Arbetsförmedlingen

Imputeringsmodell för saknad sysselsättningsstatus i arbetslöshetsdata

Författare: Petra Nilsson
Datum: 2017-03-23
Diarienummer: Af-2017/00153440

Innehåll

Sammanfattning[1]

Ofullständig data är ett fundamentalt problem vid utvärderingar av arbetsmarknadspolitiska insatser. I Arbetsförmedlingens register avslutas 20 procent av alla arbetslöshetsperioder av okänd orsak. Detta leder till en osäkerhet om vilken sysselsättningsstatus (sysselsatt eller inte sysselsatt) individen har efter sin registrerade arbetslöshetsperiod. Enligt fullständiga kombinerade administrativa data är sysselsättningsnivån bland dem som lämnat Arbetsförmedlingen av okänd orsak nära 50 procent mellan 1992 och 2006. Sedan dess har sysselsättningsnivån bland dem som lämnar Arbetsförmedlingen av okänd orsak minskat till 40 procent. Denna rapport använder en imputeringsansats för att undersöka om en imputeringsmodell kan användas för att komma åt problemet med ofullständig data. Imputeringsmodeller skattas på enkätdata från 2005/2006 och på fullständiga administrativa data från 2005/2006 och 2011/2012. Modellerna utvärderas i termer av deras förmåga att göra korrekta prediktioner. Modellerna har relativt hög prediktiv förmåga.

1 Inledning

Aktiva arbetsmarknadspolitiska åtgärder används i många länder för att bekämpa arbetslöshet. Utvärderingar av hur insatserna fungerar och i vilken utsträckning de leder till arbete är viktigt för att kunna utveckla insatsernas kvalitet och för att kunna dra policyslutsatser. Bristfälliga uppgifter om individers sysselsättningsstatus (sysselsatt eller inte sysselsatt) efter en period av arbetslöshet är ett fundamentalt problem vid effektutvärderingar. Detta framgår av exempelvis Wilke (2009) som konstruerar intervall för arbetslöshetstider utifrån brittiska data för att beskriva effekten av saknad information när det gäller längden på arbetslöshetstiden. Arntz med flera (2007) gör en liknande studie på tyska data.

I svenska data, enligt Arbetsförmedlingens register, avslutas ungefär 20 procent av arbetslöshetsperioderna av okänd orsak. De grupper av individer som är överrepresenterade bland avaktualiserade av okänd orsak är ungdomar, män, utomnordiskt födda, personer med lägre utbildningsnivå, personer som bor i storstadslän (Stockholm, Västra Götaland och Skåne) och personer som inte är medlemmar i en arbetslöshetskassa. Personer med funktionsnedsättning avregistreras av

[1] För en version av rapporten på engelska, se Nilsson, P. (2016). An Imputation Model for Dropouts in Unemployment Data. *Journal of Official Statistics*, 32(3), 719-732.

okänd orsak i mindre utsträckning än personer utan funktionsnedsättning.

Statistiska Centralbyrån (SCB) hanterar problemet med saknad information om sysselsättningsstatus genom att imputera. Den imputeringsmodell som SCB använder bygger på data från en liten undersökning från 1994, för information om modellen se Bring och Carling (2000). Föreliggande rapport är en utveckling av denna metod som presenterar imputeringsmodeller som baseras på mer omfattande och mer aktuell data, både enkätdata och registerdata. De registerdata som används är data från den longitudinella integrationsdatabasen för sjukförsäkrings- och arbetsmarknadsstudier (LISA) som innehåller information om sysselsättning för november månad varje år. De enkätdata som används kommer från en större undersökning genomförd av Arbetsförmedlingen under 2005 och 2006. Metodutvecklingen handlar om att svarsbortfallet i enkätdata imputeras innan modellerna estimeras.

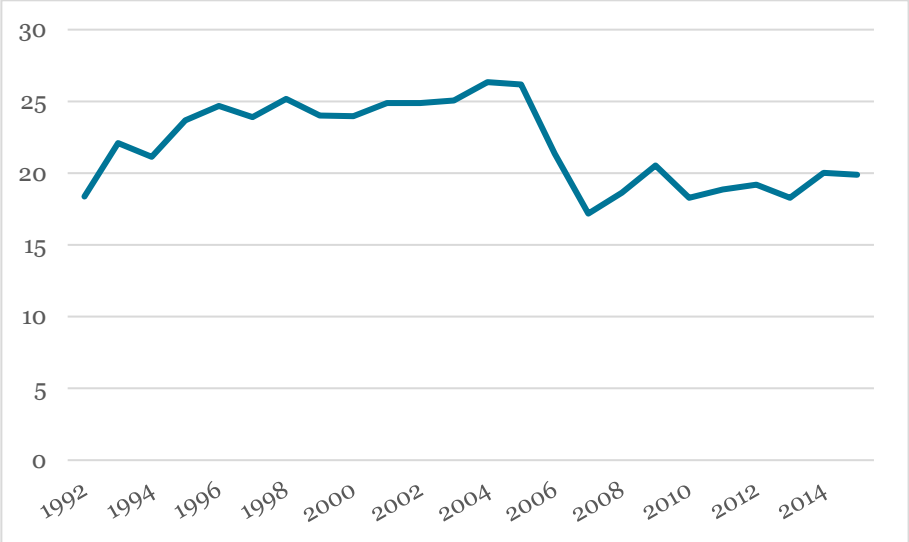
Flera olika imputeringsmodeller skattas och dessa utvärderas genom att studera överensstämmelsen mellan predikterade värden och värden enligt enkätdata och registerdata. Prediktionsförmågan för modellerna i denna studie jämförs också med prediktionsförmågan för Brings och Carlings modell och med slumpmässig imputering.

