Statistics Norway
Statistisk sentralbyrå

Anne Vedø, Jenny-Anne Sigstad Lie
og Jan Bjørnstad

Statistisk modellering i AKU
Modellstudier og modellestimering

Avdeling for samordning og utvikling/Seksjon for statistiske metoder og standarder
Emnegruppe: 00.90

Forord

Dette notatet er andre del av imputeringsprosjektet. Første del omhandler nåværende imputeringsrutiner i AKU, og utarbeides av Jan Bjørnstad og Jenny-Anne Sigstad Lie. Neste del tar for seg modellbasert imputering og estimering i AKU, med utgangspunkt i resultatene fra dette notatet. Planen er å avslutte med et notat om variansestimering i AKU.

Jan Bjørnstad har vært prosjektleder, og har skrevet delkapittel 3.1.3 om implisitt modell for responsmekanismen med nåværende imputeringsmetoder i AKU.

1. Innledning

I enhver utvalgsundersøkelse vil det være frafall, endel data man ikke har fått tak i. Det er flere tiår siden man ble klar over hvilke feil frafall kan resultere i for en statistisk analyse. Likevel er det først de senere år problemet har fått nødvendig oppmerksomhet blant statistikere.
Som i mange andre land, har svarandelene i Norge avtatt de senere år. I tillegg har man hatt et problem med manglende dokumentasjon av de imputeringsmetoder som er i bruk. Dette er bakgrunnen for imputeringsprosjektet.

I første omgang har vi valgt å bruke data fra arbeidskraftundersøkelsen (AKU), 1. kvartal 1992. Dette utvalget består av 23 923 personer, hvorav 21 900 har svart.
I 1992 trakk man AKU-utvalg som i grove trekk kan sies å være selvveiende, to-trinns utvalg. Man hadde en undersøkelsesuke hver måned.
Fra og med 1996 har man trukket et stratifisert, tilnærmet selvveiende ett-trinnsutvalg. Hver uke av året er undersøkelsesuke.

I AKU ønsker man å presentere tall for totaler (evt. andeler), f.eks. totalt antall undersysselsatte i populasjonen. Når vi skal estimere en total, og det er frafall, innebærer det at vi må trekke slutninger om elementer som er utenfor svarutvalget. Til det skal vi her bruke en modelltilnærming. Ifølge likelihood-prinsippet skal da analysen foretas gitt det observerte bruttoutvalget, basert på den underliggende populasjonsmodell og eventuelt en modell for responsmekanismen (RM). Det vil si at vi ikke kommer til å ta hensyn til utvalgsplanen i vår analyse.

Det er to typer frafall:
- enhetsfrafall, som innebærer at intervjuobjektet (IO) ikke svarer på et eneste spørsmål
- partielt frafall, som innebærer at IO lar være å svare på enkelte av spørsmålene.

AKU er en obligatorisk undersøkelse, dvs. at de som blir trukket ut til å være med, er pliktige å svare. Det gjør at andelen med enhetsfrafall blir lav. Endel av intervjuene i AKU er *indirekte intervjuer*, det vil si at IO ikke har vært tilgjengelig, og at enten IO's ektefelle/samboer eller IO's foreldre har svart for IO. Enkelte spørsmål skal ikke stilles ved indirekte intervju. Det gjelder blant annet spørsmålene om undersysselsetting og spørsmålene om ønsket arbeidstid for undersysselsatte og for arbeidssøkere uten arbeidsinntekt. Slike spørsmål vil ha et større partielt frafall enn spørsmål som kan stilles både ved direkte og indirekte intervju.

Dette notatet tar for seg modellstudier og modellestimeringer for seks utvalgte spørsmål med partielt frafall i AKU. Vi ser på utprøving av forskjellige modeller for populasjon og frafall (responsmekanisme), og maksimum likelihood estimering i modellen. Deretter kan vi studere den totale modell, inkludert frafall for alle spørsmål. Til å vurdere kvaliteten på estimeringsmetoden og modellens estimerbarhet, bruker vi simulering. Det gjøres en sammenligning med dagens implisitte «modell».

De seks spørsmålene vi tar for oss er uforandret etter omleggingen av AKU i 1996, og spørsmål om undersysselsetting og ønsket arbeidstid skal fremdeles ikke stilles ved indirekte intervju. Metoden vi bruker er generell, og kan også benyttes på andre spørsmål enn de seks vi har valgt ut.

Et senere notat vil omhandle imputering og estimering i AKU, basert på modeller som er valgt i dette notatet.

En kort oversikt over innholdet i de ulike kapitler og appendiks:

Kapittel 1 gir en oversikt over hvilke spørsmål i AKU som skal modelleres.

I **Kapittel 2** utarbeides en foreløpig populasjonsmodell ved univariate og multivariate analyser.

I **Kapittel 3** ser vi først på modellering av responsmekanismen. Deretter vurderes effekten av å tilføye frafallsmodeller. Vi tester modelltilpasningen etter at frafallsparametrene er tatt med. Kapitlet avsluttes med en gjennomgang av forslag til justering av modellen.

Appendiks A inneholder likelihoodfunksjonen for de aktuelle spørsmålene, og for undersysselsetting.

Appendiks B består av tabeller over maksimumlikelihoodestimatorene (MLE) for $P(Y=1|\mathbf{x})$, $P(R^1=1|\mathbf{z})$ og $P(R^2=1|\mathbf{z},r^1,y)$.

Appendiks C består av tabeller over gjennomsnitt og standardavvik av MLE for hver parameter, fra 1000 simuleringer.

Appendiks D inneholder en kort beskrivelse av programmer som har vært kjørt.

1.1. Spørsmål i AKU som skal modelleres

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. Tabell 1.1.2 viser antall personer i utvalget som skal svare, som har svart og som ikke har svart på hvert enkelt spørsmål.

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

19: 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: 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: Ønsket arbeidstid for deltidssysselsatte med ønske om lengre arbeidstid

Y=1 hvis ønsket arbeidstid $\geq$ 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: Faktisk arbeidstid når IO har ett arbeidsforhold
Y=1 hvis faktisk arbeidstid $\geq$ 37 timer
Y=0 hvis faktisk arbeidstid $\geq$ 0 og < 37 timer

62: Ønsket arbeidstid for arbeidsledige
Y=1 hvis ønsket arbeidstid $\geq$ 37 timer
Y=0 hvis ønsket arbeidstid > 0 og < 37 timer

De seks spørsmålene over skal stilles til ulike delgrupper av AKU-utvalget. Personer som skal svare på spørsmål j kaller vi målgruppen for spørsmål j . Målgruppene for de forskjellige spørsmålene er:

Spørsmål 18: Deltidssysselsatte

Spørsmål 19: Deltidssysselsatte med ønske om lengre arbeidstid

Spørsmål 23a: Deltidssysselsatte med ønske om lengre arbeidstid

Undersysselsetting: Deltidssysselsatte

Spørsmål 24: Deltidssysselsatte med ønske om lengre arbeidstid

Spørsmål 25: Personer med ett arbeidsforhold

Spørsmål 62: Arbeidssøkere

Vi kan i første omgang dele utvalget opp i to grupper:

1. Personer som har levert skjema, enten ved første henvendelse eller etter oppfølgning.
2. Personer som ikke har levert skjema i det hele tatt.

Av det totale utvalget på 23 923 personer er 21 900 i gruppe en, mens de resterende 2 023, dvs. omtrent 8,5 prosent, er i gruppe to. I den første gruppen kjenner vi (stort sett) hvilken målgruppe en person tilhører. Dette er ikke tilfellet i den andre gruppen. Når en person ikke har levert skjema, kan vi ikke vite om han/hun for eksempel er sysselsatt i ett arbeidsforhold. Dette gjør at vi har valgt å se bort fra gruppe to. All videre analyse foregår på personer fra gruppe en.

I modelleringen kommer vi til å bruke registervariable som forklaringsvariable. Noen få personer i utvalget mangler en eller flere av de aktuelle registervariablene, og disse kan derfor ikke brukes i analysen. I tabell 1.1.1 har vi satt opp en oversikt over antall personer i hele utvalget og antall personer som er brukt i analysen for gruppene definert i venstre kolonne.

Tabell 1.1.1

	Hele utvalget	Brukt i analysen
Svart på 1. henvendelse	20 897	20 888
Svart etter oppfølging	1 003	1 003
Svart	21 900	21 891
Ikke svart	2 023	0
Totalt	23 923	21 891

De 21 891 personene som har svart på skjemaet og som ikke mangler noen av registervariablene, deler vi videre etter målgruppe. Antall personer i målgruppen for hvert spørsmål står oppført i kolonne 2 i tabell 1.1.2 under.

I modellene kommer vi til å bruke aldersgruppe som en forklaringsvariabel, og en av aldersgruppene vil være pensjonister, dvs. personer mellom 67 og 74 år. Av grunner vi kommer nærmere inn på senere, er pensjonistene ikke brukt i analysen på spørsmål 19, 23a, 24 og 62, og heller ikke på spørsmålet om undersysselsetting sett under ett. Kolonnene 3-5 i tabell 1.1.2 inneholder, for hvert spørsmål, antall personer i målgruppen, i svarutvalget og i det partielle frafallet som er brukt i analysen.

Tabell 1.1.2

	Svart på skjema og har alle register-variable	**Brukt i analysen**		
	Målgruppe	**Målgruppe**	**Svart på spørsmål**	**Partielt frafall**
Sp 18	3 887	3 887	3 559	328
Sp 19	955	950	945	5
Sp 23a	955	950	908	42
Undersysselsetting	3 887	3 757	3 432	325
Sp 24	955	950	928	22
Sp 25	13 281	13 281	13 098	183
Sp 62	868	863	713	150

2. Populasjonsmodellering med generell populasjonsmodell

2.1. Logistiske modeller

Som en innledende modellvurdering for de seks utvalgte spørsmålene, samt den samlede undersysselsettingsvariabelen, vil vi først basere oss bare på svarutvalgene, og anta at disse er representative for populasjonen. Det vil si:
Gitt at enhet *i* er i svarutvalget, så har Y_i samme fordeling som den antatte populasjonsmodellen.

Ved å gjøre en slik antakelse, kan vi få et visst inntrykk av hva slags populasjonsmodeller vi skal vurdere når modeller for RM skal inkluderes.

For hvert spørsmål j og for hver enhet i i målgruppen for spørsmål j, definerer vi altså en 0/1-variabel Y_{ij}. For spørsmål 25 for eksempel, får vi

$$Y_{i,25} = \begin{cases} 1 \text{ hvis person } i \text{ har faktisk arb.tid} \geq 37 \text{ timer} \\ 0 \text{ ellers} \end{cases}$$

Med 0/1-variable er det vanlig å velge logistisk regresjon. Generelt er en logistisk modell på formen

$$\ln \frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

der $\mathbf{x} = (x_1, \ldots, x_n)$ er forklaringsvariable, $\pi(\mathbf{x}) = P(Y = 1 | \mathbf{x})$ og $\beta_0, \ldots \beta_n$ er ukjente parametre. Med forklaringsvariable mener vi både opprinnelige forklaringsvariable og eventuelle transformasjoner eller funksjoner av disse. Transformasjoner det er vanlig å vurdere er potenser av de opprinnelige forklaringsvariablene (f.eks. x^2, x^3) og kryssledd, dvs. produkter av de opprinnelige forklaringsvariablene (f.eks. $x_1 x_2$, $x_2 x_3$, $x_1 x_2 x_3$).

Forklaringsvariablene kan være enten nominelle eller ordinale. Ordinale variable tar verdier som kan ordnes på en meningsfull måte, som for eksempel aldersgrupper. For nominelle variable finnes det ingen meningsfull ordning av verdiene. Dette gjelder f.eks. kjønn og sivilstand.

Nominelle forklaringsvariable kodes på en annen måte enn ordinale. For å kode en nominell forklaringsvariabel x_j med k_j verdier (kategorier) innfører vi $k_j - 1$ forklaringsvariable $D_{j1}, \ldots, D_{j,k_j-1}$. D_{jl} er en indikatorvariabel for kategori l, $l = 1, \ldots k_j - 1$, dvs. at $D_{jl} = 1$ for personer i kategori l, 0 ellers. En logistisk modell der x_j er nominell vil dermed kunne skrives

$$\ln \frac{\pi(\mathrm{x})}{1-\pi(\mathrm{x})} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \sum_{l=1}^{k_j-1} \beta_{jl} D_{jl} + \cdots + \beta_n x_n$$

En av kategoriene, her k_j, har ingen korresponderende indikatorvariabel. Personer i denne kategorien kan likevel identifiseres, fordi de er de eneste som har $D_{jl} = 0$ for $l = 1, \ldots, k_j - 1$. For personer i kategori k_j blir leddet $\sum_{l=1}^{k_j-1} \beta_{jl} D_{jl}$ lik null, og k_j kalles derfor referansekategori eller referansegruppe. Det er ikke nødvendig å velge den øverste kategorien som referansegruppe. Hvis vi f.eks. ønsker å bruke kategori 2 som referansgruppe, bruker vi indikatorvariablene $D_{j1}, D_{j3}, \ldots, D_{j,k_j}$ istedenfor $D_{j1}, \ldots, D_{j,k_j-1}$.

Kryssledd mellom to nominelle forklaringsvariable x_1 og x_2, kodes ved å multiplisere alle indikatorvariablene for x_1 med alle indikatorvariablene for x_2. Hvis x_1 og x_2 har henholdsvis k_1 og k_2 kategorier, gir kryssleddet dermed opphav til $(k_1 - 1)(k_2 - 1)$ nye forklaringsvariable og parametre.

Ved valg av populasjonsmodell er det flere spørsmål vi må ta stilling til:

- Hvilke forklaringsvariable som skal være med i modellen
- Hvilke variable som skal behandles som nominelle, og hvilke som ordinale
- Hvilke transformasjoner av forklaringsvariablene skal være med, herunder
  - Høyere ordens ledd av ordinale variable
  - Kryssledd

- Hvor mange kategorier hver forklaringsvariabel skal deles inn i
- Om vi skal bruke samme modell på alle spørsmålene

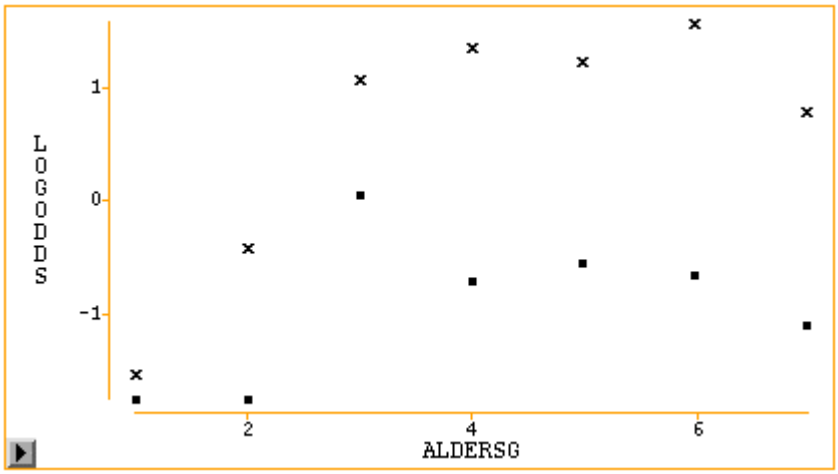
Når det gjelder det første spørsmålet, har vi i første omgang valgt å se på alder, kjønn og bostedsregion som aktuelle forklaringsvariable. Kjønn og bostedsregion betraktes som nominelle variable.

Vi vurderte først om alder kunne brukes som en ordinal variabel, dvs. om det var mulig å dele inn og ordne aldersgrupper slik at logodds av $\pi(\mathbf{x})$, $\ln\{\pi(\mathbf{x})/(1-\pi(\mathbf{x}))\}$, kunne antas å være lineær som funksjon av aldersgruppeinndelingen for *alle spørsmålene* av interesse. Spørsmål 25 ble først betraktet med følgende definisjon og ordning av syv aldersgrupper:

1: 16-19 år, 2: 67-74 år, 3: 20-24 år, 4: 25-29 år, 5: 30-39 år, 6: 40-54 år, 7: 55-66 år

Empirisk logodds for region R, kjønn K og aldersgruppe A, er gitt ved $\ln(p(R,K,A)/(1-p(R,K,A)))$, der $p(R,K,A)$ er andelen med faktisk arbeidstid større eller lik 37 timer blant personer i region R og aldersgruppe A med kjønn K i svarutvalget. Ved å plotte, for hver region og hvert kjønn, empirisk logodds mot aldersgruppe, ble det klart at den ikke er en lineær funksjon i alder i noen av strataene. Som illustrasjon vises plottet for en region i figur 1. Menn er merket med kryss, kvinner med firkant.

Figur 1 Aust- og Vestagder/Rogaland



Deretter ble aldersgruppene 3-7 slått sammen slik at det ble tre aldersgrupper i alt. Lineariteten ble bedre nå, samtidig som stigningskoeffisienten i aldersvariabelen tydelig var større for menn enn kvinner. Det ble derfor prøvd med et kryssledd mellom alder og kjønn. Denne modellen ble utprøvd på alle seks spørsmålene, både med elleve og åtte regioner. (Om valg av regioner i neste avsnitt). Det viste seg at for de fleste spørsmål fungerte modellen rimelig bra, men for spørsmål 18 var empirisk logodds for andelen med y = 1 klart ikke-lineær. Figur 2 viser plottet for en region.

Figur 2 Buskerud/Telemark

Det er ønskelig med samme type modell for alle spørsmålene, og det gjorde at vi besluttet å betrakte aldersgruppe også som en nominell variabel. Forskjellige kryssledd ble også utprøvd, noe som forutsatte en reduksjon i antall regioner. Ellers vil kryssledd som involverer region gi opphav til for mange nye parametre.

2.2. Modellevaluering

I dette delkapittelet vil vi avgjøre

- Hvilken inndeling i regioner vi vil ha
- Hvilke aldersgrupper vi skal operere med
- Hvilke kryssledd som skal være med i modellen
- Om vi skal bruke en felles modell for alle spørsmålene

En mulighet er å spesialtilpasse både regioninndeling, inndeling i aldersgrupper og antall kryssledd for hvert enkelt spørsmål. På den måten kan man oppnå en optimal modelltilpasning på hvert spørsmål. Ulempen er at det er upraktisk å operere med mange forskjellige modeller. I tillegg må vi huske på at den endelige modellvurderingen ikke kan gjøres før responsmekanismen er trukket inn i modellen. Vårt mål nå er derfor å finne en felles populasjonsmodell som, basert på informasjonen fra svarutvalget, er akseptabel for alle spørsmålene. For å oppnå dette, tar vi først for oss ett og ett spørsmål, og tilpasser en «beste» modell etter metoden beskrevet i Hosmer & Lemeshow (1989). Deretter sammenholder vi resultatene, og finner en kompromissløsning. Denne felles modellen kaller vi en prøvemodell, og den blir vårt utgangspunkt når vi skal inkludere responsmekanismen.

Etter å ha prøvet oss fram med inndeling i ulike regioner, valgte vi til slutt å bruke inndelingen i fem landsdeler fra SØS 33 (Thomsen, 1991), for alle spørsmålene. Regionene er dermed:

1. Oslo / Akershus
2. Resten av Østlandet
3. Sørlandet / Vestlandet unntatt Møre og Romsdal
4. Møre og Romsdal / Trøndelag
5. Nord-Norge

For hvert spørsmål forsøker vi både med tre og fire aldersgrupper. Inndelingen i tre aldersgrupper er de unge (16-19 år), de eldste (67-74 år) og resten (20-66 år). Ved inndelingen i fire aldersgrupper har vi delt den største gruppen i to, 20-39 år og 40-66 år. På spørsmålene 19, 23a, 24 og 62 er det ingen, eller nesten ingen av de eldste som har svart. Dette betyr at datasettet inneholder for lite informasjon

til å estimere parametre for denne aldersgruppen, og de utelates derfor fra analysene for disse spørsmålene. Her står derfor valget mellom to og tre aldersgrupper. På spørsmålet om undersysselsetting sett under ett, er det en del av de eldste som har svart, men ingen har svart at de er undersysselsatte. Den empiriske sannsynligheten for å være undersysselsatt er altså null i den eldste aldersgruppen. Dette fører til problemer med den logistiske modellen, fordi ln(p/(1-p)) går mot minus uendelig når p går mot null. Vi har derfor utelatt aldersgruppen 67-74 år i analysen av undersysselsetting, slik at valget også her står mellom to og tre aldersgrupper.

For hvert spørsmål lager vi fire tabeller: Univariat analyse, Multivariat analyse av hovedeffekter, Kryssledd og Foreløpig modell. Målet med tabellene er å finne signifikansen, eller P-verdien, til de aktuelle forklaringsvariablene, både hovedeffekter og kryssledd. Siden vi har så få som tre hovedeffekter, har vi valgt å beholde alle hovedeffektene i modellen, selv om de ikke blir signifikante i den univariate eller multivariate analysen av hovedeffekter. Av kryssleddene krever vi derimot en viss signifikans for at de skal bli tatt med i den foreløpige modellen.

I utregningen av P-verdier refererer vi ofte litt upresist til log likelihood til en modell. Som kjent er likelihoodfunksjonen en funksjon av parametrene i modellen. Med log likelihood til en modell mener vi max log likelihood, dvs. log likelihooden evaluert i maksimumlikelihood-estimatet for parametrene.

Et annet begrep vi trenger å kjenne, er antall frihetgrader til en variabel. Med dette mener vi antall parametre tilknyttet variabelen. For kategoriske hovedeffekter er dette antall kategorier minus en. For kryssledd mellom to kategoriske variable multipliseres frihetsgradene til variablene som inngår i kryssleddet.

I de neste avsnittene gir vi en nærmere beskrivelse av hva tabellene inneholder.

2.2.1. Univariat analyse

Vi tilpasser en logistisk modell med konstantledd og en variabel av gangen. Først har vi bare med konstantledd, så konstantledd og kjønn, så konstantledd og tre aldersgrupper osv. Fra teorien for logistisk regresjon har vi at dersom man utvider en modell med en variabel så vil, hvis modell uten variabel er korrekt,

G = 2(logl(modell med variabel) - logl(modell uten variabel))

være kji-kvadratfordelt med antall frihetsgrader lik antall frihetsgrader til variabelen som ble lagt til.
For hver variabel regner vi ut log likelihood til modellen med bare denne variabelen og konstantledd, og G regnes ut som

G = 2(logl(modell med konstantledd og variabel) - logl(modell med bare konstantledd))

Vi regner også ut P-verdien p, som er sannsynligheten for at G skal være større eller lik den beregnede G, gitt at modell uten variabel er sann. Hvis p er liten betyr det at variabelen er signifikant, dvs. at den bidrar til å forklare interessevariabelen.

I en univariat analyse får vi vite hvor stor forklaringskraft hver variabel har alene. I analyser med mange potensielle forklaringsvariable, kan det være aktuelt å utelukke variable med høy P-verdi i den univariate analysen fra videre analyse. Vi har imidlertid bare kjønn, alder og region som aktuelle forklaringsvariable, og disse anser vi som såpass viktige at det ikke er aktuelt å utelukke dem fra modellen, selv om de skulle komme ut med en høy P-verdi. P-verdiene for kjønn, alder og region tjener derfor hovedsakelig til orientering.

Det reelle valget som gjøres i den univariate analysen, er valget mellom tre eller fire (evt. to eller tre) aldersgrupper. Her er den interessante størrelsen

G = 2(logl(modell med fire(tre) aldersgrupper) - logl(modell med tre(to) aldersgrupper))

Denne størrelsen er, gitt modell med tre(to) aldersgrupper, kji-kvadratfordelt med en frihetsgrad. Dette er fordi modellen med fire(tre) aldersgrupper kan ses på som en utvidelse av modellen med tre(to) aldersgrupper. Vi har lagt til en ny variabel som indikerer om personen er i øvre eller nedre halvdel av aldersgruppen 20-66 år, for øvrig er indikatorvariablene like i de to modellene. P-verdien til den ekstra aldersgruppen står oppgitt under tabellen med den univariate analysen. Hvis denne er liten, fortsetter vi analysen med den ekstra aldersgruppen.

2.2.2. Multivariat analyse av hovedeffekter

Vi tilpasser en logistisk modell med alle hovedeffekter samtidig, dvs. konstantledd, kjønn, region (5 kategorier) og aldersgruppe (det antall kategorier som ble besluttet i den univariate analysen). Vi regner så ut signifikansen til en variabel V ved å se på

G = 2(logl(modell med hovedeffekter) - logl(modell med hovedeffekter unntatt V))

Under modellen med hovedeffekter unntatt V er G tilnærmet kji-kvadratfordelt, med like mange frihetsgrader som V.

I den multivariate analysen forteller P-verdien til en variabel hvor mye variabelen forklarer sammen med de andre variablene, dvs. i tillegg til de andre variablene. Som i den univariate analysen får P-verdiene for hovedeffektene ingen konsekvenser for modellvalget.

Under tabellen setter vi opp log likelihood for modellen og P-verdier for Pearsons kjikvadratobservator og Deviansen. De to siste er forskjellige mål på hvor godt modellen passer til dataene. Pearsons kjikvadratobservator er

$$\chi^2 = \sum_{i=1}^{m}\sum_{j=0}^{1}(r_{ij} - n_i p_{ij})^2 / n_i p_{ij} \qquad (2.2.1)$$

der

m er antall forskjellige kovariatvektorer
i er en indeksering av mengden av kovariatvektorer
r_{i0} er antall personer med kovariatvektor i som har svart nei på spørsmålet
r_{i1} er antall personer med kovariatvektor i som har svart ja på spørsmålet
n_i er antall personer med kovariatvektor i
p_{i1} er den estimerte sannsynligheten for at en person med kovariatvektor i svarer ja på spørsmålet
p_{i0} er $1 - p_{i1}$

Deviansen er gitt ved

$$D = 2\sum_{i=1}^{m}\sum_{j=0}^{1} r_{ij} \ln(r_{ij} / n_i p_{ij})$$

Denne størrelsen kan også defineres ved

D = 2(logl(mettet modell) - logl(nåværende modell)) (2.2.2)

der en mettet modell betyr en modell med like mange parametre som det er forskjellige kovariatvektorer, dvs. m. D sier altså noe om hvor god tilpasningen er i forhold til den beste tilpasningen det er mulig å oppnå med de forklaringsvariablene vi har valgt.

Hvis modellen er riktig er både χ^2 og D tilnærmet kji-kvadratfordelt med $m-q$ frihetsgrader, der q er antall parametre i modellen. Vi oppgir P-verdiene for disse testene. Jo større P-verdien er, jo bedre er modelltilpasningen. En P-verdi på rundt 0,10 eller over regnes som akseptabelt.

2.2.3. Kryssledd

Vi tilpasser modeller med hovedeffekter og et kryssledd av gangen. For hvert kryssledd beregnes

G = 2(logl(modell med hovedeffekter og kryssledd) - logl(modell med bare hovedeffekter))

Gitt at modellen med hovedeffekter er korrekt, så er G tilnærmet kji-kvadratfordelt med like mange frihetsgrader som kryssleddet.

Ved å se på P-verdiene finner vi ut hvilke kryssledd det er aktuelt å ta med. Vi har som hovedregel å ta med kryssledd med P-verdi under 0,10.

2.2.4. Foreløpig modell

Tilslutt tilpasser vi en modell med det antall aldersgrupper og de kryssleddene vi har kommet fram til tidligere, og oppgir P-verdiene for de forskjellige variablene i denne modellen. Vi kaller dette en foreløpig modell, fordi den kan bli endret både når vi skal finne en felles modell for alle spørsmål, og etter at responsmekanismen er inkludert. P-verdien for variabel V regnes ut på grunnlag av

G = 2(logl(foreløpig modell) - logl(foreløpig modell unntatt V)),

som, gitt foreløpig modell unntatt V, er tilnærmet kji-kvadratfordelt med like mange frihetsgrader som V.

Vi oppgir også P-verdiene for Pearsons kjikvadratobservator og Deviansen for den foreløpige modellen.

2.2.5. Spørsmål 18. Ønske om lengre arbeidstid for deltidssysselsatte

Det er 3 560 deltidssysselsatte som har svart på spørsmål 18. For en av disse mangler regionvariabelen. Analysene for spørsmål 18 er derfor basert på 3 559 personer.

Univariat analyse

Variabel	Logl (konstant og variabel)	Antall frihetsgrader	G	p
Kjønn	-2 064,4311	1	11,5526	0,0007
Aldersgruppe (3)	-2 049,1508	2	42,1132	< 0,0001
Aldersgruppe (4)	-2 026,1057	3	88,2034	< 0,0001
Region	-2 059,7131	4	21,6528	0,0002

Log likelihood for modell med bare konstantledd: -2 070,2074
P-verdi for ekstra aldersgruppe: <0,0001

Alle variablene er signifikante i den univariate analysen. Inndelingen i fire aldersgrupper er også signifikant bedre enn inndelingen i tre. Vi velger derfor å bruke fire aldersgrupper på dette spørsmålet.

Multivariat analyse av hovedeffekter

Variabel	Logl (hovedeffekter unntatt variabel)	Antall frihetsgrader	G	p
Kjønn	-2 014,4326	1	15,7018	< 0,0001
Aldersgruppe (4)	-2 053,2274	3	93,2914	< 0,0001
Region	-2 018,4273	4	23,6912	< 0,0001

Log likelihood for modell med hovedeffekter: -2 006,5817
Pearson: 0,3884
Devians: 0,2966

Alle variablene blir høyst signifikante også i den multivariate analysen.

Kryssledd

Kryssledd	Logl (hovedeffekter og kryssledd)	Antall frihetsgrader	G	p
Alder*Kjønn	-2 004,0997	3	4,9640	0,1744
Alder*Region	-2 002,2456	12	8,6722	0,7306
Kjønn*Region	-2 002,6810	4	7,8014	0,0991

Kryssleddet mellom kjønn og region er signifikant på 10 % nivå. Dette betyr at regionvariabelen påvirker menn og kvinner forskjellig med hensyn til å ønske fulltidsstilling når man har deltidsstilling. Med dette leddet blir også Pearson og Deviansen veldig gode, så vi tar dette leddet med i modellen.

Foreløpig modell

Variabel	Logl (foreløpig modell unntatt variabel)	Antall frihetsgrader	G	p
Kjønn	-2 004,4933	1	3,6246	0,0569
Aldersgruppe (4)	-2 048,1075	3	90,8530	< 0,0001
Region	-2 016,3292	4	27,2964	< 0,0001
Kjønn*Region	-2 006,5817	4	7,8014	0,0991

Log likelihood for foreløpig modell: -2 002,6810
Pearson: 0,6002
Devians: 0,4703

2.2.6. Spørsmål 19. Forsøk på å få lengre arbeidstid for deltidssysselsatte med ønske om lengre arbeidstid

Det er bare 5 personer i aldersgruppen 67-74 år som har svart på spørsmål 19. Dette er for få til å kunne behandles som en gruppe. Vi har derfor fjernet disse personene. Analysene under er basert på 945 observasjoner.

Univariat analyse

Variabel	Logl (konstant og variabel	Antall frihetsgrader	G	p
Kjønn	-640,1851	1	2,1140	0,1460
Aldersgruppe (2)	-636,2193	1	10,0456	0,0015
Aldersgruppe (3)	-632,4871	2	17,5100	0,0002
Region	-635,4809	4	11,5224	0,0213

Log likelihood for modell med bare konstantledd: -641,2421
P-verdi for ekstra aldersgruppe: 0,0063

Tre aldersgrupper er signifikant bedre enn to, så vi fortsetter med tre aldersgrupper.

Multivariat analyse av hovedeffekter

Variabel	Logl (hovedeffekter unntatt variabel)	Antall frihetsgrader	G	p
Kjønn	-625,6141	1	3,8668	0,0493
Aldersgruppe (3)	-634,7686	2	22,1758	< 0,0001
Region	-630,1460	4	12,9306	0,0116

Log likelihood for modell med hovedeffekter: -623,6807
Pearson: 0,0611
Devians: 0,0404

Kryssledd

Kryssledd	Logl (hovedeffekter og kryssledd	Antall frihetsgrader	G	p
Alder*Kjønn	-621,6059	2	4,1496	0,1256
Alder*Region	-616,3016	8	14,7582	0,0640
Kjønn*Region	-621,7075	4	3,9464	0,4133

Kryssleddet mellom alder og region er signifikant på 10 % nivå. Det hever også Pearson fra 0,0611 til 0,1863, så vi tar det med i den foreløpige modellen.

Foreløpig modell

Variabel	Logl (foreløpig modell unntatt variabel)	Antall frihetsgrader	G	p
Kjønn	-618,8322	1	5,0612	0,0245
Aldersgruppe (3)	-617,1811	2	1,7590	0,4150
Region	-620,5626	4	8,5220	0,0742
Alder*Region	-623,6807	8	14,7582	0,0640

Log likelihood for foreløpig modell: -616,3016
Pearson: 0,1863
Devians: 0,1279

2.2.7. Spørsmål 23a. Hvor raskt IO kan starte med økt arbeidstid for deltidssysselsatte med ønske om lengre arbeidstid

Personer i aldersgruppen 67-74 år er fjernet. Analysene er basert på 908 personer.

Univariat analyse

Variabel	Logl (konstant og variabel)	Antall frihetsgrader	G	p
Kjønn	-267,2340	1	2,2540	0,1334
Aldersgruppe (2)	-268,3059	1	0,1102	0,7399
Aldersgruppe (3)	-266,9330	2	2,8560	0,2398
Region	-266,1234	4	4,4752	0,3455

Log likelihood for modell med bare konstantledd: -268,361
P-verdi for ekstra aldersgruppe: 0,0975

Vi bruker tre aldersgrupper videre i analysen.

Multivariat analyse av hovedeffekter

Variabel	Logl (hovedeffekter unntatt variabel)	Antall frihetsgrader	G	p
Kjønn	-264,7962	1	3,6348	0,0566
Aldersgruppe (3)	-264,4848	2	3,0120	0,2218
Region	-265,6388	4	5,3200	0,2560

Log likelihood for modell med hovedeffekter: 262,9788
Pearson: 0,7995
Devians: 0,6242

Kryssledd

Kryssledd	Logl (hovedeffekter og kryssledd)	Antall frihetsgrader	G	p
Alder*Kjønn	-261,1575	2	3,6426	0,1682
Alder*Region	-260,2681	8	5,4214	0,7117
Kjønn*Region	-261,8699	4	2,2178	0,6958

Her er ingen av kryssleddene signifikante på 10 % nivå, så den foreløpige modellen inneholder bare hovedeffekter.

Foreløpig modell

Variabel	Logl (foreløpig modell unntatt variabel)	Antall frihetsgrader	G	p
Kjønn	-264,7962	1	3,6348	0,0566
Aldersgruppe (3)	-264,4848	2	3,0120	0,2218
Region	-265,6388	4	5,3200	0,2560

Log likelihood for foreløpig modell: 262,9788
Pearson: 0,7995

Devians: 0,6242

2.2.8. Undersysselsetting

Personer i aldersgruppen 67-74 år er fjernet. Analysene er basert på 3 432 personer.

Univariat analyse

Variabel	Logl (konstant og variabel)	Antall frihetsgrader	G	p
Kjønn	-1432,2122	1	20,3458	<0,0001
Aldersgruppe(2)	-1441,8214	1	1,1274	0,2883
Aldersgruppe(3)	-1420,7387	2	43,2928	<0,0001
Region	-1440,2117	4	4,3468	0,3611

Log likelihood for modell med bare konstantledd: -1442,3851
P-verdi for ekstra aldersgruppe: <0,0001

Variablene kjønn og alder (3 grupper) er signifikante i den univariate analysen. Inndelingen i 2 aldersgrupper er ikke signifikant, og vi velger å bruke 3 aldersgrupper når vi ser på undersysselsetting.

Multivariat analyse av hovedeffekter

Variabel	Logl (hovedeffekter untatt variabel)	Antall frihetsgrader	G	p
Kjønn	-1418,3886	1	22,1644	<0,0001
Aldersgruppe(3)	-1429,6988	2	44,7848	<0,0001
Region	-1409,8965	4	5,1802	0,2693

Log likelihood for modell med hovedeffekter: -1407,3064
Pearson: 0,1511
Devians: 0,1567

Også i den multivariate analysen er det kjønn og alder (3 grupper) som blir signifikante.

Kryssledd

Kryssledd	Logl (hovedeffekter og kryssledd)	Antall frihetsgrader	G	p
Alder*Kjønn	-1405,4822	2	3,6484	0,1614
Alder*Region	-1404,9534	8	4,7060	0,7884
Kjønn*Region	-1402,9606	4	8,6916	0,0693

Kryssleddet mellom kjønn og region er signifikant på 10 % nivå. Med dette leddet blir også Pearson god, så vi tar leddet med i modellen.

Foreløpig modell

Variabel	Logl (foreløpig modell untatt variabel)	Antall frihetsgrader	G	p
Kjønn	-1425,0572	2	44,1932	<0,0001
Aldersgruppe(3)	-1403,8731	1	1,8250	0,1767
Region	-1405,0869	4	4,2526	0,3729
Kjønn*Region	-1407,3064	4	8,6916	0,0693

Loglikelihood foreløpig modell: -1402,9606
Pearson: 0,4086
Devians: 0,1938

2.2.9. Spørsmål 24. Ønsket arbeidstid for deltidssysselsatte med ønske om lengre arbeidstid
Personer i aldersgruppen 67-74 år er fjernet. Analysene er basert på 928 personer.

Univariat analyse

Variabel	Logl (konstant og variabel)	Antall frihetsgrader	G	p
Kjønn	-629,0527	1	26,8192	< 0,0001
Aldersgruppe (2)	-615,6644	1	53,5958	< 0,0001
Aldersgruppe (3)	-612,1316	2	60,6614	< 0,0001
Region	-639,0022	4	6,9202	0,1402

Log likelihood for modell med bare konstantledd: -642,4623
P-verdi for ekstra aldersgruppe: 0,0079

Tre aldersgrupper er signifikant bedre enn to, så vi går videre med tre aldersgrupper.

