Statistics Norway

Statistisk sentralbyrå

Liv Belsby

Forbruksundersøkelsen
Vektmetoder, frafallskorrigering og intervjuer-effekt

Avdeling for samordning og utvikling / Seksjon for statistiske metoder og standarder

Innhold

1. Innledning

Siden 1974 har Statistisk sentralbyrå gjennomført årlige forbruksundersøkelser. Hovedformålet med undersøkelsene har vært å gi en detaljert oversikt over private husholdningers forbruk, som grunnlag for en ajourføring av vektgrunnlaget i konsumprisindeksen. Det er videre tatt sikte på å kartlegge forbruket i forskjellige grupper av husholdninger. Husholdningene er klassifisert etter størrelsen på husholdningen, inntekt, sosioøkonomisk status, landsdel og bostedsstrøk, størrelse av forbruksutgift mv.

Utvalgsplanen for Forbruksundersøkelsen ble endret fra og med 1992, forkortet til FU92. Derfor må metoden for å beregne vektene revurderes. Før 1992 ble utvalget trukket med familie som trekke-enhet. Fra og med 1992 er person trekke-enhet, og husholdningen etableres med basis i den trukne personen. Utvalget blir trukket fra Det sentrale personregister.

I dette notatet beskriver vi estimatoren som blir brukt for å beregne vektene. Vektene gir estimater for totalen for populasjonen og gjennomsnitt for husholdningene. Estimatoren tar hensyn til at det er skjevt frafall i FU. Frafallskorrigeringen på baseres på modellering av responssannsynligeheten. Vi beskriver tre alternative tilnærmelser. Den ene metoden fremkommer ved å formulere modelleantagelser som et ligningssystem. Vektene beregnes ved å løse dette ligningssystemet. Vektene for FU92 ble beregnet på denne måten. De to andre metodene er basert på logistiske modeller. Den enkleste er en vanlig logistisk modell. Den ble brukt til å beregne vektene for FU93. Denne beregningsalgoritmen er implementert i SAS, og en utskrift av programmet er i vedlegget. Den mest fullstendige og korrekte modellen for responssannsynligheten tar hensyn til at husholdningens størrelse er ukjent for frafallsdelen av utvalget, og at responssannsynligheten påvirkes av dette. Metoden innbærer modellering av den faktiske husholdningsstørrelsen, med utgangspunkt i prediksjonstilnærming, se (Bjørnstad,1995). Vekter basert på denne modellen er ikke tilstrekkelig utprøvd. Vi presenterer likevel metoden her fordi den teoretisk modellerer responssannsynligheten mer korrekt enn den enkle logistiske modellen.

Vi foreslår en generalisering av den logistiske modellen for responssannsynligheten. Generaliseringen tar hensyn til eventuell avhengighet mellom husholdninger, som intervjues av den samme intervjueren. Denne modellen er inneholdt i klassen av ikke-lineære modeller med tilfeldige komponenter, som omtales mye i statistisk metode litteratur for tiden. En av de tidligste artiklene er (Anderson and Aitkin,1985). De skriver om en lignende situasjon, som vi behandler i dette notatet.

Ikke-lineære modellene med tilfeldige komponeneter formulerer generelt årsakssammenhenger for data med klynge-effekter og diskrete responser. Klynge-effekter er trolig hyppige i byråets data, siden data samles inn i geografiske klynger og mange husholdninger intervjues av én og samme intervjuer. Personer innenfor en husholdning kan også i mange sammenheng utgjøre en klynge, se (Belsby og Grøtvedt, 1993). Modellene er ofte problematiske å estimere, og en stor del av artiklene i statistisk metode-litteratur, om denne typen modeller, behandler nettopp estimering.

For den vanlige logistiske modellen for responssannsynligheten benytter vi en nylig publisert testmetode, (Commenges et al., 1994) for å teste om den antatte avhengigheten er signifikant. Vi finner at avhengigheten er signifikant ved 1% nivå(og lavere). Det betyr at intervjuerne har forskjellig sannsynlighet for å få husholdningen til å akseptere å bli intervjuet.

Trolig kan denne typen testing være nyttig i flere sammenhenger. Påvist avhengighet innnenfor

klyngen gir informasjon om at flere variable knyttet til klyngen bør måles og innkluderes i analysen av årsakssammhengen. Testen har vi implementert i GAUSS. Programmet kan lett modifiseres og brukes til å teste avhengighet i data, som passer til lignende typer modeller. Utskrift av dette programmet er også vedlagt.

