Arbetssökande som lämnar Arbetsförmedlingen av okänd orsak

Petra Nilsson

Working Paper 2010:1

Arbetsförmedlingen

Arbetsförmedlingens Working Paper serie presenterar rapporter som rör analys av arbetsmarknadens funktionssätt och effekter av arbetsmarknadspolitiska insatser. Rapporterna är pågående arbete och författarna tar tacksamt emot synpunkter.

Working papers kan laddas ned från
www.arbetsformedlingen.se

Arbetsförmedlingens huvudkontor
Forskningsenheten
113 99 Stockholm
E-post: forskningsenheten@arbetsformedlingen.se

Arbetssökande som lämnar Arbetsförmedlingen av okänd orsak

-Andelen som fått arbete

-En imputeringsmodell

Petra Nilsson

3 mars 2010

Sammanfattning

Arbetsförmedlingen har som uppgift att bidra till en effektiv matchning mellan arbetssökande och lediga arbeten. Ett viktigt resultatmått för myndigheten är hur många arbetssökande som får arbete. Ett problem med detta mått är att det för en femtedel av de arbetssökande saknas information om sysselsättningsstatus, individerna har av okänd orsak avbrutit kontakten med Arbetsförmedlingen. I rapporten redovisas resultat från en urvalsundersökning från 2005 och 2006 riktad till arbetssökande som lämnat Arbetsförmedlingen av okänd orsak. Vid tidpunkten för den avbrutna kontakten hade 39 procent fått arbete. En imputeringsmodell för arbetssökande som lämnat Arbetsförmedlingen av okänd orsak presenteras. Modellen kan användas för att prediktera ifall individerna har fått arbete eller inte. Imputeringsmodellen är en logistisk regressionsmodell med ett tjugotal förklarande variabler. De förklarande variablerna är hämtade från Arbetsförmedlingens register.

Innehåll

1 Inledning

1.1 Problem att redovisa utflöde till arbete inom Arbetsförmedlingens verksamhet

Arbetsförmedlingen är en offentlig myndighet för personer som söker arbete och för arbetsgivare som söker arbetskraft. Myndigheten har som uppgift att bidra till en effektiv matchning mellan arbetssökande och lediga arbeten, att anpassa arbetsmarknadspolitiska åtgärder till efterfrågan på arbetskraft samt att korta arbetslöshetstiderna för de som står långt från arbetsmarknaden. Arbetsförmedlingen producerar en mängd statistikredovisningar och analyser för att belysa hur väl myndigheten uppfyller sina mål. Viktiga resultatmått är hur många arbetssökande som får arbete och hur lång tiden till arbete är.

Vad gäller dessa resultatmått finns det problem i Arbetsförmedlingens register. Varje arbetssökande registreras 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. Problemet är att en mycket stor andel av de arbetssökande på Arbetsförmedlingen avaktualiseras av okänd orsak. Anledningen till att koden "okänd orsak" används är att arbetsförmedlaren inte vet av vilken anledning den sökande inte längre upprätthåller kontakten med förmedlingen. Ett uteblivet besök kan till exempel vara anledning till en avaktualisering av okänd orsak.

Andelen som avregistrerats av okänd orsak har varierat över tid.[1] I början av 1990-talet var den knappt 20 procent. Trendmässigt ökade sedan denna andel för att år 2004 och 2005 vara så hög som 26 procent. Efter det minskade andelen för att 2007 vara nere på 17 procent. De senaste åren har andelen

[1] Källa: Arbetsförmedlingens datalager. Andelen som avaktualiserats till arbete under 1990- och 2000-talet varierar mellan 38 och 59 procent per år. Det finns ett negativt samband mellan andelen som avaktualiserats till arbete och andelen som avaktualiserats av okänd orsak. När andelen till arbete är hög är andelen som avbrutit kontakten av okänd orsak låg och vice versa.

återigen ökat något för att nu vara omkring 20 procent. De grupper av arbetssökande som avaktualiseras av okänd orsak är inte riktigt desamma som de som avregistreras totalt sett. Ungdomar är överrepresenterade bland avaktualiserade av okänd orsak, liksom män, utomnordiskt födda, personer med lägre utbildningsnivå, personer som bor i storstadslän (Stockholm, Västa Götaland och Skåne) samt personer som saknar arbetslöshetskassa. Funktionshindrade avregistreras av okänd orsak i mindre utsträckning än icke funktionshindrade.

Oavsett vilken tidsperiod som avses så är andelen som avaktualiseras av okänd orsak hög vilket får konsekvenser i beräkningar av utflödet till arbete. En stor del av de arbetssökande som avaktualiseras av okänd orsak har nämligen i själva verket gått till arbete. Tas inte hänsyn till detta i statistikredovisningar och i olika typer av analyser blir resultaten missvisande. I enkla redovisningar av antal personer som har lämnat förmedlingen för arbete underskattas det verkliga antalet som fått arbete. I analyser av arbetslöshetstider menar Bring och Carling (2000) att ett antagande om att övergångar till arbete är oberoende av avaktualiseringar av okänd orsak inte håller. Övergångssannolikheten till arbete underskattas med 20 procent.

1.2 Möjliga lösningar

För att inte underskatta hur många som lämnat Arbetsförmedlingen för arbete bör korrigeringar göras så att även de som fått arbete bland avaktualiserade av okänd orsak räknas med. Det är särskilt viktigt när till exempel utflödet till arbete för två grupper av arbetssökande ska jämföras och andelen avaktualiseringar 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 med att anta att en viss andel har gått till arbete. I ett datamaterial kan varje arbetssökande som avaktualiserats av okänd orsak behöva ges ett specifikt

värde på om de gått till arbete eller inte. Att ersätta saknade värden med ett nytt värde på detta sätt kallas att imputera. Det finns mer eller mindre avancerade sätt att ta fram dessa nya värden. 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. De arbetssökandes sannolikhet att gå till arbete är inte lika.

Det som istället skulle behöva användas är en imputeringsmodell. Bring och Carling (2000) argumenterar också för att en imputeringsmodell är det bästa sättet att angripa problemet. De föreslår att man gör en urvalsundersökning av de som lämnar Arbetsförmedlingen av okänd orsak så ofta som möjligt för att sedan använda informationen i en modell. Imputeringsmodellen bör innehålla alla de faktorer som påverkar övergången till arbete. Modellen kan sedan användas för att tilldela de arbetssökande med störst sannolikhet att ha fått arbete det utfallet.

1.3 Syfte

Syftet med den här rapporten är dels att undersöka hur stor andel som fått arbete bland de som avbrutit kontakten med Arbetsförmedligen, dels att utveckla en imputeringsmodell som predikterar huruvida en individ som lämnat Arbetsförmedlingen av okänd orsak har fått arbete eller inte.

1.4 Metod

En urvalsundersökning görs för att ta reda på hur stor andel som fått arbete bland avregistrerade av okänd orsak samt för att skatta en imputeringsmodell som förklarar vilka individer som fått arbete. Undersökningar har dock i de allra flesta fall ett svarsbortfall vilket riskerar att snedvrida resultaten. Även undersökningen som ligger till grund för detta arbete har ett bortfall, vilket hanteras genom att imputera värden för bortfallet.

Det kompletta datamaterialet från urvalsundersökningen med imputera-

de värden används sedan för att beräkna hur många som fått arbete och för att konstruera en imputeringsmodell som kan prediktera huruvida arbetssökande som avbrutit kontakten med Arbetsförmedlingen av okänd orsak fått arbete eller inte. Imputeringsmodellen är en logistisk regressionsmodell med arbete eller inte som utfallsvariabel och ett tjugotal variabler från Arbetsförmedlingens register som förklarande variabler.

Imputeringen av bortfallet i undersökningen går till så att en logistisk regressionsmodell som förklarar vilka som fått arbete bland de svarande i urvalsundersökningen skattas. Denna modell används sen för att prediktera huruvida individerna i bortfallet fått arbete eller inte. Multipla imputeringar för varje saknat värde görs, dvs flera imputeringar görs för varje saknat värde.

1.5 Disposition

Efter detta inledande kapitel följer ett kapitel där den data som används presenteras. Urvalsundersökningen och de registeruppgifter från Arbetsförmedligen som går att koppla till undersökningen beskrivs. I följande kapitel, kapitel 3 och 4, beskrivs de metoder som används för att imputera bortfallet i undersökningen. I kapitel 3 skattas en logistisk regressionsmodell på de svarande i undersökningen som också utvärderas mot faktiska utfall. Den logistiska regressionsmodellen används sedan i kapitel 4 för multipel imputering av bortfallet.

I kapitel 5 och 6 används det kompletta datamaterialet från undersökningen med skattade, imputerade värden. I kapitel 5 redovisas resultaten från urvalsundersökningen i form av andel som fått arbete samt konfidensintervall. I kapitel 6 presenteras imputeringsmodellen som predikterar ifall individer som avbrutit kontakten med Arbetsförmedlingen av okänd orsak har fått arbete eller inte. I det sjunde och sista kapitlet diskuteras resultaten samt hur de kan användas.

2 Data

I det här arbetet används data från en urvalsundersökning samt registeruppgifter från Arbetsförmedlingen.

2.1 Data i Arbetsförmedlingens register

En mängd registeruppgifter kan kopplas till individerna i urvalsundersökningen. Arbetsförmedlingen för ett register över individer som skrivit in sig som arbetslösa. Historiska data sedan 1992 finns samlat i ett datalager.

När en person skriver in sig på Arbetsförmedlingen registreras en mängd information om personen som kön, ålder, sökt yrke osv. Datum för inskrivning registreras liksom en så kallad sökandekategori. Kategori 11 (arbetslös) är vanligast vid inskrivning, men en arbetssökande kan också registreras som exempelvis 21 (deltidsarbetslös) eller 14 (arbetssökande med förhinder). Arbetssökande med förhinder kan vara studerande, föräldralediga eller sjuka som inte omedelbart kan tillträda ett arbete, men som ändå vill vara registrerade som arbetssökande på Arbetsförmedlingen.

Under arbetslöshetsperioden registreras sedan händelser som t ex att den arbetssökande får arbetspraktik (sökandekategori 54) eller arbetsmarknadsutbildning (sökandekategori 81). Det finns en rad olika program för de arbetssökande. Sökandekategorierna beskriver också arbetslöshetsstatus som sökandekategori 31, tillfälligt arbete t ex. Alla händelser registreras med datum.

