Statistics Norway
Statistisk sentralbyrå

Ole Klungsøyr

Markovkjede Monte Carlo i varianskomponentmodell for sysselsettingsdata

125 1876•2001 år

Seksjon for statistiske metoder og standarder
Emnegruppe: 06.90

Innhold

1. Innledning

Dette notatet er et resultat av en prosjektoppgave i et kurs på Universitetet i Oslo (ST397), høsten 2000, med professor Arnoldo Frigessi som foreleser. Innenfor både personstatistikk og økonomisk statistikk er det i SSB behov for å publisere tall på såkalte "små områder", som kan f.eks. være geografiske områder, eller forskjellige næringer i industrien. Områdene kan være små i den forstand at når dataene samles inn fra et utvalg, kan utvalget være lite der. Da kan estimering av populasjonstall for de små områdene bli unøyaktig, og forskjellige metoder kan brukes for å kompensere for dette. I litteraturen er EBLUP (Empirical Best Linear Unbiased Predictor) mye omtalt (se [2]) som en god metode. Det er en estimator som for hvert område er en veiet sum av et estimat basert kun på det bestemte området og et estimat basert på de andre områdene. Hvis estimatet for det bestemte området har stor varians, tillegges denne liten vekt og EBLUP estimatoren "låner styrke" fra de andre områdene. Men den har også sine begrensninger som metodene vi skal bruke her ikke har.

Seksjon 370 har i forbindelse med folke- og boligtellingen 2001 tatt initiativ til metodeutvikling for små områder. I dette arbeidet har data fra utvalg og registre blitt koblet sammen og dette er utgangspunktet for denne analysen. Utvalgsdataene er fra Arbeidskraftundersøkelsen (heretter kalt AKU) som er en kvartalsvis spørreundersøkelse i regi av SSB. Den brukes for å kartlegge arbeidsledighet, sysselsettningsgrad, og andel utenfor disse gruppene. Utvalget gir ikke grunnlag for å publisere tall på finere nivå enn fylke uten bruk av spesielle metoder for små områder. I dette notatet studeres metoder for å kunne publisere tall på kommunenivå. Vi ser kun på sysselsettingsandel som variabel i dette notatet, men den samme analysen passer også for ledighet. For sysselsetting og ledighet, viser det seg at EBLUP estimatoren ikke nødvendigvis fungerer så godt (se [2]). Det er skjevhet i fordelingene, spesielt kanskje i ledighetsandelene, så den tilnærmede normaliteten som EBLUP forutsetter er ikke oppfylt, og man får ofte degenererte utvalg, dvs. at for en liten kommune vil AKU-utvalget være veldig lite, og det kan fort skje at alle enten er ledige, eller sysselsatte, selv om man vet at dette ikke stemmer i virkeligheten. Da vil EBPLUP-estimatoren få problemer. Den estimerte variansen til estimatoren basert kun på det bestemte området vil kunne bli 0, med det resultat at man stoler helt på dette estimatet man vet ikke stemmer. I tillegg vil EBLUP-estimatoren uten kovariater (informasjon spesifikk for de forskjellige områdene) være følsom for såkalt "over-shrinkage", som betyr for liten variasjon mellom områdene i forhold til sann variasjon. Dette er imidlertid et generelt problem, at når man skal estimere et sett med parametre (en for hvert område), og bruker estimater basert på betinget forventning, får man mindre variasjon i settet av parameterestimatene enn i de sanne verdiene

Metodene vi skal bruke er Metropolis Hastings algoritme og Gibbs sampling (se [1]) til å simmulere fra en varianskomponentmodell. Vi vil omtale disse metodene som MCMC (Markov Chain Monte Carlo). Ideene er hentet fra Zhang ([2]), men i motsetning til han bruker vi en Bayesiansk modell som utgangspunkt. Vi ser bare på en liten del av det han gjør, og dette er først og fremst tenkt som illustrasjon på bruk av MCMC.

2. Problemformulering og data

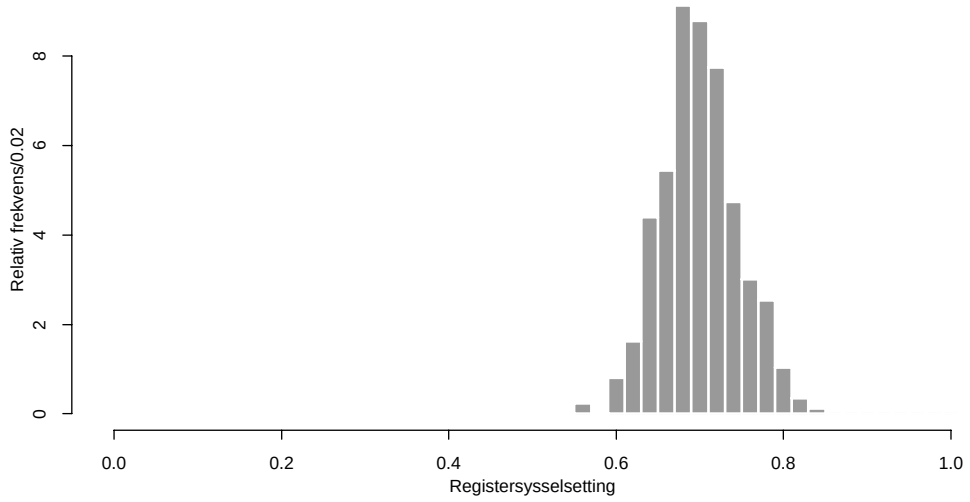
AKU-utvalget består av 24 000 personer i den yrkesaktive befolkningen mellom 16 og 74 år. Frafallet er i underkant av 10 % slik at nettoutvalget er omtrent 22 000. Hver person i utvalget deltar 8 kvartaler. Intervjuene gjennomføres over telefon. Det spørres om sysselsettings-status o.l. og denne informasjonen regner vi med som riktig her, uten målefeil, dvs. at de svarer sant. Vi har sett på 4. kvartal AKU data fra 1997 i denne analysen, og bare sysselsettingsandel (ikke ledighet, eller utenfor arbeidsstyrken).

I tillegg til denne informasjonen, har vi brukt et register, som er en sammenkobling av Arbeidstaker-registeret fra Rikstrygdeverket, Lønns- og trekkoppgaveregisteret fra Skattedirektoratet, og Ligningsregisteret, også fra Skatte-direktoratet. Dette registeret har informasjon om alle personer i

Norge angående deres sysselsettings-status (ca. 3 millioner personer) og har lang produksjonstid, det var ferdig oppdatert først i slutten av 1999, og er først og fremst interessant på årsbasis. Dette registeret er av interesse for mange anvendelser, som f.eks. folke- og boligtellinger. Derfor er det viktig å ha best mulig kvalitet på dette registeret, og vite hvor godt det er.

I tillegg til å ha registersysselsetting har vi også tilgjengelig registerinformasjonen til de som er med i AKU-utvalget. Dataene er kun tilgjengelig på kommunenivå, ikke individnivå. Det er 432 kommuner representert (mangler informasjon om 3 kommuner, som derfor er utelatt fra all analyse), så både i utvalget og i populasjonen er det 432 kommuner.

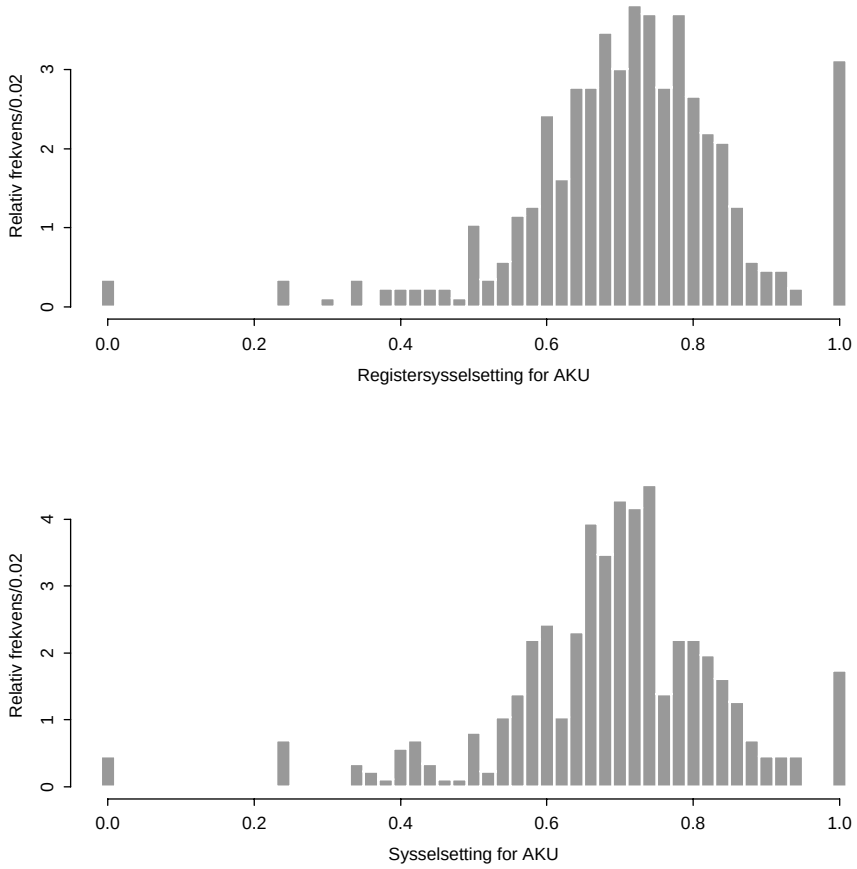
Histogrammet nedenfor viser sysselsettingsandelene i alle kommunene, basert på informasjon fra registeret, heretter kalt registersysselsettingen.



Figur 1: Histogram over registersysselsettingsandelene for alle kommuner

Vi ser at fordelingen er ganske smal og symmetrisk med gjennomsnitt 0,6994.

Neste side presenteres resten av dataene vi har brukt. Øverst ser vi et histogram over registerinformasjonen til de som er med i AKU-utvalget (for alle kommuner) og nedenfor et histogram med selve AKU-utvalgs informasjonen. Begge disse datasettene bruker vi som observasjoner.



Figur 2: Histogram over sysselsettingsandelene for alle kommuner basert på registerinformasjon om AKU-utvalget (øverst) og AKU-utvalget selv (nederst)

Vi ser tydelig større varians i utvalgs-histogrammene enn i registerhistogrammet. Det er også en del andeler som er 1 og 0, dette er først og fremst i tilfeller med et utvalgsantall på 1. Videre viser histogrammene at det er en viss forskjell på registersysselsetting for AKU-utvalget og AKU-utvalgets sysselsetting, men ikke stor. Gjennomsnitt i registersysselsetting for AKU-utvalget er 0,7141, mens gjennomsnittet av AKU sysselsettingen er 0,6897, så gjennomsnittet av AKU sysselsettingen ligger litt nærmere gjennomsnittet i registeret enn registersysselsettingen for AKU-utvalget. Dette er tilfeldig for akkurat våre observasjoner og vi skal også se dette i resultatene.

Det sentrale i estimering for små områder, er å lage gode estimater på områdenivå. Her vil det si å esimere de sanne sysselsettingsandeler (ev. ledighetsandeler) på kommunenivå, altså de andelene man ville fått hvis man spurte alle personene i hver kommune. Dette som heretter vil kalles estimert AKU sysselsetting baseres på utvalgets svar, og ev. på informasjon om hele populasjonen i registeret. Ved å bruke registerinformasjonen om AKU-utvalget som observasjoner i stedet for AKU-utvalget kan man se hvor godt metodene fungerer for da vil disse gi estimater for registersysselsettingen, noe vi også har fasiten til.

3. Modell

Vi lar $Y_i^s, i = 1, \cdots, N$ være totalt antall sysselsatte i kommune nr. i (N=432) ifølge AKU-utvalget (s er kort for "sample"). Tilsvarende for register-informasjonen; X_i^s er registerantall sysselsatte i kommune nr. i for de som er med i AKU-utvalget, og X_i^p er registerantall sysselsatte i hele populasjonen (som vi har tilgjengelig, p for populasjon). Videre er n_i antall personer i AKU-utvalget i kommune nr. i, mens N_i er antallet i populasjonen i kommune nr. i. Hvis vi tenker oss at hver person i kommune nr. i har samme sannsynlighet θ_i for å være sysselsatt, og betrakter sysselsettingen i en kommune som n_i uavhengige Bernoulli forsøk (egentlig er det ikke nødvendigvis uavhengighet) vil totalantallet sysselsatte i kommunen være binomisk fordelt, med antall sysselsatte i utvalget delt på antall i utvalget som naturlig estimator for θ_i. Da kan en enkel varianskomponentmodell for sysselsetting skrives:

(3.1) $$\theta_i = \mu_\theta + v_i,$$

og

(3.2) $$\frac{1}{n_i} Y_i^s = \theta_i + e_i$$

som sier at den sanne sysselsettingsandelen har et globalt gjennomsnitt μ_θ og lokal variasjon rundt denne, og at observasjonene man gjør er riktige bortsett fra en viss samplingsfeil e_i. v_i og e_i er begge støyledd med forventning 0 og varians σ_v^2 og σ_e^2. v_i er de såkalte "random effects" som beskriver mellom-område-variasjon og σ_v^2 er den ene varianskomponenten i modellen. Den andre varians-komponenten er σ_e^2 som beskriver samplingsfeilen, eller den såkalte innen-område-variasjonen. Settes (3.1) inn i (3.2) fås:

(3.3) $$\frac{1}{n_i} Y_i^s = \mu_\theta + v_i + e_i$$

som sier at observasjonene inneholder to typer variasjon, mellom-område-variasjon og innen-område-variasjon.

