Stockholms universitet

Skattningar av övergångssannolikheter i en demografimodell med tillämpning på Pensionsmodellen

Anders Carlsson

Masteruppsats i försäkringsmatematik
Master Thesis in Actuarial Mathematics

Masteruppsats 2019:3
Försäkringsmatematik
Juni 2019

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Matematisk statistik
Stockholms universitet
Masteruppsats **2019:3**
http://www.math.su.se/matstat

Skattningar av övergångssannolikheter i en demografimodell med tillämpning på Pensionsmodellen

Anders Carlsson*

Juni 2019

Sammanfattning

Pensionsmodellen beskriver hur den svenska befolkningen tjänar in och tar ut pension och hur detta påverkar inkomstpensionens ekonomiska ställning på sikt. Hur stor inkomst en individ kommer att ha i modellen bestäms av vilka egenskaper individen har. Exempel på dessa egenskaper är huruvida individen har barn, får socialtransfereringar, studerar, har börjat ta ut sin i pension mm. Varje individ har en viss kombination av dessa egenskaper vid varje tidpunkt. I den här uppsatsen jämför vi metoder för att skatta övergångssannolikheter mellan utfall av dessa egenskaper. Det finns redan en metod för detta i dagens pensionsmodell som i princip bygger på helt empiriska skattningar men i den här uppsatsen undersöker vi om det finns en bättre metod. De egenskaper som studeras har alla värdena sant/1 eller falskt/0. Vi betecknar antal egenskaper n_e och i vårt data är $n_e = 14$. En kombination av dessa benämns $x_\bullet = (x_1, x_2, \ldots, x_{n_e})$ och en övergång från ett år till nästa sker mellan två vektorer av egenskaper hos en individ $x_{\bullet,0} \to x_{\bullet,1}$, där vi säger att tiden 0 är starten av övergången och tiden 1 infaller ett år senare. Data som innehåller 9.7 miljoner individer där även uppgifter om kön (s), ålder (a) och grupp av födelseländer (o) finns med visar på ett stort antal möjliga övergångar. Det finns indikationer på att dessa övergångar ibland sker oberoende av varandra men att det mellan vissa egenskaper finns ett visst beroende. I den här uppsatsen kommer vi att dela upp övergångarna i en kedja av betingade sannolikheter vilket gör det möjligt att modellera ett enskilt värde $x_{k,1}$ åt gången. Detta gör att vi kan använda vanliga klassificeringsmodeller. Kedjan av betingningar ser ut på följande sätt:

- Utfallet $x_{1,1}$ modelleras betingat på s, a, o och utfallet $x_{\bullet,0}$,
- Utfallet $x_{2,1}$ modelleras betingat på s, a, o och utfallet $x_{\bullet,0}, x_{1,1}$
- Utfallet $x_{3,1}$ modelleras betingat på s, a, o och utfallet $x_{\bullet,0}, x_{1,1}, x_{2,1}$
- o.s.v.

De olika metoder som jämförs för att skatta p i dessa fördelningar är förutom de empiriska skattningarna även logistisk regression, splineregression, lassoregression och random forest. Vi finner att en modell där beroendet till kategorivariablerna modelleras med logistisk regression och beroendet till ålder modelleras med kubiska spline är ett bra alternativ.

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige.
E-post: mail@carlssonanders.se. Handledare: Mathias Lindholm.

Innehåll

1 Inledning **6**
1.1 Det allmänna pensionssystemet 6
1.2 Premiepensionen 6
1.3 Inkomstpensionen 7

2 Pensionsmodellen **8**
2.1 Notation 10
2.2 Pensionsmodellens delar 11
2.3 Typer av inkomster 12

3 Data **13**
3.1 Beroende mellan övergångar 14
3.2 Pearson χ^2-test 15
3.3 Uppdelning i träningsdata och testdata 16
3.4 Korsvalidering 16

4 Teori **18**
4.1 Dödlighet hanteras separat 18
4.2 Metod för att ta hänsyn till beroende 18
4.3 Ordningen av kateogrivariablerna 19

5 Modell **20**
5.1 Empiriska skattningar 21
5.1.1 Empriska skattning på $\mathcal{M}^{\text{train}}$ och $\mathcal{M}^{\text{test}}$ 22
5.2 Logistisk regression 22
5.2.1 Maximum Likelihood-skattning och deviance 23
5.3 Lasso 24
5.4 Kubiska spline 25
5.5 Logistisk regression med interaktioner 26
5.6 Random Forest 27
5.7 Skattade parametrar med logistisk regression 30

6 Kalibrering av sannolikheter **32**

7 Utvärdering av modeller **33**
7.1 Residual deviance 34
7.2 Fel i kronor 35

7.3 Klassificeringsfel 36
7.4 Jämförelse av modeller 37

8 Diskussion **39**

9 Appendix **40**
9.1 Diagram 40
9.1.1 Förekomster av egenskaper i data 40
9.1.2 Anpassning av spline-funktion 47
9.1.3 Uppskattning av fel 55
9.2 Omskrivning av parameterisering logit/multinomial 73

Tack

Jag vill tacka min handledare Mathias Lindhom Millberg för all hjälp och tålamod. Jag vill även tacka min arbetsgivare Pensionsmyndigheten för den tid som givits mig för att skriva här uppsatsen.

1 Inledning

Den här uppsatsen kommer att studera delar av den s.k. Pensionsmodellen som används av Pensionsmyndigheten. Pensionsmodellen används för att modellera inkomst- och premiepensions-systemets ekonomiska ställning genom att simulera framtida utfall.

1.1 Det allmänna pensionssystemet

Det svenska allmänna pensionssystemet omfattar alla som arbetar och betalar skatt i Sverige, se [4] s. 16. Så länge en individs arbetsinkomst överstiger ett tröskelvärde på 42,3 procent av ett inkomstbasbelopp räknas alla inkomster under intjänandetaket på 8,09 inkomstbasbelopp in i det s.k. pensionsunderlaget. Alla som jobbar oavsett ålder tjänar in till allmän pension, även de som börjat ta ut pensionen. På pensionsunderlaget betalas 18,5 procent in i premier (kallas för *pensionsrätter* i den allmänna pensionen) till det allmänna pensionssystemet via arbetsavgifter och den allmänna pensionsavgiften. Av dessa pengar går 16 procent till inkomstpensionen och 2,5 procent till premiepensionen. Utöver inkomstgrundad pension finns även ett grundskydd i form av garantipension, bostadstillägg och äldreförsörjningsstöd som inte grundar sig på individens tidigare inkomster utan har som syfte att ge en rimlig lägsta nivå av levnadsstandard för de med lägre inkomster.

1.2 Premiepensionen

Premiepensionssystemet liknar i hög grad en vanlig premiebestämd pensionsförsäkring. Insättningarna sätts i de fonder som individen valt eller i ett förvalsalternativ. När individen väljer att ta ut sin pension räknas månadsbeloppet ut som pensionskapitalet (kallas även pensionsbehållningen) dividerat med ett s.k. delningstal och 12. Delningstalet beror på den förväntade livslängden och även på diskonteringsräntan som kallas förskottsränta, se [4] s. 96. I premiepensionen är det alltså individen som sparar ihop till sin egen pension och pengarna öronmärks för individen, samma kapital används sedan till pensionsutbetalningar. Ett sådant system kallas för fonderat då det hela tiden finns placeringstillgångar som täcker upp den skuld som pensionssystemet i någon mening har mot individen.

1.3 Inkomstpensionen

Inkomstpensionssystemet är ett s.k. Pay-As-You-Go-system som i huvudsak är icke-fonderat. Det innebär att de pengar som betalas in visserligen skrivs upp på ett individuellt konto men själva pengarna sparas inte för individens räkning utan används för att betala ut pensioner till dagens pensionärer. Kontobehållning förräntas med hjälp av s.k. inkomstindexering vilket är ett medelvärde för hur den arbetande befolkningens inkomster förändras. Efter att individen gått i pension måste det finnas en ny arbetande befolkning som står för de pengar som behövs till pensionsutbetalningarna individen har tjänat in. I det svenska inkomstpensionssystemet finns dock s.k. buffertfonder för att balansera skillnaden i storlek mellan de olika årskullarna. Precis som i ett fonderat system finns en balansräkning som ska ge en bild över systemets finansiella ställning, se [4] s. 102. Skuldsidan består av summan av det som finns på individernas pensionskonton hos de som är förvärvsaktiva, SA, och skulden till dagens pensionärer, SP. På tillgångssidan finns dels buffertfonder, BF men den största tillgången är den så kallade avgiftstillgången, AT, vilket är en uppskattning av de framtida inbetalningarna till systemet som förväntas göras av den aktiva befolkningen. Kvoten av tillgångar och skulderna kallas för balanstal:

$$BT = \frac{BF + AT}{SA + SP}$$

Vid slutet av året 2018 räknades BT ut till:

$$BT(2018) = \frac{1383 + 8244}{5724 + 3441} = \frac{9627}{9165} = 1.05$$

där BF, AT, SA, SP angivits i enheten miljarder kronor. Är balanstalet under 1 går systemet in i s.k. *balanserings*-period vilket innebär att indexeringen av pensionerna och kontobehållningarna sänks, de kan till och med minska nominellt, detta gör att $SA + SP$ minskar och BT stiger, se [4] s. 17. Eftersom talet BT är något som kan sänka de utgående pensionerna är det viktigt att analysera. Ett typiskt fonderat system går att modellera i run-off, d.v.s. genom att anta att det inte kommer in några nya premier. Systemet ska ändå fungera genom att betala ut de innestående pengarna som pensioner. Ett icke-fonderat system som inkomstpensionssystemet upphör i någon mening att fungera om avgifterna skulle upphöra, efter att buffertfonderna tagit slut. Därför är det av extra vikt att modellera hur den förväntade strömmen

av framtida avgifter ser ut. Tidshorisonten i prognosen kan egentligen göras hur lång som helst då nya pensionsavgifter både går till att betala ut dagens pensioner men orsakar också uppbyggande av en ny skuld som framtidens arbetande befolkning får betala. Pensionsmyndigheten har ett uppdrag att årligen redovisa inkomstpensions-systemets ekonomiska ställning och göra prognoser i tre scenarier, se [4] sidan 49. Ett basscenario där så realistiska antaganden som möjligt görs, detta kompletteras med ett pessimistiskt och ett optimistiskt scenario. Exempel på orsaker till att inkomstpensionens balansräkning försämras är en svag arbetsmarknad eller låg pensionsålder.

2 Pensionsmodellen

Pensionsmyndigheten har valt att basera sina prognoser på en modell som i grunden drivs av både historisk data på individnivå såväl som externa prognoser på makronivå. Pensionsmodellen är skapad av Pensionsmyndigheten, modellen inklusive alla antagande är publik och finns utlagd på Pensionsmyndighetens hemsida, se [6]. Pensionsmodellen simulerar hur befolkningen tjänar in och tar ut allmän inkomstgrundad pension för att kunna modellera SA, SP, BF och AT i balansräkningen för inkomstpensionen. För att kunna modellera avgiftstillgången AT modellerar Pensionsmodellen hur befolkningen och dess sammansättning ändras med tiden. Detta görs genom att kombinera den data vi har på individnivå vi har från Pensionsmyndighetens register och kombinera det med externa prognoser för t.ex. arbetslöshet, dödlighet, in- och utvandring samt prognoser på t.ex. arbetslöshet. Data vi har tillgång till ser ut som följer:

- kön, betecknas $s \in S = \{man, kvinna\}$
- ålder, betecknas $a \in A = \{0, 1, \ldots, 120\}$
- grupp av födelseland, betecknas $o \in O = \{1, \ldots, 7\}$ enligt SCB:s defintion om landgrupper:
    1. Sverige
    2. Norden exkl. Sverige
    3. EU exkl. Norden
    4. Europa exkl EU

5. Länder med hög Human Development Index (HDI) exkl Europa
6. Länder med medel HDI exkl Europa
7. Länder med låg HDI exkl Europa

Se vidare i Pensionsmodellens användarmanual [5] sidan 23. Utöver dessa har en individ n_e st egenskaper som alla antar värdet 0 eller 1. De olika egenskaperna är:

- $X_1 = 0 \iff$ personer lever
- $X_2 = 1 \iff$ har minst ett barn upp till ett års ålder
- $X_3 = 1 \iff$ har minst ett barn upp till fyra års ålder
- $X_4 = 1 \iff$ har eftergymnasial utbildning
- $X_5 = 1 \iff$ intjänar pensionsrätt för barnår
- $X_6 = 1 \iff$ har ett inkomstpensionskapital större än 0
- $X_7 = 1 \iff$ intjänar pensionsrätt för studier
- $X_8 = 1 \iff$ har pensionsgrundande inkomst (PGI) från arbetsinkomster
- $X_9 = 1 \iff$ intjänar pensionsrätt baserat på sjuk- och aktivitetsersättning som grundar sig på tidigare arbetsinkomster
- $X_{10} = 1 \iff$ intjänar pensionsrätt baserat på sjuk- och aktivitetsersättning som grundar sig på ett garantibelopp
- $X_{11} = 1 \iff$ intjänar pensionsrätt från övriga transfereringar
- $X_{12} = 1 \iff$ befinner sig i Sverige
- $X_{13} = 1 \iff$ har emigrerat till Sverige
- $X_{14} = 1 \iff$ har börjat ta ut sin inkomstpension

Valet av dessa variabler ses som givet i den här uppsatsen. Den befintliga metoden använder dessa variabler och för att passa in en eventuellt uppdaterad metod i Pensionsmodellen är det praktiskt om vi använder samma lista på variabler. En variabel har inte tagits med i den här uppsatsen nämligen andel inkomstpension som tas ut, det är möjligt att ta ut delar av sitt hela månadsbeloppet (s.k. partiellt uttag) på 25, 50 och 75 procent. Utfallet att variabel $X_{14} = 1$ innebär att individen tar ut hela sitt månadsbelopp, andelen som tar ut en mindre del än 100 procent är relativt få, under 2 procent. Att ta uppdelning på olika uttagsnivåer skulle komplicera analysen och förlänga beräkningstider, men det bedöms som fullt möjligt att t.ex. parametrisera upp uttagsnivån på följande sätt:

- $X_{15} = 1 \iff$ har börjat ta ut 25 procent av sin inkomstpension eller mer
- $X_{16} = 1 \iff$ har börjat ta ut 50 procent av sin inkomstpension eller mer
- $X_{17} = 1 \iff$ har börjat ta ut 75 procent av sin inkomstpension eller mer

2.1 Notation

Som tidigare nämnts ses en individs egenskaper som utfall av stokastiska variabler $\boldsymbol{X} = \{X_1, X_2, \ldots, X_{n_e}\}$. Eftersom uppsatsen handlar om hur individer gör övergångar från en vektor till en annan kommer vi i vissa sammanhang att specificera tiden $t \in \{0, 1\}$ som subskript till den stokastiska variabeln. Tiden $t = 0$ betecknar att vi befinner oss *före* en övergång och $t = 1$ betecknar tiden *ett år senare*. För att notationen inte ska krocka med subskriptet som betecknar en enskild egenskap t.ex. X_1 (som anger huruvida en person lever eller ej), sätts symbolen $\bullet$ som det vänstra subskriptet när man anger en hel vektor, som i $\boldsymbol{x}_{\bullet,1}$ vilket anger ett utfall av en hel vektor vid tiden 1. Ibland kommer vi att vilja peka ut ett visst specifikt utfall av egenskaper för en individ, då anger vi det som ett superskript som i $\boldsymbol{x}^i$. Om en person gör en övergång från $\boldsymbol{x}^i_{\bullet,0}$ till $\boldsymbol{x}^j_{\bullet,1}$ kommer vi att förkorta detta som $\boldsymbol{x}^i \to \boldsymbol{x}^j$, pilen anger att en förflyttning ett år i tiden sker och därför utelämnar vi subskriptet som betecknar tiden. I kapitel 3 kommer vi även att gå igenom de data vi har använt i den här uppsatsen. När vi benämner utfall som finns i dessa data kommer vi att lägga till en parentes kring superskriptet så som

i $\boldsymbol{x}^{(i)}_{\bullet,0}$. Anledningen till att man behöver skilja på ett visst utfall $\boldsymbol{x}^{i}_{\bullet,1}$ och ett utfall i historiska data $\boldsymbol{x}^{(i)}_{\bullet,1}$ är p.g.a att alla utfall som våra modeller förutspår inte nödvändigtvis har förekommit i data tidigare.

2.2 Pensionsmodellens delar

Pensionsmodellen kan delas upp i tre beståndsdelar:

1. Uppgifter om antal i befolkningen $N^{s,a,o}_{\boldsymbol{x}^i}$. Det som avses är antalet personer av kön s, ålder a och födelselandgrupp o som har egenskaper som ges av $\boldsymbol{x}^i = (x^i_1, x^i_2, \ldots, x^i_{n_e})$.

2. En matris med övergångssannolikheter från tid 0 till tid 1 i projektionen. Mer formellt en samling sannolikheter som bildar en stationär markovkedja. Definiera följande:

$$p^{s,a,o}_{\boldsymbol{x}^i \to \boldsymbol{x}^j} := P(\boldsymbol{X}_{\bullet,1} = \boldsymbol{x}^j_{\bullet,1} | \boldsymbol{X}_{\bullet,0} = \boldsymbol{x}^i_{\bullet,0}, \{s, a, o\})$$

3. Med de två objekten $N^{s,a,o}_{\boldsymbol{x}^i}$ och $p^{s,a,o}_{\boldsymbol{x}^i \to \boldsymbol{x}^j}$ kan vi beräkna förväntat antal övergångar

$$T^{(s,a,o)}_{\boldsymbol{x}^i \to \boldsymbol{x}^j} := N^{s,a,o}_{\boldsymbol{x}^i} \cdot p^{s,a,o}_{\boldsymbol{x}^i \to \boldsymbol{x}^j}$$

Enligt en metod som beskrivs närmare i kapitel 6 kalibreras $p^{s,a,o}_{\boldsymbol{x}^i \to \boldsymbol{x}^j}$ till $\tilde{p}^{s,a,o}_{\boldsymbol{x}^i \to \boldsymbol{x}^j}$ och vi får då ett annat antal förväntade övergångar:

$$\tilde{T}^{(s,a,o)}_{\boldsymbol{x}^i \to \boldsymbol{x}^j} := N^{s,a,o}_{\boldsymbol{x}^i} \cdot \tilde{p}^{s,a,o}_{\boldsymbol{x}^i \to \boldsymbol{x}^j}$$

I kalibreringen är man ute efter en justering av sannolikheterna som gör så att

$$R^{(\mathcal{S},\mathcal{A},\mathcal{O})}_{\mathcal{X}^k \to \mathcal{X}^l} = \sum_{\substack{s \in \mathcal{S}, a \in \mathcal{A}, o \in \mathcal{O}, \\ \boldsymbol{x}^i \in \mathcal{X}^k, \boldsymbol{x}^j \in \mathcal{X}^l}} \tilde{T}^{(s,a,o)}_{\boldsymbol{x}^i \to \boldsymbol{x}^j} \tag{1}$$

För specifika delmängder $\mathcal{S} \subseteq S, \mathcal{A} \subseteq A, \mathcal{O} \subseteq O$ och $\mathcal{X}^i, \mathcal{X}^j \subseteq \Omega_x$.

