Interne notater

STATISTISK SENTRALBYRÅ

Nr. 86/10

28. februar 1986

NOKRE METODER FOR ESTIMERING AV REGIONALE TALL VED KOMBINASJON AV ADMINISTRATIVE REGISTER OG UTVALG.

av

Kjell Arne Brekke

Innhald

1. Innledning

Vi skal i dette notatet sjå på metoder for å kombinere informasjon frå administrative register med utvalgsundersøkelsar. Formålet er i første rekke å utvikle metoder for å bruke registeropplysingane til å lage tall på regionalt nivå. Likevel er metoden for bruk av registeret for estimering av totaltall ein sentral ide, og vi vil innledningsvis presentere denne.

Dette notatet er ein dokumentasjon til eit prosjekt som enno ikkje er fullført. Det er forholdsvis teknisk og inneheld skisser til metoder som enno ikkje er utprøvd empirisk.

Bakgrunnen for arbeidet er eit prosjekt for å bruke registeret til å lage tall for sysselsetting på fylkesnivå. Vi bruker då Arbeidskraft-undersøkelsen (AKU) kombinert med Arbeidsgivar / Arbeidstakar (A/A) registeret. Metoden er generell og den kan tenkast brukt også i andre sammenhengar. Vi vil difor gi ein generell presentasjon av den i dette notatet. Likevel vil estimering av andel lønnstakarar ved kombinasjon AKU og A/A-registeret bli brukt som eit gjenomgangseksempel.

Dette arbeidet er sterkt inspirert av Stein estimering og empirisk Bayes. Det er denne siste teknikken vi vil bruke her. Som ein motivasjon av den modellen vi vil presentere har vi tatt med eit kapittel om den historiske utviklinga av desse metodene. Dette er ikkje meint å være fullstendig, men er først å fremt ein presentasjon av enkelte interessante arbeid i denne tradisjonen.

Indeksar i tekstavsnitt er understreka.

2. Registermetoden

2.1 Estimering

Anta at vi er interessert i å estimere antall personar med kjennetekn A. I registeret er alle personane i populasjonen klassifisert med kjenneteikn A eller uten kjennetegnet ($\bar{A}$). Registeret er ofte ikkje tilstrekkelig oppdatert, eller definisjonen av kjennetegnet er forskjellig frå den vi er interessert i . Dette kan føre til at klassifiseringa i registeret er "feil". For eit utvalg på $n_{..}$ personar observerer vi så den sanne klassifiseringa A eller $\bar{A}$. Resultatet kan vi sette opp i følgande tabell:

Tabell 1. Sammenheng AKU/register.

Utvalg	Register 1. A	Register 2. $\bar{A}$	
1. A	n_{11}	n_{12}	$n_{1.}$
2. $\bar{A}$	n_{21}	n_{22}	$n_{2.}$
	$n_{.1}$	$n_{.2}$	$n_{..}$

Frå registeret får vi også klassifiseringa for heile populasjonen.

Tabell 2. Registerklassifikasjon etter strata

	Registeret. A	Registeret. $\bar{A}$
	N_1	N_2

Vi vil difor sjå på $n_{.1}$ og $n_{.2}$ som gitte. Vi antar så at n_{ij} er binomisk fordelt gitt $n_{.j}$ $i=1,2$ $j=1,2$. Vi tenker oss altså at vi har pre-stratifisert. Ofte vil vi måtte bruke etter-stratifisering, og det vil då være bedre med ein multinomisk modell. Ein slik modell er behandla i Tennebein 1970. Problemet i det tilfellet er at $n_{.j}$ blir stokastisk. Dette ville skape ein del vanskar seinare i notatet.

La P være den faktiske andelen med kjennetegn A i populasjonen. Vi estimerer no P med

$$P^* = \sum_i P_i^* N_i$$

der

$$P_i^* = \frac{n_{1i}}{n_{.i}}$$

Då P^* er ein vanlig stratifisert estimator vil den også være forventingsrett.

Vi har til no antatt at vi berre har to moglege registerklassifiseringar A og $\bar{A}$. Det er sjølvsagt ingenting i vegen for å bruke fleire klassifiseringar. For sysselsetting t.d. er også informasjon om alder, kjønn og inntekt interessant.

2.2 Varians for registermetode-estimatoren

Kva har vi så oppnådd ved å bruke denne etterstratifiseringa? Eksempelet kombinsjon av AKU og A/A registeret kan gi ein illustrasjon av gevinsten. Andelen lønnstakarar i befolkninga totalt er ca 50%. Blandt registrerte arbeidstakarar er den rundt 90% og blandt ikkje registrerte ca 15%. Då variansen på ein enkeltobservasjon er gitt ved formelen p(1-p) ser vi at variansen uten etterstratifisering er ca 0.25, medan for registrete/ikkje registrerte 0.09/0.127. Då begge gruppene er omlag like store vil variansen bli $0.109/n_{..}$ som er ein variansreduksjon på 56% i forhold til ingen stratifisering.

