ARBETSFÖRMEDLINGEN
SWEDISH PUBLIC EMPLOYMENT SERVICE

Arbetsförmedlingens nya modell för bedömning av arbetssökandes stödbehov

Funktionssätt och prestation

Af 00009 6.0 (2022-10-26)

Arbetsförmedlingen
Författare: Petter Helgesson, Bo Bekkouche, Anders Böhlmark, Emilie Videnord, Agnes Widenfalk
Datum: 2023-09-05
Diarienummer: Af-2023/0068 0370
Arbetsförmedlingen Analys 2023:11

Innehåll

Sammanfattning

I samband med att Arbetsförmedlingen införde den nya matchningstjänsten Rusta och Matcha 2 i april 2023 så infördes även en ny modell i bedömningsstödet för anvisningar till tjänsten. Denna rapport innehåller en teknisk beskrivning av modellens utformning och dess prestation. Sammanfattningsvis:

- Precis som tidigare beräknar modellen en sannolikhet till arbete eller studier. Bedömningsstödet ger utifrån denna sannolikhet en rekommendation om deltagande i tjänsten.
- För att förbättra träffsäkerheten för breda grupper av arbetssökande ingår inskrivningstid som variabel i den nya modellen. Modellen är utformad för att hantera de utmaningar som detta medför.
- I övrigt används liknande information som tidigare: till exempel utbildning, sökta yrken och erfarenhet i dessa, födelseland, kön, ålder och kod för funktionshinder som medför nedsatt arbetsförmåga.
- Modellen har jämförts med andra modeller som är utformade för liknande ändamål. I dessa jämförelser står sig modellens prestation mycket väl.

Innehållet sammanfattas något mer ingående nedan, och beskrivs mer detaljerat i huvudtexten.

Ny modell i bedömningsstöd för anvisningar till Rusta och matcha

Anvisningar till Arbetsförmedlingens tjänst Rusta och matcha görs sedan införandet 2020 med hjälp av ett statistiskt bedömningsstöd: en prediktionsmodell som skattar arbetssökandes jobbchans kombinerad med en indelning av jobbchanser i olika nivåer. Den nivå en arbetssökandes bedömda jobbchans faller inom avgör vilken anvisning verktyget rekommenderar. De anvisningar som finns är ”för nära arbetsmarknaden”, tre olika spår inom tjänsten med olika ersättningsnivå till fristående leverantörer, samt ”för långt från arbetsmarknaden för Rusta och matcha”. I samband med införandet av Rusta och matcha 2 under våren 2023 började ett nytt bedömningsstöd användas. Det nya verktyget är delvis utformat för att bemöta kritik mot det tidigare verktyget, i synnerhet riktat mot hur de arbetssökandes inskrivningstid hanterades: prediktionsmodellen var utformad för nyinskrivna och inskrivningstid ingick därmed inte i de skattade jobbchanserna utan fick hanteras som en del i nivåindelningen. Denna hantering gjorde det i praktiken svårt att få med inskrivningstidens påverkan på ett riktigt bra sätt, och det gjorde det också svårare att bedöma verktygets prestation.

Rapporten beskriver ingående hur den nya prediktionsmodellen är uppbyggd, och motiverar diverse val som gjorts längs vägen. Den innehåller även en analys av hur modellen presterar.

Modellen inkluderar inskrivningstid och hanterar insatser

För att en prediktionsmodell ska kunna bedöma avstånd till arbetsmarknaden hos arbetssökande (deras stödbehov) så behöver modellen först lära sig vilka kombinationer av individegenskaper som är förknippade med olika jobbchanser. Denna procedur brukar ofta benämnas som modellträning, vilken genomförs på historiska data som innehåller en stor mängd individer med kända arbetsmarknadsutfall. En utmaning med att träna en modell med arbetssökande som också har längre inskrivningstider är att dessa ofta får arbetsmarknadspolitiska insatser: det blir då svårt att skilja på den arbetssökandes stödbehov och påverkan av insatserna. För att hantera detta har vi valt en variant på en så kallad överlevnadsmodell med tidsvarierande individegenskaper: informationen om varje arbetssökande kan alltså variera över tid och på så vis är det möjligt att ta hänsyn till insatser under inskrivningen. Därmed försöker vi svara på frågan ”vad är sannolikheten för varaktigt arbete eller studier om vi räknar bort påverkan av insatser?”.

Genom att hantera denna utmaning kan modellen tränas med arbetssökande med varierande inskrivningstid. Till skillnad från den gamla prediktionsmodellen finns alltså inskrivningstiden nu med som en variabel i prediktionsmodellen, och påverkar den bedömda jobbchansen på ett adekvat sätt. Därmed behöver inte inskrivningstiden hanteras i samband med spårindelningen som var fallet med det tidigare verktyget.

Bortsett från inskrivningstid använder den nya modellen i stort sett samma typ av data som den tidigare modellen: information om sökta yrken, utbildningsnivå- och inriktning, kön, ålder, födelseland, kommun, funktionshinderkoder, arbetsutbud och huruvida den arbetssökande söker i ett stort geografiskt område. I samband med övergången till det nya bedömningsstödet infördes nya dataflöden för att öka systemets robusthet. Denna övergång innebär att vissa data som förut var tillgängliga för modellen inte är det nu: detta innebär framför allt att information från arbetssökandes eventuella tidigare inskrivningsperioder inte ingår i nuläget.

När modellen används för anvisningar till Rusta och matcha skattar modellen sannolikheten för varaktigt arbete eller studier inom ett år, utan arbetsmarknadspolitiska insatser. Här räknas endast sådana arbeten eller studier som leder till avregistrering från Arbetsförmedlingen, och där ingen ny inskrivning sker inom fyra månader.

Modellen presterar väl

Modellens prestation har undersökts på ett antal olika sätt för att belysa hur väl modellen rangordnar sökande och i vilken mån de skattade sannolikheterna överensstämmer med utfallen, både för populationen som helhet och för ett antal grupper. Den gruppvisa analysen är gjord för att studera likabehandling och icke-diskriminering, vilket diskuteras i rapporten. Resultaten visar att modellen presterar väl jämfört med den tidigare modellen, och med andra liknande modeller från Sverige och andra länder.

Ett perspektiv är hur modellen presterar i sin helhet om den betraktas som en modell för så kallad binär klassificering: en modell som delar in arbetssökande i två grupper: ”nära” respektive ”långt ifrån” arbetsmarknaden. Modellen är egentligen inte utformad för binär klassificering utan genererar mer detaljerade resultat, men genom att dela in de arbetssökande utifrån om deras jobbchans är under eller över ett visst tröskelvärde fås en sådan klassificering. Modellens klassificering kan sedan jämföras med faktiska utfall. Ett lättbegripligt och ofta förekommande mått är så kallad *accuracy*: andelen ”korrekta” klassificeringar: 76 procent för denna modell. Detta mått är mycket känsligt för den så kallade skevheten hos klassificeringsproblemet och är bland annat därför svårt att jämföra mot andra modeller, men när vi korrigerar för skevheten för att göra mer rättvisande jämförelser står sig modellen mycket väl mot andra modeller.

Ett annat mått är så kallad *ROC AUC,* vilket mäter en klassificeringsmodells rangordningsförmåga. Måttet jämför alla par av de som fått jobb med de som inte fått, och beräknar i hur stor andel av paren som den som fått jobb bedömdes ha större jobbchans än den som inte fick jobb. För denna modell fås 79-81 procent (beroende på om vi korrigerar för modellens tidsvarierande individegenskaper eller inte). Här står sig modellen också mycket väl mot andra modeller, och också vid jämförelser med förekommande tumregel.

Vi presenterar också resultat för *concordance*: ett mått som påminner om ROC AUC men som är mer nyanserat och är mer anpassat för överlevnadsmodeller. Det visar att tiden till jobb är rätt rangordnad av modellen i 77 procent av fallen. Här saknar vi jämförelser mot andra modeller.

Modellen har också undersökts i termer av kalibrering, det vill säga i vilken mån de skattade sannolikheterna överensstämmer med utfallen. Detta har också gjorts för ett antal grupper baserade på de diskrimineringsgrunder som kan urskiljas i tillgängliga data. För en grupp arbetssökande med näraliggande jobbchanser ger en *välkalibrerad* modell en lika stor andel som får jobb som gruppens genomsnittliga, predicerade, jobbchans. Modellen är välkalibrerad, men det finns vissa avvikelser från perfekt kalibrering som framför allt grundar sig i en begränsning i data som berör arbetssökande med mycket långa inskrivningstider.

1 Inledning

1.1 Bakgrund och syfte

Arbetsförmedlingen använder sedan 2020 ett statistiskt bedömningsverktyg som stöd för handläggarna i bedömningen av vilka arbetssökande som ska anvisas till tjänsten Rusta och matcha. Verktyget består av två huvudsakliga delar: den första delen är en prediktionsmodell som bedömer framtida jobbchans. Detta kombineras med den andra delen: en nivåindelning som avgör vilka jobbchanser som leder till Rusta och matcha, och i så fall till vilket spår. Den arbetssökandes bedömda jobbchans faller inom en nivå som leder till en av följande rekommendationer: ”för nära arbetsmarknaden”, tre olika spår inom tjänsten med olika ersättningsnivå till fristående leverantörer, samt ”för långt från arbetsmarknaden för Rusta och matcha”.[1]

Olika granskningar (Arbetsförmedlingen 2022, Arbetsförmedlingen 2021b, IFAU 2021) pekade på att den första versionen av verktyget kunde förbättras på flera punkter, bland annat:

- Den första versionen av verktyget (i bruk fram till april 2023) var anpassad för att bedöma nyinskrivna sökande, något som inte var optimalt sett till myndighetens behov av att även bedöma sökande med längre pågående inskrivningstider. Modellen överskattade jobbchansen för de med längre inskrivningstider, något som kompenserades för i efterhand med en regelbaserad lösning i spårindelningen till Rusta och matcha.
- Den första versionen av verktyget var svår att utvärdera på grund av den särskilda regelbaserade kompensationen för inskrivningstid i spårindelningen.
- Bedömd jobbchans var inte informativ om faktisk jobbchans: modellens prediktioner var inte kalibrerade mot verkliga genomsnittliga arbetsmarknadsutfall.
- Den första versionen av verktyget hade ingen rutin för att löpande kunna kontrollera hur modellen presterade.

Med anledning av de förbättringsförslag som lyfts fram har myndigheten tagit fram en ny bedömningsmodell, vilken började användas i verksamheten i april 2023. I den nya modellen är de identifierade bristerna som beskrevs ovan åtgärdade. Syftet med denna rapport är att beskriva och utvärdera den nya bedömningsmodellen. Vi noterar att det befintliga förslaget till en ny EU-förordning, som ska reglera användning och utveckling av AI, innehåller flera punkter kring transparens, dokumentation, kvalitetssäkring och krav på att åtgärda identifierade brister (EUR-Lex - 52021PC0206 - EN - EUR-Lex (europa.eu)). Ambitionen med denna rapport är att

[1] Se Rusta och matcha 2 - Arbetsförmedlingen (arbetsformedlingen.se) för mer information om tjänsten.

vara transparenta kring modell, data och urvalshantering. Modellen som beskrivs i denna rapport har genomgått en utvärdering innan driftsättning, vilken redovisas i avsnitt 4.[2] Utvärderingen följer en kvalitetsäkringsrutin som är tänkt att användas regelbundet, vid förändringar av olika slag. I nästa avsnitt beskrivs mer i detalj vad vi vill uppnå med den nya modellen.

Modeller med liknande syften används på flera andra arbetsförmedlingar runt om i världen: en sammanställning finns i Desière med flera (2019). De kan rent tekniskt fungera på något skilda sätt även om mycket ofta är gemensamt, och är i de flesta fall i huvudsak inriktade på att bedöma risken för långtidsarbetslöshet bland nyinskrivna. Den typ av modell som vi har valt är en överlevnadsmodell som ligger nära den modell som beskrivs i Bennmarker med flera (2007), och ger möjligheten att ta hänsyn till insatser mellan startdatum och utfall. Detta gör den särskilt lämplig när den inte bara ska användas för nyinskrivna, vilket beskrivs mer i avsnitt 2.

1.2 Vad vill vi uppnå med den nya modellen?

1.2.1 Träffsäker bedömning av arbetssökandes avstånd till arbetsmarknaden som tar hänsyn till insatser hos träningspopulationen

Det är inte självklart vad som menas med att bedöma avståndet till arbetsmarknaden. Här håller vi oss till att skatta sannolikheten för att få ett positivt utfall (exempelvis ett arbete med en viss hållbarhet) under en viss tidsperiod. Detta sätt att bedöma framtida arbetsmarknadsutfall hos individen är intuitivt enkelt, men det finns också viktiga avvägningar att göra. En central fråga är huruvida den skattade sannolikheten ska inkludera påverkan av insatser. En sådan sannolikhet är mer rättfram att skatta, men svarar på en annan fråga än den vi i första hand önskar svar på. I första hand är vi intresserade av vilka *förutsättningar de arbetssökande har* opåverkade av insatser. Anledningen till detta är att vi vill undvika att sammanblanda de resurser som arbetssökande tilldelats historiskt med de förutsättningar den arbetssökande har själv. Därför försöker vi hantera detta med denna modell.

Givetvis ska modellen också ha så god prediktionsförmåga som möjligt: vara ”träffsäker”. Exakt definition av detta kan utgöra en stor fråga i sig, men en modell som har goda statistiska egenskaper och tar med relevant information (givet att den är tillgänglig) har goda förutsättningar.

1.2.2 Tolkningsbara sannolikheter

Det finns goda argument för att verktyget inte bara ska kunna rangordna de arbetssökande efter stödbehov, utan att de sannolikheter som skattas också ska gå att tolka och utgöra användbar information. Det senare innebär att sannolikhetsbedömningar är välkalibrerade mot de senare realiserade

[2] En god dokumentation av modell och utvärdering är också centrala delar av den transparens kring myndigheters användande av AI som betonas i utvecklandet av den så kallade *förtroendemodellen* som tillkom genom regeringsuppdraget ”Uppdrag om att testa ny teknik vid automatisering inom offentlig förvaltning” (se Sveriges Dataportal).

genomsnittliga utfallen. Med andra ord ska skattade sannolikheter (i vårt fall jobbchanser) betyder samma sak som de utgör sig för att betyda (i vårt fall samma *genomsnittliga* faktiska jobbchans).[3] Ett viktigt exempel där det är önskvärt med tolkningsbara sannolikheter är när man ska avgöra lämpliga gränser till rekommenderade spår i insats, där olika spår ska vara riktade till arbetssökande med olika stort stödbehov. Ett annat närliggande exempel är att kunna sätta rimliga ersättningsnivåer till leverantörer som ska erbjuda, och få ersättning för, olika mycket hjälp till arbetssökande med olika stort stödbehov.

1.3 Rapportens disposition

Utöver detta inledande avsnitt innehåller rapporten tre huvudsakliga avsnitt. I nästa avsnitt beskrivs och motiveras modellens funktionssätt i grunden: bland annat hur den fungerar rent matematiskt, varför just denna typ av modell är vald, vilka variabler som ingår och en beskrivning av ett par korrektioner till den grundläggande modellen. I avsnitt 3 beskrivs mer praktiska detaljer kring hur data bearbetats och hur modellen är implementerad. Slutligen undersöks modellens prestation i avsnitt 4.

2 Modellen

2.1 En flexibel överlevnadsmodell med tidsberoende kovariater

2.1.1 En central brist hos tidigare modeller

Den modell som varit i produktion fram till den 17 april 2023 och de modeller som hittills har använts i Arbetsförmedlingens betygsmodell (Arbetsförmedlingen 2022d) och Arbetsförmedlingen (2021c) svarar alla på samma grundläggande fråga: givet kovariaterna (individegenskaper, inskrivningsstatus och historik) hos en person vid en viss tidpunkt (”startdatumet” nedan), vad är sannolikheten att hen har nått ett positivt utfall *x* månader senare? Arbetsförmedlingen (2021a) betraktar en mycket snarlik fråga: givet kovariaterna hos en person vid startdatumet, *hur lång tid förväntas det ta tills hen har nått ett positivt utfall*? I båda dessa snarlika varianter tränas modellerna på historiska data över en mängd arbetssökande, och för var och en består data av

1. Kovariaterna vid startdatumet
2. Hur det har gått x månader senare (alternativt hur lång tid det tagit till positivt utfall).

[3] Exempel på helt välkalibrerad sannolikhetsbedömning: Om man följer upp 100 sökande efter 12+4 månader som fått sin jobbchans inom 12 månader (med varaktighetskrav minst 4 månader) bedömd till att vara 80%, då ska också 80/100 av dessa personer ha ett jobb.

Något olika modeller används i de olika fallen, men den grundläggande frågan är densamma.

Ett problem med denna formulering är att vi inte riktigt får svar på den fråga vi vill ha svar på, eftersom vi inte riktigt ställer den fråga vi vill ställa. Egentligen vill vi veta: givet kovariaterna hos en person vid en viss tidpunkt, vad *skulle vara* sannolikheten att hen har nått ett positivt utfall *x* månader senare *om personen inte fick någon insats*? Problemet är alltså att under de *x* månaderna mellan startdatum och utfallsdatum så kan den arbetssökande ha fått insatser hos Arbetsförmedlingen som påverkar sannolikheten till utfall i ena eller andra riktningen (på sikt är effekten förhoppningsvis positiv, men det förekommer också inlåsningseffekter: att det är mindre sannolikt att man går ut i arbete under tiden vissa insatser pågår). Detta blir ett betydligt större problem i och med att en statistisk modell används i tilldelningen av insatser i ökande utsträckning. Detta gör att systematiken i felet (avvikelsen mellan svaret på frågan vi vill ställa och frågan vi ställer) blir större.

I diskussioner kopplade till den tidigare modellen har problemet använts som ett argument för att endast träna modellen på nyinskrivna, eftersom arbetssökande med kortare inskrivningstid får insatser i betydligt mindre utsträckning. Samtidigt ger träning endast på nyinskrivna uppenbara problem med att använda modellen för ej nyinskrivna. Det är därför önskvärt med en annan lösning.

2.1.2 Intuitiv beskrivning av hur denna nya modell kan lösa detta

På något sätt vill vi alltså hantera hur insatser mellan startdatum och utfallsdatum i modellen. Problemet med den typ av frågeformulering som beskrivs i avsnitt 2.1.1 är att man inte utan vidare kan använda information mellan startdatum och utfallsdatum: kovariater och utfall sammanblandas i så fall på ett sätt som strider mot uppställningen av problemet och kan leda till felaktiga slutsatser, se till exempel diskussion om ”Bad controls” i Angrist och Pischke (2009).

Den typ av överlevnadsmodell som beskrivs i detta dokument löser detta genom att dela in tiden från startdatum till utfall i olika intervall, och sannolikheten för ett utfall under varje intervall skattas givet kovariaternas värde vid starten av intervallet. Det blir då inget problem att låta kovariaterna variera från starten av ett intervall till starten av nästa intervall, och de kan därför innehålla information om pågående eller avslutade insatser mellan startdatumet och starten av respektive intervall. Alltså: vid inskrivningen har en person vissa värden på kovariaterna, som motsvarar en viss sannolikhet för ett positivt utfall under den första veckan efter inskrivningen. När en vecka har gått kan möjligen vissa kovariater ha ändrats på ett sätt som gör att sannolikheten för ett positivt utfall under den andra veckan kan ha påverkats. Även om kovariaterna är desamma är inte sannolikheten densamma under första och andra veckan, vilket också modelleras. Sannolikheten för att ha nått ett positivt utfall

efter *x* månader beräknas genom att ”summera” sannolikheterna för motsvarande intervall.[4]

Det föregående stycket beskriver översiktligt hur insatserna hanteras i *träningen* av modellen. När modellen används för att bedöma avståndet till arbetsmarknaden vill vi bortse från påverkan av insatser och antar därför att personen inte får någon insats mellan startdatum och utfallsdatum (vi vet heller inte vad personen kommer ha för insatser så det hade i vilket fall varit ogörligt).[5]

2.2 Matematisk formulering av vald modell

Modellen är en typ av överlevnadsmodell (Rodríguez 2007), där överlevnad i detta fall är något negativt: att kvarstå i arbetslöshet. Överlevnadsmodeller betraktar överlevnadsfunktionen $S(t)$, sannolikheten att fortfarande ”leva” vid tidpunkten t. Ett annat centralt begrepp är hasardfunktionen $\lambda(t)$, som beskriver ”risken” per tidsenhet att ”inte leva” vid tidpunkten t, *givet att man överlevt så långt*. Annorlunda uttryckt: om $f(t)$ är täthetsfunktionen för tidpunkten för ”dödsfallet” blir $\lambda(t) = f(t)/S(t)$.

I den fortsatta framställningen använder vi benämningen ”kvarstå i arbetslöshet” som motsvarar den generella modellens begrepp ”att leva”, och benämningen ”övergå till jobb” (där andra positiva utfall som utbildning för enkelhetens skull får ingå) som motsvarar den generella modellens begrepp att ”dö”.

Eftersom[6] $f(t) = -\mathrm{d}S(t)/\mathrm{d}t$ fås ett elegant samband mellan hasardfunktionen och överlevnadsfunktionen:

$$S(t) = \exp\left(-\int_0^t \lambda(\tau)\mathrm{d}\tau\right).$$

När vi söker sannolikheten för att övergå till jobb inom tiden t söker vi alltså 1 - $S(t)$. Uttrycket ovan säger dock ingenting om hur skattningen av denna sker. Ett attraktivt val är en variant på en ”piece-wise exponential model” (Rodríguez 2007), som är en typ av ”proportional hazards model” där hasardfunktionen för individ i antas följa

$$\lambda_i(t) = \lambda(t|\mathbf{x}_i) = \lambda_{0g(i,t)}(t)\exp\left(\mathbf{x}'_{ij_0}\boldsymbol{\beta}_0\right),$$

där $\lambda_{0g(i,t)}(t)$ är den så kallade baseline-hasarden för grupp g[7], $\mathbf{x}_{ij_0}$ är en vektor[8] med kovariater (individegenskaper med mera) för individ i vid tiden t_j och $\boldsymbol{\beta}_0$ är en vektor med modellparametrar som skattas från data. Modellparametrarna är lika många som kovariaterna, och enkelt uttryckt kan de beskrivas som vikter som var och en

[4] ”Summera” står inom citattecken eftersom det inte är en helt vanlig addition av sannolikheterna, men heller inte helt långt ifrån. Detaljer följer i avsnitt 2.2.
[5] I vissa sammanhang kan det också vara intressant att till exempel låta en grupp av arbetssökande få samma insats i prediktionen.
[6] $S(t) = 1 - F(t)$ där $F(t)$ är den kumulativa sannolikhetsfunktionen.
[7] I vanliga fall är baseline-hasarden samma för alla individer och betecknas endast $\lambda_0(t)$ i Rodríguez (2007): vi möjliggör här för lite extra flexibilitet i modellen genom att ha olika baseline-hasarder för ett antal olika grupper. Detta hanteras praktiskt genom interaktioner mellan tidsdummies och viktiga grupper, se avsnitt 2.5.4.
[8] $\mathbf{x}'_{ij}$ är denna vektors transponat, alltså är $\mathbf{x}'_{ij}\boldsymbol{\beta}$ skalärprodukten mellan $\mathbf{x}_{ij}$ och $\boldsymbol{\beta}$.

säger vilken påverkan det har på jobbchansen att man har en viss utbildningsnivå eller viss ålder, och så vidare. En till egenskap som definierar en ”piece-wise exponential model” är att tiden är indelad i intervall där baseline-hasarden $\lambda_{0g(i,t)}(t)$ är konstant (därav ”piece-wise”): intervallens start betecknas med t_j. Vad denna konstant *är* i varje intervall är flexibelt och utgör ytterligare modellparametrar (utöver $\boldsymbol{\beta}_0$ i ekvationen ovan). Genom att lägga till dummies (kovariater som kan anta värdet 0/1) för tidsintervallen[9] till de vanliga kovariaterna ($\mathbf{x}'_{ij_0}$ kombinerat med dummies för tidsintervallen betecknas i fortsättningen $\mathbf{x}'_{ij}$) och lägga till motsvarande modellparametrar[10] till $\boldsymbol{\beta}_0$ (kombinationen betecknas $\boldsymbol{\beta}$) kan baseline-hasarderna bakas in bland övriga modellparametrar[11]. Hasarden för individ i inom intervall j blir då

$$\lambda_{ij} = \exp\left(\mathbf{x}'_{ij}\,\boldsymbol{\beta}\right) \qquad (1)$$

och sannolikheten att övergå till jobb inom intervallet j är $1 - \exp\left(-\lambda_{ij} t^0_{ij}\right) \approx \lambda_{ij} t^0_{ij}$ där t^0_{ij} är tidsintervallets bredd i intervall j (för individ i)[12]. Eftersom vi har historiska realisationer av Bernoulli-fördelade slumpvariabler med denna sannolikhet (d_{ij}, huruvida individ i har övergått till jobb i intervall j), så har vi i princip data att anpassa denna modell till. Det visar sig dessutom att (Rodríguez 2007) likelihoodfunktionen för denna modell är proportionerlig mot likelihoodfunktionen för en viss Poisson-regression med oberoende observationer[13]: den modell där d_{ij} är observationer av en Poissonfördelad variabel med väntevärde

$$\mu_{ij} = \lambda_{ij} t_{ij} = t_{ij} \exp\left(\mathbf{x}'_{ij}\boldsymbol{\beta}\right),$$

vilket är en Poissonmodell med log-länk och $\log(t_{ij})$ som ”offset”. Här har vi introducerat *exponeringstiden* t_{ij} (utan 0 som superskript): den tid som individ i kan observeras och kvarstår som arbetslös under intervall j: om ett utfall nås eller individen censureras under tidsintervallet så markerar det exponeringstidens slut. Annars är $t_{ij} = t^0_{ij}$ (om individ i kan observeras och kvarstår under hela intervall j) eller 0 (om utfall eller censurering under tidigare intervall).

