Statistics Norway
Statistisk sentralbyrå

Anne Vedø

Notater

Estimering for undersysselsetting i AKU basert på modellbasert imputering

Stabsavdeling/Seksjon for statistiske metoder og standarder

Innhold

1. Innledning

Dette notatet er tredje del av imputeringsprosjektet. Første del omhandler nåværende imputeringsrutiner i AKU, se [1]. Andre del tar for seg modellering av populasjonen og responsmekanismen i AKU, se [2]. Her i dette notatet ser vi på imputering og estimering i AKU, basert på modellene beskrevet i [2]. Dette notatet er en fortsettelse av [2], og ser på modellbasert imputering for partielt frafall i AKU, hovedsaklig angående undersysselsetting.

1.1. Spørsmål i AKU med partielt frafall

For at ikke modellene skal bli altfor kompliserte, har vi valgt å betrakte 0/1-variable, og å gjøre om aktuelle spørsmål som har flere svaralternativ, til 0/1-variable.

Spørsmålene som skal vurderes er listet opp under. I dette notatet bruker vi den spørsmålsnummereringen som var på spørreskjemaet i 1992. Skjemaet har blitt endret etter dette, og spørsmålene har fått nye nummer. Nytt nummer står i parentes.

18 (SYS34a): Ønske om lengre arbeidstid for deltidssysselsatte
Y=1 hvis IO ønsker lengre arbeidstid
Y=0 hvis IO ikke ønsker lengre arbeidstid

19 (SYS35a) : Forsøk på å få lengre arbeidstid for deltidssysselsatte med ønske om lengre arbeidstid
Y=1 hvis IO har forsøkt å få lengre arbeidstid
Y=0 hvis IO ikke har forsøkt å få lengre arbeidstid

23a (SYS36a/SYS36c): Hvor raskt IO kan starte med økt arbeidstid, for deltidssysselsatte med ønske om lengre arbeidstid
Y=1 hvis IO kan starte med økt arbeidstid før det er gått en måned
Y=0 ellers

Spørsmål 18, 19 og 23a utgjør tilsammen et spørsmål om *undersysselsetting*. For å bli regnet som undersysselsatt, må man svare ja på alle disse spørsmålene. Det er antall undersysselsatte som er den sentrale størrelsen som skal estimeres, og ikke antallet som svarer ja på hvert enkelt av spørsmålene 18, 19 og 23a. Vi vil derfor også vurdere en direkte modell for undersysselsetting, slik at vi senere vil kunne imputere direkte for undersysselsetting, uten først å imputere for spørsmål 18, 19 og 23a.

Undersysselsetting:

Y=1 dersom IO svarer ja på spørsmål 18, 19 og 23a
Y=0 dersom IO svarer nei på minst ett av spørsmålene 18, 19 og 23a

24 (SYS34b): Ønsket arbeidstid for deltidssysselsatte med ønske om lengre arbeidstid
Y=1 hvis ønsket arbeidstid ≥ 37 timer
Y=0 hvis ønsket arbeidstid > 0 og < 37 timer

Spørsmålene om undersysselsetting henger sammen. Deltidssysselsatte blir spurt om hva de hovedsaklig betrakter seg som (spørsmål 17), og får en rekke svaralternativ (yrkesaktiv, student,...., arbeidsledig, vernepliktig). Deretter stilles spørsmål 18. Av IO som har svart på spørsmål 17 ved direkte intervju, er det ca. 60 prosent som har partielt frafall på spørsmål 18. Man regner her med at intervjueren kan ha «glemt» å krysse av på spørsmål 18, (Hobæk 1993, s. 3). Dersom spørsmål 18 ikke er besvart, så er sannsynligheten liten for at spørsmålene 19, 23a eller 24 er besvart. De fleste med uoppgitt i denne sekvensen, har uoppgitt på alle de tre spørsmålene 18, 19 og 23a.

25 (fAr11): Faktisk arbeidstid når IO har ett arbeidsforhold
Y=1 hvis faktisk arbeidstid ≥ 37 timer
Y=0 hvis faktisk arbeidstid ≥ 0 og < 37 timer

62 (ISY83): Ønsket arbeidstid for arbeidsledige
Y=1 hvis ønsket arbeidstid ≥ 37 timer
Y=0 hvis ønsket arbeidstid > 0 og < 37 timer

1.2. Populasjonsmodell

Y betegner det faktiske svaret intervjuobjektet (IO) har på et gitt spørsmål, enten IO har svart på spørsmålet eller ikke. Y er kjent i svarutvalget og ukjent i frafallet. Modellen for Y er logistisk med alder, kjønn og region som forklaringsvariable. Disse betegnes med $\mathbf{x} = (x_1, x_2, x_3)$. I tillegg til hovedeffektene har vi tatt med alle tre annen ordens kryssledd. Alder deles inn i de tre gruppene 16-19 år, 20-39 år og 40-66 år, nummerert gruppe 1, 2 og 3. Kjønn er 0 for menn og 1 for kvinner. Bostedsregion er delt i fem grupper: 1= Oslo og Akershus, 2=Resten av Østlandet, 3=Sørlandet og Vestlandet unntatt Møre og Romsdal, 4=Møre og Romsdal og Trøndelag, 5=Nord-Norge.

1.3. Enhetsfrafall ved første henvendelse

Som nevnt i [2] (side 3) er "ekte" enhetsfrafall, dvs. personer som ikke har levert skjema i det hele tatt, utelatt fra denne analysen. R^1 betegner enhetsfrafall på første henvendelse, dvs. at IO ikke svarer på noen av spørsmålene ved første henvendelse. Modellen for R^1 er logistisk med sivilstand, kjønn og region som forklaringsvariable. Disse betegnes med $\mathbf{z} = (z_1, z_2, z_3)$. Sivilstand kodes 0 for aleneboende (ugifte og før gifte) og 1 for ikke-aleneboende (gifte og samboere). Også her har vi tatt med alle tre annen ordens kryssledd i tillegg til hovedeffektene. Vi antar også at R^1 er uavhengig av Y.

1.4. Partielt frafall

R^2 betegner partielt frafall på et gitt spørsmål. Modellen for R^2 er logistisk med sivilstand, kjønn, region, Y og R^1 som forklaringsvariable. Her har vi ikke tatt med kryssledd i modellen.

1.5. Undersysselsetting

For spørsmålssekvensen om undersysselsetting, har vi undersøkt tre forskjellige modeller.

La Y_u være indikatorvariabel for undersysselsetting, og R_u^2 indikatorvariabel for svar på spørsmål om undersysselsetting. IO sies å ha svart på spørsmålet om undersysselsetting dersom det utfra IO's svarskjema er mulig å avgjøre om IO er undersysselsatt eller ikke. Dette betyr at $R_u^2 = 1$ dersom IO enten har svart ja på både spørsmål 18, 19 og 23a, eller har svart nei på minst ett av spørsmålene. I alle andre tilfeller er $R_u^2 = 0$.

Modell 1

Vi betrakter Y_u og R_u^2 som om de var enkle variable. Vi modellerer altså Y_u etter populasjonsmodellen, R^1 etter modellen for enhetsfrafall og R_u^2 gitt Y_u , R^1 etter modellen for partielt frafall. De enkelte variablene Y_{18}, Y_{19}, Y_{23}, R_{18}^2, R_{19}^2 og R_{23}^2 modelleres ikke. Modellene for Y_u og R_u^2 gjelder for deltidssysselsatte.

Denne modellen komprimerer all informasjon om Y_{18}, Y_{19}, Y_{23}, R_{18}^2, R_{19}^2 og R_{23}^2 ned til R_u^2 og Y_u. Fordeler med dette er at det er enkelt, og det blir ikke så mange parametre å estimere. En ulempe er at metoden ikke skiller mellom de forskjellige frafallsstrukturene på spørsmål 18, 19 og 23a.

Modell 2

Motivasjonen for denne modellen, er å gjøre modellen avansert nok til å kunne skille mellom de ulike frafallsmønstrene.

Y_u modelleres etter populasjonsmodellen for alle deltidssysselsatte. De enkelte variablene Y_{18}, Y_{19} og Y_{23} modelleres ikke. Vi modellerer R_u^2 etter modellen for partielt frafall for alle deltidssysselsatte. R_{18}^2 modelleres etter modellen for partielt frafall, men bare når $R_u^2 = 0$. For å få brukbare estimater måtte vi her fjerne alle forklaringsvariable unntatt y_u. R_{19}^2 modelleres etter modellen for partielt frafall når $R_u^2 = 0$ og $R_{18}^2 = 1$. Her måtte vi forenkle modellen til bare å inneholde konstantledd.

Modell 3

Her modellerer vi alle de seks basisvariablene Y_{18}, Y_{19}, Y_{23}, R_{18}^2, R_{19}^2 og R_{23}^2. Fordelingene til Y_u og R_u^2 kan da utledes fra disse modellene.

Y_{18}, Y_{19} og Y_{23} modelleres etter populasjonsmodellen, bortsett fra at vi i modellen for Y_{23} har fjernet kryssleddene. Dette er gjort fordi parametrene i modellen med kryssledd ikke lot seg estimere med vårt datasett. Y_{18} modelleres for deltidssysselsatte, Y_{19} modelleres for deltidssysselsatte med $Y_{18} = 1$, og Y_{23} modelleres for deltidssysselsatte med $Y_{18} = 1$ og $Y_{19} = 1$. Modellen for Y_u er til slutt gitt ved:

$$P(Y_u = 1 \mid \mathbf{x}) = P(Y_{18} = 1 \cap Y_{19} = 1 \cap Y_{23} = 1 \mid \mathbf{x})$$
$$= P(Y_{23} = 1 \mid Y_{19} = 1, Y_{18} = 1, \mathbf{x}) P(Y_{19} = 1 \mid Y_{18} = 1, \mathbf{x}) P(Y_{18} = 1 \mid \mathbf{x})$$

og gjelder for deltidssysselsatte.

