Stockholms universitet

Logistisk regressionsmodell - Prediktion av andel individer som har ett privat pensionssparande

Frida Öjemark

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2016:2
Matematisk statistik
Juni 2016

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholms universitet

Matematisk statistik
Stockholms universitet
Kandidatuppsats **2016:2**
http://www.math.su.se/matstat

Logistisk regressionsmodell - Prediktion av andel individer som har ett privat pensionssparande

Frida Öjemark*

Juni 2016

Sammanfattning

I denna rapport anpassas en multipel logistisk regressionsmodell som ska användas för att prediktera sannolikheten att en individ har ett privat pensionssparande. Modellen anpassas på ett urval av Min Pensions användare som loggat in och gjort pensionsprognos på Minpension.se under år 2015. De variabler vi har tillgång till per individ är bland annat kön, månadslön, yrkessektor, ålder samt civilstånd. Den slutliga modellen som väljs, med hjälp av metoden purposeful selection, består av samtliga av de ovanstående variablerna som huvudeffekter samt samtliga tvåfaktorsamspel mellan dessa, förutom de två samspelen mellan kön och månadslön samt mellan kön och civilstånd som visar sig vara icke signifikanta. Vi har tillgång till ett mycket stort antal observationer, närmare bestämt 569 270 individer, vilket gör att mycket små skillnader kommer att bli statistiskt signifikanta. Det visar sig att vi behöver logaritmera lönen samt kategorisera ålder för att modellförutsättningarna ska vara uppfyllda, men trots det visar sig modellanpassningen vara dålig. Vi prövar därför att lägga till flera kombinationer av de förklarande variablerna, men modellanpassningen blir inte bättre. Förmodligen behöver vi även tillgång till andra förklarande variabler. Modellen bekräftar ändå att kvinnor i större utsträckning än män har ett privat pensionssparande. Modellen visar också att en hög lön har stor betydelse för en hög sannolikheten att ha ett privat pensionssparande samt att ogifta i mindre utsträckning än gifta sparar privat till sin pension. Den yrkeskategori som i störst utsträckning har ett privat pensionssparande enligt modellen är de privatanställda tjänstemännen. Detta om vi bortser från de egna företagarna, som i låga löner i betydligt större utsträckning än övriga yrkeskategorier har ett privat pensionssparande. Gifta, egna företagare utmärker sig också genom att lönen inte påverkar förekomsten av att ha ett privat pensionssparande. Däremot för ogifta egna företagare kan vi se att sannolikheten minskar något i de lägre lönerna. Trots den dåliga modellanpassningen prövar vi att prediktera förekomsten av ett privat pensionssparande på ett externt dataset, med resultatet att prediktionsförmågan inte är särskilt bra.

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige.
E-post: frida.ojemark@gmail.com. Handledare: Gudrun Brattström, Jan-Olov Persson.

Abstract

In this paper a multiple logistic regression model is applied to predict the likelihood that an individual has a private pension fund. The dataset used to apply the model comes from 'Min Pensions' users who have logged into the *www.minpension.se* website and have completed a pension prognosis at any time during the year 2015. The variables that are available for each individual are: *gender*, *salary*, *sector of occupation*, *age* and *marital status*. The final model, which is chosen using the *purposeful selection method*, consists of all variables mentioned above as *main effects* and all *two factor interactions* between them, with the exceptions of the two interactions between gender and salary and between gender and marital status which are not statistically significant. We have a large sample size, more precisely 569 270 individuals, which means that even very small differences will be statistically significant. To meet the model assumptions we must use a logarithmic scale for salaries and treat age as a categorical variable, but despite these changes the model fit is poor. We then tried adding more combinations of explanatory variables to the model, but the model fit did not improve. It is most likely that we need access to other explanatory variables. Despite the poor fit the model still confirms that women are more likely to have a private pension fund than men. The model also shows that a higher salary is an important factor in the occurrence of having a private pension fund and that non married individuals are less likely to save privately to their pension. The sector of occupation where most individuals have a private pension fund, according to the model, is non-manual labour workers in the private sector if we disregard self-employed workers that have low incomes who are much more likely to have a private pension fund than all other workers. There is also a difference between married and non-married self-employed workers in the sense that salary has no effect among married self-employed workers while in low salaries the probability goes down for non-married individuals. Despite the poor model fit we try to predict the presence of a private pension fund in an external dataset, with the result that the model predicts poorly.

Förord

Denna uppsats utgör ett självständigt arbete om 15 hp och leder till en kandidatexamen i matematisk statistik. Jag vill rikta ett stort tack till mina handledare Gudrun Brattström, Docent och Universitetslektor vid Matematiska institutionen på Stockholms universitet, och Jan-Olov Persson, 1:e forskningsingenjör vid Matematiska institutionen på Stockholms universitet, för all hjälp och vägledning under arbetets gång. Jag vill även tacka alla hjälpsamma kollegor på Min Pension och Svensk Försäkring för att ni delat med er av er gedigna kunskap och erfarenhet.

Innehållsförteckning

1 Inledning

Pensionen i det svenska pensionssystemet kommer i huvudsak från tre olika delar och utgörs av *allmän pension, tjänstepension* samt *privat pensionssparande* (Pensionsmyndigheten, 2016). Pensionsmyndigheten administrerar och betalar ut den *allmänna pensionen* som är den del av pensionen man tjänar in till genom att arbeta och betala skatt, men även andra typer av inkomster, som till exempel föräldrapenning eller arbetslöshetsersättning ger rätt till allmän pension. 1999 trädde ett nytt reformerat pensionssystem i kraft och ersatte det gamla ATP systemet. Inom ATP systemet beräknas pensionen bland annat efter en individs femton bästa inkomstår. I det nya systemet är det däremot hela yrkeslivet som räknas och längre tid av deltidsarbete eller ett sent inträde på arbetsmarknaden slår hårt mot den egna pensionen. Här är det främst kvinnor som halkar efter då de tenderar att arbeta mer deltid och ha lägre löner, men även andra grupper som till exempel sent invandrade till Sverige samt egenföretagare som inte tar ut lön riskerar att få en låg allmän pension (Petterson, 2013). För de som faller allra sämst ut finns dock skyddsnät i form av till exempel garantipension (Pensionsmyndigheten, 2016). Ytterligare faktorer som påverkar hur stor den allmänna pensionen kommer att bli är den ekonomiska utvecklingen i samhället samt den ökande genomsnittliga livslängden.

Tjänstepensionen är en annan del av pensionen som arbetsgivarna betalar in till sina arbetstagare (Pensionsmyndigheten, 2016). Inom tjänstepensionen förekommer så kallade *premiebestämda* och *förmånsbestämda* försäkringar. De premiebestämda försäkringarna fungerar så att arbetsgivaren sätter av ett visst belopp motsvarande en viss procent (vanligen 4,5 procent) av den *pensionsgrundande lönen* till en tjänstepensionsförsäkring. Precis som i det reformerade pensionssystemet är det livsinkomsten som räknas här. Hur stor den framtida pensionen kommer att bli påverkas också av den ekonomiska utvecklingen i samhället och man brukar säga att risken ligger på den försäkrade. Inom förmånsbestämd tjänstepension är det istället så att arbetstagaren är garanterad ett visst belopp vid pension. Beloppet bestäms på lite olika sätt beroende på vilket avtalsområde inom tjänstepensionen individen tillhör, men principen är att beloppet beräknas som en viss procent av ett genomsnitt av de bästa inkomsterna åren innan pension samt att hänsyn tas till antalet månader som individen har arbetat med tjänstepension. På grund av att beloppet är garanterat är det arbetsgivaren som bär risken eftersom premien som arbetsgivaren betalar in till arbetstagarens försäkring justeras så att förmånen uppfylls även om den ekonomiska utvecklingen är svag. Trenden i samhället är att gå från förmånsbestämda till premiebestämda system och många yngre saknar helt förmånsbestämd pension. Den största andelen av den totala pensionen kommer för de allra flesta från den allmänna pensionen och därefter från tjänstepensionen, men variationen är oerhört stor mellan individer och det finns inga generella tumregler för hur stor pensionen kommer att bli.

Pensionen kan också komma från ett *privat pensionssparande.* Denna del av pensionen utgör oftast en mycket liten del av den totala pensionen och långt från alla har ett privat pensionssparande (Pensionsmyndigheten, 2016). För vissa grupper är dock det privata pensionssparandet mycket viktigt, till exempel om tjänstepension saknas eller om man känner att pensionen inte kommer att räcka till. För många kvinnor är det privata pensions-sparandet ett sätt att få upp pensionen. Från 2016 slopas avdragsrätten för privat pensionssparande för alla utom egna företagare och de som saknar tjänstepension. Det innebär att de som ändå fortsätter att spara privat inte får göra avdrag för det i sin deklaration och kommer därmed att bli dubbelbeskattade i och med att skatt även betalas när pengarna börjar betalas ut. Ett privat pensionssparande lönar sig alltså inte längre för de flesta och nu gäller det att hitta andra sätt att spara till sin pension på.

Minpension.se är en neutral och oberoende pensionsportal där alla som tjänat in till pension i Sverige kan se hela sin pension och göra pensionsprognoser (Min Pension, 2016). Verksamheten drivs och finansieras till hälften av staten och till hälften av pensionsbolagen. Ett av de främsta syftena med pensionsportalen är att ge enkel, individuell och samlad pensionsinformation om den egna pensionen samt att upplysa om pensionssystemet i stort. Förutom att göra tillförlitliga pensionsprognoser kan individen se så gott som hela sitt intjänande till sin pension då alla stora pensionsaktörer är anslutna och levererar pensionsuppgifter till Min Pension. Min Pension har sedan starten 2004 expanderat kraftigt för att vid årsskiftet 2015/2016 ha drygt 2,5 miljoner registrerade användare och under 2015 gjordes drygt 6,7 miljoner pensionsprognoser.

I den här studien kommer vi att ta ett urval på cirka 570 000 individer av Min Pensions användare som gjort minst en pensionsprognos under år 2015 och anpassa en multipel logistisk regressionsmodell som bland annat används för att prediktera sannolikheten att en individ har ett privat pensionssparande. De förklarande variablerna i modellen är kön, yrkessektor, ålder, civilstånd och lön. Vi kan på så sätt belysa eventuella skillnader mellan till exempel kvinnor och män samt mellan olika yrkeskategorier. Anledningen till att vi gör en prediktionsmodell är först och främst i lärande syfte till denna uppsats, men eftersom modellen bygger på Min Pensions data är resultaten direkt applicerbar på deras bestånd.

1.1 Bakgrund

Före 1976 rådde obegränsad avdragsrätt för privat pensionssparande vilket innebar att man fick dra av hela sitt privata pensionssparande i sin deklaration (Bergström, Palme, Persson, 2010). Sedan dess har avdragsrätten succesivt begränsats. År 2015 minskade avdragsrätten till högst 1 800 kronor per år för att år 2016 tas bort helt (Pensionsmyndigheten, 2016).

Svensk Försäkring har kritiserat regeringens förslag om begränsad avdragsrätt och pekat på att det saknas en analys av medborgarnas behov av ett långsiktigt sparande och hur det framöver skulle kunna stimuleras (Svensk Försäkring, 2015).

Enligt statistik från SCB (2014) har kvinnor i större utsträckning är män ett *aktivt privat pensionssparande*. 46,4 procent av kvinnorna i åldern 55-64 år och 35,3 procent av männen i samma åldersintervall har ett aktivt privat pensionssparande. Motsvarande siffror för kvinnor och män i åldersintervallet 45-54 år är 49,8 respektive 40,6 procent. Med aktivt privat pensionssparande menar vi att individen under år 2014 gjort inbetalningar till ett privat pensionssparande som de gjort avdrag för i sin deklaration.

1.2 Syfte och frågeställning

Syftet med studien är att hitta en modell som kan användas för att prediktera i vilken utsträckning olika grupper av individer har ett privat pensionssparande. Främst studeras om det är så att en större andel kvinnor än män har ett privat pensionssparande, men även om det råder skillnader beroende på lön, yrkessektor och civilstånd.

Eftersom gruppen *egen företagare* inte har tjänstepension och därmed i större utsträckning än övriga yrkeskategorier är i behov av ett privat pensionssparande, vill vi även visa hur denna grupp utmärker sig.

2 Metod

2.1 Urval och avgränsningar

Urvalet till denna studie består av Min Pensions användare. Uppgifterna på individnivå som vi har tillgång till kommer dels via pensionsaktörernas leveranser av pensionsuppgifter till Min Pension, dels genom att individen själv kan registrera vissa personuppgifter enligt följande. När en individ för första gången loggar in på Minpension.se sker en automatisk inhämtning av pensionsuppgifter från Pensionsmyndigheten, Försäkringskassan och samtliga anslutna pensionsbolag (Min Pension, 2016). Från pensionsbolagen levereras uppgifter om *nuvarande lön*, *aktivt avtalsområde inom tjänstepensionen* (det vill säga yrkessektor), *civilstånd*, *ålder*, *kön* samt *hittills intjänat pensionskapital och förmåner*. Individen kan sedan välja att bekräfta uppgifterna eller ändra vissa av dem. Till exempel kan individen välja att ändra sin lön och avtalstillhörighet. Anledningen till att individen vill ändra sina uppgifter kan vara att individen nyligen ändrat sin lön, bytt arbete eller bara vill simulera en pensionsprognos för att se vad som händer om dessa parametrar ändras. Uppgifter om en individs lön och yrkeskategori är alltså antingen de som levererats från ett försäkringsbolag eller de som individen själv angett.

Totalt under år 2015 har drygt 6,7 miljoner pensionsprognoser gjorts på Minpension.se. En och samma individ kan dock ha gjort flera prognoser. Min Pension får således tillgång till ett omfattande datamaterial till underlag för att bland annat göra analyser i pensionssystemet. Urvalet i denna studie utgörs av personer

- födda 1951 eller senare
- med en månadslön mellan 5 000 till 200 000 kronor
- som gjort pensionsprognos på Min Pension minst en gång under år 2015.

Detta urval består av cirka 570 000 unika individer och presenteras mer ingående i Avsnitt 3.1. Anledningen till att vi inte studerar samtliga individer som gjort pensionsprognos under år 2015 är att vi utgår från samma urval som Min Pension tidigare gjort i en studie med annat syfte. Där uteslöts individer med allt för osäkra prognoser, som redan börjat ta ut pension, som var folkbokförda utomlands med mera. Vi har ingen möjlighet att göra ett nytt urval specifikt för denna rapport.

Även om urvalet är stort kan vi inte dra slutsatser till hela Sveriges befolkning. Detta då det finns en risk att Min Pensions användare skiljer sig från Sveriges befolkning som helhet (K. Kamp, personlig information, 29 februari 2016). Det är till exempel känt sedan tidigare att Min Pensions användare har högre löner och i större utsträckning är privatanställda tjänstemän eller statligt anställda än i Sverige generellt. Detta gäller främst i de yngre ålderskategorierna. Det kan också förekomma annan typ av snedvridning, till exempel att individer som är mer ekonomiskt medvetna i större utsträckning gör pensionsprognoser. Denna snedvridning är större i yngre åldrar och kan leda till att andelen med privat pensionssparande överskattas kraftigt i dessa grupper. Det är också så att Min Pension av naturliga skäl har en kraftig överrepresentation av individer som närmar sig pensionsåldern. Vi kommer, på grund av begränsade resurser, inte att fördjupa oss ytterligare i dessa frågor.

2.2 Våra variabler

Den levererade uppgiften om *hittills intjänat pensionskapital* enligt Avsnitt 2.1 är den som används för att avgöra om en individ har ett privat pensionssparande eller inte. Här får vi uppgifter per individ om storleken på kapitalet från det privata pensionssparandet. Däremot kan vi inte veta om sparandet är aktivt för närvarande eller inte. Att en individ har ett privat pensionssparande innebär alltså att individen har ett kapital från ett privat pensionssparande, oavsett om sparandet är aktivt eller inte. Sannolikheten att en individ klassificeras till att ha ett privat pensionssparande kommer därmed naturligt att öka med stigande ålder om vi följer en och samma individ över tid. Däremot kan det vara så att benägenheten att spara privat till sin pension kan se olika ut i olika generationer. Om det är så kan vi inte se effekten av att sannolikheten ökar med ålder eftersom vi endast har tillgång till tvärsnittsdata där olika generationer blandas. Vi kommer också med största sannolikhet att få en högre andel med privat pensionssparande än i uppgifterna från SCB som presenterades i Avsnitt 1.1 eftersom vi, till skillnad från SCB, även inkluderar individer med ett inaktivt privat pensionssparande.

Nedan i Tabell 1 beskrivs de variabler vi kommer att ha användning av i vår studie.

Tabell 1: *Presentation av variabler.*

Variabel	**Nivå**	**Mätnivå**	**Variabeltyp**
Har privat pensions-sparande	Ja Nej	Nominalskala	Respons
Kön	Kvinna Man	Nominalskala	Förklarande
Civilstånd	Gift Ogift	Nominalskala	Förklarande
Yrkessektor	Egen företagare Kommunanställd Privatanställd arbetare Privatanställd tjänsteman Saknar förvärvsinkomst Statligt anställd Övriga	Nominalskala	Förklarande
Ålder	16-64 år	kvotskala	Förklarande
Månadslön	5 000 - 200 000 kr	kvotskala	Förklarande

Eftersom vi i studien är intresserade av att undersöka förekomsten av ett privat pensionssparande för olika grupper av individer kommer variabeln *Har privat pensionssparande* i Tabell 1 ovan att utgöra vår *responsvariabel*, det vill säga den variabel vi är intresserade av att mäta effekten på. De övriga variablerna utgör så kallade *förklarande variabler* och är variabler som används för att kunna påvisa skillnader mellan olika grupper alternativt vars effekter vi vill kontrollera för.

Att en variabel är på *nominalskalenivå* innebär att variabeln kan delas in i kategorier utan inbördes ordning (Tamhane, Dunlop, 2000). För de variabler där vi bara har två sådana kategorier räcker det att välja en av kategorierna som referensnivå *noll* och låta den andra

kategorin få värdet *ett*. Vi får en så kallad *indikatorvariabel*. För de variabler på nominalskalenivå som har fler än två kategorier, låt säga c kategorier, inför vi c-1 indikatorvariabler, $X_1, X_2, \ldots, X_{c-1}$, som antar värdet *ett* om individen tillhör kategori *i* och *noll* annars, $i = 1, 2, \ldots, c-1$. Den *c*:te kategorin utgör här referensnivå och i denna kategori befinner vi oss endast i det fall $X_1 = X_2 = \cdots = X_{c-1} = 0$. Anledningen till att vi måste utesluta referensnivån ur modellen och endast använda c-1 indikatorvariabler är för att undvika *multikollinearitet* som annars uppstår på grund av att vi får att $X_1 + X_2 + \cdots + X_c = 1$. Multikollinearitet innebär att vi kan uttrycka en variabel perfekt eller med hög korrelation i de övriga variablerna, till exempel att $X_c = 1 - X_1 - X_2 - \cdots - X_{c-1}$.

I Tabell 1 kan vi se att *Yrkessektor* är den enda variabel som har fler än två kategorier och där det är aktuellt att införa denna form av indikatorvariabler. I Tabell 2 nedan finns närmare beskrivet hur *Yrkessektor* kommer att kodas i den fortsatta analysen. Där har vi låtit de privatanställda tjänstemännen utgöra referensnivå eftersom det är den allra största yrkesgruppen i samhället.

Tabell 2: *Införande av indikatorvariabler som beskriver vilken yrkessektor individen tillhör.*

	Indikatorvariabel					
Nivå	X_1	X_2	X_3	X_4	X_5	X_6
Privatanställd tjänsteman	0	0	0	0	0	0
Egen företagare	1	0	0	0	0	0
Kommunanställd	0	1	0	0	0	0
Privatanställd arbetare	0	0	1	0	0	0
Saknar förvärvsinkomst	0	0	0	1	0	0
Statligt anställd	0	0	0	0	1	0
Övriga	0	0	0	0	0	1

Två av de förklarande variablerna, ålder och lön, är på kvotskalenivå, vilket innebär att vi kan betrakta dem som kontinuerliga. Kvotskalan kännetecknas av att ha en absolut nollpunkt (Dahmström, 2011).

