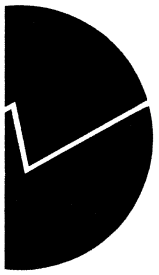
Statistics Norway



Statistisk sentralbyrå

Liv Belsby og Anne Vedø

Frafallsanalyse av Helseundersøkelsen 1995

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

Forord

Statistisk sentralbyrå gjennomførte høsten 1995 en landsdekkende undersøkelse om helse- og helseforhold. Dette notatet belyser aspekter ved frafallet på Helseundersøkelsen 1995. Seksjon for intervjuundersøkelser har gjennomført datainnsamlingen og tilrettelagt datafiler[1] og Liv Belsby og Anne Vedø fra Seksjon for statistiske metoder og standarder har gjennomført analyser på frafallet i undersøkelsen. Jorun Ramm ved Seksjon for helsestatistikk har redigert notatet og skrevet innledningen.

1 Innledning

Helseundersøkelsen 1995 baserer seg på besøksintervju, og intervjuene til undersøkelsen ble i hovedsak gjennomført i perioden september - desember 1995. Undersøkelsen omfattet også et papirskjema som skulle returneres postalt i etterkant av intervjuene.

Til undersøkelsen ble det trukket et utvalg[2] på 5 100 personer 16 år og over. Av disse gikk 150 personer til avgang (død, flytting, m.v.), mens 4 950 ble kontaktet med sikte på intervju. I tillegg til trekkepersonen skulle alle personer som tilhørte samme husholdning intervjues. Det ble også trukket et tilleggsutvalg av 1 000 personer 80 år og eldre. I dette tilleggsutvalget ble kun trekkepersonen intervjuet, ikke øvrige husholdsmedlemmer. 170 personer var ikke aktuelle pga. dødsfall, flytting mv. Bruttoutvalget til Helseundersøkelsen 1995 omfatter totalt 13 662 personer. For personer i bruttoutvalget er det koplet til opplysninger fra utdannings- og inntektsregisteret 1995. Nettoutvalget består av 10 248 personer.

I intervjudelen ble spørsmålene stilt av en intervjuer, og svarene registrert elektronisk på PC (CAPI)[3]. En del spørsmål ble imidlertid vurdert slik, at flere sannsynligvis ville svare, og man ville oppnå bedre svarkvalitet dersom spørsmålene ble stilt anonymt på papirskjema. Et skjema for selvutfylling ble delt ut til personer mellom 14 og 79 år som hadde gjennomført hovedintervjuet. Skjemaet ble enten samlet inn av intervjuer etter avsluttet intervju, eller returnert postalt. Datainnsamlingen til skjemadelen pågikk parallelt med hovedundersøkelsen. Det ble sendt èn postal purring på manglende svar. Siste purring ble sendt 12/2-1996, og sluttstrek ble satt ved påske 1996.

Frafallet for intervjudelen av undersøkelsen er 25 prosent for hele utvalget, og vel 26 prosent dersom utvalget avgrenses til å gjelde personer mellom 14 og 79 år. Av personer (14-79 år) som ble intervjuet unnlot vel 14 prosent å levere inn selvutfyllingsskjemaet til tross for fullført

---

[1] Et dokumentasjonsnotat om undersøkelsen er under utarbeidelse ved Seksjon for Intervjuundersøkelser; pers. med. Stein Opdahl

[2] Utvalget ble trukket i tråd med SSBs Standard Utvalgsplan etter revisjon i 1994/1995. Revisjonen er beskrevet i Hoel, Thomas, Jenny-Anne S. Lie og Stein Opdahl (1995): Revisjon av SSB's utvalgplan. Dokumenasjonsrapport. En utvidet utgave blir utgitt i 1998 i serien Notater.

[3] Computer Assisted Personal Interviewing

intervju. Intervjuobjektene er de samme i intervjudelen og skjemadelen. Barn og unge i alderen 0 til 14 år og eldre 80 år og over har ikke mottatt skjema.

Analyser av frafallet i intervjudelen av Helseundersøkelsen 1995 er gjort av Liv Belsby (tidligere publisert i Samfunnsspeilet 2/97) og analyser av frafallet i den postale delen av undersøkelsen er utført av Anne Vedø. Liv Belsby ser på hvordan enhetsfrafall påvirker estimatene fra intervjudelen av Helseundersøkelsen 1995. Hun fant at estimatene forandret seg lite etter korrigering for enhetsfrafall. Personer som ikke har svart på intervjudelen har heller ikke fått utlevert noe postalt skjema, og blir følgelig også enhetsfrafall på skjemadelen av undersøkelsen. Anne Vedø ser i sin artikkel på hvordan frafallet påvirker resultatene fra skjemadelen av Helseundersøkelsen 1995. Frafallet i den postale delen av undersøkelsen er høyere enn for intervjudelen.

Frafallsanalyse - intervjudelen av Helseundersøkelsen 1995

Liv Belsby

I utvalgsundersøkelser skaper frafall usikkerhet i dataene. Frafall i personutvalg kan skape skjevhet siden personer som ikke vil være med i undersøkelsen ofte atskiller seg fra resten av utvalget (Cochran 1951). Generelt har frafallet økt i spørreundersøkelser. Mange er dessuten vanskelig å få tak i eller har liten tid fordi de tilbringer mindre tid hjemme i dag enn tidligere. For eksempel er 'travle' yngre mennesker sjeldnere hjemme og tilgjengelige for intervjuing enn mer etablerte barnefamilier. Dessuten konkurrerer nå mange spørreundersøkelser om tiden til intervjupersonen. Svarprosentene i helseundersøkelsene gjenspeiler en trend mot lavere oppslutning i utvalgsundersøkelser. I 1968 var svarprosenten i helseundersøkelsen 93,61, i 1975 var den 88,55, i 1985 sank den med nesten ti prosent til 78,70, og i 1995 var svarprosenten 74,98. I tillegg til endrede holdninger og tidsknapphet, kan det voksende antallet spørsmål i spørreskjemaene (antall spørsmål for de fire helseundersøkelsene) ha påvirket responsraten.

Tabell 1 *Frafall og frafallsårsaker i Helseundersøkelsen 1975 og 1995*

	Helseundersøkelsen 1975		Helseundersøkelsen 1995	
Årsak til frafall	antall personer	prosent av utvalget	antall personer	prosent av utvalget
Ønsker ikke å delta	684	5,50	2061	15,09
Bortreist/borte pga. reiser/ arbeid/studier el.	285	2,29	165	1,21
Andre årsaker til at personen ikke er å treffe	256	2,06	178	1,30
Andre årsaker til frafall	199	1,60	1013	7,42
Totalt antall som ikke svarer	1 424	11,45	3417	25,02
Totalt antall i utvalget	12 434		13 662	

Tabellen viser frafall og oppgitte frafallsårsaker for henholdsvis Helseundersøkelsen 1975 og 1995. At man ikke ønsker å delta er den vanligste frafallsårsaken.