Talen $R^{(\mathcal{S},\mathcal{A},\mathcal{O})}_{\mathcal{X}^k \to \mathcal{X}^l}$ ges från externa prognoser på hur många som ska gå från mängden av vektorer av egenskaper $\mathcal{X}^k$ till en annan mängd $\mathcal{X}^l$ för vissa kombinationer av s, a, o. Det kan t.ex vara att man vill justera modellen för informationen om att SCB:s prognos säger att 1640 män boende i Sverige i åldern 90 förväntas att avlida år 2025. Kaliberingen görs för att framskrivningen ska vara konsistent med prognoser från SCB över antal födda, döda,

inflyttade samt prognoser på arbetslöshet från Konjunkturinsitutet. Om man inte hade lagt på några restriktioner när man gör projektionen hade det varit en stationär markovkedja d.v.s. alla sannolikheter hade varit beständiga över tid. Vi vet inte hur våra sannolikheter kommer att ändras över tid på mikronivå men vi har tillgång till externa prognoser från t.ex. SCB och konjunkturinstitutet. Det gäller alltså att ändra de sannolikheter som räknats ut från individdata så att de passar med prognoser som är på makronivå. Det finns flera lösningar på ett sådant problem . I den här uppsatsen kommer vi argumentera för en variant som är följsam för de externa prognoserna samtidigt som det i någon mening inte påverkar de andra samband som följer av de modeller som skattats för övergångssannolikheter, se kapitel 6. Det är sedan dessa övergångssannolikheter som används för att skriva fram befolkningen enligt:

$$N_{\boldsymbol{x}^j}^{(s,a+1,o)}(t+1) = \sum_{\boldsymbol{x}^i \in \Omega_x} \tilde{T}_{\boldsymbol{x}^i \to \boldsymbol{x}^j}^{(s,a,o)}(t) \tag{2}$$

Här har vi lagt till ett argument t som kan anta årtal $2019, 2020, \ldots$. Detta eftersom vi skriver fram modellen i tiden.

2.3 Typer av inkomster

Till varje kombination av $s, a, o, \boldsymbol{x}^i \to \boldsymbol{x}^j$ ordnas pensionsinkomster uppdelat på:

- Inkomst från anställning (PGII)
- Inkomst från sjuk och aktivitetsersättning, baserat på garantibelopp (SAPGB)
- Inkomst från sjuk och aktivitetsersättning, baserat på tidigare inkomst (SAPGI)
- Pensionsrätt för barnår (BGPB)
- Pensionsrätt för studier (PGBSTUD)
- Pensionsrätt för övriga transfereringar

Pensionsrätt är den allmänna pensionens ord för premie. Metoden som ordnar inkomster till dessa övergångar är inte något som tas upp i den här uppsatsen. Modellen skriver inte fram individerna var för sig utan grupperar ihop dem

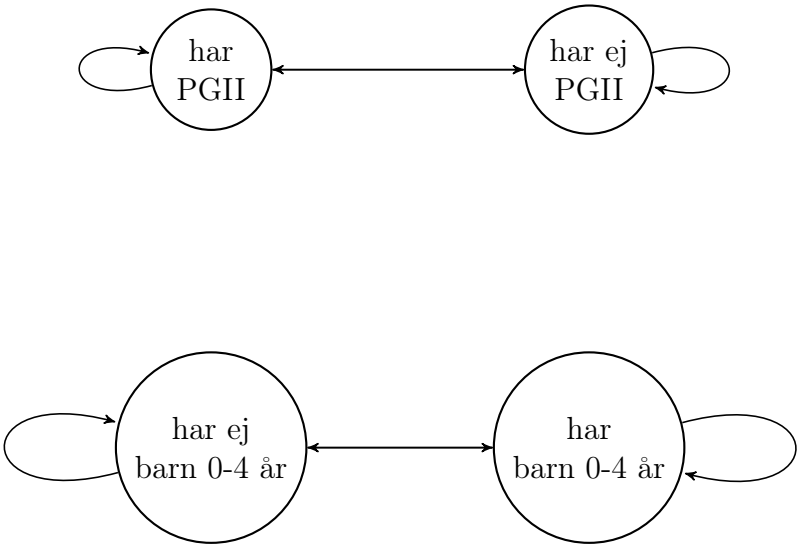
utifrån dess egenskaper. Aggregeringen på grupper görs dels för att resultatet inte ska bli känslig information, men även av prestandaskäl. Att gruppera ihop individerna till grupper av människor gör att vi inte kan ta hänsyn till en individs historia utan endast till den information som finns vid det senaste tidssteget.

3 Data

För att skapa de övergångsmatriser som Pensionsmodellen använder behövs individdata, dessa läggs inte ut på hemsidan. Individdata kommer från Pensionsmyndighetens datalager som i sin tur får data från Pensionsmyndighetens egna register men även från SCB, Försäkringskassan och Skatteverket. Befolkningen som modelleras är dels alla som finns i svenska folkbokföringen men även alla som någonsin tjänat in inkomstgrundad pension i Sverige. Vi betecknar de n_d stycken observationerna som

$$\boldsymbol{d^{(i)}} = \{s^{(i)}, a^{(i)}, o^{(i)}, \boldsymbol{x}^{(i)}_{\bullet,0}, \boldsymbol{x}^{(i)}_{\bullet,1}\}$$

för $i = 1, ..., n_d$. En observation beskriver en individ som gör en övergång över ett år i tiden. I vårt dataset är $n_d = 9.7$ miljoner och antalet distinkta observationer på $\boldsymbol{x}^i$ är 1289 till antalet. De egenskaper som utgör $\boldsymbol{x}^i_{\bullet t}$ är valda på grund av att de påverkar inkomstbildningen hos en person. I figurerna 6, 7...,16, visas hur stor andel av befolkningen som har en viss egenskap. I dessa figurer har endast befolkningen födda och boende i Sverige år 2013 använts. Man kan konstatera att vissa variabler har ett beroende till ålder som ser ut att vara kontinuerligt och ha ett enda maximum, detta gäller t.ex. för barn och pensionsrätt för studier. För variabeln transfereringar ser det ut att finnas två maximum. För andra kurvor ser man att det finns diskreta hopp t.ex. för kurvan som visar huruvida man har börjat ta ut sin inkomstpension eller ej. Detta beror på att regelsystemet säger att man inte kan ta ut sin pension innan man fyllt 61 år. På samma sätt säger reglerna att man inte kan få sjuk- och aktivitetsersättning efter 65 års ålder. Samtidigt är det en stor andel som har denna ersättning innan pensioneringen vilket gör att man ser en abrupt nedgång. Det kommer visa sig att ett sätt att hantera dessa oregelbundna beroenden till ålder är att anpassa kubiska splines till data.

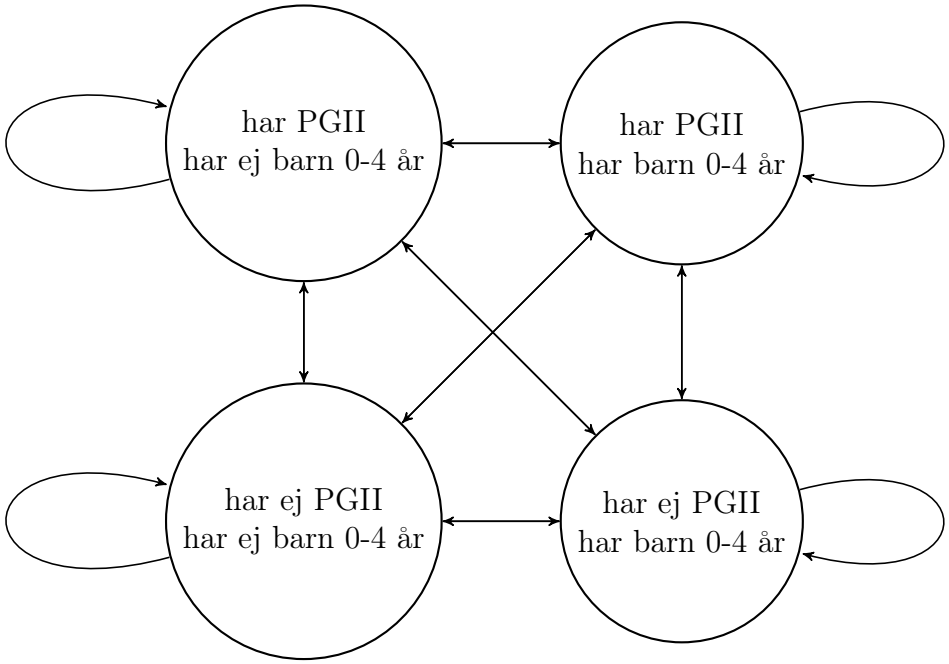


Figur 1: Övergångar mellan att ha eller inte PGII och att ha eller inte ha barn i åldern 0-4, där övergångarna sker oberoende av varandra

3.1 Beroende mellan övergångar

Innan vi går vidare behöver vi tydligare definiera vad vi menar med övergångar mellan olika tillstånd. Det finns flera sätt att betrakta dessa övergångar, antingen utgår vi ifrån att övergångarna mellan de olika egenskaperna är oberoende eller beroende av varandra. Om vi som exempel betraktar de två egenskaperna *har PGII* (har pensionsgrundande inkomst från arbete) och *har barn i åldern 0-4 år* kan vi betrakta ett schema som illustrerar skillnad mellan oberoende (se figur 1) eller beroende (se figur 2) när vi betraktar dessa egenskaper.

Om man väljer att se processen som att de två egenskaperna tillsammans utgör *ett* tillstånd och att övergångarna sker beroende av varandra ser schemat över övergångarna ut som i figur 2. Från varje tillstånd är det alltså möjligt att göra en övergång till tre andra tillstånd eller att stanna kvar i samma tillstånd. Vi betecknar i det här exemplet *har PGII* som $X_8 = 1$ och *har barn i åldern 0-4 år* som $X_3 = 1$. Om vi exempelvis ser på övergången från $(X_{3,0}, X_{8,0}) = (0, 0)$ till $(X_{3,1}, X_{8,1}) = (1, 0)$, är frågeställningen huruvida sannolikheten kan skrivas som en produkt enligt:



Figur 2: Övergångar mellan att ha eller inte PGII och att ha eller inte ha barn i åldern 0-4, där övergångarna sker beroende av varandra

$$
\begin{aligned}
&P(X_{3,1} = 1, X_{8,1} = 0 | X_{3,0} = 0, X_{8,0} = 0) \\
&\stackrel{?}{=} P(X_{3,1} = 1 | X_{3,0} = 0, X_{8,0} = 0) \cdot P(X_{8,1} = 0 | X_{3,0} = 0, X_{8,0} = 0)
\end{aligned}
$$

3.2 Pearson χ^2-test

Ett sätt att testa oberoende är att använde ett Pearson χ^2-test. Här följer ett exempel där vi har betingat på att kön är kvinna, födelseland är Sverige och att åldern är mellan 30 och 40 år. De andra x-variablerna vid tiden 0 är valda så att vi får den vanligaste kombinationen som finns i data. Vi studerar nu hur x-variablerna $X_{3,1} =$ *"har barn i åldern 0-4"* och $X_{8,1} =$ *"har PGII"* samspelar vid tiden $t = 1$. Pearsons χ^2-test ger ett p-värde på närmare 1 och därför inga indikationer på att man skulle kunna förkasta nollhypotesen som är att de två variablerna är oberoende. Nollhypotesen är samma sak som att fördelningen för två variabler kan skrivas som en produkt enligt dessa som

$X_{3,1}$	$X_{8,1} = 0$	$X_{8,1} = 1$
0	687	60347
1	10	840

ovan. Om vi istället undersöker kombinationen $X_{3,1} =$ *"har barn i åldern 0-4"* och $X_{11,1} =$ *"har transfereringar (förutom sjuk- och aktivitetsersättning)"* ger χ^2-testet ett värde nära 0. Det är alltså betydligt vanligare att ha transfereringar om man har barn än annars. Detta är naturligt då föräldrapenning är

$X_{3,1}$	$X_{11,1} = 0$	$X_{11,1} = 1$
0	46231	14803
1	466	384

pensionsgrundande och ingår i tranfereringar.

3.3 Uppdelning i träningsdata och testdata

För att kunna jämföra förmågan för de olika modellerna att göra prediktion kommer vi att dela upp datamängden i två, en för att träna modellerna och en för att testa modellernas prediktionsförmåga. Vi har valt uppdelning att 70 procent av data är träningsdata och resterande i 30 procent testdata. Formellt betecknas samtliga dataobservationer $\{\boldsymbol{d}^{(1)}, \ldots, \boldsymbol{d}^{(n_d)}\}$. Träningsdatamängden utgörs av $\mathcal{M}^{\text{train}} := \{\boldsymbol{d}^{(i)}\}_{i \in \mathcal{I}_{\text{train}}}$ och testdatamängden av $\mathcal{M}^{\text{test}} := \{\boldsymbol{d}^{(i)}\}_{i \in \mathcal{I}_{\text{test}}}$ där $\mathcal{I}_{\text{test}}$ och $\mathcal{I}_{\text{test}}$ är två disjunkta mängder av motsvarande index. Det finns givetvis andra möjliga tänkbara uppdelningar än 70/30, men det har inte undersökts vidare i den här uppsatsen.

3.4 Korsvalidering

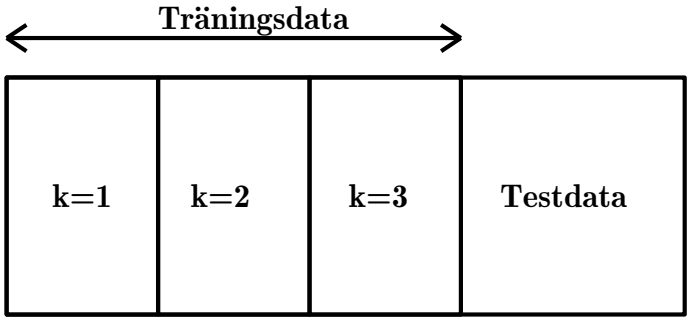
Målet med att sätta upp dessa modeller är för att kunna göra så bra prediktioner som möjligt på nytt data som inte använts tidigare. Vi använder s.k. *korsvalidering* för att skatta prediktionsfelet, se sidan 241 i [3]. I den här uppsatsen görs korsvalidering så att vi delar in $\mathcal{M}^{\text{train}}$ i K st lika stora delar. Utför följande, för varje $k = 1, 2, \ldots, K$

1. Välj ut alla data som *inte* finns i grupp k och skatta den aktuella modellen på dessa data

2. Prediktera utfall för data som finns i grupp k med modellen från punkt 1

3. Räkna ut skattningen av felet för skattningarna i punkt 2. I denna uppsats görs det med residual deviance (se kapitel 7.1) eller med mått för felklassifikation (se kapitel 7.4)

Typiskt sett görs processen om för varje värde på någon *tuning*-parameter. I fallet med lasso-regression är tuning-parametern λ, så även i fallet med att anpassa en kubiska spline. Korsvalidering går att göra med K så stort som antal data-punkter. I [3] s. 243 anges att ett K mellan 5 och 10 oftast är lämpligt. I den här uppsatsen har $K = 3$ använts. Anledningen är främst p.g.a. de redan ganska långa körtiderna på datorn som användes för att göra beräkningarna. Korsvalideringen gjordes efter 70/30-uppdelningen som nämns i kapitel 7. Korsvalideringen gjordes för att välja tuning-parametrar.



Figur 3: Data delas upp i $\mathcal{M}_{\text{train}}$ och $\mathcal{M}_{\text{test}}$. $\mathcal{M}_{\text{train}}$ delas upp i ytterligare tre delar vid korsvalideringen när λ ska väljas för lasso eller spline-regression

4 Teori

Det som studeras i den här uppsatsen är att skatta de sannolikheter som anger hur befolkningen ändrar egenskaper. Detta kan ses som att befolkningen gör övergångar mellan olika tillstånd. Dessa tillstånd utgörs av egenskaper eller kombinationer av egenskaper.

4.1 Dödlighet hanteras separat

Vi hävdade tidigare att man i princip kan röra sig från vilken vektor $\boldsymbol{x}^i_{\bullet,0}$ till vilken vektor $\boldsymbol{x}^j_{\bullet,1}$ som helst. Ett undantag är när det gäller egenskapen $X_1 = X_{\text{död}}$, när $X_1 = 1$ gäller att $X_k = 0$ för alla $k \neq 1$. Man skulle kunna tänka sig att om en person dör under ett år så kan den fortfarande ha ett intjänande under den delen av året då individen levde, men i modellen har man sen tidigare gjort den förenklingen att det inte är möjligt. Vi kan generellt skriva:

$$P(X_1 = x_1, \ldots, X_{n_e} = x_{n_e}) = P(X_1 = 1) \cdot P(X_2 = x_2, \ldots, X_{n_e} = x_{n_e} | X_1 = 1)$$

$$+P(X_1 = 0) \cdot P(X_2 = x_2, \ldots, X_{n_e} = x_{n_e} | X_1 = 0)$$

När vi undersöker övergångar mellan andra tillstånd som inte har att göra med dödlighet antar vi hädanefter att vi har betingat på $X_1 = 0$ d.v.s. att personen lever, både i tiden 0 och 1.

4.2 Metod för att ta hänsyn till beroende

De tester som gjorts i kapitel 3.2 indikerar att vissa utfall för $x_{k,1}$ är beroende andra utfall $x_{j,1}$ vid samma tidpunkt. Om man inte tar hänsyn till beroende skulle vi kunna göra regression på alla variabler $x_{k,1}$ var för sig och endast ta hänsyn till $s, a, o, \boldsymbol{x}_{\bullet,0}$. Om man däremot fullständigt skulle se de olika egenskaperna, beroende av varandra skulle det innebära att varje kombination av egenskaper $\boldsymbol{x}^i_{\bullet,1}$ blir en egen unik kategori. Vi behöver alltså en modell som i tillräcklig utsträckning tar hänsyn till beroende mellan x-variablerna i tidssteget t men som inte resulterar i en modell som är överparametriserad. Ett alternativ är att göra logistisk regression i en bestämd ordning. Alltså att man först gör regression för $x_{1,1}$ baserat på $s, a, o, x_{1,0}, x_{2,0}, ..., x_{n_e,0}$ sedan gör man en regression på $x_{2,1}$ där man betingar på utfallet av $s, a, o, x_{1,0}, x_{2,0}, ..., x_{n_e,0}, x_{1,1}$. Att detta inte medför något

särskilt antagande kommer av grundläggande sannolikhetslära, betrakta den simultana fördelningen för X_1, X_2, X_3:

$$P(X_1 = x_1 \cap X_2 = x_2, X_3 = x_3)$$

$$= P(X_1 = x_1) \cdot P(X_2 = x_2 \cap X_3 = x_3 | X_1 = x_1)$$

$$= P(X_1 = x_1) \cdot P(X_2 = x_2 | X_1 = x_1) \cdot P(X_3 = x_3 | X_2 = x_2 \cap X_1 = x_1)$$

Om vi passar in det i sammanhanget för de modeller som beskrivs i den här uppsatsen ska vi även betinga på kön, ålder, födelseland och utfallet av $\boldsymbol{X}^i_{\bullet,0}$ och vi får då:

$$p^{s,a,o}_{\boldsymbol{x}^i \to \boldsymbol{x}^j} = P(\boldsymbol{X}_{\bullet,1} = \boldsymbol{x}^j | \boldsymbol{X}_{\bullet,0} = \boldsymbol{x}^i, s, a, o) =$$
$$\prod_{k=1}^{n_e} P(X_{k,1} = x^j_{k,1} | \cap_{l=1}^{k-1} \{X_{l,1} = x^j_{l,1}\}, \boldsymbol{X}_{\bullet,0} = \boldsymbol{x}^i, s, a, o) \tag{3}$$

För att förenkla framöver införs notationen

$$p_{k|k-1} := P(X_{k,1} = 1 | \cap_{l=1}^{k-1} \{X_{l,1} = x^j_{l,1}\}, \boldsymbol{X}_{\bullet,0} = \boldsymbol{x}^i, s, a, o)$$

d.v.s.

$$p^{s,a,o}_{\boldsymbol{x}^i \to \boldsymbol{x}^j} = \prod_{k=1}^{n_e} (p_{k|k-1})^{x^j_{k,1}} (1 - p_{k|k-1})^{1 - x^j_{k,1}}$$

Det är alltså i detta avseende möjligt att dela upp sannolikheterna i produkter av sannolikheter som det går att använda exempelvis logistisk regression på. Eftersom vi förväntar oss att det kommer att finnas ett beroende mellan vissa egenskaper men inte andra kommer antagligen vissa betingade sannolikheter att förenklas.