Berre knappt 90% av lønnstakarane er registrert i A/A registeret og vel 10% av dei som ikkje er lønnstakarar er likevel registrert. Likevel er variansreduksjonen med registermetoden og dette registeret betydelig. Vi kan altså få stor gevinst med registermetoden sjølv om registeret ikkje er svært godt. Dersom registeret er svært godt vil vi dessuten neppe trenge eit utvalg.

3. Stein estimering og empirisk Bayes.

Vi skal i dette kapittelet prøve å gi eit kortfatta resyme over utvikliga av empirisk Bayes metoder. Framstillinga er ikkje meint å være fullstendig og fleire viktiga arbeid er truleg utelatt.

La $X_1,\ldots,X_k$ være uavhengige normale med ukjend forventing $\theta_1,\ldots,\theta_k$ og varians 1. W.Stein viste i ein berømt artikkel (1955) at ved kvadratsummen som tapsfuksjon er den vanlige maksimum likelihood estimatoren $X=(X_1,\ldots,X_k)'$ inadmissibel som estimator for $\theta=(\theta_1,\ldots,\theta_k)'$, når $k\geq 3$. I seinare arbeid presenterer James og Stein (1961) ein estimator som har uniformt mindre risiko enn dem vanlige maksimum likelihood estimatoren. La tapet være

$$L(\theta,a) = \sum_i L(\theta_i,a_i) = \sum_i (\theta_i - a_i)^2$$

der a er estimatet for θ. Risikoen er no definert som:

$$R(\theta,a) = E\, L(\theta,a)$$

James-Stein estimatoren er

$$\varphi_1(X) = (1-(k-2)/\|X\|^2)X$$

der

$$\|X\|^2 = \sum_i X_i^2$$

Vi vil samanlikne denne estimatoren med maksimun likelihood estimatoren.

$$\varphi_0(X) = X$$

Risikoen for dei to estimatorane er

$$R(\theta,\varphi_0) = k$$

og når $k \geq 3$:

$$R(\theta,\varphi_1) = k - E((k-2)^2/(k-2+2P)) < k$$

der P er Poisson fordelt med forventing $\|\theta\|^2/2$. Dette betyr at James-Stein estimatoren har mindre risiko uansett verdien av θ. Vi ser at James-Stein estimatoren inneberer at vi drar X mot 0. Som rimelig kan være er difor reduksjonen av risikoen størst når θ er nær 0. dvs når $\|\theta\|^2$ er liten.

Efron og Morris 1972 viser seinare at James-Stein estimatoren kan utledast som éin empirisk Bayes estimator.

La

$$X_{ij} \overset{\text{uavh}}{\sim} N(\theta_i,\sigma^2) \qquad i=1,2,\ldots,k \qquad k\geq 3 \qquad j=1,2,\ldots,n$$

Vi har altså k normalfordelingar og n observasjonar frå kvar. Ved ei enkel normering kan vi få $\sigma^2/n=1$, og vi er då i samme tilfellet som over. Vi antar at dette er gjort.

I ein vanlig Bayes modell ville vi no gi ei apriori fordeling for θ f.eks. :

$$\theta_i \overset{\text{uavh}}{\sim} N(\mu,\tau^2) \qquad i=1,2,\ldots,n$$

Bayes regelen for å estimere θ er no aposteriori forventing:

$$\delta_i^* = \mu + (1-1/(1-A))(\bar{X}_i - \mu)$$

der $A = n\tau^2/\sigma^2 = \tau^2$. Dersom vi set $\mu=0$ får vi

$$\delta_i^* = (1-1/(1-A))\bar{X}_i$$

Vi kunne no tenke oss at vi held på den parametriske forma til apriori fordelinga, men ser på τ^2 som ein ukjend som vi må estimere. No gjeld:

$$E\,\frac{(k-2)\sigma^2}{n\|\bar{X}\|^2} = 1/(1+A)$$

Ein forventingsrett estimator for $1/(1+A)$ er altså:

$$\frac{(k-2)\sigma^2}{n\|\bar{X}\|^2} = (k-2)/\|X\|^2$$

Dette gir ,innsett i δ_i^*:

$$\delta_i^* = (1 - (k-2)/\|X\|^2)\,\bar{X}_i$$

som er James-Stein estimatoren.

Ein slik metode der ein estimerer apriori fordelinga og elles bruker 'Bayes-metoder, blir kalla empirisk Bayes metode.

Efron og Morris påpeikar at sjølv om James-Stein estimatoren gir minimal risiko for gruppa totalt, så kan den individuelle risikoen bli svært stor. :

$$\max_{\theta_i} R(\theta_i, \delta_i^*) = E\,L(\theta_i, \delta_i^*) = k/4$$

medan

$$\max_{\theta_i} R(\theta_i, \bar{X}_i) = 1$$

For store k og uheldige θ kan difor den individuelle risikoen bli svært høg. Dei viser så at ein ved å tillate berre estimatorar med eit vist maksimalavvik frå X kan redusere maximal individuell risiko betydelig, uten at tapet i risiko for gruppa blir større enn 5-10 %.

