

Utveckling av estimator för skattning av antal förvärvsarbetande i olika arbetstidsklasser inom små redovisningsgrupper

Sixten Lundström



R & D Report
Statistics Sweden
Research - Methods - Development
U/STM - 40

INLEDNING

TILL

R & D report : research, methods, development, U/STM / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1987. – Nr 29-41.

Föregångare:

Promemorior från U/STM / Statistiska centralbyrån. – Stockholm : Statistiska centralbyrån, 1986. – Nr 25-28.

Efterföljare:

R & D report : research, methods, development / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1988-2004. – Nr. 1988:1-2004:2.

Research and development : methodology reports from Statistics Sweden. – Stockholm : Statistiska centralbyrån. – 2006-. – Nr 2006:1-.

R & D Report, U/STM 1987:40. Utveckling av estimator för skattning av antal förvärvsarbetande i olika arbetstidsklasser inom små redovisningsgrupper / Sixten Lundström.
Digitaliserad av Statistiska centralbyrån (SCB) 2016.

Utveckling av estimator för skattning av antal förvärvsarbetande i olika arbetstidsklasser inom små redovisningsgrupper

Sixten Lundström



R & D Report
Statistics Sweden
Research - Methods - Development
U/STM - 40

Från trycket	December 1987
Producent	Statistiska centralbyrån, Enheten för statistiska metoder
Ansvarig utgivare	Staffan Wahlström
Förfrågningar	Sixten Lundström, tel 019-176496

© 1987, Statistiska centralbyrån
ISSN 0283-8680
Printed in Sweden
Garnisonstryckeriet, Stockholm 1987

U/STM-Ö

1987-10-01

S Lundström, ILK

...

Utveckling av estimator för skattning av antal förvärvsarbetande i olika arbetstidsklasser inom små redovisningsgrupper

1 Bakgrund

Utredningen om alternativa metoder för framtida folk- och bostadsräkningar (FoBALT) låg till grund för riksdagens beslut om utformningen av 1985 års folk- och bostadsräkning (FoB 85). Beslutet innebar att datainsamlingen även denna gång huvudsakligen skulle utföras med hjälp av en enkät till alla vuxna personer i riket, dock med det undantaget att nu skulle sysselsättningsinformationen insamlas via administrativa register. Fördelarna FoBALT-utredningen såg med denna datainsamlingsmetod var att uppgiftslämnarbördan skulle minska för allmänheten och dessutom skulle det vara möjligt att tillgodose behovet av årlig sysselsättningsstatistik. Riksdagen gick i huvudsak på utredningens linje men avslog förslaget att också insamla uppgifter om arbetad tid på detta sätt. I stället säger man: "... SCB bör ... pröva möjligheterna att via s k syntetiska skattningar fördela resultatet från riksnivån på regionala data i detta avseende."

Det administrativa register som föreslogs utgöra den huvudsakliga datakällan är baserat på kontrolluppgifterna (KU)¹. Dessutom används också en del andra register. För att få en detaljerad beskrivning av systemet och på vilket sätt de förvärvsarbetande avgränsas hänvisas till (14). Systemet har getts förkortningen ÅRSYS.

I denna rapport redovisas arbetet² med att utveckla en estimator för skattning av antalet förvärvsarbetande uppdelade på arbetstidsklasser för små redovisningsgrupper.

2 Krav på estimatorn

Sysselsättningsstatistik uppdelad på arbetstidsklasser används bl a av de kommunala planerarna för barnomsorgsplanering, för pendlingsstudier och sysselsättningsprognoser. Inom barnomsorgsplaneringen vill man veta hur många mödrar till förskolebarn som arbetar mer än 16 timmar per vecka och denna statistik vill man ha uppdelad på delar av kommuner. Vid pendlingsstudier och sysselsättningsprognoser vill man ha antalet för-

1) Kontrolluppgiften innehåller bl a den anställdes inkomst under föregående år och sänds inför den årliga inkomstdeklarationen från arbetsgivaren till skattemyndigheterna (och en kopia till den anställde).

2) I projektet har följande personer deltagit: Claes Cassel, Sixten Lundström, Göran Råbäck, Jan Sävenborg, Björn Tegsjö och Staffan Wahlström, varav den "arbetande" delen har utgjorts av de tre först nämnda.

värvsarbetande som pendlar från ett litet område till ett annat litet område uppdelat på arbetstidsklasser. Det innebär att efterfrågan på statistik avser mycket små redovisningsgrupper.

Mot denna bakgrund är det naturligtvis omöjligt att utnyttja en konventionell estimator som baserar sig på ett urval. Även med ett stort urval skulle vissa redovisningsgrupper inte innehålla några observationer. För övriga redovisningsgrupper skulle urvalsfelet bli oacceptabelt stort. Ett villkor är därför att den estimator som väljs måste ge skattningar med mindre urvalsfel än konventionella skattningar. Den grupp av estimatorer man då tänker på är modellberoende eller som de ibland (något oegentligt) kallas, syntetiska estimatorer.

Dessa estimatorer bygger på ett samutnyttjande av flera datakällor och är beroende av en modell. Ofta har man ett register som ger antalet objekt i varje liten redovisningsgrupp fördelat på hjälpgrupper som kön, ålder, etc. Däremot saknas undersökningsvariabeln i registret. Från en urvalsundersökning hämtas skattningar på riksnivå fördelade på samma hjälpgrupper och undersökningsvariabelns kategorier. Modellen består vanligtvis av ett antagande om att den "struktur" som mäts på riksnivå också gäller för de små redovisningsgrupperna.

Estimatorerna kan vara mer komplicerade än så men gemensamt för hela gruppen är att de ger skattningar med små urvalsfel. Däremot kan de i vissa fall uppvisa stora modellfel. Metodarbetet består i att finna en estimator med små modell- och urvalsfel.

Modellfelen kan också förändras över tiden och därför gäller det att hitta en estimator som kan förväntas ha relativt konstanta modellfel.

3 Datakällor

Det är framför allt två datakällor som är aktuella att användas vid skattning av arbetad tid, nämligen den årliga sysselsättningsstatistiken (ÅRSYS) och arbetskraftsundersökningen (AKU). Vissa överväganden görs också kring användandet av data från FoB 80 (senaste räkning då sysselsättningsdata insamlades via enkät).

ÅRSYS ger möjlighet att avgränsa populationen av förvärvsarbetande och i registret finns även information om bl a kommun-tillhörighet, ålder, kön, näringsgren och inkomst.

I AKU finns också information om ålder, kön och näringsgren. Inkomst kan påföras personerna i urvalet. Dessutom innehåller AKU information om den "normala veckoarbetstiden". Urvalet i

AKU omfattar ca 12 000 förvärvsarbetande varje månad. Matförfarandet och definitionen av arbetstid i AKU avviker en del från tidigare folk- och bostadsräkningar. I studien görs vissa jämförelser mellan skattningar från de två datakällorna.

4 Studerade estimatorer

Det finns ett stort antal estimatorer som tillhör gruppen modellberoende estimatorer och naturligtvis kan inte alla ingå i en och samma metodstudie. Vi begränsar oss till dem som av olika skäl kan förväntas fungera bra. Dessutom vill vi ha med estimatorer som är intressanta att jämföra med.

Vi inför följande beteckningar:

Populationen Ω består av N individer ur vilken vi skall dra ett stickprov med n individer.

q är en redovisningsgrupp i Ω med N_q individer. Om vi väljer n med OSU så är $E(n_q) = N_q/N \cdot n$, där n_q är urvalsstorleken i q .

Vi önskar skatta en storhet $P_q = 100 \cdot \sum_{i=1}^{N_q} y_i / N_q$ där

$$y_i = \begin{cases} 1 & \text{om individ } i \text{ tillhör studerade arbetstidsklass} \\ 0 & \text{annars} \end{cases}$$

Till vårt förfogande finns ett antal hjälpvariabler vars värden är kända för samtliga individer i Ω . Ω kan indelas i H hjälpgrupper. Man känner då N_h och N_{hq} ; n_h betecknar urvalsstorlek i grupp h .

i) I den syntetiska estimatoren (SYNT) skattas andelen förvärvsarbetande i den studerade arbetstidsklassen i hjälpgrupp h och sedan får varje individ i hjälpgrupp h , oberoende av vilken redovisningsgrupp den tillhör, representeras av denna skattade "sannolikhet" (dvs den skattade andelen) för att tillhöra arbetstidsklassen. Det förväntade antalet med egenskapen i redovisningsgrupp q blir då

$$\sum_h N_{hq} \hat{\bar{y}}_h, \text{ där } \hat{\bar{y}}_h = \sum_i^{n_h} y_i / n_h.$$

Den syntetiska skattningen av efterfrågade parameter har följande form:

$$SYNT = \frac{100}{N_q} \sum_{h=1}^H N_{hq} \hat{\bar{y}}_h \quad (1)$$

(Referenser: (4), (13).)

ii) Jämfört med SYNT utnyttjar följande estimator, SYNTINK, en summering av årsinkomster i stället för antalsuppgifter:

$$\text{SYNTINK} = \frac{100}{N_q} \sum_{h=1}^H x_{hq} \hat{\psi}_h, \text{ där} \quad (2)$$

$$\hat{\psi}_h = \frac{\sum_{i=1}^{n_h} y_i}{\sum_{i=1}^{n_h} x_i}$$

x_i är årsinkomst för individ i och

$$x_{hq} = \sum_{i=1}^{N_{hq}} x_i$$

(Referens: (4).)

iii) Det är rimligt att anta att med stigande årsinkomst ökar sannolikheten för att vara t ex heltidsarbetande. Det är också troligt att sannolikheten som funktion av årsinkomsten ökar mer i exempelvis intervallet 60 000-80 000 kr i årsinkomst än i ett lika långt intervall för höga respektive låga inkomster. En logit-modell verkar därför rimlig.