Vi kan tenke oss akkurat samme modellen hvis vi erstatter Y_i^s med X_i^s og lar $\theta_i = \frac{X_i^p}{N_i}$, som vi har tilgjengelig som data (register-informasjonen). Histogrammet i figur 1 viser nettopp $\frac{X_i^p}{N_i}$, (tilsvarer (3.1)). Sammenlignet med det øverste histogrammet i figur 2, som er $\frac{X_i^s}{n_i}$ og tilsvarer (3.2) ser vi at dette siste histogrammet har mest variasjon, som nettopp (3.3) forteller (for Y_i^s).

I tillegg til den enkle modellen i (3.1), skal vi også se på en modell hvor vi benytter tilleggsinformasjon, altså data på område-nivå, som i vårt tilfelle er registerinformasjonen (se [2]). Da blir (3.1) på formen:

(3.5) $$\mathrm{logit}(\theta_i) = \mu_i + v_i$$

hvor $\log it(\theta_i) = \log(\frac{\theta_i}{1-\theta_i})$ brukes for å garantere at $\theta_i \in (0,1)$ og μ_i er en lineær prediktor, f.eks. $\mu_i = \beta_0 + \beta_1 x_i$ med x_i som kjent områdeinformasjon for populasjonen (også på transformert skala).

I begge våre tilfeller skal vi betrakte en Bayesiansk hierarkisk struktur på modellen. En Bayesiansk modell har apriori-fordelinger på parametrene i modellen, og man er interessert i aposteriori-fordelingen til parametrene, eller den betingede fordelingen for parametrene gitt dataene. Hvis vi tenker oss alle parametrene i en vektor $\boldsymbol{\lambda}$ og alle dataene i en vektor $\mathbf{Y}$ gjelder flg. sammenheng for aposteriorifordelingen:

$$p(\boldsymbol{\lambda} \mid \mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\lambda})p(\boldsymbol{\lambda})}{\int p(\mathbf{y}|\boldsymbol{\alpha})p(\boldsymbol{\alpha})d\boldsymbol{\alpha}} = \frac{p(\mathbf{y}|\boldsymbol{\lambda})p(\boldsymbol{\lambda})}{p(\mathbf{y})} \propto L(\boldsymbol{\lambda} \mid \mathbf{y})p(\boldsymbol{\lambda}) \tag{3.6}$$

der $L(\boldsymbol{\lambda} \mid \mathbf{y})$ er likelihooden som kan ses på som en funksjon av parametrene når dataene er gitt, og $p(\boldsymbol{\lambda})$ er apriorifordelingen for parametrene.

3.1. Uten tilleggsinformasjon

I den enkleste modellen (3.1) velger vi betafordelingen som apriorifordeling til $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_N)'$, $\theta_i \mid \nu, \omega \sim \text{Beta}(\nu, \omega)$ ("random" effektene i (3.1) blir da fordelt som beta minus en konstant). En fordel med dette valget er at med binomisk fordelte observasjoner: $Y_i \mid \theta_i \sim \text{Bin}(n_i, \theta_i)$ blir den betingede fordelingen til $\theta_i \mid y_i, \nu, \omega$ også en beta-fordeling:

$$\begin{aligned} p(\theta_i \mid y_i, \nu, \omega) &\propto p(y_i \mid \theta_i)p(\theta_i \mid \nu, \omega) = \binom{n_i}{y_i} \theta_i^{y_i}(1-\theta_i)^{n_i - y_i} \frac{1}{B(\nu,\omega)} \theta_i^{\nu-1}(1-\theta_i)^{\omega-1} \\ &\propto \theta_i^{y_i+\nu-1}(1-\theta_i)^{n_i - y_i + \omega - 1} \propto \text{Beta}(y_i + \nu, n_i - y_i + \omega) \end{aligned} \tag{3.7}$$

Der $B(\nu, \omega)$ er beta-funksjonen (se [3]). Proporsjonaliteten i (3.7) og at tettheten må integreres til 1 viser at $p(\theta_i \mid y_i, \nu, \omega) \sim \text{Beta}(y_i + \nu, n_i - y_i + \omega)$. Dette kan utnyttes i MCMC algoritmen ved å benytte Gibbs oppdatering for $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_N)'$ som nettopp krever denne betingede fordelingen på eksplisitt form, men til gjengjeld er lett å implementere i vårt tilfelle. De tilsvarende betingede fordelingene til ν og ω er ikke like enkle å utlede og derfor bruker vi Metropolis Hastings algoritme for å oppdatere disse.

Aposteriori fordelingen i dette tilfellet kan uttrykkes som:

$$p(\boldsymbol{\theta}, \nu, \omega \mid \mathbf{y}) \propto p(\mathbf{y} \mid \boldsymbol{\theta})p(\boldsymbol{\theta} \mid \nu, \omega)p(\nu)p(\omega) \tag{3.8}$$

ν og ω er positive parametre og vi har valgt gamma fordelingen som apriori fordelinger til disse. Med tidligere antatt uavhengighet, både innen Y_i-ene og innen θ_i-ene og alle parametrene i mellom kan simultanfordelingen skrives som:

$$p(\boldsymbol{\theta}, \nu, \omega \mid \mathbf{y}) \propto \prod_i \binom{n_i}{y_i} \theta_i^{y_i}(1-\theta_i)^{n_i - y_i} \prod_i \frac{1}{B(\nu,\omega)} \theta_i^{\nu-1}(1-\theta_i)^{\omega-1} \frac{\nu^{a-1}e^{-\frac{1}{b}\nu}}{b^a\Gamma(a)} \frac{\omega^{c-1}e^{-\frac{1}{d}\omega}}{d^c\Gamma(c)} \tag{3.9}$$

der a,b,c og d er parametre i apriori gammafordelingene og tenkes på som gitt (velges slik at man får høy apriori varians for ν og ω).

3.2. Med tilleggsinformasjon (registersysselsetting for populasjonen)

Med modellen fra (3.5) vil vi bruke følgende fordelinger: $\log(\frac{\theta_i}{1-\theta_i}) \sim N(\beta_0 + \beta_1 x_i, \sigma_v^2)$, og som før $Y_i \mid \theta_i \sim \text{Bin}(n_i, \theta_i)$. Aposteriori fordelingen kan nå skrives:

(3.10) $$p(\boldsymbol{\theta}, \beta_0, \beta_1, \sigma_v^2 \mid \mathbf{y}) \propto p(\mathbf{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \beta_0, \beta_1, \tfrac{1}{\sigma_v^2}) p(\beta_0) p(\beta_1) p(\tfrac{1}{\sigma_v^2})$$

Her har vi modellert $\frac{1}{\sigma_v^2}$ i stedet for σ_v^2 for å kunne bruke en gamma aprioriforfordeling ($\frac{1}{\sigma_v^2}$ må være positiv). Apriori-fordelingene til β_0 og β_1 velges til å være konstanter (=1), dvs. såkalte "improper" apriori fordelinger, som ikke integreres til 1, men slik at man ikke i det hele tatt foretrekker noen verdi for disse parametrene apriori. Fordelingen for $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_N)'$ blir nå en transformert normalfordeling. Med $z_i = \log \frac{\theta_i}{1-\theta_i}$ er $\theta_i = \frac{e^{z_i}}{1+e^{z_i}}$. Når $Z_i \sim N(\mu_i, \sigma^2)$ gir transformasjonsformelen at tettheten til θ_i (innsatt $\mu_i = \beta_0 + \beta_1 x_i$) er gitt ved:

(3.11) $$p(\theta_i \mid \beta_0, \beta_1, \sigma_v^2, x_i) = \frac{1}{\sqrt{2\pi}\sigma_v} e^{-\frac{1}{2}\left(\frac{\log(\frac{\theta_i}{1-\theta_i}) - \beta_0 - \beta_1 x_i}{\sigma_v}\right)^2} \frac{1}{\theta_i(1-\theta_i)}, \quad \theta_i \in (0,1)$$

Skrevet ut får vi:

(3.12)

$$p(\boldsymbol{\theta}, \beta_0, \beta_1, \sigma_v^2 \mid \mathbf{y}) \propto \prod_i \binom{n_i}{y_i} \theta_i^{y_i} (1-\theta_i)^{n_i - y_i} \prod_i \left(\frac{1}{\sqrt{2\pi}\sigma_v} e^{-\frac{1}{2}\left(\frac{\log(\frac{\theta_i}{1-\theta_i}) - \beta_0 - \beta_1 x_i}{\sigma_v}\right)^2} \frac{1}{\theta_i(1-\theta_i)} \right) \frac{(\frac{1}{\sigma_v^2})^{a-1} e^{-\frac{1}{b}(\frac{1}{\sigma_v^2})}}{b^a \Gamma(a)} \times 1 \times 1$$

der a og b velges slik at variansen til apriori fordelingen til $\frac{1}{\sigma_v^2}$ er stor.

4. Implementering av MCMC

4.1. Generelt

Med en spesifisert form på aposteriorifordelingen (simultanfordelingen til alle parametrene gitt data) er vi i stand til å konstruere en Markovkjede som konvergerer mot denne aposteriori fordelingen. Dette oppnås både ved Metropolis Hastings algoritme (MH) og ved Gibbs sampling (GS) som er et spesialtilfelle av MH, men på litt forskjellige måter, se [1].

GS krever at man har tilgjengelig den betingede fordelingen for en parameter, eller sett av parametre som skal oppdateres samtidig, gitt data og de andre parametrene. Oppdateringen for den aktuelle parameteren, eller settet av parametre skjer så ved å trekke fra den betingede fordelingen. Man kan oppdatere alle parametrene etter hverandre systematisk (raster scan) eller plukke ut noen tilfeldig (random scan). Hvis $\mathbf{z}^{(t)}$ er vektoren av alle parametrene ved tiden t defineres Markovkjeden ved denne. Oppdatering av en eller flere av parametrene i denne vektoren vil bringe Markovkjeden inn i en ny tilstand, og etter lang nok tid vil Markovkjeden havne i stasjonærfordelingen sin. Når alle

parametrene (eller de tilfeldige som ble trukket ut)har blitt oppdatert en gang har man gjort unna en iterasjon i algoritmen og dette gjentas mange ganger. De parametrene som betinges på beholder verdien fra forriga iterasjon, eller ev. nylig oppdatert verdi.

MH er en mer generell algoritme (som har GS som et spesialtilfelle) og hvor man ikke krever annet enn at man har spesifisert aposteriorifordelingen og at man har en god forslagsfunksjon og tilhørende akseptsannsynlighet. Forslagsfunksjonen skal foreslå nye tilstander til Markovkjeden som så skal aksepteres med akseptsannsynligheten, slik at når Markovkjeden har gjennomløpt mange nok tilstander, vil den til slutt havne i stasjonær fordelingen som nettopp er vår aposteriori fordeling. Betrakter man en parameter (variabel) for seg er dette den marginale fordelingen til denne parameteren gitt data (gjelder både MH og GS).

Man kan kombinere GS og MH, og det gjør vi i vårt tilfelle, når vi har modell (3.1). Da passer det å bruke GS for oppdatering av $\theta_1,\cdots,\theta_N$ fordi vi har den betingede fordelingen for $\theta_i \mid y_i,\nu,\omega$ tilgjengelig. For oppdatering av ν og ω har vi ikke de tilsvarende fordelingene så her bruker vi MH oppdatering.

I modell (3.5) brukes MH til oppdatering av alle parametrene.

I MH bruker vi en "random walk" forslagsfunksjon (med ev. trunkering gitt av parameterrommet), dvs. måten å foreslå en ny tilstand til Markovkjeden på er å ta utgangspunkt i den tilstanden den allerede har og så legge på litt støy som i vårt tilfelle er normalfordelt. Forslagsfunksjonen er derfor en normaltetthet med forventning lik den nåværende tilstanden og et standardavvik som bestemmer hvor mye Markovkjeden skal "hoppe" hvis den nye tilstanden blir akseptert.