Frafallsmodell

Hva karakteriserer personer som ofte svarer, eller vice versa de som ikke svarer? Opplysninger fra registre gir mulighet for å knytte informasjon til personer som ikke er med i svardelen av utvalget. Trolig ligner frafallsmønsteret i helseundersøkelsene på frafallsmønsteret i andre person- og husholdningsundersøkelser i Statistisk sentralbyrå. I en analyse av forbruksundersøkelsen fant Belsby(1995) at ènpersonhusholdninger hadde større sannsynlighet for frafall enn husholdninger med flere personer. Personer bosatt i tettbygd strøk var også mindre villige til å være med i undersøkelsen enn de som var bosatt i grisgrendte strøk. Ved en analyse av frafall i Levekårsundersøkelsen 1980, 1983 og 1987 fant Vedø (1997) bl.a. at enslige hadde lavere sannsynlighet for å svare enn andre. Bild m.fl. (1997) fant at i Statistisk sentralbyrås undersøkelse om småbarnsfamiliers erfaringer med sykdom og helsetjenester hadde personer som mottok sosialhjelp og personer som var født i den 3. verden, større sannsynlighet for frafall enn andre. Videre fant de en tendens til at personer med lav inntekt og lav utdannelse hadde større sannsynlighet for frafall.

Motivert av resultatene i disse analysene inkluderte vi følgende registervariabler i frafallsanalysen:

- antall personer i husholdningen[4]
- alder til den eldste i husholdningen
- høyeste utdannelse i husholdningen
- om noen i husholdningen mottok sosialhjelp
- overgangsstønad
- landbakgrunn etter verdensregion inndeling[5]
- tettbygd/spredtbygd bosted
- den høyeste personinntekten i husholdningen[6]

[4] For husholdninger der faktisk antall personer i husholdningen er ukjent, brukes familiestørrelse fra Det sentrale personregistret.

[5] Landbakgrunn viser til personens fødeland. Hvis barn har forskjellig fødeland fra foreldrene, blir landbakgrunn mors fødeland. For husholdninger med personer fra de forskjellige verdensregioner, brukes den verdensregionen som ligner kulturelt mest på vår norske, vesteuropeiske kultur. Er for eksempel én i husholdningen fra Vest-Europa og de andre fra Afrika, vil husholdningen bli klassifisert som vesteuropeisk.

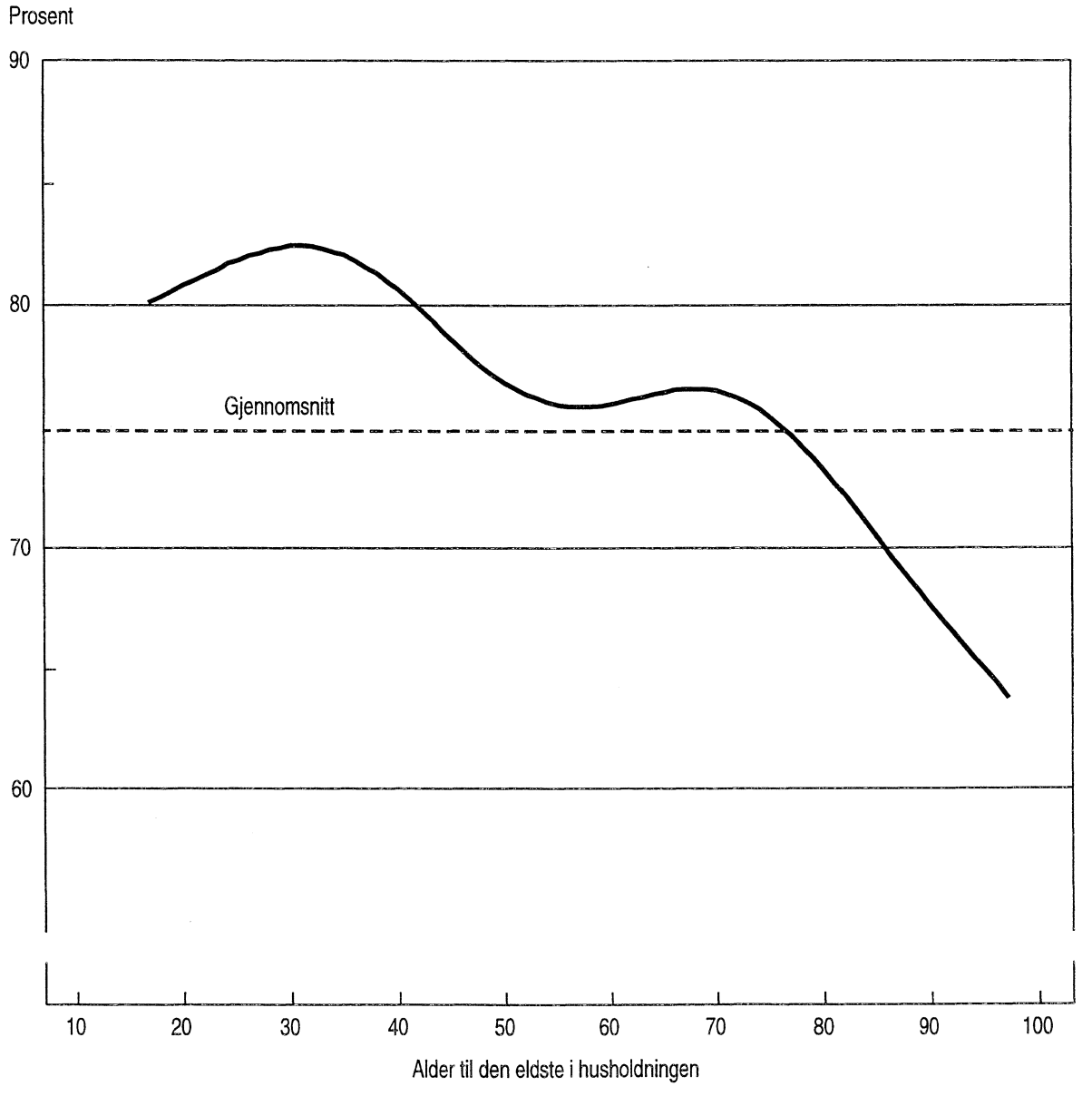
[6] Personinntekt er disponibel inntekt, som i selvangivelsen, pluss summen av utlignet skatt.

Modellen estimeres på husholdningsnivå. Hvis minst én person i husholdningen er med i undersøkelsen får responsvariabelen verdien 1, ellers får den verdien 0. Variablene for alder og inntekt blir brukt som kontinuerlige variable i modellen. De andre forklaringsvariable kodes som kategoriske variable. Det er vanlig å formulere denne typen sammenheng med en logistisk «link».

I tradisjonelle logistiske modeller må eventuelle transformasjoner av forklaringsvariablene defineres før estimeringen. For eksempel kan en ha en formening om at alder inngår med både 1. og 2. grads ledd. Ofte er det en ulempe å måtte formulere funksjonsformen på forhånd fordi en ikke har nok kunnskap om hva som er god transformasjon. Ved å velge en ''tilfeldig'' transformasjon kan en risikere å ikke finne en god tilpasning til dataene. Moderne regresjonsmetoder (Venables and Ripley 1994) beregner en ikke-parametrisk funksjon ved ''spline''-metoder. Fordelen er at man slipper å transformere basert på mer eller mindre godt funderte hypoteser om funksjonsformen. Den estimerte sammenhengen mellom alder og frafall viser at alder bør inngå som et fjerdegrads polynom i modellen.

