Notater

Statistisk sentralbyrå

93/10 Mars, 1993

Metode for å skille mellom reelle endringer og målefeil ved analyse av paneldata

av

Ib Thomsen og Dinh Quang Pham *

Avdeling for personstatistikk
Seksjon for metoder og standarder

* En kortversjon på engelsk er publisert i IASS's proceedings fra ISI's møte i Cairo 1991, vol. III

Innhold

Alle som har forsøkt å estimere bruttostrømmer mellom to tidspunkter ved hjelp av paneldata, har støtt på problemer knyttet til forekomsten av målefeil. Det er kjent at slike feil ofte gjør det umulig å estimere slike strømmer, noe som er alvorlig, da en viktig grunn for å bruke paneldata nettopp er ønske om å estimere bruttostrømmer. Nesten samtlige land som utfører arbeidskraft-undersøkelser, bruker roterende utvalg, men tall for bruttostrømmer på arbeidsmarkedet publiseres bare i noen få land, og da kombineres data fra arbeidskraftundersøkelsene med andre data, først og fremst kvalitetskontrollundersøkelser. Aaberge (1988).

Nedenfor skal det presenteres noen enkle metoder for estimering av målefeil når en har data for de samme personer for flere enn 2 tidspunkter. I avsnitt 2 gis en kort presentasjon av Markov-modeller uten målefeil. I avsnitt 3 introduseres Markov-modeller med målefeil, såkalte latente Markov-modeller. Relasjonene mellom de observerte størrelsene og parametrene settes opp. I noen tilfeller er det mulig å bestemme estimatorene for parametrene eksplisitt på grunnlag av disse relasjoner, de såkalte momentestimatorer. Disse løsninger er det vanskelig å generalisere, og vi skal derfor bruke en E-M algoritme for å finne estimatorer for parametrene.

I avsnitt 4 er metoden brukt på data fra arbeidskraftundersøkelsene og levekårsundersøkelsene. I dette notatet er de bare brukt som eksempler på bruk av metoden. I senere notater er det tanken å undersøke om metoden gir så gode resultater at de kan danne grunnlag for en løpende publisering av bruttostrømmer fra arbeidskraft-undersøkelsene. Dette vil bli undersøkt ved å bruke metoden på data fra flere år og studere stabiliteten av resultatene.

Når det gjelder levekårsundersøkelsene, er det meningen å undersøke målefeilene til samtlige spørsmål. Ettersom levekårsundersøkelsene inneholder spørsmål fra mange felter som helse, boforhold, arbeidsforhold osv., vil kvalitetsanalyse av disse spørsmålene være til nytte for en lang rekke undersøkelser.

2. En enkel Markov-modell uten målefeil

Anta vi har observasjoner for samme personer av en kategorisk variabel på flere tidspunkter. Variabelen har k forskjellige verdier, 1,2, ..., k. Når en skal studere utviklingen i en slik variabel ved hjelp av paneldata, starter en ofte med en enkel Markov-modell. La oss se på et enkelt eksempel med k=2 fra levekårsundersøkelsene. Her finnes et spørsmål om personen er plaget av mange tunge løft på arbeidet. Svarene etter to undersøkelser kan da settes opp i en enkel tabell.

Tabell 1. Personer fordelt etter om de var plaget av tunge løft i 1980 og 1983. Prosent

		TUNGE LØFT 1983	
TUNGE LØFT 1980		JA 1	NEI 2
	Ja.1	23,0	8,5
	Nei.2	8,2	60,3

Tallene i tabell 1 er estimater for parametrene

p_{ij} (1,2): Sannsynligheten for å anta verdien i på tidspunkt 1 og verdien j på tidspunkt 2. (i=1,2;j=1,2)

En Markov-modell bruker andre parametre, nemlig følgende:

p_i: Sannsynligheten for å anta verdien i på tidspunkt 1. (i=1,2)

Dessuten

m_{ij}: Sannsynligheten for å anta verdien j på tidspunkt 2, gitt at verdien var i på tidspunkt 1.