4.3 Ordningen av kateogrivariablerna

Som vi presenterade i ekvation (3) kan man dela upp den simultana fördelningen i en kedja av betingade sannolikheter. En fråga man kan ställa sig är om valet av ordningsföljd spelar någon roll för modellen. Vi kommer att undersöka detta för fallet med logistisk regression där vi inte har några interaktionstermer mellan olika x-variabler. Säg att vi har två variabler X_1, X_2, då kan vi skriva:

$$P(X_1 = x_1, X_2 = x_2) = P(X_1 = x_1) \cdot P(X_2 = x_2 | X_1 = x_1) \tag{4}$$

så väl som

$$P(X_1 = x_1, X_2 = x_2) = P(X_2 = x_2) \cdot P(X_1 = x_1 | X_2 = x_2) \tag{5}$$

Om man utgår från ekvation (4) ansätter en logistisk regressionsmodell till $P(X_1 = x_1)$ får man följande parameterisering:

$$\text{logit}(P(X_1 = 1)) = \alpha_{x_1}$$

För variabeln X_2 betingar man på att man vet utfallet av X_1 och får då:

$$\text{logit}(P(X_2 = 1 | X_1 = x_1)) = \alpha_{x_2} + \beta x_1$$

Om man utgår från ekvation (5) istället får man:

$$\text{logit}(P(X_2 = 1)) = \tilde{\alpha}_{x_2}$$

$$\text{logit}(P(X_1 = 1 | X_2 = x_2)) = \tilde{\alpha}_{x_1} + \beta x_2$$

Lägg märke till att β-parametern är den samma i bägge exemplen. I appendix 9.2 visas att detta gäller för två och tre variabler i följd och kan rimligen generaliseras till ett godtyckligt stort antal. Denna härledning tar inte hänsyn till s, a, o eller $\boldsymbol{x}_{\bullet,0}$ men man kan utgå ifrån att dessa ingår i α-parametrarna om vi antar att det inte finns några interaktionstermer.

5 Modell

För varje egenskap gäller att den har två utfall, sant/1 eller falskt/0, därför är det lämpligt att ansätta en binomialfördelning om man samtidigt antar att individerna gör dessa övergångar oberoende av andra individer. Fördelningen mellan utfallen sant och falskt förväntas bero i olika utsträckning av ålder, kön, födelseland men även på vilka andra egenskaper $\{X_{j,0}\}$ som personer kommer från innan samt även de andra egenskaperna $\{X_{j,1}\}$ som fastställts i tiden 1 som vi betingar på enligt ekvation (3).

5.1 Empiriska skattningar

Vi kommer att undersöka träffsäkerheten av de empiriska skattningarna såväl som andra parametriserade modeller. De empiriska skattningarna kommer precis som de andra modellerna skattas på $\mathcal{M}_{\text{train}}$ och utvärderas på $\mathcal{M}_{\text{test}}$. I det här kapitlet kommer vi uttrycka den empiriska skattningen mer formellt. Som utgångspunkt har vi notationen för data $\{\boldsymbol{d}^{(m)}\}$, som den uttrycks i kapitel 3. Vi börjar med att definiera den mängd av data som ska användas för en enskild skattning, vi börjar med att välja en egenskap som vi ska skatta $p_{k|k-1}$ för. Vi väljer specifika värden på $s, a, o, \boldsymbol{x}_{\bullet,0}$ och $\{x_{1,1}, \ldots, x_{k-1,1}\}$ som vi betingar på och kallar dessa för $s^i, a^i, o^i, \boldsymbol{x}^i_{\bullet,0}$ respektive $\{x^i_{1,1}, \ldots, x^i_{k-1,1}\}$. Vi väljer ut mängden $\mathcal{N}$ av alla datapunkter som har dessa värden, dvs:

$$\mathcal{N}_{k|k-1} := \{\forall d^{(m)} | s^{(m)} = s^i, a^{(m)} = a^i, o^{(m)} = o^i, \boldsymbol{x}^{(m)}_{\bullet,0} = \boldsymbol{x}^i_{\bullet,0}, \cap_{l=1}^{k-1}\{x^{(m)}_{l,1} = x^i_{l,1}\}\}$$

Vi kallar antalet observationer för

$$n_{k|k-1} := \#\{\mathcal{N}_{k|k-1}\} \tag{6}$$

Vi definierar $\mathcal{Y}_{k|k-1} \subseteq \mathcal{N}_{k|k-1}$ som de datapunkter som gav ett resultat på 1 för den aktuella egenskapen med index k:

$$\mathcal{Y}_{k|k-1} := \{\forall d^{(m)} \subseteq \mathcal{N}_k | x^{(m)}_{k,1} = 1\}$$

och antalet i denna datamängd

$$y_{k|k-1} := \#\{\mathcal{Y}_{k|k-1}\} \tag{7}$$

Den empiriska skattningen blir därför:

$$\hat{p}_{k|k-1} = \frac{y_{k|k-1}}{n_{k|k-1}} \tag{8}$$

Ett antagande i den här uppsatsen är att $y_{k|k-1}$ är binomialfördelad:

$$y_{k|k-1} \sim \text{Bin}(n_{k|k-1}, p_{k|k-1})$$

Där $p_{k|k-1}$ är en parameter vi försöker skatta som är okänd. Vi har olika $y_{k|k-1}$ och $n_{k|k-1}$ beroende på olika s, a, o men för att undvika allt för omständlig notation i det här kapitlet förenklas notationen på detta sätt. När vi har att göra med binomialfördelningen kan man visa att en logit-länk enligt $\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$ på många sätt är en praktisk länkfunktion. Det hade även gått att experimentera med andra länkfunktioner än logit men det har utlämnats i denna uppsats.

5.1.1 Empriska skattning på $\mathcal{M}^{\text{train}}$ och $\mathcal{M}^{\text{test}}$

Generellt skattas modellerna på $\mathcal{M}^{\text{train}}$ och prediktionsförmågan testas på $\mathcal{M}^{\text{test}}$. De empiriska skattningarna gjorda på $\mathcal{M}^{\text{test}}$ betraktar vi som vårt facit när vi ska beräkna träffsäkerheten. Det kan förekomma fall där uppdelningen mellan $\mathcal{M}^{\text{train}}$ och $\mathcal{M}^{\text{test}}$ gör att ett exempel på $s, a, o, \boldsymbol{x}_{\bullet,0}, \{x_{1,1}, \ldots, x_{k-1,1}\}$ finns i $\mathcal{M}^{\text{test}}$ men inte i $\mathcal{M}^{\text{train}}$, alltså att $\mathcal{N}_{k|k-1} = \varnothing$, där $\mathcal{N}_{k|k-1}$ definieras som i kapitel 5.1. Detta gör att modellen inte har någon ren empirisk skattning för denna händelse. Alltså, hade man baserat sin modell endast på empiriska skattningar, som alla varit baserade på ekvation (8) är vi inte säkra på att vi kommer ha skattningar för alla fall som kan förekomma i $\mathcal{M}^{\text{test}}$. Detta hanteras med två olika metoder:

- Utöka $\mathcal{N}_{k|k-1}$ så att mängden blir större. Här listas en succesiv utökning. Om $\mathcal{N}_{k|k-1} = \varnothing$ utökar man till en punkt längre ned på denna lista i ordning:

  1. $\mathcal{N}_{k|k-1} := \{\forall d^{(m)} | s^{(m)} = s^i, a^{(m)} = a^i, o^{(m)} = o^i, \boldsymbol{x}_{\bullet,0}^{(m)} = \boldsymbol{x}_{\bullet,0}^i\}$
  2. $\mathcal{N}_{k|k-1} := \{\forall d^{(m)} | s^{(m)} = s^i, a^{(m)} = a^i, o^{(m)} = o^i\}$
  3. $\mathcal{N}_{k|k-1} := \{\forall d^{(m)} | s^{(m)} = s^i, a^{(m)} = a^i\}$
  4. $\mathcal{N}_{k|k-1} := \{\forall d^{(m)} | s^{(m)} = s^i\}$
  5. $\mathcal{N}_{k|k-1} := \{\forall d^{(m)}\}$

  Notera att när man och utökar $\mathcal{N}_{k|k-1}$ utökas även $\mathcal{Y}_{k|k-1}$.

- Basera skattningen på modellen som använder sig av MLE + spline, se kapitel 7.4 och kapitel 5.4.

5.2 Logistisk regression

Den nämnda logit-länken tillsammans med ett helt linjärt uttryck kallas för logistisk regression som vi tidigare nämnt. Generellt ser uttrycket ut som följer:

$$\text{logit}(p) = \beta_0 + \sum_{i=1}^{p} \beta_i x_i$$

I det här fallet är alla x-variabler kategorivariabler utom en, nämligen ålder. Ett sätt att hantera beroende till ålder är att dela in åldern i olika intervall

så att man därmed transformerar den till en katogorivariabel. Processen att välja hur uppdelningen av ålder ska göras i olika intervall bedömdes dock vara allt för tidskrävande för att vara ett rimligt alternativ. Lägg därtill att denna process måste upprepas för alla n_e stycken variabler. Som ett första steg valdes att skapa en ny x-variabel som sattes lika med a^2. Uttrycket blir då med våra parametrar:

$$\begin{aligned}\text{logit}(p_{k|k-1}) = \beta_{0k} + \sum_{l=1}^{n_e} \beta_{lk}^{\boldsymbol{x}_\bullet,0} x_{l0} + \sum_{l=1}^{k-1} \beta_{lk}^{\boldsymbol{x}_\bullet,1} x_{l1} + \sum_{l=1}^{2} \beta_{lk}^{\text{sex}} I(s=l) \\ + \sum_{l=1}^{7} \beta_{lk}^{\text{origin}} I(o=l) + a\beta_{1k}^{\text{age}} + a^2\beta_{2k}^{\text{age}}\end{aligned} \tag{9}$$

Där $I(\cdot)$ är indikatorfunktionen. Anledningen till vårt val av beroendet mot ålder var att om man tittar i figurerna 6, 7...,16, ser det ut som om förekomster av olika egenskaper antingen toppar under en viss tid i livet eller antingen ökar eller avtar hela livet. Det är inte så vanligt förekommande med flera lokala maxima i graferna. Diagrammen ska dock inte tolkas som att de är grunden till beroenden till ålder, diagrammen är gjorda för alla kombinationer av de övriga x-variablerna och visar alltså inte beroendet till ålder där allt annat är lika. Vi kommer att se att den här anpassningen till ålder inte ger så bra prediktioner som andra modeller, den här metoden är mer som en första ansats. Vi kommer också att prova att parametrisera om koefficienterna för åldern a som om det vore en kategorivariabel på följande sätt:

$$\begin{aligned}\text{logit}(p_{k|k-1}) = \beta_{0k} + \sum_{l=1}^{n_e} \beta_{lk}^{\boldsymbol{x}_\bullet,0} x_{l0} + \sum_{l=1}^{k-1} \beta_{lk}^{\boldsymbol{x}_\bullet,1} x_{l1} + \sum_{l=1}^{2} \beta_{lk}^{\text{sex}} I(s=l) \\ + \sum_{l=1}^{7} \beta_{lk}^{\text{origin}} I(o=l) + \sum_{l=0}^{120} \beta_{lk}^{\text{age}} I(a=l)\end{aligned} \tag{10}$$

5.2.1 Maximum Likelihood-skattning och deviance

För att skatta parametrarna i ekvation (9) används maximum likelihood estimation (MLE). MLE innebär att man väljer de parametrar som maximerar likelihooden $\mathcal{L}$. Likelihooden för en binomialfördelning definieras

$$\mathcal{L}(p_{k|k-1}|y_{k|k-1}) = \binom{n_{k|k-1}}{y_{k|k-1}} (p_{k|k-1})^{y_{k|k-1}} (1-p_{k|k-1})^{n_{k|k-1}-y_{k|k-1}}$$

Där $n_{k|k-1}$ och $y_{k|k-1}$ definieras enligt ekvation (6) resp (7). Då vår modell är en logistisk regression där $p_{k|k-1}|y_{k|k-1}$ beror på ett antal β-parametrar kan vi skriva $\mathcal{L}(p_{k|k-1}|y_{k|k-1}) = \mathcal{L}(\{\beta_p\}|y_{k|k-1})$. Att maximera likelihooden är samma som att maximera log-likelihooden:

$$\log(\mathcal{L}(p_{k|k-1}|y_{k|k-1})) = \log\left(\binom{n_{k|k-1}}{y_{k|k-1}}\right) + y_{k|k-1}\log(p_{k|k-1}) + (n_{k|k-1} - y_{k|k-1})\log(1 - p_{k|k-1})$$

Vi passar på att definiera *deviance*-funktionen som är ett mått på avstånd mellan den empiriska modellen $\hat{p}_{k|k-1}$ enligt ekvation (8) och en annan model $\tilde{p}_{k|k-1}$. Om vi sätter in $\hat{p}_{k|k-1}$ skattade från den empiriska empiriska modellen enligt ekvation (8) i $\mathcal{L}$ har vi den allra högsta likelihooden som går att få fram utifrån data och vår modell. Deviance mäter avståndet från denna likelihooden logaritmerad enligt:

$$D(y, \tilde{p}_{k|k-1}) = -2(\log(\mathcal{L}(\hat{p}_{k|k-1}|y_{k|k-1})) - \log(\mathcal{L}(\tilde{p}_{k|k-1}|y_{k|k-1}))) \tag{11}$$

I fallet med logistisk regression har vi en samling β-parametrar, därför kan vi skriva $D(y_{k|k-1}, \tilde{p}_{k|k-1}) = D(y_{k|k-1}, \{\beta_p\})$

5.3 Lasso

Lasso-regression är en teknik där man utgår från MLE, men minskar koefficienternas storlek genom att lägga till ett villkor om att summan av koefficienterna inte får överstiga ett tal γ. Genom att prova olika värden på detta γ undersöker man om begränsningen resulterar i en modell med bättre prediktionsförmåga än vid MLE, se [2] sidan sidan 1. Vid lasso-regression kan även vissa koefficienter komma att sättas till 0. Lasso-regression utförs genom maximera likelihooden $\mathcal{L}(\{\beta_p\}|y_{k|k-1})$ under villkoret att $\sum_p |\beta_p| < \gamma$ där γ är en tuning-parameter, se [2] sidan sidan 23. R-paketet glmnet som användes i den här uppsatsen utför lasso-regression genom att lägga till strafftermen $\lambda \sum_p |\beta_p|$ till deviance-funktionen:

$$\Delta(f^{\text{lasso}}) = D(y_{k|k-1}, \{\beta_p\}) + \lambda \sum_p |\beta_p|$$

För varje λ som används söks de koeffcienter $\{\beta_p\}$ så att $\Delta(f^{\text{lasso}})$ minimeras. Att välja λ görs genom korsvalidering där modellens prediktiva förmåga mäts, se kapitel 3.4 . Lasso-regressionen förutsätter att regressionsvariablerna har standardiserats och koefficienterna har transformerats så att vi inte har något intercept, se [2] sidan 23.

5.4 Kubiska spline

Om sambandet knutet till ålder inte uppenbart följer en linjär eller andragradsfunktion kan det vara farligt att försöka ansätta ett polynom av högre grad, dessa polynom kan medföra numeriska problem och data i ett område kan få stor påverkan på hur anpassningen sker i ett helt annat område, se [1] sidan 101. Ett alternativ till anpassa åldern till ett polynom av högre grad är att utöka GLM-modellen till en s.k. *Generell Additativ Modell* (GAM) och anpassa åldern med lokala kubiska polynom som kallas *kubiska splines*. En kubisk splinefunktion är ett antal ihopsatta tredjegradspolynom som tillsammans bildar en funktion som har en kontinuerlig andra-derivata, även i knutpunkterna, se [1] sidan 104. Logit-transformationen blir då:

$$\begin{aligned}\text{logit}(p_{k|k-1}) = \beta_{0k} + \sum_{l=1}^{n_e} \beta_{lk}^{\boldsymbol{x}_\bullet,0} x_{l0} + \sum_{l=1}^{k-1} \beta_{lk}^{\boldsymbol{x}_\bullet,1} x_{l1} + \sum_{l=1}^{2} \beta_{lk}^{\text{sex}} I(s=l) \\ + \sum_{l=1}^{7} \beta_{lk}^{\text{origin}} I(o=l) + s(a)\end{aligned} \tag{12}$$

Där $s(a)$ är en kubisk spline-funktion. En spline funktion $s(a)$ definieras enligt:

$$s(a) = p_j(a), \text{ för } u_j \leq a < u_{j+1}$$

där

$$p_j(a) = \beta_{j0} + \beta_{j1} a + \beta_{j2} a^2$$

och

$$\begin{aligned} p_j(u_j) &= p_{j-1}(u_j), \text{för alla } j \\ p'_j(u_j) &= p'_{j-1}(u_j), \text{för alla } j \\ p''_j(u_j) &= p''_{j-1}(u_j), \text{för alla } j \end{aligned}$$

Utifrån u_j-värdena (som sätts till de punkter det finns a-värden för) söks parametrarna till tredjegradspolynomen så att den resulterande skattningen har bäst prediktionsförmåga med hjälp av korsvalidering, likt metoden för lasso-regression. Skattningen kommer delas upp och göras i två delar, en del för att anpassa de x-variabler som är kategorivariabler och den delen som ska anpassas med hjälp av spline. Det görs med en s.k. back fitting-algortim vilket innebär innebär att vi behöver ta fram två deviance-funktioner utfrån skattning av sannolikheten i ekvation (12). En deviance-funktion $D^{\text{kat}}(y, \{\beta_p^{\text{kat}}\}|\{\beta_p^{\text{spline}}\})$

där vi ser koefficienterna för spline-funktionen, betecknade $\{\beta_p^{\text{spline}}\}$, som givna och som hålls konstanta då vi söker $\{\beta_p^{\text{lasso}}\}$. Den andra deviance-funktionen betecknas $D^{\text{spline}}(y, \{\beta_p^{\text{spline}}\}|\{\beta_p^{\text{kat}}\})$ och för den gäller att vi håller β_p^{kat}-parametrarna konstanta när söker skattningar på $\{\beta_p^{\text{spline}}\}$. Det som ska minimeras för att hitta de sökta parametrarna är följande för kategorivariablerna:

$$\Delta(f^{\text{kat}}) = D^{\text{kat}}(y, \{\beta_p^{\text{kat}}\}|\{\beta_p^{\text{spline}}\}) + \lambda_{\text{lasso}} \sum_p |\beta_p^{\text{kat}}| \tag{13}$$

samt följande för variabeln ålder:

$$\Delta(f^{\text{spline}}) = D^{\text{spline}}(y, \{\beta_p^{\text{spline}}\}|\{\beta_p^{\text{kat}}\}) + \lambda_{\text{spline}} \int_{a_{\min}}^{a_{\max}} (s''(x))^2 dx \tag{14}$$

Anpassningen sker genom s.k. back-fitting, enligt följande:

1. Anpassa en lasso-modell genom att minimera (13). Den optimala λ-parametern hittas genom korsvalidering enligt kapitel 3.4.