Fordeling på ulike kjennetegn i modellen

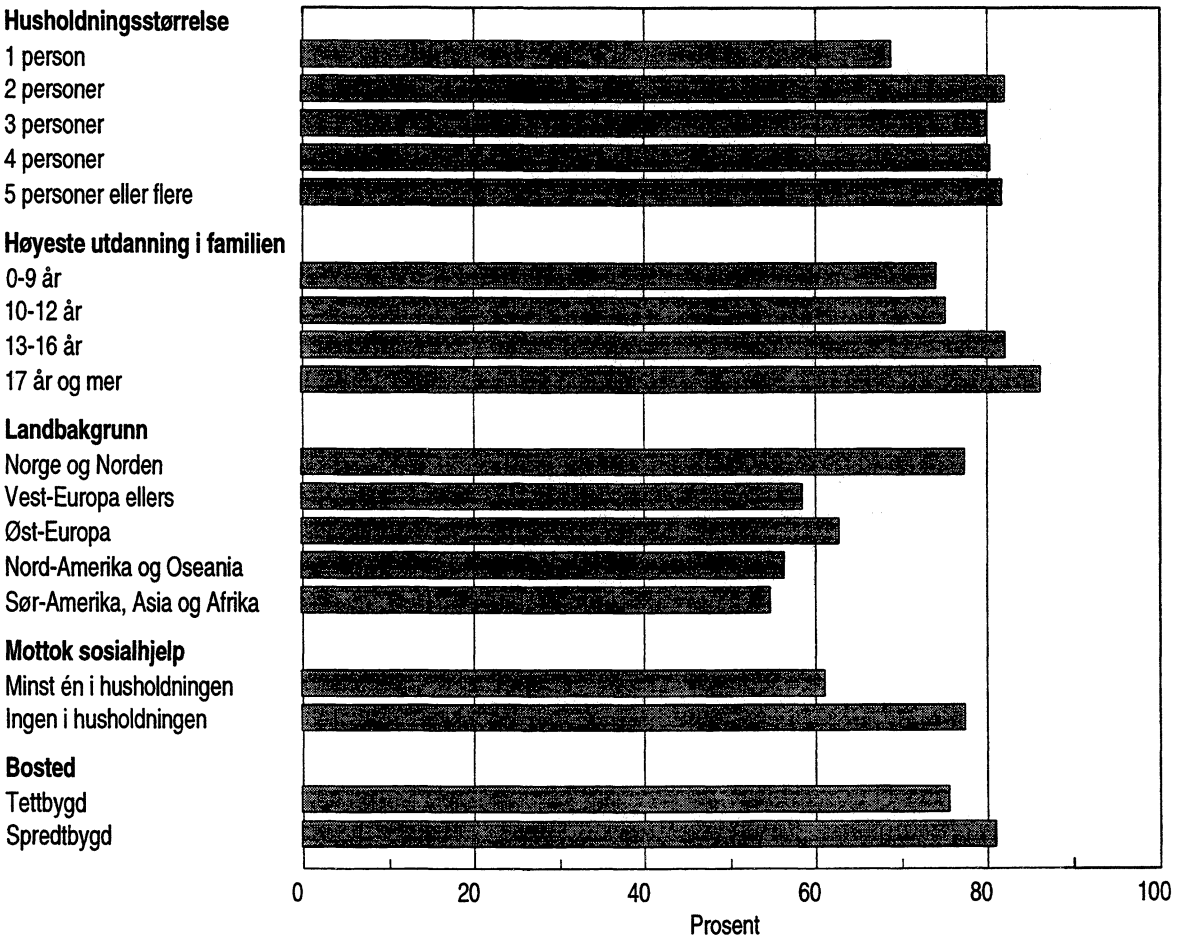
Figur 1 *Estimert sannsynlighet for å være med i Helseundersøkelsen 1995 for husholdninger, etter alder til den eldste i husholdningen. Prosent*



Kilde: Helseundersøkelsen 1995

Figuren viser andelen av husholdningene som svarer etter hvor gammel den eldste i husholdningen er. Husholdninger hvor den eldste personen er omtrent tredve år har den største svarsannsynligheten. Deretter synker responssannsynligheten frem til vel femti år, for deretter å øke igjen. Rundt sytti år er det et ikke helt tydelig toppunkt, før responssannsynligheten klart avtar med økende alder.

Figur 2 *Estimerte sannsynligheter for å være med i Helseundersøkelsen 95 for grupper med ulike kjennetegn. Prosent*



Kilde: Helseundersøkelsen 1995

Figuren viser estimerte sannsynligheter for respons på de kategoriske forklaringsvariablene som ble funnet å være signifikante ved minst 1 prosentsnivå i den logistiske modellen. Husholdninger med én person har en svarsannsynlighet på rundt 69 prosent. Husholdninger med fra to til fem eller flere personer har svarsannsynlighet rundt 80 prosent. Alder og husholdningsstørrelse er korrelert. En stor andel av èn-person husholdningene er kvinner over 80 år. Videre ser vi at husholdninger hvor personen(e) med høyest(e) utdannelse har utdannelse på universitet eller høyskole nivå har svarprosent på omtrent 86 prosent. Mens husholdninger hvor høyeste utdannelse er på grunnskolenivå ligger under 70 prosent.

Husholdninger hvor ingen er norskfødte har lavere svarsannsynlighet, rundt 55-60 prosent, enn husholdninger med minst én norsk person. Personer som får sosialhjelp har lavere svarsannsynlighet enn personer som ikke får sosialhjelp. Husholdninger i spredtbygde strøk har større svarsannsynlighet enn husholdninger i tettbygde strøk.

Økonomien i husholdet, målt med den høyeste personinntekten, påvirker ikke svarsannsynligheten noe særlig. Gruppen som mottar overgangsstønad, det vil si enslige foreldre, svarer omtrent like ofte som andre. Følgelig er disse to variablene ikke med i modellen for å beregne svarsannsynligheter.

En statistisk test viser hvilken forklaringsvariabel som alene betyr mest i modellen. Den gav følgende rekkefølge med den viktigste forklaringsvariabelen først: antall personer i husholdningen, alder til den eldste i husholdningen, høyeste utdannelse i husholdningen, landbakgrunn etter verdensregion inndeling, om noen i husholdningen mottok sosialhjelp og bosted.

Korrigering for frafall

Siden det er overhyppighet av noen kjennetegn blant de som ikke svarer, vil det kunne skape skjevhet i materialet. For å få et anslag på hvor stor skjevhet frafallet skaper, beregnes anslag på noen typer sykelighet med og uten korrigering for frafall.

Metoden som benyttes stammer fra Politz og Simmons (1949). Metoden går ut på å justere de opprinnelige vektene i helseundersøkelsen, basert på inverse trekkesannsynligheter, med den inverse svarsannsynligheten. Siden personer over 80 år er overrepresentert i helseundersøkelsen blir det benyttet to versjoner av frafallsmodellen for å beregne svarsannynligheter. Den ene modellen inkluderer alle forklaringsvariablene som påvirket frafallet signifikant, se figurene 1 og 2. I den andre modellen er ikke alder eller antall personer i husholdningen med. Antall personer er sterkt korrelert med alder idet eldre mennesker ofte bor alene. Derfor ble også denne variabelen fjernet fra modellen. Tabell 2 viser estimert andel syke uten frafallsjustering ,og justert for frafall basert på disse to modellene.