Det er lett å se at

$$p_1 = p_{11}\ (1,2) + p_{12}\ (1,2), \qquad (1)$$

$$p_2 = p_{21}\ (1,2) + p_{22}\ (1,2) = 1-p_1,$$

$$m_{11} = p_{11}\ (1,2)/p_1\ ;\ m_{12} = p_{12}\ (1,2)/p_1 \qquad (2)$$

$$m_{21} = p_{21}\ (1,2)/p_2\ ;\ m_{22} = p_{22}\ (1,2)/p_2.$$

Dessuten er

$$m_{11} + m_{12} = 1$$

og (3)

$$m_{21} + m_{22} = 1.$$

Fra tabell 1 får en følgende estimater for parametrene:

$$\hat{P}_1 = \hat{P}_{11}(1,2) + \hat{P}_{12}(1,2) = 0,315 \quad (4)$$

$$\hat{m}_{11} = \hat{P}_{11}(1,2)/\hat{P}_1 = 0,73 \quad (5)$$

$$\hat{m}_{22} = \hat{P}_{22}(1,2)/\hat{P}_2 = 0,88 \quad (6)$$

Når en velger disse parametrene gitt ved (1), (2) og (3) i Markov-modeller, er det for å kunne anvende modellen når det foreligger data fra flere tidspunkter. Markov-forutsetningen er knyttet til disse parametre, og den viktigste er at overgangssannsynlighetene er konstante over tid. Dette er en meget streng forutsetning, og det har da også vist seg at svært få prosesser i samfunnet lar seg beskrive av enkle Markov-modeller.

Fra tabell 1 lar det seg vanskelig gjøre å teste om Markov-forutsetningen er oppfylt. Fra levekårsundersøkelsene er det imidlertid mulig å lage en tilsvarende tabell for årene 1983 og 1987. Tallene i denne tabellen er praktisk talt identiske med tallene i tabell 1. Dermed blir estimatene for overgangssann-

synlighetene også de samme, og det kan se ut som om Markov-forutsetningen er oppfylt. Når en setter opp en tabell som tabell 1 for årene 1980 og 1987, altså over utviklingen i løpet av to perioder, blir resultatene igjen svært lik tallene i tabell 1. Dette er ikke konsistent med Markov-forutsetningene.

Hvis Markov-forutsetningen er oppfylt har vi

$$P_{11}(1,3) = P_1 m_{11}^2 + P_1 m_{12} m_{21},$$

hvor p_{ij} (1,3) er sannsynligheten for å anta verdien i på tidspunkt 1 og verdien j på tidspunkt 3. Innsettes estimatene for p_1, m_{11} og m_{21} i dette uttrykket, fås

$$\hat{P}_{11}(1,3) = 0,315 \cdot 0,73^2 + 0,315 \cdot 0,27 \cdot 0,12 = 0,178,$$

Som er vesentlig mindre enn den observerte verdi. Også for de andre p_{ij} (1,3) finner en lignende avvik mellom observert verdi og de forventede under en Markov-modell.

Lignende konklusjoner er gjort i en rekke andre forsøk på å bruke Markov-modeller for å forklare utviklingen i en kategorisk variabel. Det har derfor vært nødvendig å modifisere den enkle Markov-modell. I det følgende skal vi innføre målefeil i modellen.

Vi tenker oss at det er en enkel Markov-modell som "ligger bak" observasjonene, men at disse forstyrres av at respondentene svarer feil. Til slutt skal vi oppheve forutsetningen om konstante overgangssannsynligheter og likevel estimere samtlige parametre.

3. Latent Markov-modell

Parametrene i en Markov-modell er sannsynlighetene for hver av de k mulige verdier på tidspunkt 1, p_i (i=1,2,..., k). Dessuten er det overgangssannsynlighetene m_{ij} definert ovenfor.

