Statistics Norway
Statistisk sentralbyrå

Li-Chun Zhang

Domene-estimering i lønnsstatistikk

Seksjon for statistiske metoder og standarder
Emnegruppe: 00.90

Forord

Dette notat handler om noen metoder for domene estimering basert på utvalgsundersøkelser. Generelt har denne problemstilling i den senere tid blitt stadig mer aktuell i offisiell statistikk produksjon. Mer konkret stammer dette notat fra et samarbeid med seksjon 420 v/Oddbjørn Haugen, der man har mottatt forespørsel fra brukere om mer detaljert lønnsstatistikk. Metodene er beskrevet i den generelle form, og algoritmene for beregning angitt i appendiksene, slik at de lettere kan taes i bruk av andre statistikere som arbeider lignende problemer.

1 Innledning

Det har alltid eksistert et ønske om å lage statistikk for små undergrupper i populasjonen, dvs. domener, i offisiell statistikk produksjon. For eksempel kan slike domener svare til mindre geografiske områder som kommuner. Eller de kan også defineres ut ifra demografiske variabler som alder, kjønn, sivilstatus, etnisk tilhørlighet, osv. Tidligere var slike domene statistikk utelukkende basert på fulltellinger. Et godt utbygget registersystem gir oss en annen mulighet. Et tredje alternativ er domene estimering basert på utvalgsundersøkelser, med eller uten tilleggsopplysinger fra registre). Dette notat handler om noen estimeringsmetoder i det siste tilfellet.

Det er sterke tradisjoner i utvalgsundersøkelser om å basere estimering på vekting. Hovedproblemet med direkte vekting for domene estimering ligger i at antallet domener er som regel så stort at det finnes kun få observasjoner i de fleste domener, noe som gjør at den direkte estimator kan være veldig usikker der. Man kan ofte forbedre direkte vekting ved å ta i bruk modeller som gjør det mulig å bruke data på tvers av domener. Siden estimater da avhenger av data fra alle domener, betegnes de som indirekte i motsetning til direkte vekting som kun bruker data fra en bestemt domene. I dette notat skal vi konsentrere oss om en enkel lineær modell som kan ta seg av variasjoner både mellom- og innenfor domenene.

Det er nødvendig å ha noen oversiktige mål når man sammenligner alternative setter av domene estimater. Generelt kan man evaluere domene estimater i to forskjellige retninger. Den ene handler om hvor godt estimatene er for hver bestemt domene. Den andre handler om hvor godt estimatene fanger opp de fordelingsmessige egenskaper av alle domenene. Et eksempel på den siste er varians av domenegjennomsnittene i populasjonen. Et annet eksempel er variasjonsbredde i domenegjennomsnittene, dvs. forskjell mellom det største og det minste domenegjennomsnittet. Det bør bemerkes at slike fordelingsmessige karakteristikk av domenene ofte er av minst like stor interesse som domene spesifikk gjennomsnittene eller totalene i domene statistikk sammenheng.

Vi beskriver 3 domene estimatorer i avsnitt 2, nemlig

- direkte estimering,
- "empirical best linear unbiased predictor (EBLUP) under the nested-error regression model",
- simultan domene estimering basert på EBLUP.

Vi evaluerer dem mot hverandre i avsnitt 3. De empiriske data'ene er hentet fra lønnsstatistikk vha. Oddbjørn Haugen ved seksjon 420, der man vurderer å produsere domene estimater etter forespørsel fra brukere av lønnsstatistikk. Hovedkonklusjonene er følgende:

1. Modellbaserte domene estimatene har generelt mindre total usikkerhet enn estimatene basert på direkte vekting. De er også mer robuste i den forstand at de verste modellbaserte estimater typisk feiler mye mindre enn de verste direkte estimater.

2. De simultane domene estimater fanger bedre opp fordelingen til domenegjennomsnittene enn den EBLUP, mot en liten økning i den domene spesifikk feil. De representerer derfor et bedre all-round alternativ i tilfellet hensynet til domene spesifikk prediksjon ikke er utelukkende avgjørende metodevalget.

2 Metoder for domene estimering

2.1 Direkte vekting

La $s = \{1, 2, ..., n\}$ betegne det aktuelle utvalg. For hver enhet i utvalget, betegnet med $j \in s$, la y_j være verdien til interessevariabelen y av undersøkelsen. Videre la w_j betegne vekten av enhet j, som i utgangspunktet er lik den inverse trekksannsynlighet til j. Ofte er w_j i tillegg justert vha. etterstratifisering, kalibrering, rateestimering, osv. Uansett er w_j typisk beregnet for estimering på aggregerte nivåer. For eksempel er totalen og gjennomsnittet i populasjonen hhv. estimert ved

$$\hat{Y} = \sum_{j \in s} w_j y_j \qquad \text{and} \qquad \hat{\bar{Y}} = (\sum_{j \in s} w_j y_j)/(\sum_{j \in s} w_j).$$