R_{18}^2, R_{19}^2 og R_{23}^2 modelleres i utgangspunktet etter modellen for partielt frafall, men i modellene for R_{19}^2 og R_{23}^2 tar vi hensyn til frafall på foregående spørsmål. I modellene for R_{19}^2 og R_{23}^2 bruker vi bare konstantledd, igjen på grunn av at den fulle modellen ikke lar seg estimere ut fra vårt datasett.

Under bruker vi spørsmålsnummeret som indeks på parametrene, for å tydeliggjøre at parametrene har forskjellige verdier for de forskjellige spørsmålene.

Modell for R_{18}^2:

Her bruker vi modellen for partielt frafall. Modellen gjelder for deltidssysselsatte.

$$\ln \frac{P(R_{18}^2 = 1 \mid \mathbf{z}, y_{18}, y_{19}, y_{23}, R^1)}{P(R_{18}^2 = 0 \mid \mathbf{z}, y_{18}, y_{19}, y_{23}, R^1)} = \psi_{18,0} + \psi_{18,1} z_1 + \psi_{18,2} z_2 + \sum_{l=1,3,4,5} \psi_{18,3l} D_{3l} + \psi_{18,4} y_{18} + \psi_{18,5} R^1$$

Modell for R_{19}^2:

Modellen gjelder for deltidssysselsatte som ikke har svart nei på spørsmål 18.

Når $R_{18}^2 = 0$ antar vi at IO heller ikke svarer på spørsmål 19.

$$P(R_{19}^2 = 1 | R_{18}^2 = 0) = 0$$

Når $R_{18}^2 = 1 \text{ og } Y_{18} = 1$ bruker vi i utgangspunktet modellen for partielt frafall, men i vårt konkrete tilfelle den enklere versjonen:

$$\ln \frac{P(R_{19}^2 = 1 \mid \mathbf{z}, y_{19}, y_{23}, R^1, R_{18}^2 = 1, Y_{18} = 1)}{P(R_{19}^2 = 0 \mid \mathbf{z}, y_{19}, y_{23}, R^1, R_{18}^2 = 1, Y_{18} = 1)} = \psi_{19,0}$$

Modell for R_{23}^2:

Modellen gjelder for personer som hverken har svart nei på spørsmål 18 eller 19.

Hvis man ikke har svart på spørsmål 19 antar vi at sannsynligheten for å svare på spørsmål 23a er null.

$$P(R_{23}^2 = 1 \mid R_{19}^2 = 0) = 0$$

For personer som har svart ja på spørsmål 18 og 19 bruker vi generelt modellen for partielt frafall, men i vårt konkrete tilfelle den enklere versjonen:

$$\ln \frac{P(R_{23}^2 = 1 \mid \mathbf{z}, y_{23}, R^1, Y_{18} = 1, Y_{19} = 1, R_{18}^2 = 1, R_{19}^2 = 1)}{P(R_{23}^2 = 0 \mid \mathbf{z}, y_{23}, R^1, Y_{18} = 1, Y_{19} = 1, R_{18}^2 = 1, R_{19}^2 = 1)} = \psi_{23,0}$$

2. Imputeringsmetoder

Vi undersøker to imputeringsmetoder:

1. Tilfeldig trekking fra estimert fordeling
2. Imputering av forventning i estimert fordeling

Vi deler inn det partielle frafallet i grupper etter kovariater og frafallsvariable, slik at sannsynligheten for at $Y = 1$ er lik innenfor hver gruppe. I det første tilfellet trekkes det tilfeldig et svar (1 eller 0) for hver person, dermed blir antallet som får imputert svaret 1 i hver gruppe stokastisk. I den andre metoden imputeres forventet antall i hver gruppe, avrundet til nærmeste heltall, altså en deterministisk imputeringsmetode.

2.1. Tilfeldig trekking fra estimert fordeling

La $\mathbf{x}$ være kovariatvektoren i populasjonsmodellen (dvs. $\mathbf{x}$ består av alder, kjønn og region), og $\mathbf{z}$ kovariatvektoren i modellen for partielt frafall (dvs. $\mathbf{z}$ består av sivilstand, kjønn og region). For personer med partielt frafall og med $R^1 = r^1$ og kovariatverdier $\mathbf{x},\mathbf{z}$, imputerer vi $Y = 1$ med sannsynlighet $P(Y = 1 \mid R^2 = 0, R^1 = r^1, \mathbf{x}, \mathbf{z})$, ev. også betinget med hensyn på frafallsmønstret for modell 2 og 3 for undersysselsetting.

For å regne ut numeriske verdier for sannsynligheten for at $Y = 1$ gitt frafall, ev. frafallsmønster, setter vi inn MLE for sannsynlighetene som inngår i uttrykkene. Disse estimatene er regnet ut på grunnlag av den totale modellen for populasjon og frafall.

I avsnitt 2.1.1-2.1.3 uttrykker vi sannsynligheten for at $Y = 1$ gitt partielt frafall, ev. også gitt frafallsmønster for modell 2 og 3 for undersysselsetting, for vilkårlig verdi av R^1, $\mathbf{x}$ og $\mathbf{z}$, ved hjelp av sannsynligheter som inngår i modell-ligningene. Vi kan dermed plugge inn tallene vi regnet ut i [2], og få tak i imputeringssannsynlighetene.

2.1.1. Spørsmål 24, 25 og 62 og modell 1 for undersysselsetting

$$P(Y = 1 \mid R^2 = 0, R^1 = r^1, \mathbf{x}, \mathbf{z})$$

$$= \frac{P(Y = 1, R^2 = 0, R^1 = r^1 \mid \mathbf{x}, \mathbf{z})}{P(R^2 = 0, R^1 = r^1 \mid \mathbf{x}, \mathbf{z})}$$

$$= \frac{P(Y = 1, R^2 = 0, R^1 = r^1 \mid \mathbf{x}, \mathbf{z})}{\sum_{y=0,1} P(Y = y, R^2 = 0, R^1 = r^1 \mid \mathbf{x}, \mathbf{z})}$$

$$= \frac{P(R^2 = 0 \mid Y = 1, R^1 = r^1, \mathbf{z}) P(Y = 1 \mid \mathbf{x}) P(R^1 = r^1 \mid \mathbf{z})}{\sum_{y=0,1} P(R^2 = 0 \mid Y = y, R^1 = r^1, \mathbf{z}) P(Y = y \mid \mathbf{x}) P(R^1 = r^1 \mid \mathbf{z})}$$

$$= \frac{P(R^2 = 0 \mid Y = 1, R^1 = r^1, \mathbf{z}) P(Y = 1 \mid \mathbf{x})}{\sum_{y=0,1} P(R^2 = 0 \mid Y = y, R^1 = r^1, \mathbf{z}) P(Y = y \mid \mathbf{x})}$$

2.1.2. Modell 2 for undersysselsetting

Vi går frem på samme måte som over, men betinger også med hensyn på R^2_{18} og R^2_{19}.

$$P(Y = 1 \mid R^2_u = 0, R^2_{18} = r^2_{18}, R^2_{19} = r^2_{19}, R^1 = r^1, \mathbf{x}, \mathbf{z})$$

$$= \frac{P(Y = 1, R^2_u = 0, R^2_{18} = r^2_{18}, R^2_{19} = r^2_{19}, R^1 = r^1 \mid \mathbf{x}, \mathbf{z})}{P(Y = 0, R^2_u = 0, R^2_{18} = r^2_{18}, R^2_{19} = r^2_{19}, R^1 = r^1 \mid \mathbf{x}, \mathbf{z}) + P(Y = 1, R^2_u = 0, R^2_{18} = r^2_{18}, R^2_{19} = r^2_{19}, R^1 = r^1 \mid \mathbf{x}, \mathbf{z})}$$

Sannsynlighetene som inngår i brøken over kan skrives

$$P(Y=y, R_u^2=0, R_{18}^2=r_{18}^2, R_{19}^2=r_{19}^2, R^1=r^1 \mid \mathbf{x},\mathbf{z})$$
$$= P(R_{19}^2=r_{19}^2 \mid R_u^2=0, R_{18}^2=r_{18}^2) P(R_{18}^2=r_{18}^2 \mid Y=y, R_u^2=0)$$
$$\cdot P(R_u^2=0 \mid Y=y, R^1=r^1, \mathbf{z}) P(Y=y \mid \mathbf{x}) P(R^1=r^1 \mid \mathbf{z})$$

Dette er sannsynligheter vi finner fra modell-ligningene. Vi må bare huske på at $P(R_{19}^2=0 \mid R_u^2=0, R_{18}^2=0)=1$ og $P(R_{19}^2=1 \mid R_u^2=0, R_{18}^2=0)=0$.

2.1.3. Modell 3 for undersysselsetting

I modell 3 er Y_u modellert implisitt gjennom Y_{18}, Y_{19} og Y_{23}. For å regne ut estimatene våre for antall undersysselsatte, må vi finne MLE for parametrene i modell 3, basert på det komplette, imputerte utvalget. Dette betyr at det imputerte utvalget må være komplett med hensyn på både Y_{18}, Y_{19} og Y_{23}. Det holder ikke å bare imputere direkte for undersysselsetting, som vi gjorde i modell 1 og 2. Vi skal derfor imputere separat på spørsmål 18, 19 og 23. Siden undersysselsettingsvariabelen er en avledet variabel, blir det imputerte datasettet også komplett med hensyn på undersysselsetting.

