Ekonomisk statistik 4
Economic statistics 4

Masterkurs

Daniel Thorburn

Höstterminen 2008

Stockholms Universitet

Metoder för urval och estimation speciellt för företagsdata

*Indexteori
*Nationalräkenskaper
*Imputering
*Granskning
*Datainsamling

πps-urval
Samordnade urval
Permanenta slumptal
Hantering av outliers
Länkning

Urval ur företagspopulationer

- Företag och andra ekonomiska enheter brukar variera mycket i storlek och betydelse.
  - Två konsekvenser:
    - Populationerna är alltså mycket skevare
    - Man vill kunna välja ut stora företag med hög sannolikhet och små med liten.
- Företagspopulationer förändras snabbt
  - Större ramproblem

πps-urval

- Ofta delar man in företagen/ramen i tre grupper
  - Totalundersökt stratum, där alla väljs
  - Småföretag under cutoff-gränsen. (t ex < 10 anställda) Dessa väljs aldrig.
  - En mellangrupp med varierande urvalssannolikheter
- Urval med olika men förutbestämda sannolikheter för olika enheter brukar kallas πps-urval ((inclusion-) probability proportional to size).
  - Numera brukar det inte nödvändigtvis vara storlek utan kan vara andra lämpliga tal. (Jfr pps!)

πps-urval - beteckningar

- Inklusionssannolikheter är centrala vid design-baserad inferens.
- Första ordningens inklusionssannolikheter:

$$\pi_i = P(i \in S) = P(I_i = 1)$$

- Andra ordningens inklusionssannolikheter:

$$\pi_{ij} = P(i,j \in S) = P(I_{i,j} = 1)$$

speciellt: $\pi_{ii} = \pi_i$

πps-urval - skattningsformler

- Horvitz-Thompson (HT) estimator:

$$t_{y,HT} = \Sigma_S\, y_i/\pi_i = \Sigma_U\, I_i y_i/\pi_i$$

- Skrivs ofta: $t_{y,HT} = \Sigma_S\, \omega_i y_i$, där $\omega_i = 1/\pi_i$ kallas vikter (urvalsvikter)
- Väntevärdesriktig: $E(\Sigma_U\, I_i y_i/\pi_i) = \Sigma_U\, E(I_i) y_i/\pi_i = \Sigma_U\, \pi_i\, y_i/\pi_i = \Sigma_U\, y_i$
- Varians: $\Sigma\Sigma_{UU}\, ((\pi_{ij} - \pi_i\pi_j)/(\pi_i\pi_j))\, y_i\, y_j$
- Alternativ form:

$$\tfrac{1}{2}\, \Sigma\Sigma_{UU}\, (\pi_i\pi_j - \pi_{ij})\, (y_i/\pi_i - y_j/\pi_j)^2$$

  - Man ser att variansen blir noll om urvalssannolik-heten proportionell mot y.
  - Om de är mycket olika bör urvalen vara oberoende. (Jfr stratifierat urval)

πps-urval - variansskattning

- Variansskattning: $\Sigma\Sigma_{SS}\, ((\pi_{ij} - \pi_i\pi_j)/(\pi_{ij}\, \pi_i\pi_j))\, y_i\, y_j$
- Denna metod fungerar alltid för sannolikhetsurval (om $\pi_i > 0$)!

- Alternativ form: $\tfrac{1}{2}\, \Sigma\Sigma_{SS}\, ((\pi_i\pi_j - \pi_{ij})/\pi_{ij})\, (y_i/\pi_i - y_j/\pi_j)^2$
- kallas Sun-Yates-Grundy (SYG)-estimatorn och kräver att man har fix urvalsstorlek

- HT/SYG-estimatorerna är inte alltid de bästa totalsumme/varians estimatorerna men är för det mesta relativt bra eller åtminstone acceptabel för alla fall med fix stickprovsstorlek

- Första och andra ordningens inklusionssannolikheter bestämmer variansen
- När man utformar en urvalsprocedur skall man försöka ta hänsyn till båda. Tumregel
  - Välj i första hand π_i proportionella mot y_i.
  - Försök i andra hand att få π_{ij} nära $\pi_i\pi_j$ (dvs oberoende) då y_i/π_i skiljer mycket åt från y_j/π_j

- Många pratar dock om πps-urval utan att bekymra sig om andra ordningens inklusionssannolikheter.
- En relativt bra metod är systematiskt πps-urval

Neyman-allokering

- En relativt bra metod är att stratifiera efter önskad urvalssannolikhet och så att relativt få dras i varje stratum. Vid stratifierat urval bör man välja ut enheterna med sannolikheten proportionell mot $N_i \sigma_i / c_i^{½}$, där parametrarna är stratumstorlek och stratumvarians och förväntad kostnad per enhet.
- Om detta tillämpas på pps-urval och antar följande modell $E(Y_i|X_i=x_i) = f(x_i)$ och $Var(Y_i|X_i=x_i) = \sigma^2(x_i)$ och att olika enheter är oberoende av varandra givet X, så bör man välja ut enhet i med en sannolikhet proportionell mot $p_i = \sigma(x_i)/ c_i^{½}$, där c_i är den förväntade kostnaden (för insamlings-organisationen och uppgiftslämnaren)..