I avsnitt 2 og 3 beskriver vi henholdsvis estimatoren og de tre alternative metodene for å korrigere for frafall. I avsnitt 4 beskriver vi modellering av avhengigheten mellom husholdninger, som blir intervjuet av samme intervjuer. Avsnitt 5 oppsummerer notatet.

2. Estimeringsmetode

Vi velger å benytte Horwitz Thompson-estimatoren. Den estimatoren medfører at vekten for en husholdning er lik den inverse sannsynligheten for at husholdningen skal bli inkludert i utvalget. Sannsynligheten for at en husholdning i populasjoenen skal trekkes er proposjonal med antall personer i husholdningen. Hvis N er antall personer i populasjonen, n antall personer som trekkes ut og j antall personer i husholdningen, da er sannsynligheten for at en husholdning skal trekkes lik $(n/N)*j$. Merk at trekkeprosedyren gir skjevhet i utvalget ved at store husholdninger blir overrepresentert i forhold til hyppigheten i populasjonen av husholdninger.

I tillegg bidrarer det delvis skjeve frafallet også til skjevhet. Frafallet er forholdsvis stort, for eksempel ca 40% i FU93. Når vi tar hensyn til frafallet, er sannsynligheten for at en husholdning blir innkludert i utvalget uttrykt ved

$$Pr(inkludert)=Pr(trukket)*Pr(respons|trukket).$$

I neste avsnitt kommer vi tilbake til modellering av responssannsynligheten. La r_i betegne husholdning i, og sannsynligheten for at husholdning i skal bli inkludert med π_i Sannsynligheten for at en husholdning skal inkluderes er da

$$\pi_i=(n/N)*j_i*r_i.$$

La y_i være verdien av den variabelen som skal estimeres. Da kan estimatet for totalen i populasjonen skrives som

$$Y=\sum_s y_i/\pi_i.$$

Gjennomsnittlig verdi for husholdningene fremkommer ved å dividere med summen av vektene. Under forutsetning av at sannsynligheten π_i er korrekt beregnet, er estimatorene forventningsrette.

3. Responssannsynlighet

3.1 Variable som antas å påvirke responssannsynligheten

Husholdningens størrelse påvirker sannsynligheten for respons. For eksempel er husholdninger med

bare en person langt oftere i frafallsdelen enn husholdninger med mange personer. For unge enslige er den viktigste grunnen til frafallet at disse er vanskelig å treffe hjemme, og eldre enslige er ofte skeptiske og lite motiverte, (Andersen et al., 1991). Videre har frafall har vist seg å være større i byområder enn på landet. Kommuner defineres som by-kommuner hvis mer enn 90% av innbyggerne bor i tettbygde strøk. De andre kommunene defineres som landkommuner. Trolig øker frafallet i sommerhalvåret. Variabelen for tidspunktet for datainnsamling er binær delt i "sommer" og "ikke sommer". Sommer defineres som perioden fra 21.5 til 12.8.

3.2. Registerfeil

Husholdningens størrelse er gitt i Det sentrale personregister. En registrert husholdning er en familie eller partnere, som har registrert samboerskapet. En husholdning i FU består av personer som har minst et felles måltid og som bor i samme boenhet. Den faktiske størrelsen til husholdningen anses å være antallet personer i husholdningen når husholdningen defineres som i FU. Ofte er det avvik mellom den faktiske husholdningsstørrelsen og den registrerte. Den faktiske husholdningsstørrelsen er ukjent for de fleste husholdningene i frafallsdelen av utvalget. Metoden beskrevet i avsnitt **3.3** og **3.5** tar hensyn til avviket mellom den registrerte og den faktiske husholdningsstørrelsen, og søker å justere for dette. Tabell 1 nedenfor viser hvor godt registrert og faktisk husholdningsstørrelse samsvarer. Tabellen viser også respons-prosent, antall responser og totalt antall husholdninger for de forskjellige registrerte husholdningsstørrelser. Data er for FU93.

Tabell 1. Fordelingen for faktisk husholdningsstørrelse for FU93, gitt den registrerte husholdningsstørrelsen. Tallene er i gitt i % av antall registrerte husholdninger med et gitt antall personer. Husholdninger med fem og flere personer er slått sammen til en gruppe.