En arbetslöshetsperiod avslutas med en avaktualisering. Avaktualiseringsorsaker är t ex att den arbetssökande fått en tillsvidareanställning (avaktualiseringsorsak 1) eller en tidsbegränsad anställning (avaktualiseringsorsak 2). Den arbetssökande kan också lämna Arbetsförmedlingen för exempelvis föräldraledighet. Då används avaktualiseringsorsak 5, annan känd orsak. Avaktualiseringsorsak 6 används när den arbetssökande lämnat av okänd orsak.

Det finns en del fel i Arbetsförmedlingens register.[2] Dels kan uppgifterna från den arbetssökande vara felaktiga, dels kan fel värden registreras av arbetsförmedlaren. De fel som bedöms ha störst betydelse för detta arbete är problemet med överlappande inskrivningsperioder. Avaktualiserade individer kan ha samma inskrivningsdatum i den avslutade inskrivningsperioden som i den närmast föregående perioden. Det förekommer också att inskrivningsdatumet ligger mellan inskrivningsdatum och avaktualiseringsdatum i närmast föregående period. I många fall borde de båda inskrivingsperioderna egentligen vara en och samma period. Vidare är det vanligt att personer som avaktualiseras från en inskrivningsperiod snart efter detta inleder en ny inskrivningsperiod.

I denna rapport angrips problemet med överlappande inskrivningsperioder genom att vänta tio arbetsdagar för att kunna identifiera vilka individer som påbörjar en ny inskrivningsperiod. På så sätt betraktas två närliggande inskrivningsperioder som en och samma inskrivningsperiod.

Motsvarande problem finns för sökandekategoriperioderna (arbetslös, praktik, arbetsmarknadsutbildning, tillfälligt arbete osv). Den arbetssökande är alltid registrerad i en sökandekategori med ett start- och slutdatum. Negativa perioder förekommer, då ligger periodens slutdatum tidigare än startdatum. Det förekommer också att startdatum är detsamma som slutdatum, dvs att åtgärden eller arbetslöshetsstatusen varar i noll dagar. Datasystemet fungerar så att arbetsförmedlarna matar in startdatum för alla sökandekategoriperioder. Slutdatum sätts sedan automatiskt som startdatumet för den efterföljande perioden. Tidigare inmatade sökandekategoriperioder kan inte tas bort eller ändras. Det kan tänkas finnas två olika skäl till uppkomsten av negativa sökandekategoriperioder. För det första kan rena inmatningsfel förekomma. För det andra kan "korrigeringar" av tidigare inmatade uppgifter förekomma. Genom att skapa en negativ period fås två sökandekategoriperi-

[2]IFAU har dokumenterat fel i "Dataproblem vid utvärderingen av arbetsmarknadspolitik". Stencilserie 2000:5.

oder (poster i databasen) som helt eller delvis avser samma tidsperiod, men där personen är registrerad i olika sökandekategorier. Den andra perioden kan då vara avsedd som en korrigering av den första. Det här ställer förstås till problem när arbetslöshetsstatus vid en viss tidpunkt ska fastställas. I rapporten beaktas bara den senaste sökandekategoriperioden vid en viss tidpunkt.

2.2 Urvalsundersökningen

Under 2005 och 2006 riktades intervjuundersökningar till ett urval personer som lämnat Arbetsförmedlingen av okänd orsak, "okänd orsak-undersökningen", för att närmare undersöka varför kontakten med Arbetsförmedlingen upphört. Intervjuerna är gjorda vid 12 tillfällen mellan september 2005 och augusti 2006. Totalt intervjuades knappt 2500 personer.

2.2.1 Målpopulation

Inskrivningsperioder, eller arbetslöshetsperioder, som avslutas med en avaktualisering av okänd orsak och där den arbetssökande inte återkommer till Arbetsförmedlingen inom 10 arbetsdagar utgör målpopulation i undersökningen. Personer som kort efter avaktualisering återkommer till Arbetsförmedlingen anser sig vara arbetssökande och vi behöver inte fråga vad de har för sysselsättningsstatus i en intervjuundersökning. Med andra ord antas att en individ som inom kort återkommer till Arbetsförmedlingen inte väsentligen har fått en förändrad arbetslöshetssituation. Den uteblivna kontakten med Arbetsförmedlingen berodde inte på att något inträffat som gjort att individen inte längre är arbetssökande. Vilka som inom kort återkommer till Arbetsförmedlingen kan följas i Arbetsförmedlingens register.[3]

[3]I urvalsundersökningen återkom 18 procent av de avaktualiserade med okänd orsak till Arbetsförmedlingen inom 10 arbetsdagar. Vid en avaktualisering skickas ett brev till den arbetssökande där det står att personen inte längre är registrerad som arbetssökande och att Arbetslöshetskassan är informerad om detta i de fall personen har fått ersättning

De som studeras i urvalsundersökningen är således den grupp av individer som avaktualiserats av okänd orsak och som inte kommit tillbaka till Arbetsförmedlingen inom 10 arbetsdagar. Vid inferens från undersökningen går det att ta hänsyn till de som kort efter avaktualisering återkommer till förmedlingen. En andel som gäller för samtliga avaktualiserade av okänd orsak kan tas fram trots att urvalsundersökningen bara riktas till de som inte kort efter avaktualisering återkommer till Arbetsförmedlingen. Notera att en inskrivningsperiod och inte en individ utgör element i undersökningen. Ibland förekommer en individ inte bara en utan flera gånger under t ex ett år. Populationen består av unika arbetslöshetsperioder, men inte av unika individer.

2.2.2 Urvalsram och urval

Undersökningen är gjord vid 12 olika mättillfällen under ett års tid. Tanken var att på det sättet fånga upp eventuell säsongsvariation. Tidpunkten för mättillfällena samt urvalsstorleken bestämdes dock i huvudsak av tillgängliga intervjuarresurser.

De arbetslöshetsperioder som under en given vecka avslutas med en avaktualisering av okänd orsak och där den arbetssökande inte inom 10 arbetsdagar återkommer till Arbetsförmedlingen utgör population och urvalsram för respektive mättillfälle. Ett mättillfälle inkluderar således en veckas avaktualiseringar av okänd orsak. Ett obundet slumpmässigt urval om 300 arbetslöshetsperioder gjordes varje mätvecka. Totalt ingår därmed 3600 arbetslöshetsperioder i urvalet.

Urvalsundersökningen kan ses som ett stratifierat urval med mätvecka som strata. Med en stratifiering av en ändlig population $U = \{1, ..., k, ..., N\}$ menas en partitionering av U i H delpopulationer, som kallas strata och som betecknas $U_{1,}..., U_h, ..., U_h$ (Särndal, Swensson och Wretman 1992). Antalet

från en arbetslöshetskassa. Personen har möjlighet att invända mot avregistreringen och kan hävda att denna är felaktig och fallet ska då utredas. Den arbetssökande har också möjlighet att på nytt anmäla sig som arbetssökande.

element i stratum h betecknas N_h och urvalsstorleken i stratum h betecknas n_h. Sannolikheten att ett givet element ingår i urvalet, inklusionssannolikheten, ges av

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

Tabell 1 visar hur urvalet är uppdelat på olika mätveckor under andra halvåret 2005 och första halvåret 2006. Antalet avaktualiserade av okänd orsak totalt samt den delmängd som inte återkommit inom 10 arbetsdagar och som utgör populationen redovisas också. Urvalsstorleken för samtliga veckor är 300. Designvikten är det antal element i populationen som varje element i urvalet representerar och ges av inversen av inklusionssannolikheten. En designvikt lika med 10 innebär att en person i urvalet representerar 10 personer i populationen.

Eftersom en person kan ha flera arbetslöshetsperioder kan en individ förekomma flera gånger i undersökningen. Utfallet av arbetslöshetsperioder som tillhör samma individ är troligen korrelerade med varandra. Observationerna kan inte sägas vara oberoende. Av de arbetslöshetsperioder som avslutats med en avaktualisering av okänd orsak under mätveckorna tillhör 2,6 procent av perioderna inte unika personer. Av de perioder som utgör urvalsram i undersökningen, perioder där individerna inte återkommit inom 10 arbetsdagar, tillhör 1,5 procent av perioderna inte unika individer. I urvalet förekommer tre individer mer än en gång. Problemet med att samma individer förekommer flera gånger i undersökningen bedöms som litet eftersom bara någon procent av arbetslöshetsperioderna inte tillhör unika individer.

2.2.3 Intervjuerna

Undersökningen genomfördes i form av datorstödda telefonintervjuer på Arbetsförmedlingens intervjuenhet. Intervjuerna är gjorda så nära inpå avregistreringen från Arbetsförmedlingen som möjligt, två till tre veckor, detta

Tabell 1: Populationsstorlekar och designvikt per mätvecka för urvalet

Urvalsvecka	Antal avaktua-liserade med okänd orsak	Antal som inte åter-kommit efter 10 arbets-dagar	Designvikt
Vecka 34 2005 (aug)	3622	2864	9.6
Vecka 35 2005 (aug-sep)	4540	3538	11.8
Vecka 36 2005 (sep)	4403	3496	11.7
Vecka 37 2005 (sep)	4547	3706	12.4
Vecka 40 2005 (okt)	3926	3240	10.8
Vecka 3 2006 (jan)	2590	2106	7.0
Vecka 5 2006 (jan-feb)	3464	2857	9.5
Vecka 9 2006 (feb-mar)	3078	2544	8.5
Vecka 14 2006 (apr)	2473	2138	7.1
Vecka 22 2006 (maj-jun)	2927	2507	8.4
Vecka 25 2006 (jun)	2411	2041	6.8
Vecka 31 2006 (jul-aug)	2011	1735	5.8

för att intervjupersonerna skulle komma ihåg hur deras arbetssituation såg ut när de upphörde att ha kontakt med Arbetsförmedlingen. Eftersom 10 arbetsdagar inväntas för att återflödet till Arbetsförmedlingen ska kunna exkluderas är det inte möjligt att ha intervjuerna närmare inpå avregistreringen än så.

En fråga ställdes till intervjupersonerna, hur deras arbetssituation var vid intervjutillfället. För frågeformulering och svarsalternativ se nedan.[4]

Hur är Din arbetssituation i dag?

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*

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.

2.3 Bortfallet i undersökningen

Av den totala urvalsstorleken på 3600 personer svarade 2443, vilket ger en svarsfrekvens på 68 procent. Svarsbortfallet i undersökningen är således 32 procent, en relativt hög procentandel.

Det vanligaste skälet till bortfall är att det inte gått att nå urvalspersonen per telefon. En del saknar telefonnummer i Arbetsförmedlingens register och ofta har det då inte heller gått att leta fram telefonnummer i andra register.

