Statistics Norway
Statistisk sentralbyrå

Jan Henrik Wang

Frafall i konjunkturbarometeret

Notater

Avdeling for økonomisk statistikk/Seksjon for økonomiske indikatorer
Emnegruppe: 08.05.10

Innhold

1. Innledning

Dette notatet er skrevet i forbindelse med kurset Frafall og imputering (SM05). Notatet gir en empirisk analyse av ulike former for frafallsjustering basert på de metoder som ble presentert i kurset. Analysen er knyttet opp mot Konjunkturbarometeret (KBAR) for industri og bergverksdrift.

Konjunkturbarometeret er en kvalitativ undersøkelse som kartlegger bedriftsledernes vurderinger av utviklingen for kjennetegn som produksjon, kapasitetsutnytting, sysselsetting, ordretilgang etter marked, priser, generell bedømmelse av utsiktene m.m. Det norske konjunkturbarometeret ble utviklet i 1973 og satt i drift f.o.m. 1974. Gjennom det siste tiåret har det europeiske arbeidet på dette området blitt harmonisert og administreres i dag gjennom Directorate General Economic and Financial Affairs (DG ECFIN).

I kapittel 2 vil vi beskrive undersøkelsen og definere interesse variabel og analyseperiode, videre vil vi i kapittel 3 se på frafall i Konjunkturbarometeret, herunder ulike former for justering av frafall, før vi avslutter med å oppsummere resultatene i kapittel 4.

2. Om undersøkelsen

2.1 populasjon, enheter og utvalg

Enheten i undersøkelsen er definert lik *bransjeenheten*, dvs. alle bedrifter i et foretak som tilhører en og samme næringshovedgruppe, dvs. alle enheter i samme 3-sifret næring (SN94) - videre omtalt som bransje. I datafangstsammenheng blir *observasjonsenheten* satt til største bedrift i bransjeenheten, men det arbeides også med andre observasjonsenheter, f.eks. foretakets hovedkontor. Slike tilpasninger skjer som regel i samsvar med foretakenes egne ønsker, men kan også forekomme av andre årsaker. I analyse og for beregningsformål er enheten satt lik bransjeenheten.

Populasjon omfatter alle bransjeenheter i næringene bergverksdrift (SN94 10, 13-14) og industri (15 - 37). I etableringen av trekkrammen holdes enheter der største bedrift har færre enn 10 sysselsatte utenfor. Populasjonen avgrenses ved alle bransjeenheter som er omfattet i SSBs bedrifts- og foretaks-register. Bedrifts- og foretaksregisteret definert ved situasjonsfil i 2. kvartal hvert år utgjør også rammen for ajourhold av utvalg.

Den nye utvalgsplanen - tatt i bruk i 1. kvartal 1996 - ble utformet med formål å få et mest mulig heldekkende bilde av konjunktursituasjonen og -utsiktene i den enkelte bransje[1]. Bransjeenhetens sysselsetting brukes som et størrelsesmål ved stratifiseringen i utvalgsarbeidet, der hver bransjepopulasjon deles i fire strata.

Stratum 1	Enheter med flere enn 300 sysselsatte
Stratum 2	Enheter med 200 - 299 sysselsatte
Stratum 3	Enheter med 100 - 199 sysselsatte
Stratum 4	Enheter med mindre enn 100 sysselsatte

Det foretas full telling for enheter som har flere enn 300 sysselsatt (stratum 1). I øvrige strata trekkes enheter proporsjonalt med størrelsen (proporsjonal allokering). Trekkingen gjennomføres for hvert strata i hver bransje.

[1] Utvalgsplanen ble justert i 2. kvartal 1997. Det ble foretatt justeringer i stratuminndelingen, deler av opprinnelig utvalg ble rullert ut og erstattet samt foretatt en supplering. Størrelsen på bruttoutvalget ble justert opp til vel 700 enheter blant annet for å ta høyde for et forholdsvis stort frafall ved førstegangsutsendelse.

I analysedelen i dette notatet har vi forenklet noe ved å anta at trekksannsynligheten er lik i hvert stratum og at den er avhengig av dekningsgraden av sysselsatte trukket i hvert stratum. Dette for å simulere det faktum at det er en overrepresentasjon av større enheter i hvert stratum.

Bruttoutvalget dekker vel 54 prosent av populasjonssysselsettingen og noe i underkant av 62 prosent av samlet omsetning. Dekningsgraden varierer imidlertid fra bransje til bransje. På 2-sifret næringsnivå ligger dekningsprosenten fra 30 - 90. I enkelte bransjer kan imidlertid dekningsprosenten være både større og mindre enn dette.

2.2 Beregningsopplegg og vekting av svar[2]

2.2.1 På stratumnivå

Resultatene på stratumnivå beregnes ved å tildele hver aktiv enhets svar en vekt lik dens sysselsetting. Mer presist kan beregningen av svarandelen i prosent, $SY_{n,i,j,B}$, for spørsmål *n*, svaralternativ *i*, i et stratum *j* og bransje *B* formuleres i følgende tre steg:

Antall sysselsatte som er kodet til svaralternativ i er:

$$(1) \qquad Y_{n,i,j,B} = \Sigma_b (\alpha_{b,j} * \beta_{b,i} * S_{b,j,B})$$

der

- $\alpha_{b,j}$ angir om en enhet er med i utvalget i stratum j, og om den er aktiv, dvs. har besvart oppgaven, i det aktuelle kvartalet. $\alpha_{b,j}$ kan anta verdiene 0 / 1. En aktiv enhet får en verdi lik 1 - ellers 0. Enheter som ikke er i utvalget får i beregningene på stratumnivå verdi lik 0.
- $\beta_{b,i}$ kan ha verdiene 0 / 1 avhengig av hvilket svaralternativ den enkelte oppgavegiver i stratumet har valgt på det aktuelle spørsmålet. En oppgavegiver som har valgt f.eks. «større» får en faktor lik 1 når svarandelen for dette alternativet beregnes - ellers settes verdien lik 0.
- $S_{b,j,B}$ uttrykker sysselsetting for den enkelte bransjeenhet, *b*, i stratumpopulasjon *j*, i bransje *B*.

Sum sysselsatte for alle aktive bransjeenheter i stratum j er:

$$(2) \qquad SS_{n,i,j,B} = \Sigma_i \Sigma_b (\alpha_{b,j} * \beta_{b,i} * S_{b,j,B})$$

Svarandelen i prosent for alternativ i, stratum j blir da :

$$(3) \qquad SY_{n,i,j,B} = Y_{n,i,j,B} * 100 / SS_{n,i,j,B}$$

Av (1) - (3) framgår at grunnlaget for beregningen av svarandelen for et tillatt svaralternativ for spørsmål n er alle bransjeenheter som er næringskodet til populasjonen i en bransje. Ved bruk av α-faktoren tas enheter, som ikke inngår i utvalget eller som ikke er aktive (frafall) i et kvartal, ut av beregningene. Med β-faktoren grupperes de svaralternativer som aktive enheter har valgt, og gis en vekt lik bransjeenhetens sysselsetting.

[2] Forklaring av beregningsopplegget er hentet fra Andersen og Wang (2003)

Det følger av (1) - (3) at sum svarandeler i prosent for et spørsmål er lik 100, dvs. :

(4) $\Sigma_i \, SY_{n,i,j,B} = 100$

2.2.2 På aggregerte nivåer

Beregningen av svarfordelingen på bransjenivå tar utgangspunkt i svarfordelingene på stratumnivå. I overgangen fra stratum til bransje veies imidlertid stratumresultatene med populasjonssysselsettingen for å korrigere for relative forskjeller mellom strataene i en bransje. Mer presist kan beregningen av svarandelen i prosent, $SY_{n,i,B}$, for spørsmål *n*, svaralternativ *i*, i bransje *B* formuleres ved følgende sammenhenger :

(5) $SY_{n,i,B} = (\Sigma_j \, Y_{n,i,j,B} * a_{j,B}) * 100 / SS_B$

der SS_B er sum sysselsatte for alle enheter i den enkelte stratumpopulasjon i bransje *B*.

og

(6) $a_{j,B} = 1 / (SS_{n,i,j,B} / SS_{j,B})$

Formel (6) uttrykker den inverse av sum trekksannsynlighet for aktive enheter i stratum *j*, bransje *B*.

Svarandel i prosent for alternativ *i* på bransjenivå framkommer ved å summere produktet av antall sysselsatte allokert til hvert svaralternativ i stratum *j* med den inverse sum av trekksannsynlighet for aktive enheter i stratumet.

De samme prinsippene brukes også i videre aggregering.

Som det fremgår av denne gjennomgangen av beregningsopplegg og vekting av svar, beregnes det først svarandeler for nettoutvalget i hvert stratum før man beregner populasjonsandelen ved å vekte med den inverse av sum trekksannsynlighet for enheter i nettoutvalget i stratum *j*, bransje *B.* For å kunne gjennomføre analysen av frafall må vi bruke den inverse trekksannsynligheten som utvalgsvekt for hver enhet og deretter aggregere. Beregningsopplegget som er benytte i dette notatet avviker derfor fra det som benyttes i den løpende produksjon.

2.3 Interessevariabel

Skjema for Konjunkturbarometeret inneholder 28 spørsmål om ulike kjennetegn for observasjonsenhetene. For å forenkle analysen har vi konsentrert oss om et av spørsmålene; *Generell bedømmelse av utsiktene for det kommende kvartal*[3]. For dette spørsmålet er det tre svaralternativer :

- Bedre
- Uendret
- Dårligere

Videre har vi definert svaralternativet som 1 hvis enheten har besvart spørsmålet med 'Bedre' og 0 hvis det er valgt et annet alternativ. I og med at svarene vektes med enhetens sysselsetting vil interessevariabelen benyttet i frafallsanalysen bli svaralternativet multiplisert med bransjeenhetens sysselsetting.

[3] Dette er spørsmål 18 på skjema og den fullstendige spørsmålsformuleringen er : Hvordan bedømmer De - generelt for foretakets virksomhet i denne bransjen - utsiktene for kommende kvartal i forhold til situasjonen i inneværende kvartal.

I den faktiske produksjonen av statistikken beregnes det en andel for de tre svaralternativene, samt en andel som ikke har besvart spørsmålet i nettoutvalget (Partielt frafall[4]). Ut fra disse resultatene beregnes nettotall og diffusjonsindekser for de ulike spørsmål og bransjer.

Nettotall = Andel bedre - andel dårligere

Diffusjonsindeks = Andel bedre + 0,5*andel uendret

2.4 Analyseperiode

Vi vil benytte data fra undersøkelsen gjennomført for 2. kvartal 2003. Tabellen nedenfor viser antall enheter i populasjonen og utvalg i de ulike sysselsettingsstrata. Totalt sett var det 24438 enheter i populasjonen og et bruttoutvalg på 701 bransjeenheter.

Tabell 1 : Populasjon og utvalg

Sysselsettings-stratum	Populasjon	Bruttoutvalg
Større eller lik 300	159	146[5]
299 - 200	75	38
199 - 100	275	143
99 - 1	23929	374
Sum	24438	701

[4] For mer om dette se kapittel 3. Frafall i Konjunkturbarometeret.

[5] Som vi ser av oversikten er ikke alle enheter i stratumet Større eller lik 300 med i utvalget selv om trekksannsynligheten er 1 (jf. kapittel 2.1). Dette kommer av at enkelte enheter har gitt beskjed om at de ikke ønsker å delta i undersøkelsen og at de av den grunn er fjernet fra utvalget.