Vi finner sannsynligheten for å være undersysselsatt gitt forskjellige frafallsmønstre. De tre frafallsmønstrene behandles hver for seg. For enkelhet skyld har vi utelatt betingingen med hensyn på R^1, $\mathbf{x}$ og $\mathbf{z}$.

Frafallsmønster (0,0,0):

$$P(Y_u=1 \mid (R_{18}^2, R_{19}^2, R_{23}^2)=(0,0,0))$$
$$= P(Y_{18}=1, Y_{19}=1, Y_{23}=1 \mid (R_{18}^2, R_{19}^2, R_{23}^2)=(0,0,0))$$

Ifølge modellen er det sannsynlighet null for å svare på spørsmål 19 og 23 når man ikke har svart på spørsmål 18, så det holder å betinge med hensyn på at $R_{18}^2=0$.

$$P(Y_{18}=1, Y_{19}=1, Y_{23}=1 \mid R_{18}^2=0)$$
$$= \frac{P(Y_{18}=1, Y_{19}=1, Y_{23}=1, R_{18}^2=0)}{P(R_{18}^2=0)}$$
$$= \frac{P(R_{18}^2=0 \mid Y_{18}=1, Y_{19}=1, Y_{23}=1) P(Y_{23}=1 \mid Y_{18}=1, Y_{19}=1) P(Y_{19}=1 \mid Y_{18}=1) P(Y_{18}=1)}{P(R_{18}^2=0)}$$

I modellen er R_{18}^2 uavhengig av Y_{19} og Y_{23}, så vi kan forenkle til

$$\frac{P(R_{18}^2=0 \mid Y_{18}=1) P(Y_{18}=1)}{P(R_{18}^2=0)} P(Y_{23}=1 \mid Y_{18}=1, Y_{19}=1) P(Y_{19}=1 \mid Y_{18}=1)$$
$$= P(Y_{18}=1 \mid R_{18}^2=0) P(Y_{19}=1 \mid Y_{18}=1) P(Y_{23}=1 \mid Y_{18}=1, Y_{19}=1)$$

For personer med frafallmønster (0,0,0) imputerer vi derfor ja på spørsmål 18 med sannsynlighet $P(Y_{18}=1 \mid R_{18}^2=0)$. Blant dem som får imputert ja på spørsmål 18, imputerer vi ja på spørsmål 19 med sannsynlighet $P(Y_{19}=1 \mid Y_{18}=1)$. Merk at denne sannsynligheten ikke er betinget med hensyn på frafall på spørsmål 19. Dette er rimelig, for personer med frafall på spørsmål 18 blir ikke stilt

spørsmål 19. De har dermed ikke noe eksplisitt frafall på spørsmål 19. Blant dem som får imputert ja på spørsmål 19, imputeres til slutt ja på spørsmål 23 med sannsynlighet $P(Y_{23}=1 \mid Y_{18}=1, Y_{19}=1)$.

Frafallsmønster (1,0,0):
Personer med dette frafallsmønsteret som ikke har svart på undersysselsetting, må ha svart ja på spørsmål 18, så

$$P(Y_u=1 \mid R_u^2=0, (R_{18}^2, R_{19}^2, R_{23}^2)=(1,0,0))$$
$$= P(Y_{19}=1, Y_{23}=1 \mid Y_{18}=1, (R_{18}^2, R_{19}^2, R_{23}^2)=(1,0,0))$$

Modellsannsynligheten for å svare på spørsmål 23 gitt frafall på spørsmål 19 er null, så vi kan utelate R_{23}^2 fra betingelsen

$$P(Y_{19}=1, Y_{23}=1 \mid Y_{18}=1, (R_{18}^2, R_{19}^2)=(1,0))$$
$$= \frac{P(Y_{18}=1, Y_{19}=1, Y_{23}=1, (R_{18}^2, R_{19}^2)=(1,0))}{P(Y_{18}=1, (R_{18}^2, R_{19}^2)=(1,0))}$$

Telleren kan skrives

$$P(Y_{18}=1, Y_{19}=1, Y_{23}=1, (R_{18}^2, R_{19}^2)=(1,0))$$
$$= P(R_{19}^2=0 \mid Y_{18}=1, Y_{19}=1, Y_{23}=1, R_{18}^2=1) P(R_{18}^2=1 \mid Y_{18}=1, Y_{19}=1, Y_{23}=1)$$
$$\cdot P(Y_{23}=1 \mid Y_{18}=1, Y_{19}=1) P(Y_{19}=1 \mid Y_{18}=1) P(Y_{18}=1)$$

R^2-ene er uavhengige av Y-verdier på senere spørsmål, så dette kan forenkles til

$$P(R_{19}^2=0 \mid Y_{18}=1, Y_{19}=1, R_{18}^2=1) P(R_{18}^2=1 \mid Y_{18}=1)$$
$$\cdot P(Y_{23}=1 \mid Y_{18}=1, Y_{19}=1) P(Y_{19}=1 \mid Y_{18}=1) P(Y_{18}=1)$$

Nevneren kan skrives

$$P(Y_{18}=1, (R_{18}^2, R_{19}^2)=(1,0))$$
$$= P(R_{19}^2=0 \mid Y_{18}=1, R_{18}^2=1) P(R_{18}^2=1 \mid Y_{18}=1) P(Y_{18}=1)$$

Etter forkortning sitter vi igjen med

$$\frac{P(R_{19}^2=0 \mid Y_{18}=1, Y_{19}=1, R_{18}^2=1) P(Y_{23}=1 \mid Y_{18}=1, Y_{19}=1) P(Y_{19}=1 \mid Y_{18}=1)}{P(R_{19}^2=0 \mid Y_{18}=1, R_{18}^2=1)}$$
$$= P(Y_{19}=1 \mid R_{19}^2=0, Y_{18}=1, R_{18}^2=1) P(Y_{23}=1 \mid Y_{18}=1, Y_{19}=1)$$

For personer med frafallsmønster (1,0,0) imputerer vi dermed ja på spørsmål 19 med sannsynlighet $P(Y_{19}=1 \mid R_{19}^2=0, Y_{18}=1, R_{18}^2=1)$, og blant dem som får imputert ja på spørsmål 19, imputerer vi ja på spørsmål 23 med sannsynlighet $P(Y_{23}=1 \mid Y_{18}=1, Y_{19}=1)$.

Frafallsmønster (1,1,0):
Personer med dette frafallsmønsteret som ikke har svart på undersysselsetting, må ha svart ja på spørsmål 18 og 19, så

$$
\begin{aligned}
&P(Y_u = 1 \mid R_u^2 = 0, (R_{18}^2, R_{19}^2, R_{23}^2) = (1,1,0)) \\
&= P(Y_{23} = 1 \mid Y_{18} = 1, Y_{19} = 1, (R_{18}^2, R_{19}^2, R_{23}^2) = (1,1,0)) \\
&= \frac{P(Y_{18} = 1, Y_{19} = 1, Y_{23} = 1, (R_{18}^2, R_{19}^2, R_{23}^2) = (1,1,0))}{P(Y_{18} = 1, Y_{19} = 1, (R_{18}^2, R_{19}^2, R_{23}^2) = (1,1,0))} \\
&= P(Y_{23} = 1 \mid Y_{18} = 1, Y_{19} = 1, (R_{18}^2, R_{19}^2, R_{23}^2) = (1,1,0))
\end{aligned}
$$

og dette blir imputeringssannsynligheten på spørsmål 23.

2.2. Imputering av forventning i estimert fordeling

I hver gruppe (inndelt etter kovariater og frafallsvariable) imputeres forventet antall ja-svar, dvs. antall personer med partielt frafall i gruppen ganget med $P(Y = 1 \mid R^2 = 0, R^1 = r^1, \mathbf{x}, \mathbf{z})$, eventuelt også betinget med hensyn på frafallsmønster for modell 2 og 3 for undersysselsetting. Vi runder av til nærmeste heltall.

3. Estimatorer

Vi vil se på to estimatorer for totaler:

1. Prediksjonsestimatoren
2. Regresjonsestimatoren

Totalen som skal estimeres er summen av y_i-ene i hele populasjonen. I tilfellet med 0/1-variable er dette lik antall personer som har $Y = 1$ (svarer "ja").

Hver estimator har en basisversjon og en imputeringsversjon. Basisestimatoren er estimatoren slik den ser ut når den anvendes på et utvalg uten frafall. Imputeringsestimatoren er basisestimatoren anvendt på et imputert utvalg.

3.1. Prediksjonsestimatoren

Basisestimator:

$$\hat{t}_{pred} = \sum_{i \in s} y_i + \sum_{i \notin s} \hat{E}(Y_i \mid \mathbf{x}_i) = \sum_{i \in s} y_i + \sum_{i \notin s} \hat{P}(Y_i = 1 \mid \mathbf{x}_i)$$

Her er $\hat{P}(Y_i = 1 \mid \mathbf{x}_i)$ MLE for $P(Y_i = 1 \mid \mathbf{x}_i)$, basert på dataene i utvalget.

Her deler vi populasjonen i to; personene i utvalget og personene utenfor utvalget. Antall ja-svar i populasjonen er summen av antall ja-svar i utvalget og antall ja-svar utenfor utvalget. I et utvalg uten frafall kjenner vi antall ja-svar utvalget. Dette er den første summen i $\hat{t}_{pred}$. Utenfor utvalget må vi

predikere dette antallet. Dette gjør vi ved å legge sammen de estimerte sannsynlighetene for å svare ja for alle personene utenfor utvalget. Dette er den andre summen i $\hat{t}_{pred}$.