Det kan också förekomma att förklarande variabler är på *ordinalskalenivå*, vilket innebär att variabeln kan ordnas i storleksordning fast utan att differensen mellan nivåerna för den skull är meningsfull (Dahmström, 2011). Exempel på variabler som mäts på ordinalskalenivå är åsikter som kan kategoriseras från *mycket dåligt* till *mycket bra*. I Tabell 1 ovan finns inga variabler på ordinalskalenivå. Däremot om vi skulle välja att kategorisera de kontinuerliga variablerna *lön* och *ålder* kommer dessa nya variabler bli på ordinalskalenivå. Det finns metoder för att hantera sådana förklarande variabler, till exempel att välja ett visst numeriskt värde per kategori (för variabeln *ålder* skulle åldern i kategorins mitt kunna användas) och sedan behandla variabeln som en kontinuerlig variabel. Här finns dock en risk att nivåerna väljs godtyckligt beroende på vem som utför

undersökningen. Ett annat sätt är att behandla dem som kategoriska variabler på nominalskalenivå.

2.3 Modell

Vi kommer att använda en multipel logistisk regressionsmodell för att prediktera sannolikheten att en individ har ett privat pensionssparande beroende på olika bakgrundsvariabler, eller förklarande variabler, såsom kön, yrkeskategori, lön, ålder och civilstånd.

I detta kapitel förklaras inledningsvis de grundläggande förutsättningarna för att en logistisk regressionsmodell ska vara lämplig att använda. Sedan följer en förklaring till hur modellen kan tolkas med hjälp av bland annat *oddskvoten*.

2.3.1 Logistisk regressionsmodell

För en enskild individ gäller att individen antingen har ett privat pensionssparande eller inte. Det innebär att förekomsten av ett privat pensionssparande kan beskrivas med en binär stokastisk variabel, en så kallad *binär responsvariabel*, som antar värdet *ett* eller *noll* enligt följande

$$Y = \begin{cases} 1, om\ individen\ har\ ett\ privat\ pensionssparande \\ 0, annars. \end{cases}$$

Om vi antar att vi har *p* stycken förklarande variabler i vår modell, $X_1, X_2, \dots, X_p$, som antar värdena $x_1, x_2, \dots, x_p$, beskrivs den logistiska regressionsmodellen enligt

$$P(Y = 1 \mid X_1 = x_1,\ X_2 = x_2\ , \dots, X_p = x_p) = \frac{exp\{\beta_0 + \sum_{j=1}^{p} \beta_j x_j\}}{1 + exp\left\{\beta_0 + \sum_{j=1}^{p} \beta_j x_j\right\}} \qquad (1)$$

där $P\big[\, Y = 1 \mid X_1 = x_1,\ X_2 = x_2\ , \dots, X_p = x_p\big]$ är sannolikheten att Y antar värdet *ett* givet uppsättningen på de förklarande variablerna och $\beta_j,\ j = 0, 1, 2, \dots, p$ är regressionskoefficienter som ska skattas (Agresti, 2002). Observera att högerledet i (1) endast antar värden i intervallet noll till ett, vilket stämmer väl överens med värden som en sannolikhet faktiskt kan anta. Detta är en av anledningarna till att just denna modell används istället för en linjär modell då sambandet mellan en responsvariabel som endast kan anta två värden och ett antal förklarande variabler ska beskrivas (Kleinbaum, Kupper, Nizam & Muller, 2008).

För att kunna utnyttja teorin för logistiska regressionsmodeller behöver vi transformera vänsterledet i (1) så att vi får en modell som kan uttryckas linjärt i parametrarna β_j (Agresti, 2002). Vi kommer därför att använda oss av den så kallad *logitformen* av modellen. Logit är en transformation av sannolikheten $P(\,Y = 1\,)$som definieras av

$$logit[P(\,Y=1\,)] = log\,\frac{[P(\,Y=1\,)]}{1-[P(\,Y=1\,)]} \qquad (2)$$

Om vi i ovanstående ekvation betingar på $X_1, X_2, \ldots, X_p$ och sedan uttrycker $P(Y=1 \mid X_1 = x_1,\, X_2 = x_2\,, \ldots, X_p = x_p)$ i enlighet med modell (1) får vi följande samband

$$logit[P(Y=1 \mid X_1 = x_1,\, X_2 = x_2\,, \ldots, X_p = x_p)] = \beta_0 + \sum_{j=1}^{p} \beta_j x_j, \qquad (3)$$

det vill säga en linjär funktion i parametrarna $\beta_j,\ = 0, 1, 2, \ldots, p$.

2.3.2 Tolkning av den logistiska regressionsmodellen

Oddset för en händelse definieras som kvoten mellan sannolikheten att händelsen inträffar och att händelsen inte inträffar. Om vi i detta avsnitt använder beteckningar i enlighet med Agresti (2002) och sätter $\pi(x) = P(Y=1 \mid X_1 = x_1,\, X_2 = x_2\,, \ldots, X_p = x_p)$, ges oddset för händelsen $Y=1 \mid X_1 = x_1,\, X_2 = x_2\,, \ldots, X_p = x_p$ av

$$\Omega = \frac{\pi(x)}{1-\pi(x)}.$$

Vi ser att om sannolikheten att Y antar värdet *ett* är större än sannolikheten att Y antar värdet *noll* så är oddset större än ett. Ett odds på tre innebär till exempel att sannolikheten att Y är lika med *ett* är tre gånger så stor som sannolikheten att Y är lika med *noll* och efter lite räkning finner vi att $\pi(x)$ måste vara lika med 0,75 i detta fall. Det gäller också att oddset antar värden mellan noll och oändligheten eftersom $\pi(x)$ antar värden mellan noll och ett.

Vi introducerar också *oddskvoten*, θ, som kvoten mellan två odds (Agresti, 2002). Oddskvoten av oddsen Ω_1 och Ω_2 ges alltså av

$$\theta = \frac{\Omega_1}{\Omega_2} = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}.$$

För att förstå hur oddskvoten hänger samman med tolkningen av den logistiska regressionsmodellen låter vi vektorn $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ och

$\pi(\boldsymbol{x}_i) = P(Y=1 \mid X_{i1} = x_{i1},\, X_{i2} = x_{i2}\,, \ldots, X_{ip} = x_{ip}),\ i = 1,2,3, \ldots$. Det vill säga $\pi(\boldsymbol{x}_i)$ är sannolikheten att Y är lika med *ett*, givet uppsättningen $\boldsymbol{x}_i$ på de förklarande variablerna. Om vi nu utnyttjar att vi enligt den logistiska regressionsmodellen i (3) har att

$$\log\frac{\pi(\boldsymbol{x}_i)}{1-\pi(\boldsymbol{x}_i)} = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \iff$$

$$\iff \frac{\pi(\boldsymbol{x}_i)}{1-\pi(\boldsymbol{x}_i)} = exp\left\{\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}\right\}$$

får vi att (Klainbaum et al., 2008)

$$\frac{\pi(\boldsymbol{x}_i)/\big(1-\pi(\boldsymbol{x}_i)\big)}{\pi(\boldsymbol{x}_{i'})/\big(1-\pi(\boldsymbol{x}_{i'})\big)} = exp\left\{\sum_{j=1}^{p} \beta_j (x_{ij} - x_{i'j})\right\}. \qquad (4)$$

Det vill säga oddskvoten för en *grupp* som karakteriseras av att ha förklarande variabler enligt $\boldsymbol{x}_i$ och en grupp med förklarande variabler enligt $\boldsymbol{x}_{i'}$ beror bara på parametrarna β_j och differensen i värdena på de förklarande variablerna. Speciellt gäller att om inget samspel förekommer och grupperna endast skiljer sig genom att värdet på den j:te förklarande variabeln ökar med en enhet, från x_{ij} till $x_{i'j}$=$x_{ij}+1$, så blir oddskvoten lika med

$$\frac{\pi(\boldsymbol{x}_i)/\big(1-\pi(\boldsymbol{x}_i)\big)}{\pi(\boldsymbol{x}_{i'})/\big(1-\pi(\boldsymbol{x}_{i'})\big)} = exp\left\{\sum_{j=1}^{p}\beta_j(x_{ij}-x_{i'j})\right\} = e^{\beta_j}.$$

Vi ser också att om $\beta_j = 0$ för något $j = 1, 2, 3, \ldots$ så har den j:te förklarande variabeln ingen effekt på oddskvoten.

2.4 Skattningar med maximum-likelihood-metoden

2.4.1 Parameterskattningar

För att skatta parametrarna i den logistiska regressionsmodellen används *maximum-likelihood-metoden*. Metoden går ut på att hitta de parametrar som maximerar observationernas *likelihood-funktion*, eller likvärdigt maximerar *log-likelihood-funktionen*.

I vår studie kommer vi att ha ett flertal *oberoende observationer* på den binära responsvariabeln för varje uppsättning av de förklarande variablerna, $\boldsymbol{x}_i = \big(x_{i1}, x_{i2}, \ldots, x_{ip}\big)$. Om vi i enlighet med (Agresti, 2002) låter N vara antalet möjliga *unika kombinationer* av de förklarande variablerna, n_i vara det totala antalet observationer för kombination *i* samt y_i vara antalet observationer för kombination *i* där responsvariabeln antar värdet *ett*, ges log-likelihood-funktionen för den logistiska regressionsmodellen av

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{N} y_i\big(\beta_0 + \sum_{j=1}^{p}\beta_j x_{ij}\big) - \sum_{i=1}^{N} n_i log\left(1 + e^{\beta_0+\sum_{j=1}^{p}\beta_j x_{ij}}\right).$$

För mer detaljer kring log-likelihood-funktionen hänvisas till Bilaga 1. Log-likelihood-funktionen ses som en funktion av de okända parametrarna $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$. Ett sätt att maximera funktionen och få ut de skattningar som ger detta maximum är att derivera $l(\boldsymbol{\beta})$ med avseende på $\boldsymbol{\beta}$ och lösa det ekvationssystem som fås då de partiella derivatorna sätts till noll. Dessa ekvationer kommer att vara icke-linjära och kräva numeriska lösningsmetoder, till exempel Newton-Raphson metoden. Det är dock inte självklart att sådana lösningar existerar trots att funktionen är konkav och därmed inte innehåller lokala minimum som kan ställa till problem. Om det råder *perfekt diskriminering* eller *ingen överlappning*, det vill säga att värdet av responsvariabeln kan predikteras utan osäkerhet då uppsättningen på de förklarande variablerna är känd, existerar inga skattningar eller så blir skattningarna oändligt stora (Hosmer et al., 2013). Det finns även andra fall, till exempel vid *quasi-perfekt diskriminering,* där skattningarna existerar men blir oerhört stora och med stora standardfel. Quasi-perfekt diskriminering innebär att en viss överlappning finns för någon eller några få uppsättningar av de förklarande variablerna. I praktiken uppkommer dessa problem då stickprovsstorleken är liten i förhållande till antalet förklarande variabler i modellen.

2.4.2 Intervallskattningar

När vi har hittat parameterskattningarna med maximum-likelihood-metoden är det av intresse att även bestämma skattningarnas varians. Genom dem kan vi nämligen bestämma konfidensintervall för de skattade sannolikheterna. Variansen fås genom att utnyttja teorin om att maximum-likelihood-skattningar är asymptotiskt normalfördelade och väntevärdesriktiga med en varians-kovariansmatris som ges ur inversen av informationsmatrisen $\boldsymbol{I}$ (Tamhane, 2000). Om vi låter $Var(\boldsymbol{\beta})$ vara skattningarnas varians-kovariansmatris gäller alltså att

$$Var(\boldsymbol{\beta}) = \boldsymbol{I}^{-1}(\boldsymbol{\beta}).$$

Skattningarnas varians får vi som diagonalelementen i ovanstående matris. För den logistiska regressionsmodellen är det oftast inte möjligt att uttrycka skattningarnas varianser explicit (Hosmer et al., 2013). För att läsa mer om informationsmatrisen och hur den skapas hänvisas till Bilaga 1.

För att bestämma skattningarnas standardfel, $SE(\hat{\beta}_j)$, använder vi den observerade informationsmatrisen (Hosmer et al., 2013). Det skattade log-oddset i ekvation (3) i Avsnitt 2.3.1 ges av

$$\widehat{\text{logit}}(\pi(\boldsymbol{x}_i)) = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x_{ij}.$$

Vi får därför, enligt grundläggande sannolikhetsteori, att det skattade log-oddsets varians ges av

$$Var\left(\hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x_{ij}\right) = \sum_{j=0}^{p} Var(\hat{\beta}_j) x^2{}_{ij} + \sum_{j=0}^{p} \sum_{k=j+1}^{p} 2 x_j x_k Cov(\hat{\beta}_j, \hat{\beta}_k).$$

Denna varians skattas med $\widehat{Var}\left(\hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x_{ij}\right)$ genom att i uttrycket ovan byta ut $Var(\hat{\beta}_j)$ och $Cov(\hat{\beta}_j, \hat{\beta}_k)$ mot de skattningar som fås genom den observerade, inverterade informationsmatrisen. Den skattade log-oddskvotens standardfel, $SE\left(\hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x_{ij}\right)$, får vi sedan genom att ta roten ur den skattade variansen. Ett approximativt 95-procentigt konfidensintervall för log-oddset ges nu av

$$\hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x_{ij} \pm z_{0,025} \cdot SE\left(\hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x_{ij}\right)$$

där $z_{0,025}$ är normalfördelningens $(1 - 0{,}025) \cdot 100$ procent percentil.

För att förenkla beteckningarna låter vi $\hat{g}(\boldsymbol{x}_i) = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x_{ij}$. De skattade sannolikheterna kan nu skrivas som

$$\hat{\pi}(\boldsymbol{x}_i) = \frac{e^{\hat{g}(x_i)}}{1 + e^{\hat{g}(x_i)}}$$

om vi utnyttjar ekvation (1) i Avsnitt 2.3.1.

Ett approximativt 95-procentigt konfidensintervall för de skattade sannolikheterna fås nu genom att byta ut $\hat{g}(\boldsymbol{x}_i)$ mot sin lägre respektive övre gräns i konfidensintervallet (Hosmer et al., 2013). Det vill säga att om vi låter a vara konfidensintervallets lägre gräns och b den övre ges ett 95-procentigt konfidensintervall för $\pi(\boldsymbol{x}_i)$ av

$$\frac{e^a}{1 + e^a} \leq \pi(\boldsymbol{x}_i) \leq \frac{e^b}{1 + e^b}.$$

2.5 Modellutvärdering

2.5.1 Test av linearitet i kontinuerliga förklarande variabler

För att testa om de kontinuerliga förklarande variablerna *lön* och *ålder* kan anser vara linjära i log-oddset, vilket är en förutsättning för att den logistiska regressionsmodellen ska gälla, används en metod som presenteras i Hosmer et al. (2013). Den går ut på att dela den kontinuerliga variabeln i fem delar efter dess 20:e percentiler. Sedan anpassas den logistiska regressionsmodellen med den kontinuerliga förklarande variabeln sedd som en kategorisk variabel med den lägsta kategorin som referensnivå. De parameterskattningar vi får bör vara linjära i kategoriernas *mittvärden*. Som mittvärden föreslås intervallets mitt (det vill säga största plus minsta värde dividerat med två), men vi kommer istället att ta medianen som mittvärde. Detta på grund av att lönefördelningen och åldersfördelningen inte är symmetriska. Om det visar sig att den kontinuerliga variabeln inte kan anses vara linjär i log-oddset kan variabeln transformeras eller göras om till en kategorisk variabel efter att ha delats in i lämpliga intervall. Eftersom variabeln lön har ett stort spann (5 000 – 200 000 kronor i månaden) och vi har tillgång till ett stort antal observationer kommer lön att delas i 10:e percentilerna istället för 20:e percentilerna så som Hosmer et al. (2013) föreslår.

2.5.2 Likelihood kvottestet

Likelihood kvottestet kan användas för att testa om vissa av modellparametrarna är noll (Agresti, 2002). Testet går ut på att jämföra den maximerade log-likelihood-funktionen l_1 för en anpassad modell (grundmodellen) med den maximerade log-likelihood-funktionen l_0 för en modell som är ett specialfall av den anpassade modellen (hypotesmodellen), till exempel en modell som sätter en eller flera parametervärden till noll. Vi kan formulera ett hypotestest enligt följande

H_0: hypotesmodellen gäller

H_A: hypotesmodellen gäller inte till förmån för grundmodellen.

Det kan visas att teststatistikan $-2(l_0 - l_1)$ under nollhypotesen asymptotiskt följer en chi-tvåfördelning med antalet frihetsgrader, df, lika med skillnaden i antalet parametrar i grundmodellen och hypotesmodellen då antalet observationer går mot oändligheten. Därmed får vi en teststatistika som kan användas för att testa om en anpassad modell kan reduceras genom att utesluta vissa parametrar. Om vi låter

$$LR = -2(l_0 - l_1)$$

förkastar vi nollhypotesen på signifikansnivån α till förmån för grundmodellen om

$$LR > \chi^2_\alpha(df)$$

där $\chi^2_\alpha(df)$ är chi-tvåfördelningens $(1-\alpha)\cdot 100$ procent percentil då antalet frihetsgrader är *df*.

Chi-tvåfördelningen presenteras närmare i Bilaga 1.

2.5.3 Wald-test

Wald-testet är ett dubbelsidigt test som kan användas för att testa enskilda parametrar enligt följande hypoteser

$$H_0\colon \beta = \beta_0$$

$$H_A\colon \beta \neq \beta_0.$$

Baserat på de asymptotiska egenskaperna hos maximum likelihoodskattningar kan man visas att

$$Z = \frac{\hat{\beta} - \beta_0}{SE(\hat{\beta})}$$

är approximativt normalfördelad med väntevärde noll och varians ett under H_0 (Agresti, 2002). Här är $\hat{\beta}$ det skattade parametervärdet i grundmodellen och $SE(\hat{\beta})$ skattningens standardfel. Om vi kvadrerar Z får vi, för en tvåsidig alternativhypotes, att Z^2 är chi-tvåfördelad med en frihetsgrad under H_0. Vi kommer därmed att förkasta nollhypotesen på signifikansnivå α till förmån för alternativhypotesen om $Z^2 > \chi^2_\alpha(1)$. Normalfördelningen och chi-tvåfördelningen finns närmare presenterade i Bilaga 1.

2.5.4 Hosmer-Lemeshow goodness of fit test

För att utvärdera hur väl modellen passar data kan olika mått på modellanpassning användas, till exempel Pearson chi-tvåstatistikan och deviansen (Hosmer et al., 2013). Båda dessa teststatistikor är asymptotiskt chi-tvåfördelade om modellen stämmer samt under förutsättning att antalet unika kombinationer, N, av de förklarande variablerna förblir konstant då antalet observationer, n, ökar. Det sistnämnda kriteriet är sällan uppfyllt i praktiken då vi har minst en kontinuerlig förklarande variabel med i modellen, ty om n ökar kommer även N att öka. Antalet unika kombinationer av de förklarande variablerna förblir alltså inte konstant och teststatistikornas signifikansnivåer kommer att vara missvisande.

Ett sätt att komma runt detta problem presenteras av Hosmer et al. (2013) och går ut på att gruppera individerna i g olika grupper efter antingen de skattade sannolikheternas percentiler eller efter fixa nivåer på de skattade sannolikheterna. Inom varje grupp g studeras differensen mellan den observerade och den skattade frekvensen av individer med y=1 respektive y=0. Dessa differenser kvadreras och summeras enligt följande till

$$\hat{C} = \sum_{k=1}^{g} \left[\frac{(o_{1k} - \hat{e}_{1k})^2}{\hat{e}_{1k}} + \frac{(o_{0k} - \hat{e}_{0k})^2}{\hat{e}_{0k}} \right]$$

där

o_{1k} = den observerade frekvensen av individer med y=1 i grupp k

$\hat{e}_{1k}$ = den skattade frekvensen av individer med y=1 i grupp k

o_{0k} = den observerade frekvensen av individer med y=0 i grupp k

$\hat{e}_{0k}$ = den skattade frekvensen av individer med y=0 i grupp k.