Multivariat analyse av hovedeffekter

Variabel	Logl (hovedeffekter unntatt variabel)	Antall frihetsgrader	G	p
Kjønn	-609,0436	1	53,9912	< 0,0001
Aldersgruppe (3)	-626,1109	2	88,1258	< 0,0001
Region	-583,8752	4	3,6544	0,4548

Log likelihood for modell med hovedeffekter: -582,0480
Pearson: 0,0018
Devians: 0,0086

Kryssledd

Kryssledd	Logl (hovedeffekter og kryssledd)	Antall frihetsgrader	G	p
Alder*Kjønn	-574,3052	2	15,4856	0,0004
Alder*Region	-578,9855	8	6,1250	0,6332
Kjønn*Region	-580,6153	4	2,8654	0,5806

Kryssleddet mellom alder og kjønn er svært signifikant, og fører til en betydelig forbedring av modelltilpasningen . Dette betyr at det er stor forskjell på menn og kvinner når det gjelder hvordan alder påvirker ønsket arbeidstid.

Foreløpig modell

Variabel	Logl (foreløpig modell unntatt variabel)	Antall frihetsgrader	G	p
Kjønn	-583,5642	1	18,5180	< 0,0001
Aldersgruppe (3)	-582,6356	2	16,6608	0,0002
Region	-575,6280	4	2,6456	0,6188
Alder*Kjønn	-582,0480	2	15,4856	0,0004

Log likelihood for foreløpig modell: -574,3052
Pearson: 0,1394
Devians: 0,1887

2.2.10. Spørsmål 25. Faktisk arbeidstid når IO har ett arbeidsforhold

13 103 personer svarte på spørsmål 25. Fem av disse har uoppgitt region. Analysene er derfor basert på 13 098 personer.

Univariat analyse

Variabel	Logl (konstant og variabel)	Antall frihetsgrader	G	p
Kjønn	-7 950,6956	1	1 816,9962	< 0,0001
Aldersgruppe (3)	-8 616,3667	2	485,6540	< 0,0001
Aldersgruppe (4)	-8 616,3445	3	485,6984	< 0,0001
Region	-8 835,0075	4	48,3724	< 0,0001

Log likelihood for modell med bare konstantledd: -8 859,1937
P-verdi for ekstra aldersgruppe: 0,7241

Her er alle forklaringsvariablene sterkt signifikante. Det blir ingen vesentlig bedring med fire aldersgrupper i forhold til tre, så vi går videre med tre aldersgrupper.

Multivariat analyse av hovedeffekter

Variabel	Logl (hovedeffekter unntatt variabel)	Antall frihetsgrader	G	p
Kjønn	-8 590,7647	1	1 915,8558	< 0,0001
Aldersgruppe (3)	-7 917,0804	2	568,4872	< 0,0001
Region	-7 672,7624	4	79,8512	< 0,0001

Log likelihood for modell med hovedeffekter: -7 632,8368
Pearson: 0,0000
Devians: 0,0001

Kryssledd

Kryssledd	Logl (hovedeffekter og kryssledd)	Antall frihetsgrader	G	p
Alder*Kjønn	-7 626,2734	2	13,1268	0,0014
Alder*Region	7 626,1525	8	13,3686	0,0998
Kjønn*Region	7 620,4782	4	24,7172	< 0,0001

Her er både kryssleddet mellom alder og kjønn og kryssleddet mellom kjønn og region sterkt signifikante. Modelltilpasningen er imidlertid veldig dårlig i begge modellene (både Pearson og Deviansen er under 0,05 for begge modellene). Når vi tar med begge kryssleddene får vi en god modelltilpasning. Vi får seks parametre i tillegg til parametrene fra hovedeffektene, men dette skulle ikke by på problemer siden det er over 13 000 personer som har svart på spørsmål 25. Følgelig tar vi begge kryssleddene med i den foreløpige modellen.

Foreløpig modell

Variabel	Logl (foreløpig modell unntatt variabel)	Antall frihetsgrader	G	p
Kjønn	-7 672,0485	1	116,0510	< 0,0001
Aldersgruppe (3)	-7 674,5428	2	121,0396	< 0,0001
Region	-7 653,8976	4	79,7492	< 0,0001
Alder*Kjønn	-7 620,4782	2	12,9104	0,0016
Kjønn*Region	-7 626,2734	4	24,5008	< 0,0001

Log likelihood for foreløpig modell: -7 614,0230
Pearson: 0,4246
Devians:0,3302

2.2.11. Spørsmål 62. Ønsket arbeidstid for arbeidsledige

Personer i aldersgruppen 67-74 år er fjernet. Analysene er basert på 713 personer.

Univariat analyse

Variabel	Logl (konstant og variabel)	Antall frihetsgrader	G	p
Kjønn	-338,6339	1	81,9710	< 0,0001
Aldersgruppe (2)	-340,5236	1	78,1916	< 0,0001
Aldersgruppe (3)	-338,8211	2	81,5966	< 0,0001
Region	-368,6064	4	22,0260	0,0002

Log likelihood for modell med bare konstantledd: 379,6194
P-verdi for ekstra aldersgruppe: 0,0650

Vi fortsetter med tre aldersgrupper.

Multivariat analyse av hovedeffekter

Variabel	Logl (hovedeffekter unntatt variabel)	Antall frihetsgrader	G	p
Kjønn	-329,6533	1	87,9216	< 0,0001
Aldersgruppe (3)	-327,6155	2	83,8460	< 0,0001
Region	-295,1173	4	18,8496	0,0008

Log likelihood for modell med hovedeffekter: 285,6925
Pearson: 0,0024
Devians: 0,0041

Kryssledd

Kryssledd	Logl (hovedeffekter og kryssledd)	Antall frihetsgrader	G	p
Alder*Kjønn	-280,0007	2	11,3836	0,0034
Alder*Region	-281,4526	8	8,4798	0,3881
Kjønn*Region	-280,7358	4	9,9134	0,0419

Her har vi også to signifikante kryssledd. Vi tilpasset derfor en modell med begge disse leddene i tillegg til hovedeffektene, og endte opp med en akseptabel verdi på Pearson.

Foreløpig modell

Variabel	Logl (foreløpig modell unntatt variabel)	Antall frihetsgrader	G	p
Kjønn	-278,8753	1	9,2380	0,0024
Aldersgruppe (3)	-286,6455	2	24,7784	< 0,0001
Region	-278,9654	4	9,4182	0,0515
Alder*Kjønn	-280,7358	2	12,9590	0,0015
Kjønn*Region	-280,0007	4	11,4888	0,0216

Log likelihood for foreløpig modell: -274,2563
Pearson: 0,2966
Devians: 0,1944

2.2.12. Oppsummering

Vi lager en oversikt over hvilke aldersgrupper og hvilke kryssledd vi kom fram til på hvert spørsmål.

Spørsmål	Aldersinndeling	Kryssledd
18	Fineste	Kjønn*Region
19	Fineste	Alder*Region
23a	Fineste	Ingen
Undersysselsetting	Fineste	Kjønn*Region
24	Fineste	Alder*Kjønn
25	Groveste	Alder*Kjønn, Kjønn*Region
62	Fineste	Alder*Kjønn, Kjønn*Region

Fineste aldersinndeling betyr at gruppen 20-66 år er splittet i to, 20-39 og 40-66.

Alle spørsmål unntatt spørsmål 25 modelleres best med den fineste inndelingen av aldersgrupper. Kryssleddet mellom alder og kjønn er med på tre av spørsmålene. Kryssleddet mellom kjønn og region er med på tre av spørsmålene, og i tillegg på spørsmålet om undersysselsetting sett under ett. Alder*Region er bare med på ett spørsmål.

Vi ønsker å undersøke om vi kan oppnå en akseptabel felles modell for alle spørsmål ved å bruke den fineste aldersinndelingen og alle tre kryssledd, evt. to kryssledd (alder*kjønn og kjønn*region).

I tabellen under har vi funnet P-verdiene for Pearson (øverst) og Deviansen (nederst) for modellen med fineste aldersinndeling og to kryssledd og for modellen med fineste aldersinndeling og alle kryssleddene. I kolonnen lengst til høyre har vi til sammenligning satt opp P-verdiene for den foreløpige modellen.

Spørsmål	**2 kryssledd**	**3 kryssledd**	**Foreløpig modell**
18	0,5536 0,5654	0,3828 0,3184	0,6002 0,4703
19	0,0935 0,0663	0,2885 0,2742	0,1863 0,1279
23a	0,6208 0,5938	0,5602 0,4823	0,7995 0,6242
Undersyssel-setting	0,5423 0,2684	0,1941 0,0638	0,4086 0,1938
24	0,0083 0,1158	0,0071 0,0166	0,1394 0,1887
25	0,1273 0,0946	0,0883 0,0626	0,4246 0,3302
62	0,2966 0,1944	0,5110 0,4617	0,2966 0,1944

Modellen med tre kryssledd gir akseptable verdier på Pearson for spørsmålene 18, 19, 23a og 62, og for spørsmålet om undersysselsetting. Problemet med denne modellen er at vi får dårlig tilpasning på spørsmål 24 og 25.

Hovedproblemet på spørsmål 25 er ikke antall kryssledd, men antall aldersgrupper. Dette er det eneste spørsmålet der den groveste aldersinndelingen passer best. Hvis vi tilpasser en modell med groveste aldersinndeling (tre grupper) og tre kryssledd, blir Pearson og Deviansen på henholdsvis 0,7527 og 0,6287. Dette gir altså enda bedre tilpasning enn den foreløpige modellen med groveste aldersinndeling og to kryssledd.

På spørsmål 24 er de to kryssleddene som inneholder region ikke særlig signifikante. Når disse leddene blir lagt til, reduseres Pearsons kjikvadratobservator relativt lite i forhold til den foreløpige modellen (fra 26,85 til 21,03). Antall frihetsgrader reduseres derimot fra 20 til 8. Dette gjør at P-verdien blir mye mindre.

Alle modeller vil bli vurdert igjen etter at responsmekanismen er tatt med, så vi bestemmer oss for å bruke modellen med tre kryssledd og fineste aldersinndeling som felles prøvemodell, med unntak av spørsmål 25. På spørsmål 25 bruker vi tre kryssledd og groveste aldersinndeling. Resultatet blir dermed:

Alder (x_1) deles inn i følgende grupper:

1: 16-19 år
2: 20-39 år

3: 40-66 år
4: 67-74 år

På spørsmål 25 slår vi sammen gruppe 2 og 3, og på spørsmål 19, 23a , 24 og 62, og på spørsmålet om undersysselsetting, utelater vi personer i gruppe 4 fra analysen. Alder behandles som en nominell variabel, og kodes med indikatorvariable D_{11}, D_{12}, D_{14} (spørsmål 18) eller D_{11}, D_{12} (spørsmål 19, 23a, 24 og 62 og spørsmålet om undersysselsetting) eller D_{11}, D_{14} (spørsmål 25). Aldersgruppen 40-66 år er referansegruppe (20-66 år for spørsmål 25).

Variabelen kjønn (x_2) er kodet 0 for menn og 1 for kvinner.

Bostedsregion (x_3) er delt inn i følgende grupper:

1: Oslo, Akershus
2: Resten av Østlandet
3: Sørlandet, Vestlandet (unntatt Møre og Romsdal)
4: Møre og Romsdal, Trøndelag
5: Nord-Norge

Bostedregion kodes med fire indikatorvariable $D_{31}, D_{33}, D_{34}, D_{35}$. Region 2 er referansegruppe.

La $\mathbf{x} = (x_1, x_2, x_3)$ og $\pi(\mathbf{x}) = P(Y = 1|\mathbf{x})$.

Prøvemodellen vår er:

$$\ln\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})} = \beta_0 + \sum_{l=1,2,4}\beta_{1l}D_{1l} + \beta_2 x_2 + \sum_{l=1,3,4,5}\beta_{3l}D_{3l} + \sum_{l=1,2,4}\beta_{4l}x_2 D_{1l} + \sum_{l=1,3,4,5}\beta_{5l}x_2 D_{3l} + \sum_{l=1,2,4}\sum_{m=1,3,4,5}\beta_{6lm}D_{1l}D_{3m} \qquad (2.2.3)$$

For spørsmål 19, 23a, 24 og 62, og for spørsmålet om undersysselsetting, må indeks 4 utelates fra summene som inneholder D_{1l}, og for spørsmål 25 må indeks 2 utelates fra de samme summene.

3. Modellering av populasjon og frafall

3.1. Modellering av responsmekanismen

Når frafall inntreffer må vi anta en modell for responsmekanismen. Om IO returnerer skjemaet ved første henvendelse eller ikke, antar vi er uavhengig av svarene han måtte ha på spørsmålene. Det vil si vi antar *ignorerbar responsmekanisme* for de som har enhetsfrafall ved første henvendelse. I vårt utvalg er det 20 897 personer som svarte ved første henvendelse, og 1 003 som svarte etter oppfølgning.

I forbindelse med *partielt frafall*, vil vi lage en modell for hvert av de spørsmålene som har størst partielt frafall. Vi antar *ikke-ignorerbar responsmekanisme* ved partielt frafall, dvs. at hvorvidt en person svarer eller ikke på et spørsmål, avhenger av hvilket svar han har på spørsmålet.

I AKU har vi et begrenset antall registervariable. I tillegg til de tre som er nevnt under populasjonsmodellen, har vi sivilstand. Vi antar at registervariablene kjønn, bosted (region) og

sivilstand er med på å avgjøre både om personen svarer på et bestemt spørsmål, og om han har enhetsfrafall ved første henvendelse. Sivilstand påvirker enhetsfrafall ved at enslige som regel er i jobb, og derfor kan være vanskeligere å få tak i. Alder er ikke tatt med ved modellering av responsmekanismen. Dette er fordi vi må forsøke å begrense antall forklaringsvariable, og vi tror sivilstand har mer å si for frafallssannsynligheten enn alder.

Sivilstand omfatter egentlig fire grupper: ugifte, gifte, samboere og før gifte. For ikke å få for mange strata, har vi valgt å redusere til to grupper:

0: aleneboende (ugifte og før gifte)
1: ikke-aleneboende (gifte og samboere)

Vi bruker altså følgende tilleggsvariabler i RM-modellen:

$z_1 = \text{sivilstand}$
$z_2 = x_2 = \text{kjønn}$
$z_3 = x_3 = \text{bostedregion}$

$\mathbf{z} = (z_1, z_2, z_3)$

3.1.1. Enhetsfrafall ved første henvendelse

For hver enhet i utvalget defineres en tilfeldig variabel:

$$R_i^1 = \begin{cases} 1 \text{ hvis person } i \text{ svarer ved første henvendelse} \\ 0 \text{ hvis person } i \text{ svarer etter oppfølgning} \end{cases}$$

Vi antar ignorerbar stratifisert responsmekanisme, og velger en logistisk modell.

$$\ln \frac{P(R^1 = 1 \mid \mathbf{z}, y)}{P(R^1 = 0 \mid \mathbf{z}, y)} = \varphi_0 + \varphi_1 z_1 + \varphi_2 z_2 + \sum_{l=1,3,4,5} \varphi_{3l} D_{3l} + \varphi_4 z_1 z_2 + \sum_{l=1,3,4,5} \varphi_{5l} z_1 D_{3l} + \sum_{l=1,3,4,5} \varphi_{6l} z_2 D_{3l} \tag{3.1.1}$$

Som i populasjonsmodellen har vi tatt med tre kryssledd. Vi har brukt SAS til å tilpasse denne modellen, og vi får en meget god tilpasning. Under har vi også satt opp signifikansnivået til de forskjellige variablene i modellen. Siden R^1 er en variabel som er relevant for alle målgrupper, er denne analysen basert på alle personer som har levert skjemaet og som ikke mangler verdi på noen av forklaringsvariablene, dvs. 21 891 personer.

Variabel	**Logl (modell unntatt variabel)**	**Antall frihetsgrader**	**G**	**p**
Kjønn	-4 009,6068	1	0,8194	0,3654
Sivilstand	-4 010,4892	1	2,5842	0,1079
Region	-4 021,1086	4	23,8230	0,0001
Sivilstand*Kjønn	-4 009,7753	1	1,1564	0,2822
Sivilstand*Region	-4 012,9757	4	7,5572	0,1092
Kjønn*Region	-4 009,8613	4	1,3284	0,8565

Log likelihood: -4 009,1971
Pearson: 0,8206
Devians: 0,8205

3.1.2. Partielt frafall

For hver person i målgruppen for spørsmål j , og hvert spørsmål j defineres en tilfeldig variabel:

$$R_{ij}^2 = \begin{cases} 1 \text{ hvis person } i \text{ svarer på spm. } j \\ 0 \text{ ellers} \end{cases}$$

Vi vil forsøke forskjellige måter å modellere R^2 -ene på. Først betrakter vi spørsmålene hver for seg, og modellerer R^2 -ene uavhengig av hverandre. Deretter betrakter vi spørsmålene sammen, og vurderer tre ulike modeller.

Spørsmålene betraktet hver for seg

Vi antar en ikke-ignorerbar responsmekanisme, så vi tar med y som forklaringsvariabel i tillegg til registervariablene kjønn, sivilstand og region. Vi antar også at om en person svarer på et spørsmål, avhenger av om han svarte med en gang, eller om han først svarte etter purring. Vi forsøker med en logistisk modell uten kryssledd, og vi bruker samme modell på alle spørsmål.

$$\ln \frac{P(R^2 = 1 \mid \mathbf{z}, y, R^1)}{P(R^2 = 0 \mid \mathbf{z}, y\ , R^1)} = \psi_0 + \psi_1 z_1 + \psi_2 z_2 + \sum_{l=1,3,4,5} \psi_{3l} D_{3l} + \psi_4 y + \psi_5 R^1 \qquad (3.1.2)$$

For hvert spørsmål er altså den samlede modellen for populasjon og responsmekanisme gitt ved (2.2.3), (3.1.1) og (3.1.2).

Spørsmålene betraktet sammen

I prinsippet er det mulig å tenke seg en modell der alle spørsmålene er sett i sammenheng, og der all relevant informasjon fra et spørsmål blir utnyttet i modelleringen av partielt frafall på de etterfølgende spørsmålene. I praksis blir en slik modell for komplisert, så vi har valgt å se de tre spørsmålene som til sammen utgjør spørsmålet om undersysselsetting (spørsmål 18, 19 og 23a) i sammenheng, mens spørsmålene 24, 25 og 62 betraktes som selvstendige spørsmål.

Vi undersøker tre forskjellige modeller.

La Y_u være indikatorvariabel for undersysselsetting, og R_u^2 indikatorvariabel for svar på spørsmål om undersysselsetting, dvs.

$Y_u = 1$ dersom IO er undersysselsatt
$Y_u = 0$ dersom IO ikke er undersysselsatt

Som nevnt tidligere regnes man som undersysselsatt hvis og bare hvis man svarer ja på både spørsmål 18, 19 og 23a. Svarer man nei på ett eller flere av spørsmålene, er man altså ikke undersysselsatt. Y_u er dermed en funksjon av Y_{18}, Y_{19} og Y_{23}.

$R_u^2 = 1$ dersom IO har svart på spørsmålet om undersysselsetting

$R_u^2 = 0$ dersom IO ikke har svart på spørsmålet 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. I klartekst betyr dette 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, dvs. dersom IO har unnlatt å svare på minst ett av spørsmålene, og samtidig har svart ja på de spørsmålene han har svart på, er $R_u^2 = 0$. R_u^2 er dermed en funksjon av Y_{18}, Y_{19}, Y_{23}, R_{18}^2, R_{19}^2, R_{23}^2.

Modell 1

Vi betrakter Y_u og R_u^2 på samme måte som vi betraktet Y_j og R_j^2 da vi modellerte spørsmålene hver for seg. Vi modellerer altså Y_u etter (2.2.3), R^1 etter (3.1.1) og R_u^2 gitt Y_u, R^1 etter (3.1.2). 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.

$$\ln \frac{P(R_u^2 = 1 \mid \mathbf{z}, y_u, R^1)}{P(R_u^2 = 0 \mid \mathbf{z}, y_u, R^1)} = \psi_0 + \psi_1 z_1 + \psi_2 z_2 + \sum_{l=1,3,4,5} \psi_{3l} D_{3l} + \psi_4 y_u + \psi_5 R^1$$

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. Det er rimelig å anta at en person som har svart ja både på spørsmål 18 og 19, men ikke har svart på spørsmål 23a, oftere er undersysselsatt enn en person som hverken har svart på spørsmål 18, 19 eller 23a. Når man skal imputere en verdi for Y_u blant personer som ikke har svart på spørsmålet om undersysselsetting, bør man derfor ideelt sett imputere $Y_u = 1$ oftere blant personer som har svart ja på spørsmål 18 og 19 enn blant personer som ikke har svart på noen av spørsmålene. Dette lar seg ikke gjøre under modell 1. Vi må legge $P(Y_u = 1 | R_u^2 = 0)$ til grunn for imputeringen uansett frafallsmønster. Vi vil dermed antageligvis imputere for få undersysselsatte blant dem som har svart ja på spørsmål 18 og 19, og for mange blant dem som ikke har svart på noen av spørsmålene.

Motivasjonen for modell 2 under, er nettopp å gjøre modellen avansert nok til å kunne skille mellom de ulike frafallsmønstrene.

Modell 2

For å utvide modell 1 til en modell som gjør det mulig å skille mellom frafallsmønstre, er den mest nærliggende ideen å modellere de enkelte frafallsvariablene R_{18}^2, R_{19}^2 og R_{23}^2, samt Y_u. Dette gjør vi i modell 2a under. Her gjør vi også rede for en del vanskeligheter som viser seg å være forbundet med denne modellen. Disse problemene leder oss til å konstruere modell 2b, som blir den versjonen av modell 2 som vi går videre med.

Modell 2a

Vi modellerer her R_{18}^2, R_{19}^2 og R_{23}^2 etter (3.1.2), men med y_u istedenfor y som forklaringsvariabel. Y_u modelleres etter (2.2.3). De enkelte variablene Y_{18}, Y_{19} og Y_{23} modelleres ikke. Modellene gjelder for deltidssysselsatte.

Modell for R_{18}^2:

Her har vi en felles modell for alle deltidssysselsatte.

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

Modell for R_{19}^2 :

Her deler vi de deltidssysselsatte i to grupper avhengig av verdien på R_{18}^2.

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 $R_{18}^2=1$ bruker vi (3.1.2):

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

Kommentar: Da vi definerte variabelen R_{19}^2 gjorde vi det kun for personer i målgruppen for spørsmål 19, dvs. deltidssysselsatte med ønske om lengre arbeidstid ($Y_{18}=1$). Dette er fordi partielt frafall på spørsmål 19 bare gir mening for personer som faktisk har blitt spurt/burde ha blitt spurt spørsmål 19. En person som ikke har svart på spørsmål 19 fordi han/hun har svart nei på spørsmål 18, kan ikke sies å ha partielt frafall på spørsmål 19. Personen har da fylt ut skjemaet korrekt, det er avgjort at personen ikke er undersysselsatt, og det er ikke aktuelt å imputere noe svar på spørsmålet om undersysselsetting.

I den nåværende modellen ser vi at R_{19}^2 er definert for alle deltidssysselsatte, dvs. R_{19}^2 har en verdi (enten 0 eller 1) for alle deltidssysselsatte. Dette er fordi vi ikke har mulighet til å skille ut målgruppen for spørsmål 19 når variabelen Y_{18} ikke er med i modellen. Når R_{19}^2 er null, kan det altså enten dreie seg om "ekte" partielt frafall på spørsmål 19, dvs. at personen er i målgruppen for spørsmål 19, men likevel har unnlatt å svare på spørsmål 19, eller det kan være en person som er utenfor målgruppen, og derfor ikke har blitt stilt spørsmål 19.

Modell for R_{23}^2:

Her skiller vi mellom de deltidssysselsatte som har svart på spørsmål 19 (og derfor også på spørsmål 18), og de som ikke har svart på spørsmål 19.

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

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

For personer som har svart på spørsmål 18 og 19 bruker vi (3.1.2).

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

I modell 2a komprimerer vi informasjonen om Y_{18}, Y_{19} og Y_{23} sammen til Y_u, men vi tar vare på opplysningene om partielt frafall for hvert spørsmål. Det er nå lett å få inntrykk av at variablene som modelleres i modell 2a (dvs. $R_{18}^2, R_{19}^2, R_{23}^2, Y_u$) inneholder ekte mer informasjon enn variablene i den enkle modell 1 (dvs. R_u^2, Y_u). Slik er det imidlertid ikke, for det finnes tilfeller der (R_u^2, Y_u) ikke lar seg utlede fra ($R_{18}^2, R_{19}^2, R_{23}^2, Y_u$). Eksempel:

Anta at $(R_{18}^2, R_{19}^2, R_{23}^2, Y_u) = (1,0,0,0)$. For å unngå forvirring minner vi nå om at Y_u står for personens faktiske verdi på undersysselsettingsvariabelen, og ikke den observerte verdien. Y_u er alltid enten 0 eller 1, mens den observerte verdien i tillegg kan være missing.

Hvis personen har svart nei på spørsmål 18, er spørsmålet om undersysselsetting avgjort, og vi får $R_u^2 = 1$. Hvis personen har svart ja på spørsmål 18, er spørsmålet om undersysselsetting ikke avgjort, og vi får $R_u^2 = 0$. Altså er R_u^2 ukjent i dette tilfellet. Det samme problemet dukker opp i tilfellet $(R_{18}^2, R_{19}^2, R_{23}^2, Y_u) = (1,1,0,0)$.

Det er en klar svakhet ved modell 2a at en så sentral størrelse som R_u^2 ikke lar seg uttrykke ved hjelp av variablene som er med. Når vi senere skal imputere for undersysselsetting er det jo kun blant personer som ikke har svart på spørsmålet om undersysselsetting vi ønsker å imputere. Ideen med å utvide modell 1 til modell 2a, var å kunne imputere $Y_u = 1$ med forskjellig sannsynlighet avhengig av frafallsmønsteret. Ved nærmere ettersyn viser det seg imidlertid at modell 2a ikke er noen ordentlig forfining av modell 1. Vi får evnen til å betinge med hensyn på det eksakte frafallsmønsteret, men må gi fra oss muligheten til å betinge med hensyn på at personen ikke har svart på spørsmålet om undersysselsetting. For at det skal være akseptabelt å bare betinge med hensyn på frafallsmønsteret ved imputeringen må vi anta at

$$P(Y_u = 1 \mid \text{frafallsmønster}) = P(Y_u = 1 \mid \text{frafallmønster og } R_u^2 = 0)$$

for alle frafallsmønstre der imputering kan være aktuelt. Dette gjelder mønstrene $(0,0,0)$, $(1,0,0)$ og $(1,1,0)$. Ligningen gjelder for frafallsmønsteret $(0,0,0)$ for da er R_u^2 automatisk null. For de to andre mønstrene er den mer tvilsom. De aller fleste med frafallsmønster $(1,0,0)$ har dette mønsteret fordi de har svart nei på spørsmål 18 (i vårt datasett 2 496 av 2 501). Disse har svart på spørsmålet om undersysselsetting, og de har svart nei. Sannsynligheten på venstre side blir derfor fullstendig dominert av disse personene, og blir derfor svært liten. Det kan godt være at sannsynligheten på høyre side også er liten, men personer med frafallsmønster $(1,0,0)$ og $R_u^2 = 0$ må ha svart ja på spørsmål 18. De oppfyller dermed det første kravet til undersysselsetting. Dette gjør det rimelig å anta at sannsynligheten på høyre side er større enn den på venstre, og at vi følgelig vil imputere for få undersysselsatte i gruppen med dette frafallsmønsteret.

Situasjonen for frafallsmønsteret $(1,1,0)$ er tilsvarende. De aller fleste med dette mønsteret har svart nei på spørsmål 19 (392 av 401).

Modell 2a gir også problemer med simuleringene. Siden R_u^2 ikke alltid lar seg utlede fra variablene i modellen, kan vi komme i den uvanlige situasjonen at det ikke er mulig å avlede det observerte

datasettet fra det simulerte. Hvis vi simulerer $(R_{18}^2, R_{19}^2, R_{23}^2, Y_u) = (1,0,0,0)$ vet vi ikke om vi her ville observert null eller missing på undersysselsetting. Vi vet dermed ikke om den simulerte vektoren skal gi opphav til faktoren

$$P(R_{18}^2 = 1, R_{19}^2 = 0, R_{23}^2 = 0, Y_u = 0 | \mathbf{x}, \mathbf{z})$$

eller faktoren

$$P(R_{18}^2 = 1, R_{19}^2 = 0, R_{23}^2 = 0 | \mathbf{x}, \mathbf{z})$$

i likelihoodfunksjonen.

Vi må huske at målet med modell 2a var å kunne betinge med hensyn på frafallsmønsteret ved imputeringen. Etter å ha sett vanskelighetene med modell 2a er det fristende å nærme seg dette fra en litt annen vinkel. Vi innser nå at den viktigste frafallsvariabelen er R_u^2. Det er først når denne er null at det aktuelt å imputere. I såfall kan vi begynne å spørre etter den mer detaljerte informasjonen som det eksakte frafallsmønsteret gir. Når $R_u^2 = 0$ er det tre mulige verdier av vektoren $(R_{18}^2, R_{19}^2, R_{23}^2)$, nemlig $(0,0,0)$, $(1,0,0)$ og $(1,1,0)$. Siden $R_{23}^2 = 0$ i alle disse tilfellene, er frafallsmønstret entydig gitt ved R_{18}^2 og R_{19}^2 når $R_u^2 = 0$. Disse observasjonene leder oss til å formulere modell 2b.

Modell 2b

Y_u modelleres etter (2.2.3) for alle deltidssysselsatte. De enkelte variablene Y_{18}, Y_{19} og Y_{23} modelleres ikke. Vi modellerer R_u^2 etter (3.1.2) for alle deltidssysselsatte. R_{18}^2 modelleres bare når $R_u^2 = 0$, og da etter (3.1.2). R_{19}^2 modelleres etter (3.1.2) når $R_u^2 = 0$ og $R_{18}^2 = 1$.

Modell for R_u^2:

Modellen gjelder for alle deltidssysselsatte.

$$\ln \frac{P(R_u^2 = 1 | \mathbf{z}, y_u, R^1)}{P(R_u^2 = 0 | \mathbf{z}, y_u, R^1)} = \psi_{u,0} + \psi_{u,1} z_1 + \psi_{u,2} z_2 + \sum_{l=1,3,4,5} \psi_{u,3l} D_{3l} + \psi_{u,4} y_u + \psi_{u,5} R^1$$

Modell for R_{18}^2:

Modellen gjelder for deltidssysselsatte med $R_u^2 = 0$. I utgangspunktet er modellen for R_{18}^2 analog med de tidligere modellene for partielt frafall:

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

Med vårt konkrete datasett fikk vi imidlertid ikke brukbare parameterestimater for denne modellen. For å lette estimeringen, har vi derfor fjernet leddene knyttet til sivilstand, kjønn, region og R^1. Modellen for R_{18}^2 er dermed:

$$\ln \frac{P(R_{18}^2 = 1 \mid \mathbf{z}, y_u, R^1, R_u^2 = 0)}{P(R_{18}^2 = 0 \mid \mathbf{z}, y_u, R^1, R_u^2 = 0)} = \psi_{18,0} + \psi_{18,4} y_u$$

Modell for R_{19}^2:

Modellen gjelder for deltidssysselsatte med $R_u^2 = 0$.

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

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

Når $R_{18}^2 = 1$ bruker vi i utgangspunktet (3.1.2):

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

Som i modellen for R_{18}^2 måtte vi også her forenkle modellen betydelig for å få brukbare parameterestimater. Modellen vi bruker inneholder bare konstantledd:

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

Vi kan merke oss at R_{19}^2 i denne modellen er udefinert for personer som har oppgitt svaret nei på spørsmål 18, siden $R_u^2 = 0$ utelukker denne muligheten.

Heretter kaller vi modell 2b for modell 2.

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 (2.2.3), 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 $Y_{19} = 1$. Legg merke til at vi her modellerer Y_{23} bare når $Y_{19} = 1$, og ikke for hele målgruppen for spørsmål 23a. Dette er fordi målsettingen her er å modellere undersysselsettingsvariabelen Y_u. Fra ligningen under ser vi at det er sannsynligheten for å svare ja på spørsmål 23a gitt at man har svart ja på spørsmål 19 som er den interessante størrelsen i denne sammenhengen. Modellen for Y_u er til slutt gitt ved:

$$\begin{aligned} &P(Y_u = 1 \mid \mathbf{x}) = P(Y_{18} = 1 \cap Y_{19} = 1 \cap Y_{23} = 1 \mid \mathbf{x}) \\ &= P(Y_{23} = 1 \mid Y_{19} = 1, Y_{18} = 1, \mathbf{x}) P(Y_{19} = 1 \mid Y_{18} = 1, \mathbf{x}) P(Y_{18} = 1 \mid \mathbf{x}) \end{aligned} \qquad (3.1.3)$$

og gjelder for deltidssysselsatte.

R_{18}^2, R_{19}^2 og R_{23}^2 modelleres i utgangspunktet etter (3.1.2), 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 modell (3.1.2) 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 (3.1.2). Modellen gjelder for deltidssysselsatte.

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

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 $R_{18}^2=1$ og $Y_{18}=1$ bruker vi i utgangspunktet (3.1.2), men i vårt konkrete tilfelle den enklere versjonen:

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

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.

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

For personer som har svart ja på spørsmål 18 og 19 bruker vi generelt (3.1.2), men i vårt konkrete tilfelle den enklere versjonen:

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

3.1.3. Implisitt modell for responsmekanismen med nåværende imputeringsmetoder i AKU

Responsmodellene for R^1 og R^2 vil i et senere notat bli brukt for å utvikle nye imputeringsmetoder for partielt frafall. Nåværende imputeringsrutiner for partielt frafall varierer en del fra spørsmål til

spørsmål. Vi skal i dette delkapittel beskrive hovedtrekkene ved den nåværende imputeringen i AKU. Til sammenligning med den ikke-ignorerbare frafallsmodellen for (R^1,R^2) ovenfor skal vi se hva slags frafallsmodell som ligger under dagens imputeringsmetoder. Imputeringen i AKU for alle variablene vi ser på gjøres uavhengig av R^1, dvs. at fordelingen til R^1 er uavhengig av y, og dermed ignorerbar. I tillegg er R^2 stokastisk uavhengig av R^1.

Når det gjelder R^2, la oss først betrakte undersysselsettingsvariabelen Y_u. AKU imputerer kun for Y_u, og ikke for de separate spørsmålene 18, 19 og 23. Imputeringsmetoden er delvis basert på etterstratifisering etter $\mathbf{z} = (z_1,z_2)$ hvor z_1 = kjønn (= 0/1 hvis mann/kvinne) og z_2 = aldersgruppe (= 1, . . ., 5 etter alder 16-19, 20-24, 25-39, 40-54, 55-74). Spørsmålene om undersysselsetting gis bare til deltidssysselsatte og ved direkte intervju av IO. La s_f være frafallsgruppen for deltidssysselsatte med direkte intervju. Alle deltidssysselsatte skal først besvare spørsmål 17 om hvilken hovedvirksomhet de har. For de IO med direkte intervju som svarer på spørsmål 17, men har frafall på alle undersysselsettingsspørsmålene 18, 19 og 23 så imputeres verdien $Y_u^* = 0$ (man antar at intervjueren har glemt å krysse av for nei i spørsmål 18). Denne gruppen kan vi betegne med s_{rf}. AKU betrakter egentlig ikke s_{rf} som en del av frafallsgruppen, men det er den selvsagt. Siden dette vanligvis er en stor del av det partielle frafallet ved direkte intervju (i perioden 1988-1992 ca. 60% av det totale frafallet blant IO med direkte intervju) så kan AKU's imputeringsmetode medføre at antall undersysselsatte blir under-
estimert. Vi ser at $s_{rf} = \{i \in s_f : R_{i17}^2 = 1, \mathbf{R}_i^2 = 0\}$ hvor $\mathbf{R}_i^2 = (R_{i18}^2, R_{i19}^2, R_{i23}^2)$. Dette betyr at man med dagens imputeringsrutine antar at

$$P(Y_u = 1 \mid \mathbf{z}, s_{rf}) = 0,$$

$$\text{og dermed : } P(R_{17}^2 = 1, \mathbf{R}^2 = \mathbf{0} \mid \mathbf{z}, D = 1, y_u = 1) = 0.$$

Her angir indikatorvariabelen D at det er direkte intervju av IO. Dvs., D = 1/0 hvis direkte/indirekte intervju. Utenfor s_{rf} antar AKU essensielt tilfeldig frafall innen hvert stratum $\mathbf{z}$. Skjematisk kan vi presentere imputeringen i AKU som i figuren nedenfor.

Figur 3. Utvalget av deltidssysselsatte med imputerte data y_u^*, og $q(\mathbf{z})$ = observert andel undersysselsatte i stratum $\mathbf{z}$.

<table>
<tr><th>Stratum</th><th>Svarut-valget</th><th colspan="4">Frafall</th></tr>
<tr><td></td><td></td><td>indirekte intervju</td><td colspan="3">direkte intervju , s_f</td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td></td><td>s_{rf}</td></tr>
<tr><td></td><td>$R_u^2 = 1$</td><td>$\mathbf{R}^2 = \mathbf{0}$</td><td>$R_{17}^2 = 0, \mathbf{R}^2 = \mathbf{0}$</td><td>$\mathbf{R}^2 = (1,0,0)$
$R_u^2 = 0$</td><td>$\mathbf{R}^2 = (1,1,0),$
$R_u^2 = 0$</td><td>$R_{17}^2 = 1, \mathbf{R}^2 = \mathbf{0}$</td></tr>
<tr><td>.</td><td>.</td><td>.</td><td></td><td></td><td></td><td>.</td></tr>
<tr><td>.</td><td>.</td><td colspan="2">antall: m_0</td><td>antall: m_1</td><td>antall: $m_{1,1}$</td><td>.</td></tr>
<tr><td>$\mathbf{z}$</td><td>$q(\mathbf{z})$</td><td colspan="2">Tilfeldig utvalg av personer gis $y_u^* = 1$
Antall: $m_0 \cdot q(\mathbf{z}) - (m_1 + m_{1,1})$</td><td>$y_u^* = 1$</td><td>$y_u^* = 1$</td><td>$y_u^* = 0$</td></tr>
<tr><td>.</td><td>.</td><td colspan="2">.</td><td>.</td><td>.</td><td>.</td></tr>
</table>

Med $\bar{s}_{rf}$ lik gruppen av IO i det *totale* utvalget som ikke er med i s_{rf} har vi følgende "AKU-imputerings-modell":

$$P(Y_u = 1|\mathbf{z}, \bar{s}_{rf}, R_u^2 = 0) = P(Y_u = 1|\mathbf{z}, \bar{s}_{rf}, R_u^2 = 1) = P(Y_u = 1|\mathbf{z}, \bar{s}_{rf}).$$

Legg merke til at $P(Y_u = 1|\mathbf{z}, \bar{s}_{rf}, R_u^2 = 1) = P(Y_u = 1|\mathbf{z}, R_u^2 = 1)$ hvilket betyr at imputeringen utenfor s_{rf} er basert på den observerte andel undersysselsatte innen etterstrata bestemt av **z**. Dvs., Y_u og R_u^2 er stokastisk uavhengige, gitt ($\mathbf{z}, \bar{s}_{rf}$), slik at RM-modellen for AKU med hensyn til undersysselsetting i $\bar{s}_{rf}$ blir:

$$P(R_u^2 = 1|\mathbf{z}, \bar{s}_{rf}, y_u) = P(R_u^2 = 1|\mathbf{z}, \bar{s}_{rf}) \text{ for } y_u = 0,1.$$

og dermed også: $P(R_u^2 = 0|\mathbf{z}, \bar{s}_{rf}, y_u) = P(R_u^2 = 0|\mathbf{z}, \bar{s}_{rf})$ for $y_u = 0,1$.