3. Frafall i Konjunkturbarometeret

Konjunkturbarometeret er en frivillig undersøkelse (ikke underlagt statistikkloven) og man opplever derfor et noe større frafall en ved andre pliktige konjunkturundersøkelser. Ser vi på andre undersøkelser rettet mot samme populasjon (industri og bergverk), som f.eks. Kvartalsvis investeringsstatistikk eller Ordre- og lagerstatistikk, som er pliktige, har vi en svarandel opp mot 98 prosent.

Enhetsfrafall for Konjunkturbarometeret – dvs. enheter som er trukket, men som ikke har sendt inn skjema – er ganske stabilt og ligger på om lag 15 prosent. Det gir en gjennomsnittlig svarandel i de siste kvartaler på 85 prosent.

Partielt frafall – dvs. manglende verdi på enkelte av spørsmålene – varierer mellom de ulike spørsmålene. En oversikt viser at det partielle frafallet i 2. kvartal 2003 varierer fra 9,7 til 0,1 prosent. Grunnen til at dette varierer så mye mellom de ulike spørsmål er at enkelte spørsmål passer dårlig for enkelte bransjer, og de gir av den grunn ikke svar på disse spørsmål. De spørsmål det fokuseres på ved publisering har imidlertid et lavt partielt frafall.

Det kan være ulike kilder og årsaker til frafall i konjunkturbarometeret. Som tidligere nevnt er undersøkelsen frivillig og av den grunn er det enkelte som gir tilbakemelding om at de ikke ønsker å delta i undersøkelsen. Utvalget baserer seg på et panel der det årlig suppleres for avgang grunnet konkurser og nedleggelser. I tillegg rulleres enheter som ikke har svart de to siste kvartaler ut av undersøkelsen. Undersøkelsen er postal. Følgende årsaker kan identifiseres som grunner til frafall:

Enhetsfrafall :

- Registerfeil. Det kan være enheter som er trukket i utvalget fra trekkpopulasjonen, men som reelt sett ikke har produksjon eller er nedlagt. I de tilfeller hvor vi får tilbakemelding, kan vi fjerne enheten fra populasjon og utvalg, men i mange tilfeller blir vi ikke informert av respondenten og fanger derfor ikke opp slike feil.
- Ønsker ikke å delta. De fleste respondenter i grunnlagspopulasjonen industri og bergverk har en rekke pliktige undersøkelser de må besvare. Det er derfor en del som unnlater å svare da denne undersøkelsen er frivillig og da det oppfattes som en for stor oppgavebyrde.
- Når ikke frem til kontaktperson. I enkelte tilfeller er kontaktpersonen til enheten sluttet eller ikke til stede, slik at skjema ikke når frem til riktig person.
- Skjema har ikke blitt trykket for alle enheter
- Skjema blir ikke registrert under datafangstarbeidet
- Feil i beregningsmetoder. Registrerte skjema blir ikke inkludert i beregningen av aggregatene

Av erfaring og kontroller, som gjennomføres i de ulike delene av statistikkproduksjonen, er det uvilje til å fylle ut skjema som virker som den største frafallsårsak. Vi gjennomfører krysskontroller mellom enheter i utvalget til Konjunkturbarometeret mot utvalgene til andre statistikker, med samme populasjon, for å kontrollere om skjema blir sendt til riktig sted og person, og i de fleste tilfeller mottar vi skjema for de pliktige undersøkelsene, mens vi mangler svar for Konjunkturbarometeret selv om respondenten er den samme.

Partielt frafall:

- Irrelevant spørsmål. Det er det samme skjema som sendes til alle respondenter uavhengig av hvilken næring de tilhører. Dette fører til at det for enkelte oppgavegivere føles vanskelig å besvare alle spørsmål. Man har forsøkt å rette på dette ved å innføre et eget svaralternativ som er *'Ikke relevant'* , men ikke for alle spørsmål, da det antas at enkelte vil føle seg fristet til å bruke dette alternativet for ofte.
- Feil kontaktperson. Spørsmålene i Konjunkturbarometeret forutsetter at respondenten har inngående kjennskap til en rekke økonomiske forhold knyttet til driften. I enkelte tilfeller

kjenner respondenten kun en del av de kjennemerker vi etterspør og vi kan av den grunn oppleve partielt frafall.
- Skjønner ikke spørsmålet. Respondenten forstår ikke enkelte spørsmål, og lar av den grunn være å svare på enkelt spørsmål.
- Feil registrering. Skjema leses i de fleste tilfeller optisk. I disse tilfeller er det sjelden feil. Imidlertid blir skjema som ikke kan verifiseres (kopier, faks m.m) registrert manuelt. I denne prosessen kan det bli feilregistrering eller at den som registrerer hopper over et svar.

For spørsmålet vi har valgt å konsentrere oss om; Generell bedømmelse av utsiktene, fordeler svarene og frafallet seg som i tabell 2.

Tabell 2 : Svarfordeling og frafall i de ulike sysselsettingsstrata

Sysselsettings-stratum	Bedre	Uendret	Dårligere	Netto utvalg	Partielt frafall	Enhetsfrafall	Bruttoutvalg
Større elller lik 300	27	79	19	125	1	20	146
299 - 200	7	18	9	34	1	3	38
199 - 100	25	69	32	126	1	16	143
99 - 0	84	166	74	324	1	49	374
Sum	143	332	134	609	4	88	701

Av tabellen ser vi at det partielle frafallet fordeler seg med en enhet i hvert strata og utgjør til sammen et frafall på 0,6 prosent i forhold til bruttoutvalget. I den videre analysen vil vi betrakte det partielle frafallet sammen med enhetsfrafall slik at det totale frafallet blir på 92 enheter. Dette gir en svarandel på 86,9 prosent. Ser vi nærmere på frafallet innenfor hvert sysselsettingsstratum får vi følgende svarandeler :

Tabell 3 : Svarandeler

Sysselsettings-stratum	Svarandel
Større elller lik 300	85,6
299 - 200	89,5
199 - 100	88,1
99 - 0	86,6
Totalt	86,9

3.1 Justering for frafall

I beregningsopplegget som benyttes for Konjunkturbarometeret er det antatt helt tilfeldig frafall. Frafall imputeres implisitt ved at man betrakter nettoutvalget som bruttoutvalg når man beregner populasjonsandel. På den måten vil frafallsenheter ikke inngå i utvalget når raten som benyttes for estimering av populasjonsandel beregnes.

I den videre analysen skal vi se nærmere på om denne antakelsen holder, eller om det er grunnlag for å vurdere en mer kompleks modellering av frafallet. Som vi ser av tabell 3 er det ikke noe som tyder på stor skjevhet i frafallet mellom de ulike sysselsettingsstrata. Dette vil vi undersøke nærmere når vi prøver ut ulike frafallsmodeller. Vi skal benytte ulike imputeringsmetoder og frafallsmodeller, beskrevet i Zhang (2003) og gjennomgått i kurset Frafall og imputering (SM05), for å analysere effekten av frafall.

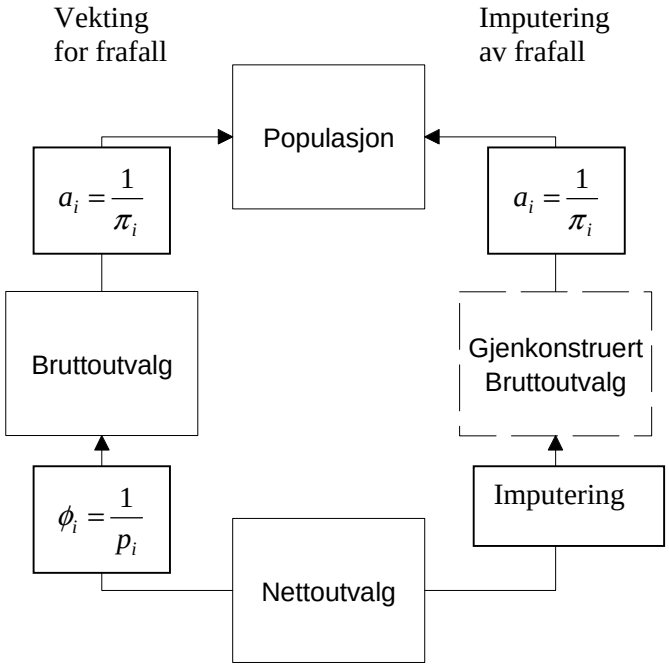
I første del av analysen skal vi benytte ulike frafallsmodeller (vekting) for å justere for frafall. I den andre delen skal vi teste ut ulike metoder for imputering av frafall. I beregningene benyttes en rekke SAS-makroer skrevet for kurset av Anna-Karin Mevik.

Følgende notasjon vil bli benyttet:

- U = {1,...,N} => Populasjon & i = indeksen til enheter
- s = (brutto-)utvalg & s_r = nettoutvalg (svarutvalg) & s_m = enhetsfrafall
- r_i er responsvariabel slik at r_i =1 hvis $i \in s_r$ & r_i =0 hvis $i \in s_m$
- π_i er trekksannsynlighet & $a_i = 1/\pi_i$ => utvalgsvekt
- p_i er svarsannsynlighet & $\phi_i = 1/p_i$ => frafallsvekt
- $w_i = a_i\phi_i = (\pi_i p_i)^{-1}$ => designvekt for $i \in s_r$
- y_i er interessevariabel & $Y = \sum_{i \in U} y_i$ => total av y_i i populasjonen

Figuren nedenfor illustrerer forskjellen mellom vekting og imputering

Fig 1 Vekting og imputering



Som vi ser av figuren ligger forskjellen mellom vekting og imputering i at ved vekting gjennomfører vi en oppblåsing (tilsvarende som fra bruttoutvalg til populasjon) fra nettoutvalg til bruttoutvalg før man beregner populasjonsnivå, mens ved imputering legger vi inn estimerte/antatte verdier for alle svar som mangler før man beregner populasjonsnivå. Ved vekting vil produktet av utvalgsvekt og frafallsvekt gi designvekten :

$$w_i = a_i\phi_i = (\pi_i p_i)^{-1}$$

I analysen som følger vil vi se på den andelen som mener at de generelle utsiktene har blitt *bedre*. Vi vil også forenkle ved å kun beregne resultater for industri og bergverksdrift samlet sett, ikke fordelt på de ulike bransjer.

For å foreta en strukturert implementering av modellene og imputeringsmetodene har vi tilpasset data fra Konjunkturbarometeret på en måte som gjør at vi kan benytte de SAS-makroer som er laget til kurset. Programmet som omstrukturerer data er gjengitt i vedlegg 1. Programmet kan enkelt tilpasses for å kunne analysere andre spørsmål eller perioder for Konjunkturbarometeret.

Interessevariabelen er i analysen definert som

(1) $y_i = \beta_i * S_i$

Der $\beta_i = \begin{cases} 1 & \text{Hvis enhet } i \text{ har valgt } \textit{'bedre'} \\ 0 & \text{Hvis enhet } i \text{ har valgt et annet svaralternativ} \end{cases}$

S_i er enhetens sysselsetting.