Vi betegner nåværende tilstand til Markovkjeden for $\mathbf{z}^{(t)}$ (igjen vektoren som består av alle parametrene som inngår i aposteriori fordelingen og oppdateres iterativt) og den foreslåtte nye tilstanden for $\mathbf{z}^*$. Videre kaller vi forslagsfunksjonen for $q(z_j^{(t)}, z_j^*)$ og mener med dette forslagsfunksjonen som tar utgangspunkt i nåværende tilstand $z_j^{(t)}$ for komponent j i $\mathbf{z}^{(t)}$, og foreslår en ny tilstand z_j^* for denne. Generelt kan forslagsfunksjonen være en multivariat fordeling, slik at flere parametre oppdateres samtidig, men ikke i våre tilfeller. Da får vi følgende uttrykk for $q(z_j^{(t)}, z_j^*)$:

$$(4.1) \qquad q(z_j^{(t)}, z_j^*) = \frac{1}{\sqrt{2\pi}\sigma_q\left(\Phi\left(\frac{c_ø - z_j^{(t)}}{\sigma_q}\right) - \Phi\left(\frac{c_n - z_j^{(t)}}{\sigma_q}\right)\right)} e^{-\frac{1}{2}\left(\frac{z_j^* - z_j^{(t)}}{\sigma_q}\right)^2}, \quad c_n \le z_j^* \le c_ø$$

som er en trunkert normalfordeling med forventning i nåværende tilstand, $\Phi(\ )$ er den kumulative standard normalfordelingen. Trunkeringen avgjøres av parameterrommet. Hvis man f.eks. skal oppdatere en parameter som må være positiv settes $c_n = 0$ og $c_ø = \infty$. σ_q^2 velges spesifikt for hver parameter slik at akseptraten ligger mellom 20 og 70 %.

Akseptsannsynligheten er gitt ved uttrykket:

$$(4.2) \qquad a(\mathbf{z}^{(t)}, \mathbf{z}^*) = \min\left\{1, \frac{p(\mathbf{z}^*)q(z_j^*, z_j^{(t)})}{p(\mathbf{z}^{(t)})q(z_j^{(t)}, z_j^*)}\right\}$$

der $p(\ )$ er aposteriorifordelingen og $\mathbf{z}^*$ er vektoren hvor komponent j er satt lik den nye foreslåtte verdien og alle andre parametre er som før (dette siden vi kun endrer en parameter adgangen). Etter at akseptsannsynligheten er utregnet velges den nye tilstanden for komponent j slik:

(4.3)
$$z_j^{(t+1)} = \begin{cases} z_j^* & \text{med sannsynlighet } a(\mathbf{z}^{(t)}, \mathbf{z}^*) \\ z_j^{(t)} & \text{med sannsynlighet } \left(1 - a(\mathbf{z}^{(t)}, \mathbf{z}^*)\right) \end{cases}$$

og resten av komponentene blir værende uforandret. Vi oppdaterer alle parametrene systematisk en etter en ("raster scan").

Vi trenger også fornuftige valg av initialverdier til parametrene (selv om stasjonærfordelingen til Markovkjeden er uavhengig av initialverdien) og en formening om hvor mange iterasjoner man trenger. Vi velger initialverdiene fra dataene ved vanlige moment / ML -estimater.

Flere ting må tas hensyn til for å beregne nødvendig antall iterasjoner, vi har latt ressursene og tiden bestemme, men med tanken om at jo flere iterasjoner, jo bedre resultat. Dette er fordi man ikke bare er interessert i at Markovkjeden er i stasjonærfordelingen, men også at asymptotikken i store talls lov skal få" virke så lenge som mulig" slik at gjennomsnittet som brukes som estimat for forventningen i de forskjellige punktestimatene vil ha minst mulig varians. Programpakken S+ er brukt for beregninger og begge de omtalte algorimene er tidkrevende pr. iterasjon. Det er av stor betydning å programmere fornuftig, for å f.eks. ungå flere løkker inni hverandre. Vi har derfor kjørt algoritmene så lenge som det har vært praktisk mulig, og så gjort beregninger om nøyaktigheten etterpå.

Hvis vi betrakter Markovkjeden for komponent j i vektoren $\mathbf{z}^{(t)}$, er de etterfølgende tilstandene avhengige. Likevel gjelder sentralgrenseteoremet på formen:

(4.4)
$$\sqrt{T}\left(\tfrac{1}{T}\sum_{t=1}^{T} f\left(z_j^{(t)}\right) - E_p\left[f\left(z_j\right)\right]\right) \xrightarrow{\ d\ } N\left(0, Var_p\left[f(z_j)\right]\tau^2\right)$$
$$\text{der} \quad \tau^2 = \sum_{k=-\infty}^{\infty} \rho\left(|k|\right)$$

der $f\left(z_j^{(t)}\right)$ er en vilkårlig funksjon av vår parameter, p betegner stasjonær fordelingen, d betegner konvergens i fordeling, τ^2 er korreksjonsfaktoren som avhengigheten forårsaker , T er antall iterasjoner, og ρ er autokorrelasjonen.

Vi bruker det vanlige variansestimatet for $Var_p\left[f(z_j)\right]$:

(4.5)
$$Var_p\left(f(z_j)\right) \approx \frac{1}{T}\sum_{t=1}^{T}\left(f(z_j^{(t)}) - \overline{f}_T\right)^2$$

Et approksimativt 95 % konfidensintervall for $E_p\left(f(z_j)\right)$ vil dermed være gitt ved:

(4.6)
$$\left[\tfrac{1}{T}\sum_{t=1}^{T} f(z_j^{(t)}) - 1.96\sqrt{\frac{\left(Var_p\left(f(z_j)\right)\right)\tau^2}{T}}, \tfrac{1}{T}\sum_{t=1}^{T} f(z_j^{(t)}) + 1.96\sqrt{\frac{\left(Var_p\left(f(z_j)\right)\right)\tau^2}{T}}\right]$$

Det som gjenstår for å kunne beregne (4.6) er et estimat for τ^2. Vi setter estimert autokorrelasjon inn i (4.4) og må da bare finne ut hvor mange ledd som skal være med i summen. Her bruker vi Geyers metode (se [1]), som sier at hvis man starter i stasjonærfordelingen og betrakter funksjonen $\Gamma(k) = \gamma(2k) + \gamma(2k+1)$ der $\gamma(\ )$ er autokovariansfunksjonen så vil denne være positiv for små k og avtagende (estimeres ved å sette inn estimert autokovariansfunksjon). Da vil man kunne estimere τ^2 ved:

$$\widehat{\tau^2} = \sum_{k=-k_0}^{k_0} \hat{\rho}(|k|) \tag{4.7}$$

der k_0 er den første k slik at $\hat{\Gamma}(k) < 0$, $\hat{\rho}$ er estimert autokorrelasjon. Siden dette forutsetter at man allerede er i stasjonærfordelingen må man først prøve å få en oppfatning av hvor fort konvergensen foregår (ved å se på plottet av parametrene som funksjon av antall iterasjoner), og så kjøre Markov-kjeden T iterasjoner etter at man mener man er i stasjonærfordelingen. De første iterasjonene før man er i stasjonærfordelingen kalles "burn-in", betegnes med T_0 og tas ikke vare på.

4.2. Uten tilleggsinformasjon

I modellen fra avsnitt 3.1 bruker vi som nevnt både GS og MH oppdatering av parametrene. Vi oppdaterer $\theta_1, \cdots, \theta_{432}$ med GS, og ν og ω ved MH. Initialverdier for parametrene velger vi fra dataene.

Oppdatering av $\theta_1, ..., \theta_{432}$:
Initielt setter vi disse parametrene lik de andelene vi har i utvalget (histogrammene i figur 2), alt etter hva vi bruker som observasjoner. Bruker vi AKU-utvalget som observasjoner settes også initialverdiene for $\theta_1, \cdots, \theta_{432}$ lik andelene vi har AKU-utvalget ($\frac{Y_i^s}{n_i}$). Bruker vi register-informasjonen om AKU-utvalget som observasjoner settes initialverdiene for $\theta_1, \cdots, \theta_{432}$ lik $\frac{X_i^s}{n_i}$.

For hver θ_i oppdateres verdien med å trekke fra fordelingen:

$$p(\theta_i^{(t+1)} \mid y_i, \nu^{(t)}, \omega^{(t)}) \sim \text{Beta}(y_i + \nu^{(t)}, n_i - y_i + \omega^{(t)}) \tag{4.8}$$

der (t) betegner nåværende tilstand og (t+1) neste tilstand. Vi setter altså inn eksisterende tilstand for ν og ω. Alle θ_i-ene er uavhengige av hverandre, så dette kan gjøres på vektorform (samtidig for alle θ_i-ene).

Oppdatering av ν og ω:
Initialverdiene til disse setter vi lik momentestimatene i en betafordeling (se [3]), gitt ved:

$$\nu_0 = \frac{\bar{\theta}\left[\bar{\theta}(1-\bar{\theta}) - s_0^2\right]}{s_0^2}, \qquad \omega_0 = \frac{(1-\bar{\theta})\left[\bar{\theta}(1-\bar{\theta}) - s_0^2\right]}{s_0^2} \tag{4.9}$$

der s_θ^2 er det vanlige variansestimatet for initialverdiene til $\theta_1, \cdots, \theta_{432}$. Siden både ν og ω er positive parametre benytter vi en trunkert forslagsfunksjon (4.1) med $c_n = 0, c_\omega = \infty$. I utregningen av akseptsannsynligheten (4.2) kan mye forkortes vekk. For akseptsannsynligheten for ν fås:

(4.10) $$a_\nu = \min\left\{1, \frac{\prod_i \frac{1}{B(\nu^*,\omega^{(t)})}\theta_i^{(t+1)(\nu^*-1)}\nu^{*(a-1)}e^{-\frac{1}{b}\nu^*}\left[1-\Phi\left(-\frac{\nu^{(t)}}{\sigma_{q,\nu}}\right)\right]}{\prod_i \frac{1}{B(\nu^{(t)},\omega^{(t)})}\theta_i^{(t+1)(\nu^{(t)}-1)}\nu^{(t)(a-1)}e^{-\frac{1}{b}\nu^{(t)}}\left[1-\Phi\left(-\frac{\nu^*}{\sigma_{q,\nu}}\right)\right]}\right\}$$

Legg merke til at nå benyttes de sist oppdaterte verdiene for $\theta_1,\cdots,\theta_{432}$ (betegnes med (t+1)). Tilsvarende for ω:

(4.11) $$a_\omega = \min\left\{1, \frac{\prod_i \frac{1}{B(\nu^{(t+1)},\omega^*)}\left(1-\theta_i^{(t+1)}\right)^{(\omega^*-1)}\omega^{*(c-1)}e^{-\frac{1}{d}\omega^*}\left[1-\Phi\left(-\frac{\omega^{(t)}}{\sigma_{q,\omega}}\right)\right]}{\prod_i \frac{1}{B(\nu^{(t+1)},\omega^{(t)})}\left(1-\theta_i^{(t+1)}\right)^{(\omega^{(t)}-1)}\omega^{(t)(c-1)}e^{-\frac{1}{d}\omega^{(t)}}\left[1-\Phi\left(-\frac{\omega^*}{\sigma_{q,\omega}}\right)\right]}\right\}$$

Og siden vi nå har oppdatert $\theta_1,\cdots,\theta_{432}$ og ν brukes de sist oppdaterte verdiene for disse.

4.3. Med tilleggsinformasjon (registersysselsetting for populasjonen)

I modellen fra avsnitt 3.2 er det mest naturlig med MH oppdatering av alle parametrene, siden vi ikke har tilgjengelig de betingede fordelingene for en parameter gitt de andre.

Oppdatering av $\theta_1,...,\theta_{432}$:

Initialverdier til disse parametrene velges som i forrige avsnitt lik de andelene vi har i utvalget (histogrammene i figur 2), alt etter hva vi bruker som observasjoner, bruker vi AKU-utvalget som observasjoner settes også initialverdiene for $\theta_1,\cdots,\theta_{432}$ lik andelene i AKU-utvalget, og ellers settes initialverdiene for $\theta_1,\cdots,\theta_{432}$ lik $\frac{x_i^s}{n_i}$.