Genom vägd minska-kvadrat-anpassning skattas koefficienterna i regressionslinjen

$$u_h = \alpha + \beta_1 x_1 + \dots + \beta_m x_m + \varepsilon$$

där

$$u_h = e^{\log \frac{P_h}{1-P_h}}$$

P_h = sannolikheten att en person i grupp h tillhör studerade arbetstidsklass

och

$\underline{x}' = (x_1, \dots, x_m)$ betecknar en uppsättning dummyvariabler för de kategoriska variablerna kön, ålder (klassindeldad) och näringsgren samt en variabel som anger medelvärdet för varje inkomstklass. Vid skattning av koefficienterna utnyttjas andelen i grupp h som observerats höra till studerade arbetstidsklass som skattning av P_h .

Estimatorn, LOGIT, har följande utseende:

$$\text{LOGIT} = \frac{100}{N_q} \sum_{h=1}^H N_{hq} P_h^* \quad (3)$$

där P_h^* betecknar den genom modellen predikterade andelen i grupp h för arbetstidsklassen och beräknas som

$$P_h^* = e^{u_h^*} / (1 + e^{u_h^*})$$

$$u_h^* = \hat{\alpha} + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_m x_m$$

Anmärkning: I metodstudien beräknas LOGIT för olika arbetstidsklasser. Varje gång ses det som ett 2-kategori-problem: studerade arbetstidsklass respektive övriga.

(Referenser: 2 och 3.)

iv) SCB har tidigare provat en modellberoende estimator vid skattnings av antalet förvärvsarbetande mellan folk- och bostadsräkningar. Denna SPINK-estimator¹ är konstruerad på följande sätt:

Q förutsättes sorterad efter variabeln x , årsinkomst. Då är estimatorn

$$\text{SPINK} = \frac{100}{N_q} \sum_{h=1}^H \sum_{i=1}^{N_{hq}} z_{hi} \quad (4)$$

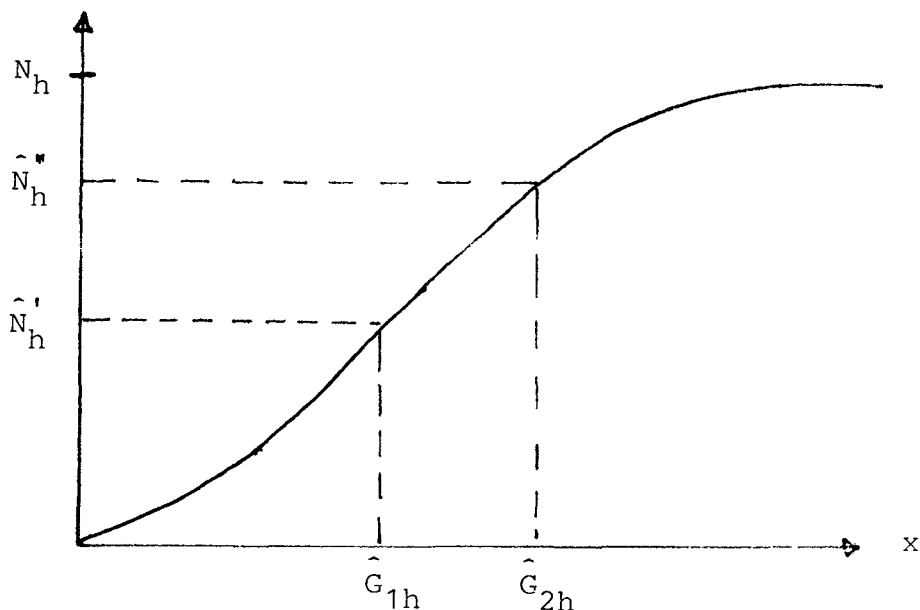
där

$$z_{hi} = \begin{cases} 1 & \text{om individ } i \text{ tillhör hjälpgrupp } h \text{ och har en årsinkomst } (x) \text{ som ligger i intervallet } [\hat{G}_{1h}, \hat{G}_{2h}] \\ 0 & \text{f ö} \end{cases}$$

$\hat{G}_{1h}$ och $\hat{G}_{2h}$ är undre och övre gränsen för vad en individ i hjälpgrupp h "bör" tjäna för att rimligtvis tillhöra den aktuella arbetstidsklassen. Hur dessa gränser skattas framgår av figur 1 och efterföljande förklaring.

1) Den SPINK-estimator som SCB har använt är mindre effektiv än den vi använder. Vi beaktar nämligen också interaktioner mellan variablerna som bygger upp hjälpgrupperna (se (15)):

Antal individer
i hjälpgrupp h



Figur 1 Antal individer i hjälpgrupp h som har en årsinkomst mindre än x.

N_h' betecknar det från urvalet skattade antalet individer som tillhör sysselsättningsklasserna under den aktuella. Antalet individer som tillhör den aktuella sysselsättningsklassen, N_{Gh} , skattas med hjälp av:

$$\hat{N}_{Gh} = \left[\begin{array}{c} n_h \\ \sum_{i=1} y_i / n_h \end{array} \right] N_h$$

och $\hat{N}_h''$, som betecknar en skattning av antalet individer vars arbetstid högst tillhör den aktuella klassen, kan beräknas som $\hat{N}_h'' = \hat{N}_h' + \hat{N}_{Gh}$. Inkomstgränserna $\hat{G}_{1h}$ och $\hat{G}_{2h}$ skattas sedan genom projektion på inkomst-axeln.

(Referens: (15).)

v) Särndal (13) har via generaliserad regressionsestimation utvecklat en estimator som är asymptotiskt väntevärdesriktig. Den består av en syntetisk estimator (se i)) minus en skattning av biasen för den syntetiska estimatoren. Den senare delen blir vanligtvis behäftad med stora urvalsfel. Cassel (1) har därför utvecklat en estimator där den senare delen multipliceras med en konstant (α). I denna studie utnyttjar vi en skattning av optimalt (minimerar Mean Square Error under vissa antaganden) α -värde. Estimatoren benämnes:

"Optimalt korrigerad syntetisk skattning" (KORRSYNT)

$$\text{KORRSYNT} = 100 \left(\sum_{h=1}^H N_{hq} \hat{\bar{Y}}_h - \hat{\alpha} \frac{N}{n} \sum_{i=1}^{n_q} e_i \right) / N_q \quad (5)$$

$$e_i = (y_i - \hat{\bar{Y}}_i) = (y_i - \hat{\bar{Y}}_h)$$

$$\hat{\alpha} = [1 - \text{var}(\hat{T}_c) / \hat{T}_c^2] \text{ där } \hat{T}_c = \frac{N}{nN_q} \sum_{i=1}^{n_q} e_i$$

(Referenser: (1), (13).)

vi) Tidigare nämnda estimatorer är alla modell-beroende och därför också biased. För att få en konventionell unbiased estimator att jämföra med ingår även följande direkta estimator (DIR). Den utgörs av medelvärdet bland observationerna i q utan användning av hjälpinformation.

$$\text{DIR} = 100 \sum_{i=1}^{n_q} y_i / n_q \quad (6)$$

5 Problem i metodstudien

En evalvering av modellberoende estimatorer vid skattningar för små redovisningsgrupper innehåller många svårigheter. För att kunna skatta felmått som bias, varians och Mean Square Error krävs att man har tillgång till populationsvärdena. Estimatorerna är också i många fall så komplicerade att simuleringar får genomföras. Någon form av "nedtransformering" av de stora mängderna är också nödvändiga av kostnadsskäl. Det är inte möjligt att göra ett stort antal upprepande urval (av, som i detta fall, 35 000 personer) ur en mycket stor population och sedan skatta felen för det stora antalet små redovisningsgrupper.

I föreliggande studie är det ytterligare två problem:

1) Definitionen av arbetad tid avviker något mellan AKU och FoB. Hur skall effekten av detta mätas?