Det vi ønsker å estimere er da andelen sysselsatte som mener de generelle utsiktene er bedre for det kommende kvartal, $\bar{Y}$, gitt ved formel (2)

(2) $\bar{Y} = (\sum_{i \in U} y_i) / \sum_{i \in U} S_i$

Fra populasjonsfilen har vi at sum sysselsetting $S = \sum_{i \in U} S_i$ = 292940

3.1.1 Vekting for enhetsfrafall

3.1.1.1 Direkte vekting

I dette avsnittet vil vi anta at frafallet er helt tilfeldig og benytte fremgangsmåten for direkte vekting. Vi betrakter her frafall som en tilleggsfase i sannsynlighetsbasert utvalgstrekking. Den inverse svarsannsynlighet benyttes som frafallsvekt. Designvekten er da produktet av utvalgs- og frafallsvekt.

Et estimat for $\bar{Y}$, som andelen av de sysselsettingsveide svarene for de som har svart *'bedre'* på spørsmålet om de generelle utsiktene, kan da skrives som

(3) $\hat{\bar{Y}} = (\sum_{i \in s_r} w_i y_i) / (\sum_{i \in U} S_i)$

For å finne w_i må vi beregne svarsannsynlighetene, p_i, og frafallsvektene, ϕ_i, slik at vi kan beregne designvekten gitt ved (4):

(4) $w_i = a_i \phi_i = (\pi_i p_i)^{-1}$ der $\phi_i = (p_i)^{-1} = \left(\frac{n}{n+m} \right)^{-1}$

og n er antall enheter i s_r, og m antall enheter i s_m.

Ved å benytte direkte vekting med helt tilfeldig frafall vil ϕ_i være en konstant, dvs. at svarsannsynligheten er den samme uansett hvilken enhet det dreier seg om. Med disse antakelsene får vi følgende estimat

(3) $$\hat{\bar{Y}} = (\sum_{i \in s_r} w_i y_i) / (\sum_{i \in U} S_i) = \frac{68372,7}{292940} = 0,233$$

Det gir altså at innenfor industri og bergverk vurderer 23,3 prosent de generelle utsiktene for det kommende kvartal som bedre. Programmet som er benyttet i beregningen er gjengitt i vedlegg 2.

3.1.1.2 Estimering under en ikke-informativ SHG-modell

Vi skal nå ta utgangspunkt i en ikke-informativ SHG[6]-modell. Med denne modellen forsøker man å dele utvalget inn i grupper som man antar har ulike frafallsmekanismer. Denne modellen vil kunne justere for skjevheter som kommer av at frafallet er vesentlig større innenfor enkelte grupper av utvalget. Disse gruppene kan defineres som enheter innenfor samme sysselsettingsstratum eller innenfor samme næring eller andre konstellasjoner der man kan anta at frafallet er avhengig sammensettingen av gruppene. Målet med å dele inn i slike svarhomogene grupper er å gjøre svarsannsynligheten p_i mest mulig lik innen hver gruppe, samtidig som den er mest mulig ulik mellom gruppene. Generelt kan modellen fremstilles på følgende måte:

- Vi antar at utvalget er delt inn i *G* SHG'er, betegnet med s_g for *g = 1,...,G*. La s_{rg} inneholde svarenheter i s_g, og la s_{mg} inneholde frafallsenheter i s_g slik at $s_g = s_{rg} \cup s_{mg}$
- Vi lar n_g være antall enheter i s_{rg}, og m_g antall enheter i s_{mg}. Vi kan da beregne svarsannsynligheten, p_i, for $i \in s_g$ som

(5) $$p_i = n_g / (n_g + m_g)$$

Ved å benytte (5) i (4) får vi beregnet designvekten til hver enhet avhengig av hvilken SHG enheten er klassifisert under. Videre aggregering blir som i (3).

Under denne modellen har vi valgt å se på to mulige inndelinger av de svarhomogene gruppene. i *a)* har vi valgt å gruppere enheter innenfor samme sysselsettingsstrata og i *b)* har vi valgt å dele inn i grupper avhengig av hvilken næring enheten tilhører.

Antakelsen under *a)* impliserer at det er større frafall blant de minste enhetene i utvalget i forhold til de største. Av tabell 3 ovenfor ser det imidlertid ikke ut til at det er noen større forskjell mellom frafallet blant store og små enheter, der størrelsen er målt i enhetens sysselsetting.

Antakelsen under *b)* impliserer at frafallet kan være større innen enkelte næringer og at det på den måten er en systematisk skjevhet i frafallet.

a) Gruppering etter sysselsettingsstrata

Ved å sette en SHG-indeks lik variabelen for sysselsettingsstrata i programmet benyttet for helt tilfeldig frafall får vi beregnet et estimat basert på en ikke-informativ SHG-modell. Modellen vil i dette tilfellet ha 4 svarhomogene grupper som tilsvarer sysselsettingsstrata gjengitt i tabell 4 (*g=1,2,3,4).*

[6] SHG = Svarhomogene grupper

Tabell 4 : SHG = Sysselsettingsstrata

SHG	Sysselsettingsstrata
1	Større eller lik 300
2	299 - 200
3	199 - 100
4	99 - 1

Under denne modellen får vi følgende estimat

$$(3) \qquad \hat{\bar{Y}} = (\sum_{i \in s_r} w_i y_i)/(\sum_{i \in U} S_i) = \frac{68646,2}{292940} = 0,234$$

Programmet som er benyttet i beregningen er gjengitt i vedlegg 3. Vi ser at estimatet blir tilnærmet helt likt som ved antakelsen om helt tilfeldig frafall. Dette er ingen overraskelse da svarandelen i de ulike sysselsettingsstrata var omtrent like (jf. tabell 3).

b) Gruppering etter næring
I dette tilfellet velger vi å gruppere de svarhomogene gruppene etter hvilken næring enheten tilhører. For å unngå for mange grupper vil vi gruppere enhetene etter publiseringsnivå. Tabell 5 viser sammenhengen mellom NACE 2-nivå og stratum. I tabellen har vi også tatt med svarandelen i hvert stratum.

Tabell 5 : SHG = Næringsgruppe

SHG	Næringsgrupper[7]	Svarandel
1	10, 13 -14	91,7
2	15 - 16	85,5
3	17 - 19	76,7
4	20	89,3
5	21	95,5
6	22	89,2
7	23 - 24	87,1
8	25	81,0
9	26	81,8
10	27	95,5
11	28	93,1
12	29	77,6
13	30 - 33	88,5
14	34 - 35	83,6
15	36 - 37	92,5

Av tabell 5 ser vi at svarandelen varierer mellom de ulike SHG'er og at spesielt *g=3* (NACE 17 -19 ; Tekstil og bekledning) og *g=12* (NACE 29; Produksjon av maskiner og utstyr) har lavere svarandel enn de andre SHG'er.

Under denne modellen får vi samme estimat som hvis vi benytter SHG = Sysselsettingsstratum :

$$(3) \qquad \hat{\bar{Y}} = (\sum_{i \in s_r} w_i y_i)/(\sum_{i \in U} S_i) = \frac{68643,0}{292940} = 0,234$$

[7] Tallene i kolonnen samsvarer med 2-siffret NACE

Programmet som er benyttet i beregningen er det samme som benyttet i *a)* (gjengitt i vedlegg 3) bortsett fra at vi har byttet ut *g=x* med *g=x2* (*g* angir SHG-indeks, x er sysselsettingsstratum og x2 er indekseringen av næringsgruppene).

Av disse beregningene ser det ikke ut til at det er noen klar korrelasjon mellom de SHG'er vi har forutsatt og frafallet. I hvert fall ikke på en slik måte at det påvirker estimatet.

Som vi så av tabell 5 er det enkelte næringer som har lavere svarandel enn andre, men andelen av sysselsatte varierer kraftig mellom de ulike næringsgruppene. Hvis vi ser på næringsgruppen Tekstil og bekledning; *g=3*, så vil denne gruppen få en større frafallsvekt enn de andre SHG'ene. Dette har imidlertid liten betydning for det totale sysselsettingsveide estimatet da gruppen har en svært liten andel av sysselsettingen for industrien totalt sett.

3.1.1.3 Estimering under en enkel informativ SHG-modell

Modellen vi nå skal se på forutsetter at frafallet er korrelert med interessevariabelen. Dvs. at det antas at frafallet er større eller mindre blant de som velger et svaralternativ fremfor et annet.

- vi definerer SHG'er s_g for $g = 1,...,G$ som blant annet avhenger av interessevariabelen
- videre lages *tilleggsklasser* s_h for $h=1,...,H$ basert på variabler som er kjente i hele utvalget
- anta videre at svarsannsynligheten til i er uavhengig av at $i \in s_h$ gitt at $i \in s_g$

Vi antar at frafallet er homogent blant de som svarer hhv. *bedre, uendret* eller *dårligere* og at svaralternativene definerer s_g for $g=1,..,3$. Da vi ikke kjenner gruppetilhørigheten til frafallet, s_{mg}, er vi nødt til å etablere tilleggsklasser for å estimere denne tilhørigheten. Vi antar derfor at vi har tilleggsklassene s_h for $h=1,...,4$, definert ved de fire sysselsettingsstrata som vi kjenner for hele utvalget $s = s_r + s_m$. Til slutt antar vi at frafallet er uavhengig av sysselsettingsstrata gitt svaralternativet.

For å estimere svarsannsynlighetene må vi også estimere gruppetilhørigheten til frafallet.

- Vi lar s_{rgh} betegne delutvalget $s_{rg} \cap s_{rh}$, dvs. svarenheter som tilhører både s_g og s_h. Videre lar vi s_{mgh} betegne delutvalget $s_{mg} \cap s_{mh}$, dvs. frafall som tilhører både s_g og s_h.
- Vi betegner størrelsen til s_{rgh}, som er kjent i utvalget, med n_{gh}. Videre lar vi m_{gh} være størrelsen til s_{mgh}, som er ukjent bortsett fra at $m_h = \sum_{g=1}^{G} m_{gh}$ siden s_h er kjent.
- Gitt estimat for m_{gh}, betegnet $\hat{m}_{gh}$, kan vi estimere svarsannsynligheten med

$$(6) \qquad \hat{p}_i = \frac{n_g}{n_g + \hat{m}_g} = \frac{\sum_{h=1}^{H} n_{gh}}{\sum_{h=1}^{H} n_{gh} + \sum_{h=1}^{H} \hat{m}_{gh}} \qquad \text{for } i \in s_g$$

For å estimere m_{gh} benytter vi en iterativ algoritme :

1. Velg initiale verdier for m_{gh}, betegnet med $m_{gh}^{(0)}$, som f.eks.

$$m_{gh}^{(0)} = \frac{m_h n_{gh}}{n_h} = \frac{m_h n_{gh}}{\sum_{g=1}^{G} n_{gh}}$$

2. For k=1,2,...., beregn

$$w_{gh}^{(k)} = \frac{(n_{gh} + m_{gh}^{(k-1)})(\sum_h m_{gh}^{(k-1)})}{\sum_h n_{bh} + \sum_h m_{gh}^{(k-1)}} \qquad \text{og} \qquad m_{gh}^{(k)} = \frac{m_h w_{gh}^{(k)}}{\sum_{g=1}^{G} w_{gh}^{(k)}}$$

3. Vi stopper algoritmen etter 40 iterasjoner (k=40), og bruker $\hat{m}_{gh} = m_{gh}^{(k)}$ som estimat for m_{gh}.

For å forsøke å estimere svarsannsynlighetene under den informative SHG-modellen har vi benyttet SAS-makroen 'frafall'. Med de SHG'er og tilleggsklassene som er beskrevet over, får vi ingen konvergens. Vi har forsøkt å definere *h* og *g* på ulike måter for å forsøke å få algoritmen til å konvergere uten hell. For å fullføre analysen har vi valgt å benytte SAS-makroen 'svarsh' som gir svarsannsynligheter uansett om algoritmen konvergerer eller ikke. Resultatene av denne analysen må derfor tolkes med tanke på at de beregnede svarsannsynligheter kan være feil. Med 'svarsh' får vi følgende svarsannsynligheter, $\hat{p}_i$, for de 3 svaralternativene :

Tabell 6 : Estimerte svarsannsynligheter i prosent

	Bedre	**Uendret**	**Dårligere**
$\hat{p}_i$	85,3	86,0	91,0

Av tabellen med de estimerte svarsannsynlighetene for de tre informative SHG'ene, ser vi at estimatene for svarsannsynlighetene er tilnærmet like for de som svarer *bedre* eller *uendret*, mens den er noe høyere for de som svarer *dårligere*.