Eftersom likelihoodfunktionerna är proportionerliga för den egentliga modellen och den ekvivalenta Poissonmodellen maximeras de för samma parametrar och därför kan den senare modellen användas för att ge maximum-likelihood-skattningar för den förra.

Den ekvivalenta Poissonmodellen är en generaliserad linjär modell med så kallad kanonisk länk, vilket innebär att $\sum_{ij} \mu_{ij}\,\mathbf{x}_{ij} = \sum_{ij} d_{ij}\,\mathbf{x}_{ij}$ för alla komponenter av $\mathbf{x}_{ij}$

[9] Här lägger vi också till interaktionerna mellan tidsdummies och vissa grupper.
[10] En parameter utanför exponentialfunktionen kan flyttas in genom att logaritmera den: $\lambda e = e^{\log \lambda}$.
[11] Detta – att baka in tidsdummiesarna bland övriga kovariater – görs något implicit i framställningen i Rodríguez (2007): här försöker vi göra det något explicitare för att det är så det i praktiken implementeras. Det leder dock till viss skillnad i notationen.
[12] Tidsintervallen är lika breda för alla individer, men den besläktade exponeringstiden som introduceras något längre ned kan variera mellan individer.
[13] I den *ekvivalenta* Poissonmodellen är alltså d_{ij} oberoende för olika tidsintervall för samma individ, vilket nästan kan beskrivas som en lycklig slump. I den ordinarie modellen kan inte d_{ij} vara oberoende eftersom det inte går att ”dö” två gånger. Oberoendet gör saker betydligt lättare.

(varje kovariat), det vill säga antalet som övergår till jobb är lika många i data som i prediktioner gjorda av modellen på samma data[14], och detta gäller även antalet kvinnor som övergår till jobb, antalet med gymnasieutbildning som övergår till jobb, och så vidare[15]. Denna garanti, som inte finns hos modeller i allmänhet, är en egenskap som ger goda förutsättningar för en välkalibrerad modell.

Eftersom tiden är indelad i intervall (och eftersom resultaten som beskrivs ovan inte involverar några antaganden om att kovariaterna ska vara konstanta, (Rodríguez 2007)) är det rättframt att låta kovariaterna variera mellan varje intervall. På samma sätt som ”vanligt” får inte eventuella utfall under intervallet sammanblandas med kovariaterna, så dessa bör sättas till värdet vid början av intervallet.

När alla modellparametrar $\boldsymbol{\beta}$ är skattade har vi också en skattning $\hat{\lambda}_{ij}$ av varje λ_{ij} genom att sätta in skattningen $\widehat{\boldsymbol{\beta}}$ i Ekvation 1. Då kan överlevnadsfunktionen för individ i skattas som

$$\hat{S}_i(t) = \exp\left(-\int_0^t \hat{\lambda}_i(\tau)\mathrm{d}\tau\right) = \exp\left(-\sum_j \hat{\lambda}_{ij} t_{ij}\right),$$

där j går över alla intervall där individ i kan observeras. Övergången från integral till summation är möjlig eftersom hasarden är konstant inom respektive intervall. Sannolikheten för ett utfall inom tiden t skattas med $1 - \hat{S}_i(t)$.

2.2.1 Varför just denna typ av modell?

Vi vill ha möjligheten att hantera insatser mellan startdatum och utfall. I någon mån bör en godtyckligt vald modell (exempelvis neurala nätverk eller random forests) kunna anpassas så att kovariaterna kan ändras i början av varje intervall, men för vår valda modell finns en stark teoretisk underbyggnad för hur detta påverkar skattningarna (Rodríguez 2007) där bland annat beroendet mellan olika tidsintervall hanteras (genom en lycklig ”slump”). Dessutom är modellen relativt lättimplementerad och har en del attraktiva egenskaper i och med kopplingen till en generaliserad linjär modell med kanonisk länk. Detta gör det *relativt* lätt att

1. Tolka hur väl modellen förklarar data
2. Jämföra olika varianter på modellen
3. Analysera modellen: vilka variabler är viktiga?
4. Rapportera vilka variabler som är viktiga för en enskild arbetssökande.

Resultaten i Salganik med flera (2020), som diskuteras i Arbetsförmedlingen (2021a), tyder också på att det exakta modellvalet har mycket begränsad betydelse i sammanhang som dessa: där man i någon mening försöker prognosticera hur en individs liv ska utvecklas. Denna fråga om det exakta modellvalet är en annan sak än

[14] Givet att modellen innehåller ett intercept.
[15] Givet att modellen innehåller dummyvariabler för dessa exempel.

den *förändring i problemformulering* som denna modell innebär, alltså att kovariaterna kan ändras mellan startdatum och utfall i modellträningen. Därför finns det ett stort värde i att byta problemformuleringen, och därmed modell, men sannolikt inte särskilt stort värde i att implementera andra typer av modeller för denna problemformulering. Slutligen kan nämnas att det är en modell av just detta slag som används av IFAU i Bennmarker med flera (2007). Denna studie är inte helt färsk, men uppvisar en träffsäkerhet som står sig mycket väl även idag, och har också utvecklingspotential: det går att förfina modellen på flera sätt, till exempel genom att öka flexibiliteten i beroendet av inskrivningstid och inkludera mer information om insatser.

2.2.2 Censurering

Det finns i princip två olika sätt som ”tiden kan upphöra” för en individ i modellen, antingen genom att individen övergår till jobb eller genom att vi inte längre kan observera vad som händer med individen. Det senare kan inträffa i några olika fall: antingen genom att tideräkningen tar slut (datainsamlingstillfället har passerat) eller genom att personen försvinner ur data av någon annan anledning. Om det går att anta att försvinnandet ur data inte säger någonting om hur det gått för personen så är det lätt att hantera genom att individen *censureras*: i denna modell görs då ingenting annat än att konstatera att exponeringstiden tar slut. Hur censurering praktiskt används i den implementerade modellen beskrivs i avsnitt 2.4.

2.2.3 Särskild åtskillnad mellan träning och prediktion (användning)

När modellen ”tränas” – anpassas till data – varierar kovariaterna över tiden mellan inskrivning och utfall på ett naturligt sätt: de antar helt enkelt de värden de råkar ha vid början av varje tidsintervall. När modellen används för att predicera sannolikheten för ett utfall för respektive individ känner vi i de flesta fall inte till kovariaternas värden framöver: vi vet inte på förhand vilken sökandekategori en arbetssökande kommer tillhöra om sju månader, till exempel. Därför sätts de flesta kovariater konstant lika med sina värden vid prediktionstillfället under hela prediktionshorisonten. Det finns dock undantag, till exempel ålder och tid på året – dessa kan räknas upp. Mer om vad som görs praktiskt i modellen i avsnitt 2.5.1.

2.2.4 Tidsintervall

Tidsintervallen bör vara valda så att det är någorlunda rimligt att modellera både baseline-hasarden och kovariaterna som konstanta i respektive intervall. Samtidigt bör det inte vara onödigt finfördelat för att undvika onödig förlust av frihetsgrader och onödigt tung modell. Det finns också praktiska skäl att använda sig av ”naturliga” intervall och därför håller vi oss till veckor eller månader.

För att ge en bild av om det är rimligt att använda veckor eller månader visar Figur 1 **Fel! Hittar inte referenskälla.**inskrivningslängder i veckor[16], indelat i intervall som är en eller fyra veckor långa. De första 8 veckorna är förändringarna ganska stora och skillnaden mellan veckovis och månadsvis indelning sannolikt inte endast

[16] Alla inskrivna enligt tabellen Insper i Arbetsförmedlingens datalager.

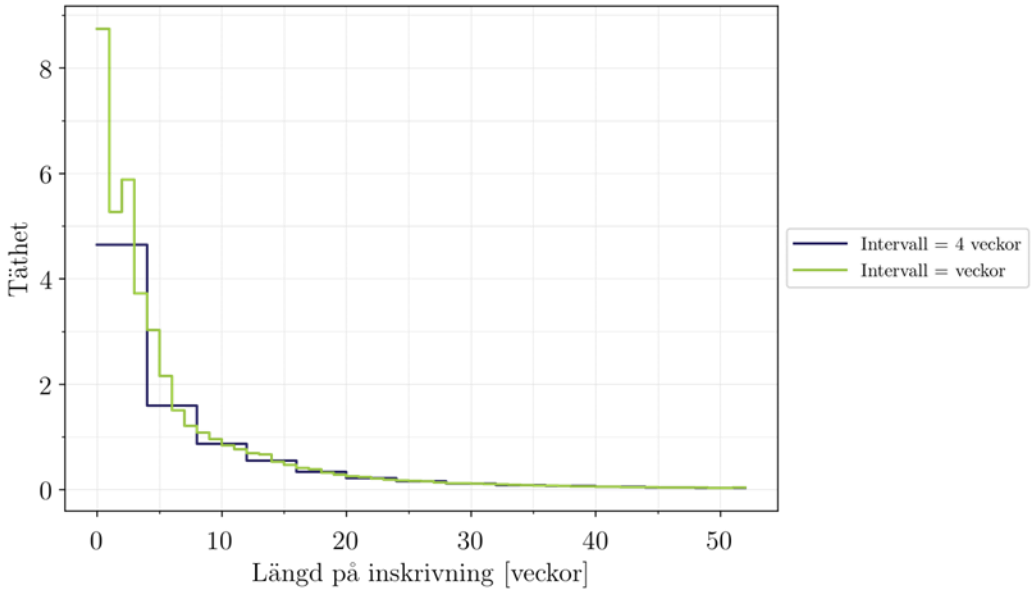
slumpmässig. Senare är förändringarna mindre och skillnaden mellan veckovis och månadsvis indelning mindre. Vi antar också att kovariaterna typiskt förändras enligt ett liknande mönster – mer i början och mindre senare.

Baserat på detta använder vi veckolånga intervall för de 13 första veckorna (tre månader), och sedan månader. För att kunna uttrycka tiden i såväl månader som år utan att förlora exakthet används ”månader” som antingen är 30 eller 31 dagar långa genom att ta gränsen för början på månad x som det antal dagar som ligger närmast $365{,}25\,x\,/\,12$.

Utifrån ovanstående fås intervallsgränser vid dag 0, 7, 14, …, 77, 84, 91, 122, 152, 182, 213, 243, 274, 304, 335, 365, 396, 426, 457, ….

Slutligen, för att det inte ska finnas tidsintervall som är alltför ovanliga, slås de intervall (på slutet) med färre än 1 000 förekomster i data ihop till ett intervall.

Figur 1. Längd på inskrivningar kortare än ett år som påbörjades under 2018.



2.3 Utfall

Syftet med modellen är i nuläget att producera en träffsäker bedömning av den arbetssökandes avstånd till arbetsmarknaden (utan påverkan av insatser). För att modellen ska svara på den fråga som vi vill besvara måste utfallet konstrueras så att det fångar modellens syfte. För att sannolikheten att nå ett utfall på ett någorlunda entydigt sätt ska beskriva avstånd till arbetsmarknaden väljer vi att inte ta med arbete med stöd bland utfallen: dessa ska oftast tilldelas personer som har längre avstånd till arbetsmarknaden och blir därför tvetydiga. Av liknande skäl (om än mindre tydliga) räknas inte heller andra anställningar där man kvarstår som registrerad hos Arbetsförmedlingen (till exempel sökandekategorierna för timanställningar och deltidsanställningar). Slutligen är vi i nuläget begränsade till information i

Arbetsförmedlingens register. Ett positivt utfall baseras därför på avregistreringar[17] från Arbetsförmedlingen till osubventionerad anställning eller studier: avregistreringsorsak 1, 2, 3 eller 7.

För att en avregistrering med relevanta orsaker ska räknas som ett positivt utfall måste avregistreringen hålla i sig i minst 122 dagar. Detta varaktighetskrav kan också motiveras med att ett mer varaktigt utfall är mer entydigt: att den arbetssökande verkligen har kommit ut på arbetsmarknaden. Att få ett jobb för att sedan komma tillbaka som inskriven hos Arbetsförmedlingen kort därefter är inte ett lika entydigt utfall. Att gränsen för varaktighet satts till just 122 dagar baseras på att den tidigare tjänsten rusta och matcha (KROM) definierade en övergång till arbete och studier som varaktig om övergången hållit i sig i minst 4 månader (122 dagar). I den nya tjänsten är resultatersättningen tvådelad, efter 3 respektive 6 månader, vilket inte ger en lika tydlig gräns för varaktighet. För enkelhetens skull behåller vi därför varaktighetskravet på 4 månader som ligger mellan de två nya gränserna för resultatersättning.

Genom att tillgängliga data om utfall i nuläget är begränsad till Arbetsförmedlingens register innebär det att det finns en ganska stor andel individer som vi inte vet det verkliga utfallet för, se avsnitt 2.4.2 (och i viss mån 2.4.3).

2.4 Censurering

2.4.1 Tiden tar slut

Om tiden då vi kan följa en individ tar slut vet vi endast att hen inte nått ett utfall fram till denna tidpunkt, och detta är precis vad som antas vid censurering: censurering kan därför användas utan *större* betänkligheter. Vi låter dock censureringen dröja så länge som möjligt, vilket gör att censureringen implicit innehåller viss information om när individen skrevs in. Strikt sett är detta inte helt oproblematiskt, men vi använder ändå detta tillvägagångssätt för att kunna följa deltagare med långa inskrivningstider, se avsnitt 3.4.

2.4.2 Hantering av avregistrering av okänd orsak (orsak 6) och orsak 8

Precis som för de positiva utfallen räknas bara avregistreringar som sedan håller i sig i minst 122 dagar.

Varaktiga avregistreringar av orsak 6 censureras vid tidpunkten för avregistreringen. Detta bygger på ett antagande om att censureringen är icke-informativ, vilket inte är riktigt sant eftersom sannolikheten för att man gått ut i arbete i närheten av denna tidpunkt är större än vid någon slumpvis vald tidpunkt[18]. Vi bedömer dock att detta är det bästa vi kan göra i nuläget, men förhoppningsvis kan inkomstuppgifter användas i träning av senare versioner av modellen.

[17] Att avregistreras från Arbetsförmedlingen innebär att man slutar stå som inskriven. Detta kan göras av olika orsaker, bland annat att man fått en anställning utan stöd eller påbörjat studier.
[18] Detta kan man se genom att använda SCB-data på inkomster som i Arbetsförmedlingen (2022b).

Avregistreringsorsak 8 (avliden) hanteras som orsak 6. För plötsliga dödsfall (typ trafikolycka) är detta ett helt korrekt förfarande; för dödsfall som följer på en periods sjukdom eller liknande kunde en hantering liknande den i avsnitt 2.4.3 övervägas, men dessa olika typer går inte att skilja åt i data och i det senare fallet är det troligt att avregistreringsorsaken ändå är en annan. Avregistreringsorsak 8 är lyckligtvis mycket ovanlig så den exakta hanteringen har inte någon större påverkan på modellens prediktioner.

2.4.3 Hantering av avregistrering av orsak 5 och orsak 4

Vid varaktig avregistrering av orsak 5 (*känd* orsak men ej arbete eller studier) censureras egentligen den arbetssökande i data, i den meningen att vi inte längre kan följa dessa individer efter en sådan avregistrering. I detta fall är det dock mindre rimligt att anta att censureringen är icke-informativ än vid orsak 6: orsaken *är* känd och inte något som ska räknas som ett positivt utfall: på kort sikt vet vi alltså att personen inte nått ett positivt utfall. Vad som händer på längre sikt vet vi mindre om. Därför låtsas vi som att vi har personen kvar i data, med oförändrade kovariater (förutom de som ändras i och med tidens gång), under 122 dagar. Efter detta censureras personen, det vill säga vi antar att vi inte vet något mer om utvecklingen, givet att personen inte påbörjat en ny inskrivning. Valet av 122 dagar är någorlunda godtyckligt.

Avregistreringsorsak 4 (fått anställning inom Samhall; mycket ovanlig) hanteras på samma sätt som orsak 5. Detta motiveras av att endast arbete utan stöd räknas som positiva utfall.

2.5 Kovariater (alias förklarande variabler eller features)

2.5.1 Hantering av kovariater i prediktion

I träning fungerar i princip alla kovariater likadant: de värden de har vid början av varje tidsintervall används. Vid prediktion däremot är det inte riktigt lika självklart: i allmänhet vet vi inte vid prediktionstillfället vilka värden kovariaterna kommer ha under tiden som prediktionen gäller.

En del kovariater kan med god noggrannhet antas faktiskt vara konstanta: födelseland och kön.

De flesta kovariater kan ändra värde men vi kan inte veta hur på förhand och vi antar att de är konstant lika med de värden de har vid prediktionstillfället: utbildningsnivå, sökta yrken, sökandekategori, med flera.

Vid prediktion antas ingen insats för samtliga. Är man i en betydelsefull insats är det sällan aktuellt att profileras för att eventuellt ges en ny insats, så detta antagande är oftast mer av akademisk karaktär[19], men när man mäter modellens prediktionsförmåga kan det däremot vara rimligt att ha med insatser som en

[19] Att *ha haft* insats är en annan sak och ingår i planerad vidareutveckling.

deltagare har vid prediktionstillfället. Detta gäller även om modellen används i vissa andra ändamål, till exempel för att göra bedömningar över förväntade flöden för myndighetens planering och styrning.

En del kovariater kan vi på ett trivialt sätt veta vad de ska anta för värden i framtiden, detta gäller ålder och dag på året (tänk kalendermånad för den senare). Dessa räknas då upp även vid prediktion.

2.5.2 Kovariatgrupper (för förklaringsmodell och beskrivning av modellen)

I Tabell 1 syns också en kolumn med ”grupp” för varje variabel. Denna gruppindelning används för att tolka vilka kovariater som är viktiga i modellen och presenteras för handläggarna efter genomförd profilering, se avsnitt 2.9.

2.5.3 Definitioner av grundläggande kovariater

Tabell 1 sammanfattar de grundläggande variabler som används som kovariater i modellen. Texten i detta avsnitt beskriver lite mer detaljer kring variablerna, även om vissa förkunskaper om variablerna i Arbetsförmedlingens datalager förutsätts för att förstå alla detaljer.

Dummyvariabler för vanliga värden hos nominalvariabler
Många av de variabler som vi har att tillgå beskriver olika kategorier. I den numeriska behandlingen översätts dessa till dummyvariabler – ett antal variabler som antar värdet ett eller noll. Många kan anta ett ganska stort (tio- eller hundratals) antal värden och det är ett tidsödande arbete att fundera över exakt vilka av dem som ska få en egen kolumn (de slutgiltiga variablerna bildar en kolumn i en matris, där raderna representerar en individ vid en tidpunkt). Det är inte oproblematiskt att skapa en kolumn för precis varje förekommande värde: en del kan förekomma i endast ett fåtal fall och detta kan med lite otur leda till linjärt beroende kolumner (alltså att en kolumn går att skriva som en viktad summa av andra kolumner)[20], vilket gör problemet olösligt. Även om inte linjära beroenden uppstår kan dummyvariabler för alltför ovanliga värden också bidra till problem med multikollinearitet[21]. Därför skapas endast egna kolumner för de värden som är ”vanliga”, och övriga värden buntas ihop till en övrigt-kategori (förutom referenskategorin, se nedan).

Gränsen för vad som räknas som tillräckligt vanligt är satt till 1000 förekomster vid den senaste tidpunkten för varje individ.

För dummyvariabler krävs det också i de flesta fall ett referensvärde, det vill säga ett värde som, även om det förekommer ofta, inte får någon kolumn. I stället innebär nollor *i alla andra* dummykolumner härrörande från denna variabel att variabelvärdet är lika med referensvärdet. Anledningen till att man måste göra på

[20] Alternativt så kan det leda till numeriskt linjärt beroende (inte exakt linjärt beroende, men i praktiken) kolumner i en matris som dyker upp i lösningsalgoritmen. Detta riskerar att hända i några fall ändå och hanteras i lösningsalgoritmen.
[21] Multikollinearitet är särskilt problematiskt om man är intresserad av parameterskattningar, vilket vi *inte* är i första hand. Vid prediktion (det vi i första hand är intresserade av) kan det leda till onödigt stora osäkerheter i vissa predicerade värden, men om kombinationer av variabelvärden som är ovanliga i träningsdata också är det i de data som prediktionerna görs på är dessa fall ovanliga.

detta sätt är återigen att linjärt beroende mellan kolumnerna måste undvikas. I vissa specialfall är de olika kolumnerna *inte* ömsesidigt uteslutande och då krävs inte något explicit referensvärde (se ”notera”-punkterna 2-4 nedan för exempel). Av i princip samma anledning som det är klokt att ha en gräns för hur sällsynta värden som ska få en egen kolumn så är det bra om referensvärdet är vanligt förekommande. Referensvärdena som valts nedan är valda med detta i åtanke och dessutom så att de ska vara kompatibla med varandra – en person med enbart referensvärden ska gå att föreställa sig.

Notera:

1. Sökandekategorierna (skat) som motsvarar betydande insatser (subventionerade anställningar och arbetsmarknadsutbildningar) ignoreras bland skat-variablerna eftersom de ingår bland insats-variablerna, som bedöms tydligare[22]. Att inkludera båda riskerar att skapa alltför starkt korrelerade kovariater och gör dessutom parameterskattningarna mer svårtolkade. Utöver dessa sökandekategorier ignoreras tillfälliga sökandekategorier (95-98). Om man bara tittar på dummyvariablerna för sökandekategorier tolkar man det som att dessa personer har referensvärdet (öppet arbetslös, skatkod 11).[23] Referensen betyder i det här fallet alltså att personen antingen har skat = 11, eller någon av de ovan nämnda. Därför behöver dessa kolumner tolkas tillsammans med insatserna.

2. Insatserna hämtas från beslut. Eftersom man kan ha flera insatser samtidigt är referensvärde inte relevant.

3. För funktionshinderkoder och sökta yrken (SSYK-2012) finns tre respektive fyra variabler i registren, detta för att möjliggöra för flera funktionshinderkoder respektive sökta yrken. Dessa kombineras så att dummies skapas som säger huruvida någon *har* respektive sökt yrke eller funktionshinderkod, det vill säga oberoende av om det angivits som sökt yrke 1, 2, 3, eller 4, till exempel. Utöver detta skapas dummies som beskriver hur *många* sökta yrken eller funktionshinderkoder en arbetssökande har (dessa hör till kategorin av dummies som beskrivs i nästa stycke).

4. SSYK-koderna består av 4 siffror i en hierarki där första siffran utgör en grov gruppindelning av olika yrken, andra siffran en något mindre grov indelning och så vidare, ned till ganska detaljerade yrkesbeskrivningar när alla fyra siffror är inkluderade. Vanliga yrken inkluderas på 4-siffernivå, men i stället för att bunta ihop alla andra yrken till en och samma övrigt-kategori utnyttjas hierarkin till att inkludera dummies på 3-siffernivå för de 3-siffergrupper som är tillräckligt vanliga (utan de yrken som representerats på 4-siffernivå), och på samma sätt med två siffror och slutligen en siffra. Eftersom detta förfarande i sig leder till mindre omfattande övrigt-grupper, och eftersom de

[22] Man kan bara ha en sökandekategori samtidigt och program-skat övertrumfar andra skat, vilket gör att sökandekategorier inte alltid fångar insatserna.
[23] Vid prediktion hanteras dessa skat-koder likadant, dvs skat 95-98 samt koder som motsvarar betydande insatser tolkas som att dessa personer har referensvärdet skat 11.

fyra positionerna ger utrymme för fler yrken per person, är det rimligt att sätta gränsen för vad som är tillräckligt vanligt för att få en egen dummy något högre än i det allmänna fallet och i nuläget är denna gräns satt till 10 000.

5. Det sökta yrke som anges på plats ett av fyra i registren är förknippat med vissa förväntningar på att aktivt söka jobb inom detta yrke. Det är dessutom obligatoriskt att ange ett sökt yrke på denna första position men inte på de senare. Detta gör att det angivna yrket på position ett, fokusyrket, sannolikt har större betydelse än övriga angivna sökyrken. Därför skapas separata dummies för angivet fokusyrke. Genom att utnyttja hierarkin i SSYK-koderna skapas variabler på 4-, 3-, 2- eller 1-siffernivå enligt samma procedur som för SSYK-koderna i sin helhet. I och med att dessa variabler endast baseras på en variabel (jämfört med fyra variabler för har_SSYK) finns färre förekomster av varje SSYK-kod. Därför används den generella gränsen på 1 000 förekomster för att fastställa vilka värden som är tillräckligt vanliga för att få en egen dummy.

6. Det finns två speciella koder som förekommer bland SSYK-variablerna: X21 (ej identifierbart yrke) och X33 (kan ej ta tidigare yrke). Om SSYK angivits som X21 tolkas det på samma sätt som att angivet SSYK saknas helt (vilket är *mycket* sällsynt). Koden X33 anges i princip uteslutande på position ett, dvs som ett fokusyrke. En variabel ”har_fokusyrkeX33” ingår i modellen så X21 och X33 skiljs åt. För de få individer som har X33 på en annan position än den första sorteras X33 in i övrigt-kategorin.

7. Till varje sökt yrke hör två variabler som säger huruvida den sökande har erfarenhet respektive utbildning i yrket. Dessa kombineras med respektive sökt yrke så att det skapas dummies på exakt samma sätt som SSYK-variablerna, men där endast de sökta yrken där den sökande har erfarenhet respektive utbildning räknas. På samma sätt kombineras erfarenhet och utbildning med det angivna sökyrket på position ett, fokusyrket.