I tillegg innføres nå svarsannsynligheter på følgende måte:

La

q_{ij}: Sannsynligheten for å svare j gitt at den sanne (latente) verdi er i. (i=1,2,...,k; j=1,2,...,k).

Dvs. at $\sum_{j=1}^{K} q_{ij} = 1$ *for* $i=1,2,\ldots,k$.

Parametrene m_{ij} (i=1,2,...k; j=1,2,...k) beskriver strømmene mellom de latente verdier, mens q_{ij} (i=1,2,...,k; j=1,2,...k) beskriver sammenhengene mellom latent verdi og observert, manifest, verdi. I avsnittet foran ble det vist hvordan m_{ij} og p_i (i=1,2,...,k; j=1,2,...,k) lett lot seg estimere i en enkel Markov-modell ut fra estimater for

$p_{ij}(1,2)$: Andelen som svarer i på tidspunkt 1 og tilstand j på tidspunkt 2.

For k=2 er sammenhengene gitt ved (4), (5) og (6). Spørsmålet er nå om det er mulig å sette opp en sammenheng mellom p_{ij} (1,2) og parametrene i en Markov-modell med målefeil. For å sette opp sammenhengen må vi gjøre visse forutsetninger, som dermed blir modellen vår. Foreløpig skal vi anta at sjansen for å svare en bestemt verdi bare avhenger av den sanne verdi på samme tidspunktet. Modellen forutsetter altså at sjansen for å svare feil på ett tidspunkt ikke er influert av om en har svart feil på et tidligere tidspunkt.

I første omgang skal vi sette opp sammenhengene mellom parametrene og de observerte størrelsene når vi har målinger på to tidspunkter og k=2. Enkle regler fra sannsynlighetsregningen gir da følgende 4 ligninger:

I $$p_{11}(1,2) = p_1 q_{11} m_{11} q_{11} + p_1 q_{11} m_{12} q_{21} + p_2 q_{21} m_{21} q_{11} + p_2 q_{21} m_{22} q_{21}$$

II $$p_{12}(1,2) = p_1 q_{11} m_{11} q_{12} + p_1 q_{11} m_{12} q_{22} + p_2 q_{21} m_{21} q_{12} + p_2 q_{21} m_{22} q_{22}$$

III $$p_{21}(1,2) = p_1 q_{12} m_{11} q_{11} + p_1 q_{12} m_{12} q_{21} + p_2 q_{22} m_{21} q_{11} + p_2 q_{22} m_{22} q_{21}$$

IV $$p_{22}(1,2) = p_1 q_{12} m_{11} q_{12} + p_1 q_{12} m_{12} q_{22} + p_2 q_{22} m_{21} q_{12} + p_2 q_{22} m_{22} q_{22}$$

Ligningene I-IV er noe mer kompliserte enn de tilsvarende ligninger satt opp i avsnittet foran, hvor det ikke er målefeil i modellen. Likevel er ligningene lett å tolke. Ligning I, f.eks., sier at sannsynlighetene for å svare 1 både på tidspunkt 1 og 2 kan skrives som en sum av sannsynligheter til fire disjunkte begivenheter.

De fire disjunkte begivenheter kommer av at vi har fire latente verdier etter to tidspunkter, (1,1), (1,2), (2,1) og (2,2). For de latente verdier gjelder en enkel Markov-modell, og sannsynlighetene for dem, L(i,j), kan derfor fås som i avsnitt 2:

$$L(1,1) = p_1 m_{11} \; ; \; L(1,2) = p_1 m_{12}$$

$$L(2,1) = p_2 m_{21} \; ; \; L(2,2) = p_2 m_{22}.$$

For hver av disse latente verdier er det nå en sannsynlighet for å observere (1,1). Sannsynligheten for å observere (1,1) gitt at latent verdi er (1,1) er q_{11}^2, sannsynligheten for å observere (1,1) gitt latent verdi er (1,2) er $q_{11}\, q_{21}$, sannsynligheten for å observere (1,1) gitt at latent verdi er (2,1) er $q_{21}\, q_{11}$ og sannsynligheten for å observere (1,1) gitt at latent verdi er (2,2) er q_{21}^2. Den samlede sannsynlighet for å observere (1,1), blir derfor