Tabell 2 *Tall for sykelighet for Helseundersøkelsen 1995 beregnet med, og uten, to alternative modeller for å korrigere for frafall. Prosent*

Sykdom	Ingen justering for frafall	Modell for svarsannsynlighet	
		Inkluderer høyeste utdannelse i husholdet, landbakgrunn, sosial-hjelp, tettbygd/ spredtbygd	Inkluderer alder til eldste i husholdet, antall personer i husholdet, høyeste utdannelse i husholdet, landbakgrunn, sosialhjelp, tettbygd/ spredtbygd
Varig sykdom	57,24	57,46	58,04
Nervøse lidelser	6,77	6,87	7,01
Hjerte-/karsykdommer	14,35	14,43	14,92
Sykdom i åndedrettsorganer	17,44	17,54	17,50
Sykdom i hud/underhud	4,46	4,48	4,50
Skjelett-/muskelsykdommer	24,79	25,34	24,93

Tallene i tabellen tyder på at frafallet bare i liten grad påvirker sykelighetstallene for hele befolkningen. Det kan tenkes at andre typer sykelighet eller undergrupper eller geografiske områder påvirkes mer. De som ikke svarer ser imidlertid ut til å være litt mer syke enn de som er med i undersøkelsen.

Metoden for å justere for frafall forutsetter homogenitet i grupper med lik verdi på forklaringsvariablene i frafallsmodellen. Gitt utdannelsesnivå, landbakgrunn i husholdningen, om noen i husholdningen mottar sosialhjelp eller ikke og om husholdningen bor i tettbygd eller spedbygd strøk, eventuelt også alder til eldste i husholdningen og husholdningens størrelse, forutsettes sannsynligheten for sykdom å være den samme for personer som svarer og personer som ikke svarer. For å undersøke om denne forutsetningen holder, måtte personer som ikke svarer bli oppsøkt og på en eller annen måte overtalt til å være med i helseundersøkelsen.

Denne analysen tyder på at en frafallsprosent på rundt 25 prosent ikke svekker kvaliteten særlig på helseundersøkelsen. Likevel ville det naturligvis være ønskelig å øke deltagelsen. I den amerikanske National Health Interview Survey lyktes det å holde svarprosenten på 95 prosent i perioden 1985 til 1989 (Couper 1996). I den samme perioden økte gjennomsnittlig antall forsøk på å komme i kontakt med intervju-objektet fra 2,56 til 2,95. Kanskje burde man for den neste helseundersøkelsen anstrenge seg ytterligere for å øke svarprosenten?

Frafallsanalyse - skjemadelen av Helseundersøkelsen 1995

Anne Vedø

Liv Belsby (Samfunnsspeilet 2/97) har sett på hvordan enhetsfrafall påvirker estimatene fra intervjudelen i Helseundersøkelsen 1995. Vi skal her forsøke å finne ut noe om hvordan enhetsfrafall påvirker resultatene fra skjemadelen. Merk at vi i denne analysen kun tar for oss personer som skal svare på det postale skjemaet, nemlig de mellom 14 og 79 år. Blant personer mellom 14 og 79 år var det 26,3 prosent enhetsfrafall på intervjudelen. Personer som ikke har svart på intervjudelen får ikke utlevert noe postalt papirskjema, og blir følgelig også enhetsfrafall på skjemadelen. Av 73,7 prosent som svarte på intervjudelen var det 14,5 prosent som ikke returnerte det postale skjemaet. Dette gir et samlet enhetsfrafall på 37 prosent på skjemadelen. 37 prosent av det opprinnelige bruttoutvalget svarte ikke på skjemaet enten fordi de ikke svarte på intervjudelen eller fordi de ikke leverte/sendte inn skjemaet etter intervjuet. Modellen vi bruker i analysen under tar bare hensyn til enhetsfrafall. Partielt frafall blir bare kort kommentert.

Helseundersøkelsen foregår på personnivå, og filene inneholder en record for hver person. Frafallsmodelleringen under er imidlertid foretatt på husholdningsnivå. Dette er fordi modellen vi bruker forutsetter uavhengige observasjoner, dvs. at det den ene enheten gjør ikke skal påvirke de andre enhetenes beslutning i noen retning. Dette er ikke oppfylt på personnivå. Det viser seg at husholdningsmedlemmer stort sett tar samme beslutning når det gjelder om de skal svare eller ikke svare på undersøkelsen. I 56 prosent av husholdningene deltok alle medlemmene, og i 30,8 prosent deltok ingen. I de resterende 13,2 prosent av husholdene var det noen som svarte og noen som ikke svarte. I skjemadelen definerer vi enhetsfrafall på

husholdningsnivå ved at en husholdning defineres som 'med' i undersøkelsen dersom *minst* ett medlem har svart på skjemaet, og som frafall dersom ingen medlemmer har svart. Likeledes regnes et husholdningsmedlem som 'med' i undersøkelsen dersom han/hun har svart på *minst* ett spørsmål på skjemaet. Når vi regner på denne måten får vi 30,8 prosent enhetsfrafall på husholdningsnivå.

I likhet med Liv Belsby anses følgende åtte registervariable som aktuelle forklaringsvariable:

- antall personer i husholdningen
- alder til den eldste i husholdningen
- høyeste utdanning i husholdningen
- om noen i husholdningen mottok sosialhjelp
- om noen i husholdningen mottok overgangsstønad
- landbakgrunn etter verdensregioninndeling
- tettbygd/spredtbygd bosted
- husholdningens samlede inntekt

For husholdninger som ikke har svart på undersøkelsen brukes familieopplysninger fra Det sentrale personregisteret. I vedlegg 1 gis en oversikt over verdiene disse variablene kan ha, både på opprinnelig personnivå og på husholdningsnivå slik vi har brukt dem i analysen.

Frafallsmodell

Vi modellerer sannsynligheten for at en husholdning svarer på skjemadelen, p, med en logistisk modell:

$$\ln(\frac{p}{1-p}) = \beta_0 + \beta_{1,x_1} + \beta_{2,x_2} + \cdots + \beta_{n,x_n}, \text{ der}$$

$x_1,\ldots,x_n$ er husholdningens verdier på forklaringsvariablene som er med i modellen, β_{i,x_i} er parametre (ukjente tall) knyttet til verdien x_i på variabel nummer i, $i = 1,\ldots,n$.

Dette er en logistisk modell med kategoriske forklaringsvariable $x_1,\ldots,x_n$. Til hver forklaringsvariabel x_i hører det like mange parametre som antall kategorier x_i er delt inn i. Hvis vi for eksempel har to forklaringsvariable, si alder (x_1) og utdanning (x_2), og alder deles inn i tre grupper (1,2,3) og utdanning i fire (1,2,3,4), ville modellen bli

$$\ln(\frac{p}{1-p}) = \beta_0 + \beta_{1,x_1} + \beta_{2,x_2}$$

Vi ville få tre aldersparametre $\beta_{1,1},\beta_{1,2},\beta_{1,3}$ og fire utdanningsparametre $\beta_{1,1},\beta_{2,2},\beta_{2,3},\beta_{2,4}$, i tillegg til konstantleddet β_0. For en husholdning der eldste person er i aldersgruppe 1 og høyeste utdanning er i utdanningskategori 3 får vi altså

$$\ln(\frac{p}{1-p}) = \beta_0 + \beta_{1,1} + \beta_{2,3}$$

Parametrene (β-ene) er ukjente tall som må estimeres på grunnlag av dataene. Vi gjør dette i SAS, som benytter en numerisk metode til å finne maximum-likelihood-estimatorer.