2) Man kan förvänta sig att det finns ett samband mellan arbetad tid per vecka och årsinkomst. Problemet är att det finns personer som endast arbetar en del av året och därmed får en låg årsinkomst men som arbetar heltid under den period de är sysselsatta. På KU finns en fråga om vilken period inkomstuppgifter avser men för närvarande är den dåligt ifylld. Den upp-

giften förväntas bli betydligt bättre i framtiden och därmed kan angivna inkomst lätt transformeras om till en inkomst som skulle uppstått om sysselsättningen pågått hela året. Hur skall metodstudien kunna göras realistisk under dessa förutsättningar?

För att i möjligaste mån klara av dessa problem har metodstudien lagts upp i olika etapper. Dessa framgår av de kommande avsnitten.

6 Simuleringsstudier

6.1 Felmått och metod för skattning av dessa

Ett vanligt felmått för en estimator $\hat{T}_q$ är Mean Square Error (MSE). Det innehåller både slumpfel och bias enligt följande:

$$\text{MSE}(\hat{T}_q) = \text{Var}(\hat{T}_q) + B^2(\hat{T}_q), \quad (7)$$

där $\text{Var}(\hat{T}_q)$ är variansen för $\hat{T}_q$

och $B(\hat{T}_q)$ är biasen för $\hat{T}_q$.

Om $B(\hat{T}_q) = 0$ brukar man ange konfidensintervall för skattningen, $\hat{T}_q \pm 2 \sqrt{\text{Var}(\hat{T}_q)}$ (detta är ett 95 procentigt konfidensintervall). När $B(\hat{T}_q) \neq 0$ finns inget motsvarande sannolikhetsmått men brukligt är att ge måttet samma dimension, dvs beräkna $\text{RMSE} = \sqrt{\text{MSE}(\hat{T}_q)}$.

I studien skattas förutom RMSE också $\text{Var}(\hat{T}_q)$, $B(\hat{T}_q)$ och den procentuella andelen som absolutvärdet av biasen utgör av roten ur MSE (Andel-B = $100 \cdot |B(\hat{T}_q)| / \sqrt{\text{MSE}(\hat{T}_q)}$). Dessutom redovisas också vissa medelvärden (över redovisningsgrupperna).

Dessa värden skattas i studien med hjälp av simuleringar. För varje upprepning r ($r = 1, \dots, R$) beräknas estimatet ($\hat{T}_{q(r)}$) och när alla R upprepningar är gjorda beräknas:

$$\hat{B}(\hat{T}_q) = \left(\sum_{r=1}^R \hat{T}_{q(r)} / R \right) - T_q \quad (8)$$

och

$$\widehat{\text{Var}}(\hat{T}_q) = \left[\sum_{r=1}^R \hat{T}_{q(r)}^2 - \left(\sum_{r=1}^R \hat{T}_{q(r)} \right)^2 / R \right] / (R-1) \quad (9)$$

6.2 Överväganden kring urvalsstorlek

De modellberoende estimatorerna utnyttjar urvalsinformationen till att skatta andelen förvärvsarbetande i en viss arbetstidsklass i en viss hjälpgrupp för hela riket. Det innebär att urvalsfelet kan förväntas vara litet även om urvalsstorleken är liten, säg 1 000. Det stora bidraget till felet förväntas komma från biasen.

Estimatoren DIR baserar sig enbart på de observationer som hamnar i redovisningsgruppen och blir därför mycket mer beroende av den totala urvalsstorleken.

I studien används två urvalsstorlekar, nämligen 1 000 respektive 2 500 individer. Båda alternativen ger i nästan alla studierna något för stora urvalsfel för de modellberoende estimatorerna men för små urvalsfel för estimatoren DIR jämfört med en verklig situation då AKU-data utnyttjas.

6.3 Simulering inom AKU-population

Populationen i denna etapp bildas av de svarande individerna i AKU-urvalen för oktober, november och december 1980 samt av de utgående panelerna för övriga månader under året. För att även få med variabeln arbetstidsklass från FoB matchas beståndet mot FoB 80. Dessutom matchas det mot KU-registret samt SCBTB-registret för att erhålla enbart helårsarbetande anställda personer. Det är känt att kategorin företagare har ett mycket avvikande samband mellan inkomster och arbetad tid jämfört med övriga. Det är oklart om denna kategori i framtiden skall särredovisas eller kanske inte redovisas alls. Denna grupp får även en särbehandling i ARSYS. Det visar sig att det råder stor skillnad avseende sambanden mellan årsinkomst och arbetstidsklass mellan åldersgrupperna 16-24 år och 25-64 år. Eftersom vi har ett relativt litet antal individer i den lägre åldersgruppen begränsas studien till gruppen 25-64 år. Dessutom begränsas populationen till personer som är förvärvsarbetande både enligt AKU och FoB.

Simuleringspopulationen består efter dessa reduceringar av knappt 35 000 förvärvsarbetande. Den är för liten för att en uppdelning på kommuner skall bli meningsfull, därför får länen utgöra redovisningsgrupper. För varje person finns uppgift om kön, ålder, årsinkomst, arbetstidsklass enligt AKU respektive FoB samt uppgift om länstillhörighet.

I tabellerna 1 a-3, bilaga 1, redovisas en del av resultaten¹ för simuleringarna inom AKU-populationen. Hjälpgrupperna har bildats av kön och 8 inkomstklasser (i tusentals kr: 0-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-).

Tabellerna 1 a-1 c visar att med urvalsstorleken 1 000 är SYNT den estimator som fungerar bäst för alla arbetstidsklasser. SYNTINK fungerar nästan lika bra för klasserna 1-19 timmar/vecka och 20-34 timmar/vecka men för klassen 35-w timmar/vecka är SYNT överlägsen. För denna senare klass är även LOGIT bättre än SYNTINK. De två estimatorerna DIR och KORRSYNT som är

1) I (5) redovisas samtliga resultat.

unbiased respektive asymptotiskt unbiased uppvisar stora fel. Huvuddelen av dessa fel utgörs av urvalsfel, vilket framgår av tabellerna 2 a-2 c. För övriga estimatorer är biasen dominerande även med så litet urval som 1 000 individer. En beräkning visar att medelvärdet över andel-B (se definition i avsnitt 6.1) är 62, 55 respektive 56 procent för de tre arbetstidsklasserna.

Tabell 3 visar RMSE vid skattning av procentuella andelen förvärvsarbetande 35-w timmar/vecka när stickprovsstorleken är 2 500 och med samma hjälpgrupper som för tidigare tabeller. Jämför man tabell 1 c och 3 ser man som väntat att det är DIR och KORRSYNT som uppvisar större reduceringar av felen vid ökning av urvalsstorleken än övriga estimatorer. Med urvalsstorleken 2 500 är medelvärdet över andel-B 72, 64 respektive 63 procent för de tre arbetstidsklasserna.

Simuleringar har också genomförts för hjälpgrupper bestämda av kön och 14 inkomstklasser (i tusentals kr: 0-19, 20-24, 25-29, ..., 75-79, 80-). De visar att ingen förbättring åstadkoms med denna utökning av inkomstklasserna.

I bilaga 2 redovisas skillnaderna mellan AKU-uppgiften och FoB-uppgiften om arbetad tid i AKU-populationen. Stora avvikelser förekommer speciellt vad gäller den lägsta arbetstidsklassen (underskattning med 24,6 procent i AKU). Dessa avvikelser påverkar naturligtvis skattningarna.

6.4 Simulering inom FoB-population över kommuner i Örebro län respektive Norrbottens län

Syftet med denna etapp är i första hand att utreda om slutsatserna från simulering i AKU-populationen står sig när kommunen utgör redovisningsgrupp och när värdena på undersökningsvariabeln hämtas från FoB 80.

Populationen utgörs av de som är förvärvsarbetande enligt FoB 80 och helårsarbetande anställda enligt skatteadministrationens register. Storleken på populationen är ca 90 000.

Tabellerna 4 a-4 c i bilaga 1 visar skattningar av RMSE för de tre arbetstidsklasserna när urvalsstorleken är 1 000. Hjälpgrupperna är bestämda av kön och 14 inkomstklasser (samma som anges i avsnitt 6.3).

Slutsatsen blir att den enkla estimatoren SYNT även i denna simuleringsstudie uppvisar de minsta felen.

En motsvarande simulering genomfördes även för kommuner i Norrbottens län. Också den studien visade att SYNT-estimatorn fungerade bäst.

Mot bakgrund av dessa studier valdes SYNT-estimatoren. Förutom att den nästan genomgående fungerar bäst så är den också enkel, vilket gör det lätt att beskriva den för statistikanvändarna.

7 Förbättring av SYNT-estimatoren

Hittills har bara hjälpgrupper bildade av kön och inkomstklasser utnyttjats. Alder och näringsgren samvarierar troligen också med undersökningsvariabeln. Dessutom har tidigare arbeten inom detta område (12) visat att det kan vara möjligt att kraftigt reducera biasen genom att gruppera kommunerna och utnyttja estimatoren inom varje grupp.