$$p_{11}(1,2) = p_1 m_{11} q_{11}^2 + p_1 m_{12} q_{11} q_{21} + p_2 m_{21} q_{21} q_{11} + p_2 m_{22} q_{22}^2 \; ,$$

som er identiske med ligning I ovenfor. De øvrige ligninger framkommer på lignende måte.

Fordi ligning IV følger av ligningene I-III og det faktum at summen av sannsynlighetene på venstre side er lik 1, har vi egentlig bare 3 ligninger. Det er imidlertid 5 ukjente, p_1, q_{11}, q_{22}, m_{11} og m_{22}. Det er derfor nødvendig å gjøre visse forutsetninger i tillegg for å finne de ukjente parametre som funksjon av $P_{ij}(1,2)$. I Wiggin (1973) er det foreslått noen tilleggsbetingelser som er rimelige for visse anvendelser, og samtidig gjør det mulig å løse ligningene I-IV. Se også Chua and Fuller (1987) og Poterba and Summers (1986).

Vi skal imidlertid gå en annen vei. Vi skal anta det er observasjoner på et tredje tidspunkt, og at m_{ij} og q_{ij} er konstante. I tilfeller med tre tidspunkter og k=2, observeres de åtte størrelsene

$P_{ijk}(1,2,3)$: Sannsynligheten for å svare verdi i på tidspunkt 1, verdi j på tidspunkt 2 og verdi k på tidspunkt 3.
$(i=1,2;j=1,2;k=1,2)$

Mellom de observerte sannsynligheter og parametrene p_1, m_{ij} og q_{ij} $(i=1,2;j=1,2)$ fås nå et ligningssett på samme måte som ligningene I-IV, men nå består ligningssettet av 8 (7 uavhengige) ligninger med 5 ukjente.

På konsentrert form kan ligningene skrives på følgende måte:

$$p_{ijk}(1,2,3) = \sum_{a=1}^{2}\sum_{b=1}^{2}\sum_{c=1}^{2} p_a q_{ai} m_{ab} q_{bj} m_{bc} q_{ck}$$

$$(i=1,2;J=1,2;k=1,2). \qquad (7)$$

Da ligningene ikke er lineære, er det ikke helt enkelt å løse dem. Lazarsfeld og Henry (1968) viste at parametrene kan finnes ut fra ligningene. Imidlertid er det vanskelig å generalisere resultatene til variable med flere enn 3 tidspunkter. Dessuten kan en få løsninger som gir negative sannsynligheter. Det finnes derfor ikke lett tilgjengelig soft-ware for disse løsninger.

En annen måte er å bruke sannsynlighetsmaksimeringsprinsippet. (Bye and Schechter (1986), Haberman (1978)). Denne metoden består av å finne de verdier på parametrene som gir størst sannsynlighet for de observerte verdiene. Det er ikke mulig å sette opp eksplisitte, analytiske uttrykk for sannsynlighetsmaksimerings-estimatorene, men ved hjelp av vanlige metoder for numerisk bestemmelse av løsninger i ikke lineære ligninger er det mulig å bestemme dem.

I eksemplene i neste avsnitt skal vi anvende en litt annen teknikk, nemlig den såkalte E-M-algoritmen. Det er en iterativ prosedyre for å bestemme sannsynlighetsmaksimeringsestimatoren, som er meget enkel å anvende. Problemet sammenlignet med andre numeriske metoder er at den konvergerer nokså langsomt. For de enkle modeller som betraktes her, er imidlertid dette ikke noe problem.

Langeheine og Van de Pol (1990), har utviklet et meget generelt program for estimering innen latente Markov-modeller. Den delen av algoritmen som brukes her, er nærmere beskrevet i vedlegg 1.