I rapporten presenteras dessutom sysselsättningsnivån över tid för dem som lämnat Arbetsförmedlingen av okänd orsak baserat på fullständiga kombinerade administrativa data för november månad respektive år.

2 Beskrivning av problemet och tillvägagångssättet

På Arbetsförmedlingen registreras de arbetssökande när de påbörjar en arbetslöshetsperiod och följs till dess att arbetslöshetsperioden avslutas med en så kallad avaktualisering. En avaktualisering innebär att den arbetssökande avregistreras och inte längre är aktuell som aktivt arbetssökande hos Arbetsförmedlingen. Om en arbetssökande inte upprätthåller kontakten med Arbetsförmedlingen och arbetsförmedlaren inte vet anledningen till det avaktualiseras den arbetssökande av okänd orsak. Figur 1 visar andelen arbetslöshetsperioder som avslutats av okänd orsak mellan 1992 och 2015.

Figur 1 Andel arbetslöshetsperioder som avslutats av okänd orsak av samtliga avslutade arbetslöshetsperioder, 1992-2015, procent.



Andelen avaktualiseringar av okänd orsak ökade trendmässigt fram till 2005. Under 2006 och 2007 sjönk andelen för att de senaste åren ha legat kring 20 procent. Nedgången beror på bättre administrativa rutiner på Arbetsförmedlingen. Arbetsförmedlarna följer i högre grad upp varför de arbetssökande inte upprätthåller kontakten med förmedlingen och avaktualiserar därmed de arbetssökande av okänd orsak i lägre utsträckning.

Tidigare undersökningar visar att en stor del av de arbetssökande som avaktualiserats av okänd orsak har gått till arbete.[2] I Arbetsförmedlingens resultatredovisningar rapporteras bara kända övergångar till arbete vilket innebär att antalet övergångar till arbete underskattas. Vid analyser av utflödet till arbete bör korrigeringar göras så att även de som fått arbete bland avaktualiserade av okänd orsak

[2] Bring och Carling (2000) och Nilsson (2010).

räknas med. Det är särskilt viktigt när till exempel utflödet till arbete för två grupper av arbetssökande jämförs och andelen med avaktualisering av okänd orsak är olika stor för grupperna. Ett enkelt sätt att justera är att anta att en viss andel av de avaktualiserade av okänd orsak har gått till arbete.

I mer avancerade analyser av data är det inte säkert att det räcker att anta att en viss andel har gått till arbete. I effektutvärderingar kan varje enskild arbetssökande som avaktualiserats av okänd orsak behöva tilldelas en sysselsättningsstatus. Att ersätta saknade värden med ett nytt värde på detta sätt kallas att imputera. Det finns mer eller mindre avancerade sätt att imputera. Enklast möjliga är att låta slumpen avgöra vilka av de arbetssökande som ska ges utfallet arbete. Mycket tyder dock på att det inte är ett särskilt bra tillvägagångssätt eftersom de arbetssökandes sannolikhet att gå till arbete inte nödvändigtvis är lika.

Imputeringsmodellerna som skattas i denna rapport är logistiska regressionsmodeller som ger en sannolikhet för varje individ att denne har fått ett arbete beroende på socioekonomiska och arbetsmarknadsrelaterade variabler som till exempel ålder, utbildningsnivå och antal tidigare övergångar till arbete. För att kunna konstruera imputeringsmodellerna behövs fullständig data med information om sysselsättningsstatus för ett urval individer som lämnat Arbetsförmedlingen av okänd orsak. Information om sysselsättningsstatus hämtas dels från en enkätundersökning där vi frågar individerna om deras sysselsättningsstatus några veckor efter det att de avaktualiserats, dels hämtas den från SCB:s registerdata där sysselsättningsstatus i november månad varje år finns registrerad.

3 Data

Både de enkätdata och de registerdata som används i denna studie kan innehålla mätfel. Bound, Brown and Mathiowetz (2001) undersöker orsaker till mätfel i enkätdata. De finner att ju längre sedan något inträffade och ju mer vanligt förekommande desto svårare är det att samla informationen. Socialt oönskade händelser tenderar att inte rapporteras medan det omvända gäller för socialt önskade händelser. Liknande resultat finns i till exempel Pyy-Martikanen och Rendtel (2009) som använder länkade finska enkät- och registerdata för att analysera mätfel i enkätdata. I registerdata om sysselsättning i LISA kan det finnas icke observerad eller felregistrerad information eftersom inte alla individuella uppgifter om sysselsättning registreras i den administrativa processen.