[4] I bilaga A redovisas svarsfördelningen.

Om bortfallet består av en särskild grupp personer som fått arbete i lägre eller högre grad än de som svarat i undersökningen blir resultaten missvisande. Nedan följer en analys av bortfallet.

2.3.1 Deskriptiv statistik

Tabell 2 visar individegenskaperna för bortfallet och för de som svarat. Män, äldre, utlandsfödda, lågutbildade, personer som saknar arbetslöshetskassa samt personer som bor i storstadslän är överrepresenterade i bortfallet. Det finns risk att dessa personer har fått arbete i lägre utsträckning än de som har svarat i undersökningen.[5]

2.3.2 Jämförelse av återflödet till arbetslöshet

En analys av i hur stor utsträckning personer i bortfallsgruppen respektive svarandegruppen återkommer till Arbetsförmedlingen kan visa om grupperna skiljer sig åt vad gäller återflödet till arbetslöshet. Tidpunkten för när en individ avaktualiseras av okänd orsak och tidpunkten för när individen eventuellt återkommer till Arbetsförmedlingen kan användas för att studera hur lång tid individerna inte är arbetssökande.

Med överlevnadsanalys studeras tider tills något inträffar (Collett 2003). Överlevnadstiden för en individ, t, kan ses som ett värde på en variabel, T, som kan anta vilket icke negativt värde som helst. De olika värdena på T har en sannolikhetsfördelning och T är den slumpmässiga variabeln förknippad med överlevnadstiden. Den slumpmässiga variabeln T har en sannolikhetsfördelning med den underliggande täthetsfunktionen $f(t)$. Fördelningssfunktionen för T ges av

$$F(t) = P(T < t) = \int_0^t f(u)\mathrm{d}u$$

[5] Se t ex Bennmarker, Carling och Forslund (2007) för vilka kategorier av arbetssökande som riskerar långtidsarbetslöshet.

Tabell 2: Individkaraktäristika för svarande respektive bortfall, i procent

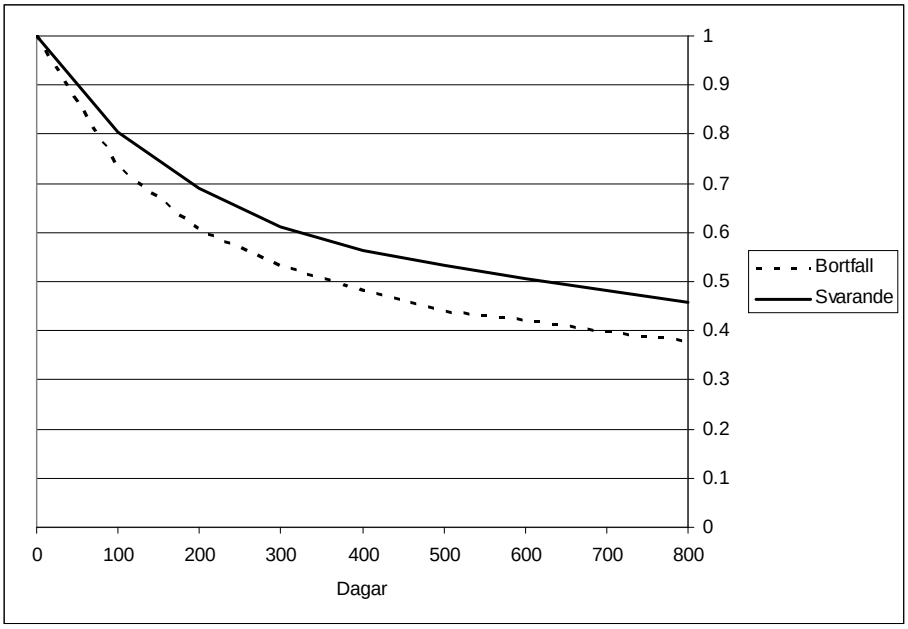
Variabel	Svarande	Bortfall
Kön		
Kvinna	48.3	41.8
Man	51.7	58.2
Ålder		
16-24 år	47.8	42.4
25-34 år	28.4	29.9
35-44 år	14.2	16.1
45-66 år	9.5	11.7
Födelseland		
Sverige	78.6	62.7
Utanför Sverige	21.4	37.3
Funktionshinder		
Nej	97.6	95.5
Ja	2.4	4.5
Utbildningsnivå		
Grundskola	21.5	35.2
Gymnasium	56.4	46.7
Högskola <= 2 år	6.5	6.2
Högskola > 2 år	15.6	11.8
Tillhör Arbetslöshetskassa		
Ja	58.9	45.4
Nej	41.1	54.6
Region		
Storstadslän (Stockholm, Västra Götaland och Skåne)	54.4	58.5
Skogslän (Värmland, Dalarna, Gävleborg, Västernorrland, Jämtland, Västerbotten och Norrbotten)	19.1	16.4
Övriga län	26.5	25.1

och representerar sannolikheten att överlevnadstiden är mindre än ett visst värde t. Överlevnadsfunktionen, $S(t)$, är definierad som sannolikheten att överlevnadstiden är större än eller lika med t:

$$S(t) = P(T \geq t) = 1 - F(t).$$

Överlevnadsfunktionen kan användas för att visa andelen arbetssökande som inte återkommit till Arbetsförmedlingen vid olika tidpunkter. I figur 1 visas överlevnadsfunktionen för de som svarat i undersökningen och för bortfallet.

Figur 1: Överlevnadsfunktioner för svarande respektive bortfall



Överlevnadsfunktionerna skiljer sig signifikant från varandra. De som svarat i undersökningen har vid varje tidpunkt större sannolikhet att inte ha återkommit till förmedlingen. Bortfallsgruppen återkommer snabbare till

Arbetsförmedlingen. Detta skulle kunna bero på att de svarande i undersökningen har fått ett mer varaktigt arbete jämfört med bortfallsgruppen som kanske bara fått korta, tillfälliga jobb. Det skulle också kunna bero på att bortfallsgruppen i mindre utsträckning fått jobb överhuvudtaget.

Slutsatsen från den här bortfallsanalysen är att inferens inte kan dras bara från de som svarat undersökningen. Det skulle ge snedvridna resultat. Bortfallet skiljer sig från de svarande med avseende på individkaraktäristika. Detta visar sig också ha betydelse för återflödet till arbetslöshet, vilket tyder på att bortfallet inte fått arbete i lika stor utsträckning som de som svarat. I kapitel 3 och 4 beskrivs hur bortfallet hanteras för att vi inte ska få snedvridna resultat.

3 Modell för att imputera bortfallet i undersökningen

Ett sätt att hantera bortfall i undersökningar är att imputera värden. De imputerade värdena måste tas fram med någon metod eller modell. Eftersom en mängd registeruppgifter kan kopplas till individerna i urvalsundersökningen kan en modell skattas som förklarar vilka kategorier av individer bland de svarande som har störst sannolikhet att få ett arbete. Denna modell kan sedan användas för att prediktera sannolikheten för arbete för de individer det inte finns svar från i undersökningen. Modellen kan således användas för att imputera utfallsvärden för bortfallet i undersökningen. För att det ska fungera på ett bra sätt måste modellen skattad på de svarande också gälla för bortfallet givet de förklarande variablerna. Modellen måste fånga in de faktorer som förklarar vilka som fått arbete även i bortfallsgruppen. Om modellen inte på ett bra sätt lyckas fånga in vilka som fått arbete i bortfallsgruppen kommer felaktiga värden imputeras i många fall. Hur väl imputeringen av bortfallet med hjälp av modellen fungerar diskuteras i senare avsnitt.

En modell som ska användas för imputering bör rent generellt innehålla

så många variabler som möjligt (Rubin 1996). Bernard och Meng (1999) menar dock att antalet variabler måste hållas under kontroll. För imputering av en viss variabel borde modellen innehålla alla de variabler som sedan används för att analysera datamaterialet, de variabler som är korrelerade med den imputerade variabeln samt variabler som förklarar vilka observationer det saknas data för (Schafer (1997) och van Buuren, Boshuizen och Knook (1999)).

Den imputerade variabeln är i det här fallet utfall arbete eller inte. De variabler som är korrelerade med den bör användas i modellen som ska imputera bortfallet i undersökningen. Förutom dessa variabler bör också variabler som förklarar vilka individer som svarat i undersökningen och vilka som inte gjort det vara med i modellen. Logistisk regression kan användas för att förklara vilka som fått arbete bland de svarande samt för att förklara vilka som svarat i undersökningen. I nästa avsnitt beskrivs logistisk regression teoretiskt och därefter följer en genomgång av variabler som kan vara användbara i modellen som ska imputera bortfallet i undersökningen.

3.1 Logistisk regression

De utfallsvariabler (beroende variabler) som studeras är om den arbetssökande efter avslutad arbetslöshetsperiod fått arbete eller inte samt om individen svarat i undersökningen eller inte. Utfallsvariablerna är binära. Vissa av de förklarande variablerna är klassindelade, andra är kontinuerliga. Logistiska regressionsmodeller passar bra för att modellera denna typ av data.

McCullagh och Nelder (1989) beskriver modeller för binära utfall. När sambandet mellan en utfallssannolikhet, π, och kovariatvektorn $(x_1, ..., x_p)$ ska studeras är det lämpligt att konstruera en formell modell som beskriver effekten på π när $(x_1, ..., x_p)$ förändras. Modellen bör vara konsistent med kända teoretiska samband mellan utfallsvariabeln och kovariatvektorn.

Sambandet mellan π och $(x_1, ..., x_p)$ sker genom den linjära kombinationen

$$\eta = \sum_{j=1}^{p} x_j \beta_j$$

för okända kofficienter $\beta_1, ..., \beta_p$. Utan restriktioner på β så $-\infty < \eta < \infty$. Detta är inte konsistent med att en sannolikhet för det binära utfallet ligger mellan 0 och 1. En transformation $g(\pi)$ kan användas för att transformera utfallet så att den kan anta alla värden på den reella tallinjen $(-\infty, \infty)$. Detta leder till en generaliserad linjär modell där den systematiska delen är

$$g(\pi) = \eta = \sum_{j=1}^{p} x_j \beta_j.$$

Den logistiska länkfunktionen definieras som

$$g(\pi) = \log \left\{ \pi/(1-\pi) \right\}.$$

Om en linjär logistisk modell med två kovariater, x_1 och x_2, används har vi modellen

$$\log(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

för log odds för ett positivt utfall. Detta ger att sannolikheten för ett positivt utfall är

$$\pi = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2)}.$$

Modellen tolkas som att effekten av en enhets förändring i x_2 ökar log oddsen för ett positivt utfall med en faktor β_2. Detta ger att effekten av en enhets förändring i x_2 ökar oddsen för ett positivt utfall multiplikativt med en faktor $\exp(\beta_2)$.