Registrert	Faktisk husholdningsstørrelse					% respons	Antall respons	totalt antall
	1	2	3	4	≥ 5			
1	53	27	13	5	2	51	220	434
2	4	72	20	2	1	56	296	525
3	5	19	58	18	0	61	247	403
4	1	3	7	82	7	67	348	516
≥ 5	3	3	3	6	86	66	199	302
totalt antall	151	336	260	359	204	60	1310	2180

3.3. Modellantagelser formulert ved et ligningssystem

Denne algoritmen ble brukt i vektberegningene for FU92. For å beregne vektene for FU93 brukte vi derimot metoden som baseres på responsmodelleringen beskrevet i **3.4**, siden den er enklere å implementere. I motsetning til responsmodellene beskrevet nedenfor, gir denne metoden heller ikke umiddelbart mulighet til å teste hvilke variable som påvirker responssannsynligheten signifikant.

Algoritmen bygger på at husholdningenes størrelse, bosted inndelt i by og landsbygd og tidspunktet for intervjuet påvirker responssannsynligheten. Denne metoden tar sikte på å korrigere for registerfeilen for husholdningens størrelse. Den bygger på forutsetningen at når faktisk husholdningsstørrelse er gitt, er frafallssannsynligheten uavhengig av den registrerte husholdnings-

størrelsen.

Algoritmen forklarer vi uten å ta hensyn til stratifisering på andre variable enn husholdningensstørrelse. Det forenkler notasjonen og prinsippet fokuseres. La x_{ij} betegne antall husholdninger i frafallsgruppen med registerte husholdningsstørrelse i og faktisk j. Da er

$$\sum_j x_{ij}=n_i, \qquad (1)$$

hvor n_i er antall i frafallsgruppen med registert husholdningsstørrelse i. Denne størrelsen er kjent, men x_{ij} er ukjent. I nettoutvalget, dvs den delen av utvalget som svarer, er både registert og faktisk husholdningstørrelse kjent. La n_{ij} være antallet husholdninger i nettoutvalget med kjennetegn ij. Forutsetningen om ved kjent faktisk husholdningsstørrelse er frafallssannsynligheten uavhengig av den registrerte husholdnings-størrelse. betyr at

$$\frac{x_{ij}}{x_{ij}+n_{ij}}=\frac{x_{kj}}{x_{kj}+n_{kj}}.$$

Denne relasjonen forenkler ligningssystem (1) til antall ukjente og ligninger lik antall husholdningsgrupper. For FU92 samlet vi husholdninger med størrelse seks og større til en gruppe. Ved å løse ligningssystemet beregnes x_{ij}. Estimert frafallssannsynlighet blir da $x_{ij}/(x_{ij}+n_{ij})$ og responssannsynligheten naturligvis $1-x_{ij}/(x_{ij}+n_{ij})$. Merk at forutsetning ovenfor implisere at denne brøken blir den samme for alle i, registert husholdningsstørrelse, når j, faktisk husholdningsstørrelse, holdes fast. Det betyr at estimert responssannsynlighet er den samme for husholdninger med lik faktisk husholdningsstørrelse selv om de registrert husholdningsstørrelsene er forskjellige.

3.4. Logistisk modellering av responssannsynligheten

Denne metoden ble brukt i FU93. Den er enkel å implementere og gir umiddelbart mulighet for å teste hvilke av de potensielle forklaringsvariblene som er signifikante. En ulempe i forhold til tilnærmelsene beskrevet i **3.3** og **3.5** er at denne modellen ikke tar hensyn til at den registrerte husholdninsstørrelsen for husholdningene i frafallsgruppen ofte avviker fra den faktiske.

Motivert av erfaringer fra tidligere undersøkelser antas responssannsynligheten å være en funksjon av husholdningens størrelse, tidspunktet for intervjuet delt i perioden fra 21.5 til 12.8 og utenom denne tidsperioden, og bosted delt i by og land. Med andre ord innkluderes de samme variablene i denne modellen som i algoritmen beskrevet i **3.3**. Vi antar en logistisk link-funksjon. Siden faktisk husholdningsstørrelse er ukjent for frafallsdelen, bruker vi den registerte husholdningsstørrelsen for disse husholdningene. Estimeringen (maximum likelihood) viser at bosted og husholdningstørrelse er signifikant, med gruppering 1, 2, og 3. Den estimerte modellen er

$$\Pr(\text{respons})=\frac{1}{1+\exp(-0.77+0.92z_1+0.24z_2+0.37b)},$$

hvor z_j er 1 når husholdningensstørrelsen er j, og 0 ellers ($j=3$ og større blir tatt hånd om av konstantleddet), b er 0 for husholdninger i landkommuner og 1 for bykommuner.