Ved å benytte de estimerte svarsannsynlighetene i formel (6) får vi beregnet designvekten til hver enhet, som vil avhenge av hvilket svaralternativ enheten har valgt.

(7) $$\hat{w}_i = a_i \hat{\phi}_i = (\pi_i \hat{p}_i)^{-1}$$

Ved å benytte de estimerte designvektene fra (7) i formel (3), får vi et estimat på andelen som vurderer de generelle utsiktene som bedre

(8) $$\hat{\bar{Y}} = (\sum_{i \in s_r} \hat{w}_i y_i)/(\sum_{i \in U} S_i) = \frac{69675{,}5}{292940} = 0{,}238$$

Som vi ser gir denne modellen et marginalt høyere estimat på andelen som mener utsiktene er bedre. Dette følger av de estimerte svarsannsynlighetene. Den estimerte frafalls vekten, $\hat{\phi}_i = 1/\hat{p}_i$, vil bli større for enheter som svarer *bedre* enn de som svarer dårligere. På den måten blåses andelen *bedre* opp i forhold til de andre alternativene – da det antas at frafallet er større blant denne gruppen. Det er viktig å ta med i betraktningen at vi ikke kan si at det estimerte svarsannsynlighetene er korrekte, da algoritmen, basert på mine antakelser, ikke konvergerte.

Det er vanskelig å gi noen god tolkning på hvorfor frafallet skulle være mindre for enheter som svarer *dårligere* i forhold til enheter som svarer *bedre.* En mulig årsak kan være at det ved nedgangskonjunktur – i bransjen enheten virker i – er et større behov for å klage (via offentlig statistikk) enn når man er i en oppgangskonjunktur. For å underbygge denne hypotesen kunne man gjort en analyse på enhetsfrafallet over tid, for å undersøke om svarprosenten er korrelert med konjunkturbildet. Det vil imidlertid føre for langt å gå nærmere inn på dette i denne analysen. Programmet som er benyttet i estimeringen av den informative SHG-modellen er gjengitt i vedlegg 4.

3.1.1.4 Kalibrering av direkte vekting ved rateestimering

De modellene vi nå har testet ut benytter kun informasjon i utvalget. Ved å benytte tilleggsinformasjon fra populasjonen kan man forbedre det direkte veide estimat. Kurset Frafall og imputering gjennomgår tre ulike typer for kalibrering; etterstratifisering, rateestimering og regresjonsestimering. I dette notatet har vi valgt å se nærmere på kalibrering ved rateestimering.

Vi lar sysselsetting, S_i, være en tilleggsvariabel. Med tanke på variansreduksjon og justering for frafall, er det ønskelig at å bruke en tilleggsvariabel som er høyt korrelert med interessevariabelen. Vi har ikke noe belegg for å kunne anta en slik korrelasjon mellom svaralternativer og antall sysselsatte, men av mangel på andre registervariable som kunne tenkes brukt benytter vi sysselsetting. Rateestimatoren er da gitt som

$$(9) \qquad w_{i,rat} = w_i(S/\hat{S}) = \frac{(Sw_i)}{\sum_{i\in s_r} w_i S_i} \qquad \text{der} \qquad S = \sum_{i\in U} S_i \qquad \text{og} \qquad \hat{S} = \sum_{i\in s_r} w_i S_i$$

Totalt antall sysselsatte, *S*, i populasjonen er kjent: $S = \sum_{i\in U} S_i = 292940$

og vi kjenner S_i for enheter som har svart på undersøkelsen.

Vi skal benytte denne kalibreringsmetoden på de tre modellene vi har sett på under vekting for enhetsfrafall:

a) Direkte vekting (helt tilfeldig frafall)
b) Ikke-informativ SHG-modell
c) Informativ SHG-modell

SAS-makroen 'rate' er benyttet i programmene som estimerer de kalibrerte resultatene.

a) Direkte vekting (helt tilfeldig frafall)

Ved å bruke den kalibrerte designvekten, $w_{i,rat}$, fra (9) i formel (3) får vi følgende kalibrerte rateestimat på andelen som svarer *'bedre'*

$$(10) \qquad \hat{\bar{Y}}_{rat} = (\sum_{i\in s_r} w_{i,rat} y_i)/(\sum_{i\in U} S_i) = (\sum_{i\in s_r} w_i(S/\hat{S}) y_i)/(\sum_{i\in U} S_i) = \frac{72840{,}1}{292940} = 0{,}249$$

I estimeringen finner vi at raten $S/\hat{S}$ er 1,065. På den måten justeres det sysselsettingsveide estimatet noe opp i forhold til direkte vekting på grunn av underdekning av sysselsetting i utvalget grunnet frafall. Uten kalibrering vil alle frafallsenheter ha samme vekt $\phi = p_i^{-1} = (n/(n+m))^{-1}$. Da vi ønsker at enheter med større sysselsetting skal telle mer enn enheter med liten sysselsetting vil vi med rateestimeringen kompensere for dette. Programmet som benyttet er gjengitt i vedlegg 5.

b) Ikke-informativ SHG-modell
På samme måte som under antakelsen om helt tilfeldig frafall beregner vi den kalibrerte designvekten, $w_{i,rat}$, i formel (9), men nå beregnes det en rate pr. SHG. I dette eksempelet vil vi gjøre beregningene både for a) og b)

a) SHG'er lik sysselsettingsstratum

b) SHG'er lik næring

Ved å benytte formel (9) og (3) får vi følgende kalibrerte rateestimat på andelen som svarer *'bedre'*

(10a) $$\hat{\bar{Y}}_{rat} = (\sum_{i \in s_r} w_{i,rat} y_i)/(\sum_{i \in U} S_i) = (\sum_{i \in s_r} w_i (S/\hat{S}) y_i)/(\sum_{i \in U} S_i) = \frac{72920{,}2}{292940} = 0{,}249$$

(10b) $$\hat{\bar{Y}}_{rat} = (\sum_{i \in s_r} w_{i,rat} y_i)/(\sum_{i \in U} S_i) = (\sum_{i \in s_r} w_i (S/\hat{S}) y_i)/(\sum_{i \in U} S_i) = \frac{73145{,}6}{292940} = 0{,}250$$

Også under den ikke-informative SHG-modellen blir designvektene kalibrert med raten $S/\hat{S}$ = 1,065. På den måten justeres de sysselsettingsveide estimatene også her noe opp, i forhold til den ikke-informative SHG-modellen uten kalibrering. I vedlegg 6 er programmet som ble benyttet for (10a) gjengitt. Programmet for (10b) er tilsvarende, bare med en annen SHG-indeks.

c) Informativ SHG-modell
Tilsvarende som under den ikke-informative SHG-modellen skal vi beregne den kalibrerte designvekten, men her benytter vi den estimerte svarsannsynligheten der frafallet er korrelert med interessevariabelen. Den kalibrerte estimerte designvekten betegnes som $\hat{w}_i = a_i \hat{\phi}_i = (\pi_i \hat{p}_i)^{-1}$. Ved å benytte dette estimatet i formel (9) og (3) får vi følgende uttrykk for det kalibrerte estimatet under en informativ SHG-modell.

(11) $$\hat{\bar{Y}}_{rat} = (\sum_{i \in s_r} \hat{w}_{i,rat} y_i)/(\sum_{i \in U} S_i) = (\sum_{i \in s_r} \hat{w}_i (S/\hat{S}) y_i)/(\sum_{i \in U} S_i) = \frac{74166{,}3}{292940} = 0{,}253$$

Tilsvarende som under direkte vekting med helt tilfeldig frafall og ikke-informativ SHG-modell kalibreres det sysselsettingsveide estimatet med raten $S/\hat{S}$ = 1,065, noe som fører til en oppjustering i forhold til estimatet uten kalibrering. Programmet som er benyttet i estimeringen er gjengitt i vedlegg 7.

3.1.2 Imputering for partielt frafall
I dette kapittelet skal vi se på metoder for imputering av frafall. I motsetning til vekting vil vi her forsøke å fylle inn svarverdier for frafallsenhetene, og på den måten lage et fullstendiggjort datasett for bruttoutvalget (jf. Fig 1).

To typer imputering:

- Deterministisk : Samme verdier imputeres ved gjentakelse av imputeringsprosessen
- Stokastisk : Ulike verdier kan imputeres ved gjentakelse, og man vil på den måten kunne få ulikt resultat hver gang imputeringsprosessen gjennomføres

3.1.2.1 Imputering fra 'nærmeste nabo'

Dette er en deterministisk metode som estimerer svaralternativer basert på en metrikk funksjon som benytter tilleggsvariabler for å måle 'avstanden' mellom en frafallsenhet og en giver[8]. Som tilleggsvariabel har vi benyttet sysselsetting, S_i . Vi får da at avstanden mellom en frafallsenhet og en giver blir :

(12) $\delta_{ij} = |S_i - S_j|$

På den måten imputeres svaralternativet som gir minst mulig δ_{ij} mellom en frafallsenhet og en giver. Dvs. giveren som har antall sysselsatte nærmest antall sysselsatte til frafallsenheten. Vi antar derfor med denne modellen at det er en sammenheng mellom hvilket svaralternativ som velges og hvor mange sysselsatte enheten har.

Fra (1) har vi interessevariabelen $y_i = \beta_i * S_i$

Der $\beta_i = \begin{cases} 1 & \text{Hvis enhet } i \text{ har valgt } \textit{'bedre'} \\ 0 & \text{Hvis enhet } i \text{ har valgt et annet svaralternativ} \end{cases}$

og S_i er enhetens sysselsetting.

Med imputerte verdier $\beta_i^* = \beta_j$ der $\delta_{ij} = |S_i - S_j|$ er minimert får vi følgende sammenheng

$$\tilde{\beta}_i = \begin{cases} \beta_i & i \in s_r \\ \beta_i^* & i \in s_m \end{cases}$$

Fra denne sammenhengen får vi

(13) $\tilde{y}_i = \tilde{\beta}_i * S_i$

Med de imputerte verdier har vi nå en verdi for alle enheter i bruttoutvalget. I estimatet blir derfor designvekten lik utvalgsvekten og frafallsvekten, p_i blir 1

(14) $w_i = a_i \phi_i = (\pi_i p_i)^{-1} = (\pi)^{-1} = a_i$

Får å estimere den sysselsettingsveide andelen, $\hat{\bar{Y}}_{imp}$, benytter vi (13), (14) og (3) og får

(15) $\hat{\bar{Y}}_{imp} = (\sum_{i \in s} a_i \tilde{y}_i)/(\sum_{i \in U} S_i) = \frac{67574{,}1}{292940} = 0{,}231$

I estimeringen av β_i^* for $i \in s_m$ har vi benyttet makroen 'nabo'. Av (15) ser vi at det frafallsjusterte estimatet basert på imputering ved hjelp av 'nærmeste nabo' gir et noe lavere estimat enn de vi fikk under modeller for vekting for enhetsfrafall. Andelen av de imputerte verdiene som ble gitt verdien $\beta_i^* = 1$ var 0,239, men på grunn av at svaralternativene vektes med sysselsettingen blir altså det

[8] Enhet som imputeringsverdi hentes fra.

sysselsettingsveide estimatet lavere. Dette tyder på at det var en overrepresentasjon av større enheter (enheter med mange sysselsatte) som fikk imputert $\beta_i^* = 0$. Tabell 7 kan illustrere dette.