Övriga dummyvariabler
För en del nominalvariabler finns ett begränsat antal möjliga värden, och i dessa fall beskrivs dummy-variablerna mer explicit i Tabell 1. Här ingår till exempel variabler som beskriver hur många sökta yrken och funktionshinderkoder den sökande har samt utbildningsnivå. För alla dessa används en dummy per förekommande värde förutom referensen, med ett undantag: 2 och 3 funktionshinderkoder buntas ihop.

Notera: den lägsta utbildningsnivån (0) och saknad uppgift buntas ihop till en dummy.

Interpolationsvariabler
För en variabel som till exempel ålder kan man i princip utnyttja att variabeln är en ordinalvariabel (ett numeriskt värde där värdet i sig har en betydelse). För att modellera åldersberoendet flexibelt kan man ändå välja att använda dummies, där

varje dummy beskriver tillhörighet till en åldersgrupp. Detta innebär dock att beroendet blir en trappstegsfunktion. Med en minimal ökning av antalet variabler kan man göra om trappstegsfunktionen till en kontinuerlig funktion genom att i stället skapa variabler som utgör interpolationsvikter för ett antal fördefinierade värden. Till exempel kan åldern 27 år beskrivas som (0,6 · 25 + 0,4 · 30) år.

Precis som för de flesta dummyvariablerna behövs här en referensvariabel för att inte riskera att skapa linjära beroenden.

Tabell 1. Sammanfattning av variabler som är med i nuvarande version av modellen.

Variabel	Referens	Grupp
Sökandekategori	11 (Öppet arbetslös)	skat
Insats	-	insats
Födelseland	None (Sverige)	födland
Har SSYK X (1-4 siffror)	-	ssyk
Har erf. i SSYK X (1-4 siffror)	-	ssyk
Har utb. i SSYK X (1-4 siffror)	-	ssyk
Har SSYK X som fokusyrke (pos. 1)	-	ssyk
Har erf. i fokusyrke	-	ssyk
Har utb. i fokusyrke	-	ssyk
Har funktionshinderkod X	-	fkod
A-kassa	00 (saknar) eller saknat värde	a-kassa
Kommun	0180 (Stockholm)	bostadsort
Utbildningsinriktning	010 (Bred, generell, typiskt låg)	utbildning
Kvinna	0 (man)	kön
Antal SSYK = 0, 2, 3, 4	1	ssyk
Antal yrken med erf. = 1, 2, 3, 4	0	ssyk
Antal yrken med utb. = 1, 2, 3, 4	0	ssyk
Antal funktionshiderkoder = 1, > 1	0	fkod
Utbildningsnivå = 0/None, 1, 2, 4, 5, 6	3 (Gymnasial utbildning)	utbildning
Interlokalt sökande	0 (Nej)	övrigt
Sökt arbetstid = 1, 2	12 (hel- eller deltid) eller saknas	övrigt
Inskrivningstid (intervall)	_to (första veckan sedan inskr.)	inskrivningstid
Ålder	40	ålder
Dag på året	61 (2:a mars, 1:a vid skottår)	kalender

2.5.4 Särskilda interaktionsvariabler

För vissa variabler används interaktioner med grupptillhörighet: i nuläget för insatser och inskrivningstid. Detta eftersom vi förväntar oss att insatser har olika effekter på olika grupper (se exempelvis Arbetsförmedlingen 2023), och eftersom beroendet av inskrivningstid kan förväntas vara olika i olika grupper (olika baseline-hasarder). De olika grupperna som används är i grunden alla kombinationer av kön, inrikes-/utrikesfödd, lågutbildad/högutbildad[24] och ung/mellanålder/äldre[25]. Totalt resulterar detta i 24 grupper. Varje kombination, dvs grupptillhörighet, skapar en egen dummy. Eftersom grupptillhörigheterna är ömsesidigt uteslutande måste en grupptillhörighet användas som referens för att undvika linjärt beroende. Därför är det endast 23 dummies som inkluderas i modellen: grupptillhörigheten man/lågutbildad/inrikesfödd/mellanålder används som referens.

När grupptillhörigheterna interageras med inskrivningstid respektive insatser måste hänsyn återigen tas till problemet med linjära beroenden (och multikollinearitet). Därför skapas inte interaktioner med alla möjliga kombinationer av insatser/inskrivningstid och grupptillhörighet. I stället delas populationen stegvis upp i mindre och mindre grupper så länge det finns minst 1 000 förekomster i varje grupp som har ett nollskilt värde på respektive variabel. Till exempel, för en viss insats finns det för få förekomster för att kunna ha med ålder i interaktionen. Därför skapas en ny gruppering där alla ålderskategorier slås ihop och därmed återstår interaktioner med alla möjliga kombinationer av insatsen tillsammans kön, inrikes-/utrikesfödd och låg-/högutbildad. På detta sätt grupperas interaktionerna stegvis och om nödvändigt slås alla grupper ihop.

2.6 Olika definitioner av inskrivningstid

I träningen av modellen har vi i förekommande fall tillgång till flera inskrivningsperioder för samma individ, men i produktion har vi i nuläget endast tillgång till data för pågående inskrivningsperiod. I träningen av modellen slås inskrivningsperioder ihop, främst på grund av utfallets varaktighetskrav, och i överlevnadsmodellen är det tiden sedan starten på den sammanslagna inskrivningsperioden som utgör tiden som ger upphov till olika tidsintervall och till exponeringstid. Däremot bestäms värdet på den tidsdummy som kopplar hasarderna till inskrivningstid utifrån den systemtidsstämpel som vid tidsintervallets start motsvarar den senast påbörjade inskrivningsperioden – detta för att bättre matcha det som finns tillgängligt i produktion.

[24] Upp till och med gymnasieutbildad räknas som lågutbildad.
[25] Gränserna är < 25 år för ung och > 55 år för äldre. Övriga räknas i detta sammanhang som mellanålder.

Dessutom är inte systemtidsstämpeln för inskrivningen i nuläget tillgänglig i produktion, utan det manuellt inmatade inskrivningsdatumet, se avsnitt 3.2.1.

2.7 Hantering av ”för” långa inskrivningstider

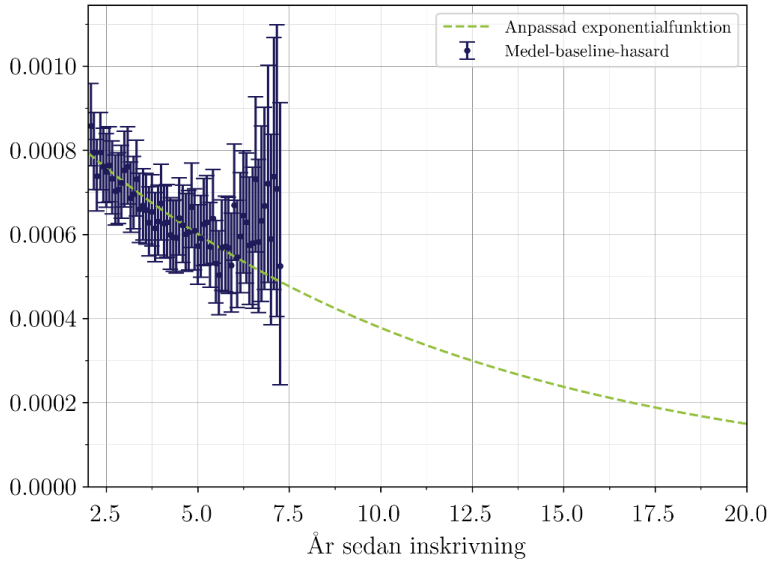
Modellen är tränad på inskrivningar som började som tidigast 2015-01-01, av skäl som beskrivs i avsnitt 3.4. Detta innebär att det finns en begränsning i hur långa inskrivningstider som modellen är tränad på – modellen ”vet” därför inte utan vidare hur den ska hantera inskrivningstider som är längre än knappt sju och ett halvt år (2 678 dagar), vilket påverkar alla prediktioner (med avseende på 365 dagar) som görs vid mer än 2 678 – 365 = 2 313 dagars inskrivning. I nuvarande version av modellen hanteras detta genom att skriva ned inskrivningstiden till 2313 dagar vid prediktioner för någon med längre inskrivning än så, men vi korrigerar också för en generell nedåtgående trend i hasarderna när inskrivningstiden ökar.

Korrektionen görs baserat på en medelvektor $\bar{\boldsymbol{\lambda}}_{\text{baseline}}$ av de gruppvisa baseline-hasarderna (som fångar beroendet av endast inskrivningstid), där medelvärdet är viktat i proportion till gruppernas storlek i träningsdata[26]. Baseline-hasarderna för utrikesfödda har en uppgång vid två års inskrivning som sannolikt har att göra med etableringsprogrammets längd[27] – vilket gör att det samma gäller för medelvärdet: på grund av detta används $\bar{\boldsymbol{\lambda}}_{\text{baseline}}$ för mer än 730 dagars inskrivning för att beräkna trenden, mer exakt bestämma den funktion som används för extrapoleringen.

Trenden bestäms genom att anpassa en linje till $\log \bar{\boldsymbol{\lambda}}_{\text{baseline}}$ med hjälp av generaliserad minsta kvadrat-anpassning (eng: *Generalized Least Squares*, GLS): en minsta kvadrat-anpassning som tar hänsyn till både varians och korrelationer hos $\bar{\boldsymbol{\lambda}}_{\text{baseline}}$.[28] Enklare uttryckt anpassas en exponentialfunktion $A \exp(-t/\theta)$ till $\bar{\boldsymbol{\lambda}}_{\text{baseline}}$: detta är ett enkelt val som garanterar att kurvan håller sig positiv och som även reproducerar $\bar{\boldsymbol{\lambda}}_{\text{baseline}}$ någorlunda väl. Resultatet syns i Figur 2. Observera att de senare värdena i $\bar{\boldsymbol{\lambda}}_{\text{baseline}}$ har betydligt större osäkerhet än de tidigare, och att de därför påverkar anpassningen betydligt mindre.[29]

[26] Viktningen görs i proportion till gruppernas storlek totalt sett: vikterna är alltså konstanta över tid. Detta för att fånga baseline-hasardernas förändring utan att blanda in sammansättningens förändring. Vikterna är baserade på de totala livslängderna för varje grupp (alltså antalet tidsintervall viktat med tidsintervallens längd).
[27] Detta ska inte övertolkas åt ena eller andra hållet: modellparametrarna kopplade till bland annat etableringen och andra ramprogram behöver tas i beaktning för att göra en riktig tolkning av detta.
[28] Denna kovarians beräknas genom att propagera kovariansen hos modellens parameterskattningar genom att utnyttja att $\operatorname{cov}(\log \bar{\boldsymbol{\lambda}}_{\text{baseline}}) \approx T cov(\hat{\boldsymbol{\beta}}) T'$ där T är det viktade medelvärdet av de matriser T_g som uppfyller $\log \boldsymbol{\lambda}_{\text{baseline},g} = T_g \hat{\boldsymbol{\beta}}$, se Bilaga 2.
[29] Denna påverkan av osäkerheterna – att de senare punkterna påverkar anpassningen mindre – förstärks av att osäkerheterna för de tidigare punkterna också är betydligt starkare korrelerade (osäkerheten är mer beroende av ett fåtal modellparametrar), vilket i hög grad låser lutningen på den linje som anpassas (till $\log \bar{\boldsymbol{\lambda}}_{\text{baseline}}$).

Figur 2. Medel-baseline-hasarderna $\bar{\boldsymbol{\lambda}}_{\text{baseline}}$ för mer än 730 dagar sedan inskrivning, med exponentialfunktion anpassad med generaliserad minsta kvadrat-anpassning.



För de individer som har längre inskrivningstid än 2313 dagar görs sedan korrektionen genom att multiplicera hasarden med två faktorer: först den faktor som flyttar ned medelvärdet hos $\bar{\boldsymbol{\lambda}}_{\text{baseline}}$ för de använda tidsperioderna (från och med 2313 dagar) till den anpassade exponentialfunktionen.[30] Den andra faktorn korrigerar för att vi flyttat oss bakåt och uppåt längs kurvan: faktorn är $\exp\left[-\frac{t-t_0}{\theta}\right]$ där t_0 är 2313 dagar och t är den faktiska inskrivningstiden.

2.8 Konjunkturkorrektioner

Hur konjunkturen ser ut under den arbetssökandes inskrivningsperiod påverkar individens chanser att komma ut på arbetsmarknaden. Av praktiska skäl (dataflöden och kvalitetskontroll av dessa) har kovariater kopplade till konjunkturläget inte kunnat inkluderas i själva överlevnadsmodellen i denna version. Nedan ges en beskrivning av ett problem detta medför samt den lösning som har implementerats för att åtgärda detta, utan att behöva ta in nya dataflöden i produktionsmiljön.

2.8.1 Översiktlig beskrivning av den implementerade lösningen

Eftersom konjunkturläget inte är inkluderat i nuvarande version av själva överlevnadsmodellen kommer de sannolikheter som modellen ger spegla en typ av genomsnittlig konjunktur under träningsperioden. Om konjunkturen vid tillfället för prediktion avviker från detta genomsnitt (vilket den alltid kommer göra i någon mån) så kommer *sannolikheterna* att vara något missvisande: de genomsnittliga andelarna som faktiskt fått ett positivt utfall kommer avvika från de genomsnittliga sannolikheterna. I vissa sammanhang är detta av mindre betydelse: själva rangordningen av arbetssökande påverkas inte mycket av normala förändringar i

[30] Denna första faktor används för att inte de osäkrast skattade baseline-parametrarna ska få särskilt stort genomslag.

konjunkturen.[31] I andra sammanhang är man verkligen intresserad av att själva sannolikheten är rättvisande.

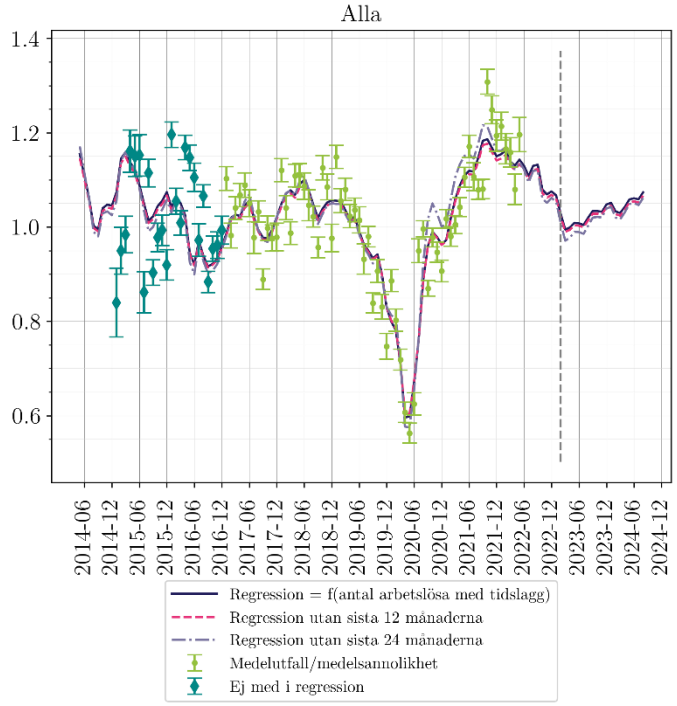
Figur 3 illustrerar problemet (och även en del av lösningen). Punkterna i figuren utgör kvoten mellan genomsnittliga utfall och genomsnittliga sannolikheter, månad för månad, för träningspopulationen. Under de första åren, 2015–2019, rör sig kvoten någorlunda nära 1 (centrerat något högre) men med signifikanta avvikelser. I samband med pandemin 2020 går kvoten ned kraftigt och vänder omkring 0,6, för att under senare delen av 2021 och under 2022 ligga närmare 1,2.

Den implementerade lösningen går kortfattat ut på att hitta och utnyttja samband mellan arbetslöshetsnivå och kvoten mellan genomsnittliga utfall och genomsnittliga sannolikheter. På så sätt fås en uppskattning av hur kvoten varierar över tid *som också har giltighet i framtiden*[32], med användning av prognosticerade arbetslöshetssiffror. Den modellerade kvoten används sedan till att korrigera sannolikheterna: om kvoten är 1,1 så kommer en predicerad sannolikhet på 0,33 korrigeras till 0,33/1,1 = 0,30.

Innan vi går vidare med detaljer i följande avsnitt är det värt att nämna att hela korrektionen görs grupp för grupp i 24 olika grupper av arbetssökande: figurerna med alla arbetssökande är bara illustrationer för att beskriva hur korrektionerna är gjorda.

[31] Olika grupper påverkas olika av förändrad konjunktur, vilket i viss mån påverkar rangordningen. Detta är något som delvis hanteras av korrektionen som beskrivs i detta avsnitt, eftersom den görs gruppvis.
[32] Eftersom vi vill använda modellen för prediktioner av sannolikheter för utfall i framtiden så är det essentiellt att korrektionen är meningsfull för framtida månader.

Figur 3. Kvoten mellan genomsnitten för andel positiva utfall och skattade sannolikheter månad för månad, i träningspopulationen.[33]



2.8.2 Grupper

Korrektionerna görs gruppvis i de 24 grupper som modellen använder i interaktioner med tid och insatser, se avsnitt 2.5.4. Anledningen till just detta val är delvis att det ger en enkel möjlighet till implementering genom att modifiera modellparametrar, men det är också ett sätt att fånga hur olika viktiga grupper påverkas olika.

2.8.3 Detaljer om använd arbetslöshetsdata

Antal i arbetskraften

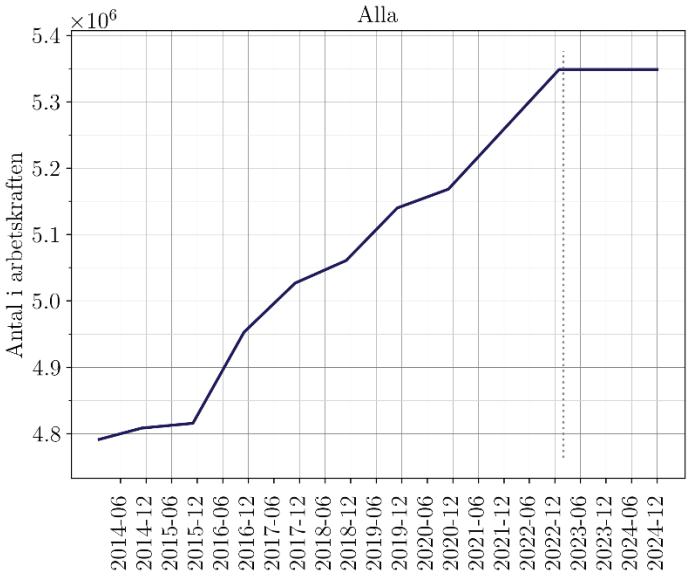
För varje grupp används siffror på antal i arbete från Statistiska centralbyråns årliga arbetskraftsundersökning[34] och lägger till arbetslöshetssiffror från Arbetsförmedlingen vid samma tillfälle (november för det aktuella året). Mellan de olika tillfällena för arbetskraftsundersökningen används linjär interpolation. Den senast tillgängliga arbetskraftsundersökningen är från november 2021: från denna månad fram till januari 2023 (denna månad var den sista med tillgängliga månadssiffror då korrektionerna beräknades) används extrapolation med konstant lutning, där lutningen hämtas från det sista tillgängliga året. Efter januari 2023

[33] Observationerna visas som punkter med 95-procentiga konfidensintervall. Kurvorna är en form av modellanpassningar till data som beskrivs i avsnitt 2.8.4.
[34] Dessa siffror är hämtade från tabellen martman.mprepl_dm i Arbetsförmedlingens datalager. Denna tabell gäller egentligen arbetskraften beräknad enligt metoden i officiell statistik: månadens arbetslöshetssiffror adderas till siffrorna från senast tillgängliga arbetskraftsundersökning, det vill säga arbetslöshetssiffrorna för varje månad 2022 adderas till arbetskraftsundersökningen från november 2020, till exempel. Genom att subtrahera arbetslöshetssiffrorna från godtycklig månad (januari) respektive år fås siffrorna för senast tillgängliga arbetskraftsundersökning. Att inte de officiella arbetskraftssiffrorna används direkt är för att ta bort missmatchningen mellan tiden för arbetskraftsundersökningen och arbetslöshetssiffrorna.

används prognosticerade arbetslöshetssiffror: under denna period (februari 2023 - december 2024) antas arbetskraften vara konstant.

Resultatet för totalen syns i Figur 4.

Figur 4. Antal i arbetskraften beräknat enligt den använda metoden.



Antal och andel arbetslösa
Antalet arbetslösa i varje grupp är hämtat från månadsdata i martman.sokande_vy i Arbetsförmedlingens datalager. För de totala siffrorna är överensstämmelsen med officiella siffror exakt under åren 2017-2022.[35]

I officiella arbetslöshetssiffror har det betraktade åldersspannet ändrats från 16-64 år till 16-65 år vid årsskiftet 2022/2023[36]. Den äldre definitionen behålls i de siffror som används här (både för arbetskraft och antal arbetslösa), eftersom vi är intresserade av hur konjunkturläget förändras, mätt så konsekvent som möjligt över tid. Den prognosticerade arbetslösheten (se avsnitt 2.8.5) skalas ned med den faktor som skiljer mellan den nya och den äldre definitionen i januari 2023.

I beräkningarna används *relativ* arbetslöshet i varje grupp, eftersom det säger mer om hur konjunkturläget är än det faktiska *antalet* arbetslösa vilket ju också påverkas av befolkningsunderlaget, speciellt när man tittar gruppvis. Andelen är alltså beräknad som antalet arbetslösa i gruppen delat med storleken på arbetskraften framtagen enligt föregående avsnitt.

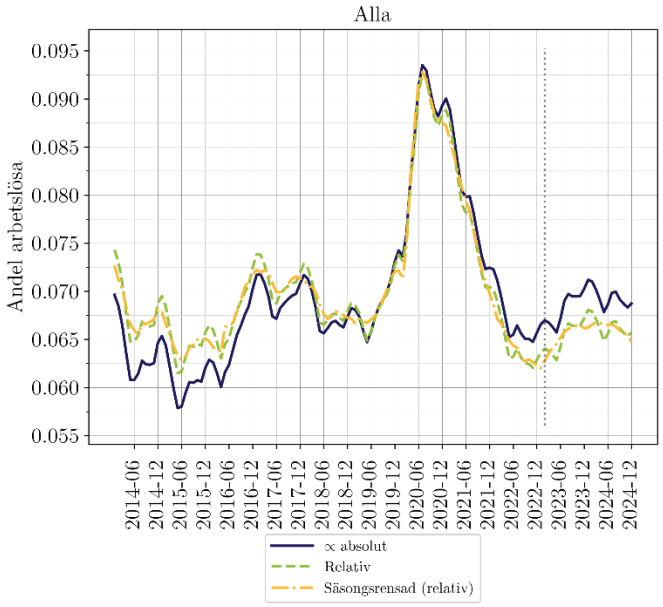
I regressionssteget som beskrivs nedan används säsongsrensad[37] (relativ) arbetslöshet. Detta görs av två anledningar: dels eftersom intentionen är att fånga

[35] Siffrorna före 2017 används inte i beräkningen av korrektionerna, vilket framgår nedan. Före 2017 finns små skillnader som sannolikt beror på förändringar i sökandekategorier som vi inte tagit hänsyn till. För 2023 finns skillnad på grund av förändrad pensionsålder som kommenteras i nästa stycke.
[36] Se Höjd pensionsålder påverkar månadsstatistik - Arbetsförmedlingen (arbetsformedlingen.se).
[37] Säsongsrensningen görs genom att göra en regression (OLS) med månadsdummies, och avvikelsen från medelvärdet av denna regression subtraheras från arbetslöshetssiffrorna.

säsongsberoende i modellen, dels eftersom vi är intresserade av större förändringar än säsongsvariationer i detta sammanhang.

I Figur 5 syns den säsongsrensade relativa totala arbetslösheten, tillsammans med ej säsongsrensad dito. En jämförelse med absolut arbetslöshet syns också.

Figur 5. Total relativ arbetslöshet över tid enligt beräkningarna för konjunkturkorrektionen.[38]



2.8.4 Regression som utnyttjar samband mellan ”kvoten” och arbetslöshetsnivå

Kärnan i hur korrektionerna är gjorda består av OLS-regressioner där logaritmen av kvoten mellan genomsnittliga utfall och genomsnittliga sannolikheter är utfallet, och den relativa säsongsrensade arbetslöshetsnivån ligger till grund för ett fåtal förklarande variabler. Samma regression görs för var och en av de 24 grupperna.

De förklarande variablerna består av arbetslöshetsnivån vid tre tillfällen: vid samma månad (som kvoten), två månader tidigare, och fyra månader senare. Modellen innehåller också ett intercept (en konstant). Att arbetslösheten två månader tidigare ingår grundar sig i en observation att sambandet var starkare till *förändringen* i arbetslöshetsnivån snarare än den faktiska nivån, vilket motiverar ett ”tidslagg”, följt av observationen att ett tidslagg på två månader ger en högre förklaringsgrad än en månad (kan tolkas som att förändringar över två månader säger mer då de är mindre känsliga för tillfälliga förändringar, eller säsongsbetonade förändringar som inte rensats bort). Arbetslösheten fyra månader senare motiveras i grunden av varaktighetskravet på fyra månader. Några olika varianter på regressionsmodeller

[38] Den gröna streckade kurvan visar den totala relativa arbetslösheten. Den gula, punktstreckade visar säsongsrensad motsvarighet och den heldragna blå kurvan är proportionerlig mot absolut arbetslöshet (och saknar skala) för att visa påverkan av förändringen i arbetskraftens storlek (som är betydligt större än så här för vissa grupper). Den vertikala punktade linjen markerar starten på prognosticerad data (se avsnitt 2.8.5).