Uttrykket for responssannsynligheten viser at husholdninger med bare en person har mindre sannsynlighet for respons enn husholdninger med flere personer. Dessuten har husholdninger i byene lavere sannsynlighet for respons enn husholdningene i landkommuner.

I dette tilfellet hvor alle forklaringsvariablene er klassevariable gir det omtrent samme estimater for responssannsynlighet som en etterstratifisering ville ha gitt. En fordel med denne typen tilnærmelse er at den er enklere å generalisere til mer generelle modeller, som for eksempel den beskrevet i **3.5** nedenfor. Dessuten er uttesting av hvilke forklaringsvarable som er signifikante teknisk enklere enn ved etter-stratifisering.

For å vurdere vektene sammenligner vi den estimerte husholdningsfordelingen for populasjonen med den estimerte fordelingen fra Kvalitetsundersøkelsen for Folke- og boligtellingen i 1990, FoB90. I Fob90 er en husholdning definert som en personer som bor sammen. Tilsvarende estimater finnes ikke for 1993. Derfor sammenligner vi med denne fordelingen. Estimert antall husholdninger av en gitt størrelse er summen av vektene for husholdninger med denne størrelsen i svardelen av utvalget. Husholdninger med seks og flere personer er slått sammen til en gruppe ved beregning av trekkesannsynlighet. Responsmodellen er den som det er gitt estimater for ovenfor, dvs at husholdninger med 3 og flere personer er slått sammen til en gruppe.

Tabell 2. Estimert husholdningsfordeling ved metoden beskrevet i dette avsnittet og estimert fra Kvalitetsundersøkelsen for Folke- og boligtellingen i 1990.

Husholdnings-størrels	Estimert antall husholdninger	
	Metoden i *3.4*	KU
1 person	680429	718000
2 personer	535435	500000
3 og flere	628149	643000
Total antall	1844012	186100

Avviket kan kanskje delvis skyldes definisjonsforskjellen mellom husholdning i Forbruksundersøkelsen og i Fob90. I Fob90 er en husholdning definert som de personene som.... Dessuten ignorerer, som nevnt ovenfor, modellen at faktisk størrelse er ukjent for husholdningene i frafallsdelen. Dette kan muligens bidra til at den estimerte husholdningsfordelingen avviker fra den estimerte i KU.

3.5. Prediksjonstilnærmelse ved modellering

Metoden i dette avsnittet er et forslag motivert av resultatene fra frafallsmodellering i FU. Metoden har ennå ikke blitt brukt til å beregne vekter. Med andre ord er den heller ikke tilstrekkelig utprøvd. Men teoretisk gir den en mer korrekt modellering av responssannsynligheten, idet den tar hensyn til at

faktisk husholdningsstørrelse er ukjent for de (fleste av) husholdningene i frafallsgruppen. Videre bygger den på en modellering av faktisk husholdningsstørrelse, som tar hensyn til at frafallet er skjevt. Modellen er mer fullstendig beskrevet i (Belsby, Bjørnstad, 1995). Den observerte husholdningsfordelingen betraktes som en del av det stokastiske utfallet, og innkluders derfor i likelihooden La y_i være den faktiske husholdningsstørrelsen for husholdning i, og $p_{y|x}$ sannsynligheten for faktisk husholdningsstørrelse y når den registererte er x. For husholdningene som inkluderes i utvalget er responssannsynligheten

$$\begin{aligned} \Pr(R_i=1, Y=y_i \mid x_i, b_i) &= \Pr(R_i=1 \mid x_i, y_i, b_i) * \Pr(y_i \mid x_i, b_i) \\ &= \Pr(R_i=1 \mid y_i, b_i) * p_{y_i \mid x_i}, \end{aligned}$$

Siden sannsynligheten for respons er uavhengig av registrert husholdningsstørrelse når den faktiske er kjent. Dessuten avhenger sannsynligheten for faktisk husholdningsstørrelse ikke av bosted. Linkfunksjonen for responssannsynligheten den logistiske. Sannsynligheten for frafallsdelen er imidlertid

$$\begin{aligned} \Pr(R_i=0 \mid x_i, b_i) &= \sum_y \Pr(R=0 \mid y, x_i, b_i) * \Pr(y \mid X_i) \\ &= \sum_y \Pr(R=0 \mid y, b_i) * p_{y \mid x_i} \\ &= \sum_y (1-\Pr(R=1 \mid y) * p_{y \mid x}. \end{aligned}$$

Parameterne i modellen estimeres ved å maksimere likelihooden. Under forutsetning om at observasjonene er uavhengige betyr det beregne de parameterne som maksimere logaritmen til sannsynligeheten

$$L=\prod_{i \epsilon S_r} \Pr(R_i=1, y_i \mid x_i, b_i) \prod_{i \epsilon S-s_r} \Pr(R=0 \mid y_i, x_i, b_i).$$

Maksimeringen kan for eksempel gjøres med programmet TSP. Det finner analytiske første og annen deriverte, og gir t-verdier basert på Fishers informasjonsmatrise, se for eksempel appendix A, (McCullagh og Nelder, 1990). Forutsetningen om uavhengighet kommer vi tilbake til i neste avsnitt.