I neste avsnitt skal gis noen eksempler på hvordan algoritmen brukes på data fra arbeidskraftundersøkelsene og levekårs-undersøkelsene. I det første eksemplet må vi "myke opp" forutsetningen om at m_{ij} er konstante over tid. En nærmere analyse av ligningene (7) vil viser at vi da har 7 ligninger med 7 ukjente. Estimeringsmetoden er gitt i vedlegget.

I det andre eksemplet viser det seg at m_{ij} er konstant over tid, og estimeringen er derfor gjort med den forutsetningen at m_{ij} er konstant.

4. Eksempler

4.1 Estimering av målefeil i arbeidskraftundersøkelsene (AKU)

Som i mange andre land, følges utviklingen på arbeidsmarkedet ved hjelp av løpende utvalgsundersøkelser, Arbeidskraftundersøkelsene, AKU. Undersøkelser utføres hver måned med roterende utvalg, slik at en viss del av utvalget ett bestemt kvartal er med i kvartalet før og det samme kvartal året før. Dette gir i prinsippet mulighet for å studere strømmene inn og ut av arbeidsstyrken fra ett kvartal til samme kvartal året før. I praksis har det vist seg vanskelig å få god nok kvalitet på tall for slike strømmer, og de fleste land avstår fra å publisere dem. (Poterba and Summers (1986). I USA benyttes resultater fra kvalitetskontrollundersøkelser for å

justere for målefeil.

I dette eksemplet skal vi anvende metoden presentert i avsnitt 3 ovenfor. Vi bruker en variabel som er lik 1 hvis personen tilhører arbeidsstyrken, og lik 2 hvis personen er utenfor. I AKU er hver pulje med i utvalget 4 ganger. I dette eksempel har vi bare brukt observasjoner for 3 tidspunkter. Tidsavstanden mellom de to første tidspunkter er et kvartal, mens den er 3 kvartaler for det 2. og 3. tidspunkt. Det er derfor urimelig å anta at m_{ij} er konstant. Som nevnt tidligere må vi derfor innføre to nye parametre, og får derved 7 parametre å estimere. Imidlertid består (7) også av 7 ligninger. For å undersøke om parametrene varierer mellom forskjellige befolkningsgrupper, er parametrene estimert for 5 forskjellige grupper, som er:

Gruppe 1: Personer 16-19 år
Gruppe 2: Personer 20-24 år
Gruppe 3: Menn 25-54 år
Gruppe 4: Kvinner 25-54 år
Gruppe 5: Personer 55-74 år

Resultatene er gitt i tabell 2 og tabell 3.
Det er stor usikkerhet i resultatene. Usikkerheten kan estimeres ved å bruke den vanlige metoden for sannsynlighetsmaksimeringsestimatorer. Dette er ikke gjort her fordi det kreves et omfattende regnearbeid. Dessuten mener vi at en får et bedre bilde av usikkerhetene ved å gjenta de samme beregninger for alle puljer i AKU. Hvis vi på denne måten ser et klart mønster i estimatene for målefeilene, vil vi tolke dette som en god indikator på om vi har funnet meningsfulle estimater for målefeilene.

Tabell 2. Manifeste og latente andeler i arbeidsstyrken samt svarsannsynligheter. 5 grupper. 2. kvartal 1988. Prosent

Gruppe	Manifeste* andeler i arbeids-styrken	Latente andeler i arbeids-styrken	Svarsannsynlig-heter	
			q_{11}	q_{22}
16-19 år	49,50	47,0	92,5	85,6
20-24 år	81,3	77,6	97,1	73,6
Menn 25-54 år	94,5	93,9	99,9	89,6
Kvinner 25-54 år	79,8	80,1	98,0	93,8
55-74	44,7	44,9	97,7	98,6

*) Disse andeler er ikke identiske med de tilsvarende tall publisert i offisiell statistikk. De er basert på ca 1/4 av utvalget i AKU og ikke justert for effekter av frafallet.