Hur gör man πps-urval i praktiken

Det finns två i princip olika sätt att dra urval med givna inklusionssannolikheter

- Härma obundet slumpmässigt urval. För inte in mer struktur en vad som minimalt krävs. Dvs försök få π_{ij} då nära $\pi_i*\pi_j$ som möjligt. Om en sådan procedur tillämpas på fallet lika sannolikheter vill man få OSU.
- Använd all tillgänglig information. Även inklusionssannolikheterna kan innehålla mer information. (T ex som kan användas för stratifiering eller ordna för systematiskt urval).

Att dra πps-urval
Problem med enkla förslag!

- Det är enkelt om urvalsfraktionen är liten. Då kan man dra med återläggning och bortse från möjligheten att få samma enhet två gånger.
- Lahiris metod säger att man skall dra med återläggning men om samma enhet finns två gånger skall den också få dubbla vikten i skattningen. Tyvärr är det en ineffektiv metod.

- Men utan återläggning blir de flesta naturliga metoder fel. Vi skall se några exempel

Drag 2 eheter från en population med 5 enheter med sannolikheterna
0,7, 0,3, 0,5, 0,2 och 0,3.

- Första försöket. Drag först en med halava denna sannolikhet 0,7/2=0,35, 0,15, 0,25, 0,1 0ch 0,15. Drag sedan den andra med sannolikheten proportionell mot p för de kvarvarande. Det ger $p_1 = 0.35+2*0.15*0.7/1.7+0.25*0.7/1.5+0.1*0.7/1.8 = 0.629$
- Andra försöket. Betingad Poisson. Gå igenom enheterna en och drag dem med sannolikheten p oberoende av varandra. Om stickprovet innehåller för många eller för få enheter förkasta det och försök egen nytt stickprov. Det ger
  $p_1 = P(1€S|2$ enheter dragna) = 0.7*(2*0.3*(1-0.5)*(1-0,2)*(1-0,3)+ (1 -0.7)*0.5*(1-0.8)*(1-0.7)+(1-0.7)*(1-0.5)*0.2*(1-0.7)/ (0.7*(2*0.3*0.5*0.8*0.7+0.7*0.5*0.8*0.7+0.7*0.5*0.2*0.7)+2*0.3*0. 3*(0.5*0.8*0.7+0.5*0.2*0.7+0.5*0.8*0.3)+0.3*0.7*0.5*0.2 *0.7) = 0.744
  Vid stora stickprov kan det ta lång tid.

Tredje försöket.

- Order sampling (Distributional Poisson)
  - Drag ett likformigt tal U_i för varje enhet.
  - Transformera på något sätt $W_i = g(\pi_i , U_i)$ så att $P(W_i > 1) = \pi_1$.
  - Välj ut de n enheter som har störst transformerat tal.

  Inte exakt korrekt men nästan för lämpligt valt g. (Det bör bli $\Sigma_S \pi_i$ = n enheter större än 1.)

- vid Pareto Poisson väljs $g(\pi_i , U_i) = \pi_i(1- U_i)/ U_i(1- \pi_i)$. Då följer W_i en Pareto fördelning. (Metoden har vissa optimalitetsegenskaper i order samplingskategorin. Den används vid SCB. Den ger dock inte inklusionssannolikheter som är exakt rätt, men i tillräckligt bra).
  - Exempel: Drag ett stickprov med n=2 med inklusionss-sannolikheterna 0,7, 0,3 0,5, 0,2, 0,3
  - Ta fem slumptal: 0,825, 0,168, 0,254, 0,688, 0,902
  - Beräkna g(.): 0,50 (0,175*0,7/0,825*0,3), 2,12, 2,96, 0,14, 0,05
  - Välj de två största. I detta fall nummer 2 och 3

Andra metoder som härmar OSU

- Många andra metoder har föreslagits t ex Sampford (som kanske är den mest använda, används t ex i USA vid inte alltför stora stickprov), Hajek, Sunter. Alla dessa är rätt komplicerade.
- Några personer har som mål att uppnå högsta tänkbara ”entropy”, dvs lägga in maximal mängd slump givet inklusionssannolikheterna. (Sampford ger maximal entropy). Det finns de som hävdar att maximal entropi ger den bästa variansskattningen.