Merk at vi modellerer sannsynligheten for at en husholdning svarer på skjemadelen direkte, uten å ta hensyn til om husholdet har svart på intervjudelen. En annen mulighet er å modellere p som $p_1 . p_2$, der p_1 er sannsynligheten for at husholdet svarer på intervjudelen, og p_2 sannsynligheten for at husholdet svarer på skjemadelen, gitt at den har svart på intervjudelen, og deretter tilpasse to logistiske modeller, en for p_1 og en for p_2. Til slutt i denne analysen har vi laget en slik totrinnsmodell for å sammenligne, og det viste seg at vi fikk nesten nøyaktig de samme frafallskorrigerte estimatene.

Totrinnsmodellen gir likevel noe ekstra informasjon, ettersom den skiller mellom de variablene som har betydning for hvorvidt en husholdning svarer på intervjudelen, og de som har betydning for hvorvidt husholdningen svarer på skjemaet når den har svart på intervjudelen.

Univariate analyser

Vi kjører først logistisk regresjon med én og én forklaringsvariabel. På denne måten finner vi ut hvilke variable som alene har signifikant betydning for enhetsfrafall.

P-verdier for forklaringsvariablene:

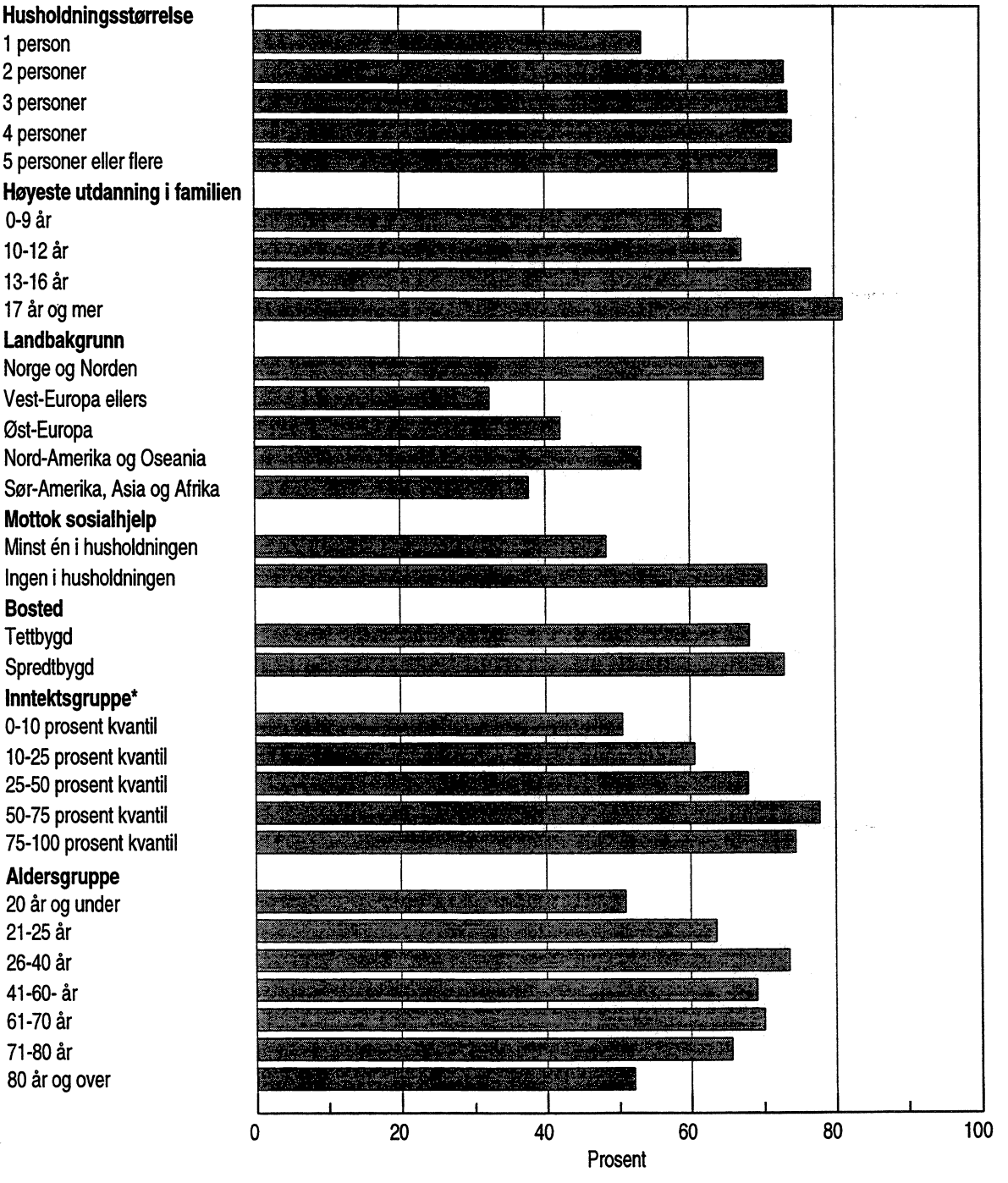
Variabel	P-verdi
Landbakgrunn	0,0001
Bosted	0,0001
Overgangsstønad	0,8209
Inntekt	0,0001
Alder	0,0001
Sosialbidrag	0,0001
Utdanning	0,0001
Husholdningsstørrelse	0,0001

P-verdien sier hvor signifikant variabelen er. Jo lavere p-verdien er, desto mer signifikant er variabelen. Tallet 0,0001 må tolkes som mindre eller lik 0,0001. Vi ser at alle variablene unntatt overgangsstønad er sterkt signifikante. Overgangsstønad har ingen betydning for enhetsfrafall, og vil ikke bli tatt med i analysen videre.

Gruppering av alder og inntekt

Opprinnelig ble samlet husholdningsinntekt delt inn i seks grupper, der kvantiler ble brukt som skille mellom gruppene. Vi skilte ved 10, 25, 50, 75 og 90 prosent kvantilene. Fordi svarandelen var omtrent lik i de to øverste inntektsgruppene ble disse slått sammen til en gruppe. lder på den eldste i husholdningen ble først delt i 9 grupper: 20 år eller yngre, 21-25 år, 26-30 år, 31-40 år, 41-50 år, 51-60 år, 61-70 år, 71-80 år og 80 år eller eldre. Etter å ha sett på svarsannsynlighetene slo vi sammen gruppene 26-30 år og 31-40 år til en gruppe 26-40 år, og gruppene 41-50 år og 51-60 år til en gruppe 41-60 år. Under følger en figur som viser andelen av husholdningene som svarer, gitt forskjellige verdier på de signifikante forklaringsvariablene.

Figur 3 *Andeler av husholdninger i bruttoutvalget som har svart på skjemadelen av undersøkelsen. Prosent*



* Bortsett fra i det første intervallet er venstre endepunkt ikke med i intervallene. Høyre endepunkt er med i alle intervaller. F.eks. vil gruppe 2 inneholde alle med inntekt over 10 prosent kvantil og under eller lik 25 prosent kvantil.

Kilde: Helseundersøkelsen 1995

Multivariat analyse

Vi tilpasser nå en logistisk modell der alle forklaringsvariablene er med samtidig.