Ser en på estimatene for svarsannsynlighetene, q_{11} og q_{22} i tabell 2, viser det seg at q_{11} er svært lik 1. Det betyr at sannsynligheten for å svare "i arbeidsstyrken" når en faktisk

tilhører arbeidsstyrken, er meget høy. Derimot er q_{22} vesentlig mindre enn 1, spesielt for de yngre, dessuten er den litt høyere for kvinner enn for menn. Det betyr at det er en klar tendens til å svare "i arbeidsstyrken" når en faktisk er utenfor arbeidsstyrken. Det synes altså å være et press på ungdom til å svare "i arbeidsstyrken".

Ser en på forskjellen mellom manifeste og latente andeler i arbeidsstyrken, fører svarsannsynlighetene til at manifeste andeler er større enn latente for unge og for menn. Det ser altså ut som om målefeil i arbeidskraftundersøkelsene fører til en overestimering av arbeidsstyrken. Bortsett fra aldersgruppen 20-24 år synes imidlertid overestimeringen å være liten. At de lave svarsannsynligheter ikke fører til større skjevheter, skyldes at skjevheten er en funksjon av både svarsannsynlighetene og andelen i og utenfor arbeidsstyrken.

Tabell 3. Latente og manifeste overgangssannsynligheter i perioden 2. kvartal 1988 til 2. kvartal 1989

	2. kvartal 1988	2. kvartal 1989			
		Manifest		Latent	
		1	2	1	2
16-19 år	1. I arbeidsstyrken	0,71	0,29	0,81	0,19
	2. Utenfor arbeidsstyrken	0,35	0,65	0,25	0,75
20-24 år	1. I arbeidsstyrken	0,85	0,15	0,89	0,11
	2. Utenfor arbeidsstyrken	0,40	0,60	0,09	0,91
Menn 25-54 år	1. I arbeidsstyrken	0,98	0,02	0,99	0,01
	2. Utenfor arbeidsstyrken	0,40	0,60	0,32	0,68
Kvinner 25-54 år	1. I arbeidsstyrken	0,94	0,06	0,97	0,03
	2. Utenfor arbeidsstyrken	0,24	0,76	0,13	0,87
55-74 år	1. I arbeidsstyrken	0,87	0,13	0,91	0,09
	2. Utenfor arbeidsstyrken	0,01	0,99	0,00	1,00

I tabell 3 kan det sees at forskjellen mellom latente og manifeste overgangssannsynligheter er til dels meget store. Stort sett gjelder det at de manifeste overgangene utenfor diagonalen er større enn de latente. Dette skyldes at mange av de observerte overgangene fra ett tidspunkt til det neste skyldes målefeil. Når disse fjernes, står en igjen med de latente overganger, som viser mye større stabilitet enn de manifeste. Usikkerhetene på tallene er store, lignende analyser må gjøres på flere tidspunkter før en kan tenke på å publisere bruttostrømmer fra AKU, men metoden synes å ha et godt potensial.

4.2 Estimering av målefeil i Levekårsundersøkelsene

Statistisk sentralbyrå har gjennomført 4 levekårsundersøkelser i 1973, 1980, 1983 og 1987. I de tre siste har en brukt et panel på ca. 1000 personer.

Dette gir rik anledning til å undersøke målefeilene til en lang rekke spørsmål som dekker mange emner om økonomiske, sosiale og helsemessige forhold. I dette eksemplet skal vi ta for oss tre spørsmål knyttet til arbeidsmiljø. De tre spørsmålene dreier seg om respondenten er plaget av tunge løft, ensidige bevegelser og belastende arbeidsstilling i sitt arbeid. Hvis "ja" er svaret satt lik 1, hvis "nei" lik 2. Metoden fra avsnitt 3 er nå brukt for å estimere latente parametre på grunnlag av undersøkelsene 1980, 1983 og 1987. Avstanden mellom de to første undersøkelser er 3 år, mens den er 4 år for de to siste. Dette er det ikke tatt hensyn til her. Resultatene er gitt i tabell 4a og 4b.