Sampford

- Ta en enhet med sannolikhet proportionell mot π_i (dvs med sannolikhet $\pi_i/\Sigma \pi_j$)

0,35, 0,15, 0,25, 0,1, 0,15

Säg att vi får nr 3

- Ta n-1 enheter med återläggning med sannolikheter proportionella mot $\pi_i/(1-\pi_i)$ (dvs med sannolikhet $\pi_i/(1-\pi_i) / \Sigma (\pi_j/(1-\pi_j))$

Proportionellt mot 2,33, 0,43, 1, 0,25, 0,43

dvs. 0,52, 0,10, 0,22, 0,06, 0,10

Säg att vi får nr 1

- Om stickprovet nu innehåller exakt n olika enheter detta är vårt slutliga stickprov.

Stickprovet innehåller 3 och 1, evs två olika. Vi stannar här

- Annars börja om från enhet 1.
- Gör om tills man får ett stickprov med rätt storlek

- Det är inte enkelt att visa att detta ger rätt inklusionssannolikheter, men det gör det.
- För stora stickprov kommer metoden att bli alltmer ineffektiv eftersom det blir för många försök innan man får ett stickprov av rätt storlek.

Det finns mer struktur (eller härma inte OSU)

- Ibland vill man utnyttja sin information bättre. I dessa fall är systematiskt πps-urval bra. Det ger en mycket bra täckning över populationen, men är inte alltid ett riktigt sannolikhetsurval.
- Stratifierat urval med små strata är ett annat alternativ (kan inte göras för alla inklusionssannolikheter, men bra som en approximation)
- Systematiskt urval. Ordna enheterna i urvalet på något lämpligt sätt (t ex geografiskt, någon viktig variabel i ramen eller urvalssannolikheterna.

Systematiskt Πps-urval

- Ordna enheterna på något vettigt sätt efter den information man vill utnyttja (t ex geografiskt, någon viktig bakgrundsvariabel eller urvalssannolikheterna själva)
- Beräkna de kumulativa inklusionssannolikheterna $\Pi_k = \Sigma_{i<k+1} \pi_i$
- Välj en godtycklig, U, startpunkt likformigt i $(0, \Pi_N/n)$
- Välj ut alla enheter i på avståndet Π_N/n (dvs alla enheter där den kumulerade summan $\Pi_{i-1} < U + k\Pi_N/n < \Pi_i$) för något $0<k<n$.
- Denna procedur ger en bra spridning över den variabel man ordnat efter och ger därför nästan alltid lägre varians än vanliga pps-metoder (om det inte finns periodicitet)

Systematiskt urval

- Variansen under systematiskt urval kan inte skattas väntevärdesriktigt
- Den variansen man skulle fått med vanliga pps-metoder kan däremot lätt skattas (tom bättre än med PPS-urvalet)
- Det är känt att det är en överskattning, så alla intervall kommer att bli konservativa. (Täcker alltid rätt värde med en sannolikhete som är större än angiven konfidensgrad).
- Variansen kan skattas om man istället tar t ex M oberoende systematiska stickprov, dvs M startpunkter $< Mk\Pi_N/n$ och sedan enheter på avståndet $Mk\Pi_N/n$.

Samordnade urval

- Typ av samordning
  - Positive coordination with large overlap
  - Negativt koordination betyder lite överlapp mellan undersökningar
- Varför samordna urval?
  - Uppgiftslämnarinsatser
    - Fördela den mer jämnt
    - Minska dem
  - Få information om samband
    - Över tiden, longitudinella studier
    - Mellan undersökningar. Om samma företag deltar i två olika undersökningar kan man studera samband, t ex mellan lönsamhet och arbetsmiljö.