Utenfor s_{rf} er R_u^2 stokastisk uavhengig av Y_u, gitt stratum definert av **z**, slik at innenfor hvert stratum i $\bar{s}_{rf}$ så er sannsynligheten for frafall den samme blant undersysselsatte som de andre, dvs. uavhengig av om de er undersysselsatt eller ikke. AKU's imputeringsmetode antar altså implisitt at det partielle frafallet for undersysselsettingsvariabelen er ignorerbart utenfor s_{rf}. Dette er det essensielle ved RM-modellen for undersysselsetting som AKU implisitt benytter.

Vi har også at $P(\mathbf{R}^2 = \mathbf{1}|\mathbf{z}, \bar{s}_{rf}, y_u = 1) = P(R_u^2 = 1|\mathbf{z}, \bar{s}_{rf}, y_u = 1) = P(R_u^2 = 1|\mathbf{z}, \bar{s}_{rf})$. Dette betyr at sannsynligheten for svar på *alle* undersysselsettingsspørsmål blant de undersysselsatte i stratum **z** er lik sannsynligheten for å gi et svar om undersysselsetting i hele stratum **z** utenfor s_{rf}.

Det kan også være av interesse å se på *frafallsmønsteret* for undersysselsettingsvariabelen. Som vist i figur 3, er situasjonen da ikke så enkel. Når det gjelder frafallsmønsteret på de 3 spørsmålene så har vi ikke ignorerbarhet gitt stratum **z**, heller ikke utenfor s_{rf}. Dvs. sannsynlighetene for de forskjellige frafallsmønstre gitt stratum **z** i $\bar{s}_{rf}$ er avhengig av y_u-verdien. Legg merke til at, siden imputeringen i AKU gjøres direkte på y_u, så vil den implisitte frafallsmodellen i AKU kun være avhengig av y_u, og ikke de enkelte spørsmålsvariablene. Vi bemerker at frafallssannsynlighetene for $\mathbf{R}^2$ = (1,0,0) og $\mathbf{R}^2$ = (1,1,0) er typisk nær 0, av størrelsesorden mindre enn 0,0005, så disse er ikke så interessante. I tillegg så has at

$$P(\mathbf{R}^2 = \mathbf{0}|\mathbf{z}, \bar{s}_{rf}, y_u = 0) \approx P(R_u^2 = 0|, \bar{s}_{rf}, y_u = 0) = P(R_u^2 = 0|\mathbf{z}, \bar{s}_{rf}).$$

Hvis vi antar $\mathbf{R}^2 = \mathbf{0}$ i *hele* frafallsgruppen (som vanligvis holder tilnærmet), så imputerer AKU etter følgende modell for Y_u:

$$(I.a) P(Y_u = 1|\mathbf{z}, D = 0, \mathbf{R}^2 = \mathbf{0}) = P(Y_u = 1|\mathbf{z}, D = 1, R_{17}^2 = 0, \mathbf{R}^2 = \mathbf{0}) = P(Y_u = 1|\mathbf{z}, R_u^2 = 1)$$
$$(I.b) P(Y_u = 1|\mathbf{z}, s_{rf}) = 0.$$

Merk at nå has $\{R_u^2 = 0\} = \{\mathbf{R}^2 = \mathbf{0}\}$ og dermed: $P(\mathbf{R}^2 = \mathbf{0}|\mathbf{z}, \bar{s}_{rf}, y_u) = P(\mathbf{R}^2 = \mathbf{0}|\mathbf{z}, \bar{s}_{rf})$ for $y_u = 0,1$.

(Selv om vi har at $\mathbf{R}^2 = \mathbf{0}$ er eneste mulige frafallsmulighet, så er sannsynlighetene for de forskjellige *svarmønstre* under $\{ R_u^2 = 1 \}$ avhengig av y_u-verdien. Dette er imidlertid ikke relevant for imputeringsmetoden, selvsagt.)

Vi har også partielt frafall på undersysselsetting på grunn av indirekte intervju. Hvis $\mathbf{R}^2 = \mathbf{0}$ i hele frafallsgruppen så har vi at $P(D = 0|\mathbf{z}, \bar{s}_{rf}, y_u) = P(D = 0|\mathbf{z}, \bar{s}_{rf})$, for $y_u = 0,1$, dvs., D, Y_u *er stokastisk uavhengige utenfor* s_{rf}. Vi har da også $P(D = 1|\mathbf{z}, \bar{s}_{rf}, y_u) = P(D = 1|\mathbf{z}, \bar{s}_{rf})$. Vi har altså ignorerbarhet for det partielle frafallet pga. indirekte intervju i $\bar{s}_{rf}$-gruppen når $\mathbf{R}^2 = \mathbf{0}$ er eneste mulighet i frafallsgruppen. Generelt, imidlertid, har vi, gitt stratum $\mathbf{z}$ i $\bar{s}_{rf}$, at det partielle frafallet pga. indirekte intervju er ikke-ignorerbart.

Tilslutt for undersysselsetting ser vi på RM blant de IO med direkte intervju. Hvis $\mathbf{R}^2 = \mathbf{0}$ er eneste frafallsmulighet så has at $P(R_u^2 = 0|\mathbf{z}, \bar{s}_{rf}, D = 1, y_u) = P(R_u^2 = 0|\mathbf{z}, \bar{s}_{rf}, D = 1)$, uavhengig av y_u. Dvs. den betingede sannsynligheten for partielt frafall for undersysselsetting for direkte intervju-gruppen, gitt stratum $\mathbf{z}$ i $\bar{s}_{rf}$, er uavhengig av y_u-verdi. Det partielle frafallet for "undersysselsetting" i "direkte-gruppen" er ignorerbart utenfor s_{rf}. Dette gjelder, imidlertid, ikke generelt.

De tre andre variablene som AKU imputerer for har vi modifisert til 0/1-variable. Logistiske modeller for RM er derfor selvsagt ikke relevant for AKU. I AKU gjøres imputeringen stort sett ved etterstratifisering, noe som betyr at RM antas ignorerbar.

Når det gjelder spørsmål 24, "ønsket arbeidstid for undersysselsatte", så stratifiserer AKU etter $\mathbf{z} = (z_1, z_2, z_3)$, hvor z_1, z_2 er som for undersysselsetting, og z_3 er avtalt arbeidstid (gruppene 1-9, 10-19, 20-29 og 30-36 timer). For alle i samme etterstratum $\mathbf{z}$ imputeres verdien : avtalt arbeidstid + gjennomsnittlig observert differanse mellom ønsket og avtalt arbeidstid i etterstratum $\mathbf{z}$. Dette betyr at AKU implisitt antar at i hvert etterstratum $\mathbf{z}$ så er de betingede fordelingene til Y gitt $R^2 = 0$ og $R^2 = 1$ like, dvs., $f(y|\mathbf{z}, R^2 = 0) = f(y|\mathbf{z}, R^2 = 1)$. Også for spørsmål 62, "ønsket arbeidstid for arbeidssøkere" imputeres etter modellen $f(y|\mathbf{z}, R^2 = 0) = f(y|\mathbf{z}, R^2 = 1)$, hvor nå $\mathbf{z}$ består av alder, kjønn og virksomhet. Dette betyr at for spørsmålene 24 og 62 så antas Y og R^2 stokastisk uavhengige, gitt etterstratum $\mathbf{z}$. Sagt på en annen måte, $P(R^2 = 1|\mathbf{z}, y) = P(R^2 = 1|\mathbf{z})$, uavhengig av y, hvilket betyr at AKU's imputering for disse spørsmål antar at responsmekanismen er ignorerbar.

Angående spørsmål 25, "faktisk arbeidstid" for personer med ett arbeidsforhold så imputeres enten personens avtalte eller gjennomsnittlige arbeidstid hvis det er oppgitt. Hvis ikke, etterstratifiserer man etter næring og kjønn hvis mulig, eller bare kjønn, og gjennomsnittlig observert faktisk arbeidstid i etterstrata brukes som imputeringsverdi. Dette betyr at RM er ignorerbar på tilsvarende måte som for spørsmålene 24 og 62, såfremt partielt frafall på spørsmålene om avtalt og gjennomsnittlig arbeidstid er stokastisk uavhengig av Y og R^2 på spørsmål 25.

For en mer detaljert beskrivelse av de implisitte RM-modellene i AKU henviser vi til notatet om de nåværende imputeringsrutiner i AKU som er under utarbeidelse.

3.2. Effekten av å tilføye frafallsmodell

Når modellene for Y, R^1 og R^2 er valgt, beregner vi maksimumlikelihoodestimater (MLE) for de ukjente parametrene. MLE fremkommer ved å maksimere likelihoodfunksjonen, som er skrevet ut i Appendiks A. Siden funksjonen er komplisert, bruker vi en numerisk rutine (E04KCF i NAG) til å

finne maksimumspunktet. MLE for hver enkelt parameter er tabulert i Appendiks C. Vi har også beregnet MLE for sannsynlighetene i modell-ligningene, dvs. $P(Y=1|\mathbf{x})$, $P(R^1=1|\mathbf{z},y)$ og $P(R^2=1|\mathbf{z},y,R^1)$ (Appendiks B). Dette gjøres ved å sette inn MLE for enkeltparametrene i henholdsvis (2.2.3), (3.1.1) og (3.1.2), og deretter løse sannsynlighetene ut av ligningene.

I kapittel 2 modellerte vi bare Y-ene, og parameterestimatene var basert på svarutvalget. I dette kapittelet modellerer vi fremdeles Y-ene, men vi har i tillegg føyd til modeller for enhetsfrafall på første henvendelse og partielt frafall. Vi kan få et inntrykk av effekten av frafallsmodellering ved å sammenligne estimatene for $P(Y=1|\mathbf{x})$ i tilfellene med og uten frafallsmodeller. Dette gjøres i tabellene under.

Estimater for $P(Y=1|\mathbf{x})$ basert på den samlede populasjons- og frafallsmodellen er det første tallet i hver celle, og de tilsvarende estimatene basert kun på populasjonsmodellen anvendt på svarutvalget er det andre tallet i hver celle. Stor ulikhet vitner om at frafallet har stor betydning.

a står for aldersgruppe. For definisjon av aldersgrupper og boregioner, se avsnitt 2.2.12.

Spørsmål 18

Menn

a	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
1	0,365 0,344	0,375 0,334	0,357 0,319	0,370 0,332	0,433 0,365
2	0,488 0,457	0,408 0,384	0,411 0,378	0,548 0,521	0,583 0,557
3	0,309 0,293	0,254 0,240	0,287 0,260	0,353 0,335	0,298 0,279
4	0,070 0,054	0,013 0,011	0,010 0,009	0,024 0,017	0,000 0,000

Kvinner

a	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
1	0,197 0,195	0,371 0,340	0,288 0,277	0,306 0,284	0,326 0,279
2	0,242 0,229	0,345 0,327	0,284 0,273	0,415 0,398	0,409 0,392
3	0,154 0,152	0,241 0,232	0,218 0,211	0,281 0,273	0,203 0,196
4	0,156 0,138	0,070 0,062	0,042 0,042	0,095 0,078	0,000 0,000

Vi ser ingen særlig store forskjeller her. Den største differansen (første tall minus andre tall) er 0,068 (menn i aldersgruppe 1, boregion 5). Ellers kan vi legge merke til alle differansene er positive, dvs. at sannsynligheten for å ønske lengre arbeidstid blir estimert større når frafallsmodellen er med.

Spørsmål 19

Menn

a	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
1	0,696 0,596	0,183 0,304	0,361 0,361	0,146 0,213	0,313 0,215
2	0,829 0,887	0,759 0,700	0,757 0,750	0,629 0,591	0,556 0,593
3	0,900 0,867	0,652 0,659	0,701 0,714	0,518 0,545	0,746 0,547

Kvinner

a	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
1	0,705 0,533	0,373 0,501	0,528 0,468	0,365 0,423	0,514 0,387
2	0,583 0,680	0,697 0,652	0,629 0,622	0,610 0,578	0,444 0,541
3	0,678 0,589	0,525 0,558	0,509 0,525	0,446 0,480	0,603 0,443

Forskjellene her er noe større enn på spørsmål 18. Den største forskjellen finner vi blant menn i aldersgruppe 3, boregion 5, der differansen er nesten 0,2.

Spørsmål 23a

Menn

a	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
1	0,875 0,874	0,925 0,927	0,906 0,901	1,000 1,000	0,733 0,734
2	0,952 0,950	0,940 0,940	0,917 0,912	0,952 0,952	0,990 0,989
3	0,971 0,969	0,959 0,957	0,953 0,943	0,975 0,972	0,990 0,987

Kvinner

a	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
1	0,937 0,930	0,989 0,988	0,987 0,986	1,000 1,000	0,700 0,700
2	0,804 0,786	0,914 0,909	0,894 0,891	0,888 0,886	0,897 0,888
3	0,873 0,872	0,941 0,940	0,939 0,935	0,940 0,938	0,897 0,880

Her er det svært små forskjeller.

Undersysselsetting

Menn

a		boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
1	**Modell 1**	0,220	0,088	0,123	0,081	0,163
	Modell 2	0,220	0,088	0,124	0,081	0,165
	Modell 3	0,277	0,084	0,157	0,066	0,173
	Svarutvalg	0,210	0,079	0,112	0,073	0,138
2	**Modell 1**	0,376	0,306	0,286	0,345	0,329
	Modell 2	0,373	0,306	0,287	0,345	0,330
	Modell 3	0,424	0,341	0,352	0,368	0,328
	Svarutvalg	0,355	0,293	0,267	0,326	0,311
3	**Modell 1**	0,245	0,152	0,159	0,168	0,181
	Modell 2	0,240	0,152	0,157	0,165	0,180
	Modell 3	0,297	0,194	0,245	0,205	0,245
	Svarutvalg	0,233	0,146	0,146	0,159	0,171

Kvinner

a		boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
1	**Modell 1**	0,145	0,125	0,142	0,109	0,153
	Modell 2	0,146	0,125	0,142	0,109	0,153
	Modell 3	0,145	0,156	0,162	0,127	0,204
	Svarutvalg	0,144	0,115	0,137	0,103	0,135
2	**Modell 1**	0,129	0,212	0,162	0,230	0,158
	Modell 2	0,129	0,211	0,162	0,231	0,156
	Modell 3	0,142	0,248	0,180	0,258	0,166
	Svarutvalg	0,123	0,202	0,157	0,221	0,150
3	**Modell 1**	0,089	0,119	0,101	0,123	0,094
	Modell 2	0,089	0,119	0,101	0,124	0,093
	Modell 3	0,103	0,133	0,115	0,128	0,120
	Svarutvalg	0,088	0,115	0,098	0,121	0,092

De estimerte sannsynlighetene for å være undersysselsatt under modell 1, 2 og 3, er alle større eller lik de tilsvarende sannsynlighetene estimert bare på grunnlag av svarutvalget. Det er som forventet at sannsynligheten for å være undersysselsatt anslås større når man tar hensyn til frafall, ettersom undersysselsatte antageligvis er overrepresentert i frafallet. Modell 1 og 2 gir svært like estimater. Dette er ikke så overraskende, for de to modellene bruker samme modell for både Y_u og R_u^2. Forskjellen er bare at modell 2 i tillegg inneholder modeller for R_{18}^2 og R_{19}^2. Sannsynlighetene estimert under modell 1 og 2, er bare litt større enn sannsynlighetene basert på svarutvalget. Sannsynlighetene estimert under modell 3 er gjennomgående en god del større enn de for modell 1 og 2. Det kan tolkes som at modell 3 korrigerer best for frafall.

Ellers ser vi at det er stor forskjell på menn og kvinner angående undersysselsetting. Bortsett fra i aldersgruppe 1, har menn gjennomgående en høyere sannsynlighet for å være undersysselsatt enn kvinner.

Spørsmål 24

Menn

a	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
1	0,134 0,135	0,096 0,101	0,087 0,092	0,221 0,220	0,324 0,319
2	0,838 0,842	0,914 0,917	0,874 0,885	0,890 0,891	0,964 0,964
3	0,717 0,733	0,812 0,823	0,727 0,757	0,718 0,739	0,852 0,862

Kvinner

a	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
1	0,298 0,298	0,126 0,133	0,141 0,139	0,333 0,334	0,256 0,261
2	0,517 0,522	0,525 0,529	0,474 0,478	0,520 0,521	0,595 0,602
3	0,529 0,532	0,488 0,491	0,425 0,429	0,421 0,433	0,402 0,418

For dette spørsmålet blir forskjellene små.

Spørsmål 25

Menn

a	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
1	0,094 0,094	0,204 0,204	0,139 0,139	0,154 0,154	0,150 0,150
2	0,822 0,822	0,771 0,771	0,785 0,785	0,787 0,787	0,736 0,736
3	0,463 0,463	0,425 0,426	0,339 0,339	0,220 0,220	0,413 0,413

Kvinner

a	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
1	0,067 0,067	0,127 0,127	0,074 0,075	0,077 0,077	0,135 0,135
2	0,506 0,506	0,381 0,381	0,370 0,370	0,354 0,354	0,443 0,443
3	0,167 0,167	0,124 0,124	0,080 0,079	0,042 0,042	0,174 0,174

For spørsmål 25 er estimatene nærmest identiske.

Spørsmål 62

Menn

a	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
1	0,192 0,199	0,199 0,209	0,353 0,361	0,737 0,738	0,446 0,500
2	0,876 0,878	0,949 0,952	0,926 0,928	0,990 0,990	1,000 1,000
3	0,824 0,830	0,963 0,966	0,913 0,917	0,993 0,993	1,000 1,000

Kvinner

a	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
1	0,422 0,426	0,192 0,194	0,367 0,369	0,328 0,328	0,000 0,000
2	0,715 0,718	0,673 0,679	0,604 0,612	0,667 0,674	0,952 0,955
3	0,417 0,419	0,556 0,565	0,357 0,362	0,548 0,557	0,783 0,800

Også her er forskjellene små.

3.3. Simuleringsstudier av MLE

Det er nå viktig å undersøke om maksimumlikelihoodestimatorene er forventningsrette, og hvilket standardavvik de har. Dette gjør vi ved hjelp av en metode som kalles parametrisk bootstrapping. Vi lager 1000 simulerte datasett på følgende måte:

Vi tar utgangspunkt i de 21 900 personene som har levert skjemaet, enten ved første henvendelse eller etter oppfølging. Verdiene på forklaringsvariablene, altså alder, kjønn, region og sivilstand, lar vi være uforandret, mens vi simulerer nye verdier for R^1, Y-ene og R^2-ene. I det opprinnelige datasettet er det 9 personer som mangler verdi på en eller flere av forklaringsvariablene. Ettersom vi trenger forklaringsvariablene i simuleringsprosessen fjernes disse personene. Hvert av de simulerte datasettene består derfor av 21 891 personer.

Først simulerer vi R^1 etter (3.1.1), der vi har satt inn MLE fra det ekte datasettet for φ-ene. Dette gjør vi for alle personene. En alene-boende mann i boregion 1 (Oslo, Akershus) vil f.eks. få simulert $R^1 = 1$ med sannsynlighet 0,904 (se Appendiks B). Deretter plukker vi ut målgruppene for hvert enkelt spørsmål, og simulerer Y og R^2 for disse personene. Y simuleres først, etter (2.2.3), med MLE fra det ekte datasettet innsatt for β-ene. Når vi har simulert Y og R^1, kan vi til slutt simulere R^2 etter (3.1.2), også her med MLE fra det ekte datasettet innsatt for de ukjente ψ-ene.

Vi understreker at vi ikke simulerer kovariatverdier. Alle de simulerte datasettene er altså like med hensyn til forklaringsvariable. Variasjonen mellom de simulerte datasettene reflekterer derfor modellvariansen gitt et bestemt personnutvalg, og ikke variasjonen i gjentatte personutvalg.

For hvert simulert utvalg maksimeres likelihoodfunksjonen. Gjennomsnitt og standardavvik av MLE for hver parameter beregnes fra de 1000 simuleringene. Disse gjennomsnittene og standardavvikene regner vi som gode estimater for maksimumlikelihoodestimatorenes forventning og standardavvik. Resultatene av beregningene er tabulert i Appendiks C, og de er litt varierende. Parametre som gjelder små grupper, som f.eks. pensjonister i en bestemt boregion, blir gjerne dårlig estimert. Frafallsparameteren ψ_4 (koeffisienten foran y), er også gjennomgående problematisk.

Det at visse enkeltparametre blir dårlig estimert, trenger imidlertid ikke å ha avgjørende betydning. For oss er det viktigere hvor godt selve sannsynlighetene i modell-ligningene, dvs. $P(Y = 1 | \mathbf{x})$, $P(R^1 = 1 | \mathbf{z}, y)$ og $P(R^2 = 1 | \mathbf{z}, y, R^1)$, estimeres. Empirisk gjennomsnitt og standardavvik for MLE

for de tre sannsynlighetene, basert på de 1000 simuleringene, er tabulert i Appendiks B. Resultatene her er jevnt over gode. Estimatorene er tilnærmet forventningsrette, og standardavvikene er stort sett nokså små.

3.4. Testing av modelltilpasning

Vi trenger å teste modelltilpasning også etter at alle frafallsparametrene er tatt med i modellen. Slike tester finnes ikke i standard programvare, så vi må definere våre egne tester, som vi så implementerer i FORTRAN. Vi tar utgangpunkt i de to observatorene (2.2.1) og (2.2.2) beskrevet i avsnitt 2.2.2. og generaliserer disse.

3.4.1. Pearsons kjikvadrattest

La som før x_1 stå for alder, x_2 for kjønn, x_3 for bostedsregion og z_1 for sivilstand. Essensen i en kji-kvadrattest er å dele opp datasettet i forskjellige grupper for å måle forskjellen mellom det faktiske antall observasjoner og forventet antall observasjoner under modellen i hver gruppe. Gruppeinndelingen skjer på grunnlag av kovariatvektorer (forklaringsvariable) og en eller flere stokastiske variable.

Når vi i det følgende snakker om spørsmål j, R_j^2 eller Y_j inkluderer dette spørsmålet om undersysselsetting sett under ett, R_u^2 og Y_u.

For personer i målgruppen til spørsmål j er de stokastiske variablene R^1, R_j^2 og Y_j. R^1 indikerer om IO har svart på skjemaet ved første henvendelse, R_j^2 indikerer om IO har svart på spørsmål j (enten ved første henvendelse eller etter oppfølgning) og Y_j er svaret på spørsmål j. Forklaringsvariablene er x_1, x_2, x_3 og z_1.

For personer utenfor målgruppen til spørsmål j er det bare R^1 som er relevant som stokastisk variabel, og forklaringsvariablene er z_1, x_2 og x_3.

Vi lager to testobservatorer, en for personer i målgruppen og en for personer utenfor, og legger dem sammen til slutt.

Personer i målgruppen

En gruppe er her bestemt ved verdien av kovariatvektoren (z_1, x_1, x_2, x_3) og verdien av den stokastiske vektoren (R_j^2, R^1, Y_j). Når R_j^2 er null, dvs. når vi har partielt frafall på spørsmål j, kjenner vi ikke verdien av Y_j. Blant personer i målgruppen er det derfor seks mulige verdier av vektoren (R_j^2, R^1, Y_j), nemlig de som er listet opp under.

Tabell 3.4.1

l	R_j^2	R^1	Y_j
1	0	0	-
2	0	1	-
3	1	0	0
4	1	0	1
5	1	1	0
6	1	1	1

Her betyr "-" at verdien mangler.

Vi indekserer de forskjellige kovariatvektorene (z_1, x_1, x_2, x_3) som forekommer blant personer i målgruppen fra $i = 1$ til m_1, og de seks mulige verdiene av den stokastiske vektoren (R_j^2, R^1, Y_j) fra $l = 1$ til 6.

La nå

r_{il} være antall personer i målgruppen med kovariatvektor i og stokastisk vektor l

n_i være antall personer i målgruppen med kovariatvektor i,

p_{il} være modellsannsynligheten for at en person med kovariatvektor i har stokastisk vektor l

$\hat{p}_{il}$ være MLE for p_{il} under modellen

For å regne ut $\hat{p}_{il}$, finner vi først p_{il} uttrykt ved β-ene, φ-ene og ψ-ene. Her må vi skille mellom to tilfeller:

1. Spørsmålene betraktet hver for seg og modell 1 og 2 for spørsmålet om undersysselsetting sett under ett:
For $i = (z_1, x_1, x_2, x_3)$ og l på formen $(1, r^1, y_j)$ får vi

$$p_{il} = P(R_j^2 = 1, R^1 = r^1, Y_j = y_j \mid z_1, x_1, x_2, x_3)$$

$$= P(R_j^2 = 1 \mid R^1 = r^1, Y_j = y_j, z_1, x_1, x_2, x_3) \cdot P(R^1 = r^1 \mid z_1, x_2, x_3) \cdot P(Y_j = y_j \mid x_1, x_2, x_3) \qquad (3.4.1)$$

For $i = (z_1, x_1, x_2, x_3)$ og l på formen $(0, r^1, -)$ får vi

$$p_{il} = \sum_{y_j=0}^{1} P(R_j^2 = 0, R^1 = r^1, Y_j = y_j \mid z_1, x_1, x_2, x_3)$$

$$= \sum_{y_j=0}^{1} P(R_j^2 = 0 \mid R^1 = r^1, Y_j = y_j, z_1, x_1, x_2, x_3) P(R^1 = r^1 \mid z_1, x_2, x_3) \cdot P(Y_j = y_j \mid x_1, x_2, x_3) \qquad (3.4.2)$$

Vi minner om at Y_j modelleres etter (2.2.3), R^1 etter (3.1.1) og R_j^2 gitt Y_j og R^1 etter (3.1.2). Sannsynlighetene som inngår i (3.4.1) og (3.4.2) får vi ved å løse dem ut av disse modell-ligningene.

Alle modell-ligningene er på formen

$$\ln\left(\frac{P(V=1)}{1-P(V=1)}\right) = hs,$$

der V er en 0/1-variabel, hos oss Y_j, R^1 eller R_j^2, og hs (høyre side) er en funksjon av kovariater og parametre. Dette gir

$$P(V=1)=\frac{e^{hs}}{1+e^{hs}} \text{ og } P(V=0)=\frac{1}{1+e^{hs}}$$

Dette kan sammenfattes til $P(V=v)=\frac{e^{v\cdot hs}}{1+e^{hs}}$ for $v=0,1$ (3.4.3)

Ved hjelp av (3.4.1), (3.4.2) og (3.4.3) kan vi nå skrive p_{il} som en funksjon av β-ene, φ-ene og ψ-ene. Vi finner $\hat{p}_{il}$ ved å erstatte disse parametrene med sine respektive maksimum-likelihood-estimater i denne funksjonen.

2. Modell 3 for spørsmålet om undersysselsetting sett under ett:
For å uttrykke p_{il} som en funksjon av β-ene, φ-ene og ψ-ene, har vi her tatt utgangspunkt i de 16 gruppene beskrevet i tabell A1.2 i Appendiks A. De 14 første gruppene utgjør målgruppen. Under utarbeidingen av likelihoodfunksjonen for modell 3, har vi funnet modellsannsynlighetene for å være i hver enkelt gruppe, gitt kovariatverdier. Disse sannsynlighetene har vi kalt $P_1,\ldots,P_{14}$, og de står skrevet ut i Appendiks A. Vi finner så p_{il}- ene ved å legge sammen sannsynlighetene (P-ene) til grupper med samme l-verdi. For eksempel, for å finne p_{i1} må vi først finne hvilke grupper som har $l=1$, dvs. $R_u^2=0$, $R^1=0$ og $Y_u=-$. Fra tabell A1.2 ser vi at dette må være gruppe 1, 3 og 4. Dermed er $p_{i1}=P_1+P_3+P_4$.

Pearsons kjikvadratobservator for målgruppen er gitt ved

$$\chi_m^2=\sum_{i=1}^{m_1}\sum_{l=1}^{6}(r_{il}-n_i\hat{p}_{il})^2/n_i\hat{p}_{il}$$

For å kunne bruke denne observatoren til modelltesting, må vi vite hvilken fordeling den har når modellen er riktig. Ut i fra standard teori kan vi si at dersom

1. $\hat{p}_{il}$-ene er ML-estimater basert kun på observasjonene i målgruppen, og
2. Forventet antall observasjoner i hver gruppe, altså $n_i p_{il}$, er større enn 5 i de fleste gruppene,

så er χ_m^2 tilnærmet kjikvadratfordelt med antall frihetsgrader lik $5m_1$ minus antall ukjente parametre i modellen. De to forutsetningene over er imidlertid ikke oppfyllt i vårt tilfelle. Forutsetning 1 er brutt fordi vi bruker hele datasettet når vi maksimerer likelihood-funksjonen, ikke bare data fra målgruppen. Personene utenfor målgruppen påvirker estimatene for φ-ene, og dermed også $\hat{p}_{il}$-ene. Vi regner med at dette fører til at χ_m^2 får en noe høyere forventning, fordi når φ-ene er estimert fra hele utvalget så kan disse parametrene nærmest regnes som kjente konstanter. Dermed reduseres antall "ukjente" parametre og antall frihetsgrader (som er lik forventningen til χ_m^2) øker.

Vi får også problemer med forutsetning 2, men det kan avhjelpes endel ved å slå sammen grupper.

Personer utenfor målgruppen
Her er gruppene delt inn etter kovariatvektoren (z_1, x_2, x_3) og den stokastiske variabelen R^1.
Vi indekserer de forskjellige kovariatvektorene (z_1, x_2, x_3) som forekommer blant personer utenfor målgruppen fra $k = 1$ til m_2, og lar

s_{kh} være antall personer utenfor målgruppen med kovariatvektor k og $R^1 = h$, $h = 0,1$

o_k være antall personer utenfor målgruppen med kovariatvektor k

q_{kh} være modellsannsynligheten for at en person med kovariatvektor k har $R^1 = h$

$\hat{q}_{kh}$ være MLE for q_{kh} under modellen

Vi finner $\hat{q}_{kh}$ på samme måte som $\hat{p}_{il}$:

$q_{kh} = P(R^1 = h | k = (z_1, x_2, x_3))$, som kan uttrykkes som en funksjon av φ-ene ved hjelp av (3.4.3). Vi bytter så ut φ-ene med de tilsvarende ML-estimatene.

Pearsons kjikvadratobservator for personer utenfor målgruppen er gitt ved

$$\chi^2_{mc} = \sum_{k=1}^{m_2} \sum_{h=0}^{1} (s_{kh} - o_k \hat{q}_{kh})^2 / o_k \hat{q}_{kh}$$

Hvis $\hat{q}_{kh}$-ene er ML-estimater basert kun på observasjonene utenfor målgruppen, og hvis $o_k q_{kh}$ er større enn 5 i de fleste gruppene, vil χ^2_{mc} være tilnærmet kjikvadratfordelt med antall frihetsgrader lik m_2 minus antall parametre i modellen for R^1. Også her er den første av disse forutsetningene brutt, siden data fra målgruppen påvirker $\hat{\varphi}$-ene. Den andre forutsetningen er imidlertid oppfylt her.

Samlet observator
Vi legger sammen observatorene i og utenfor målgruppen

$$\chi^2 = \chi^2_m + \chi^2_{mc}$$

Hvis vi antar at χ^2_m og χ^2_{mc} virkelig er kjikvadratfordelte, er χ^2 tilnærmet kjikvadratfordelt med antall frihetsgrader lik summen av frihetsgradene til χ^2_m og χ^2_{mc}. Siden vi ikke er sikre på fordelingene til χ^2_m og χ^2_{mc}, kommer vi til å gjøre simuleringsstudier for å sjekke fordelingen til χ^2.

3.4.2. Likelihood-kvote test
Observatoren i denne testen er deviansen,

$$D = -2 \ln\left(\frac{\text{likelihood nåværende modell}}{\text{likelihood mettet modell}} \right)$$

Likelihood-ene i uttrykket skal evalueres i maximum-likelihood-estimatet.

Vi må finne likelihooden i nåværende og mettet modell. Som tidligere indekserer vi de forskjellige kovariatvektorene (z_1, x_1, x_2, x_3) som forekommer i målgruppen fra $i = 1,\ldots,m_1$, og de seks mulige verdiene for vektoren (R_j^2, R^1, Y_j) fra $l = 1,\ldots,6$, i samme rekkefølge som de står listet i tabell 3.4.1. Kovariatvektorene på formen (z_1, x_2, x_3) som forekommer utenfor målgruppen indekserer vi fra $k = 1$ til m_2.

Vi minner om at

p_{il} er sannsynligheten for at en person i målgruppen med kovariatvektor i har $(R_j^2, R^1, Y_j) = l$
r_{il} er antall personer i målgruppen med $(R_j^2, R^1, Y_j) = l$ og kovariatvektor i.
n_i er antall personer i målgruppen med kovariatvektor i.

q_{kh} er sannsynligheten for at en person med kovariatvektor k har $R^1 = h$
s_{kh} er antall personer utenfor målgruppen med $R^1 = h$ og kovariatvektor k.

Likelihoodfunksjonen kan da skrives som

$$l() = \prod_{i=1}^{m_1}\prod_{l=1}^{6}(p_{il})^{r_{il}} \cdot \prod_{k=1}^{m_2}\prod_{h=0}^{1}(q_{kh})^{s_{kh}}$$

Likelihood for nåværende modell fås ved å erstatte p_{il} og q_{kh} med maximum-likelihood-estimatene for p_{il} og q_{kh} under nåværende modell, kall disse $\hat{p}_{il}$ og $\hat{q}_{kh}$.

MLE i mettet modell:

Vi maksimerer $\ln(l)$ under følgende bibetingelser:

Opplagte bibetingelser:

$A_i : \sum_{l=1}^{6} p_{il} = 1$ for alle i, (m_1 bibetingelser)

$B_k : q_{k0} + q_{k1} = 1$ for alle k, (m_2 bibetingelser)

I tillegg må vi ha bibetingelser som sikrer at sannsynligheten for at $R^1 = 1$ gitt (z_1, x_2, x_3) er den samme i og utenfor målgruppen, og uavhengig av alder:

For hver $i = (z_1, x_1, x_2, x_3)$ skal vi ha

$C_i : p_{i2} + p_{i5} + p_{i6} = q_{k(i),1}$, der $k(i) = (z_1, x_2, x_3)$, (m_1 bibetingelser)

Vi bruker Lagranges metode og maksimerer funksjonen

$$F_\lambda(\mathbf{p},\mathbf{q}) = \ln(l(\mathbf{p},\mathbf{q})) + \sum_{i=1}^{m_1} \lambda_{A_i} (\sum_{l=1}^{6} p_{il} - 1) + \sum_{k=1}^{m_2} \lambda_{B_k} (q_{k0} + q_{k1} - 1) + \sum_{i=1}^{m_1} \lambda_{C_i} (p_{i2} + p_{i5} + p_{i6} - q_{k(i),1})$$

Her er $\mathbf{p} = (p_{il} : i = 1,\ldots,m_1, l = 1,\ldots,6)$ og $\mathbf{q} = (q_{kh} : k = 1,\ldots,m_2, h = 0,1)$.

Vi deriverer F_λ:

$$\frac{\partial F_\lambda}{\partial p_{il}} = \frac{r_{il}}{p_{il}} + \lambda_{A_i} + \lambda_{C_i}, \text{ for } l = 2,5,6$$

$$\frac{\partial F_\lambda}{\partial p_{il}} = \frac{r_{il}}{p_{il}} + \lambda_{A_i}, \text{ for } l = 1,3,4$$

$$\frac{\partial F_\lambda}{\partial q_{k0}} = \frac{s_{k0}}{q_{k0}} + \lambda_{B_k}$$

$$\frac{\partial F_\lambda}{\partial q_{k1}} = \frac{s_{k1}}{q_{k1}} + \lambda_{B_k} - \sum_{i:k(i)=k} \lambda_{C_i}$$

Vi setter de deriverte lik null, og ender opp med følgende sett av likninger:

$$D1_{il}: r_{il} + p_{il}(\lambda_{A_i} + \lambda_{C_i}) = 0, \text{ for } l = 2,5,6$$

$$D2_{il}: r_{il} + p_{il}\lambda_{A_i} = 0, \text{ for } l = 1,3,4$$

$$D3_{k0}: s_{k0} + q_{k0}\lambda_{B_k} = 0$$

$$D4_{k1}: s_{k1} + q_{k1}(\lambda_{B_k} - \sum_{i:k(i)=k} \lambda_{C_i}) = 0$$

Vi skal finne $\mathbf{p},\mathbf{q}$ slik at likningene over, samt bibetingelsene A_i, B_k og C_i er oppfylt.

La k være gitt.