$\hat{C}$ är Hosmer-Lemeshow goodness of fit statistika som, under antagande om att modellen är korrekt, väl approximeras med en chi-tvåfördelad stokastisk variabel med *g-2* frihetsgrader. Vi kommer att pröva att använda *g* lika med 10, 20 och 100 i våra beräkningar av $\hat{C}$. Vi kommer alltså att förkasta hypotesen om att modellen passar data om

$$\hat{C} > \chi^2_\alpha(g-2)$$

där $\chi^2_\alpha(g-2)$ är chi-tvåfördelningens $(1-\alpha)\cdot 100$ procent percentil då antalet frihetsgrader är *g-2*.

2.5.5 AIC – Akaike Information Criterion

Liksom likelihood kvottestet kan *Akaike information criterion*, AIC, användas för att jämföra olika modeller med varandra. En stor skillnad mellan dessa metoder är dock att i likelihood kvottestet måste den mindre modellen vara ett specialfall av den större modellen, se Avsnitt 2.5.2, medan det kriteriet inte måste vara uppfyllt för att kunna använda AIC (Gujarati, Porter, 2009). AIC inkluderar även ett straff för antalet förklarande variabler i modellen och på så sätt kan överparametrisering undvikas genom att studera AIC. Måttet AIC definieras som

$$AIC = -2 \cdot l + 2 \cdot (1 + p)$$

där l är modellens maximerade log-likelihoodfunktion och *p* är antalet regressions-koefficienter för icke-konstanta förklarande variabler (Hosmer et al., 2013). Det finns inget statistiskt test för att utvärdera AIC kopplat till olika modeller, utan den modell som har lägst AIC är den som anses vara bättre.

2.5.6 ROC – Receiver Operating Characteristic

Ett sätt att mäta den anpassade modells prediktiva förmåga är att beräkna arean under ROC-kurvan (Hosmer, et al., 2013). ROC står för receiver-operating-characteristic. För att göra det definierar vi *sensitiviteten* och *specificiteten* enligt följande

$$sensitiviteten = P(\hat{y}_i = 1 |\, individen\ har\ egenskapen)$$

$$specificiteten = P(\hat{y}_i = 0 |\, individen\ har\ inte\ egenskapen)$$

där $\hat{y}_i$ är det enligt modellen predikterade värdet på responsvariabeln (i vårt fall förekomsten av ett privat pensionssparande eller inte) för en individ med förklarande variabler enligt $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip})$. Eftersom modellen endast skattar sannolikheterna, $\hat{\pi}(\boldsymbol{x}_i)$, kommer vi att behöva använda ett så kallat *tröskelvärde*, *c*, för att avgöra om vi ska tilldela värdet *ett* eller *noll* till $\hat{y}_i$. Tröskelvärdet används enligt följande

$$\hat{y}_i = \begin{cases} 1, & \hat{\pi}(\boldsymbol{x}_i) > c \\ 0, & \hat{\pi}(\boldsymbol{x}_i) \leq c \end{cases}$$

Om *c* till exempel är 0,7 kommer vi att anse att $\hat{y}_i$ är lika med *ett* ifall $\hat{\pi}(\boldsymbol{x}_i)$ är större än 0,7 och noll annars. Väljer vi ett högre värde på *c* kommer vi att få färre observationer som anses vara *ett (*om det inte är så att modellen har väldigt bra prediktiv förmåga) och *sensitiviteten* minskar samtidigt som motsatt samban gäller för *specificiteten.*

Kurvan bestäms sedan genom att plotta *sensitiviteten* mot 1-*specificiteten* för olika nivåer på *tröskelvärdet*, *c*.

ROC-kurvan kommer att visa modellens förmåga att *diskriminera* mellan de två grupperna (det vill säga mellan de som har egenskapen och de som inte har det). Idealiskt är både sensitiviteten och specificiteten hög vilket innebär en större area under ROC-kurvan (arean varierar mellan 0,5 och 1). Det finns inga generella regler för hur stor arean ska vara, men en tumregel som ges är

$$om \begin{cases} arean \in [0{,}5; 0{,}7), & Dålig\ eller\ ingen\ prediktionsförmåga \\ arean \in [0{,}7; 0{,}8), & Acceptabel\ prediktionsförmåga \\ arean \in [0{,}8; 0{,}9), & Utmärkt\ prediktionsförmåga \\ arean \in [0{,}9; 1], & Enastående\ prediktionsförmåga. \end{cases}$$

2.5.7 Test på extern data samt praktisk signifikans

Eftersom urvalet består av alla Min Pensions användare som gjort minst en pensionsprognos under år 2015 och uppfyller urvalskriterierna beskrivna i Avsnitt 2.1 kommer vi ha ett stort antal individer till underlag för vår logistiska prediktionsmodell (totalt har vi tillgång till cirka 570 000 individer). Därmed finns möjlighet att använda så kallad *extern validering* för att avgöra modellens anpassning och prediktionsförmåga utanför den population modellen anpassats på. Ett sätt är att dela urvalet i två delar, varav den ena delen används för att bygga modellen efter och den andra delen används för att testa modellanpassningen och den prediktiva förmågan på. Vi kommer därför att göra ett *systematiskt urval,* som delar individerna i två lika stora delar på ett sådant sätt att vi får lika fördelning i de båda halvorna vad avser kön, lön samt ålder, det vill säga de förklarande variabler som vi anser vara viktigast. Det systematiska urvalet dras på ett sätt som beskrivs enligt Dahmström (2011). Först sorteras individerna i *urvalsramen (*det vill säga individerna i vårt ursprungliga urval enligt Avsnitt 2.1) efter kön, lön och ålder. Sedan dras urvalet till den halva efter vilken modellen ska anpassas genom att välja varannan individ i urvalsramen, där slumpen får avgöra om vi startar med individ nummer ett eller två. På detta sätt kommer vi få ett *stratifierat urval* som innebär att vi får lika fördelning av kön, lön och ålder i de båda halvorna.

Eftersom vi även efter delning av det ursprungliga urvalet kommer att ha ett mycket stort antal individer att anpassa modellen efter, kommer vi att kunna statistiskt säkerställa mycket små skillnader mellan olika grupper på grund av mycket små standardfel i skattningarna (Tamhane, et al., 2000). Det kommer därför att vara nödvändigt att inte bara se till den statistiska signifikansen i parameterskattningarna utan även till vad som är signifikant i verkligheten eller med andra ord signifikans i praktiken. Det kommer inte heller att uppstå problem med för få individer, utan vi kan kosta på oss att inte slå ihop grupper i onödan.

För att avgöra hur bra modellanpassningen är utanför det data modellen anpassats på kan vi använda en modifiering av Hosmer-Lemeshow testet som presenterades i Avsnitt 2.5.4 (Hosmer et al., 2013). Modifieringen går ut på att vi nu istället studerar antalet utfall av y=1 och y=0 i de *g* grupperna i det externa datasetet (se Avsnitt 2.5.4 för information om hur indelningen i grupperna sker). Låt nu n_k vara antalet observationen i det externa datasetet som tillhör grupp *k*, o_k vara antalet observationer i grupp *k* som antar värdet *ett* och $e_k = \sum m_j \pi_j$ vara det förväntade antalet ettor i grupp *k* under förutsättning att modellen är korrekt, k =1,2,…,*g*. Här är m_j antalet observationer för den unika kombinationen $\boldsymbol{x}_j$ av de förklarande variablerna och π_j är sannolikheten att en observation med kombination $\boldsymbol{x}_j$ av de förklarande variablerna antar värdet *ett.* Då ges Hosmer-Lemeshow statistikan av

$$C_v = \sum_{k=1}^{g} \frac{(o_k - e_k)^2}{n_k \overline{\pi}_k (1 - \overline{\pi}_k)}$$

där $\overline{\pi}_k = \sum \frac{m_j \pi_j}{n_k}$. Under antagandet att modellen är korrekt följer att C_v är approximativt chi-tvåfördelad med *g-2* frihetsgrader. Detta innebär att vi förkastar hypotesen att modellanpassningen är god i det externa datasetet om

$$C_v > \chi^2_\alpha(g-2)$$

där $\chi^2_\alpha(g-2)$ är chi-tvåfördelningens $(1-\alpha)\cdot 100$ procent percentil då antalet frihetsgrader är *g-2*. Detta test kommer med största sannolikhet att visa på en sämre modellanpassning än om vi utgår från det data som modellen anpassats på.

Det vi främst är intresserade av är att testa modellens prediktiva förmåga på det externa datasetet. Här använder vi oss istället av arean under ROC-kurvan så som beskrivs i Avsnitt 2.5.6, men med skillnaden att vi tittar på de predikterade sannolikheterna i det externa datasetet som vi får om vi utgår från den anpassade modellen.

2.5.8 Residualer och inflytelserika observationer

I datamaterialet kan det förekomma extrema observationer vars riktighet kan ifrågasättas. För att upptäcka dessa observationer kan bland annat Pearson residualer eller deviance residualer studeras (Agresti, 2002). Om det förekommer kontinuerliga variabler i modellen kommer vi i många fall endast att ha en observation per unik kombination av de förklarande variablerna, vilket gör att informationen från residualerna inte blir särskilt användbar. Ett sätt att komma runt detta problem är att kategorisera den kontinuerliga variabeln och på så sätt få flera observationer per unik kombination av de förklarande variablerna. Pearson residualerna ges sedan av

$$e_i = \frac{y_j - m_j\hat{\pi}_j}{\sqrt{m_j\hat{\pi}_j(1-\hat{\pi}_j)}}$$

där y_j är antalet observationer som antar värdet *ett* för kombination *j* av de förklarande variablerna, m_j är antalet observationer som har kombination *j* av de förklarande variablerna och $\hat{\pi}_j$ är den skattade sannolikheten att en observation med kombination *j* av de förklarande variablerna antar värdet *ett.*

Ett sätt att studera inflytelserika observationer är att se hur modellanpassning samt skattningar ändras då en variabel i taget utesluts och modellen anpassas på nytt (Hosmer et al., 2013). I vårt fall är detta inte meningsfullt då vi har ett mycket stort antal observationer.

2.6 Val av modell

Det finns flera olika standardmetoder som kan används för att avgöra vilka förklarande variabler som ska ingå i modellen och vilka som kan uteslutas. Att ta med många förklarande variabler i modellen gör visserligen att modellanpassningen förbättras (modellen kommer att passa våra data bättre), men andra problem kommer att uppstå (Hosmer et al., 2013). Till exempel minskar modellens förmåga att fånga intressanta samband eftersom modellen även fångar upp en stor del av den rent slumpmässiga variationen. Vi kommer att få stora standardfel i våra skattningar och modellen kommer att passa dåligt utanför våra data.

Den metod vi kommer att använda för modellanpassning beskrivs av Hosmer et al. (2013) och kallas *purposeful selection.* Det är en metod i sju steg som beskrivs nedan.

1. Den logistiska regressionsmodellen anpassas för var och en av de förklarande variablerna var för sig. Variabelns signifikans testas genom att studera p-värdet i likelihood kvottestet för modellen med respektive utan den förklarande variabeln, se Avsnitt 2.5.2 för mer detaljer om testet. Då den förklarande variabeln har *k* kategorier kommer teststatistikan att vara chi-tvåfördelad med *k*-1 frihetsgrader. Variabler med p-värden större än eller lika med 0,25 utesluts, men kommer att utvärderas igen i ett senare steg.

2. En logistisk regressionsmodell anpassas med alla de förklarande variabler som i steg 1 fick p-värden mindre än 0,25. Vi får på så sätt en första modell där bara huvudeffekter ingår. Nu studeras Wald-statistikan, se Avsnitt 2.5.3, och alla förklarande variabler med p-värden större än 0.05 utesluts. Om förklarande variabler utesluts i detta steg får vi en reducerad modell. Denna reducerade modell jämförs med den större modellen genom ett likelihood kvottest, se Avsnitt 2.5.2, för att se vilken modell vi ska gå vidare med.

3. I detta steg antar vi att vi valt den reducerade modellen i steg två. Nu jämförs skattningarna i denna modell med de skattningar vi fick i den större modellen i steg 2 för att se om stora förändringar skett. Om skattningarna ändrats med mer än 20 procent anses förändringen vara stor. Den uteslutna variabeln som orsakade förändringen bör då inkluderas i modellen igen eftersom den behövs för att kontrollera för effekten av andra, inkluderade variabler.

4. Nu går vi tillbaka till de variabler som uteslöts i steg 1. Dessa läggs till en i taget samtidigt som deras signifikans testas med Wald-statistikan. De förklarande variabler som nu blir statistiskt signifikanta inkluderas i modellen igen. Detta steg är viktigt för att hitta variabler som ensamma inte är statistiskt signifikanta utan endast är det i kombination med andra förklarande variabler.

5. Nu har vi hittat en modell med endast huvudeffekterna inkluderade och det är dags att utvärdera om de kontinuerliga förklarande variablerna kan anses vara lineära i log-oddset, se Avsnitt 2.5.1. Vi utreder också om de kategoriska variablerna kan kategoriseras på ett bättre sätt, till exempel genom att ändra eller slå ihop kategorier.

6. I detta steg läggs samspel mellan de förklarande variablerna till. Samspel innebär att effekten av en förklarande variabel beror på värdet eller nivån på en annan förklarande variabel. Ett exempel på samspel i vår modell skulle kunna vara om lönen bidrar positivt till att ha ett privat pensionssparande för män, medan den för kvinnor istället bidrar negativt. Det vill säga att effekten av lönen är olika beroende på om vi studerar kvinnor eller män. Här är det viktigt att tänka på att samspelen ska ha någon form av mening i praktiken, så som beskrevs i avsnittets inledning. Samspelen testas en och en genom att de enskilt läggs till i modellen med huvudeffekterna. Samtidigt testas den statistiska signifikansen med likelihood kvottestet. Observera att kategoriska variabler med fler än två kategorier kommer att resultera i att fler än ett samspel läggs till samtidigt. Till exempel är den kategoriska variabeln yrkeskategori beskriven med sex indikatorvariabler som alla kommer att kombineras samtidigt med en vald annan förklarande variabel. Alla samspel som visar sig vara signifikanta inkluderas i modellen och vi går tillbaka till steg 2 för att se om modellen kan reduceras, men vi tar då endast bort samspelen och låter huvudeffekterna vara kvar. Modellen som blir kvar är vår slutligt anpassade modell.

7. Den slutligt anpassade modellen i steg 6 utvärderas i enlighet med Avsnitt 2.5.

Eftersom vi har ett mycket stort antal observationer, relativt få förklarande variabler och en sned fördelning av till exempel kön inom olika yrkeskategorier samt åldrar kommer vi att starta modellanpassningen med en modell där samtliga huvudeffekter ingår. Steg ett i

purposeful selection kommer nämligen med största sannolikhet visa att alla våra variabler är statistisk signifikanta, alternativt kan vi inte lita på ett icke signifikant resultat på grund av att vi på förhand vet att det förmodligen är urvalets sammansättning som är med och påverkar. Dessutom vill vi ha med just dessa förklarande variabler i vår modell då vi på förhand tror oss veta att just dessa variabler påverkar förekomsten av att ha ett privat pensionssparande. Vi kommer således att börja modellanpassningen med steg 5, vilket innebär att vi först testar de kontinuerliga förklarande variablernas linearitet i log-oddskvoten. Efter det går vi till steg 6 där samspelen testas enskilt, för att sedan gå igenom steg 2-4 för att avgöra vilka samspel som ska tas med i den slutligt valda modellen (här ses huvudeffekterna som fixa). Den slutligt valda modellen utvärderas sedan enligt steg 7.

3 Resultat

Inledningsvis i detta kapitel beskrivs fördelningen i urvalet med avseende på de olika variabler vi har tillgång till. I Avsnitt 3.2, kommer vi att testa de kontinuerliga variablernas linearitet för att se om modellförutsättningarna är uppfyllda. Sedan, i Avsnitt 3.3, går vi vidare med att försöka hitta en slutlig modell som kan användas för att prediktera förekomsten av ett privat pensionssparande. I Avsnitt 3.4, utvärderas modellen och vi testar hur bra den predikterar observationerna i det externa datasetet vi har tillgång till. I Avsnitt 3.5, tolkar vi vår slutligt valda modell och använder den för att skatta sannolikheten att ha ett privat pensionssparande i olika grupper. Avslutningsvis, i Avsnitt 3.6, presenteras vilken storlek på pensionen från det privata pensionssparandet som olika grupper förväntas att få.

3.1 Fördelning i urvalet

Urvalet som utgör underlag för den anpassade modellen fås bland Min Pensions användare i enlighet med Avsnitt 2.1 samt 2.5.7 och består av 284 635 individer (efter att urvalet delats i två). Vi börjar med att presentera urvalets fördelning med avseende på kön. I Tabell 3.1 kan vi se att vi har en större andel män med i urvalet. Cirka 58 procent är män, vilket stämmer överens med att Min Pension faktiskt har fler manliga användare. Kvinnorna utgör cirka 42 procent av urvalet.

Tabell 3.1: *Könsfördelningen i urvalet.*

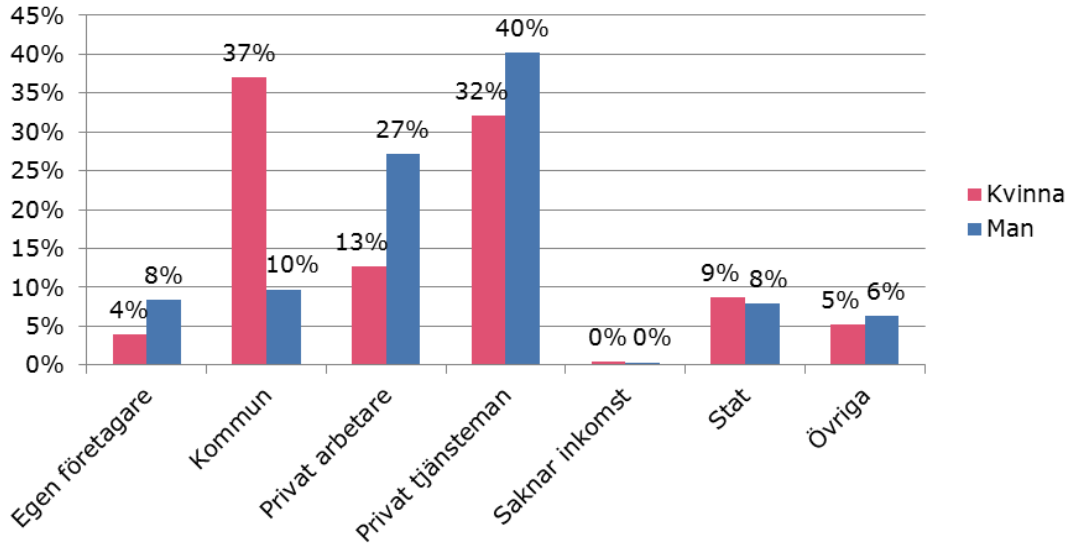
Kön	**Antal**	**Andel**
Kvinnor	119 054	42 %
Män	165 581	58 %
Totalt	284 635	100 %

Vi fortsätter med att studera urvalet uppdelat på yrkeskategori, se Tabell 3.2. Där kan vi se att det finns flest privatanställda tjänstemän med i urvalet, närmare bestämt ca 37 procent. Sedan kommer de kommunanställda och de privatanställda arbetarna, som utgör cirka 21 procent av urvalet. Att vi har flest individer i dessa yrkeskategorier går helt i linje med att det faktiskt är dessa yrkesgrupper som utgör störst andel av Sveriges befolkning (SCB, 2016). De statligt anställda, egna företagarna respektive gruppen övriga utgör cirka 8,2, 6,5 respektive 5,9 procent av urvalet. Allra minst är gruppen som saknar förvärvsinkomst, som utgör cirka 0,4 procent av urvalet.