P-verdier for forklaringsvariablene:

Variabel	P-verdi
Landbakgrunn	0,0060
Bosted	0,0021
Inntekt	0,0001
Alder	0,0001
Sosialbidrag	0,0001
Utdanning	0,0001
Husholdningsstørrelse	0,0001

Pearson Chi-square: 0,1232

Alle variablene er sterkt signifikante også i den multivariate modellen. Dette betyr at hver variabel bidrar til å forklare enhetsfrafall, selv etter at de andre variablene har forklart sitt. Pearson Chi-square-observatoren måler hvor god modelltilpasningen er i forhold til en såkalt full modell. Med full modell menes en modell som har like mange parametre som det er ulike kovariatvektorer (dvs. kombinasjoner av forklaringsvariable) i datasettet. Observatoren ligger alltid mellom 0 og 1, og høyt tall indikerer god modelltilpasning. Som tommelfingerregel kan vi si at verdier over 0,10 regnes som akseptabelt. Vi lar derfor dette være den endelige modellen.

Korrigering for frafall

Vi er nå ferdige med modelleringsfasen, og går over til å behandle dataene på personnivå. Vi korrigerer for frafall ved å multiplisere de opprinnelige personvektene med den inverse av svarsannsynligheten til den husholdningen som personen bor i:

$$\text{nyvekt} = \text{gammel vekt} \cdot (1 / p)$$

Siden p er mellom 0 og 1, vil $1 / p$ alltid være større enn 1, så alle vektene blir større på denne måten. Personer som bor i husholdninger med lav svarsannsynlighet (iflg. modellen) vil imidlertid få økt sine vekter mer enn personer som bor i husholdninger med høy svarsannsynlighet. Dette er rimelig, fordi det fører til at svarene fra underrepresenterte grupper blir «blåst opp».

Vi valgte seks spørsmål, og lagde frekvenstabeller med både gamle og nye vekter. Vi har også regnet ut det partielle frafallet på hvert spørsmål, dvs. antall som ikke svarte på spørsmålet delt på antall som leverte skjema. For spørsmål 29, om alvorlige motsetninger i hjemmet under oppveksten, regner vi ut denne andelen blant personer over 16 år, siden dette spørsmålet bare

skal besvares av personer 16 år og over. Det var 6 478 personer som leverte skjemaet, og 6 263 av disse var 16 år eller eldre.

Tabell 3 *Estimater for seks utvalgte variable i skjemadelen av Helseundersøkelsen 1995, med og uten korreksjon for enhetsfrafall. Prosent*

Spørsmål	Uten korreksjon	Med korreksjon	Partielt frafall, prosent og antall
Spm. 2. Søvnproblemer			7,38
Ikke plaget	72,6	71,6	(478)
Litt plaget	20,1	20,6	
Ganske mye plaget	5,6	6,0	
Veldig mye plaget	1,6	1,7	
Spm. 4. Røyker du sigaretter			2,22
Daglig	28,4	29,1	(144)
Av og til	9,2	9,0	
Nei	62,5	61,9	
Spm. 14. Hvor ofte trener eller mosjonerer du vanligvis			2,22 (144)
Aldri	15,1	15,6	
Sjeldnere enn 1 dag pr uke	21,8	21,5	
1 dag pr uke	20,6	20,3	
2-3 dager pr uke	28,3	28,1	
4-7 dager pr uke	14,3	14,5	
Spm. 17. Hvilken type melkeprodukt bruker du vanligvis mest av			3,44 (223)
Kulturmelk, kefir, yoghurt	4,4	4,5	
Helmelk	11,9	12,2	
Lettmelk	53,6	53,6	
Skummet melk, skummet kulturmelk	21,5	20,9	
Bruker sjelden eller aldri melk	8,6	8,8	
Spm. 23. Har du noen du kan snakke helt fortrolig med			3,38 (219)
Nei	9,7	9,9	
Ja, en	33,8	34,0	
Ja, flere	56,5	56,1	
Spm. 29. Alvorlige motsetninger i hjemmet under oppveksten			1,20 (75)
Ja	14,1	14,3	
Nei	83,6	83,3	
Vet ikke	2,3	2,4	

Det er små forskjeller på de korrigerte og de ukorrigerte tallene. Frafallet ser altså ikke ut til å ha noen særlig betydning for estimatene. Vi må likevel huske på at korrigeringen vi har gjort er basert på at personer som har de samme verdiene på forklaringsvariablene i modellen også har samme sannsynlighet for å (f.eks.) ha søvnproblemer, enten de er i frafallsgruppen eller svargruppen. Det er mulig å omgå dette problemet ved å bruke såkalt latent modellering, men det ville bli for omfattende å gjøre her.

Bortsett fra spørsmålet om søvnproblemer har alle spørsmålene over lavt partielt frafall, og det er ikke grunn til å tro at estimatene hadde endret seg vesentlig selv om vi hadde hatt de manglende svarene. Dersom alle de 144 personene som ikke har svart på spørsmål 4 er dagligrøykere, ville de korrigerte estimatene for røyking blitt 30,8 (daglig), 8,7 (av og til) og 60,4 (aldri). Dersom ingen av de 144 røyker ville vi fått 28,4 (daglig), 8,7 (av og til) og 62,8 (aldri). Vi kan gjøre et tilsvarende eksperiment på spørsmål 2, der det partielle frafallet er størst. Det verst tenkelige tilfellet er når det partielle frafallet skiller seg maksimalt fra dem som har svart på spørsmålet, dvs. dersom alle de 478 personene med partielt frafall på dette spørsmålet er veldig mye plaget. I så fall ville de korrigerte estimatene ha blitt 66,0 (ikke plaget), 19,0 (litt plaget), 5,5 (ganske mye plaget) og 9,4 (veldig mye plaget). Dette skiller seg ganske mye fra estimatene i tabellen, men det er jo også et helt hypotetisk tilfelle. Dersom hele det partielle frafallet ikke har søvnproblemer ville vi ha fått 73,8 (ikke plaget), 19,0 (litt plaget), 5,5 (ganske mye plaget) og 1,6 (veldig mye plaget), et resultat nokså likt tabellen.

Totrinnsmodell

Modellen vi har brukt hittil omfatter ikke variabelen som forteller om en husholdning har svart på intervjudelen. Den registrerer bare om en husholdning har svart på skjemaet eller ikke, og tar ikke hensyn til at frafall på skjemadelen kan ha vært forårsaket av frafall på intervjudelen. Vi skal nå se på en litt mer sofistikert modell som utnytter informasjonen om svar på intervjudelen. I tillegg til å anvende mer av informasjonen i datasettet har den nye totrinnsmodellen den fordelen at vi får se at det er ett sett av registervariable som har betydning for frafall på intervjudelen, og et litt annet sett som har betydning for frafall på skjemadelen når en husholdning allerede har svart på intervjudelen.

La p_1 være sannsynligheten for at en husholdning svarer på intervjudelen av undersøkelsen, og la p_2 være sannsynligheten for at husholdet svarer på skjemaet gitt at den har svart på intervjudelen. Da er den ubetingede sannsynligheten for at husholdet svarer på skjemaet, p, lik $p_1 \cdot p_2$. Vi har tilpasset logistiske modeller for p_1 og p_2 med de samme forklaringsvariablene og de samme gruppeinndelingene som tidligere.

