Ekonomisk statistik 2
Economic statistics 2

Masterkurs

Daniel Thorburn

Höstterminen 2008

Stockholms Universitet

Imputering

Ekonomisk statistik

Höstterminen 2008

Stockholms Universitet

Saknade värden

- Totalt bortfall
  - dvs inga uppgifter för enheten
- Partiellt bortfall
  - Man har vissa uppgifter för enheten. Det
    - Det kan vara det vanliga att enheten inte ”svarar” på vissa frågor.
    - Men lika gärna att man har vissa uppgifter som samlats in från register eller att man inte frågat om den
- Strukturellt saknade värden
  - Variabler som logiskt sett inte kan finns.
    - Om ett företag inte har egna byggnader kan man inte svara på frågan om arean eller uppvärmningssätt
- Mörkertal
  - Man vet inte ens att enheten finns.
    - Kallas ramfel om registret används för dragning . Men mörkertal om antalet företag i Sverige

Rubins indelning

Se Little RJA and Rubin DB (2002) *Statistical Analysis with Missing Data*, 2nd edition. New York : John Wiley.

- MCAR, Missing Completely at Random
  - Att ett värde eller en enhet saknas är helt slumpmässigt och säger inget om företaget
  - Formellt B och Y är oberoende där B är bortfallsmekanismen och Y den studerade variabeln
- MAR, Missing at Random
  - Att ett värde eller en enhet saknas beror bara på förhållanden som vi redan vet och säger inget ytterligare om företaget.
  - Formellt B och Y är oberoende givet X där X är kända hjälpvariabler
  - Bortfallssannolikheten kan tex variera mellan olika branscher eller antal anställda, men branschen och antal anställda känner vi redan från ramen
- NMAR,
  - Not Missing at Random, Bortfallet beror på det vi studerar.
  - Man tror t ex att expansiva enmansföretag svarar mindre eftersom företagaren inte har tid att svara.

Sätt att hantera missing data

- Glöm bort dem!
  - Om man dragit ett stickprov med 100 företag och man får kompletta uppgifter från 90, låtsas man att man bara tänkt undersöka 90.
    - Urvalssannolikheterna/uppräkningsfaktorerna måste givetvis justeras ändras, tex genom efterstratifiering eller kalibrering, så att totalantalet och kanske andra kända proportioner blir rätt, tex antalet i olika branscher (med klok ändring av vikter kan man åstadkomma mycket, men osäkerhetsberäkningar blir lätt fel)
    - Detta går inte att göra vid räkningar/totalundersökningar
    - Detta är dumt vid partiellt bortfall eftersom man inte utnyttjar de uppgifter man har.
  - Om detta skall funger måste man anta MAR/MCAR
- Imputera! Dvs sätt in rimliga värden istället för de saknade!

Imputera

- Även här antar man att man har MAR eller MCAR (annars kommer det inte att fungera).
- Man använder den information man har för att välja vilka rimliga värden man skall sätta in (eller hur vikterna skall ändras). Beror värdet av förhållanden som man inte känner till blir det sällan bra.
- Då måste man göra specialstudier av bortfallet.
- De metoder vi kommer att prata om nedan fungerar bra om man har nästan MCAR.

Dokumentera

Ofta gjordes (görs?) imputeringen av kunniga personer som valde mycket rimliga värden. Det leder ofta till att varje företag får rimliga värden, men det leder till att ingen vet riktigt vad som görs och om det i slutändan blev bra. Det är viktigt att all imputering (och granskning) görs på ett väl dokumenterat och objektivt sätt så att kvaliteten på statistiken kan bedömas.