Tabell 7 Fordeling av imputerte verdier

Imputert verdi β_i^*	**Antall**	**Sum sysselsatte**
1	22	2984
0	70	19400
Sum	92	22384
Andel	0,239	0,133

Av tabellen ser vi at andelen av de som fikk imputert verdien 1 er 0,239, mens hvis vi ser på andelen av de sysselsettingsveide imputerte svaralternativene som fikk verdien 1 er denne kun 0,133. Programmet som er benyttet er gjengitt i vedlegg 8.

3.1.2.2 Stokastisk imputering under ikke-informativ SHG-modell (hot-deck)

I motsetning til imputering med 'nærmeste nabo' er denne imputeringsformen stokastisk. Dvs. at ved gjentatte simuleringer av imputering vil vi få ulike resultater. Hot-deck imputering går ut på å forsøke å gruppere sammen enheter som på en eller annen måte ligner på hverandre.

For å gruppere enhetene har vi valgt SHG = Næringsgruppe (definert i tabell 5). Grunnen til at vi velger denne grupperingen er en antakelse om at enheter som tilhører samme næringsgruppe har større sannsynlighet for å ha samme konjunkturutvikling enn enheter som tilhører ulike næringsgrupper. På den måten vil vi anta at giver trekkes fra samme næringsgruppe => Frafall i tekstil industrien dekkes ved imputering fra en giver fra tekstil industrien.

Imputeringsmetoden går ut på å imputere verdien β_i^* fra en tilfeldig trukket giver innen samme SHG. SAS-makroen 'hotdeck' er benyttet for å imputere verdiene. På samme måte som under 'nærmeste nabo' får vi sammenhengen

$$\tilde{\beta}_i = \begin{cases} \beta_i & i \in s_r \\ \beta_i^* & i \in s_m \end{cases}$$

Får å estimere den sysselsettingsveide andelen, $\hat{\bar{Y}}_{imp}$, benytter vi (13), (14) og (3) og får igjen

$$(15) \qquad \hat{\bar{Y}}_{imp} = (\sum_{i \in s} a_i \tilde{y}_i)/(\sum_{i \in U} S_i)$$

Da imputeringsmetoden er stokastisk vil estimatet variere ved gjentakelser av imputeringen. Som estimat for det stokastiske estimat har vi valgt å kjøre 20 simuleringer for deretter å ta forventningen gitt ved gjennomsnittet av de stokastiske estimatene.

$$(16) \qquad \mathrm{E}(\hat{\bar{Y}}_{imp}) \approx \frac{\sum_{i \in N} \hat{\bar{Y}}_{imp,i}}{N} = 0{,}233 \qquad N=(1,...,20)$$

Som vi ser av (16) gir gjennomsnittet av de 20 simuleringene det samme sysselsettingsveide estimatet som under direkte vekting med helt tilfeldig frafall, men usikkerheten i estimatet har økt pga. den

stokastiske prosessen. Resultatene fra de 20 simuleringene er gjengitt i tabell 8. Som vi ser av tabellen varierer estimatet justert for frafall med hot-deck imputering fra 0,220 til 0,257. Grunnen til dette er at i og med at vi trekker tilfeldig innenfor hver SHG, vil vi ikke få de samme givere ved hver simulering.

Tabell 8 Resultater fra Hot-deck imputering

Nr	$\hat{\bar{Y}}_{imp}$
1	0,233
2	0,233
3	0,235
4	0,231
5	0,232
6	0,250
7	0,226
8	0,241
9	0,239
10	0,237
11	0,231
12	0,222
13	0,257
14	0,231
15	0,232
16	0,229
17	0,234
18	0,220
19	0,222
20	0,230
Gj. sn	0,233
St. dv	0,009

Programmet som er benyttet i estimeringen er gjengitt i vedlegg 9.

3.1.2.3 Kalibrering av estimat under imputeringsmodeller ved rateestimering

Som vi så nærmere på under kalibrering av direkte vekting ved rateestimering, kan vi også under imputeringsmodellene gjennomføre en kalibrering basert på tilleggsinformasjon fra populasjonen. I dette tilfellet vil det ikke være designvektene, w_i, som kalibreres, men utvalgsvektene, a_i.

Fra (14) har vi at $w_i = a_i$ i tilfellet med imputering.

Ved å bruke (9) kan vi definere den kalibrerte utvalgsvekten som

(17) $$a_{i,rat} = a_i(S/\tilde{S}) = \frac{(Sa_i)}{\sum_{i\in s} a_i S_i} \qquad \text{der} \qquad S = \sum_{i\in U} S_i \qquad \text{og} \qquad \tilde{S} = \sum_{i\in s} a_i S_i$$

Totalt antall sysselsatte, *S*, i populasjonen er kjent: $S = \sum_{i\in U} S_i = 292940$

og vi kjenner S_i for alle enheter i bruttoutvalget.

Fra (17) og (15) kan vi da definere det kalibrerte sysselsettingsveide estimatet basert på imputering som

(18) $$\hat{\tilde{Y}}_{imp,rat} = (\sum_{i \in s} a_{i,rat} \tilde{y}_i) / (\sum_{i \in U} S_i) = (\sum_{i \in s} a_i (S / \tilde{S}) \tilde{y}_i) / (\sum_{i \in U} S_i)$$

Vi skal benytte denne kalibreringsmetoden på de to imputeringsmetodene vi beskrev ovenfor:

a) Nærmeste nabo
b) Hot-deck

a) Nærmeste nabo
Ved kalibrering av det sysselsettingsveide estimatet under denne imputeringsmetoden får følgende resultat

(18) $$\hat{\tilde{Y}}_{imp,rat} = (\sum_{i \in s} a_i (S / \tilde{S}) \tilde{y}_i) / (\sum_{i \in U} S_i) = 0{,}241$$

I estimeringen finner vi at raten $S / \tilde{S}$ er 1,045. Vi får også her en kalibrering av det sysselsettingsveide estimatet oppover, men med en mindre faktor enn under vekting ($S / \hat{S}$ =1,065). Dette gir sammenhengen

(19) $$\tilde{S} = \sum_{i \in s} a_i S_i > \sum_{i \in s_r} w_i S_i = \hat{S}$$

Dvs. at summen av produktene av utvalgsvekten og sysselsettingen for alle enheter i bruttoutvalget er større enn summen av produktene av designvekten og sysselsettingen for alle enheter i nettoutvalget. Programmet som er benyttet i beregningen er gjengitt i vedlegg 10.

b) Hot-deck
I stokastisk imputering under ikke-informativ SHG-modell (hot-deck), kan vi også gjennomføre en kalibrering av det sysselsettingsveide estimatet basert på rateestimatoren. Vi benytter som i eksempelet med hot-deck uten kalibrering SHG = Næringsgrupper (definert i tabell 5).

Ved å bruke (18) og (16) kan vi beregne et gjennomsnitt av de kalibrerte stokastiske estimatene ved å kjøre 20 simuleringer, for deretter å ta gjennomsnittet av de stokastiske estimatene

(18) $$\hat{\tilde{Y}}_{imp,rat} = (\sum_{i \in s} a_i (S / \tilde{S}) \tilde{y}_i) / (\sum_{i \in U} S_i)$$

(20) $$\mathrm{E}(\hat{\tilde{Y}}_{imp,rat}) \approx \frac{\sum_{i \in N} \hat{\tilde{Y}}_{imp,rat,i}}{N} = 0{,}244 \qquad N=(1,...,20)$$

Som vi ser av (20) gir gjennomsnittet av de 20 simuleringene noe høyere estimat enn ved hot-deck imputering uten kalibrering ved rateestimator. Dette fordi raten, $S / \tilde{S}$, blir 1,045. Denne vil være konstant (ikke stokastisk) i og med at raten ikke avhenger av svaralternativene, $\tilde{\beta}_i$, og raten vil være lik som i tilfellet med imputering ved 'nærmeste nabo'.

Grunnen til at raten blir den samme som i tilfellet med imputering ved 'nærmeste nabo' er at bruttoutvalget, utvalgsvektene og sysselsettingen er de samme i de to tilfellene, og uavhengig av $\tilde{\beta}_i$

(jf. formel (17)). Forskjellen i estimatene vil kun ligge i hvilke verdier som imputeres for frafallsenhetene.

Resultatene fra de 20 simuleringene er gjengitt i tabell 9. Som vi ser av tabellen varierer estimatet , justert for frafall med hot-deck imputering kalibrert ved rateestimering, fra 0,226 til 0,258. På samme måte som under hot-deck uten kalibrering vil vi få en stokastisk prosess i og med at vi trekker tilfeldig innenfor hver SHG ved hver simulering.

Tabell 9 Resultater fra rate-kalibrert hot-deck imputering

Nr	$\hat{\bar{Y}}_{imp,rat}$
1	0,226
2	0,246
3	0,250
4	0,239
5	0,258
6	0,233
7	0,241
8	0,240
9	0,249
10	0,247
11	0,243
12	0,230
13	0,239
14	0,258
15	0,236
16	0,246
17	0,256
18	0,247
19	0,255
20	0,240
Gj. sn	0,244
St. dv	0,009

Programmet som er benyttet i estimeringen er gjengitt i vedlegg 11.

3.1.3 Effekten av kalibrering

Får å se nærmere på effekten vi oppnår med kalibrering ved hjelp av rateestimatoren, kan vi analysere variansen til det sysselsettingsveide estimatet med og uten kalibrering. For at kalibrering med rateestimatoren skal ha noen variansreduserende effekt er vi avhengig av at tilleggsvariabelen, sysselsatte, er korrelert med interessevariabelen. Da interessevariabelen y_i er produktet av sysselsettingen til enheten og svaret på spørsmålet (1 eller 0) er ikke dette en urimelig forutsetning. For å se om rateestimatoren vi har benyttet i kalibreringen har noen variansreduserende effekt vil vi måle effekten av tilleggsinformasjon betinget på justering for frafall, dvs. frafallsvektene. Fra Zhang (2003) har vi at man kan beskrive et variansestimat for den direkte veide estimator $\hat{\bar{Y}}$, der vi antar konstant varians, som

$$(21) \qquad v_1 = \frac{1}{n}(1 + c_w^2)s_y^2$$

der c_w^2 er varianskoeffisienten til w_i over s_r , og var(y_i) , betegnet s_y^2 , kan skrives som

(22) $$s_y^2 = \frac{1}{n-1}\sum_{i\in s_r}(y_i - \bar{y})^2$$

der y_i er det sysselsettingsveide svaret definert som i formel (1), og gjennomsnittet $\bar{y}$ er

(23) $$\bar{y} = \frac{1}{n}\sum_{i\in s_r} y_i$$

Dette vil gjelde uansett om frafallsmodellen er informativ eller ikke.