I modellen antas residualerna vara oberoende och likafördelade med konstant varians. De oberoende variablerna antas ge en linjär skattning.

Logistisk regression på data från en urvalsundersökning kan analyseras i proceduren *surveylogistic* i SAS. I den proceduren är det möjligt att ange designvikt, strata och populationsstorlek.

3.2 Empirisk strategi

Den logistiska regressionsmodellen som ska användas för att imputera bortfallet bör innehålla alla de variabler som förklarar vilka som fått arbete samt alla de variabler som förklarar vilka som svarat i undersökningen. I följande avsnitt (3.2.1) diskuteras vilka variabler som kan tänkas förklara vilka som fått arbete. I avsnittet därefter (3.2.2) diskuteras vilka variabler som kan tänkas förklara vilka som svarat i undersökningen. I avsnitt 3.3 presenteras sedan den modell som inkluderar alla relevanta variabler och som kan användas för att imputera bortfallet.

3.2.1 Variabler som förklarar vilka som fått arbete

Bring och Carling (2000) gjorde en urvalsundersökning liknande okänd orsak-undersökningen 1994. Undersökningen riktades till arbetssökande som avaktualiserats av okänd orsak och de arbetssökande fick frågan om de hade eller väntade på att tillträda ett arbete. Undersökningen från 1994 är mycket mindre än okänd orsak-undersökningen från 2005/2006. Urvalsstorleken var 200 och svarsfrekvensen 80 procent, vilket gav 156 observationer. Bring och Carling skattade en logistisk regressionsmodell för att förklara vilka som fått arbete på dessa 156 individer. Modellen innehåller ålder, kön, medborgarskap, arbetslivserfarenhet, utbildning och region. Att inte fler variabler inkluderades i modellen beror antagligen på att datamaterialet var litet. Fler variabler än så kan vara användbara för att förklara utfall av en arbetslöshetsperiod (se t ex Bennmarker, Carling och Forslund (2007), de Luna, Forslund och Liljeberg (2008) och Okeke (2005)).

De variabler som brukar användas testas som förklarande variabler i mo-

dellen som syftar till att imputera bortfallet i undersökningen (för lista över testade variabler samt procentandelar och medelvärden se bilaga B). Variablerna kan delas in i tre olika slag. För det första personliga karaktäristika, för det andra variabler som beskriver arbetslöshetssituationen och till sist variabler som beskriver individens arbetslöshetshistorik.

Till personliga karaktäristika räknas kön, ålder, medborgarskap, födelseland, förekomsten av funktionshinder, utbildningsnivå samt utbildningsinriktning. Vissa av dessa variabler kan ändra värde under en inskrivningsperiods gång som ålder och medborgarskap till exempel. Då gäller att värdena är beräknade vid tidpunkten för avaktualisering. Ålder har testats både som en kontinuerlig variabel och som en klassindelad variabel. Variabeln utbildningsnivå är klassificerad som grundskola, gymnasium, högskola mindre än eller lika med två år samt högskola mer än två år. Indelningen i utbildningsinriktningar grundar sig på svensk utbildningsnomenklatur, SUN, som är en standard för klassificering av utbildningar. På 1-siffernivå finns nio olika inriktningar vilket är det som används. I datamaterialet är enbart grundskoleutbildade klassificerade som att de har allmän utbildning i SUN. För gymnasieutbildade förekommer alla olika inriktningar förutom pedagogik/lärarutbildning. För högskoleutbildade förekommer samtliga inriktningar förutom allmän utbildning.

Variablerna erfarenhet av sökta yrken, önskad arbetstid, söker arbete längre bort än pendlingsavstånd, ersättning från arbetslöshetskassa, deltagande i aktivitetsgarantin (ett arbetsmarknadspolitiskt paraplyprogram för långtidsinskrivna[6]), sysselsättningsstatus enligt Arbetsförmedlingens register vid tidpunkten för avaktualisering av okänd orsak, vilket län individen bor i, kvoten mellan vakanser och arbetslöshet, andel avregistrerade av okänd orsak relativt samtliga avregistreringar samt månad för avregistreringen räknas till variabler som beskriver arbetslöshetssituationen. Vad gäller ersättning från arbetslöshetskassa har två olika definitioner testats. Dels huruvida individen

[6] Aktivitetsgarantin ersattes den 2 juli 2007 av Jobb- och utvecklingsgarantin.

tillhör en arbetslöshetskassa eller inte. Dels vilken nivå på arbetslöshetsersättningen som individen erhållit uppdelat på ingen ersättning, grundbelopp samt inkomstrelaterad ersättning. Sysselsättningsstatus enligt Arbetsförmedlingens register vid tidpunkten för avaktualiseringen har testats på två olika sätt. Dels indelningen arbete[7], arbetslös samt övriga. Dels indelningen deltid/timanställd, tillfälligt arbete, arbetslös samt övriga. Målpopulationen inkluderar samtliga individer oavsett deras arbetslöshetsstatus vid tidpunkten för avaktualisering. Det innebär att personer som redan har någon form av arbete (deltidarbete, timanställning eller tillfälligt arbete) vid tidpunkten för avaktualisering inkluderas. För dessa är det inte säkert att den uteblivna kontakten med Arbetsförmedligen beror på att de fått arbete i större utsträckning än de hade innan.

Länen har delats in i storstadslän (Stockholm, Västra Götaland och Skåne), skogslän (Värmland, Dalarna, Gävleborg, Västernorrland, Jämtland, Västerbotten och Norrbotten) samt övriga län. Kvoten mellan vakanser och arbetslöshet är testad både som ett månadsgenomsnitt och som ett kvartalsgenomsnitt. Med vakanser avses antalet lediga platser registrerade på Arbetsförmedlingen och med arbetslöshet avses arbetslösa och programdeltagare registrerade på Arbetsförmedlingen. Andelen avregistrerade av okänd orsak relativt samtliga avregistreringar är testad som ett månadsgenomsnitt, kvartalsgenomsnitt och som ett årsgenomsnitt. Den är även testad som ett årsgenomsnitt per Arbetsförmedlingskontor. Vad gäller månad för avregistrering finns inga observationer för månaderna november och december. Månad för avregistrering är även testad med en ihopslagning av månaderna maj/juni/juli, augusti/september samt övriga månader. Under sommaren får många arbetssökande sommarjobb och i augusti/september börjar många utbildning.

Vad gäller variabler som beskriver individens arbetslöshetshistorik avses inskrivningstid på Arbetsförmedlingen, tid i arbetslöshet, antal arbetslöshets-

[7] Sökandekategori deltidsarbetslösa, timanställda och tillfälligt arbete.

perioder, antal programperioder samt antal övergångar till arbete föregående fem år samt tid i arbetslöshet/program den senaste inskrivningsperioden. De historiska uppgifterna kan indirekt fånga upp personliga karaktäristika som vi inte har information om. Individernas motivation och ambition påverkar exempelvis sannolikt chanserna att få ett arbete och kan indirekt fångas upp i information om tidigare arbetslöshetsperioder. Inskrivningstid, arbetslöshetstid samt tid i arbetslöshet/program är kontinuerliga variabler mätta i antal dagar. Med inskrivningstid avses den totala tiden som en arbetssökande varit inskriven på Arbetsförmedligen. Med tid i arbetslöshet avses antal dagar individen varit registrerad i sökandekategori 11, arbetslös. Med tid i arbetslöshet/program avses antal dagar individen varit arbetslös eller deltagit i ett arbetsmarknadspolitiskt program.

Därutöver har interaktionstermer mellan ålder och arbetslöshetskassa, ålder och erfarenhet samt ålder och utbildningsnivå testats. Det kan vara så att benägenheten att ha fått arbete beroende på utbildningsnivå och om individen har arbetslöshetsersättning eller erfarenhet är olika för olika åldersgrupper.

3.2.2 Variabler som förklarar vilka som svarat i undersökningen

Vad gäller variabler som kan tänkas förklara vilka som svarat i undersökningen kan de flesta av variablerna som beskrivs i föregående avsnitt vara aktuella. Att personliga karaktäristika kan förklara svarsbenägenheten är troligt. Detsamma gäller för variabler som beskriver individens arbetslöshetshistorik. Även variabler som beskriver arbetslöshetssituationen kan tänkas förklara huruvida individen svarar eller inte, möjligen med undantag för variablen som beskriver arbetsmarknadsläget (kvoten mellan vakanser och arbetslöshet) och andelen avregistrerade av okänd orsak relativt samtliga avregistreringar. Dessa båda variabler beskriver inte individuella förutsättningar utan situationen på arbetsmarknaden och förhållanden på Arbetsförmedlingen.

3.3 Modellen och tolkningar

I modellen som ska användas för att imputera bortfallet i undersökningen testas alla de variabler som redogörs för i tidigare avsnitt. De variabler som inkluderats i modellen motiveras utifrån sinsemellan teoretiskt motiverade samband mellan variablerna, att de är viktiga förklaringsfaktorer för utflödet till arbete, eller utifrån vilka variabler som signifikant förklarar vilka som svarat i undersökningen.[8] Vid val av referenskategori har den kategori valts som flest individer i urvalsundersökningen tillhör. Tabell 3 visar resultatet.

[8]En separat logistisk regression som förklarar vilka som svarat i undersökningen har skattats. Resultatet redovisas i bilaga C.

Tabell 3: Modell för att imputera bortfallet i undersökningen. Det datamaterial som används är de svarande i urvalsundersökningen. Utfallsvariabel är huruvida individen fått arbete eller inte.