2. Anpassa en modell där vi har en kubisk spline för beroendet till ålder genom att minimera 14 utifrån de parametrar som togs fram i steg 1

3. Upprepa steg 1 och 2 tills parametrarna ser ut att ha konvergerat

I de modeller som där vi använder MLE för kategorivariablerna sätter vi $\lambda = 0$. Spline-anpassningarna visas i figur 17 till 30. Kurvornas utseende visar att det är svårt att ansätta ett enda polynom för att få bra passform då sambandet är så pass oregelbundet. Om man jämför med figurerna 6 till 16 som visar andelen som har respektive egenskap i befolkningen ser man att det är ganska stor skillnad på dessa figurer. Det beror på att splineanpassningen är gjord utifrån att effekterna som har att göra med alla de andra kategorivariablerna redan är omhändertagna. Spline-figurerna visar alltså effekten av åldern givet *allt annat lika*. Det gör inte figur 6 till 16, dessa visar endast andelen i befolkningen med dessa egenskaper, utan att ta hänsyn till andra faktorer som kan påverka förutom ålder.

5.5 Logistisk regression med interaktioner

I uttrycken som ges i kapitel 7.4 och 5.3 innehåller inga interaktioner av typen $\text{logit}(p_{k|k-1}) = \cdots + x_{0,2}x_{0,3}\beta_{0,2,3} + \ldots$, alltså sådana termer där fler

än en enda x-variabel förekommer. I utvärderingen av modellerna kommer vi även att prova att lägga till dessa typer av beroenden. När vi skriver interaktionstermer av andra ordningen menar vi sådana där exakt två stycken x-variabler förekommer, när vi skriver interaktionstermer av tredje ordningen menar vi sådana där exakt tre stycken x-variabler förekommer. Av någon anledning lyckades inte beräkningen av MLE med interaktionstermer, beräkningarna hade svårt att konvergera. Däremot genomfördes beräkningarna vid lasso-regression utan problem, iaf för med interaktionstermer av andra graden. Beräkningar av modeller som innehöll interaktionstermer av tredje graden lyckades i vissa fall, men inte alla. Beräkningarna tog helt enkelt för lång tid, ibland över 8 timmar utan att bli färdiga.

5.6 Random Forest

Som ett komplement till de logistiska modellerna testar vi även modellen Random Forest som är en typ av trädmodell, se [3] sidan 587. Generellt i denna uppsats har vi modeller som predikterar utfall för variabeln $x_{k,1}$ och som förklarande regressionsvariabler har vi $s, a, o, \boldsymbol{x}_{\bullet,0}, x_{1,1}, \ldots, x_{k-1,1}$ tillgängliga. En trädmodell är en algoritm som rekursivt delar in data i halvor för en x-variabel i taget. Innan vi går vidare och beskriver algoritmen behöver vi parametrisera om variabeln o. Vi kodar om o från $o \in \{1, \ldots, 7\}$ till $o_1, \ldots, o_7 \in \{0, 1\}$ för att på så sätt hantera att o är en variabel där värdena inte har någon ordinalitet. Kalla antalet regressionsvariablerna som kommer användas $N_{RF} := 2 + 7 + n_e + (k - 1)$ där 2:an står för de två variablerna s och a och 7:an står för $o_1, \ldots, o_7$. Kalla dessa regressionvariabler för $\tilde{x} = (\tilde{x}_1, \ldots, \tilde{x}_{N_{RF}})$, där $\tilde{x}_1 = s, \tilde{x}_2 = a, \tilde{x}_3 = o, \tilde{x}_4 = x_{1,0}, , \ldots, \tilde{x}_{N_{RF}=x_{k-1,1}}$, detta underlättar notationen senare. Den data vi har tillgång till för att skatta modellen är $\mathcal{M}^{\text{train}}$, se 3.3. En trädmodell delar in $\mathcal{M}^{\text{train}}$ i disjunkta områden $\mathcal{M}_1^{\text{train}}, \ldots, \mathcal{M}_r^{\text{train}}$ där prediktionen för $p_{k:1-k}$ är c_j för alla $\tilde{x}^i \in \mathcal{M}_j^{\text{train}}$. En Random Forest-modell är en metod för att sätta ihop B st trädmodeller till en enda modell genom att ta medelvärdet av dess prediktioner. För b=$1, 2, \ldots, B$ och träningsdata som har $n_{\text{train}} := \#\{\mathcal{M}^{\text{train}}\}$ antal observationer sker följande:

1. Dra ett urval $\tilde{\mathcal{M}}^{\text{train}}$ av storlek n_{train} med återläggning från $\mathcal{M}^{\text{train}}$.

2. Skapa ett beslutsträd genom att:

(a) Dra ett utval av storlek $n_{RF} = \text{avrunda}(\sqrt{N_{RF}})$ utan återläggning från de tillgängliga regressionsvariablerna $\tilde{x}$ ur $\tilde{\mathcal{M}}^{\text{train}}$. Kalla dessa variabler för $\tilde{x}^{\subset} = \tilde{x}_1^{\subset}, \ldots, \tilde{x}_{n_{RF}}^{\subset}$

(b) Hitta det $i \in \{1, \ldots, n_{RF}\}$ och det $j \in \{1, \ldots, n_{\text{train}}\}$ som minimerar summan av *Mis-classification error* (MCE):

$$MCE(R_1(i,j)) + MCE(R_2(i,j))$$

där

$$R_1(i,j) = \{\tilde{x}^{\subset} | \tilde{x}_j^{\subset} \leq \tilde{x}_j^{\subset,i}\}$$
$$R_2(i,j) = \{\tilde{x}^{\subset} | \tilde{x}_j^{\subset} > \tilde{x}_j^{\subset,i}\}$$

och där

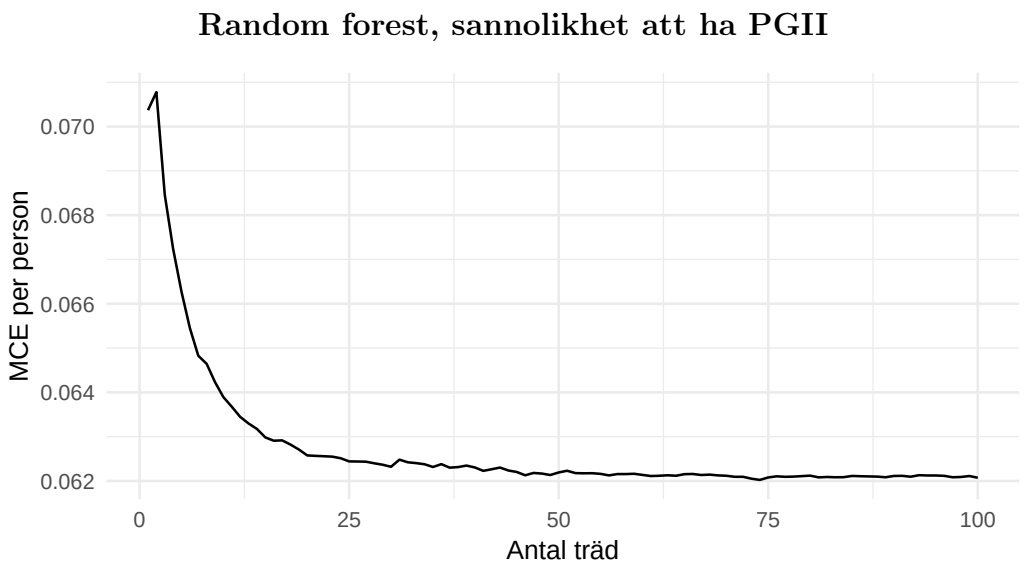
$$MCE(R) = 1 - \frac{1}{\#\{R\}} \sum_{x_{k,1}^i \in R} I(x_{k,1}^i = c(R))$$

där $c(R)$ i vår modell är det värde på $x_{k,1} \in \{0,1\}$ som är vanligast förekommande i området R:

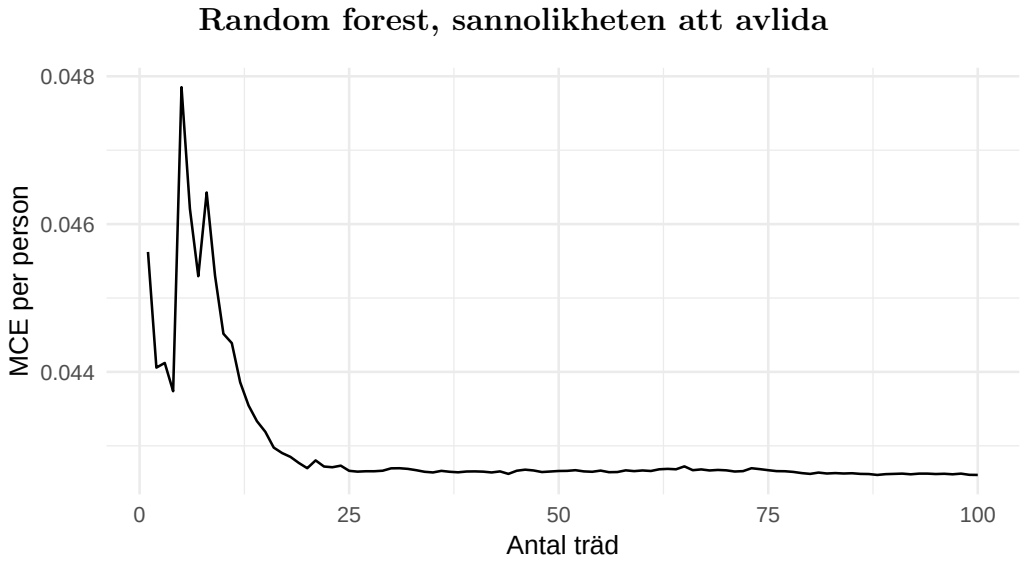
$$c(R) = \arg\max_x (\#I\{x_{k,1}^i \in R | x_{k,1}^i = x\})$$

Om $\#\{R_1(i,j)\} > 1$, börja om från punkt 2a med $\tilde{\mathcal{M}}^{\text{train}} := R_1(i,j)$, på samma sätt för $R_2(i,j)$.

När man byggt upp B st beslutsträd tar man medelvärdet av dess klassificeringar $c(R)$ för att få prediktionen. Ju fler träd man bygger desto bättre tenderar skattningarna att bli. I figur 4 ser vi hur MCE sjunker ju fler träd vi tar med. Tyvärr gick inte R-paket randomForest att beräkna för fler än 30-40 träd då hela $\mathcal{M}_{\text{train}}$ användes p.g.a. datorns begränsningar i kapacitet. För att kunna jämföra hur Random Forest presterade då man anpassar fler träd gjordes en anpassning till ett urval på en tiondel av datamängden ihop med jämförelser för två andra modeller, resultaten syns i figur 57 till 62. Det mindre urvalet gjordes för att datorkapaciteten skulle räcka till. Dessa figurer visar att Random Forest tycks ha ungefär samma prediktionsförmåga som de två andra modellerna då vi anpassar 100 träd. I figur 4 och 5 visas hur prediktionsfelet sjunker ju fler träd som anpassas och det ser ut som att prediktionsfelet inte går ner så mycket efter att antalet träd har överstigit 30. Hur figurerna hade sett ut om hela $\mathcal{M}^{\text{train}}$ hade använts i beräkningen går förstås inte helt att förutse.



Figur 4: MCE minskar ju fler träd som används i en Random Forest. Figuren visar modellen över sannolikheten att ha PGII.



Figur 5: Figuren visar MCE för modellen över sannolikheten att avlida.

5.7 Skattade parametrar med logistisk regression

I tabell 5.7 listas alla värden på β-parametrarna då man använder den kedja av betingningar som beskrivits genom ekvation (3) och använder lasso-regression. Man ska läsa tabellen som att man gör regression för de variabler som står som kolumnrubriker. Exempelvis den första variabeln *"har barn upp till ett års ålder"* beror på alla variabler vid tiden 0, alltså tiden före övergången. Nästa variabel *"har barn upp till åldern fyra år"* beror på den nyss nämnda samt alla variabler i tiden 0 o.s.v. I den här tabellen utelämnas variabeln *"har avlidit"* eftersom den hanteras separat enligt det som beskrivits i kapitel 4.1. I tabellen utelämnas kön, ålder och födelseland även de i själva verket är med i modellen. Eftersom lasso-regression använts har vissa β-parametrar krympts till noll, detta ger en indikation på att det inte finns något beroende. Om man tar som exempel, variabeln PGII som betyder *"har pensionsrätt från arbetsinkomst"* har ett beroende motsvarande ett β-värde på 4.17 för om man hade PGII året innan. Det finns också ett svagt positivt samband med om personen var boende i Sverige året innan. Vi betingar alltså inte på huruvida personen är i Sverige vid tiden 1 eftersom det ligger längre fram i kedjan av betingningar. Om man istället betraktar variabeln *bor i Sverige"* vid tiden 1 har den en samvariation med *"har PGII"* i tiden 1 som är av storleken 2.74. Vi kan konstatera att metoden med en kedja av betingningar fungerar så tillvida att det är en förhållandevis automatisk metod som ser till att de beroenden som finns tas med i modellen och beroenden som är svagare eller icke existerande ser ut att reduceras till en β-parameter med värdet 0.

koeff	tid	Barn01	Barn04	Educ	BPGB	IpCap	stud	PGII	SAPGB	SAPGI	TRANS	inSwe	isImig	isRet
Barn01	0	-1.24	-5.60	-0.20	0.02	0.00	-0.46	0.00	-0.38	0.00	1.43	0.07	0.10	0.00
Barn04	0	1.15	7.78	0.00	5.94	0.00	-0.20	0.00	-0.57	0.00	0.37	0.06	0.62	0.00
Educ	0	0.22	0.08	12.41	0.00	0.07	-2.60	0.00	0.00	0.00	0.00	-3.89	-0.34	-0.07
BPGB	0	-0.08	5.90	-0.04	2.15	1.40	0.56	0.00	0.00	0.00	0.11	0.00	0.00	0.00
IpCap	0	0.12	0.14	-0.81	0.08	7.12	0.00	0.32	0.45	0.00	0.15	-0.23	0.08	1.18
PGBSTUD	0	-0.38	-0.05	1.67	0.22	1.45	3.62	0.00	0.00	0.00	0.03	-0.46	0.39	0.00
PGII	0	0.48	0.00	0.38	-0.48	2.14	-0.50	4.17	-2.88	-1.12	0.00	-0.27	-1.06	0.00
SAPGB	0	0.00	-0.21	-0.28	0.00	0.00	0.00	0.00	9.56	-2.43	-0.03	0.00	0.00	0.08
SAPGI	0	0.00	-0.78	-0.01	0.00	0.00	0.00	0.00	0.00	10.43	-0.10	0.00	0.00	0.00
TRANS	0	0.01	1.41	-0.81	0.74	0.25	-0.51	0.00	2.31	1.09	2.72	0.00	-0.52	0.29
inSwe	0	0.36	0.13	0.57	1.05	0.75	1.41	1.83	0.00	0.00	1.16	11.22	-13.78	1.87
isImig	0	0.00	0.00	-0.98	0.08	0.00	0.68	-0.42	0.00	0.00	0.15	-2.50	21.57	-0.15
isRet	0	0.00	0.00	-0.00	0.00	0.00	0.00	0.00	-0.13	-0.32	-0.11	0.32	0.02	4.15
Barn01	1		0.09	-0.16	0.23	0.00	-0.37	0.00	-0.01	0.00	2.95	0.36	0.58	0.00
Barn04	1			0.00	0.06	0.00	-0.06	0.19	-0.43	0.00	0.78	0.14	0.19	0.00
Educ	1				0.00	0.00	3.37	0.24	-0.35	-0.02	-0.16	3.78	1.68	0.00
BPGB	1					0.09	0.45	0.00	0.00	0.00	0.00	0.00	-0.77	0.00
IpCap	1						0.04	0.32	0.25	0.00	0.32	-1.20	1.15	0.70
PGBSTUD	1							0.00	-0.41	0.00	-0.39	-0.15	0.00	0.00
PGII	1								3.39	12.14	1.18	2.29	2.74	-0.37
SAPGB	1									10.50	0.00	0.00	0.00	0.00
SAPGI	1										2.46	-0.30	0.00	-0.14
TRANS	1											0.64	0.00	0.00
inSwe	1												8.28	0.03
isImig	1													0.00

Tabell 1: Skattade värden på koefficienter med lasso-regression. I varje kolumn finns koefficienter för utfallet av $x_{k,1} = 1$ för ett k per kolumn. Varje rad innebär ett beroende till ett $x_{k,0}$ (den översta hälften av raderna), eller ett beroende till $x_{k,1}$ (de nedersta hälften av raderna). Egenskaperna listas i 2

6 Kalibrering av sannolikheter

Som vi tidigare nämnt används de olika regressionsmodellerna för att skapa en uppsättning $p^{s,a,o}_{\boldsymbol{x}^k \to \boldsymbol{x}^l}$ som anger sannolikheten att gå från ett tillstånd till ett annat mellan tiden 0 och tiden 1. Använder man modellen med dessa sannolikheter får man en stationär markovkedja. Det som önskas i Pensionsmodellen är sannolikheter som ändras i takt med de externa prognoser som vi vill ska styra utfallet framåt i tiden. I ekvation (1) står det att vi vill hitta ett $\tilde{p}^{s,a,o}_{\boldsymbol{x}^k \to \boldsymbol{x}^l}(t)$ så att

$$R^{(\mathcal{S},\mathcal{A},\mathcal{O})}_{\mathcal{X}^k \to \mathcal{X}^l} = \sum_{\substack{s \in \mathcal{S}, a \in \mathcal{A}, o \in \mathcal{O}, \\ \boldsymbol{x}^i \in \mathcal{X}^k, \boldsymbol{x}^j \in \mathcal{X}^l}} \tilde{T}^{(s,a,o)}_{\boldsymbol{x}^i \to \boldsymbol{x}^j}$$