Intervjudelen

For husholdninger som skal svare på skjemadelen av undersøkelsen (har minst ett medlem mellom 14 og 79 år) tilpasser vi først en logistisk modell for sannsynligheten for å svare på intervjudelen:

$$\ln(\frac{p_1}{1-p_1}) = \beta_0 + \beta_{1,x_1} + \beta_{2,x_2} + \cdots + \beta_{n,x_n}$$

P-verdier for forklaringsvariablene:

Variabel	P-verdi
Landbakgrunn	0,0172
Bosted	0,0006
Inntekt	0,0115
Alder	0,0001
Sosialbidrag	0,0001
Utdanning	0,0001
Husholdningsstørrelse	0,0001

Pearson Chi-square: 0,0765

Vi gjorde også et forsøk på å ta med overgangsstønad i modellen, men denne variabelen ble ikke signifikant (p-verdi lik 0,7825) og Pearson Chi-square-observatoren ble mindre (0,0195).

Skjemadelen

Så tilpasser vi en logistisk modell for sannsynligheten for at husholdningen svarer på skjemadelen gitt at den har svart på intervjudelen:

$$\ln(\frac{p_2}{1-p_2}) = \alpha_0 + \alpha_{1,x_1} + \alpha_{2,x_2} + \cdots + \alpha_{n,x_n}$$

P-verdier for forklaringsvariablene:

Variabel	P-verdi
Landbakgrunn	0,2898
Bosted	0,6706
Inntekt	0,0023
Alder	0,0013
Sosialbidrag	0,0155
Utdanning	0,0179
Husholdningsstørrelse	0,0050

Pearson Chi-square: 0,3981

Vi forsøkte også her å ta med overgangsstønad i modellen, men denne variabelen var heller ikke her signifikant (p-verdi lik 0,8462) og Pearson Chi-square-observatoren ble noe mindre (0,3445). Landbakgrunn og særlig bosted kommer ikke signifikant ut i denne modellen. Vi tilpasset en modell uten bosted, og en modell der vi fjernet både bosted og landbakgrunn, men i begge tilfeller fikk vi dårligere modelltilpasning (0,3410 og 0,1723 hhv.), så vi lot disse variablene være med likevel.

Korrigering for frafall

Vi korrigerte for frafall på samme måte som før, bortsett fra at vi brukte de nye modellsannsynlighetene. Med to ubetydelige unntak ble de nye korrigerte prosentene, med en desimals nøyaktighet, akkurat de samme som før. Unntakene var at prosenten som røyker daglig endret seg fra 29,1 til 29,2, og prosenten som foretrakk helmelk ble 12,3 mot tidligere 12,2.

Konklusjon

Vi har sett på to forskjellige modeller for enhetsfrafall på skjemadelen av helseundersøkelsen. I begge modellene var det de åtte registervariablene listet på side 10 som var aktuelle som forklaringsvariable.

Den første modellen var en enkel logistisk modell, der sannsynligheten for at husholdet svarer på skjemadelen ble modellert direkte, uten å ta hensyn til om husholdningen hadde svart på intervjudelen av undersøkelsen. Her fant vi at de syv variablene *landbakgrunn, bosted, husholdningsinntekt, alder til den eldste i husholdningen, om husholdningen mottar sosialbidrag, høyeste utdanning i husholdningen* og *husholdningsstørrelse* hadde betydning for en husholdnings svarsannsynlighet på skjemadelen. Overgangsstønad hadde ikke signifikant betydning.

Den andre modellen var en totrinnsmodell. Her tenkte vi oss at frafallet på skjemadelen foregår i to trinn, først de husholdningene som faller fra allerede på intervjudelen, og derfor også på skjemadelen, og deretter de husholdningene som blir intervjuet, men som ikke sender inn skjemaet. Totrinnsanalysen ga et litt mer nyansert bilde av hvordan de syv variablene nevnt over påvirker frafallssannsynligheten på skjemadelen. Vi fant at alle syv hadde betydning for første trinn i prosessen (frafall på intervjudelen), mens alle syv unntatt landbakgrunn og bosted hadde betydning for andre trinn (frafall på skjemadelen gitt svar på intervjudelen). Overgangsstønad hadde ikke betydning for noen av trinnene.

Først brukte vi den enkleste modellen til å korrigere vektene i undersøkelsen, og vi sammenlignet frekvenstabeller basert på opprinnelige vekter og frafallskorrigerte vekter for seks spørsmål fra skjemaet. Vi fant at det var liten forskjell på de korrigerte og de ukorrigerte tallene. Dette betyr at *dersom modellen vår er riktig* ser ikke frafallet ut til å være noe stort problem på de seks spørsmålene vi så på.

Vi lagde også korrigerte vekter basert på totrinnsmodellen, og det viste seg at de korrigerte tallene forble så å si uendret.

Vedlegg 1

Variable som inngår i analysen av skjemadelen

Personnivå

Filen med personer inneholder 10 288 records. Dette er husholdningene til personer mellom 14-79 år i bruttoutvalget.

Variabel-navn	Forklaring	Verdier	Antall missing
MED	Enhetsfrafall på skjemadelen på personnivå	1=Personen har svart på minst ett av spørsmålene på skjemaet 0=Personen har ikke svart på noen av spørsmålene på skjemaet	0
SVARINT	Enhetsfrafall på intervjudelen på personnivå	1=Personen har svart på intervjudelen 0=Personen har ikke svart på intervjudelen	0
LAND	Personens landbakgrunn, verdensregioninndeling.	0=Norge 1=Norden ellers 2=Vest-Europa ellers 3=Øst-Europa 4=Nord-Amerika og Oseania 5=Sør-Amerika, Asia og Afrika	9
BY	Om personen bor i tettbygd eller spredtbygd strøk	0=Spredtbygd 1=Tettbygd	24
OVERGST	Om personen mottar overgangsstønad	0=Får ikke overgangsstønad 1=Får overgangsstønad	5
SAMINNT	Personens samlede inntekt	Kontinuerlig variabel	5
ALDER	Personens alder	Kontinuerlig variabel	0
SOSBIDR	Om personen mottar sosialbidrag	0=Får ikke sosialbidrag 1=Får sosialbidrag	5
UTDAN	Personens utdanningsnivå	1=0-9 år 2=10-12 år 3=13-16 år 4=17 år og mer	828

På personnivå er det 37,0 % enhetsfrafall på skjemadelen og 26,3 % enhetsfrafall på intervjudelen. Blant personer som svarte på intervjudelen er det 14,5 % frafall på skjemadelen.

Husholdningsnivå

Filen med husholdninger inneholder 4 773 records.