3.1 Enkätdata

3.1.1 Urvalsdesign och mätmetod

Enkätundersökningen genomfördes vid tolv olika mättillfällen mellan september 2005 och augusti 2006. Varje mättillfälle inkluderade avregistreringar av okänd orsak under en vecka, där den arbetssökande inte återvände till Arbetsförmedlingen inom två veckor. Detta för att kunna exkludera återflödet i urvalet. Ett obundet slumpmässigt urval av 300 arbetslöshetsperioder gjordes varje mätvecka. Det totala urvalet var 3 600 arbetslöshetsperioder. För att beakta säsongseffekter inkluderades mätveckor från så många olika perioder som möjligt under ett år.

Undersökningen kan ses som ett stratifierat urval med mätvecka som strata. En stratifiering av en ändlig population $U = \{1, \ldots, k, \ldots, N\}$ betyder att U delas in i H delmängder av populationen (Lundström and Särndal 2001). Antalet element i stratum h benämns N_h och urvalsstorleken i stratum h benämns n_h. Sannolikheten att ett givet element finns i urvalet, inklusionssannolikheten, ges av

$$\pi_k = \frac{n_h}{N_h}.$$

Låt

$$d_k = \frac{1}{\pi_k}$$

stå för designvikten för element *k*.

Notera att det är en arbetslöshetsperiod och inte en individ som utgör ett element i undersökningen. Populationen består av unika arbetslöshetsperioder, men inte av unika individer eftersom vissa personer förekommer i data flera gånger. Utfallen av arbetslöshetsperioder för samma person är antagligen korrelerade med varandra och kan inte antas vara oberoende. Detta problem studeras till exempel i den ekonomiska litteraturen (Lancaster 1979; Heckman and Singer 1982).

Korrelationsstrukturen ignoreras i denna studie vilket kan leda till en underskattning av variabiliteten för imputeringsmodellerna. En mycket stor andel av arbetslöshetsperioderna refererar till unika individer. 97,4 procent av arbetslöshetsperioderna som avslutades av okänd orsak under mätveckorna gäller unika personer. Av de perioder som utgör urvalsram för undersökningen, perioder där individerna inte har återkommit inom två veckor, refererar 98,5 procent till unika personer. I urvalet är 99,8 procent unika individer. Tre individer återfinns två gånger.

Enkätundersökningen genomfördes som datorstödda telefonintervjuer på Arbetsförmedlingens intervjuenhet. För att intervjupersonerna skulle komma ihåg hur deras arbetssituation såg ut när de upphörde att ha kontakt med Arbetsförmedlingen gjordes intervjuerna så nära inpå avregistreringen från Arbetsförmedlingen som möjligt, två till tre veckor. Eftersom två veckor inväntades för att återflödet till Arbetsförmedlingen skulle kunna exkluderas var det inte möjligt att ha intervjuerna närmare inpå avregistrering än så.

En fråga ställdes till individerna i undersökningen (”Hur är din arbetssituation idag?”). Svarsalternativen var följande:

1. Har arbete (heltid)
2. Har arbete (deltid)
3. Studerar/går i utbildning
4. Deltar i arbetsmarknadspolitiskt program
5. Har startat eget
6. Sjukskriven/sjukledig/föräldraledig
7. Arbetslös/söker jobb
8. Annat

Individer som svarade enligt alternativ 1,2 eller 5 definierades som att de fått arbete. Intervjuarna läste inte upp svarsalternativen till frågan. I de fall intervjupersonen hade svårt att ge något entydigt svar som passade med svarsalternativen hjälpte intervjuaren till att tolka syftet med frågan.

3.1.2 Hantering av svarsbortfall

Av den totala urvalsstorleken på 3 600 perioder fick vi svar för 2443, vilket ger en svarsfrekvens på 68 procent. Svarsbortfallet i undersökningen är således 32 procent. Äldre, utlandsfödda, lågutbildade och personer utan medlemskap i en arbetslöshetskassa är överrepresenterade bland bortfallet. Dessa personer har antagligen fått jobb i mindre utsträckning än de som har svarat i undersökningen, se till exempel Bennmarker, Carling and Forslund (2007) för vilka grupper av arbetslösa som riskerar långtidsarbetslöshet.

Det finns en risk för skevhet i form av en systematisk överskattning om bara data från de svarande används. Detta hanteras genom att imputera värden för bortfallet i undersökningen. Eftersom både socio-ekonomiska och arbetsmarknadsrelaterade variabler kan länkas till de individer som lämnar Arbetsförmedlingen av okänd orsak kan avsaknad information om sysselsättningsstatus imputeras med hjälp av en logistisk regressionsmodell som förklarar vilka kategorier av individer som har den största sannolikheten att ha fått arbete. Rubin (1996) rekommenderar att en imputeringsmodell innehåller så många relevanta variabler som möjligt och i denna studie används variabler som Bennmarker, Carling och Forslund (2007) har funnit förklarar sannolikheten att ha fått arbete.[3]