	Modell-skattning	Std-fel	P-värde	Odds-kvot
Intercept	-0.6555	0.2359	0.0055	
Man (ref)				
Kvinna	-0.3050	0.1047	0.0036	0.7370
16-24 år (ref)				
25-34 år	-0.0020	0.2828	0.9944	0.9980
35-44 år	-0.6193	0.4907	0.2069	0.5380
45-66 år	-1.3887	0.5485	0.0113	0.2490
Svenskfödda (ref)				
Utlandsfödd Norden	-0.7283	0.3353	0.0299	0.4830
Utlandsfödd övriga världen	-0.5342	0.1234	<.0001	0.5860
Inte funktionshinder (ref)				
Funktionshinder	-0.3777	0.3844	0.3258	0.6850
Grundskola	0.1886	0.1924	0.3269	1.2080
Gymnasium (ref)				
Högskola <= 2 år	0.0573	0.2012	0.7758	1.0590
Högskola > 2 år	0.4797	0.1604	0.0028	1.6160
Bred, generell eller okänd utbildningsinriktning (ref)				
Pedagogik och lärarutbildning	0.4663	0.3616	0.1971	1.5940
Humaniora och konst	0.4811	0.2302	0.0367	1.6180
Samhällsvetenskap, juridik, handel och administration	0.4054	0.2027	0.0455	1.5000
Naturvetenskap, matematik och data	0.1019	0.2500	0.6836	1.1070
Teknik och tillverkning	0.7407	0.2159	0.0006	2.0970
Lant- och skogsbruk samt djursjukvård	1.0244	0.4852	0.0347	2.7850
Hälso- och sjukvård samt social omsorg	0.5859	0.2302	0.0109	1.7970
Tjänster	0.7552	0.2459	0.0021	2.1280
Har inte erfarenhet av sökta yrken (ref)				
Har erfarenhet av sökta yrken	0.2247	0.1569	0.1521	1.2520
Söker inte bara arbete på heltid (ref)				
Söker bara heltidsarbete	0.0728	0.1017	0.4743	1.0760

forts. nästa sida

Tabell 3 – forts. föregående sida

	Modell-skattning	Std-fel	P-värde	Odds-kvot
Söker inte arbete längre bort än pendlings-avstånd (ref)				
Söker arbete längre bort än pendlings-avstånd	-0.1792	0.1533	0.2425	0.8360
Tillhör inte arbetslöshetskassa (ref)				
Tillhör arbetslöshetskassa	0.0588	0.1456	0.6864	1.0610
Deltar inte i aktivitetsgarantin (ref)				
Deltar i aktivitetsgarantin	-1.0433	0.4444	0.0189	0.3520
Status arbetslös innan avregistrering (ref)				
Status arbete innan avaktualisering	0.9303	0.1246	<.0001	2.5350
Status övriga innan avaktualisering	-0.4270	0.1304	0.0011	0.6520
Storstadslän (ref)				
Skogslän	-0.4812	0.1304	0.0002	0.6180
Övriga län	0.1143	0.1139	0.3153	1.1210
Antal arbetslöshetsperioder föregående 5 år	-0.1473	0.0292	<.0001	0.8630
Antal övergångar till arbete föregående 5 år	0.3151	0.0581	<.0001	1.3700
Alla månader utom maj, juni och juli (ref)				
Maj, juni eller juli	0.5616	0.1265	<.0001	1.7530
25-34 år och tillhör arbetslöshetskassa	0.8229	0.2532	0.0012	2.2770
35-44 år och tillhör arbetslöshetskassa	1.3397	0.3603	0.0002	3.8180
45-66 år och tillhör arbetslöshetskassa	0.6217	0.4191	0.1379	1.8620
25-34 år och har erfarenhet	-0.2092	0.2875	0.4668	0.8110
35-44 år och har erfarenhet	0.1336	0.4478	0.7654	1.1430
45-66 år och har erfarenhet	1.0555	0.5154	0.0406	2.8730

Modellen är signifikant med ett p-värde < 0.0001. En oddskvot för en variabel som är lika med 1 betyder att variablen inte har någon inverkan. En oddskvot mindre än 1 betyder att benägenheten för att ha fått ett arbete är mindre, medan en oddskvot större än 1 betyder att benägenheten är större. För kvinnor är exempelvis oddskvoten för att ha fått ett arbete 0.74 jämfört med männen. Det betyder att benägenheten att ha fått ett arbete är mindre för kvinnor än för män med liknande förutsättningar. Modellen frigör de olika faktorernas inverkan från varandra. Jämförelsen mellan till exempel kvinnor och män sker mellan kvinnor och män som liknar varandra i fråga om övriga faktorer i modellen.

Vidare har äldre en mindre benägenhet att ha fått arbete jämfört med referenskategorin 16-24 år. Koefficienterna för åldersgrupperna 25-34 år och 35-44 år är dock inte signifikant skilda ens på 20 procents nivå från åldersgruppen 16-24 år. Modellen är ändå specificerad med dessa åldersgrupper, vid studier av arbetslöshet brukar ålder förklara en del av variationen i utfall. Dessutom visar analysen av vad som förklarar vilka som svarar i undersökningen att samtliga åldersgrupper skiljer sig signifikant från referenskategorin 16-24 år i det avseendet (se bilaga C).

Utlandsfödda har en lägre oddskvot än svenskfödda, medan variablen funktionshindrad inte signifikant bidrar till förklaringsgraden. Utbildningsnivåerna mäts mot referenskategorin gymnasieutbildade. Samtliga utbildningsnivåer har en oddskvot större än 1 jämfört med gymnasieutbildning. En förklaring till det kan vara att gymnasieutbildade i större utsträckning lämnar Arbetsförmedlingen för att studera. Det är dock bara för högskoleutbildade mer än 2 år som koefficienten är signifikant skild från koefficienten för gymnasieutbildade.

De olika utbildningsinriktningarna jämförs med att ha en bred, generell eller okänd utbildningsinriktning. Samtliga specifika utbildningsinriktningar ger en förhöjd benägenhet för arbete. Huruvida individerna söker arbete bara på heltid eller inte samt om de söker arbete utanför pendlingsavstånd är intressant teoretiskt även om inte de variablerna är signifikanta i det här fallet. Att vara deltagare i aktivitetsgarantin ger en oddskvot på 0.35 för arbete. För de som hade någon form av arbete redan innan de avaktualiserades av okänd orsak är benägenheten att ha ett arbete 2.5 gånger så stor jämfört med de som stod registrerade som arbetslösa innan avregistrering. De som var registrerade i övriga sökandekategorier har en mindre benägenhet att ha fått arbete jämfört med arbetslösa.

De som bor i skogslän har en mindre benägenhet för arbete jämfört med de som bor i storstadslän. Ju fler arbetslöshetsperioder en arbetssökande har bakom sig desto mindre benägenhet för att ha fått ett arbete, medan

ju fler övergångar till arbete en individ har haft tidigare desto mer ökar benägenheten för arbete. Att avregistreras under sommarmånaderna maj, juni och juli är positivt, sommarjobben ökar benägenheten för arbete.

Ett antal interaktionstermer finns med i modellen. Oddskvoterna för äldre med arbetslöshetskassa är större än 1 vilket innebär effekten av att ha arbetslöshetskassa inte riktigt är lika stor för unga som för övriga åldersgrupper. Detsamma gäller för äldre med erfarenhet, oddskvoterna är större än 1, vilket innebär att den positiva effekten av att ha erfarenhet är större för äldre än för yngre.

Ett antal testade variabler är inte med i modellen. Variabeln medborgarskap är inte med, i analysen visar det sig att variablen utlandsfödd fungerar bättre. Variablerna vu-kvot och andelen avregisterade av okänd orsak visade sig inte vara signifikanta. Antagligen finns det för lite variation i de variablerna, den studerade tidsperioden på 1 år är relativt kort. Vad gäller ersättning från arbetslöshetskassa fungerade definitionen tillhör arbetslöshetskassa eller inte bättre än definitionen inte erhållit någon ersättning under arbetslöshetsperioden, erhållit grundbelopp eller erhållit inkomstrelaterad ersättning. Den senare kan tyckas mer informativ men problemet med den är att en tidpunkt måste bestämmas för när ersättning ska gått ut till den arbetssökande för att det ska definieras som att individen fått arbetslöshetskassa. Vad denna tidpunkt ska vara är inte givet.

Bland de testade variablerna som beskriver arbetslöshetshistorik är det många som inte är med i modellen. Det beror till stor del på att många av dem mäter ungefär samma sak. Antal arbetslöshetsperioder beskriver bäst hur belastad individen är av tidigare arbetslöshet. Total inskrivningstid, tid i arbetslöshet och antal programperioder under de senaste 5 åren samt tid i arbetslöshet/program i den senaste inskrivningsperioden fungerar sämre än antal arbetslöshetsperioder de senaste 5 åren.

3.4 Modellens prediktionsförmåga

Den logistiska regressionsmodellen som ska användas för att imputera bortfallet kan utvärderas genom att jämföra vad modellen predikterar med faktiskt utfall. Modellen ger för varje individ en sannolikhet att få ett arbete. Två fel kan uppstå, dels kan modellen prediktera att en arbetssökande gått till arbete fast personen inte gjort det, dels kan modellen prediktera att den arbetssökande inte fått ett arbete trots att så är fallet. Två begrepp används (Bennmarker, Carling och Forslund (2007)):

Sensitivitet = Sannolikheten att en individ prognostiseras som att ha gått till arbete givet att individen också har gått till arbete.

Specificitet = Sannolikheten att en individ prognostiseras som att inte ha gått till arbete givet att individen inte heller har gått till arbete.

Dessa begrepp kan användas för att beskriva hur väl modellen kan göra prediktioner. Sensitivitet och specificitet ska vara så hög som möjligt. Vad som kan anses vara tillräckligt högt går inte att säga rent allmänt utan varierar med vad prediktionerna ska användas till. En jämförelse i det här fallet är att slumpen skulle ge 50 procent korrekta prediktioner.

En jämförelse mellan modellens prediktioner och faktiskt utfall enligt svaren i undersökningen görs. Modellen predikterar ett värde mellan 0 och 1 och en gräns måste sättas för hur prediktionerna ska klassificeras, dvs vid vilken sannolikhet för arbete individen ska prognostiseras att ha gått till arbete.

Andelen som fått arbete bland de svarande i urvalsundersökningen är 51 procent. Arbete definieras här som att i undersökningen svarat att man idag har ett arbete på hel- eller deltid eller att man startat eget företag. Används ett gränsvärde som ger ungefär 51 procent till arbete i svarandegruppen enligt modellen uppgår värdet för sensitivitet till 71 och värdet för specificitet till 70. Detta innebär att 71 procent av de som prognostiseras ha gått till arbete

bland de svarande också har gjort det. 70 procent av de som prognostiseras inte ha gått till arbete har heller inte gjort det. Vid en jämförelse av vad slumpen skulle prognostisera ger den logistiska regressionsmodellen bättre prediktioner.

I nästa kapitel imputeras bortfallet med hjälp av modellen så att ett komplett datamaterial med skattade värden för bortfallet erhålls. Imputering beskrivs teoretiskt.