Individstatistik

- Bland SCBs metodstatistiker fanns tidigare än stor misstro mot imputering. Det finns en lag som säger att man inte får lägga in felaktiga värden i en individdatabas. Men man får lägga in en kolumn med ett annat namn t ex modellvärde.
- Inom företagstatistiken har man alltid imputerat, men statistikerna ägnade sig inte åt det problemet. Det var inte fint och blev därför mycket ad hoc. Nu tittar även metodstatistikerna på hur imputeringen görs, men fortfarande är SCB dåliga på imputering, teoretiskt sett (Praktiskt görs en del bra).
- Det som följer är mer en genomgång av den internationella forskningens nivå än vad som görs i Sverige

Imputering

- Det finns några olika sätt att dela in metoderna.
- En vanlig indelning är Real donor eller Model donor.
  - Real donor - värdet som sätts in hämtas från en existerande enhet
  - Model donor - värdet som sätts in konstrueras med hjälp av en modell
  - Men det finns mellanlägen

Real eller model donor - exempel

Data

Nr	var	Kön	Inkomst
1	17	2	14180
2	16	1	-
3	14	-	27690
4	10	1	-
5	18	2	16189
7	10	2	23457

...

Real donor

Nr	var	Kön	Inkomst
1	17	2	14180
2	16	1	14180
3	14	1	27690
4	10	1	23457
5	18	2	16189
7	10	2	23457

...

”Nearest Neighbour”

Model donor

Nr	var	Kön	Inkomst
1	17	2	14180
2	16	1	20379
3	14	1,6	27690
4	10	1	20379
5	18	2	16189
7	10	2	23457

Medelvärdesimputering

Real donor - några exempel

- Hot - Cold deck?
  - Hot deck - imputering Värdet hämtas från den aktuella datamängden
  - Cold deck - imputering Värdet hämtas från ett annat dataset (t ex förra årets värde på antal anställda. Denna typ är vanlig vid företagsdata)
- Var hämta värdet vid hot deck?
  - Från den enhet som mest liknar denna på relevanta variabler.
    - Hur välja avstånd?
    - Psykologer väljer gärna euklidiskt avstånd efter att ha standardiserat till variansen 1.
    - I ekonomiska studier är dock vanligen vissa variabler viktigare än andra och man kan använda olika pattern recognition metoder
    - Även neurala nätverk har föreslagits och fungerar hyfsat om ...
  - Från någon av de 10 mest liknande (slumpmässig imputering)
  - Från föregående i löpnummer

Model Donor - exempel

X	17	16	14	10	18	10	15	22	37	48
Y	235	-	173	-	163	142	315	277	-	423

Medelvärde 247 standardavvikelse 100 konfidensintervall 287 +/- 92

enligt standardrutiner baserat på sju värden

Medelvärdesimputering

Y	142	247	277	247	163	235	315	173	247	423

Medelvärde 247 standardavvikelse 83 konfidensintervall 247 +/- 52

Ger bra skattningar av medelvärden/totaler men för liten varians

Regressionsimputering

Y	142	240	277	208	163	235	315	173	350	423

Medelvärde 253 standardavvikelse 89 konfidensintervall 253 +/- 56

Det blir alltså för lite variation i datamängden och vanliga analysprogram tror att vi har fler observationer än vad vi ha. Även korrelationer och samband kommer att bli starkare eller mer signifikanta med denna imputeringsteknik

Model Donor - exempel

X 17 16 14 10 18 10 15 22 37 48

Y 235 - 173 - 163 142 315 277 - 423

Medelvärde 247 standardavvikelse 100 konfidensintervall 287 +/- 92

Regressionsimputering

Y 142 240 277 208 163 235 315 173 350 423

Medelvärde 253 standardavvikelse 89 konfidensintervall 253 +/- 56

Det blir alltså för lite variation i datamängden och vanliga analysprogram tror att vi har fler observationer än vad vi har

Regression plus slump

Y 142 337 277 221 163 235 315 173 327 423

Medelvärde 261 standardavvikelse 92 konfidensintervall 260 +/- 58

Om det görs rätt blir det väntevärdesriktiga skattning av medelvärde och varians men resultatet blir mycket slumpmässigt och kan inte upprepas (om MAR/MCAR)

Rätt sätt är multipel imputering enligt Rubin

Ett annat sätt att få mer slump än regressionsimputering är att använda real donor med lämpligt avståndsmått. Men fortfarande hade precisionsskattningarna blivit för bra

Multipel imputering

För att lösa problemet med att det tillkommer slump har Rubin föreslagit multipel imputering. Gör om imputeringen många (B) gånger så att man får många skattningar T_1, T_2, …, T_B, med uppskattade varianser S_1^2, S_2^2, …, S_B^2

- medelvärdet uppskattas med medelvärdet av medelvärdena $\Sigma T_i/B$
- variansen med medelvärdet av varianserna + variansen av medelvärdena

  $\Sigma S_1^2/B + \Sigma (T_1 - \Sigma T_1/B)^2 / (B-1)$
- Stora talens lag garanterar att det fungerar. Men i praktiken räcker det i allmänhet med ca tio upprepningar (vid stora bortfall krävs fler)

Den multipla imputeringen måste göras rätt

Det finns alldeles för många som försöker, men inte klarar att modellera slumpen rätt.

