Ekonomisk statistik
Economic statistics

Masterkurs

Daniel Thorburn

Höstterminen 2010

Stockholms Universitet

1 Sampling
1.1 Allmänt om urval

Daniel Thorburn

Ekonomisk statistik

Höstterminen 2010

Sampling

- Man har tillgång till ett register ”en ram” över alla intressanta element (t ex företag) i en population.
- Där kan finnas bakgrundsvariabler (X) (t ex förra årets omsättning enligt deklarationen eller inbetald preliminärskatt)
- Man är intresserad av några andra uppgifter (Y). (t ex orderingång eller miljöinvesteringar). Syftet är i allmänhet att beräkna summan av alla företagens värde på Y.
- Välj ut en del av företagen i registret (med hjälp av X) och ta reda på deras värden på Y.
- Hur skall urvalet dras? Hur skall uppgifterna sedan sammanställas.

- Sannolikhetsurval
  - T ex stratifierade urval
- Balanserade urval
  - T ex systematiska urval
- Bekvämlighetsurval
  - T ex Cut off gränser eller bara de som är villiga att svara
- Representativa urval
  - Inexakt ord för bra urval som ger väntevärdesriktiga skattningar
- Vetenskapliga urval
  - T ex sannolikhetsurval men också flera andra t ex matched pairs etc

- Sannolikhetsurval
  - Varje element i populationen har en positiv sannolikhet att bli vald
  - Varje element i stickprovet har känd urvalssannolikhet.
- Då kan man väntevärdesriktigt skatta totaler och medelvärden
  - Ibland varje par av element i stickprovet har känd positiv sannolikhet att ingå båda två (och normalt varje egenskap t ex median skattas konsistent)
- Då kan också variansen skattas

- Tänk på ramen!
- Ekonomer vid universitet drar ofta från börslistan. Men är detta en bra ram?
  - Jag såg nyligen ett projekt i Lund där man studerade ”Corporate management in Sweden” och hade valt denna ram (large cap och midcap).
  - Nämn några stora svenska företag som inte kommer med då

Mer komplicerade ramar

- Multiframe (man har flera listor och drar från alla. Samma företag kan finnas på flera listor. Ibland t o m har en lista arbetsställen och en annan juridiska enheter. Då kan det bli komplicerat
- Hierarkisk sampling (Clusterurval) Välj juridiska enheter, lista alla arbetsställen, välj på den listan. Då kan man bara räkna ut urvalssannolikheter för valda enheter
- …

1.2 Urval – ramar, företagsregister

Daniel Thorburn

Ekonomisk statistik

Höstterminen 2010

Företagsregister -
Centrala Företagsregistret

- 2007 fanns det 945801 företag med 1021083 arbetsställen.
- 538 101 var fysiska personer och 270 084 aktiebolag
- Resten är en blandning av handelsbolag, ideella föreningar, bostadsrättsföreningar, kommuner, stiftelser, ekonomiska föreningar,, utländska juridiska personer etc.
- De huvudsakliga källorna till SCBs företagsregister är skattelagstiftningen
- När skattereglerna ändras ändras antalet företag i Sverige. 1996 sänktes gränsen för att vara registrerad i MOMS-registret till minst 1 krona i moms.
- Då ökade antalet företag med ungefär 200 000.

Företagsregister

- Vad är ett företag? Ekonomisk enhet, Juridisk enhet, Koncern, Arbetsställe eller …
- Hur upptäcka nystartade företag? Veta när ett företag är nerlagt? Om ett företag blir uppköpt upphör det eller skall det anses leva vidare?
- Företag är branschklassificerade. Maximikriteriet efter omsättning. Om huvuddelen av verksamheten är bilar anses hela företaget vara ett bilföretag även om man har en stor dataavdelning och även är Sveriges största företag på flyg- och marinmotorer.
- Uppdatera branschändringar. Ex: Man vill göra en undersökning av företag i t ex IT-branschen. Företag som klassats som IT väljs ut. De som bytt bransch från IT tas bort - men de som bytt till IT hittas inte.