Detta förbättringsarbete är möjligt att utföra utan att genomföra simuleringar, p g a att såväl formeln för (förväntad) varians som för biasen är möjliga att härleda för SYNT-estimatorn (se bilaga 3):

Förväntad varians:

$$E [\text{Var} (\text{SYNT})] = \left(\frac{100}{N_q}\right)^2 \sum_h^H N_{hq}^2 \frac{\bar{y}_h (1 - \bar{y}_h)}{n w_h} \cdot K_h, \quad (10)$$

$$\text{där } K_h = 1 + \frac{1 - w_h}{n w_h} \text{ och } w_h = \frac{N_h}{N}$$

Bias:

$$B (\text{SYNT}) = \frac{100}{N_q} \sum_h^H N_{hq} (\bar{y}_h - \bar{y}_{hq}) \quad (11)$$

Genom att vi i studien känner $\bar{y}_h$, $\bar{y}_{hq}$ och N_{hq} kan vi för godtycklig urvalsstorlek (n) beräkna förväntad varians och bias.

Felmått som används i studien är följande:

Roten ur (förväntad) Mean Square Error:

$$\text{RMSE} = \{ E [\text{Var} (\text{SYNT})] + B^2 (\text{SYNT}) \}^{1/2} \quad (12)$$

Relativ bias:

$$\text{Rel-B} = 100 \cdot B (\text{SYNT}) / P_q \quad (13)$$

Relativ RMSE:

$$\text{Rel-RMSE} = 100 \cdot \text{RMSE} (\text{SYNT}) / P_q \quad (14)$$

Beräkningarna utförs för samtliga kommuner i FoB 80. För att reducera datamängden gjordes analyser inom Örebro- respektive Norrbottens-populationer av bl a hur inkomsterna skulle klassificeras (10). Ett problem med övergången till rena FoB-data är att årsdeltidare inte kan selekteras bort utan tillgång till mycket stora resurser. Därför gjordes inom Norrbottens-populationen en studie av hur felen avvek med och utan denna selektering (8). Felen ökade när delårsarbetande var med men det såg inte ut att påverka valet av hjälpgrupper och därför ansåg vi att förbättringsarbete utifrån FoB-data skulle fungera bra.

Ett viktigt krav på estimatoren är att den fungerar bra över tiden. Därför studerade vi också hur en estimator utvecklad på FoB 80-data fungerar på FoB 75-data.

I bilaga 4 redovisas medelvärdet (över samtliga kommuner) av Rel-RMSE för olika hjälpgruppsindelningar. Urvalsstorleken är 35 000 vilket ungefär motsvarar antalet förvärvsarbetande i tre månads-AKU. De inkomst- och näringsgrensklassificeringar som är gjorda bygger på beräkningar vars detaljer inte redovisas här (se (5)-(7)).

Tabell A i bilaga 4 visar att de genomsnittligt minsta felen erhålls då alla fyra variablerna, KÖN, ÅLDER, SNI5 och INK4 (se förklaringar av variablerna i bilaga 4) används vid hjälpgruppsindelningen. Detta gäller dock enbart arbetstidsklassen 35-w timmar/vecka. För klassen 20-w timmar/vecka spelar inte hjälpgruppsindelningen nämnvärd roll. ÅLDER kan dock tas bort utan betydande ökning av felen.

För att få en uppfattning om effekten av hjälpinformation har vi beräknat medelvärdet av Rel-RMSE för en syntetisk estimator utan hjälpinformation (dvs den skattade andelen förvärvsarbetande i studerade arbetstidsklass för hela riket används i varje kommun). Baserat på FoB 80-data och för klassen 35-w timmar/vecka erhålls värdet 2.22 som kan jämföras med 1.54 för estimatoren med den "optimala" hjälpinformationen. Detta kan också jämföras med DIR-estimatorn som ger värdet 7.7.

Tabell A visar att tidsstabiliteten är god för klassen 35-w timmar/vecka men något sämre för klassen 20-w timmar/vecka.

En tänkbar estimator är "SYNT minus bias vid FoB 80-tidpunkten". Den är naturligtvis meningslös att utvärdera mot FoB 80-data (endast det obetydliga urvalsfelet är kvar) men däremot kan det vara intressant att utreda om den är stabil över tiden. Mellan FoB 75 och FoB 80 har det skett en del kommunförändringar och därför jämförs bara de kommuner (275) som kan identifieras (är oförändrade eller delar kan summeras) vid

båda tillfällena. Först har biasen vid FoB 80-tillfället beräknats för hjälpgruppsindelningen SNI5*INK6 och därefter utnyttjas angiven estimator vid FoB 75-tillfället. Medelvärde av Rel-RMSE är 1.07 respektive 1.12 för respektive arbetstidsklasser 20-w timmar/vecka och 35-w timmar/vecka. Det innebär att på 5 år har denna estimator försämrats kraftigt.

Vi har också gjort försök att gruppera kommunerna efter olika kriterier (se (11)) men i samtliga fall erfarit att tidsstabiliteten är svag för dessa estimatorer. En gruppering av kommuner leder också till att AKU-urvalet splittras upp på grupperna med följd att urvalsfelet ökar.

Även en s k SPREE-skattning (se (12)) med utnyttjande av FoB 80-data har testats på FoB 75-data men visat sig ge dåligt resultat.

Slutsatsen av dessa studier är att estimatorn bör endast basera sig på aktuella data.

För att ge en bild av kvaliteten har vardera urvalet dragits från respektive FoB 75 och FoB 80 och en syntetisk skattning av procentuella andelen förvärvsarbetande 35-w timmar/vecka för varje kommun har beräknats. I bilaga 5 redovisas en del av resultaten. Skattningarna ligger för de flesta kommuner relativt nära FoB-värdet men för vissa kommuner har vi stora skillnader (tabell B och C i bilaga 5). Det största felet ser vi för Oxelösunds kommun (0481) där överskattningen är 9,1 procentenheter i FoB 75 och 11,6 procentenheter i FoB 80! Felet i skattningen av förändringen mellan FoB 75 och FoB 80 blir däremot inte så stort.

7 Sammanfattning

I riksdagens beslut om utformningen av 1985 års folk- och bostadsräkning sägs att sysselsättningsdata i fortsättningen skall insamlas via skatteadministrationen (systemet benämns nu för ARSYS). Undantag utgör uppdelningen av de förvärvsarbetande i arbetstidsklass, där riksdagen föreslår att SCB prövar s k syntetiska skattningar.

Omfattande studier har nu genomförts och därvid har en enkel syntetisk estimator befunnits fungera bäst. Två olika datakällor används i estimatorn: AKU ger en skattning av andelen förvärvsarbetande i viss arbetstidsklass för hjälpgrupper bestämda av kön, ålder, näringsgren och årsinkomst och ARSYS ger antalet förvärvsarbetande i varje redovisningsgrupp och hjälpgrupp. I den syntetiska estimatorn antas att varje person som tillhör en viss hjälpgrupp har samma sannolikhet (skattas av nämnda andel i AKU) för att tillhöra studerad arbetstids-

klass. Skattningen av antalet förvärvsarbetande i arbetstidsklassen bildas sedan genom att summera sannolikheterna för personerna som tillhör redovisningsgruppen.

Den syntetiska estimatorn har alltså visat sig fungera bäst av studerade estimatorer men ger den skattningar av tillräckligt god kvalitet? Den frågan besvaras inte av genomfört arbete. Naturligtvis är det svårt att svara på en sådan fråga eftersom svaret beror på vad statistiken skall användas till. Vid användningen av konventionella estimatorer kan slumpfelet beräknas utifrån informationen i urvalet och därmed får konsumenten underlag för att själv svara på frågan. Med en syntetisk skattning är biasen vanligtvis det stora felet och det kan inte uppskattas (med rimlig precision) utifrån tillgänglig information. De minsta redovisningsgrupperna i metodstudien är kommuner, medan behovet avser betydligt mindre grupper. Även för mindre redovisningsgrupper finns det skäl att tro att den syntetiska estimatorn är den bästa estimatorn. Däremot blir ovanstående fråga ändå mer relevant.

För närvarande finns inga andra alternativ att erhålla dessa skattningar på. Förhoppningsvis ger de mer information än konsumenternas egna subjektiva skattningar.

Referenser

- (1) Cassel, C: Optimalt val av α . Opublicerat manus SCB 1984-08-30.
- (2) Cassel, C, Lundström, S och Råbäck, G: Syntetisk estimation - Resultat av simuleringsstudier för att skatta arbetad tid med hjälp av kontrolluppgiften på deklARATIONEN. Utkast 1986-04-08.
- (3) Hannshek, E. A. and Jackson, J. E. (1977): Statistical Methods for Social Scientists. Academic Press, New York.
- (4) Hidioglou, M. A., Morry, M., Dagum, E. B., Rao, J. N. K, and Särndal, C. E. (198): Evaluation of Alternative Small Area Estimators Using Administrative Records. (1985) Proc. Soc. Statis. Sec., American Statistical Association.
- (5) Lundström, S: Resultat från simuleringsstudie inom AKU-populationen. Tre rapporter: Utkast 1986-02-25; 1986-03-03; 1986-03-06.
- (6) Lundström, S: Resultat från simuleringsstudie inom FoB 80-populationen. Två rapporter: Utkast 1986-04-03; 1986-04-03.