La $i = 1, ..., m$ betegne de aktuelle domener i populasjonen. La y_{ij} være y-verdien til enhet j fra domene i, betegnet med $j \in s_i$ og $s_i = \{1, ..., n_i\}$ der n_i er antallet utvalgte enheter fra domene i. La w_{ij} betegne vekten til $j \in s_i$. Den direkte domene estimator for domene-totalen og -gjennomsnittet til domene i er da gitt som

$$\hat{Y}_i = \sum_{j \in s_i} w_{ij} y_{ij} \qquad \text{and} \qquad \hat{\bar{Y}}_i = (\sum_{j \in s_i} w_{ij} y_{ij})/(\sum_{j \in s_i} w_{ij}).$$

Den direkte estimator $\hat{Y}_i$ er forventningsrett for domenetotalen, slik $\hat{Y}$ er for populasjonstotalen. I tilfellet w_{ij} er justert mht. tilleggs opplysinger på populasjonsnivået derimot, kan $\hat{Y}_i$ inneholde skjevheter dersom tilleggsvariablene inneholder sterke domene-til-domene variasjoner. I så fall er det nødvendig at man i stedet justerer vektene mht. domene spesifikk tilleggs opplysninger. Til slutt estimeres variansen til $\hat{Y}_i$ på samme måte som variansen til $\hat{Y}$.

Som et eksempel på domene estimering basert på direkte vekting kan man nevne fylkestallet i AKU. I utgangspunktet er vekten i AKU justert mht. etterstrata definert på landsnivået. For å gi bedre tall på fylkesnivået, er vekten kalibrert fylkesvis deretter[1]. I tillegg justerer man utvalgsstørrelsene i fylkene slik at, istedenfor å være omtrent proporsjonelle med populasjonsstørrelsene, små fylkene har relativt høyere trekkandeler enn de store.

2.2 EBLUP

2.2.1 Modell

EBLUP betyr empirisk best lineær forventningsrett prediktor. Hvordan den konkret ser ut avhenger av den antatte modell. Den er forventningsrett mht. modellen, og ikke mht. gjentatt utvalgstrekking fra den samme populasjonen etter den samme utvalgsplan. I dette notat konsentrerer vi oss om følgende "nested-error (NE) regression model"

$$y_{ij} = \mu + u_{ij} \qquad \text{and} \qquad u_{ij} = v_i + e_{ij} \tag{1}$$

[1] Zhang, L-.C. (1998). *Dokumentasjonsrapport: Den nye estimeringsmetoden for Arbeidskraftunderskelsen (AKU) — Fylkesvis kalibrering med i landsetterstratifiserte vekter som startverdier.* Notater 98/1, SSB.

der μ er en overall gjennomsnitt, og v_i representerer en tilfeldig domene effekt, mens e_{ij} representerer en tilfeldig individ effekt. Det totale avvik $y_{ij} - \mu$ er på den måte delt i en mellom-domene variasjon (dvs. v_i) og en innen-domene varaisjon (dvs. e_{ij}), siden forskjellen mellom y_{ij} og y_{ik} fra den samme domene er nå gitt som $e_{ij} - e_{ik}$, dvs. uavhengig av domene effekt v_i. Alternativt kan vi skrive

$$y_{ij} = \theta_i + e_{ij} \qquad \text{and} \qquad \theta_i = \mu + v_i$$

der θ_i er det teoretiske gjennomsnitt for domene i, og μ er igjen det teoretiske gjennomsnitt av alle domenegjennomsnitter. Legg merke til at θ_i ikke er det samme som det faktiske gjennomsnitt for domene i, betegnet med $U_i = \{1, 2, ..., N_i\}$, som er gitt som

$$\bar{Y}_i = \sum_{j \in U_i} y_{ij}/N_i = \theta_i + \sum_{j \in U_i} e_{ij}/N_i \neq \theta_i.$$

Videre antar man at v_i og e_{ij} er uavhengige av hverandre og har forventning lik null. La

$$\sigma_v^2 = Var(v_i) \qquad \text{and} \qquad \sigma_e^2 = Var(e_{ij})$$

betegne variansen til v_i og e_{ij}. Legg merke til at σ_e^2 antaes å være konstant fra domene til domene. Til slutt antar man at e_{ij} og e_{ik} fra den samme domene er uavhengige av hverandre.

Modellen overfor er den enklest NE modell. Mer generelt kan en NE modell inkludere tilleggs variabler, betegnet med x_i. Da erstatter man μ i ligning (1) med $x_i^T\beta$, der β representerer faste effekter som kan forklares av x_i. I tillegg er det mulig å utvide antallet tilfeldige effekter, dersom data inneholder variasjon på flere nivåer enn to.