Som ett exempel säger SCB:s befolkningsprognos från 2018 att 1 640 män kommer avlida vid 90 års ålder år 2025. Med våra beteckningar blir det

$$\mathcal{S} = \{man\}, \mathcal{A} = \{90\}, \mathcal{O} = \{1, 2, ..., 7\}$$

$$\mathcal{X}^i = \{\forall \boldsymbol{x}_k | \boldsymbol{x}^k_{\text{lever}} = 1, \boldsymbol{x}^k_{\text{bor i Sverige}} = 1\}$$

$$\mathcal{X}^j = \{\forall \boldsymbol{x}_k | \boldsymbol{x}^k_{\text{lever}} = 0\}$$

$$R^{(\mathcal{S},\mathcal{A},\mathcal{O})}_{\mathcal{X}^i \to \mathcal{X}^j}(t) = 1640$$

Eftersom $R^{(\mathcal{S},\mathcal{A},\mathcal{O})}_{\mathcal{X}^i \to \mathcal{X}^j}(t)$ och $N^{s,a,o}_{\boldsymbol{x}^i}(t)$ är givna är det bara $\tilde{p}^{s,a,o}_{\boldsymbol{x}^k \to \boldsymbol{x}^l}(t)$ som går att ändra. Det vi gör är att dela upp $p^{s,a,o}_{\boldsymbol{x}^k \to \boldsymbol{x}^l}(t)$ enligt ekvation (3). Därefter justeras beståndsdelarna $p_{k|k-1}$ genom:

$$p^{\text{just}}_{k|k-1} = f_{adj}\Big(p_{k|k-1}, \delta\Big)$$

där

$$f_{adj}(p, \delta) := \text{logit}^{-1}(\text{logit}(p) + \delta)$$

och $-\infty < \delta < \infty$. Funktionen $f_{adj}()$ är vald så att den kan appliceras på alla sannolikheter skilda från 0 och 1 med samma δ. Om en sannolikhet är precis 0 eller 1 motsvaras det av $-\infty$ resp $+\infty$ vilket inte går att justera på det viset. Ta som exempel att man vill höja sannolikheterna, det kan vara lockande att höja sannolikheterna med en viss faktor, säg v, enligt $\tilde{p} = p \cdot v$. Det motsvarar i någon mening att justera linjärt på log-skalan istället för att justera linjärt på logit-skalan. Ett problem med den metoden är att man

riskerar att vissa sannolikheter hamna över 1. Logit-funktionen är en mer praktisk funktion att använda eftersom den gör att en sannolikheten p alltid är i intervallet $-\infty < p < \infty$. Att justera med funktionen $f_{adj}()$ blir samma sak som att höja intercept-parametern i den aktuella logit-modellen. För att veta vilket δ som löser den aktuella ekvationen behöver man prova sig fram, t.ex. med Newton-Rapson-metoden tills man hittar ett värde som gör att differensen till ens målvärde är inom den tolerans man har bestämt sig för. Om man justerar variabeln X_k så att förekomsten av $X_k = 1$ ändras kommer utfallet av alla variabler X_l där $l > k$ att ändras för vilka det finns $\beta_k \neq 0$. Rent praktiskt är det alltså lämpligt att börja justera i ordningen $X_1, X_2, ...$ När man är nöjd med utfallet av $X_1 = 1$ kalibrerar man utfallet av $X_2 = 1$ o.s.v. När man justerar utfallet av $X_k = 1$ kan man dock vara säker på att det inte påverkar utfallet av $X_m = 1$ för $m < k$. Det är svårt att argumentera för varför just detta sätt att justera skulle vara mer lämpligt än något annat. Ett argument för denna metod är att det är mer naturligt att justera α-parametrarna än att ändra t.ex. β-parametrarna vilket skulle ändra beroenden mellan de olika utfallen. Andra metoder för att empiriskt testa hur rimlig denna justering är har inte undersökts i denna uppsats.

7 Utvärdering av modeller

För att jämföra de olika modellerna vill vi utvärdera hur bra de är på att prediktera utfall som modellerna inte har anpassats till. Processen sker i två steg.

1. Anpassa modellen på $\mathcal{M}_{\text{train}}$. För lasso, spline och Random Forest sker detta flera gånger. För lasso och spline sker korsvalidering där det mest passande λ-värdet hittas och sedan skattas modellen utefter det. För lasso och spline-modellerna görs detta utifrån måttet binomial deviance, detta är inbyggt i paketen glmnet som användes för lasso och mgcv som användes för splines.

2. Modellen som skattades i punkt 1 används på data i $\mathcal{M}_{\text{test}}$. I det här läget kan vi välja mått på felet som modellerna ger mera fritt än vi kan i de inbyggda funktionerna i glmnet, mgcv och randomForest. Vi har valt att titta på tre mått med betoning på det första för att utvärdera modellerna.

(a) Residual deviance, beskrivet i kapitel 7.1. Eftersom lasso och spline-modellerna har använt samma mått för att välja ut den bästa modellen är det intressant att utvärdera de olika varianterna sida vid sida då vi använder detta mått på $\mathcal{M}_{\text{test}}$.

(b) Fel i kronor, hela poängen med modellen är att modellera vilka inkomster som kommer in i pensionssystemet och hur folks pensioneringsmönster ser ut. Till detta har vi konstruerat ett mått som ska ge en uppskattning på vad felen i skattningarna gör för skillnad i kronor. Metoden är beskriven i kapitel 7.2

(c) Mis-classification error (MCE), beskrivet i kapitel 7.4. Vi undersöker detta mått därför att samma mått används då Random Forest-modellerna skapas.

Teoretiskt skulle man kunna välja modell i punkt 2 efter något annat mått än de som är förvalda i R-paketen glmnet och mgcv, på liknande sätt hade man kunnat utföra partitioneringarna i Random Forest efter ett annat mått än MCE. Det är dock inte möjligt att välja något annat i de färdigbyggda paketen. Om vi hade programmerat våra egna implementationer av lasso-regression, spline-regression och Random Forest hade vi haft frihet att helt välja mått men det bedömdes som allt för tidskrävande för att rymmas i den här uppsatsen. Av den anledningen valdes detta upplägg där man gör korsvalidering på en mängd och en slags validering på en helt annat mängd. Då kan man i valideringen utvärdera modellerna med vilket mått som helst utan att vara beroende av de R-paketen som valt modellerna. Det är dock troligt att det spelar roll för utvärderingen vilket mått som är valt då modellerna skattades.

7.1 Residual deviance

Deviance, definierades i kapitel 5.2.1 som:

$$\begin{aligned} D(\tilde{p}_{k|k-1}, y) &= -2(\log(\mathcal{L}(\hat{p}_{k|k-1}|y)) - \log(\mathcal{L}(\tilde{p}_{k|k-1}|y))) \\ &= -2(y \cdot \log(\hat{p}_{k|k-1}) + (n-y) \cdot \log(1-\hat{p}_{k|k-1}) \\ &\quad -y \cdot \log(\tilde{p}_{k|k-1}) + (n-y) \cdot \log(1-\tilde{p}_{k|k-1}) \end{aligned} \tag{15}$$

Termen $\binom{n}{y}$ försvinner eftersom den finns med i båda log-likelihood-termerna. För att kunna utvärdera prediktionsförmågan skattas $\tilde{p}_{k|k-1}$ på $\mathcal{M}_{\text{train}}$ och $\hat{p}_{k|k-1}$ på $\mathcal{M}_{\text{test}}$. När $D(\tilde{p}_{k|k-1}, y)$ ska beräknas är y hämtat från $\mathcal{M}_{\text{test}}$. Om

en skattning $\tilde{p}_{k|k-1}$ blir exakt 0 eller 1 tillsammans med $y > 0$ respektive $y = 0$ är deviance-måttet odefinierat. Det är rent teoretiskt inget problem när vi har logit-modeller där $\tilde{p}_{k|k-1}$ aldrig riktigt blir exakt 0 eller 1. Däremot har modellen som bygger rakt av på empiriska sannolikheter möjlighet att producera skattningar som är exakt 0 eller 1, det gäller även i vissa fall för Random Forest. För att kunna jämföra alla modeller med varandra har vi därför valt att sätta en begränsning på skattningarna enligt:

$$\tilde{p}_{k|k-1} = \begin{cases} \text{logit}^{-1}(5) & , \tilde{p}_{k|k-1} > \text{logit}^{-1}(5) \\ \text{logit}^{-1}(-5) & , \tilde{p}_{k|k-1} < \text{logit}^{-1}(-5) \\ \tilde{p}_{k|k-1} & , \text{annars} \end{cases}$$

Där $\text{logit}^{-1}(x) = \frac{\exp(x)}{1+\exp(x)}$. I diagram med udda nummer 31, 33, ..., 55 syns skattningarna av residual deviance baserat på testdata. I figurer med jämna nummer 32 till 56 ser vi felen uträknade i ekonomiska termer utifrån det vi kommer beskriva i nästa kapitel, 7.2.

7.2 Fel i kronor

Vi påminner oss om att Pensionsmodellens syfte är att prognostisera pensionsinkomster och även pensionsutbetalningar. Pensionsutbetalningarna är helt regelstyrda utifrån individens pensionskapital och ålder, inkomsterna som ligger till grund för pensionsrätterna är däremot slumpmässiga och det är därför vi intresserar oss för att bygga modeller för dem. Hittills har vi inte undersökt vilken skillnad de olika modellerna gör i ekonomiska termer, vi har enbart studerat hur antalet i $\mathcal{M}_{\text{test}}$ avviker från det som vi erhåller från modellernas prediktioner. Beräkningen av residual deviance på $\mathcal{M}_{\text{test}}$ ger en fingervisning om hur träffsäkra modellerna är, vi vill dock komplettera med ett mått som säger hur mycket modellernas felmarginal påverkar i pengar. Mer specifikt betraktar vi en individ eller en grupp av individer som delar samma $s, a, o, \boldsymbol{x}_{\bullet,0}, x_{1,1}, \ldots, x_{k-1,1}$ och ser hur stora inkomster $m_{k|k-1}$ de får. Subskriptet $k|k-1$ i $m_{k|k-1}$ betyder att vi precis som förut betingar på $s, a, o, \boldsymbol{x}_{\bullet,0}, x_{1,1}, \ldots, x_{k-1,1}$. Utifrån detta beräknas inkomsten $m_{k|k-1}$ uppdelat på utfallet på variabeln $x_{k,1}$ enligt:

$$\hat{m}_{k|k-1} := \hat{p}_{k|k-1}\hat{m}_{k_1|k-1} + (1 - \hat{p}_{k|k-1})\hat{m}_{k_0|k-1}$$

Där $\hat{m}_{k_1|k-1}$ betecknar inkomsterna för de fall då $x_{k,1} = 1$ och $\hat{m}_{k_0|k-1}$ betecknar de inkomsterna då $x_{k,1} = 0$. Skattningarna $\hat{p}_{k|k-1}, \hat{m}_{k_0|k-1}, \hat{m}_{k_1|k-1}$ är

från $\mathcal{M}_{\text{test}}$. Vi definierar nu skattningen där vi har bytt ut $\hat{p}_{k|k-1}$ mot $\tilde{p}_{k|k-1}$:

$$\tilde{m}_{k|k-1} := \tilde{p}_{k|k-1}\hat{m}_{k_1|k-1} + (1 - \tilde{p}_{k|k-1})\hat{m}_{k_0|k-1}$$

Där $\tilde{p}_{k|k-1}$ är beräknat på $\mathcal{M}_{\text{train}}$ och $\hat{p}_{k|k-1}$ på $\mathcal{M}_{\text{test}}$. Detta blir grunden i hur vi beräknar felet i ekonomiska termer då vi beräknar differensen mellan de olika skattningarna av $m_{k|k-1}$:

$$\begin{aligned}\tilde{m}_{k|k-1} - \hat{m}_{k|k-1} &= \tilde{p}_{k|k-1}\hat{m}_{k_1|k-1} + (1 - \tilde{p}_{k|k-1})\hat{m}_{k_0|k-1} \\ &\quad -(\hat{p}_{k|k-1}\hat{m}_{k_1|k-1} + (1 - \hat{p}_{k|k-1})\hat{m}_{k_0|k-1}) \\ &= (\tilde{p}_{k|k-1} - \hat{p}_{k|k-1})(\hat{m}_{k_1|k-1} - \hat{m}_{k_0|k-1})\end{aligned} \tag{16}$$

Måttet för avvikelse valdes till en kvadratsumma. Det finns andra mått på avvikelser men detta bedömdes som ett rimligt sådant, andra typer av mått har inte testats i den här uppsatsen. När vi beräknat denna residualkvadratsumma skalas den med antalet personer n och vi tar roten ur för att ha ett mått där man eventuellt kan få en uppfattning av felet på samma skala som de ursprungliga avvikelserna. Kalla detta mått för ϵ_k, vilket alltså beräknas enligt:

$$\epsilon_k := \sqrt{\frac{\sum\left((\tilde{p}_{k|k-1} - \hat{p}_{k|k-1})(\hat{m}_{k_1|k-1} - \hat{m}_{k_0|k-1})\right)^2}{n}}$$

Summering görs över alla $s, a, o, \boldsymbol{x}_{\bullet,0}, x_{1,1}, \ldots, x_{k-1,1}$. Detta värde visas i diagram med jämna nummer 32, 34, ..., 56.

7.3 Klassificeringsfel

I R-paketet randomForest byggs träden upp efter att man räknat ut hur många $x_{k,1}$-värden som är i majoritet i en viss mängd R, se kapitel 5.6. Enligt [3] sidan 309 finns flera mått som man kan använda för att utvärdera en skattning är och som därmed kan används för att bestämma vid vilken punkt man ska göra uppdelningen i halvor vid skapande av en Random Forest. Förutom mis-classification error (MCE) nämns även Gini-koefficent och deviance. Det hade varit intressant att välja deviance-måttet som mått för att dela upp beslutsträdet eftersom samma mått används i korsvalideringen i R-paketen som gör lasso- och spline-regression. I den här uppsatsen har vi dock inte lyckats ändra så att R-paketet som användes delade upp R efter deviance istället för MCE. Detta skulle kunna vara en nackdel när man jämför

prediktionsförmågan för Random Forest-modellerna eftersom vi i huvudsak använder residual deviance för att jämföra modellerna. Vi har dock beräknat miss-classification error på ett stickprov på en tiondel av den totala datamängden, ställt in antalet träd till 100 och jämfört Random Forest-modellen med lasso + spline-modellen samt modellen där empiriska sannolikheter har använts i kombination med MLE+spline. Resultatet i figurerna 57, 59 och 61 visar att modellerna har överlag ganska lika prediktionsförmåga även då vi utvärderar dem med mis-classification error vilket borde vara mer till fördel för Random Forest modellen eftersom partitionerna som ligger till grund för modellen är gjorda efter detta mått.

7.4 Jämförelse av modeller

Om man granskar diagram med udda nummer 31, 33, ..., 55 som innehåller skattningarna av residual deviance baserat på testdata och i figurer med jämna nummer 32 till 56 ser vi felen i ekonomiska termer. De modeller som jämförs är:

- *MLE*: logistisk regression med beroendet till ålder i form av ett andragradspolynom enligt kapitel .
- *Lasso med ålder*: logistisk regression med beroendet till ålder i form av ett andragradspolynom enligt kapitel 5.3.
- *Lasso + spline*: logistisk lasso-regression med beroendet till ålder i form av en kubisk splinefunktion enligt kapitel 5.4.
- *Lasso + spline 2*: logistisk lasso-regression med beroendet till ålder i form av en kubisk splinefunktion. Till kategorivariablerna har interaktion av andra ordningen lagts till enligt kapitel 5.5.
- *Lasso + spline 3*: logistisk lasso-regression med beroendet till ålder i form av en kubisk splinefunktion. Till kateogrivariablerna har interaktion av tredje ordningen lagts till enligt kapitel 5.5. Denna modell visade sig så tunga att beräkna att den har inte anpassats för alla k i $k_{k,1}$.
- *Empiriska skattningar*: Empiriska skattningar enligt kapitel 5.1. I de fall då modellen behöver producera en skattning på en övergång som inte

fanns i observationerna i $\mathcal{M}_{\text{train}}$ utvidgades skattningen enligt schemat i lista 5.1.1.

- *Empiriska skattningar + MLE +spline*: Empiriska skattningar enligt kapitel 5.1. I de fall då modellen behöver producera en skattning på en övergång som inte fanns i observationerna i $\mathcal{M}_{\text{train}}$ användes en MLE + spline-modell.

- *Random Forest*: Random Forest-modell med 30 st träd enligt kapitel 5.6. Vi hade helst gjort en Random Forest-modell med fler träd än så, men det visade sig att datorn inte räckte till. Vi undersökte även prediktionsförmågan då ett urval på en tiondel användes. Resultatet visas i figurerna 57 till 62.

När vi granskar resultatet ser man för det första att det är svårt att utse en tydlig vinnare. Följande modeller ger liknande prediktionsförmåga överlag:

- MLE + spline
- Lasso + spline
- Lasso med interaktioner + spline
- Den empiriska modellen + MLE + spline.

Vilken man väljer av dessa beror på hur man prioriterar saker som modellens enkelhet, en praktiskt struktur och beräkningstider då modellerna ska skattas.

8 Diskussion

Vi har gjort numeriska utvärderingar av några olika parametriska modeller inklusive den med maximalt antal parametrar, som vi kallar den empiriska modellen. Ett skäl som talar emot modellen som har helt empiriska skattningar är att det ibland dyker upp exempel på data som den inte sett förut när den ska användas på nytt data. De logistiska modellerna däremot har fördelen att de i kan producera prediktioner även i exempel på övergångar som inte fanns med i historiska data. Vidare föredras en modell som ger en bra prediktiva förmåga parat med en så enkel struktur som möjligt. Om man ska försöka jämföra den prediktiva förmågan hos modellerna som jämförs går det inte att se något tydligt svar på vilken av modellerna som är bäst i alla situationer. Följande modeller uppvisar de bästa resultaten på prediktiv förmåga:

- MLE + spline
- Lasso + spline
- Lasso med interaktioner + spline
- Den empiriska modellen + MLE + spline (för de fall där den empiriska modellen inte kan ge någon prediktion).