4 Bortfallsimputeringen

4.1 Imputering

Lundström och Särndal (2001) beskriver imputering. En metod är att imputera värden utifrån en logistisk regressionsmodell, vilket kommer att göras här. Det imputerade värdet för element k är då

$$\hat{y}_k = \mathbf{z}_k^{'}\hat{\boldsymbol{\beta}}$$

där $\mathbf{z}_k$ är värdet på imputeringsvektorn för element k, och

$$\hat{\boldsymbol{\beta}} = (\sum_r q_k \mathbf{z}_k \mathbf{z}_k^{'})^{-1} \sum_r q_k \mathbf{z}_k y_k.$$

$\hat{\boldsymbol{\beta}}$ är en vektor av regressionskoefficienter som erhållits från en multipel regression på data $(y_k, \mathbf{z}_k)$ tillgängliga för $k \in r$ och viktade med q_k. I specialfallet där

$$\mathbf{z}_k = (1, z_k)'$$

tar det imputerade värdet formen

$$\hat{y}_k = \hat{\alpha} + \hat{\beta} z_k$$

som motsvarar anpassningen av en enkel linjär regression med ett intercept.

Regressionsimputering är deterministisk eftersom den ger samma imputeringsvärde ifall den upprepas. Metoden kan dock göras stokastisk genom att addera en slumpmässigt utvald residual. Då blir det imputerade värdet för element k

$$\hat{y}_k = \mathbf{z}_k'\hat{\boldsymbol{\beta}} + e_k^*$$

där

$$\hat{\boldsymbol{\beta}} = (\sum_r q_k \mathbf{z}_k \mathbf{z}_k^{'})^{-1} \sum_r q_k \mathbf{z}_k y_k$$

som förut och e_k^* är en slumpmässigt utvald residual från datasetet med beräknade residualer $\{e_k : k \in r\}$, där

$$e_k = y_k - \mathbf{z}_k^{'}\hat{\boldsymbol{\beta}}.$$

Att addera en sådan residual har den effekten att den gör datamaterialet mer realistiskt. Ett komplett datamaterial som innehåller regressionsimputerade värden

$$\hat{y}_k = \mathbf{z}_k'\hat{\boldsymbol{\beta}}$$

tenderar att ha mindre varians än ett datamaterial med observerade värden y_k.

4.2 Multipel imputering

Vid multipel regression görs flera imputeringar för varje saknat värde. På så sätt skapas flera kompletta datamaterial. y_k-värdena för de svarande är samma i alla dataset medan de imputerade värdena är olika. En av fördelarna med multipel imputering är att variansestimationen blir mycket enkel eftersom det finns flera datamaterial.

Det behövs bara ett litet antal imputeringar per saknat värde (Rubin 1987). En estimators effektivitet baserad på m imputeringar är approximativt

$$(1 + \frac{\gamma}{m})^{-1}$$

där γ är andelen saknade värden.

Om γ är 0.3 och m 10 så är effektiviteten 97. Om m ökas till 20 är effektiviteten 99.

I proceduren *mi* i SAS kan multipla imputeringar göras. I det här arbetet har varje saknat värde imputerats 20 gånger. Det ger 20 olika datamaterial. För de observationer som inte saknar värden på utfallsvariabeln replikeras observationerna 20 gånger med samma värden. För varje observation som saknar värde på utfallsvariabeln imputeras 20 värden. Med dessa kompletta datamaterial med skattade, imputerade värden kan inferens dras från urvalsundersökningen.

4.3 Kombinera parameterskattningar vid multipel imputering

De 20 olika datamaterialen ger 20 olika parameterskattningar. Dessa måste kombineras för att ge en parameterskattning. Rubin (1987) beskriver hur olika parameterskattningar vid multipel imputering kan kombineras. Vid m imputeringar kan m olika punkt- och variansskattningar för en parameter Q beräknas. Anta att $\hat{Q}_i$ och $\hat{U}_i$ är punkt- och variansskattningarna från det $i$te imputerade datasetet, $i = 1, 2, ..., m$. Då är den kombinerade punktskattningen Q för multipel imputering medelvärdet av estimaten från de m kompletta datamaterialen:

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

Anta att $\bar{W}$ är inom-imputeringsvariansen, vilket är medelvärdet av estimaten från de m kompletta datamaterialen:

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

och B är mellan-imputeringsvariansen

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

då är variansestimatet associerat med $\bar{Q}$ den totala variansen

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

Proceduren *mianalyze* i SAS kombinerar parameterskattningar vid multipel imputering. En jämförelse av hur många som fått arbete i bortfallsgruppen jämfört med svarandegruppen i det imputerade materialet från urvalsundersökningen visar att det är en skillnad. 40 procent av bortfallsgruppen har gått till arbete, motsvarande andel för svarandegruppen är 51 procent. Att bortfallet har fått arbete i mindre utsträckning än de svarande är konsistent med resultaten från bortfallsanalysen.

Med det kompletta datamaterialet med skattade, imputerade värden kan inferens dras från urvalsundersökningen. Först beräknas andelen som fått arbete (kapitel 5) och sedan skattas en imputeringsmodell för arbetssökande som lämnat Arbetsförmedlingen av okänd orsak (kapitel 6).

5 Andelen som fått arbete

Proceduren *surveymeans* i SAS kan användas för att beräkna andelar och standardfel för stratifierade urval. Det datamaterial som används är de 20 imputerade materialen från urvalsundersökningen som beskrivs i föregående kapitel. 20 separata skattningar av andelen och standardfelet görs, dvs en för varje imputerat material. Dessa parameterskattningar kombineras sedan enligt metoden som beskrivs i avsnitt 4.3.

Resultatet är att 38.6 procent av avaktualiserade av okänd orsak har fått ett arbete. Eftersom den procentandelen baseras på resultat från en urvalsundersökning med imputerade värden är den exakta andelen något osäker. Konfidensintervallet för andelen är 37.0 till 40.2 procent. Det innebär att med 95 procents trolighet ligger procentandelen för populationen någonstans mellan 37 och 40.

Slutsatsen är att det vid statistikredovisningar av hur många som lämnat Arbetsförmedlingen för arbete kan antas att 39 procent av de som avaktualiserats av okänd orsak har gått till arbete. Den andelen gäller för samtliga som avbrutit kontakten med Arbetsförmedlingen av okänd orsak under en viss tidsperiod.

Bring och Carling (2000) finner att 45 procent har gått till arbete i undersökningen från 1994. En lägre procentandel har således fått arbete i den här nyare undersökningen jämfört med undersökningen från 1994.

Undersökningarna är dock inte helt jämförbara, frågorna skiljer sig något åt. I undersökningen från 1994 frågas om individen hade eller väntade på att tillträda ett arbete. En högre procentandel erhålls förstås om en framtida jobbstart räknas in jämfört med om bara dagens situation undersöks som i föreliggande rapport. Skillnaden i om det räknas in eller inte borde dock inte vara så stor, enligt en undersökning är den ca 1 procentenhet.[9] Ytterligare en sak som skiljer undersökningarna åt är att en bortfallskorrigering inte är gjord på undersökningen från 1994. Bortfallet i den undersökningen är 20 procent. Det kan vara så att en något lägre andel har fått arbete i bortfallsgruppen vilket i så fall skulle innebära en något lägre andel i arbete totalt sett i studien från 1994.

Jämförelsen av undersökningarna tyder dock ändå på att andelen som fått arbete bland avaktualiserade av okänd orsak faktiskt har ändrats över tid, skillnaden är så pass stor. Detta kan förklaras av att det 2006 infördes

[9] Arbetsförmdlingens lämnat-undersökning för 2009. Resultatet avser bara de som lämnat förmedligen av okänd orsak.

ett uttalat mål i Arbetsförmedlingens verksamhet att en viss procentandel av de arbetssökande skulle ha fått arbete efter 90 dagars arbetslöshet. Redan hösten 2005 förbereddes organisationen för införandet av detta nya mål. Det ledde antagligen till att arbetsförmedlarna i större utsträckning än tidigare undersökte vart de av okänd orsak som troligtvis hade fått ett jobb hade tagit vägen för att kunna registrera arbete i de fall personerna hade fått arbete. Förändringen kan avläsas i andelen som avregistreras av okänd orsak relativt det totala antalet avregistreringar. År 2004 och 2005 var den procentandelen 26 medan den år 2006 sjönk till 21.

6 Imputeringsmodell för arbetssökande som lämnar Arbetsförmedlingen av okänd orsak

6.1 Modellen och tolkningar

Med det kompletta, imputerade materialet från undersökningen kan en imputeringsmodell för arbetssökande som lämnar Arbetsförmedlingen av okänd orsak skattas. Beroende variabel i modellen är huruvida individerna i undersökningen fått arbete eller inte. Ett tjugotal förklarande variabler från Arbetsförmedlingens register används. De variabler som signifikant förklarar vilka som fått arbete eller är teoretiskt motiverade är med i modellen.

Separata modeller skattas för varje imputerat material, dvs 20 separata modeller skattas. De olika parameterskattningarna kombineras sedan till en. Tabell 4 visar imputeringsmodellen.

Tabell 4: Imputeringsmodell för arbetssökande som lämnar Arbetsförmedlingen av okänd orsak. Det datamaterial som används är urvalsundersökningen med skattade, imputerade värden för bortfallet. Utfallsvariabel är huruvida individen fått arbete eller inte.