2.2.2 Estimering av parametre

Modell (1) inneholder 3 parametre, nemlig μ, σ_v^2 og σ_e^2. Først la

$$z_{ij} = y_{ij} - \bar{y}_i = e_{ij} - \bar{e}_i \qquad \text{der} \quad \bar{y}_i = \sum_{j \in s_i} y_{ij}/n_i \quad \text{og} \quad \bar{e}_i = \sum_{j \in s_i} e_{ij}/n_i$$

som kun avhenger av individ variasjoner e_{ij}. En forventningsrett estimator for σ_e^2 er nå gitt som

$$\hat{\sigma}_e^2 = \nu_1^{-1}(\sum_{i,j} z_{ij}^2) \qquad \text{der} \quad \nu_1 = \sum_{i;n_i>1} (n_i - 1). \tag{2}$$

Legg merke til at man ser bort fra alle domener med kun 1 oberservasjon, siden $z_{ij} \equiv 0$ når $n_i = 1$. En forventningsrett estimator for σ_v^2 er nå gitt som

$$\hat{\sigma}_v^2 = \eta_1^{-1}\{\sum_{i,j}(y_{ij} - \bar{y})^2 - (n-1)\hat{\sigma}_e^2\} \qquad \text{der} \quad \eta_1 = \sum_i n_i(1 - n_i/n) \tag{3}$$

og $\bar{y} = \sum_{i,j} y_{ij}/n$ er det observerte gjennomsnitt i utvalget. Til slutt estimeres μ med

$$\hat{\mu} = \frac{\sum_i n_i(1-\gamma_i)\bar{y}_i}{\sum_i n_i(1-\gamma_i)} \qquad \text{der} \quad \gamma_i = \frac{\hat{\sigma}_v^2}{\hat{\sigma}_v^2 + \hat{\sigma}_e^2/n_i}.$$

Vi minner om at $\hat{\mu}$ er en estimator for det teoretiske gjennomsnitt av domenegjennomsnittene under modellen. Generelt skal man ikke betrakte den som estimator for det faktiske gjennomsnitt av domenegjennomsnittene i populasjonen.

2.2.3 Prediksjon av domenegjennomsnitt

Det faktiske domenegjennomsnitt kan skrives som

$$\bar{Y}_i = (n_i/N_i)\bar{y}_i + (1 - n_i/N_i)\bar{y}_i^c \qquad \text{der} \quad \bar{y}_i^c = \sum_{i=n_i+1}^{N_i} y_i/(N_i - n_i)$$

er gjennomsnittet av alle enheter fra domene i som ikke er med i utvalget. Det først ledd er observert, slik at vi trenger bare å predikere det siste ledd. Den EBLUP for $\bar{y}_c$ er gitt som

$$\hat{\bar{y}}_i^c = \hat{\theta}_i = \hat{\mu} + \hat{v}_i \qquad \text{der} \quad \hat{v}_i = \gamma_i(\bar{y}_i - \hat{\mu})$$

er den predikerte domene effekt. Den EBLUP for $\bar{Y}_i$ blir derfor

$$\hat{\bar{Y}}_i = f_i\bar{y}_i + (1 - f_i)\hat{\bar{y}}_i^c \qquad \text{der} \quad f_i = n_i/N_i$$

er trekkandelen i domene i. Forskjellen mellom $\bar{y}_i^c$ og $\hat{\bar{Y}}_i$ er derfor ubetydelig kun hvis f_i er ubetydelig. Legg merke til at trekkandelen er langt fra ubetydelig i de fleste bedriftsundersøkelser i SBB. Skillet mellom $\hat{\bar{Y}}_i$ og $\hat{\bar{y}}_i^c$ kan være avgjørende der.

Et spesielt problem gjelder trekkandelen f_i. Legg merke til at f_i her referer til andelen av statistiske enheter som er med i utvalget. Den er ikke alltid den samme som andelen trekkenheter. For eksempel er de to som regel forskjellige når populasjonen av interesse betår av varer, tjenester, transaksjoner, osv. Ta f.eks. lønnsstatistikk som vi skal se på i avsnitt 3. Den statistiske populasjon består av alle heltidsstillinger i løpet av et år. Mens enhetene i trekking er bedrifter. Siden man ikke har oversikt over hvor mange heltidsstillinger som finnes, kjenner man heller ikke til f_i nøyaktig. I slike tilfeller er det nødvendig å estimere f_i.