Tabell 3.2*: Fördelningen över avtalsområden i urvalet.*

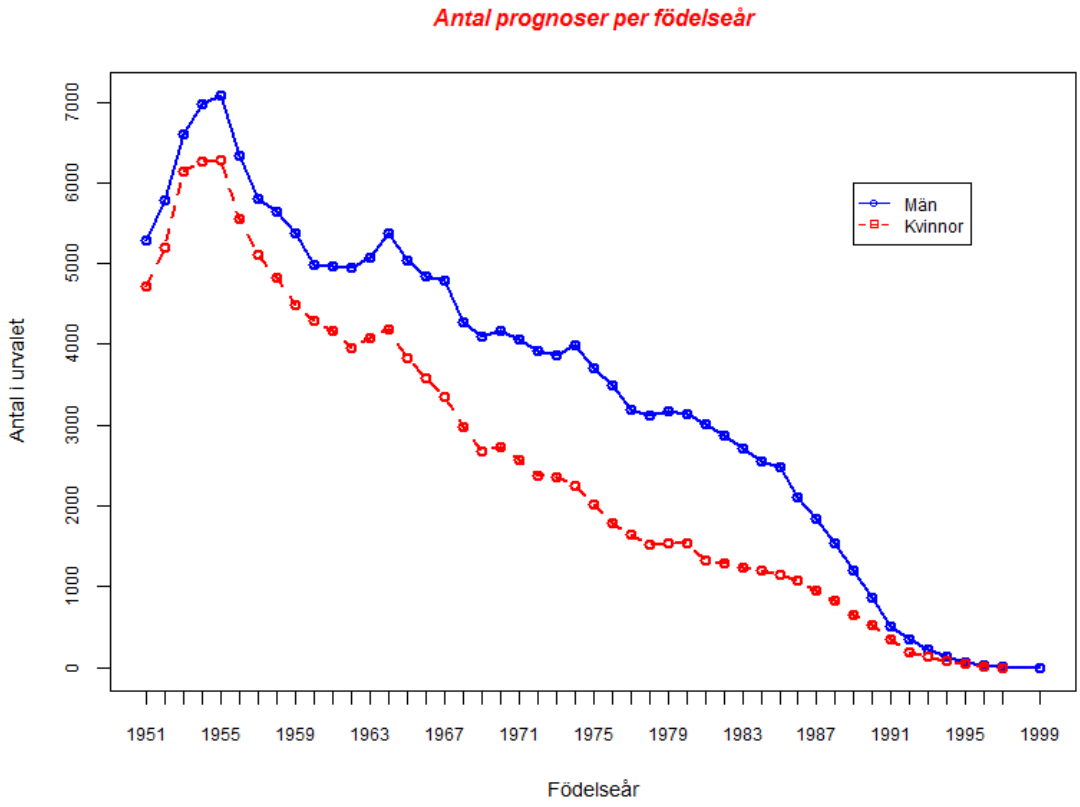
Kategori	Antal	Andel
Egen företagare	18 508	6,5 %
Kommunanställd	60 200	21 %
Privat arbetare	59 992	21 %
Privat tjänsteman	104 742	37 %
Saknar förvärvsinkomst	1 124	0,4 %
Statligt anställd	23 392	8,2 %
Övriga	16 677	5,9 %
Totalt	284 635	100 %

Om vi vidare delar upp urvalet på kön och studerar fördelningen över yrkeskategorierna, se Figur 3.1, finner vi stora skillnader mellan könen. Till exempel ser vi att cirka 37 procent av kvinnorna i urvalet är kommunanställda medan bara cirka 10 procent av männen är det. De yrkeskategorier där männen dominerar är bland egna företagare, privatanställda arbetare samt privatanställda tjänstemän. Cirka 27 procent av männen i urvalet är privatanställda arbetare medan bara cirka 13 procent av kvinnorna är det. Att könsfördelningen skiljer sig åt inom de olika yrkeskategorierna beror till största del på att det är så det ser ut i Sverige (SCB, 2016). Vi kommer inte vidare att fördjupa oss i jämförelser mellan Min Pensions användare och Sveriges befolkning som helhet utan hänvisar den intresserade till SCB.

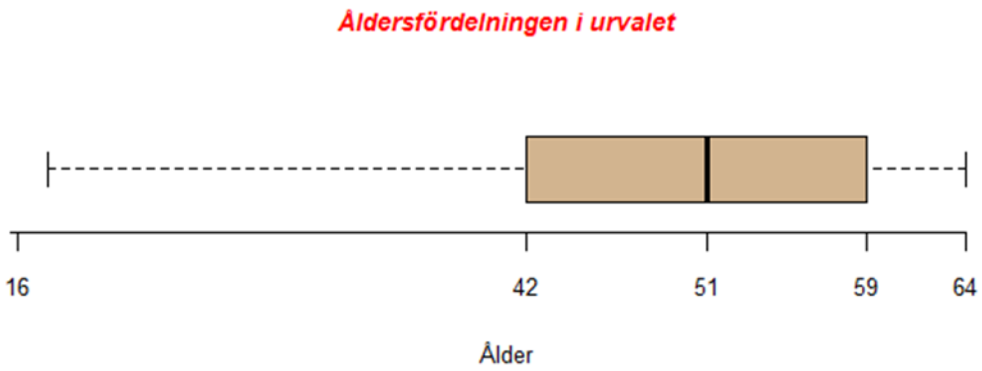


Figur 3.1: *Fördelningen över yrkeskategorier, uppdelat på kvinnor och män.*

Som nämnts i Avsnitt 2.1 har Min Pension en övervägande andel användare som närmar sig pensionsåldern. Detta beror på att intresset för den egna pensionen naturligt ökar med åldern. Figur 3.2 visar hur många individer i olika åldrar som finns med i urvalet, uppdelat på kön. Där kan vi se att vi har flest individer som är födda 1955 med i urvalet, närmare bestämt 13 365 stycken, samt väldigt få unga. Totalt finns till exempel endast en sextonåring, ingen sjuttonåring och sju artonåringar med i urvalet. I lådagrammet i Figur 3.3 kan vi se åldersfördelningen i urvalet. Där kan vi se att medianåldern är 51 år och att de 50 procent *mittersta* individerna är mellan 42 och 59 år gamla. Med mittersta menas här de individer som har en ålder mellan den 75:e och den 25:e percentilen.

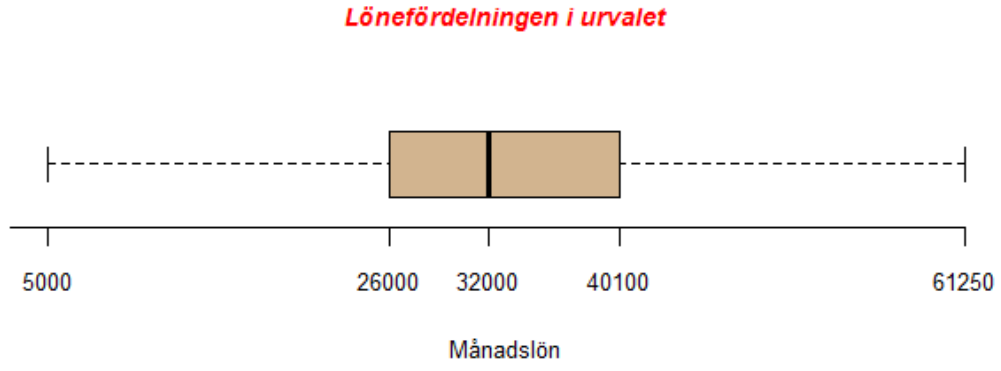


Figur 3.2: *Antalet individer i urvalet per födelseår, uppdelat på kvinnor och män.*



Figur 3.3: *Åldersfördelningen i urvalet.*

Även lönefördelningen åskådliggörs i ett lådagram, se Figur 3.4. Där kan vi se att medianlönen är 32 000 kronor i månaden och att de 50 procent *mittersta* individerna har löner mellan 26 000 och 40 100 kronor i månaden. De så kallade *whiskers* eller *morrhår* som dras ut från lådan visar mellan vilka nivåer som lönen kan anses vara normal i någon bemärkelse. Dessa morrhår dras ut 1,5 gånger kvartilavståndet (differensen mellan lönens 75:e och 25:e percentil) från lådans sidor och vi kan se att månadslöner mellan 5 000 och 61 250 kronor kan anses ligga inom normalspannet i urvalet. Den horisontella axeln har begränsats till att endast visa månadslöner upp till 61 250 kronor, men i själva verket varierar månadslönen mellan 5 000 och 200 000 kronor.



Figur 3.4: *Lönefördelningen i urvalet.*

Till sist studeras hur många gifta respektive ogifta individer som ingår i urvalet. I Tabell 3.3 kan vi se att vi har en något större andel gifta, cirka 54 procent, än ogifta, cirka 46 procent.

Tabell 3.3: *Fördelningen över civilstånd i urvalet.*

Civilstånd	**Antal**	**Andel**
Gift	152 802	54 %
Ogift	131 833	46 %
Totalt	284 635	100 %

Eftersom vi studerar förekomsten av ett privat pensionssparande presenteras även antalet och andelen som har ett privat pensionssparande i olika grupper. I Tabell 3.4 kan vi se att andelen kvinnor som har ett privat pensionssparande inom var och en av yrkeskategorierna är större än andelen män som har det. I Tabell 3.5 kan vi se att gifta i större utsträckning har ett privat pensionssparande både vad gäller kvinnor och män.

Tabell 3.4: *Andel och antal som har ett privat pensionssparande i olika grupper efter kön och yrkeskategori.*

Kön	**Yrke**	**Andel med Pps**	**(Antal med Pps/Totalt antal)**
Kvinna	Företagare	0,74	(3436/4650)
	Kommun	0,66	(29104/44040)
	Privat arbetare	0,57	(8637/15071)
	Privat tjm.	0,68	(25893/37986)
	Saknar inkomst	0,52	(271/521)
	Stat	0,67	(6910/10318)
	Övriga	0,59	(3664/6178)
	Samtliga	0,66	(77915/118764)
Man	Företagare	0,62	(8576/13822)
	Kommun	0,57	(9181/16010)
	Privat arbetare	0,53	(23705/44905)
	Privat tjm.	0,62	(40945/65582)
	Saknar inkomst	0,45	(269/594)
	Stat	0,61	(7874/13002)
	Övriga	0,53	(5386/10243)
	Samtliga	0,58	(95936/164158)

Tabell 3.5: *Antal och andel med ett privat pensionssparande uppdelat i grupper efter kön och civilstånd.*

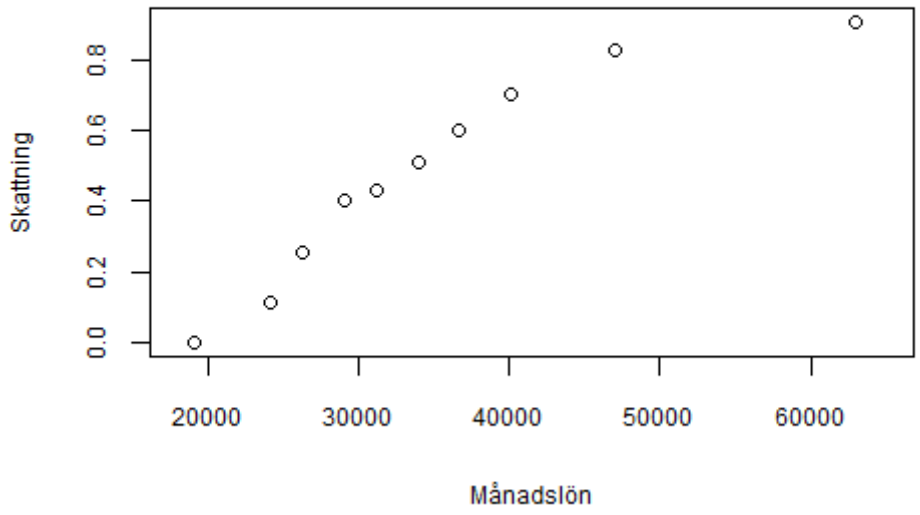
Kön	Civilstånd	Andel med Pps	(Antal med Pps/Totalt antal)
Kvinna	Gift	0,69	(44713/64861)
	Ogift	0,62	(33202/53903)
	Totalt	0,57	(8637/15071)
Man	Gift	0,63	(54744/86639)
	Ogift	0,53	(41102/77429)
	Totalt	0,67	(6910/10318)

3.2 Test av linearitet i kontinuerliga förklarande variabler

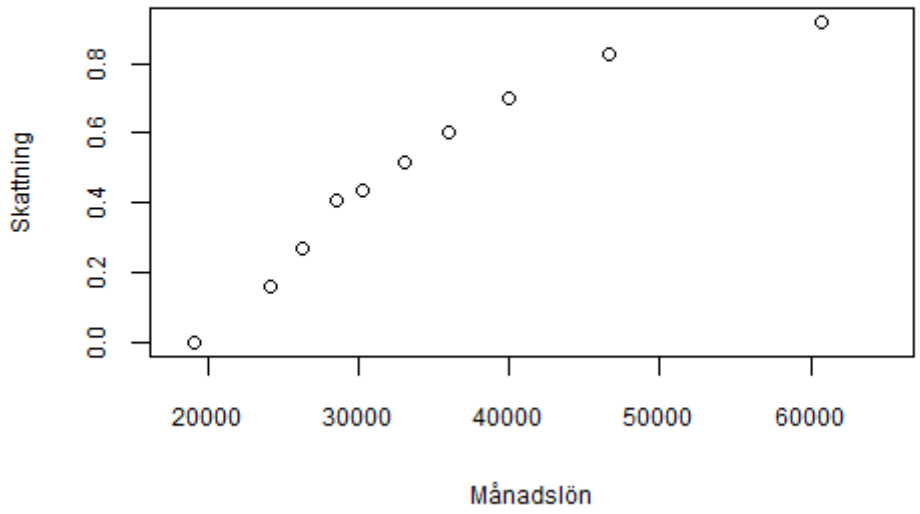
För att anpassa en slutlig logistisk regressionsmodell används delar av metoden *purposeful selection,* se Avsnitt 2.6. Där framgår att vi börjar med att titta på en modell där endast huvudeffekterna ingår innan vi går vidare och lägger till samspel. Den första modellen består alltså av de förklarande variablerna *kön*, *lön, ålder, yrkeskategori* och *civilstånd*. Resultatet presenteras i Tabell 1a-b i Bilaga 2. Där kan vi som väntat se att samtliga förklarande variabler är statistiskt signifikanta på signifikansnivån 0,05 om vi ser till Waldstatistikan, se Avsnitt 2.5.3. Vi kan också se att Hosmer-Lemeshow goodness of fit statistikans p-värde är mindre än 0,0001 samt att arean under ROC-kurvan endast är 0,6478 vilken tyder på en dålig modellanpassning och prediktiv förmåga, se Avsnitt 2.5.4 samt 2.5.6.

Det är nu dags att testa om de kontinuerliga förklarande variablerna *lön* och *ålder* kan anses vara lineära i log-oddset, se Avsnitt 2.5.1. Vi börjar med att studera variabeln lön och anpassar en logistisk regressionsmodell där samtliga av de fem ovanstående förklarande variablerna ingår, fast med lön som en kategorisk variabel indelad efter sina 10:e percentiler. Parameterskattningarna för de olika lönekategorierna plottas mot löneintervallens median och resultatet presenteras i Figur 3.4. Vi kan se att sambandet ser ganska linjärt ut för löner upp till en viss nivå. Vi har ju månadslöner ända upp till 200 000 kronor även om ytterst få individer (cirka 0,6 procent av urvalet) har månadslöner som är större än 100 000 kronor. Det finns metoder (till exempel med hjälp av indikatorvariabler) att behandla individer med höga löner i en separat grupp, men vi kommer istället pröva att utesluta individerna och se om vi får en bättre modellanpassning på så sätt. Resultatet, efter uteslutning av individer med månadslöner större än 100 000 kronor, presenteras i Figur 3.5. Vi kan se att resultatet är bättre, men fortfarande förekommer en krökning i de högre löneintervallen. Till viss del kan detta bero på att höga löneintervall blir bredare på grund av att vi har färre observationer där och på så sätt blir resultatet mer känsligt för val av mittpunkt. Vi väljer att gå vidare med analysen efter att höga månadslöner större än 100 000 kronor uteslutits.

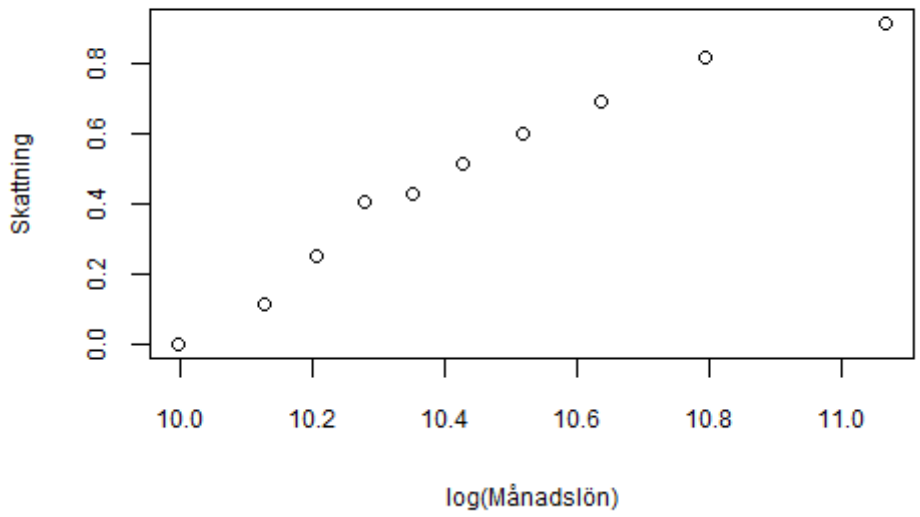
På grund av det krökta sambandet vi kan se prövar vi även att logaritmera lönen. Resultatet presenteras i Figur 3.6 och vi kan se en förbättrad linearitet, om än en svag sådan. Vi väljer därför att gå vidare med den logaritmerade lönen.



Figur 3.4: *Samband mellan medianen i olika lönekategorier indelade efter percentiler och parameterskattningarna för respektive kategori.*

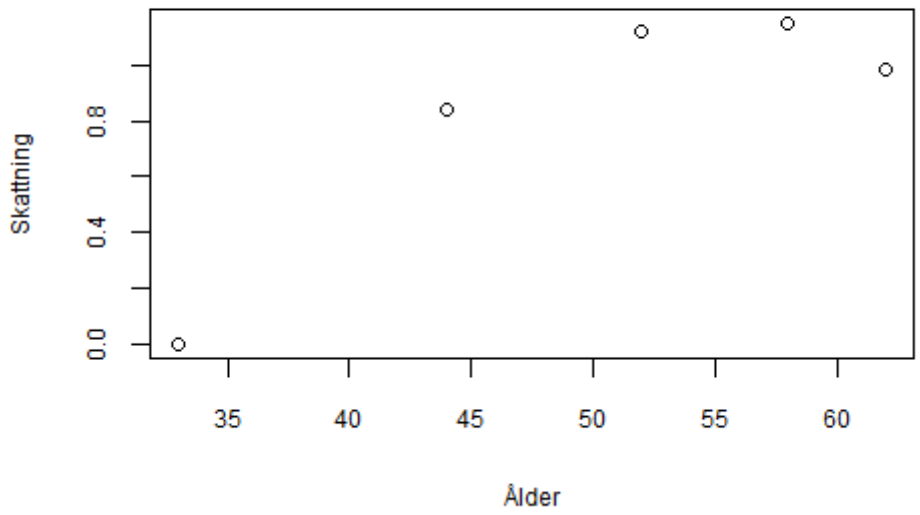


Figur 3.5: *Samband mellan medianen i olika lönekategorier indelade efter percentiler och parameterskattningarna för respektive kategori. Individer med månadslöner större än 100 000 kronor har uteslutits.*



Figur 3.6: *Samband mellan medianen i olika logaritmerade lönekategorier indelade efter percentiler och parameterskattningarna för respektive kategori. Individer med månadslöner större än 100 000 kronor har uteslutits.*

Vi går vidare på samma sätt och testar om ålder kan anses linjär. Den logistiska regressionsmodellen som anpassas innehåller samtliga förklarande variabler (log-lön som kontinuerlig variabel), fast ålder har delats in i kategorier efter sina 10:e percentiler. Parameterskattningarna som fås i den anpassade modellen plottas mot åldersintervallens median och presenteras i Figur 3.7. Vi kan se att sambandet inte ser linjärt ut. Sambandet är inte heller monotont och det är därför inte möjligt att transformera åldern och på så sätt försöka hitta ett samband som kan anses linjärt. Vi kommer istället att behålla ålder som den kategoriska variabel vi får efter att ha delat in åldrarna efter percentilerna. Den nya, kategoriserade variabeln ålder presenteras i Tabell 3.6. Även om ålder nu kommer att vara på ordinalskalenivå kommer vi att behandla den som en kategorisk variabel på nominalskalenivå, enligt argumentationen i Avsnitt 2.2. Vi kommer att välja åldrarna 61-64 år som referensnivå på grund av att resultaten är minst snedvridna i höga åldrar, se Avsnitt 2.1.