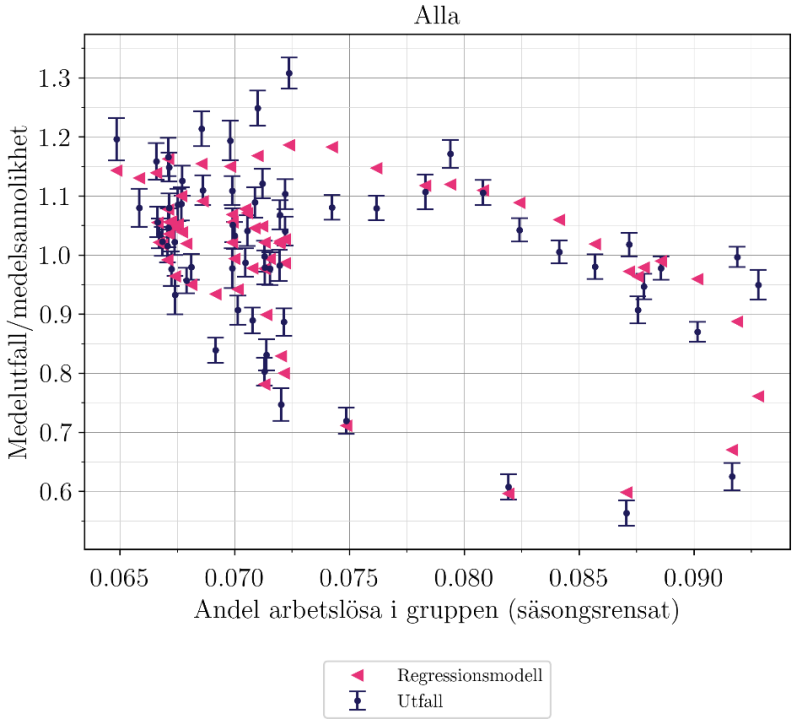
har prövats, och denna variant fick på totalen den högsta förklaringsgraden med ett så pass litet antal parametrar. Modeller med fler parametrar tillförde mycket lite, knappt något vad gäller justerad förklaringsgrad, och för giltighet i framtiden är det klokt att hålla ned antalet parametrar.

I regressionen används bara data för 2017 och framåt, eftersom sammansättningen av arbetssökande i träningspopulationen är så pass annorlunda i början, då (sammanslagna) inskrivningar som påbörjats före 2015 inte finns med.

Figur 6 visar observerade kvoter för alla jämfört med regressionsmodellens prediktioner av samma kvoter. Det finns i grunden ett negativt samband mellan kvot och arbetslöshetsnivå. Dessutom kan man följa kvotens rörelse genom pandemin motsols i en ring i figuren: vid ökande arbetslöshet ligger kvoten lägre än vid minskande arbetslöshet. En mindre ring från tidigare upp- och nedgångar syns längre till vänster i figuren.

En god överensstämmelse mellan utfall och regressionsmodellen syns, vilket också reflekteras av regressionsmodellens förklaringsgrad på 85 procent. Återigen är detta (sammantagna resultat för alla grupper av arbetssökande) bara en sammanfattande illustration, regressionen som används i korrektionerna är gjord grupp för grupp. De gruppvisa regressionerna ser liknande ut, men med större slumpmässiga osäkerheter eftersom grupperna är mindre.

Figur 6. Kvoten mellan genomsnittliga utfall och genomsnittliga sannolikheter för olika arbetslöshetsnivåer, och resultaten från regressionsmodellen, från och med januari 2017.



Resultaten från regressionen syns också i Figur 3, fast då som funktion av tid, där också prognosticerade resultat framåt syns. Dessutom visar Figur 3 motsvarande

kurvor där de sista 12 respektive 24 månaderna är borttagna ur regressionen, för att undersöka hur robusta resultaten är. Överensstämmelsen mellan de olika kurvorna är mycket god. För de gruppvisa regressionerna är skillnaden också väldigt liten när de sista 12 månaderna tas bort, medan det för 5 av 24 grupper uppstår lite större skillnader när de sista 24 månaderna tas bort. Detta kan förklaras av att nedgången i arbetslöshet efter pandemin då inte kommer med, vilket gör att särskilt viktig information inte kommer med.

2.8.5 Prognoser

Prediktionerna ska göras framåt i tiden, och korrektionerna använder arbetslöshetssiffror för tiden för prediktionen (och dessutom två månader innan och fyra månader efter). Därför krävs arbetslöshetssiffror för framtida månader, det vill säga prognosticerad arbetslöshet.

Vi använder i grunden samma arbetslöshetsprognos som använts som underlag i Arbetsförmedlingens senaste utgiftsprognos: därifrån fås prognosticerad total arbetslöshet från och med februari 2023 (när korrektionerna togs fram var januari 2023 senast tillgängliga med månadsdata). Eftersom vi använder gruppvis data behöver den prognosticerade arbetslösheten fördelas på de olika grupperna. Här gör vi denna fördelning på ett ganska enkelt sätt som ändå är ämnat att fånga det faktum att förändringar i arbetslöshet fördelas olika på olika grupper.

Genom att titta på hur kvoten mellan relativ arbetslöshet i gruppen och relativ arbetslöshet totalt förändras när den totala relativa arbetslösheten förändras ser vi ett tydligt samband: under senare år, åtminstone från och med 2020, är sambandet närapå linjärt. Exempel för två grupper syns i Figur 7 och Figur 8. Vi tar fram lutningen k_g i detta samband för alla grupper g, med regression (OLS) av en linje.

Vi antar i ett första steg att detta samband fortsätter gälla för förändringar i arbetslöshet under den prognosticerade perioden, så att kvoten mellan gruppens relativa arbetslöshet $a_g(t)$ och den totala relativa arbetslösheten $a_{\text{tot}}(t)$ följer

$$\frac{a_g(t)}{a_{\text{tot}}(t)} = \frac{a_g(t_0)}{a_{\text{tot}}(t_0)} + k_g[a_{\text{tot}}(t) - a_{\text{tot}}(t_0)],$$

där t_0 är sista månaden med observerade data: januari 2023. Genom att multiplicera båda sidor med $a_{\text{tot}}(t)$ fås ett uttryck för gruppens relativa arbetslöshet:

$$a_g(t) = a_{\text{tot}}(t) \cdot \left(\frac{a_g(t_0)}{a_{\text{tot}}(t_0)} + k_g[a_{\text{tot}}(t) - a_{\text{tot}}(t_0)] \right).$$

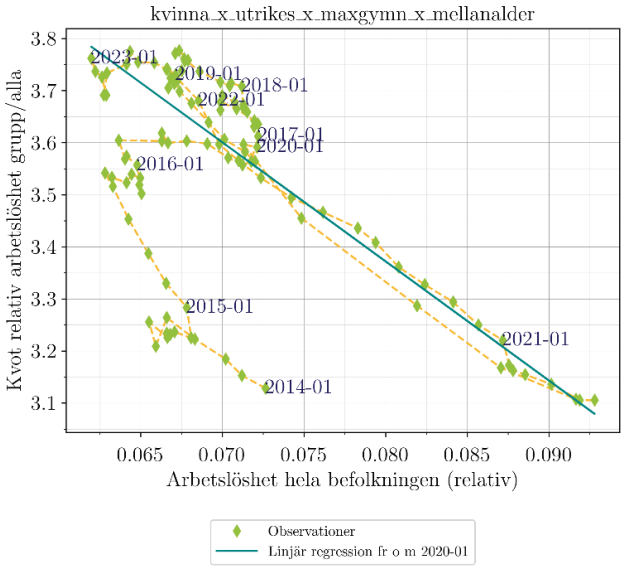
Detta leder till att summan av gruppernas absoluta arbetslöshet blir aningen skild från den totala absoluta arbetslösheten. Slutligen multipliceras därför de gruppvisa siffrorna månadsvis med den faktor $\eta(t)$ som gör att denna summa blir lika med totalen, alltså med

$$\eta(t) = \frac{A_{\text{tot}}(t)}{\sum_g A_g(t)},$$

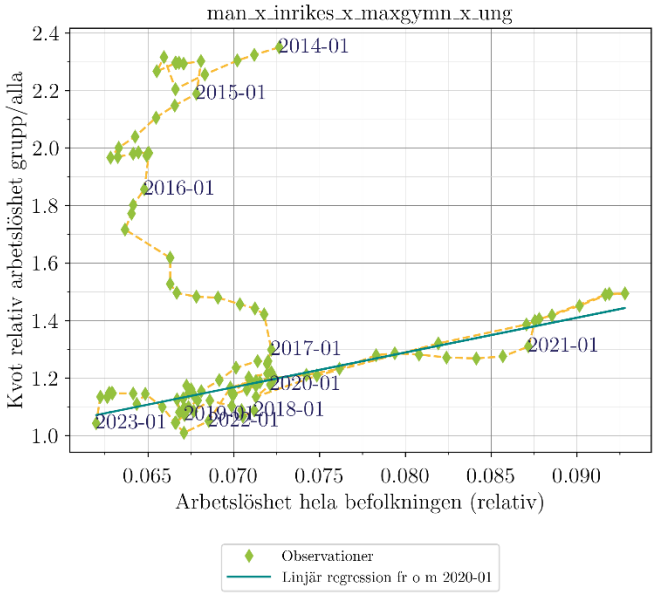
där $A_{\text{tot}}(t)$ och $A_g(t)$ är de absoluta motsvarigheterna till de relativa arbetslöshetssiffrorna $a_{\text{tot}}(t)$ respektive $a_g(t)$. Det kan vara värt att notera att $\eta(t) \approx 1$.

Prognosen för totalen läggs till före säsongsrensningen.

Figur 7. Kvoten mellan gruppens relativa arbetslöshet och total relativ arbetslöshet, för olika värden på total relativ arbetslöshet, där gruppen är utrikesfödda kvinnor på 25-55 år med högst gymnasieutbildning.



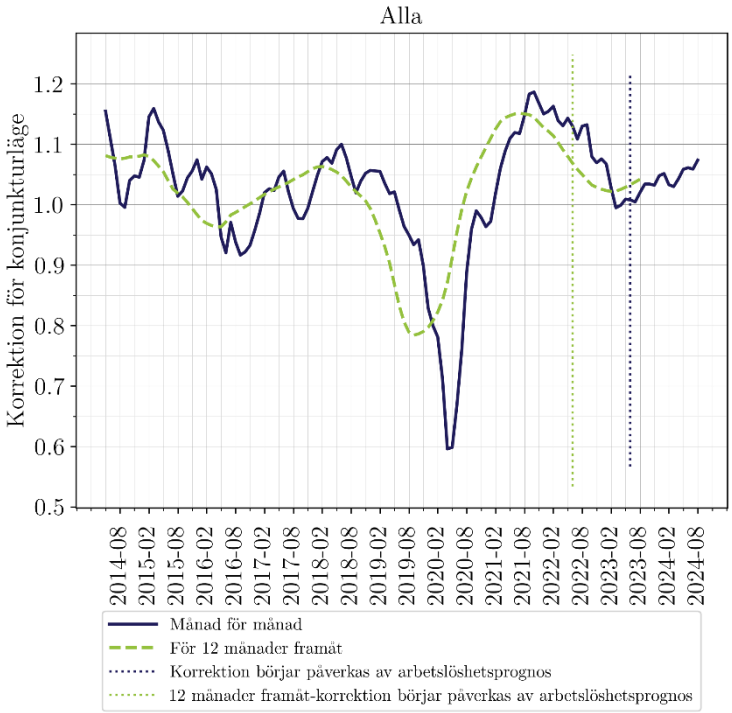
Figur 8. Kvoten mellan gruppens relativa arbetslöshet och total relativ arbetslöshet, för olika värden på total relativ arbetslöshet, där gruppen är inrikesfödda män under 25 år med högst gymnasieutbildning.



2.8.6 Resulterande korrektioner

De resulterande månatliga korrektionerna för alla syns dels i Figur 3 högre upp, dels i Figur 9 nedan. I Figur 9 syns också, vid varje månad där det går, ett genomsnitt för 12 månader framåt. Denna korrektion används när prediktioner görs för 12 månader framåt.

Figur 9. Resulterande korrektioner för alla över tid, både månad för månad och genomsnitt för 12 månader framåt.[39]



2.8.7 Hur implementeras korrektionerna i modellen?

Två sätt att implementera korrektionerna i modellen har använts och beskrivs i följande stycken. Det första är något mer rättframt och exakt, medan det senare av praktiska skäl används i produktion.

Direkt skalning av sannolikheter i analyssyfte
Korrektionerna kan användas direkt på de hasarder λ_{ij} som ligger till grund för beräkningen av sannolikheter för arbete/studier som modellen ger, genom att för varje grupp multiplicera motsvarande hasard värdet som gäller vid tidsintervallet. Korrektionerna är beräknade månadsvis; den månad som mittpunkten för respektive tidsintervall ligger i används.

Detta angreppssätt går inte att använda i befintlig modell genom att modifiera befintliga modellparametrar, men det går att använda i analyssyfte.

[39] Genomsnittet för 12 månader framåt används när modellparametrarna justeras för att inkludera korrektionen.

Via modellparametrar för användning i produktion
Beräkningen av hasarderna λ_{ij} kan skrivas som

$$\lambda_{ij} = \exp(\mathbf{x'}_{ij}\boldsymbol{\beta}) = \exp\left(\sum_k (x_{ij})_k \beta_k\right) = \prod_k \exp\left((x_{ij})_k \beta_k\right).$$

I befintlig modell finns det en parameter för varje grupp som korrektionerna är framtagna för, förutom för referensgruppen (svenskfödd man mellan 25 och 55 år med högst gymnasieutbildning). För att ändra hasarderna för alla individer i referensgruppen med en faktor a_{ref} får man utnyttja parametern för interceptet:

$$\lambda_{ij}^{\text{ny}} = a_{\text{ref}}\lambda_{ij} = a_{\text{ref}} \prod_k \exp\left((x_{ij})_k \beta_k\right)$$

$$= a_{\text{ref}} \exp\left((x_{ij})_{\text{intercept}} \beta_{\text{intercept}}\right) \cdot \prod_{k \neq \text{intercept}} \exp\left((x_{ij})_k \beta_k\right)$$

$$= \exp(\log a_{\text{ref}} + \beta_{\text{intercept}}) \cdot \prod_{k \neq \text{intercept}} \exp\left((x_{ij})_k \beta_k\right),$$

Alltså kan man ersätta modellparametern för interceptet med:

$$\beta_{\text{intercept}}^{\text{ny}} = \log a_{\text{ref}} + \beta_{\text{intercept}}.$$

Denna förändring ökar *alla* hasarder med faktorn a_{ref}, vilket man behöver ta hänsyn till när man modifierar parametrarna för övriga grupper. När interceptet är justerat behöver vi nu därför justera övriga grupper med a_g/a_{ref}, där a_g är korrektionsfaktorn för grupp g. Med precis samma härledning som ovan fås då att nya parametern för grupp g blir

$$\beta_g^{\text{ny}} = \log(a_g/a_{\text{ref}}) + \beta_g.$$

I nuvarande version av modellen är parametrarna korrigerade med avseende på 12-månadersprediktioner gjorda i april 2023 och bör uppdateras under HT 2023. Om de egentliga korrektionerna förändras kraftigt (på grund av kraftigt förändrat arbetsmarknadsläge jämfört med prognosen) kan det utgöra skäl för tidigare uppdatering.

2.9 Förklaringsmodell

För att modellens bedömningar ska vara transparenta kompletteras sannolikheterna från modellen med resultat från en *förklaringsmodell*. Dessa resultat beskriver, enkelt uttryckt, vilka ”egenskaper” hos den arbetssökande som bidrog mest till att bedömningen blev som den blev.

2.9.1 Gruppering av kovariater

Många ”egenskaper” (som kan sammanfattas på ett begripligt sätt för den arbetssökande) är uppdelade på flera kovariater, inte minst på grund av interaktionerna med grupptillhörighet som i sig byggs upp av flera grundläggande kovariater: kön, födelseland, utbildningsnivå och ålder – påverkan av dessa kovariater blir därmed utspridda på olika modellparametrar. Detta måste hanteras: annars kan sådana egenskaper hamna långt ned på listan på ett obefogat sätt (jämför med argument för detta i variansdekomponeringen i Bennmarker med flera (2007)).

Därför är de grundläggande kovariaterna indelade i grupper som framgår i Tabell 1.

2.9.2 Förklaringsmodellen beskriver påverkan på marginalen

Förklaringsmodellen beräknar hur stor påverkan respektive kovariatgrupp har på jobbchansen för en viss arbetssökande, givet att alla andra kovariater har de värden de har. Med påverkan menas här hur stort bidrag denna kovariatgrupp ger till sannolikheten jämfört med om alla kovariater i kovariatgruppen hade satts till sina medelvärden[40] i en referenspopulation (som beskrivs i avsnitt 2.9.3).

Bidraget till sannolikheten för individ *i* från kovariatgruppen *g* beräknas alltså som skillnaden mellan sannolikheten för individ *i* och den sannolikhet som man *skulle ha fått* om alla kovariater i gruppen *g* sattes till sina medelvärden (samtidigt som alla andra lämnades oberörda):

$$b_{ig} = p_i - p_{i,\text{medel}(g)}.$$

Detta innebär att påverkan av till exempel kön blir olika beroende på vilka andra egenskaper individen har: inskrivningstid, födelseland, utbildningsnivå, med mera. Förklaringsmodellen svarar inte *explicit* på hur olika kombinationer av egenskaper bidrar, men om kombinationen födelseland, utbildningsnivå och kön bidrar starkt så kommer alla dessa grupper hamna högt på listan över viktiga grupper: om kombinationen är viktig så kommer var och en av dessa grupper vara viktiga på marginalen. Detta eftersom det just är den sista gruppen som läggs till som fullbordar kombinationen, och när vi tittar på förändringar på marginalen så läggs var och en av grupperna till som den sista.

2.9.3 Referenspopulationen för medelvärdesberäkning

Medelvärdena på de kodade kovariaterna är beräknade på alla tidsintervall i träningspopulationen för sökandekategorier som är mer aktuella för profilering[41] under det sista tillgängliga året före censureringen: 2021-05-31 till 2022-05-31, detta för att få så aktuella medelvärden som möjligt men ändå inkludera ett helt år.

[40] Medelvärdena beräknas efter kodning av data. De flesta kovariater är då dummyvariabler (1/0), och medelvärdesbildningen innebär att värdena nu ligger mellan 0 och 1, exempelvis blir medelvärdet för kvinna lika med andelen kvinnor i referenspopulationen.
[41] Sökandekategorier som inte innebär subventionerad anställning eller arbetsmarknadsutbildning är borttagna. Det har (efter implementering i denna version) visat sig att detta inte stämmer *exakt* överens med de sökandekategorier som är aktuella för profilering (vilka framgår i avsnitt 4.3), men skillnaden på medelvärdena är liten.

Medelvärdet är viktat med avseende på exponeringstiden[42]. Detta innebär att medelvärdet är taget från något som ligger nära ett årligt genomsnitt på den del av sökandestocken som är aktuell för profilering.

Här kan också nämnas att modellparametrarna som används i förklaringsmodellen är desamma som i sannolikhetsberäkningen, det vill säga konjunkturkorrigerade för april 2023. Skillnad i konjunktur från referenspopulationens tidsperiod är alltså inte med i förklaringen och det är avsiktligt: det är inget speciellt med den valda tidsperioden vi vill fånga förutom att den är så aktuell som möjligt och sträcker sig över ett år.

2.9.4 Vad rapporteras till handläggaren?

Modellen returnerar bidragen b_{ig} för alla kovariatgrupper g för den aktuella individen i. På grund av gränssnittet mot handläggaren så rapporteras endast en 10 i topp-lista där de olika kovariatgrupperna sorteras med sjunkande absolutbelopp på b_{ig}: detta svarar alltså endast på frågan i vilken ordning bidragens *storlek* ordnas, och inte i vilken riktning bidragen går (även om det ofta går att gissa), eller hur stora skillnaderna i bidrag är. Det finns 13 kovariatgrupper, så de tre minst bidragande utelämnas (”insats” utelämnas i nuläget alltid eftersom de är satta till noll i prediktionerna).

De 13 kovariatgrupperna beskrivs för handläggarna i gränssnittet med hjälp av en kort förklarande text. Tabell 2 visar hur kovariatgrupperna kopplas till de förklarande texter som visas för handläggaren.

[42] För att vara exakt med avseende på tidsintervallens längd justerat för censurering utan positivt utfall, av praktiska skäl. Skillnaden mot att använda den faktiska exponeringstiden i detta sammanhang blir mycket liten.

Tabell 2. Förklaringar av kovariatgrupperna som visas för handläggarna.

Modellens gruppnamn	Handläggaren ser
skat	Sökandekategori
insats	Deltagande i insatser
fodland	Födelseland
ssyk	Sökta yrken samt erfarenhet och utbildning kopplade till dessa
fkod	Eventuell rätt till särskilda insatser på grund av en funktionsnedsättning registrerad hos Arbetsförmedlingen (alternativt avsaknad av sådana)
a-kassa	Information angående a-kassa
utbildning	Din utbildning
bostadsort	Förutsättningar i din kommun
kön	Kön
alder	Ålder
kalender	Tidpunkt på året för inskrivning
inskrivningstid	Din inskrivningstid
ovrigt	Angiven arbetstid och arbetssökande i andra geografiska områden

3 Dataanvändning och anpassning av modellen

3.1 Översikt

I detta avsnitt beskrivs det data som använts vid träning av modellen samt hur modellen har anpassats till data. Denna rapport beskriver endast data som använts för träning av modellen. Av juridiska skäl används ett separat dataflöde för

prediktion i produktionsmiljön. Hur själva prediktionen görs är däremot lika i både tränings- och produktionsmiljö.

Översiktligt kan träningen av modellen delas in i två steg:

1. Generering av grundläggande paneldataset.
2. Anpassning av modellen till data.

Det viktigaste grundläggande datasetet består av olika variabelvärden för en stor mängd personer vid starten av de olika tidsintervall som beskrivs i avsnitt 2.2.4. Det finns även ett dataset som innehåller information om och när personerna har nått ett utfall eller censurerats enligt beskrivningar i avsnitt 2.3 respektive 2.4.

När modellen är tränad kan den användas för prediktion. Då krävs samma data för varje person som vid träningen men endast vid en tidpunkt, i övrigt finns all information i modellparametrarna och själva specifikationen (i princip tredje ekvationen i avsnitt 2.2 och en koppling mellan indata och parametrar). Se kortfattad beskrivning av prediktion med tränad modell i avsnitt 3.6.

3.2 Generering av grundläggande paneldataset

Det data som används för träning kommer från Arbetsförmedlingens register via Data Warehouse Classic (DWC). Enligt utlåtande från Rättsavdelningen får inskrivningar som legat aktiva i AIS någon gång under de 10 senaste åren plus innevarande år användas för modellträning (Arbetsförmedlingen 2022c). Detta innebär att vi även får med de inskrivningsperioder som startat tidigare än för 10 år sedan men som fortfarande är aktiva efter 2012-01-01[43]. I och med att arbetet med data och modell gjordes under hösten 2022 var senast tillgängliga data 2022-09-20, det vill säga data till och med tredje kvartalet 2022 har använts.

3.2.1 Grundstädning

För att informationen från DWC ska vara optimalt anpassad för syftet att träna den modell som beskrivs i denna rapport genomförs en grundstädning av aktuella data. Huvudsyftet med grundstädningen är att på ett enhetligt sätt definiera inskrivningsperioder samt perioder med olika insatser respektive skat-koder.

Om en individ har en inskrivningsperiod som tar slut för att sedan påbörja en ny inskrivningsperiod mindre än 30 dagar från att den föregående avslutades, slås perioderna ihop. Fler än två perioder kan slås ihop givet att det är mindre än 30 dagar från slutet på en period till början på nästa period för samtliga perioder som slås ihop.

De städade inskrivningsperioderna utgår ifrån systemtidsstämpel för inskrivningen, vilket för ca 3% av inskrivningsraderna innebär en differens i tid jämfört med

[43] Från och med 1 december 2022 gäller inte längre begränsningen om att data för statistikframställning ska gallras efter 10 år. Så när själva modellen tränades fanns inte 10-års begränsningen. Dock användes endast data från och med 2015-01-01, se avsnitt 3.4.

manuellt rapporterat inskrivningsdatum. För hälften av raderna med differens, är denna skillnad under en vecka, med mönstret att manuellt rapporterat inskrivningsdatum ligger före systemtidsstämpeln (retroaktiv inskrivning). De inskrivningar med större differenser än så har i regel mycket långa inskrivningstider, på så vis att differensen ändå framstår som begränsad i sammanhanget.

Anledningen till att vi här väljer att gå på systemtidsstämpeln och inte den manuellt rapporterade inskrivningstid som vi får in i produktion är att de flesta av de andra uppgifter vi använder i träning utgår från just systemtidsstämplar. Går vi på manuellt rapporterad inskrivningstid i träningen innebär det i typfallet med en differens mellan tidsstämplarna att vi inte har information om andra kovariater för starten av inskrivningstiden, vilket skapar problem vid modellträning.[44]

För insatserna görs städningen av data framför allt för att rensa bort rader som innehåller felregistreringar samt för att få fram en korrekt start och ett korrekt slut för varje period i en insats. Detta görs bland annat genom att kontrollera så att statuskoderna indikerar att ärendet är giltigt samt att plocka ut rader där det gått minst en dag mellan insatsens start och insatsens slut.