Longitudinella studier

Longitudinella studier är en samlingsnamn för studier där man följer samma enheter över tiden

Exempel: Roterande urval

Varje företag är med i studien ett i förväg bestämt antal gånger t ex fyra år och en fjärdedel byts varje år

- Det minskar uppgiftslämnarbördan
  - Första gången man är med är alltid jobbigast
  - Basfrågor ställs endast en gång
  - Men samtidigt gör rotationen att ingen är med för alltid, vilket skulle anses orättvist
- Lägre spårnings- och kontaktkostnader efter första gången
- Det ger möjlighet att studera utvecklingen mellan åren

Fyra aktiva roterande paneler

Ett enkelt exempel

Time Panel	2000	2001	2002	2003	2004	2004	2005	2006	2007
A	X								
B	X	X							
C	X	X	X						
D	X	X	X	X					
E		X	X	X	X				
F			X	X	X	X			
G				X	X	X	X		
H					X	X	X	X	
I						X	X	X	X
J							X	X	X
K								X	X
L									X

Roterande paneler med k aktiva paneler

- Låt X_{ti} vara medelvärdesskattningen från den i:te panelen
- Antag att dessa skattningar har variansen σ^2 och att korrelationen mellan tidpunkter inom panel är $\rho^{|t1-t2|}$
- En enkel skattning av medelnivån vid tidpunkt t är då medelnivån i alla paneler $\Sigma_i X_{ti}/k$ med variansen σ^2/k
- Variansen för skillnaden mellan två följande tidpunkter, t och t+1, blir då $2(1 - ((k-1)/k) \rho) \sigma^2/k$
- Slumpfelet minskar med antalet paneler, dvs rotationsperioden. Variensen utan överlapp skulle varit $2 \sigma^2/k$
- T ex med k = 4 och $\rho = 0.9$ blir vinsten en faktor 0.325

- Men man kan göra något ännu bättre (men det görs sällan)
- Skillnaden mellan första och andra tidpunkten kan skattas på två sätt:
  - Skillnaden mellan de k-1 gemensamma panelerna $D_1 = \Sigma_{i=2}^{k} (X_{2i} - X_{1i})/(k-1)$ med varians $2(1-\rho) \sigma^2/(k-1)$
  - Skillnaden mellan dn nya och gamla panelen $D_2 = (X_{2k+1} - X_{11})$ med varians $2\sigma^2$.
- Om dessa vägs samman optimalt får man $(D_1 + (1-\rho)/(k-1) D_2)/(1 + (1-\rho)/(k-1))$ with the variance $2\sigma^2/(1 + (k-1)/(1-\rho))$.
- Med k = 4 och = 0.9 blir vinsten 0.129

- Varför tror ni att detta sällan utnyttjas?

- Man vill inte ändra skattningar som en gång publicerats.
- Och då är det naturligt att vilja skatta nivån ett visst år från de värden som samlats in det året
- Man vill ha konsistens dvs nivåförändringen skall vara lika med skillnaden mellan nivåskattningarna. Men som synes tappar men precision på detta

Underurval - Screening

- Ibland samordnar man urval genom att undersöka ett underurval
  - t ex vart femte företag får dessutom svara på några extrafrågor, om urvalet i det andra ledet inte behöver vara så stort. I skattningsfasen bör man här utnyttja att det utgör ett delurval. Man kan t ex kalibrera efter några viktiga basfrågor.
  - Screening. Man vill specialundersöka företag med vissa egenskaper t ex som genomfört arbetsmiljåtgärder eller har kvinnlig VD. I huvudundersökningen har man en fråga om detta - och sedan återkommer man med nästa. Planerade urval välj lika många med kvinnlig som manlig VD. Leder till effektivare jämförelser

Permanenta Slumptal

Alla företag i en ram (t ex Centrala Företagsregistret) ges ett rektangelfördelat slumptal, så fort det kommer in i registret, U_i, (i intervallet (0,1)). Detta slumptal behålls så länge företaget finns kvar.

Ett enkelt sätt att dra ett urval är då att ta alla företag med slumptal i intervallet (a,b). Urvalet täcker då (ungefär) andelen b-a av populationen, och är ett OSU oberoende av när vi drar urvalet.

Det är enkelt att ta urval med valfri grad av överlappning. Välj bara intervallen lämpligt överlappande.

- Man kan t ex välja disjunkta urval då deltar varje företag i högst en undersökning.
- Genom att ta ett annat intervall t ex ((a+b)/2, c) får man t ex med hälften av företagen i den första undersökningen

