg på transformerte data. Vi har imidlertid sekvensielt estimert modeller for transportmiddelvalg og destinasjonsvalg på data med og uten transformasjoner. Også modellene for destinasjonsvalg estimert på transformerte data viser bedre statistiske egenskaper enn tilsvarende modeller uten transformasjon.

Vi har på ingen måte kommet helt i mål når det gjelder hypotesen om hetero-skedastisitet i nyttefunksjoner. Vi mener imidlertid at arbeidet knyttet til skalering av nyttefunksjonene viser lovende resultater. I de aggregerte modellene ble det imidlertid arbeidet på et svært høyt aggregeringsnivå, mens forsøkene med de disaggregerte data ikke var så omfattende (bare ett reiseformål og én type transformasjon). Når det gjelder estimering på disaggregerte data, spiller segmentering en stor rolle (for eksempel etter varighet på destinasjonen, reise-avstand m m). Den nye reisevaneundersøkelsen (1997/98) kan danne utgangspunkt for videre analyser på dette felt.

Alternative modellstrukturer og formuleringer av langdistansemodeller

I dette arbeidet har vi tatt for oss to av reiseformålene i dagens versjon av Den Nasjonale Transportmodellen, fritidsreiser og private reiser. Disse to reiseformålene kan hevdes å være litt mer heterogene i "sin natur" enn de tre resterende formålene som er tjenestereiser, pendling og besøksreiser. De private reisene er en samlekategori for innkjøpsreiser, reiser knyttet til medisinske tjenester, reiser hvor man skal følge andre, og andre private reiser, mens fritidsreisene omfatter reiser i forbindelse med fornøyelse/underholdning, organisert fritidsaktivitet (idrett, politikk, religion etc), ferie og fritidsreiser.

For begge de to reiseformålene er det estimert modeller for valget av transportmiddel. I estimeringsarbeidet har vi spesielt sett på behandlingen av intervjuobjektenes inntekt i transportmiddelvalget, formuleringen av variable for avgangsfrekvens for kollektive transportmidler, alternative modellstrukturer og segmentering etter reiseavstand. For alle disse problemstillingene har vi også studert de implisitte tidsverdiene i de ulike modellene. Det er flere årsaker til at man ønsker å ha med både inntekt og reisekostnad i modeller for reisemiddelvalg. Her skal vi kort nevne tre av dem.

- Inntekten kan påvirke trafikantenes reisemiddelvalg dels fordi verdsettingen av spart reisetid øker med inntekten, og dels fordi betalingsevnen øker.
- Transformering av kostnader med inntekt kan bidra til å bryte en vanligvis sterk korrelasjon mellom reisekostnad og reisetid.

Det første punkt er en mer prinsipiell begrunnelse, mens det siste er et rent modellteknisk ”triks” for å oppnå mer presist bestemte parametre og bedre modellegenskaper. I de reisevaneundersøkelser som benyttes til å estimere denne type modeller blir respondentene vanligvis bedt om å oppgi personlig brutto årsinntekt og/eller brutto inntekt for husstanden samlet. Problemet er her at brutto inntekt sjelden vil være noe godt mål på hvor mye et individ eller en husstand har disponibelt. Bruken av inntekt som variabel i modeller for valg av transportmiddel kan derfor gjøre det vanskelig å oppnå realistiske implistitte tidsverdier i modellene. I dagens langdistansemodeller i NTM4c inngår variable hvor reisekostnader er dividert med ulike inntektsbegrep. I mange av modellene blir derfor tidsverdiene svært høye. I dette arbeidet har vi derfor undersøkt om det er mulig å formulere modellene på en slik måte at man både kan inkludere inntekt og få mer rimelige tidsverdier. Arbeidet har vist at dette er mulig. Spesielt viser formuleringer av typen inntekt minus reisekostnad lovende resultater.