Hur regressionsimputering går till beskrivs i Lundström och Särndal (2001). I denna studie utgör en inskrivningsperiod ett element. y_k anger om en inskrivningsperiod k har avslutats på grund av arbete eller inte, $y_k = 1$ om period k avslutats på grund av arbete och $y_k = 0$ om period k inte avslutats på grund av arbete. Enkätsvar finns för elementen i ett set som benämns *r*. Regressionen ger ett imputerat värde för element k enligt

$$\hat{y}_k = z_k' \hat{\beta}$$

[3] De förklarande variablerna är: kvinna, 16-24 år, 35-44 år, 45-66 år, född i Norden, född utomlands, funktionsnedsättning, grundskola, högre utbildning <= 2 år, högre utbildning > 2 år, erfarenhet inom sökt yrke, söker bara heltid, söker arbete utanför pendlingsavstånd, medlem i en arbetslöshetskassa, deltar i aktivitetsgarantin, status arbete före avregistrering, annan status innan avregistrering (inte arbete eller arbetslös), skogslän, övriga län (inte skogs- eller storstadslän), antal inskrivningsperioder under en femårsperiod, antal övergångar till arbete under en femårsperiod, månaderna maj till juli och interaktionstermerna 16-24 år och medlem i en arbetslöshetskassa, 35-44 år och medlem i en arbetslöshetskassa, 45-66 år och medlem i en arbetslöshetskassa, 16-24 år och erfarenhet, 35-44 år och erfarenhet and slutligen 45-66 år och erfarenhet.

där z_k är värdet på imputeringsvektorn för element k enligt $z_k = (z_{1k},...,z_{jk},...,z_{Jk})'$, en kolumnvektor med J förklarande variabler, där z_{jk} är värdet för element k, för den j^{te} förklarande variabeln och

$$\hat{\beta} = (\sum_r d_k z_k z_k')^{-1} \sum_r d_k z_k y_k .$$

$\hat{\beta}$ är en vektor av regressionskoefficienter efter anpassning av en regressionsmodell på data (y_k, z_k) tillgängliga för $k \in r$ och viktade med d_k.

Metoden är deterministisk och ger samma imputerade värde när den upprepas. Dataset med regressionsimputerade värden tenderar att ha lägre varians än data med observerade värden, y_k. Det är dock möjligt att addera en slumpmässigt vald residual. Då blir det imputerade värdet för element k

$$\hat{y}_k = z_k'\hat{\beta} + e_k^*$$

där e_k^* representerar en slumpmässigt vald residual från ett dataset som innehåller beräknade residualer $\{e_k : k \in r\}$, där

$$e_k = y_k - z_k'\hat{\beta}.$$

Vid multipel imputering imputeras flera olika värden för varje saknat värde. På så sätt skapas flera fullständiga dataset. Värdena y_k för de svarande är desamma i alla dataset medan de imputerade värdena är olika. Svarsbortfallet i enkätundersökningen imputeras 20 gånger vilket innebär att vi skapar 20 fullständiga dataset.

3.2 Registerdata

För att skapa fullständig data baserat på registerdata kombineras Arbetsförmedlingens register med uppgifter i LISA. Registeruppgifterna om sysselsättningsstatus i LISA avser november månad vilket betyder att de kombinerade dataseten också är begränsade till denna månad. Information om sysselsättningsstatus i november månad i LISA används för att bestämma sysselsättningsstatus för personer som lämnar Arbetsförmedlingen av okänd orsak i november varje år.

LISA-databasen innehåller årliga register och omfattar alla individer över 16 år som är registrerade i Sverige. Individerna klassificeras som

sysselsatta om de kan antas ha arbetat minst fyra timmar under november månad. Skattningen är modellbaserad där korrelationen mellan flera variabler, till exempel information om utbetalningar från arbetsgivare, används för klassificeringen. Det finns mätfel i form av felregistreringar i data. Risken för felregistreringar är större för personer som bara arbetar delar av året och för personer med en svagare anknytning till arbetsmarknaden. Felregistreringar beror delvis på fel i modellen, men också på ofullständig information.

Imputeringsmodeller skattas på kombinerade, fullständiga administrativa data för åren 2005/2006 respektive åren 2011/2012. Åren 2005/2006 är valda för att kunna jämföra en modell som bygger på registerdata med en modell som utgår från enkätdata för samma år. Åren 2011/2012 är valda för att kunna jämföra modeller som skattas på data från 2005/2006 med modeller som skattas på data från senare år.

Tabell 1 beskriver de kombinerade administrativa dataseten för åren 2005/2006 och 2011/2012. För vissa individer finns ingen information om sysselsättningsstatus i LISA. Detta partiella bortfall uppgår till 1,5 procent av individerna för åren 2005/2006 och till 2 procent för åren 2011/2012. Partiellt bortfall för sysselsättningsstatus i LISA är mer vanligt för personer i åldersgruppen 55-66 år, för utomeuropeiskt födda och för lågutbildade.

Tabell 1 Beskrivning av kombinerad administrativ data november 2005/2006 respektive november 2011/2012.