Variabel-navn	Forklaring	Verdier	Antall missing
MEDHUSH	Enhetsfrafall på skjemadelen på husholdningsnivå	1=Minst en person i husholdningen har levert skjemaet (dvs. har MED=1) 0=Ingen i husholdningen har levert skjemaet	0
INTHUSH	Enhetsfrafall på intervjudelen på husholdningsnivå	1=Minst en person i husholdningen har svart på intervjudelen 0=Ingen i husholdningen har svart på intervjudelen	0
HUSHLAND	Den «vestligste» landbakgrunnen som forekommer i husholdningen. Vi rangerer regionene etter vestlighet på følgende måte: 0,2,4,3,5.	0=Norge og Norden 2=Vest-Europa ellers 3=Øst-Europa 4=Nord-Amerika og Oseania 5=Sør-Amerika, Asia og Afrika	1
MEANBY	Om husholdningen bor i tettbygd eller spredtbygd strøk	0=Spredtbygd 1=Tettbygd	16
HUSHOVER	Om husholdningen mottar overgangsstønad	0=Ingen i husholdningen får overgangsstønad 1=Minst en i husholdningen får overgangsstønad	0
INNTGR	Summen av husholdningsmedlemmenes inntekt, gruppert etter 10 %, 25 %, 50 % og 75 % kvantilene	1=0-102 593 kr. 2=102 594-177 038 kr. 3=177 039-293 692 kr. 4=293 693-422 196 kr. 5=Over 422 197 kr.	0
ALDGR	Alder til den eldste i husholdningen, gruppert i 7 grupper	1=Under 20 år 2=21-25 år 3=26-40 år 4=41-60 år 5=61-70 år 6=71-80 år 7=Over 80 år	0
HUSHSOS	Om husholdningen mottar sosialbidrag	0=Ingen i husholdningen får sosialbidrag 1=Minst en i husholdningen får sosialbidrag	0
MAX-UTDAN	Høyeste utdanning i husholdningen	1=0-9 år 2=10-12 år 3=13-16 år 4=17 år og mer	69
ANTP	Antall personer i husholdningen	1,2,3,4 eller 5. Flere enn fem personer regnes som fem.	0

På husholdningsnivå er det 30,8 % enhetsfrafall på skjemadelen og 22,0 % enhetsfrafall på intervjudelen. Blant husholdninger som svarte på intervjudelen er det 11,3 % frafall på skjemadelen.

Referanser

Belsby, Liv (1995): Forbruksundersøkelsen. Vektmetoder, frafallskorrigeringer og intervjuer-effekt, Notat 95/18, Statistisk sentralbyrå.

Bild, Hanna, Jon Erik Finnvold, Kari Kveim Lie, Rannveig Nordhagen og Arnfinn Schjalm (1996): Brukererfaringer med forebyggende helsetjenester, Rapport 98/(in prep.). Statistisk sentralbyrå.

Couper, Mick P. (1996): *Theoretical and Practical Aspects of Survey Nonresponse.* Metodekonferanse på SCB, Stockholm september 1996.

Hoel, Thomas, Jenny-Anne S. Lie og Stein Opdahl (1995): Revisjon av SSB's utvalgplan. Dokumenasjonsrapport, Statistisk sentralbyrå.

Politz, A., and W. Simmons (1949). An attempt to get not-at-homes into the sample without callbacks. *Journal of the American Statistical Association* 44, s. 9-31.

SSB's Standard Utvalgsplan (1977), Samfunnsøkonomiske studier nr. 33, Statistisk sentralbyrå.

Vedø, Anne (1997): Frafall i levekårspanelet 80, 83 og 87, Notat 97/31, Statistisk sentralbyrå

Venables, Bill N., Brian D. Ripley (1994): *Modern Applied Statistics with S-plus.* Springer-Verlag.

De sist utgitte publikasjonene i serien Notater

97/49 H.M. Edvardsen, J. Mønnesland og K.Ø. Sørensen: Regional arbeidsdeling: Sogn og Fjordanes plass i norsk verdiskaping. 35s.

97/50 O. Rognstad: SSBs forslag til landbrukstelling 1999. 65s.

97/51 J.E. Sivertsen: Flyktninger og arbeidsmarkedet 4. kvartal 1996. 38s.

97/52 J. Nordøy: Nytten av forventningsbaserte konjunkturindekser ved predikering av konsum. 36s.

97/53 S. Hansen og T. Skoglund: Sammenligning av data for sysselsetting og lønn fra ulike kilder. 30s.

97/54 S. Blom: Holdning til innvandrere og innvandringspolitikk: Spørsmål i SSBs omnibus i mai/juni 1997. 39s.

97/55 K. Mork: SSB-AVLØP: Fylkeshefte 1996. 203s.

97/56 Opplysninger om inntekt, formue og skatt i forløpsdatabasen Trygd-fobhistorie: Tilrådinger fra et utvalg. 52s.

97/57 E.J. Fløttum: Ordliste og definisjoner i økonomisk statistikk: Engelsk - bokmål - nynorsk. 166s.

97/58 T. Dale: Samordnet levekårsundersøkelse 1997 - panelundersøkelsen: Dokumentasjonsrapport. 87s.

97/59 H. Høie og A. Grønlund: Driftstypemodellen: Modell for tilrettelegging av jordbruksstatistikk for beregning av tap av næringsstoffer fra jordbruksarealene: Dokumentasjon. 37s.

97/60 A. Sundvoll: Undersøkelse om mødre med nyfødte barn. 36s.

97/61 S. Todsen: Nasjonalregnskap: Beregning av realkapitalbeholdninger og kapitalslit. 34s.

97/62 K. Mork: Utslepp og rensing av avløpsvatn: Datakvalitet og berekningsmåter. 64s.

97/63 S. Stamnes og B.L. Western: Inntekts- og kostnadsundersøkelse for privatpraktiserende psykologer 1996: Dokumentasjon. 26s.

97/64 H.M. Teigum: Barns helse og velferd 1996: Dokumentasjon og frafallsanalyse. 39s.

97/65 F. Gjertsen: Dødsårsaksregistret i Statistisk sentralbyrå: Rapport om virksomheten i 1996. 56s.

97/66 B. Olsen: Prøveundersøkelse om 1-3 dagers sykefravær i sentral sykefraværsstatistikk: Dokumentasjon. 15s.

97/67 S. Nygårdseter: Prisindeks for engroshandel. 22s.

97/68 R. Johansen: REGARD - Modell for regional analyse av arbeidsmarked og demografi: Teknisk dokumentasjon. 212s.

97/69 A.A. Ritland: Inntekts- og formuesundersøkelsen 1996: Dokumentasjonsrapport. 21s.

97/70 B. Bye: Imperfeksjoner i arbeidsmarkedet: Konsekvenser for velferdseffekter av en grønn skattereform. 18s.

97/72 E.J. Fløttum: Grupperinger av næringer i offisiell statistikk - revidert utgave. 41s.

97/73 L. Solheim og D.Q. Pham: Prekorrigering av påskeeffekten for detaljvolumindeksen 1979-1997. 58s.

97/74 D. Roll-Hansen: Lesernes mening om avisen Forskning. 45s.

97/75 M.V. Dysterud og E. Engelien: Tettstedsavgrensing og arealbruksstatistikk for tettsteder 1997: Dokumentasjon av metode og programmering. 61s.

98/1 L.C. Zhang: Dokumentasjonsrapport: Den nye estimeringsmetoden for Arbeidskraftundersøkelsen (AKU): Fylkesvis kalibrering med landsetterstratifiserte vekter som startverdier. 18s.

98/4 H.M. Teigum: Omnibusundersøkelsene 1997: Dokumentasjonsrapport. 138s.

98/5 Metodevalg og kostnader ved etablering og drift av et boligregister. Revidert forslag: Rapport fra en arbeidsgruppe som har revidert og oppdatert planene for opprettelse av et boligregister. 31s.

Notater

PORTO BETALT
VED
INNLEVERINGA
P.P.
NORGE/NOREG

Tillatelse nr.
159 000/502

B *Returadresse:*
Statistisk sentralbyrå
Postboks 8131 Dep.
N-0033 Oslo

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

ISSN 0806-3745

Statistisk sentralbyrå
Statistics Norway