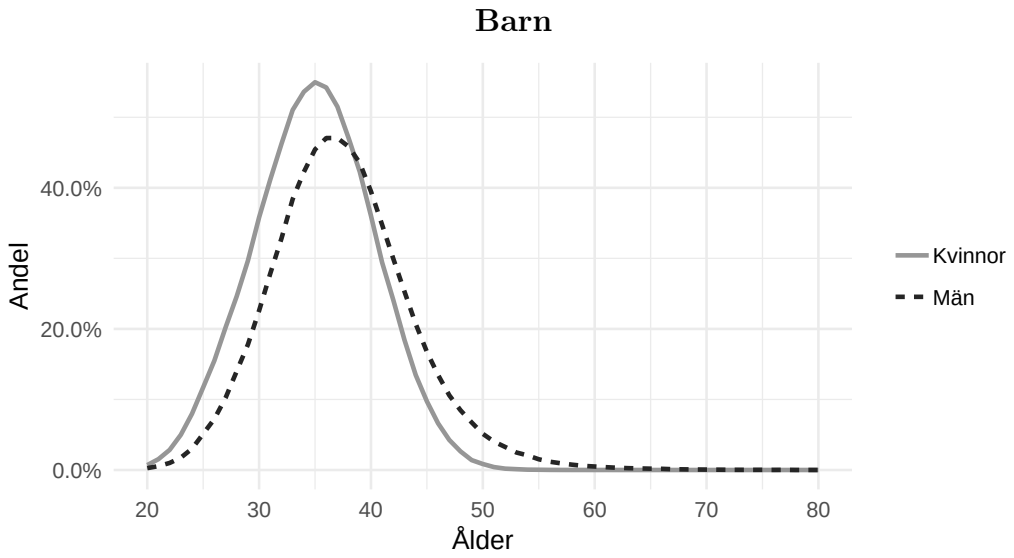
Det går förstås att välja olika typer av modeller för olika variabler. T.ex. ger lasso-modellen där åldern modelleras som en kategori-variabel bra resultat när det gäller sannolikheten att gå i pension (se figur 55), men inte i så hög utsträckning för de andra x-variablerna. Det kan bero på att det finns en hård 61-årsgräns inbyggd i pensionssystemet (se figur 16), det går inte att ta ut sin inkomstpension innan man fyllt 61 år. Spline-funktionerna som har en mer kontinuerlig karaktär ser ut att ge sämre resultat vid prediktion för denna x-variabel. En idé är att utöver spline-funktionen lägga till en indikatorvariabel för att ha uppnått åldern 61 år och kanske även ytterligare en indikatorvariabel för att ha uppnått 65 års ålder vilket är en annan gräns i pensionssystemet. Det kanske skulle kunna ge modellen möjlighet att ta hänsyn till en tröskel som är diskret och inte någon slät funktion. Vi ser det som en fördel om samma typ av modell kan väljas till alla x-variablerna. Mest för överskådlighetens skull men även för att det blir enklare i t.ex. en situation då man vill programmera om allt i ett nytt programspråk. Av dessa skäl tycker jag att MLE + spline är det rimligaste alternativet. När

man ser till prediktionsförmågan är denna modell aldrig det alternativ som ger lägst uppskattat fel, men nästan alltid bland de bättre och en av de enklare modellerna i sin uppbyggnad bland de som jämförs. Det går även att undersöka parametrarna för att dra slutsatser som finns om beroende mellan x-variablerna vilket kan vara av intresse. Ytterligare en fördel är att det skulle kunna gå att justera parametrarna manuellt i modellerna för att hantera vissa problem som kan uppstå t.ex. p.g.a. bristande datakvalitet.

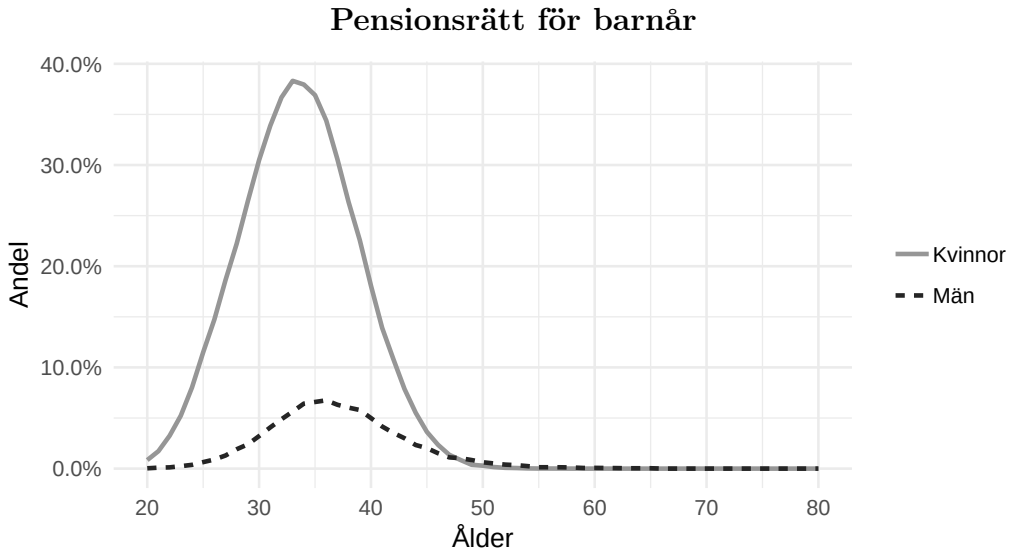
9 Appendix

9.1 Diagram

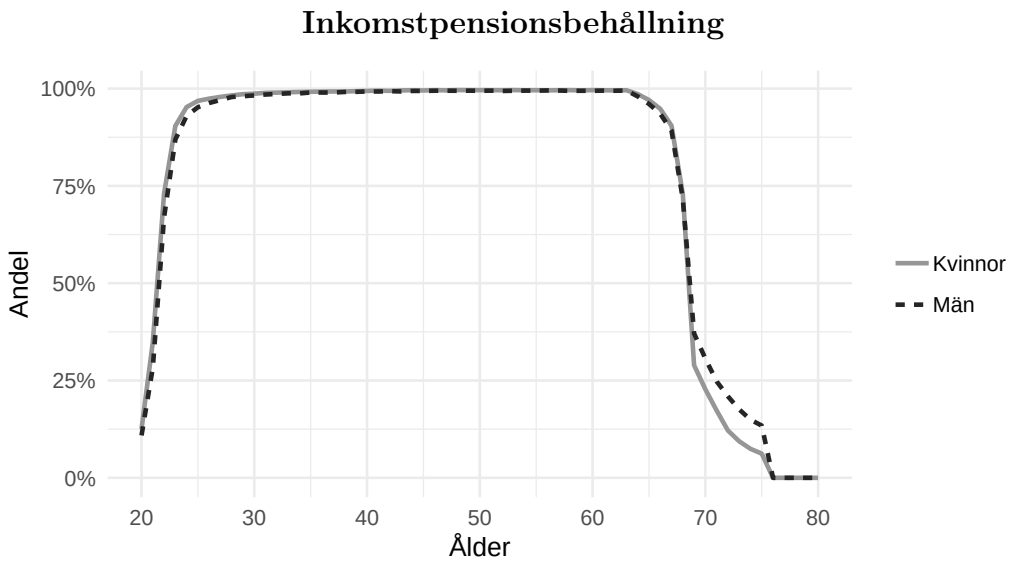
9.1.1 Förekomster av egenskaper i data



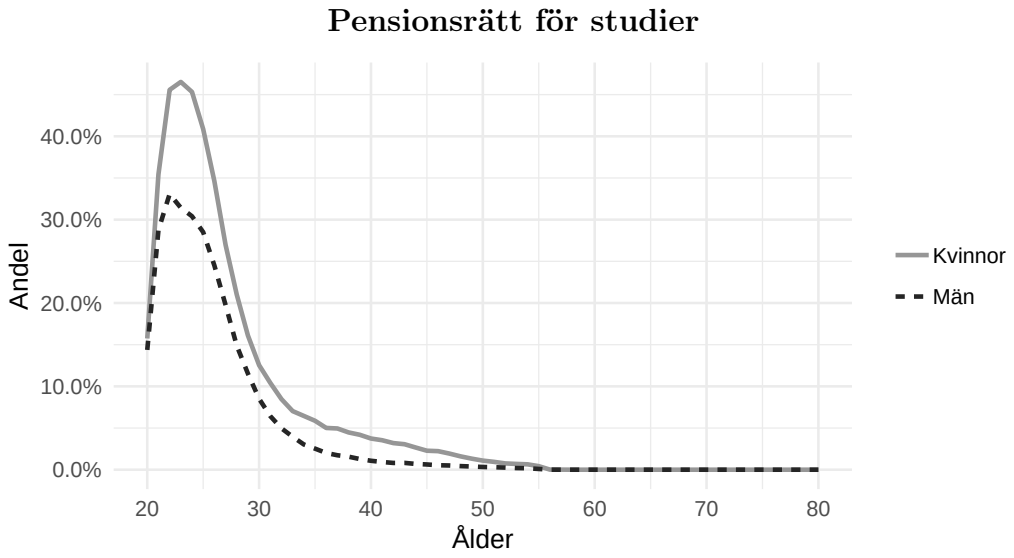
Figur 6: Andel av befolkningen, uppdelat på kön och ålder, som har barn upp till ett års ålder.



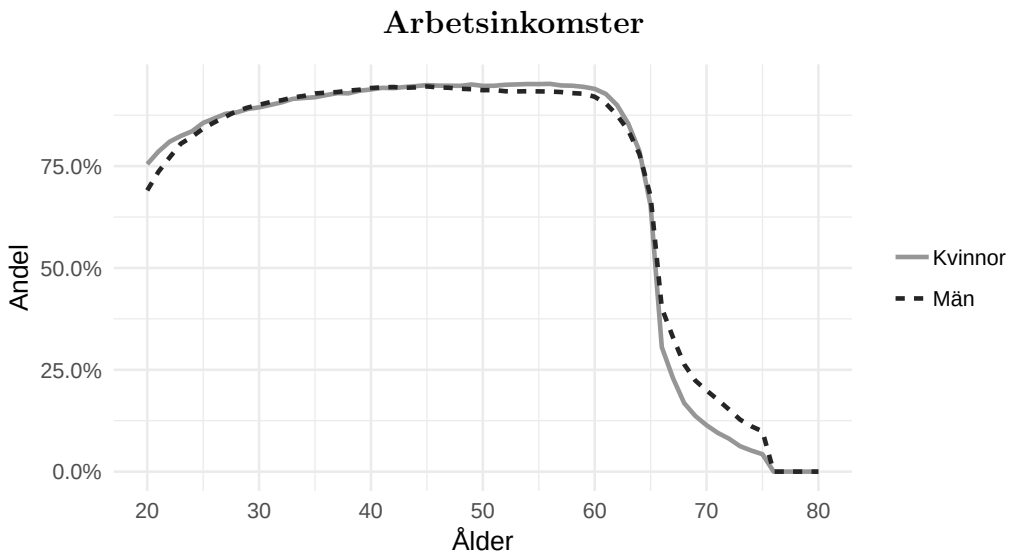
Figur 7: Andel av befolkningen, uppdelat på kön och ålder, som har pensionsrätt för barnår. Pensionsrätt är detsamma som premie. Pensionsrätten tillfaller i regel den förälder som hade lägst inkomst innan barnet föddes.



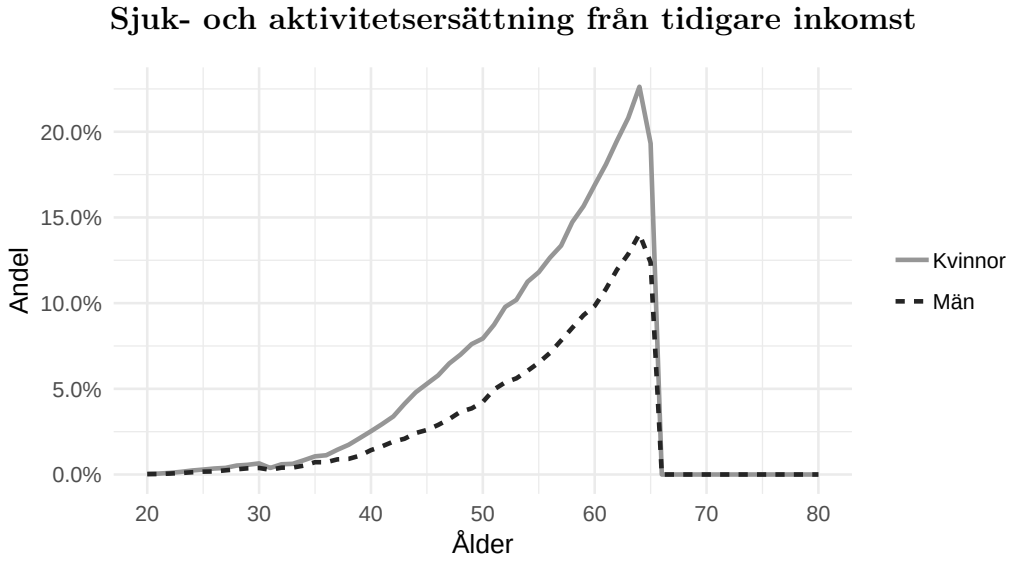
Figur 8: Andel av befolkningen, uppdelat på kön och ålder, som har ett inkomstpensionskapital.



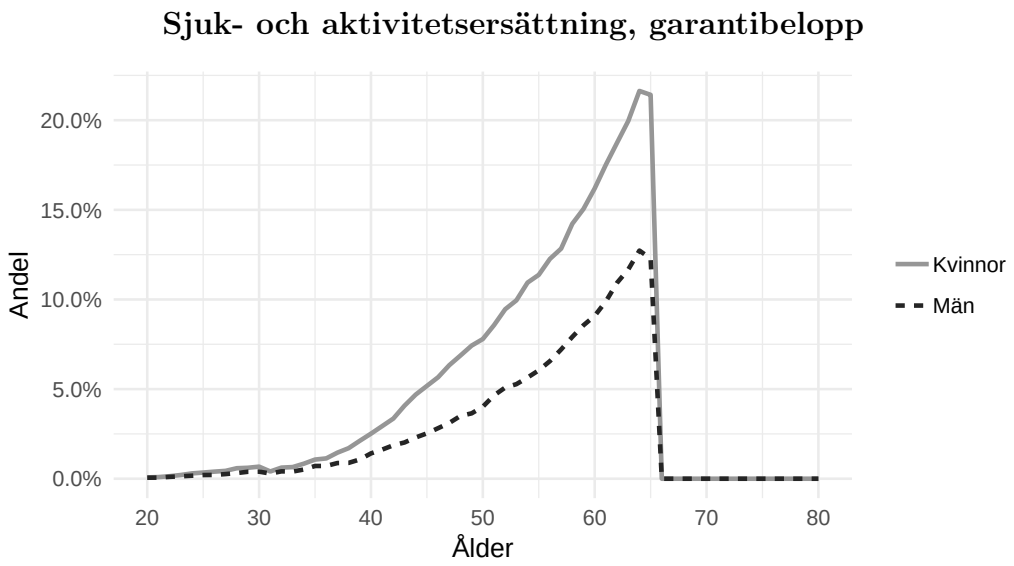
Figur 9: Andel av befolkningen, uppdelat på kön och ålder, som har pensionsrätt för studier.



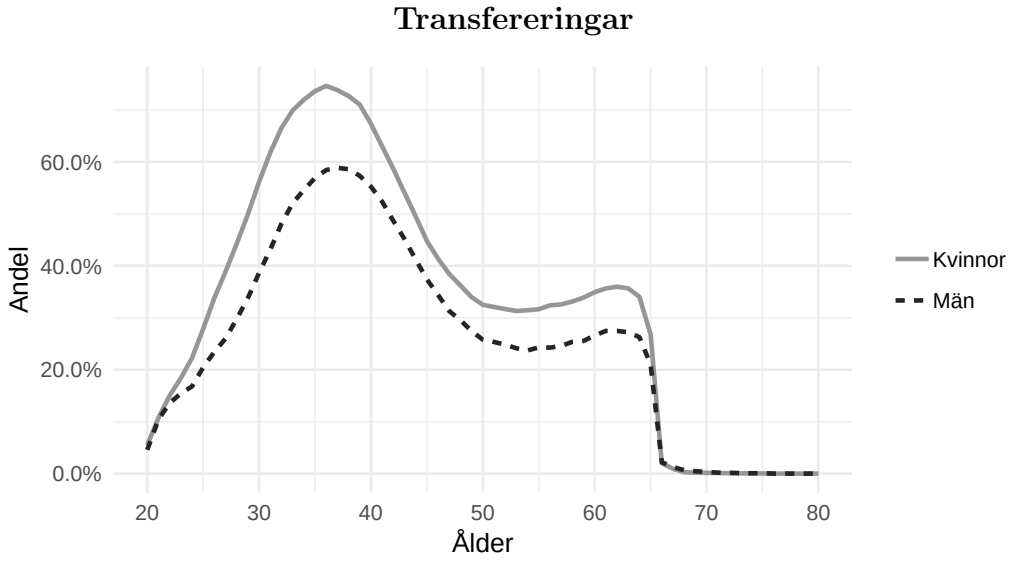
Figur 10: Andel av befolkningen, uppdelat på kön och ålder, som har pensionsrätt från arbetsinkomster.



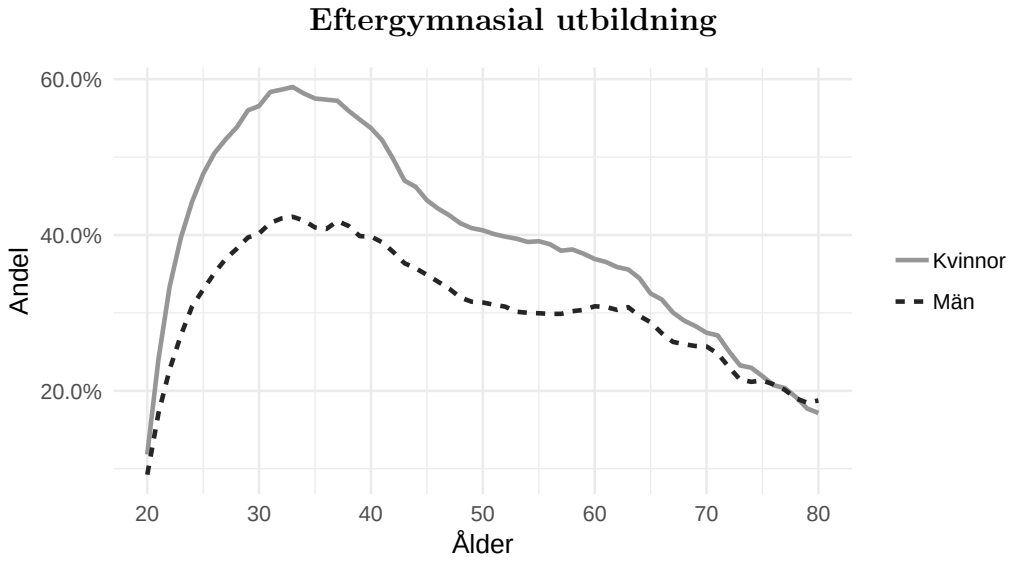
Figur 11: Andel av befolkningen, uppdelat på kön och ålder, som har pensionsrätt från sjuk- och aktivitetsersättning baserat på tidigare inkomst.



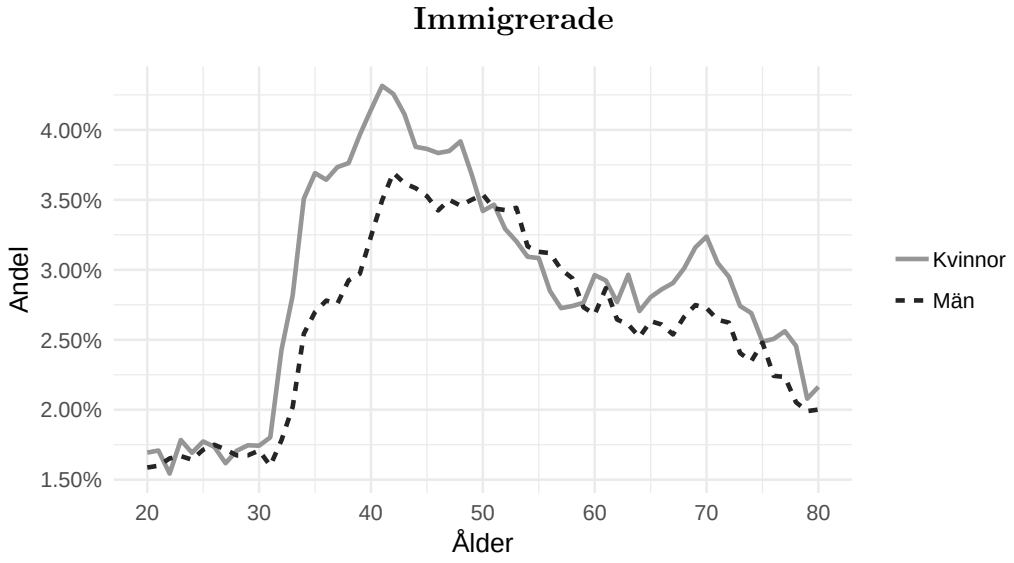
Figur 12: Andel av befolkningen, uppdelat på kön och ålder, som har pensionsrätt från sjuk- och aktivitetsersättning baserat garantibelopp.