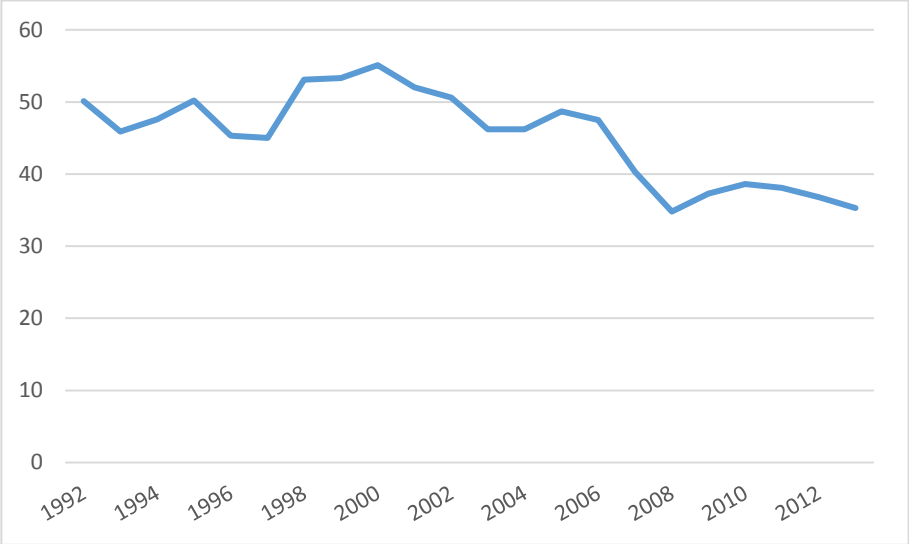
	2005/2006 Fullständig data	Partiellt bortfall	2011/2012 Fullständig data	Partiellt bortfall
Antal observationer	20 566	311	14 375	282
16-24 år (%)	45,7	15,8	43,8	20,9
55-66 år (%)	3,6	7,4	4,6	8,2
Född utanför Europa (%)	17,2	24,4	29,5	44
Funktionsnedsättning (%)	3,6	2,9	6,6	3,6
Grundskola (%)	26,2	52,8	29,5	44,3

4 Sysselsättningsnivå

4.1 Sysselsättningsnivå i november månad för åren 1992 till 2013 baserat på kombinerade administrativa data

Figur 2 visar andelen sysselsatta av dem som lämnat Arbetsförmedlingen av okänd orsak i november varje år beräknade utifrån fullständiga kombinerade administrativa data för åren 1992 till 2013.[4] Mellan åren 1992 och 2006 är andelen sysselsatta ungefär 50 procent. För åren 2007 och därefter har andelen sysselsatta sjunkit till knappt 40 procent.

Figur 2 Andelen sysselsatta av dem som lämnat Arbetsförmedlingen av okänd orsak i november varje år enligt fullständiga kombinerade administrativa data, 1992–2013, procent.



Nedgången från ungefär 50 procent till knappt 40 procent beror på de bättre administrativa rutinerna på Arbetsförmedlingen. Arbetsförmedlarna följer i högre grad upp varför de arbetssökande inte upprätthåller kontakten med förmedlingen och identifierar i högre utsträckning vilka som fått arbete.

4.2 Sysselsättningsnivå år 2005/2006 baserat på enkätdata

En skattad andel sysselsatta bland avregistrerade av okänd orsak tas också fram baserat på enkätdata från 2005/2006. Den skattade andelen bygger i detta fall på de 20 bortfallsimputerade replikaten av data, det vill säga 20 separata estimat skattas. De olika parameterestimaten kombineras sedan enligt Rubin (1987).

[4] Arbetsförmedlingen har inte tillgång till data om sysselsättningsstatus i LISA senare än november 2013.

Anta att $\hat{Q}_i$ är ett estimat av en skalär kvantitet av intresse, från dataset $i,\ i = 1, 2, \ldots, m$ och att $\hat{W}_i$ är variansen kopplat till $\hat{Q}_i$. Det övergripande estimatet är genomsnittet av de individuella estimaten från de m fullständiga dataseten:

$$\overline{Q} = \frac{1}{m}\sum_{i=1}^{m}\hat{Q}_i .$$

Anta att $\overline{W}$ är inom-imputering-varians, vilket är medelvärdet av estimaten från de m fullständiga dataseten,

$$\overline{W} = \frac{1}{m}\sum_{i=1}^{m}\hat{W}_i$$

och att B är mellan-imputering-varians,

$$B = \frac{1}{m-1}\sum_{i=1}^{m}(\hat{Q}_i - \overline{Q})^2 ,$$

då är den totala variansen

$$T = \overline{W} + (1 + \frac{1}{m})B .$$

Tabell 2 Skattad sysselsättningsnivå baserat på enkätundersökningen 2005/2006, totalt samt uppdelat på svarande respektive bortfallsgrupp.

	Estimat	Stand. fel	95 % Konfidens-intervall
Baserat både på de svarande och på bortfallsgruppen	47,3	1,1	(45,1; 49,4)
Baserat enbart på de svarande	50,7	1,0	(48,6; 52,7)
Baserat enbart på bortfallsgruppen (imputerade värden)	39,9	2,6	(34,8; 45,1)

Tabell 2 visar den skattade andelen sysselsatta baserat på enkätundersökningen från 2005/2006.

Den skattade andelen sysselsatta är 47,3 procent. Ett 95-procentigt konfidensintervall för andelen är plus/minus 2 procentenheter. Den skattade sysselsättningsnivån baserad på undersökningen ligger nära sysselsättningsnivån enligt fullständiga, kombinerade administrativa data för åren 2005 och 2006, vilken är 48,1 procent. Tabell 2 visar också den skattade sysselsättningsnivån baserat enbart på de svarande i undersökningen samt enbart på imputerade värden för bortfallsgruppen.

Den skattade andelen sysselsatta i 1994 års undersökning är 44, 7 procent (Bring och Carling 2000). Enligt fullständiga, kombinerade administrativa data för november 1994 är sysselsättningsnivån 47, 6 procent.