Tabell 4.a. Manifeste og latente andeler og svarsannsynligheter. Arbeidsmiljøspørsmål. 1980

	Manifest andel	Latent andel	Svarsannsynligheter	
	%	%	q_{11}	q_{22}
Tunge løft	31,5	31,3	0,909	0,955
Ensidige bevegelser	42,0	48,1	0,833	0,967
Belastende arbeids-stilling	35,9	35,1	0,876	0,921

Tabell 4.b. Manifeste og latente overgangssannsynligheter

	Manifest			Latent		
		1	2		1	2
Tunge løft	1	0,73	0,27	1	0,88	0,12
	2	0,12	0,88	2	0,05	0,95
Ensidige bevegelser	1	0,67	0,33	1	0,82	0,18
	2	0,23	0,77	2	0,16	0,84

Sammenligning mellom manifest og latent andel viser at for ett av spørsmålene, nemlig det om ensidige bevegelser, blir andelen som svarer "ja" noe underestimert på grunn av målefeil. For samtidig spørsmål er sjansen for å si "nei" når det riktige svar er "ja" ganske betydelig, samtidig som det er en mye mindre sjanse for å si "ja" når det riktige er "nei". Det synes som om vi har en tendens til å beskrive vårt arbeidsmiljø litt bedre enn det egentlig er. Noe som stemmer godt med hva en skulle vente seg. Det overraskende er at dette ikke får større konsekvenser for forskjellene mellom latente og manifeste andeler.

Tabell 4.b viser samme klare tendens som tabell 3, nemlig at de latente overgangssannsynlighetene langs diagonalen er mye større enn de tilsvarende manifeste overgangssannsynligheter.

Referanser

Anderson, T.W. (1954): Probability Models for Analyzing Time Change in Attitudes. pp 17-66 in P.F. Lazarsfeld (ED), Mathematical Thinking in Social Sciences. New York: Free Press.

Chua, T.C. and Fuller, W.A. (1987): A Model for Multinomial Response Error Applied to Labour Flows. ASA Vol 872 No. 397, pp. 46-51.

Langeheine, Rolf and Pol Frank van de (1990): A Unifying Framework for Markov Modelling in Discrete Space and Discrete Time. Soc. Methods and Research. 18 pp. 416-441.

Lazarsfeld, P.F. and Henry N.W. (1968): Latent Structure Analysis. Boston: Houghton-Miffin.

Haberman, S.J. (1979): Analysis of Quantitative Data. Vol.2, New Developments. New York. Academic Press.

Bye, B.V. and Schechter, E.S: A latent Markov Model Approach to the estimation of response error in multivariate panel data. J. of Amer. Stat. Assn.81. pp. 375-380.

Poterba, J.M. and Summers, L.H. (1986): Reporting Errors and Labour Market Dynamics. Econometrica, Vol. 54 No. 6.

Wiggins, L.M. (1973): Panel Analysis. Amsterdam: Elsevier.

Aaberge, R. (1988): Bruttostrømmer på arbeidsmarkedet.
NAUT rapport 1988: 3. Nordisk ministerråd.

ESTIMERING VED HJELP AV E-M ALGORITMER

I modellen vår er det følgende parametre:

p_i: Sannsynligheten for k mulige verdier på tidspunkt 1. (i=1,2,...k)

m_{ij}: Sannsynligheten for å anta verdien j på tidspunkt (t+1), gitt at verdien var i på tidspunkt t. (i=1,2,...k; j=1,2,...k)

q_{ij}: Sannsynligheten for å svare verdi j, gitt at sann verdi er i. (i=1,2,...k; j=1,2,...k).