Forslagsfunksjonen for hver θ_i er nå trunkert både under 0 og over 1 for å gi lovlig verdi på forslaget. Mye kan forkortes her også i beregningen av akseptsannsynligheten, og etter forenkling får man uttrykket:

(4.12)
$$a_\theta = \min\left\{1, \frac{\theta_i^{*y_i}\left(1-\theta_i^*\right)^{(n_i-y_i)}e^{-\frac{1}{2}\left(\frac{1}{\sigma^2}\right)^{(t)}\left(\log\left(\frac{\theta_i^*}{1-\theta_i^*}\right)-\beta_0^{(t)}-\beta_1^{(t)}x_i\right)^2}\theta_i^{(t)}\left(1-\theta_i^{(t)}\right)\left[\Phi\left(\frac{1-\theta_i^{(t)}}{\sigma_{q,\theta}}\right)-\Phi\left(-\frac{\theta_i^{(t)}}{\sigma_{q,\theta}}\right)\right]}{\theta_i^{(t)y_i}\left(1-\theta_i^{(t)}\right)^{(n_i-y_i)}e^{-\frac{1}{2}\left(\frac{1}{\sigma^2}\right)^{(t)}\left(\log\left(\frac{\theta_i^{(t)}}{1-\theta_i^{(t)}}\right)-\beta_0^{(t)}-\beta_1^{(t)}x_i\right)^2}\theta_i^*\left(1-\theta_i^*\right)\left[\Phi\left(\frac{1-\theta_i^*}{\sigma_{q,\theta}}\right)-\Phi\left(-\frac{\theta_i^*}{\sigma_{q,\theta}}\right)\right]}\right\}$$

Igjen settes parameterverdiene til alle andre parametre enn θ_i som skal oppdateres til de sist oppdaterte verdiene, og alle θ_i-ene er uavhengige av hverandre, så dette kan gjøres på vektorform (samtidig for alle θ_i-ene). x_i er områdeinformasjonen og er lik $x_i = \log\left(\frac{\frac{x_i^p}{N_i}}{1-\frac{x_i^p}{N_i}}\right)$ (se kapittel 3).

Oppdatering av β_0 og β_1:

Initialverdier for β_0 og β_1 settes til 1, pga. sine "improper" apriori fordelinger. Mht forslagsfunksjonene blir ingen trunkering nødvendig siden alle verdier her er tillatt. Dermed forkortes alt som har med forslagsfunksjonen bort i utregning av akseptsannsynligheten. Vi får for β_0:

(4.13)
$$a_{\beta_0} = \min\left\{1, \frac{\prod_i e^{-\frac{1}{2}\left(\frac{1}{\sigma^2}\right)^{(t)}\left[\log\left(\frac{\theta_i^{(t+1)}}{1-\theta_i^{(t+1)}}\right)-\beta_0^{*}-\beta_1^{(t)}x_i\right]^2}}{\prod_i e^{-\frac{1}{2}\left(\frac{1}{\sigma^2}\right)^{(t)}\left[\log\left(\frac{\theta_i^{(t+1)}}{1-\theta_i^{(t+1)}}\right)-\beta_0^{(t)}-\beta_1^{(t)}x_i\right]^2}}\right\}$$

og legger merke til at $\theta_1,\cdots,\theta_{432}$ er oppdatert og kan brukes i tilstand (t+1). Helt tilsvarende er det for β_1 som får akseptsannsynlighet:

(4.14)
$$a_{\beta_1} = \min\left\{1, \frac{\prod_i e^{-\frac{1}{2}\left(\frac{1}{\sigma^2}\right)^{(t)}\left[\log\left(\frac{\theta_i^{(t+1)}}{1-\theta_i^{(t+1)}}\right)-\beta_0^{(t+1)}-\beta_1^{*}x_i\right]^2}}{\prod_i e^{-\frac{1}{2}\left(\frac{1}{\sigma^2}\right)^{(t)}\left[\log\left(\frac{\theta_i^{(t+1)}}{1-\theta_i^{(t+1)}}\right)-\beta_0^{(t+1)}-\beta_1^{(t)}x_i\right]^2}}\right\}$$

hvor nye tilstander for $\theta_1,\cdots,\theta_{432}$ og β_0 kan brukes.

Oppdatering av $\frac{1}{\sigma^2}$:

Initialverdi for denne parameteren velges lik den inverse av estimert varians til registerandelene på logit-skala (inverse av $\mathrm{Var}\left(\log\left(\frac{\frac{x_i^p}{N_i}}{1-\frac{x_i^p}{N_i}}\right)\right)$). Denne variansen er selvfølgelig altfor stor, siden den tilsier at all variasjonen kommer fra "random effects" og ikke noe fra selve observasjonene eller kovariatene, men likevel fungerer som initialverdi. Forslagsfunksjonen må være trunkert under 0. Vi får følgende uttrykk for akseptsannsynligheten:

(4.15)
$$a_{\frac{1}{\sigma^2}} = \min\left\{1, \frac{\left(\left(\frac{1}{\sigma^2}\right)^{*}\right)^{\frac{N}{2}}\prod_i e^{-\frac{1}{2}\left(\frac{1}{\sigma^2}\right)^{*}\left[\log\left(\frac{\theta_i^{(t+1)}}{1-\theta_i^{(t+1)}}\right)-\beta_0^{(t+1)}-\beta_1^{(t+1)}x_i\right]^2}\left(\frac{1}{\sigma^2}\right)^{*(a-1)}e^{-\frac{1}{b}\left(\frac{1}{\sigma^2}\right)^{*}}\left[1-\Phi\left(-\frac{\left(\frac{1}{\sigma^2}\right)^{(t)}}{\sigma_{q,\frac{1}{\sigma^2}}}\right)\right]}{\left(\left(\frac{1}{\sigma^2}\right)^{(t)}\right)^{\frac{N}{2}}\prod_i e^{-\frac{1}{2}\left(\frac{1}{\sigma^2}\right)^{(t)}\left[\log\left(\frac{\theta_i^{(t+1)}}{1-\theta_i^{(t+1)}}\right)-\beta_0^{(t+1)}-\beta_1^{(t+1)}x_i\right]^2}\left(\frac{1}{\sigma^2}\right)^{(t)(a-1)}e^{-\frac{1}{b}\left(\frac{1}{\sigma^2}\right)^{(t)}}\left[1-\Phi\left(-\frac{\left(\frac{1}{\sigma^2}\right)^{*}}{\sigma_{q,\frac{1}{\sigma^2}}}\right)\right]}\right\}$$

hvor alle de andre parametrene allerede har blitt oppdatert og inngår med siste verdi.

5. Resultater

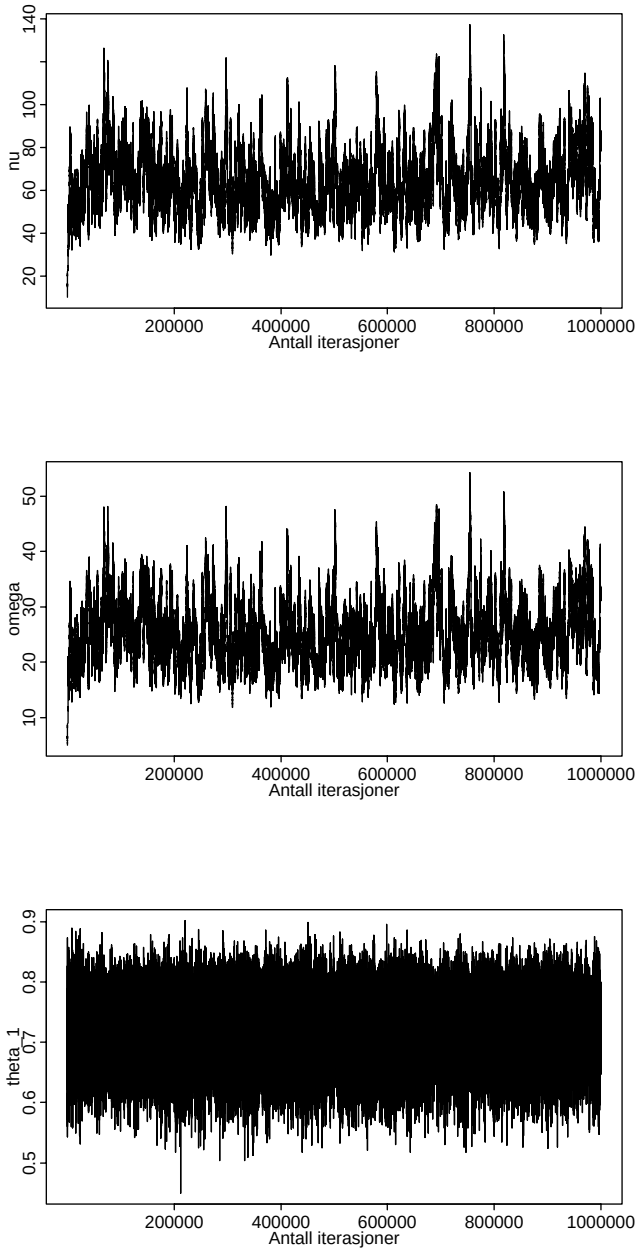
Antall iterasjoner i den første simuleringen er 1×10^6. Det tok mellom 10 og 14 timer. Gjentatte forsøk med så mange iterasjoner førte til problemer med S+. Hver iterasjon brukte lengre og lengre tid, så for å begrense tidsbruken er alle bortsett fra den første simuleringen gjennomført med 500 000 iterasjoner i stedet.

Konvergens til Markovkjeden illustreres ved å plotte parametrene mot antall iterasjoner. Vi har valgt en burn-in periode på $T_0 = 100\,000$ iterasjoner i alle simuleringene.

5.1. Uten tilleggsinformasjon

Først bruker vi registersysselsettingen for AKU-utvalget som observasjoner, altså $\frac{X_i^s}{n_i}$, og modellen fra avsnitt 3.1. Vi bruker disse observasjonene først og fremst for sammenligning med det andre tilfellet når AKU-utvalget selv er observasjoner, og for å kontrollere kvaliteten på estimatene i forhold til fasiten, gitt ved $\frac{X_i^p}{N_i}$.

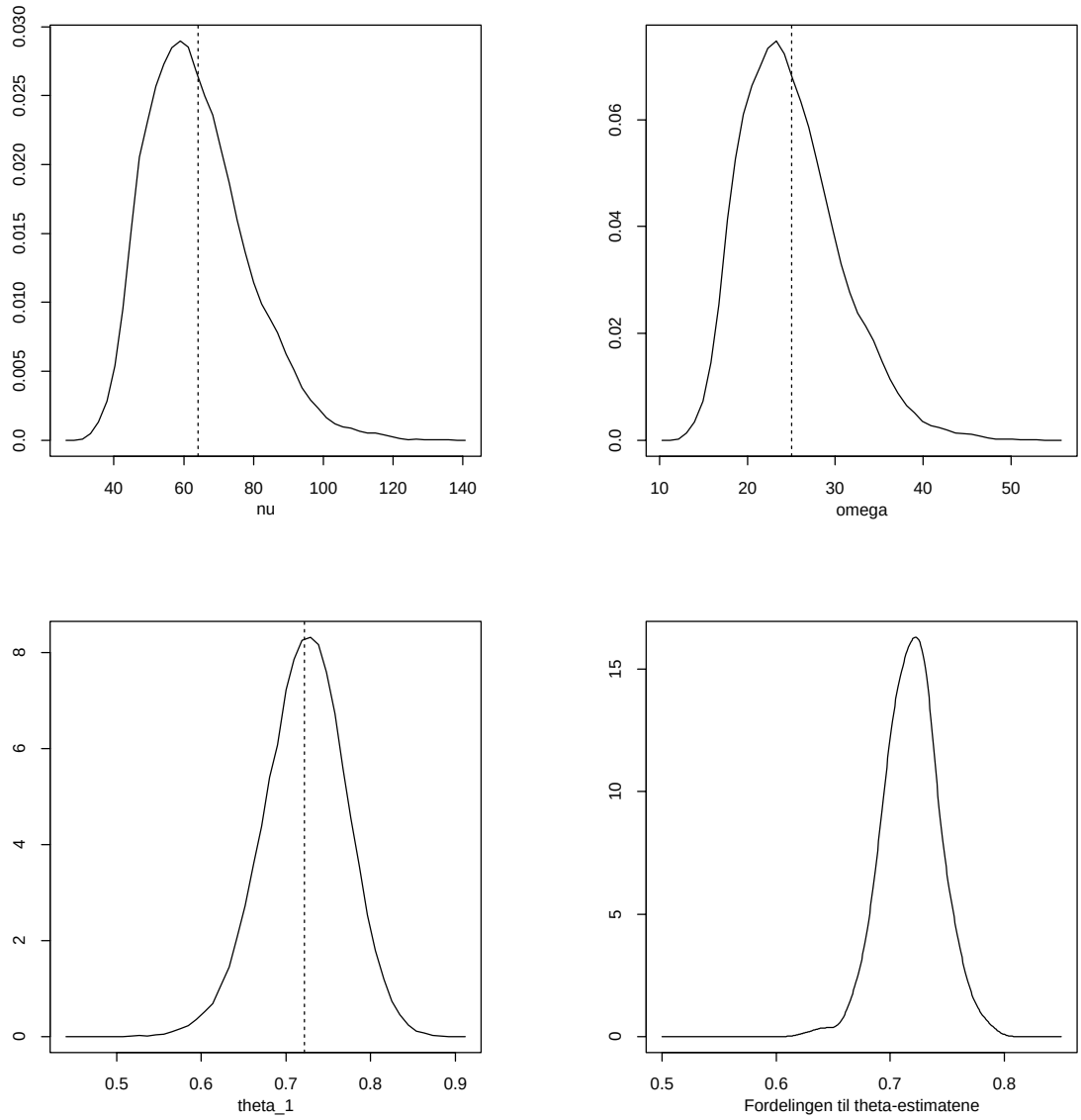
Som den eneste gangen er det her kjørt 1×10^6 iterasjoner. Akseptraten for parametrene er : $AR_\nu = 50,7\%$ og $AR_\omega = 42,6\%$. For $\theta_1,\cdots,\theta_{432}$ er det 100% aksept fordi disse oppdateres ved GS, som betyr å trekke direkte fra den betingede fordelingen. Neste side viser figuren hvordan de forskjellige parametrene varierer som funksjon av antall iterasjoner.