Et enkelt variansestimat for den kalibrerte estimator, under tilsvarende forutsetninger, har følgende generelle form

(24) $$v_2 = \frac{1}{n}(1 + c_w^{*2})s_e^2$$

der c_w^{*2} er varianskoeffisient til de kalibrerte vekter, og s_e^2 er variansen til kalibreringsresidualene. Definisjonen av kalibreringsresidualene under rateestimering er gitt som

(25) $$e_i = y_i - x_i\beta = y_i - x_i\frac{\hat{Y}}{\hat{X}} = y_i - x_i\frac{\sum_{i\in s_r} w_i y_i}{\sum_{i\in s_r} w_i x_i}$$

I formel (25) er tilleggsvariabelen sysselsetting som benyttes i raten betegnet x_i. Variansen til kalibreringsresidualene kan da beregnes med følgende formel

(26) $$s_e^2 = \frac{1}{n-1}\sum_{i\in s_r}\left(y_i - x_i\frac{\sum_{i\in s_r} w_i y_i}{\sum_{i\in s_r} w_i x_i}\right)^2$$

Med disse sammenhengene kan vi måle effekten av tilleggsinformasjon vha. raten

(27) $$\eta = \frac{v_2}{v_1} = \frac{1 + c_w^{*2}}{1 + c_w^2}.\frac{s_e^2}{s_y^2}$$

Under forutsetning om helt tilfeldig frafall, og at varianskoeffisienten til de kalibrerte vektene er tilnærmet lik varians koeffisienten til designvektene uten kalibrering, $c_w^* \approx c_w$, får vi redusert raten til $\eta \approx s_e^2 / s_y^2$. Denne raten har vi beregnet for modellen med direkte vekting og helt tilfeldig frafall, med og uten kalibrering. Variansestimeringen er gjengitt til slutt i programmene i vedlegg 2 og 5. Dette gir oss følgende resultat

(28) $$\eta \approx \frac{s_e^2}{s_y^2} = \frac{33320}{44982} = 0{,}74$$

Med andre ord vil vi redusere variansen i det sysselsettingsveide estimatet, ved hjelp av kalibrering med rateestimator, med 26 prosent.

4. Sammendrag

I dette notatet har vi sett nærmere på frafallet i Konjunkturbarometeret, og da spesielt for spørsmål 18; Generell bedømmelse av utsiktene for det kommende kvartal. I kapittel 3 har vi prøvd å gi en generell beskrivelse av mulige former for enhetsfrafall og partielt frafall. Det er også beregnet aggregerte svarsannsynligheter for de 4 sysselsettingsstrata (se tabell 3).

I kapittel 3.1, Justering for frafall, har vi justert det sysselsettingsveide estimatet, for andelen som mener de generelle utsiktene er bedre, ved hjelp av ulike metoder for vekting for frafall og to ulike imputeringsmetoder. I tillegg har vi kalibrert disse estimatene ved hjelp av rateestimering. I kapittel 3.1.3 har vi sett nærmere på effekten av kalibrering vha. rateestimatoren, og da spesielt om den har noen variansreduserende effekt.

Estimatet med direkte vekting under antakelsen om helt tilfeldig frafall, kalibrert ved rateestimering, gir om lag det samme estimatet som vi får med det beregningsopplegget som benyttes i den løpende produksjonen. Der antar vi helt tilfeldig frafall og i rateestimeringen inngår kun nettoutvalget. En forskjell er at vi i den løpende produksjonen beregner en egen andel for det partielle frafallet (betegnet som andel *'uoppgitt'*), samt at vi kalibrerer med raten for hvert sysselsettingsstrata i hver bransje.

Ser vi på estimatene fra den ikke-informative SHG-modellen, er resultatene tilnærmet like de vi får ved antakelsen om helt tilfeldig frafall. Dette gjelder både med og uten kalibrering ved rateestimering. Dette tyder på at definisjonen av de svarhomogene gruppene (sysselsettingsstratum og næringsgruppering) ikke gir grupper der det er ulikt frafallsmønster mellom gruppene, og dermed heller ingen justering av estimatet. Her kan man tenke seg at man kan benytte andre former for inndeling av SHG'ene, slik at svarsannsynligheten blir ulik mellom gruppene, og så lik som mulig innad i gruppen. Vi har imidlertid ikke klart å finne en slik inndeling.

Resultatene fra den informative SHG-modellen ligger noe høyere enn de andre estimatene, noe som skulle tyde på at frafallet er større blant de enheter som forventer en bedre utvikling for de kommende kvartal. Da algoritmen benyttet i estimeringen ikke konvergerer vil disse resultatene preges av dette, og det blir vanskelig å trekke noen konklusjon.

I avsnitt 3.1.2, Imputering for partielt frafall, benyttet vi to ulike former for imputering; en deterministisk (Nærmeste nabo) og en stokastisk metode (Hot-deck). Av resultatet med imputering med 'nærmeste nabo' ser vi at dette estimatet blir noe lavere enn under frafallsmodellene med vekting. Dette tyder på at det er var en overrepresentasjon av større enheter som fikk imputert verdien 0 (svaralternativ *'uendret'* eller *'mindre'*). På den måten ble det sysselsettingsveide estimatet for andelen som vurderte de generelle utsikte som *'bedre'* noe lavere. Når det gjelder den stokastiske imputeringen under ikke-informativ SHG-modell (Hot-deck), ser vi at det gjennomsnittlige estimatet basert på 20 simuleringer er det samme som under antakelsen om helt tilfeldig frafall. Da SHG lik næringsgruppe viste seg å ha liten effekt under frafallsmodellen, vil denne imputeringsmetoden gi et resultat tilsvarende imputering ved tilfeldig trekking innenfor hele nettoutvalget (ingen SHG-indeks). I og med at disse resultatene er like vil frafallsmodellen med helt tilfeldig frafall være å foretrekke, da denne vil ha lavere varians enn den stokastiske imputeringen.

Vi ser videre at med kalibrering ved rateestimering, får vi et lavere estimat med imputeringsmetodene enn ved frafallsmodelleringen. Tabell 10 oppsummerer de ulike estimatene vi har beregnet.

Tabell 10 Resultater fra justering av frafall ved frafallsmodeller og imputeringsmetoder

		Frafallsmodell				Imputeringsmetode	
		Helt tilfeldig frafall	Ikke informativ SHG		Informativ SHG	Nærmeste nabo	Hot-deck (SHG=Næring)
			Syss. stratum	Næringsgruppe			
Kalibrering ved rateestimering	Nei	0,233	0,234	0,234	0,238	0,231	0,233
	Ja	0,249	0,249	0,250	0,253	0,241	0,244

Fra de beregningene som er gjennomført er det vanskelig å trekke noen slutning om at det er en skjevhet i fordelingen av frafallet i Konjunkturbarometeret, men vi kan ikke utelukke at det finnes frafallsmekanismer som vi ikke har funnet og som gir systematisk skjevhet.

Kalibrering av estimatene med rateestimator gir gjennomgående et høyere estimat på andelen som mener de generelle utsiktene er bedre. Beregning av effekten av kalibrering vha. rateestimering viser at dette gir estimater med lavere varians. Dette styrker troen på at den estimatoren som benyttes i den løpende produksjonen gir mer effisiente estimater enn uten kalibrering, og at det er en fornuftig å kalibrere estimatene på denne måten.

Vi har gjort en rekke forenklinger av problemstillingen som f.eks. at vi kun ser på et svaralternativ og et spørsmål. Mange av spørsmålene i Konjunkturbarometeret er korrelert med hverandre, og det vil ha innvirkning på metoder man eventuelt skulle valgt for å justere for frafall. Fordelen med dagens beregningsopplegg med forutsetningen om tilfeldig frafall er at det gir en oversiktlig og enkel sammenstilling av data for samtlige spørsmål sett under ett.

I et eventuelt videre arbeid med analyse av frafallet i Konjunkturbarometeret vil det være interessant å se nærmere på ulike former for imputering. En imputeringsmetode vil enkelt kunne tilpasses dagens beregningsopplegg, mens en frafallsmodell basert på vekting vil bety en total omlegging av statistikken beregningsrutiner.

Referanser

Andersen og Wang (2003) : Konjunkturbarometeret. Statistisk sentralbyrå, Rapporter 2003/10, Tom Langer Andersen og Jan Henrik Wang

Zhang (2003) : SM05 - Innføring i justering for frafall, upublisert kursnotat August 2003, Li-Chun Zhang,

Vedlegg 1. Tilpassning av data

```
*************************************************************************************************;
* PROGRAM FOR Å TILPASSE DATA FOR FRAFALLSANALYSE                                               *;
*************************************************************************************************;

* BEREGNER TREKKSANNSYNLIGHETER PÅ BAKGRUNN AV NACE3 X SYSS STR:;
* Tar utgangspunkt i populasjonsfil og utvalgsfil fra beregningsopplegget til Konjunkturbarometeret;

proc sort data=kurs.utvalg03k02; by b_enhet;
run;

proc sort data=kurs.pop03k02; by b_enhet;
run;

* Slår sammen populasjon og utvalgsfil og døper om enkelte variable;

Data filpop_utv (drop=orgnr foretaksnr syssel: ar kvartal stratum_utv  );
merge kurs.pop03k02 (in=pop) kurs.utvalg03k02 (in=utv);
by b_enhet;
if N > 1 then psyssel=sysselutv;
rename s18=y1;
run;

* Lager responsvariable og sletter enheter som ikke har sysselsetting i pop;
* Sletter også næring 11 : Olje og gasutvinning;

data data_s18 (keep=nace3 b_enhet utvalg psyssel stratum y1 r frafall);
set filpop_utv;
if utvalg=1 and y1 > 0  then r=1;
else if utvalg=1 and y1 =< 0 then r=0;
if y1=0 and r=0 then frafall='P';
else if y1=. and r=0 then frafall='E';
label frafall='P=partielt fraf. E=enhetsfraf.'
      r='Responsvariabel';
if psyssel=0 then delete;
nace2=substr(nace3,1,2);
if nace2='11' then delete;
run;

* Beregner trekksannsynlighetene. Ulike for alle stratum X nace3;

proc sort data=data_s18; by utvalg stratum nace3;

proc means data=data_s18 noprint;
class utvalg stratum nace3;
output out=summer sum(psyssel)=sumsyss;
run;

data fil1;
set summer;
if _TYPE_=3;
rename _freq_=N_pop
      sumsyss=sumsyss_pop;
drop utvalg  _type_;
run;

data fil2;
set summer;
if utvalg=1 and _type_=7;
rename _freq_=N_utv
      sumsyss=sumsyss_utv;
drop utvalg _type_;
run;

proc sort data=fil1; by stratum nace3;
```

```
proc sort data=fil2; by stratum nace3;

data fil3;
merge fil1 fil2;
by stratum nace3;
run;

proc sort data=fil3; by stratum nace3;
proc sort data=data_s18; by stratum nace3;
run;

* Legger til variabelen trekksh til alle enheter i utvalget;

data data_s18;
merge data_s18 fil3;
by stratum nace3;
trekksh=sumsyss_utv/sumsyss_pop;
run;

* Henter ut data fra utvalget som skal benyttes i analysen;

data utvalg_s18;
set data_s18;
if utvalg=1;
if y1=1 then y2=1;
else y2=0;
y=y2;
if frafall ne ' ' then y= . ;
nace2=substr(nace3,1,2);
run;

proc sort data=utvalg_s18; by b_enhet;

data kurs.oppgave (keep= ident trekksh x x2 x3 z y r nace2);
set utvalg_s18;
by b_enhet;
retain ident 0;
ident+1;
Label stratum='Syss. Stratum'
        psyssel='Sysselsetting'
        y='Svar: 1=bedre'
        y1='Svar;
rename psyssel=z
          stratum=x
          y1=x3;
if nace2 in ('10','13','14') then x2=1;
else if nace2 in ('15','16') then x2=2;
else if nace2 in ('17','18','19') then x2=3;
else if nace2 ='20' then x2=4;
else if nace2 ='21' then x2=5;
else if nace2 ='22' then x2=6;
else if nace2 in ('23','24') then x2=7;
else if nace2 ='25' then x2=8;
else if nace2 ='26' then x2=9;
else if nace2 ='27' then x2=10;
else if nace2 ='28' then x2=11;
else if nace2 ='29' then x2=12;
else if nace2 in ('30','31','32','33') then x2=13;
else if nace2 in ('34','35') then x2=14;
else if nace2 in ('36','37') then x2=15;
label x2='Næringsgruppe';
run;
```