2.2.4 Estimering av modellbasert bruttovarians

Bruttovariansen, dvs. "mean squared error (MSE)", til EBLUP $\hat{\bar{y}}_i^c$ består av 3 komponenter

$$\text{MSE}(\bar{y}_i^c) = g_{1i}(\psi) + g_{2i}(\psi) + g_{3i}(\psi) \qquad \text{for} \quad \psi = (\sigma_v^2, \sigma_e^2)$$

der

$$g_{1i} = \gamma_i\sigma_e^2/n_i = O(1)$$

skyldes estimeringen av domene effekt v_i, og

$$g_{2i} = (1 - \gamma_i)^2\sigma_e^2/\{\sum_i n_i(1 - \gamma_i)\} = O(1/m)$$

skyldes estimeringen av μ, og

$$g_{3i} = n_i^{-2}(\sigma_v^2 + \sigma_e^2/n_i)^{-3}h(\psi)$$

skyldes estimeringen av varianskomponentene, der

$$h(\psi) = \sigma_e^4 Var_1(\hat{\sigma}_v^2) + \sigma_v^4 Var_1(\hat{\sigma}_e^2) - 2\sigma_v^2\sigma_e^2 Cov_1(\hat{\sigma}_e^2, \hat{\sigma}_v^2)$$

og Var_1 og Cov_1 betyr asymptotisk varians og kovarians. Vi har

$$\begin{aligned} Var_1(\hat{\sigma}_v^2) &= 2\eta_1^{-2}\{\nu_1^{-1}(n-1-\nu_1)(n-1)\sigma_e^2 + \eta_2\sigma_v^4 + 2\eta_1\sigma_e^2\sigma_v^2\} \\ Var_1(\hat{\sigma}_e^2) &= 2\nu_1^{-1}\sigma_e^4 \\ Cov_1(\hat{\sigma}_v^2, \hat{\sigma}_e^2) &= -2\eta_1^{-1}\nu_1^{-1}(n-1-\nu_1)\sigma_e^4 \end{aligned}$$

hvor ν_1 er gitt overfor i (2), og η_1 er gitt overfor i (3), og $\eta_2 = \sum_i n_i^2(1-n_i/n) + (\sum_i n_i^2/n)^2$.

Bruttovariansen overfor er korrekt til den 2. orden, dvs. feilen er på størrelsen $O(1/m^2)$. Et estimat for MSE fåes ved å sette inn $\hat{\psi} = (\hat{\sigma}_v^2, \hat{\sigma}_e^2)$ for ψ. Men dette medfører en skjevhet som er på størrelsen $O(1/m)$. For å få en 2. ordens korrekt MSE estimator, må vi derfor korrigere for denne skjevhet. Den endelige MSE estimator er nå gitt som

$$\widehat{\mathsf{MSE}}(\hat{\bar{y}}_i^c) = g_{1i}(\hat{\psi}) + g_{2i}(\hat{\psi}) + 2g_{3i}(\hat{\psi}).$$

Studier basert på simulering i litteraturen tyder på at g_{1i} typisk utgjør over 90% av hele MSE. Til slutt estimerer vi bruttovariansen til $\hat{\bar{Y}}_i$ som

$$\widehat{\mathsf{MSE}}(\hat{\bar{Y}}_i) = (1-f_i)^2\widehat{\mathsf{MSE}}(\hat{\bar{y}}_i^c) + (1-f_i)\hat{\sigma}_e^2/N_i.$$

2.3 Simultan estimering

Både den direkte estimator og EBLUP retter seg mot domene spesifikk prediksjon. Men minst like ofte er man interessert i fordelingsmessige karakteristikk til domenene. F.eks. kan man være opptatt av å vite hvor mange domener som har et gjennomsnitt som ligger under et visst nivå, eller hvor stor forskjell er mellom det høyest og det lavest domenegjennomsnitt. Det kan vises[2] at direkte estimatene typisk har for stor spredning enn de sanne domenegjennomsnitter i populasjonen, mens EBLUP estimatene typisk har for liten spredning. Her skal vi beskrive en metode av simultan estimering som bedre fanger opp fordelingen til domenegjennomsnitter.

Vi tar utgangspunktet i EBLUP. Anta negligibel trekkandelen f_i, slik at $\hat{\bar{Y}}_i$ er lik $\hat{\theta}_i = \hat{\mu} + \hat{v}_i$. Den empiriske varians til EBLUP estimatene er derfor lik

$$\tau_v^2 = \sum_i (\hat{v}_i - \hat{\bar{v}}_i)^2/(m-1) \qquad \text{der} \quad \hat{\bar{v}} = \sum_i \hat{v}_i/m.$$

Når vi sammenligner τ_v^2 med $\hat{\sigma}_v^2$, som er den estimerte varians til v_i, finner vi ofte at τ_v^2 ligger langt under halvparten av $\hat{\sigma}_v^2$. Dette er da et tegn på at EBLUP resulterer i for liten spredning. Vi kan ganske enkelt justere spredningen opp på følgende måte. La

$$\tilde{v}_i = \hat{\bar{v}} + (\hat{v}_i - \hat{\bar{v}})\hat{\sigma}_v/\tau_v \tag{4}$$

[2] Zhang, L.-C. (2003) Simultaneous estimation of the mean of a binary variable from a large number of small areas. *Journal of Official Statistics*, vol. 19, 253 - 363.