Vi summerer $D1$ over $l = 2,5,6$ og over de i som er slik at $k(i) = k$:

$$\sum_{i:k(i)=k} \sum_{l=2,5,6} r_{il} + \sum_{i:k(i)=k} (\lambda_{A_i} + \lambda_{C_i}) \sum_{l=2,5,6} p_{il} = 0$$

Fra C_i har vi at $\sum_{l=2,5,6} p_{il} = q_{k(i),1}$, så

$$\sum_{i:k(i)=k}\sum_{l=2,5,6} r_{il} + q_{k1}\sum_{i:k(i)=k}(\lambda_{A_i} + \lambda_{C_i}) = 0 \qquad (3.4.4)$$

På samme måte summerer vi $D2$ over $l = 1,3,4$ og over de i slik at $k(i) = k$, og bruker at $\sum_{l=1,3,4} p_{il} = q_{k(i),0}$ (Dette er en konsekvens av bibetingelsene).

$$\sum_{i:k(i)=k}\sum_{l=1,3,4} r_{il} + q_{k0}\sum_{i:k(i)=k}\lambda_{A_i} = 0 \qquad (3.4.5)$$

Summerer $D3$ og $D4$, og bruker at $q_{k0} + q_{k1} = 1$:

$$s_{k0} + s_{k1} + \lambda_{B_k} - q_{k1}\sum_{i:k(i)=k}\lambda_{C_i} = 0 \qquad (3.4.6)$$

Summerer (3.4.4), (3.4.5) og (3.4.6):

$$\sum_{i:k(i)=k}\sum_{l=1}^{6} r_{il} + s_{k0} + s_{k1} + \lambda_{B_k} + \sum_{i:k(i)=k}\lambda_{A_i} = 0$$

Multipliserer med q_{k0}:

$$q_{k0}\left[\sum_{i:k(i)=k}\sum_{l=1}^{6} r_{il} + s_{k0} + s_{k1}\right] + q_{k0}\lambda_{B_k} + q_{k0}\sum_{i:k(i)=k}\lambda_{A_i} = 0$$

Fra $D3$ har vi at $q_{k0}\lambda_{B_k} = -s_{k0}$, og fra (3.4.5) får vi at $q_{k0}\sum_{i:k(i)=k}\lambda_{A_i} = -\sum_{i:k(i)=k}\sum_{l=1,3,4} r_{il}$, så

$$q_{k0}\left[\sum_{i:k(i)=k}\sum_{l=1}^{6} r_{il} + s_{k0} + s_{k1}\right] = s_{k0} + \sum_{i:k(i)=k}\sum_{l=1,3,4} r_{il}$$

Løser ut q_{k0}:

$$q_{k0} = \frac{s_{k0} + \sum_{i:k(i)=k}\sum_{l=1,3,4} r_{il}}{\left[\sum_{i:k(i)=k}\sum_{l=1}^{6} r_{il} + s_{k0} + s_{k1}\right]}$$

Vi ser at q_{k0} er andelen med $R^1 = 0$ blant de med kovariat (z_1, x_2, x_3) lik k, i hele utvalget.

Fra B_k får vi da at

$$q_{k1} = 1 - q_{k0}$$

Finner så p_{il}-ene:

Løser ut p_{il} fra $D2$:

$$p_{il} = -\frac{r_{il}}{\lambda_{A_i}}, \text{ for } l = 1,3,4 \tag{3.4.7}$$

Summerer $D2$ over $l = 1,3,4$:

$$\sum_{l=1,3,4} r_{il} + q_{k(i),0}\lambda_{A_i} = 0$$

Løser ut λ_{A_i} og setter inn i (3.4.7):

$$p_{il} = q_{k(i),0}\frac{r_{il}}{\sum_{l=1,3,4} r_{il}}, \text{ for } l = 1,3,4$$

Denne løsningen gjelder bare når $\sum_{l=1,3,4} r_{il} \neq 0$. Hvis $\sum_{l=1,3,4} r_{il} = 0$, er $r_{il} = 0$ for $l = 1,3,4$, og leddet $p_{il}^{r_{il}}$ i likelihoodfunksjonen er 1. Likelihooden er dermed konstant i p_{il}, og vi kan velge en hvilken som helst løsning som oppfyller bibetingelsene, f.eks.

$$p_{il} = \frac{1}{3}q_{k(i),0}, \text{ for } l = 1,3,4 \text{ når } \sum_{l=1,3,4} r_{il} = 0$$

Tilsvarende får vi at

$$p_{il} = q_{k(i),1}\frac{r_{il}}{\sum_{l=2,5,6} r_{il}}, \text{ for } l = 2,5,6 \text{ når } \sum_{l=2,5,6} r_{il} \neq 0$$

$$p_{il} = \frac{1}{3}q_{k(i),1}, \text{ for } l = 2,5,6 \text{ når } \sum_{l=2,5,6} r_{il} = 0$$

Kall maksimum-likelihood-estimatene fra den mettede modellen for $\tilde{p}_{il}, \tilde{q}_{kh}$. Vi får da at

$$\frac{\text{likelihood nåværende modell}}{\text{likelihood mettet modell}} = \frac{\prod_{i=1}^{m_1}\prod_{l=1}^{6}(\hat{p}_{il})^{r_{il}} \cdot \prod_{k=1}^{m_2}\prod_{h=0}^{1}(\hat{q}_{kh})^{s_{kh}}}{\prod_{i=1}^{m_1}\prod_{l=1}^{6}(\tilde{p}_{il})^{r_{il}} \cdot \prod_{k=1}^{m_2}\prod_{h=0}^{1}(\tilde{q}_{kh})^{s_{kh}}} = \prod_{i=1}^{m_1}\prod_{l=1}^{6}(\frac{\hat{p}_{il}}{\tilde{p}_{il}})^{r_{il}} \cdot \prod_{k=1}^{m_2}\prod_{h=0}^{1}(\frac{\hat{q}_{kh}}{\tilde{q}_{kh}})^{s_{kh}}$$

$$D = -2\ln\left(\frac{\text{likelihood nåværende modell}}{\text{likelihood mettet modell}}\right) = -2\left(\sum_{i=1}^{m_1}\sum_{l=1}^{6} r_{il}\ln(\frac{\hat{p}_{il}}{\widetilde{p}_{il}}) + \sum_{k=1}^{m_2}\sum_{h=0}^{1} s_{kh}\ln(\frac{\hat{q}_{kh}}{\widetilde{q}_{kh}})\right)$$

Hvis noen av $\widetilde{p}_{il}$-ene eller $\widetilde{q}_{kh}$-ene er null, vil formlene over inneholde inneholde udefinerte uttrykk av typen $\frac{\hat{p}_{il}}{0}$. Dette er bare et notasjonsproblem, fordi dersom $\widetilde{p}_{il}$ (evt. $\widetilde{q}_{kh}$) er null, må r_{il} (evt. s_{kh}) også være null, og grupper der r_{il} (evt. s_{kh}) er null trenger ikke å være representert med noe ledd i likelihoodfunksjonen i utgangspunktet. Vi tolker derfor udefinerte uttrykk slik at de ikke får noen innvirkning på uttrykket, dvs. $0\ln(\frac{\hat{p}_{il}}{0})$ tolkes som 0.

Vi antar foreløpig at D er tilnærmet kji-kvadratfordelt med samme antall frihetsgrader som χ^2. Denne antagelsen vil bli sjekket ved simulering.

3.4.3. Modifiserte tester for modelltilpasning

I praksis har vi måttet modifisere testene over litt, fordi vi fikk mange grupper med få observasjoner. Utenfor målgruppen blir gruppene store nok, det er innenfor målgruppen oppsplittingen blir for stor. For hver i er det her i utgangspunktet seks mulige verdier på l. I de modifiserte testene slår vi sammen noen av disse verdiene. Hvor mye som slås sammen, avhenger av antall personer med kovariatvektor i. For lesbarhetens skyld gjentar vi her tabell 3.4.1:

l	R_j^2	R^1	Y_j
1	0	0	-
2	0	1	-
3	1	0	0
4	1	0	1
5	1	1	0
6	1	1	1

I tabellen under har vi laget en oversikt over hvilke l-verdier som slås sammen.

Antall personer med kovariatvektor i	**Slår sammen l -verdier**	
	Sp. 18 og sp. om undersysselsetting	**Sp. 19, 23a, 24, 25 og 62**
1-30	1,2,3,4,6	1,2,3,4,5
31-100	1,2,3,4	1,2,3,4
101 eller mer	1,3,4	1,3,4

Det er få personer som har svart etter oppfølgning (l lik 1, 3 eller 4), så disse slår vi sammen uansett hvor mange personer som har kovariatvektor i. I strata med mellom 1 og 30 personer skiller vi bare mellom den største gruppen og de andre. For spørsmål 18 og for spørsmålet om undersysselsetting sett under ett er den største gruppen de som har svart nei på spørsmålet ved første henvendelse (l lik 5). For de andre spørsmålene er det de som har svart ja på spørsmålet ved første henvendelse (l lik 6). I strata med mellom 31 og 100 personer skiller vi i tillegg ut de som har svart ja på spørsmålet ved første henvendelse (l lik 6) for spørsmål 18 og for spørsmålet om undersysselsetting sett under ett, og de som har svart nei på første henvendelse (l lik 5) for de andre spørsmålene. I de største strataene kan vi skille mellom alle de tre gruppene som har $R^1 = 1$.

3.4.4. Utregning av antall frihetsgrader

Når vi regner ut antall frihetsgrader for χ^2_m, χ^2_{mc} og χ^2, antar vi at forutsetningene for at disse skal være kji-kvadratfordelt er oppfyllt. Vi regner altså som om $\hat{p}_{il}$-ene var ML-estimater bare basert på målgruppen, $\hat{q}_{kh}$-ene var ML-estimater bare basert på observasjonene utenfor målgruppen, og forventet antall observasjoner var større enn fem i alle grupper. Det antallet vi kommer frem til på denne måten, kaller vi observatorens teoretiske antall frihetsgrader. I neste avsnitt vil vi gjøre simuleringer for å undersøke om observatorenes faktiske fordeling ligger nær den teoretiske.

Tabellene under inneholder antall frihetsgrader for de forskjellige testene, og størrelsene vi trenger for å regne dem ut. Etter tabellene er det en forklaring på hva kolonnene inneholder.

Spørsmål 18

	Antall mulige kovariat-komb.	**Antall kov.komb. i datasettet**	**Antall utfall i hver rekke**	**Antall grupper**	**Antall frie para-metre**	**Antall uavh. grupper**	**Teoretisk antall frihets-grader**
Målgr. , umod.	80	73	6	438	53	365	312
Målgr., mod.	80	45 18 10	2 3 4	184	53	111	58
Utenfor målgr.	20	20	2	40	16	20	4
Totalt, umod.							316
Totalt, mod							62

Spørsmål 19, 23a og 24

	Antall mulige kovariat-komb.	**Antall kov.komb. i datasettet**	**Antall utfall i hver rekke**	**Antall grupper**	**Antall frie para-metre**	**Antall uavh. grupper**	**Teoretisk antall frihets-grader**
Målgr. , umod.	60	51	6	306	45	255	210
Målgr., mod.	60	44 7 0	2 3 4	109	45	58	13
Utenfor målgr.	20	20	2	40	16	20	4
Totalt, umod.							214
Totalt, mod							17

Undersysselsetting, modell 1

	Antall mulige kovariat-komb.	**Antall kov.komb. i datasettet**	**Antall utfall i hver rekke**	**Antall grupper**	**Antall frie para-metre**	**Antall uavh. grupper**	**Teoretisk antall frihets-grader**
Målgr. , umod.	60	54	6	324	45	270	225
Målgr., mod.	60	26 18 10	2 3 4	146	45	92	47
Utenfor målgr.	20	20	2	40	16	20	4
Totalt, umod.							229
Totalt, mod							51

Undersysselsetting, modell 2

	Antall mulige kovariat-komb.	Antall kov.komb. i datasettet	Antall utfall i hver rekke	Antall grupper	Antall frie para-metre	Antall uavh. grupper	Teoretisk antall frihets-grader
Målgr. , umod.	60	54	6	324	48	270	222
Målgr., mod.	60	26 18 10	2 3 4	146	48	92	44
Utenfor målgr.	20	20	2	40	16	20	4
Totalt, umod.							226
Totalt, mod							48

Undersysselsetting, modell 3

	Antall mulige kovariat-komb.	Antall kov.komb. i datasettet	Antall utfall i hver rekke	Antall grupper	Antall frie para-metre	Antall uavh. grupper	Teoretisk antall frihets-grader
Målgr. , umod.	60	54	6	324	75	270	195
Målgr., mod.	60	26 18 10	2 3 4	146	75	92	17
Utenfor målgr.	20	20	2	40	16	20	4
Totalt, umod.							199
Totalt, mod							21

Spørsmål 25

	Antall mulige kovariat-komb.	Antall kov.komb. i datasettet	Antall utfall i hver rekke	Antall grupper	Antall frie para-metre	Antall uavh. grupper	Teoretisk antall frihets-grader
Målgr. , umod.	60	58	6	348	45	290	245
Målgr., mod.	60	32 6 20	2 3 4	162	45	104	59
Utenfor målgr.	20	20	2	40	16	20	4
Totalt, umod.							249
Totalt, mod							63

Spørsmål 62

	Antall mulige kovariat-komb.	Antall kov.komb. i datasettet	Antall utfall i hver rekke	Antall grupper	Antall frie para-metre	Antall uavh. grupper	Teoretisk antall frihets-grader
Målgr. , umod.	60	49	6	294	45	245	200
Målgr., mod.	60	43 6 0	2 3 4	104	45	55	10
Utenfor målgr.	20	20	2	40	16	20	4
Totalt, umod.							204
Totalt, mod							14

Antall mulige kovariatkombinasjoner:
I målgruppen er forklaringsvariablene sivilstand, alder, kjønn og region. På spørsmål 18 har vi 4 aldersgrupper, så vi får 2x4x2x5=80 mulige kombinasjoner. På de andre spørsmålene har vi tre alders-

grupper, så vi får 2x3x2x5=60 mulige kovariatkombinasjoner. Utenfor målgruppen er forklaringsvariablene sivilstand, kjønn og region, som gir 2x2x5=20 mulige kombinasjoner for alle spørsmålene.

Antall kovariatkombinasjoner i datasettet:
Antall kovariatkombinasjoner som er representert i datasettet. For den modifiserte testen oppgir vi hvor mange av kombinasjonene som har hhv. 1-30 personer, 31-100 personer og over 100 personer.

Antall utfall i hver rekke:
Hver kovariatkombinasjon som forekommer i datasettet kan ses på som en forsøksrekke, der personene med den bestemte kovariatkombinasjonen representerer forsøkene. I den umodifiserte testen har hvert forsøk seks mulige utfall, et for hver verdi av l. I den modifiserte testen er antall mulige utfall 2, 3 eller 4 for kovariatkombinasjoner med hhv. 1-30 personer, 31-100 personer og over 100 personer. Utenfor målgruppen er det to mulige utfall, $R^1 = 0$ og $R^1 = 1$.

Antall grupper:
Dette er (antall forsøksrekker) x (antall utfall i hver rekke). For spørsmål 18 får vi for eksempel i den umodifiserte testen 73x6=438 grupper, og i den modifiserte testen 45x2+18x3+10x4=184 grupper.

Antall frie parametre:
For hver modell teller vi opp hvor mange frie parametre som hører til hver variabel.
Når vi modellerer spørsmålene enkeltvis får vi:

	Populasjons-modell sp. 18	**Populasjons-modell sp. 19, 23a, 24, 25 og 62**	**Modell for R^1**	**Modell for R^2**
Konstantledd	1	1	1	1
Alder	3	2	-	-
Kjønn	1	1	1	1
Region	4	4	4	4
Sivilstand	-	-	1	1
y	-	-	-	1
R^1	-	-	-	1
Alder*Kjønn	3	2	-	-
Alder*Region	12	6	-	-
Kjønn*Region	4	4	4	-
Sivilstand*Kjønn	-	-	1	-
Sivilstand*Region	-	-	4	-
Sum	28	20	16	9

Dette gir totalt 28+16+9=53 frie parametre for spørsmål 18, og 20+16+9=45 frie parametre for de andre spørsmålene.

I de tre modellene for undersysselsetting brukes det tre aldersgrupper i populasjonsmodellen. For modell 1 får vi da 20+16+9=45 frie parametre. I modell 2 har vi de samme parametrene som i modell 1, og i tillegg har vi to parametre i modellen for R^2_{18}, og en i modellen for R^2_{19}, så vi får 45+3=48 frie parametre. Modell 3 har 20 frie parametre i hver av modellene for Y_{18} og Y_{19}, og 8 frie parametre i modellen for Y_{23} (siden kryssleddene er utelatt). Det er 9 frie parametre i modellen for R^2_{18}, og 1 i hver av modellene for R^2_{19} og R^2_{23}. Dermed får vi 20+20+8+9+1+1+16=75 frie parametre.

Antall uavhengige grupper:
Hvis vi har en forsøksrekke der hvert forsøk har n mulige utfall, og dersom vi vet hvor mange forsøk som resulterte i utfall $1,2,\ldots,n-1$, vet vi at resten av forsøkene resulterte i utfall n. Denne avhengigheten må vi ta hensyn til når vi regner på frihetsgradene. Med antall uavhengige grupper mener vi (antall forsøksrekker) x (antall utfall i hver rekke-1). For spørsmål 18 får for eksempel for den umodifiserte testen 73x(6-1)=365, og for den modifiserte 45x(2-1)+18x(3-1)+10x(4-1)=111.

Antall frihetsgrader:
Dette er antall uavhengige grupper minus antall frie parametre. Vi finner frihetsgradene til den samlede observatoren χ^2 ved å legge sammen frihetsgradene i og utenfor målgruppen. For spørsmål 18 får vi da for den umodifiserte testen 312+4=316, og for den modifiserte 58+4=62.

3.4.5. Fordelingen til testobservatorene

Spørsmål 18
For å sjekke fordelingen til testobservatorene under modellen, har vi brukt de tusen simulerte datasettene beskrevet i avsnitt 3.3, og regnet ut både den umodifiserte og den modifiserte testen for hvert av dem. Vi ønsker også å undersøke om fordelingen til de forskjellige testobservatorene endrer seg når parameterverdiene i (2.2.3), (3.1.1) og (3.1.2) endres. Vi har derfor på tilsvarende måte simulert 1000 datasett med to andre sett av parametre i modellen: parametersettet som fremkommer ved å legge til 0,1 på alle ML-estimatene fra det opprinnelige datasettet og parametersettet som fremkommer ved å trekke fra 0,1 på alle ML-estimatene fra det opprinnelige datasettet. Tabellene under gir en oversikt over empirisk gjennomsnitt og standardavvik for de forskjellige testobservatorene for hvert av de tre parametersettene. Disse sammenligner vi med forventing og standardavvik i de teoretiske fordelingene.

Simuleringsparametre = MLE fra opprinnelig datasett

	Teoretisk		**Pearson**		**Devians**	
	For-ventning	Std.	Gj.snitt	Std.	Gj.snitt	Std.
Umodifisert test (total)	316	25,14	320,56	66,83	208,39	18,89
Modifisert test (total)	62	11,14	90,24	13,10	65,55	12,27

Simuleringsparametre = MLE fra opprinnelig datasett + 0.1

	Teoretisk		**Pearson**		**Devians**	
	For-ventning	Std.	Gj.snitt	Std.	Gj.snitt	Std.
Umodifisert test (total)	316	25,14	322,17	84,16	196,74	18,25
Modifisert test (total)	62	11,14	88,61	13,41	66,24	12,39

Simuleringsparametre = MLE fra opprinnelig datasett - 0.1

	Teoretisk		**Pearson**		**Devians**	
	For-ventning	Std.	Gj.snitt	Std.	Gj.snitt	Std.
Umodifisert test (total)	316	25,14	314,12	48,07	214,33	19,40
Modifisert test (total)	62	11,14	91,92	12,97	65,36	12,36

Ingen av de umodifiserte testobservatorenes fordeling er særlig nær den teoretiske fordelingen. Pearsons observator har altfor høyt standardavvik, og for deviansen sin del er både forventning og standardavvik langt unna. Deviansen kan heller se ut til å være tilnærmet kji-kvadratfordelt med omkring 200 frihetsgrader snarere enn 316. Med 200 frihetsgrader blir standardavviket 20

($=\sqrt{2 \cdot 200}$). Når det gjelder de modifiserte testene, ser vi at Pearsons observator har mye høyere forventning enn den teoretiske. En forventning på 90 medfører et standardavvik på 13,4, så en kjikvadratfordeling med 90 frihetsgrader er en mulig tilnærmet fordeling. De beste resultatene oppnår vi imidlertid med den modifiserte deviansen. Her er den empiriske forventningen og det empiriske standardavviket bare litt høyere enn de tilsvarende teoretiske verdiene. Forventning og standardavvik ser også ut til å være nokså uavhengige av parametrene i modellen. Vi kommer derfor heretter til å bruke den modifiserte deviansen som observator i modelltilpasningstestene.

Spørsmål 19, 23a, 24, 25, 62 og undersysselsetting sett under ett
For disse spørsmålene undersøker vi bare fordelingen til den modifiserte deviansen når simuleringsparametrene er lik MLE fra opprinnelig datasett.

Spørsmål	**Teoretisk**		**Empirisk**	
	Forventning	Std.	Gj.snitt	Std.
19	17	5,83	33,98	9,02
23a	17	5,83	29,75	9,08
Undersyss., Mod. 1	51	10,10	44,69	10,81
Undersyss., Mod. 2	48	9,80	55,58	11,08
Undersyss., Mod. 3	21	6,48	54,63	10,52
24	17	5,83	35,70	9,41
25	63	11,22	56,20	10,41
62	14	5,29	36,45	9,19

På spørsmål 18 var gjennomsnitt og standardavvik for de modifiserte deviansene beregnet på de tusen simulerte datasettene nokså like henholdsvis teoretisk forventning og teoretisk standardavvik. Som det fremgår av tabellen over er dette ikke tilfellet for de andre spørsmålene. En årsak til dette kan være at vi får mange små grupper til tross for at vi har slått sammen mye. Vi kan følgelig ikke anta at den modifiserte deviansen følger den teoretiske fordelingen. Når vi skal teste modelltilpasningen på det ekte datasettet må vi derfor bruke empiriske P-verdier. Vi tar da utgangspunkt i de tusen simulerte datasettene. Hvis den observerte verdien av den modifiserte deviansen er D, er den empiriske P-verdien lik andelen simulerte datasett som har en modifisert devians større eller lik D.

3.4.6. Resultat av modelltilpasningstestene
Tabellen under viser de modifiserte deviansene beregnet på det ekte datasettet, og de tilhørende empiriske og teoretiske P-verdiene.

Spørsmål	**Modifisert devians**	**Empirisk P-verdi**	**Teoretisk P-verdi**
18	93,49	0,0160	0,0061
19	31,03	0,6120	0,0198
23a	61,76	0,0020	$5{,}368 \times 10^{-7}$
Undersyss., Mod. 1	84,71	0,0010	0,0021
Undersyss., Mod. 2	84,79	0,0050	$8{,}351 \times 10^{-4}$
Undersyss., Mod. 3	79,88	0,0130	$8{,}456 \times 10^{-9}$
24	75,26	0,0000	$2{,}631 \times 10^{-9}$
25	162,74	0,0000	$9{,}167 \times 10^{-11}$
62	33,28	0,5990	0,0026

På spørsmål 19 og 62 er P-verdiene høye, mens på de andre spørsmålene får vi svært små P-verdier. Små P-verdier er imidlertid ikke unormalt for denne type modeller. Modellen kan likevel være egnet som et utgangspunkt for imputering. Vi må også huske på at lav P-verdi ikke nødvendigvis betyr dår-

lig tilpasning når utvalgsstørrelsen er stor. Ved sammenligning av modellene 1, 2 og 3, ser modell 3 ut til å gi litt bedre tilpasning.

I neste avsnitt undersøker vi om vi kan oppnå noen vesentlig forbedring av tilpasningen ved å gjøre endringer i populasjonsmodellen og modellen for partielt frafall.

3.5. Modellvurderinger. Konklusjon.

Vi tar her for oss ett og ett spørsmål, og undersøker ulike alternative modeller. For hvert alternativ regner vi ut modifisert devians og empirisk P-verdi. Fra populasjonsmodellene fjerner vi de kryssleddene som ikke kom med i den foreløpige modellen for spørsmålet (kap 2.2). Fra modellene for partielt frafall fjerner vi (som hovedregel) en forklaringsvariabel av gangen. For hvert spørsmål presenteres resultatene i en tabell. Første linje i tabellen inneholder den opprinnelige modellen, altså den samlede modellen for populasjon og responsmekanisme slik den er beskrevet i avsnitt 3.1.2.

Spørsmål 18

Her har vi, i tillegg til de empiriske P-verdiene, også tatt med en kolonne med teoretiske P-verdier. Disse forutsetter at observatorene er kjikvadratfordelte med det antall frihetsgrader som står i fjerde kolonne.

Parametre som fjernes fra pop. modellen	**Parametre som fjernes fra modell for part. frafall**	**Modifisert devians**	**Antall frihets-grader**	**Teoretisk P-verdi**	**Empirisk P-verdi**
ingen	ingen	93,49	62	0,0059	0,016
alder*boreg kjønn*boreg	ingen	112,45	78	0,0064	0,008
alder*boreg kjønn*boreg	sivilstand	117,85	79	0,0030	0,003
alder*boreg kjønn*boreg	kjønn	119,85	79	0,0020	0,003
alder*boreg kjønn*boreg	boregion	114,50	82	0,0103	0,008
alder*boreg kjønn*boreg	y-verdi	114,449	79	0,0056	0,009
alder*boreg kjønn*boreg	sivilstand boregion	119,60	83	0,0053	0,002

Ingen av endringene gir noen markant forbedring av den teoretiske P-verdien. Størst forbedring får vi ved å fjerne boregion fra R^2-modellen, i tillegg til de to kryssleddene vi fjerner fra populasjonsmodellen.
Ser vi på de empiriske P-verdiene, kan det se ut som den modellen vi har valgt er bedre enn alle alternativene.

Spørsmål 19

Parametre som fjernes fra pop. modellen	Parametre som fjernes fra modell for part. frafall	Modifisert devians	Empirisk P-verdi
ingen	ingen	31,03	0,612
alder*kjønn kjønn*boreg	ingen	41,47	0,410
alder*kjønn kjønn*boreg	sivilstand	41,40	0,426
alder*kjønn kjønn*boreg	kjønn	41,53	0,409
alder*kjønn kjønn*boreg	boregion	43,49	0,401
alder*kjønn kjønn*boreg	y-verdi	41,43	0,415

Ingen av de alternative modellene gir bedre tilpasning enn modellen den opprinnelige modellen.

Spørsmål 23a

Parametre som fjernes fra pop. modellen	Parametre som fjernes fra modell for part. frafall	Modifisert devians	Empirisk P-verdi
ingen	ingen	61,76	0,002
alder*kjønn kjønn*boreg alder*boreg	ingen	67,11	0,006
alder*kjønn kjønn*boreg alder*boreg	sivilstand	68,08	0,005
alder*kjønn kjønn*boreg alder*boreg	kjønn	66,90	0,007
alder*kjønn kjønn*boreg alder*boreg	boregion	67,15	0,010
alder*kjønn kjønn*boreg alder*boreg	y-verdi	66,64	0,009

Her ser det ut til at de alternative modellene gir litt bedre tilpasning.

Undersysselsetting, modell 1

Parametre som fjernes fra pop. modellen	Parametre som fjernes fra modell for part. frafall	Modifisert devians	Empirisk P-verdi
ingen	ingen	84,71	0,001
alder*kjønn alder*boreg	ingen	89,79	0,002
alder*kjønn alder*boreg	sivilstand	93,29	0,002
alder*kjønn alder*boreg	kjønn	95,64	0,002
alder*kjønn alder*boreg	boregion	95,74	0,001
alder*kjønn alder*boreg	y-verdi	92,67	0,002

Når vi gjør 1000 simuleringer, blir den empiriske P-verdien alltid et heltallig antall tusendeler. En empirisk P-verdi på 0,002 betyr for eksempel at 2 av de simulerte datasettene har en modifisert devians som er større eller lik den modifiserte deviansen i det reelle datasettet. Hadde vi gjort 1000 andre simuleringer kunne vi fått en empirisk P-verdi på 0,001 eller 0,003. Tusen simuleringer er derfor ikke nok til å skille mellom P-verdier som er så små som her.

Det vi kan si er at de empiriske P-verdiene er av samme størrelsesorden for alle alternativene. Vi har derfor ingen grunn til å anta at noen av alternativene er vesentlig bedre enn den opprinnelige modellen.

Undersysselsetting, modell 2

Modellene for R^2_{18} og R^2_{19} inneholder allerede så få parametre, så vi fjerner ingen ledd fra disse modellene.

Parametre som fjernes fra pop. modellen	Parametre som fjernes fra modell for R^2_u	Modifisert devians	Empirisk P-verdi
ingen	ingen	84,79	0,005
alder*kjønn alder*boreg	ingen	89,78	0,021
alder*kjønn alder*boreg	sivilstand	93,55	0,006
alder*kjønn alder*boreg	kjønn	95,85	0,005
alder*kjønn alder*boreg	boregion	95,66	0,010
alder*kjønn alder*boreg	y-verdi	92,71	0,011

Det gir en viss forbedring å fjerne to kryssledd fra populasjonsmodellen.

Undersysselsetting, modell 3

Modellene for R_{19}^2 og R_{23}^2 inneholder bare konstantledd, så det er ikke aktuelt å fjerne noen ledd fra disse modellene. I modellen for Y_{23} er kryssleddene allerede fjernet, så vi fjerner ingen flere ledd fra denne modellen heller.

Parametre som fjernes fra modell for Y_{18}	**Parametre som fjernes fra modell for Y_{19}**	**Parametre som fjernes fra modell for R_{18}^2**	**Modifisert devians**	**Empirisk P-verdi**
ingen	ingen	ingen	79,88	0,013
alder*boreg kjønn*boreg	alder*kjønn kjønn*boreg	ingen	94,50	0,004
alder*boreg kjønn*boreg	alder*kjønn kjønn*boreg	sivilstand	98,09	0,001
alder*boreg kjønn*boreg	alder*kjønn kjønn*boreg	kjønn	98,50	0,003
alder*boreg kjønn*boreg	alder*kjønn kjønn*boreg	boregion	97,54	0,004
alder*boreg kjønn*boreg	alder*kjønn kjønn*boreg	y-verdi	97,80	0,006

De alternative modellene gir ingen bedring i tilpasningen.

Spørsmål 24

Parametre som fjernes fra pop. modellen	**Parametre som fjernes fra modell for part. frafall**	**Modifisert devians**	**Empirisk P-verdi**
ingen	ingen	75,26	0,000
alder*boreg kjønn*boreg	ingen	83,39	0,000
alder*boreg kjønn*boreg	sivilstand	84,64	0,000
alder*boreg kjønn*boreg	kjønn	84,38	0,000
alder*boreg kjønn*boreg	boregion	82,62	0,000
alder*boreg kjønn*boreg	y-verdi	83,74	0,000

Vi er her i situasjon der alle de empiriske P-verdiene blir 0. Med tusen simuleringer får vi som tidligere nevnt bare tre desimaler i P-verdien, og det er ikke nok til å se forskjell her.

Spørsmål 25

Parametre som fjernes fra pop. modellen	Parametre som fjernes fra modell for part. frafall	Modifisert devians	Empirisk P-verdi
ingen	ingen	162,74	0,000
alder*boreg	ingen	173,70	0,000
alder*boreg	sivilstand	186,19	0,000
alder*boreg	kjønn	176,39	0,000
alder*boreg	boregion	202,92	0,000
alder*boreg	y-verdi	173,57	0,000

Også her blir alle de empiriske P-verdiene 0.

Spørsmål 62

Parametre som fjernes fra pop. modellen	Parametre som fjernes fra modell for part. frafall	Modifisert devians	Empirisk P-verdi
ingen	ingen	33,28	0,599
alder*boreg	ingen	45,26	0,404
alder*boreg	sivilstand	45,21	0,455
alder*boreg	kjønn	45,87	0,428
alder*boreg	boregion	54,17	0,215
alder*boreg	y-verdi	45,26	0,421

Vi ser at den opprinnelige modellen gir den beste tilpasningen.

Konklusjon
De alternative modellene gir liten eller ingen bedring i tilpasningen, så vi holder fast på den opprinnelige modellen for populasjon og frafallsmekanisme, slik den er beskrevet i kapittel 3.1.

Vi må også foreta et valg mellom modellene 1, 2 eller 3 for undersysselsetting. Det er en fordel at modellen er så enkel som mulig, men det viktigste er hvordan den fungerer ved imputering. Som nevnt tidligere, vil vi ta opp imputering i et senere notat. Vi venter derfor med denne avgjørelsen til da.

Referanser

[1]: Hobæk, Tone: «Justering av partielt frafall i arbeidskraftundersøkelsen (AKU)» Notater 93/34
[2]: Hosmer, D. & Lemeshow, S.: «Applied Logistic Regression» Wiley 1989

[3]: Thomsen, Ib: «Prinsipper og metoder for statistisk sentralbyrås utvalgsundersøkelser», SØS 33, 1991

Litteratur

[1]: Håland, I., Hobæk, T. & Bø,T.P.: «Dokumentasjon av arbeidskraftundersøkelsen (AKU)» Notater 93/27
[2]: Ihå, JHE & TPB: «Dokumentasjon av arbeidskraftundersøkelsen (AKU). Beskrivelse av fil for tabellkjøringer og dokumentasjon av variable» Upublisert notat 2.7.1991
[3]: Boije, L.: «Estimering og Etterstratifisering i AKU» (boi,25.03.96). Upublisert notat.

[4]: Zhang, L.-C: «Dokumentasjonsrapport: Den nye estimeringsmetoden for Arbeidskraftundersøkelsen(AKU)» Notater 98/1
[5]: Sigstad, J.A.: «Stratifisering og imputering» Universitetet i Trondheim, 1993

Appendiks A. Likelihoodfunksjoner

A.1. Likelihoodfunksjonen for enkeltspørsmål

Etter å ha valgt populasjonsmodell og modell for RM ved enhetsfrafall og partielt frafall, kan vi finne maksimum likelihood estimatene (MLE) for de ukjente parametrene ved å maksimere likelihoodfunksjonen, $l()$.

La M_j være målgruppen for spørsmål j. Likelihoodfunksjonen vil være et produkt av fem faktorer, som svarer til følgende fem grupper av personer:

I) Personer i målgruppen for spørsmål j, som svarer ved første henvendelse, og som også svarer på spørsmål j.

$$\text{Gr I} = \{i \in M_j : R_i^1 = 1, R_{ij}^2 = 1\}$$

II) Personer i målgruppen for spørsmål j, som svarer ved første henvendelse, men ikke på spørsmål j. (Partielt frafall på j).

$$\text{Gr II} = \{i \in M_j : R_i^1 = 1, R_{ij}^2 = 0\}$$

III) Personer i målgruppen for spørsmål j, som svarer etter oppfølgning, også på spørsmål j.

$$\text{Gr III} = \{i \in M_j : R_i^1 = 0, R_{ij}^2 = 1\}$$

IV) Personer i målgruppen for spørsmål j, som svarer etter oppfølgning, men ikke på spørsmål j.

$$\text{Gr IV} = \{i \in M_j : R_i^1 = 0, R_{ij}^2 = 0\}$$

V) Personer utenfor målgruppen for spørsmål j.

$$\text{Gr V} = M_j^c$$

Personene i målgruppen for spørsmål j (Gr I-Gr IV) gir informasjon om β_j-ene, ψ_j-ene og φ-ene. Personene utenfor målgruppen (Gr V) gir bare informasjon om φ-ene.

Likelihoodfunksjonen skriver vi i fem linjer, en for hver faktor svarende til de fem gruppene I-V. Vi holder nå j fast, og dropper for enkelhets skyld denne indeksen, slik at $R_{ij}^2 = R_i^2$.

$$l = \prod_{i \in \text{Gr I}} P(Y_i = y_i \cap R_i^1 = 1 \cap R_i^2 = 1 | \mathbf{x}_i, \mathbf{z}_i)$$

$$* \prod_{i \in \text{Gr II}} P(R_i^1 = 1 \cap R_i^2 = 0 | \mathbf{z}_i)$$

$$* \prod_{i \in \text{Gr III}} P(Y_i = y_i \cap R_i^1 = 0 \cap R_i^2 = 1 | \mathbf{x}_i, \mathbf{z}_i)$$

$$* \prod_{i \in \text{Gr IV}} P(R_i^1 = 0 \cap R_i^2 = 0 \mid \mathbf{z}_i)$$

$$* \prod_{i \in \text{Gr V}} P(R_i^1 = r_i^1 | \mathbf{z}_i)$$

som videre kan skrives:

$$l = \prod_{i \in \text{Gr I}} P(Y_i = y_i | \mathbf{x}_i) * P(R_i^1 = 1 | \mathbf{z}_i) * P(R_i^2 = 1 | Y_i = y_i, R_i^1 = 1, \mathbf{z}_i)$$

$$* \prod_{i \in \text{Gr II}} P(R_i^1 = 1 | \mathbf{z}_i) * \sum_{y=0}^{1} P(R_i^2 = 0 | Y_i = y, R_i^1 = 1, \mathbf{z}_i) * P(Y_i = y | \mathbf{x}_i)$$

$$* \prod_{i \in \text{Gr III}} P(Y_i = y_i | \mathbf{x}_i) * P(R_i^1 = 0 | \mathbf{z}_i) * P(R_i^2 = 1 | R_i^1 = 0, Y_i = y_i, \mathbf{z}_i)$$

$$* \prod_{i \in \text{Gr IV}} P(R_i^1 = 0 | \mathbf{z}_i) * \sum_{y=0}^{1} P(R_i^2 = 0 | Y_i = y, R_i^1 = 0, \mathbf{z}_i) * P(Y_i = y | \mathbf{x}_i)$$

$$* \prod_{i \in \text{Gr V}} P(R_i^1 = r_i^1 | \mathbf{z}_i)$$

Legg merke til at l kan skrives som produktet av likelihoodfunksjonen for R^1-observasjonene i hele utvalget og likelihoodfunksjonen for Y- og R^2- observasjonene i målgruppen for spørsmål j, dvs.