Imputeringsestimator:
La s_r betegne svarutvalget, $s - s_r$ frafallet, og la y_i^*, $i \in s - s_r$, betegne de imputerte verdiene i frafallet. Vi får imputeringsestimatoren ved å bruke basisestimatoren på det imputerte utvalget.

$$\hat{t}_{I,pred} = \sum_{i \in s_r} y_i + \sum_{i \in s - s_r} y_i^* + \sum_{i \notin s} \hat{P}^I(Y_i = 1 \mid \mathbf{x}_i)$$

Her er $\hat{P}^I(Y_i = 1 \mid \mathbf{x}_i)$ MLE for $P(Y_i = 1 \mid \mathbf{x}_i)$, basert på det imputerte datasettet.

3.2. Regresjonsestimatoren

Basisestimator:

$$\hat{t}_{reg} = \sum_{i=1}^{N} \hat{P}(Y_i = 1 \mid \mathbf{x}_i)$$

Forskjellen mellom regresjonsestimatoren og prediksjonsestimatoren er at regresjonsestimatoren predikerer antall ja-svar i *hele* populasjonen, også i utvalget. Når utvalget er en liten andel av populasjonen, blir de to estimatorene tilnærmet like.

Imputeringsestimator:

$$\hat{t}_{I,reg} = \sum_{i=1}^{N} \hat{P}^I(Y_i = 1 \mid \mathbf{x}_i)$$

4. Estimeringsmetoder

4.1. Med imputering

Hver imputeringsmetode kan kombineres med hver estimator. Vi ser på to imputeringsmetoder og to estimatorer. Dette gir dermed opphav til fire estimeringsmetoder.

4.2. Uten imputering

Det aller enkleste er å betrakte svarutvalget som et komplett utvalg, og benytte basisversjonene av prediksjons- og regresjonsestimatoren på dette utvalget. $\hat{P}(Y_i = 1 \mid \mathbf{x}_i)$ er da MLE for $P(Y_i = 1 \mid \mathbf{x}_i)$ under populasjonsmodellen alene, altså en ordinær logistisk modell.

Vi kan forbedre metoden over ved å la $\hat{P}(Y_i = 1 \mid \mathbf{x}_i)$ være MLE for $P(Y_i = 1 \mid \mathbf{x}_i)$ i den totale modellen, altså den simultane modellen for populasjon og frafall. På denne måten får vi tatt hensyn til at frafallsgruppen kan skille seg fra svarutvalget, uten å gå om imputering.

4.3. Ignorerbar frafallsmodell og etterstratifisering

For modell 1 og 3 for undersysselsetting har vi i tillegg regnet ut:

- Etterstratifiseringsestimatoren på imputerte datasett, ved hjelp av vektene som ligger på utvalgsfila. Dette er oppblåsningsfaktorene som ble brukt i produksjonen av offisielle tall. Disse er basert på 105 etterstrata, som er konstruert på bakgrunn av alder, kjønn, registrert sysselsetting, næring, bostedskommunens sentralitet og uførepensjon (se NOS C 87, Arbeidsmarkedstatistikk 1992).
- Estimater basert på en ignorerbar versjon av modellen, altså en modell der koeffisienten foran y i modellen for partielt frafall er satt lik 0.

5. Resultater

Vi har regnet ut både prediksjons- og regresjonsestimatoren, og tallene blir så å si identiske. Derfor viser vi bare prediksjonsestimatoren i tabellene under.

Tabellene inneholder prediksjonsestimatoren regnet ut på det reelle datasettet, og standardavvik og konfidensintervall basert på 1 000 simulerte datasett. De simulerte datasettene er de samme som vi brukte i [2]. De stokastiske variablene, dvs. Y-ene, R^1 og R^2-ene er simulert etter modellene beskrevet i [2], med ML-estimater basert på det reelle datasettet satt inn for parametrene i modellen. Kovariatverdiene, dvs. kjønn, alder, sivilstand og region er de samme i alle de 1 000 simuleringene, og lik kovariatverdiene i det reelle datasettet.

Datagrunnlag
Svarene på spørsmål 18, 19 og 23 fra det reelle datasettet for de 3757 personene som er brukt i analysen av undersysselsetting:

sp. 18	Ja	Nei	Frafall	Totalt
	950	2496	311	3757

De med ja på 18:

sp. 19	Ja	Nei	Frafall	Totalt
	553	392	5	950

De med ja på 19:

sp. 23	Ja	Nei	Frafall	Totalt
	510	34	9	553

5.1. Undersysselsetting

Den aktuelle populasjonen for spørsmålssekvensen om undersysselsetting er deltidssysselsatte. De to estimatorene våre forutsetter at antall personer i populasjonen er kjent i hvert stratum. Dette er ikke tilfellet for denne populasjonen. Vi estimerer antallet i populasjonen i hvert stratum ved hjelp av de samme oppblåsningsfaktorene som ble brukt til å lage offisielle tall for 1. kvartal 92. På denne måten blir stratumantallene våre konsistente med de publiserte tallene over deltidssysselsatte dette kvartalet. Det var 545 000 deltidssysselsatte i alderen 16 til 74 år 1. kvartal 92. I alderen 16 til 66 år var det 525 600 deltidssysselsatte.

Publisert antall undersysselsatte for 1. kvartal 92 er 77 000 (Ukens Statistikk nr. 30/31, 1993). Dette gjelder i aldersgruppen 16 til 74 år. Vi har utelatt personer mellom 67 og 74 år fra vår analyse av undersysselsetting, fordi det var for få personer i denne aldersgruppen i datasettet til å kunne estimere modellparametrene. Våre estimater gjelder derfor for aldersgruppen 16 til 66 år. Tallene er dermed ikke helt sammenlignbare, men det er veldig få undersysselsatte mellom 67 og 74, faktisk ingen i utvalget.

5.1.1. Modell 1

I denne modellen modelleres undersysselsettingsvariabelen Y_u, frafall på undersysselsetting (R_u^2) og frafall på første henvendelse (R^1).

Antall undersysselsatte 16-66 år

PREDIKSJONS-ESTIMATOREN	Uten imputering		Med imputering	
	Uten frafallsmodell	Med frafallsmodell	Metode a: Stokastisk fra estimert fordeling	Metode b: Forventet antall
Reelle data	79 311	82 731	83 491	81 806
Std. (1 000 sim.)	3 208	6 885	7 011	7 059
95% konf. int. (sim)	(73 738,86 269)	(70 846,98 031)	(70 725,98 567)	(70 283,98 015)

De metodene som tar hensyn til frafall (de tre siste kolonnene) gir litt høyere estimater enn metoden som bare benytter populasjonsmodellen på svarutvalget. Alle metodene gir høyere estimater enn det offisielle tallet. Dette stemmer godt overens med vurderingen gjort i [2], kap.3.1.3, som viser at AKU's imputeringsmetode høyst sannsynlig leder til underestimering av antall undersysselsatte. Fra konfidensintervallene ser vi, imidlertid, at estimatforskjellene ikke er statistisk signifikante.

Standardavviket for metode b (forventet antall) blir litt større enn standardavviket for metode a (tilfeldig trekking). Dette er merkelig, siden metode b er en deterministisk imputeringsmetode, mens metode a er stokastisk, med samme forventning som b i hvert stratum. Det kommer antakeligvis av at vi i metode b avrunder forventet antall til nærmeste heltall i hvert stratum. Vi benytter en logistisk modell i prediksjonsestimatoren, og den forutsetter strengt tatt at undersysselsettingsvariabelen er enten null eller en for alle personer. Det er likevel mulig å plugge inn data uten heltallig sum i hvert stratum i likelihoodfunksjonen. En slik gjennomkjøring av metode b uten avrunding gir et estimat fra de reelle dataene på 82 753 og et standardavvik på 6 919, altså mindre enn standardavviket for metode a.

I tabellen under har vi også til sammenligning regnet ut etterstratifiseringsestimatoren for de imputerte datasettene. Denne estimatoren gir en vekt til hver observasjon, så det er uproblematisk å bruke imputerte verdier forskjellig fra 0 og 1. Vektene er oppblåsningsfaktorene som ligger på datafila fra Seksjon 260.

ETTER-STRATIFISERING	Metode a	Metode b	Metode b uten avrunding
Reelle data	83 660	81 989	82 840
Std. (1 000 sim.)	7 013	7 071	6 925

Standardavviket for metode b uten avrunding blir også her litt mindre enn ved metode a. Etterstratifiseringsestimatoren og prediksjonestimatoren er svært like, både når det gjelder punktestimat og varians.

Det er også interessant å se hvor mye den ikke-ignorerbare delen av modellen, altså koeffisienten foran y, påvirker estimatene. I tabellen under har vi regnet ut en ignorerbar versjon av modell 1.

Antall undersysselsatte 16-66 år

PREDIKSJONS-ESTIMATOREN	Uten imputering		Med imputering	
	Uten frafallsmodell	Med frafallsmodell	Metode a	Metode b
Reelle data	79 311	79 321	80 526	78 550
Std. (1 000 sim.)	3 219	3 222	3 329	3 252
95 % konf. int. (sim)	(73 414,85 908)	(73 357,85 948)	(73 063,85 803)	(72 620,85 396)

Vi ser at de estimatene som tar hensyn til frafall blir redusert med rundt 3 000 personer, og standardavvikene blir mer enn halvert. Det blir nå liten forskjell mellom estimater som tar hensyn til frafall (de tre siste kolonnene) og det som ikke gjør det (første kolonne). Det at det gjør så stor forskjell å ekskludere koeffisienten foran y, tyder på at vi har ikke-ignorerbart frafall for undersysselsetting.

5.1.2. Modell 2

Til forskjell fra modell 1, skiller modell 2 mellom de forskjellige frafallsmønstrene blant personer med frafall på undersysselsetting.