4. Intervjueren og responssannsynligheten

4.1. Å modellere intervjuer-effekten med tilfeldige komponenter

Det virker rimelig at intervjueren påvirker sannsynligheten for respons. Grunner kan for eksempel være at erfaringen den enkelte intervjuer har og belastningen ved mange intervjuobjekter varierer. Andre mer personlige egenskaper kan også tenkes å påvirke responssannsynligheten. Den variasjonen som skyldes at noen intervjuer husholdninger på landsbygda og andre intervjuer husholdninger i byene tas hånd om av modellene beskrevet i **3.4**. Det samme gjelder variasjonen som skyldes husholdningenes størrelse. I dette avsnittet fokuserer vi på den "uforklarte" delen av variasjonen knyttet til intervjueren. Heretter betegner vi den "personlige" påvirkningen intervjueren har på intervjuobjektet for intervjuer-effekten. Hvis det hadde vært få intervjuere og mange intervju-objekter per intervjuer, ville den

enkleste måten å tatt hensyn til intervjuer-effekten på vært å utvide modellen med et konstant-ledd. Konstant-ledd ville generelt kunne ha forskjellige verdier for de forskjellige intervjuerne. Antallet intervjuere er for stort til at så mange parametre kunne estimeres. For eksempel er det vel 300 intervjuere i FU93.

En populær måte å modellere denne typen effekter på er å inkludere en tilfeldig variabel. I motsetning til konstant-leddet, nevnt ovenfor, estimeres ikke verdiene til den tilfeldige variabelen for de forskjellige intervjuerne. Derimot er variansen til den tilfeldige variabelen av interesse. Den tilfeldige variabelen forutsettes å ha samme verdi for alle observasjonene innen gruppen med antatt korrelerte observasjoner. Her betyr dette at variabelen antas å ha samme verdi for alle husholdninger som intervjues av en og samme intervjuer. Med den tilfeldige komponenten α_j inkludert, er modellen for husholdning i intervjuet av intervjuer j

$$\Pr(R_{ij}=1 \mid x_{ij}) = \frac{1}{1+\exp(\alpha_j+\beta' x_{ij})},$$

hvor x_{ij} betegner forklaringsvariablene. Denne typen modeller kalles modeller med tilfeldige komponenter eller flernivå-modeller. I data fra Forbruksundersøkelsen bygges den hierarkiske strukturen opp av husholdning og intervjuer, dvs. to nivåer. Modellen for responssannsynligheten er ikke- lineær siden respons måles med en binær variabel.

4.2. Er intervjuer-effekten signifikant?

Daniel Commenges et al.(1990) beskriver en test på om det er uavhengighet innen grupper for modeller av typen beskrevet i **4.1.** Testen er en modifisering av en test foreslått av Liang (1987). I vårt tilfelle består en gruppe av de husholdningene som intervjues av en og samme intervjuer. Testen er en type "score-test", se for eksemple (McCullagh og Nelder, 1990). Den er basert på den deriverte til likelihooden under H_0. I dette tilfellet er H_0 at det ikke er noen intervjuer-effekt, eller med andre ord at den tilfeldige komponenten ikke varierer fra intervjuer til intervjuer. Intuitivt kan teststørrelsen begrunnes med at hvis denne verdien er i nærheten av den sanne verdien, vil den deriverte av likelihooden være i nærheten av null. Standard statistisk teori sier nemlig at

$$\frac{d}{d\theta}\log f(x;\theta)=0.$$

Eller med ord; den deriverte av likelihooden med hensyn på den sanne parameterverdien er 0. Teststørrelsen fremkommer ved å dividere med variansen til den deriverte likelihooden, fortsatt under nullhypotesen. Under H_0 er teststørrelsen asymptotisk standard normalfordelt. Den beregnes til å være 24, som indikerer at det faktisk er en intervjuereffekt.