Datakällor

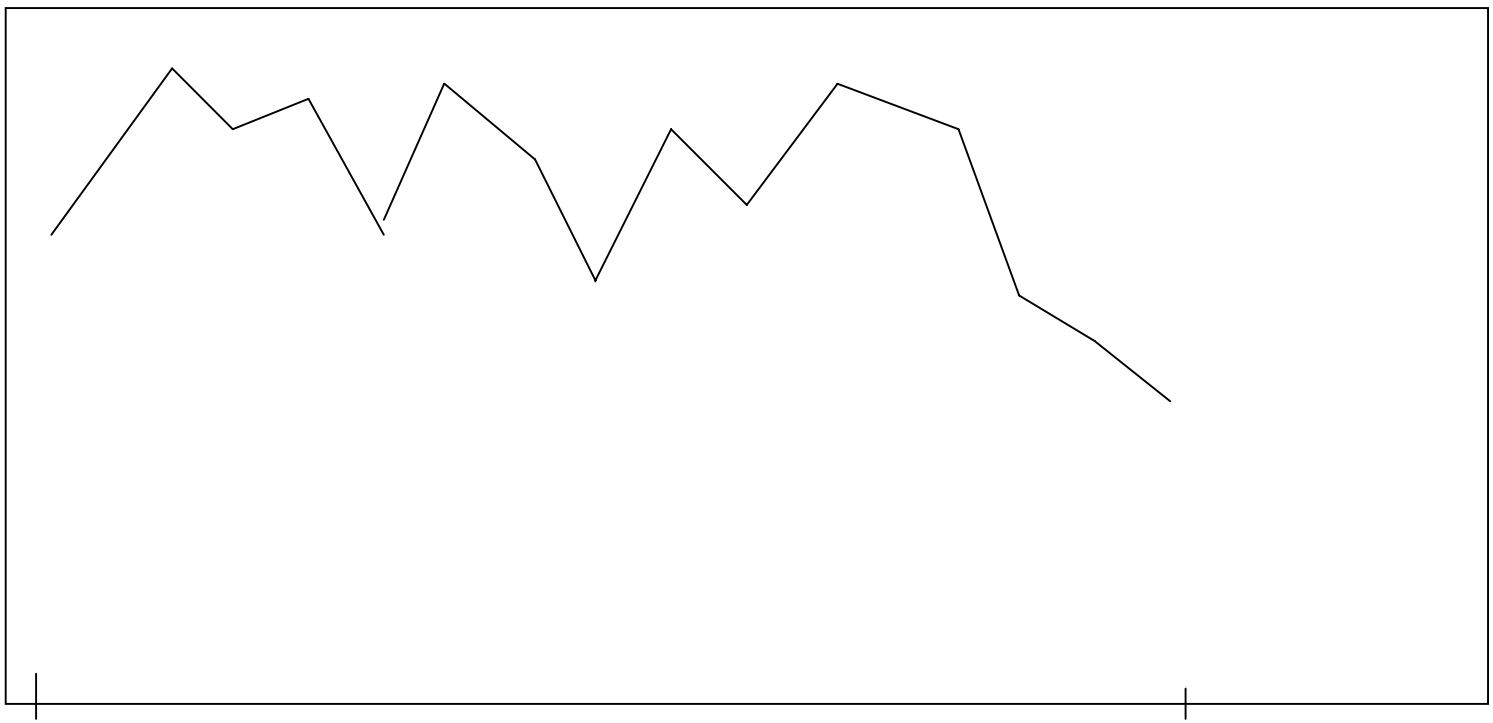
- Register
- Momsuppgifter
- Löneuppgifter (arbetsgivaravgifter, preliminärskatt)
- Deklarationer, bokslut, årsredovisningar
- Intrastat (Utrikeshandel)
- Urvals- och andra specialundersökningar

Registret uppdateras löpande. Om man i en studie upptäcker t ex ett branschbyte förs det omedelbart in. Det betyder att uppgifterna blir aktuellare för stora företag. Det betyder också att ordningen på undersökningar kan påverka resultatet. Studie av kemisktekniska branschen först och sedan verkstadsindustrin. I den andra kommer som bytt inriktning från kemiskteknisk industri med. Medan de som bytt till kemiskteknisk industri från verkstad kommer inte med i någon.