Vi ser at James-Stein estimatoren er ei veiing mellom $\bar{X}$ og 0. 0 er her eigenlig eit vikårlig valgt punkt, og estimatoren ville få dei samme egenskapane om vi hadde valgt eit anna punkt μ. No er sjølvsagt reduksjonen i risiko størst dersom vi vel μ nær ξ. Risikoen avtar med størrelsen:

$$\|\xi-\mu\|^2 = (\xi-\mu)'(\xi-\mu)$$

I ein empirisk Bayes modell er det difor rimelig å prøve å estimere μ. I anvendelsar blir dette også gjort.

Ein av dei første anvendelsane av empirisk Bayes metoder i litteraturen er eit arbeid av Fay og Herriot 1979. Dei skulle estimere intekt "per capita" (IPC) i 1972 for ca 39 000 forskjellige områder i USA. Resultatet var grunnlag for fordelig av offentlige midlar. "Lokale sensorar" ville difor ha stor interesse av talla.

Frå ei folketelling i 1970 hadde dei intektsopplysingar for eit utvalg på 20%. Då 15 000 av områda hadde færre enn 500 innbyggarar var likevel utvalgsvariansen betydelig. Ein empirisk Bayes estimator vart nytta for å forsøke å forbedre estimata.

La X_i være ein forventingsrett estimator for IPC i område i. Faktisk IPC i område i er ξ_i. Vi kjenner også verdien av ein forklaringsvariabel z_i i område i. Vi antar no at:

$$X_i \overset{\text{uavh}}{\sim} N(\xi_i, D_i)$$

$$\xi_i \overset{\text{uavh}}{\sim} N(z_i'\beta, A)$$

Der A ikkje avheng av i. β blir estimert som ved vanlig regresjon, og A kan vi estimere ved å løyse likninga:

$$\sum_i \frac{(X_i - z_i'\beta^*)^2}{D_i + A} = I - k$$

der $I \approx 39000$ er antall regionar og k er antall uavhengige forklaringsvariable. Bayes estimatoren ved denne apriori fordeliga er:

$$\xi_i^* = w_i X_i + (1-w_i) z_i'\beta^*$$

der

$$w_i = A/(A+D_i)$$

For å avgrense individuell risiko vart estimatet avgrensa til å ligge innan to standardavvik frå X_i.

For å tese estimata vart det foretatt ei spesiell tellig i ein del av

dei minste områda. Denne vart foretatt i 1973 og samla informasjon om intekt i 1972. (1970 tellinga gjalt inntekt i 1969). Når ein ser på resultatet av denne totaltellinga som "fasit" fant ein at gjenomsnittlig prosentavik vart redusert frå 28.6% for X_i til 22.0% for ξ^*_i i område med færre enn 500 innbyggarar. Då ein del av avvika er differansar over tid er truleg risikoreduksjonen endå større enn dette tyder på.

Ein meir fullstendig teoretisk diskusjon av slike regresjon / empirisk Bayes modellar finn ein i Morris 1983 .

4. Nokre metoder for beregning av regionale tall

Som nevt er formålet med denne estimeringa å lage best moglege regionale tall. Vi vil prøve å anvende empirisk Bayes ideen. Vi vil då lage ein estimator som berre baserer seg på observasjonar frå den regionen vi er interessert i. Denne estimatoren kallar vi den direkte estimatoren. I tillegg vil vi lage ein estimatore som baserer seg på utvalget frå heile landet. Denne estimatoren vil vi likevel sjå på som ein etimator for den regionale parameteren. T.d. vil ein estimator for andelen lønnstakarar i heile landet bli sett på som ein estimator for andelen lønnstakarar i eit spesielt fylke. Denne estimatoren kallar vi den syntetiske estimatoren. Den systetiske estimatoren har sjølvsagt langt mindre varians enn den direkte, men kan på den andre sida være forventingsskjev, noko den direkte ikkje er. For å få det beste ut av begge vil vi bruke ein lineærkombinasjon av dei.

I resten av dette arbeidet vil vi la kjennetegnet A være sysselsatt.

4.1 Ein empirisk Bayes modell

I dette avsnittet vil vi presentere nokre hovudidear for estimering av ein parameter på regionalt nivå. Vi tenker oss no at vi skal estimere ein parameter P_f ,andelen sysselsatt for region f=1,..,F. Der F er antall regionar. Då det er samme parameter vi skal estimere i alle regionane er det rimelig å prøve ein empirisk Bayes estimator. Då vi ikkje har normalfordelte variable vi vi avgrense oss til lineære estimatorar. Denne metoden blir gjerne kalla lineær empirisk Bayes.