För skat-koderna finns liknande typ av dilemma som för inskrivningsperioderna, systemtidsstämpeln och den manuellt satta tidsstämpeln stämmer inte överens i alla fall. En ytterligare problematik med just skat-koder är att de ofta ändras i efterhand. Efter kontakt med handläggare samt genomgång av handläggarstöd för registrering av programkoder är bilden att de flesta av skat-ändringarna är motiverade. Det vill säga, en ögonblicksbild av vad som ligger aktivt i systemet en viss dag innehåller en massa felaktiga skat-koder. Skat-koden uppdateras sedan av handläggare när felet upptäckts, exempelvis i kontakt med den sökande. Innan skarp profilering för Rusta och matcha ska handläggare gå igenom de individrelaterade posterna som går in i profileringen noggrant, och uppdatera dessa om fel noterats. Givet den arbetshistoriken används det manuellt registrerade datumet för skat-koden snarare än när förändringen registrerades i systemen.

3.2.2 Paneldata

Med hjälp av den data som städats enligt beskrivningen ovan konstrueras ett paneldataset med alla individer och dess inskrivningsperioder enligt avsnitt 3.4, vid alla tidpunkter för starten av ett nytt tidsintervall enligt avsnitt 2.2.4. Varje individ och inskrivningsperiod har alltså en rad per tidsintervall. Genom att använda information från städade data med perioder i en insats respektive perioder med skat-koder innehåller varje rad också information om skat-kod samt eventuell insats under varje tidsintervall. Detta skapar det huvudsakliga datasetet för träning av modellen.

[44] Ett alternativ som hade möjliggjort användning av det manuellt satta inskrivningsdatumet hade varit att bakåtfylla information för tidiga delar av inskrivningen för att täcka luckan mellan manuellt inskrivningsdatum och systemtidsstämpel. Detta har dock bedömts som för riskfyllt och har därmed inte implementerats i denna version.

De städade inskrivningsperioderna med tillhörande avregistreringsorsaker behandlas sedan för att koda fram utfall enligt definitionen i avsnitt 2.3. Varaktighetskravet på 122 dagar kontrolleras och avregistreringar som inte uppfyller kravet tas bort. Censurering för vissa avregistreringsorsaker enligt beskrivning i avsnitt 2.4 läggs till. Detta förfarande genererar ytterligare ett dataset, med en rad per inskrivningsperiod i huvudsakliga datasetet, som används för att anpassa modellen till data.

3.3 Uppdelning av arbetssökande i delpopulationer

Den population av arbetssökande som är tillåten att använda för modellträning delas innan träning upp i tränings-, validerings- och utvärderingspopulation. Detta görs för att kunna träna modellen och sedan kunna utvärdera den tränade modellen på ett så rättvisande sätt som möjligt. Uppdelningen är kopplad till den specifika personen, alltså kommer alla inskrivningsperioder tillhörande samma individ hamna i samma grupp.

Populationen är uppdelad fyra grupper:

1. En slutgiltig utvärderingspopulation som ska inte användas innan det verkligen är dags. Innehåller 15% av den totala populationen.

2. En utvärderingspopulation i reserv, som också kan inkluderas i framtida träningspopulation om det så småningom bedöms som en bra idé. Innehåller 15% av den totala populationen.

3. En population som är tänkt att använda för validering. Innehåller 20% av den totala populationen.

4. En population avsedd för träning av modellen. Innehåller 50% av den totala populationen.

Ytterligare detaljer om varför uppdelningen behövs och motivering till uppdelningen presenteras i avsnitt 3.3.1.

3.3.1 Motivering av en strikt reserverad utvärderingspopulation

Personer som inte får användas i modellutveckling
För att minska risken för att överanpassning (eng: ”overfitting”) gör att modellens prediktionsförmåga överskattas bör inte de individer som ingått i modellens träning användas i utvärdering av modellen. Överanpassning kan ske på flera sätt: både genom direkt överanpassning av parametrar (ett enkelt exempel är anpassning av ett n:e-gradspolynom där n är stort) och genom överanpassning av modellval: till exempel val av interaktionstermer eller modellspecifikation mer generellt. I modellutveckling prövas ofta en del olika varianter mer eller mindre systematiskt (det kan även handla om rättningar av implementeringsfel som ett exempel på mindre systematiskt testande). Därför bör man i princip se till att den population som ska användas i utvärderingssyfte är strikt reserverad för detta syfte, så att inte specifikt denna data guidar val som görs. För att vara noggrann med detta bör samma

population vara reserverad hela tiden, så att inte en liten grad av manuell överanpassning smyger sig in över tid.

Motivering av storlek
Den reserverade utvärderingspopulationen ska vara så liten som möjligt för att kunna utnyttja så mycket data som möjligt i modellträning och modellutveckling. Samtidigt behöver den vara stor nog så att tillräckligt liten slumpmässig osäkerhet nås i de utvärderingar som görs. Den analys som vid beslut om denna storlek var påtänkt och som skulle vara den mest begränsande i detta avseende är en undersökning av kalibrering på gruppnivå, där grupperna bestäms av diskrimineringsgrunderna.

I denna analys bortser vi från hur insatser mellan start- och utfallsdatum påverkar hur kalibreringen undersöks. Varje grupp (baserad på diskrimineringsgrunderna) delas in i $m = 10$ lika stora delgrupper[45] sorterade efter predicerad sannolikhet att ha nått det aktuella utfallet, och andelen positiva utfall i delgruppen bestäms. Nollhypotesen som (mer eller mindre explicit) testas är att den predicerade sannolikheten är korrekt i varje grupp. Om gruppen består av n individer består delgruppen av n/m individer, och om vi för enkelhets skull antar att den predicerade sannolikheten för positivt utfall i delgruppen är konstant[46] lika med p fås, under nollhypotesen, att antalet positiva utfall är fördelat enligt $\mathrm{Bin}(n/m, p)$, alltså binomialfördelat med gruppstorlek n/m och utfallssannolikhet p. Den slumpmässiga variansen på andelen blir då (Alm och Britton 2008)

$$\sigma^2 = \frac{p(1-p)}{n/m} = \frac{mp(1-p)}{n}.$$

Detta maximeras för $p = 0{,}5$ och då fås $\sigma^2 = m/(4n)$. För de gruppstorlekar som är aktuella kan normalapproximation användas[47], så att halva bredden på ett 95-procentigt konfidensintervall blir $1{,}96\sigma$. Om vi för de minsta grupperna tolererar en halv bredd på 0,05 fås $\sigma^2 \leq (0{,}05/1{,}96)^2$, vilket ger

$$n \geq \frac{m \cdot 0{,}05^2}{4 \cdot 1{,}96^2} \approx 4\,000.$$

Detta är ett riktmärke för den minsta gruppstorlek som tolereras i denna analys. En av de minsta grupperna baserat på diskrimineringsgrunderna och som vi kan hitta i data (endast en diskrimineringsgrund i taget) är de med funktionshinder. Det är sannolikt att man till exempel vill kunna studera denna grupp bland de som vid ett givet tillfälle är i garantierna. Bland inskrivna den sista februari 2022 uppfylls dessa villkor av 32 000 personer. För att utvärderingspopulationen i sökandestocken vid detta tillfälle ska innehålla 4 000 sådana personer kan vi, med lite marginal, välja att

[45] När denna analys gjordes användes 10 delgrupper: i underökningen av kalibrering i avsnitt 4 används 25. Båda siffrorna är godtyckliga och blir en avvägning mellan hur finfördelad bild man får och hur stora osäkerheterna blir.
[46] Egentligen varierar den predicerade sannolikheten något inom gruppen, men de ligger nära varandra. Analysen är ungefärlig.
[47] De aktuella gruppstorlekarna är långt större än de som anges i den tumregel för normalapproximation av binomialfördelningen som presenteras i Alm och Britton (2008).

andelen i utvärderingspopulationen är 15 procent. I nuläget är ytterligare 15 procent reserverade som en mindre strikt utvärderingspopulation.

3.3.2 Träningspopulation och valideringspopulation för framtida modellutveckling

Vid träning av denna modell har 50 procent av individerna har använts som träningspopulation. Detta kan eventuellt utökas något i senare versioner, men det är bra att behålla mer än bara utvärderingspopulationen utanför träningspopulationen. Ytterligare en population kan användas för att förbättra vissa modellval, till exempel: vilka parametrar (inklusive interaktioner) tillför prediktiv förmåga? I nuläget finns en valideringspopulation på 20 procent reserverad för att kunna använda för sådana empiriskt grundade val. Eftersom den inte har använts i detta syfte ännu, har denna population använts i utvärderingen (den har inte på något sätt påverkat träning eller modellval) av nuvarande modell som presenteras i avsnitt 4.

3.4 Populationsurval: Individer och tid

Häften av alla individer är slumpvis utvalda till att kunna ingå i träningspopulationen, se avsnitt 3.3.2. För dessa individer används alla inskrivningsperioder[48] som påbörjas mellan 2015-01-01 och 2021-05-31. Data som är tillgänglig i projektet gäller från och med 2012-01-01 till och med 2022-09-30. Vid kvalitetskontroll av tillgängliga data upptäcktes brister i data mellan 2012-01-01 och 2014-12-31. Bland annat skedde ett byte av yrkesklassificeringen enligt SSYK år 2014 och det förekom även problem med tidsstämplar före 2015.[49] Därför används endast inskrivningsperioder som påbörjats från och med 2015-01-01 trots att det enligt juridiska bedömningar hade varit möjligt att använda även äldre inskrivningsperioder. På grund av varaktighetskravet censureras alla kvarstående individer 122 dagar tidigare än sista tillgängliga datum, det vill säga 2022-05-31.

Som nämnts i avsnitt 2.4.1 om censurering då tiden tar slut, kommer då viss information om när inskrivningen påbörjades med i tidpunkten för censureringen (någon som censureras efter 3 år måste vara inskriven senast 2019-05-31, till exempel). Alltså blir inte censureringen helt icke-informativ. Ett alternativ skulle vara att följa alla individer under en begränsad tid – säg två år – och sedan censurera oavsett om möjligheten att följa individen kvarstår, vilket är mer likt det som gjorts i Bennmarker med flera (2007). Detta gör verkligen censureringen icke-informativ, men förfarandet skulle innebära en stor nackdel i praktiken: det skulle sätta en stark begränsning på hur långa inskrivningar som kunde följas, om inte endast någorlunda föråldrade data används. Detta skulle kräva att vi modellerade påverkan av längre

[48] Inskrivningsperioderna avser de städade inskrivningsperioderna som beskrivs i avsnitt 0, och är här hopslagna dels om det är mindre än 30 dagar mellan två inskrivningsperioder, dels om avbrott mellan dem inte uppfyller varaktighetskravet: alltså om de inte är längre än 122 dagar.
[49] Att använda två olika varianter av ssyk-klassificeringen hade varit möjligt men det hade krävt användning av nycklar mellan de båda varianterna. Bedömningen gjordes att det var för svårt att avgöra kvaliteten hos dessa nycklar och därför är det mest lämpligt att endast använda data med samma variant av ssyk.

inskrivningstider (och inte bara *riktigt* långa inskrivningstider, se avsnitt 2.7), och vi föredrar alternativet att censureringen inte är helt icke-informativ.

3.5 Anpassning av modellen till data

För att anpassa modellen behövs det grundläggande paneldatasetet, med en rad per tidsintervall för varje person och inskrivningsperiod samt utfallsdatasetet som beskrivits i avsnitt 3.2.2.

3.5.1 Expansion av data

Paneldatasetet kompletteras och kombineras sedan med utfallsdatasetet så att:

1. Det finns en rad per ocensurerat tidsintervall för respektive individ
2. Det finns en dummy-kolumn per tidsintervall förutom det första
3. Det finns en kolumn som beskriver tiden som arbetssökande i intervallet (lika med längden på intervallet om inte individen får ett positivt utfall eller censureras i just detta intervall)
4. Det finns en kolumn med indikator för positivt utfall i tidsintervallet
5. Sorterar om så att det är sorterat efter i första hand identifikationsnyckel för respektive hopslagen inskrivningsperiod och i andra hand tidsintervallets start för individen

När paneldatat transformerats enligt ovan så räknas vissa kolumner upp (som ålder och tidpunkt på året), och data kodas om så att dummy-variabler skapas enligt beskrivning i avsnitt 2.5.3. Utöver kodning till dummies ”ackumuleras” insatserna, det vill säga det skapas variabler som säger huruvida någon *har haft* respektive insats, och för hur länge sedan det var. Insatserna interageras med grupptillhörighet, enligt de grupper som beskrivs i avsnitt 2.5.4. Även inskrivningstiden interageras med grupperna.

3.5.2 Anpassning med iterativ viktad minsta kvadrat-anpassning

Enligt beskrivning i avsnitt 2.2 kan modellen beskrivas som en Poisson-regression. Ett Poisson-regressionsproblem kan lösas genom att maximera log-likelihood-funktionen numeriskt, och det finns stöd för detta i Pythonpaketen scikit-learn. Denna implementering använder dock en onödigt tung beräkning eftersom hela det likelihoodbesläktade ”deviance” beräknas i varje steg. Det finns dock en effektivare typ av algoritm (Rodríguez 2007): iterativ viktad minsta kvadrat-anpassning, som dessutom ger en god skattning av kovariansmatrisen för parametrarna på köpet. Denna iterativa viktade minsta kvadratanpassning är implementerad för att hitta en anpassning av modellen till data.

Eftersom algoritmen kan göra så att vissa kolumner blir numeriskt sett linjärt beroende så kontrollerar funktionen i varje iteration om detta har hänt: i så fall

exkluderas en eller flera kolumner ur algoritmen i fortsättningen. För att inte startgissningen ska påverka vilka kolumner som tas bort körs hela algoritmen en gång till, med resultatet i förra omgången som startgissning på modellparametrarna. Detta görs om tills samma kolumner tas bort i två efterföljande iterationer.

När metoden konvergerat sparas modellparametrarna tillsammans med kolumnnamn, datatyper med mera för att snabbt och enkelt kunna läsas in igen när modellen används för prediktion. När konjunkturkorrektionen är genomförd sparas både de korrigerade och de okorrigerade parametrarna. De korrigerade parametrarna används för prediktion men de okorrigerade parametrarna behövs när konjunkturkorrektionen ska göras om för att vara bättre anpassad till en annan tidsperiod.

Den hantering av långa inskrivningstider som beskrivs i avsnitt 2.7 regleras genom att spara de två faktorer som resulterar från den anpassade exponentialfunktionen, för att sedan kunna hämta och applicera dem när modellen läses in inför prediktion.

3.6 Prediktion

Vid prediktion med redan tränad modell behövs motsvarande paneldata som vid träningen, men endast för ett tillfälle, och ingen utfallsdata.

3.6.1 Expansion av data

Samma expansion av data som för träningen görs, med skillnaden att utvidgningen av antalet rader blir mycket större: det läggs till rader för alla tidsintervall som kan behövas i prediktionen. För de flesta kolumner är innehållet detsamma som vid prediktionstillfället, men inskrivningstid, ålder och tidpunkt på året räknas upp. Kodningen till dummies matchas de dummies som skapats under modellträningen. Det vill säga, vid prediktion sker kodningen till dummies genom att skapa samma dummies som skapades vid modellträningen.

3.6.2 Själva prediktionen av sannolikheter

Vid prediktion av sannolikheter ges ett expanderat prediktionsdata enligt ovan samt ett antal dagar. De resulterande sannolikheterna anger sannolikhet för ett utfall, enligt avsnitt 2.3, inom det specificerade antalet dagar. Hasarden i varje tidsintervall beräknas baserat på kovariaterna i varje tidsintervall enligt beskrivning i avsnitt 2.2, varefter korrektionen för långa inskrivningstider görs enligt beskrivning i avsnitt 2.7. Därefter beräknas sannolikheterna. För att beräkningen av sannolikheten ska kunna göras krävs att expansionen i föregående steg gjorts tillräckligt långt. Det vill säga, det finns tidsintervall som sträcker sig fram till det angivna antalet dagar för beräkning av sannolikhet.

De sannolikheter som modellen i nuläget skattar är sannolikheten för varaktig (minst 122 dagar) avregistrering till arbete (orsak 1,2 eller 3) eller studier (orsak 7) som påbörjas inom 365 dagar från prediktionstillfället.

3.7 Jämförelser mellan produktions- och träningsmiljö

I och med att modellträningen och prediktionen från modellen som kommuniceras till handläggare ligger i olika miljöer, har det varit viktigt att jämföra olika bitar av miljöerna med varandra.

Jämförelser av träningsdata från DWC med data från produktionsmiljön har genomförts för att säkerställa att variablerna i träning till sitt format överensstämmer med motsvarande fält i produktionsdataflödet.

Efter att modellen tränats färdigt och implementerats för prediktion i produktionsmiljön jämfördes prediktionsresultat från träningsmiljön med prediktionsresultat från produktionsmiljön. Testet är genomfört innan produktionssättning och visar på fullständig överensstämmelse mellan produktions- och träningsmiljö. Testet ska återupprepas vid varje förändring av filerna med modellkoden.

4 Hur presterar modellen?

Detta avsnitt beskriver hur modellens prestation undersöktes före produktionssättning. Vid uppdateringar av modellen ska en liknande undersökning göras, liksom den kontroll av överensstämmelse mellan tränings- och produktionsmiljö som beskrivs i avsnitt 3.7.

4.1 Kvalitetskriterier

4.1.1 Välkalibrerade bedömningar

I en välkalibrerad modell betyder jobbchansen vad den utger sig att betyda. Exempel: Om 100 sökande har fått sin jobbchans bedömd att vara 20% (inom t.ex. 12 månader) så kommer 20 av dessa 100 sökande också hitta arbete (inom 12 månader). Det finns argument för att välkalibrerade bedömningar är viktiga för Arbetsförmedlingens bedömningsmodell: 1) Bedömningar av jobbchans som är informativa om genomsnittlig faktisk jobbchans kan användas för att göra informerade beslut om anvisningsgränser till olika nivåer av insatser; 2) Det skapar en transparens kring de genomsnittliga förutsättningarna bland deltagare i de olika insatsnivåerna (spåren). Detta borde vara särskilt viktigt i samarbetet med de fristående aktörerna, och i utformningen av ersättningsmodell.

4.1.2 Likabehandling och att undvika diskriminering

Litteraturen beskriver ett flertal olika definitioner och perspektiv på rättvisa och likabehandling vid automatiserade bedömningar (se t.ex. Noriega-Campero, Bakker, Garcia-Bulle & Pentland, 2018). Vilken definition av likabehandling som man kräver att en bedömningsmodell ska leva upp till beror på sammanhanget. Olika definitioner av likabehandling kan vara matematiskt oförenliga, därför behöver man välja vilket

perspektiv som är viktigast för det sammanhang som bedömningsverktyget används (se t.ex. Kleinberg, Mullainathan & Raghavan, 2016).

Låt oss utgå från Arbetsförmedlingens bedömningsmodell och de i data observerbara diskrimineringsgrunderna i diskrimineringslagen: ålder, kön, utländsk-/svensk bakgrund, funktionsnedsättning. Likabehandling på gruppnivå kan då definieras på flera olika sätt, här är två vanliga definitioner:

1. Att bedömningarna i genomsnitt betyder samma sak och är lika informativa oberoende av grupptillhörighet (välkalibrerade jobbchanser på gruppnivå).

2. Att chansen till att korrekt bedömas ha hög/låg jobbchans, om personen har hög/låg faktisk jobbchans, är oberoende av grupptillhörighet.

Kleinberg, Mullainathan & Raghavan, 2016 visar att 1 & 2 tyvärr är svåra att förena (undantaget vid perfekt prediktion eller när grupper som jämförs har samma riskfördelning), trots att båda definitionerna av likabehandling är högst önskvärda att uppfylla.[50] Definition 1 är särskilt viktigt om bedömningarna används till att fatta informerade beslut (anvisningsgränser, ersättningssystem) och där t.ex. handläggare och fristående leverantörer behöver kunna lita på att prediktionerna stämmer i genomsnitt. Samtidigt borde 2 vara viktigt för att leva upp till myndighetens uppgift att fördela stöd rättvist efter individuellt stödbehov. Vi har valt att i första hand säkerställa att modellen lever upp till likabehandlingskriterium 1. En motivering till detta val finns i slutet av detta avsnitt.

Vi kan först notera att likabehandling och icke-diskriminering enligt diskrimineringslagen inte nödvändigtvis är samma sak. En vanlig strategi för att undvika att bedömningsverktyg diskriminerar är att definiera olika ”skyddade grupper” som inte får missgynnas. Detta synsätt sammanfaller inte nödvändigtvis med strikt likabehandling. Det är till exempel inte nödvändigtvis förenligt med kriterium 2 ovan, vilket säger att individer med samma stödbehov ska ges samma möjligheter till stöd oavsett vilken grupp individen tillhör. En möjlig definition av ”skyddade grupper” i Arbetsförmedlingens verksamhet är de grupper som myndigheten har identifierat med lägre konkurrensförmåga historiskt *och* som är en diskrimineringsgrund i lagen[51]. ”Skyddade grupper” skulle då vara äldre arbetssökande (55+), arbetssökande med utländsk bakgrund samt arbetssökande med funktionsnedsättning.[52]

[50] Debatten i USA kring verktyget COMPAS för att bedöma återfallsrisk i brott illustrerar detta dilemma. Verktygets utvecklare har säkerställt välkalibrerade bedömningar på gruppnivå, nödvändiga för att domare ska kunna lita på att bedömningar betyder samma sak för tex svarta och vita åtalade. Samtidigt ledde journalistiskt arbete till att man kunde belysa att verktyget missgynnade svarta åtalade genom att det andra likabehandlingskriteriet inte var uppfyllt (se t.ex. Dieterich, Mendoza and Brennan, 2016).
[51] Lågutbildade ingår inte trots att de har lägre konkurrensförmåga, eftersom låg utbildning inte utgör en diskrimineringsgrund.
[52] Vi argumenterar för att det ligger i bedömningsverktygets konstruktion (givet att man inte kräver likabehandling enligt 2 ovan) att dessa ”skyddade grupper” inte kommer att missgynnas -i meningen lägre chans till anvisning till stöd. Denna slutsats stöds också empiriskt i Arbetsförmedlingen (2021) som utvärderade bedömningsstödet till den första versionen av Rusta och matcha.

En vanlig uppfattning är att profileringsmodeller riskerar att reproducera historisk diskriminering. Det stämmer delvis eftersom profileringsverktyg utvecklas (tränas) med hjälp av historiska data, vilket innebär att eventuell historisk diskriminering reflekteras i bedömningarna, men huruvida det är diskriminerande beror också på vad verktyget används till. Arbetsförmedlingens bedömningsmodell innehåller t.ex. uppgift om utländsk bakgrund. Vid en bedömning av en sökandes jobbchans får en sökande med utländsk bakgrund (allt annat lika!) en lägre skattad jobbchans än sökande med svensk bakgrund. Det beror på att indikatorn utländsk bakgrund (allt annat lika!) fångar att utlandsfödda i genomsnitt har fler icke-observerade faktorer som är förknippade med lägre jobbchanser än vad svenskfödda har. Exempel på sådana faktorer är lägre nivå av kunskaper i det svenska språket och högre risk för att utsättas för diskriminering i anställningsprocessen. Alltså, om utlandsföddas jobbchanser historiskt har påverkats negativt av diskriminering i anställningsprocessen så är det något som bidrar till att en sökande med utländsk bakgrund tenderar att bedömas ha lägre jobbchans (allt annat lika) när de profileras. *Det genuina* stödbehovet tenderar i så fall att vara överskattat bland utlandsfödda[53], vilket innebär en högre chans att tilldelas stöd (men också högre risk att felaktigt anvisas stöd i onödan bland utlandsfödda med lågt stödbehov).

Allt annat lika, bedöms alltså arbetssökande som tillhör en observerbar grupp som står längre ifrån arbetsmarknaden att ha en lägre jobbchans. Om det är bra eller dåligt ur individens synpunkt beror på vad jobbchansbedömningen används till. I fallet med Arbetsförmedlingens bedömningsmodell används den till att rangordna sökande efter jobbchans, så att sökande med lägre jobbchanser får mer av rustande och matchande insatser eller fördjupat stöd. Om dessa insatser är något som gynnar individens möjligheter till arbete och inkomst är alltså en lägre bedömd jobbchans något som ökar möjligheterna på arbetsmarknaden.[54] På det sätt som dagens bedömningsstöd används har alltså sökande som tillhör grupper med lägre konkurrensförmåga större möjligheter till stöd vid en jämförelse med sökande med liknande stödbehov som tillhör grupp med högre genomsnittlig konkurrensförmåga (se Arbetsförmedlingen 2021 för evidens). Risken för diskriminering i lagens mening – att eventuellt missgynnande är kopplat till en diskrimineringsgrund – bör alltså vara låg i de rekommenderade anvisningar som bedömningsstödet genererar.

Valet att säkerställa likabehandlingskriterium 1, det vill säga att bedömningarna i genomsnitt betyder samma sak och är lika informativa oberoende av grupptillhörighet (välkalibrerade jobbchanser på gruppnivå), är förenligt med ovanstående slutsats kring icke-diskriminerande rekommenderade anvisningar (i

[53] Huruvida det genuina stödbehovet överskattas beror på vilken typ av diskriminering det handlar om. Är det frågan om statistisk diskriminering från arbetsgivarnas sida, som bygger på en reell genomsnittlig skillnad i produktivitet mellan utlandsfödda och svenskfödda, så är det inte lika uppenbart att det genuina stödbehovet på gruppnivå är överskattat bland utlandsfödda.