Antall undersysselsatte 16-66 år

PREDIKSJONS-ESTIMATOREN	Uten imputering		Med imputering	
	Uten frafallsmodell	Med frafallsmodell	Metode a	Metode b
Reelle data	79 311	82 505	82 640	81 247
Std. (1 000 sim.)	3 177	6 628	6 729	6 818
95% konf. int. (sim)	(72 878,85 365)	(70 519,96 352)	(70 448,96 740)	(69 833,96 181)

Disse resultatene skiller seg lite fra modell 1. Dette er som forventet, for av de 325 personene som har frafall på undersysselsetting, har de aller fleste (311) frafallsmønsteret (0,0,0). Bare 5 personer har mønsteret (1,0,0) og 9 har mønsteret (1,1,0). Vi har derfor ikke gått videre i analysen av modell 2.

5.1.3. Modell 3

Dette er den mest detaljerte modellen. Her modelleres Y og R^2 for alle de tre spørsmålene som inngår i spørsmålssekvensen om undersysselsetting.

Antall undersysselsatte 16-66 år

PREDIKSJONS-ESTIMATOREN	Uten imputering		Med imputering	
	Uten frafallsmodell	Med frafallsmodell	Metode a	Metode b
Reelle data	80 785	94 515	96 836	96 422
Std. (1 000 sim.)	3 056	4 790	4 994	4 836
95% konf. int. (sim)	(74 598,86 595)	(83 653,102 578)	(83 096,102 968)	(84 504,103 672)

Her er det stor forskjell mellom de metodene som tar hensyn til frafall (de tre siste kolonnene) og den som ikke gjør det. Standardavvikene til estimatene i de tre siste kolonnene er også mindre enn de tilsvarende tallene for modell 1 og 2.

ETTER-STRATIFISERING	Metode a	Metode b
Reelle data	96 956	96 426
Std. (1 000 sim.)	5 002	4 846

Også her blir etterstratifiseringsestimatoren tilnærmet lik prediksjonsestimatoren.

Vi undersøker så hva som skjer hvis vi fjerner koeffisienten foran y i modellen for partielt frafall, slik at vi får en ignorerbar frafallsmodell.

Antall undersysselsatte 16-66 år

PREDIKSJONS-ESTIMATOREN	Uten imputering		Med imputering	
	Uten frafallsmodell	Med frafallsmodell	Metode a	Metode b
Reelle data	80 785	80 763	82 224	81 865
Std. (1 000 sim.)	3 088	3 094	3 236	3 151
95% konf. int. (sim)	(74 650,87 084)	(74 522,87 113)	(74 023,87 136)	(74 735,87 426)

Her får vi helt andre tall. Det er antagelsen om ikke-ignorerbarhet som får estimatene med og uten frafallsjustering til å sprike så mye i modell 3. Forskjellen mellom ignorerbar og ikke-ignorerbar modell er mye større enn den var for modell 1. Dette kommer av koeffisienten foran y i modellen for partielt frafall på spørsmål 18, som er stor og negativ (omtrent lik -3). Det fører til at den betingede sannsynligheten for å ha svaret ja på spørsmål 18 (dvs. ønske lengre arbeidstid) gitt frafall på spørsmål 18 blir veldig høy, og dermed blir det også høy sannsynlighet for å være undersysselsatt. I vedlegg A er en tabell som viser sannsynligheten for å imputere ja på spørsmålet om undersysselsetting i modell 1 og 3, og også antall imputerte undersysselsatte med metode b. Her kan man også finne de ubetingede sannsynlighetene for at Y=1 for Y_{18}, Y_{19}, Y_{23} og Y_u, samt den betingede sannsynligheten for at $Y_{18} = 1$ gitt partielt frafall på spørsmål 18.

5.2. Spørsmål 24

Målgruppen for spørsmål 24, dvs. de som får dette spørsmålet, er deltidssysselsatte med ønske om lengre arbeidstid. Vi bruker derfor denne gruppen som populasjon ved beregning av totaler. Vi anslår antallet i populasjonen i hvert stratum ved hjelp av AKUs oppblåsningsfaktorer. Dette gir en populasjon på 133 378 personer. I AKU publiseres det ikke tall for ønsket arbeidstid for denne målgruppen, bare for undersysselsatte.

Antall personer som ønsker fulltid, blant deltidssysselsatte med ønske om lengre arbeidstid, 16-66 år

PREDIKSJONS-ESTIMATOREN	Uten imputering		Med imputering	
	Uten frafallsmodell	Med frafallsmodell	Metode a	Metode b
Reelle data	69 624	68 968	68 708	68 464
Std. (1 000 sim.)	1 950	2 034	2 043	2 071
95% konf. int (sim)	(66 072,73 540)	(65 128,72 967)	(65 018,72 855)	(65 129,72 913)

Det blir liten forskjell mellom estimeringsmetodene her. Det partielle frafallet på dette spørsmålet er bare 2,3 prosent.

5.3. Spørsmål 25

Målgruppen for spørsmål 25 er personer med ett arbeidsforhold. Vi anslår antallet i populasjonen i hvert stratum som før. Dette gir en populasjon på 1 885 998 personer. Heller ikke her er det publisert noen sammenlignbare tall.

Antall personer med faktisk arbeidstid >= 37 timer, blant personer med ett arbeidsforhold, 16-74 år

PREDIKSJONS-ESTIMATOREN	Uten imputering		Med imputering	
	Uten frafallsmodell	Med frafallsmodell	Metode a	Metode b
Reelle data	1 099 188	1 099 208	1 099 555	1 098 701
Std. (1 000 sim.)	7 289	7 948	7 934	8 041
95% konf. int. (sim)	(1 084 540, 1 114 437)	(1 083 361, 1 114 877)	(1 083 557, 1 114 470)	(1 082 742, 1 114 578)

Det partielle frafallet på spørsmål 25 er 1,4 prosent, så bruk av frafallsmodell gir ikke utslag.

5.4. Spørsmål 62

Målgruppen for spørsmål 62 er arbeidsledige. Vi anslår antallet i populasjonen i hvert stratum ved hjelp av AKUs oppblåsningsfaktorer, som før. Ifølge US nr. 30/31, 1993, var det 132 000 arbeidssøkere mellom 16 og 74 år i 1. kvartal 1992. Av disse ønsket 101 000 en arbeidstid på 37 timer eller mer.
Antall arbeidsledige mellom 16 og 66 år var 131 100, så hvis det hadde blitt publisert et tall for antall personer med ønsket arbeidstid på 37 timer eller mer blant arbeidsledige i aldersgruppen 16-66 år, ville det ligget litt under 101 000.

Antall personer med ønsket arbeidstid >= 37 timer, blant arbeidsledige 16-66 år

PREDIKSJONS-ESTIMATOREN	Uten imputering		Med imputering	
	Uten frafallsmodell	Med frafallsmodell	Metode a	Metode b
Reelle data	103 590	103 072	103 552	103 438
Std. (1 000 sim.)	1 635	2 151	2 237	2 267
95% konf. int. (sim)	(100 606,106 795)	(98 840,107 200)	(98 374,107 446)	(98 637,107 560)

Det partielle frafallet på dette spørsmålet er på 17,4 prosent. Likevel blir estimatene med og uten frafallsmodell nesten like. Vi får litt høyere tall enn de offisielle.

6. Konklusjoner

Modell 1 og 2 for undersysselsetting gir omtrent like resultater. Disse gir en del høyere estimater når vi tar hensyn til frafall enn når ikke gjør det, men konfidensintervallene overlapper hverandre mye. Da vi prøvde en ignorerbar versjon av frafallsmodellen i modell 1, ble estimatene veldig like versjonen helt uten frafallsmodell.

Med modell 3 får vi veldig stor forskjell på estimater med og uten frafallsmodell, og versjonen med frafallsmodellering gir mye høyere estimater for undersysselsetting enn modell 1. Standardavvikene blir mindre enn for modell 1. Det er likevel overlapp mellom konfidensintervallene fra modell 1 og 3. Når vi tilpasser en ignorerbar versjon av modell 3, blir estimatene som tar hensyn til frafall mye mindre, og vi får en tabell som ligner mer på den tilsvarende tabellen for modell 1, om enn med noe høyere tall.

Alle de tre modellene for undersysselsetting gir høyere estimater enn det offisielle tallet på 77 000 undersysselsatte, og for modell 3 med frafallsmodellering er 77 000 utenfor konfidensintervallene. De ikke-ignorerbare versjonene av modellene gir også flere undersysselsatte enn de ignorerbare. Vi kan ikke vite sikkert hvilken av modellene som best beskriver virkeligheten, men det ser ut til at antall undersysselsatte har blitt underestimert, og at vi har ikke-ignorerbart frafall.

På spørsmålene 24, 25 og 62 fikk vi lite utslag av frafallsmodellen. På spørsmål 24 og 25 var det veldig lavt partielt frafall, så det kan være årsaken til at det ikke ble noen forskjell.

Når det gjelder imputeringsmetodene er metode a (trekke 0 eller 1 på hver person uavhengig av hverandre) kanskje litt bedre enn b (sette inn forventet antall ja-svar i hver gruppe). Den er en tanke lettere å implementere, og avrunding til nærmeste heltall i hver gruppe gjør at metode b ikke gir lavere varians enn metode a.

7. Referanser

[1] Jan Bjørnstad, "Imputering i AKU for undersysselsetting", Notater 2006/70

[2] Anne Vedø, Jenny-Anne Sigstad Lie og Jan Bjørnstad, "Statistisk modellering i AKU, modellstudier og modellestimering", Notater 2000/21

Vedlegg A: Imputeringssannsynligheter i modell 1 og 3

Sammenligning av imputeringssannsynlighetene og antall imputerte undersysselsatte for modell 1 og 3

I denne tabellen sammenligner vi sannsynligheten for å imputere ja på spørsmålet om undersysselsetting i modell 1 og 3 (impsannsmod1 og impsannsmod3). Vi har også sammenlignet antall imputerte undersysselsatte med metode b under de to modellene (Antimpmod1 og Antimpmod3). Modell 1 skiller ikke mellom de forskjellige frafallsmønstrene. Når vi oppgir antall imputerte undersysselsatte, har vi derfor slått sammen disse frafallsmønstrene. Tall med grå bakgrunn er totalen for de to grå gruppene. Tabellen inneholder bare de kombinasjonene som finnes i utvalget.