I NTM4c inngår avgangsfrekvens for kollektive transportmidler inklusive fly, med lineære spesifikasjoner. Dette innebærer at en ekstra avgang får samme effekt uavhengig av hvor mange avganger som går fra før. Sannsynligvis får vi dermed for små effekter på etterspørselen av en økning i frekvens der frekvensene er lave i utgangspunktet, og for store etterspørselseffekter på det høyfrekvente transporttilbud. Analysene som er gjennomført i dette arbeidet tyder på at det med enkle grep er mulig å formulere mer realistiske funksjoner for avgangsfrekvens. Dagens lineære formulering ser imidlertid ut til å gi modeller med de beste statistiske egenskaper. Modellene for reisemiddelvalg i NTM4c har en enkel multinomisk struktur. Bak denne strukturen ligger en forutsetning om at nyttefunksjonene for alle alternativer, dvs transportmidler, har identisk varians. Dette er en kritisk forutsetning, da vi som oftest vil ha ikke-observerte faktorer som i noen grad er korrelert mellom alternativene. Det kan for eksempel være at tog og buss ”likner” mer på hverandre og har større konkurranseflater enn bil i gitte valgsituasjoner, mens tog, buss og bil kanskje igjen er mer innbyrdes like enn fly.

Den beste strategi i slike tilfeller vil selvsagt være å forsøke å observere disse faktorene eller forholdene slik at de kan inngå i den multinomiske modellen tilknyttet de aktuelle transportmidler. Alternativt kan man ta hensyn til dette ved å undersøke andre modellstrukturer enn den multinomiske. ”*Nested logit*”, eller på norsk strukturerte logitmodeller, er et viktig hjelpemiddel i så henseende. I disse modellene

kan flere transportmiddel defineres som egne alternativer, og dermed danne egne sammensatte alternativer. Forsøkene med strukturerte logitmodeller i dette arbeidet viser at det kan være relativt mye å hente både når det gjelder bedre statistiske egenskaper og mer realistiske tidsverdier ved å estimere denne type modeller.

Til slutt i dette notatet har vi forsøkt med ulike segmentinger i datagrunnlaget med hensyn på reiseavstand. Vi gjør dette for å gå litt dypere inn i problemet med varierende restleddsvarians, som vi har vært opptatt av i hele dette notatet. Vi tror som nevnt at de problemer vi kan se i Den Nasjonale Persontransportmodellen skyldes at man i spesielt i Norge kan gjennomføre lange reiser i et stort avstandsintervall. Vår hypotese har hele tiden vært at variansen til det stokastiske restleddet i nyttefunksjonen ikke er identisk fordelt men til en viss grad varierer med reiseavstand. I tidligere arbeid har vi forsøkt å ta hensyn til dette ved å transformere alle variable med en funksjon av avstand, med brukbare resultater.

I dette arbeidet skal vi se nærmere på hva som skjer når vi heller deler inn datamaterialet etter hvor langt respondentene har reist. Vi har gjort dette på to ulike måter. For det første har vi fysisk sett delt datamaterialet inn i to grupper, en gruppe som har reist kortere enn 250 km og en som har reist lengre, og estimert egne modeller for hver av gruppene. Den andre formen for segmentering innebærer at vi estimerer en parameter for kostnad på inntekt for de som har reist kortere og en parameter for de som har reist lengre enn 250 km. I det siste tilfellet beholdes altså datamaterialet samlet. Resultatene fra arbeidet indikerer ganske sterkt at det er problematisk å estimere modeller som dekker alle reiser fra 100 til 3000 km, slik modellene i dagens versjon av transportmodellen (NTM4c) er estimert. Å segmentere etter reiseavstand kan være en av flere mulige strategier for å komme unna dette problemet. En annen strategi kan være å transformere nyttefunksjonene med en funksjon av reiseavstand slik vi har vært inne på tidligere. En tredje mulig strategi vil være å undersøke om det er alternative måter å dele inn datamaterialet på. Som nevnt er datamaterialet som ligger til grunn for dagens nasjonale modeller inndelt etter reiseformål. I Sverige har man ved estimeringen av langdistansemodellene i SAMPERS, heller forøkt å inndele datamaterialet etter varigheten på bestemmelsesstedet. Dette kan kanskje være veien å gå også i norske modeller.