En faktor som påverkar jämförbarheten mellan enkätdata och registerdata är att registerdata refererar till november månad medan enkätdata refererar till januari och februari (1994 års undersökning) eller till 12 olika perioder under ett år (2005/2006 års undersökning).

Tabell 3 visar den skattade sysselsättningsnivån per mätvecka enligt enkätundersökningen från 2005/2006 tillsammans med standardavvikelser och 95 procentiga konfidensintervall.

Tabell 3 Resultat från enkätundersökningen 2005/2006, uppdelat på mätvecka.

	Estimat	Stand. fel	95 % Konfidens-intervall
Vecka 34 2005 (aug)	44.3	3.1	(38.1; 50.5)
Vecka 35 2005 (aug-sep)	44.8	3.1	(38.7; 50.9)
Vecka 36 2005 (sep)	43.0	3.1	(36.8; 49.1)
Vecka 37 2005 (sep)	47.1	3.4	(40.3; 53.9)
Vecka 40 2005 (okt)	46.6	3.4	(39.9; 53.4)
Vecka 3 2006 (jan)	43.9	3.3	(37.4; 50.5)
Vecka 5 2006 (jan-feb)	44.8	3.3	(38.4; 51.3)
Vecka 9 2006 (feb-mar)	48.6	3.3	(42.1; 55.0)
Vecka 14 2006 (apr)	52.4	3.2	(46.1; 58.8)
Vecka 22 2006 (maj-jun)	56.3	3.3	(49.9; 62.8)
Vecka 25 2006 (jun)	54.5	3.6	(47.3; 61.7)
Vecka 31 2006 (jul-aug)	45.6	3.2	(39.2; 51.9)

Den skattade andelen sysselsatta varierar mellan 43,0 och 56,3 beroende på mätvecka. Andelen sysselsatta är lägre i början av höst- och vårterminen eftersom arbetssökande då lämnar förmedlingen för studier i högre utsträckning. Andelen sysselsatta är istället högre i början av sommaren då många arbetssökande lämnar förmedlingen för sommarjobb.

5 Imputeringsmodeller för saknad sysselsättningsstatus

5.1 Modellernas parameterestimat

Imputeringsmodeller för saknad sysselsättningsstatus i arbetslöshetsdata skattas på enkätdata från 2005/2006 och på kombinerade, fullständiga administrativa dataset för 2005/2006 och 2011/2012. Modellerna är logistiska regressionsmodeller och beroende variabel är om individen är sysselsatt eller inte. Ekvationen

$$\hat{P}(Y_k = 1 | z_k) = \frac{1}{1 + \exp(-z_k \hat{B}')}$$

skattas där ålder, födelseland, funktionsnedsättning, utbildning, medlemskap i en arbetslöshetskassa, sysselsättningsstatus innan avregistrering och erfarenhet i sökt yrke är förklarande variabler. Ambitionen är att inkludera så många relevanta variabler som möjligt för att förbättra modellernas prediktionsförmåga samtidigt som modellerna ska vara så enkla som möjligt att använda. Variabler som används i Bring och Carling (2000) och i Bennmarker, Carling och Forslund (2007) har testats och variabler som har ett p-värde mindre än 0,05 vid skattningar på registerdata används slutligen.

Eftersom enkätdata innehåller observationer från olika perioder under året är det möjligt att inkludera månad för avregistrering som förklarande variabel i en modell som skattas på enkätdata. Två olika modeller estimeras på enkätdata, en med (modell 1) och en utan (modell 2) månad för avregistrering.

För enkätdata skattar vi imputeringsmodellerna på de 20 bortfallsimputerade dataseten, vilket betyder att 20 separata modeller skattas. De olika parameterestimaten kombineras sedan enligt Rubin (1987). Tabell 4 visar imputeringsmodellerna baserat på enkätdata.

Tabell 4 Imputeringsmodeller baserat på enkätdata från 2005/2006, 11 (modell 1) respektive 12 (modell 2) förklarande variabler.

	Modell 1			Modell 2		
	Estimat	Stand. fel	p-värde	Estimat	Stand. fel	p-värde
Intercept	0.14	0.14	0.33	0.07	0.14	0.61
16-24 år	-0.37	0.11	0.00	-0.37	0.11	0.00
55-66 år	-0.36	0.18	0.05	-0.37	0.18	0.05
Född utanför Europa	-0.53	0.12	<.0001	-0.53	0.12	<.0001
Funktionsnedsättning	-0.37	0.31	0.23	-0.38	0.31	0.22
Grundskola	-0.30	0.12	0.02	-0.31	0.13	0.02
Medlem i en arbetslöshetskassa	0.52	0.11	<.0001	0.52	0.11	<.0001
Status arbete innan avregistrering	0.84	0.14	<.0001	0.85	0.14	<.0001
Status övrig innan avregistrering	-0.48	0.13	0.00	-0.50	0.13	<.0001
Antal arbetslöshetsperioder	-0.16	0.03	<.0001	-0.16	0.03	<.0001
Antal övergångar till arbete	0.31	0.06	<.0001	0.31	0.06	<0.001
Ingen erfarenhet	-0.15	0.12	0.21	-0.15	0.12	0.21
Avregistrering i maj, juni eller juli				0.50	0.12	<.0001

Parameterestimaten visar att sysselsättningsnivån bland personer som avregistreras av okänd orsak är lägre för äldre personer, för lågutbildade, för personer födda utanför Europa och för personer med en funktionsnedsättning. Personer registrerade som deltidsanställda eller tillfälligt anställda innan avregistreringen är sysselsatta i högre grad än personer som är registrerade som arbetslösa. Personer med många tidigare övergångar till arbete och medlemmar i en arbetslöshetskassa är också sysselsatta i högre grad. Den alternativa modellen (modell 2) där månad för avregistrering inkluderas visar att de som avregistrerats i maj, juni eller juli är sysselsatta i högre grad.