Vedlegg 2. Helt tilfeldig frafall

```
*************************************************************************************;
* KURSOPPG #1                                                                       *;
*************************************************************************************;

*Direkte veid estimat : Helt tilfeldig frafall;

*Setter SHG indeks til 1 for alle enheter og beregner interessevariabel t=syss.*svar;

Data kbar;
set kurs.oppgave;
if r=1 then t=z*y;
else t= .;
g=1;
run;

* Estimerer svarsannsynlighetene og lager en variabel p med de estimerte verdiene;

proc means data = kbar noprint nway;
class g;
var r;
output out= aggr_kbar (drop= _type_ rename= (_freq_=brutto_g))
        sum=n_g;
run;

proc sort data=kbar; by g;
proc sort data=aggr_kbar; by g;

data kbar;
merge kbar aggr_kbar;
by g;
p=n_g/brutto_g;
label p = 'p=est.svarsh.';
run;

* Beregner designvekten;

data kbar;
set kbar;
w=1/(trekksh*p);
label w='W=designvekt';
run;

* lager en variabel wt som er produktet av designvekten og interessevariabelen;

data kbar;
set kbar;
wt=w*t;
run;

* Estimerer andelen ved først å estimere Y (konjunkitursysselsettingen);
* og så dele på totalt antall sysselsatte i populasjonen;

proc means data=kbar noprint nway;
var wt;
output out= estimat (drop=_freq_ _type_)
        sum=sum_wt
run;

data estimat;
set estimat;
Y_andel=sum_wt/292940;
run;

* Variansestimering av y;
```

```
proc means data=kbar noprint nway;
var t;
output out=var1 sum=sum_t;
run;

data var1 (drop=_freq_ _type_);
set var1;
g=1;
run;

data kbar1;
merge kbar var1;
by g;
t_gj=(1/n_g)*sum_t;
e=(t-t_gj)**2;
if t=. then delete;
run;

proc means data=kbar1 noprint nway;
var e;
output out=var2 sum=sum_e;
run;

data var2 (drop=_freq_ _type_);
set var2;
g=1;
run;

data kbar1;
merge kbar1 var2;
by g;
var_y=(1/(n_g-1))*sum_e;
run;
```

Vedlegg 3. Ikke-informativ SHG-modell

```
*************************************************************************************;
* KURSOPPG #2                                                                       *;
*************************************************************************************;

* Ikke-informativ SHG-frafallsmodell;

*Setter SHG indeks til sysselsettingsstratum =>g=x;

Data kbar;
set kurs.oppgave;
if r=1 then t=z*y;
else t= .;
g=x;
run;

* Estimerer svarsannsynlighetene og lager en variabel p med de estimerte verdiene;
* I dette tilfellet får vi 4 ulike svarsannsynligheter avhengig av hvilket sysselsettingsstratum enheten er i;

proc means data = kbar noprint nway;
class g;
var r;
output out= aggr_kbar (drop= _type_ rename= (_freq_=brutto_g))
        sum=n_g;
run;

proc sort data=kbar; by g;
proc sort data=aggr_kbar; by g;

data kbar;
merge kbar aggr_kbar;
by g;
p=n_g/brutto_g;
label p = 'p=est.svarsh.';
run;

* Beregner designvekten;

data kbar;
set kbar;
w=1/(trekksh*p);
label w='W=designvekt';
run;

* lager en variabel wt som er produktet av w og t;

data kbar;
set kbar;
wt=w*t;
run;

* Estimerer andelen ved først å estimere Y (konjunkitursysselsettingen);
* og så dele på totalt antall sysselsatte i populasjonen;

proc means data=kbar noprint nway;
var wt;
output out= estimat (drop=_freq_ _type_)
        sum=sum_wt;
run;

data estimat;
set estimat;
T=sum_wt/292940;
run;
```

Vedlegg 4. Informativ SHG-modell

```
*************************************************************************************;
* KURSOPPG #3                                                                       *;
*************************************************************************************;

*Informativ SHG-modell;

*Setter SHG indeks lik svaralternativ;
*Tilleggsklasse er definert som sysselsettingsstratum;

Data kbar;
set kurs.oppgave;
if r=1 then t=z*y;
else t= .;
format h 4.
       g 4.;
h=x;
g=x3;
if g < 1 then g= .;
run;

* Estimerer svarsannsynlighetene og lager en variabel p med de estimerte verdiene;

%include '$KONJBAR/sas/prog/kurs/makro/svarsh.sas'/source2;
%let ant_it=40;
%svarsh(kbar);

proc sort data=kbar; by g;
proc sort data=svarsh; by g;

data kbar;
merge kbar svarsh;
by g;
run;

* Beregner designvekten;

data kbar;
set kbar;
w=1/(trekksh*p);
label w='W=designvekt';
run;

* lager en variabel wt som er produktet av w og t;

data kbar;
set kbar;
wt=w*t;
run;

* Estimerer andelen ved først å estimere Y og så dele på total;

proc sort data=kbar; by x;

proc means data=kbar noprint nway;
var wt;
output out= estimat (drop=_freq_ _type_)
       sum=sum_wt;
run;

data estimat;
set estimat;
T=sum_wt/292940;
run;
```

Vedlegg 5. Rate-kalibrert direkte vekting

```
*************************************************************************************;
* KURSOPPG #4                                                                       *;
*************************************************************************************;

*Rate-kalibrert direkte veid estimat : Helt tilfeldig frafall;

*Setter SHG indeks til 1 for alle enheter og beregner interessevariabel t=syss.*svar;

Data kbar (drop=x);
set kurs.oppgave;
if r=1 then t=z*y;
else t= .;
g=1;
run;

data kbar;
set kbar;
x=z;
run;

* Estimerer svarsannsynlighetene og lager en variabel p med de estimerte verdiene;

proc means data = kbar noprint nway;
class g;
var r;
output out= aggr_kbar (drop= _type_ rename= (_freq_=brutto_g))
        sum=n_g;
run;

proc sort data=kbar; by g;
proc sort data=aggr_kbar; by g;

data kbar;
merge kbar aggr_kbar;
by g;
p=n_g/brutto_g;
label p = 'p=est.svarsh.';
run;

* Beregner designvekten;

data kbar;
set kbar;
w=1/(trekksh*p);
label w='W=designvekt';
run;

* BRUKER MAKRO FOR Å BEREGNE RATE;

%include '$KONJBAR/sas/prog/kurs/makro/rate.sas' / source2;

%let X=292940;
%let imputering=nei;

%rate(kbar);


* lager en variabel kwt som er produktet av den rate-kalibrerte designvekten og interessevariabelen;

data kbar;
set kbar;
kwt=kw*t;
run;
```

```
* Estimerer andelen ved først å estimere Y (konjunktursysselsettingen);
* og så dele på totalt antall sysselsatte i populasjonen;

proc means data=kbar noprint nway;
var kwt;
output out= estimat (drop=_freq_ _type_)
        sum=sum_kwt
run;

data estimat;
set estimat;
Y_andel=sum_kwt/292940;
run;


* Variansestimering av residualene;

proc means data=kbar noprint nway;
output out=var1 sum(wt)=sum_wt sum(wx)=sum_wx;
run;

data var1 (drop=_freq_ _type_);
set var1;
g=1;
q=sum_wt/sum_wx;
run;

data kbar1;
merge kbar var1;
by g;
f=q*x;
e=(t-f)**2;
if t=. then delete;
run;

proc means data=kbar1 noprint nway;
var e;
output out=var2 sum=sum_e;
run;

data var2 (drop=_freq_ _type_);
set var2;
g=1;
run;

data kbar1;
merge kbar1 var2;
by g;
var_y=(1/(n_g-1))*sum_e;
run;
```

Vedlegg 6. Rate-kalibrert ikke-informativ SHG-modell

```
*************************************************************************************;
* KURSOPPG #5                                                                       *;
*************************************************************************************;

*Rate-kalibrert ikke-informativ SHG-modell;

*Setter SHG indeks til sysselsettingsstrata og beregner interessevariabel t=syss.*svar;

Data kbar (drop=x);
set kurs.oppgave;
if r=1 then t=z*y;
else t= .;
g=x;
run;

data kbar;
set kbar;
x=z;
run;

* Estimerer svarsannsynlighetene og lager en variabel p med de estimerte verdiene;

proc means data = kbar noprint nway;
class g;
var r;
output out= aggr_kbar (drop= _type_ rename= (_freq_=brutto_g))
        sum=n_g;
run;

proc sort data=kbar; by g;
proc sort data=aggr_kbar; by g;

data kbar;
merge kbar aggr_kbar;
by g;
p=n_g/brutto_g;
label p = 'p=est.svarsh.';
run;

* Beregner designvekten;

data kbar;
set kbar;
w=1/(trekksh*p);
label w='W=designvekt';
run;

* BRUKER MAKRO FOR Å BEREGNE RATE;

%include '$KONJBAR/sas/prog/kurs/makro/rate.sas' / source2;

%let X=292940;
%let imputering=nei;

%rate(kbar);

* lager en variabel kwt som er produktet av den rate-kalibrerte designvekten og interessevariabelen;

data kbar;
set kbar;
kwt=kw*t;
diff=kw-w;
rat=kw/w;
run;
```

```
* Estimerer andelen ved først å estimere Y (konjunkitursysselsettingen);
* og så dele på totalt antall sysselsatte i populasjonen;

proc means data=kbar noprint nway;
var kwt;
output out= estimat (drop=_freq_ _type_)
        sum=sum_kwt
run;

data estimat;
set estimat;
Y_andel=sum_kwt/292940;
run;

* Skriver ut raten som benyttes i kalibreringen;

proc freq data=kbar;
tables rat;
run;
```