4.3. Klynge-effekt i kommunene

Vi har hittil ikke kommentert at det kunne også tenkes å være en klynge-effekt, dvs. at husholdninger fra samme kommune har tilsvarende korrelasjon som husholdninger intervjuet av samme intervjuer. I FU93 er det 146 kommuner. Færre enn 10% av husholdningene kommer fra kommuner hvor det er bare en intervjuer. Det er praktiske grunner til at vi ikke har modellert denne mulige korrelasjonsstrukturen simultant med intervjuer-effekten. Det ville gjøre testingen meget komplisert.

For å få en indikasjoen på hvor viktig det er å ta hensyn til klynge-effekten, tester vi denne på tilsvarende måte som for intervjuer-effekten. Nå ignorerer vi intervjuer-effekten. Teststørrelsen er da 0.60, dvs. ingen indikasjon på avhengighet innen kommunen. Naturligvis har vi nå gjort den feilen at vi ikke har tatt hånd om den faktiske intervjuer-effekten. Med bakrunn i den store forskjellen på teststørrelsene, mener vi likevel at testresultatet kan tolkes som at det er en signifikant intervjuer-effekt. Mens den det ikke ser ut til å være noen sterk klynge-effekt i kommunene. Naturligvis kunne det tenkes en annen definisjon av klynger, kunne ha gitt en signifikant klyngeeffekt.

4.4. Estimering av parameterne og intervjuer-effekten.

Ikke-lineære modeller med tilfeldige komponenter er vanligvis kompliserte å estimere. I (Anderson and Aitkin, 1985) foreslås en metode for logistiske modeller med tilfeldige komponenter. Metoden bygger på å integrere numerisk over de tilfeldige komponentene, og maksimere den marginale likelihooden (som ikke inneholder den tilfeldige komponenten) ved hjelp av Newton-Raphson algoritmen. Goldstein(1987) beskriver en alternativ metode basert på linearisering av den ikke-lineære funksjons-sammenhengen mellom responsen og de påvirkende variablene. Deretter beregnes minstekvadratersestimatorer ved hjelp av den iterative generaliserte minstekvadrartersmetoden, (IGLS). Simuleringsbaserte teknikker, (Zeger and Karim, 1991) peker seg ut som de mest robuste, se (Rodriguez and Goldman, 1995), spesielt for mer kompliserte modeller. Vi vil ikke gå nøyere inn på estimeringen her, men trekke fram at flere artikler tyder på at regresjonskoeffesientene ikke endres særlig ved å ta hensyn til avhengigheten introdusert ved den tilfeldige komponenten, se (Anderson and Aitkin, 1985), (Rodriguez and Goldman, 1995) og (Commenges et al., 1994). Vi er først og fremst interessert i å estimere disse, og ikke så mye i variansensene til de estimerte regresjonskoeffesientene. Derfor foretrekker vi å ignorere avhengigheten for å kunne bruke standard programvare. Det er vesentlig mindre ressurskrevende.

5. Noen konkluderende kommentarer

Vi mener at sammenlignigen med estimater fra KU tyder på at metoden beskrevet i avsnitt **3.4** gir tilfredstillende god kvalitet på vektene. En fordel med vektmetoden er at den er enkel å tilpasse ved bruk av deler av FU og et annet frafall enn for hele undersøkelsen. Det er gjort for undersøkelsen Energiforbruk i husholdningene 1993, (Djupskås, Nesbakken, 1993). (Djupskås, Nesbakken, 1993) sammenliger estimater beregnet ved denne oppblåsningsmetoden og tall fra FoB90. De finner at fordeling for husholdningstørrelse, hustype og region samsvarer rimelig bra. Men at fordelingen for byggeår tyder på en overvurdering av nye hus og hus med store arealer.

Med noe mer arbeid kan trolig vektene forbedre ved å ta utgangspunkt i modelltilnærmelsen beskrevet i **3.5.** En annen mulighet er å bruke informasjonen på populasjonsnivå til å justere vektene, -- kalibrere --, se Heldal(1992). Vi brukte denne metoden på FU92, og kalibrerte med hensyn på antall registrerte husholdninger i populasjonen. Dette ga ikke forbedrede vekter, men metoden kunne evenetuelt utprøves med andre kalibreringsvariable, som for eksempel aldersfordeling.

Referanser

Anderson, A. & Aitkin, M. (1985). Variance component models with binary response: Interviewer variability. Journal of the Royal Statistical Society, Ser. B, 47, no 2, 203-210.