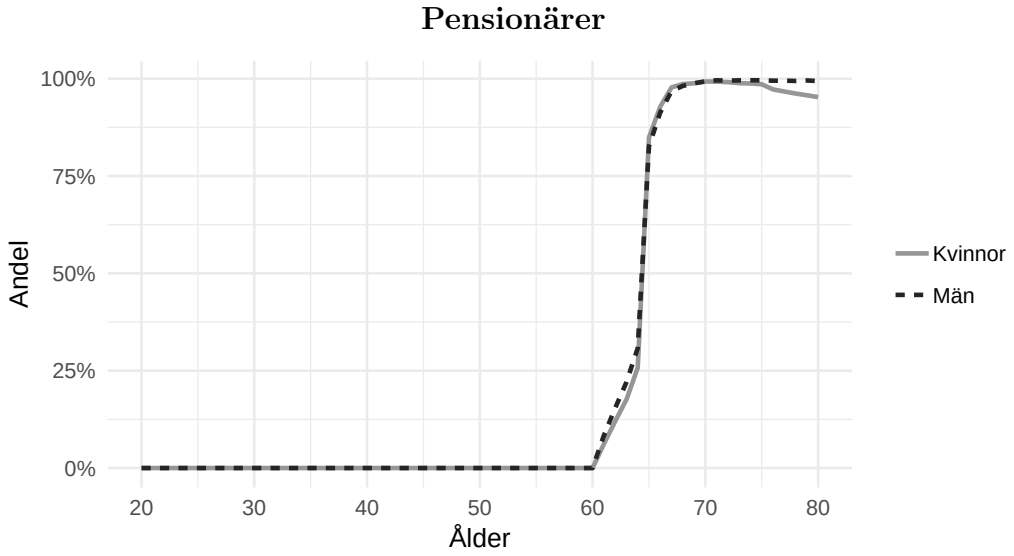
Figur 13: Andel av befolkningen, uppdelat på kön och ålder, som har pensionsrätt från andra transfereringar än sjuk- och aktivitetsersättning.



Figur 14: Andel av befolkningen, uppdelat på kön och ålder, som har en eftergymnasial utbildning.

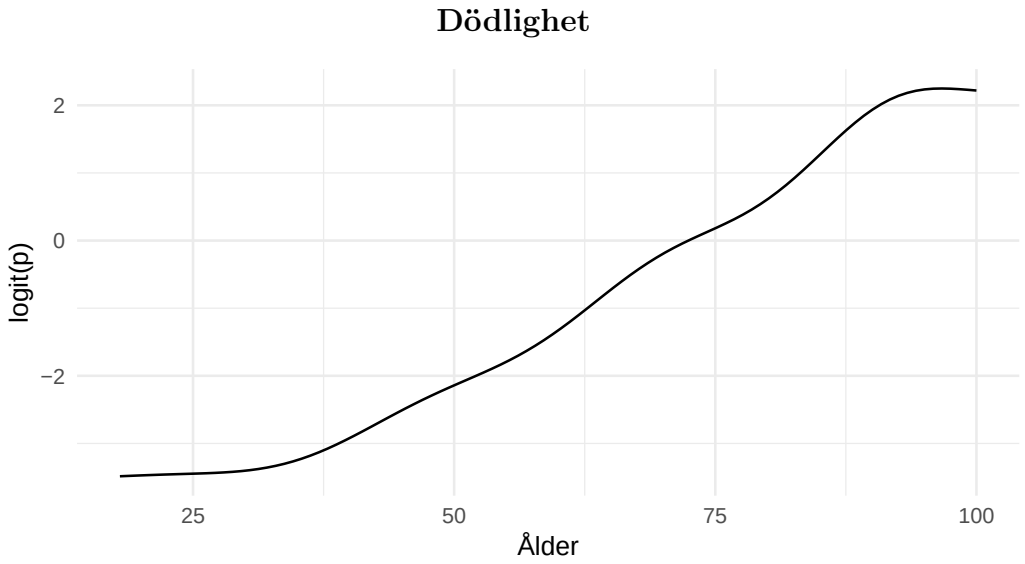


Figur 15: Andel av befolkningen (födda i Sverige), uppdelat på kön och ålder, som har immigrerat tillbaka till Sverige.



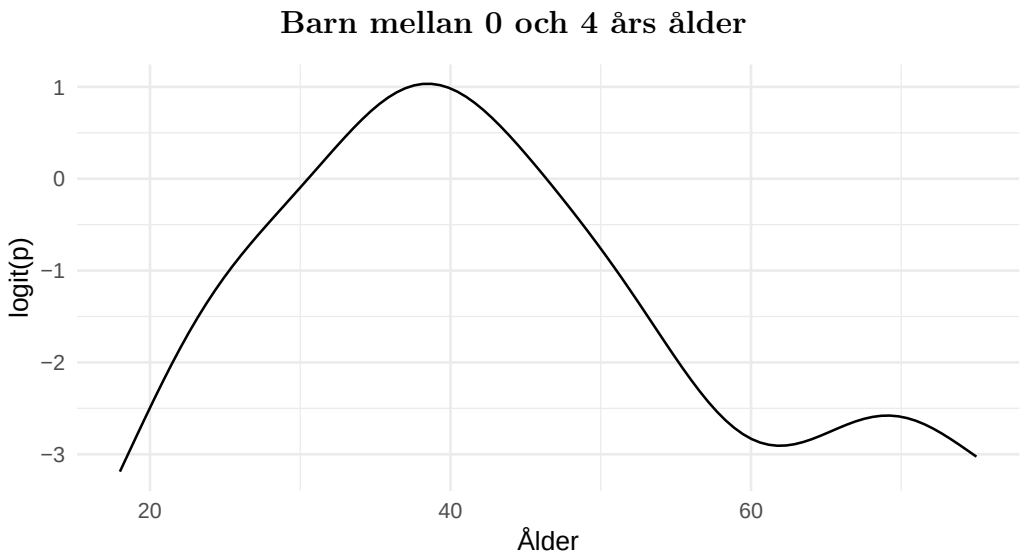
Figur 16: Andel av befolkningen, uppdelat på kön och ålder, som har börjat ta ut sin inkomstpension. Nästan alla pensionärer har tjänat in till inkomstpension i någon utsträckning. För de som har låg eller ingen inkomstpension finns garantipension och bostadstillägg.

9.1.2 Anpassning av spline-funktion

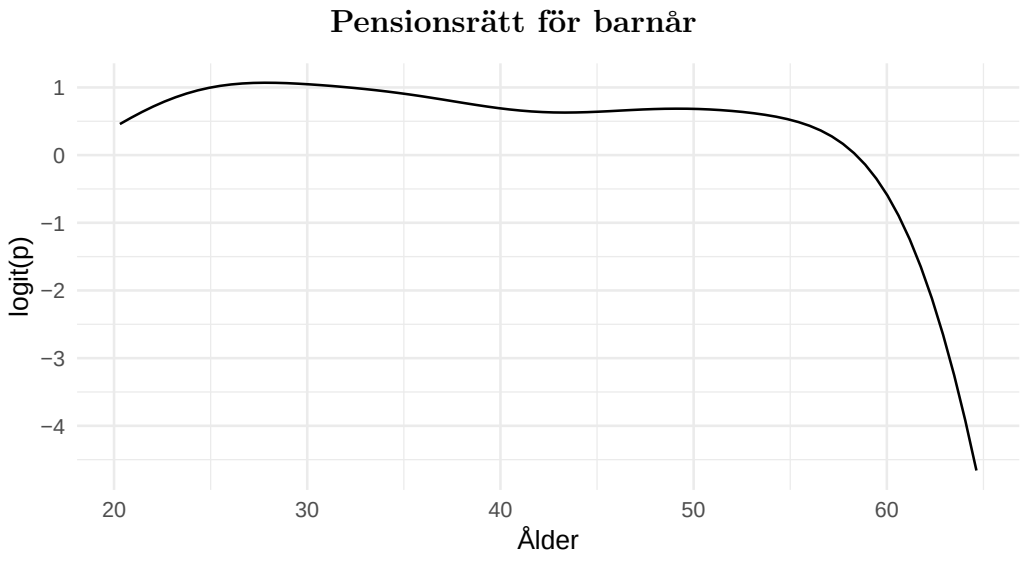


Figur 17: Anpassning av kubisk splinefunktion till åldern (a) då effekten av kategorivariablerna modellerats med logistisk regression

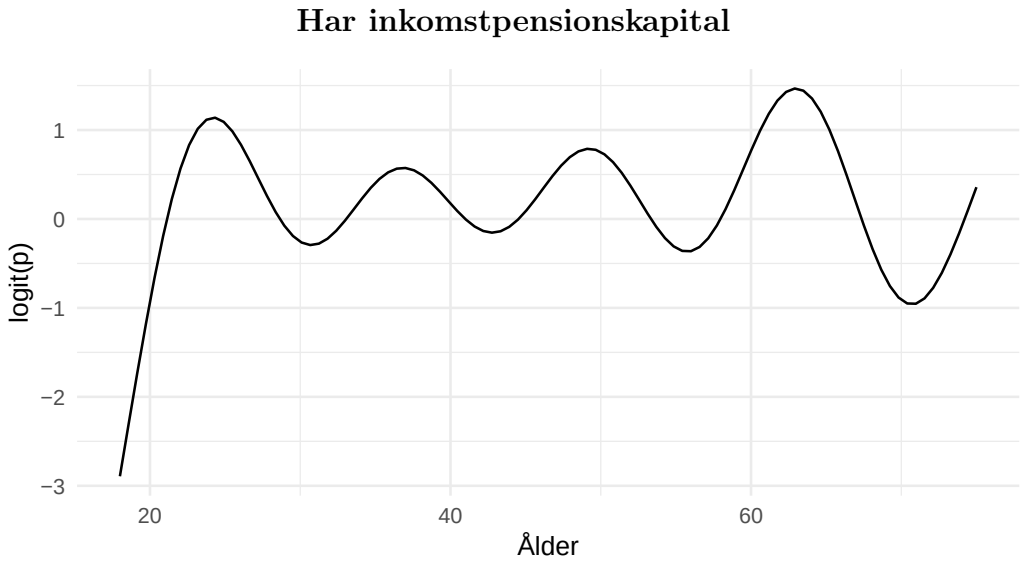
Figur 18: Anpassning av kubisk splinefunktion till åldern (a) då effekten av kategorivariablerna modellerats med logistisk regression



Figur 19: Anpassning av kubisk splinefunktion till åldern (a) då effekten av kategorivariablerna modellerats med logistisk regression

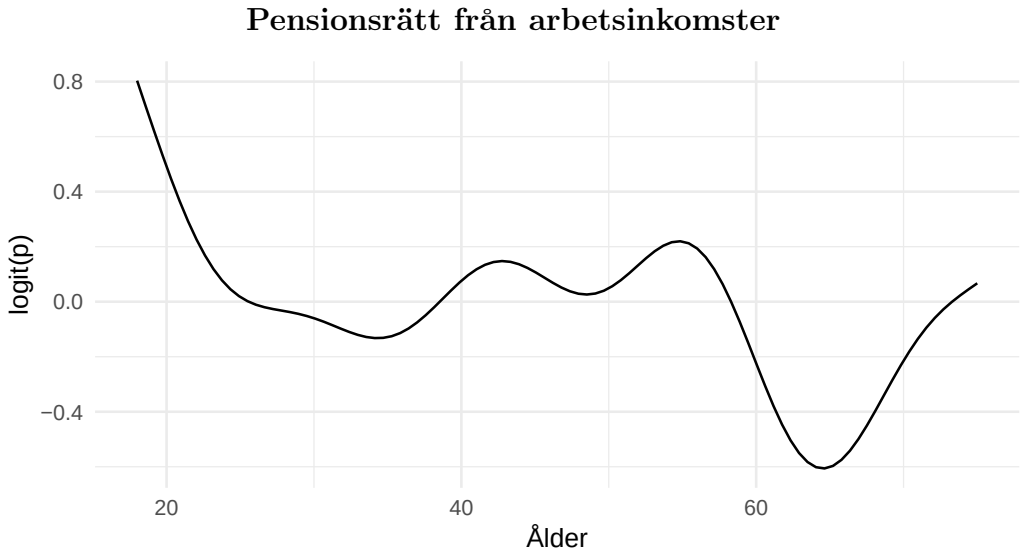


Figur 20: Anpassning av kubisk splinefunktion till åldern (a) då effekten av kategorivariablerna modellerats med logistisk regression

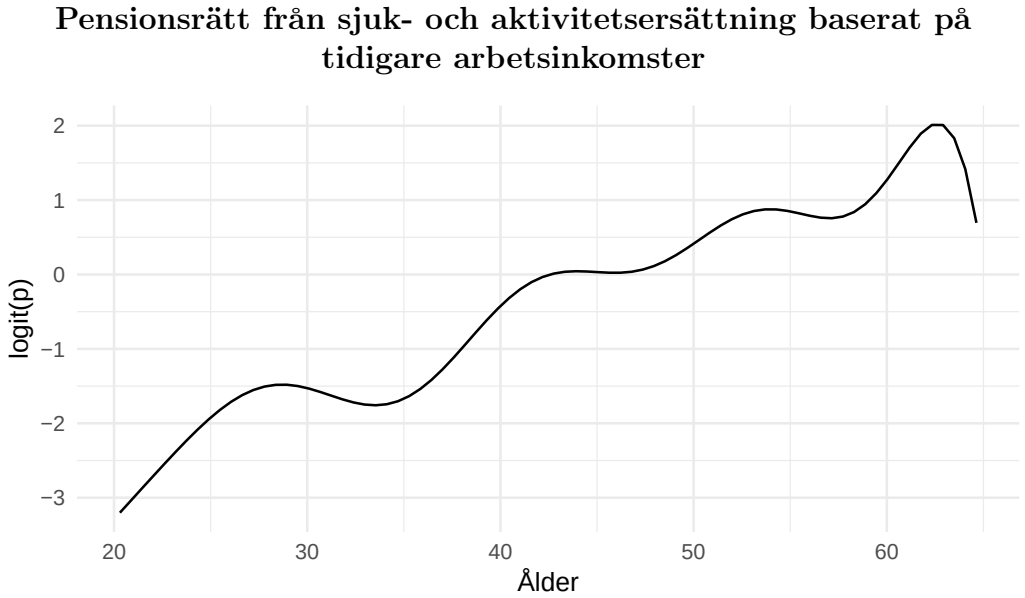


Figur 21: Anpassning av kubisk splinefunktion till åldern (a) då effekten av kategorivariablerna modellerats med logistisk regression

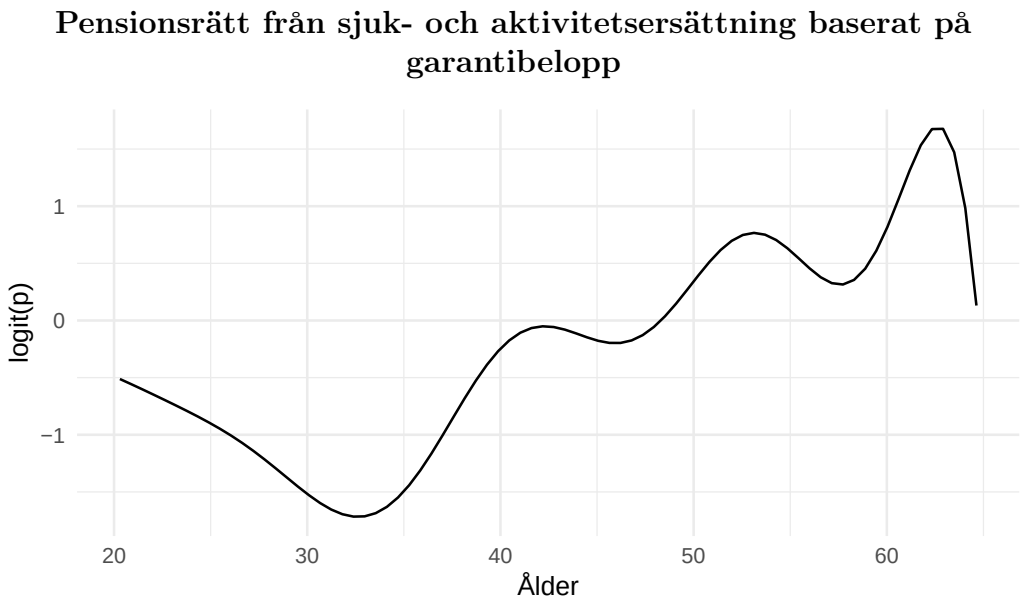
Figur 22: Anpassning av kubisk splinefunktion till åldern (a) då effekten av kategorivariablerna modellerats med logistisk regression



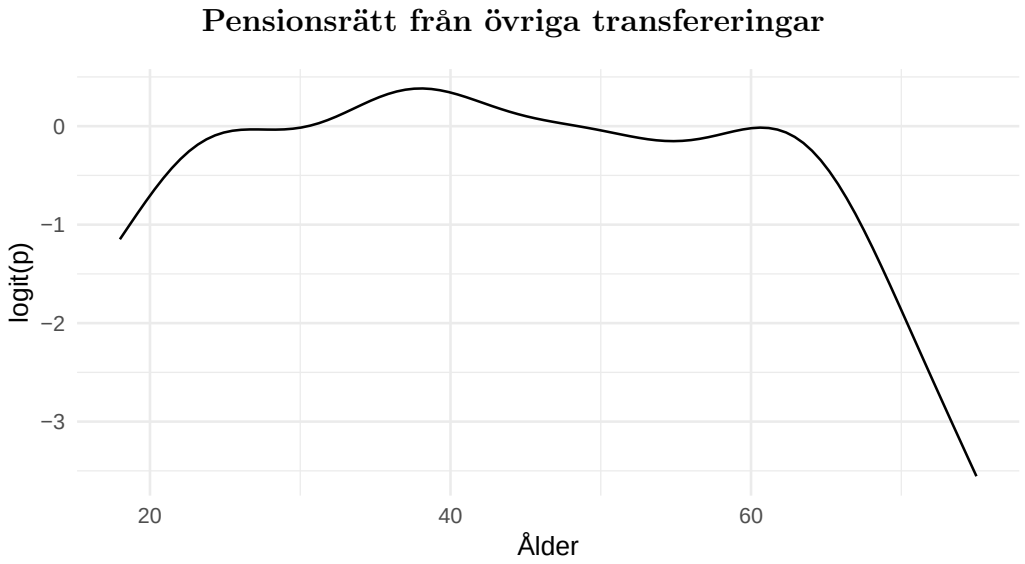
Figur 23: Anpassning av kubisk splinefunktion till åldern (a) då effekten av kategorivariablerna modellerats med logistisk regression