Modellen vår er:

$$E P_f = P$$

og

$$\mathrm{Var}\, P_f = A$$

Frå kvart fylke har vi trekkt eit tilfeldig utvalg av $n_{.f}$ personar. Vi observerer så for kvar enkelt av desse om dei er lønnstakarar. La n_{1f} være antall lønnstakarar. No vil

$$P^*_f = n_{1f} / n_{.f}$$

være ein forventingsrett estimator for P_f. D.v.s.:

$$E\, P^*_f = P_f$$

Vi antar at n_{if} er binomisk. Vi har her i røynda betinga m.h.p. $n_{.f}$ som ofte vil være stokastisk. Dessuten ,då vi trekker frå ei endelig befolkning burde n_{if} strengt tatt være hypergeometrisk. Så lenge vi held oss til små utvalg vil dette likevel ikkje ha særlig betydning.

Variansen til P^*_f er:

$$\text{Var } P^*_f = V_f = P_f (1-P_f)/n_{.f}$$

No blir:

$$E\,(P^*_f - P_f)^2 = V_f + A_f$$

La P^* være minste kvadraters estimatoern for P, dvs den P^* som minimerer

$$\sum_f \frac{(P^*_f - P^*_f)^2}{V_f + A_f}$$

Det er lett å vise at

$$P^* = C \sum_f P^*_f (V_f + A_f)^{-1}$$

der

$$C^{-1} = \sum_f (V_f + A_f)^{-1}$$

Ein hovudide er at vi no kan sjå på P^* som ein estimator for P_f. Denne estimatoren vil vi kalle den syntetiske estimatoren i motsetning til P^*_f som vi kallar den direkte estimatoren.

Vi ser bort frå kovariansen mellom P^* og P^*_f . Dersom regionane er mange og små vil dette ikkje være av særlig betydning. I tilfeller med

få regionar kan vi oppnå uavhengighet ved å la den syntetiske estimatoren basere seg berre på observasjonar frå andre regionar.

Vi søker no ein James-Stein liknande estimator som drar P^*_f mot forventinga i apriori-fordelinga P^*. Vi avgrensar oss til å sjå på lineære kombinasjonar

$$\theta^*_f = (1-w_f)P^*_f + w_f P^*_f$$

Vi ønsker å minimere

$$g(w_f) = E(\theta^*_f - P_f)^2 = (1-w_f)^2 V_f + w_f^2 (G_f + A_f)$$

der

$$G_f = E(P^*_f - P_f)^2$$

er prediksjonfeilen ved minste kvadraters estimering. Vi finn at g blir minimert av

$$w_f = V_f / (V_f + G_f + A_f)$$

Vi kan sjå på P^*_f som ein estimator for P_f og tolke

$$G_f + A_f = E(P^*_f - P_f)^2$$

som variansen til P^*_f. Resultatet over seier då at det er optimalt å gi dei to alternative estimatorane for P_f vekter som er proporsjonale med variansen.

4.2 Ein modell med fleire strata og kovarians

Den modellen vi såg på i forrige avsnitt hadde ingen stratainndeling. Vi tar på den måten ikkje hensyn til den informasjon vi har om ein region når vi kjenner størrelsen på ulike strata. I vårt tilfelle ville dette bety at vi ikkje tok hensyn til andelen registrert i A/A registeret når vi laga den syntetiske estimatoen. Vi misser då verdifull informasjon.

Vi kan no tenke oss at vi kan dele variasjonen i P_f i to delar. Den eine delen er variasjon som vi kan observere i registeret. den andre delen er variasjon innen samme gruppe i registeret. Når vi t.d. ser på sysselsetting vil vi ha variasjon i registrert sysselsetting (som vi kjenner) og variasjon i andelen faktisk sysselsatte balndt registrerte/ ikkje registrerte. Desse andelane kjenner vi ikkje , men

empirien tyder på at variasjonen i desse er langt mindre enn variasjonen i P_f. Dersom P_{if} er andelen sysselsatte med registerklassifikasjon i i fylke f og R_{if} er andelen registrert i i i fylke f så er

$$P_f = \Sigma_i P_{if} R_{if}$$

R_{if} kjenner vi og P_{if} varierer mindre enn P_f mellom fylka.

Vi vil no gjere som i modellen vi laga i 4.1. Denne gongen vil vi derimot sjå på andelane innen ulike strata som realisajonar av stokastiske variable og ikkje total-andelane. Då desse varierer mindre enn totalandelane er dei også lettare å predikere.

Problemet når vi bruker fleire strata er at dei ulike P_{if} i samme fylke ikkje kan sjåast på som realisasjonar av uavhengige variable. Vi må difor eksplisitt regne med ein kovarians. Denne kovariansen ser ut til å ha stor betydning for resultatet.

I mange tilfeller kan vi no trenge å estimere mange kovariansar i apriori-fordeliga. Dette er ofte ikkje mogleg i ein undersøkelse med lite utvalg. Ofte vil det då bli nødvendig å bruke data frå andre kelder. For AKU t.d. vil vi måtte estimere denne kovariansen på data frå Folke og Boligtellinga (FOB) 1980. Vi nærmar oss då ein rein Bayes modell der delar av apriori-fordelinga er bestemt før vi gjer observasjonane.