$l = l_1(\vec{\phi}) \cdot l_2(\vec{\beta}, \vec{\psi})$, der

$l_1(\vec{\phi}) = \prod_{i \in s} P(R_i^1 = r_i^1 | \mathbf{z}_i)$, der s er unionen av gruppene I-V, og

$$l_2(\vec{\beta}, \vec{\psi}) = \prod_{i \in \text{Gr I}} P(Y_i = y_i | \mathbf{x}_i) * P(R_i^2 = 1 | Y_i = y_i, R_i^1 = 1, \mathbf{z}_i)$$

$$* \prod_{i \in \text{Gr II}} \sum_{y=0}^{1} P(R_i^2 = 0 | Y_i = y, R_i^1 = 1, \mathbf{z}_i) * P(Y_i = y | \mathbf{x}_i)$$

$$* \prod_{i \in \text{Gr III}} P(Y_i = y_i | \mathbf{x}_i) * P(R_i^2 = 1 | R_i^1 = 0, Y_i = y_i, \mathbf{z}_i)$$

$$* \prod_{i \in \text{Gr IV}} \sum_{y=0}^{1} P(R_i^2 = 0 | Y_i = y, R_i^1 = 0, \mathbf{z}_i) * P(Y_i = y | \mathbf{x}_i)$$

Siden $l_1(\vec{\phi})$ er lik for alle spørsmål, blir MLE for ϕ-ene uavhengig av spørsmål.

Under den videre utledningen av likelihoodfunksjonen, og ved programmeringen, holder vi oss av praktiske årsaker til det opprinnelige uttrykket for l, vi faktoriserer altså ikke ut $l_1(\vec{\phi})$. Vi utnytter likevel observasjonen over, ved at vi kun beregner MLE for φ-ene én gang, nemlig under maksimeringen av likelihoodfunksjonen for spørsmål 18. Likelihoodfunksjonen for de andre spørsmålene maksimeres så under bibetingelsen at φ-ene skal være like de φ-ene vi fikk på spørsmål 18.

Før vi kan skrive ut loglikelihoodfunksjonen, må vi definere noen størrelser, som angir antall personer i ulike strata:

yja(x_1, x_2, x_3) = antall personer i gruppe I, i aldersgruppe x_1, med kjønn x_2, i boregion x_3, som svarer ja(=1) på spørsmålet.

ynei(x_1, x_2, x_3) = antall personer i gruppe I , i aldersgruppe x_1 , med kjønn x_2, i boregion x_3, som svarer nei(=0) på spørsmålet.

yja2(z_1, x_2, x_3) = antall personer i gruppe I, med sivilstand z_1, kjønn x_2, i boregion x_3, som svarer ja på spørsmålet.

ynei2(z_1, x_2, x_3) = antall personer i gruppe I, med sivilstand z_1, kjønn x_2, i boregion x_3 som svarer nei på spørsmålet.

pf1(z_1, x_1, x_2, x_3) = antall personer i gruppe II, med sivilstand z_1, i aldersgruppe x_1, med kjønn x_2, i boregion x_3.

jato(x_1, x_2, x_3) = antall personer i gruppe III, i aldersgruppe x_1, med kjønn x_2, i boregion x_3, som svarer ja på spørsmålet.

neito(x_1, x_2, x_3) = antall personer i gruppe III, i aldersgruppe x_1, med kjønn x_2, i boregion x_3, som svarer nei på spørsmålet.

jato2(z_1, x_2, x_3) = antall personer i gruppe III, med sivilstand z_1, kjønn x_2, i boregion x_3, som svarer ja på spørsmålet.

neito2(z_1, x_2, x_3) = antall personer i gruppe III, med sivilstand z_1, kjønn x_2, i boregion x_3, som svarer nei på spørsmålet.

pfo(z_1, x_1, x_2, x_3) = antall personer i gruppe IV, med sivilstand z_1, i aldersgruppe x_1, med kjønn x_2, i boregion x_3.

grfem1(z_1, x_2, x_3) = antall personer i gruppe V, med sivilstand z_1, kjønn x_2, i boregion x_3, som svarer på 1. henvendelse.

grfemo(z_1, x_2, x_3) = antall personer i gruppe V, med sivilstand z_1, kjønn x_2, i boregion x_3, som svarer etter oppfølgning.

La nå

$$\beta(x_1,x_2,x_3)=\beta_0+\sum_{l=1,2,4}\beta_{1l}D_{1l}+\beta_2x_2+\sum_{l=1,3,4,5}\beta_{3l}D_{3l}+\sum_{l=1,2,4}\beta_{4l}x_2D_{1l}+\sum_{l=1,3,4,5}\beta_{5l}x_2D_{3l}+\sum_{l=1,2,4}\sum_{m=1,3,4,5}\beta_{6lm}D_{1l}D_{3m}$$

$$\varphi(z_1,x_2,x_3)=\varphi_0+\varphi_1z_1+\varphi_2x_2+\sum_{l=1,3,4,5}\varphi_{3l}D_{3l}+\varphi_4z_1x_2+\sum_{l=1,3,4,5}\varphi_{5l}z_1D_{3l}+\sum_{l=1,3,4,5}\varphi_{6l}x_2D_{3l}$$

$$\psi(z_1,x_2,x_3,y,r^1)=\psi_0+\psi_1z_1+\psi_2x_2+\sum_{l=1,3,4,5}\psi_{3l}D_{3l}+\psi_4y+\psi_5R^1$$

Tilsvarende som for likelihoodfunksjonen, skriver vi loglikelihoodfunksjonen ut som fem ledd. Vi studerer hvert av de fem leddene for seg:

log(l) = (I) + (II) + (III) + (IV) +(V)

(I) = $\sum_{i\in \mathrm{Gr\,I}}\log[P(Y_i=y_i\mid\mathbf{x}_i)*P(R_i^1=1\mid\mathbf{z}_i)*P(R_i^2=1\mid R_i^1=1,Y_i=y_i,\mathbf{z}_i)]$

$$=\sum_{i\in \mathrm{Gr\,I}}\{\log[P(Y_i=y_i\mid\mathbf{x}_i)]+\log[P(R_i^1=1\mid\mathbf{z}_i)]+\log[P(R_i^2=1\mid R_i^1=1,Y_i=y_i,\mathbf{z}_i)]\}$$

$$=\sum_{\{i\in \mathrm{Gr\,I}:y_i=1\}}\log[P(Y_i=1\mid\mathbf{x}_i)]+\sum_{\{i\in \mathrm{Gr\,I}:y_i=0\}}\log[P(Y_i=0\mid\mathbf{x}_i)]+\sum_{i\in \mathrm{Gr\,I}}\{\log[P(R_i^1=1\mid\mathbf{z}_i)]+\log[P(R_i^2=1\mid Y_i=y_i,\mathbf{z}_i)]\}$$

$$=\sum_{x_1}\sum_{x_2}\sum_{x_3}\{yja(x_1,x_2,x_3)*[\beta(x_1,x_2,x_3)-\log(1+\exp(\beta(x_1,x_2,x_3)))]$$
$$+\,ynei(x_1,x_2,x_3)*[0-\log(1+\exp(\beta(x_1,x_2,x_3)))]\}$$

$$-\{\sum_{z_1}\sum_{x_2}\sum_{x_3}[yja2(z_1,x_2,x_3)+ynei2(z_1,x_2,x_3)]*\log(1+\exp(-\varphi(z_1,x_2,x_3)))\}$$

$$-\{\sum_{z_1}\sum_{x_2}\sum_{x_3}yja2(z_1,x_2,x_3)*\log[1+\exp(-\psi(z_1,x_2,x_3,1,1))]$$
$$+\,ynei2(z_1,x_2,x_3)*\log[1+\exp(-\psi(z_1,x_2,x_3,0,1))]\}$$

$$=\sum_{x_1}\sum_{x_2}\sum_{x_3}\{yja(x_1,x_2,x_3)*(\beta(x_1,x_2,x_3)-[yja(x_1,x_2,x_3)+ynei(x_1,x_2,x_3)]*\log(1+\exp(\beta(x_1,x_2,x_3)))\}$$

$$-\{\sum_{z_1}\sum_{x_2}\sum_{x_3}[yja2(z_1,x_2,x_3)+ynei2(z_1,x_2,x_3)]*\log(1+\exp(-\varphi(z_1,x_2,x_3)))\}$$

$$-\{\sum_{z_1}\sum_{x_2}\sum_{x_3}yja2(z_1,x_2,x_3)*\log[1+\exp(-\psi(z_1,x_2,x_3,1,1))]$$
$$+\,ynei2(z_1,x_2,x_3)*\log[1+\exp(-\psi(z_1,x_2,x_3,0,1))]\}$$

(II) = $\sum_{i\in \mathrm{Gr\,II}} \log[P(R_i^1 = 1 \mid \mathbf{z}_i) * \sum_{y=0}^{1} P(R_i^2 = 0 \mid Y_i = y, R_i^1 = 1, \mathbf{z}_i) * P(Y_i = y \mid \mathbf{x}_i)]$

$= \sum_{z_1}\sum_{x_1}\sum_{x_2}\sum_{x_3} pf1(z_1, x_1, x_2, x_3)$

$* \log\{[\frac{1}{1+\exp(-\varphi(z_1, x_2, x_3))}] * \sum_{y=0}^{1}[\frac{1}{1+\exp(\psi(z_1, x_2, x_3, y, 1))} * \frac{\exp(y * \beta(x_1, x_2, x_3))}{1+\exp(\beta(x_1, x_2, x_3))}]\}$

(III) = $\sum_{i\in \mathrm{Gr\,III}} \log[P(Y_i = y_i \mid \mathbf{x}_i) * P(R_i^1 = 0 \mid \mathbf{z}_i) * P(R_i^2 = 1 \mid R_i^1 = 0, Y_i = y_i, \mathbf{z}_i)]$

$= \sum_{x_1}\sum_{x_2}\sum_{x_3}\{jato(x_1, x_2, x_3) * \beta(x_1, x_2, x_3) - [jato(x_1, x_2, x_3) + neito(x_1, x_2, x_3)] * \log(1 + \exp(\beta(x_1, x_2, x_3)))\}$

$-\{\sum_{z_1}\sum_{x_2}\sum_{x_3}[jato2(z_1, x_2, x_3) + neito2(z_1, x_2, x_3)] * \log(1 + \exp(\varphi(z_1, x_2, x_3)))\}$

$-\{\sum_{z_1}\sum_{x_2}\sum_{x_3} jato2(z_1, x_2, x_3) * \log[1 + \exp(-\psi(z_1, x_2, x_3, 1, 0))]$

$+ neito2(z_1, x_2, x_3) * \log[1 + \exp(-\psi(z_1, x_2, x_3, 0, 0))]\}$

(IV) = $\sum_{i\in \mathrm{Gr\,IV}} \log[P(R_i^1 = 0 \mid \mathbf{z}_i) * \sum_{y=0}^{1} P(R_i^2 = 0 \mid Y_i = y, R_i^1 = 0, \mathbf{z}_i) * P(Y_i = y \mid \mathbf{x}_i)]$

$= \sum_{z_1}\sum_{x_1}\sum_{x_2}\sum_{x_3} pfo(z_1, x_1, x_2, x_3)$

$* \log\{[\frac{1}{1+\exp(\varphi(z_1, x_2, x_3))}] * \sum_{y=0}^{1}[\frac{1}{1+\exp(\psi(z_1, x_2, x_3, y, 0))}] * [\frac{\exp(y * \beta(x_1, x_2, x_3))}{1+\exp(\beta(x_1, x_2, x_3))}]\}$

(V) = $\sum_{i\in \mathrm{Gr\,V}} \log[P(R_i^1 = r_i^1 \mid \mathbf{z}_i)]$

$= \sum_{z_1}\sum_{x_2}\sum_{x_3}\{grfem1(z_1, x_2, x_3) * (\varphi(z_1, x_2, x_3))$

$-[grfem1(z_1, x_2, x_3) + grfemo(z_1, x_2, x_3)] * \log(1 + \exp(\varphi(z_1, x_2, x_3)))\}$

A.2. Likelihoodfunksjonen for undersysselsetting

Likelihoodfunksjonen for enkeltspørsmål ble satt opp som et produkt av fem faktorer, svarende til fem grupper av personer. Den siste gruppen omfatter personer utenfor målgruppen for aktuelle spørsmål. De fire første gruppene representerer de ulike kombinasjoner av R^1 og R_j^2 – verdier.

Likelihoodfunksjonen for undersysselsetting lages på liknende måte.

For å bli regnet som undersysselsatt, må intervjuobjektet ha svart ja på tre spørsmål, nemlig 18, 19 og 23a. (Det vil si på spørsmål 23a må han ha svart at han kan begynne med økt arbeidstid innen en måned.)

Modell 1

Her modelleres Y_u og R_u^2 på nøyaktig samme måte som da Y_j og R_j^2 ble modellert på hvert enkelt spørsmål. Vi kan derfor bruke likelihoodfunksjonen for enkeltspørsmål i dette tilfellet.

Modell 2

I tabellen under har vi delt datasettet inn i 12 grupper etter *observerte* verdier på variablene i modellen. De 10 første gruppene utgjør til sammen de deltidssysselsatte, de to siste utgjør resten.

Tabell A1.1

Gruppe	R^1	R^2_u	R^2_{18}	R^2_{19}	Y_u
1	0	0	0	0	-
2	0	0	1	0	-
3	0	0	1	1	-
4	0	1			0
5	0	1			1
6	1	0	0	0	-
7	1	0	1	0	-
8	1	0	1	1	-
9	1	1			0
10	1	1			1
11	0				
12	1				

I gruppe 4, 5, 9 og 10 har vi ikke oppgitt verdier for R_{18}^2 og R_{19}^2. Dette er fordi disse variablene bare modelleres når $R_u^2 = 0$ i modell 2. Verdiene er derfor uinteressante når $R_u^2 = 1$. En person med $R^1 = 0$, $R_u^2 = 1$ og $Y_u = 0$ havner i gruppe 4, og gir samme bidrag til likelihoodfunksjonen, uansett verdi på R_{18}^2 og R_{19}^2.

Kommentar: I gruppe 5 og 10 vet vi faktisk at både R_{18}^2 og R_{19}^2 av logiske årsaker må være 1. Dette er fordi vi her har *observert* verdien 1 på undersysselsetting. I gruppe 4 og 9 kan vi utlede at R_{18}^2 må være 1, og at R_{19}^2 enten må være 1 (hvis personen svarte ja på spørsmål 18) eller udefinert (hvis personen svarte nei på spørsmål 18). Alt dette har imidlertid ingen betydning for likelihoodfunksjonen.

La $const(i,\mathbf{x},\mathbf{z})$, $i=1,\ldots,12$ være antall personer i gruppe i med tilleggsvariable $\mathbf{x},\mathbf{z}$, og la $P_i(\mathbf{x},\mathbf{z})$ være modellsannsynligheten for å havne i gruppe i gitt tilleggsvariablene $\mathbf{x},\mathbf{z}$.

*Log*likelihoodfunksjonen blir da:

$$\log(l)=\sum_{i=1}^{12}\sum_{\mathbf{x},\mathbf{z}} const(i,\mathbf{x},\mathbf{z})*\log[P_i(\mathbf{x},\mathbf{z})]$$

Vi skriver ut $P_i(\mathbf{x},\mathbf{z})$, $i=1,\ldots,12$ under modell 2:

$$\begin{aligned}
&P_1(\mathbf{x},\mathbf{z})=P\left(R^1=0,R_u^2=0,R_{18}^2=0,R_{19}^2=0\mid\mathbf{x},\mathbf{z}\right)\\
&=P\left(R^1=0,R_u^2=0,R_{18}^2=0\mid\mathbf{x},\mathbf{z}\right)\\
&=P\left(R^1=0,R_u^2=0,R_{18}^2=0,Y_u=0\mid\mathbf{x},\mathbf{z}\right)+P\left(R^1=0,R_u^2=0,R_{18}^2=0,Y_u=1\mid\mathbf{x},\mathbf{z}\right)\\
&=P\left(R_{18}^2=0\mid R^1=0,R_u^2=0,Y_u=0,\mathbf{z}\right)P\left(R_u^2=0\mid R^1=0,Y_u=0,\mathbf{z}\right)P\left(Y_u=0\mid\mathbf{x}\right)P\left(R^1=0\mid\mathbf{z}\right)\\
&+P\left(R_{18}^2=0\mid R^1=0,R_u^2=0,Y_u=1,\mathbf{z}\right)P\left(R_u^2=0\mid R^1=0,Y_u=1,\mathbf{z}\right)P\left(Y_u=1\mid\mathbf{x}\right)P\left(R^1=0\mid\mathbf{z}\right)
\end{aligned}$$

$$\begin{aligned}
&P_2(\mathbf{x},\mathbf{z})=P\left(R^1=0,R_u^2=0,R_{18}^2=1,R_{19}^2=0\mid\mathbf{x},\mathbf{z}\right)\\
&=P\left(R^1=0,R_u^2=0,R_{18}^2=1,R_{19}^2=0,Y_u=0\mid\mathbf{x},\mathbf{z}\right)+P\left(R^1=0,R_u^2=0,R_{18}^2=1,R_{19}^2=0,Y_u=1\mid\mathbf{x},\mathbf{z}\right)\\
&=P\left(R_{19}^2=0\mid R^1=0,R_u^2=0,R_{18}^2=1,Y_u=0,\mathbf{z}\right)P\left(R_{18}^2=1\mid R^1=0,R_u^2=0,Y_u=0,\mathbf{z}\right)\\
&*P\left(R_u^2=0\mid R^1=0,Y_u=0,\mathbf{z}\right)P\left(Y_u=0\mid\mathbf{x}\right)P\left(R^1=0\mid\mathbf{z}\right)\\
&+P\left(R_{19}^2=0\mid R^1=0,R_u^2=0,R_{18}^2=1,Y_u=1,\mathbf{z}\right)P\left(R_{18}^2=1\mid R^1=0,R_u^2=0,Y_u=1,\mathbf{z}\right)\\
&*P\left(R_u^2=0\mid R^1=0,Y_u=1,\mathbf{z}\right)P\left(Y_u=1\mid\mathbf{x}\right)P\left(R^1=0\mid\mathbf{z}\right)
\end{aligned}$$

$$\begin{aligned}
&P_3(\mathbf{x},\mathbf{z})=P\left(R^1=0,R_u^2=0,R_{18}^2=1,R_{19}^2=1\mid\mathbf{x},\mathbf{z}\right)\\
&=P\left(R^1=0,R_u^2=0,R_{18}^2=1,R_{19}^2=1,Y_u=0\mid\mathbf{x},\mathbf{z}\right)+P\left(R^1=0,R_u^2=0,R_{18}^2=1,R_{19}^2=1,Y_u=1\mid\mathbf{x},\mathbf{z}\right)\\
&=P\left(R_{19}^2=1\mid R^1=0,R_u^2=0,R_{18}^2=1,Y_u=0,\mathbf{z}\right)P\left(R_{18}^2=1\mid R^1=0,R_u^2=0,Y_u=0,\mathbf{z}\right)\\
&*P\left(R_u^2=0\mid R^1=0,Y_u=0,\mathbf{z}\right)P\left(Y_u=0\mid\mathbf{x}\right)P\left(R^1=0\mid\mathbf{z}\right)\\
&+P\left(R_{19}^2=1\mid R^1=0,R_u^2=0,R_{18}^2=1,Y_u=1,\mathbf{z}\right)P\left(R_{18}^2=1\mid R^1=0,R_u^2=0,Y_u=1,\mathbf{z}\right)\\
&*P\left(R_u^2=0\mid R^1=0,Y_u=1,\mathbf{z}\right)P\left(Y_u=1\mid\mathbf{x}\right)P\left(R^1=0\mid\mathbf{z}\right)
\end{aligned}$$

$$\begin{aligned}
&P_4(\mathbf{x},\mathbf{z})=P\left(R^1=0,R_u^2=1,Y_u=0\mid\mathbf{x},\mathbf{z}\right)\\
&=P\left(R_u^2=1\mid R^1=0,Y_u=0,\mathbf{z}\right)P\left(R^1=0\mid\mathbf{z}\right)P\left(Y_u=0\mid\mathbf{x}\right)
\end{aligned}$$

$$\begin{aligned}
&P_5(\mathbf{x},\mathbf{z})=P\left(R^1=0,R_u^2=1,Y_u=1\mid\mathbf{x},\mathbf{z}\right)\\
&=P\left(R_u^2=1\mid R^1=0,Y_u=1,\mathbf{z}\right)P\left(R^1=0\mid\mathbf{z}\right)P\left(Y_u=1\mid\mathbf{x}\right)
\end{aligned}$$

$P_6(\mathbf{x},\mathbf{z})$ til $P_{10}(\mathbf{x},\mathbf{z})$ blir helt analoge. Den eneste forskjellen er at $R^1=0$ byttes ut med $R^1=1$ overalt. $P_{11}(\mathbf{x},\mathbf{z})$ og $P_{12}(\mathbf{x},\mathbf{z})$ er også helt rett frem.

For å få uttrykt likelihoodfunksjonen som en funksjon av parametrene i modellen, er det praktisk å innføre kortere notasjon for størrelser som ofte går igjen.

La nå

$$\underline{\beta_u(x_1,x_2,x_3)} = \beta_{u,0} + \sum_{l=1,2}\beta_{u,1,l}D_{1,l} + \beta_{u,2}x_2\sum_{l=1,3,4,5}\beta_{u,3,l}D_{3,l} + \sum_{l=1,2}\beta_{u,4,l}x_2D_{1,l} + \sum_{l=1,3,4,5}\beta_{u,5,l}x_2D_{3,l}$$
$$+ \sum_{l=1,2}\sum_{m=1,3,4,5}\beta_{u,6,l,m}D_{1,l}D_{3,m}$$

$$\underline{\psi_u(z_1,z_2,z_3,y_u,R^1)} = \psi_{u,0} + \psi_{u,1}z_1 + \psi_{u,2}z_2 + \sum_{l=1,3,4,5}\psi_{u,3,l}D_{3,l} + \psi_{u,4}y_u + \psi_{u,5}R^1$$

$$\underline{\psi_{18}(z_1,z_2,z_3,y_u,R^1)} = \psi_{18,0} + \psi_{18,1}z_1 + \psi_{18,2}z_2 + \sum_{l=1,3,4,5}\psi_{18,3,l}D_{3,l} + \psi_{18,4}y_u + \psi_{18,5}R^1$$

$$\underline{\psi_{19}(z_1,z_2,z_3,y_u,R^1)} = \psi_{19,0} + \psi_{19,1}z_1 + \psi_{19,2}z_2 + \sum_{l=1,3,4,5}\psi_{19,3,l}D_{3,l} + \psi_{19,4}y_u + \psi_{19,5}R^1$$

$$\underline{\varphi(z_1,z_2,z_3)} = \varphi_0 + \varphi_1z_1 + \varphi_2x_2 + \sum_{l=1,3,4,5}\varphi_{3,l}D_{3,l} + \varphi_4z_1z_2 + \sum_{l=1,3,4,5}\varphi_{5,l}z_1D_{3,l} + \sum_{l=1,3,4,5}\varphi_{6,l}z_2D_{3,l}$$

La videre

$$t1 = P(R^1 = 0 \mid \mathbf{z}) = 1/(1+\exp(\varphi(\mathbf{z})))$$
$$t2 = P(Y_u = 1 \mid \mathbf{x}) = 1/(1+\exp(-\beta_u(\mathbf{x})))$$
$$t3 = P(R_u^2 = 1 \mid R^1 = 0, Y_u = 0, \mathbf{z}) = 1/(1+\exp(-\psi_u(\mathbf{z},0,0)))$$
$$t4 = P(R_u^2 = 1 \mid R^1 = 0, Y_u = 1, \mathbf{z}) = 1/(1+\exp(-\psi_u(\mathbf{z},1,0)))$$
$$t5 = P(R_{18}^2 = 1 \mid R^1 = 0, R_u^2 = 0, Y_u = 0, \mathbf{z}) = 1/(1+\exp(-\psi_{18}(\mathbf{z},0,0)))$$
$$t6 = P(R_{18}^2 = 1 \mid R^1 = 0, R_u^2 = 0, Y_u = 1, \mathbf{z}) = 1/(1+\exp(-\psi_{18}(\mathbf{z},1,0)))$$
$$t7 = P(R_{19}^2 = 1 \mid R^1 = 0, R_u^2 = 0, R_{18}^2 = 1, Y_u = 0, \mathbf{z}) = 1/(1+\exp(-\psi_{19}(\mathbf{z},0,0)))$$
$$t8 = P(R_{19}^2 = 1 \mid R^1 = 0, R_u^2 = 0, R_{18}^2 = 1, Y_u = 1, \mathbf{z}) = 1/(1+\exp(-\psi_{19}(\mathbf{z},1,0)))$$

$t101 - t108$ svarer til $t1 - t8$, bortsett fra at $R^1 = 0$ er byttet ut med $R^1 = 1$:

$$t101 = P(R^1 = 1 \mid \mathbf{z}) = 1/(1+\exp(-\varphi(\mathbf{z}))) = (1 - t1)$$
$$t102 = P(Y_u = 1 \mid \mathbf{x}) = 1/(1+\exp(-\beta_u(\mathbf{x}))) = t2$$
$$t103 = P(R_u^2 = 1 \mid R^1 = 1, Y_u = 0, \mathbf{z}) = 1/(1+\exp(-\psi_u(\mathbf{z},0,1)))$$
$$t104 = P(R_u^2 = 1 \mid R^1 = 1, Y_u = 1, \mathbf{z}) = 1/(1+\exp(-\psi_u(\mathbf{z},1,1)))$$
$$t105 = P(R_{18}^2 = 1 \mid R^1 = 1, R_u^2 = 0, Y_u = 0, \mathbf{z}) = 1/(1+\exp(-\psi_{18}(\mathbf{z},0,1)))$$
$$t106 = P(R_{18}^2 = 1 \mid R^1 = 1, R_u^2 = 0, Y_u = 1, \mathbf{z}) = 1/(1+\exp(-\psi_{18}(\mathbf{z},1,1)))$$
$$t107 = P(R_{19}^2 = 1 \mid R^1 = 1, R_u^2 = 0, R_{18}^2 = 1, Y_u = 0, \mathbf{z}) = 1/(1+\exp(-\psi_{19}(\mathbf{z},0,1)))$$
$$t108 = P(R_{19}^2 = 1 \mid R^1 = 1, R_u^2 = 0, R_{18}^2 = 1, Y_u = 1, \mathbf{z}) = 1/(1+\exp(-\psi_{19}(\mathbf{z},1,1)))$$

Vi kan nå uttrykke $P_1(\mathbf{x},\mathbf{z}),\ldots,P_{12}(\mathbf{x},\mathbf{z})$ ved hjelp av $t1,\ldots,t8,t103,\ldots,t108$.

$$P_1(\mathbf{x},\mathbf{z}) = (1-t5)(1-t3)(1-t2)t1 + (1-t6)(1-t4)t2t1$$
$$P_2(\mathbf{x},\mathbf{z}) = (1-t7)t5(1-t3)(1-t2)t1 + (1-t8)t6(1-t4)t2t1$$
$$P_3(\mathbf{x},\mathbf{z}) = t7t5(1-t3)(1-t2)t1 + t8t6(1-t4)t2t1$$
$$P_4(\mathbf{x},\mathbf{z}) = t3(1-t2)t1$$
$$P_5(\mathbf{x},\mathbf{z}) = t4t2t1$$

$$P_6(\mathbf{x},\mathbf{z}) = (1-t105)(1-t103)(1-t2)(1-t1) + (1-t106)(1-t104)t2(1-t1)$$
$$P_7(\mathbf{x},\mathbf{z}) = (1-t107)t105(1-t103)(1-t2)(1-t1) + (1-t108)t106(1-t104)t2(1-t1)$$
$$P_8(\mathbf{x},\mathbf{z}) = t107t105(1-t103)(1-t2)(1-t1) + t108t106(1-t104)t2(1-t1)$$
$$P_9(\mathbf{x},\mathbf{z}) = t103(1-t2)(1-t1)$$
$$P_{10}(\mathbf{x},\mathbf{z}) = t104t2(1-t1)$$

$$P_{11}(\mathbf{x},\mathbf{z}) = t1$$
$$P_{12}(\mathbf{x},\mathbf{z}) = 1-t1$$

Disse uttrykkene vil bli brukt under programmeringen av likelihoodfunksjonen.

Modell 3
Vi går frem på samme måte som under modell 2, og deler datasettet inn i 16 grupper, etter observerte verdier på variablene i modell 3. Gruppe 1-14 er de deltidssysselsatte, gruppe 15 og 16 er personene utenfor målgruppen.

Tabell A1.2

Gruppe	R^1	R^2_{18}	R^2_{19}	R^2_{23}	Y_{18}	Y_{19}	Y_{23}	Y_u	R^2_u
1	0	0	0	0	-	-	-	-	0
2	0	1			0			0	1
3	0	1	0	0	1	-	-	-	0
4	0	1	1	0	1	1	-	-	0
5	0	1	1		1	0		0	1
6	0	1	1	1	1	1	1	1	1
7	0	1	1	1	1	1	0	0	1
8	1	0	0	0	-	-	-	-	0
9	1	1			0			0	1
10	1	1	0	0	1	-	-	-	0
11	1	1	1	0	1	1	-	-	0
12	1	1	1		1	0		0	1
13	1	1	1	1	1	1	1	1	1
14	1	1	1	1	1	1	0	0	1
15	0								
16	1								

Blanke ruter betyr her at variabelen ikke modelleres i gruppen. En "-" betyr partielt frafall.

Kommentar: I det observerte datasettet kan vi ikke alltid se forskjell på ekte partielt frafall og det at en person er utenfor målgruppen for et spørsmål. I gruppe 1 for eksempel vil noen personer ha svaret nei på spørsmål 18, resten vil ha svaret ja, men ingen av dem har oppgitt svaret. De førstnevnte personene er utenfor målgruppen for spørsmål 19, og har derfor egentlig ikke ekte partielt frafall på

spørsmål 19. Vi har likevel satt "-" i ruten for Y_{19} i gruppe 1, selv om det er litt upresist. Dette er fordi det ikke spiller noen for likelihoodfunksjonen. Det samme fenomenet dukker også opp i gruppe 3, og dermed også i gruppene 8 og 10.

Som for modell 2 lar vi $const(i,\mathbf{x},\mathbf{z})$, $i=1,\ldots,16$ være antall personer i gruppe i med tilleggsvariable $\mathbf{x},\mathbf{z}$, og $P_i(\mathbf{x},\mathbf{z})$ være modellsannsynligheten for å havne i gruppe i gitt tilleggsvariablene $\mathbf{x},\mathbf{z}$.

*Log*likelihoodfunksjonen blir da:

$$\log(l)=\sum_{i=1}^{16}\sum_{\mathbf{x},\mathbf{z}} const(i,\mathbf{x},\mathbf{z})*\log\left[P_i(\mathbf{x},\mathbf{z})\right]$$

der $const(i,\mathbf{x},\mathbf{z})$ leses inn fra en datafil produsert av SAS.

Hver av de 16 sannsynlighetene kan, som under modell 2, uttrykkes som et produkt av betingede sannsynligheter. For eksempel faktoriserer vi P_6 på følgende måte:

$$\begin{aligned}
&P_6(\mathbf{x},\mathbf{z})=P\left(R^1=0,R_{18}^2=1,Y_{18}=1,R_{19}^2=1,Y_{19}=1,R_{23}^2=1,Y_{23}=1\mid\mathbf{x},\mathbf{z}\right)\\
&=P\left(R_{23}^2=1\mid R^1=0,R_{18}^2=1,Y_{18}=1,R_{19}^2=1,Y_{19}=1,Y_{23}=1,\mathbf{z}\right)\\
&*P\left(Y_{23}=1\mid R_{18}^2=1,Y_{18}=1,R_{19}^2=1,Y_{19}=1,\mathbf{x}\right)\\
&*P\left(R_{19}^2=1\mid R^1=0,R_{18}^2=1,Y_{18}=1,Y_{19}=1,\mathbf{z}\right)\\
&*P\left(Y_{19}=1\mid R_{18}^2=1,Y_{18}=1,\mathbf{x}\right)\\
&*P\left(R_{18}^2=1\mid R^1=0,Y_{18}=1,\mathbf{z}\right)\\
&*P\left(Y_{18}=1\mid\mathbf{x}\right)\\
&*P\left(R^1=0\mid\mathbf{z}\right)
\end{aligned}$$

I grupper med ekte partielt frafall (gruppe 1, 3 og 4 og 8, 10 og 11) må sannsynligheten splittes opp i en sum av to ledd, en for hver y-verdi (Y_{18} for gruppe 1 og 8, Y_{19} for gruppe 3 og 10 og Y_{23} for gruppe 4 og 11). For gruppe 1 får vi f.eks.:

$$\begin{aligned}
P_1(\mathbf{x},\mathbf{z})&=P(R^1=0,R_{18}^2=0\mid\mathbf{x},\mathbf{z})\\
&=P\left(R^1=0,R_{18}^2=0,Y_{18}=0\mid\mathbf{x},\mathbf{z}\right)+P\left(R^1=0,R_{18}^2=0,Y_{18}=1\mid\mathbf{x},\mathbf{z}\right)
\end{aligned}$$

Hvert av de to leddene faktoriseres på tilsvarende måte som P_6.

På samme måte som under modell 2, definerer vi følgende funksjoner for å forenkle notasjonen når vi skal skrive ut de betingede sannsynlighetene:

$$\underline{\beta_{sp}(x_1,x_2,x_3)}=\beta_0+\sum_{l=1,2}\beta_{sp,1,l}D_{1,l}+\beta_{sp,2}x_2\sum_{l=1,3,4,5}\beta_{sp,3,l}D_{3,l}+\sum_{l=1,2}\beta_{sp,4,l}x_2D_{1,l}+\sum_{l=1,3,4,5}\beta_{sp,5,l}x_2D_{3,l}$$
$$+\sum_{l=1,2}\sum_{m=1,3,4,5}\beta_{sp,6,l,m}D_{1,l}D_{3,m}$$

Funksjonen ovenfor er lik i form for alle spørsmålene, men $\beta(\ )$ vil variere fra spørsmål til spørsmål, og vi bruker derfor spørsmålsnummeret som indeks for β : $\beta_{18}(\ ),\beta_{19}(\ ),\beta_{23}(\ )$.

$$\underline{\psi_{18}(z_1,z_2,z_3,y_{18},R^1)} = \psi_{18,0} + \psi_{18,1}z_1 + \psi_{18,2}z_2 + \sum_{l=1,3,4,5}\psi_{18,3,l}D_{3,l} + \psi_{18,4}y_{18} + \psi_{18,5}R^1$$

$$\underline{\psi_{19}(z_1,z_2,z_3,y_{19},R^1)} = \psi_{19,0} + \psi_{19,1}z_1 + \psi_{19,2}z_2 + \sum_{l=1,3,4,5}\psi_{19,3,l}D_{3,l} + \psi_{19,4}y_{19} + \psi_{19,5}R^1$$

$$\underline{\psi_{23}(z_1,z_2,z_3,y_{23},R^1)} = \psi_{23,0} + \psi_{23,1}z_1 + \psi_{23,2}z_2 + \sum_{l=1,3,4,5}\psi_{23,3,l}D_{3,l} + \psi_{23,4}y_{23} + \psi_{23,5}R^1$$

$$\underline{\varphi(z_1,z_2,z_3)} = \varphi_0 + \varphi_1 z_1 + \varphi_2 x_2 + \sum_{l=1,3,4,5}\varphi_{3,l}D_{3,l} + \varphi_4 z_1 z_2 + \sum_{l=1,3,4,5}\varphi_{5,l}z_1 D_{3,l} + \sum_{l=1,3,4,5}\varphi_{6,l}z_2 D_{3,l}$$

Som under modell 2, setter vi navn på hver av de betingede sannsynlighetene, og uttrykker dem ved hjelp av funksjonene over. Under modell 2 brukte vi t -er, her bruker vi s -er.

For å kunne uttrykke $P_1 - P_7$ og P_{15}, trenger vi 10 slike s-er:

$$s1 = P(R^1 = 0 \mid \mathbf{z}) = 1/(1+\exp(\varphi(\mathbf{z})))$$
$$s2 = P(Y_{18} = 1 \mid R^1 = 0, \mathbf{x}) = P(Y_{18} = 1 \mid \mathbf{x}) = 1/(1+\exp(-\beta_{18}(\mathbf{x})))$$
$$s3 = P(R_{18}^2 = 1 \mid R^1 = 0, Y_{18} = 0, \mathbf{z}) = 1/(1+\exp(-\psi_{18}(\mathbf{z},0,0)))$$
$$s4 = P(R_{18}^2 = 1 \mid R^1 = 0, Y_{18} = 1, \mathbf{z}) = 1/(1+\exp(-\psi_{18}(\mathbf{z},1,0)))$$
$$s5 = P(Y_{19} = 1 \mid R^1 = 0, R_{18}^2 = 1, Y_{18} = 1, \mathbf{x}) = 1/(1+\exp(-\beta_{19}(\mathbf{x})))$$
$$s6 = P(R_{19}^2 = 1 \mid R^1 = 0, R_{18}^2 = 1, Y_{18} = 1, Y_{19} = 1, \mathbf{z}) = 1/(1+\exp(-\psi_{19}(\mathbf{z},1,0)))$$
$$s7 = P(R_{19}^2 = 1 \mid R^1 = 0, R_{18}^2 = 1, Y_{18} = 1, Y_{19} = 0, \mathbf{z}) = 1/(1+\exp(-\psi_{19}(\mathbf{z},0,0)))$$
$$s8 = P(R_{23}^2 = 1 \mid R^1 = 0, R_{18}^2 = 1, Y_{18} = 1, R_{19}^2 = 1, Y_{19} = 1, Y_{23} = 0, \mathbf{z}) = 1/(1+\exp(-\psi_{23}(\mathbf{z},0,0)))$$
$$s9 = P(R_{23}^2 = 1 \mid R^1 = 0, R_{18}^2 = 1, Y_{18} = 1, R_{19}^2 = 1, Y_{19} = 1, Y_{23} = 1, \mathbf{z}) = 1/(1+\exp(-\psi_{23}(\mathbf{z},1,0)))$$
$$s10 = P(Y_{23} = 1 \mid R^1 = 0, R_{18}^2 = 1, Y_{18} = 1, R_{19}^2 = 1, Y_{19} = 1, \mathbf{x}) = 1/(1+\exp(-\beta_{23}(\mathbf{x})))$$

For å uttrykke $P_8 - P_{14}$ og P_{16} trenger vi nye 10 s-er, som tilsvarer de ovenstående, bortsett fra at $R^1 = 0$ er erstattet med $R^1 = 1$. Vi kaller disse $s101 - s110$, slik at $s1$ svarer til $s101$, osv. Da har vi, f.eks.:

$$s101 = P(R^1 = 1 \mid \mathbf{z}) = 1/(1+\exp(-\varphi(\mathbf{z})))$$
$$\vdots$$
$$s109 = P(R_{23}^2 = 1 \mid R^1 = 1, R_{18}^2 = 1, Y_{18} = 1, R_{19}^2 = 1, Y_{19} = 1, Y_{23} = 1, \mathbf{z}) = 1/(1+\exp(-\psi_{23}(\mathbf{z},1,1)))$$
$$s110 = P(Y_{23} = 1 \mid R^1 = 1, R_{18}^2 = 1, Y_{18} = 1, R_{19}^2 = 1, Y_{19} = 1, \mathbf{x}) = 1/(1+\exp(-\beta_{23}(\mathbf{x})))$$

Hvis vi ser tilbake på eksempelet der P_6 ble faktorisert, ser vi at vi nå kan uttrykke P_6 ved hjelp av parametrene i modellen: $P_6 = s9 * s10 * s6 * s5 * s4 * s2 * s1$.