- (7) **Lundström, S:** Analys av hjälpinformation till den syntetiska estimatoren vid skattning av arbetad tid - förstudie i Örebro och Norrbottens län. Två rapporter: Utkast 1986-08-22; 1986-10-09.
- (8) **Lundström, S:** Kan vi testa tidsstatistiken för den syntetiska estimatoren genom att utnyttja material från FoB 75 och FoB 80? Utkast 1986-10-14.
- (9) **Lundström, S:** Kort beskrivning av arbetet med skattning av veckoarbetstiden i ÅRSYS. 1986-10-21.
- (10) **Lundström, S:** Hur skall klassindelningen av inkomst göras för att fungera bra vid både FoB 75 och FoB 80? Utkast 1986-10-29.
- (11) **Lundström, S:** Utveckling av syntetisk estimator för arbetad tid - sammanfattning. Utkast 1987-03-10.
- (12) **Lundström, S:** Estimating Population Characteristics and Households in Swedish Municipalities Using Survey and Register Data. Proceedings of the Second Annual Research Conference, March 23-26, 1986. Bureau of the Census.
- (13) **Särndal, C.E. (1980):** When robust estimation is not the obvious answer: The case of the synthetic estimator versus alternatives for small areas. University of Montréal.
- (14) **Tegsjö, B:** SYSS-nytt. REKO nr 85.4.
- (15) **Tegsjö, B:** SPINK - En metod att skatta förvärvsintensiteter med hjälp av inkomststatistik. Rapport 1981-12-14.

U/STM-Ö

1987-10-01

S Lundström, ILK

Resultat från simuleringsstudierna

Tabell 1 a

Population:	AKU-populationen (se avsnitt 6.3).
Parameter:	Procentuella andelen förvärvsarbetande i arbetstidsklassen 1-19 timmar/vecka (storleksordning: 4-9).
Urvalsstorlek:	1 000.
Hjälpgrupper:	Kön * inkomstklasser (8).
Felmått:	Roten ur Mean Square Error.

LÄN	SYNT	SYNTINK	LOGIT	SPINK	KORRSYNT	DIR
01	.95	.92	.74	1.69	1.23	1.85
03	.80	.78	1.09	1.19	2.41	4.53
04	1.25	1.29	.81	1.73	2.29	4.54
05	.75	.77	1.52	.91	2.04	4.08
06	.88	.89	1.43	1.68	2.42	4.97
07	1.58	1.77	2.24	1.77	2.32	4.39
08	.78	.81	1.63	.91	2.65	5.36
09	2.77	2.85	3.85	4.05	3.21	4.48
10	2.40	2.52	3.48	2.31	2.67	4.31
11	1.48	1.29	2.20	2.95	2.34	5.08
12	.78	.96	1.52	.90	1.71	3.05
13	1.05	1.08	.86	1.11	2.54	5.04
14	2.18	2.15	1.38	2.62	2.30	3.42
15	.96	1.10	1.51	1.59	2.28	4.55
16	.95	1.00	1.48	1.60	2.30	5.04
17	1.15	1.33	1.92	1.58	2.61	4.73
18	1.33	1.35	2.24	1.33	2.18	4.15
19	1.30	1.22	2.32	1.03	2.32	4.01
20	.74	.79	1.21	1.29	2.39	4.78
21	.71	.74	1.11	.84	2.35	4.66
22	.70	.75	1.26	1.29	2.51	4.76
23	1.12	1.00	1.99	1.47	2.33	4.72
24	1.04	1.14	.86	.99	2.48	4.85
25	2.04	1.95	1.27	2.53	2.95	4.95
Medel- värde	1.24	1.27	1.66	1.64	2.37	4.43

Tabell 1 b

Population: AKU-populationen (se avsnitt 6.3).

Parameter: Procentuella andelen förvärvsarbetande i arbetstidsklassen 20-34 timmar/vecka (storleksordning: 17-25).

Urvalsstorlek: 1 000.

Hjälpgrupper: Kön * inkomstklasser (8).

Felmått: Roten ur Mean Square Error.

LÄN	SYNT	SYNTINK	LOGIT	SPINK	KORRSYNT	DIR
01	1.01	1.04	1.56	1.75	1.43	2.82
03	1.74	1.67	1.39	1.16	3.56	7.54
04	.98	1.00	1.37	1.08	3.03	6.67
05	1.44	1.43	1.09	1.18	3.32	7.15
06	1.04	1.04	1.95	1.29	3.31	6.86
07	1.39	1.38	1.13	1.16	3.40	7.36
08	1.25	1.27	1.12	1.79	3.30	7.64
09	1.43	1.42	2.52	2.12	3.66	7.15
10	2.28	2.25	1.13	2.00	3.61	7.47
11	1.26	1.33	1.60	1.22	2.93	7.46
12	1.25	1.21	1.14	1.17	2.48	4.79
13	1.05	1.03	2.24	1.27	3.30	7.24
14	1.52	1.56	2.54	1.13	2.44	4.85
15	1.81	1.91	3.50	1.72	2.99	6.24
16	1.51	1.46	3.54	1.24	3.48	6.95
17	1.05	1.05	1.74	1.77	3.07	6.71
18	1.00	1.01	1.64	1.02	2.98	7.04
19	1.70	1.75	1.06	1.49	3.11	6.26
20	2.02	1.99	3.43	2.11	3.53	6.48
21	1.04	1.04	2.33	1.27	3.42	7.20
22	1.59	1.58	1.96	2.26	3.18	7.07
23	1.47	1.47	2.91	1.55	3.33	7.35
24	1.02	1.02	2.04	1.21	3.13	6.90
25	2.81	2.73	1.57	3.16	3.99	7.85
Medel- värde	1.44	1.44	1.94	1.55	3.16	6.71

Tabell 1 c

Population:	AKU-populationen (se avsnitt 6.3).
Parameter:	Procentuella andelen förvärvsarbetande i arbetstidsklassen 35-w timmar/vecka (storleksordning: 68-80).
Urvalsstorlek:	1 000.
Hjälpgrupper:	Kön * inkomstklasser (8).
Felmått:	Roten ur Mean Square Error.

LÄN	SYNT	SYNTINK	LOGIT	SPINK	KORRSYNT	DIR
01	1.59	4.68	1.42	3.36	1.88	3.29
03	2.18	2.98	1.98	1.99	3.69	8.13
04	1.71	1.34	1.42	1.95	3.46	7.85
05	1.47	1.30	1.34	1.45	3.28	7.43
06	1.32	1.37	1.56	1.88	2.72	7.46
07	1.32	1.63	1.44	2.06	3.07	7.59
08	1.42	1.32	1.36	1.72	3.40	8.51
09	3.82	4.19	4.51	5.84	4.85	8.13
10	1.29	2.04	2.00	1.41	3.37	8.03
11	1.36	1.71	1.75	2.66	2.98	8.30
12	1.33	1.56	1.31	1.39	2.44	5.26
13	1.33	1.53	1.27	1.54	3.19	8.51
14	1.58	2.36	1.32	2.31	2.63	5.67
15	2.41	2.44	3.01	2.79	3.39	7.15
16	2.03	2.64	2.76	2.41	3.88	8.29
17	1.71	2.74	1.90	3.00	3.32	7.72
18	1.63	1.84	2.26	1.68	2.79	7.61
19	1.30	1.31	1.49	1.52	2.99	6.82
20	2.15	2.76	2.90	2.73	3.93	7.86
21	1.28	1.39	1.80	1.42	3.28	7.68
22	1.66	2.21	1.93	1.60	3.00	7.90
23	2.29	2.57	3.25	2.70	3.52	7.97
24	1.49	1.40	1.30	1.45	3.67	7.87
25	4.76	3.71	3.60	5.60	5.02	8.52
Medel- värde	1.85	2.21	2.04	2.35	3.32	7.48

Tabell 2 a

Population: AKU-populationen (se avsnitt 6.3).

Parameter: Procentuella andelen förvärvsarbetande i arbetstidsklassen 1-19 timmar/vecka (storleksordning: 4-9).

Hjälpgrupper: Kön * inkomstklasser (8).

Felmått: Bias.