La P_{if} være andel lønnstakarar i strata i og fylke f. Vi antar at P_{if} er ein stokastisk variabel med forventing

$$E P_{if} = P_i \qquad 4.2.1$$

og kovariansmatrise:

$$E (P_{if} - P_i)(P_{jf} - P_j) = A_{ij} \qquad 4.2.2$$

Frå strata i og fylke f har vi trekkt tilfeldig $n_{.if}$ personar. For kvar person observerer vi om vedkomande er lønnstakar eller ikkje. Antall lønnstakarar er n_{1if}. Ein forventingsrett observator for P_{if} er no

$$P^*_{if} = n_{1if} / n_{.if}$$

$$EP^*_{if} = P_{if} \qquad 4.2.3$$

Denne estimatoren er tilnerma binomisk fordelt med varians

$$\text{Var } P^*_{if} = V_{if} = P_{if}(1-P_{if})/n \qquad 4.2.4$$

No gjeld:

$$E(P^*_{if} \mid P_i) = P_i$$

og

$$E(P^*_{if} - P_i)^2 = V_{if} + A_{ii}$$

Som i forrige avsnitt finn vi minste kvadraters estimatoren for P_i basert på observasjonar frå andre fylker:

$$P^*_i = \Sigma_f P^*_{if} (V_{if} + A_{ii})^{-1} C \qquad 4.2.5$$

der

$$C^{-1} = \Sigma_f (V_{if} + A_{ii})^{-1}$$

Vi er no intessert i størrelsen

$$\theta_f = \Sigma_i P_i R_{if}$$

der R_{if} er andelen i stratum i fylke f.

Den syntetiske estimatoren blir i dette tilfellet

$$\theta^*_f = \Sigma_i P^*_{if} R_{if}$$

og den direkte er

$$\theta^-_f = \Sigma_i P^*_i R_{if}$$

Vi kunne no bruke ein estimator av typen

$$v_f = (1-w_f)\,\theta^*_f + w_f\,\theta^-_f$$

Ulempa med denne estimatoren er at den ikkje tar hensyn til at informasjonen frå andre fylker kan ha ulik verdi i ulike strata. T.d. er det rimeleg å tru $P^*_{\underline{i}}$ er ein betre estimator for $P_{\underline{if}}$ når vi ser på etablerte (25-55 år) registrerte arbeidstakarar enn om vi ser på ungdommar (16-24 år). $P^*_{\underline{i}}$ bør difor ha større vekt i den første gruppa. Vi vil derfor tillate vektene å være forskjellige i forskjellige stratum. Dette gir oss ein estimator av typen:

$$\mu_f = \Sigma_i [(1-w_{if}) P^*_{if} + w_{if} P^*_i] * R_{if}$$

Som vi skal sjå treng vi mange parametrar for å kunne estimere $\mu_{\underline{f}}$. Estimatoren $v_{\underline{f}}$ er difor eit interessant alternativ då vi treng lang ferre parametrar til den. Det er også viktig å merke seg at fordi vi treng ferre parametrer vil det oftst være mogleg å estimere alle på data frå den aktuelle undersøkelsen. Vi kan på den måten halde oss til ein rein empirisk Bayes modell. For å kunne velge mellom dei vil det være nyttig med meir innsikt i det datamaterialet vi skal nytte dei på. Vi skal i det følgande konstentrere oss om $\mu_{\underline{f}}$, men vil likevel ta med eit resultat om $v_{\underline{f}}$ i teorem 1.

Vi ønsker no å bestemme $w_{\underline{f}} = (w_{1\underline{f}},\ldots,w_{I\underline{f}})$ (I - antall strata) slik at kvadratavviket blir minst mogleg. Dvs vi vil minimere:

$$g(w_f) = E (\mu_f - \theta_f)^2$$

set

$$G_{if} = E (P^*_{if} - P_i)^2$$

Vi har no følgande teorem:

Teorem 1.
Under modellen 4.2.1 - 4.2.5 vil Vektene som minimerer forventa kvadratavvik være gitt ved likningssystemet:

$$-2(1-w_{if})R_{if}^2 V_{if} + 2 w_{if} R_{if}^2 (G_{ii} + A_{ii}) + 2 \sum_{j \neq i} w_{jf} R_{if} R_{jf} A_{ij} = 0$$

Under bibetingelsen $w_{if} = W_f$ vil den W_f som minimerer kvadratavvikert være:

$$W_f = (\Sigma R_{if}^2 V_{if})/(\Sigma R_{if}^2 (V_{ii} + G_{ii} + A_{ii}) + \Sigma_{i \neq j} R_{if} R_{jf} A_{ij})$$

Bevis:

$$g(w_f) = E(\mu_f - \theta_f)^2$$

$$= \Sigma_i (1-w_{if})^2 R_{if}^2 V_{if} + \Sigma_i w_{if}^2 R_{if}^2 (G_{ii} + A_{ii}) + 2 \Sigma_{i<j} w_{if} w_{jf} R_{if} R_{jf} A_{ij}$$

Dei partielle deriverte av g blir:

$$\partial g/\partial w_{if} = -2(1-w_{if})R_{if}^2 V_{if} + 2 w_{if} R_{if}^2 (G_{ii} + A_{ii}) + 2 \sum_{j \neq i} w_{jf} R_{if} R_{jf} A_{ij}$$

Første ordens betingelse for minimum :

$$\partial g/\partial w_{if} = 0 \quad i=1,\ldots,I$$

gir no det nevnde likningssystem for å bestemme w_{if} .