Imputeringssannsynlighetene er gjennomgående høyere for modell 3 enn modell1, og vi får imputert mange flere undersysselsatte.

Sivst	Aldgr	Kjønn	Region	R1	R2_18	R2_19	R2_23	Antall	Impsannsmod1	Impsannsmod3	Antimpmod1	Antimpmod3
0	1	0	1	1	0	0	0	5	0,297	0,622	1	3
0	1	0	2	1	0	0	0	10	0,122	0,167	1	2
0	1	0	3	1	0	0	0	11	0,172	0,327	2	4
0	1	0	4	1	0	0	0	4	0,114	0,134	0	1
0	1	0	5	1	0	0	0	4	0,218	0,297	1	1
0	1	1	1	1	0	0	0	1	0,208	0,549	0	1
0	1	1	2	1	0	0	0	8	0,179	0,334	1	3
0	1	1	3	1	0	0	0	4	0,202	0,450	1	2
0	1	1	4	1	0	0	0	2	0,158	0,322	0	1
0	1	1	5	1	0	0	0	4	0,215	0,444	1	2
0	2	0	1	1	0	0	0	9	0,475	0,749	4	7
0	2	0	2	1	0	0	0	5	0,390	0,682	2	4
0	2	0	3	1	0	0	0	9	0,371	0,684	3	6
0	2	0	4	0	0	0	0	1	0,422	0,582	0	1
0	2	0	4	1	0	0	0	4	0,436	0,592	2	3
0	2	0	5	1	0	0	0	3	0,413	0,511	1	2
0	2	1	1	1	0	0	0	11	0,187	0,436		
0	2	1	1	1	1	0	0	1	0,187	0,504	2	5
0	2	1	2	0	0	0	0	1	0,281	0,566	0	1
0	2	1	2	1	0	0	0	14	0,290	0,584		
0	2	1	2	1	1	1	0	1	0,290	0,927	4	9
0	2	1	3	0	0	0	0	1	0,222	0,486	0	1
0	2	1	3	1	0	0	0	7	0,228	0,504	2	4
0	2	1	4	0	0	0	0	2	0,305	0,525	1	1
0	2	1	4	1	0	0	0	5	0,314	0,535	2	3
0	2	1	5	0	0	0	0	2	0,212	0,338	0	1
0	2	1	5	1	0	0	0	1	0,220	0,347	0	0
0	3	0	1	1	0	0	0	1	0,327	0,769	0	1
0	3	0	2	1	0	0	0	3	0,206	0,543	1	2
0	3	0	3	1	0	0	0	3	0,218	0,606	1	2
0	3	0	4	1	0	0	0	2	0,228	0,462	0	1
0	3	0	5	1	0	0	0	1	0,241	0,621	0	1
0	3	1	2	1	0	0	0	2	0,171	0,420	0	1
0	3	1	3	0	0	0	0	1	0,142	0,377	0	1

0	3	1	3	1	0	0	0	3	0,147	0,396		
0	3	1	3	1	1	1	0	1	0,147	0,933	1	2
0	3	1	4	0	0	0	0	2	0,172	0,361	0	1
0	3	1	4	1	0	0	0	1	0,178	0,374	0	0
0	3	1	5	1	1	0	0	1	0,136	0,519	0	1
1	1	0	3	1	0	0	0	1	0,177	0,331	0	0
1	2	0	1	1	0	0	0	2	0,483	0,754		
1	2	0	1	1	1	1	0	1	0,483	0,962	1	3
1	2	0	2	1	0	0	0	2	0,400	0,690	1	2
1	2	0	3	1	0	0	0	3	0,380	0,690	1	2
1	2	0	4	1	0	0	0	1	0,446	0,595	0	1
1	2	0	5	1	0	0	0	4	0,424	0,514	2	2
1	2	1	1	1	0	0	0	6	0,190	0,441	1	3
1	2	1	2	0	0	0	0	2	0,290	0,578	1	1
1	2	1	2	1	0	0	0	26	0,295	0,590		
1	2	1	2	1	1	1	0	2	0,295	0,927	8	18
1	2	1	3	0	0	0	0	2	0,228	0,498	0	1
1	2	1	3	1	0	0	0	14	0,232	0,510	3	7
1	2	1	4	0	0	0	0	2	0,313	0,532	1	1
1	2	1	4	1	0	0	0	10	0,319	0,538	3	5
1	2	1	5	1	0	0	0	6	0,225	0,349		
1	2	1	5	1	1	0	0	1	0,225	0,375	2	3
1	3	0	1	1	0	0	0	2	0,334	0,779		
1	3	0	1	1	1	0	0	1	0,334	0,872	1	3
1	3	0	2	1	0	0	0	7	0,213	0,555	1	4
1	3	0	3	0	0	0	0	1	0,217	0,595	0	1
1	3	0	3	1	0	0	0	6	0,224	0,615		
1	3	0	3	1	1	1	0	1	0,224	0,980	2	5
1	3	0	4	1	0	0	0	1	0,235	0,467		
1	3	0	4	1	1	1	0	1	0,235	0,987	0	2
1	3	1	1	1	0	0	0	7	0,134	0,477	1	3
1	3	1	2	0	0	0	0	1	0,170	0,413	0	1
1	3	1	2	1	0	0	0	24	0,174	0,427		
1	3	1	2	1	1	1	0	1	0,174	0,943	4	11
1	3	1	3	1	0	0	0	16	0,150	0,402		
1	3	1	3	1	1	1	0	1	0,150	0,933	3	7
1	3	1	4	0	0	0	0	1	0,178	0,369	0	0
1	3	1	4	1	0	0	0	8	0,181	0,377	1	3
1	3	1	5	0	0	0	0	2	0,136	0,418	0	1
1	3	1	5	1	0	0	0	7	0,139	0,433		
1	3	1	5	1	1	0	0	1	0,139	0,519	1	4
Sum								325			72	175

Sannsynlighetene for å svare ja på undersysselsettingsspørsmålene i modell 3

Her angis, under modell3, den betingede sannsynligheten for at $Y_{18} = 1$ gitt partielt frafall på spørsmål 18 (betpy18), og de ubetingede sannsynlighetene for at $Y_{18} = 1$ (py18), $Y_{19} = 1$ (py19) og $Y_{23} = 1$ (py23). I tillegg angis $P(Y_u = 1)$ under modell 1 og 3 (pymod1 og pymod3).

Det er stor forskjell på den betingede og den ubetingede sannsynligheten for spørsmål 18. Det vil si at sannsynligheten for å ønske lengre arbeidstid gitt at man ikke har svart på spørsmål 18, er større enn den ubetingede sannsynligheten for å ønske lengre arbeidstid. Det er dette som gjør at vi får så høye imputeringssannsynligheter i modell 3. Det er ikke så stor forskjell på de ubetingede sannsynlighetene for å være undersysselsatt under modell 1 og 3, selv om modell 3 også her ligger noe høyere.

Sivst	Aldgr	Kjønn	Region	R1	betpy18	py18	py19	py23	pymod1	pymod3
0	1	0	1	1	0,907	0,405	0,696	0,986	0,220	0,277
0	1	0	2	1	0,918	0,463	0,183	0,992	0,088	0,084
0	1	0	3	1	0,913	0,439	0,362	0,990	0,123	0,157
0	1	0	4	1	0,921	0,453	0,146	0,994	0,081	0,066
0	1	0	5	1	0,941	0,550	0,321	0,985	0,163	0,174
0	1	1	1	1	0,819	0,217	0,706	0,950	0,145	0,145
0	1	1	2	1	0,920	0,430	0,374	0,972	0,125	0,156
0	1	1	3	1	0,879	0,317	0,530	0,966	0,142	0,162
0	1	1	4	1	0,898	0,354	0,366	0,978	0,109	0,127
0	1	1	5	1	0,918	0,421	0,510	0,950	0,153	0,204
0	2	0	1	1	0,943	0,533	0,826	0,962	0,376	0,424
0	2	0	2	1	0,916	0,457	0,761	0,979	0,306	0,341
0	2	0	3	1	0,924	0,476	0,759	0,974	0,286	0,352
0	2	0	4	0	0,937	0,592	0,631	0,984	0,345	0,368
0	2	0	4	1	0,954	0,592	0,631	0,984	0,345	0,368
0	2	0	5	1	0,954	0,613	0,556	0,962	0,329	0,328
0	2	1	1	1	0,865	0,281	0,576	0,876	0,129	0,142
0	2	1	2	0	0,876	0,383	0,697	0,927	0,212	0,248
0	2	1	2	1	0,904	0,383	0,697	0,927	0,212	0,248
0	2	1	3	0	0,845	0,312	0,630	0,914	0,162	0,180
0	2	1	3	1	0,876	0,312	0,630	0,914	0,162	0,180
0	2	1	4	0	0,911	0,448	0,610	0,944	0,230	0,258
0	2	1	4	1	0,929	0,448	0,610	0,944	0,230	0,258
0	2	1	5	0	0,902	0,443	0,429	0,875	0,158	0,166
0	2	1	5	1	0,924	0,443	0,429	0,875	0,158	0,166
0	3	0	1	1	0,882	0,341	0,899	0,971	0,245	0,297
0	3	0	2	1	0,849	0,303	0,651	0,983	0,152	0,194
0	3	0	3	1	0,882	0,357	0,700	0,980	0,159	0,245
0	3	0	4	1	0,905	0,402	0,517	0,987	0,168	0,205
0	3	0	5	1	0,873	0,344	0,733	0,970	0,181	0,245
0	3	1	2	1	0,848	0,268	0,526	0,943	0,119	0,133
0	3	1	3	0	0,791	0,241	0,510	0,933	0,101	0,115
0	3	1	3	1	0,832	0,241	0,510	0,933	0,101	0,115
0	3	1	4	0	0,845	0,300	0,447	0,956	0,124	0,128
0	3	1	4	1	0,874	0,300	0,447	0,956	0,124	0,128
0	3	1	5	1	0,822	0,232	0,575	0,901	0,094	0,120