være det simultane estimat for v_i, som gir oss

$$\tilde{\theta}_i = \hat{\mu} + \tilde{v}_i.$$

Det følger av (4) at den empiriske varians til simultane estimatene $\hat{\theta}_i$ nødvendigvis er lik $\hat{\sigma}_v^2$. Samtidig beholder et simultan estimat den samme rang blant alle simultane estimatene som det EBLUP estimat for den samme domene gjør blant alle EBLUP estimatene. Forskjellen mellom EBLUP og simultan estimering er lettest å se når vi antar at alle modell parametre er kjente. I dette tilfellet er $\hat{\theta}_i = \mu + \hat{v}_i$, som kun avhenger av direkte data fra domene i. Mens siden τ_v^2 avhenger av alle EBLUP estimatene og dermed data fra alle domenene, gjør $\tilde{\theta}_i$ også det , noe som er grunnen til navnet simultan estimering.

For å estimere MSE til $\tilde{\theta}_i$, bruker vi en bootstrap metode. Først fastsetter vi μ til den estimerte verdi. Deretter generer vi bootstrap utvalg i domene i som følgende:

1. la $\theta_i^* = \hat{\mu} + \tilde{v}_i^*$, der $\tilde{v}_i^*$ er tilfeldig trukket blant $(\tilde{v}_1, ..., \tilde{v}_m)$,

2. la $y_{ij}^* = \theta_i^* + e_{ij}^*$, der e_{ij}^* er tilfeldig trukket blant $\tilde{e}_{ij}$ for alle (i, j), og

$$\tilde{e}_{ij} = \hat{e}_{ij}\hat{\sigma}_e/\tau_e \qquad \text{der} \quad \hat{e}_{ij} = y_{ij} - \theta_i$$

   og τ_e^2 er den empiriske varians til $\hat{e}_{ij}$.

Etter å gjort det samme i alle domenene, beregner vi $\tilde{\theta}_i^*$ basert på bootstrap utvalget på akkurat den samme måte som $\tilde{\theta}_i$ basert på det opprinnelige utvalg. Gitt negligibel f_i, så er $\tilde{\theta}_i^* - \theta_i^*$ et bootstrap replikat av $\tilde{\theta}_i - \theta_i$. Basert på et visst antall, betegnet med B, slike bootstrap replikater, kan vi approksimere bootstrap MSE til $\tilde{\theta}_i$ med

$$\widehat{\mathrm{MSE}}_{boot}(\tilde{\theta}_i) = \sum_{k=1}^{B} (\tilde{\theta}_i^* - \theta_i^*)^2/B.$$

I tilfellet ikke-negligibel f_i, må vi istedet se på $\tilde{\bar{Y}}_i^* - \bar{Y}_i^*$. Men fremgangsmåten er helt lik.

3 Eksempel: Lønnsstatisktikk

3.1 Problemstilling

Lønnstatistikk er en årlig utvalgsbasert statistikk om lønnsnivå og lønnsendring. Utvalgsenheter er bedrifter, stratifisert mht. antallet sysselsatte. I stratumet av de største bedrifter gjennomføres fulltelling. Trekkandelen er lavest i stratumet av de minste bedrifter. Den statistiske enhet er alle heltidsstillinger. Utvalget dekker litt under 60% av alle heltidsstillinger sammenlignet med anslaget i Nasjonalregnskapet. I statistikken er man først og fremst interessert i det gjennomsnittlige månedslønn etter kjennemerke næring, yrke, kjønn, utdanning osv.

Selv om det finnes en del opplysinger om lønnstrekk fra LTO, er det vanskelig å bruke disse som tilleggs informasjon av to grunner:

- vi kjenner ikke til enhetene i den statistiske populasjon unntatt de som er med i utvalget;

- vi vet ikke om et lønnstrekk i LTO gjelder en heltidsstilling eller ikke, heller ikke varigheten til stillingen det gjelder i tilfellet den ikke varer for hele året.

Mangelen på identifikasjon av den statistiske populasjon skaper også problemer for bruk av andre tilgjengelige personlige opplysinger som finnes i registersystemet. Vi har derfor sett bort fra alle mulige tilleggsvariabler i dette studie. Som det empiriske datagrunnlag bruker vi delutvalget av alle som har yrkeskode 5 i næring 52 i år 2000, 2001, og 2002. Vi setter opp det gjennomsnittlige månedslønn i alle kommuner som er representert i utvalget som målet for estimering.