Permanenta slumptal för roterande urval

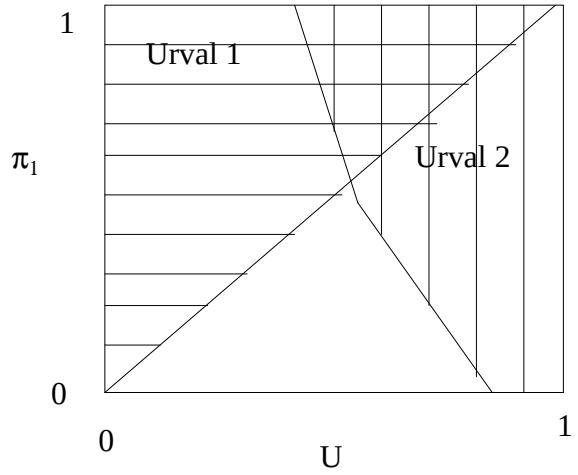
- Ett problem vid roterande urval är att en del av urvalet är draget ur en gammal ram.
  - Ett år gamla företag kan bara finnas med i senaste panelen
  - Men vi vill att urvalet vid varje tidpunkt skall vara representativt för dagens population. Detta löser man idag med permanenta slumptal
- Man kan flytta intervallet en lämplig bit åt höger (t ex (b-a)/4) varje år.
  - Då får man ett roterande urval.
  - Och genom att ramen uppdaterats mellan åren får man automatiskt med rätt andel nya företag. (Vissa av de nystartade kommer dock bara med kortare tid än 4 år)

Permanenta Slumptal

- Denna diskussion gällde OSU. Men vad göra vid πps-urval?

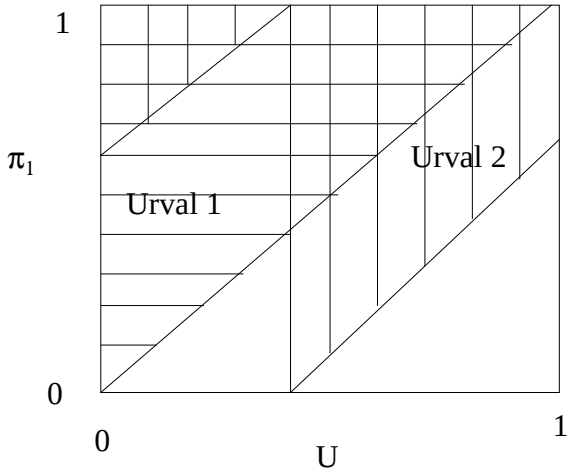
Permanenta slumptal

Samordnade urval med varierande urvalssannolikheter
Svårt rita i tre dimensioner - därför bara två - men ofta beror urvalssannolkiheterna av samma variabel



Permanenta slumptal

Samordnade urval med varierande urvalssannolikheter
Samma urvalssannolikhet



Permanenta slumptal

- Använd Pareto πps när man vill dra med varierande sannolikheter
  - beräkna $\pi_i(1-U_i)/ (U_i(1-\pi_i))$ och ta de n största (Ungefär de med ett värde större än 1)
  - Nästa undersökning. Ersätt U_i med ett annat tal
    - t ex $U_i - \pi_{i1}$. ($U_i - \pi_{i1}$ +1 om $U_i - \pi_{i1}$ är negativ) och använd i formeln $\pi_{i2}(1-U_i + \pi_{i1})/ ((U_i - \pi_{i1})(1-\pi_{i2}))$. (Även $U_i - \pi_{i1}$ är ett likformig fördelat slumptal). Detta ger inget överlappning. Med $(U_i - \pi_i/4)$ får man en rotation på fyra år.
  - I allmänhet gör man det enklare för sig genom att bara ändra startpunkt $U_i - b$, där b väljs större än de flesta urvalssannolikheter
  - Ibland ändrar man tecken istället U_i -> 1- U_i

Föränderliga populationer

- Är alltid besvärliga.
- Registret uppdateras ständigt. Om man gör ett antal undersökningar under året så har man olika populationer
- Om vi då skattar samma variabler t ex förra årets omsättning i branschen så blir skattningen olika beroende på att populationen förändrats. Vi kan alltså få flera olika skattningar av 2007 års omsättning.
- Man diskuterar olika sätt att lösa detta t ex genom att
  - dra stickproven samtidigt för undersökningar som går vid olika tidpunkter
  - justera uppräkningsfaktorerna (kalibrering) så att skattningen alltid blir rätt

Outliers

- Outliers är värden som kraftigt avviker från vad man väntat sig.
  - De kan vara fel (bra granskningsprocedurer behövs)
  - De kan vara riktiga - men påverkar skattningen orimligt
- Exempel:
  - I registret finns ett brevlådeföretag med urvalssannolikhet 1/1000. Det blir utvalt.
  - Det har under året gjort en nyemission på 100 miljoner och lånat upp 900 miljoner samt köpt ett existerande pappersbruk och fått en omsättning på 1200 miljoner.
  - När totalomsättningen skattas bidrar detta företag (som har uppräkningstalet 1000) med 1200 miljarder dvs 1 fjärdedel av Sveriges BNP..

Vad gör man då?