Figur 3.7: *Samband mellan medianen i olika ålderskategorier indelade efter percentiler och parameterskattningarna för respektive kategori.*

Tabell 3.6: *Presentation av variabeln ålder efter att den kategoriserats.*

Variabel	**Nivå**	**Mätnivå**	**Variabeltyp**
Ålder	16-39	Ordinal	Förklarande
	40-48		
	49-54		
	55-60		
	61-64		

3.3 Den anpassade modellen

Vi går vidare med metoden *purposeful selection,* se Avsnitt 2.6, för att se vilka samspel som bör läggas till i modellen. Signifikansnivån i testerna väljs till 0,05. Vi utgår från modellen med endast huvudeffekterna, som beskriver det linjära sambandet mellan *log-oddset av responsvariabeln Y* och de fem förklarande variablerna *kön, lön, yrkeskategori, ålder* samt *civilstånd*. Ålder har delats in i kategorier och lönen har logaritmerats, se Avsnitt 3.2.

Nu är det dags att testa ett samspel i taget genom att lägga till samspelet i fråga till modellen med bara huvudeffekterna och samtidigt testa dess signifikans. Totalt har vi tio olika samspel att testa. Detta görs med likelihood kvottestet, se Avsnitt 2.5.2. Resultatet presenteras i Tabell 3.7. Där kan vi se vilka samspel som testas samt teststatistikans storlek, antal frihetsgrader och p-värde. Vi kan även se mått på modellanpassning efter att det enskilda samspelet i fråga lagts till. De mått som presenteras är AIC samt Hosmer-Lemeshow goodness of fit statistika $\hat{C}$, se Avsnitt 2.5.4 och 2.5.5. Däremot presenteras inte parameterskattningarna för de tio olika modellerna i någon tabell, även fast vi samtidigt som vi testar ett enskilt samspel håller ett öga på huvudeffekterna så att inga drastiska förändringar sker när samspelet inkluderas. Om något utmärkande sker redovisas det dock i den löpande texten.

Vi börjar med att inkludera samspelet mellan kön och logaritmen av lön. Vi får en likelihood kvotstatistika på 0,04 och ett högt p-värde på 0,843, se Tabell 3.7. Därmed är samspelet inte statistiskt signifikant vilket innebär att vi tror på hypotesen att samspelet inte har någon effekt på responsvariabeln. Samtidigt kan vi se att huvudeffekten av responsvariabeln kön ändras från -0,36 till -0,31 och att den inte längre är signifikant, med ett p-värde från Wald-testet på 0,190 (detta presenteras inte i någon tabell). Eftersom huvudeffekterna i detta skede ses som fixa kommer vi endast att utesluta samspelet mellan kön och logaritmen av lön. Detta samspel kommer i enlighet med sjustegsmetoden i *purposeful selection* att utvärderas i ett senare skede.

Vi går vidare med samspelet mellan kön och yrkeskategori. Här måste vi komma ihåg att vi har sju yrkeskategorier representerade med sex indikatorvariabler, se Avsnitt 2.2. Vi kommer att lägga till samtliga sex möjliga samspel samtidigt till modellen och testa denna modells signifikans mot modellen med endast huvudeffekterna. Likelihood kvotstatistikan är således chi-tvåfördelad med sex frihetsgrader. Vi ser att teststatistikan antar värdet 86,0 och p-värdet är mindre än 0,001, se Tabell 3.7. Därmed förkastar vi hypotesen att samspelet inte har någon effekt.

Vi går vidare på samma sätt och testar de återstående tvåfaktorsamspelens signifikans. Vi kan se i Tabell 3.7 att samtliga av dessa samspel är signifikanta då p-värdena från likelihood kvottesten är mindre än 0,001, vilket innebär att samtliga av dessa samspel inkluderas.

Tabell 3.7: *Resultatet då ett samspel i taget testas genom att lägga till det i modellen med samtliga huvudeffekter. Hypotesen som prövas med likelihood kvottestet är att samspelet i fråga inte har någon effekt. Låga p-värden leder till att vi förkastar hypotesen.* I tabellen redovisas även AIC samt Hosmer-Lemeshow goodness of fit statistika $\hat{C}$ som fås efter att det enskilda samspelet lagts till i modellen med bara huvudeffekterna.

Samspel som lagt till	**Antal para-metrar i modellen**	**AIC**	**$\hat{C}$ med 8 f.g. (p-värde)**	**LR-statistika (p-värde) (f.g)**
Inget	14	358 305	41,1 (***)	-
Kön*log(Lön)	15	358 307	41,2 (***)	0,04 (0,843) (1)
Kön*Yrke	20	358 231	50,3 (***)	86,0 (***) (6)
Kön*Civilstånd	15	358 295	46,0 (***)	12,5 (***) (1)
Kön*Ålder	18	358 149	8,8 (0,358)	164 (***) (4)
log(Lön)*Yrke	20	357 786	72,4 (***)	531 (***) (6)
log(Lön)*Civilstånd	15	358 265	50,1 (***)	42,1 (***) (1)
log(Lön)*Ålder	18	358 119	45,7 (***)	195 (***) (4)
Yrke*Civilstånd	20	358 251	30,1 (***)	66,4 (***) (6)
Yrke*Ålder	38	357 852	13,0 (0,111)	502 (***) (24)
Civilstånd*Ålder	18	358 160	78,3 (***)	153 (***) (4)

Sammantaget får vi att detta steg i *purposeful selection* ger att samtliga tvåfaktorsamspel, förutom samspelet mellan kön och logaritmen av lön, är signifikanta. Vi anpassar därför en modell med samtliga huvudeffekter och alla tvåfaktorsamspel mellan kön, logaritmen av lön, yrkeskategori, ålder samt civilstånd, förutom samspelet mellan kön och logaritmen av lön som visade sig vara icke signifikant och uteslöts. Totalt kommer vi alltså ha en modell med 56 olika samspel, 13 olika huvudeffekter samt 1 intercept, det vill säga en modell med 70 parametrar. Vi kallar denna modell för Modell 1. Modellen finns presenterad i Tabell 2a-b i Bilaga 2.

Vi kommer nu att gå tillbaka till steg 2 i *purposeful selection* för att testa om något av samspelen kan uteslutas. Som vi kan se i Tabell 2a i Bilaga 2 är det bara samspelet mellan kön och civilstånd som inte är signifikant. Wald-statistikan antar värdet -1,2 med ett p-värde på 0,238. Därmed utesluts detta samspel ur modellen. Låt oss kalla denna reducerade modell för Modell 2. Modell 2 finns presenterad i Tabell 3a-b i Bilaga 2.

Vi jämför nu parameterskattningarna i Modell 1 och Modell 2 för att se om det råder stora skillnader. Vi kan se att parameterskattningen för huvudeffekten *saknar förvärvsinkomst* ändras från -0,001476 till -0,011721 samt att parameterskattningen för samspelet mellan att vara kommunanställd och civilstånd ändras från -0,000678 till 0,004989. Dessa förändringar är de allra största (på 694 respektive 836 procent), men eftersom parameterskattningarna är nära noll kommer effekten av förändringarna endast att orsaka

en marginell påverkan på responsvariabeln. Vidare ser vi att de tre olika parameterskattningarna för samspelen mellan att vara privatanställd arbetare och logaritmen av lön, att vara man och i åldern 55-60 år samt att vara man och sakna förvärvsinkomst också ändras med mellan 22 och 52 procent. Även dessa förändringar kommer att påverka responsvariabeln marginellt då parameterskattningarna är små (-0,001950, -0,001957 samt 0,015641). Vi bortser därför från dessa förändringar och utesluter samspelet mellan kön och civilstånd ur modellen. Vi väljer alltså att gå vidare med Modell 2.

Vi har nu kommit till steg fyra i *purposeful selection* och testar om samspelet mellan kön och logaritmen av lön nu har blivit signifikant om det läggs till i Modell 2. Ett likelihood kvottest visar att samspelet fortfarande inte kan anses signifikant då teststatistikan antar värdet 1,2 med en frihetsgrad och p-värdet är 0,238. Därmed väljer vi Modell 2 som vår slutligt anpassade modell att gå vidare med.

I vår slutliga modell, se Tabell 3a i Bilaga 2, kan vi se att parameterskattningen framför huvudeffekten *egen företagare* är cirka 8,17. Detta är en stor ökning om vi jämför med motsvarande skattning som är 0,10 i modellen med endast huvudeffekterna, se Tabell 1a i Bilaga 2. Det som orsakar ökningen är då vi lägger till samspelet mellan *logaritmen av lön* och *yrkeskategori* till modellen. Om vi bortser från de övriga samspelen kan vi felaktigt dra slutsatsen att *egna företagare* har ett log-odds som är 8,17 enheter större jämfört med referensnivån som är *privatanställda tjänstemän*. Nu är det emellertid så att vi har en negativ samspelseffekt på cirka -0,79 mellan *logaritmen av lön* och att vara *egen företagare*, vilket innebär att effekten på responsvariabeln samtidigt minskar. Vi åskådliggör detta med ett exempel där gruppen egna företagare med en månadslön på 30 000 kronor jämförs med privatanställda tjänstemän med samma lön. I modellen med endast huvudeffekterna får vi en ökning av responsvariabeln på 0,10 medan vi i modellen med samspelen får en ökning på 8,17 från huvudeffekten och samtidigt en minskning på $0{,}79 \cdot \log(30\,000) \approx 8{,}14$ från samspelseffekten. Modellen med samspelen inkluderade ger alltså sammantaget en ökning på $8{,}17 - 8{,}14 \approx 0{,}03$, vilket är mindre än 0,10.

3.4 Modellutvärdering

Vi har nu valt en slutlig modell som finns presenterad i Tabell 3a-b i Bilaga 2 och det är dags att utvärdera modellen samt testa dess prediktiva förmåga på det externa datasetet.

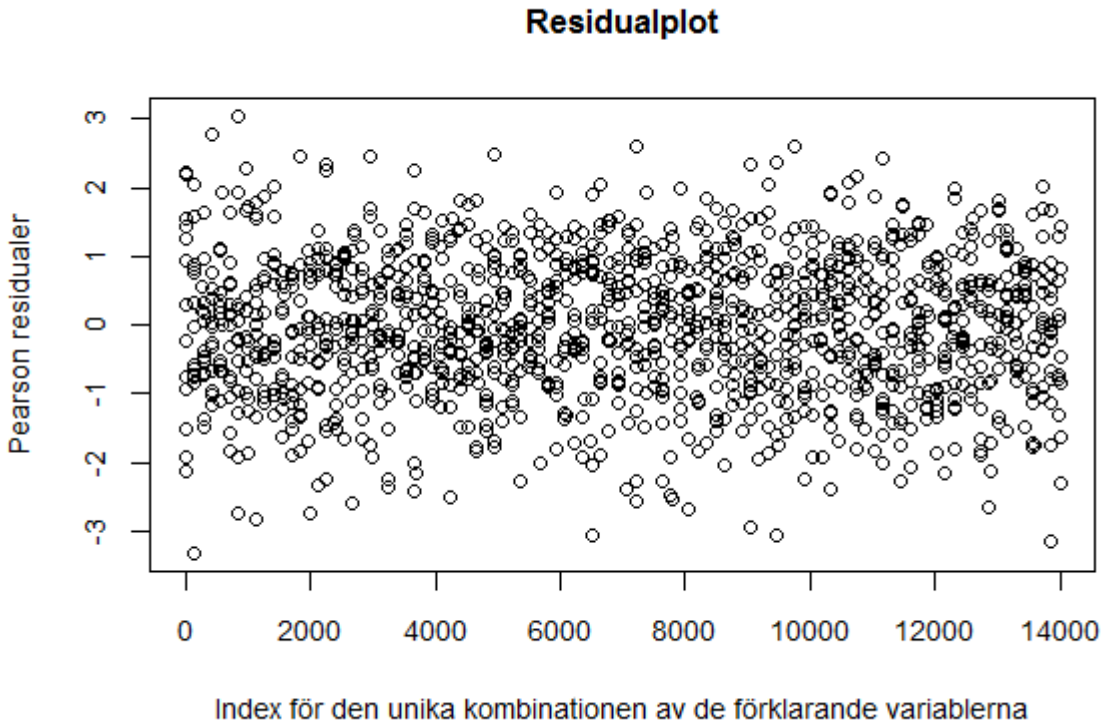
Vi kan se i Tabell 3b i Bilaga 2 att modellanpassningen inte är särskilt bra. Hosmer-Lemeshow goodness of fit statistika, $\hat{C}$, antar ett värde på 17,3 med åtta frihetsgrader och ett p-värde på 0,028 samt att arean under ROC-kurvan endast är 0,6534 vilket visar på en dålig modellanpassning, se Avsnitt 2.5.4 och 2.5.6. I Figur 1 i Bilaga 2 visas också ROC-kurvan. Uppenbarligen finns variation som vi inte lyckas fånga med vår modell, antingen på grund av att vi har fel förklarande variabler eller för att modellen stämmer dåligt på annat sätt. Om vi testar att öka antalet grupper, *g*, till 20 respektive 100 får vi ännu sämre resultat med p-värden på cirka 0,00052 respektive 0,00056. Vi använder även Hosmer-Lemeshow goodness of fit statistikan, C_v, för att testa hur väl vår modell passar i vårt externa dataset, se Avsnitt 2.5.7 för detaljer kring testet. Vi får att teststatistikan, C_v, är cirka 27,4 med åtta frihetsgrader och ett p-värde på cirka 0,00060 då *g* är lika med tio. Att resultatet blir sämre är väntat eftersom vi här testar modellen på data som inte använts för att anpassa modellen efter. Vi prövar om den dåliga modellanpassningen beror på att vi använt för få kombinationer av de förklarande variablerna genom att se hur bra modellen blir om vi anpassar en modell med samtliga huvudeffekter, tvåfaktorsamspel och

trefaktorsamspel. Resultatet från modellutvärderingen av en sådan modell presenteras i Tabell 3.8 nedan. Vi kan se att modellen inte blir nämnvärt bättre om vi lägger till fler parametrar till modellen. Arean under ROC-kurvan ökar bara från 0,6534 till 0,6541, AIC minskar från 356 857 till 356 809 samt att p-värdet från Hosmer-Lemeshow goodness of fit statistika ökar från 0,028 till 0,048 (här använder vi g=10). Vi drar därför slutsatsen att vi inte får bättre modellanpassning genom att lägga till ytterligare parametrar.

Tabell 3.8: *Modellutvärdering för modellen med samtliga huvudeffekter, tvåfaktor-samspel och trefaktorsamspel.*

Antal parametrar i modellen	**AIC**	$\hat{C}$ **med 8 f.g. (p-värde)**	**Area under ROC**
168	356 809	15,6 (0,048)	0,6541

Eftersom vår modell verkar passa dåligt försöker vi utreda om det är för någon eller några kombinationer av de förklarande variablerna som modellanpassningen är särskilt dålig. Vi utgår från teorin som presenteras i Avsnitt 2.5.8 och studerar Pearson residualerna för varje unik kombination av de förklarande variablerna. Logaritmen av lön är vår enda kontinuerliga variabel och den kategoriseras i tio kategorier efter sina 10:e percentiler. På så sätt får vi 1 400 unika kombinationer av de förklarande variablerna att studera. I Figur 3.8 nedan plottas residualerna efter att de sorterats i stigande löneintervall. Det innebär att residualerna med lägst index tillhör den lägsta lönekategorin och residualerna med högst index tillhör den högsta lönekategorin. Vi kan inte se någon speciell kombination som är mycket sämre än de övriga. Vi sorterar även residualerna efter kön, yrkeskategori, civilstånd och ålder, men även här syns ingen grupp som sticker ut (dessa residualplottar redovisas inte då de ser ungefär likadana ut som Figur 3.8).



Figur 3.8: *Pearson residualerna inom varje unik kombination av de förklarande variablerna. Residualerna är sorterade efter stigande löneintervall.*

Trots vår dåliga modellanpassning testar vi, med hjälp av arean under ROC-kurvan, hur väl vår modell predikterar utfallen i vårt externa dataset, se Avsnitt 2.5.7 för detaljer kring testet. Vi får att arean under ROC-kurvan är 0,6525 vilken visar på en dålig prediktiv förmåga.

3.5 Tolkning av modellen

Nu har vi valt en slutlig logistisk regressionsmodell och det är dags att tolka resultatet av modellen. Enligt Avsnitt 2.3.2 ekvation (4) gäller att $e^{\hat{\beta}_j}$ är oddskvoten mellan två kombinationer av de förklarande variablerna där endast den j:te förklarande variabeln ökats med en enhet. I vår modell förekommer samspel och då är det viktigt att tänka på att det inte är möjligt att ändra på endast en förklarande variabel i taget. Till exempel om vi ändrar den förklarande variabeln *kön* kommer inte bara en påverkan från huvudeffekten för kön utan även från samtliga samspel där kön ingår. Vi måste därför, som ekvation (4) visar, även ta med dessa samspelsskattningar i exponenten.

Den största parameterskattningen är den för huvudeffekten *egen företagare* som antar ett värde på cirka 8,17. Effekten av att vara egen företagare, jämfört med referensnivån (som är privatanställd tjänsteman), minskas dock på grund av den negativa samspelseffekten mellan egen företagare och lön, se Avsnitt 3.3 där vi mer i detalj visar detta. Andra stora parameterskattningar är de negativa skattningarna för huvudeffekterna att vara *kommunanställd*, *statligt anställd* och *ogift* samt de positiva skattningarna för

åldersintervallen. Alla dessa skattningar ligger i absolutbelopp mellan cirka 1,14 och 4,30. När vi tolkar parameterskattningarna utgår vi från den förklarande variabelns referensnivå.

Ser vi till parametersskattningen för månadslönen är den på cirka 0,80 vilket innebär att om den logaritmerade månadslönen höjs med en enhet kommer log-oddset att öka med 0,80 om vi utgår från referensnivån. Om månadslönen däremot höjs med flera enheter, från till exempel referensnivån (där månadslönen är noll kronor) till 30 000 kronor, ökar log-oddset med $0{,}80 \cdot \log(30\,000) \approx 8{,}21$. Det vill säga att vi får en stor oddskvot på $e^{8{,}21} \approx 3\,693$. Mer vanligt är att vi vill veta skillnaden mellan en grupp som till exempel har 25 000 kronor i månadslön med en som har 30 000 kronor i månadslön. Vi utnyttjar då ekvation (4) och får att log-oddskvoten är $0{,}80 \cdot [log(30\,000) - \log(25\,000)] \approx 0{,}15$. Det vill säga att oddskvoten är $e^{0{,}15} \approx 1{,}16$. Det innebär att oddset att ha ett privat pensionssparande då månadslönen är 30 000 kronor är 1,16 gånger så stort som oddset att ha ett privat pensionssparande då månadslönen är 25 000 kronor, allt annat lika.