3.2 Domene estimering

Først ser vi på estimater for modell parametre i år 2000 til 2002 (Tabell 3.2). Domenene her er men eller kvinner som arbeider heltid i næring 52 og har yrkeskode 5 i alle utvalgte kommuner. Vi ser at estimatet for det teoretiske gjennomsnitt μ er som forventet stigende fra 2000 til 2002 for både menn og kvinner. Mens de estimerte varianskomponentene varierer en del fra år til år. Spesielt er estimatet for σ_v^2 kjent for å ha en relativt stor varians, noe vi kommer tilbake til i simuleringen etterpå. En mulig måte å redusere MSE på kan derfor være å plugge inn en forhåndsbestemt σ_v^2. Vi skal imidlertid ikke undersøke mer denne mulighet her.

Tabell 1. Modell parametre i år 2000 - 2002.

	Menn: næring 52 & yrke 5				
År	Antall observasjoner	Antall kommuner	σ_e^2	σ_v^2	$\hat{\mu}$
2002	6062	316	4601.5	573.4	20954
2001	6353	305	4589.3	780.1	19768
2000	5444	303	3916.6	623.9	19318
	Kvinner: næring 52 & yrke 5				
År	Antall observasjoner	Antall kommuner	σ_e^2	σ_v^2	$\hat{\mu}$
2002	10424	365	3494.1	359.6	19318
2001	10659	269	4556.9	1025.0	18475
2000	9025	366	2949.2	503.9	17552

Tabell 3.2 oppsummerer den estimerte MSE for det teoretiske domenegjennomsnitt θ_i med direkte estimator og EBLUP. Her beregner vi den relative rot MSE i forholdet til $\hat{\mu}$, nemlig $\sqrt{\widehat{\mathrm{MSE}}(\hat{\theta}_i)}/\hat{\mu}$, i alle domenene. EBLUP reduserer betydelig usikkerheten i domeneestimatene sammenlignet med den direkte estimator. I tillegg har MSE til EBLUP en mye mindre variasjon fra domene til domene. Spesielt har den en mye mindre maximum MSE enn den direkte estimator. På denne måte er EBLUP en mye mer robust metode.

3.3 Simulering

For å befeste funnene overfor kjører vi her en simulering basert på data i år 2001. Vi tar de observerte domenegjennomsnitter som de sanne teoretiske domenegjennomsnitt θ_i. Vi setter det observerte gjennomsnitt til θ_i som modell parameter μ, og den observerte varians til θ_i som σ_v^2. Til slutt fastsetter vi den observerte varians til $e_{ij} = y_{ij} - \bar{y}_i$ som modell parameter σ_e^2.

Vi generer nå data under den antatte modell på følgende måte:

Tabell 2. Estimert relativ rot MSE i år 2000 - 2002, i prosent av domenegjennomsnitt

	Direkte estimator for menn i næring 52 og yrke 5					
År	Minimum	1. Kvartil	Median	Gjennomsnitt	3. Kvartil	Maximum
2002	0.7	6.1	9.8	11.5	15.5	22.0
2001	0.7	6.2	10.4	11.9	16.4	23.2
2000	0.7	5.9	9.1	10.8	14.3	20.3
	EBLUP for menn i næring 52 og yrke 5					
År	Minimum	1. Kvartil	Median	Gjennomsnitt	3. Kvartil	Maximum
2002	0.7	2.6	2.7	2.6	2.7	2.8
2001	0.7	3.4	3.8	3.5	3.9	3.9
2000	0.7	2.9	3.1	3.0	3.2	3.2
	Direkte estimator for kvinner i næring 52 og yrke 5					
År	Minimum	1. Kvartil	Median	Gjennomsnitt	3. Kvartil	Maximum
2002	0.4	4.0	6.8	7.6	10.4	18.1
2001	0.6	5.5	10.1	11.1	14.2	24.7
2000	0.4	3.9	6.9	7.7	9.7	16.8
	EBLUP for kvinner i næring 52 og yrke 5					
År	Minimum	1. Kvartil	Median	Gjennomsnitt	3. Kvartil	Maximum
2002	0.4	1.7	1.8	1.8	1.9	1.9
2001	0.6	4.0	4.9	4.5	5.2	5.5
2000	0.4	2.4	2.7	2.5	2.8	2.9

1. trekk θ_i^*, for $i = 1, ..., m$, tilfeldig og med tilbakelegging fra $\{\theta_i; i = 1, ..., m\}$;

2. trekk e_{ij}^*, for $i = 1, ..., m$ og $j = 1, ..., n_i$, tilfeldig og med tilbakelegging fra alle $\{e_{ij}; i = 1, ..., m \cap j = 1, ..., n_i\}$, og sett $y_{ij}^* = \theta_i^* + e_{ij}^*$.

Basert på $\{y_{ij}^*; i = 1, ..., m \cap j = 1, ..., n_i\}$, estimerer vi θ_i som basert på et vanlig utvalg. La $\hat{\theta}_i^*$ betegne det estimerte teoretiske domenegjennomsnitt. Differansen $\hat{\theta}_i^* - \theta_i^*$ er da feilen av estimeringen på det simulerte utvalg. Ved å gjenta simuleringen mange ganger, kan vi evaluere usikkerheten til en bestemt estimator under den antatte modell.