- Det vanligaste är att hantera detta är trimning (trimming) dvs att låta företaget få uppräkningstalet 1 (dvs bara representera sig själv. För att få totalantalet företag rätt justerar man även vikterna för övriga i urvalet).
  - Ett problem är att avgöra när denna lösning skall tillgripas.
  - Ofta knyter man det till hur stor del av totalskattningen som företaget svarar t ex om enheten svarar för mer än 5% av totalen eller om enheten svarar för mer än hälften av stratumtotalen. Detta kan göras mekaniskt och därför i någon mening objektivt.
  - Det vanligaste är dock att fatta beslutet subjektivt
- Ett annat sätt är winsorisering (Winsorizing) dvs dela företaget i två
  - ett med uppräkningsfaktorn 1
  - ett med uppräkningsfaktorn w-1 som har y-värdet lika med t ex övre 5% percentilen i stratat

Fler förslag

- Ett tredje sätt är att säga att om företaget i ramen sett ut som det gör nu hade det fått urvalssannolikheten π. Därför skall den nu ha uppräkningstalet $1/\pi$ (och justerar övriga så att totalen blir rätt).
- Alla dessa sätt ger i medeltal för små skattningar (bias) - alla justeringar sker neråt. Men i allmänhet blir medelkvadratfelet minst med den första metoden.
- Det finns förslag på att jämna ut över åren. Dvs se vad som skulle ha hänt under en tioårsperiod över hela näringslivet och sedan justera varje års skattningar med detta. (Det innebär att nivån kommer att korrigeras uppåt litegrann när inga outlier finns. De skulle ju kunnat göra det). Man kan då få väntevärdesriktighet över en längre tidsperiod.

- Ett fjärde sätt är att inte använda designbaserade skattningsmetoder i fördelningssvansen (övre 5 eller 10 procenten eller ungefär 10-20 observationer) utan modellbaserade. Använd t.ex. Pareto, Weibull eller lognormalfördelningen, som alla brukar vara bra för skeva populationer. Detta skall göras vare sig det finns outlier eller ej.
- $(Y_i|Y_i > c)$ € Pareto(a,b,c). Täthet:

$$f(x) = b\frac{(c-a)^b}{(x-a)^{b+1}} \quad x > c > a$$

- c är cutoffgränsen (eller skattas med $\min(x_i)$). Om a är känt så ges ML-skattningen av b av $\frac{n}{\Sigma \ln(x_i - a) - \ln(c - a)}$ (vid lika urvalssannolikheter).
- Även a kan lätt skattas numeriskt med ML-metoden (eller t o m sättas till noll)

När väl parametrarna är skattade får man skattningen av totalen av populationen genom att ge observationer

- mindre än c* sin vanliga vikt
- större än c* vikten 1
- Återstående vikt blir summan av vikterna för observationer tal större än c - antalet större än c. Dessa beskrivs som oberoende dragningar från (i exemplet) Pareto(a*,b*,c*), dvs
  - med väntevärdet $E(Y|a^*, b^*, c^*) = c^* + (c^* - a^*)/(b^*-1)$
    - lika många som antalet observationer, n', större än c och med vikterna $1/\pi$ - 1.
  - med ungefärlig (modell)-varians kring detta värde med

$$\text{Var}(Y|a^*,b^*,c^*) = \left(\frac{1}{n'} - \frac{1}{\pi n'}\right)\frac{b^*(c^*-a^*)^2}{(b^*-2)(b^*-1)^2}$$

Länkning

Ekonomisk statistik

Höstterminen 2008

Stockholms Universitet

Länkning

- När statistiken läggs om uppstår ofta språng, t ex. när
  - näringsgrensindelningen moderniseras
  - man byter insamlingssystem t ex från disketter till Internet
  - en cutoff-gräns ändras.
  - Man formulerar om en fråga
- Vid länkning försöker man räkna om serien bakåt som om det nya insamlingssystemet hade gällt. Bra för personer som arbetar med tidsserier.
- Samma problem kan uppstå vid ändringar i samhället.
  - Följa sjuklighetens utveckling i samhället, när antalet karensdagar ändras.
  - Vad händer med data över hushållens förmögenhet när förmögenhetsskatten avskaffas?
- Vid omläggningar anses länkningen inte alltid som en uppgift för statistikproducenten utan snarare tidsserieanalytikern. Men länkning vid metodskift är något för statistikproducenten.

- Vid omläggningar bör statistikern planera för länkning. Vid reformer i samhället bör politikerna planera så att reformen kan utvärderas.
- Det finns två typer av länkning (med mellanformer)
  - Mekanisk länkning (Den enda källa du har är serien själv och det gäller att uppskatta språnget)
  - Länkning med hjälp av speciella data
    - Specialundersökningar
    - Andra relevanta data
    - Gamla rådata, blanketter

Mekanisk länkning ?