[54] Ofta förekommer uppfattningen att uppgifter om kön, ålder, födelseland och funktionshinderkod bör utelämnas i en profileringsmodell för att undvika diskriminerande utfall. Med syftet att bedöma stödbehov så bra som möjligt skulle dock en sådan strategi vara svår att försvara. Låt säga att man skulle välja att utelämna informationen om att en arbetssökande har en funktionsnedsättning. Dessa individer skulle då likställas med övriga grupper och riskera att gå miste om det stöd de kan ha rätt till.

enlighet med diskrimineringslagen). De argument som varit vägledande i detta val är annars i huvudsak:

1. Om handläggare kan lita på att en viss bedömning (eller rekommenderad anvisning) betyder ungefär samma sak för alla arbetssökande som fått den bedömningen, är risken lägre för att handläggare väger in grupptillhörighet när de eventuellt gör annan bedömning än bedömningsstödet. Om det motsatta gäller, att samma bedömda jobbchans betyder till exempel 40% faktisk jobbchans för grupp X och 60% för grupp Y, så är risken stor att handläggare i stället försöker kompensera för denna snedvridning i den samlade bedömningen.

2. Om fristående aktör kan lita på att en viss bedömning (eller rekommenderad anvisning) betyder ungefär samma sak för alla arbetssökande som fått den bedömningen, är chansen större att alla deltagare i tjänsten erbjuds lämpligt stöd. Om det faktiska stödbehovet däremot skiljer sig kraftigt åt mellan deltagare från olika grupper som borde ha samma stödbehov (till exempel i samma spår i Rusta och matcha), finns större risk för oälkommen behandling så som parkering av arbetssökande på grundval av deras grupptillhörighet.

4.1.3 Träffsäkerhet

Ett vanligt sätt att utvärdera en modell är att använda utvärderingsdata med ett stort antal individer som går att följa över tid och jämföra prediktioner och verkliga utfall *för varje individ*. Det vill säga att jämföra prediktion gjord vid tidpunkt t för individ i:s framtida okända utfall vid tidpunkt $t + m$ med individ i:s realiserade utfall vid tidpunkt $t + m$. Arbetsförmedlingens bedömningsmodell har vid prediktion ett okänt binärt utfall: varaktigt jobb eller studier inom 12 månader, eller inte. Modellen gör sannolikhetsbedömningar för att tillhöra kategorin ”jobb”==1 (den positiva klassen), alltså jobbchansen (på skala 0-100%). Det finns dock inga motsvarande realiserade ”jobbchanser” på individnivå att jämföra mot för att utvärdera modellens träffsäkerhet. Verkliga sysselsättningsutfall är oftast enklast att hantera som binära. Antingen är man i sysselsättning (tillhör positiva klassen) eller så är man det inte (tillhör negativa klassen).

Problemet att jämföra predicerade sannolikheter med binära realiserade utfall kan man hantera vid utvärdering genom att testa hur bra modellen är på just binär klassificering. Man väljer en gräns i fördelningen av prediktioner, och – med vårt exempel – klassificerarar arbetssökande med en predicerad jobbchans under gränsen som jobb==0 (”negativa”), och observationer över gränsen som jobb==1 (”positiva”). Gränsen kan väljas godtyckligt, exempelvis till 50%, eller empiriskt baserat på lämpliga kriterier. Vanligt är att sätta gränsen så att man minimerar falska positiva eller falska negativa klassificeringar (tillhör negativa/positiva klassen men prediceras ”positiva”/”negativa”). De senare strategierna är viktiga i många sammanhang, särskilt vid potentiellt livsavgörande bedömningar inom, till exempel, medicin eller rättsväsende. I den här utvärderingen använder vi alltså binära prediktioner på

individnivå för att kunna jämföra mot binära faktiska utfall på individnivå, något som är nödvändigt för att kunna använda vanliga mått på träffsäkerhet.

Ett annat sätt att utvärdera är att testa modellens förmåga att rangordna individerna korrekt. Då kan man jämföra rangordningen efter skattade sannolikheter – som jobbchansen i Arbetsförmedlingens modell – med faktiska utfall, binära eller icke-binära. Man testar då alltså inte om en prediktion var ”sann” eller ”falsk” på individnivå, utan i vilken utsträckning sannolikheternas rangordning är korrekt. I fallet med ett binärt faktiskt utfall, som jobb==1/0, kan ett sådant mått (som ROC AUC) säga vad sannolikheten är att en slumpmässigt vald individ med jobb==1 rankas högre av modellens jobbchans än en slumpmässigt vald individ med jobb==0.

Det finns ett stort antal mått på träffsäkerhet och rangordningsförmåga hos bedömningsmodeller. Förenklat kan man säga att det är ganska enkelt att välja ett lämpligt mått som har en hög intern validitet. Med intern validitet menar vi att måttet är användbart för att välja mellan olika varianter av modeller applicerade på det specifika problem, och det specifika datamaterial, man arbetar med. Och vidare till att mäta prestationsutveckling över tid för sitt specifika verktyg, när man gör modellutveckling eller vid förändringar i data över tid.

En utmaning är dock att det kan vara svårt att relatera testvärden till någon extern referensram, för att avgöra vad som är ett högt eller lågt testresultat i utgångsläget. Anledningen till det är att svårighetsgraden i bedömningsuppgiften har stor betydelse för vilken träffsäkerhet som är möjlig att uppnå. Ett exempel: Om vi bygger en modell för att bedöma om det blir vackert väder eller inte imorgon, baserat på ingående data från olika vädertjänster med hög kvalitet, så kan vi förvänta oss en väldigt hög träffsäkerhet. Vi kommer nästan alltid att predicera rätt. Om vi däremot bygger en modell för att bedöma individers chanser till ett varaktigt jobb inom 12 månader är osäkerheterna mycket större, och vi kan inte förvänta oss att predicera rätt i lika hög grad.

Bedömningens svårighetsgrad beror dels på uppgiften i vid bemärkelse, som väder imorgon jämfört med jobb inom ett år, men det beror även på hur den underliggande fördelningen i data ser ut. Om vi fortsätter med väderexemplet: i geografiska områden med mycket stabilt väder - t.ex. solsken 350 av 365 dagar om året, är det enkelt att uppnå väldigt hög träffsäkerhet med ett primitivt verktyg som är tränat att predicera solsken varje dag alla dagar om året. På motsvarande sätt är det enklare att predicera jobbchans i en skev fördelning av faktiska okända jobbchanser, till exempel om de flesta antingen har väldigt låg eller väldigt hög jobbchans.

Sammanfattningsvis finns det flera olika mått på hur bra en modell är, som säger något skilda saker. För en viss specifik situation är det relativt lätt att avgöra vilken modell som är bäst – den som har högst värde på ett lämpligt mått – men det är svårare att jämföra modeller som har applicerats på skilda situationer, eller generellt sätt uttrycka vilka testvärden som är bra (oavsett situation). Vi återkommer till val av mått i avsnitt 4.2 och till diskussionen om vilka testvärden som är bra i avsnitt 4.4.4.

4.2 Val av utvärderingsmått

I rapporten har vi beskrivit överlevnadsmodellen som är tränad att bedöma stödbehov i termer av individens jobbchanser med hänsyn tagen till olika insatser som individen eventuellt har fått. En komplikation vid utvärdering är dock att predicerade jobbchanser rensade för effekterna av deltagande i insatser inte har någon direkt motsvarighet i realiserade utfall. Realiserade utfall kan vara påverkade av insatser. Denna komplikation kan hanteras på olika sätt:

1. Inte göra något alls – alltså jämföra modellens jobbchanser mot realiserade utfall. Detta är rättframt att utföra och lätt att beskriva, men kan vara orättvist till nackdel för en modell som försöker hantera insatser mellan start och utfall. Ett sådant konservativt mått på modellens prestationsförmåga är dock att föredra framför ett mått som överdriver modellens förmåga.

2. Betrakta prediktioner och utfall tidsintervall för tidsintervall. Detta är relativt lätt att beskriva, och blir korrekt och rättvisande. För vissa mått blir det dock svårt att jämföra med andra studier, eftersom det blir ett mycket skevt klassificeringsproblem: sannolikheterna för utfall inom respektive intervall är små. För andra mått, som huruvida modellen är välkalibrerad, är detta emellertid det bästa angreppsättet.

3. Använda utvärderingsmått särskilt lämpade för överlevnadsmodell.

I utvärdering som redovisas i denna rapport använder vi samtliga tre angreppssätt. Nedan relaterar vi till dessa angreppsätt som 1, 2 och 3.

4.2.1 Kalibrering och likabehandling

Det finns formella test för kalibrering men de är svåra att tolka i en situation där man har mycket data och inte kan förvänta sig en *helt* perfekt kalibrering i teoretisk mening[55]. Vi bedömer därför att det är tillräckligt med en grafisk redovisning och inspektion. För att studera kalibrering på totalen och på gruppnivå för män/kvinnor; utlandsfödda/sverigefödda; äldre/yngre samt för funktionsnedsatta/inte funktionsnedsatta använder vi figurer där vi visar relationen mellan bedömda jobbchanser och faktiska genomsnittliga utfall. Vi gör detta i enlighet med angreppsätt 2 ovan.

4.2.2 Träffsäkerhet

Modellen ska vara bra på att träffsäkert predicera och rangordna stödbehov. Det kanske vanligast förekommande måttet på träffsäkerhet är Accuracy som mäter andelen korrekta klassificeringar (sanna positiva + sanna negativa) av alla klassificeringar (sanna positiva + sanna negativa + falska positiva + falska negativa). Måttet är attraktivt för att det är enkelt och lätt att kommunicera. Det är också användbart eftersom det ofta redovisas i utvärderingar, och därmed möjligt att jämföra med flera olika studier. Det har dock begränsningen att det är enkelt att

[55] I en sådan situation skulle storleken på testdatasetet avgöra vad det formella testet resulterar i, oavsett om avvikelsen är av praktisk betydelse eller inte.

uppnå höga värden även för simpla modeller när klassificeringsproblemet är snedfördelat, dvs när det är enkelt att gissa rätt eftersom de allra flesta som bedöms tillhör en och samma kategori.

Vi utvärderar modellen med Accuracy som vårt första mått på träffsäkerhet. Vi använder då angreppssätt 1 ovan (konservativt, dvs det missgynnar modellen).

Det finns attraktiva mått för att avgöra hur bra modellen är på att korrekt rangordna individerna efter – i vårt fall – jobbchans (eller omvänt stödbehov). Ett sådant vanligt mått är ROC AUC. ROC kan grafiskt illustreras med en kurva (ROC) som visar modellens ”andel sanna positiva” mot ”andel falska positiva” för olika gränsvärden för klassificering. AUC är arean under denna kurva som i ett värde sammanfattar hur bra modellen är på att rangordna. Detta värde är ekvivalent med rangkorrelationen mellan prediktioner och verkliga utfall. Det har därmed en enkel tolkning. Måttet säger vad sannolikheten är att en slumpmässigt vald individ med jobb==1 (dvs, med faktiskt positivt utfall) rankas högre av modellens jobbchans än en slumpmässigt vald individ med jobb==0 (dvs, med faktiskt negativt utfall).

Genom tolkningen som rangkorrelation har ROC AUC ett nära besläktat mått som brukar användas för att utvärdera överlevnadsmodeller där man har tid till positivt (eller negativt) utfall som utfallsmått: Concordance. Detta mått utgörs av sannolikheten att längden på två slumpvis valda arbetslöshetsperioder (vars längd går att åtskilja[56]) är korrekt ordnade av modellen. Detta mått går att generalisera så att tidsvarierande kovariater hanteras (Kremers 2007): jämförelsen görs då vid den senaste tidpunkt (mätt i tid sedan start på arbetslöshetsperioderna) då båda arbetslöshetsperioderna går att följa i data. Precis samma generalisering går också att göra för ROC AUC: jämförelsen av de två individernas skattade sannolikheter görs vid tidpunkten för det positiva utfallet för den individ som fick ett sådant.

Concordance-måttet är mer nyanserat än måtten som går via binär klassificering: det behövs inget val av varken tid till utfall eller av tröskelvärde. Dessutom kan måttet hantera de som avregistrerats av okänd orsak på ett ”mjukare” sätt. Anta till exempel att vi har ett par av arbetslöshetsperioder där den ena individen avregistrerats till arbete efter 200 dagar och den andra avregistrerats av okänd orsak efter 300 dagar. Då vet vi att det gått bättre för den individ som fått det positiva utfallet, information som också räknas i Concordance-måttet: detta par skulle varit rätt ordnat om sannolikheten efter 200 dagar var högre för individen som nådde utfallet då. Vi kompletterar därför också med detta mått. ROC AUC för binär klassificering, liksom Concordance-mått för överlevnadsanalys, antar vanligtvis värden mellan 0,5 och 1, där ett värde på 0,5 indikerar att modellen inte bidrar till att rangordna bättre än slumpen och ett värde på 1 indikerar att den rangordnar perfekt.

[56] Det går inte alltid att skilja arbetslöshetsperiodernas längd åt på grund av censurering: om båda är censurerade eller den ena censureras före det att den andra får ett positivt utfall. Detta kan jämföras med att endast par av individer där den ena fått positivt utfall och den andra inte (med avseende på en viss tid) jämförs i ROC AUC (vilket alltså utesluter betydligt fler par från jämförelsen).

Vi utvärderar modellen med ROC AUC samt Concordance som mått på rangordningsförmåga. Detta är ett exempel på angreppssätt 3 ovan.

Flertalet träffsäkerhetsmått bygger på komponenterna ”sanna positiva”-, ”sanna negativa”-, ”falska positiva”- och ”falska negativa”-klassificeringar. Dessa komponenter brukar sammanfattas i en så kallad ”Confusion Matrix”, en fyrfältsmatris med faktiska värden på y-axeln och predicerade värden på x-axeln. Med dessa komponenter kan granskare använda sina favorit-mått för att studera hur Arbetsförmedlingens bedömningsmodell presterar. Matrisen ger också information om hur snedfördelat klassificeringsproblemet är.

Vi redovisar ”The Confusion Matrix” med angreppssätt 1 ovan (konservativt, dvs det missgynnar modellen).

4.3 Utvärderingsdata

Syftet med utvärderingen är att säkra att det statistiska verktyget håller hög kvalitet för de sökande som bedöms. Målpopulationen för utvärderingen är alltså den kategori av sökande som för närvarande är föremål för bedömning i produktion.[57] Utvärderingsdata består av ett slumpmässigt urval av individobservationer som har reserverats strikt för utvärderingen[58]. Dessa data har inte i något skede använts för modellträning. I dessa data kan vi följa de sökandes verkliga arbetsmarknadsutfall över tid. Storleken på utvärderingsdata har avgjorts enligt avsnitt 3.3.1. Se nedan för antal observationer och kort beskrivande statistik. Urvalet har dragits bland alla sökande i den aktuella målpopulationen som var inskrivna hos Arbetsförmedlingen vid ett tänkt bedömningstillfälle som är 2021-03-01. Bedömningstillfället hat valts så att vi har möjlighet att observera faktiska utfall om varaktigt jobb eller studier något längre än 12 månader från utvärderingens bedömningstillfälle 2021-03-01, se avsnitt 4.4.4.

Antal observationer = 97 268[59]

Beskrivande statistik:

- Andel med varaktigt jobb inom 12 månader: **24,9 %**
- Andel varaktig avregistrering med okänd orsak inom 15 månader[60]: **13,0 %**

[57] Den aktuella målgruppen för bedömning i produktion är arbetssökande med Skat-koder: 11, 14, 15, 23, 28, 34, 46, 56, 59, 67, 68, 69, 70, 71, 72, 73, 75, 80, 81, 83, 85, 86, 87, 88, 95, 96, 97, 98. Notera att Skat-koder 21 och 22, tim- och deltidsanställda, inte finns med i aktuell målgrupp i denna utvärdering. Dessa sökande har kunnat profileras för Rusta och matcha 1 (de är med i aktuell målgrupp fram till april 2023), men de ska inte kunna ta del av Rusta och matcha 2.

[58] Den använda populationen utgör 20 procent av alla individer som skulle kunna komma i fråga givet övriga begränsningar. Den population som slutgiltigt har reserverats för utvärdering utgör 15 procent, men ”sparas” i nuläget eftersom en mindre strikt reserverad population ännu ej använts i modellträning.

[59] Detta är ett baserat på ett slumpmässigt urval på 20 procent av sökandestocken vid tillfället, där ingen individ är densamma som ingått i träningen av modellen. De aktuella 20 procenten av sökandestocken bestod enligt tabellen hist_aktso vid tillfället av 137 088 personer varav 136 301 av praktiska skäl kunde användas i utvärderingen enligt den rättade data som används i projektet (de allra flesta faller bort på grund av retroaktiva avregistreringar). Av dessa hade 97 268 de aktuella sökandekategorierna vid datumet.

[60] I stora delar av analysen tas dessa individer bort (det framgår när), likt det som gjorts i Arbetsförmedlingen (2021) för den tidigare modellen. Anledningen till det är att vi inte vet om utfallen egentligen är positiva eller inte. Det är vanligt att man tappar kontakten med Arbetsförmedlingen på grund av att man fått ett arbete eller

- Andel kvinnor: **47,4 %**
- Andel utlandsfödda**: 52,9 %**
- Andel äldre än 55 år**: 17,6 %**
- Andel med kod för funktionsnedsättning: **14,6 %**

4.4 Resultat

4.4.1 Resultat: Kalibrering på totalen

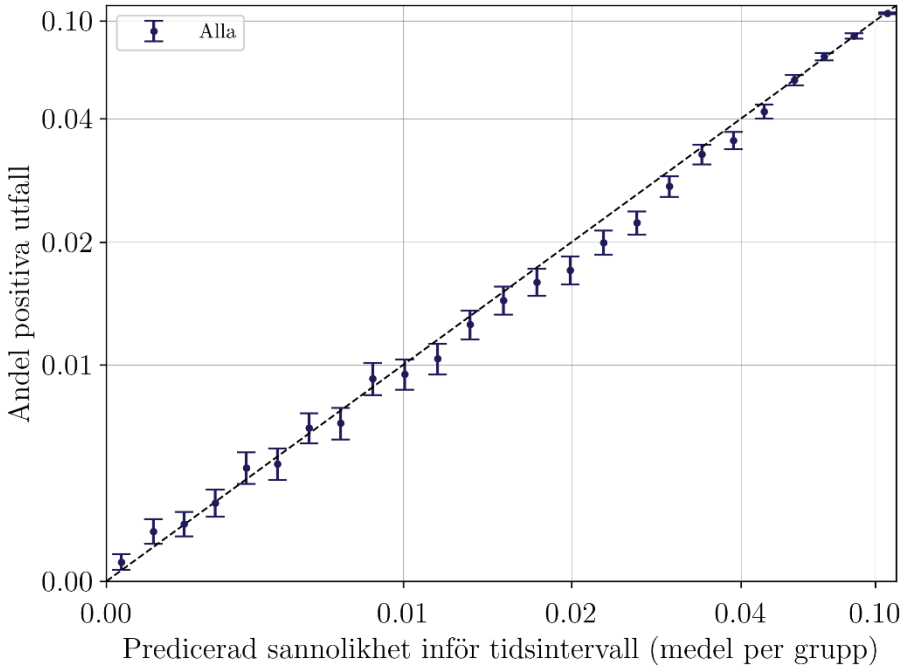
Figur 10 visar en kalibreringskurva för aktuell målpopulation för bedömning. Eftersom vi utvärderar en överlevnadsmodell så är det viktigt att först notera att varje arbetssökandes sannolikhet till varaktigt arbete eller studier inom 12 månader skattas genom att dela upp dessa 12 månader i tidsintervall med separata prediktioner (med uppdaterad information) inför varje sådant tidsintervall. Dessa intervall har olika längd. Under de 13 första veckorna av en inskrivning används intervall med en längd på 7 dagar och därefter används intervall med en längd på 30–31 dagar (se avsnitt 2.2.4). Eftersom de flesta inskrivningsperioder är betydligt längre än 13 veckor är den typiska längden på intervallen 30–31 dagar. Exempelvis innebär detta att en arbetssökande som har ett års inskrivning vid profileringstillfället och som inte får positivt utfall inom 12 månader finns med i 12 tidsintervall med upprepade prediktioner vid starten av varje intervall. Resultaten i Figur 10 är baserade på alla observerbara tidsintervall för varje individ som börjar tidigast 2021-03-02[61] och senast 2022-03-01, så att samma individ kan återkomma i flera olika intervall (en panel med individer över tid). Dessa observationer av tidsintervall är ordnade efter ökande sannolikhet och sedan grupperade i 25 grupper, med lika många observationer i varje grupp. Punkterna som utgör kurvan i figuren anger andel faktiska positiva utfall inom dessa 25 grupper, där grupperna är ordnade efter den genomsnittliga predicerade jobbchansen bland de arbetssökande inför starten av respektive tidsintervall.[62] Punkterna kommer med 95%-konfidensintervall. En perfekt kalibrering indikeras av den streckade 45-graderslinjen.

på grund av att man faller ur arbetskraften (Arbetsförmedlingen 2022). Avregistreringar av orsak 4 (till Samhall) och 8 (avliden) behandlas likadant som avregistreringsorsak 6.

[61] Ett villkor för att vara med i populationen är att man är inskriven 2021-03-01, så intervall som inkluderar detta datum är borttagna från denna del för att inte detta villkor ska påverka resultaten.

[62] Tidsintervallens längd har i utvärderingen bestämts som hela tidsintervallets längd med undantag för om den arbetssökande blivit censurerad (utan positivt utfall) under tidsintervallet: då räknas tidsintervallets längd fram till dagen för censurering. Baserat på denna längd på tidsintervallet (upp till 31 dagar) har sannolikheten för utfall i tidsintervallet beräknats. Detta är aningen annorlunda från hur det görs i modellträning och i normala prediktioner: i träning hanteras tidsintervallens längd på ett speciellt sätt på grund av matematisk-tekniska detaljer och i normala prediktioner tittar vi på en längre tidsperiod där vi inte vet något om censurering. Här vill vi beräkna sannolikheten före tidsintervallet, utan påverkan av hur det går, med justering för censurering utan positivt utfall eftersom sådan censur omöjliggör ett positivt utfall under den delen av intervallet. Tidsintervallen har delats in i 25 grupper som är lika stora så när som på avrundning, viktat med avseende på tidsintervallens längd. Summan av tidsintervallens längd i varje grupp är alltså lika stora i den mån det är möjligt. Viktningen är främst gjort för att inte 7-dagarsintervallen i början av inskrivningarna ska få större tyngd än deras bidrag till helheten. Medelsannolikheterna i varje grupp är också viktade med tidsintervallens längd, liksom andelarna positiva utfall.

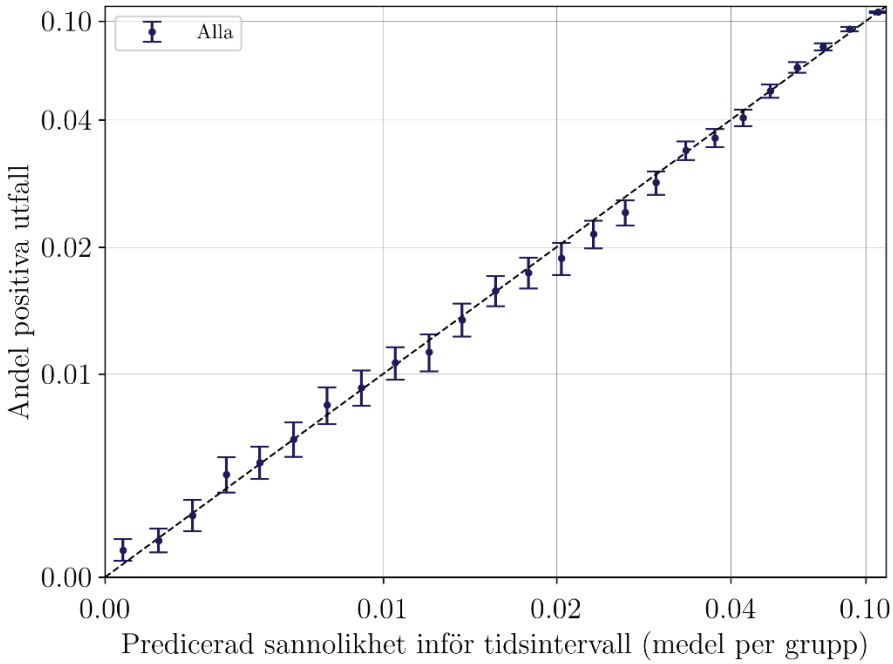
Figur 10. Kalibreringskurva för aktuell målpopulation



Resultaten visar att kalibreringen på totalen i huvudsak är tillfredsställande. Det finns dock statistiskt signifikanta avvikelser från 45-graderslinjen för grupper med medelhög sannolikhet (jobbchans), för vilka värdena ligger något lågt. Storleksmässigt innebär den största av dessa avvikelser att gruppen med predicerad jobbchans på knappt 2,6 procent har en faktisk realiserad jobbchans på 2,2 procent. Dessa avvikelser beror på att modellen är tränad på individer som är inskrivna som tidigast 2015 och sökande med längre inskrivningstider hanteras på ett något grövre sätt, samtidigt som modellen här utvärderas för alla individer. Anledningen till denna restriktion i träningen är bland annat ett byte av SSYK-standard 2014 som skulle göra det utmanande att träna modellen på data med gammal klassificering för att använda till att bedöma sökande med ny yrkesklassificering (se avsnitt 3.4).

Figur 11 visar att kalibreringen förbättras om de som skrivits in före 2015 (motsvarande 11%) tas bort från utvärderingsdata. Denna figur visas här för att illustrera förklaringen ovan till att kalibreringen inte är helt perfekt i figur 10.