Figur 3: Konvergens for ν, ω og θ_1

Vi ser at ν og ω konvergerer seinere enn θ_1 og at $T_0 = 100\,000$ ser ut til å være stor nok til å gi en stasjonær fordeling for $T > T_0$.

Figuren nedenfor viser tetthetsestimater for de marginale aposteriori fordelingene til parametrene. Gjennomsnittet er markert med prikket linje siden dette brukes som estimat for forventningen. Fordelingen til $\hat{\theta}_1, \cdots, \hat{\theta}_{432}$ er tatt med nederst til høyre.



Figur 4: Marginale aposteriori fordelinger for ν, ω og θ_1 (de prikkede linjene er gjennomsnitt), og fordelingen til $\hat{\theta}_1,...,\hat{\theta}_{432}$. Registersysselsettingen for AKU-utvaleget er observasjonene ($\frac{X_i^s}{n_i}$).

Vi legger merke til at fordelingen til $\theta_1 \mid X_1^s$ som er vist nederst til venstre, ligner nokså mye på fordelingen til $\hat{\theta}_1,\cdots,\hat{\theta}_{432}$ nederst til høyre, som er estimatet på den ubetingede fordelingen til θ_i, som $\theta_1,\cdots,\theta_{432}$ er uavhengige identisk fordelte realisasjoner fra. Grunnen til likheten er at observasjonen, i dette tilfellet er $X_1^s = 1$ og $n_1 = 1$. Det betyr at dette området får 100 % sysselsetting i utvalget på bakgrunn av en person som tilfeldigvis er sysselsatt. Med så liten informasjon i observasjonen forandres ikke den betingede fordelingen til θ_1 seg så mye fra den ubetingede.

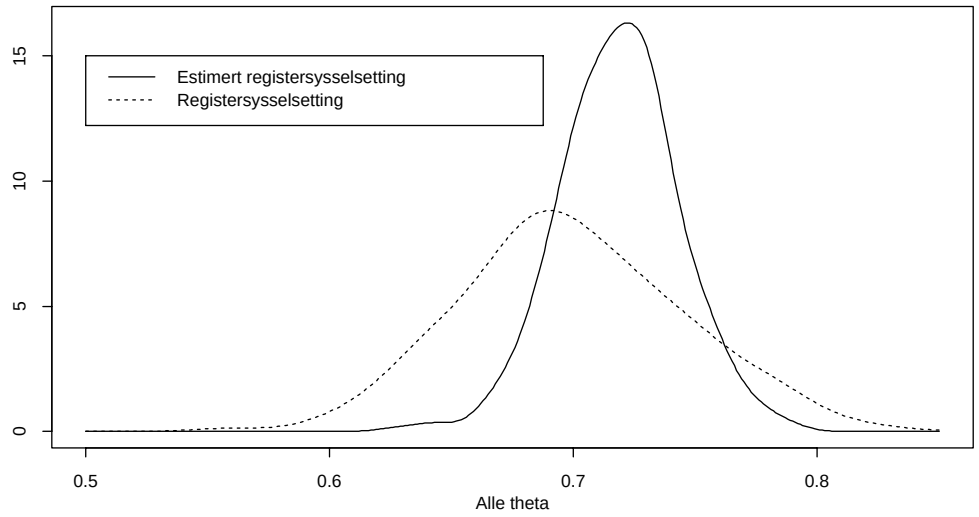
Nedenfor viser tabellen punktestimatene (gj.snittene) for parametrene og tilnærmede 95% konfidens-intervaller for forventningene, som er beregnet ved uttrykkene i (4.5), (4.6) og (4.7).

PARAMETER	PUNKTESTIMAT (GJENNOMSNITT)	APPR. 95% KONF. INTERVALL
ν	64,05	62,25_65,86
ω	25,08	24,38_25,78
θ_l	0,7219	0,7216_0,7222

Tabell 1: Parameter estimater og approksimative 95% konfidensintervaller for forventningen

Disse resultatene er nokså like de Zhang har i sin analyse (se [2]).

Figur 5 viser tetthetsestimater av registersysselsettingen $\frac{X_i^p}{N_i}$, og estimert registersysselsetting $\hat{\theta}_i$, $i = 1, \cdots, 432$. Siden vi bruker registersysselsettingen til AKU-utvalget som observasjoner ($\frac{X_i^s}{n_i}$), vil $\hat{\theta}_i$ være estimert registersysselsetting til hele populasjonen som vi dermed kan sammenligne med fasiten, $\frac{X_i^p}{N_i}$.

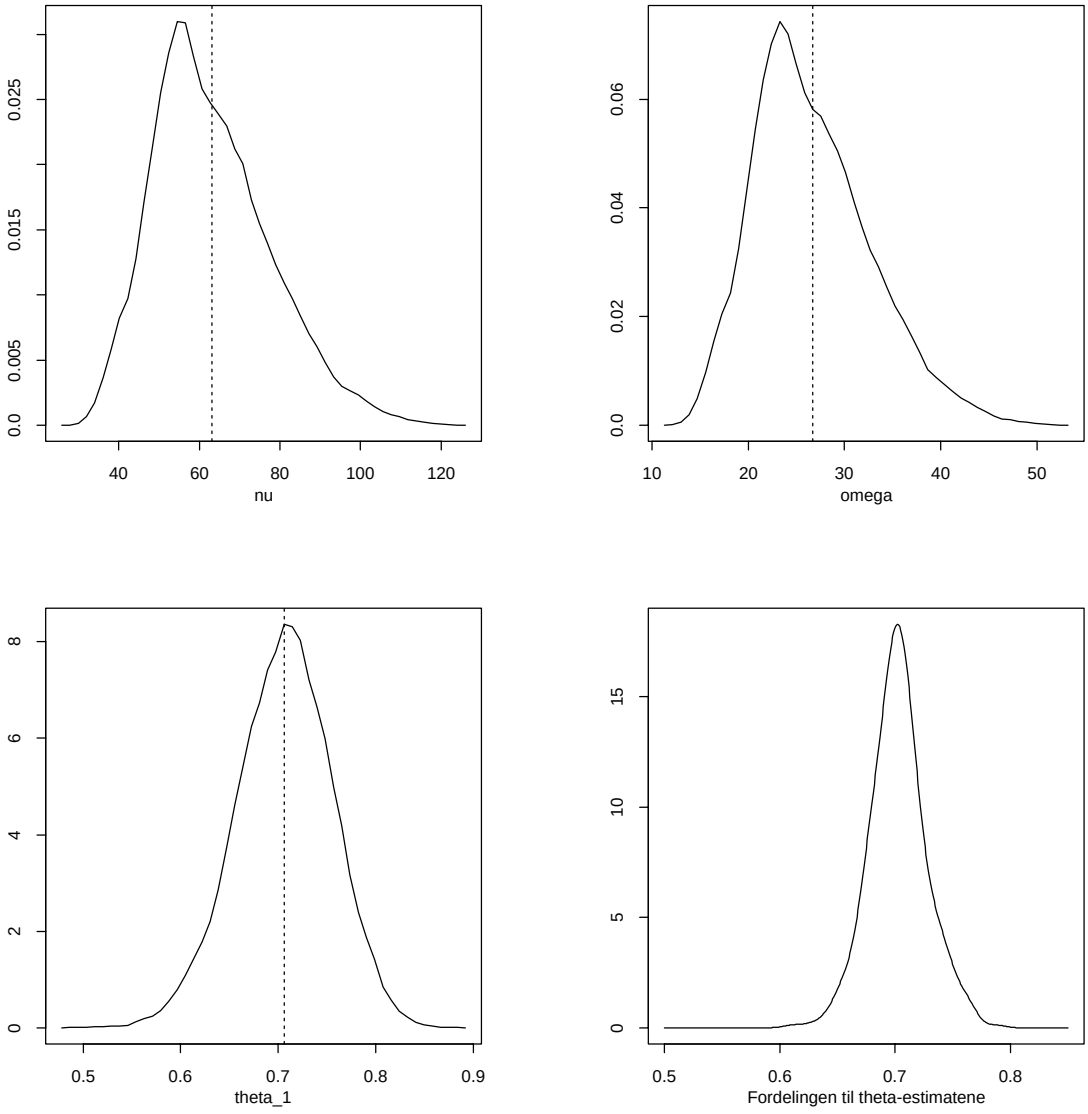


Figur 5: Tetthetsestimater av registerersysselsetting ($\frac{X_i^p}{N_i}$) og estimert registersysselsetting ($\hat{\theta}_i$) når observasjonene er registersysselsettingen for AKU-utvalget ($\frac{X_i^s}{n_i}$)

Vi legger merke til at fordelingen til $\hat{\theta}_i$-ene er forskjøvet litt til høyre for fordelingen til $\frac{X_i^p}{N_i}$, og dette var synlig allerede i begynnelsen da vi betraktet dataene (figur 1 og figur 2) og fant at gjennomsnittet i $\frac{X_i^s}{n_i}$ var høyere enn gjennomsnittet i $\frac{X_i^p}{N_i}$. Videre ser vi tydelig en "overshrinkage" altså at variasjonen i estimatene er for liten i forhold til den virkelige variasjonen. Dette er et resultat av at esimatene er gitt ved den betingede forventningen $\hat{\theta}_i = E\left(\theta_i \mid X_i^s\right)$. $\hat{\theta}_i$-ene dras for mye mot det globale gjennomsnittet.

Helt tilsvarende som i figur 3, 4 og 5 har vi brukt samme modellen, men byttet ut observasjonene med AKU-utvalget selv ($\frac{Y_i^s}{n_i}$). Resultatene følger nedenfor.

Pga. tidsbruken er antall iterasjoner halvert til 500 000, mens $T_0 = 100\,000$ fremdeles. Akseptraten for parametrene er : $AR_\nu = 49,4\%$ og $AR_\omega = 44\%$. For $\theta_1, \cdots, \theta_{432}$ er det som tidligere akseptrate på 100%. Første figuren nedenfor tilsvarer figur 4 og viser de marginale aposteriori fordelingene til parametrene.



Figur 6: Marginale aposteriori fordelinger for ν, ω og θ_1 (de prikkede linjene er gjennomsnitt), samt fordeling til $\hat{\theta}_1, ..., \hat{\theta}_{432}$. AKU-utvalget er observasjoner ($\frac{Y_i^s}{n_i}$).

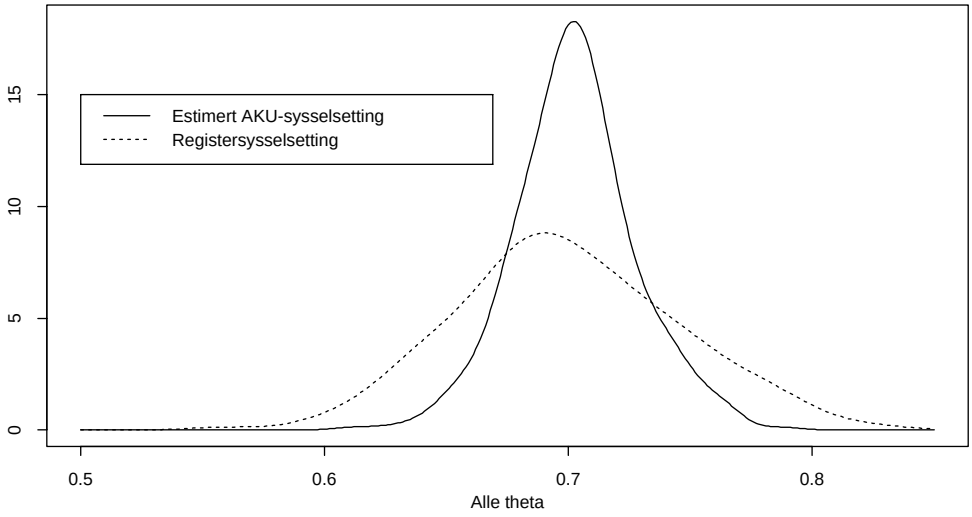
Vi ser at figuren er nokså lik figur 4. En tabell for punktestimater og konfidensintervall er vist på neste side.

PARAMETER	PUNKTESTIMAT (GJENNOMSNITT)	APPR. 95% KONF. INTERVALL
ν	63,1	60,46_65,74
ω	26,69	25,59_27,79
θ_1	0,706	0,7055_0,7065

Tabell 2: Parameter estimater og approksimative 95% konfidensintervaller for forventningen

Sammenlignet med tabell 1 er konfidensintervallene her litt bredere pga. lavere antall iterasjoner.