Ved å bruke samme faktoriseringsmåte som for P_6, får vi:

$$P_1 = s1 * s2 * (1 - s4) + s1 * (1 - s2) * (1 - s3)$$
$$P_2 = s3 * (1 - s2) * s1$$
$$P_3 = s1 * s2 * s4 * s5 * (1 - s6) + s1 * s2 * s4 * (1 - s5) * (1 - s7)$$
$$P_4 = s1 * s2 * s4 * s5 * s6 * (1 - s10) * (1 - s8) + s1 * s2 * s4 * s5 * s6 * s10 * (1 - s9)$$
$$P_5 = s1 * s2 * s4 * (1 - s5) * s7$$
$$P_6 = s1 * s2 * s4 * s5 * s6 * s10 * s9$$
$$P_7 = s1 * s2 * s4 * s5 * s6 * (1 - s10) * s8$$

$$P_{15} = s1$$

$P_8 - P_{14}$ og P_{16} blir da som følger:

$$P_8 = s101 * s102 * (1 - s104) + s101 * (1 - s102) * (1 - s103)$$
$$P_9 = s101 * (1 - s102) * s103$$
$$P_{10} = s101 * s102 * s104 * s105 * (1 - s106) + s101 * s102 * s104 * (1 - s105) * (1 - s107)$$
$$P_{11} = s101 * s102 * s104 * s105 * s106 * (1 - s110) * (1 - s108) + s101 * s102 * s104 * s105 * s106 * s110 * (1 - s109)$$
$$P_{12} = s101 * s102 * s104 * (1 - s105) * s107$$
$$P_{13} = s101 * s102 * s104 * s105 * s106 * s110 * s109$$
$$P_{14} = s101 * s102 * s104 * s105 * s106 * (1 - s110) * s108$$

$$P_{16} = s101$$

På denne måten er alle P-ene uttrykt ved hjelp av parametrene i modellen, som er det vi trenger for å optimere likelihoodfunksjonen.

Appendiks B. Simulering av MLE i tabeller over $P(Y=1|\mathrm{x})$, $P(R^1=1|\mathrm{z})$ og $P(R^2=1|\mathrm{z},r^1,y)$.

B.1. Populasjonsmodellen

I hvert stratum har vi beregnet MLE for $P(Y=1|\mathbf{x})$ på grunnlag av det reelle datasettet. I tabellene nedenfor står dette lengst til venstre i hver celle. Vi har også beregnet MLE for samme sannsynlighet for hvert av de 1000 simulerte datasettene. Tallet i midten er gjennomsnittsverdien til disse estimatene. Tallet lengst til høyre er tilhørende standardavvik. a står for aldersgruppe. Vi minner om kodene for alder og boregion:

Alder deles inn i følgende grupper:

1: 16-19 år, 2: 20-39 år, 3: 40-66 år, 4: 67-74 år

På spørsmål 25 slår vi sammen gruppe 2 og 3 til en ny gruppe 2, og gruppe 4 blir dermed gruppe 3. På spørsmål 19, 23a ,24 og 62, og på spørsmålet om undersysselsetting, utelater vi personer i gruppe 4 fra analysen.

Bostedsregion er delt inn i følgende grupper:

1: Oslo, Akershus, 2: Resten av Østlandet, 3: Sørlandet, Vestlandet (unntatt Møre og Romsdal), 4: Møre og Romsdal, Trøndelag, 5: Nord-Norge

Spørsmål 18

Menn

a	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
1	0,365 0,363 0,07	0,375 0,375 0,07	0,357 0,355 0,06	0,370 0,365 0,10	0,433 0,424 0,11
2	0,488 0,487 0,06	0,408 0,408 0,06	0,411 0,409 0,06	0,548 0,546 0,08	0,583 0,580 0,08
3	0,309 0,309 0,06	0,254 0,255 0,05	0,287 0,286 0,05	0,353 0,354 0,07	0,298 0,298 0,08
4	0,070 0,071 0,07	0,013 0,014 0,02	0,010 0,011 0,02	0,024 0,026 0,04	0,000 0,000 0,01

Kvinner

a	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
1	0,197 0,198 0,05	0,371 0,369 0,06	0,288 0,288 0,06	0,306 0,301 0,09	0,326 0,322 0,10
2	0,242 0,240 0,03	0,345 0,344 0,03	0,284 0,283 0,02	0,415 0,413 0,04	0,409 0,407 0,05
3	0,154 0,153 0,02	0,241 0,241 0,02	0,218 0,218 0,02	0,281 0,281 0,03	0,203 0,202 0,04
4	0,156 0,165 0,10	0,070 0,068 0,06	0,042 0,043 0,04	0,095 0,097 0,08	0,000 0,002 0,02

MLE for $P(Y=1|\mathbf{x})$ er tilnærmet forventningsrette. De fleste av standardavvikene er små. For menn er de største standardavvikene i aldersgruppe 1 (16-19 år) og boregionene 4 og 5 (fra Møre og Romsdal og nordover). For kvinner er standardavvikene størst i aldersgruppe 4 (67-74 år) og boregion 1 (Oslo/Akershus), og i aldersgruppe 1 og boregion 5 (Nord-Norge). Alle de nevnte gruppene har få observasjoner.

Spørsmål 19

Menn

a	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
1	0,696 0,686 0,11	0,183 0,183 0,07	0,361 0,360 0,10	0,146 0,145 0,10	0,313 0,306 0,16
2	0,829 0,829 0,06	0,759 0,763 0,06	0,757 0,759 0,07	0,629 0,627 0,09	0,556 0,544 0,11
3	0,900 0,895 0,05	0,652 0,655 0,08	0,701 0,704 0,08	0,518 0,513 0,11	0,746 0,729 0,11

Kvinner

a	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
1	0,705 0,698 0,12	0,373 0,375 0,10	0,528 0,537 0,11	0,365 0,351 0,16	0,514 0,496 0,18
2	0,583 0,577 0,06	0,697 0,697 0,04	0,629 0,629 0,04	0,610 0,612 0,05	0,444 0,430 0,07
3	0,678 0,674 0,07	0,525 0,524 0,04	0,509 0,511 0,05	0,446 0,443 0,05	0,603 0,594 0,08

Også her er MLE tilnærmet forventningsrette, og standardavvikene er stort sett relativt små. De største standardavvikene finner vi blant unge kvinner i region 4 og 5, og blant unge menn i region 5.

Spørsmål 23a

Menn

a	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
1	0,875 0,860 0,09	0,925 0,918 0,07	0,906 0,894 0,07	1,000 0,998 0,02	0,733 0,712 0,20
2	0,952 0,946 0,04	0,940 0,936 0,04	0,917 0,908 0,05	0,952 0,950 0,05	0,990 0,986 0,03
3	0,971 0,966 0,04	0,959 0,957 0,04	0,953 0,948 0,05	0,975 0,971 0,03	0,990 0,986 0,03

Kvinner

a	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
1	0,937 0,928 0,10	0,989 0,986 0,02	0,987 0,982 0,03	1,000 0,999 0,02	0,700 0,671 0,22
2	0,804 0,790 0,06	0,914 0,909 0,02	0,894 0,884 0,03	0,888 0,884 0,03	0,897 0,886 0,05
3	0,873 0,864 0,06	0,941 0,937 0,02	0,939 0,939 0,03	0,940 0,936 0,03	0,897 0,885 0,07

For både menn og kvinner er variansen større i aldersgruppe 1 (under 20 år) enn i aldersgruppene 2 og 3. Dette skyldes antakelig at det er få observasjoner i aldersgruppe 1.

Undersysselsetting, modell 1

Menn

a	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
1	0,220 0,218 0,07	0,088 0,091 0,04	0,123 0,125 0,05	0,081 0,081 0,06	0,163 0,162 0,10
2	0,376 0,377 0,06	0,306 0,308 0,06	0,286 0,285 0,06	0,345 0,341 0,08	0,329 0,326 0,09
3	0,245 0,247 0,06	0,152 0,154 0,04	0,159 0,159 0,04	0,168 0,165 0,05	0,181 0,182 0,07

Kvinner

a	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
1	0,145 0,146 0,05	0,125 0,126 0,05	0,142 0,146 0,05	0,109 0,108 0,06	0,153 0,156 0,09
2	0,129 0,130 0,02	0,212 0,213 0,02	0,162 0,162 0,02	0,230 0,232 0,03	0,158 0,160 0,04
3	0,089 0,090 0,02	0,119 0,120 0,02	0,101 0,102 0,02	0,123 0,124 0,02	0,094 0,095 0,03

Undersysselsetting, modell 2

Menn

a	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
1	0,220 0,217 0,06	0,088 0,091 0,04	0,124 0,125 0,05	0,081 0,082 0,05	0,165 0,167 0,09
2	0,373 0,377 0,06	0,306 0,305 0,06	0,287 0,286 0,06	0,345 0,342 0,08	0,330 0,331 0,09
3	0,240 0,242 0,06	0,152 0,153 0,04	0,157 0,156 0,04	0,165 0,165 0,05	0,180 0,183 0,07

Kvinner

a	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
1	0,146 0,144 0,05	0,125 0,125 0,04	0,142 0,144 0,05	0,109 0,108 0,06	0,153 0,155 0,08
2	0,129 0,129 0,02	0,211 0,210 0,02	0,162 0,162 0,02	0,231 0,230 0,03	0,156 0,158 0,04
3	0,089 0,089 0,02	0,119 0,121 0,02	0,101 0,101 0,02	0,124 0,124 0,02	0,093 0,094 0,02

Undersysselsetting, modell 3

Menn

a	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
1	0,277 0,274 0,07	0,084 0,085 0,04	0,157 0,155 0,05	0,066 0,069 0,05	0,174 0,172 0,10
2	0,424 0,424 0,06	0,341 0,338 0,05	0,352 0,350 0,05	0,368 0,361 0,07	0,328 0,321 0,08
3	0,297 0,294 0,06	0,194 0,192 0,04	0,245 0,243 0,05	0,205 0,202 0,06	0,245 0,235 0,07

Kvinner

a	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
1	0,145 0,142 0,05	0,156 0,153 0,05	0,162 0,160 0,05	0,127 0,124 0,06	0,204 0,200 0,10
2	0,142 0,141 0,02	0,248 0,246 0,02	0,180 0,178 0,02	0,258 0,258 0,03	0,166 0,165 0,04
3	0,103 0,101 0,02	0,133 0,132 0,02	0,115 0,113 0,02	0,128 0,127 0,02	0,120 0,118 0,03

MLE er omtrent forventningsrette, og standardavvikene er ganske like i alle de tre modellene.

Spørsmål 24

Menn

a	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
1	0,134 0,133 0,07	0,096 0,097 0,06	0,087 0,088 0,05	0,221 0,219 0,14	0,324 0,330 0,18
2	0,838 0,840 0,06	0,914 0,914 0,03	0,874 0,874 0,05	0,890 0,888 0,06	0,964 0,961 0,03
3	0,717 0,713 0,10	0,812 0,812 0,07	0,727 0,726 0,09	0,718 0,714 0,10	0,852 0,846 0,10

Kvinner

a	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
1	0,298 0,299 0,14	0,126 0,128 0,06	0,141 0,149 0,08	0,333 0,323 0,16	0,256 0,252 0,17
2	0,517 0,518 0,06	0,525 0,525 0,04	0,474 0,474 0,04	0,520 0,523 0,05	0,595 0,598 0,07
3	0,529 0,530 0,07	0,488 0,487 0,04	0,425 0,427 0,05	0,421 0,419 0,05	0,402 0,403 0,08

For spørsmål 24 er variansen for menn større i aldersgruppene 1 og 3 enn i aldersgruppe 2. For kvinner er det de yngste som har høyest varians. Som for spørsmål 23a, er det grupper med få observasjoner som har de største variansene.

Spørsmål 25

Menn

a	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
1	0,094 0,093 0,04	0,204 0,204 0,05	0,139 0,140 0,04	0,154 0,151 0,06	0,150 0,149 0,07
2	0,822 0,821 0,01	0,771 0,772 0,01	0,785 0,785 0,01	0,787 0,787 0,01	0,736 0,735 0,02
3	0,463 0,467 0,08	0,425 0,428 0,08	0,339 0,339 0,08	0,220 0,213 0,08	0,413 0,410 0,18

Kvinner

a	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
1	0,067 0,067 0,03	0,127 0,127 0,03	0,074 0,076 0,03	0,077 0,077 0,04	0,135 0,136 0,06
2	0,506 0,506 0,01	0,381 0,381 0,01	0,370 0,370 0,01	0,354 0,355 0,02	0,443 0,445 0,02
3	0,167 0,165 0,06	0,124 0,124 0,05	0,080 0,078 0,03	0,042 0,042 0,03	0,174 0,184 0,12

Både for kvinner og menn er variansen lavest i aldersgruppe 2. I aldersgruppene 1 og 3 har menn litt høyere varians enn kvinner. For både kvinner og menn er variansen størst i aldersgruppe 3 (over 66 år) og region 5 (Nord-Norge).

Spørsmål 62

Menn

a	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
1	0,192 0,192 0,11	0,199 0,201 0,12	0,353 0,352 0,13	0,737 0,734 0,20	0,446 0,445 0,29
2	0,876 0,875 0,05	0,949 0,949 0,02	0,926 0,925 0,03	0,990 0,990 0,01	1,000 1,000 0,00
3	0,824 0,827 0,07	0,963 0,963 0,03	0,913 0,911 0,05	0,993 0,992 0,01	1,000 1,000 0,00

Kvinner

a	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
1	0,422 0,413 0,16	0,192 0,194 0,11	0,367 0,378 0,15	0,328 0,321 0,21	0,000 0,000 0,04
2	0,715 0,718 0,07	0,673 0,674 0,07	0,604 0,604 0,07	0,667 0,669 0,08	0,952 0,952 0,05
3	0,417 0,416 0,13	0,556 0,555 012	0,357 0,359 0,10	0,548 0,546 0,14	0,783 0,782 0,17

For menn er variansen størst i aldersgruppe 1, der standardavviket tar verdier mellom 0,11 og 0,29. Lavest varians har aldersgruppe 2 (20-66 år), der standardavviket varierer mellom 0,00 og 0,05. For kvinner er variansen noe høyere enn for menn i alle aldersgrupper, bortsett fra i boregion 5, aldersgruppe 1. Også for kvinner er variansen størst i aldersgruppe 1 (standardavvik fra 0,04 til 0,21) og minst i aldersgruppe 2, (standardavviket varierer mellom 0,05 og 0,08).

B.2. Modell for R^1

Modellen for R^1 bruker alle observasjonene i utvalget. Estimatene for $P(R^1 = 1 | \mathbf{z}, y) = P(R^1 = 1 | \mathbf{z})$ vil være de samme for alle spørsmålene. I de neste tabellene ser vi hvordan estimert sannsynlighet for å svare ved første henvendelse varierer med sivilstand, kjønn og bosted.

Som i tabellene over, er de tre tallene i hver celle henholdsvis: MLE basert på svarutvalget, gjennomsnitt basert på 1000 simulerte "utvalg", og tilhørende standardavvik. R^1 er ikke avhengig av y, og tabellene viser sivilstand mot boregion.

$s = 1$ betyr aleneboende, $s = 2$ betyr gift eller samboende.

Menn

s	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
1	0,904 0,904 0,01	0,951 0,951 0,01	0,944 0,943 0,01	0,938 0,937 0,01	0,930 0,930 0,01
2	0,950 0,950 0,01	0,961 0,961 0,00	0,964 0,964 0,00	0,956 0,956 0,01	0,962 0,962 0,01

Kvinner

s	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
1	0,914 0,913 0,01	0,957 0,957 0,01	0,948 0,948 0,01	0,940 0,940 0,01	0,923 0,923 0,01
2	0,960 0,960 0,00	0,970 0,970 0,00	0,971 0,971 0,00	0,963 0,963 0,01	0,963 0,963 0,01

MLE er forventningsrette. Standardavviket er ikke høyere enn 0,01 i noe stratum.

De betingede sannsynligheter for å svare ved første henvendelse, er lavere for aleneboende enn for IO som er gift eller samboende. Dette gjelder både for kvinner og menn, og virker rimelig: Det er flere gifte og samboende enn aleneboende som er hjemme og kan treffes ved første henvendelse. Sannsynlighetene varierer fra 0,90 (aleneboende menn i boregion 1) til 0,97 (gifte eller samboende kvinner i boregion 2).

B.3. Modell for R^2

De neste tabellene tar for seg MLE for $P\left(R^2 = 1 \mid \mathbf{z}, r^1, y\right)$. Det første tallet i hver celle er MLE basert på det reelle datasettet. Tallet i midten er gjennomsnittsverdien av tilsvarende estimat, basert på de 1000 simulerte datasettene beskrevet i kapittel 3.3. Lengst til høyre står tilhørende standardavvik. Det partielle frafallet avhenger av både R^1 og Y . Hver tabell viser y-verdi mot boregion.

I modell 2 og 3 for undersysselsetting har vi flere R^2 -er i modellen. Modell 2 har R^2_u , R^2_{18} og R^2_{19}, og modell 3 har R^2_u , R^2_{18}, R^2_{19} og R^2_{23}. I disse tilfellene lager vi en tabell for hver R^2 -variabel.

Spørsmål 18

Mann, aleneboende, r1=0 (svarte etter purring)

y	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
0	0,792 0,791 0,06	0,724 0,724 0,07	0,759 0,759 0,06	0,748 0,745 0,07	0,719 0,719 0,08
1	0,637 0,647 0,09	0,547 0,562 0,10	0,591 0,606 0,10	0,577 0,590 0,10	0,541 0,557 0,10

Mann, gift el. samboer, r1=0

y	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
0	0,867 0,865 0,04	0,817 0,816 0,05	0,843 0,842 0,04	0,835 0,832 0,05	0,814 0,812 0,06
1	0,750 0,754 0,08	0,673 0,682 0,09	0,712 0,720 0,08	0,700 0,707 0,08	0,668 0,678 0,09

Mann, aleneboende, r1=1 (svarte ved 1. henvendelse)

y	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
0	0,890 0,888 0,03	0,848 0,846 0,04	0,870 0,869 0,03	0,863 0,859 0,04	0,845 0,842 0,05
1	0,788 0,792 0,05	0,719 0,727 0,06	0,755 0,761 0,06	0,744 0,749 0,06	0,714 0,732 0,06

Mann, gift eller samboer, r1=1

y	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
0	0,932 0,931 0,02	0,905 0,903 0,02	0,919 0,919 0,02	0,915 0,912 0,02	0,903 0,900 0,03
1	0,864 0,865 0,04	0,814 0,818 0,05	0,840 0,843 0,04	0,832 0,835 0,04	0,810 0,815 0,05

Kvinne, aleneboende, r1=0

y	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
0	0,884 0,884 0,03	0,840 0,840 0,04	0,863 0,863 0,04	0,856 0,854 0,04	0,837 0,836 0,05
1	0,779 0,782 0,07	0,707 0,716 0,08	0,744 0,751 0,08	0,733 0,739 0,07	0,703 0,712 0,08

Kvinne, gift eller samboende, r1=0

y	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
0	0,929 0,928 0,02	0,900 0,899 0,03	0,915 0,915 0,02	0,919 0,909 0,03	0,898 0,897 0,03
1	0,857 0,858 0,05	0,805 0,809 0,06	0,832 0,835 0,06	0,824 0,826 0,06	0,801 0,806 0,06

Kvinne, aleneboende, r1=1

y	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
0	0,942 0,941 0,01	0,918 0,917 0,02	0,930 0,930 0,01	0,926 0,925 0,02	0,916 0,915 0,02
1	0,882 0,883 0,03	0,837 0,841 0,04	0,860 0,863 0,04	0,853 0,856 0,04	0,834 0,838 0,04

Kvinner, gift eller samboende, r1=1

y	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
0	0,965 0,965 0,01	0,950 0,950 0,01	0,958 0,958 0,01	0,955 0,955 0,01	0,949 0,948 0,01
1	0,927 0,927 0,02	0,898 0,899 0,03	0,913 0,915 0,02	0,908 0,910 0,02	0,895 0,897 0,03

MLE er tilnærmet forventningsrette. Standardavvikene er stort sett små. De største standardavvikene finner vi for aleneboende menn, som først svarte etter purring, men da svarte ja på spørsmålet. I denne gruppen er det bare 9 personer, fordelt på 5 boregioner.

Vi ser også at det er stor forskjell mellom svarsannsynlighetene for $y = 0$ og $y = 1$. Dette viser viktigheten av å ha med y i modellen.

Spørsmål 19

Mann, aleneboende, r1=0 (svarte etter purring)

y	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
0	1,000 0,999 0,00	1,000 1,000 0,00	1,000 1,000 0,00	1,000 1,000 0,00	1,000 1,000 0,00
1	1,000 1,000 0,00	1,000 1,000 0,00	1,000 1,000 0,00	1,000 1,000 0,00	1,000 1,000 0,00

Mann, gift el. samboer, r1=0

y	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
0	1,000 0,999 0,00	1,000 1,000 0,00	1,000 1,000 0,00	1,000 1,000 0,00	1,000 0,999 0,02
1	1,000 1,000 0,00	1,000 1,000 0,00	1,000 1,000 0,00	1,000 1,000 0,00	1,000 0,999 0,02

Mann, aleneboende, r1=1 svarte ved 1. henvendelse)

y	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
0	1,000 0,985 0,03	1,000 1,000 0,00	1,000 1,000 0,00	1,000 1,000 0,00	1,000 0,979 0,04
1	0,985 0,990 0,01	1,000 1,000 0,00	1,000 1,000 0,00	1,000 1,000 0,00	0,956 0,972 0,04

Mann, gift eller samboer, r1=1

y	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
0	1,000 0,978 0,06	1,000 1,000 0,00	1,000 1,000 0,00	1,000 1,000 0,00	1,000 0,974 0,05
1	0,986 0,990 0,01	1,000 1,000 0,00	1,000 1,000 0,00	1,000 1,000 0,00	0,961 0,972 0,04

Kvinne, aleneboende, r1=0

y	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
0	1,000 0,999 0,00	1,000 1,000 0,00	1,000 1,000 0,00	1,000 1,000 0,00	1,000 0,999 0,03
1	1,000 0,999 0,01	1,000 1,000 0,00	1,000 1,000 0,00	1,000 1,000 0,00	1,000 0,999 0,01

Kvinne, gift eller samboende, r1=0

y	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
0	1,000 0,999 0,00	1,000 1,000 0,00	1,000 1,000 0,00	1,000 1,000 0,00	1,000 0,998 0,03
1	1,000 0,999 0,02	1,000 1,000 0,00	1,000 1,000 0,00	1,000 1,000 0,00	1,000 0,997 0,04

Kvinne, aleneboende, r1=1

y	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
0	1,000 0,981 0,03	1,000 1,000 0,00	1,000 1,000 0,00	1,000 1,000 0,00	1,000 0,970 0,05
1	0,972 0,982 0,02	1,000 1,000 0,00	1,000 1,000 0,00	1,000 1,000 0,00	0,922 0,950 0,06

Kvinner, gift eller samboende,r1=1

y	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
0	1,000 0,982 0,02	1,000 1,000 0,00	1,000 1,000 0,00	1,000 1,000 0,00	1,000 0,974 0,03
1	0,974 0,984 0,02	1,000 1,000 0,00	1,000 1,000 0,00	1,000 1,000 0,00	0,929 0,957 0,05

I utvalget er det partielle frafallet for spørsmål 19 lavt. Det samme gjelder for simuleringstilfellene:
I de fleste strata er den gjennomsnittlige, betingede sannsynlighet for å svare på spørsmål 19, lik en.

Spørsmål 23a

Mann, aleneboende, r1=0 (svarte etter purring)

y	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
0	1,000 1,000 0,00	1,000 1,000 0,00	1,000 1,000 0,00	1,000 1,000 0,00	1,000 1,000 0,00
1	1,000 1,000 0,00	1,000 1,000 0,00	1,000 1,000 0,00	1,000 1,000 0,00	1,000 1,000 0,00

Mann, gift el. samboer, r1=0

y	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
0	1,000 1,000 0,00	1,000 1,000 0,00	1,000 1,000 0,00	1,000 1,000 0,00	1,000 1,000 0,00
1	1,000 1,000 0,00	1,000 1,000 0,00	1,000 1,000 0,00	1,000 1,000 0,00	1,000 1,000 0,00

Mann, aleneboende, r1=1 (svarte ved 1. henvendelse)

y	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
0	1,000 0,909 0,13	1,000 0,926 0,12	1,000 0,900 0,15	1,000 0,946 0,10	1,000 0,880 0,17
1	0,923 0,937 0,03	0,958 0,963 0,02	0,934 0,941 0,03	0,970 0,974 0,02	0,903 0,916 0,04

Mann, gift eller samboer, r1=1

y	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
0	1,000 0,919 0,13	1,000 0,934 0,12	1,000 0,911 0,14	1,000 0,952 0,09	1,000 0,894 0,17
1	0,940 0,950 0,03	0,968 0,970 0,02	0,949 0,953 0,02	0,977 0,979 0,01	0,925 0,933 0,03

Kvinne, aleneboende, r1=0

y	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
0	1,000 1,000 0,00	1,000 1,000 0,00	1,000 1,000 0,00	1,000 1,000 0,00	1,000 1,000 0,00
1	1,000 1,000 0,00	1,000 1,000 0,00	1,000 1,000 0,00	1,000 1,000 0,00	1,000 1,000 0,00

Kvinne, gift eller samboende, r1=0

y	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
0	1,000 1,000 0,00	1,000 1,000 0,00	1,000 1,000 0,00	1,000 1,000 0,00	1,000 1,000 0,00
1	1,000 1,000 0,00	1,000 1,000 0,00	1,000 1,000 0,00	1,000 1,000 0,00	1,000 1,000 0,00

Kvinne, aleneboende, r1=1

y	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
0	1,000 0,921 0,11	1,000 0,937 0,10	1,000 0,912 0,13	1,000 0,955 0,07	1,000 0,895 0,14
1	0,918 0,935 0,04	0,956 0,963 0,02	0,930 0,940 0,03	0,969 0,973 0,02	0,897 0,914 0,05

Kvinner, gift eller samboende, r1=1

y	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
0	1,000 0,933 0,10	1,000 0,947 0,08	1,000 0,925 0,11	1,000 0,962 0,06	1,000 0,909 0,13
1	0,937 0,950 0,03	0,966 0,972 0,01	0,946 0,954 0,02	0,976 0,980 0,01	0,920 0,933 0,03

Estimatene er stort sett gode. De største skjevhetene og standardavvikene finner vi i strata med r1=1 og y = 0.

Undersysselsetting, modell 1

Mann, aleneboende, r1=0 (svarte etter purring)

y	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
0	0,781 0,783 0,06	0,696 0,701 0,08	0,732 0,735 0,07	0,721 0,726 0,08	0,671 0,676 0,08
1	0,688 0,710 0,14	0,586 0,620 0,16	0,628 0,655 0,15	0,615 0,648 0,15	0,558 0,595 0,17

Mann, gift el. samboer, r1=0

y	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
0	0,865 0,865 0,05	0,805 0,806 0,06	0,831 0,832 0,05	0,823 0,825 0,06	0,786 0,788 0,06
1	0,798 0,808 0,10	0,718 0,736 0,13	0,752 0,764 0,12	0,742 0,759 0,12	0,694 0,713 0,14

Mann, aleneboende, r1=1 (svarte ved 1. henvendelse)

y	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
0	0,871 0,871 0,03	0,813 0,814 0,04	0,838 0,838 0,04	0,831 0,832 0,04	0,795 0,795 0,05
1	0,807 0,815 0,09	0,729 0,743 0,11	0,762 0,771 0,11	0,752 0,766 0,11	0,706 0,720 0,13

Mann, gift eller samboer, r1=1

y	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
0	0,924 0,924 0,02	0,887 0,887 0,03	0,903 0,903 0,02	0,898 0,899 0,03	0,875 0,874 0,03
1	0,883 0,885 0,06	0,829 0,835 0,08	0,852 0,855 0,08	0,845 0,851 0,08	0,812 0,817 0,09

Kvinne, aleneboende, r1=0

y	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
0	0,876 0,877 0,04	0,819 0,822 0,04	0,844 0,845 0,04	0,836 0,839 0,05	0,801 0,804 0,05
1	0,813 0,818 0,10	0,737 0,749 0,13	0,769 0,775 0,12	0,760 0,771 0,12	0,714 0,727 0,14

Kvinne, gift eller samboende, r1=0

y	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
0	0,927 0,928 0,02	0,891 0,892 0,03	0,907 0,908 0,03	0,902 0,904 0,03	0,879 0,880 0,04
1	0,887 0,887 0,07	0,834 0,838 0,09	0,857 0,857 0,09	0,850 0,854 0,09	0,818 0,820 0,10

Kvinne, aleneboende, r1=1

y	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
0	0,930 0,931 0,02	0,896 0,897 0,02	0,911 0,912 0,02	0,907 0,908 0,02	0,885 0,885 0,03
1	0,892 0,891 0,06	0,842 0,843 0,08	0,863 0,862 0,08	0,857 0,859 0,08	0,826 0,826 0,10

Kvinne, gift eller samboende, r1=1

y	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
0	0,960 0,961 0,01	0,939 0,940 0,01	0,949 0,949 0,01	0,946 0,947 0,01	0,932 0,933 0,02
1	0,937 0,936 0,04	0,905 0,905 0,05	0,919 0,917 0,05	0,915 0,915 0,05	0,895 0,893 0,06

Det er litt mindre forskjell mellom svarsannsynlighetene for $y = 0$ og $y = 1$ enn det var på spørsmål 18.

Undersysselsetting, modell 2

Tabell over $P(R_u^2 = 1 \mid \mathbf{z}, y_u, R^1)$

Mann, aleneboende, r1=0 (svarte etter purring)

y	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
0	0,779 0,784 0,06	0,695 0,701 0,07	0,730 0,734 0,06	0,720 0,725 0,07	0,670 0,677 0,08
1	0,692 0,710 0,13	0,591 0,619 0,16	0,633 0,654 0,15	0,620 0,647 0,15	0,563 0,595 0,17

Mann, gift el. samboer, r1=0

y	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
0	0,864 0,866 0,04	0,804 0,807 0,05	0,830 0,832 0,05	0,822 0,825 0,05	0,785 0,789 0,06
1	0,802 0,809 0,10	0,723 0,736 0,13	0,756 0,764 0,12	0,746 0,759 0,12	0,699 0,715 0,13

Mann, aleneboende, r1=1 (svarte ved 1. henvendelse)

y	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
0	0,870 0,872 0,03	0,812 0,814 0,04	0,837 0,838 0,03	0,830 0,831 0,04	0,794 0,796 0,05
1	0,810 0,817 0,09	0,733 0,745 0,11	0,766 0,773 0,10	0,756 0,767 0,10	0,710 0,724 0,12

Mann, gift eller samboer, r1=1

y	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
0	0,924 0,924 0,02	0,886 0,887 0,03	0,903 0,903 0,02	0,898 0,898 0,03	0,874 0,875 0,03
1	0,885 0,887 0,06	0,832 0,836 0,08	0,855 0,856 0,07	0,848 0,853 0,07	0,815 0,820 0,09

Kvinne, aleneboende, r1=0

y	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
0	0,875 0,877 0,03	0,818 0,821 0,04	0,843 0,844 0,04	0,835 0,838 0,04	0,800 0,804 0,05
1	0,816 0,817 0,10	0,741 0,747 0,13	0,773 0,774 0,12	0,763 0,770 0,12	0,718 0,727 0,14

Kvinne, gift eller samboende, r1=0

y	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
0	0,926 0,928 0,02	0,890 0,892 0,03	0,906 0,907 0,02	0,901 0,903 0,03	0,878 0,881 0,03
1	0,889 0,887 0,07	0,837 0,837 0,09	0,860 0,856 0,08	0,853 0,853 0,08	0,821 0,821 0,10

Kvinne, aleneboende, r1=1

y	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
0	0,930 0,931 0,02	0,895 0,897 0,02	0,911 0,911 0,02	0,906 0,907 0,02	0,884 0,885 0,03
1	0,894 0,892 0,06	0,845 0,844 0,08	0,866 0,863 0,07	0,860 0,860 0,07	0,829 0,828 0,09

Kvinne, gift eller samboende, r1=1

y	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
0	0,960 0,961 0,01	0,939 0,940 0,01	0,948 0,949 0,01	0,946 0,946 0,01	0,932 0,933 0,01
1	0,938 0,937 0,04	0,907 0,906 0,05	0,921 0,918 0,05	0,917 0,916 0,05	0,897 0,895 0,06

For y = 0 er sannsynligheten for å svare litt mindre under modell 2 enn modell 1 (forskjellen er stort sett 0,001). For y = 1 er svarsannsynligheten noe høyere under modell 2 enn modell 1 (forskjellene varierer mellom 0,002 og 0,005). Dette betyr at forskjellen mellom svarsannsynlighetene for y = 0 og y =1 blir litt mindre for modell 2 enn for modell 1.

Tabell over $P(R_{18}^2 = 1 | \mathbf{z}, y_u, R^1, R_u^2 = 0)$

Siden leddene med $\mathbf{z}$ og R^1 er fjernet, er denne sannsynligheten bare avhengig av y_u.

y	MLE reellt datasett	Gjennom-snitt (1000)	Standard-avvik (1000)
0	0,051	0,043	0,03
1	0,019	0,094	0,21

Sannsynligheten for å ha svart på spørsmål 18 når undersysselsetting er ubesvart, er liten. I vårt datasett har vi 325 personer som ikke har svart på spørsmålet om undersysselsetting, og bare 14 av disse har svart på spørsmål 18. Ellers ser vi at standardavviket til MLE blir høyt for $y_u = 1$.

Tabell over $P(R_{19}^2 = 1 | \mathbf{z}, y_u, R^1, R_u^2 = 0, R_{18}^2 = 1)$

Modellen inneholder bare konstantledd.

MLE reellt datasett	Gjennom-snitt (1000)	Standard-avvik (1000)
0,643	0,641	0,13

Vi har 14 personer med $R_u^2 = 0$ og $R_{18}^2 = 1$ i datasettet. 9 av disse har svart på spørsmål 19. Dette svarer til en andel på 9/14 = 0,643. MLE faller sammen med denne andelen fordi modellen bare inneholder konstantledd.

Undersysselsetting, modell 3

I dette tilfellet er ikke R_u^2 modellert direkte, så vi starter med å finne $P(R_u^2 = 1 | \mathbf{z}, R^1, y_{18}, y_{19}, y_{23})$ under modell 3. I tabellen under har vi satt opp alle variabelkombinasjoner som er mulige under modell 3. Y - ene betegner her de faktiske verdiene, og ikke de observerte. I kolonnen lengst til høyre har vi satt opp R_u^2.

R_{18}^2	Y_{18}	R_{19}^2	Y_{19}	R_{23}^2	Y_{23}	R_u^2
0	0	0	ikke definert	0	ikke definert	0
0	1	0	0	0	ikke definert	0
0	1	0	1	0	0	0
0	1	0	1	0	1	0
1	0	ikke definert	ikke definert	ikke definert	ikke definert	1
1	1	0	0	0	ikke definert	0
1	1	0	1	0	0	0
1	1	0	1	0	1	0
1	1	1	0	ikke definert	ikke definert	1
1	1	1	1	0	0	0
1	1	1	1	0	1	0
1	1	1	1	1	0	1
1	1	1	1	1	1	1

Fire av kombinasjonene gir $R_u^2 = 1$. Vi kan dermed splitte begivenheten $R_u^2 = 1$ opp i de fire gjensidig utelukkende begivenhetene som svarer til disse variabelkombinasjonene. De to nederste kombinasjonene kan slås sammen til begivenheten $R_{18}^2 = 1 \cap Y_{18} = 1 \cap R_{19}^2 = 1 \cap Y_{19} = 1 \cap R_{23}^2 = 1$. Vi får da at under modell 3 er

$$
\begin{aligned}
&P(R_u^2 = 1 \mid \mathbf{z}, R^1, y_{18}, y_{19}, y_{23}) \\
&= P(R_{18}^2 = 1 \cap Y_{18} = 0 \mid \mathbf{z}, R^1, y_{18}, y_{19}, y_{23}) \\
&+ P(R_{18}^2 = 1 \cap Y_{18} = 1 \cap R_{19}^2 = 1 \cap Y_{19} = 0 \mid \mathbf{z}, R^1, y_{18}, y_{19}, y_{23}) \\
&+ P(R_{18}^2 = 1 \cap Y_{18} = 1 \cap R_{19}^2 = 1 \cap Y_{19} = 1 \cap R_{23}^2 = 1 \mid \mathbf{z}, R^1, y_{18}, y_{19}, y_{23})
\end{aligned}
$$

Vi skriver ut denne sannsynligheten for de mulige verdiene av (y_{18}, y_{19}, y_{23}).