For å estimere disse parametrene, skal vi nå anta at vi på tre tidspunkter observerer både de sanne verdier og verdiene med målefeil. Vi innfører derfor følgende størrelse:

$\xi_{\alpha\beta\gamma jik}$: Antall personer med sann verdi α, β og γ på tidspunktene 1, 2 og 3 henholdsvis, samt observert verdi i, j og k på tidspunktene 1, 2 og 3 henholdsvis.

Nå er det lett å se at dersom en har observert $\xi_{\alpha\beta\gamma ijk}$, er sannsynlighetsmaksimeringsestimatorene for parametrene gitt ved

$$p_\alpha = \frac{\xi_{\alpha.....}}{\xi_{.....}}; \qquad (1)$$

$$m_{\alpha\beta}^{12} = \frac{\xi_{\alpha\beta....}}{\xi_{\alpha.....}} \text{ og } m_{\beta\gamma}^{23} = \frac{\xi_{.\beta\gamma...}}{\xi_{.\beta....}}, \qquad (2)$$

hvor $m_{\alpha\beta}^{st}$ betegner overgangssannsynlighetene fra tidspunkt s til t. Hvis $m_{\alpha\beta}^{st} = m_{\alpha\beta}$, altså uavhengig av tidspunktet er

$$m_{\alpha\beta} = \frac{\xi_{\alpha\beta....} + \xi_{.\beta\gamma...}}{\xi\alpha..... + \xi_{.\beta....}}. \qquad (3)$$

$$q_{\alpha i} = \frac{\xi_{\alpha..i..} + \xi_{.\alpha..i.} + \xi_{..\alpha..i}}{\xi_{\alpha.....} + \xi_{.\alpha....} + \xi_{..\alpha...}}. \qquad (4)$$

Nå er det slik at $\xi_{\alpha\beta\gamma ijk}$ er ikke observert, men vi må sette inn rimelige anslag for parametrene, p_α, $m_{\alpha\beta}$ og $q_{\alpha i}$. Deretter beregnes forventete verdier for alle $\xi_{\alpha\beta\gamma ijk}$, gitt disse første anslag. Denne blir

$$\theta_{\alpha\beta\gamma ijk} = p_\alpha q_{\alpha i} m_{\alpha\beta} q_{\beta j} m_{\beta\gamma} q_{\gamma k} \qquad \begin{matrix} \alpha=1,2,..,k \\ \beta=1,2,..,k \\ \gamma=1,2,..,k \\ i=1,2,..,k \\ j=1,2,..,k \\ k=1,2,..,k \end{matrix} \qquad (5)$$

Disse er ikke identiske med $\xi_{\alpha\beta\gamma ijk}$ gitt ovenfor, fordi de er basert på anslag.

Fra undersøkelsene har vi observasjoner av $\xi_{...ijk}$, og på grunnlag av våre anslag kan vi regne ut $\theta_{...ijk}$. Hvis avvikene mellom $\xi...ijk$ og $\theta...ijk$ er meget små, har vi valgt gode anslag i utgangspunktet. Hvis ikke regnes ut et nytt sett av verdier

$$\xi_{\alpha\beta\gamma ijk} = \frac{\xi_{...ijk}}{\theta_{...ijk}} \theta_{\alpha\beta\gamma ijk}, \qquad (6)$$

som utgjør første iterasjon. For disse nye verdier av $\xi_{\alpha\beta\gamma ijk}$ gjentas beregningene (1)-(5), og nye $\xi_{\alpha\beta\gamma ijk}$ beregnes ved hjelp av (6). Dette fortsetter inntil koeffisienten på høyre side av (6) blir nær 1. (Formelt regnes sannsynligheten for de observerte verdiene ut etter hver iterasjon, og iterasjonen stopper når økningen i denne sannsynligheten blir negliserbar.) For en nærmere diskusjon av valg av startverdier og stoppkriterium henvises til Langheine and Pol (1990), som også har utviklet den algoritmen som er brukt her.