Nedenfor vises tetthetsestimater av registersysselsettingen og $\hat{\theta}_1,\cdots,\hat{\theta}_{432}$ når selve AKU-utvalget er observasjoner ($\frac{Y_i^s}{n_i}$).

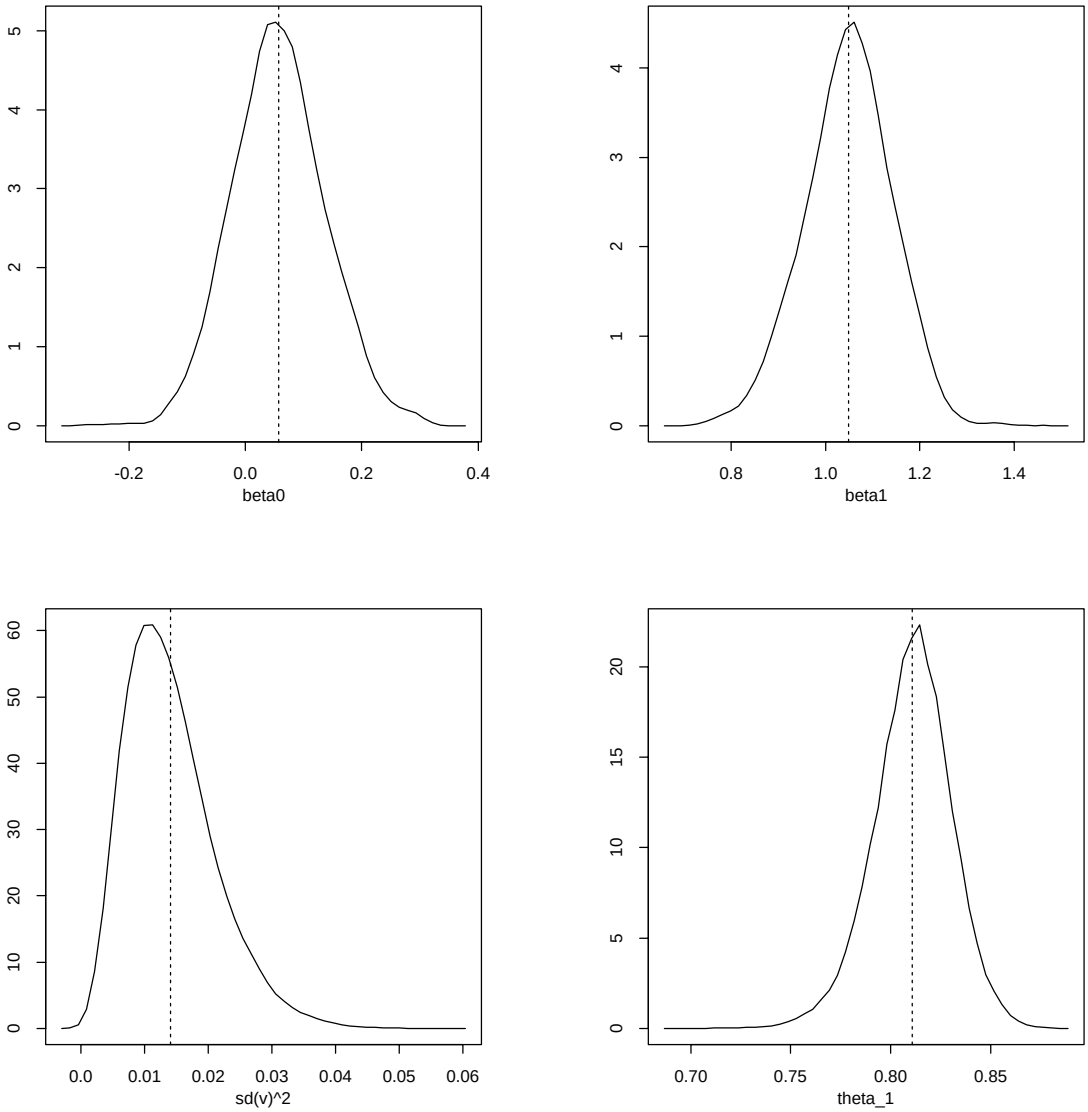


Figur 7: Tetthetsestimater av registersysselsetting ($\frac{X_i^p}{N_i}$) og estimert AKU-sysselsetting ($\hat{\theta}_i$) når observasjonene er AKU-utvalget ($\frac{Y_i^s}{n_i}$)

Sammenlignet med figur 5 ser vi at forskyvningen i forhold til registeret er omtrent borte, men dette er selvfølgelig tilfeldig og pga. vårt spesielle AKU-utvalg, som ligner mer på registeret enn $\frac{X_i^s}{n_i}$ gjør. Vi ser også at fordelingen til $\hat{\theta}_1,\cdots,\hat{\theta}_{432}$ er litt mer symmetrisk enn i figur 5. Ellers ser vi samme grad av "overshrinkage" her.

5.2. Med tilleggsinformasjon (registersysselsetting for populasjonen)

Nå bruker vi modellen fra avsnitt 3.2 og først med registerisysselsettingen for AKU-utvalget som observasjoner ($\frac{X_i^s}{n_i}$). Antall iterasjoner er 500 000. Akseptraten for parametrene er : $AR_{\beta_0} = 56,5\%$, $AR_{\beta_1} = 52,3\%$, $AR_{\frac{1}{\sigma^2}} = 30,2\%$. For $\theta_1, \cdots, \theta_{432}$ er det varierende akseptrate som ligger mellom 37% og 73%. Figuren nedenfor viser marginale aposteriorifordelinger som i avsnitt 5.1.



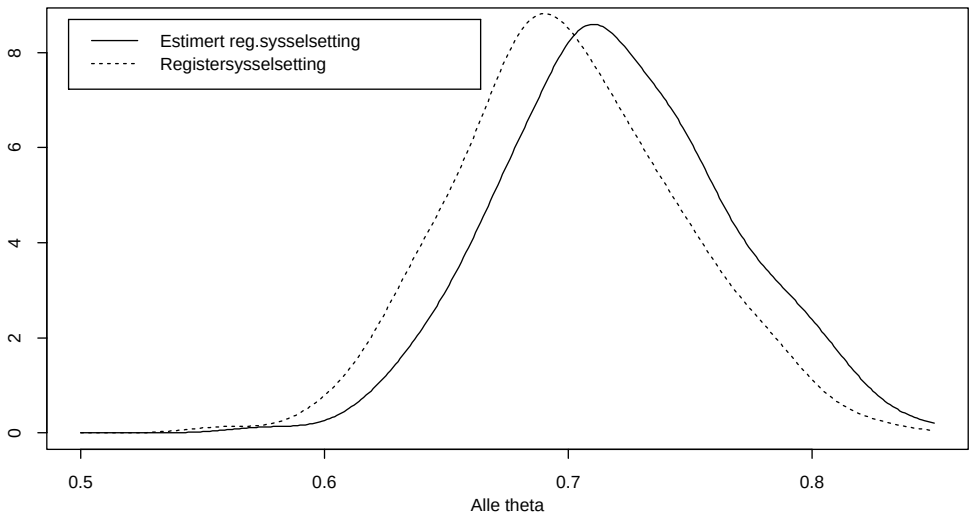
Figur 8: Marginale aposteriori fordelinger for $\beta_0, \beta_1, \frac{1}{\sigma^2}$ og θ_1, med $\frac{X_i^s}{n_i}$ som observasjoner og $\log\left(\frac{\frac{x_i^p}{N_i}}{1-\frac{x_i^p}{N_i}}\right)$ som kovariat.

Vi legger merke til en stor økning i $\hat{\theta}_1$ i forhold til tidligere. Tabellen på neste side viser punktestimater for parametrene og tilnærmede konfidensintervaller for forventningen til disse.

PARAMETER	PUNKTESTIMAT (GJENNOMSNITT)	APPR. 95% KONF. INTERVALL
β_0	0,0575	0,0478_0,0672
β_1	1,0485	1,0373_1,0597
σ_v^2	0,0141	0,0137_0,0145
θ_1	0,8109	0,8101_0,8117

Tabell 3: Parameter estimater og tilnærmede 95% konfidensintervaller for forventningen

Tabellen viser signifikant positiv forskyvning (β_0) mellom kovariat og $\hat{\theta}_i$ (på logit-skala) og en proporsjonalitet (β_1) signifikant større enn 1 (så vidt). Man kan argumentere for å utelate β_1 fra modellen (se [2]). Det er verdt å legge merke til den nokså store varianskomponenten til tross for at observasjonene kommer fra registeret selv. Denne variansen påvirker forventningen til $\hat{\theta}_i$ pga. logit-skalaen (som er ikke-lineær) og tolkningen av størrelsen på den er ikke opplagt, men i det minste en bekreftelse på "random effects".



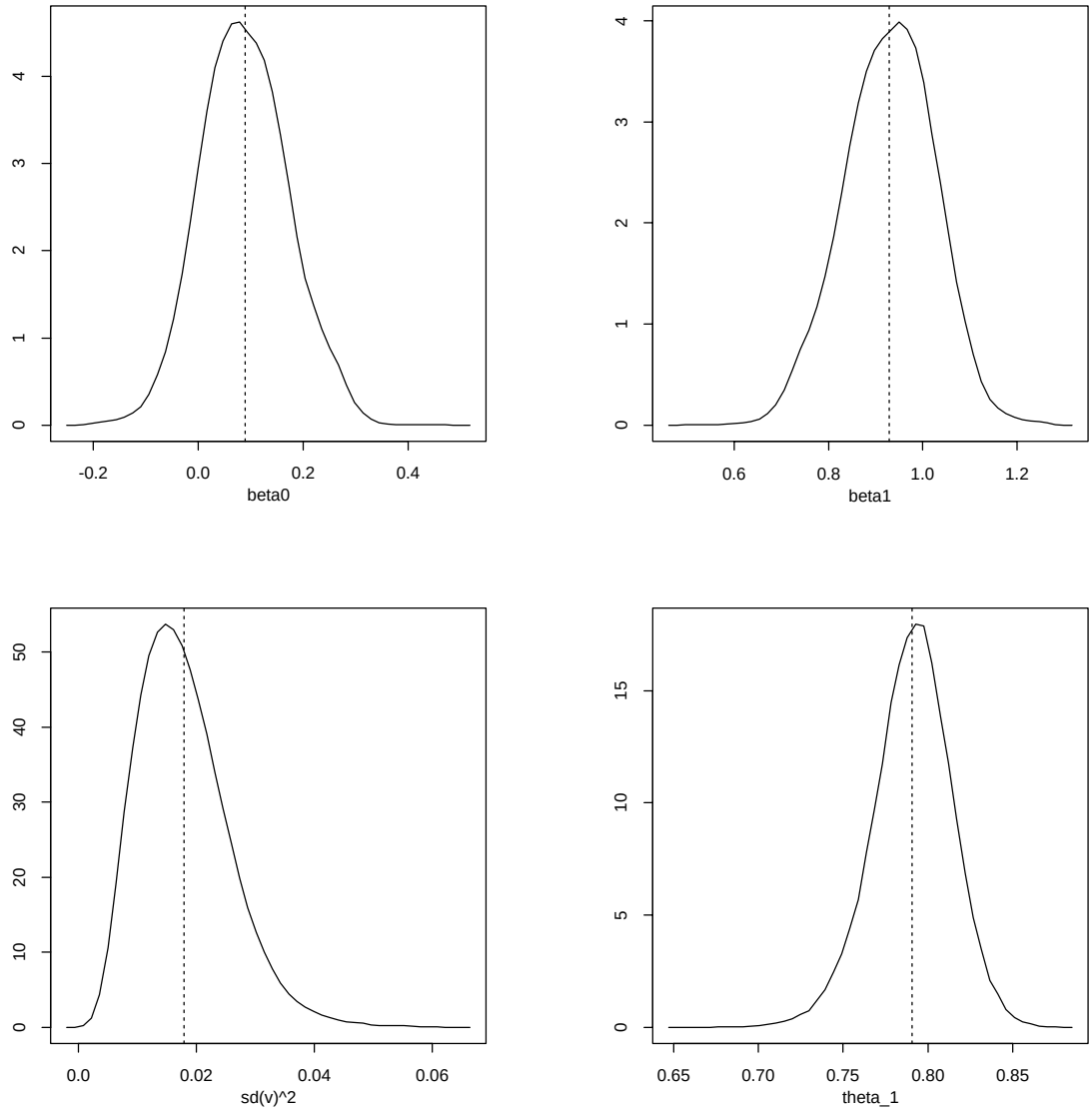
Figur 9: Tetthetsestimater av $\frac{X_i^p}{N_i}$ og $\hat{\theta}_i$ (estimert registersysselsetting)

Figuren viser forskyvningen akkurat som i forrige avsnitt. Vi ser også tydelig at det ikke lengre er synlig "overshrinkage" i $\hat{\theta}_1,\cdots,\hat{\theta}_{432}$. Dette er fordi at kovariaten tillater mye større variasjon i estimatene enn når det kun er "random effects" som gir variasjonen. Vi kan nå betrakte estimatet som betinget på kovariaten også slik at estimatet er gitt ved: $\hat{\theta}_i = E\left(\theta_i \mid X_i^s, X_i^p\right)$. Denne formen forklarer at vi i figur 8 ser at fordelingen til θ_1 (betingede) nå er langt mer forskjellig fra den ubetingede fordelingen til θ_i som er estimert i figur 9, enn det som var tilfellet uten kovariat (figur 4). Ved å beregne fra den tilpassede modellen $\widehat{Var}\left(\log it\left(\hat{\theta}_i\right) - \hat{\beta}_0 - \hat{\beta}_1 \log it\left(X_i^p\right)\right) = 0,00122$ (tabell 3) og se at denne er mye mindre enn $\hat{\sigma}_v^2 = 0,0141$ viser at det fremdeles er "overshrinkage". Å estimere registersysselsettingen ved å bruke seg selv både som observasjoner og som kovariat gir ikke så mye

informasjon annet enn som grunnlag for sammenligning når vi nedenfor skal bruke AKU-utvalget som observasjoner.