LÄN	SYNT	SYNTINK	LOGIT	SPINK	KORRSYNT	DIR
01	-.73	-.66	.15	-1.60	-.61	-.20
03	-.37	-.22	.74	-.91	-.55	.12
04	-1.05	-1.09	-.04	-1.58	-1.07	.27
05	.22	.20	1.26	-.30	-.29	-.43
06	.38	.30	1.12	1.21	-.04	.17
07	1.39	1.56	2.07	1.50	.28	-.04
08	.04	.02	1.37	-.27	-.17	.55
09	2.65	2.72	3.74	3.93	.90	.05
10	2.29	2.41	3.38	2.18	.74	.38
11	1.26	.98	2.01	2.75	.39	.46
12	-.31	.56	1.28	.39	-.13	-.05
13	.70	-.71	-.08	.33	-.92	-.23
14	-2.08	-2.03	-1.13	-2.50	-1.40	-.07
15	.57	.71	1.23	1.23	-.17	-.05
16	.48	.48	1.17	1.34	.04	-.07
17	.88	1.06	1.72	1.29	.13	-.00
18	1.13	1.14	2.09	1.01	.39	.21
19	1.11	1.01	2.17	.75	.14	-.03
20	-.04	.15	.86	.60	-.30	-.03
21	-.16	-.00	.74	-.14	-.59	-.05
22	-.14	-.26	.95	-1.10	-.56	-.15
23	.84	.63	1.78	1.25	-.08	-.05
24	-.70	-.81	.06	-.03	-.54	.31
25	-1.94	-1.83	-1.01	-2.45	-1.43	.22
Medel- värde av ab- solut- värden	.89	.90	1.34	1.28	.49	.17

Tabell 2 b

Population: AKU-populationen (se avsnitt 6.3).

Parameter: Procentuella andelen förvärvsarbetande i arbetstidsklassen 20-34 timmar/vecka (storleksordning: 17-25).

Hjälpgrupper: Kön * inkomstklasser.

Felmått: Bias.

LÄN	SYNT	SYNTINK	LOGIT	SPINK	KORRSYNT	DIR
01	-.10	.05	1.18	-1.50	-.12	.11
03	-1.42	-1.33	-.93	-.66	-.81	.05
04	-.13	-.23	.93	.18	-.33	.50
05	-.99	-.98	-.22	.13	-.52	-.49
06	.03	-.09	1.62	.11	.00	.21
07	-.95	-.94	.44	.07	-.60	-.60
08	-.69	-.73	-.25	1.50	-.59	.12
09	.92	.90	2.24	1.72	.49	-.57
10	-2.05	-2.01	-.45	-1.71	-1.34	.24
11	-.75	-.87	1.18	-.41	-.62	.37
12	-.76	-.68	.48	-.41	-.44	-.11
13	.24	.18	1.98	.28	-.43	-.22
14	1.15	1.20	2.33	.55	.42	-.24
15	1.48	1.59	3.33	1.08	.70	.31
16	1.09	1.02	3.36	.70	.40	-.67
17	.27	.28	1.39	1.33	-.36	-.52
18	-.10	-.18	1.28	.09	-.30	-.61
19	-1.37	-1.42	-.26	-.86	-1.23	-.84
20	1.76	1.72	3.27	1.71	.59	-.35
21	.36	.37	2.10	.73	.22	.35
22	1.21	1.19	1.66	2.01	.42	-.51
23	1.02	1.03	2.70	.94	.54	.64
24	-.10	-.14	1.74	.20	-.36	.03
25	-2.64	-2.55	-1.21	-2.97	-1.42	.65
Medel- värde av ab- solut- värden	.90	.90	1.52	.91	.55	.39

Tabell 2 c

Population: AKU-populationen (se avsnitt 6.3).

Parameter: Procentuella andelen förvärvsarbetande i arbetstidsklassen 35-w timmar/vecka (storleksordning: 68-80).

Hjälpgrupper: Kön * inkomstklasser (8).

Felmått: Bias.

LÄN	SYNT	SYNTINK	LOGIT	SPINK	KORRSYNT	DIR
01	.83	4.45	-.28	3.10	.55	.09
03	1.78	2.65	1.50	1.57	1.04	-.17
04	1.17	.45	.52	1.39	.72	-.77
05	.76	-.03	.27	.17	.51	.92
06	-.41	-.49	-.87	-1.32	-.24	-.39
07	-.45	-.98	-.66	-1.57	.04	.63
08	.65	.02	.26	-1.23	.27	-.68
09	-3.58	-3.95	-4.30	-5.66	-1.71	.52
10	-.25	-1.54	-1.49	-.47	-.06	-.62
11	-.52	-1.06	-1.18	-2.34	-.37	-.83
12	.44	.73	-.23	.01	.34	.15
13	.45	.80	-.01	-.62	.76	.45
14	.92	1.93	.17	-1.95	.66	.30
15	-2.05	-2.07	-2.72	-2.32	-.82	-.27
16	-1.58	-2.30	-2.43	-2.05	-.59	.74
17	-1.15	-2.40	-1.37	-2.63	-.26	.52
18	-1.04	-1.32	-1.85	-1.11	-.59	.40
19	.25	.12	-.69	.11	.45	.86
20	-1.73	-2.42	-2.57	-2.31	-.60	.37
21	-.21	-.45	-1.24	-.59	.04	-.30
22	-1.08	-1.81	-1.41	-.91	-.43	.66
23	-1.87	-2.19	-2.96	-2.20	-1.05	-.60
24	.79	-.57	.08	-.17	.64	-.34
25	4.57	3.46	3.34	5.41	2.27	-.88
Medel- värde av ab- solut- värden	1.19	1.59	1.35	1.72	.63	.52

Tabell 3

Population: AKU-populationen (se avsnitt 6.3).

Parameter: Procentuella andelen förvärvsarbetande i arbetstidsklassen 35-w timmar/vecka (storleksordning: 68-80).

Urvalsstorlek: 2 500.

Hjälpgrupper: Kön * inkomstklasser (8).

Felmått: Roten ur Mean Square Error.

LÄN	SYNT	SYNTINK	LOGIT	SPINK	KORRSYNT	DIR
01	1.31	4.33	1.11	3.27	1.44	2.23
03	2.00	2.61	1.86	1.80	2.36	4.52
04	1.51	1.07	1.18	1.85	2.14	4.59
05	1.24	1.05	1.09	1.15	2.08	4.81
06	1.05	1.23	1.29	1.63	2.09	4.85
07	1.09	1.56	1.18	1.84	1.93	5.11
08	1.16	1.02	1.09	1.61	2.27	5.04
09	3.74	4.27	4.36	5.79	3.43	5.16
10	1.05	2.03	1.76	1.16	2.21	4.88
11	1.15	1.63	1.51	2.55	2.19	5.03
12	1.09	1.17	1.03	1.07	1.64	3.38
13	1.05	1.19	1.01	1.28	2.09	5.04
14	1.35	2.03	1.08	2.22	1.60	3.33
15	2.29	2.47	2.85	2.63	2.36	4.32
16	1.88	2.66	2.58	2.32	2.35	5.01
17	1.54	2.79	1.66	2.85	2.25	5.17
18	1.46	1.83	2.07	1.56	2.08	4.93
19	1.03	1.05	1.24	1.15	2.12	4.67
20	2.03	2.80	2.71	2.57	2.56	4.54
21	1.03	1.22	1.58	1.24	2.26	5.06
22	1.51	2.25	1.71	1.44	2.33	4.78
23	2.14	2.55	3.08	2.52	2.85	5.25
24	1.26	1.26	1.03	1.16	2.42	5.12
25	4.65	3.43	3.57	5.56	4.01	5.12
Medel- värde	1.65	2.06	1.82	2.17	2.29	4.66

Tabell 4 a

Population: Örebro län (se avsnitt 6.4).
 Parameter: Procentuella andelen förvärvsarbetande i arbetstidsklassen 1-19 timmar/vecka (storleksordning: 4-9).
 Urvalsstorlek: 1 000.
 Hjälpgrupper: Kön * inkomstklasser (14).
 Felmått: Roten ur Mean Square Error.

KOMMUN	SYNT	SYNTINK	LOGIT	SPINK	KORRSYNT	DIR
1860	.68	.79	1.95	.97	1.72	4.01
1861	.64	.67	1.75	.81	1.26	2.81
1862	.73	.74	1.23	.81	1.66	3.59
1863	.71	.72	1.51	.97	1.71	3.73
1864	.94	.92	2.49	1.51	2.19	4.91
1880	.62	.68	1.90	.79	.65	1.15
1881	.61	.70	1.71	.79	1.22	2.90
1882	.66	.72	1.51	.87	1.92	4.08
1883	.52	.56	1.79	.90	.87	1.73
1884	1.08	1.15	.97	.97	2.05	4.22
1885	.83	.87	1.35	.88	1.27	2.58
Medel- värde	.73	.77	1.65	.93	1.50	3.25

Tabell 4 b

Population: Örebro län (se avsnitt 6.4).

Parameter: Procentuella andelen förvärvsarbetande i arbetstidsklassen 20-34 timmar/vecka (storleksordning: 17-25).