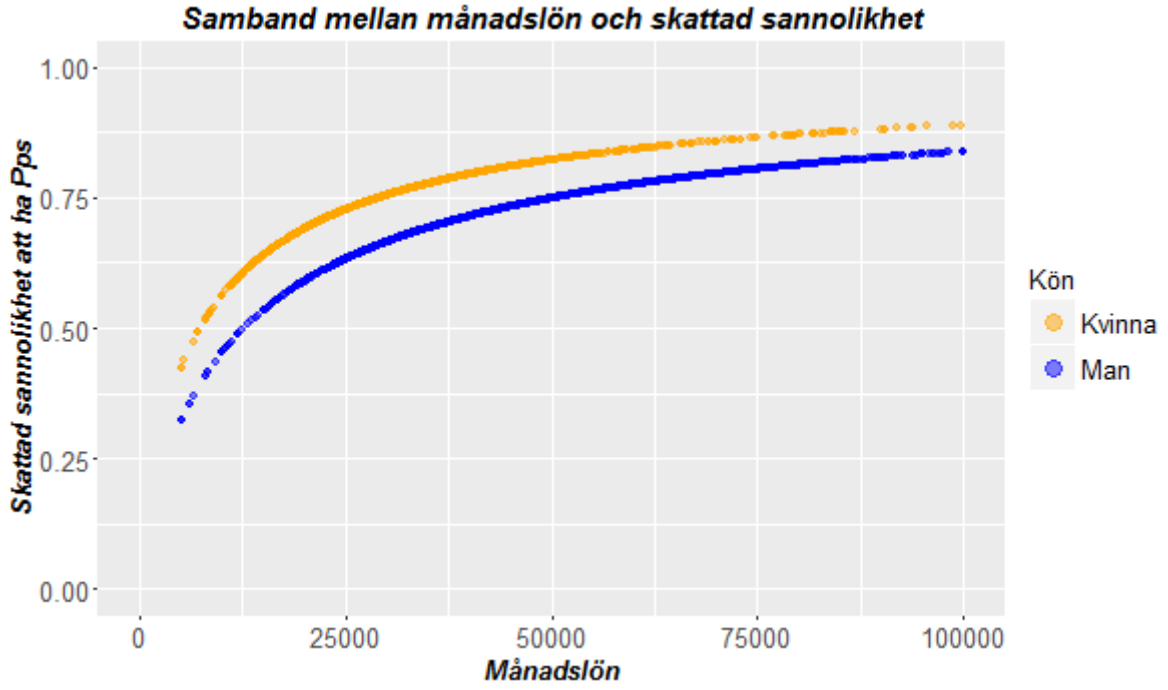
Kön anses vara en viktig förklarande variabel för att skatta sannolikheten att ha ett privat pensionssparande. Vi kan se att parameterskattningen från huvudeffekten är cirka -0,44. Eftersom kvinnor utgör referensnivå innebär den negativa skattningen att män har ett lägre odds än kvinnor att ha ett privat pensionssparande (om vi utgår från referensnivån och allt annat än kön hålls lika). Den negativa effekten, då män jämförs med referensnivån kvinnor, förstärks i de flesta fall något genom samspelen med yrkeskategori. Till exempel kan vi se att en manlig, egen företagare har en samspelseffekt på cirka -0,13.

Om vi ser till civilstånd har vi en stor negativ parameterskattning på cirka -2,30. Om vi jämför gruppen ogifta med referensnivån som är gifta innebär det att oddskvoten är $e^{-2{,}30} \approx 0{,}10$. Ogifta har alltså 0,10 gånger lägre odds än gifta att ha ett privat pensionssparande. Nu kan vi se att samspelet mellan civilstånd och logaritmen av lön är positivt och antar ett värde på cirka 0,20, vilket gör att effekten av att vara ogift minskar. Till exempel är samspelseffekten för en individ med 30 000 kronor i månadslön $0{,}20 \cdot \log(30\,000) \approx 0{,}90$. Vi har också positiva samspelseffekter mellan många av yrkeskategorierna.

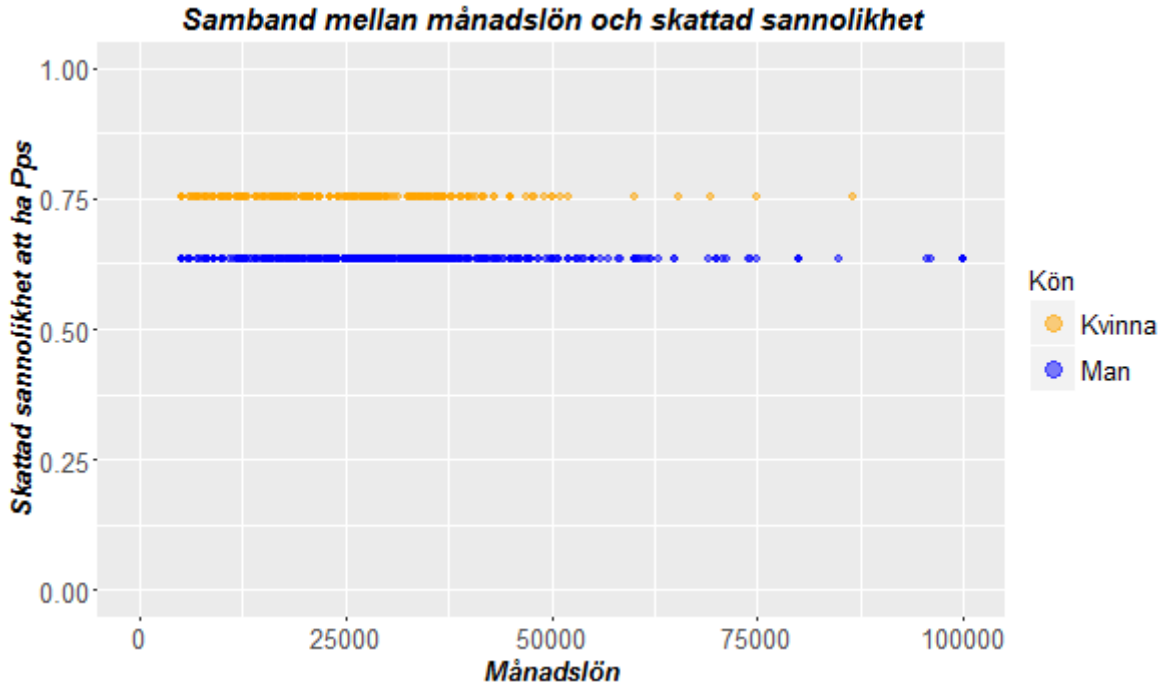
Vi utnyttjar nu ekvation (1) i Avsnitt 2.3.1 och beräknar den skattade sannolikheten att ha ett privat pensionssparande för olika intressanta grupper av individer. I Figur 3.9 presenteras resultatet för gruppen gifta, privatanställda tjänstemän som är 61-64 år, uppdelat på kön och för olika månadslöner. Vi kan se att den skattade sannolikheten att ha ett privat pensionssparande är större för kvinnor än män oavsett lön. Till exempel är den skattade sannolikheten att en kvinna med en månadslön på 40 000 kronor har ett privat pensionssparande cirka 0,79 medan den för en man med samma lön är cirka 0,71. Vi ser också att de skattade sannolikheterna avtar då lönen minskar. Då lönen är 25 000 kronor i månaden är de skattade sannolikheterna för kvinnor respektive män cirka 0,73 respektive 0,63.

Om vi tittar på gruppen gifta, egna företagare som är 61-64 år ser vi ett annat samband, se Figur 3.10. Kvinnorna har fortfarande högre skattade sannolikheter, men lönen verkar inte spela någon roll (sannolikheterna har en mycket liten variation som inte går att urskilja i figuren). Samma svaga samband ser vi för de övriga ålderskategorierna. Tittar vi däremot på ogifta, egna företagare som är 61-64 år avtar den skattade sannolikheten då lönen minskar, se Figur 3.11.

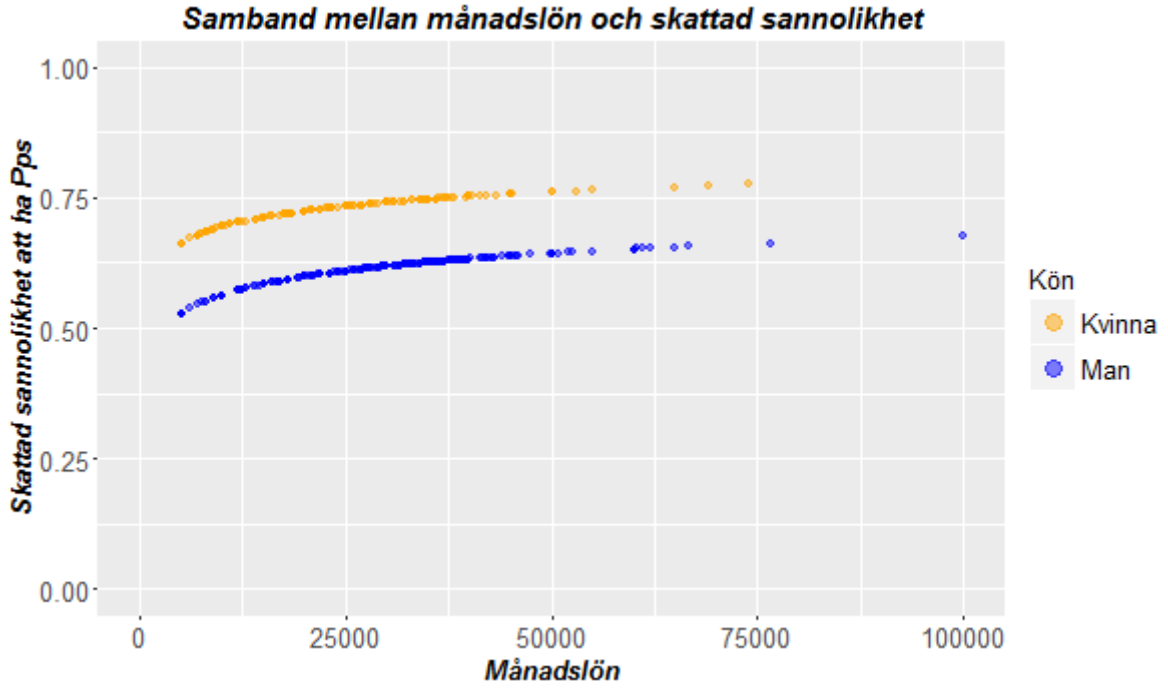
Resultaten för de övriga yrkeskategorierna finns presenterade i Bilaga 3 i Figur 2-6. Där kan vi se att sambandet ser ut ungefär som för de privatanställda tjänstemännen, förutom att de skattade sannolikheterna antar lägre nivåer.



Figur 3.9: *Skattad sannolikhet att ha ett privat pensionssparande för gruppen gifta, privatanställda tjänstemän som är 61-64 år, uppdelat på kön.*



Figur 3.10: *Skattad sannolikhet att ha ett privat pensionssparande för gruppen gifta, egna företagare som är 61-64 år, uppdelat på kön.*



Figur 3.10: *Skattad sannolikhet att ha ett privat pensionssparande för gruppen ogifta, egna företagare som är 61-64 år, uppdelat på kön.*

Om vi istället studerar gruppen ogifta kan vi se att samma samband som i figurerna råder mellan månadslönen och de skattade sannolikheterna, men att nivåerna är lägre (förutom i gruppen egen företagare där sambandet skiljer sig enligt Figur 3.10 och 3.11). Till exempel är den skattade sannolikheten för kvinnliga, ogifta kommunanställda med en månadslön på 25 000 kronor cirka 0,62 medan den för de som är gifta är cirka 0,67. Motsvarande siffror för män är 0,49 respektive 0,55. Om lönen går upp till 35 000 kronor i månaden är motsvarande siffror för kvinnor 0,70 respektive 0,74 och för män 0,59 respektive 0,63. I Tabell 4 i Bilaga 3 presenteras fler skattade sannolikheter för olika grupper efter kön, yrkeskategori och civilstånd för åldrarna 61-64 år och med olika löner. Här visas också skattningarnas konfidensintervall som beräknas enligt teorin i Avsnitt 2.4.2.

3.6 Storleken på pensionen från det privata pensionssparandet

Vi har hittills i rapporten lagt fokus på att analysera förekomsten av ett privat pensionssparande i olika grupper av individer. Det visar sig till exempel att kvinnor har högre sannolikhet att ha ett privat pensionssparande än män och att privatanställda tjänstemän i större utsträckning har ett privat pensionssparande än övriga yrkeskategorier (förutom i låga löneintervall där de egna företagarna har en större sannolikhet att ha ett privat pensionssparande). Vi vet dock fortfarande ingenting om vilken pension dessa grupper i genomsnitt förväntas att få. Det skulle till exempel kunna vara så att kvinnor i större utsträckning sparar privat till sin pension, men i mindre belopp än män. I detta kapitel tittar vi därför på den förväntade pensionen från det privata pensionssparandet för olika grupper av individer.

Vi kommer att simulera en pensionsprognos per individ och sedan studera median-pensionen från det privata pensionssparandet inom olika grupper. För att läsa mer om hur prognoserna beräknas, hänvisas till Bilaga 4. Medianen är beräknad efter de individer inom respektive grupp som faktiskt har ett privat pensionssparande och resultatet presenteras i Tabell 4 i Bilaga 3. Ett samband vi kan se är att medianpensionen ökar med stigande löneintervall. Däremot kan vi i de flesta grupper inte se något tydligt samband mellan kön och civilstånd vad gäller storleken på den prognostiserade pensionen från det privata pensionssparandet. I gruppen egna företagare i det högsta löneintervallet i tabellen, 42 500 till 47 500 kronor i månaden, visar det sig dock att män har en högre medianpension än kvinnor i samma löneintervall. Männens medianpension i denna grupp är även högre än i alla övriga grupper och antar värdena 1 615 respektive 2 151 kronor i månaden för gruppen gifta respektive ogifta. För kvinnliga, egna företagare i samma löneintervall är medianpensionen 1 132 respektive 1 134 kronor i månaden för gruppen gifta respektive ogifta. Dessa nivåer är ungefär samma som för kvinnliga privatanställda tjänstemän och statligt anställda i samma löneintervall. Medianen för de prognostiserade pensionerna för individer med månadslöner i intervallet 22 500 till 27 500 är mellan 557 och 840 kronor i månaden, om vi bortser från de egna företagarna där medianen varierar mellan 1 091 och 1 531 kronor i månaden. I detta löneintervall har gifta kvinnor i samtliga yrkeskategorier, utom statligt anställda, en något högre medianpension, medan motsatt samband gäller för män. I löneintervallet 32 500 till 37 500 varierar medianerna mellan 629 och 1 032 kronor i månaden, om vi även här bortser från de egna företagarna där medianen varierar mellan 1 268 och 1 632 kronor i månaden. Se vidare tabellen i bilagan för ytterligare jämförelser av nivåer.

4 Analys och diskussion

Ett av syftena med denna rapport är att visa på skillnader mellan olika grupper av individer vad gäller förekomsten av ett privat pensionssparande. Det visar sig att vi med våra data kan bekräfta teorin om att kvinnor, gifta samt egna företagare i större utsträckning har ett privat pensionssparande. I studien utgår vi från data från Min Pensions användare och kan därmed inte dra slutsatser om hela Sveriges befolkning, eftersom en snedvridning i urvalet förmodligen förekommer. Detta även fast vi har ett stort antal individer till vårt förfogande. Ett sätt att ta reda på hur stor snedvridningen är kan vara att till exempel göra noggranna analyser av hur Min Pensions användare skiljer sig från Sveriges befolkning som helhet. I denna studie har inte detta gjorts och vi har ingen möjlighet att kvantifiera på vilket sätt resultatet skulle ändras i ett representativt urval. Detta är dock en mycket intressant fråga. En ytterligare begränsning i denna studie är att vi inte kan skilja på individer som har ett aktivt privat pensionssparande som det betalas in premier till idag från de som har ett privat pensionssparande som det inte längre betalas in premier till. I och med att avdragsrätten för privat pensionssparande upphört är dock denna fråga inte längre lika intressant annat än i syfte att studera om det är så att vissa grupper fortsätter att spara privat till sin pension trots att det inte lönar sig längre (detta gäller inte gruppen egna företagare och de som saknar tjänstepension då avdragsrätten fortfarande gäller för dem).

Det visar sig att modellanpassningen för vår slutligt valda modell inte är särskilt bra. En av anledningarna kan vara att vi inte har tillgång till förklarande variabler som kan vara viktiga, så som till exempel intresse för privatekonomiska frågor, utbildning, bostadsregion med mera. Även här finns förbättringspotential för kommande studier. Det finns också möjligheter att behandla de kontinuerliga variablerna lön och ålder med andra metoder än de som använts i denna studie och på så sätt förhoppningsvis få en något bättre modellanpassning. Till exempel kan det visa sig lämpligt att införa indikatorvariabler för vissa lönenivåer där avvikelser från linearitet förekommer alternativt använda så kallade splinefunktioner. Även om vi försöker förbättra modellanpassningen med olika metoder är det förmodligen ändå så att huvudorsaken till vår dåliga modellanpassning är att det förekommer variation i data som våra förklarande variabler inte lyckas fånga.

Referenser

Agresti, A. (2002). *Categorical Data Analysis.* New Jersey: Wiley Series in Probability and Statistics.

Bergström, Å.-P.J., Palme, M., & Persson, M. (2010). *Beskattning av privat pensionssparande.* (Rapport till Expertgruppen för studier i offentlig ekonomi, 2010:2). Regeringskansliet, Finansdepartementet.

Dahmström, K. (2011). *Från datainsamling till rapport.* Lund: Studentlitteratur.

Gujarati, D.N., & Porter, D.C. (2009). *Basic Econometrics.* New York: McGraw-Hill/Irwin.

Hosme, D.W., Jr., Lemeshow, S., & Sturdivant, R.X. (2013). *Applied Logistic Regression.* New Jersey: Wiley Series in Probability and Statistics.

Kleinbaum, D.G., Kupper, L.L., Nizam, A., & Muller, K.E. (2008). *Applied Regression Analysis and Other Multivariable Methods.* Brooks/Cole Cengage Learning.

Petterson, M. (2013): *Är du en person i riskgruppen för att få låg pension?* Pensionsmyndigheten. Hämtad 2016-02-27. http://www.pensionsmyndigheten.se/6187.html

Ross, S. (2006). *A First Course in Probability.* Upper Saddle River, NJ: Pearson Prentice Hall.

Svensk Försäkring. (2015). *Verksamhetsberättelse 2014.* Hämtad 2016-02-29. http://www.svenskforsakring.se/Huvudmeny/Press/#/documents/verksamhetsberaverksve-2014-45458

Tamhane, A.C., & Dunlop, D.D. (2000). *Statistics and Data Analysis: from Elementary to Intermediate.* Upper Saddle River, NJ: Prentice Hall.

Bilaga 1

Här ges en kort introduktion till likelihoodfunktionen och hur den kan användas för att hitta parameterskattningar, chi-tvåfördelningen, normalfördelningen samt informations-matrisen. Teorin går att läsa om i till exempel Ross (2006) samt Tamhane et al. (2000).

Likelihoodfunktionen

Likelihooden för ett stickprov som utgörs av *n* observationer $y_i, = 1,2, \dots, n,$ av den stokastiska variabeln $Y_i,\ i = 1,2, \dots$, ges av stickprovets simultana täthetsfunktion.

Om vi låter $\boldsymbol{\beta}$ vara den uppsättning parametrar som definierar den sannolikhetsfördelning som observationerna kommer från och $L(\boldsymbol{\beta})$ beteckna likelihoodfunktionen gäller att

$$L(\boldsymbol{\beta}) = f_{Y_1,Y_2,\dots,Y_n}(y_1, y_2, \dots, y_n|\boldsymbol{\beta})$$

där $f_{Y_1,Y_2,\dots,Y_n}(y_1, y_2, \dots, y_n|\boldsymbol{\beta})$ är observationernas simultana täthetsfunktion givet parameter-värden enligt $\boldsymbol{\beta}$.

Om observationerna kan anser vara oberoende, vilket våra observationer anses vara, kommer enligt grundläggande sannolikhetsteori

$$L(\boldsymbol{\beta}) = f_{Y_1}(y_1|\boldsymbol{\beta}) \cdot f_{Y_2}(y_2|\boldsymbol{\beta}) \cdot ,,,\cdot f_{Y_n}(y_n|\boldsymbol{\beta}).$$

Det vi gör är att, istället för att betrakta observationerna som slumpmässiga utfall från en fördelning med känd täthetsfunktion, betraktar den simultana täthetsfunktionen som en funktion av den okända parametervektorn $\boldsymbol{\beta}$. Att maximera $L(\boldsymbol{\beta})$ innebär då att vi hittar de parametervärden som gör stickprovets utfall så troligt eller sannolikt som möjligt.