Helt tilsvarende som i figur 8 og 9 har vi brukt samme modellen, men byttet ut observasjonene med AKU-utvalget selv ($\frac{Y_i^s}{n_i}$). Resultatene følger nedenfor. Antall iterasjoner er igjen 500 000.

Akseptraten for parametrene er nå: $AR_{\beta_0} = 52,2\%$, $AR_{\beta_1} = 47,9\%$, $AR_{\frac{1}{\sigma^2}} = 37,8\%$. For $\theta_1, \cdots, \theta_{432}$ er det varierende akseptrate og ligger mellom 36,5% og 69,7%. Figuren nedenfor viser marginale aposteriori fordelinger som i figur 8.



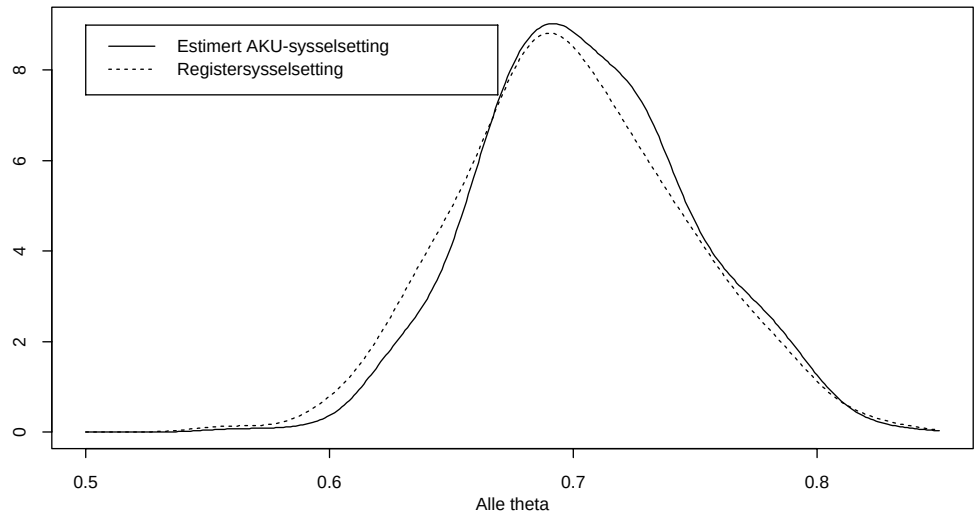
Figur 10: Marginale aposteriori fordelinger for $\beta_0, \beta_1, \frac{1}{\sigma^2}$ og θ_1, med $\frac{Y_i^s}{n_i}$ som observasjoner og $\log\left(\frac{\frac{x_i^p}{N_i}}{1-\frac{x_i^p}{N_i}}\right)$ som kovariat.

Figuren viser stor likhet med figur 8. Tabellen nedenfor viser punktestimater for parametrene og tilnærmede konfidensintervaller for forventningen.

PARAMETER	PUNKTESTIMAT (GJENNOMSNITT)	APPR. 95% KONF. INTERVALL
β_0	0,0893	0,0794_0,0993
β_1	0,9287	0,9172_0,9402
σ_v^2	0,0179	0,0175_0,0184
θ_1	0,7905	0,7896_0,7912

Tabell 4: Parameter estimater og approksimative 95% konfidensintervaller for forventningen

Tabellen viser at i forhold til tabell 3 har vi fremdeles signifikant positiv og litt høyere forskyvning (β_0) mellom kovariat og $\hat{\theta}_1,\cdots,\hat{\theta}_{432}$ (på logit-skala) og en proporsjonalitet (β_1) signifikant mindre enn 1 (så vidt) i stedet for større enn 1 i tabell 3. Som i sted er det likevel kanskje bedre å utelate β_1 fra modellen, pga. at hvis den er forskjellig fra 1 er det vanskelig å tolke den. Det sier nemlig at forskjellen mellom registersysselsetting og tenkt AKU-sysselsetting i populasjonen varierer med størrelsen på sysselsettingen, noe som er rart. Ved å gi modellen en mulighet for å bestemme denne forskjellig fra 1 ut fra våre spesifikke data, er ikke nødvendigvis en bekreftelse på modellens gyldighet. Videre ser vi at varianskomponenten har økt litt i forhold til i tabell 3 (fremdeles er signifikant større enn 0). En slik økning er naturlig siden vi ikke lengre bruker en del av registeret som observasjoner, og det tolker vi igjen som en bekreftelse på "random effects".



Figur 11: Tetthetsestimater av $\frac{X_i^p}{N_i}$ og $\hat{\theta}_i$ som nå er estimert AKU-sysselsetting

I forhold til figur 9 ser vi at forskyvningen mellom $\frac{X_i^p}{N_i}$ og $\hat{\theta}_i$ ikke lengre er til stede og som i figur 9 er det heller ikke her synlig "overshrinkage". Forskyvningen er jo som beskrevet tidligere pga. at AKU-utvalget ligner i gjennomsnittet mer på registeret enn registersysselsettingen til AKU-utvalget.

6. Diskusjon og mulige utvidelser

Figur 11 viser at $\hat{\theta}_i$ ligner mye på $\frac{X_i^p}{N_i}$. Når man bruker AKU-utvalget som observasjoner (som man antar er helt riktige) og registersysselsetting som kovariat får man estimert sysselsetting som ligner mye på selve registeret. Dette sier noe om at kvaliteten på registeret er bra, uten at vi har klart å kvantifisere dette, og heller ikke klart å si så mye om kvaliteten til $\hat{\theta}_1, \cdots, \hat{\theta}_{432}$.

Vi kan likevel få en indikasjon på om man skal bruke estimatene fra vår varianskomponentmodell eller utnytte informasjonen fra registeret mer direkte ved å sammenligne vår "random effect" modell med en tilsvarende "fixed effect" modell, som f.eks. kan bestå i akkurat den samme modellen, men uten støyleddet v_i (en GLM modell). Hvis vi for illustrasjonens skyld lager en slik modell og bruker den tilpassede modellen fra tabell 4 (uten σ_v^2), så vil vårt estimat fra denne "fixed effect" modellen kunne skrives: $\log it\left(\hat{\xi}_i\right) = \hat{\beta}_0 + \hat{\beta}_1 \log it\left(X_i^p\right)$. Dette estimatet må nødvendigvis ha en skjevhet for mange kommuner, siden vi vet med sikkerhet at denne deterministiske sammenhengen ikke er eksakt. Derimot slipper vi varians siden støyleddet er borte (ser for enkelthets skyld bort fra variansen i $\hat{\beta}_0$ og $\hat{\beta}_1$ siden denne uansett er veldig liten). Dette estimatet kan vi nå sammenligne med vårt betingede forventningsestimat gitt ved $\hat{\theta}_i = E\left(\theta_i \mid Y_i^s, X_i^p\right)$ som under modellen er forventningsrett, men som på sin side har en varians. Siden de marginale aposteriorifordelingene for $\theta_1, \cdots, \theta_{432}$ som vi har generert med vår Markovkjede nettopp er de betingede fordelingene til $\theta_i \mid Y_i^s, X_i^p$ kan vi estimere denne variansen.

Dermed kan vi sammenligne approksimert Mean Square Errror (MSE) til de to forskjellige estimatene for alle kommuner. Hvis vi antar at vårt betingede forventningsestimat er riktig blir $\widehat{MSE}\left(\hat{\xi}_i\right) = \left(\hat{\theta}_i - \hat{\xi}_i\right)^2$, mens $\widehat{MSE}\left(\hat{\theta}_i\right) = \widehat{MSE}\left(\theta_i \mid Y_i^s, X_i^p\right) = \widehat{Var}\left(\theta_i \mid Y_i^s, X_i^p\right)$. Vi har gjort dette for θ_1 som er presentert tidligere (utvalget består av kun en person) og θ_{432} som er for en stor kommune med mye mindre varians i estimatet $\frac{1}{n_i} Y_i^s$ fra utvalget , og får at: $\widehat{MSE}\left(\hat{\xi}_1\right) = 2{,}16 \times 10^{-8}$, mens $\widehat{MSE}\left(\hat{\theta}_1\right) = 5{,}5 \times 10^{-4}$ som viser at for kommune nr. 1 gir ikke varianskomponentmodellen noen gevinst pga. at observasjonene er for dårlig. Derimot for kommune nr. 432 er det motsatt: $\widehat{MSE}\left(\hat{\xi}_{432}\right) = 1{,}54 \times 10^{-3}$ mens $\widehat{MSE}\left(\hat{\theta}_{432}\right) = 8{,}78 \times 10^{-5}$, så her er observasjonene gode nok til at varianskomponentmodellen er best.

En naturlig fortsettelse av dette arbeidet, er å gjøre det samme for ledighetsandelene som vi har gjort for sysselsetting. Deretter kan man utvide modellen til en multivariat modell som tar for seg sysselsetting og ledighet samtidig. Dette kan gi en bedre modell fordi den da tar hensyn til korrelasjonen (negativ) som man vet eksisterer mellom ledighet og sysselsetting. Man kan kontrollere om den multivariate modellen er bedre ved å sammenligne med de to univariate modellene, og bruke $\frac{X_i^s}{n_i}$ som observasjoner siden vi da har fasiten. En slik multivariat modell kan konstrueres ved å erstatte den binomiske fordelingen i observasjonene med en multinomisk modell. Da er observasjonene en vektor med tre komponenter $\left(Y_{l,i}, Y_{s,i}, Y_{u,i}\right)'$ som betegner antall ledige, antall sysselsatte og antall utenfor i kommune nr. i. Hvis vi lar γ_i betegne ledighetsandelen får vi den multinomiske tettheten

$\frac{n_i!}{y_{1,i}!y_{s,i}!y_{u,i}!}\gamma_i^{y_{1,i}}\theta_i^{y_{s,i}}\left(1-\gamma_i-\theta_i\right)^{y_{u,i}}$ som erstatter den binomiske alle steder der denne inngår. Man får 432 flere parametre i modellen.

En annen utvidelse av modellen er å prøve å utnytte alle dataene når man skal estimere. Altså bruke $\frac{X_i^p}{N_i}$ som kovariat, utnytte at man har $\frac{X_i^s}{n_i}$ og så bruke $\frac{Y_i^s}{n_i}$ som observasjoner (som før). Det er imidlertid litt usikkert hvordan dette skal gjøres best. Det er f.eks. ikke naturlig å bruke $\frac{X_i^s}{n_i}$ som en ekstra kovariat, da det er intuitivt lite rimelig å la registerinformasjon på utvalgsnivå være forklaring til AKU-andeler på populasjonsnivå.

Modellene som er behandlet her, både i analysen og i forslag til utvidelser er for kompliserte til å regne på, uten bruk av simulering. Da er MCMC metoder et alternativ. Vi har sett at de også takler kompliserende faktorer som høyt antall parametre og tilfeller med degenererte utvalg.