Tabell 5 visar imputeringsmodeller estimerade på fullständiga, kombinerade administrativa data för 2005/2006 (modell 3) respektive 2011/2012 (modell 4).

Tabell 5 Imputeringsmodeller baserade på fullständiga, kombinerade administrativa data för 2005/2006 (modell 3) respektive 2011/2012 (modell 4).

	Modell 3			Modell 4		
	Estimat	Stand. fel	p-värde	Estimat	Stand. fel	p-värde
Intercept	-0.50	0.05	<.0001	-0.67	0.06	<.0001
16-24 år	0.22	0.04	<.0001	0.42	0.05	<.0001
55-66 år	-0.43	0.09	<.0001	-0.48	0.10	<.0001
Född utanför Europa	-0.50	0.05	<.0001	-0.20	0.05	<.0001
Funktionsnedsättning	-0.52	0.10	<.0001	-0.50	0.10	<.0001
Grundskola	-0.54	0.04	<.0001	-0.64	0.05	<.0001
Medlem i en arbetslöshetskassa	1.22	0.04	<.0001	1.26	0.05	<.0001
Status arbete innan avregistrering	1.66	0.05	<.0001	1.86	0.07	<.0001
Status övrig innan avregistrering	-0.24	0.05	<.0001	-0.45	0.07	<.0001
Antal arbetslöshetsperioder	-0.24	0.01	<.0001	-0.24	0.02	<.0001
Antal övergångar till arbete	0.34	0.02	<.0001	0.40	0.03	<.0001
Ingen erfarenhet	-0.41	0.04	<.0001	-0.64	0.06	<.0001

Tolkningen av estimaten är i stort sett desamma som för tabell 4. En skillnad är att estimatet för interceptet och estimatet för personer 16-24 år har bytt tecken jämfört med modellerna baserade på enkätdata. Standardfelen för modellerna skattade på de kombinerade administrativa dataseten är ungefär hälften så stora som för modellerna skattade på enkätdata.

5.2 Modellernas prediktionsförmåga

Imputeringsmodellernas prediktionsförmåga ger information om modellerna kan användas för att imputera ett värde för sysselsättning för dem som lämnar Arbetsförmedlingen av okänd orsak. Med prediktionsförmåga menas andelen korrekta prediktioner.

Prediktionsförmågan jämförs för varje modell i ovanstående tabell 4 och tabell 5 (modell 1-4) för både enkätdata från 2005/2006 och för fullständiga, kombinerade administrativa data från 2005/2006 och 2011/2012. Prediktionsförmågan för respektive modell relateras också till prediktionsförmågan för slumpmässig imputering och för Bring och Carlings modell som bygger på enkätdata från 1994. Dessutom utvärderas modellerna mot samma data som de skattats på genom att

skatta modellerna på 60 procent av data och sedan göra prediktioner för återstående 40 procent.

Imputeringsmodellerna ger en sannolikhet mellan 0 och 1 för att individerna har fått ett arbete. Imputering kräver ett gränsvärde för hur prediktionerna ska klassificeras, det vill säga vilka prediktioner som ska klassificeras som att individen har fått arbete, $\hat{y}_k = 1$, eller inte fått arbete, $\hat{y}_k = 0$. För varje modell används ett gränsvärde så att imputeringen ger samma andel sysselsatta som i data. För enkätdata är sysselsättningsgraden 47,3 medan den för administrativa data från 2005/2006 är 48,1 och för 2011/2012 är 37,4.

Tabell 6 visar prediktionsförmågan för varje modell för olika data. För enkätdata från 2005/2006 (rad 1 i tabellen) har slumpmässig imputering (modell 6) lägst prediktionsförmåga och 50 procent korrekta prediktioner. Imputeringsmodellen som utgår från enkätdata från 1994 (modell 5) har 54 procent korrekta prediktioner. Imputeringsmodellerna som skattats i denna studie (modell 1-4) har högre prediktionsförmåga, 68 procent för alla modeller.

För fullständiga administrativa data från 2005/2006 (rad 2 i tabellen) har modellerna som bygger på administrativa data högre prediktionsförmåga än modellerna som använder enkätdata. Modellen som är baserad på registerdata från 2005/2006 har högst prediktionsförmåga, 76 procent.

För registerdata från 2011/2012 (rad 3 i tabellen) har också modellerna som använder registerdata högre prediktionsförmåga jämfört med modellerna som utgår från enkätdata. Modellerna som använder registerdata från 2005/2006 respektive 2011/2012 har en prediktionsförmåga på 74 procent.

Tabell 6 Andelen korrekta prediktioner beroende på modell för olika dataset.