Det gjenstå no berre å vise at g faktisk har eit minimum. Det er då nok å vise at Hessematrisa H(g) er positivt definit. Set:

$$z' = (x_1 R_{11}, \ldots, x_I R_{II})$$

då har vi for $x \neq 0$ at $z \neq 0$ og

$$x'H(g)x = z'(A+Diag(V_1+G_1,\ldots,V_I+G_I))z > 0$$

Beviset for det siste punktet er heilt analogt og vi difor bli utelatt.

QED.

Vi kan no bruke dette resultatet til å vurdere betydninga av kovariansen mellom prediksjonane innan fylket. Vi antar då at $A_{ij} = cr$ $A_{ii} = c$ der $r \varepsilon (0,1)$ og c er konstantar og $j \neq i$. Vi antar og at $V_{if}=V$ uavhengig av i. og at $R_{if}=R$ uavhengig av i. Vi antar også at G_{if} er så liten i forhold til V og c at vi kan sette den lik 0. Då blir vektene dei samme innen alle strata og dei blir bestemt av likninga.

$$R^2V = w\,R^2V + w\,R^2c + \sum_{j\neq i} w\,R^2r\,c$$

Med H strata gir dette :

$$w = V / (V + (1+(H-1)r)c)$$

r er her korrelasjonskoeffisienten. Vi ser at dersom vi ser bort frå denne korrelasjonen vil w bli kraftig overestimert med mindre :

$$(H-1)r << 1$$

Dersom vi har t.d. 3 aldersgrupper og to registerklassifikasjonar og to kommunetyper gir dette H=12 og med r=0.1 blir (H-1)r=1.1 . Dette illustrerer at sjølv små korrelasjonar kan det være viktig å ta hensyn til. Rett nok er det ikkje rimelig å anta ar korrelasjonen er den samme mellom alle strata, men dette skulle ikkje bety noko for hovudkonklusjonen.

Vi bør difor kunne forkaste hypotesa r>0.1 før vi kan sjå bort frå denne korrelasjonen.

Vi vil prøve å finne ut om dette er eit intuitivt rimelig resultat. Vi tenker oss då at vi deler populasjonen i eit fylke i to like store delar, slik at delane er mest mogleg like. Utvalgsvariansen blir dobbla innan kvar av desse to delane,

$$V_{if} = 2\,V_f \qquad i=1,2$$

men det er ikkje rimelig å anta at A blir dobbla :.

$$A_{ii} < 2A \qquad i=1,2$$

(Tenk deg befolkninga i Oslo tilfeldig delt i to. Dette vil ikkje endre den syntetiske estimatoren særleg, det er heller ikkje grunn til å tru at P_{if} er særleg forskjellig frå P_f. Altså $A_{ii} \approx A$.)

Dersom vi veier optimalt innen kvart strata uten å ta hensyn til kovariansen får vi :

$$w_{if} = V_{if} / (V_{if} + A_{ii}) > V_f / (V_f + A) = w_f$$

Dersom vi så summerer over dei to strata får vi

$$\mu_f = (1-w_{if})P^*_{if} + w_{if} P_f$$

men dette er ikkje dei vektene som minimerer kvadratavviket. Dersom vi ikkje tar hensyn til kovariansen får vi altså ei suboptimal forskyvning mot den syntetiske estimatoren.

4.3 Modell med tidsperspektiv

Vi vil no tenke oss at vi har ein løpande utvalgsundersøkelse. Til no har vi berre glatta over rom, ved å lage ein syntetisk estimator som brukar informasjon frå andre regionar. Ein nærliggande tanke er no å prøve å glatte også over tid.

I eksempelet med AKU og A/A registeret kunne vi observere at den syntetiske estimatoren i enkelte fylker lå konsekvent over eller under den direkte estimatoren over fleire kvartal. Vi ønsker å lage ein estimator som brukar slik informasjon frå tidligare perioder til å forbetre den syntetiske estimatoren.

Vi vil i dette avsnittet predikere differansen mellom den syntetiske estimatoren og den vireklige andelen , v.h.a tidligar observasjonar. Dette brukar vi til å lage ein ny syntetisk estimator. Til slutt finn vi eit likningssystem som bestemmer dei optimale vektene.