Andersen, A., Opdahl, S. & Aasness, J. (1991). Nytte og kostnader ved alternative opplegg for SSB's forbruksundersøkelser. Interne Notater 91/22, Statistisk sentralbyrå.

Belsby, L. & Grøtvedt, L. (1993). Sammenhengen mellom barns sykelighet og miljøet. Norsk Epidemiologi; 3:3.

Belsby, L. & Bjørnstad, F. J. (1995). Nonresponse and householdsize in the Norwegian household expenditure survey. Foreløpig notat Statistisk sentralbyrå.

Bjørnstad, F. J. (1995). Utvalgsundersøkelser og prediksjon. Universitetet i Trondheim - AVH.

Commenges, D., Letenneur, L., Jacqmin, H., Moreau, T. & Dartigues, J. (1994) Test of homogenity of binary data with explanatory variables. Biomterics 50, 613-620.

Dempster, A. P., Laird, N.M. & Rubin, D.B. (1977). Journal of the Royal Statistical Society, B, 39, 1-39.

Djupskås, O. T. & Nesbakken, R. (1995) Energiforbruk i husholdningene 1993. Data fra forbruksundersøkelsen 1993. Rapport 95/10, Statistisk sentralbyrå.

Duncan, D. B. & Horn, S. D. (1972). Linear dynamic recursive estimation from the viewpoint of regression analysis. Journal of the American Statistical Assosiation, 72, 815- 821.

Goldstein, H. (1986). Multilevel mixed linear models analysis using iterativ generalized least squares. Biometrika, 73, 43-56.

Goldstein, H. (1987). Multilevel models in educational research. London: Griffin.

Heldal, J.,(1992) A method for Calibration of Weights in Sample Survey, Methods for Collections and Analysis., Working paper from Department for Statistics on Individuals and Households. Statistics Norway.

Kalman, R.E. (1960). A new approach to linear filtering and prediction problems. ASME Transactions, Part D (Journal of Basic Engineering), 82, 35-45.

Kalman, R.E. & Bucy, R. S. (1961). New results in linear filtering and prediction theory. ASME Transactions, Part D (Journal of Basic Engineering), 83, 95-198.

Liang, K. (1987). A locally most powerful test for homogeneity. Biometrika, 74, 259-64.

Rodriguez. G. & Goldman. N. (1995). An assessment of estimation procedures for multilevel models with binary responses. Journal of the Royal Statistical Society, A, 158, part 1, 73-89.

Rosner, B. (1984). Multivariate methods in ophthalmology with application to other paired-data situations. Biometrics 40, |025-1035.

Shumway, R. H. & Stoffer, D. S. (1982) An approach to time series smoothing and forecasting using the EM-algorithm. Journal of time series analysis Vol 3, No 4.