1	1	0	3	1	0,923	0,439	0,362	0,990	0,123	0,157
1	2	0	1	1	0,949	0,533	0,826	0,962	0,376	0,424
1	2	0	2	1	0,927	0,457	0,761	0,979	0,306	0,341
1	2	0	3	1	0,933	0,476	0,759	0,974	0,286	0,352
1	2	0	4	1	0,959	0,592	0,631	0,984	0,345	0,368
1	2	0	5	1	0,960	0,613	0,556	0,962	0,329	0,328
1	2	1	1	1	0,874	0,281	0,576	0,876	0,129	0,142
1	2	1	2	0	0,894	0,383	0,697	0,927	0,212	0,248
1	2	1	2	1	0,913	0,383	0,697	0,927	0,212	0,248
1	2	1	3	0	0,865	0,312	0,630	0,914	0,162	0,180
1	2	1	3	1	0,886	0,312	0,630	0,914	0,162	0,180
1	2	1	4	0	0,923	0,448	0,610	0,944	0,230	0,258
1	2	1	4	1	0,934	0,448	0,610	0,944	0,230	0,258
1	2	1	5	1	0,931	0,443	0,429	0,875	0,158	0,166
1	3	0	1	1	0,893	0,341	0,899	0,971	0,245	0,297
1	3	0	2	1	0,867	0,303	0,651	0,983	0,152	0,194
1	3	0	3	0	0,867	0,357	0,700	0,980	0,159	0,245
1	3	0	3	1	0,895	0,357	0,700	0,980	0,159	0,245
1	3	0	4	1	0,915	0,402	0,517	0,987	0,168	0,205
1	3	1	1	1	0,782	0,169	0,676	0,902	0,089	0,103
1	3	1	2	0	0,833	0,268	0,526	0,943	0,119	0,133
1	3	1	2	1	0,860	0,268	0,526	0,943	0,119	0,133
1	3	1	3	1	0,844	0,241	0,510	0,933	0,101	0,115
1	3	1	4	0	0,864	0,300	0,447	0,956	0,124	0,128
1	3	1	4	1	0,882	0,300	0,447	0,956	0,124	0,128
1	3	1	5	0	0,806	0,232	0,575	0,901	0,094	0,120
1	3	1	5	1	0,836	0,232	0,575	0,901	0,094	0,120

Vedlegg B: Filstruktur på Unix

Programmene for del 3 av imputeringsprosjektet ligger på området $METODER/imput/prog/impest.

Området impest består av flg. 6 områder:

sp24, sp25, sp62, mod1, mod2, mod3

Områdene mod1, mod2 og mod3 gjelder undersysselsetting. De 6 områdene har samme struktur, så det holder å skissere strukturen for spørsmål 24. Modell 2 og 3 for undersysselsetting skiller seg litt fra spørsmål 24 noen steder. Der dette er tilfellet, har vi tatt med en egen beskrivelse av disse programmene og datafilene.

Området sp24 er delt i to områder: ekte og sim. På "ekte" ligger programmer og datafiler som har med det ekte datasettet å gjøre. På "sim" ligger tilsvarende programmer og datafiler for de 1000 simulerte datasettene.

B.1. Programmene og datafilene på sp24/ekte

B.1.1. Programmer

lags24.sas
Input: SAS-datasettet $METODER/imput/kvartal1.ssd01
Output: Flat fil s24.dat. Se beskrivelse under s24.dat.

imputab24.f:
Input: Et utvalg med partielt frafall, i dette tilfellet s24.dat. Utvalget må være på formen beskrevet under s24.dat. MLE for β-ene (populasjonsmodell) og ψ-ene (modell for R^2) som ligger på hhv. $METODER/imput/sp24/mlebeta24.dat og $METODER/imput/sp24/mlepsi24.dat.

Output: De fire filene sr24.dat, impa24.dat, impb24.dat og mlepytotmod24.dat. Sr24.dat inneholder bare svarutvalget. Impa24.dat og impb24.dat inneholder utvalget imputert på to forskjellige måter, og er på samme form som sr24.dat, beskrevet i avsnittet om datafiler.

De to imputeringsmetodene er:

a) Tilfeldig trekking fra estimert fordeling (impa24.dat)
b) Imputere forventet antall fra estimert fordeling (impb24.dat)

Mlepytotmod24.dat inneholder MLE for $P(Y = 1 \mid \mathbf{x})$, basert på total modell. Det er praktisk å skrive disse sannsynlighetene ut til fil her, siden vi likevel må regne dem ut i forbindelse med imputeringen.

mlepy24.f
Input: Utvalg uten partielt frafall, her sr24.dat, impa24.dat og impb24.dat. Utvalgene må være på formen beskrevet under sr24.dat.

Output: De fire filene mlesr24.dat, mlea24.dat og mleb24.dat. Disse inneholder MLE for $P(Y = 1 \mid \mathbf{x})$, basert på hvert av de komplette utvalgene.

lagantpop24.sas
Input: SAS-datasettene $METODER/imput/kvartal1.ssd01 og $METODER/imput/bebas/akupop.ssd01.

Output: Flat fil antpop24.dat. Inneholder antall personer i populasjonen, fordelt på strata (alder, kjønn, region).

I hvert stratum (alder, kjønn, region) anslår vi populasjonsantallet, dvs. totalt antall deltidssysselsatte med ønske om lengre arbeidstid. Dette gjøres ved å gange antall personer i BEBAS i stratumet med andelen av deltidssysselsatte med ønske om lengre arbeidstid i stratumet, der andelen regnes ut i utvalget.

predreg24.f
Input: De tre komplette utvalgene sr24.dat, impa24.dat og impb24.dat, og de fire filene med ML-estimater, mlesr24.dat, mlea24.dat, mleb24.dat og mlepytotmod24.dat. Filen antpop24.dat, som inneholder antall i populasjonen fordelt på strata (alder, kjønn, region).

Output: Filen predreg24.dat. Denne filen inneholder prediksjonsestimatoren og regresjonsestimatoren, regnet ut på sr24.dat (kombinert både med mlesr24.dat og mlepytotmod24.dat), impa24.dat og impb24.dat.

Modell 3 for undersysselsetting

imputabmod3.f
Her imputerer vi på spørsmål 18, 19 og 23 hver for seg. Først imputeres det verdier på de personene som ikke har svart på spørsmål 18. Ved imputering på spørsmål 19 bruker vi de som enten har svart ja eller har fått imputert ja på spørsmål 18 som målgruppe. De i målgruppen som ikke har svart på spørsmål 19 får en imputert verdi. På tilsvarende måte tar vi utgangspunkt i de som enten har svart ja på spørsmål 19 eller har fått imputert ja på spørsmål når vi imputerer på spørsmål 23. Når vi har imputert på alle tre spørsmål, har vi implisitt imputert på undersysselsettingsvariabelen Y_u, siden denne er en funksjon av Y_{18}, Y_{19} og Y_{23}.

mlepymod3.f
Output blir her $P(Y_{18}=1|\mathbf{x})$, $P(Y_{19}=1|Y_{18}=1,\mathbf{x})$ og $P(Y_{23}=1|Y_{18}=1,Y_{19}=1,\mathbf{x})$ istedenfor $P(Y_u=1|\mathbf{x})$

predregmod3.f
Vi tar utgangspunkt i utvalgene som er komplette mhp. på alle de tre undersysselsettingsvariablene, og utleder utvalg som er komplette med hensyn på Y_u. Vi finner $P(Y_u=1|\mathbf{x})$ fra $P(Y_{18}=1|\mathbf{x})$, $P(Y_{19}=1|Y_{18}=1,\mathbf{x})$ og $P(Y_{23}=1|Y_{18}=1,Y_{19}=1,\mathbf{x})$. Deretter beregnes regresjons- og prediksjonsestimatoren som før.

B.1.2. Datafiler

s24.dat: Det opprinnelige utvalget med frafall.
Form: Antall i stratum (k,u,v,w,r1,yobs) for k=0,1 (sivilstand) u=1,..., antald (alder) v=0,1 (kjønn) w=1,...5 (boregion) r1=0,1 (enhetsfrafall på 1. henv.) yobs=0,1,2 (observert verdi på spørsmålet), der 2 står for missing. Ett tall på hver linje. 2 x antald x 2 x 5 x 2 x 3 linjer.

sr24.dat: Svarutvalget.
Form: Antall i stratum (u,v,w,y) for u=1,..., antald (alder) v=0,1 (kjønn) w=1,...5 (boregion) y=0,1 (svar på spørsmålet). Ett tall på hver linje. antald x 2 x 5 x 2 linjer.

impa24.dat, impb24.dat: Imputert versjon av det opprinnelige utvalget, etter hhv. metode a og b.
Form: Som sr24.dat

mlesr24.dat, mlea24.dat, mleb24.dat: MLE for $P(Y=1 \mid \mathbf{x})$, basert på hhv. sr24.dat, impa24.dat og impb24.dat.
Form: Ett tall på hver linje, sortert på alder, kjønn og region i denne rekkefølgen.

mlepytotmod24.dat: MLE for $P(Y=1 \mid \mathbf{x})$, basert på total modell, altså med frafallsmodell.
Form: Ett tall på hver linje, sortert på alder, kjønn og region i denne rekkefølgen.

antpop24.dat: Antall personer i populasjonen, fordelt på strata (alder, kjønn, region). For spørsmål 24 er populasjonen deltidssysselsatte med ønske om lengre arbeidstid.
Form: Ett tall på hver linje, sortert på alder, kjønn og region i denne rekkefølgen.

predreg24.dat: Prediksjonsestimatoren og regresjonsestimatoren regnet ut på grunnlag av sr24.dat/mlesr24.dat, impa24.dat/mlea24.dat, impb24.dat/mleb24.dat, og sr24.dat/mlepytotmod24.dat
Form: En linje for hver kombinasjon av utvalg og MLE, i samme rekkefølge som over. To tall på hver linje.