- Tänk t ex på att även osäkerheten i skattningen av regressionslinjen skall med
- När man bara är ute efter medelvärden eller totaler kan man använda normalmodell även om modellen data inte är (bivariat) normalfördelade men inte annars.

Modellimputering

- Vid regressionsimputeringen användes linjär regression. Logistisk regression skulle fungerat bra om den studerade variabeln var dikotom
- Vid slump-imputeringen kan man välja metod så att de imputerade värdena blir realistiska.
  - Om man inte gör multipel (eller slumpmässig) imputering finns dock en stor risk för systematiska fel.
  - Om t ex P(man|kända värden) = 0,7 och man väljer att imputera mest sannolika värde blir alla med imputerat värde män. (Istället för 70 %)
- Det finns användbara modeller för multivariat logistiska fördelningar vid flera klasser

•

- Vid företagsdata har man ofta tillgång till föregående års uppgifter. Dessa kan alltså användas som hjälpvariabler vid imputering. Mycket vanligt är att använd föregående års värden plus uppräkning med inflation.
- Dessutom imputerar man ofta saknade värden som går att räkna ut logiskt. (Ibland skall vissa summor gå ihop, men ofta kontrollringer man även på sådan uppgifter)
- Imputering med extra slump används vad jag vet inte inom SCBs företagstatistik. Det är en alltför ny metod för att ha nått ut. För många användare är osäkerhetsintervallen helt ointressanta och det gör att imputering med mest troligt värde ofta ger bra resultat.

Länkning

Ekonomisk statistik

Höstterminen 2008

Stockholms Universitet

Länkning

När statistiken läggs om uppstår ofta språng. T ex när näringsgrensindelningen moderniseras eller när man byter insamlingssystem från disketter till internet eller ändrar en cutoffgräns. Vid länkning försöker man räkna om serien bakåt som om det nya insamlingssystemet hade gällt. Bra för personer som arbetar med tidsserier.

Samma problem vid ändringar i samhället. Följ sjuklighetens utveckling i samhället, när antalet karensdagar ändras. Detta brukar ses som en uppgift för dem som analyserar tidsserier. Medan länkning är något för statistikproducenten.

Vid omläggningar bör statistikern planera för länkning. Vid reformer i samhället bör politikerna planera så att reformen kan utvärderas.

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

År 1 2 3 4 5 6 7 8 9 10 11 12 13

35 36 39 39 38 40 42 42 43 45 44 43 47

32 34 38 38 36 39 42 42 43 45 44 43 47

och ändra värdena innan motsvarande (additivt resp multiplikativt).

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

32 33 36 36 35 37 39 42 43 45 44 43 47

Andra modeller kan användas. Regression mot en logistisk linje hade förmodligen varit ännu bättre, men görs aldrig vad jag vet

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

33 34 37 37 36 38 40 42 43 45 44 43 47

Ger lite större osäkerhet än svar 3.

•

- Detta var mekaniska länkningsmetoder för en enda serie där det antogs att man inte hade mer information.
- En vanlig rekommendation är att om det är möjligt göra undersökningen på båda sätten parallellt vid ett tillfälle eller göra andr andra metodstudier (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. Slumpmässiga halvor.
- När större omläggningar 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

•

- Ibland har man annan information. 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 dne större kategorin. Då får man lägga in en subjektiv gissning. T ex andel 0 tio år tidigare och sedan en linjär ökning av andelen till den första observerade nivån.

•

- Om man länkat en dataserie, skall den som använder serien alltid kunna se vad som hänt. Orsaken till språnget och hur det eliminerats.
- Dokumentera vad du gör