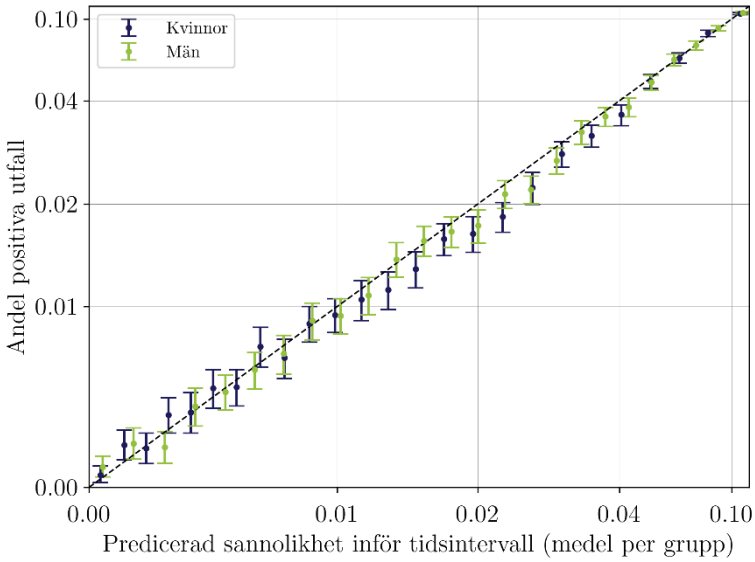
Figur 11. Kalibreringskurva för aktuell målpopulation, minus sökande som skrev in sig före 2015



4.4.2 Resultat: Kalibrering på gruppnivå

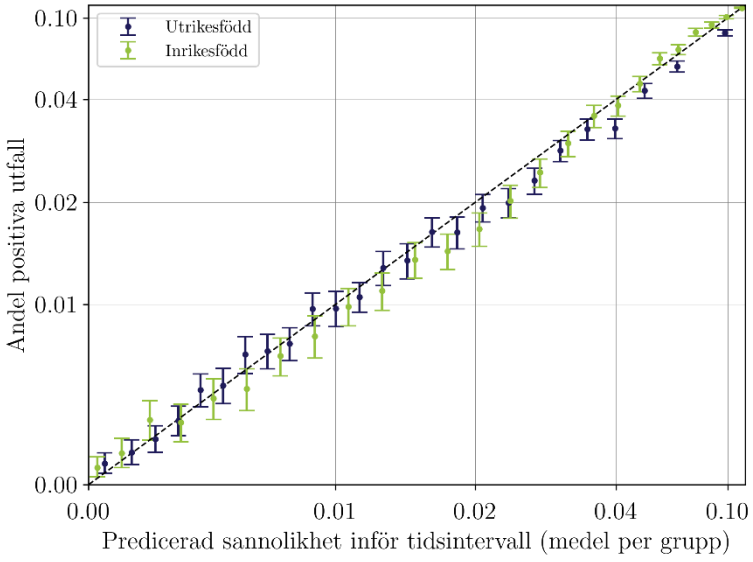
I detta avsnitt studerar vi kalibreringskurvor separat för olika grupper definierade utifrån de observerbara diskrimineringsgrunderna i diskrimineringslagen.

Figur 12. Kalibreringskurvor för kvinnor och män



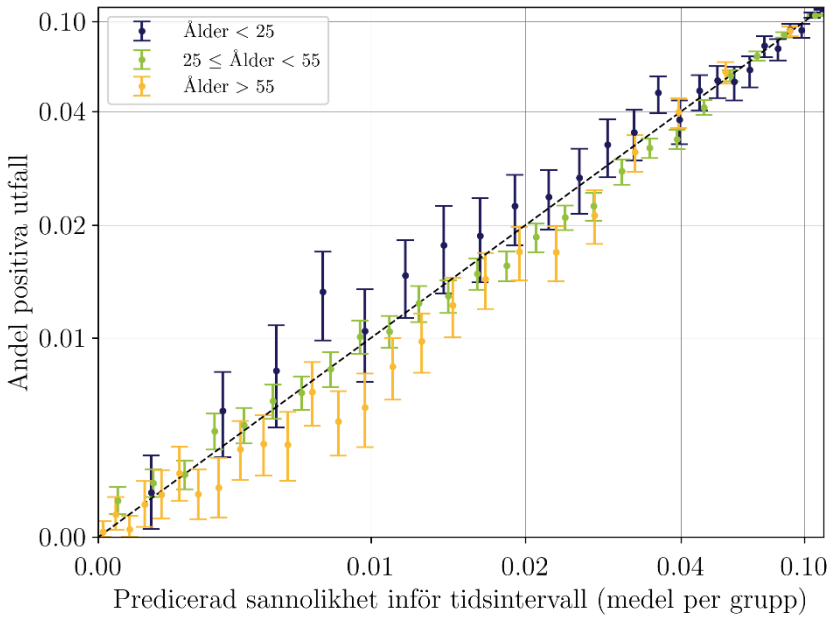
Figur 12 visar att det inte verkar finnas några systematiska skillnader mellan män och kvinnor i hur prediktionerna är kalibrerade mot verkliga utfall. Det vill säga, mönstret i Figur 10 för alla sökande i aktuell målpopulation är väldigt lika även för kvinnor och män separat.

Figur 13. Kalibreringskurvor för utrikes- och inrikesfödda



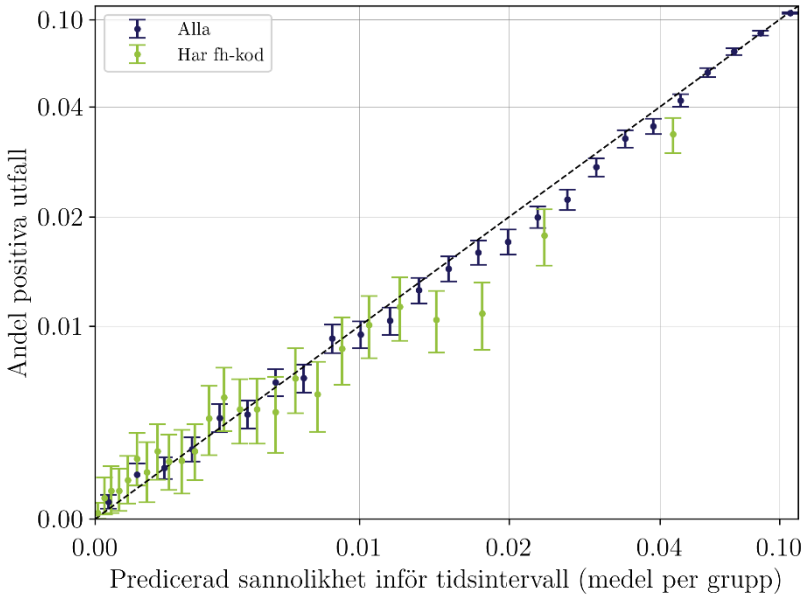
Figur 13 visar motsvarande för utrikes- och inrikes- födda: vi ser inga tydliga systematiska skillnader i kalibrering. Det finns en viss tendens till att de utrikesfödda med allra högst skattade sannolikheter är något överskattade, och tvärtom för de inrikesfödda.

Figur 14. Kalibreringskurvor för olika åldersgrupper



Figur 14 visar resultaten för olika åldersgrupper separat. Här noterar vi vissa skillnader, och särskilt att den äldre gruppens jobbchanser tenderar att vara något överskattade vid låga sannolikheter.

Figur 15. Kalibreringskurvor för sökande med funktionshinderkod och för totalen



Figur 15 visar resultaten separat för sökande med och utan funktionshinderkod. Här finns det vissa skillnader, där jobbchanserna är överskattade vid medelhöga sannolikheter för sökande med funktionshinderkod.

Sammantaget visar resultaten att kalibreringen inte skiljer sig nämnvärt åt mellan män och kvinnor eller för utlandsfödda jämfört med svenskfödda, men att det finns vissa skillnader i en jämförelse av sökande med och utan funktionshinderkod, samt vid en jämförelse mellan sökande i olika åldersgrupper. De förekommande avvikelserna har till största delen att göra med hanteringen av de som har mycket långa inskrivningstider: de med funktionshinderkod och de äldre är överrepresenterade bland de som skrevs in innan 2015, vilka modellen inte har kunnat tränats för att bedöma på ett optimalt sätt. Motsvarande figurer utan de inskrivna före 2015 finns i bilaga 2. De figurerna visar som väntat mindre skillnader mellan de olika grupperna.

4.4.3 Resultat: Träffsäkerhet och rangordningsförmåga

Måtten vi använder är utformade för att utvärdera modeller som används till binär klassificering (undantaget *Concordance*). Bedömningsmodellens syfte är egentligen inte att göra binär klassificering, det är de predicerade jobbchanserna som används i syfte att rangordna arbetssökande efter stödbehov. Men genom att välja ett tröskelvärde med avseende på skattad sannolikhet (jobbchans) kan modellens skattningar grovt delas in i grupperna ”nära” respektive ”långt ifrån” arbetsmarknaden. Vi väljer tröskelvärde så att en lika stor andel bedöms vara ”nära” som den andel som faktiskt påbörjat ett varaktig arbete eller studier inom 365 dagar[63], vilket är 28,7 procent. Tröskelvärdet blir då en sannolikhet (jobbchans) på 42,3 procent. Sökande med en högre bedömd jobbchans än 42,3 procent klassas därmed som ”nära” (alltså 28,7 procent av alla sökande i utvärderingsdata), och de med en lägre bedömd jobbchans än 42,3 klassas som ”långt ifrån” arbetsmarknaden i testerna.[64]

Accuracy

Tabell 3 visar träffsäkerheten i modellen mätt med Accuracy, det vill säga andelen korrekta prediktioner till kategorierna nära eller långt ifrån arbetsmarknaden. Notera igen att måttet Accuracy baseras på ”angreppssätt 1” (se ovan). Vi jämför alltså modellens förmåga att klassificera rätt, baserat på predicerade jobbchanser rensade för effekterna av deltagande i insatser, i förhållande till realiserade utfall som är påverkade av insatser. Vi kan observera att andelen korrekta prediktioner är nära 76 procent, trots att det är en konservativ skattning. Som en jämförelse skulle slumpmässig klassificering ge 59 procent korrekta prediktioner. Notera att detta är en ”kvalificerad slump” som tar hänsyn till att själva prediktionsproblemet är snedfördelat, och att det därmed är enklare att gissa rätt än om det skulle vara lika många sökande som har positiva som negativa utfall. Den kvalificerade

[63] I normalfallet används gränsen 365 dagar – detta varieras dock nedan för att göra bättre jämförelser med studier av andra verktyg.
[64] I dessa resultat är de som varaktigt avregistrerats av okänd orsak inom 15 månader borttagna ur analysen[64], förutom då vi använder Concordance-måttet. Det är 15 månader och inte 12 för att förenkla jämförelser med andra studier genom att variera tiden då utfallet räknas. Det finns också en annan poäng: avregistreringar av orsak 6 sker ofta med en viss fördröjning.

slumpjämförelsen är gjord för att inte överdriva modellens mervärde: det är enklare att träffa rätt när prediktionsproblemet är snedfördelat.

Tabell 3. Accuracy: Andel korrekta prediktioner till kategorierna nära eller långt från arbetsmarknaden.

Denna modell	Slump
75,9 %	**59,1 %**

ROC AUC

Eftersom modellens viktigaste uppgift är att rangordna de sökandes jobbchanser korrekt går vi vidare och utvärderar modellen med mått som fångar rangordningsförmågan. Tabell 4 visar resultat för måttet ROC AUC. Värdet 79,3 procent kan tolkas som att sannolikheten är drygt 79 procent för att en slumpmässigt vald individ med positivt realiserat utfall (jobb/studier==1) rankas högre av modellens klassificering än en slumpmässigt vald individ med jobb/studier==0. I tabellens andra kolumn visas testvärdet med angreppsätt 3, det vill säga att måttet har anpassats till att överlevnadsmodellen rensar för insatser. Vi kan notera att det konservativa testvärdet med angreppsätt 1 underskattar modellens rangordningsförmåga med ca 2 procentenheter.

Tabell 4. ROC AUC: ekvivalent med rangkorrelationen mellan prediktioner (till kategorierna nära eller långt från arbetsmarknaden) och faktiska utfall.

Ej korrigerat (angreppsätt 1)	Korrigerat (angreppsätt 3)	Slump
79,3 %	**81,5 %**	**50 %**

The Confusion Matrix

Tabell 5 visar "the confusion matrix". Vi kan se att prediktionsproblemet är snedfördelat: 71,3 procent (12,0+59,3) får inte arbete eller studier[65]. Matrisens komponenter med andelarna "sanna positiva", "sanna negativa", "falska positiva" och "falska negativa" kan användas till att ta fram flera olika mått på modellens träffsäkerhet. Notera, till exempel, att modellens Accuracy helt enkelt är summan av andelarna sanna positiva och sanna negativa prediktioner. Notera också igen att

[65] Det kan vara värt att påminna sig om definitionen av utfallet som används: avregistrering till arbete eller studier inom 12 månader, som varar i minst 4 månader. De 12 månaderna startar vid profileringstillfället (inte vid inskrivningen).

matrisen är framtagen med angreppssätt 1, vilket innebär att modellens förmåga att klassificera rätt är underskattad.

Tabell 5. The confusion matrix.

	Bedömts nära	Bedömts långt ifrån
Får arbete/studier	16,7 %	12,0 %
Får ej arbete/studier	12,0 %	59,3 %

Concordance
Tabell 6 visar testvärdena för Concordance-måttet. Måttet är anpassat för att utvärdera överlevnadsmodeller och det är justerat för att modellen rensar för insatser, det vill säga att angreppsätt 3 används. Vi finner att sannolikheten är ca 77 procent att längden på två slumpvis valda arbetslöshetsperioder (vars längd går att åtskilja) är korrekt ordnade av modellen. Värdet påverkas inte mycket av huruvida sökande som avregistreras av okänd orsak ingår eller inte.

Tabell 6. Concordance.

Inklusive avors 6	Avors 6 borttaget	Slump
76,9 %	**77,3 %**	**50 %**

4.4.4 Att jämföra kvalitet: Vilka testvärden är bra?

I föregående avsnitt använde vi slumpmässig allokering som referenspunkt till resultaten för träffsäkerhet/rangordningsförmåga. Vi noterar att modellen bidrar avsevärt till förbättrad träffsäkerhet och korrekt rangordning av prediktionerna jämfört med en ”kvalificerad slump”. Den kvalificerade slumpen är anpassad så att modell och slump står inför lika svåra uppgifter, så att inte modellen får ett mekaniskt övertag genom att prediktionsproblemet är snedfördelat. Andelen korrekta prediktioner (Accuracy) är till exempel 16,9 procentenheter (eller 29 procent) högre än vad den kvalificerade slumpen genererar. Slumpmässig allokering ger en slags referenspunkt till att bedöma om testvärdena är bra eller dåliga. I detta avsnitt går vi vidare och gör andra jämförelser. Först jämför vi testresultaten mot motsvarande resultat från föregående utvärdering. Vi jämför sedan våra testresultat mot ”tumregler” som omnämns i litteraturen, samt mot motsvarande resultat i liknande tillämpningar från andra länder.

Jämförelse mot föregående utvärdering
Den nya bedömningsmodellen skiljer sig väsentligt från den gamla. Den största skillnaden är att inskrivningstid ingår i den nya modellen och att modellen är tränad på sökandestocken[66] och inte enbart på nyinskrivna. I den gamla modellen kompenserades bedömningsmodellens brister vad gäller att bedöma andra än nyinskrivna med en regelbaserad lösning i bedömningsverktygets spårindelning. En rättvis jämförelse måste därför vara att jämföra träffsäkerheten i den nya bedömningsmodellen mot det gamla bedömningsstödet i dess helhet, för att inte mekaniskt missgynna den gamla modellen i jämförelsen. Passande nog finns en sådan tidigare utvärdering av det gamla bedömningsstödet i Arbetsförmedlingen (2021). Det gamla verktyget hade en accuracy på 68 %, och med ett snedfördelat prediktionsproblem som innebar att den kvalificerade slumpen genererade en accuracy på 57 %. För att göra prediktionsproblemets svårighetsgrad mer jämförbart med det för det gamla verktyget kan vi förlänga tiden för vilken vi räknar positiva utfall (i normalfallet 365 dagar) till 423 dagar. Denna justering är gjord för att ge en kvalificerad slump-accuracy på 57 %, det vill säga samma som i utvärderingen av den gamla modellen. Då fås ett tröskelvärde på 44,6 %, och en accuracy på 74,9 %. I dessa någorlunda jämförbara situationer fås alltså att träffsäkerheten mätt som accuracy är ca 7 procentenheter (eller ca 10 procent) bättre i den nya modellen än i det gamla verktyget (jobbchans + korrigering för inskrivningstid i spårindelning). Det är en avsevärd förbättring. Förbättringen 10 procent motsvarar till exempel mer än halva förbättringen som det gamla verktyget bidrog med jämfört med slumpmässig allokering (vilket var totalt ca 19 procent).

Jämförelse mot tumregel
Tumregel för ROC AUC enligt Hosmer och Lemeshow (2000):

- AUC < 0.5 — Sämre än slumpen
- AUC = 0.5 — Vad slumpmässig klassificering skulle ge
- 0.7 ≤ AUC < 0.8 — Acceptabel
- 0.8 ≤ AUC < 0.9 — Utmärkt
- AUC ≥ 0.9 — Enastående

Vi kan notera att det konservativa testvärde som vi får med angreppsätt 1 är 0,793, det vill säga i den övre delen av intervallet ”acceptabelt”. Vi vet att detta estimat underskattar modellens rangordningsförmåga. Vi kan vidare notera att testvärdet med angreppsätt 3, det vill säga när måttet har anpassats till att överlevnadsmodellen rensar för insatser, är 0,815, det vill säga i nedre delen av intervallet ”utmärkt”.

[66] En population som ligger ”nära” sökandestocken, på så vis att en stor variation av inskrivningstider ingår och inga andra begränsningar gjorts. Detaljer återfinns i avsnitt 3.4.

Jämförelse med utvärderingar av liknande tillämpningar

I detta avsnitt börjar vi med att jämföra träffsäkerheten på ett övergripande plan mot publicerade siffror som rapporterats från liknande modeller i Sverige och från andra länder. Här ska man vara medveten om att svårighetsgraden i det som modellerna predicerar kan skilja sig åt avsevärt. Det handlar till exempel oftast om att predicera utfallen för nyinskrivna arbetssökande till skillnad från att predicera för alla arbetssökande oavsett inskrivningstid. Och vi vet att måtten är känsliga för hur pass snedfördelade prediktionsproblemen är, något som kan variera stort mellan dessa olika tillämpningar. Jämförelsen ger dock en uppfattning om huruvida testresultaten i vår utvärdering sticker ut åt något håll. Vi kan konstatera att Arbetsförmedlingens nya modell ligger högt i jämförelsen av ROC-AUC och kanske mer åt medelhögt i jämförelsen av Accuracy. Nedan går vi vidare med att göra mer rättvisande jämförelser som bättre tar hänsyn till skillnader i prediktionsproblemens svårighetsgrad, och kommer då till lite andra slutsatser, särskilt vad gäller Accuracy.

Tabell 7. Jämförelser av träffsäkerhet mot andra modeller för att bedöma avstånd till arbetsmarknaden (avrundade siffror i procent)

Land	Accuracy	ROC-AUC
Sverige AF 2023	75,9 (konservativt)	79,3 (konserv.) 81,5 (justerat)
Sverige (AF 2020-2023)	68	-
Sverige (IFAU, 2007)	69	-
UK	-	80
Tyskland	84 - 85	70 – 77
Irland	69 - 86	-
Nya Zeeland	-	63 – 83
Nederländerna	70	-
Belgien	67	76
Österrike	80 - 85	-

Notera: Uppgifterna från andra länder än Sverige finns sammanfattade i Desiere m.fl. (2019)

Vi går vidare med att göra mer rättvisande jämförelser med två modeller i tabellen ovan, modellerna från Tyskland och Irland. Att vi valt just dessa två beror på att nödvändiga uppgifter finns dokumenterade.

- **Tyskland**

Kern m.fl. (2021) har ett mycket skevt prediktionsproblem: 12,8 procent av arbetslöshetsperioderna 2016 är faktiskt långtidsarbetslösa. Man vill med sin modell predicera de 10 procent med högst risk för långtidsarbetslöshet. Om man skulle låta göra dessa prediktioner med hjälp av en kvalificerad slumpmässig allokering skulle man få en Accuracy (andel sanna positiva + andel sanna negativa) på 0,128 × 0,1 + (1 – 0,128) × (1 – 0,1) = 79,8 %. Modellens till synes höga Accuracy på 84 – 85 innebär alltså en förbättring jämfört med slumpen på ca 4 – 5 procentenheter. Samma slump-Accuracy (79,8%) uppnås för vårt prediktionsproblem genom att flytta gränsen för när vi observerar utfall från 365 dagar till 97 dagar. Detta ger i sin tur ett tröskelvärde på 25,8 procents jobbchans. Med detta tröskelvärde följer en skattad Accuracy på 85,4 % för Arbetsförmedlingens nya bedömningsmodell (och okorrigerat och korrigerat ROC-AUC på 79,9 % respektive 80,9 %).

Denna jämförelse visar tydligt att man bör vara försiktig med att jämföra träffsäkerheten rakt av mellan olika länder och modeller när prediktionsproblemets svårighetsgrad varierar stort, och då måtten på träffsäkerhet ofta är känsliga för detta. Jämförelsen visar att den höga rapporterade träffsäkerheten för den tyska modellen på Accuracy = 84 – 85 %, jämfört med Arbetsförmedlingens modells konservativa Accuracy = 76 % inte innebär att den tyska modellen är bättre. Arbetsförmedlingens modell har en Accuracy på 85,4 procent (konservativt räknat), det vill säga i samma nivå som de rapporterade värdena i den tyska studien, när man försöker ta hänsyn till prediktionsproblemets svårighetsgrad. Värdet på ROC-AUC är några procentenheter högre för Arbetsförmedlingens modell.

- **Irland**

I O'Connell m.fl. (2012) varierar svårighetsgraden i de olika prediktionsproblem man tittar på: man skiljer i vissa delar endast på de som bedöms riktigt nära från de som står riktigt långt ifrån, och de högre träffsäkerhetssiffrorna i intervallet Accuracy = 69 - 86 procent är för enklare prediktionsproblem. Accuracy = 69% uppnås i studien i den situation som är mest jämförbar med vår: modellen utvärderas för alla personer. För att beräkna Accuracy som vad en kvalificerad slumpmässig allokering skulle ge kan vi först observera att med studiens tröskelvärde på 50 procent prediceras 38,7 procent av de arbetssökande som långtidsarbetslösa. Och i studien redovisas att 39,0 procent är långtidsarbetslösa när man tittar på faktiska utfall. Detta motsvarar en slump-Accuracy på 52,5 %. Jämfört med slumpen förbättrar den irländska modellen prediktionernas Accuracy med ca 16 procentenheter.

Vi kan inte nå denna slump-Accuracy (52,5 %) med våra data genom att flytta gränsen på 365 dagar, eftersom vi inte kan följa individerna mer än 455 dagar. Det närmaste vi kommer är just 455 dagar, vilket ger tröskelvärdet 45,6 procents jobbchans. Med detta blir slump-Accuracy= 56,0 %. Och med detta tröskelvärde är Arbetsförmedlingens modells Accuracy 74,6 % (och okorrigerat och korrigerat ROC-AUC på 79,2 % respektive 81,7 %). Situationerna går alltså inte på rak arm att göra jämförbara, men det ser ut som att Arbetsförmedlingens modell har en högre Accuracy än den irländska modellens 69% när man försöker ta hänsyn till prediktionsproblemets svårighetsgrad.

4.4.5 Modell tränad under tidigare tidsperiod

Resultaten ovan gäller den modell som är i användning sedan 17 april 2023, vilken har tränats på data från en tidsperiod som innehåller tidsperioden för utvärdering (men inte samma individer). I praktiken används ju modellen för prediktioner för en tidsperiod som ligger *efter* tidsperioden för träning, vilket gör att det också är intressant att titta på resultat för en modell som är tränad på data enbart från en tidsperiod som ligger före utvärderingsperioden. I bilaga 3 återfinns sådana resultat. Dessa resultat är väldigt lika resultaten ovan: på marginalen är prestationen i allmänhet något sämre, framför allt vad gäller kalibreringskurvorna.

4.5 Slutsatser från genomförd utvärdering

Vi finner att den utvärderade modellen uppvisar en hög träffsäkerhet/rangordningsförmåga. Vi finner också att modellen är välkalibrerad, men att det förekommer mindre avvikelser från perfekt kalibrering. Avvikelserna från perfekt kalibrering som vi observerar beror på en begränsning i möjligheten att träna modellen på arbetssökande med riktigt långa inskrivningstider: Modellen är tränad på individer som är inskrivna som tidigast 2015, och inskrivningstiden för de som har för lång inskrivningstid sätts i nuläget till den längsta tillåtna. En modell optimerad även för sökande med över sju års inskrivningstid hade alltså behövt träningsdata längre tillbaka i tiden än 2015. Anledningen till denna restriktion i träningen är framför allt ett byte av SSYK-standard 2014 som skulle göra det utmanande att träna modellen på data med gammal klassificering för att använda till att bedöma sökande med ny yrkesklassificering. På längre sikt finns det förbättringspotential vad gäller kalibrering för grupper med mycket långa inskrivningstider. Detta kan vara relevant om bedömningsmodellen kommer att användas för att även bedöma och rangordna stödbehovet sinsemellan bland sökande långt från arbetsmarknaden, vilket dessa oftast är.