	Modell-skattning	Std-fel	P-värde	Odds-kvot
Intercept	-0.6496	0.2260	0.0043	
Man (ref)				
Kvinna	-0.3030	0.1033	0.0037	0.7386
16-24 år (ref)				
25-34 år	0.0104	0.2944	0.9720	1.0104
35-44 år	-0.7479	0.5074	0.1441	0.4734
45-66 år	-1.4737	0.6478	0.0253	0.2291
Svenskfödda (ref)				
Utlandsfödd Norden	-0.7165	0.3993	0.0765	0.4885
Utlandsfödd övriga världen	-0.5181	0.1308	0.0001	0.5956
Inte funktionshinder (ref)				
Funktionshinder	-0.3851	0.3372	0.2551	0.6804
Grundskola	0.1656	0.1959	0.3992	1.1800
Gymnasium (ref)				
Högskola <= 2 år	0.0646	0.2118	0.7609	1.0667
Högskola > 2 år	0.4757	0.1580	0.0028	1.6091
Bred, generell eller okänd utbildningsinriktning (ref)				
Pedagogik och lärarutbildning	0.4643	0.3533	0.1895	1.5908
Humaniora och konst	0.4601	0.2342	0.0508	1.5842
Samhällsvetenskap, juridik, handel och administration	0.4117	0.1996	0.0400	1.5093
Naturvetenskap, matematik och data	0.1008	0.2471	0.6834	1.1061
Teknik och tillverkning	0.7188	0.2128	0.0008	2.0520
Lant- och skogsbruk samt djursjukvård	1.0088	0.5044	0.0463	2.7422
Hälso- och sjukvård samt social omsorg	0.6185	0.2261	0.0066	1.8562
Tjänster	0.7183	0.2654	0.0078	2.0510
Har inte erfarenhet av sökta yrken (ref)				
Har erfarenhet av sökta yrken	0.2127	0.1478	0.1509	1.2370
Söker inte bara arbete på heltid (ref)				

forts. nästa sida

Tabell 4 – forts. föregående sida

	Modell-skattning	Std-fel	P-värde	Odds-kvot
Söker bara heltidsarbete	0.0540	0.0951	0.5709	1.0555
Söker inte arbete längre bort än pendlings-avstånd (ref)				
Söker arbete längre bort än pendlings-avstånd	-0.1525	0.1428	0.2865	0.8586
Tillhör inte arbetslöshetskassa (ref)				
Tillhör arbetslöshetskassa	0.0835	0.1457	0.5671	1.0871
Deltar inte i aktivitetsgarantin (ref)				
Deltar i aktivitetsgarantin	-1.0543	0.3938	0.0083	0.3484
Status arbetslös innan avregistrering (ref)				
Status arbete innan avaktualisering	0.9520	0.1278	$<.0001$	2.5910
Status övriga innan avaktualisering	-0.4257	0.1297	0.0013	0.6533
Storstadslän (ref)				
Skogslän	-0.4749	0.1214	0.0001	0.6219
Övriga län	0.1108	0.1089	0.3100	1.1171
Antal arbetslöshetsperioder föregående 5 år	-0.1449	0.0287	$<.0001$	0.8651
Antal övergångar till arbete föregående 5 år	0.3064	0.0605	$<.0001$	1.3585
Alla månader utom maj, juni och juli (ref)				
Maj, juni eller juli	0.5525	0.1419	0.0002	1.7376
25-34 år och tillhör arbetslöshetskassa	0.8054	0.2637	0.0028	2.2376
35-44 år och tillhör arbetslöshetskassa	1.2921	0.3542	0.0004	3.6403
45-66 år och tillhör arbetslöshetskassa	0.7275	0.4421	0.1039	2.0698
25-34 år och har erfarenhet	-0.2202	0.2886	0.4464	0.8024
35-44 år och har erfarenhet	0.2965	0.4671	0.5269	1.3452
45-66 år och har erfarenhet	1.0333	0.6008	0.0880	2.8104

Benägenheten att ha fått ett arbete är mindre för kvinnor än för män med liknade förutsättningar. Detsamma gäller för äldre jämfört med 16-24-åringar samt utlandsfödda jämfört med svenskfödda. Högskoleutbildning mer än 2 år ger en förhöjd benägenhet för arbete jämfört med gymnasieutbildning. En specifik utbildningsinriktning jämfört med en bred, generell eller okänd inriktning ger en ökad benägenhet för arbete. Variablerna yrkeserfarenhet, önskad arbetstid samt om arbete söks utanför pendlingsavstånd är inte signfikanta. Att vara deltagare i aktivitetsgarantin ger en sådan låg oddskvot för arbete som 0.35. För de som hade någon form av arbete redan innan de

avaktualiserades av okänd orsak är benägenheten att ha ett arbete 2.6 gånger så stor jämfört med de som stod registrerade som arbetslösa.

De som bor i skogslän har en mindre benägenhet för arbete jämfört med de som bor i storstadslän. Ju fler arbetslöshetsperioder en arbetssökande har bakom sig desto mindre benägenhet för att ha fått ett arbete, medan ju fler övergångar till arbete en individ har haft tidigare desto mer ökar benägenheten för arbete. Att avregistreras under sommarmånaderna maj, juni och juli är positivt. Oddskvoterna för interaktionstermerna ålder och arbetslöshetskassa är större än 1 för äldre vilket kan tolkas som att effekten av att ha arbetslöshetskassa inte riktigt är lika stor för unga som för övriga åldersgrupper.

Jämfört med den modell som är skattad bara på de som svarat i undersökningen (se kapitel 3) blir det störst skillnad på parameterskattningarna för åldersgruppen 35-44 år, utbildningsinriktningarna Hälso- och sjukvård samt Tjänster, status arbete innan avregistrering samt interaktionstermerna där åldersgruppen 35-44 år ingår.

Aktivitetsgarantin ersattes under 2007 av Jobb- och utvecklingsgarantin. Vid användande av imputeringsmodellen för senare år måste därför deltagande i Jobb- och utvecklingsgarantin istället för deltagande i aktivitetsgarantin användas som förklarande variabel. Det skulle förstås kunna vara så att parameterestimatet för deltagande i aktivitetsgarantin inte rättvisande beskriver sambandet mellan deltagande i Jobb- och utvecklingsgarantin och utflödet till arbete. Mycket talar dock för att de båda programmen är likvärdiga. Båda vänder sig till personer som varit inskrivna på Arbetsförmedlingen länge och volymerna i de båda programmen har varit ungefär lika stora.

6.2 Modellens prediktionsförmåga

Imputeringsmodellen kan utvärderas mot en delpopulation utanför det urval modellen är skattad på. Sedan 2006 genomför Arbetsförmedlingen två gånger per år en undersökning riktad mot arbetslösa som lämnat Arbetsförmedling-

en, ”lämnat-undersökningen”. Urvalet består av 3000 sökande som lämnat förmedlingen under en tvåveckorsperiod och som inte återkommit under den närmaste tiodagarsperioden. Urvalet är ett obundet slumpmässigt urval och svarsfrekvensen har i runda tal uppgått till knappt 75 procent. Syftet med undersökningen är att ta reda på vad dessa personer gör, om de fått ett arbete och i så fall om Arbetsförmedlingen bidragit till detta och om de är nöjda med den hjälp de fått från förmedlingen.

Bland urvalspersonerna finns individer som lämnat Arbetsförmedlingen av okänd orsak. Dessa observationer skulle kunna användas för att testa modellens prediktionsförmåga utanför det material den är skattad på. Totalt finns det 1684 observationer från lämnat-undersökningen där individerna lämnat Arbetsförmedlingen av okänd orsak och där det finns svar på frågan om vad personerna gör idag. I lämnat-undersökningen ingår bara arbetslösa och inte de som hade någon form av arbete eller någon övrig arbetslöshetsstatus innan avaktualisering. För den delpopulationen i okänd orsak-undersökningen, nämligen arbetslösa som lämnat ett svar, har 43 procent gått till arbete.

För ett gränsvärde som ger ungefär den andelen till arbete för individerna i lämnat-undersökningen uppgår värdet på sensitivitet till 62 och värdet på specificitet till 61. Det innebär att 62 procent av de som prognostiseras ha gått till arbete bland de svarande i lämnat-undersökningen också har gjort det. 61 procent av de som prognostiseras att inte ha gått till arbete har heller inte gjort det. Alltför stora slutsatser bör inte dras från denna utvärdering, lämnat-undersökningen avser bara en delpopulation och bortfallet är stort. Lämnat-undersökningen kan dock användas för att jämföra prognosförmågan över tid. Det skulle kunna vara så att modellen fungerar sämre på senare data än på data kring år 2005 och 2006 som modellen skattats på. Tabell 5 visar modellens prognosförmåga på data från lämnat-undersökningen uppdelat på år.

Tabellen visar för det första att andelen som fått arbete i lämnat-undersökningen

Tabell 5: Prognosförmåga uppdelat på år

	2006	2007	2008	2009
Antal i lämnat-undersökningen som lämnat Arbetsförmedlingen av okänd orsak	704	711	834	724
Andel som svarat	56.4	54.6	57.6	56.9
Andel i arbete	34.8	28.6	32.9	25.6
Sensitivitet	54.0	69.4	62.9	56.6
Specificitet	59.6	58.1	58.3	71.8

varierar mellan 26 och 35 procent. Bortfallet är stort, bara drygt hälften har svarat. Sensitivitet och specificitet varierar, men det är inte så att prognosförmågan är sämre för 2007 till 2009 jämfört med för 2006, snarare tvärtom.

Imputeringsmodellen kan också utvärderas genom att skatta imputeringsmodellen på ett urval av observationer från det kompletta undersökningsmaterialet med skattade, imputerade värden. Vad den modellen predikterar för den del av materialet som inte är med i skattningarna jämförs med hur individerna svarade i undersökningen eller mot det imputerade värde de fått i de fall de inte svarat. Vid en skattning på 60 procent av materialet och en utvärdering mot resterande 40 procent uppgår värdet på sensitivitet till 72 och värdet på specificitet till 68. Om en jämförelse görs bara mot de som svarat i undersökningen uppgår sensitivitet och specificitet till ungefär detsamma.

Det är svårt att säga exakt hur bra prognosförmågan är för imputeringsmodellen utifrån dessa utvärderingar. Utvärderingen på en modell skattad på en del av materialet ger att sensitivitet och specificitet är relativt högt för den del av materialet som inte är med i skattningen av modellen. En jämförelse är att en slumpmässig tilldelning av värdena arbete, ej arbete skulle ge att 50 procent av de som prognostiseras ha gått till arbete också har gjort det. Sett till det fungerar modellen bättre än slumpen för att prognostisera ifall individerna har fått arbete eller inte. Jämförelsen av prognosförmågan över tid indikerar dessutom att imputeringsmodellen inte fungerar sämre för

senare år än för de år den är skattad på.

7 Avslutande diskussion

Resultaten i den här rapporten kan användas för att göra antaganden om hur många som fått arbete av de som av okänd orsak avbrutit kontakten med Arbetsförmedlingen. Vid enklare statistikredovisningar kan det antas att 39 procent av dem har gått till arbete. I mer avancerade analyser av utflöde till arbete kan imputeringsmodellen användas för att prediktera utfall för de individer det saknas information om. Imputeringsmodellen bör appliceras på ett datamaterial där återflödet till Arbetsförmedlingen inom 10 arbetsdagar har beaktats. Detta torde inte innebära några problem. Individer i Arbetsförmedligens register analyseras oftast en tid efter avregistrering. Vid imputering kan bara ett värde per saknat värde imputeras eller så kan flera värden imputeras. Rekommendationen är att göra en multipel imputering, 5 eller 10 imputeringar per saknat värde kan dock räcka. Vid multipel imputering kan beskrivningen av metoden i kapitel 4 fungera som vägledning.