	Modell 1 Enkät 2005/2006 11 variabler	Modell 2 Enkät 2005/2006 12 variabler	Modell 3 Register 2005/2006	Modell 4 Register 2011/2012	Modell 5 Enkät 1994	Modell 6 Slump
Data: Enkät 2005/2006	68	68	68	68	54	50
Data: Register 2005/2006	74	74	76	75	58	50
Data: Register 2011/2012	72	72	74	74	61	53

6 Slutsatser och avslutande diskussion

Imputering kan användas vid analyser av arbetslöshetsdata där information om sysselsättningsstatus saknas. Två imputeringsmodeller som bygger på enkätdata och två modeller baserade på kombinerade administrativa data har skattats och utvärderats. Modellerna med registerdata har en något högre prediktionsförmåga än modellerna som använder enkätdata. Standardfelen för modellerna skattade på registerdata är ungefär hälften så stora som för modellerna skattade på enkätdata.

De två imputeringsmodeller som använder enkätdata från 2005/2006 baseras på mer data och har högre prediktionsförmåga än den modell som är skattad på enkätdata från 1994. Dessutom har multipel imputering använts vilket innebär att svarsbortfallet hanterats på ett mer tillfredsställande sätt. Den estimerade andelen sysselsatta för dem som lämnar Arbetsförmedlingen av okänd orsak är 47 procent baserat på enkätdata från 2005 och 2006 vilket är konsistent med kombinerade administrativa data för samma år.

En skillnad mellan enkätdata och kombinerade administrativa data är att administrativa data avser uppgifter för november månad respektive år medan enkätdata refererar till 12 olika mättillfällen under 2005 och 2006. Det finns ingen information om imputeringsmodellernas prognosförmåga när det gäller alla arbetssökande som avregistreras av okänd orsak under ett år. Det går därför inte att med säkerhet säga vilken modell som är bäst. Men antagligen spelar det inte så stor roll vilken av modellerna som används, prognosförmågan är ungefär lika stor oavsett modell för de tre olika datamaterial som studeras här. Ett förslag är att använda imputeringsmodellen baserad på registerdata från november 2011/2012.

Referenser

Arntz, M., S. Lo, och R. Wilke. 2007. “Bounds Analysis of Competing Risks: A Nonparametric Evaluation of the Effect of Unemployment Benefits on Migration in Germany.” ZEW - Centre for European Economic Research Discussion Paper No. 07-049. Doi: http://dx.doi.org/10.2139/ssrn.1010286.

Bennmarker, H., K. Carling, och A. Forslund. 2007. Vem blir långtidsarbetslös? Rapport 2007:29. Uppsala: Institutet för arbetsmarknads- och utbildningspolitisk utvärdering (IFAU). Hämtad från: http://www.ifau.se/globalassets/pdf/se/2007/r07-20.pdf (2016-04-01).

Bound, J., C. Brown, och N. Mathiowetz. 2001. “Measurement error in survey data.” I J. Heckman och E. Leamer, redaktörer. Handbook of econometrics, vol. 5, 3705–3833. Amsterdam: Elsevier. Hämtad från: http://www.psc.isr.umich.edu/pubs/pdf/rr00-450.pdf (2016-04-01).

Bring, J. och K. Carling. 2000. “Attrition and Misclassification of Drop-outs in the Analysis of Unemployment Duration.” Journal of Official Statistics 16: 321–330. Hämtad från: http://www.jos.nu/Articles/abstract.asp?article¼4164321 (2016-04-01).

Heckman, J. och B. Singer. 1982. “Population Heterogeneity in Demographic Models.” I K. Land och A. Rogers, redaktörer. Multidimensional Mathematical Demography, 567–599. New York: Academic Press. Hämtad från: http://www.popline.org/node/410098 (2016-04-01).

Lancaster, T. 1979. “Econometric Methods for the Duration of Unemployment.” Econometrica 47: 939–956. Doi: http://dx.doi.org/10.2307/1914140.

Lundström, S. och C-E. Särndal. 2001. Estimation in the Presence of Nonresponse and Frame Imperfections. Örebro: SCB-Tryck.

Nilsson, P. 2010. ”Arbetssökande som lämnar Arbetsförmedlingen av okänd orsak.” Stockholm: Arbetsförmedlingen. (Working paper 2010:1). Hämtad från: http://www.arbetsformedlingen.se/download/18.749929de128499e606080006854/1401114885581/workingpaper10_1.pdf (2016-04-01).

Pyy-Martikainen, M. och U. Rendtel. 2009. “Measurement Errors in Retrospective Reports of Event Histories. A Validation Study with Finnish Register Data.” Survey Research Methods 3: 139–155. Doi: http://dx.doi.org/10.18148/srm/2009.v3i3.2372.

Rubin, D.B. 1987. Multiple Imputation for Nonresponse in Surveys. New York: John Wiley.

Rubin, D.B. 1996. “Multiple Imputation After 18 Years (with discussion).” Journal of the American Statistical Association 91: 473–489. Doi: http://dx.doi.org/10.1080/01621459.1996.10476908.

Wilke, R. 2009. “Unemployment Duration in the United Kingdom: An Incomplete Data Approach.” Doi: http://dx.doi.org/10.2139/ssrn.1348019.