Tänk er att man har en serie som ser ut så här

(Andel positiva till statliga investeringsgarantier %.)

År 1 2 3 4 5 6 7 8 9 10 11 12 13

23 24 27 27 26 28 30 42 43 45 44 43 47

År 8 lades undersökningen om och frågan formulerades om. Om man vill analysera utvecklingen vill man ha en serie utan språng. Hur skall man göra?

•

- Svar 1 o 2: Enklast är att sätta värdena år 7 och 8 lika och ändra värdena innan motsvarande (additivt resp multiplikativt).

År	1	2	3	4	5	6	7	8	9	10	11	12	13
Urspr.	23	24	27	27	26	28	30	42	43	45	44	43	47
add	35	36	39	39	38	40	42	42	43	45	44	43	47
Mult	32.2	33,6	37.8	37.8	36.4	39.2	42	42	43	45	44	43	47

Problem: Detta säger inget om skillnaden mellan år 7 och 8.

- Svar 3: (För enkelhets skull ges lösningen bara additivt)

Mer avancerat gör regression $y_t = a + b*I(t<8)+c*t$ och justera med b

34,9 35,9 38,9 38,9 37,9 39,9 41.9 42 43 45 44 43 47

Andra modeller kan användas. Regression mot en logistisk linje hade förmodligen varit ännu bättre.

$p_t = \exp(a + b * I(t<8) + c * t)) / (1+ \exp(a + b * I(t<8) + c * t)))$

$y_t = p_t + \varepsilon_t$; $\quad Var(y_t) = d * p_t *(1- p_t)$

Problem: Man vill kunna göra länkningen redan första året efter tidsseriebrottet (dvs år 8).

- Svar 4: Använd samma modell redan år 8 och håll sedan modellen fix

  34,6 35,6 38,6 38,6 37,6 39,6 41.6 42 43 45 44 43 47

Ger lite större osäkerhet än svar 3.

- Svar 5: Ännu mer komplicerad modeller kan användas t ex en ARIMA-process eller en dynamisk modell (Kalmanfilter). T.ex
  - $X_t = a*X_{t-1} + \varepsilon_t$ X är en latent serie som mäts med slumpfel:
  - $Y_t = X_t + \eta_t$ före brott; $Y_t = X_t + b + \eta_t$ efter brott;
  - Y + Predikterat värde av b redovisas före brottet
  - Den dynamiska modellen ger
  - 34,9 35,9 38,9 38,9 37,9 39,9 41,9 42 43 45 44 43 47

- Mekaniska modeller som 3-5 ger möjlighet att ge ett osäkerhetsintervall för justeringens storlek. I så fall bör man ha en relativt flexibel modell som den dynamiska modellen.
  - Medelfel i språnguppskattningen i vårt exempel var 1,8 i svar 5.
  - Osäkerhet i språnget blir +/-3,6.
- Det är dock mycket ovanligt att man anger osäkerheten i uppskattningen av länkningsfelet, när den görs mekaniskt, som här.

- Men i allmänhet försöker man välja så enkla modeller som möjligt.

Hjälpserier

- I diskussionen ovan antogs att man inte hade någon annan kunskap. Ofta har man det.
- T ex kan man ha tillgång till serier för en eller flera andra relaterade variabler (I vårt fall t ex branschindelning, partisympatier eller investeringsvilja). Då kan dessa variabler också användas i regressionsmodellerna. (eller göra om ARIMA-modellerna till VARIMA)
- Ibland kan denna typ av modeller leda till att man inte justerar alla åren bakåt i tiden lika mycket. Man kan t ex komma fram till att alla moderater skall justeras med 7%, övriga borgerliga med 10% och övriga med 12%. Sedan justerar man bakåt med de tidigare observerade partisympatierna. Eller efter branschernas historiska andel av totalen.

Länkning med andra data

Allt tidigare var s.k. mekaniska länkningsmetoder. Om möjligt bör man göra specialstudier vid alla större omläggningar.

- En vanlig rekommendation är att om det är möjligt göra undersökningen på båda sätten parallellt vid ett tillfälle
- T ex När ULF (undersökningen om levnadsförhållanden) gick över från besök till telefon-undersökning som bas gjorde man under övergångsåret halva urvalet besök och halva per telefon. (Delarna valdes slumpmässigt, men varje intervjuare fick bara använda en metod, ny eller gammal).
- Men vanligare är att man gör mindre specialundersökningar av olika typer.
- Ibland har man tom gjort experiment i förväg för att finna en bra ny uppläggning. Då kan man i underlaget för beslutet om omläggning även ha tillräckliga data för att uppskatta språngets storlek.