Tabell 3. Estimering av modell parametre under simulering

	Menn			Kvinner		
Modell parameter	μ	σ_e	σ_v	μ	σ_e	σ_v
Antatt verdi	19852	4500	2679	18287	4493	1515
Gjennomsnitt til estimator	19856	4475	2592	18285	4474	1425
Relativt standard avvik (%)	1.0	2.6	27.1	0.7	8.2	28.4

I vårt tilfelle har vi kjørt denne simulering 1000 ganger. I Tabell 3.3 sammenligner vi de antatte modell parametre med gjennomsnittet til de 1000 setter av estimatene for disse parametre. Estimeringen av μ og σ_e kan sies å være forventningsrett. Mens $\hat{\sigma}_v$ ser ut til å ha en liten negativ skjevhet. Legg merke til at estimatorene for varianskomponenter, spesielt σ_v^2, er mye mer usikker enn $\hat{\mu}$. I tillegg noterer vi at gjennomsnittet til det empiriske standardavvik til direkte estimatene over de 1000 simuleringer er 3755 for menn og 2724 for kvinner. Mens det tilsvarende gjennomsnitt for EBLUP er 2018 for menn og 972 for kvinner. Mao. har direkte estimatene for stor spredning i

forholdet til spredningen i domenegjennomsnittene i populasjonen, mens EBLUP estimatene har for liten spredning.

Nest sammenligner vi den direkte estimator, den EBLUP og den simultane estimator for prediksjon av domenegjennomsnittet. Tabell 3.3 viser forventet gjennomsnittlig og maximum, *absolutt relativ feil (ARE)*, dvs. hhv.

$$m^{-1}\sum_{i=1}^{m}|\hat{\theta}_i^*/\theta_i^*-1| \qquad \text{og} \qquad \max_{i=1,\ldots,m}|\hat{\theta}_i^*/\theta_i^*-1|.$$

Deretter ser vi på hvor godt de 3 estimatorene fanger opp fordelingen til θ_i. Tabell 3.3 viser forventet gjennomsnittlig *absolutt relativ fordelings feil (ARDE)*, dvs.

$$m^{-1}\sum_{i=1}^{m}|\hat{\theta}_{(i)}^*/\theta_{(i)}^*-1|$$

der $\theta_{(i)}^*$ er den $i$te minst verdi blant alle θ_i^*, mens $\hat{\theta}_{(i)}^*$ er den $i$te minst verdi blant alle $\hat{\theta}_i^*$. I tillegg viser den forventet *relativ feil (RE)* i variasjonsbredde, dvs.

$$\{\max_i(\hat{\theta}_i^*)-\min_i(\hat{\theta}_i^*)\}/\{\max_i(\theta_i^*)-\min_i(\theta_i^*)\}-1.$$

Tabell 4. Simulerings resultater for direkte, EBLUP og simultan estimator

	Menn			Kvinner		
Forventning	Direkte	EBLUP	Simultan	Direkte	EBLUP	Simultan
Gjennomsnittlig ARE (%)	8.6	5.7	6.4	6.8	3.9	4.4
Maximum ARE (%)	89.7	42.0	46.5	116.4	39.2	45.8
Gjennomsnittlig ARDE (%)	4.5	2.1	2.0	4.1	2.0	1.2
RE i variasjonsbredde (%)	47.9	-17.1	6.3	121.7	-27.6	6.2

Simuleringen bekrefter resultatene i avsnitt 3.2, nemlig at den EBLUP generelt reduserer den totale usikkerhet i estimering i forholdet til den direkte estimator. Spesielt har modellbasert estimatene en bedre evne i å begrense de største feiler blant domenene. Når det gjelder de fordelingsmessige karakteristikk, ser vi at (i) modellbasert estimatene er bedre den direkte estimator, (ii) den simultane estimator er bedre enn den EBLUP — forbedringen er spesielt klar mht. variasjonsbredden. Alt sett under ett, synes den simultane estimator å være det best generelle alternativ, mens den EBLUP er best hvis hensynet til domene spesifikk prediksjon er den eneste kriterium for metodevalg.