Urvalsstorlek: 1 000.

Hjälpgrupper: Kön * inkomstklasser (14).

Felmått: Roten ur Mean Square Error.

KOMMUN	SYNT	SYNTINK	LOGIT	SPINK	KORRSYNT	DIR
1860	2.07	2.06	4.97	1.90	3.79	7.42
1861	1.07	1.07	2.66	1.47	2.30	5.28
1862	1.04	1.03	2.95	1.27	2.85	6.37
1863	1.04	1.03	1.92	1.71	3.12	6.94
1864	1.15	1.15	3.32	1.64	4.03	8.87
1880	1.39	1.35	1.65	1.70	1.41	2.44
1881	1.69	1.72	3.76	1.73	2.91	5.36
1882	1.50	1.45	3.86	1.89	2.80	6.56
1883	1.02	1.06	2.07	1.04	1.65	3.65
1884	1.10	1.11	2.59	1.39	2.91	7.02
1885	1.37	1.35	2.98	2.17	2.01	4.22
Medel- värde	1.31	1.31	2.98	1.63	2.71	5.83

Tabell 4 c

Population: Örebro län (se avsnitt 6.4).

Parameter: Procentuella andelen förvärvsarbetande i arbetstidsklassen 35-w timmar/vecka (storleksordning: 68-80).

Urvalsstorlek: 1 000.

Hjälpgrupper: Kön * inkomstklasser (14).

Felmått: Roten ur Mean Square Error.

KOMMUN	SYNT	SYNTINK	LOGIT	SPINK	KORRSYNT	DIR
1860	2.57	3.06	4.75	2.62	3.66	7.82
1861	1.42	2.61	2.17	1.94	2.43	5.92
1862	1.36	1.82	2.11	1.65	2.70	6.96
1863	1.35	1.95	1.73	1.77	2.85	7.65
1864	1.82	2.80	3.37	2.82	3.97	9.38
1880	1.51	2.61	1.72	1.99	1.55	2.87
1881	1.96	2.79	3.29	2.17	2.90	5.91
1882	1.71	2.46	2.78	2.53	2.89	7.34
1883	1.35	1.96	2.21	1.65	1.70	3.92
1884	1.79	1.87	1.68	2.03	3.00	7.81
1885	1.40	1.96	2.21	2.15	2.12	4.95
Medel- värde	1.66	2.35	2.55	2.12	2.71	6.41

Klassificeringsfel i AKUs uppgift om arbetstidsklass inom
AKU-populationen.

Tabell A

FoB 80s uppgift antas vara riktig.

Procentuell andel av FoB 80s skattning	Arbetstidsklass (timmar/vecka)		
	1-19	20-34	35-w
a) Felaktigt medtagna	24.6	28.1	6.5
b) Felaktigt uteslutna	49.2	19.8	6.2
c) Bruttofel	73.8	47.9	12.7
d) Rätt klassificerade	50.8	80.2	93.4
e) Nettofel	-24.6	8.3	0.2

Anmärkning: En del av felen kan bero på att uppgifterna i AKU
avser annan tidpunkt än uppgifterna i FoB.

Härledning av Mean Square Error för den syntetiska estimatoren

Den syntetiska estimatoren har följande utseende:

$$\hat{\bar{Y}}_{\cdot q} = \frac{100}{N_{\cdot q}} \sum_h N_{hq} \bar{Y}_{h\cdot}, \text{ där}$$

$\hat{\bar{Y}}_{h\cdot}$ är skattade andelen personer förvärvsarbetande ett visst antal timmar av alla förvärvsarbetande i hjälpgrupp h

och

N_{hq} är antalet personer i populationen som tillhör hjälpgrupp h och kommun q

Mean Square Error för $\hat{\bar{Y}}_{\cdot q}$ har följande allmänna utseende:

$$\text{MSE}(\hat{\bar{Y}}_{\cdot q}) = \text{Var}(\hat{\bar{Y}}_{\cdot q}) + B^2(\hat{\bar{Y}}_{\cdot q})$$

a) Härledning av variansen:

Om $n_{h\cdot}$ är fix och osv gäller:

$$\text{Var}(\hat{\bar{Y}}_{\cdot q}) = \frac{100^2}{N_{\cdot q}^2} \sum_h N_{hq}^2 \cdot \frac{\bar{Y}_{h\cdot} (1 - \bar{Y}_{h\cdot})}{n_{h\cdot}}$$

(ändlighetsfaktorn försumbar)

Men

$$E \left[\frac{1}{n_{h\cdot}} \right] = \frac{1}{n W_h} + \frac{1 - W_h}{n^2 W_h^2}, \text{ där } W_h = \frac{N_{h\cdot}}{N_{\cdot\cdot}}$$

$$E[\text{Var}(\hat{\bar{Y}}_{\cdot q})] = \frac{100^2}{N_{\cdot q}^2} \sum N_{hq}^2 \cdot \frac{\bar{Y}_{h\cdot} (1 - \bar{Y}_{h\cdot})}{n\bar{W}_h} \cdot K_h ,$$

$$\text{där } K_h = 1 + \frac{1 - \bar{W}_h}{n\bar{W}_h}$$

b) Härledning av biasen för $\hat{\bar{Y}}_{\cdot q}$

$$\begin{aligned} B(\hat{\bar{Y}}_{\cdot q}) &= E(\hat{\bar{Y}}_{\cdot q}) - \frac{100}{N_{\cdot q}} \sum N_{hq} \bar{Y}_{hq} = \\ &= \frac{100}{N_{\cdot q}} \sum N_{hq} \bar{Y}_{h\cdot} - \frac{100}{N_{\cdot q}} \sum N_{hq} \bar{Y}_{hq} = \\ &= \frac{100}{N_{\cdot q}} \sum N_{hq} (\bar{Y}_{h\cdot} - \bar{Y}_{hq}) \end{aligned}$$

Medelvärde av Rel-RMSE för SYNT-estimatorn över samtliga kommuner

Urvalsstorleken är 35 000 individer.

Estimatorn utvecklas på FoB 80-data och tidsstabiliteten testas genom att beräkna felen i FoB 75 för samma estimator.

Variablerna som bestämmer hjälpgruppsindelningen har följande betydelse:

Variabel-namn	Kategorier
KÖN	Man, kvinna
ÅLDER	20-24, 25-64 år
SNI 5	1 Ospecificerat + jordbruk + byggnadsverksamhet 2 Brytning av mineral + El-, gas, och vattenförs 3 Tillverkning + Varuhandel, restaurang- och hotellverksamhet 4 Samfärdsel + Bank o förs 5 Offentlig förvaltning
INK 4	0-49, 50-59, 60-69, 70- tkr
INK 6	0-29, 30-39, 40-49, 50-59, 60-69, 70- tkr

Tabell A

Medelvärde av Rel-RMSE för SYNT-estimatorn.

Hjälpgrupper ¹	FoB 80		FoB 75	
	20-w	35-w	20-w	35-w
KÖN * ÅLDER * INK 4 * SNI 5	0.91	1.54	1.07	1.29
KÖN * INK 4 * SNI 5	0.92	1.55	1.10	1.34
ÅLDER * INK 4 * SNI 5	0.89	1.94	1.05	1.72
INK 4 * SNI 5	0.91	1.90	1.09	1.76

1) Vid beräkningen av Rel-RMSE i FoB 75 används INK 6 i stället för INK 4.

Syntetisk skattning av procentuella andelen förvärvsarbetande 35-w timmar/vecka i varje kommun vid FoB 75- respektive FoB 80-tidpunkten

- Förutsättning** Med hjälp av SYNT-estimatoren med "optimal" hjälpgruppsindelning skattas procentuella andelen förvärvsarbetande 35-w timmar/vecka. Ett obundet slumpmässigt urval (OSU) dras från FoB 75 och ett annat från FoB 80. Urvalsstorleken är 12 000.
- Realismen i skattningarna** Definitionsskillnaden mellan AKU och FoB kommer inte med vilket talar för en underskattning av felen. Å andra sidan över-skattas felen p g a att större urval kommer att användas i verkligheten och att störningen från årsdeltidare blir mindre i verkligheten. En sammanfattande bedömning är att realismen är ganska stor.
- Resultatredovisning** Det tar för stort utrymme att redovisa resultaten för samtliga kommuner, därför görs en begränsning till ett slumpmässigt urval av kommuner (tabell A) och en redovisning av kommuner med de största felen (tabell B och C).

Tabell A

Slumpmässigt urval av kommuner.

Syntetisk skattning av procentuella andelen förvärvsarbetande 35-w timmar/vecka jämfört med FoB skattningen.