Resultaten i den här rapporten är förenade med viss osäkerhet. Exempel på felkällor är urvalsfel, bortfallsfel, bearbetningsfel och mätfel. Vad gäller fel som uppstått i urvalsundersökningen har bortfallsfelet troligtvis störst betydelse. Vid imputering av bortfallet införs istället ett imputeringsfel som liknar mätfel i det att det riktiga värdet inte har mätts. Det är känt att ett imputerat värde är mer eller mindre fel. Hur väl bortfallsimputeringen lyckas är svårt att uttala sig om. Imputeringen ger dock att en lägre andel har fått arbete i bortfallsgruppen jämfört med svarandegruppen vilket är konsistent med resultaten från bortfallsanalysen.

Mätfel som uppstått i undersökningen genom att fråga och svar missuppfattas av intervjuaren och av den svarande antas vara liten. Intervjuarna har instruerats i hur fråga och svarsalternativ ska tolkas och hjälpte också respondenten med tolkningar i de fall det behövdes. I och med att telefo-

nintervjuerna är datorstödda sker dataregistreringen direkt under intervjun vilket minimerar risken för bearbetningsfel i samband med uppgiftsinsamlingen. Bearbetningsfel kan också uppstå i kodningen av variabler. Uppgifterna i Arbetsförmedlingens register kodas i många fall om enligt hur de förklarande variablerna är definierade. I det steget kan bearbetningsfel uppstå. Risken bedöms som liten, kodningen är kontrollerad.

Däremot finns det en hel del fel i Arbetsförmedlingens register. Om den bristande registerkvaliteten påverkar resultaten slumpmässigt ingår den i beräknade konfidensintervall. Om den bristande registerkvaliteten däremot orsakar systematiska fel är den besvärligare att uppskatta storleksmässigt. Det finns dock inget särskilt som skulle tala för att mätfelen i Arbetsförmedlingens register snedvrider resultaten på ett systematiskt sätt.

Rekommendationen är att imputeringsmodellen används vid analyser av utflöde till arbete där en delmängd av individerna har avaktualiserats av okänd orsak. Alternativet är att använda en slumpmodell och allt talar för att imputeringsmodellen ger bättre prediktioner. Det går dock inte att säga exakt hur väl imputeringsmodellen fungerar. Att nya undersökningar genomförs med jämna mellanrum är önskvärt för att kontrollera giltigheten i rapportens resultat samt för att eventuellt modifiera imputeringsmodellen och rekommendationerna om vilka antaganden om andelen till arbete som kan användas.

Referenser

[1] Barnard J & Meng X.L (1999), "Applications of Multiple Imputation in Medical Studies: From AIDS to NHANES", Statistical Methods in Medical Research, 8, 17 - 36.

[2] Bennmarker H, Carling K & Forslund A (2007), Vem blir långtidsarbetslös?, Rapport 2007:29, Institutet för arbetsmarknadspolitisk uvärdering (IFAU), Uppsala.

[3] Bennmarker H, Davidsson L, Forslund A, Hemström M, Johansson E, Larsson L, Martinson S & Persson K (2000), "Dataproblem vid utvärderingen av arbetsmarknadspolitik", Stencilserie 2000:5, Institutet för arbetsmarknadspolitisk uvärdering (IFAU), Uppsala.

[4] Bring J & Carling K (2000), "Attrition and Misclassification of Dropouts in the Analysis of Unemployment Duration", Journal of Official Statistics, vol. 16, No.4: 321-330.

[5] Collett D (2003), Modelling Survival Data in Medical Research, Chapman & Hall, London.

[6] de Luna X, Forslund A & Liljeberg L (2008), Effekter av yrkesinriktad arbetsmarknadsutbildning under perioden 2002-04, Rapport 2007:29, Institutet för arbetsmarknadspolitisk uvärdering (IFAU), Uppsala.

[7] Lundström S & Särndal C-E (2001), Estimation in the presence of Nonresponse and Frame Imperfections, Statistics Sweden, SCB-Tryck, Örebro.

[8] McCullagh P & Nelder J.A (1989), Generalized Linear Models, Capman & Hall, USA.

[9] Okeke S (2005), Arbetsmarknadsutbildningens effekter för individen, Ura 2005:6, Arbetsförmedlingen, Stockholm.

[10] Rubin D.B (1987), Multiple Imputation for Nonresponse in Surveys, John Wiley & Sons, New York.

[11] Rubin D.B (1996), "Multiple Imputation After 18+ Years (with discussion)". Journal of the American Statistical Association, 91, 473-489.

[12] Schafer J.L (1997), Analysis of Incomplete Multivariate Data, Chapman and Hall, New York.

[13] Särndal C-E, Swensson B & Wretman J (1992), Model Assisted Survey Sampling, Springer Verlag, New York.

[14] van Buuren S, Boshuizen H.C, and Knook D.L (1999), "Multiple Imputation of Missing Blood Pressure Covariates in Survival Analysis", Statistics in Medicine, 18, 681 - 694.

A Svarsfördelning på frågan om dagens arbetssituation i undersökningen

Tabell 6: Svarsfördelning på frågan "Hur är din arbetssituation idag?". Beräknad enbart på de svarande i undersökningen.

Svarsalternativ	Procentandel
Har arbete (heltid)	29.5
Har arbete (deltid)	19.3
Studerar / går i utbildning	14.2
Deltar i arbetsmarknadspolitiskt program	1.9
Har startat eget	1.9
Sjukskriven / sjukledig / föräldraledig	5.4
Arbetslös / söker jobb	24.8
Annat	3.1

B Beskrivande statistik för de förklarande variablerna

Tabell 7: Beskrivande statistik för de förklarande variablerna. Procentandelar för klassindelade variabler, medelvärden för kontinuerliga.

Variabel	Statistika
Kön	
Kvinna	46.2
Man	53.8
Ålder	
16-24 år	46.0
25-34 år	28.9
35-44 år	14.8
45-66 år	10.2
Medborgarskap	
Sverige	86.1
Norden	2.0
Övriga världen	11.9
Födelseland	
Sverige	73.5
Norden	1.8
Övriga världen	24.7
Funktionshinder	
Nej	96.9
Ja	3.1
Utbildningsnivå	
Grundskola	25.9
Gymnasium	53.3
Högskola <= 2 år	6.4
Högskola > 2 år	14.4
Utbildningsinriktning	
Pedagogik	2.3
Humaniora	9.8
Samhälls- och beteendevetenskap	16.7

forts. nästa sida

Tabell 7 – forts. föregående sida

Variabel	Statistika
Naturvetenskap	7.0
Teknik	13.4
Lantbruk	1.0
Medicin	9.3
Personliga tjänster	6.5
Bred, generell utbildning eller okänd	34.0
Erfarenhet	
Nej	19.6
Ja	80.4
Önskad arbetstid	
Heltid	32.1
Heltid eller deltid, Deltid	67.9
Söker jobb längre bort än pendlingsavstånd	
Nej	87.8
Ja	12.3
Tillhör a-kassa	
Nej	45.5
Ja	54.5
Deltar i aktivitetsgarantin	
Nej	97.9
Ja	2.1
Sysselsättningsstatus	
Arbete	27.3
Arbetslös	51.3
Övriga	21.4
Region	
Storstadslän	55.8
Skogslän	18.2
Övriga län	26.0
Inskrivningstid föregående 5 år (dagar)	*488.0*
Arbetslöshetstid föregående 5 år (dagar)	*209.8*
Tid i arbetslöshet/program den senaste inskrivningsperioden (dagar)	*65.1*
Antal arbetslöshetsperioder föregående 5 år	*2.8*
Antal programperioder föregående 5 år	*0.7*
Antal övergångar till arbete föregående 5 år	*0.7*
Kvoten mellan vakanser och arbetslöshet	*0.1*

forts. nästa sida

Tabell 7 – forts. föregående sida

Variabel	Statistika
Andel avregistrerade av okänd orsak relativt samtliga avregistreringar	*22.8*
Månad maj, juni eller juli	
Nej	81.1
Ja	18.9

C Modell som förklarar vilka som svarat i undersökningen

Tabell 8: Modell som förklarar vilka som svarat i undersökningen. Det datamaterial som används är urvalsundersökningen. Utfallsvariabel är huruvida individen svarat i undersökningen eller inte.

	Modell-skattning	Std-fel	P-värde	Odds-kvot
Intercept	1.3583	0.0988	<.0001	
16-24 år (ref)				
25-34 år	-0.3446	0.1410	0.0145	0.7090
35-44 år	-0.8736	0.1849	<.0001	0.4170
45-66 år	-0.6812	0.2245	0.0024	0.5060
Svenskfödda (ref)				
Utlandsfödd Norden	-0.5549	0.2896	0.0553	0.5740
Utlandsfödd övriga världen	-0.5924	0.0906	<.0001	0.5530
Inte funktionshinder (ref)				
Funktionshinder	-0.3346	0.2057	0.1038	0.7160
Grundskola	-0.3674	0.0941	<.0001	0.6930
Gymnasium (ref)				
Högskola <= 2 år	0.0104	0.1639	0.9493	1.0100
Högskola > 2 år	0.1405	0.1261	0.2651	1.1510
Söker inte arbete längre bort än pendlingsavstånd (ref)				
Söker arbete längre bort än pendlingsavstånd	-0.2275	0.1180	0.0539	0.7970
Tillhör inte arbetslöshetskassa (ref)				
Tillhör arbetslöshetskassa	0.0736	0.1233	0.5509	1.0760
Deltar inte i aktivitetsgarantin (ref)				
Deltar i aktivitetsgarantin	-0.6297	0.2614	0.0160	0.5330
Status arbetslös innan avregistrering (ref)				
Status arbete innan avaktualisering	0.2557	0.1072	0.0171	1.2910
Status övriga innan avaktualisering	-0.0172	0.0981	0.8611	0.9830
Antal arbetslöshetsperioder föregående 5 år	-0.1283	0.0179	<.0001	0.8800

forts. nästa sida

Tabell 8 – forts. föregående sida

	Modell-skattning	Std-fel	P-värde	Odds-kvot
Antal övergångar till arbete föregående 5 år	0.1792	0.0441	<.0001	1.1960
Alla månader utom maj, juni och juli (ref)				
Maj, juni eller juli	-0.3394	0.0953	0.0004	0.7120
25-34 år och tillhör arbetslöshetskassa	0.4294	0.1899	0.0237	1.5360
35-44 år och tillhör arbetslöshetskassa	1.0924	0.2416	<.0001	2.9810
45-66 år och tillhör arbetslöshetskassa	0.6120	0.2800	0.0288	1.8440

Arbetsförmedlingen