Modell 2 for undersysselsetting

smod2.dat: Det opprinnelige utvalget med frafall.
Form: Antall i stratum (k,u,v,w,r1,yobs,r218,r219) for k=0,1 (sivilstand), u=1,...,3 (alder), v=0,1 (kjønn), w=1,...5 (boregion), r1=0,1 (enhetsfrafall på 1. henv.), yobs=0,1,2 (observert verdi på spørsmålet), der 2 står for missing, r218=0,1 og r219=0,1. Ikke alle kombinasjoner av yobs, r218 og r219 er mulige. Når yobs er 0 eller 1 er r218 og r219 uinteressante. Når yobs er 2, kan (r218, r219) ta verdiene (0,0), (1,0) og (1,1). Det blir dermed 5 mulige kombinasjoner av yobs, r218 og r219, som skrives ut i flg. rekkefølge: yobs=0, yobs=1, yobs=2 og (r218, r219)=(0,0), yobs=2 og (r218, r219)=(1,0), yobs=2 og (r218, r219)=(1,1). Fila har ett tall på hver linje, dvs. det blir 2 x 3 x 2 x 5 x 2 x 5 = 600 linjer.

Modell 3 for undersysselsetting

smod3.dat: Det opprinnelige utvalget med frafall.
Form: Data for spørsmål 18, 19 og 23 ligger etter hverandre. Først fordeles målgruppen til spørsmål 18 (deltidssysselsatte unntatt pensjonister) på strata, så fordeles målgruppen til spørsmål 19 (de som har svart ja på spørsmål 18), og til slutt fordeles målgruppen til spørsmål 23 (de som har svart ja på spørsmål 19).

Vi bruker her litt andre målgrupper enn de vi har definert tidligere. Pensjonister er fjernet fra målgruppen for spørsmål 18, og målgruppen for spørsmål 23 er bare de som har svart ja på spørsmål 19, i motsetning til tidligere alle som har svart ja på spørsmål 18. Dette er fordi det er disse målgruppene som er relevante i forbindelse med modell 3 for undersysselsetting.

For hvert spørsmål skriver vi ut antall i stratum (k,u,v,w,r1,yobs) for k=0,1 (sivilstand) u=1,...,3 (alder) v=0,1 (kjønn) w=1,...5 (boregion) r1=0,1 (enhetsfrafall på 1. henv.) yobs=0,1,2 (observert verdi på spørsmålet), der 2 står for missing. Ett tall på hver linje. Dette gir 2 x 3 x 2 x 5 x 2 x 3 = 360 linjer for hvert av de tre spørsmålene og totalt 1080 linjer.

srmod3.dat: Svarutvalget.
Form: Data for spørsmål 18, 19 og 23 ligger etter hverandre. For hvert spørsmål skriver vi ut antall i stratum (u,v,w,y) for u=1,..., 3 (alder) v=0,1 (kjønn) w=1,...5 (boregion) y=0,1 (svar på spørsmålet). Ett tall på hver linje. Dette gir 3 x 2 x 5 x 2 = 60 linjer for hvert spørsmål, og totalt 180 linjer.

mlesrmod3.dat, mleamod3.dat, mlebmod3.dat: MLE for $P(Y_{18}=1 \mid \mathbf{x})$, $P(Y_{19}=1 \mid Y_{18}=1, \mathbf{x})$ og $P(Y_{23}=1 \mid Y_{18}=1, Y_{19}=1, \mathbf{x})$, basert på srmod3.dat, impamod3.dat og impbmod3.dat.

Form: Data for et spørsmål av gangen, først 18, så 19 og til slutt 23. Ett tall på hver linje, sortert på alder, kjønn og region i denne rekkefølgen. Det er 30 strata, så vi får totalt 30 x 3 = 90 linjer.

mlepytotmodmod3.dat: MLE for $P(Y_{18}=1|\mathbf{x})$, $P(Y_{19}=1|Y_{18}=1,\mathbf{x})$ og $P(Y_{23}=1|Y_{18}=1,Y_{19}=1,\mathbf{x})$, basert på total modell, altså med frafallsmodell.
Form: Som mlesrmod3.dat, mleamod3.dat og mlebmod3.dat.

B.2. Programmene og datafilene på sp24/sim

Filene her er analoge med filene på sp24/ekte. Filnavnene er de samme, bare med "sim" som prefiks. Litt upresist kan vi si at programmene er de samme, bare satt inn i en løkke fra 1 til 1000, og at datafilene er like, bare med 1000 utgaver etter hverandre.

B.2.1. Programmer

lagsims24.f
Input: 1000 simulerte utvalg fra fila r2yr1.dat, som skal ligge på $METODER/imput/sp24/sim.
Output: Fila sims24.dat. Inneholder de samme 1000 simulerte utvalgene, men på en annen form. Se beskrivelse under sims24.dat.

simimputab24.f:
Input: 1000 utvalg med partielt frafall, i dette tilfellet sims24.dat. Utvalgene må være på formen beskrevet under sims24.dat. MLE for β-ene (populasjonsmodell), φ-ene (modell for R^1) og ψ-ene (modell for R^2) i de 1000 simulerte utvalgene. Dette ligger på $METODER/imput/sp24/sim/ml.dat.

Output: De fire filene simsr24.dat, simimpa24.dat, simimpb24.dat og simmlepytotmod24.dat. Simsr24.dat inneholder de 1000 svarutvalgene. Simimpa24.dat og simimpb24.dat inneholder de 1000 utvalgene imputert etter metode a og b, og er på formen beskrevet under simsr24.dat i avsnittet om datafiler. Simmlepytotmod24.dat inneholder MLE for $P(Y=1|\mathbf{x})$, basert på total modell, for de 1000 simulerte utvalgene.

simmlepy24.f
Input: 1000 utvalg uten partielt frafall, her simsr24.dat, simimpa24.dat og simimpb24.dat. Utvalgene må være på formen beskrevet under simsr24.dat.

Output: De tre filene simmlesr24.dat, simmlea24.dat og simmleb24.dat. Disse inneholder 1000 sett med MLE for $P(Y=1|\mathbf{x})$, basert på hvert av de fire settene med 1000 komplette utvalg.

simpredreg24.f
Input: De tre komplette utvalgssettene simsr24.dat, simimpa24.dat og simimpb24.dat, og de fire filene med ML-estimater, simmlesr24.dat, simmlea24.dat, simmleb24.dat og simmlepytotmod24.dat. Populasjonstotalene fra fila antpop24.f, på sp24/ekte.

Output: Filen simpredreg24.dat. Denne filen inneholder:

1) Gjennomsnitt og standardavvik for 1000 observasjoner av prediksjonsestimatoren
2) Gjennomsnitt og standardavvik for 1000 observasjoner av regresjonsestimatoren

regnet ut på utvalgssettene simsr24.dat (kombinert både med simmlesr24.dat og simmlepytotmod24.dat), simimpa24.dat og simimpb24.dat.

B.2.2. Datafiler

sims24.dat: De 1000 simulerte utvalgene med frafall.
Form: Hvert utvalg har samme form som s24.dat. De 1000 utvalgene ligger etter hverandre.

simsr24.dat: De 1000 simulerte svarutvalgene.
Form: Hvert utvalg har samme form som sr24.dat. De 1000 utvalgene ligger etter hverandre.

simimpa24.dat, simimpb24.dat: Imputerte versjoner av de 1000 simulerte utvalgene, etter hhv. metode a og b.
Form: Som simsr24.dat

simmlesr24.dat, simmlea24.dat, simmleb24.dat: 1000 sett av MLE for $P(Y = 1 \mid \mathbf{x})$, basert på hhv. simsr24.dat, simimpa24.dat og simimpb24.dat.
Form: Ett tall på hver linje, i samme rekkefølge som de tilsvarende filene på "ekte". De 1000 settene etter hverandre.

simmlepytotmod24.dat: 1000 sett av MLE for $P(Y = 1 \mid \mathbf{x})$, basert på den totale modellen med frafall.
Form: Ett tall på hver linje, i samme rekkefølge som de tilsvarende filene på "ekte". De 1000 settene etter hverandre.

simpredreg24: Gjennomsnitt og standardavvik for prediksjonsestimatoren og regresjonsestimatoren, regnet ut på grunnlag av simsr24.dat/simmlesr24.dat, simimpa24.dat/simmlea24.dat, simimpb24.dat/simmleb24.dat og simsr24.dat/simmlepytotmod24.dat.
Form: Fire tall på hver linje (gj.snitt pred, st.avvik pred, gjsnitt reg, st.avvik reg), en linje for hver kombinasjon av utvalg og MLE, i samme rekkefølge som over.

Filstruktur på UNIX

Gjelder for spørsmål 24. De andre spørsmålene er analogt konstruert. Hvis ingen sti er nevnt gjelder $METODER/imput/prog/impest/sp24/ekte.