Vi tar derfor eksplisitt inn i modellen at vi har ein tidsserie av observasjonar. La P_{ift} være andel lønnstakrar i strata i , fylke f og tidspunkt t. P_{ift} er ein stokastisk variabel og vi antar at:

$$P_{ift} = P_{it} + D_{ift}$$

og

P_{it} er ein konstant

$$E\,D_{ift} = 0$$

$$E\,D_{ift}\,D_{jft} = A_{ij} .$$

Vi treng ikkje spesifisere den stokastiske prosessen D_{ift} nærmare her, men vi må anta at den er stasjonær.

Som før har vi ein binomisk observator P^*_{ift} for P_{ift}. Vi brukar den samme estimator for P_{it} som før:

$$P^*_{it} = \sum_f P^*_{ift} \,(V_{ift} + A_{ift})^{-1}\, C_{ii}$$

$$C^{-1} = \sum_f (V_{ift} + A_{ift})^{-1}$$

I tillegg vil vi estimere D_{ift} med

$$D^*_{ift} = \sum_{\tau=0}^{t-1} e_\tau (P^*_{if\tau} - P^*_{i\tau})$$

der

$$\sum_{\tau=0}^{t-1} e_\tau = 1$$

Då vi ikkje har spesifisert den stokastiske prosessen nærmare, står vi her fritt til å bruke kva tidsrekkemodell vi vil til å estimere e_τ. Den einaste avgrensinga er at vi må anta at tidsrekka er stasjonær.

I mange undersøkelsar blir det brukt roterande utvalg. Dette fører til at observasjonar på ulike tidspunkt kan være korrelerte. Først og fremst vil P^*_{ift} og D^*_{ift} være korrelerte. Men også P^*_{it} og D^*_{ift} er korrelerte. Desse korrelasjonane kan vi beregne når vi kjenner e_τ, $\mathrm{Corr}(P^*_{ift}, P^*_{if\tau})$ og $\mathrm{Corr}(P^*_{it}, P^*_{i\tau})$.

Set

$$Q^*_{ift} = P^*_{it} + D^*_{ift}$$

Vi ser på estimatorar av forma:

$$\theta^*_{ft} = \sum_i [(1-w_{if})P^*_{ift} + w_{if} Q^*_{ift}]\, R_{ift}$$

Som før ser vi på:

$$g(w_f) = E(\theta^*_{ft} - \theta_{ft})^2$$

Set:

$$B_{ift} = G_{ift} + H_{ift} + J_{ift}$$

$$= E(P^*_{it} - P_{it})^2 + E(D^*_{ift} - D_{ift})^2 + E(P^*_{it} - P_{it})(D^*_{ift} - D_{ift})$$

$$C_{ift} = E(P^*_{ift} - P_{ift})(D^*_{ift} - D_{ift})$$

Vi har no følgande teorem:

Teorem 2.

Dersom $2V_{ift} + 2B_{ift} - A_{ii} - 2C_{ift} \geq 0$

så vil vektene w_{if} som minimerer kvadratvviket være bestemt av likningssytemet:

$$-2(1-w_{if})R^2_{if}\,V_{ift} + 2\,w_{if}\,R^2_{if}\,B_{ift}$$

$$+ (1-2w_{if})R^2_{if}\,C_{ift} + 2\sum_{i\neq j} w_{jf} R_{jf} R_{if} A_{ij} = 0$$

Bevis: Sjå Appendix.

4.4 Estimering av parametrane i modellen

Det er svært mange parametrar i modellen . For kvar periode er det $4IF + I^2$ parametrar. Det er difor viktig at vi klan lage gode estimat for desse parametrane.

$$V_{ift} = P_{ift}(1-P_{ift}) / n$$

Vi treng her ein estimator for P$\underline{ift}$. $P^{*}\underline{ift}$ vil i mange tilfelle basere seg på svært få observasjonar og vil difor være svært ustabil. Vi vil difor heller bruke

$$P^{**}_{ift} = P^{*}_{it} + D^{*}_{ift}$$

Matrisa A er det største problemet. Dersom vi har mange strata vil det bli svært mange kovariansar å estimere. Ofte vil vi måtte bruke data frå andre kjelder for å få estimert den. Vi må då gjerne anta at A er tidsinvariant. I eksempelet AKU og A/A registeret kunne vi tenke oss å estimere A på FOB 1980. Vi ville då som tidligare nevnt forlate den reine empiriske Bayes tenkegangen og bruke ein meir Bayes liknande modell.

$$C_{ift} = E\,(P^{*}_{ift} - P_{ift})(D^{*}_{ift} - D_{ift})$$

$$= \sum_{\tau=0}^{t-1} e_{\tau}\, E(P^{*}_{if\tau} - P_{if\tau})(P^{*}_{ift} - P_{ift})$$

$$\approx V_{ift} \sum_{\tau=0}^{t-1} e_{\tau}\, Corr(P^{*}_{if\tau}, P^{*}_{ift})$$

Denne korrelasjonen kan vi anta berre avheng av kor stor del av panelet som er felles og av t-τ. Den blir då uavhengig av t og kan estimerast på dei tidsseriane vi har. Truleg vil den bli nesten neglisjerbar.