Zeger, S.L. & Karim, M.R. (1991) Generalized linear models with random effects; A Gibbs sampling approach. Journal of the American Statistical Assosiation, 86, 79- 86.

```
/* Program for å lage vekter for Forbruksundersøkelsen    */
/*                                                        */
/*   laget av Liv Belsby                                  */
/*                                                        */
/*   Metoden er dokumentert i notat  95/18                */



************************  INPUT  ***************************

/* datafil med variablene

id: identifikasjonsnr
rh: respons ikke respons   0 ved respons 1 ved ikke respons
f:  antall personer i husholdningen fra utvalgsregisteret
j:  antall personer i husholdningen etter innsjekking
s:  sommer ikke sommer sommer 0 er ikke sommer 1 er sommer
b:  bosted 0 er by 1 er ikke by

************** ENDRINGER SOM MÅ GJØRES I PROGRAMMET  **********

Gi navn på infile (rett nedenfor her )       */

data brutto;

/*   datafile navn  */
infile 'hel93';

input id rh f j s b;
jorig=j;
if jorig > 6 then jorig=6;

/*  skriv inn antall i bruttoutvalget/antall i populasjonen */
pistar=(2180/4131851)*jorig;

if f > 3 then f=3;
if j > 3 then j=3;
if rh= 0 then r=1;
if rh=1 then r=0;
* rh =0 ved respons 1 ved ikke respons;

options ps=200;
data netto;
set brutto; if r=1;

title1 'frafallsmodell ';
title2  'rh er 0 ve respons og e eller';
title3  'stort  koeff, større sh for resp';
proc probit data=brutto covout lackfit;
class id rh b j f ;
model rh=j b /D=logistic;
output out =ut p=prob;

proc sort data=netto; by id;
proc sort data=ut; by id;


data netto;
merge netto(in=x) ut(in=y);
by id;
if x and y;
pistar=pistar*prob;

/*  antall i bruttoutvalget i stedet for 2180   */
khjelp=1/(2180*jorig*prob);
```

```
proc means noprint data=netto;
var khjelp;
output out=ut
sum=k;

data netto;
merge netto ut;
 retain k kny 0;
 if k ne . then kny=k;
 else k=kny;

data netto;
set netto;
pi=k*pistar;

* proc print;
* var id b jorig pi;

data dat1;
set netto;
* if r=1;
vektma=1/pi;
invpist=1/pistar;

title 'estimert antall husholdner i populasjonen';
proc means data=dat1 sum;
var invpist;
output out=dat2
sum=sumv;

data netto;
merge dat1 dat2;
 retain sumv sumvny 0;
 if sumv ne . then sumvny=sumv;
 else sumv=sumvny;


data netto;
set netto;
vektmi=vektma/sumvny;

title1 'sum vekter for mikro og makro, vektmi vektma';
title2 'sum mikro er estimert gjsn. hush.str.';
title3 'sum makro er totalt antall i populasjonen';
proc means sum n mean min max data=netto;
var vektmi vektma;

proc sort data=netto;
by b jorig;

proc means noprint;
by b jorig;
var vektmi vektma;
output out=ut
max =vmikro vmakro;

data ut;
set ut;
bosted=b;
hushstr=jorig;


label
hushstr='innsjekket husholdn. str'
vmikro ='vekt for gjennomsnittstall'
```

```
vmakro='vekt for å få totaltall';

title1 'vekter for mikro- og makrotall';
title2 'bosted, 0 er by, 1 er ikke-by';
proc print;
var bosted hushstr vmikro vmakro;
run;
```

```
/* program som beregner s-test observartor */
/* se artikkel: Daniel Commenges et. al. biometrics 1994 */
/* variable:
1: gruppe  her kommune
2: y, dvs observert 1 betyr respons 0 betyr ikke respons
3: p estimert sh for y=1

forklaringsvariable:

4: z1: 1 hvis husholdningsstr = 1 0 ellers
5: z2: 1 hvis husholdningsstr= 2 0 ellers
6: bosted 0 betyr by 1 betyr land

datafilen må ha to kopier av den siste observasjonen
må være sortert på variabelen som definerer gruppene.
denne variablene må ligge i første kolonne på datafilen.
*/


load dat[2181,6]=homtest.dat;
output file=homtest.out reset;
gruppe=dat[.,1];
y=dat[.,2];
p=dat[.,3];

/*  forklaringsvariable  */

x=dat[.,4:6];
dat=0;

i=1;

s=0;
itt=0;
igg=0;
itg=0;
do while i< rows(y);
   ii=i+1;
   y_p=y[i]-p[i];
   sump=p[i];
   pq=p[i]*(1-p[i]);
   pq6=p[i]*(1-p[i])*(1-6*p[i]*(1-p[i]));
   pqzz=p[i]*(1-p[i])*(x[i,.]'.*x[i,.]);
   pqz=p[i]*(1-p[i])*(0.5-p[i])*x[i,.];
   do while (ii< rows(y) and gruppe[ii]==gruppe[i]);
       format /rd 10,4;
       y_p=y_p+y[ii]-p[ii];
       pq=pq+p[ii]*(1-p[ii]);
       pq6=pq6+p[ii]*(1-p[ii])*(1-6*p[ii]*(1-p[ii]));
       pqzz=pqzz+p[ii]*(1-p[ii])*(x[ii,.]'*x[ii,.]);
       pqz=pqz+p[ii]*(1-p[ii])*(0.5-p[ii])*x[ii,.];
       sump=sump+p[ii];
       i=ii;
       ii=ii+1;
   endo;
   s=s+y_p^2-pq;
   itt=itt+pq6+2*(pq^2);
   igg=igg+pqzz;
   itg=itg+pqz;
   i=i+1;
endo;
"s" .5*s;
"itt" itt/4;
"igg " igg;
"igt" itg;
i=itt-itg*inv(igg)*itg';
```

```
"i" i;
"test:" s/sqrt(i);
```

Statistisk sentralbyrå

Oslo
Postboks 8131 Dep.
0033 Oslo

Telefon: 22 86 45 00
Telefaks: 22 86 49 73

Kongsvinger
Postboks 1260
2201 Kongsvinger

Telefon: 62 88 50 00
Telefaks. 62 88 50 30

Statistisk sentralbyrå
Statistics Norway