$Y_{18} = 0$: I dette tilfellet er Y_{19} og Y_{23} udefinerte. De to siste sannsynlighetene i summen over blir null, fordi de inneholder begivenheten $Y_{18} = 1$.

$$
\begin{aligned}
&P(R_u^2 = 1 \mid \mathbf{z}, R^1, Y_{18} = 0) \\
&= P(R_{18}^2 = 1 \cap Y_{18} = 0 \mid \mathbf{z}, R^1, Y_{18} = 0) + 0 + 0 \\
&= P(R_{18}^2 = 1 \mid \mathbf{z}, R^1, Y_{18} = 0)
\end{aligned}
$$

$Y_{18} = 1, Y_{19} = 0$: I dette tilfellet er Y_{23} udefinert.

$$
\begin{aligned}
&P(R_u^2 = 1 \mid \mathbf{z}, R^1, Y_{18} = 1, Y_{19} = 0) \\
&= 0 + P(R_{18}^2 = 1 \cap R_{19}^2 = 1 \mid \mathbf{z}, R^1, Y_{18} = 1, Y_{19} = 0) + 0 \\
&= P(R_{19}^2 = 1 \mid \mathbf{z}, R^1, Y_{18} = 1, Y_{19} = 0, R_{18}^2 = 1) P(R_{18}^2 = 1 \mid \mathbf{z}, R^1, Y_{18} = 1)
\end{aligned}
$$

$(y_{18}, y_{19}, y_{23}) = (1,1,0)$:

$$P(R_u^2 = 1 \mid \mathbf{z}, R^1, Y_{18} = 1, Y_{19} = 1, Y_{23} = 0)$$
$$= 0 + 0 + P(R_{18}^2 = 1 \cap R_{19}^2 = 1 \cap R_{23}^2 = 1 \mid \mathbf{z}, R^1, Y_{18} = 1, Y_{19} = 1, Y_{23} = 0)$$
$$= P(R_{23}^2 = 1 \mid \mathbf{z}, R^1, Y_{18} = 1, Y_{19} = 1, Y_{23} = 0, R_{18}^2 = 1, R_{19}^2 = 1)$$
$$* P(R_{19}^2 = 1 \mid \mathbf{z}, R^1, Y_{18} = 1, Y_{19} = 1, R_{18}^2 = 1) P(R_{18}^2 = 1 \mid \mathbf{z}, R^1, Y_{18} = 1)$$

$(y_{18}, y_{19}, y_{23}) = (1,1,1)$:

$$P(R_u^2 = 1 \mid \mathbf{z}, R^1, Y_{18} = 1, Y_{19} = 1, Y_{23} = 1)$$
$$= 0 + 0 + P(R_{18}^2 = 1 \cap R_{19}^2 = 1 \cap R_{23}^2 = 1 \mid \mathbf{z}, R^1, Y_{18} = 1, Y_{19} = 1, Y_{23} = 1)$$
$$= P(R_{23}^2 = 1 \mid \mathbf{z}, R^1, Y_{18} = 1, Y_{19} = 1, Y_{23} = 1, R_{18}^2 = 1, R_{19}^2 = 1)$$
$$* P(R_{19}^2 = 1 \mid \mathbf{z}, R^1, Y_{18} = 1, Y_{19} = 1, R_{18}^2 = 1) P(R_{18}^2 = 1 \mid \mathbf{z}, R^1, Y_{18} = 1)$$

Vi tabulerer MLE for $P(R_u^2 = 1 \mid \mathbf{z}, R^1, \mathbf{y})$. Her står $\mathbf{y}$ for vektoren (y_{18}, y_{19}, y_{23}). $\mathbf{y} = (0, ,)$ betyr at $y_{18} = 0$ og at Y_{19} og Y_{23} er udefinerte. $\mathbf{y} = (1,0,)$ betyr at $y_{18} = 1$, $y_{19} = 0$ og Y_{23} udefinert.

Mann, aleneboende, r1=0 (svarte etter purring)

y	**boreg = 1**	**boreg = 2**	**boreg = 3**	**boreg = 4**	**boreg = 5**
(0, ,)	0,951 0,942 0,06	0,933 0,919 0,08	0,940 0,928 0,07	0,948 0,933 0,07	0,935 0,920 0,08
(1,0,)	0,486 0,499 0,10	0,403 0,414 0,09	0,433 0,445 0,09	0,470 0,479 0,09	0,413 0,424 0,09
(1,1,0)	0,478 0,491 0,10	0,397 0,407 0,09	0,426 0,437 0,09	0,462 0,472 0,09	0,406 0,417 0,09
(1,1,1)	0,478 0,491 0,10	0,397 0,407 0,09	0,426 0,437 0,09	0,462 0,472 0,09	0,406 0,417 0,09

Mann, gift el. samboer, r1=0

y	**boreg = 1**	**boreg = 2**	**boreg = 3**	**boreg = 4**	**boreg = 5**
(0, ,)	0,969 0,963 0,04	0,957 0,947 0,06	0,962 0,953 0,05	0,967 0,957 0,05	0,959 0,948 0,06
(1,0,)	0,603 0,614 0,10	0,521 0,531 0,09	0,551 0,562 0,09	0,588 0,596 0,09	0,531 0,540 0,10
(1,1,0)	0,594 0,604 0,09	0,513 0,522 0,09	0,542 0,552 0,09	0,579 0,586 0,09	0,522 0,531 0,09
(1,1,1)	0,594 0,604 0,09	0,513 0,522 0,09	0,542 0,552 0,09	0,579 0,586 0,09	0,522 0,531 0,09

Mann, aleneboende, r1=1 (svarte ved 1. henvendelse)

y	**boreg = 1**	**boreg = 2**	**boreg = 3**	**boreg = 4**	**boreg = 5**
(0, ,)	0,979 0,973 0,03	0,970 0,961 0,04	0,974 0,966 0,04	0,977 0,968 0,04	0,971 0,961 0,04
(1,0,)	0,688 0,698 0,06	0,613 0,621 0,05	0,641 0,650 0,05	0,674 0,682 0,05	0,622 0,630 0,06
(1,1,0)	0,677 0,687 0,06	0,603 0,611 0,05	0,630 0,640 0,05	0,663 0,671 0,05	0,612 0,619 0,06
(1,1,1)	0,677 0,687 0,06	0,603 0,611 0,05	0,630 0,640 0,05	0,663 0,671 0,05	0,612 0,619 0,06

Mann, gift eller samboer, r1=1

y	**boreg = 1**	**boreg = 2**	**boreg = 3**	**boreg = 4**	**boreg = 5**
(0, ,)	0,987 0,983 0,02	0,981 0,975 0,03	0,983 0,979 0,02	0,986 0,980 0,02	0,982 0,976 0,03
(1,0,)	0,780 0,788 0,05	0,718 0,726 0,05	0,741 0,750 0,05	0,769 0,776 0,04	0,725 0,732 0,05
(1,1,0)	0,767 0,775 0,05	0,706 0,714 0,05	0,729 0,737 0,04	0,756 0,763 0,04	0,713 0,720 0,05
(1,1,1)	0,767 0,775 0,05	0,706 0,714 0,05	0,729 0,737 0,04	0,756 0,763 0,04	0,713 0,720 0,05

Kvinne, aleneboende, r1=0

y	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
(0, ,)	0,971 0,966 0,03	0,959 0,952 0,05	0,964 0,958 0,04	0,969 0,961 0,04	0,961 0,952 0,05
(1,0,)	0,616 0,625 0,10	0,534 0,543 0,10	0,564 0,574 0,10	0,600 0,608 0,09	0,544 0,553 0,10
(1,1,0)	0,606 0,615 0,10	0,525 0,534 0,09	0,555 0,564 0,10	0,591 0,598 0,09	0,535 0,543 0,10
(1,1,1)	0,606 0,615 0,10	0,525 0,534 0,09	0,555 0,564 0,10	0,591 0,598 0,09	0,535 0,543 0,10

Kvinne, gift eller samboende, r1=0

y	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
(0, ,)	0,982 0,979 0,02	0,974 0,969 0,03	0,977 0,973 0,03	0,980 0,975 0,03	0,975 0,970 0,03
(1,0,)	0,720 0,726 0,08	0,648 0,655 0,09	0,675 0,682 0,09	0,707 0,712 0,08	0,657 0,664 0,09
(1,1,0)	0,708 0,714 0,08	0,638 0,644 0,09	0,664 0,671 0,08	0,696 0,700 0,08	0,646 0,653 0,09
(1,1,1)	0,708 0,714 0,08	0,638 0,644 0,09	0,664 0,671 0,08	0,696 0,700 0,08	0,646 0,653 0,09

Kvinne, aleneboende, r1=1

y	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
(0, ,)	0,987 0,985 0,02	0,982 0,978 0,02	0,984 0,981 0,02	0,986 0,982 0,02	0,983 0,978 0,02
(1,0,)	0,788 0,795 0,05	0,728 0,735 0,05	0,751 0,758 0,05	0,777 0,784 0,04	0,735 0,742 0,06
(1,1,0)	0,775 0,782 0,05	0,716 0,723 0,05	0,738 0,746 0,05	0,765 0,771 0,04	0,723 0,729 0,06
(1,1,1)	0,775 0,782 0,05	0,716 0,723 0,05	0,738 0,746 0,05	0,765 0,771 0,04	0,723 0,729 0,06

Kvinne, gift eller samboende, r1=1

y	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
(0, ,)	0,992 0,991 0,01	0,989 0,986 0,01	0,990 0,988 0,01	0,992 0,989 0,01	0,989 0,986 0,01
(1,0,)	0,856 0,862 0,03	0,811 0,817 0,03	0,828 0,835 0,03	0,848 0,854 0,03	0,816 0,822 0,04
(1,1,0)	0,842 0,847 0,03	0,797 0,804 0,03	0,815 0,821 0,03	0,834 0,840 0,03	0,803 0,808 0,04
(1,1,1)	0,842 0,847 0,03	0,797 0,804 0,03	0,815 0,821 0,03	0,834 0,840 0,03	0,803 0,808 0,04

Sannsynligheten for å svare på undersysselsetting gitt at man er undersysselsatt (nederste linje i hver tabell), er gjennomgående lavere under modell 3 enn under modell 1 og 2.

Vi kan legge merke til at sannsynligheten for å besvare spørsmålet om undersysselsetting synker når antall spørsmål som må stilles øker. De laveste svarsannsynlighetene finner vi i de to nederste linjene, som er like i alle tabellene. De gjelder personer som har svaret "ja" både på spørsmål 18 og 19, og som derfor må besvare tre spørsmål for å svare på spørsmålet om undersysselsetting. I nest øverste linje er sannsynlighetene litt høyere enn i de to nederste linjene. Denne linja gjelder personer med "ja" på spørsmål 18 og "nei" på spørsmål 19, så de har svart på undersysselsetting etter to spørsmål. I den øverste linja er sannsynlighetene vesentlig større. Dette er personer med "nei" på spørsmål 18, og disse trenger derfor bare å svare på ett spørsmål i sekvensen om undersysselsetting.

Ellers ser vi at estimatene er tilnærmet forventnigsrette, og standardavvikene er relativt små.

Tabell over $P(R_{18}^2 = 1 \mid \mathbf{z}, y_{18}, R^1)$

Mann, aleneboende, r1=0 (svarte etter purring)

y	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
0	0,951 0,942 0,06	0,933 0,919 0,08	0,940 0,928 0,07	0,948 0,933 0,07	0,935 0,920 0,08
1	0,489 0,502 0,10	0,406 0,416 0,09	0,435 0,447 0,09	0,472 0,482 0,09	0,415 0,426 0,09

Mann, gift el. samboer, r1=0

y	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
0	0,969 0,963 0,04	0,957 0,947 0,06	0,962 0,953 0,05	0,967 0,957 0,05	0,959 0,948 0,06
1	0,607 0,617 0,10	0,524 0,534 0,09	0,554 0,565 0,09	0,591 0,599 0,09	0,534 0,543 0,10

Mann, aleneboende, r1=1 (svarte ved 1. henvendelse)

y	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
0	0,979 0,973 0,03	0,970 0,961 0,04	0,974 0,966 0,04	0,977 0,968 0,04	0,971 0,961 0,04
1	0,692 0,702 0,06	0,616 0,624 0,05	0,644 0,654 0,05	0,678 0,686 0,05	0,625 0,633 0,06

Mann, gift eller samboer, r1=1

y	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
0	0,987 0,983 0,02	0,981 0,975 0,03	0,983 0,979 0,02	0,986 0,980 0,02	0,982 0,976 0,03
1	0,784 0,792 0,05	0,721 0,729 0,05	0,745 0,754 0,05	0,773 0,780 0,04	0,729 0,736 0,05

Kvinne, aleneboende, r1=0

y	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
0	0,971 0,966 0,03	0,959 0,952 0,05	0,964 0,958 0,04	0,969 0,961 0,04	0,961 0,952 0,05
1	0,619 0,628 0,10	0,537 0,546 0,10	0,567 0,576 0,10	0,604 0,611 0,09	0,546 0,555 0,10

Kvinne, gift eller samboende, r1=0

y	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
0	0,982 0,979 0,02	0,974 0,969 0,03	0,977 0,973 0,03	0,980 0,975 0,03	0,975 0,970 0,03
1	0,724 0,730 0,08	0,652 0,659 0,09	0,679 0,686 0,09	0,711 0,716 0,08	0,661 0,667 0,09

Kvinne, aleneboende, r1=1

y	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
0	0,987 0,985 0,02	0,982 0,978 0,02	0,984 0,981 0,02	0,986 0,982 0,02	0,983 0,978 0,02
1	0,792 0,799 0,05	0,732 0,739 0,05	0,755 0,762 0,05	0,782 0,788 0,04	0,739 0,745 0,06

Kvinner, gift eller samboende, r1=1

y	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
0	0,992 0,991 0,01	0,989 0,986 0,01	0,990 0,988 0,01	0,992 0,989 0,01	0,989 0,986 0,01
1	0,860 0,866 0,03	0,815 0,822 0,03	0,832 0,839 0,03	0,852 0,858 0,03	0,821 0,826 0,04

Vi får andre tall enn vi gjorde under den separate analysen av spørsmål 18. Dette er fordi vi har fjernet pensjonistene når vi ser på spørsmålet om undersysselsetting under ett.

Sammenligner vi første linje i tabellene over (y = 0) med første linje i tabellene for R_u^2 (y = (0, ,)), ser vi at de er like. Dette er ingen tilfeldighet, fordi når $y_{18} = 0$, er det å svare på spørsmål 18 det samme som å svare på spørsmålet om undersysselsetting.

Siste linje i tabellene over er ikke direkte sammenlignbar med noen linje i tabellene for R_u^2, siden alle de tre nederste linjene i tabellene for R_u^2 gjelder $y_{18} = 1$. Vi kan likevel legge merke til at sannsynlighetene i nederste linje i tabellene over er litt større enn de tilsvarende sannsynlighetene i alle de tre nederste linjene i tabellene for R_u^2. Dette er rimelig, for når $y_{18} = 1$ er det mer krevende å svare på undersysselsetting enn å bare svare på spørsmål 18.

Tabell over $P(R_{19}^2 = 1 \mid \mathbf{z}, y_{19}, R^1, Y_{18} = 1, R_{18}^2 = 1)$

Modellen inneholder bare konstantledd.

MLE reellt datasett	Gjennom-snitt (1000)	Standard-avvik (1000)
0,9947	0,9948	0,0023

De som har svart ja på spørsmål 18 svarer nesten alltid på spørsmål 19 (945 av 950 i vårt datasett).

Tabell over $P(R_{23}^2 = 1 \mid \mathbf{z}, y_{23}, R^1, Y_{18} = 1, Y_{19} = 1, R_{18}^2 = 1, R_{19}^2 = 1)$

Modellen inneholder bare konstantledd.

MLE reellt datasett	Gjennom-snitt (1000)	Standard-avvik (1000)
0,9837	0,9833	0,0055

De som har svart ja på spørsmål 18 og 19 svarer nesten alltid på spørsmål 23 (544 av 553 i vårt datasett).

Spørsmål 24

Mann, aleneboende, r1=0 (svarte etter purring)

y	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
0	0,935 0,936 0,07	0,922 0,919 0,04	0,865 0,872 0,11	0,870 0,880 0,11	0,828 0,837 0,14
1	0,977 0,977 0,03	0,959 0,960 0,05	0,950 0,950 0,06	0,952 0,953 0,06	0,935 0,939 0,06

Mann, gift el. samboer, r1=0

y	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
0	0,963 0,959 0,06	0,934 0,925 0,08	0,920 0,913 0,09	0,923 0,918 0,09	0,897 0,885 0,13
1	0,987 0,987 0,02	0,976 0,976 0,03	0,971 0,970 0,04	0,972 0,972 0,04	0,962 0,964 0,04

Mann, aleneboende, r1=1 (svarte ved 1. henvendelse)

y	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
0	0,955 0,955 0,07	0,922 0,919 0,05	0,905 0,904 0,05	0,909 0,908 0,06	0,878 0,870 0,09
1	0,984 0,984 0,02	0,972 0,973 0,02	0,966 0,965 0,03	0,967 0,968 0,03	0,955 0,957 0,04

Mann, gift eller samboer, r1=1

y	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
0	0,975 0,971 0,03	0,955 0,947 0,04	0,945 0,937 0,04	0,947 0,940 0,05	0,928 0,912 0,08
1	0,991 0,991 0,01	0,984 0,985 0,01	0,980 0,980 0,02	0,981 0,982 0,02	0,974 0,975 0,04

Kvinne, aleneboende, r1=0

y	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
0	0,970 0,970 0,03	0,946 0,948 0,05	0,934 0,938 0,05	0,936 0,942 0,06	0,914 0,919 0,08
1	0,990 0,987 0,02	0,981 0,978 0,03	0,976 0,976 0,04	0,978 0,974 0,04	0,969 0,966 0,04

Kvinne, gift eller samboende, r1=0

y	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
0	0,983 0,982 0,02	0,969 0,969 0,03	0,962 0,963 0,04	0,963 0,965 0,03	0,950 0,950 0,05
1	0,994 0,993 0,01	0,989 0,988 0,02	0,986 0,985 0,02	0,987 0,986 0,02	0,982 0,981 0,02

Kvinne, aleneboende, r1=1

y	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
0	0,980 0,980 0,02	0,963 0,964 0,02	0,955 0,956 0,02	0,957 0,959 0,03	0,941 0,941 0,04
1	0,993 0,991 0,01	0,987 0,985 0,01	0,984 0,980 0,02	0,985 0,982 0,02	0,979 0,977 0,02

Kvinne, gift eller samboende,r1=1

y	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
0	0,988 0,988 0,01	0,979 0,980 0,01	0,974 0,975 0,01	0,976 0,976 0,01	0,950 0,949 0,05
1	0,996 0,996 0,01	0,993 0,993 0,01	0,991 0,990 0,01	0,991 0,991 0,01	0,982 0,981 0,02

MLE er tilnærmet forventningsrette, og standardavvikene er stort sett små.

Målgruppen for dette spørsmålet er deltidssysselsatte med ønske om lengre arbeidstid. Vi ser at svarsannsynlighetene er litt høyere for y = 1 (ønsket arbeidstid større eller lik 37 timer) enn for y = 0 (ønsket arbeidstid mindre enn 37 timer).

Spørsmål 25

Mann, aleneboende, r1=0 (svarte etter purring)

y	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
0	0,950 0,950 0,03	0,946 0,945 0,03	0,906 0,904 0,04	0,911 0,910 0,04	0,863 0,863 0,06
1	0,950 0,952 0,01	0,946 0,947 0,02	0,905 0,906 0,03	0,911 0,912 0,03	0,863 0,866 0,04

Mann, gift el. samboer, r1=0

y	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
0	0,971 0,970 0,02	0,960 0,967 0,02	0,944 0,941 0,03	0,947 0,945 0,03	0,917 0,914 0,04
1	0,971 0,972 0,01	0,968 0,968 0,01	0,943 0,944 0,02	0,947 0,947 0,02	0,917 0,917 0,02

Mann, aleneboende, r1=1 (svarte ved 1. henvendelse)

y	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
0	0,985 0,984 0,01	0,946 0,945 0,03	0,983 0,983 0,00	0,972 0,970 0,01	0,955 0,954 0,02
1	0,950 0,952 0,01	0,946 0,947 0,01	0,970 0,969 0,01	0,972 0,972 0,01	0,955 0,955 0,01

Mann, gift eller samboer, r1=1

y	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
0	0,991 0,991 0,00	0,990 0,990 0,00	0,983 0,982 0,01	0,983 0,983 0,01	0,974 0,973 0,01
1	0,991 0,991 0,00	0,990 0,990 0,00	0,983 0,983 0,00	0,984 0,984 0,00	0,974 0,974 0,01

Kvinne, aleneboende, r1=0

y	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
0	0,980 0,980 0,01	0,980 0,980 0,01	0,960 0,961 0,01	0,963 0,963 0,01	0,941 0,943 0,02
1	0,980 0,980 0,01	0,978 0,977 0,01	0,960 0,958 0,02	0,963 0,960 0,02	0,940 0,938 0,03

Kvinne, gift eller samboende, r1=0

y	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
0	0,988 0,989 0,01	0,987 0,987 0,01	0,987 0,977 0,01	0,978 0,978 0,01	0,965 0,966 0,01
1	0,988 0,988 0,01	0,987 0,987 0,01	0,977 0,975 0,01	0,978 0,977 0,01	0,965 0,963 0,02

Kvinne, aleneboende, r1=1

y	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
0	0,994 0,994 0,00	0,993 0,994 0,00	0,988 0,988 0,00	0,987 0,989 0,00	0,982 0,982 0,01
1	0,994 0,994 0,03	0,993 0,993 0,00	0,988 0,987 0,00	0,989 0,987 0,01	0,981 0,981 0,01

Kvinne, gift eller samboende,r1=1

y	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
0	0,997 0,997 0,00	0,996 0,996 0,00	0,993 0,993 0,00	0,993 0,994 0,00	0,989 0,990 0,00
1	0,997 0,997 0,00	0,996 0,996 0,00	0,993 0,993 0,00	0,993 0,993 0,00	0,989 0,989 0,01

På dette spørsmålet har vi mange observasjoner, og sannsynlighetene i modellen for partielt frafall blir meget godt estimert.

Estimatene er svært like for y = 0 og y = 1, dvs. sannsynligheten for å svare på spørsmålet om faktisk arbeidstid er uavhengig av hva den faktiske arbeidstiden er.

Spørsmål 62

Mann, aleneboende, r1=0 (svarte etter purring)

y	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
0	0,875 0,876 0,07	0,743 0,746 0,12	0,808 0,808 0,10	0,873 0,867 0,08	0,669 0,674 0,15
1	0,897 0,899 0,05	0,783 0,787 0,08	0,839 0,840 0,06	0,895 0,895 0,05	0,715 0,720 0,10

Mann, gift el. samboer, r1=0

y	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
0	0,857 0,855 0,09	0,712 0,714 0,14	0,782 0,780 0,11	0,854 0,845 0,10	0,632 0,639 0,16
1	0,881 0,883 0,06	0,754 0,758 0,09	0,817 0,817 0,07	0,879 0,879 0,06	0,681 0,687 0,10

Mann, aleneboende, r1=1 (svarte ved 1. henvendelse)

y	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
0	0,866 0,866 0,06	0,729 0,727 0,10	0,796 0,792 0,08	0,865 0,865 0,08	0,652 0,649 0,13
1	0,890 0,891 0,03	0,769 0,770 0,03	0,829 0,828 0,04	0,888 0,888 0,03	0,699 0,698 0,05

Mann, gift eller samboer, r1=1

y	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
0	0,847 0,845 0,07	0,696 0,692 0,12	0,769 0,763 0,10	0,845 0,834 0,90	0,615 0,612 0,15
1	0,873 0,875 0,04	0,740 0,740 0,04	0,805 0,803 0,04	0,871 0,870 0,03	0,665 0,662 0,06

Kvinne, aleneboende, r1=0

y	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
0	0,923 0,926 0,04	0,832 0,837 0,08	0,878 0,881 0,06	0,922 0,921 0,05	0,775 0,780 0,11
1	0,937 0,936 0,04	0,860 0,860 0,07	0,899 0,897 0,05	0,936 0,935 0,04	0,810 0,813 0,80

Kvinne, gift eller samboende, r1=0

y	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
0	0,911 0,913 0,05	0,809 0,812 0,09	0,860 0,861 0,07	0,909 0,907 0,06	0,746 0,751 0,12
1	0,927 0,926 0,04	0,840 0,840 0,07	0,884 0,881 0,06	0,926 0,924 0,04	0,785 0,788 0,09

Kvinne, aleneboende, r1=1

y	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
0	0,917 0,920 0,03	0,821 0,825 0,06	0,870 0,873 0,04	0,916 0,915 0,04	0,762 0,763 0,09
1	0,932 0,932 0,03	0,851 0,851 0,04	0,892 0,890 0,04	0,931 0,931 0,02	0,799 0,798 0,05

Kvinne, gift eller samboende, r1=1

y	boreg = 1	boreg = 2	boreg = 3	boreg = 4	boreg = 5
0	0,904 0,908 0,04	0,797 0,799 0,07	0,851 0,853 0,05	0,903 0,901 0,04	0,732 0,732 0,10
1	0,922 0,922 0,03	0,830 0,830 0,04	0,876 0,874 0,04	0,920 0,920 0,03	0,772 0,772 0,05

MLE er tilnærmet forventningsrette, og standardavvikene er stort sett små.

Som for spørsmål 24, er svarsannsynlighetene litt høyere blant arbeidsledige som ønsker full arbeidstid enn blant dem som ønsker kortere arbeidstid.

Appendiks C. Simulering av MLE. Modellparametre.

I dette appendikset blir gjennomsnitt og standardavvik for MLE fra de 1000 simuleringene, beskrevet i kapittel 3.3, tabulert for hver parameter.

For variabelen *region* er gruppe 2 (Østlandet uten Oslo og Akershus) valgt som referansegruppe. For *kjønn* er det menn som er referansegruppen, og for *sivilstand* er alene referansegruppe. Aldersgruppen 40-66 år er referansegruppe for alle spørsmålene så nær som spørsmål 25, der referansegruppen er 20-66 år.

Referansegruppene er utelatt fra tabellene. Når det gjelder kryssleddene, vil også de verdiene utelates som inneholder referansegrupper. For β_4 (kjønn*alder) for eksempel, vil vi utelate verdiene som er knyttet til menn, og dessuten verdiene knyttet aldersgruppen 40-66 år.

I modellen for partielt frafall vil parametrene ψ_4 og ψ_5 være knyttet til hhv. y-verdien og R_1-verdien. Referanseverdien for y er 'nei' på spørsmålet. Vi tabulerer derfor parameteren knyttet til 'ja'. Tilsvarende vil referanseverdien for R_1 være 0 (dvs. at IO ikke svarer ved første henvendelse) og vi tabulerer parameteren for R_1=1.

Fylkenes inndeling i boregioner:

1: Oslo, Akershus
2: Hedmark, Oppland, Buskerud, Telemark, Østfold, Vestfold
3: Aust-Agder, Vest-Agder, Rogaland, Hordaland, Sogn og Fjordane
4: Møre og Romsdal, Sør Trøndelag, Nord Trøndelag
5: Nordland, Troms, Finnmark

Spørsmål 18

Populasjonsmodell

		MLE reelt datasett	gj.snitt (1000)	SD
β_0		-1,0757	-1,0879	0,2557
β_1	**aldersgruppe**			
	16-19	0,5679	0,5678	0,3351
	20-39	0,7037	0,7130	0,2445
	67-74	-3,2245	-19,7068	19,1700
β_2	**kjønn**			
	kvinner	-0,0690	-0,0625	0,2421
β_3	**boregion**			
	1	0,2740	0,2685	0,3226
	3	0,1681	0,1611	0,2948
	4	0,4736	0,4752	0,3501
	5	0,2191	0,2034	0,4022
β_4	**kjønn*alder**			
	kvinne, 16-19 år	0,0503	0,0373	0,3153
	kvinne, 20-39 år	-0,1981	-0,2082	0,2301
	kvinne, 67-74 år	1,7959	8,7179	13,7461
β_5	**kjønn*region**			
	kvinne, 1	-0,8281	-0,8334	0,2971
	kvinne, 3	-0,2982	-0,2885	0,2818
	kvinne, 4	-0,2682	-0,2693	0,3478
	kvinne, 5	-0,4369	-0,4370	0,3818
β_6	**alder*boregion**			
	16-19, 1	-0,3191	-0,3235	0,4332
	16-19, 3	-0,2458	-0,2473	0,3754
	16-19, 4	-0,4962	-0,5363	0,4907
	16-19, 5	0,0216	-0,0095	0,5875
	20-39, 1	0,0538	0,0569	0,2699
	20-39, 3	-0,1522	-0,1562	0,2211
	20-39, 4	0,0939	0,0879	0,2532
	20-39, 5	0,4914	0,5031	0,3118
	67-74, 1	1,4451	9,1412	15,6636
	67-74, 3	-0,3996	-0,3751	18,8811
	67-74, 4	0,1220	2,4638	18,4680
	67-74, 5	-27,0173	-17,9129	12,9662

Parametrene som angår aldersgruppen 67-74 år blir dårlig estimert. Gjennomsnittene fra simuleringene ligger stort sett langt unna MLE fra det reelle datasettet, og standardavvikene er svært store. Dette skyldes antageligvis at vi har relativt få deltidssysselsatte i denne aldersgruppen (130 personer). Når disse fordeles på fem regioner blir det små grupper, og estimeringen blir upresis. De øvrige estimatene er gode.

Modell for partielt frafall (R_2)

		MLE reelt datasett	gj.snitt (1000)	SD
ψ_0		0,9663	1,0297	1,0128
ψ_1	**sivilstand**			
	ikke-alene	0,5360	0,5391	0,1267
ψ_2	**kjønn**			
	kvinner	0,6962	0,7049	0,1353
ψ_3	**boregion**			
	1	0,3742	0,3763	0,1855
	3	0,1825	0,1893	0,1582
	4	0,1243	0,1186	0,1858
	5	-0,0233	-0,0187	0,2171
ψ_4	**y-verdi**			
	ja	-0,7770	-0,7508	1,2278
ψ_5	**R_1-verdi**			
	1	0,7536	0,7445	0,2381

Parametrene i modellen for partielt frafall blir godt estimert.

Spørsmål 19

Populasjonsmodell

		MLE reelt datasett	gj.snitt (1000)	SD
β_0		0,6286	0,6689	0,3892
β_1	**aldersgruppe**			
	16-19	-2,1184	-2,2539	0,6264
	20-39	0,5235	0,5341	0,4188
β_2	**kjønn**			
	kvinner	-0,5253	-0,5679	0,3952
β_3	**boregion**			
	1	1,5711	1,6105	0,6114
	3	0,2268	0,2367	0,4974
	4	-0,5535	-0,6118	0,5162
	5	0,4496	0,4110	0,6617
β_4	**kjønn*alder**			
	kvinne, 16-19 år	1,4981	1,6169	0,5719
	kvinne, 20-39 år	0,2075	0,2073	0,3900
β_5	**kjønn*region**			
	kvinne, 1	-0,9284	-0,9641	0,5534
	kvinne, 3	-0,2902	-0,2900	0,4784
	kvinne, 4	0,2367	0,2819	0,4980
	kvinne, 5	-0,1319	-0,1144	0,5835
β_6	**alder*boregion**			
	16-19, 1	0,7484	0,8723	1,6016
	16-19, 3	0,6963	0,7500	0,6617
	16-19, 4	0,2831	-0,4350	3,8747
	16-19, 5	0,2578	-0,0156	2,5958
	20-39, 1	-1,1398	-1,1686	0,4828
	20-39, 3	-0,2390	-0,2531	0,3704
	20-39, 4	-0,0676	-0,0474	0,3750
	20-39, 5	-1,3762	-1,4245	0,5271

Maksimumlikelihoodestimatorene ser ut til å være omtrent forventningsrette, bortsett fra estimatene for β_6 i den yngste aldersgruppen. I målgruppen til spørsmål 19 er det 93 personer i denne aldersgruppen. Dette blir nok i minste laget når modellen inneholder tre kryssledd.

Modell for partielt frafall (R_2)

		MLE reelt datasett	gj.snitt (1000)	SD
ψ_0		95,6601	92,0122	40,5362
ψ_1	**sivilstand**			
	ikke-alene	0,1118	-3,8564	16,4183
ψ_2	**kjønn**			
	kvinner	-0,6259	-10,3898	22,6924
ψ_3	**boregion**			
	1	-40,4803	-35,3520	18,2595
	3	-9,9410	-6,5499	12,0348
	4	-12,7243	-6,2841	11,0508
	5	-41,5453	-38,6968	17,3203
ψ_4	**y-verdi**			
	ja	-30,6451	-5,8535	23,3192
ψ_5	**R_1-verdi**			
	1	-20,3799	-21,6169	6,7801

Estimatene for ψ -ene er helt ubrukelige på dette spørsmålet. Dette har sammenheng med at vi har problemer med å identifisere målgruppen på spørsmål 19. Målgruppen for spørsmål 19 er deltidssysselsatte med ønske om lengre arbeidstid. Informasjon om hvorvidt en person ønsker lengre arbeidstid får vi fra spørsmål 18. Det betyr at hvis en deltidssysselsatt har partielt frafall på spørsmål 18, kan vi ikke vite om personen er i målgruppen for spørsmål 19 eller ikke. Vi har valgt å bare regne deltidssysselsatte som har svart ja på spørsmål 18 til målgruppen for spørsmål 19. Med denne definisjonen av målgruppen er det bare 5 personer med partielt frafall på spørsmål 19. Det er klart at dette er altfor lite til å estimere parametrene i frafallsmodellen.

Spørsmål 23a

Populasjonsmodell

		MLE reelt datasett	gj.snitt (1000)	SD
β_0		3,1641	7,2680	11,2428
β_1	**aldersgruppe**			
	16-19	-0,6448	2,0084	17,9904
	20-39	-0,4204	-3,3633	9,5845
β_2	**kjønn**			
	kvinner	-0,3874	-4,4718	11,2350
β_3	**boregion**			
	1	0,3606	1,9937	10,8000
	3	-0,1594	-0,4787	7,3748
	4	0,5064	9,6382	18,8095
	5	1,5038	18,2466	24,1300
β_4	**kjønn*alder**			
	kvinne, 16-19 år	2,3423	18,8027	22,4736
	kvinne, 20-39 år	0,0103	2,9085	9,5843
β_5	**kjønn*region**			
	kvinne, 1	-1,2061	-2,7546	10,5961
	kvinne, 3	0,1132	0,4808	7,1090
	kvinne, 4	-0,5364	-9,3517	18,5334
	kvinne, 5	-2,1138	-17,3779	22,6036
β_6	**alder*boregion**			
	16-19, 1	-0,9304	-6,4368	18,9810
	16-19, 3	-0,0960	-3,3288	17,8494
	16-19, 4	12,1832	14,1765	19,3100
	16-19, 5	-3,0147	-21,1599	23,4073
	20-39, 1	-0,1079	-0,2161	2,2731
	20-39, 3	-0,1884	-0,2657	1,8263
	20-39, 4	-0,2707	-0,5478	3,1774
	20-39, 5	0,4051	-0,8363	8,0785

Parametrene i populasjonsmodellen blir dårlig estimert på dette spørsmålet. Som for spørsmål 19 skiller parametre som angår den yngste aldersgruppen seg særlig negativt ut, men her får vi dårlige estimater også for alle de andre parametrene.

Modell for partielt frafall (R_2)

		MLE reelt datasett	gj.snitt (1000)	SD
ψ_0		112,3787	38,5560	16,9854
ψ_1	**sivilstand**			
	ikke-alene	0,2792	0,2455	0,3712
ψ_2	**kjønn**			
	kvinner	-0,0581	0,0167	0,4473
ψ_3	**boregion**			
	1	-0,6503	-0,5101	0,5621
	3	-0,4842	-0,4921	0,4457
	4	0,3565	0,7548	2,7900
	5	-0,9027	-0,8356	0,5433
ψ_4	**y-verdi**			
	ja	-66,3090	-11,9344	17,0970
ψ_5	**R_1-verdi**			
	1	-42,9392	-22,6500	5,2986

Målgruppen for spørsmål 23a er den samme som for spørsmål 19, så vi får de samme problemene med identifisering av målgruppen som vi gjorde på spørsmål 19. Estimeringen går likevel endel bedre på spørsmål 23a, selv om estimatene for ψ_0, ψ_4 og ψ_5 er svært dårlige. Det er 42 personer i målgruppen som har partielt frafall på spørsmål 23a.

Undersysselsetting, modell 1

Populasjonsmodell

		MLE reelt datasett	gj.snitt (1000)	SD
β_0		-1,7188	-1,7391	0,3118
β_1	**aldersgruppe**			
	16-19	-0,6256	-0,7104	1,1434
	20-39	0,8981	0,9127	0,2918
β_2	**kjønn**			
	kvinner	-0,2819	-0,2668	0,2899
β_3	**boregion**			
	1	0,5914	0,5959	0,3581
	3	0,0531	0,0378	0,3545
	4	0,1194	0,0678	0,4203
	5	0,2103	0,1562	0,5071
β_4	**kjønn*alder**			
	kvinne, 16-19 år	0,6827	0,6817	0,4489
	kvinne, 20-39 år	-0,2111	-0,2219	0,2704
β_5	**kjønn*region**			
	kvinne, 1	-0,9108	-0,9218	0,3362
	kvinne, 3	-0,2396	-0,2220	0,3294
	kvinne, 4	-0,0748	-0,0359	0,4043
	kvinne, 5	-0,4704	-0,4418	0,4654
β_6	**alder*boregion**			
	16-19, 1	0,4887	0,5292	1,1708
	16-19, 3	0,3310	0,3908	1,1715
	16-19, 4	-0,2034	-1,6435	6,5136
	16-19, 5	0,4958	-0,2267	4,8076
	20-39, 1	-0,2757	-0,2799	0,3171
	20-39, 3	-0,1451	-0,1499	0,2738
	20-39, 4	0,0624	0,0780	0,2984
	20-39, 5	-0,1028	-0,0858	0,4230

Bortsett fra parametre som angår den yngste aldersgruppen, blir parametrene i populasjonsmodellen godt estimert.