KOMMUNER	FoB 75-TIDPUNKT		FoB 80-TIDPUNKT		FÖRÄNDRING MELLAN FoB 75 OCH FoB 80	
	VERKLIG	SKATTAD	VERKLIG	SKATTAD	VERKLIG	SKATTAD
0123	78.3	78.9	75.3	76.5	-3.0	-2.4
0191	76.8	77.5	74.7	74.3	-2.1	-3.2
0319	77.4	79.3	72.0	74.0	-5.4	-5.3
0617	73.9	76.2	69.4	70.3	-4.5	-5.9
1162	79.6	80.9	74.1	74.4	-5.5	-6.5
1165	80.4	78.9	75.7	74.1	-4.7	-4.8
1264	79.3	76.6	75.2	72.8	-4.1	-3.8
1382	75.7	76.0	69.3	69.9	-6.4	-6.1
1482	75.9	76.4	71.6	72.9	-4.3	-3.5
1563	75.9	75.5	71.1	69.8	-4.8	-5.7
1661	77.1	77.9	70.9	71.3	-6.2	-6.6
1682	75.9	76.3	69.3	70.1	-6.6	-6.2
1715	78.8	77.7	72.8	72.2	-6.0	-5.5
1784	75.7	75.8	69.9	70.4	-5.8	-5.4
2029	75.0	74.7	69.7	68.5	-5.3	-6.2
2031	74.7	73.4	68.2	68.0	-6.5	-5.4
2082	76.1	74.0	71.5	68.8	-4.6	-5.2
2101	78.8	77.7	72.8	71.1	-6.0	-6.6
2510	79.6	78.3	73.2	73.5	-6.4	-4.8
2580	77.5	79.1	72.8	76.0	-4.7	-3.1

Tabell B

Kommuner där utvecklingen mellan FoB 75 och FoB 80 överskattas med mer än två procentenheter.

KOMMU- NER	FoB 75-TIDPUNKT		FoB 80-TIDPUNKT		FÖRÄNDRING MELLAN FoB 75 OCH FoB 80	
	VERKLIG	SKATTAD	VERKLIG	SKATTAD	VERKLIG	SKATTAD
0509	74.2	76.6	71.4	70.3	-2.8	-6.3
0767	74.6	76.7	71.4	71.0	-3.2	-5.7
1265	79.2	77.7	75.9	70.9	-3.3	-6.8
1504	72.7	74.5	69.9	67.8	-2.8	-6.7
1561	77.5	75.9	72.8	69.1	-4.7	-6.8
1565	76.2	74.7	73.0	69.3	-3.2	-5.4
1730	76.8	76.9	74.5	70.0	-2.3	-6.9
2303	77.4	74.8	74.5	69.5	-2.9	-5.3
2422	75.6	75.4	73.1	70.0	-2.5	-5.4
2462	76.3	75.0	74.0	70.4	-2.3	-4.6
2521	74.5	75.1	71.9	70.0	-2.6	-5.1

Tabell C

Kommuner där utvecklingen mellan FoB 75 och FoB 80 underskattas med mer än två procentenheter.

KOMMU- NER	FoB 75-TIDPUNKT		FoB 80-TIDPUNKT		FÖRÄNDRING MELLAN FoB 75 OCH FoB 80	
	VERKLIG	SKATTAD	VERKLIG	SKATTAD	VERKLIG	SKATTAD
0125	77.1	77.8	72.5	75.7	-4.6	-2.1
0481	71.1	80.2	63.4	75.0	-7.7	-5.2
1401	76.7	76.7	70.9	73.0	-5.8	-3.7
2023	78.6	74.5	71.0	69.4	-7.6	-5.1
2417	79.0	77.1	72.8	73.5	-6.2	-3.6

Tidigare nummer av Promemorior från U/STM:
NR

- 1 Bayesianska idéer vid planeringen av sample surveys. Lars Lyberg (1978-11-01)
- 2 Litteraturförteckning över artiklar om kontingenstabeller. Anders Andersson (1978-11-07)
- 3 En presentation av Box-Jenkins metod för analys och prognos av tidsserier. Åke Holmén (1979-12-20)
- 4 Handledning i AID-analys. Anders Norberg (1980-10-22)
- 5 Utredning angående statistisk analysverksamhet vid SCB: Slutrapport. P/STM, Analysprojektet (1980-10-31)
- 6 Metoder för evalvering av noggrannheten i SCBs statistik. En översikt. Jörgen Dalén (1981-03-02)
- 7 Effektiva strategier för estimation av förändringar och nivåer vid föränderlig population. Gösta Forsman och Tomas Garås (1982-11-01)
- 8 How large must the sample size be? Nominal confidence levels versus actual coverage probabilities in simple random sampling. Jörgen Dalén (1983-02-14)
- 9 Regression analysis and ratio analysis for domains. A randomization theory approach. Eva Elvers, Carl Erik Särndal, Jan Wretman och Göran Örnberg (1983-06-20)
- 10 Current survey research at Statistics Sweden. Lars Lyberg, Bengt Swensson och Jan Håkan Wretman (1983-09-01)
- 11 Utjämningsmetoder vid nivåkorrigering av tidsserier med tillämpning på nationalräkenskapsdata. Lars-Otto Sjöberg (1984-01-11)
- 12 Regressionsanalys för f d statistikstuderande. Harry Lütjohann (1984-02-01)
- 13 Estimating Gini and Entropy inequality parameters. Fredrik Nygård och Arne Sandström (1985-01-09)
- 14 Income inequality measures based on sample surveys. Fredrik Nygård och Arne Sandström (1985-05-20)
- 15 Granskning och evalvering av surveymodeller, tiden före 1960. Gösta Forsman (1985-05-30)
- 16 Variance estimators of the Gini coefficient - simple random sampling. Arne Sandström, Jan Wretman och Bertil Waldén (Memo, Februari 1985)
- 17 Variance estimators of the Gini coefficient - probability sampling. Arne Sandström, Jan Wretman och Bertil Waldén (1985-07-05)
- 18 Reconciling tables and margins using least-squares. Harry Lütjohann (1985-08-01)
- 19 Ersättnings och uppgiftslämnarbördans betydelse för kvaliteten i undersökningarna om hushållens utgifter. Håkan L. Lindström (1985-11-29)
- 20 A general view of estimation for two phases of selection. Carl-Erik Särndal och Bengt Swensson (1985-12-05)
- 21 On the use of automated coding at Statistics Sweden. Lars Lyberg (1986-01-16)
- 22 Quality Control of Coding Operations at Statistics Sweden. Lars Lyberg (1986-03-20)
- 23 A General View of Nonresponse Bias in Some Sample Surveys of the Swedish Population. Håkan L. Lindström (1986-05-16)
- 24 Nonresponse rates in 1970 - 1985 in surveys of Individuals and Households. Håkan L. Lindström och Pat Dean (1986-06-06)
- 25 Two Evaluation Studies of Small Area Estimation Methods: The Case of Estimating Population Characteristics in Swedish Municipalities for the Intercensal Period. Sixten Lundström (1986-10-14)
- 26 A Survey Practitioner's Notion of Nonresponse. Richard Platek (1986-10-20)
- 27 Factors to be Considered in Developing a Reinterview Program and Interviewer Debriefings at SCB. Dawn D. Nelson (1986-10-24)
- 28 Utredning kring bortfallet i samband med den s k Metropolitdebatten. Sixten Lundström (1986-11-10)
- 29 Om HINK-undersökningens design- och allokeringsproblematik. Bengt Rosén (1986-10-06)
- 30 Om HINK-undersökningens estimationsförfarande. Bengt Rosén (1987-01-28)
- 31 Program för minskat bortfall i SCBs individ- och hushållsundersökningar - förslag från aktionsgruppen för uppgiftslämnarfrågor. Lars Lundgren (1987-02-20)
- 32 Practical estimators of a population total which reduce the impact of large observations. Jörgen Dalén (87-03-03)
- 33 A Multivariate Approach to Social Inequality. M Ribe (1987-05-05)
- 34 BORTFALLSBAROMETERN NR 2. Bortfall vid I-avdelningen, AKU och F-avdelningen t o m 1986. Håkan L. Lindström, Peter Lundquist och Tomas Garås (1987-08-01)
- 35 Use of Laboratory Methods and Cognitive Science for the Design and Testing of Questionnaires. Handbook of Methods. Judith T. Lessler (June 1987)

- 36 Some Optimal Problems When Estimating Household Data on the Basis of a "Primary", Stratified Sample of Individuals. Bengt Rosén (1987-08-10)
- 37 Early Survey Models and Their Impact on Survey Quality Work. Gösta Forsman (1987-09-17)
- 38 A Generalized Perks Formula for Old-Age Mortality. Sten Martinelle (1987-08-17)
- 39 Beskrivning av arbetstidens förklaringsgrad på arbetsinkomstens ojämnheter 1975-1984. Arne Sandström (1987-03-31)

Kvarvarande exemplar av ovanstående promemorior kan rekvideras från Elseliv Lindfors, U/STM, SCB, 115 81 Stockholm, eller per telefon 08 7834178

G-90693-★1016 A -001-K-H

CATARINA ELFFORS

A/BF-S