7. Referanser

[1]: Frigessi, Arnoldo
Forelesningsnotater i kurset ST397, høsten 2000

[2]: "Simultaneous estimation of the mean of a binary variable from a large number of small areas", Zhang, Li-Chun,
Discussion Paper SSB, høsten 2000

[3]: "Continuous univariate distributions, 1-2", Johnson, N., Kotz, S.
Wiley, 1970

8. Appendix - Splus kode

8.1. For modell uten tilleggsinformasjon (avsnitt 3.1)

```
# Gibbs sampler oppdatering for theta og Metropolis Hastings oppdatering for nu og omega

    id <- s.a.a == 0 | s.a.a == 1
    theta.0 <- s.a.a
    theta.0[id] <- mean(s.a.r)
    n.i <- n.aku
    ant.syss <- syss.aku
    nu.0 <- 10
    omega.0 <- 5

#Parametrene til Gamma priorene

    a1 <- 0.01; b1 <- 1/a1
```

```
c1 <- 0.01; d1 <- 1/c1
m <- 500000
theta.g <- theta.0
nu.g <- nu.0
omega.g <- omega.0
stor.n <- length(theta.g)
simsd.nu <- 1.4
simsd.omega <- 0.7
ar.nu <- 0
ar.omega <- 0
thetasum <- theta.0

write(c(nu.g,omega.g,theta.g),file="ut0601.dta",ncolumns=stor.n+2,append=F)

for (j in 1:m) {

#Oppdater alle theta-komponentene fra betingede fordelingen

st.n <- length(theta.g)
theta.ny <- rbeta(st.n,(ant.syss+nu.g),(n.i-ant.syss+omega.g))

#Oppdater nu

nu.x <- rnorm(20,nu.g,simsd.nu)
nu.ny <- nu.x[nu.x>0][1]

log.t.nu <- sum(log(dbeta(theta.ny,nu.ny,omega.g)))+(a1-1)*log(nu.ny)-(nu.ny/b1)+log((1-
pnorm(-nu.g/simsd.nu)))

log.n.nu <- sum(log(dbeta(theta.ny,nu.g,omega.g)))+(a1-1)*log(nu.g)-(nu.g/b1)+log((1-
pnorm(-nu.ny/simsd.nu)))

broek.nu <- exp(log.t.nu - log.n.nu)
a.nu <- min(1,broek.nu)
nu.valg <- c(nu.g,nu.ny)
nu.ny <- sample(nu.valg,size=1,prob=c((1-a.nu),a.nu))
if (nu.ny != nu.g) ar.nu <- ar.nu+1

#Oppdater omega

omega.x <- rnorm(20,omega.g,simsd.omega)
omega.ny <- omega.x[omega.x>0][1]

log.t.omega <- sum(log(dbeta(theta.ny,nu.ny,omega.ny)))+(c1-1)*log(omega.ny)-
(omega.ny/d1)+log((1-pnorm(-omega.g/simsd.omega)))

log.n.omega <- sum(log(dbeta(theta.ny,nu.ny,omega.g)))+(c1-1)*log(omega.g)-
(omega.g/d1)+log((1-pnorm(-omega.ny/simsd.omega)))

broek.omega <- exp(log.t.omega - log.n.omega)
a.omega <- min(1,broek.omega)
omega.valg <- c(omega.g,omega.ny)
omega.ny <- sample(omega.valg,size=1,prob=c((1-a.omega),a.omega))
if (omega.ny != omega.g) ar.omega <- ar.omega+1
```

```
        if (j/10 == trunc(j/10))
write(c(nu.ny,omega.ny,theta.ny),file="ut0601.dta",ncolumns=stor.n+2,append=T)
        theta.g <- theta.ny
        nu.g <- nu.ny
        omega.g <- omega.ny
        if (j/2000 == trunc(j/2000)) cat("j=",j,"\n")
        if (j > 100000) thetasum <- thetasum + theta.ny

        }
```

8.2. For modell med tilleggsinformasjon (avsnitt 3.2)

```
#Med kovariat informasjon (registerinformasjon)

        id <- s.a.ar == 0 | s.a.ar == 1

        theta.0 <- s.a.ar
        theta.0[id] <- mean(s.a.r)
        n.i <- n.akureg
        ant.syss <- syss.akureg

        x.i <- log(s.a.r/(1-s.a.r))

        beta0.0 <- 1
        beta1.0 <- 1

        sd2inv.0 <- 4.5


#Parametrene til Gamma priorene

        a1 <- 0.01; b1 <- 1/a1
        c1 <- 0.01; d1 <- 1/c1

        m <- 500000

        .Random.seed <- mitt.seed

        theta.g <- theta.0
        beta0.g <- beta0.0
        beta1.g <- beta1.0
        sd2inv.g <- sd2inv.0

        stor.n <- length(theta.g)

        simsd.theta <- 0.027
        simsd.beta0 <- 0.01
        simsd.beta1 <- 0.013
        simsd.sd2inv <- 17


        ar.beta0 <- 0
```

```
ar.beta1 <- 0
ar.sd2inv <- 0
ar.theta <- rep(0,stor.n)

write(c(beta0.g,beta1.g,sd2inv.g,theta.g),file="ut100101.dta",ncolumns=stor.n+3,append=F)

theta.ny <- theta.g
sd2inv.ny <- sd2inv.g
thetasum <- rep(0,stor.n)

for (j in 1:m) {


#Oppdater alle theta-komponentene individuelt, men på vektorform

th.ny <- rnorm(stor.n,0,simsd.theta)+theta.g
id1 <- th.ny > 0 & th.ny < 1

a.th <- rep(0,stor.n)
log.t.th <- rep(0,stor.n)
log.n.th <- rep(0,stor.n)

log.t.th[id1] <- ant.syss[id1]*log(th.ny[id1])+(n.i[id1]-ant.syss[id1])*log(1-th.ny[id1])-
0.5*sd2inv.g*((log(th.ny[id1]/(1-th.ny[id1]))-beta0.g-beta1.g*x.i[id1])^2)+log(theta.g[id1])+log(1-
theta.g[id1])

log.n.th[id1] <- ant.syss[id1]*log(theta.g[id1])+(n.i[id1]-ant.syss[id1])*log(1-theta.g[id1])-
0.5*sd2inv.g*((log(theta.g[id1]/(1-theta.g[id1]))-beta0.g-beta1.g*x.i[id1])^2)+log(th.ny[id1])+log(1-
th.ny[id1])

broek.th <- exp(log.t.th - log.n.th)
en.v <- rep(1,stor.n)
id2 <- broek.th < en.v
a.th <- en.v
a.th[id2] <- broek.th[id2]
u.v <- runif(stor.n)
id3 <- u.v < a.th
a.id <- id1 & id3
theta.ny[a.id] <- th.ny[a.id]
ar.theta <- ifelse (theta.ny != theta.g, ar.theta+1, ar.theta)


#Oppdater beta0

beta0.ny <- rnorm(1,beta0.g,simsd.beta0)

logdiff <- 0.5*sd2inv.g*(sum((log(theta.ny/(1-theta.ny))-beta0.g-beta1.g*x.i)^2)-
sum((log(theta.ny/(1-theta.ny))-beta0.ny-beta1.g*x.i)^2))

broek.beta0 <- exp(logdiff)
a.beta0 <- min(1,broek.beta0)
beta0.valg <- c(beta0.g,beta0.ny)
beta0.ny <- sample(beta0.valg,size=1,prob=c((1-a.beta0),a.beta0))
if (beta0.ny != beta0.g) ar.beta0 <- ar.beta0+1
```

```
#Oppdater beta1

        beta1.ny <- rnorm(1,beta1.g,simsd.beta1)

        logdiff2 <- 0.5*sd2inv.g*(sum((log(theta.ny/(1-theta.ny))-beta0.ny-beta1.g*x.i)^2)-
sum((log(theta.ny/(1-theta.ny))-beta0.ny-beta1.ny*x.i)^2))

        broek.beta1 <- exp(logdiff2)
        a.beta1 <- min(1,broek.beta1)
        beta1.valg <- c(beta1.g,beta1.ny)
        beta1.ny <- sample(beta1.valg,size=1,prob=c((1-a.beta1),a.beta1))
        if (beta1.ny != beta1.g) ar.beta1 <- ar.beta1+1

#Oppdater sd2inv

        sd2inv.x <- rnorm(1,sd2inv.g,simsd.sd2inv)
        if (sd2inv.x > 0) {
        sd2inv.ny <- sd2inv.x

        logdiff <- ((stor.n/2)+(a1-1))*(log(sd2inv.ny)-log(sd2inv.g))+(0.5*sum((log(theta.ny/(1-
theta.ny))-beta0.ny-beta1.ny*x.i)^2)+(1/b1))*(sd2inv.g-sd2inv.ny)
        broek.sd2inv <- exp(logdiff)
        a.sd2inv <- min(1,broek.sd2inv)
        sd2inv.valg <- c(sd2inv.g,sd2inv.ny)
        sd2inv.ny <- sample(sd2inv.valg,size=1,prob=c((1-a.sd2inv),a.sd2inv))
        if (sd2inv.ny != sd2inv.g) ar.sd2inv <- ar.sd2inv+1

        }

        if (j/10 == trunc(j/10))
write(c(beta0.ny,beta1.ny,sd2inv.ny,theta.ny),file="ut100101.dta",ncolumns=stor.n+3,append=T)

        theta.g <- theta.ny
        beta0.g <- beta0.ny
        beta1.g <- beta1.ny
        sd2inv.g <- sd2inv.ny

        if (j/2000 == trunc(j/2000)) cat("j=",j,"\n")

        if (j > 100000) thetasum <- thetasum + theta.ny

        }
```

De sist utgitte publikasjonene i serien Notater

2000/64 R. N. Johnsen: Undersøking om foreldrebetaling i barnehagar, august 2000. 36s.

2000/65 A. Thomassen: Byggekostnadsindeks for rørleggerarbeid i kontor- og forretningsbygg. 14s.

2000/67 A.G. Hustoft og G. Olsen: Metadata for statistikk om personer og husholdninger : Forprosjektrapport. 34s.

2000/68 A. Bruvoll, K. Flugsrud og H. Medin: Dekomponering av endringer i utslipp til luft i Norge - dokumentasjon av data. 19s.

2000/69 M. Vik Dysterud og E. Engelien: Tettstedsavgrensing: Teknisk dokumentasjon 2000. 53s.

2000/70 A. Akselsen, G. Dahl, J. Lajord og Ø. Sivertstøl: FD - Trygd: Variabelliste. 48s.

2000/71 B.O. Lagerstrøm: Kompetanse i grunnskolen , del 2: Dokumentasjonsrapport. 19s.

2000/72 B.O. Lagerstrøm: Kompetanse i grunnskolen: Hovdresultater 1999/2000 170s.

2000/73 J.H. Wang: Kvartalsvis investeringsstatistikk. 57s.

2000/74 P.O. Lande og T. Hoel: Dødsårsaksregisteret: Systemdokumentasjon. 90s.

2000/75 A.G. Pedersen, P.O. Lande og T. Hoel: Dødsårsaksregisteret: Brukerdokumentasjon. 99s.

2000/76 A.G. Hustoft, B. Vannebo: En undersøkelse av frafallet i utvalgsundersøkelser i perioden 1997-2000. 56s.

2000/77 P.O. Lande og J. Kittelsen: Forbruksundersøkinga 2000. Innlasting/Innsjekking: Brukardokumentasjon. 17s.

2000/78 J. Fosen, A.K. Johnsen og G. Røyne: Frafall blant innvandrere. En undersøkelse av frafall i Utdanningsundersøkelsen 1999 og i valgundersøkelser blant innvandrere. 53s.

2000/79 J. Kittelsen og P.O. Lande: OPPSLAG - Forbruksundersøkelsen. Brukerdokumentasjon. 39s.

2000/80 J. Kittelsen og P. O. Lande: Forbruksundersøkinga 2000. Systemdokumentasjon. 156s.

2000/81 J.T. Lind: Testing av stokastiske individuelle effekter i paneldatamodeller. 17s.

2001/2 D.Q. Pham: Innføring i tidsserier - sesongjustering og X-12-AMIRA. 110s.

2001/3 O. Rognstad: Eiendomsomsetning. Dokumentasjon av datagrunnlag og bearbeidingsrutine. 72s.

2001/4 T. Nøtnæs: Innføring i kognitiv kartlegging. 20s.

2001/5 T. Bye, M. Hansen og B. Strøm: Hvordan framskrive utslipp av klimagasser? 16s.

2001/6 A. Langørgen og R. Aaberge: KOMMODE II estimert på data for 1998. 16s.

2001/7 B.R. Joneid og J. Lajord: FD - Trygd: Dokumentasjonsrapport. Stønader til enslig forsørger. 1992-1999. 39s.

2001/8 T. Karlsen, E. Karstensen og E. Evensen: Beregningsrutiner og teknisk programstruktur for fylkesfordelt nasjonalregnskap. 27s.

2001/9 L. Rognstad, N.M. Stølen, T. Jakobsen og P. Schøning: Regional statistikk og analyse - strategi og prioriteringer. 45s.

2001/10 A. Akselsen og B.R. Joneid: FD - Trygd: Dokumentasjonsrapport. Pensjoner. Grunn- og hjelpestønader. 1992-1998. 94s.

2001/11 B. Mathisen: Flyktninger og arbeidsmarkedet 4. kvartal 1999. 34s.

2001/12 A. Rognan og N. Barrabés: NUS2000. Dokumentasjonsrapport. 36s.

2001/13 K.I. Bøe, J. Johansen og Ø. Sivertstøl: FD - Trygd: Dokumentasjonsrapport. Attføringspenger, 1992-1998. 88s.