Modell for partielt frafall (R_2)

		MLE reelt datasett	gj.snitt (1000)	SD
ψ_0		0,8302	0,8828	0,3940
ψ_1	**sivilstand**			
	ikke-alene	0,5876	0,5925	0,1272
ψ_2	**kjønn**			
	kvinner	0,6800	0,6820	0,1564
ψ_3	**boregion**			
	1	0,4408	0,4450	0,1940
	3	0,1746	0,1703	0,1573
	4	0,1216	0,1307	0,1885
	5	-0,1153	-0,1180	0,2092
ψ_4	**y-verdi**			
	ja	-0,4816	0,7186	5,5352
ψ_5	**R_1-verdi**			
	1	0,6412	0,6210	0,2548

Parametrene i frafallsmodellen blir svært godt estimert, unntatt ψ_4.

Undersysselsetting, modell 2

Populasjonsmodell

		MLE reelt datasett	gj.snitt (1000)	SD
β_{u0}		-1,7217	-1,7455	0,3074
β_{u1}	**aldersgruppe**			
	16-19	-0,6112	-0,6636	0,5245
	20-39	0,9039	0,9063	0,2886
β_{u2}	**kjønn**			
	kvinner	-0,2798	-0,2488	0,3032
β_{u3}	**boregion**			
	1	0,5697	0,5754	0,3788
	3	0,0398	0,0187	0,3641
	4	0,1030	0,0719	0,4218
	5	0,2040	0,1697	0,5272
β_{u4}	**kjønn*alder**			
	kvinne, 16-19 år	0,6677	0,6485	0,4391
	kvinne, 20-39 år	-0,2235	-0,2413	0,2747
β_{u5}	**kjønn*region**			
	kvinne, 1	-0,8889	-0,9234	0,3544
	kvinne, 3	-0,2300	-0,2235	0,3527
	kvinne, 4	-0,0587	-0,0470	0,4223
	kvinne, 5	-0,4752	-0,4781	0,4936
β_{u6}	**alder*boregion**			
	16-19, 1	0,4964	0,5048	0,5839
	16-19, 3	0,3365	0,3807	0,5493
	16-19, 4	-0,1984	-2,0206	7,6804
	16-19, 5	0,5051	-0,1318	4,6366
	20-39, 1	-0,2732	-0,2492	0,3245
	20-39, 3	-0,1331	-0,1141	0,2841
	20-39, 4	0,0722	0,0906	0,3139
	20-39, 5	-0,0942	-0,0666	0,4342

Med få unntak blir parametrene bra estimert.

Modell for R_u^2

		MLE reelt datasett	gj.snitt (1000)	SD
ψ_{u0}		0,8213	0,8757	0,3455
ψ_{u1}	**sivilstand**			
	ikke-alene	0,5884	0,5954	0,1292
ψ_{u2}	**kjønn**			
	kvinner	0,6818	0,6783	0,1505
ψ_{u3}	**boregion**			
	1	0,4401	0,4496	0,1962
	3	0,1754	0,1673	0,1593
	4	0,1210	0,1244	0,1933
	5	-0,1149	-0,1121	0,2061
ψ_{u4}	**y-verdi**			
	ja	-0,4526	0,0929	2,9987
ψ_{u5}	**R_1-verdi**			
	1	0,6416	0,6240	0,2507

Parametrene estimeres stort sett godt, bortsett fra ψ_{u4}.

I modellen for R_{18}^2 har vi med konstantledd og koeffisisenten foran y_u. I modellen for R_{19}^2 har vi bare tatt med konstantledd.

Modell for R_{18}^2

		MLE reelt datasett	gj.snitt (1000)	SD
$\psi_{18,0}$		-2,9337	-5,5502	7,6470
$\psi_{18,4}$		-1,0024	-7,0400	20,2683

Modell for R_{19}^2

		MLE reelt datasett	gj.snitt (1000)	SD
$\psi_{19,0}$		0,5878	0,8141	2,3586

Parametrene i disse to modellene blir dårlig estimert, særlig $\psi_{18,4}$, koeffisienten foran y_u.

Undersysselsetting, modell 3

Modell for Y_{18}

		MLE reelt datasett	gj.snitt (1000)	SD
β_0		-0,8328	-0,8667	0,2311
β_1	**aldersgruppe**			
	16-19	0,6832	0,6963	0,3073
	20-39	0,6617	0,6743	0,2303
β_2	**kjønn**			
	kvinner	-0,1701	-0,1485	0,2143
β_3	**boregion**			
	1	0,1728	0,1849	0,3088
	3	0,2424	0,2460	0,2721
	4	0,4353	0,4494	0,3189
	5	0,1890	0,1695	0,3825
β_4	**kjønn*alder**			
	kvinne, 16-19 år	0,0368	0,0147	0,2697
	kvinne, 20-39 år	-0,1342	-0,1467	0,2188
β_5	**kjønn*region**			
	kvinne, 1	-0,7647	-0,7552	0,2901
	kvinne, 3	-0,3885	-0,3915	0,2624
	kvinne, 4	-0,2790	-0,2874	0,3197
	kvinne, 5	-0,3847	-0,3866	0,3578
β_6	**alder*boregion**			
	16-19, 1	-0,4099	-0,4196	0,4113
	16-19, 3	-0,3392	-0,3223	0,3408
	16-19, 4	-0,4755	-0,4843	0,4621
	16-19, 5	0,1603	0,1709	0,5320
	20-39, 1	0,1292	0,1319	0,2482
	20-39, 3	-0,1688	-0,1703	0,2083
	20-39, 4	0,1097	0,1157	0,2344
	20-39, 5	0,4403	0,4698	0,2992

Modell for Y_{19}

		MLE reelt datasett	gj.snitt (1000)	SD
β_0		0,6227	0,6813	0,4399
β_1	**aldersgruppe**			
	16-19	-2,1163	-2,2547	0,6874
	20-39	0,5362	0,5256	0,4681
β_2	**kjønn**			
	kvinner	-0,5185	-0,5802	0,4567
β_3	**boregion**			
	1	1,5605	1,6611	1,2543
	3	0,2268	0,2375	0,5528
	4	-0,5535	-0,6208	0,5842
	5	0,3879	0,3658	0,7793
β_4	**kjønn*alder**			
	kvinne, 16-19 år	1,4967	1,6031	0,6329
	kvinne, 20-39 år	0,1927	0,2066	0,4467
β_5	**kjønn*region**			
	kvinne, 1	-0,9291	-0,9765	0,6704
	kvinne, 3	-0,2904	-0,2973	0,5476
	kvinne, 4	0,2371	0,2967	0,5797
	kvinne, 5	-0,1885	-0,1377	0,7004
β_6	**alder*boregion**			
	16-19, 1	0,7609	0,8942	2,1719
	16-19, 3	0,6982	0,7126	0,7223
	16-19, 4	0,2821	-0,9818	6,5103
	16-19, 5	0,3543	-0,0905	4,5440
	20-39, 1	-1,1586	-1,1963	1,1389
	20-39, 3	-0,2386	-0,2339	0,3880
	20-39, 4	-0,0680	-0,0540	0,4207
	20-39, 5	-1,3202	-1,3642	0,5870

Modell for Y_{23}

		MLE reelt datasett	gj.snitt (1000)	SD
β_0		4,0847	6,4558	8,0046
β_1	**aldersgruppe**			
	16-19	0,7239	10,8696	13,0915
	20-39	-0,2656	-0,2581	0,3900
β_2	**kjønn**			
	kvinner	-1,2795	-3,5627	8,0071
β_3	**boregion**			
	1	-0,5869	-0,5787	1,0144
	3	-0,1773	-0,2136	0,5275
	4	0,2823	0,6752	3,2399
	5	-0,5929	-0,0326	3,9745

I denne modellen er kryssleddene utelatt.

I modellene for Y_{18} og Y_{19} er estimatene stort sett bra, mens de er vesentlig dårligere i modellen for Y_{23}.

Modell for R^2_{18}

		MLE reelt datasett	gj.snitt (1000)	SD
ψ_0		2,6312	9,9463	12,5817
ψ_1	**sivilstand**			
	ikke-alene	0,4793	0,4900	0,1464
ψ_2	**kjønn**			
	kvinner	0,5304	0,5425	0,1638
ψ_3	**boregion**			
	1	0,3363	0,3580	0,2209
	3	0,1208	0,1294	0,1790
	4	0,2719	0,2744	0,2077
	5	0,0383	0,0397	0,2335
ψ_4	**y-verdi**			
	ja	-3,0136	-10,2949	12,6888
ψ_5	**R_1-verdi**			
	1	0,8546	0,8641	0,2905

Også her får vi problemer med estimeringen av ψ_4.

I modell 3 har vi bare tatt med konstantleddet i modellene for R^2_{19} og R^2_{23}:

Modell for R^2_{19}

		MLE reelt datasett	gj.snitt (1000)	SD
ψ_0		5,2417	5,5178	2,0328

Modell for R^2_{23}

		MLE reelt datasett	gj.snitt (1000)	SD
ψ_0		4,1017	4,1681	1,0018

Spørsmål 24

Populasjonsmodell

		MLE reelt datasett	gj.snitt (1000)	SD
β_0		1,4658	1,5467	0,5185
β_1	**aldersgruppe**			
	16-19	-3,7046	-4,5682	4,5385
	20-39	0,9074	0,9571	0,5165
β_2	**kjønn**			
	kvinner	-1,5135	-1,5958	0,5239
β_3	**boregion**			
	1	-0,5342	-0,5705	0,6301
	3	-0,4838	-0,5148	0,6492
	4	-0,5281	-0,5618	0,6887
	5	0,2907	2,0676	7,3641
β_4	**kjønn*alder**			
	kvinne, 16-19 år	1,8204	1,9810	1,2124
	kvinne, 20-39 år	-0,7589	-0,8035	0,4846
β_5	**kjønn*region**			
	kvinne, 1	0,6982	0,7442	0,6230
	kvinne, 3	0,2309	0,2675	0,6423
	kvinne, 4	0,2605	0,2795	0,6897
	kvinne, 5	-0,6378	-2,4228	7,3648
β_6	**alder*boregion**			
	16-19, 1	0,9118	1,1293	5,7082
	16-19, 3	0,3834	0,4601	5,8974
	16-19, 4	1,5083	1,3058	6,7129
	16-19, 5	1,2152	-0,8093	10,0328
	20-39, 1	-0,1928	-0,2017	0,4574
	20-39, 3	0,0484	0,0390	0,3764
	20-39, 4	0,2498	0,2711	0,3867
	20-39, 5	0,6310	0,6595	0,5494

Også her blir det visse problemer med parametrene for den yngste aldersgruppen, men de største vanskelighetene har vi med parametre som angår region 5, dvs. Nordland, Troms og Finnmark. Det er relativt få (94) personer som bor i denne regionen i målgruppen for spørsmål 24. I tillegg har vi fra analysene på svarutvalget på spørsmål 24 (avsnitt 2.2.9), at hverken region eller kryssledd med region ble signifikante, og at prøvemodellen ikke passet så godt.

Modell for partielt frafall (R_2)

		MLE reelt datasett	gj.snitt (1000)	SD
ψ_0		2,0729	7,0812	11,4541
ψ_1	**sivilstand**			
	ikke-alene	0,5877	0,5713	0,4864
ψ_2	**kjønn**			
	kvinner	0,7915	0,8244	0,5770
ψ_3	**boregion**			
	1	0,6007	2,9888	7,9111
	3	-0,2110	-0,2457	0,5956
	4	-0,1686	0,2202	3,3014
	5	-0,4944	-0,2952	2,5480
ψ_4	**y-verdi**			
	ja	1,0880	3,6558	9,1470
ψ_5	**R_1-verdi**			
	1	0,4035	-4,1828	10,6359

Målgruppen for spørsmål 24 er den samme som for spørsmål 19 og 23a, og også her blir frafallsparametrene nokså dårlig estimert, men ikke så dårlig som på spørsmål 19. Det er 22 personer i målgruppen som har partielt frafall på spørsmål 24. Som på spørsmål 23a er det særlig ψ_0, ψ_4 og ψ_5 som peker seg ut i negativ retning.

Spørsmål 25

Populasjonsmodell

		MLE reelt datasett	gj.snitt (1000)	SD
β_0		1,2158	1,2193	0,0528
β_1	**aldersgruppe**			
	16-19	-2,5772	-2,6083	0,3161
	67-74	-1,5160	-1,5195	0,3516
β_2	**kjønn**			
	kvinner	-1,7020	-1,7056	0,0717
β_3	**boregion**			
	1	0,3143	0,3104	0,0823
	3	0,0817	0,0773	0,0782
	4	0,0910	0,0866	0,0920
	5	-0,1909	-0,1979	0,0998
β_4	**kjønn*alder**			
	kvinne, 16-19 år	1,1333	1,1263	0,3403
	kvinne, 67-74 år	0,0485	-0,0438	0,4476
β_5	**kjønn*region**			
	kvinne, 1	0,1962	0,2018	0,1106
	kvinne, 3	-0,1277	-0,1235	0,1053
	kvinne, 4	-0,2060	-0,1985	0,1297
	kvinne, 5	0,4498	0,4615	0,1369
β_6	**alder*boregion**			
	16-19, 1	-1,2137	-1,3049	1,1092
	16-19, 3	-0,5405	-0,5451	0,4413
	16-19, 4	-0,4361	-0,6042	1,7014
	16-19, 5	-0,1828	-0,4212	2,2506
	67-74, 1	-1,1635	-0,1480	0,4618
	67-74, 3	-0,4505	-0,4676	0,4879
	67-74, 4	-1,0538	-1,3649	2,4965
	67-74, 5	0,1390	-0,7065	4,9439

Parametrene i populasjonsmodellen blir stort sett godt estimert. De største skjevhetene og variansene finner vi under alder*boregion, i de gruppene som har med region 4 og 5 å gjøre. Dette er også det kryssleddet som ble minst signifikant i analysen på svarutvalget.

Modell for partielt frafall (R_2)

		MLE reelt datasett	gj.snitt (1000)	SD
ψ_0		2,8619	3,0443	1,5546
ψ_1	**sivilstand**			
	ikke-alene	0,5578	0,5511	0,1595
ψ_2	**kjønn**			
	kvinner	0,9207	0,9273	0,2952
ψ_3	**boregion**			
	1	0,1020	0,1160	0,2942
	3	-0,6000	-0,6161	0,2215
	4	-0,5272	-0,5400	0,2581
	5	-1,0195	-1,0272	0,2482
ψ_4	**y-verdi**			
	ja	0,0042	-0,1158	1,5968
ψ_5	**R_1-verdi**			
	1	1,2149	1,1961	0,2284

Bortsett fra ψ_4 blir frafallsparametrene svært godt estimert på dette spørsmålet.

Spørsmål 62

Populasjonsmodell

		MLE reelt datasett	gj.snitt (1000)	SD
β_0		3,2722	3,4996	0,8084
β_1	**aldersgruppe**			
	16-19	-4,6633	-5,7470	4,9715
	20-39	-0,3463	-0,4340	0,8035
β_2	**kjønn**			
	kvinner	-3,0475	-3,2623	0,7730
β_3	**boregion**			
	1	-1,7291	-1,8590	0,8316
	3	-0,9161	-0,9904	0,8671
	4	1,6724	9,1950	14,4958
	5	34,1486	38,5586	15,5253
β_4	**kjønn*alder**			
	kvinne, 16-19 år	3,0026	3,2288	0,9352
	kvinne, 20-39 år	0,8431	0,9410	0,6931
β_5	**kjønn*region**			
	kvinne, 1	1,1675	1,2553	0,7323
	kvinne, 3	0,1027	0,1432	0,7232
	kvinne, 4	-1,7042	-9,2010	14,4426
	kvinne, 5	-33,0922	-31,4432	11,8013
β_6	**alder*boregion**			
	16-19, 1	1,6846	2,4274	5,1715
	16-19, 3	1,7025	2,5697	5,0018
	16-19, 4	0,7514	-1,1831	10,5562
	16-19, 5	-32,9729	-37,2626	14,3094
	20-39, 1	0,7584	0,8233	0,8232
	20-39, 3	0,5127	0,5345	0,7594
	20-39, 4	0,0033	-0,0132	1,2651
	20-39, 5	1,2158	4,2411	15,9236

Vi får store standardavvik på parametre som har med region 4 og 5 å gjøre. Også i den yngste aldersgruppen er estimeringen nokså dårlig. Når vi fordeler på både alder og region får vi enkelte veldig små grupper. I aldersgruppen 16-19 år er det for eksempel bare 9 personer i region 4, og det samme i region 5. Kryssleddet mellom alder og region var ikke signifikant i analysen på svarutvalget på spørsmål 62.

Modell for partielt frafall (R_2)

		MLE reelt datasett	gj.snitt (1000)	SD
ψ_0		1,0642	1,3190	1,7682
ψ_1	**sivilstand**			
	ikke-alene	-0,1583	-0,1670	0,1965
ψ_2	**kjønn**			
	kvinner	0,5361	0,5616	0,2823
ψ_3	**boregion**			
	1	0,8814	0,9321	0,3523
	3	0,3731	0,3716	0,2495
	4	0,8645	0,8846	0,3357
	5	-0,3627	-0,3713	0,2859
ψ_4	**y-verdi**			
	ja	0,2164	0,0568	1,7291
ψ_5	**R_1-verdi**			
	1	-0,0753	-0,1522	0,4792

Også her blir frafallsparametrene akseptabelt estimert, med unntak av ψ_4.

Modell for svar på første henvendelse (R_1)

		MLE reelt datasett	gj.snitt (1000)	SD
φ_0		2,9865	2,9882	0,1200
φ_1	**sivilstand**			
	ikke-alene	0,2377	0,2328	0,1460
φ_2	**kjønn**			
	kvinner	0,1384	0,1402	0,1494
φ_3	**boregion**			
	1	-0,7350	-0,7316	0,1600
	3	-0,1641	-0,1535	0,1675
	4	-0,2636	-0,2630	0,1903
	5	-0,3945	-0,3848	0,2061
φ_4	**kjønn*sivilstand**			
	kvinne, ikke-alene	0,1407	0,1451	0,1345
φ_5	**kjønn*region**			
	kvinne, 1	0,4610	0,4696	0,1875
	kvinne, 3	0,2434	0,2406	0,1929
	kvinne, 4	0,1348	0,1423	0,2166
	kvinne, 5	0,4165	0,4267	0,2454
φ_6	**sivilstand*boregion**			
	ikke alene, 1	-0,0257	-0,0330	0,1926
	ikke alene, 3	-0,0607	-0,0661	0,1932
	ikke alene, 4	-0,0995	-0,0947	0,2108
	ikke alene, 5	-0,2498	-0,2558	0,2389

Siden estimeringen av φ-ene skjer på grunnlag av hele datasettet, og ikke bare på en spesiell målgruppe, hadde vi ventet oss forventningsrette og presise estimater. Parametrene i modellen for R_1 estimeres da også godt. Det er liten forskjell mellom gjennomsnittene fra simuleringene og MLE fra det reelle datasettet, og det er ingen utpreget store standardavvik. Vi hadde likevel håpet på noe lavere standardavvik med et så stort datagrunnlag.

Appendiks D. Program.

Programmene i forbindelse med dette prosjektet er dels skrevet som SAS-programmer, dels i programmeringsspråket Fortran. Fortran er nødvendig ved bruk av de numeriske NAG-rutiner, som har vært benyttet til maksimering av loglikelihoodfunksjoner. Følgende NAG-rutiner er benyttet:

E04KCF: Modifisert Newton-algoritme til å finne minimum av en funksjon $F(x_1, x_2, ..., x_n)$, når det er satt faste øvre og nedre grenser på de uavhengige variable $x_1, x_2, ..., x_n$, og når funksjonens første-deriverte er kjent. Algoritmen er laget for kontinuerlige funksjoner, med kontinuerlige første- og annenderiverte, (men vil vanligvis virke også når de deriverte har enkelte diskontinuiteter).

G05EDF: Lager en referansevektor R til en binomisk fordeling av antallet suksesser ved N forsøk, når hvert forsøk har suksessannsynlighet P.

G05EYF: Inneholder et tilfeldig heltall fra en diskret fordeling, som er definert ved en referansevektor R.

Til å finne loglikelihoodfunksjonenes første deriverte, har vi brukt MAPLE, som er et symbolsk matematikkprogram.

Programmene er kjørt på UNIX.

D.1. Spørsmål 18

Følgende Fortranprogrammer er brukt i forbindelse med simulering for spørsmål 18:

funct2.f: Egentlig en subrutine, hvor loglikelihoodfunksjonen til spørsmål 18 spesifiseres. funct2.f kalles opp fra maksimeringsprogrammene hoved18.f og max18.f.

hoved18.f: Maksimerer loglikelihoodfunksjonen til spørsmål 18, og gir MLE for hver parameter. Programmet henter inn en rekke datafiler som inneholder ulike konstanter. F.eks. fila yja18.dat, som inneholder antall IO som svarte ved første henvendelse, er deltidssysselsatte og ønsker lengre arbeidstid, sortert etter aldersgruppe, kjønn og boregion. (Disse filene er produsert av sas-program.)

r1r2y18.f: Simulerer fordelinger.

- Leser inn simuleringsverdier for beta, fi og psi, og beregner $P(Y = 1 \mid \mathbf{x})$, $P(R^1 = 1 \mid \mathbf{z})$ og $P(R^2 = 1 \mid \mathbf{z}, y, r^1)$.

- For hver av de 1000 simuleringene gjøres følgende:

  - Leser fra fil antall IO som har levert skjema, i hvert stratum, m( ). Stratum er her definert ved sivilstand, alder, kjønn, boregion, om deltidssysselsatt eller ikke, om arbeidsledig eller ikke og om ett arbeidsforhold eller flere.

  - Simulerer antall IO med $R^1 = 1$ i hvert av strataene ovenfor:
  antallr1_1( ) antas binomisk (m( ), $P(R^1 = 1 \mid \mathbf{z})$)

- De deltidssysselsatte, i hvert av strataene over, lagres på en egen fil. På samme måte lager vi en fil for arbeidsledige og en for IO med ett arbeidsforhold.

- Alle deltidssysselsatte skal svare på spørsmål 18, og fila med deltidssysselsatte blir derfor brukt til å simulere antall IO med $Y = 1$, i hvert stratum av alder, kjønn, sivilstand, boregion og R^1. Dette antallet vil være binomisk fordelt, med $P(Y = 1 | \mathbf{x})$ som suksessannsynlighet.

- Simulerer antall IO med $R^2 = 1$, i hvert stratum av alder, kjønn, sivilstand, boregion, R^1 og Y. Antallet vil være binomisk fordelt, med suksessannsynlighet $P\left(R^2 = 1 | \mathbf{z}, y, r^1\right)$.

max18.f: Maksimerer loglikelihoodfunksjonen til hvert av de simulerte datasett fra programmet over. Som startverdier for maksimeringsrutinen har vi brukt verdien 0,0 for alle parametrene. Henter inn fila funct2.f.
Til slutt regner programmet ut gjennomsnitt og standardavvik av MLE fra de 1000 simuleringer, for hver parameter.

sdpverdi.f: Finner gjennomsnitt og standardavvik for $P\left(Y = 1|\mathbf{x}\right), P\left(R^1 = 1|\mathbf{z}\right)$ og $P\left(R^2 = 1|\mathbf{z}, y, r^1\right)$ fra de 1000 simuleringer.

LR18.f: Beregner deviansen (umodifisert test) for spørsmål 18.

LR18mod.f: Bergner deviansen (modifisert test) for spørsmål 18. (*)

chisq18.f: Beregner umodifisert kjikvadrattest for spørsmål 18.

chisq18mod.f: Beregner modifisert kjikvadrattest for spørsmål 18.

simLR18.f: Beregner umodifisert devians for de simulerte datasettene.

simLR18mod.f: Beregner modifisert devians for de simulerte datasettene. (*)

simchisq18.f: Beregner umodifisert kjikvadrattest for de simulerte datasettene.

simchisq18mod.f: Beregner modifisert kjikvadrattest for de simulerte datasettene.

D.2. Spørsmål 19, 23a, 24, 25, 62 og undersysselsetting

For spørsmålene 19, 23a ,24 og undersysselsetting lages det tilsvarende programmer som for spørsmål 18, bortsett fra programmene for modelltesting, som for disse spørsmålene reduseres til programmer for modifisert devians, (det vil si de to programmene som er markert med stjerne). Simuleringsprogrammene r1r2y19.f, r1r2y23.f og r1r2y24.f tar utgangspunkt i fila over simulerte datasett av deltidssysselsatte som svarer ja på spørsmål 18.

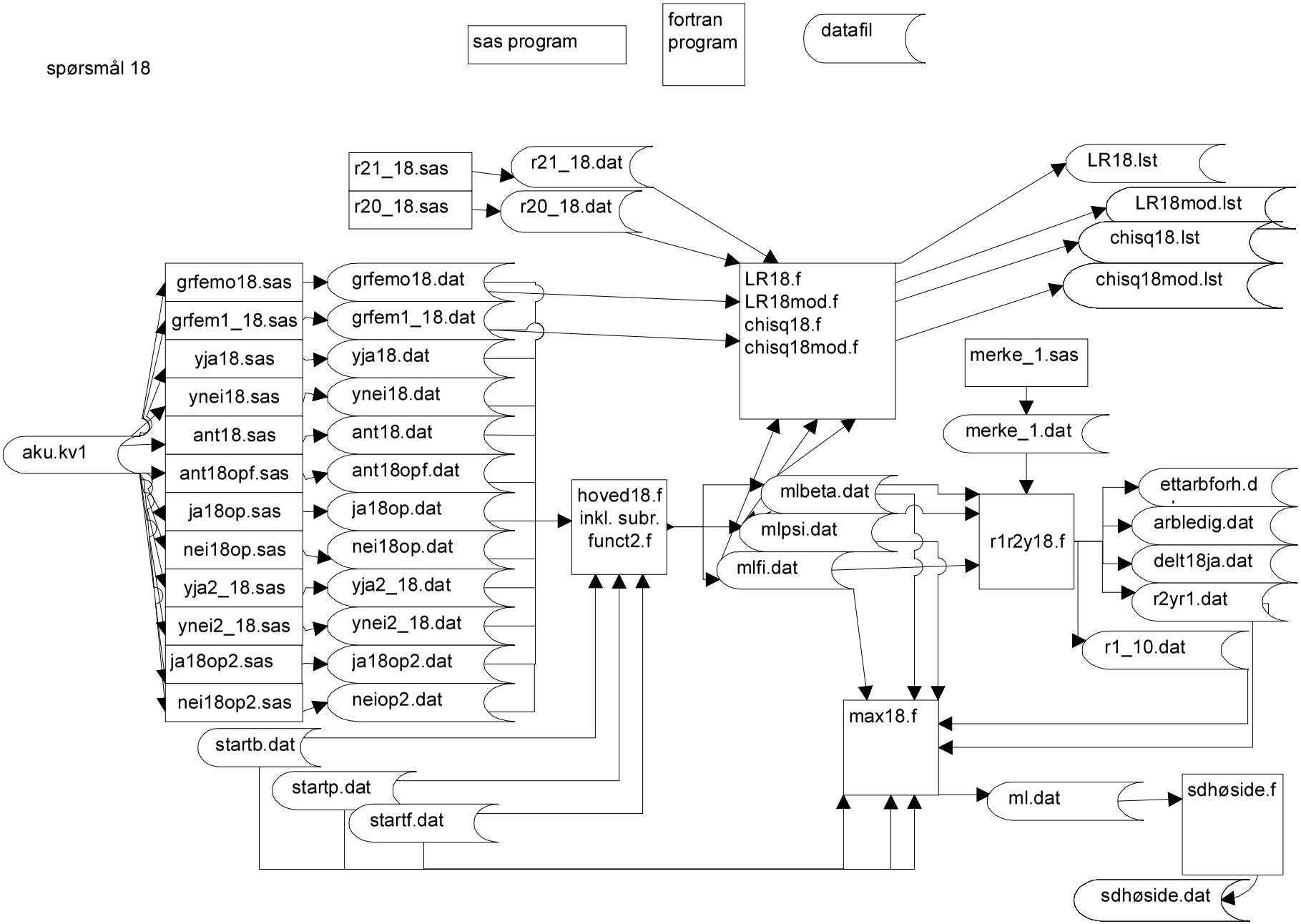
For spørsmål 25 tar simuleringsprogrammet (r1r2y25.f) utgangspunkt i fila med simulert datasett av personer med ett arbeidsforhold.

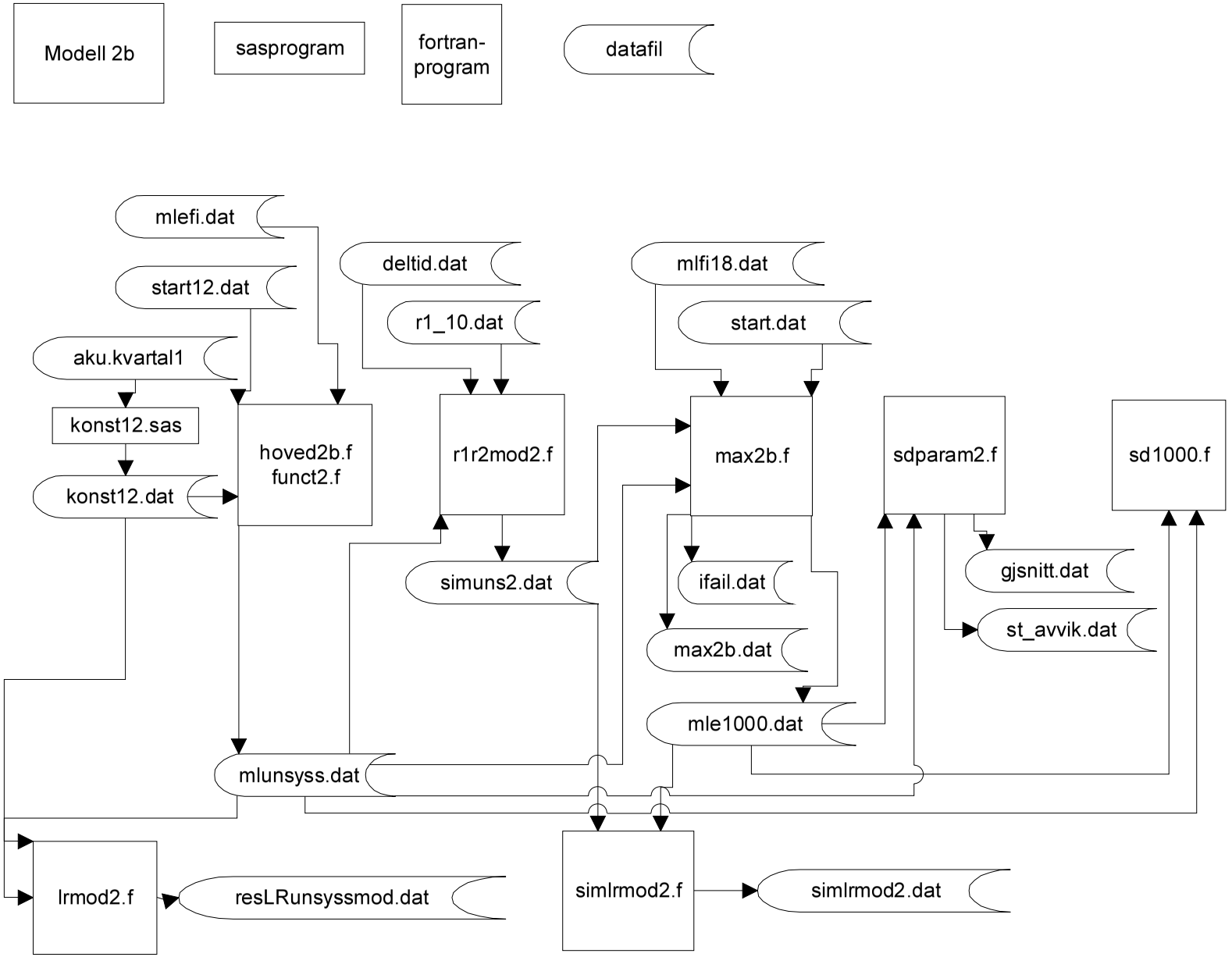
For spørsmål 62 tar simuleringsprogrammet (r1r2y62.f) utgangspunkt i fila med simulert datasett av arbeidsledige personer.

For undersysselsetting tar simuleringsprogrammene i hver av de tre modellene utgangspunkt i fila over simulerte datasett av deltidsysselsatte.

Til slutt tar vi med dataflytdiagrammer for spørsmål 18, og for undersysselsetting, modell 2. En del forenklinger er gjort fra spørsmål 18, og fram til undersysselsetting, men fortranprogrammene svarer til hverandre.

spørsmål 18


sas program
fortran program
datafil
r21_18.sas
r20_18.sas
r21_18.dat
r20_18.dat
LR18.lst
LR18mod.lst
chisq18.lst
chisq18mod.lst
LR18.f
LR18mod.f
chisq18.f
chisq18mod.f
grfemo18.sas
grfem1_18.sas
yja18.sas
ynei18.sas
ant18.sas
ant18opf.sas
ja18op.sas
nei18op.sas
yja2_18.sas
ynei2_18.sas
ja18op2.sas
nei18op2.sas
grfemo18.dat
grfem1_18.dat
yja18.dat
ynei18.dat
ant18.dat
ant18opf.dat
ja18op.dat
nei18op.dat
yja2_18.dat
ynei2_18.dat
ja18op2.dat
neiop2.dat
aku.kv1
merke_1.sas
merke_1.dat
hoved18.f
inkl. subr.
funct2.f
mlbeta.dat
mlpsi.dat
mlfi.dat
r1r2y18.f
ettarbforh.d
arbledig.dat
delt18ja.dat
r2yr1.dat
r1_10.dat
max18.f
startb.dat
startp.dat
startf.dat
ml.dat
sdhøside.f
sdhøside.dat


Modell 2b
sasprogram
fortran-
program
datafil
mlefi.dat
start12.dat
aku.kvartal1
konst12.sas
konst12.dat
hoved2b.f
funct2.f
deltid.dat
r1_10.dat
r1r2mod2.f
simuns2.dat
mlfi18.dat
start.dat
max2b.f
ifail.dat
max2b.dat
mle1000.dat
sdparam2.f
gjsnitt.dat
st_avvik.dat
sd1000.f
mlunsyss.dat
lrmod2.f
resLRunsyssmod.dat
simlrmod2.f
simlrmod2.dat

De sist utgitte publikasjonene i serien Notater

1999/82 Ø. Kleven, E. Dalheim og D. Roll-Hansen: Innvandreres utdanning: - en pilotundersøkelse. 61s.

1999/83 E. Fidjestøl og I. Håland: Yrkeskatalog: Pr. desember 1999. 136s.

1999/84 T. Solberg: Virkning av revisjon på Avlingsstatistikk for jordbruksvekster i 1998. 24s.

1999/85 R. Choudhury, T. Eika og L. Haakonsen: KVARTS i praksis II: Systemer og rutiner i den daglige driften. 66s.

1999/86 G. Frøiland: Økonometrisk modellering av husholdningenes konsum i Norge: Demografi og formueseffekter. 55s.

1999/87 Y. Li: Beregning av elementæraggregater i konsumprisindeksen ved hjelp av generalisert gjennomsnitt. 41s.

1999/88 L. Rogstad og S.T. Vikan: Kobling av adresseregistrene i DSF og GAB 1999: Dokumentasjon av samsvar og avvik. 31s.

1999/89 E. Dalheim, J-A. S. Lie og D. Roll-Hansen: En skjemabasert komplettering av registeret over befolkningens høyeste utdanning - forprosjekt med fokus på innvandrere. 60s.

1999/90 K-A. Hovland og Å. Nossum: Flyreiser i konsumprisindeksen. 39s.

2000/1 E. Rønning: Utenlandske statsborgere og kommunestyrevalget 1999: Dokumentasjonsrapport. 34s.

2000/2 M. Bråthen: Personer registrert som yrkeshemmet i SOFA-søkerregisteret. 25s.

2000/3 A.K. Johnsen og Ø. Hokstad: FoB2001: Kvalitativ testing av boligskjema - prøveundersøkelse 1999: Dokumentasjonsnotat. 32s.

2000/4 C. Hendriks, Ø. Hokstad og R. Sønsterudbråten: FoB2001: Boligtelling - prøveundersøkelse 1999: Dokumentasjonsnotat. 60s.

2000/5 K. Bjønnes, G. Dahl og B.R. Joneid: FD - Trygd: Dokumentasjonsrapport: Økonomisk sosialhjelp 1992-1997. 31s.

2000/6 B.R. Joneid og J. Lajord: FD - Trygd: Dokumentasjonsrapport: Demografi 1992-1997. 117s.

2000/7 J. Heldal: Kalibrering av AKU: Dokumentasjon av metode og program. 28s.

2000/8 H. Hågård og L. Rogstad: FoB2001: Adresser i folkeregisteret og GAB: Rapport fra en arbeidsgruppe for adressesamordning og utredning av elektronisk datautveksling mellom DSF og GAB. 51s.

2000/9 B. Sundby: Rutiner for produksjon av statistikk over pleie- og omsorgstjenestene i kommunene 1997. 84s.

2000/10 E. Aas: På leting etter målefeil - en studie av pleie- og omsorgssektoren. 31s.

2000/11 I. Øyangen: Lokalvalgsundersøkelsen 1999: Dokumentasjonsrapport. 36s.

2000/12 E. Engelien: Arealbruksstatistikk for tettsteder: Dokumentasjon av arbeid med metodeutvikling 1999. 50s.

2000/13 F. Gundersen og A.E. Hustad: Statistikk over anmeldte lovbrudd og registrerte ofre: Dokumentasjon. 51s.

2000/14 T. Martinsen: Prosjekt over industriens energibruk. 58s.

2000/15 R.Ragnarsøn: Harmonisert produksjonsstatistikk for industrien. 39s.

2000/16 B. Halvorsen og R. Nesbakken: Fordelingseffekter av økt elektrisitetsavgift for husholdningene. 74s.

2000/17 J. Fosen og L. Solheim: Avledede variable i registerstatistikk: To metoder for klassifikasjon av sysselsettingsstatus. 43s.

2000/18 K. Myklebust: Rapport fra seminar om stedfesting av bedrifter. Oslo 1. desember 1999. 73s.