Observera att registret 1 sept 2009 kan ha minst två betydelser

- Vi tar gör statistik på dem som detta datum finns i registret
- Men registret lever vidare och uppdateras. Företag som nystartas i augusti kommer inte in förrän senare. Branschbyten, nedläggningar, fusioner etc kommer också till
- Om vi i december gör statistik på alla som enligt registret fanns 1 september, har många ändringr hunnit föras in och statistiken blir annorlunda

Typisk bild över antal nystartade företag i Sverige månadsvis
Varför denna minskning i slutet?



- ”Stora företag kan inte vara nystartade”. Det visar sig oftast vare t ex en avknoppad verksamhet eller en ny ägare.

- Det gör att SCB speciellt noga utreder alla stora som uppger sig vara nystartade

- Statistiken på stora nystartade företag får därför ännu större fördröjning (kan vara upp till tre år). Men företaget finns naturligtvis med men startår är okänt

Andra företagsregister

- Finns andra aktörer. Störst är
- RATOS-koncernen som har ett konglomerat av företag inom företagsinformationssektorn
  - Dunn och Bradstreet (kredituppgifter), MM-analys (företagsanalys t ex identifiera potentiella kunder eller växande företag där det kan vara värt att komma in och få leverera.
- Har en större databas än SCB
  - Säkert eftersom man köper hela CFAR från SCB
  - Har egen avdelning som stansar alla bokslut som kommer in till Patentverket. Innehåller även t ex uppgifter om styrelse och VD
  - De för på alla betalningsanmärkningar hos kronofogden,
  - Historiskt sett sämre - lyder under se KUL / PUL (kreditupplysnings/personuppgiftslagarna. SCB har lite större möjligheter genom statistiklagen)
- UC (Upplysningscentralen) stor inom kreditupplysning
  - Antalet konkurser tar de regelmässigt fram, som ni kanske hör i media

1.3 Urval och estimation – företagsurval, πps

Daniel Thorburn

Ekonomisk statistik

Höstterminen 2010

Urval ur företagspopulationer

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

Urval

Mycket skeva populationer - man väljer företagen med varierande urvalssannolikheter.

- Små företag med liten sannolikhet och stora med stor
- Den som väljs med sannolikheten 1 på 5 räknas upp med en faktor 5, den som väljs med sannolikheten 1 på 1000 räknas upp med faktorn 1000.
- Totalstratum
- Cutoffurval

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 π är centrala vid design-baserad inferens. Inklusionsindikator I=1 anger om enheten finns i stratum (I=0 annars)
- 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 urvalssannolikheten proportionell mot y (sista parentesen = 0).
  - Om de är mycket olika bör urvalen vara oberoende (första parentesen = 0). (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: $\frac{1}{2}\ \Sigma\Sigma_{SS} ((\pi_i\pi_j\ - \pi_{ij})/\pi_{ij})\ (y_i/\pi_i\ - y_j/\pi_j)^2$

  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 acceptabla 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 πps-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)..

$$t_{y,HT} = \Sigma_S \omega_i y_i$$

- När man har (total-)bortfall ändrar man ofta vikterna, så att vikterna blir större där bortfallet är stort

$$t_{y,HT} = \Sigma_S d_i y_i$$

- Det kallas kompensationsvägning
- Fungerar bra om man har ”MAR” (Missing at random). Men ofta beror bortfallet på viktiga företagsegenskaper och då blir det inte så bra.

1.4 Hur gör man πps-urval i praktiken

Daniel Thorburn

Ekonomisk statistik

Höstterminen 2010

1.4 Hur gör man πps-urval i praktiken

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

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

1. 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 enheter 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 halva 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 t ex 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 oberoende och med återläggning med sannolikheter proportionella mot $\pi_i/(1-\pi_i)$ (dvs med sannolikheten $\pi_i/(1-\pi_i) / \Sigma\ (\pi_j/(1-\pi_j))$

  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, dvs två olika. Vi stannar här

- Annars börja om från enhet 1. (dvs i detta fall t ex element 1 två gånger)

  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 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 πps-metoder (om det inte finns periodicitet)

Systematiskt urval

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