A Algoritme for EBLUP

```
#########################################################################
# input:
# data = data matrix where each record corresponds to an observation
# which contains the domain index "i" and the variable of interest "y"
# notation: y_ij = the j-th observed value of "y" in domain i
#
# output:
# (hat.mu, hat.sigma2.e, hat.sigma2.v) = estimates of model parameters
# hat.v = vector of EBLUP estimates of random domain effects
# hat.theta = vector of EBLUP estimates of teoretical domain means
#########################################################################

eblup <- function(data)
{
        m = number of domains
        n_i = sample size of domain i
        n = sum_i n_i = total sample size
        y_i = sample mean within domain i
        e_ij = y_ij - y_i = deviation from the within-domain sample mean
        y.bar = sum_{i,j} y_ij/n = overall sample mean

        Q = set of index for all domains with n_i > 1
        alpha = number of domains with n_i > 1
        sse1 = sum_{i,j} e_ij^2
        nu1 = (sum_{i in Q} n_i) - alpha
        hat.sigma2.e = sse1 / nu1

        sse2 = sum_{i,j} (y_i - y.bar)^2
        eta1 = sum_i (n_i - n_i^2 / n)
        hat.sigma2.v = {sse2 - hat.sigma2.e * (n - m)} / eta1

        gamma_i = hat.sigma2.v / (hat.sigma2.v + hat.sigma2.e/n_i)
        hat.mu = {sum_i n_i * (1-gamma_i) * y_i} / {sum_i n_i * (1-gamma_i)}
        hat.v_i = g_i * (y_i - hat.mu)
        hat.theta_i = hat.mu + hat.v_i

        output (hat.mu, hat.sigma2.e, hat.sigma2.v, hat.v, hat.theta)
}
```

B Estimering av MSE

```
######################################################################
# input:
# (sigma2.e, sigma2.v) = estimates of varaince components
# n = vector of domain sample sizes
# gamma = vector of estimated domain shringkage factors
#
# output:
# mse = vector of estimated mse of theoretical domain means
######################################################################

mse.est <- function(sigma2.v, sigma2.e, n, gamma)
{
        n = sum_i n_i
        Q = set of index for all domains with n_i > 1
        alpha = number of domains with n_i > 1
        nu1 = (sum_{i in Q} n_i) - alpha
        eta1 = sum_i (n_i - n_i^2 / n)
        eta2 = sum_i n_i^2 * (1 - n_i / n) + (sum_i n_i^2)^2 / n^2

        g1_i = gamma_i * sigma2.e / n
        g2_i = sigma2.e * (1 - gamma_i)^2 / {sum_i n_i * (1 - gamma_i)}

        s2 = sigma2.v * sigma2.e
        psi = {(n - 1 - nu1) * (n - 1) * sigma2.e^2} / nu1
        phi = eta2 * sigma2.v^2 + 2 * eta1 * s2
        g31_i = 2 * sigma2.e^2 * (psi + phi) / eta1^2
        g32_i = 2 * s2^2 / nu1
        g33_i = 4 * s2 * (n - 1 - nu1) * sigma2.e^2 / (nu1 * eta1)
        kappa_i = n_i^2 * (sigma2.v + sigma2.e / n_i)^3
        g3_i = (g31_i + g32_i + g33_i) / kappa_i

        mse_i = g1_i + g2_i + 2 * g3_i

        output mse
}
```

De sist utgitte publikasjonene i serien Notater

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

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

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

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

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

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

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

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

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

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

2003/77 E.E: Eibak: Undersøking om foreldre-betaling i barnehagar, august 2003. 46s.

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

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

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

2003/81 J.H. Wang: Frafall i konjunkturbarometeret. 45s.

2003/82 P. Holmen og K.Lorentzen: Dokumentasjon av etableringen av UT - populasjonen - konsentrasjon om store enheter og stabilitet over tid. 49s.

2003/83 T.H. Christensen: Boligprisindeksen. Datagrunnlag og beregningsmetode. 20s.

2003/84 G. Dahl: Enslige forsørgere med overgangsstønad. Økonomisk situasjon etter avsluttet stønad. 74s.

2003/85 T.M. Normann: Omnibusundersøkelsen august/september 2003. Dokumentasjonsrapport. 36s.

2003/86 T. Eika og T. Skjerpen: Hvitevarer 2004. Modell og prognose. 19s.

2003/87 S. Blom og B. Lie: Holdningen til innvandrere og innvandring. Spørsmål i SSBs omnibus i august/september 2003. 58s.

2003/88 A. Holmøy: Undersøkelse om livsløp, aldring og generasjon (LAG). Dokumentasjonsrapport. 135s.

2003/89 Ø. Kleven og E. Wedde: Medieundersøkelsen 2002. Dokumentasjonsrapport. 43s.

2003/90 S. Derakhshanfar, S. Lien og C. Nordseth: FD - Trygd. Dokumentasjonsrapport. Barnetrygd. 1996-2002. 44s.

2003/91 J. Larsson og K. Telle: Dokumentasjon av DEED . En database over bedriftspesifikke miljødata og økonomiske data for forurensende norske industribedrifter. 16s.

2003/92 J.I. Hamre: Undersøkelsen om legemeldt sykefravær. Dokumentasjon av utvalgsplan, trekking og rullering for 2003. 37s.

2004/1 manuell A.G. Pedersen: Sammenligning av og auomatisert metode ved koding av dødsårsak. 22s.

2004/2 T.M. Köber: Registerbasert sysselsettingsstatistikk for helse og sosialhjelp. 42s.