Tilsvarande finn vi J$\underline{ift}$.

$$J_{ift} \approx V_{it} \sum_{\tau=0}^{t-1} e_{\tau}\, Corr(P^{*}_{i\tau}, P^{*}_{it})$$

der V$\underline{it}$ er variansen til $P^{*}\underline{it}$. Då denne variansen er langt mindre enn V$\underline{ift}$, vil også J$\underline{ift}$ være langt mindre enn C$\underline{ift}$. I praksis kan vi derfor sjå bort frå denne kovariansen.

G$\underline{ift}$ kan estimerast med standard resultat for minste kvadraters estimatorar. Det gjenstår då berre å estimere H$\underline{ift}$. Då vi truleg kan dra nytte av observasjonane frå tidligare kvartal , er det rimeleg å tru at A$\underline{ii}$ er ei øvre grense for denne størrelsen. Elles kan vi observere at

$$E\,(P^{*}_{ift} - Q^{*}_{ift})^2 = V_{ift} + G_{ift} + H_{ift} + 2C_{ift}$$

Dersom vi antar at

$$H_{ift} = H_{i}$$

dvs. uavhengig av fylke og tidspunkt, kan vi estimere G_i med

$$H^{*}_{it} = F^{-1} \sum_{f=1}^{F} \{(P^{*}_{ift} - Q^{*}_{ift})^2 - V_{ift} - G_{ift} - 2C_{ift}\}$$

og

$$H^{*}_{i} = T^{-1} \sum_{t=0}^{T} H^{*}_{it}$$

4.5 Appendix

Bevis for teorem 2.

Set:

$$g(w_f) = E\,(\theta^{*}_{ft} - \theta_{ft})^2$$

$$= E\{\sum_i [(1-w_{if})(P^{*}_{ift} - P_{ift})R_{ift} + w_{if}((P^{*}_{it} - P_{it}) + (D^{*}_{ift} - D_{ift}))R_{ift}]\}^2$$

$$= \sum_i (1-w_{if})^2 R^2_{ift}\, E(P^{*}_{ift} - P_{ift})^2$$

$$+ \sum_i w^2_{if} R^2_{if}\, [E(P^{*}_{it} - P_{it})^2 + E(D^{*}_{ift} - D_{ift})^2]$$

$$+ \sum_i w_{if}(1-w_{if})R^2_{if}\, E(P^{*}_{it} - P_{it})(D^{*}_{ift} - D_{ift})$$

$$+ 2 \sum_{i<j} w_{if} w_{jf} R_{ift} R_{jft} E(D_{ift} D_{jft})$$

Når vi set inn

$$V_{ift} = E(P^{*}_{ift} - P_{ift})^2$$

og

$$B_{ift} = G_{ift} + H_{ift} = E(P^{*}_{it} - P_{it})^2 + E(D^{*}_{ift} - D_{ift})^2$$

$$C_{ift} = E(P^{*}_{ift} - P_{ift})(D^{*}_{ift} - D_{ift})$$

$$A_{ij} = E(D_{ift,\,jft})$$

og deriverer får vi:

$$\partial g/\partial w_{if} = -2(1-w_{if})R^2_{if} V_{ift} + 2 w_{if} R^2_{if} B_{ift}$$

$$+ (1-2w_{if})R^2_{if} C_{ift} + 2 \sum_{i\neq j} w_{if} R_{if} R_{jf} A_{ij}$$

Når vi set alle deriverte lik 0 får vi den ønska første ordens betingelse.

Det gjenstår å vise at g har eit minimum. Vi viser då at Hessematrisa H(g) er positivt definit. L x være ein vilkålig I-vektor x≠0. Set

$$z' = (x_1 R_1, \ldots, x_I R_I)$$

No er z≠0 og:

$$x'H(g)x = z'(A-\mathrm{Diag}(A_{ii}) + 2\,\mathrm{Diag}(V_{ift} + B_{ift} - C_{ift}))z$$

det siste leddet her er opplagt positivt når $2(V_{ift} + B_{ift} - C_{ift}) > A_{ii}$.

QED

Referanser:

Efron B. og Morris C. 1972: 'Limiting the Risk of Bayes and Empirical Bayes Estimators - Part II : The Empirical Bayes Case.' JASA 130-139.

Fay . og Herriot . 1979 :'Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data.' JASA Vol 74 269-277.

James W. og Stein C. 1961: 'Estimation with Quadratic Loss.' Proceedings of the Fourth Berkeley Symposium, Vol 1 , Berkeley: University of California Press. 361-79.

Morris C.N. 1983 : 'Parametric Empirical Bayes Inference: Theory and Application.' JASA Vol78 47-55 (Comments 55-65).

Stein C. 1955: 'Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution.' Proceedings of the Third Berkeley Symposium, Vol 1 , Berkeley: University of California Press. 197-206.

Tennenbein A. 1970: 'A Double Sampling Sheme for Estimating from Binomial Data with Missclassifications.' JASA 1350-61.