Referenser

Arbetsförmedlingen (2021a). *En undersökning av förutsättningarna för statistiska bedömningar av avstånd till arbetsmarknaden, med fokus på betydelsen av inskrivningstid*. Dnr Af-2020/0046 8022.

Arbetsförmedlingen (2021b). *Träffsäkerhet och likabehandling vid automatiserade anvisningar inom Rusta och matcha – En kvalitetsgranskning*. Dnr Af-2020/0046 7913.

Arbetsförmedlingen (2021c). *Träffsäkerhet i bedömningen av arbetssökande. En jämförelse av arbetsförmedlare och en statistisk modell*. Dnr Af-2020/0046 7620.

Arbetsförmedlingen (2022a). *Internrevisionsrapport – Arbetsmarknadspolitisk bedömning*. Dnr Af-2021/0077 2672.

Arbetsförmedlingen (2022b). *Nya metoder för att hantera avaktualiseringsorsak*. Dnr Af-2021/0049 6067.

Arbetsförmedlingen (2022c). *VUM Beslutsstöd – Vilka personuppgifter får behandlas i modellträning*. Dnr *Af-2023/0057 9220*.

Arbetsförmedlingen (2022d). *Betygsmodellen i Rusta och matcha*. Dnr Af-2022/0006 9925.

Arbetsförmedlingen (2023). *Effekter av tre arbetsmarknadspolitiska program 2010–2020*. Dnr Af-2023/0003 2435.

Alm, S. och Britton, T. (2008). *Stokastik*. Liber.

Angrist, J.D. och Pischke, J-S. (2009). *Mostly Harmless Econometrics.* Princeton University Press.

Bennmarker, H., Carling, K. och Forslund, A (2007). *Vem blir arbetslös?* IFAU rapport 2007:20.

Desière, S., Langenbucher, K., and Struyven, L. (2019). *Statistical profiling in public employment services*. OECD Social, Employment and Migration Working Papers, No. 224.

Dieterich, William, Christina Mendoza and Tim Brennan (2016), *COMPAS risk scales: Demonstrating accuracy equity and predictive parity*. Technical report, Northpointe, July 2016.

IFAU (2021), Rusta och Matcha – erfarenheter från en ny matchningstjänst med fristående leverantörer inom arbetsmarknadspolitiken. Rapport 2021:7.

Hosmer, D. W., Jr., and S. Lemeshow (2000), *Applied Logistic Regression*. 2nd ed. New York: Wiley.

Kern, Christoph, Ruben L. Bach, Hannah Mautner and Frauke Kreuter (2021), *Fairness in Algorithmic Profiling: A German Case Study*, arXiv:2108.04134v1 [cs.CY] 4 Aug 2021.

Kleinberg, Jon, Sendhil Mullainathan and Manish Raghavan (2016), *Inherent Trade-Offs in the Fair Determination of Risk Scores*, arXiv:1609.05807.

Kremers, Walter, K. (2007), *Concordance for Survival Time Data: Fixed and Time-Dependent Covariates and Possible Ties in Predictor and Time*. Mayo Foundation Technical Report Series #80.

Noriega-Campero, Alejandro, Michiel A. Bakker, Bernardo Garcia-Bulle och Alex Pentland (2018), *Active Fairness in Algorithmic Decision Making*, arXiv:1810.00031v2.

O'Connell, P.J., McGuiness S. & Kelly E. (2012). *The Transition from Short- to Long-Term Unemployment: A Statistical Profiling Model for Ireland*. The Economic and Social Review, 43(1), 135–164.

Rodríguez, G. (2007). *Lecture Notes on Generalized Linear Models*. URL: https://data.princeton.edu/wws509/notes/.

Salganik med flera (2020). *Measuring the probability of life outcomes with a scientific mass collaboration*. Proceedings of the National Academy of Sciences 117(15), 8398-8403.

Bilagor

Bilaga 1 – Härledning av kovarians för medel-baseline-hasard

I avsnitt 2.7 utnyttjas att $\text{cov}(\log \bar{\boldsymbol{\lambda}}_{\text{baseline}}) \approx T cov(\widehat{\boldsymbol{\beta}}) T'$, vilket härleds i detta avsnitt.

Till att börja med kan medel-baseline-hasarden uttryckas som

$$\bar{\boldsymbol{\lambda}}_{\text{baseline}} = \sum_g a_g \boldsymbol{\lambda}_{g,\text{baseline}} = \sum_g a_g \exp(\hat{\beta}_{\mathbf{0}}) \exp(T_g \widehat{\boldsymbol{\beta}} - \mathbf{1}\hat{\beta}_0),$$

där a_g är andelen av grupp g i medelvärdesbildningen, $\boldsymbol{\lambda}_{g,\text{baseline}}$ är baseline-hasarden för grupp g, $\hat{\beta}_0$ är interceptet, T_g är matrisen som uppfyller $\boldsymbol{\lambda}_{g,\text{baseline}} = \exp(T_g \widehat{\boldsymbol{\beta}})$, och $\mathbf{1}$ är en vektor med 1:or med samma längd som $\boldsymbol{\lambda}_{g,\text{baseline}}$. Eftersom interceptet är gemensamt för alla grupper kan $\exp(\hat{\beta}_{\mathbf{0}})$ brytas ut ur summan:

$$\bar{\boldsymbol{\lambda}}_{\text{baseline}} = \exp(\hat{\beta}_{\mathbf{0}}) \sum_g a_g \exp(T_g \widehat{\boldsymbol{\beta}} - \mathbf{1}\hat{\beta}_0).$$

Definiera vidare vektorn $\mathbf{Y}_g = T_g \widehat{\boldsymbol{\beta}} - \mathbf{1}\hat{\beta}_0$. $\mathbf{Y}_g$ är någorlunda nära 0, så att första två termerna i Taylorutvecklingen av exponentialfunktionen runt 0 kan användas (detta brukar man vanligtvis göra i propagering av kovarianser genom en ickelinjär funktion (Alm och Britton, 2008)). Då fås

$$\bar{\boldsymbol{\lambda}}_{\text{baseline}} \approx \exp(\hat{\beta}_{\mathbf{0}}) \sum_g a_g (\mathbf{1} + \mathbf{Y}_g) = \exp(\hat{\beta}_{\mathbf{0}}) \left(\mathbf{1} + \sum_g a_g \mathbf{Y}_g \right),$$

där det sista steget utnyttjar att $\sum_g a_g = 1$. Logaritmering ger

$$\log \bar{\boldsymbol{\lambda}}_{\text{baseline}} \approx \hat{\beta}_{\mathbf{0}} + \log \left(\mathbf{1} + \sum_g a_g \mathbf{Y}_g \right) \approx \hat{\beta}_{\mathbf{0}} + \sum_g a_g \mathbf{Y}_g,$$

där första två termerna i Taylorutvecklingen av $\log(1 + x)$ runt $x = 0$ används (konstanten är 0). Interceptet kan nu återinföras innanför summan, på motsatt sätt som tidigare, vilket ger

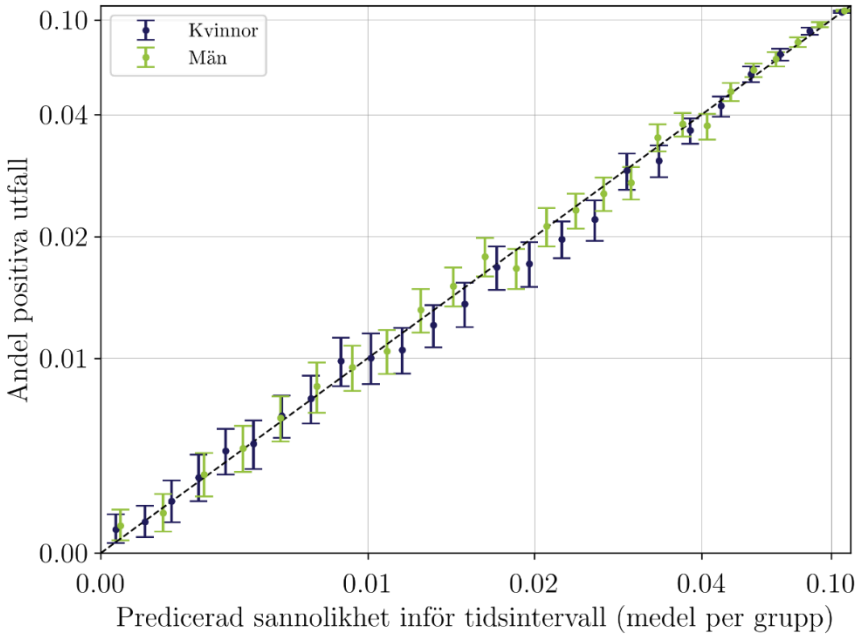
$$\log \bar{\boldsymbol{\lambda}}_{\text{baseline}} \approx \sum_g a_g (\mathbf{1}\hat{\beta}_0 + \mathbf{Y}_g) = \sum_g a_g T_g \widehat{\boldsymbol{\beta}} = \left(\sum_g a_g T_g \right) \widehat{\boldsymbol{\beta}} = T \widehat{\boldsymbol{\beta}},$$

där näst sista steget utnyttjar att $\widehat{\boldsymbol{\beta}}$ inte beror på g och att matrismultiplikation är en linjär operation så att linjärkombinationen av matrisprodukterna $T_g \widehat{\boldsymbol{\beta}}$ kan uttryckas som motsvarande linjärkombination av matriserna T_g multiplicerat med vektorn $\widehat{\boldsymbol{\beta}}$. Med vanliga regler för kovariansen av en matrisprodukt fås slutligen

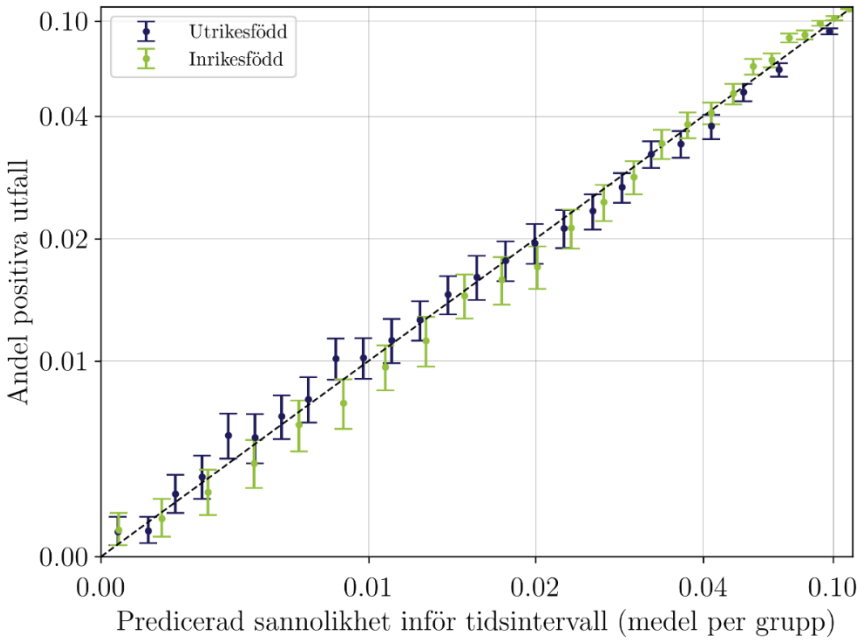
$$\text{cov}(T \widehat{\boldsymbol{\beta}}) = T cov(\widehat{\boldsymbol{\beta}}) T'.$$

Bilaga 2 – Kalibrering på gruppnivå där sökande med mycket långa inskrivningstider (inskrivna innan 2015) har tagits bort från utvärderingsdata.

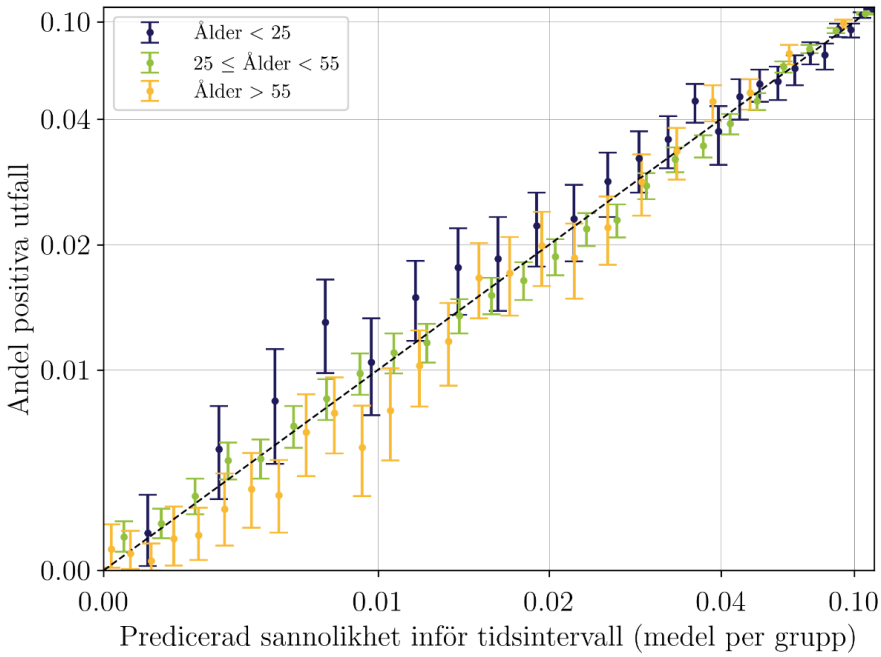
Figur 16. Separata kalibreringskurvor för kvinnor och män för den del av utvärderingspopulationen som är inskrivna sedan 2015



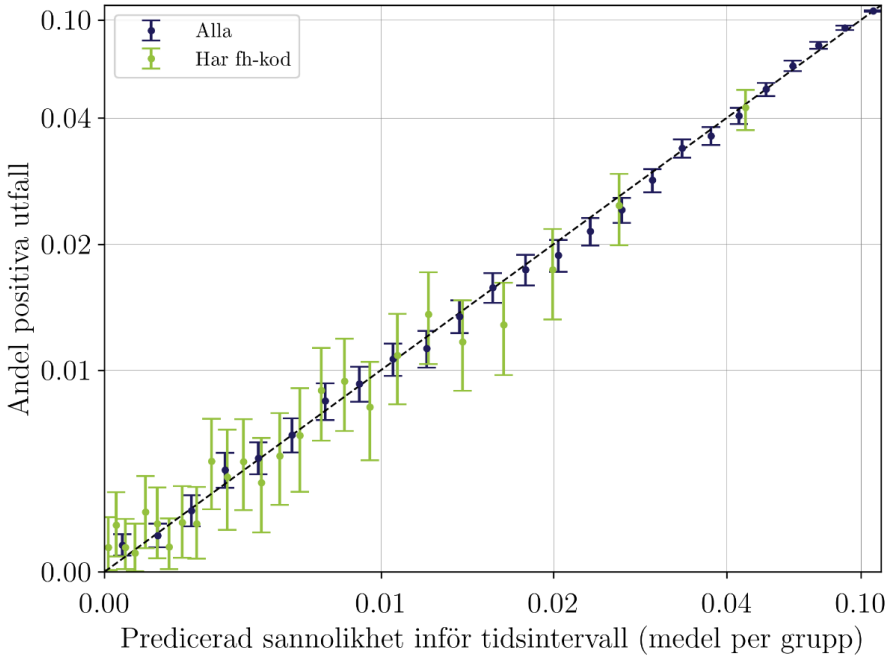
Figur 17. Separata kalibreringskurvor för inrikes- och utrikesfödda för den del av utvärderingspopulationen som är inskrivna sedan 2015



Figur 18. Separata kalibreringskurvor för olika åldersgrupper för den del av utvärderingspopulationen som är inskrivna sedan 2015



Figur 19. Separata kalibreringskurvor för sökande med funktionshinderkod och totalen, för den del av utvärderingspopulationen som är inskrivna sedan 2015

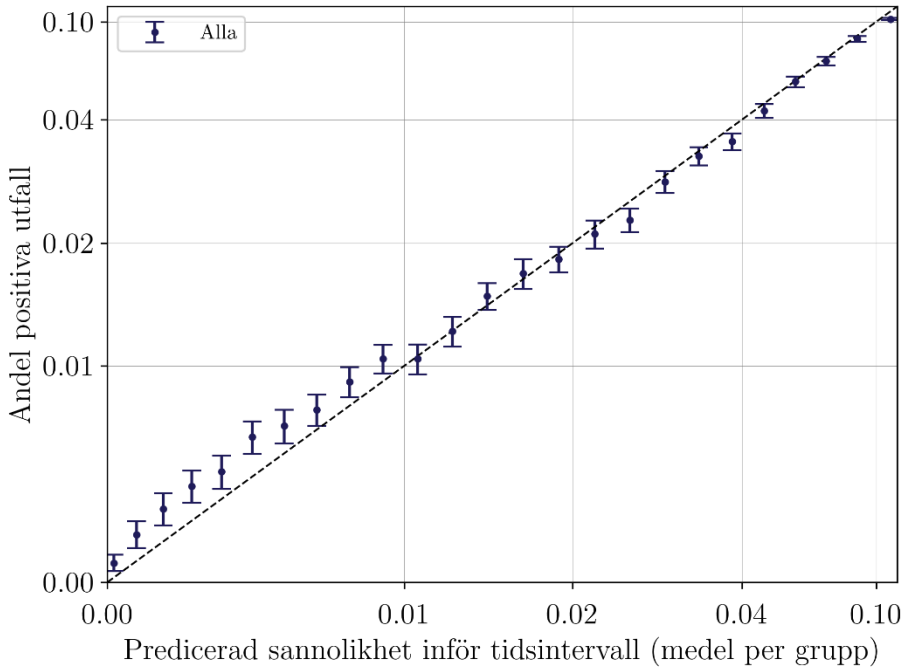


Bilaga 3 – Resultat för modell tränad på tidigare tidsperiod

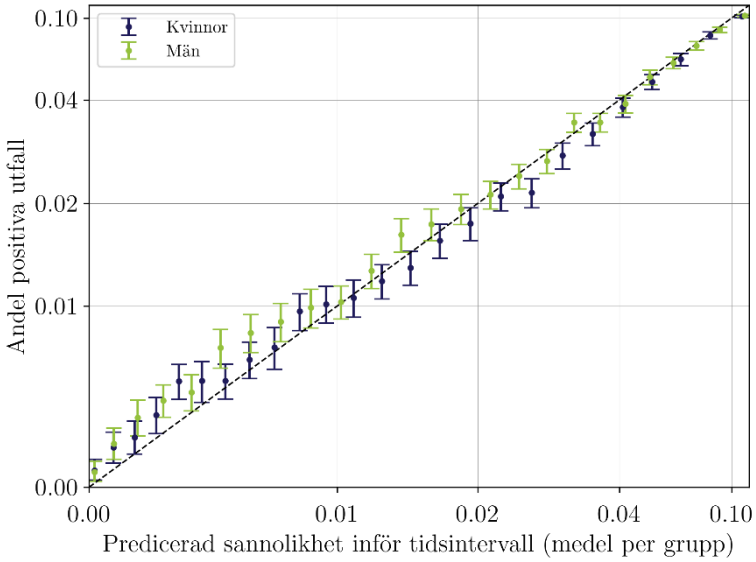
I denna bilaga återfinns resultat motsvarande de i avsnitt 4.4, men för en modell tränad på inskrivna 2015-01-01 till 2020-02-29, och med censurerad data efter 2021-02-29. Utvärderingspopulationen är densamma som i avsnitt, det vill säga ett stort slumpmässigt urval av aktuella sökandekategorier i sökandestocken 2021-03-01.

Kalibrering

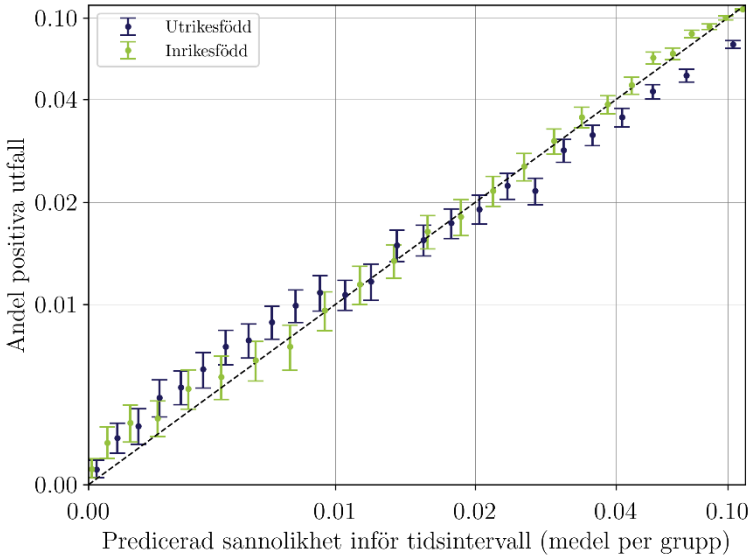
Figur 20. Kalibreringskurva för hela aktuella målpopulationen, för modell tränad på tidigare tidsperiod.



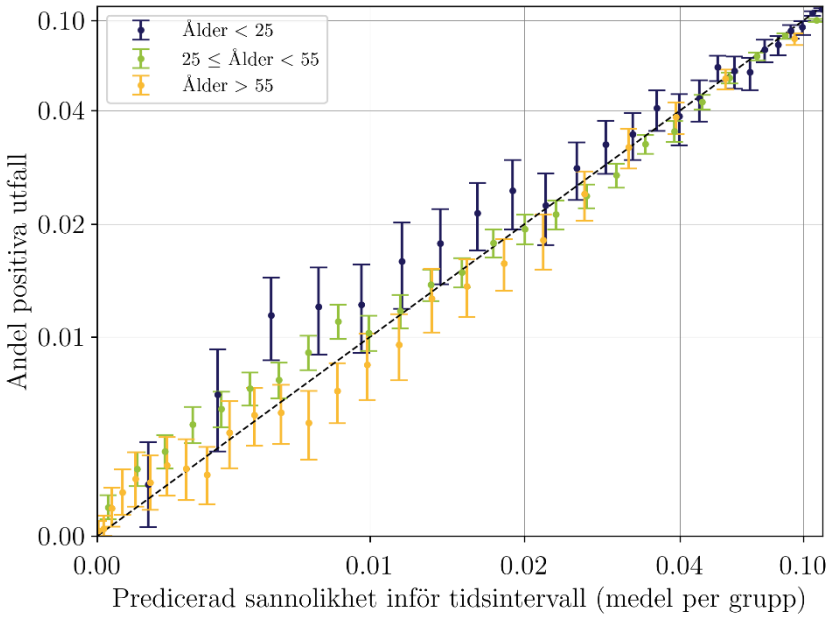
Figur 21. Kalibreringskurvor för kvinnor och män, för modell tränad på tidigare tidsperiod.



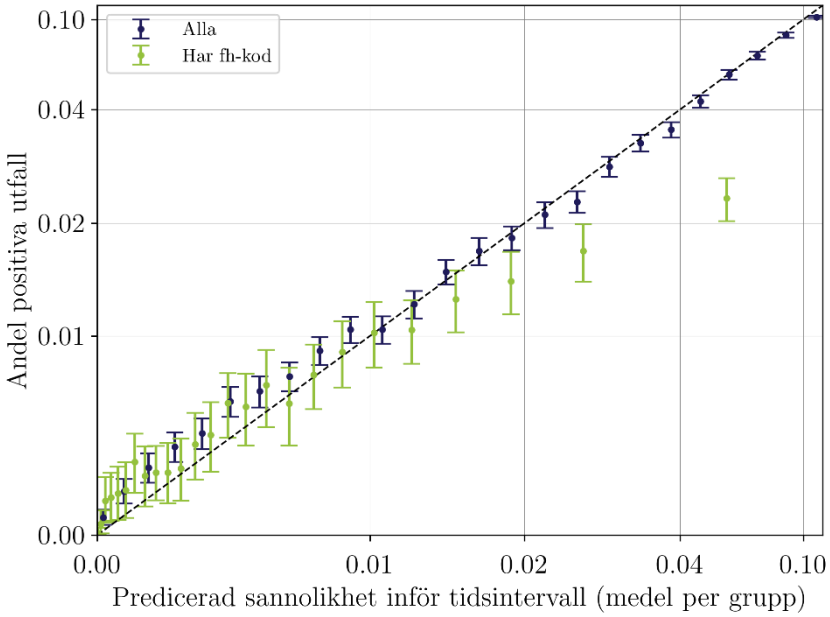
Figur 22. Kalibreringskurvor för utrikes- och inrikesfödda, för modell tränad på tidigare tidsperiod.



Figur 23. Kalibreringskurvor för olika åldersgrupper, för modell tränad på tidigare tidsperiod.



Figur 24. Kalibreringskurvor för sökande med funktionshinderkod och för totalen, för modell tränad på tidigare tidsperiod.



Sammanfattande träffsäkerhetsmått för modell tränad på tidigare tidsperiod

Tabell 7. Sammanfattande träffsäkerhetsmått för modell tränad på tidigare tidsperiod. För accuracy och ROC-AUC avses klassificering med gränsen 365 dagar. Resultat för modellen tränad på hela tidsperioden inom parentes.

	Ej korrigerat	Korrigerat	Slump
Accuracy	75,7 % (75,9 %)	-	59,1 %
ROC-AUC	79,2 % (79,3 %)	80,9 % (81,5 %)	50 %
Concordance inkl avors 6	-	76,4 % (76,9 %)	50 %
Concordance exkl avors 6	-	76,9 % (77,3 %)	50%