Vedlegg 7. Rate-kalibrert informativ SHG-modell

```
*************************************************************************************************;
* KURSOPPG #6                                                                                   *;
*************************************************************************************************;

* Rate-kalibrert informativ SHG-modell;

*Setter SHG indeks til svaralternativ;
*Tilleggsklasse er definert som sysselsettingsstratum;

Data kbar (drop=x);
set kurs.oppgave;
if r=1 then t=z*y;
else t= .;
format h 4.
          g 4.;
h=x;
g=x3;
if g < 1 then g= .;
run;

data kbar;
set kbar;
x=z;
run;

* Estimerer svarsannsynlighetene og lager en variabel p med de estimerte verdiene;

%include '$KONJBAR/sas/prog/kurs/makro/svarsh.sas'/source2;
%let ant_it=40;
%svarsh(kbar);

proc sort data=kbar; by g;
proc sort data=svarsh; by g;

data kbar;
merge kbar svarsh;
by g;
run;

* Beregner designvekten;

data kbar;
set kbar;
w=1/(trekksh*p);
label w='W=designvekt';
run;

* BRUKER MAKRO FOR Å BEREGNE RATE;

%include '$KONJBAR/sas/prog/kurs/makro/rate.sas' / source2;

%let X=292940;
%let imputering=nei;

%rate(kbar);

* lager en variabel kwt som er produktet av den rate-kalibrerte designvekten og interessevariabelen;

data kbar;
set kbar;
kwt=kw*t;
diff=kw-w;
rat=kw/w;
run;
```

```
* Estimerer andelen ved først å estimere Y (konjunkitursysselsettingen);
* og så dele på totalt antall sysselsatte i populasjonen;

proc means data=kbar noprint nway;
var kwt;
output out= estimat (drop=_freq_ _type_)
        sum=sum_kwt
run;

data estimat;
set estimat;
Y_andel=sum_kwt/292940;
run;

* Skriver ut raten som benyttes i kalibreringen;

proc freq data=kbar;
tables rat;
run;
```

Vedlegg 8. Imputering fra 'nærmeste nabo'

```
*************************************************************************************;
* KURSOPPG#7                                                                        *;
*************************************************************************************;

* Imputering : nærmeste nabo, ikke kalibrert;

data kbar;
set kurs.oppgave;
run;

%include '$KONJBAR/sas/prog/kurs/makro/nabo.sas' / source2;
%nabo (kbar);

* Lager variabelen tt fot t est;

data kbar;
set kbar;
if r=1 then tt=z*y;
else tt=z*imp;
run;

* Beregner utvalgsvekten;

data kbar;
set kbar;
a=1/(trekksh);
label a='A=utvalgsvekt';
run;

* lager en variabel att som er produktet av a og tt;

data kbar;
set kbar;
att=a*tt;
run;

* Estimerer andelen ved først å estimere Y (konjunkitursysselsettingen);
* og så dele på totalt antall sysselsatte i populasjonen;

proc means data=kbar noprint nway;
var att;
output out= estimat (drop=_freq_ _type_)
        sum=sum_att
run;

data estimat;
set estimat;
Y_andel=sum_att/292940;
run;
```

Vedlegg 9. Imputering med Hot-deck under ikke-inf. SHG

```
*************************************************************************************;
* KURSOPPG #8                                                                       *;
*************************************************************************************;

* Hot-deck imputering under ikke-informativ SHG-frafallsmodell;

*Setter SHG indeks til næringsgruppe;

Data kbar;
set kurs.oppgave;
g=x2;
run;

* Imputerer med Hot-Deck metode innenfor antakelsen om ikke informative SHG'er
* definert med næringsgruppe;

%include '$KONJBAR/sas/prog/kurs/makro/hotdeck.sas' / source2;

%hotdeck(kbar);


* Beregner utvalgsvekten a;

data kbar;
set kbar;
a=1/(trekksh);
label a='a=utvalgsvekt';
run;

* lager en variabel ayy som er produktet av a og y est;

data kbar;
set kbar;
ayy=a*yy;
run;

* Beregner de sysselsettingsveide estimatene som er produktet av ayy og z=sysselsetting;

data kbar;
set kbar;
ayyt=ayy*z;
run;

* Estimerer andelen ved først å estimere Y og så dele på total;

proc sort data=kbar; by x;

proc means data=kbar noprint nway;
var ayyt;
output out= estimat (drop=_freq_ _type_)
        sum=sum_ayyt;
run;

data estimat;
set estimat;
T=sum_ayyt/292940;
run;
```

Vedlegg 10. Rate-kalibrert med imputering fra 'nærmeste nabo'

```
*************************************************************************************;
* KURSOPPG#9                                                                        *;
*************************************************************************************;

* Imputering : nærmeste nabo, kalibrert ved rateestimator;

data kbar (drop=x);
set kurs.oppgave;
run;

data kbar;
set kbar;
x=z;
run;

%include '$KONJBAR/sas/prog/kurs/makro/nabo.sas' / source2;
%nabo (kbar);

* Lager variabelen tt fot t est;

data kbar;
set kbar;
if r=1 then tt=z*y;
else tt=z*imp;
run;

* Beregner utvalgsvekten;

data kbar;
set kbar;
a=1/(trekksh);
label a='A=utvalgsvekt';
run;

* BRUKER MAKRO FOR Å BEREGNE RATE;

%include '$KONJBAR/sas/prog/kurs/makro/rate.sas' / source2;

%let X=292940;
%let imputering=ja;

%rate(kbar);


* lager en variabel katt som er produktet av den rate-kalibrerte utvalgsvekten og interessevariabelen;

data kbar;
set kbar;
katt=ka*tt;
diff=ka-a;
rat=ka/a;
run;


* Estimerer andelen ved først å estimere Y (konjunkitursysselsettingen);
* og så dele på totalt antall sysselsatte i populasjonen;

proc means data=kbar noprint nway;
var att;
output out= estimat (drop=_freq_ _type_)
        sum=sum_att
run;

data estimat;
```

```
set estimat;
Y_andel=sum_att/292940;
run;

proc freq data=kbar;
tables rate;
run;
```

Vedlegg 11. Rate-kalibrert hot-deck imputering

```
*************************************************************************************;
* KURSOPPG #10                                                                      *;
*************************************************************************************;

* Rate-kalibrert hot-deck imputering under ikke-informativ SHG-frafallsmodell;

*Setter SHG indeks til næringsgruppe;

Data kbar (drop=x);
set kurs.oppgave;
g=x2;
run;

data kbar;
set kbar;
x=z;
run;

* Imputerer med Hot-Deck metode innenfor antakelsen om ikke informative SHG'er;
* definert med næringsgruppe;

%include '$KONJBAR/sas/prog/kurs/makro/hotdeck.sas' / source2;

%hotdeck(kbar);


* Beregner utvalgsvekten a;

data kbar;
set kbar;
a=1/(trekksh);
label a='a=utvalgsvekt';
run;


* BRUKER MAKRO FOR Å BEREGNE RATE;

%include '$KONJBAR/sas/prog/kurs/makro/rate.sas' / source2;

%let X=292940;
%let imputering=ja;

%rate(kbar);

* lager en variabel kayy som er produktet av de kalibrerte utvalgsvektene ka og y est;

data kbar;
set kbar;
kayy=ka*yy;
run;

* Beregner de sysselsettingsveide estimatene som er produktet av kayy og z=sysselsetting;

data kbar;
set kbar;
kayyt=kayy*z;
diff=ka-a;
rat=ka/a;
run;

* Estimerer andelen ved først å estimere Y og så dele på total;

proc sort data=kbar; by x;
```

```
proc means data=kbar noprint nway;
var kayyt;
output out= estimat (drop=_freq_ _type_)
        sum=sum_kayyt;
run;

data estimat;
set estimat;
T=sum_kayyt/292940;
run;

proc freq data=kbar;
tables rat;
run;
```

De sist utgitte publikasjonene i serien Notater

2003/51 C. Wiecek: Undersøkelse om fremtidsplaner, familie og samliv. Dokumentasjonsrapport. 59s.

2003/52 KOSTRA: Arbeidsgrupperapporter 2003. 153s.

2003/53 A. Haglund: Rapport fra arbeidsgruppa om forslag til arbeidsdeling mellom Brønnøysundregistrene (BR) og Statistisk sentralbyrå (SSB). 40s.

2003/54 E. Eng Eibak: Forventningsindikator - konsumprisene. Mai - november 2003. 19s.

2003/55 G. Daugstad: Levekår for ungdom i større byer. 80s.

2003/56 A. Vedø og D. Rafat: Sammenligning av utvalgsplaner i AKU. 17s.

2003/57 L. Belsby: Frafall og vekter i Tidsbruksundersøkelsen 2000-2001. 20s.

2003/58 L.Belsby: Vekter i Forbruksundersøkelsen. 28s.

2003/59 M. Mogstad og L.C. Zhang: På veien fra familie- til husholdningsregister. En metode for prediksjon av samboere uten barn .53s

2003/60 A. Vedø og D. Rafat: Redigering av husholdningsfilen fra Kvalitetsundersøkelsen. 13s.

2003/61 M. Mogstad: Analyse av fattigdom basert på register- og folketellingsdata. 75s.

2003/62 T. Eika og J.A. Jørgensen: Makroøkonomiske virkninger av høye strømpriser i 2003. En analyse med den makroøkonometriske modellen KVARTS.16s

2003/63 B. Mathisen: Flyktninger og arbeidsmarkedet 4. kvartal 2001. 32s.

2003/64 E. Røed Larsen og D.E. Sommervoll: Til himmls eller utfor stupet? En katalogisering av forklaringer på stigende boligpriser. 31s.

2003/65 P.E. Tønjum: Tilbakemelding/ dokumentasjon av prosjektet: Avstemming av KNR mot nye årstall ifølge tallrevisjonen.43s.

2003/66 B.A. Holth: Arbeids- og bedriftsundersøkelsen 2003. Dokumentasjon. 67s.

2003/67 H. Tønseth: Kommuneale helseforskjeller -de finnes, men kan de måles? 15s.

2003/68 T.M. Normann: Omnibusundersøkelsen mai/juni 2003. Dokumentasjonsrapport. 50s.

2003/69 KOSTRA (Kommune- Stat- Rapportering) Rutinebeskrivelse og dokumentasjon. 60s.

2003/70 E. Holmøy og B. Strøm: Fordeling av tjenesteproduksjon mellom offentlig og privat sektor i MSG-6. 25s.

2003/71 J.K. Dagsvik: Hvordan skal arbeidstilbudseffekter tallfestes? en oversikt over den mikrobaserte arbeidstilbudsforskningen i Statistisk sentralbyrå. 67s.

2003/72 A. Steinkellner: Inntektsstatistikk for personer og familier 1999-2001. Dokumentasjon av datagrunnlag og produksjonsprosess. 43s.

2003/73 F. Tverå, I. Sagelvmo: Beregning av næringene fiske eget bruk, fiske og fangst og fiskeoppdrett i nasjonalregnskapet. 19s.

2003/74 K.H. Grini: Lønnsstatistikk privat sektor 1997-2001. Dokumentasjon av utvalg og beregning av vekter. 36s.

2003/75 A.H. Foss: Grafisk revisjon av nøkkeltallene i KOSTRA. 16s.

2003/76 K. Hansen: Ideelle organisasjoner i nasjonalregnskapet. 30s.

2003/77 E.E: Eibak: Undersøking om foreldrebetaling i barnehagar, august 2003. 46s.

2003/78 A.H. Foss: Kvaliteten i husholdningsdelen i Folke- og boligtellingen 2001. 31s.

2003/79 O. Villund: Yrke i Arbeidstakerregisteret. 31s.

2003/80 O. Villund: Partielt frafall av yrkesdata i Arbeidstakerregisteret. 18s.