Det visar sig att det i praktiken är enklare att maximera den logaritmerade likelihood-funktionen, den så kallade log-likelihoodfunktionen $l(\boldsymbol{\beta})$. Log-likelihoodfuntionen ges således av

$$l(\boldsymbol{\beta}) = \log\left[f_{Y_1}(y_1|\boldsymbol{\beta}) \cdot f_{Y_2}(y_2|\boldsymbol{\beta}) \cdot ,,,\cdot f_{Y_n}(y_n|\boldsymbol{\beta})\right] = logf_{Y_1}(y_1|\boldsymbol{\beta}) + logf_{Y_2}(y_2|\boldsymbol{\beta}) + \cdots + logf_{Y_n}(y_n|\boldsymbol{\beta}).$$

Maximeringen sker genom att hitta lösningen på det ekvationssystem som sätter de partiella derivatorna av $l(\boldsymbol{\beta})$ lika med noll.

Chi-tvåfördelningen

Chi-tvåfördelningen är en kontinuerlig sannolikhetsfördelning som beskrivs av täthetsfunktionen

$$f(x) = \frac{1}{2^{\upsilon/2}\Gamma(\frac{\upsilon}{2})} e^{-x/2} \cdot x^{\upsilon/2-1}, \qquad \geq 0, \quad \upsilon \in \mathbb{N},$$

där υ är antalet frihetsgrader och

$$\Gamma\left(\frac{\upsilon}{2}\right) = \int_0^\infty e^{-y} y^{\upsilon/2-1} dy.$$

Om vi låter X vara en chi-tvåfördelad stokastisk variabel med υ frihetsgrader gäller att dess väntevärde och varians är

$$E[X] = \upsilon \text{ och}$$
$$Var(X) = 2\upsilon.$$

Vi skriver att $X \sim \chi^2(\upsilon)$.

Normalfördelningen

Normalfördelningen är en kontinuerlig sannolikhetsfördelning som beskrivs av täthetsfunktionen

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \qquad -\infty \leq x \leq \infty,$$

där μ och σ är givna storheter och $\sigma > 0$.

Om vi låter X vara en normalfördelad stokastisk variabel med denna täthetsfunktion gäller att dess väntevärde och varians är

$$E[X] = \mu \text{ och}$$
$$Var(X) = \sigma^2.$$

Vi skriver att $X \sim N(\mu, \sigma^2)$. Speciellt gäller att

$$Z = \frac{X-\mu}{\sigma} \sim N(0,1)$$

och vi säger att Z har den standardiserade normalfördelningen.

Informationsmatrisen

Informationsmatrisen är en matris som består av log-likelihoodfunktionens partiella andraderivator. Dessa partiella derivator har formen

$$\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \beta_j^2} = -\sum_{i=1}^{n} x_{ij}^2 \, \pi_i(\boldsymbol{x}_i)(1 - \pi_i(\boldsymbol{x}_i))$$

och

$$\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \beta_j \partial \beta_l} = -\sum_{i=1}^{n} x_{ij} x_{il} \, \pi_i(\boldsymbol{x}_i)(1 - \pi_i(\boldsymbol{x}_i))$$

för $j, l = 0, 1, 2, \ldots, p$. Här är x_{ij} den i:te observationens j:te förklarande variabel.

Bilaga 2

Här presenteras olika modeller med parameterskattningar, skattningarnas standardfel, teststatistika från Wald-testet som testar om parameterskattningen kan antas vara noll, p-värde från Wald-testet samt olika mått på modellanpassning. I tabellerna framgår vilka kategorier för respektive kategorisk förklarande variabel som utgör referensnivå genom att parameterskattningen där är satt till 0. Vi ser att referensnivån utgörs av kvinnliga, gifta, privatanställda tjänstemän som är 61 till 64 år. Eftersom månadslönen är en kontinuerlig variabel kommer referensnivån här att vara en månadslön på noll kronor. Vi skulle kunna välja en annan referensnivå genom att subtrahera lönen med till exempel medelvärdet, men eftersom vi logaritmerar vill vi inte att negativa löner ska förekomma.

Ett exempel på hur modellen med bara huvudeffekterna ska tolkas:

$logit(P(Y_i = 1|individ\ i\ tillhör\ referensgruppen) = \beta_0$

$logit(P(Y_i = 1|individ\ i\ skiljer\ sig\ från\ referensgruppen\ genom\ att\ kön = man) = \beta_0 + \beta_{Man}$

Om det förekommer samspel i modellen, till exempel mellan kön och civilstånd blir exempelvis:

$logit(P(Y_i = 1|individ\ i\ tillhör\ referensgruppen) = \beta_0$

$logit(P(Y_i = 1|individ\ i\ skiljer\ sig\ från\ referensgruppen\ genom\ att\ kön = man) = \beta_0 + \beta_{Man}$

$logit(P(Y_i = 1|individ\ i\ skiljer\ sig\ från\ referensgruppen\ genom\ att\ civilstånd = ogift) = \beta_0 + \beta_{Ogift}$

$logit(P(Y_i = 1|individ\ i\ skiljer\ sig\ från\ referensgruppen\ genom\ att\ kön = man\ och\ civilstånd = ogift) = \beta_0 + \beta_{Man} + \beta_{Man*Ogift}$

I Tabell 2a kan vi se ett exempel på när skattningarna av $\beta_{Man} = -0{,}428886$, $\beta_{Ogift} = -2{,}375686$ och $\beta_{Man*Ogift} = -0{,}022037$.

Tabell 1a: *Resultatet av den logistiska regressionsmodellen med samtliga huvudeffekter med i modellen. P-värden mindre än 0,001 är markerade med ***.*

Variabel	Skattning ($\hat{\beta}_j$)	Standardfel	Wald-statistika	p-värde
(Intercept)	-5,487538	0,1317	-41,7	***
Kön				
Kvinna	0	-	-	-
Man	-0,362144	0,0091	-39,8	***
log(Lön)	0,626461	0,0127	49,5	***
Yrkeskategori				
Privat tjänsteman	0	-	-	-
Egen företagare	0,103631	0,0177	5,8	***
Kommun	-0,160082	0,0118	-13,6	***
Privat arbetare	-0,200068	0,0118	-16,9	***
Saknar inkomst	-0,369847	0,0627	-5,9	***
Stat	-0,109253	0,0156	-7,0	***
Övriga	-0,331959	0,0177	-18,8	***
Civilstånd				
Gift	0	-	-	-
Ogift	-0,175587	0,0082	-21,4	***
Ålder				
61-64 år	0	-	-	-
16-39 år	-0,916560	0,0134	-68,6	***
40-48 år	-0,095270	0,0133	-7,2	***
49-54 år	0,175991	0,0137	12,8	***
55-60 år	0,154822	0,0130	11,9	***

Tabell 1b: *Antal modellparametrar och mått på modellanpassning för modellen med samtliga huvudeffekter.*

Antal parametrar i modellen	AIC	$\hat{C}$ med 8 f.g. (p-värde)	Area under ROC
14	358 305	41,1 (***)	0,6478

Tabell 2a: *Resultatet av den logistiska regressionsmodellen med samtliga huvudeffekter och alla tvåfaktorsamspel utom mellan kön och logaritmen av lön. P-värden mindre än 0,001 är markerade med ***.*

Variabel	Skattning	Standardfel	Wald-statistika	p-värde
(Intercept)	-7,057116	0,3820	-18,5	***
Kön				
Kvinna	0			
Man	-0,428886	0,0279	-15,4	***
Log(Lön)	0,792903	0,0370	21,4	***
Yrkeskategori				
Privat tjänsteman	0			
Egen företagare	8,160644	0,4452	18,3	***
Kommunanställd	-1,428929	0,3811	-3,7	***
Privat arbetare	-0,476683	0,4111	-1,2	0,246
Saknar inkomst	-0,001476	1,1850	0,0	0,999
Stat	-2,849873	0,5876	-4,9	***
Övriga	0,050320	0,4456	0,1	0,910
Civilstånd				
Gift	0			
Ogift	-2,375686	0,2747	-8,6	***
Ålder				
61-64 år	0			
16-39 år	1,272856	0,4588	2,8	0,006
40-48 år	4,306534	0,4307	10,0	***
49-54 år	2,723263	0,4412	6,2	***
55-60 år	1,147471	0,4079	2,8	0,005
Kön*Yrkeskategori				
Man*Egen företagare	-0,127351	0,0419	-3,0	0,002
Man*Kommun	-0,084663	0,0256	-3,3	***
Man*PrivatArb	0,024956	0,0258	1,0	0,333
Man*Saknar inkomst	-0,015641	0,1303	-0,1	0,904
Man*Stat	-0,022756	0,0333	-0,7	0,494
Man*Övriga	-0,015690	0,0389	-0,4	0,686
Kön*Ålder				
Man*16-39 år	0,250531	0,0307	8,2	***
Man*40-48 år	0,190808	0,0304	6,3	***
Man*49-54 år	0,031611	0,0316	1,0	0,318
Man*55-60 år	-0,001957	0,0302	-0,1	0,948
Kön*Civilstånd				
Man*Ogift	-0,022037	0,0187	-1,2	0,238
Yrkeskategori*Ålder				
Egen företagare*16-39 år	0,352094	0,0655	5,4	***

Fortsättning på nästa sida
Fortsättning på Tabell 2a

Kommun*16-39 år	0,194786	0,0403	4,8	***
Privat arbetare*16-39 år	0,494982	0,0390	12,7	***
Saknar inkomst*16-39 år	-0,197284	0,2433	-0,8	0,417
Stat*16-39 år	0,000916	0,0539	0,0	0,986
Övriga*16-39 år	-0,007597	0,0619	-0,1	0,902
Egen företagare*40-48 år	0,158216	0,0598	2,6	0,008
Kommun*40-48 år	0,086372	0,0385	2,2	0,025
Privat arbetare*40-48 år	0,322598	0,0403	8,0	***
Saknar inkomst*40-48 år	0,101220	0,2190	0,5	0,644
Stat*40-48 år	0,045521	0,0528	0,9	0,389
Övriga*40-48 år	-0,022791	0,0616	-0,4	0,712
Egen företagare*49-54 år	0,058271	0,0610	1,0	0,339
Kommun*49-54 år	0,053272	0,0398	1,3	0,181
Privat arbetare*49-54 år	0,197200	0,0414	4,8	***
Saknar inkomst*49-54 år	0,181628	0,2086	0,9	0,384
Stat*49-54 år	0,014131	0,0559	0,3	0,800
Övriga*49-54 år	0,025688	0,0654	0,4	0,694
Egen företagare*55-60 år	0,034062	0,0595	0,6	0,567
Kommun*55-60 år	0,063961	0,0371	1,7	0,085
Privat arbetare*55-60 år	0,114217	0,0395	2,9	0,004
Saknar inkomst*55-60 år	0,061292	0,1583	0,4	0,699
Stat*55-60 år	0,045686	0,0533	0,9	0,391
Övriga*55-60 år	0,059355	0,0646	0,9	0,358
Log(Lön)*Yrkeskategori				
Log(Lön) *Egen företagare	-0,792208	0,0438	-18,1	***
Log(Lön)*Kommun	0,116691	0,0370	3,2	0,002
Log(Lön)*Privat arbetare	-0,001950	0,0406	0,0	0,962
Log(Lön)*Saknar inkomst	-0,033235	0,1220	-0,3	0,785
Log(Lön)*Stat	0,258409	0,0562	4,6	***
Log(Lön)*Övriga	-0,032697	0,0436	-0,7	0,453
Yrkeskategori*Civilstånd				
Egen företagare*Ogift	0,133346	0,0364	3,7	***
Kommun*Ogift	-0,000678	0,0243	0,0	0,978
PrivatArb*Ogift	0,088368	0,0243	3,6	***
Saknar inkomst*Ogift	-0,036527	0,1295	-0,3	0,778
Stat*Ogift	0,012982	0,0324	0,4	0,689
Övriga*Ogift	-0,064347	0,0368	-1,8	0,080
Log(Lön)*Ålder				
Log(Lön)*16-39 år	-0,230886	0,0441	-5,2	***
Log(Lön)*40-48 år	-0,447353	0,0413	-10,8	***
Log(Lön)*49-54 år	-0,258795	0,0424	-6,1	***
Log(Lön)*55-60 år	-0,104007	0,0394	-2,6	0,008

Fortsättning på nästa sida

Fortsättning på Tabell 2a

Log(Lön)*Civilstånd				
Log(Lön)*Ogift	0,212263	0,0265	8,0	***
Civilstånd*Ålder				
Ogift*16-39 år	-0,219421	0,0278	-7,9	***
Ogift*40-48 år	0,030283	0,0274	1,1	0,269
Ogift*49-54 år	0,065005	0,0283	2,3	0,022
Ogift*55-60 år	0,037325	0,0269	1,4	0,165

Tabell 2b: *Antal modellparametrar och mått på modellanpassning för modellen med samtliga huvudeffekter.*

Antal parametrar i modellen	**AIC**	**$\hat{C}$ med 8 f.g. (p-värde)**	**Area under ROC**
70	356 858	16,1 (0,042)	0,6534

Tabell 3a: *Resultatet av den logistiska regressionsmodellen med samtliga huvudeffekter och alla tvåfaktorsamspel, utom mellan kön och logaritmen av lön samt kön och civilstånd. P-värden mindre än 0,001 är markerade med ***.*

Variabel	Skattning	Standardfel	Wald-statistika	p-värde
(Intercept)	-7,092214	0,3809	-18,6	***
Kön				
Kvinna	0			
Man	-0,437600	0,0269	-16,3	***
Log(Lön)	0,796815	0,0369	21,6	***
Yrkeskategori				
Privat tjänsteman	0			
Egen företagare	8,165896	0,4452	18,3	***
Kommunanställd	-1,434111	0,3811	-3,8	***
Privat arbetare	-0,483341	0,4111	-1,2	0,2397
Saknar inkomst	-0,011721	1,1851	0,0	0,9921
Stat	-2,848873	0,5876	-4,8	***
Övriga	0,048642	0,4456	0,1	0,9131
Civilstånd				
Gift	0			
Ogift	-2,297237	0,2666	-8,6	***
Ålder				
61-64 år	0			
16-39 år	1,253943	0,4585	2,7	0,0062
40-48 år	4,302093	0,4307	10,0	***
49-54 år	2,716927	0,4412	6,2	***
55-60 år	1,143971	0,4079	2,8	0,0050
Kön*Yrkeskategori				
Man*Egen företagare	-0,126706	0,0419	-3,0	0,0025
Man*Kommun	-0,084627	0,0256	-3,3	***
Man*Privat arbetare	0,023017	0,0257	0,9	0,3708
Man*Saknar inkomst	-0,019016	0,1303	-0,1	0,8840
Man*Stat	-0,022877	0,0333	-0,7	0,4919
Man*Övriga	-0,015793	0,0389	-0,4	0,6844
Kön*Ålder				
Man*16-39 år	0,245528	0,0304	8,1	***
Man*40-48 år	0,189739	0,0304	6,3	***
Man*49-54 år	0,030265	0,0316	1,0	0,3385
Man*55-60 år	-0,002460	0,0302	-0,1	0,9351
Yrkeskategori*Ålder				
Egen företagare*16-39 år	0,353548	0,0655	5,4	***
Kommun*16-39 år	0,193839	0,0403	4,8	***
Privat arbetare*16-39 år	0,496635	0,0390	12,7	***

Fortsättning på nästa sida

Fortsättning på Tabell 3a

Saknar inkomst*16-39 år	-0,196872	0,2433	-0,8	0,4184
Stat*16-39 år	0,000801	0,0539	0,0	0,9881
Övriga*16-39 år	-0,007417	0,0619	-0,1	0,9046
Egen företagare*40-48 år	0,158555	0,0598	2,7	0,0080
Kommun*40-48 år	0,086337	0,0385	2,2	0,0249
Privat arbetare*40-48 år	0,322968	0,0403	8,0	***
Saknar inkomst*40-48 år	0,100853	0,2190	0,5	0,6451
Stat*40-48 år	0,045578	0,0528	0,9	0,3881
Övriga*40-48 år	-0,022800	0,0616	-0,4	0,7114
Egen företagare*49-54 år	0,058679	0,0610	1,0	0,3361
Kommun*49-54 år	0,053184	0,0398	1,3	0,1814
Privat arbetare*49-54 år	0,197590	0,0414	4,8	***
Saknar inkomst*49-54 år	0,182133	0,2085	0,9	0,3825
Stat*49-54 år	0,014115	0,0559	0,3	0,8006
Övriga*49-54 år	0,025629	0,0654	0,4	0,6950
Egen företagare*55-60 år	0,034247	0,0595	0,6	0,5651
Kommun*55-60 år	0,064016	0,0371	1,7	0,0845
Privat arbetare*55-60 år	0,114309	0,0395	2,9	0,0038
Saknar inkomst*55-60 år	0,061055	0,1583	0,4	0,6997
Stat*55-60 år	0,045766	0,0533	0,9	0,3906
Övriga*55-60 år	0,059314	0,0646	0,9	0,3584
Log(Lön)*Yrkeskategori				
Log(Lön)*Egen företagare	-0,792570	0,0438	-18,1	***
Log(Lön)*Kommun	0,116950	0,0370	3,2	0,0016
Log(Lön)*Privat arbetare	-0,000944	0,0406	0,0	0,9815
Log(Lön)*Saknar inkomst	-0,031855	0,1220	-0,3	0,7940
Log(Lön)*Stat	0,258280	0,0562	4,6	***
Log(Lön)*Övriga	-0,032519	0,0436	-0,7	0,4559
Yrkeskategori*Civilstånd				
Egen företagare*Ogift	0,127960	0,0361	3,5	***
Kommun*Ogift	0,004989	0,0238	0,2	0,8338
PrivatArb*Ogift	0,082870	0,0239	3,5	***
Saknar inkomst*Ogift	-0,040558	0,1294	-0,3	0,7540
Stat*Ogift	0,014146	0,0324	0,4	0,6623
Övriga*Ogift	-0,064936	0,0368	-1,8	0,0773
Log(Lön)*Ålder				
Log(Lön)*16-39 år	-0,228822	0,0441	-5,2	***
Log(Lön)*40-48 år	-0,446933	0,0413	-10,8	***
Log(Lön)*49-54 år	-0,258180	0,0424	-6,1	***
Log(Lön)*55-60 år	-0,103678	0,0394	-2,6	0,0084
Log(Lön)*Civilstånd				
Log(Lön)*Ogift	0,203474	0,0254	8,0	***

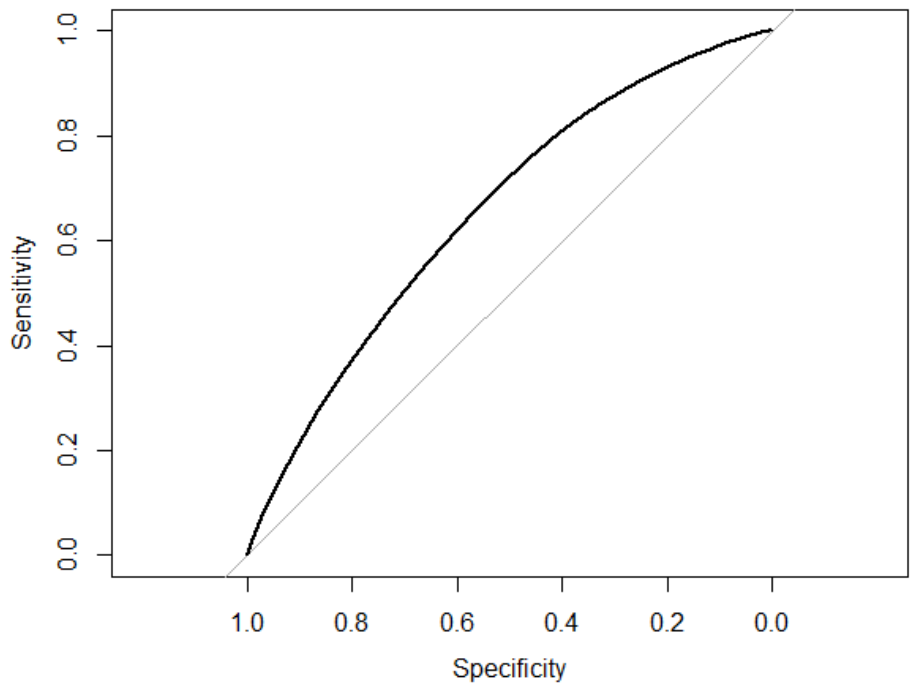
Fortsättning på nästa sida

Fortsättning på Tabell 3a

Civilstånd*Ålder				
Ogift*16-39 år	-0,220141	0,0278	-7,9	***
Ogift*40-48 år	0,030475	0,0274	1,1	0,2659
Ogift*49-54 år	0,065556	0,0283	2,3	0,0205
Ogift*55-60 år	0,037453	0,0269	1,4	0,1640