Länkning med underlag

- ”Inbäddade experiment” är en speciell typ av studier. De görs som en del i den normala statistikproduktionen
- När uppskattningen av hoppet görs med specialundersökningar är det vanligt med feluppskattningar
- När större omläggningar av klassificeringar görs, t ex av SNI-kod försöker man
  - koda uppgifterna på båda sätten under en period eller
  - gå igenom gamla statistikunderlag (blanketter) och koda om företagen enligt den nya mallen och därmed få jämförbara data för tidigare år.

- Dessa metoder ger en uppskattning av språnget - men inte över hur man skall ändra längre bakåt
- Vid bl a omklassificering, t ex omläggning av SNI, kanske man vet summan av två grupper men det tillkommer en uppdelning.
  - T ex när Dator och IT-konsulter bröts ur kategorin annan konsultverksamhet. Då visste man att den inte under de senaste tio åren haft samma andel av den större kategorin.
- Då gäller det att rekonstruera en uppdelning bakåt. Om man inte har data lägger man ibland in en subjektiv gissning. T ex andel 0 tio år tidigare och sedan en linjär ökning av andelen till den första observerade nivån. Eller samma andel hela tiden bakåt (Dumt i just detta fall)
- Ett problem med länkning är att serien blir utjämnad. (Inte för producenten men ekonometrikern, som använder data). Ekonometrikern kommer att underskattar slumptermen storlek när han sedan använder t ex vid ARIMA-modeller. Det betyder bl a att hans prognoser blir för säkra.

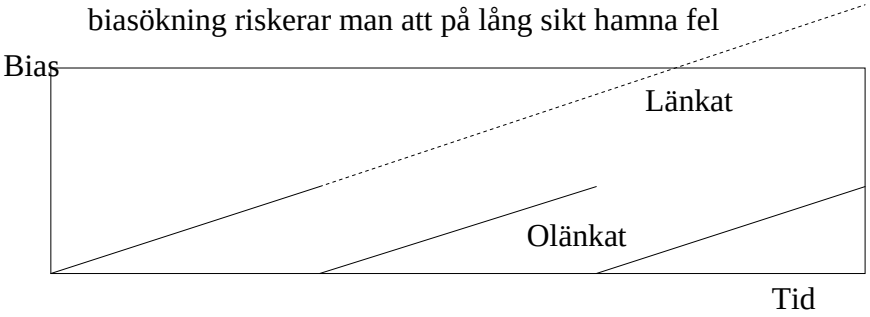
Flera serier

- Vid flera omlagda serier som skall summera till en total kan man
  - länka delserierna och sedan summera
  - länka totalerna och delserierna var för sig
    - Då får man inte konsistens
  - länka totalen och sedan länka delserierna
    - med bivillkor att länkningen skall vara konsistent
- Man talar om indirekt och direkt länkning
- Vilket som bör väljas beror på vad man tror om stabiliteten
  - t ex om man har två serier varav en lagts om och inte den andra och man skall göra mekanisk länkning är det alltid bättre att länka den serie som förändrats
  - Men om man har en total som beror av två delar och omläggningen främst består av omklassificeringar så bör man länka totalen först.

- Även vid flera serier är det ibland lämpligt att använda modeller. Vi diskuterade univariata serier - men man kan lika gärna använda multivariata modeller.

- Om man gör specialstudier av länkningsfelet för några delserier är det naturligt att länka dessa delserier först. Det är inte givet hur studierna skall användas när man länkar de andra delserierna eller totalen. Men ofta kan man utnyttja informationen från de studerade delserierna även för övriga.

- I praktiken görs ofta små ändringar i statistikinsamlingen. I princip skulle de kvalificera för en länkning, men om effekten är liten struntar man ofta i det. Felet i uppskattningen av språnget blir större en språnget själv.
- Ofta kan man säga att när man gjort en stor omläggning innehåller studien bara små fel. Sedan förändrar sig verkligheten och studiens relevans blir sämre och sämre.
- Efter tio år, säg, blir man tvungen att göra en ny stor omläggning. Om man bar justerar för en långsam biasökning riskerar man att på lång sikt hamna fel



- Om man länkat en dataserie, skall den som använder serien alltid kunna läsa om vad som hänt. Om orsaken till språnget och hur det eliminerats.

- Dokumentera vad du gör. Detta är en allmän regel, som man alltid säger, men många slarvar med.