Tabell 3b: *Antal modellparametrar och mått på modellanpassning för modellen med samtliga huvudeffekter.*

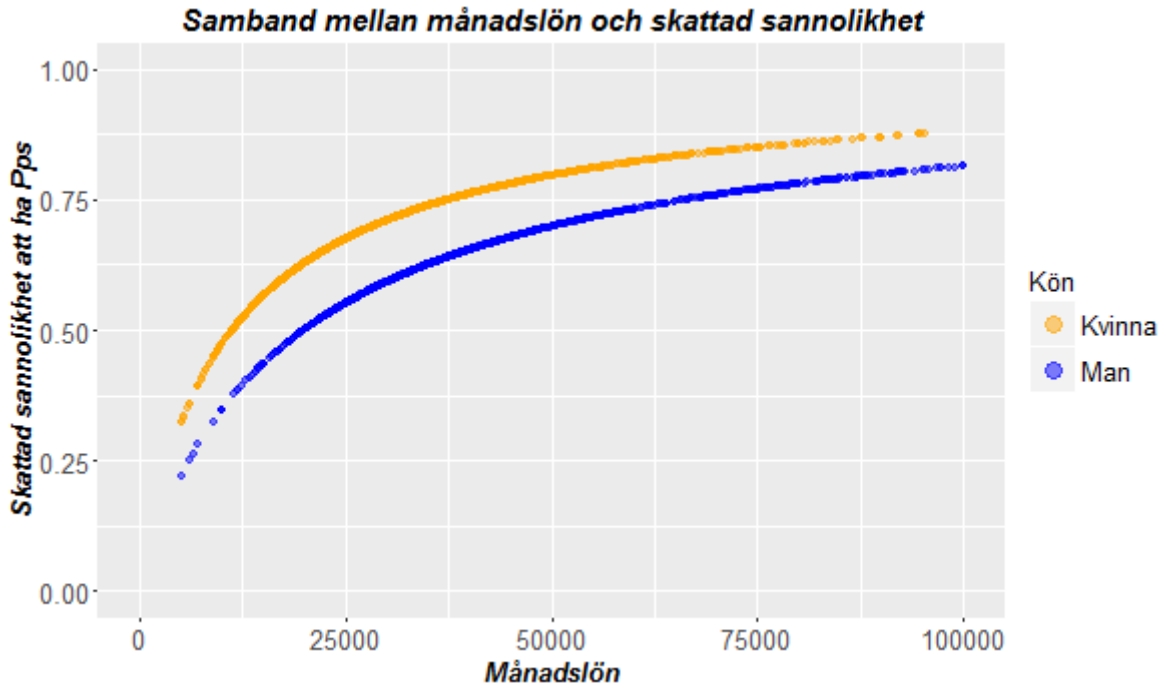
Antal parametrar i modellen	**AIC**	$\hat{C}$ **med 8 f.g. (p-värde)**	**Area under ROC**
69	356 857	17,3 (0,028)	0,6534



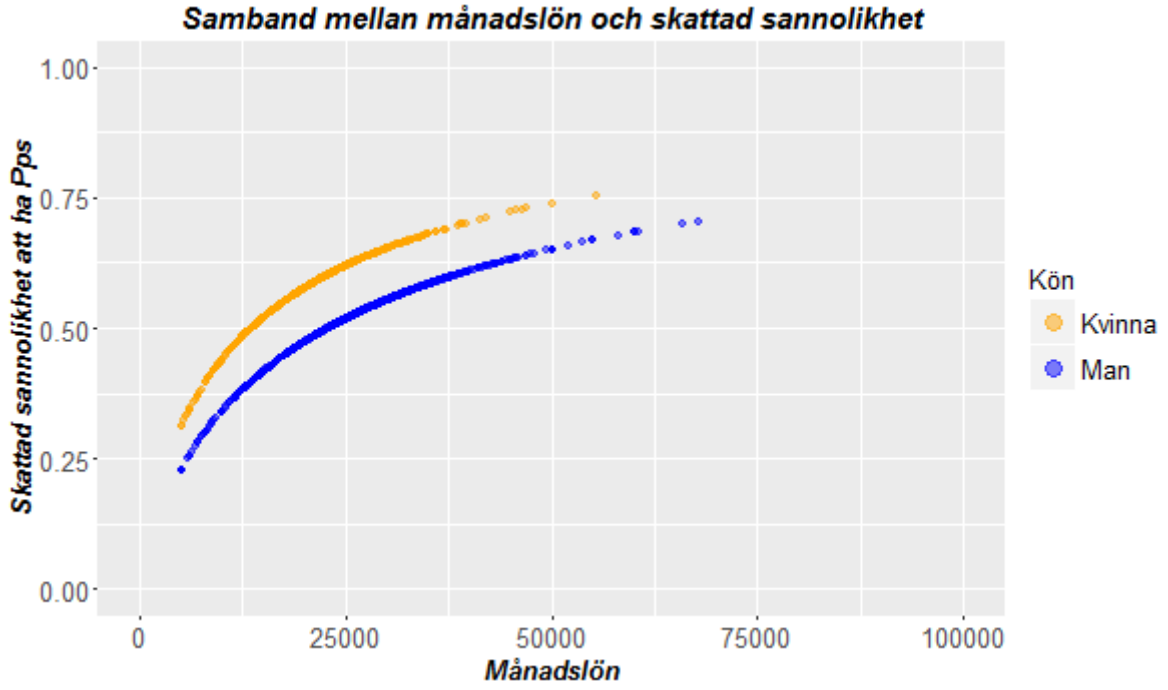
Figur 1: *Sensitiviteten plottad mot specificiteten för modellen i Tabell 3a.*

Bilaga 3

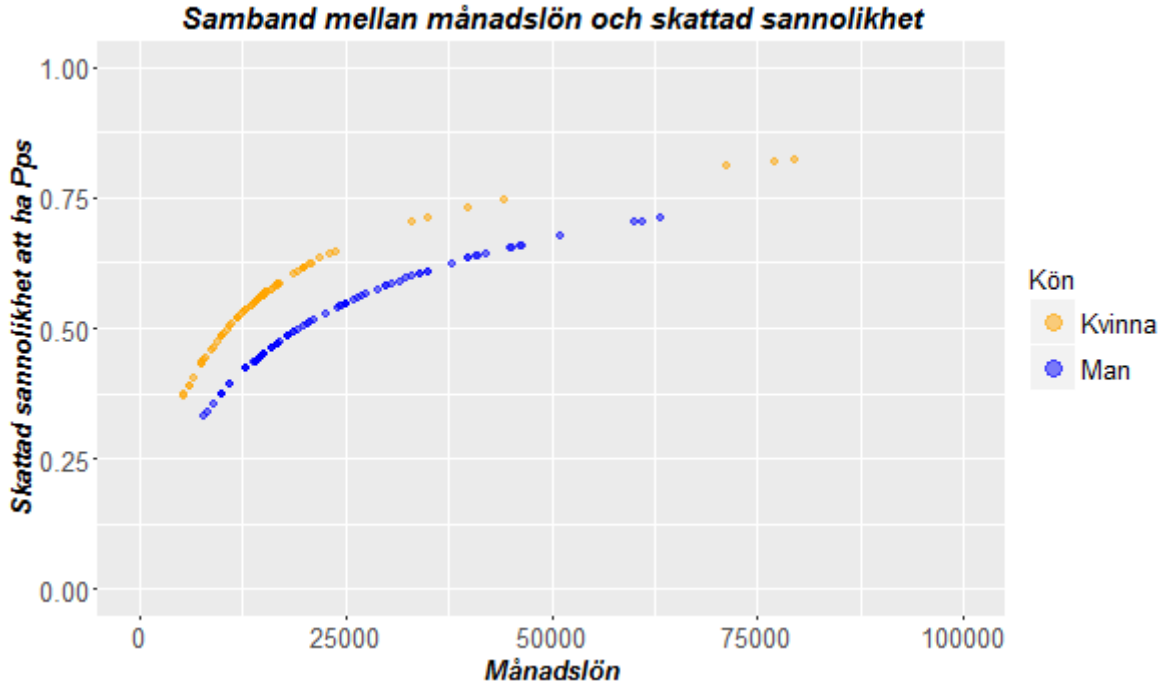
Här presenteras olika plottar som beskriver den skattade sannolikheten att ha ett privat pensionssparande i olika grupper, uppdelade på kvinnor och män.



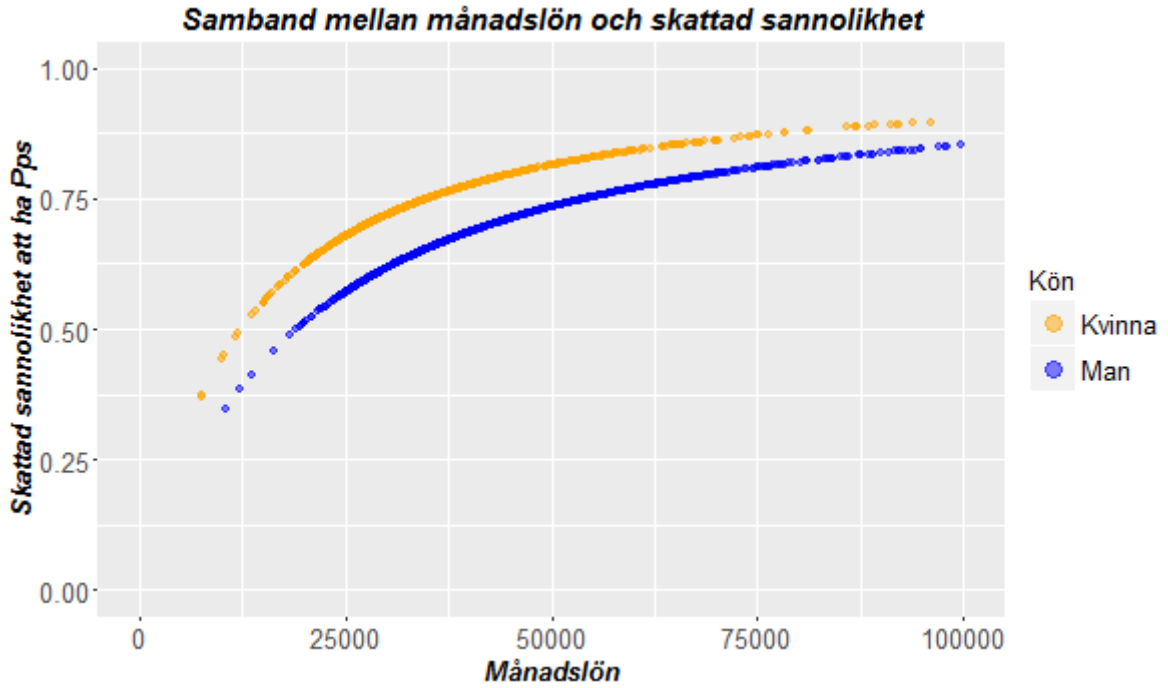
Figur 2: *Skattad sannolikhet att ha ett privat pensionssparande (Pps) för gruppen gifta, kommunanställda som är 61-64 år, uppdelat på kön.*



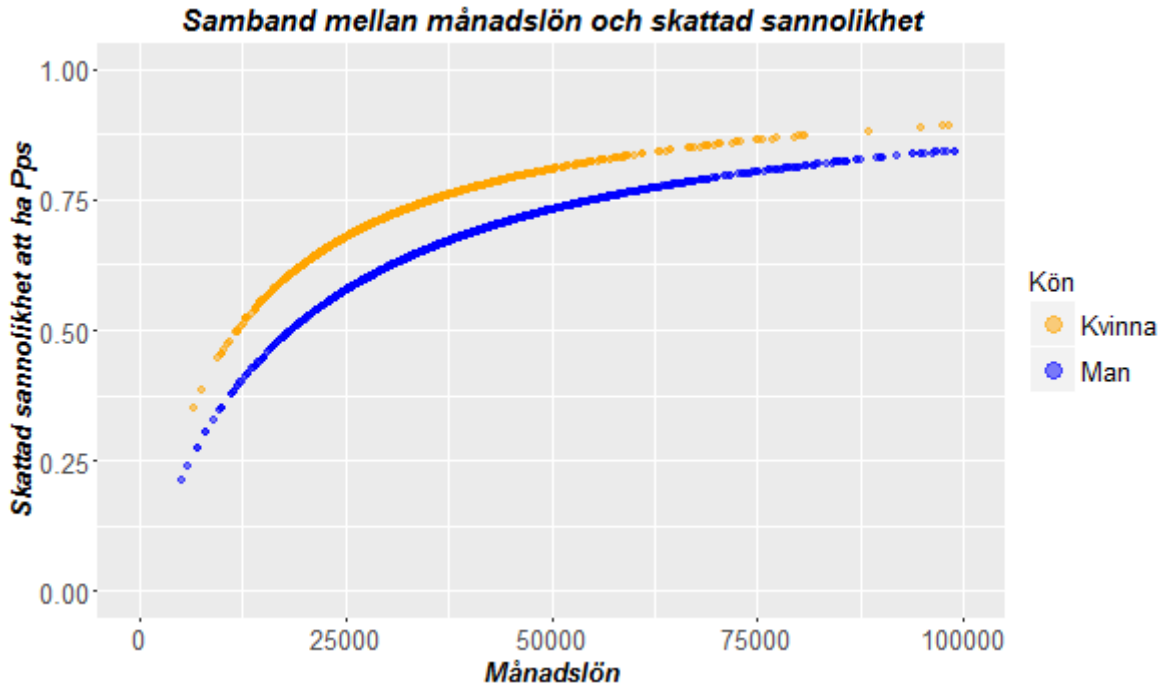
Figur 3: *Skattad sannolikhet att ha ett privat pensionssparande (Pps) för gruppen gifta, privatanställda arbetare som är 61-64 år, uppdelat på kön.*



Figur 4: *Skattad sannolikhet att ha ett privat pensionssparande (Pps) för gruppen gifta som saknar förvärvsinkomst och är 61-64 år, uppdelat på kön.*



Figur 5: *Skattad sannolikhet att ha ett privat pensionssparande (Pps) för gruppen gifta, statligt anställda som är 61-64 år, uppdelat på kön.*



Figur 6: *Skattad sannolikhet att ha ett privat pensionssparande (Pps) för gruppen ogifta, privatanställda tjänstemän som är 61-64 år, uppdelat på kön.*

Tabell 4: *Skattade sannolikheter att ha ett privat pensionssparande samt skattningens 95 procentiga konfidensintervall i olika grupper av individer efter yrkeskategori, månadslön, kön och civilstånd. Medianen för de som har ett privat pensionssparande inom respektive grupp beräknas och redovisas som Median Pps. Här är medianen beräknad för månadslöner i ett intervall om +/- 2 500 från den nivå som redovisas i tabellen.*

Yrke	Månadslön	Kön	Civilstånd	$\hat{\pi}$ (K.I)	Median Pps
Privat arb.	25 000	Kvinna	Gift	0,62 (0,60; 0,63)	680
			Ogift	0,58 (0,57; 0,60)	557
		Man	Gift	0,52 (0,51; 0,53)	558
			Ogift	0,48 (0,47; 0,49)	598
	35 000	Kvinna	Gift	0,68 (0,66; 0,70)	629
			Ogift	0,66 (0,64; 0,68)	780
		Man	Gift	0,58 (0,57; 0,60)	777
			Ogift	0,56 (0,55; 0,58)	741
Kommun	25 000	Kvinna	Gift	0,67 (0,66; 0,68)	700
			Ogift	0,62 (0,61; 0,63)	619
		Man	Gift	0,55 (0,54; 0,57)	620
			Ogift	0,49 (0,48; 0,51)	662
	35 000	Kvinna	Gift	0,74 (0,73; 0,75)	994
			Ogift	0,70 (0,69; 0,72)	836
		Man	Gift	0,63 (0,61; 0,64)	834
			Ogift	0,59 (0,57; 0,60)	936
Privat tjm.	25 000	Kvinna	Gift	0,73 (0,72; 0,74)	840
			Ogift	0,68 (0,66; 0,69)	677
		Man	Gift	0,63 (0,62; 0,64)	648
			Ogift	0,58 (0,56; 0,59)	730
	35 000	Kvinna	Gift	0,78 (0,77; 0,79)	958
			Ogift	0,75 (0,73; 0,76)	891
		Man	Gift	0,69 (0,68; 0,70)	846
			Ogift	0,65 (0,64; 0,67)	990
	45 000	Kvinna	Gift	0,81 (0,80; 0,82)	1 175
			Ogift	0,79 (0,78; 0,80)	1 187
		Man	Gift	0,73 (0,72; 0,74)	1 127
			Ogift	0,71 (0,70; 0,72)	1 146

Fortsättning på nästa sida

Fortsättning på Tabell 4

Stat	25 000	Kvinna	Gift	0,68 (0,66; 0,70)	673
			Ogift	0,63 (0,61; 0,65)	689
		Man	Gift	0,57 (0,55; 0,59)	738
			Ogift	0,52 (0,49; 0,54)	573
	35 000	Kvinna	Gift	0,75 (0,73; 0,77)	856
			Ogift	0,72 (0,70; 0,74)	837
		Man	Gift	0,65 (0,64; 0,67)	863
			Ogift	0,62 (0,60; 0,64)	1 032
	45 000	Kvinna	Gift	0,80 (0,78; 0,81)	1 246
			Ogift	0,78 (0,76; 0,80)	1 005
		Man	Gift	0,71 (0,69; 0,73)	1 034
			Ogift	0,69 (0,67; 0,71)	1 310
Företag	25 000	Kvinna	Gift	0,75 (0,73; 0,77)	1 352
			Ogift	0,73 (0,71; 0,75)	1 091
		Man	Gift	0,63 (0,61; 0,65)	1 342
			Ogift	0,61 (0,59; 0,63)	1 531
	35 000	Kvinna	Gift	0,75 (0,73; 0,77)	1 262
			Ogift	0,75 (0,72; 0,77)	1 515
		Man	Gift	0,64 (0,61; 0,66)	1 638
			Ogift	0,63 (0,60; 0,65)	1 629
	45 000	Kvinna	Gift	0,75 (0,73; 0,78)	1 132
			Ogift	0,76 (0,73; 0,78)	1 134
		Man	Gift	0,64 (0,61; 0,66)	1 615
			Ogift	0,64 (0,61; 0,66)	2 151

Bilaga 4

Här presenteras kort hur Min Pension beräknar den pensionsprognos som används för att bestämma den förväntade pensionen från det privata pensionssparandet.

I prognosberäkningarna följer Min Pension den *Standard för pensionsprognoser* som finns utarbetad av pensionsbranschen som ett led i arbetet att hitta ett gemensamt sätt att beräkna pensionsprognoser. Standarden ska svara på frågan "Om jag fortsätter att arbeta och spara som i dag, hur mycket pension per månad kan jag då förväntas få?". I huvudsak är tanken att den inkomst och det sparande som en individ har idag fortsätter att gälla fram till pension. Resultatet av pensionsprognosen presenteras i dagens pris- och löneläge vilket innebär att inflation och inkomsttillväxt är satt till noll. Detta görs för att den prognostiserade pensionen ska blir lättare att tolka och inte missuppfattas av användaren. Vidare är kapitalavkastningen, utöver löneutvecklingen, satt till 2,1 procent. Standarden och dess förarbeten finns att hämta på bland annat pensionsmyndigheten.se för den som är intresserad av att läsa mer.

Innan själva prognossimuleringen startar sätter vi

- kapitalavkastningen till 2,1 procent
- inkomstutvecklingen till noll procent
- livsvarig eller längsta uttagstid på alla försäkringar
- pensionsålder 65 år.

Att alla individer har samma inställningar gör att vi lättare kan jämföra grupper med varandra.

Prognoserna kommer att presenteras uppdelade på *allmän pension*, *tjänstepension* och *pension från det privata pensionssparandet*. I denna studie kommer vi endast att titta på den sist nämnda pensionsformen.