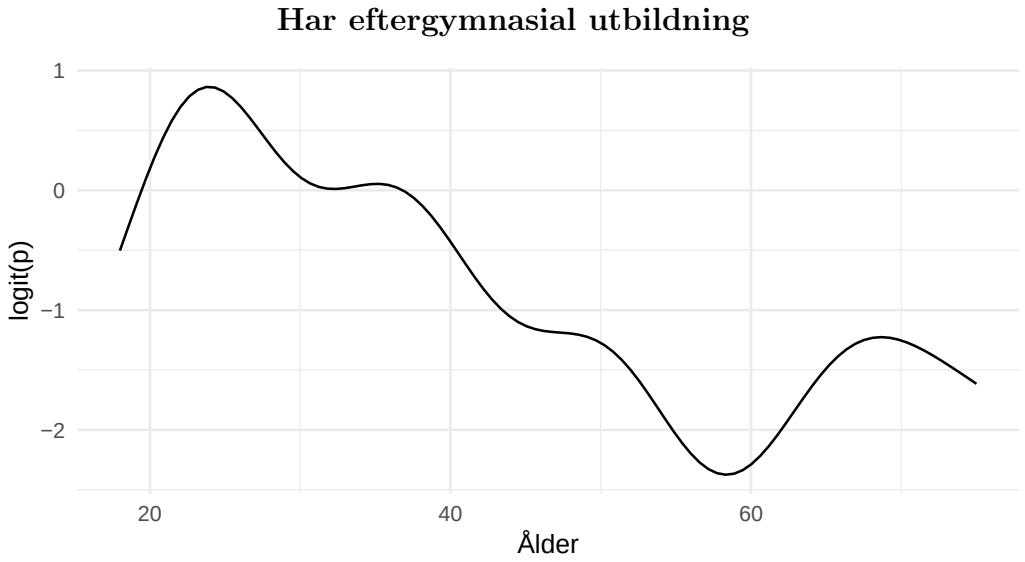
Figur 24: Anpassning av kubisk splinefunktion till åldern (a) då effekten av kategorivariablerna modellerats med logistisk regression



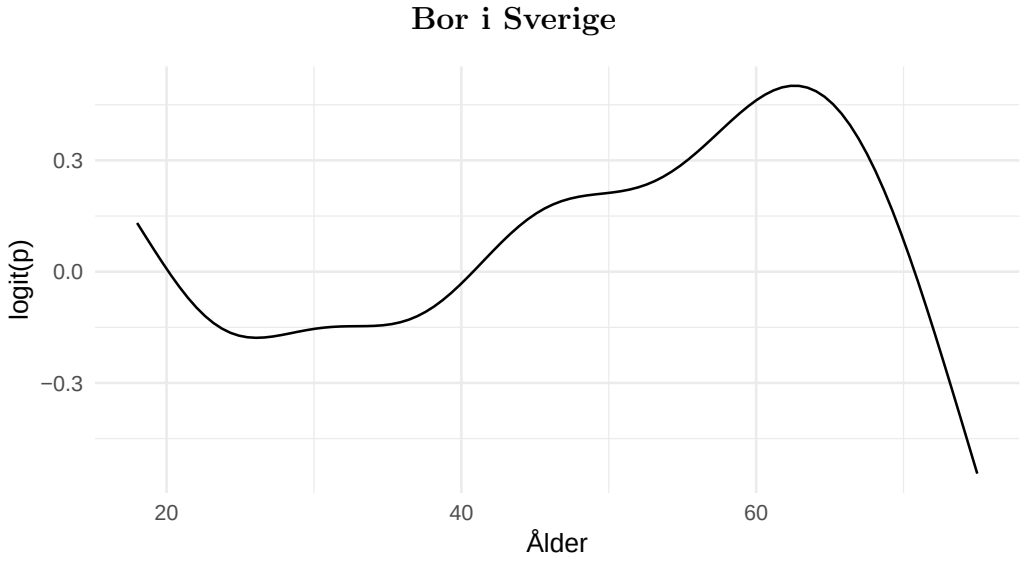
Figur 25: Anpassning av kubisk splinefunktion till åldern (a) då effekten av kategorivariablerna modellerats med logistisk regression



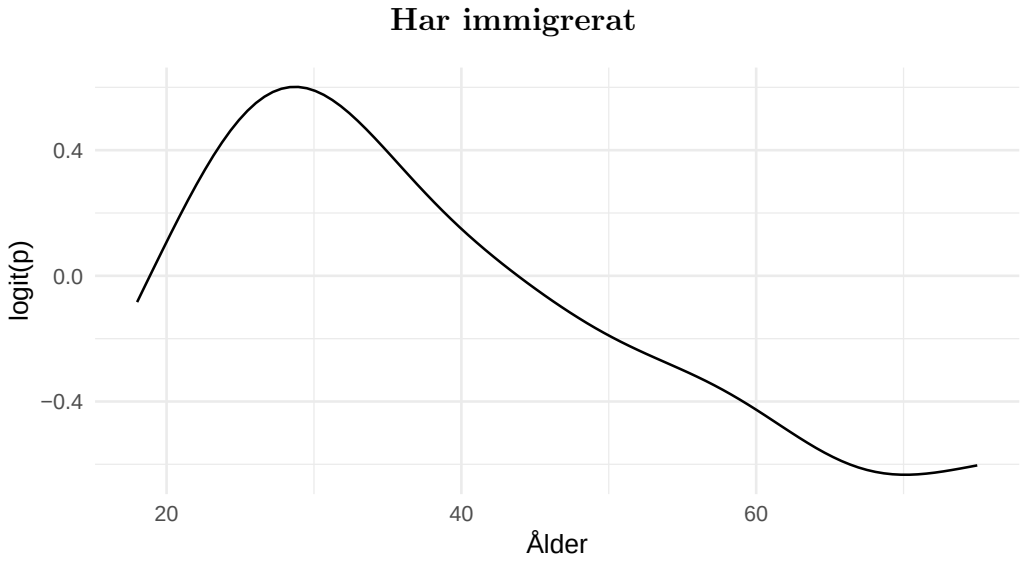
Figur 26: Anpassning av kubisk splinefunktion till åldern (a) då effekten av kategorivariablerna modellerats med logistisk regression



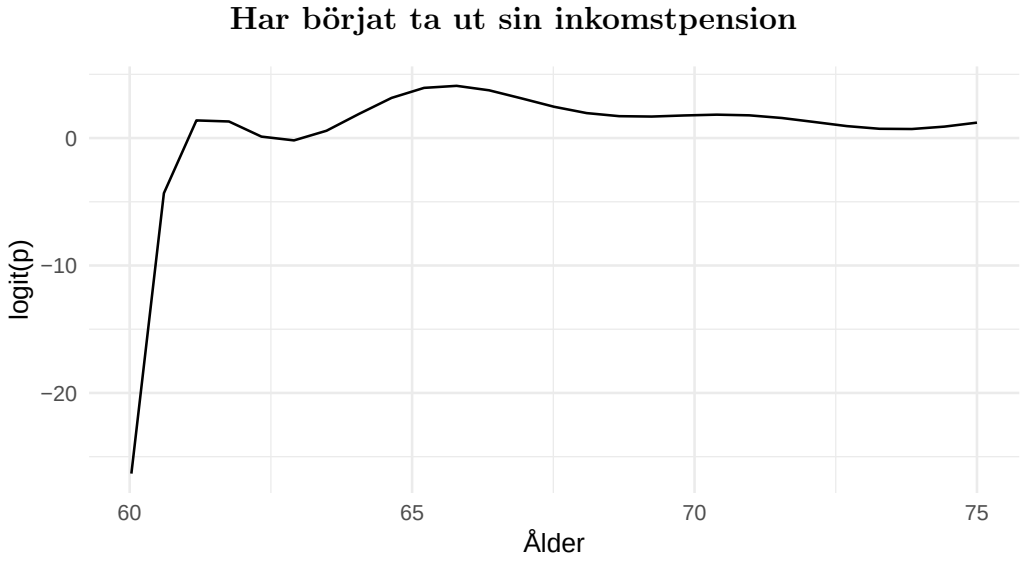
Figur 27: Anpassning av kubisk splinefunktion till åldern (a) då effekten av kategorivariablerna modellerats med logistisk regression



Figur 28: Anpassning av kubisk splinefunktion till åldern (a) då effekten av kategorivariablerna modellerats med logistisk regression

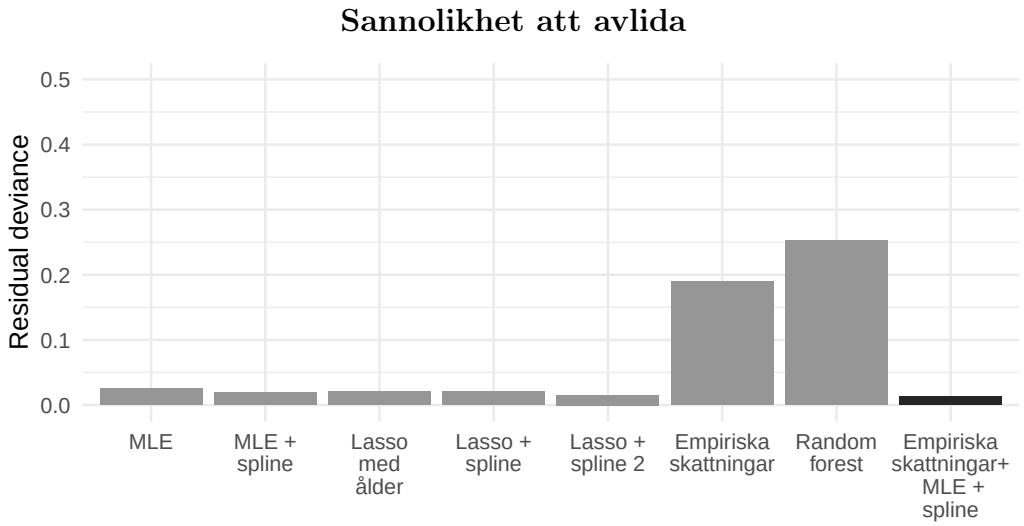


Figur 29: Anpassning av kubisk splinefunktion till åldern (a) då effekten av kategorivariablerna modellerats med logistisk regression

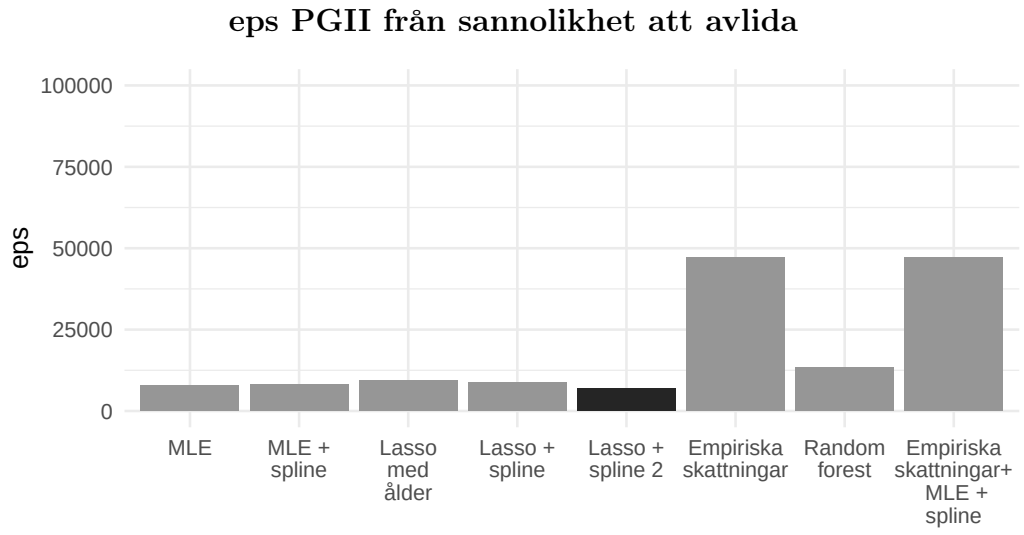


Figur 30: Anpassning av kubisk splinefunktion till åldern (a) då effekten av kategorivariablerna modellerats med logistisk regression

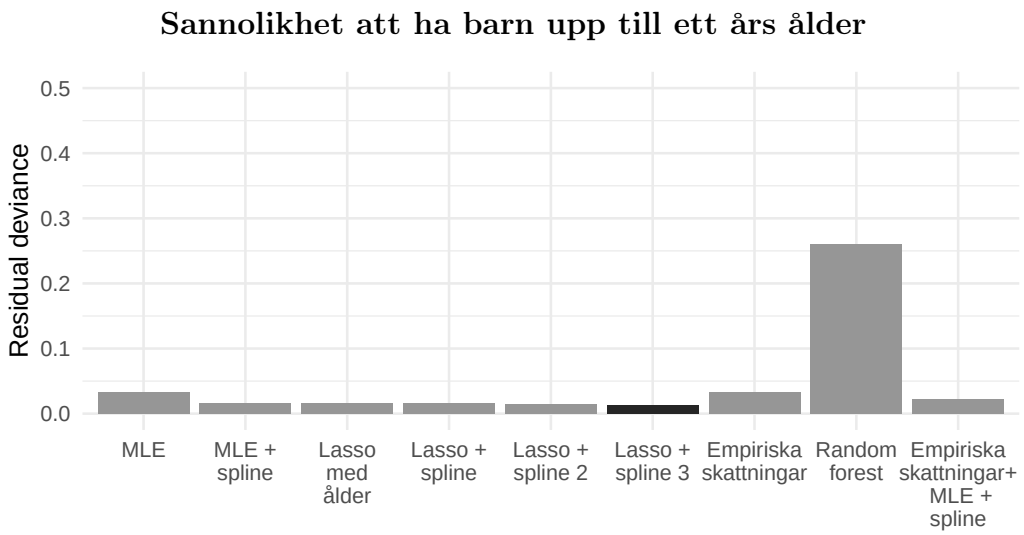
9.1.3 Uppskattning av fel



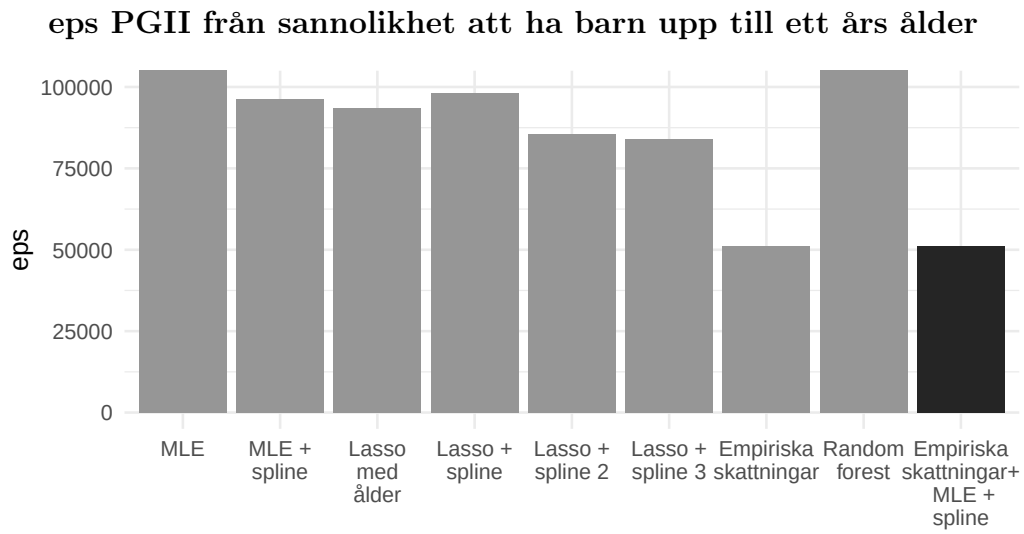
Figur 31:



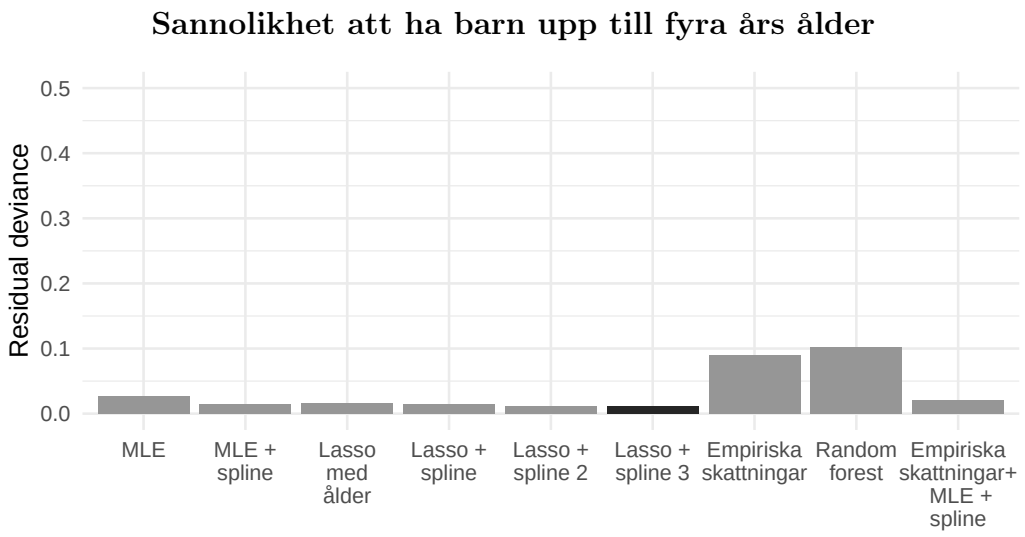
Figur 32:



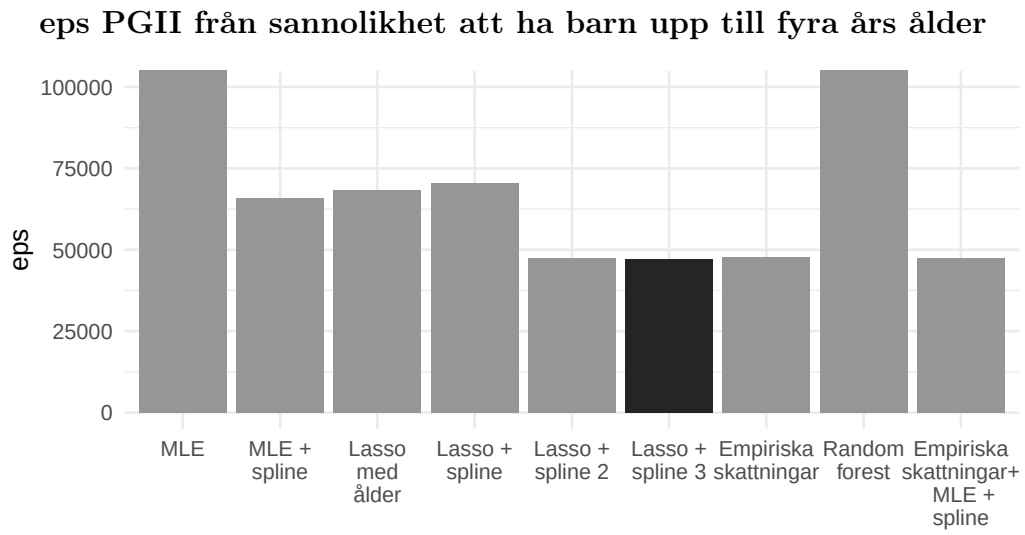
Figur 33:



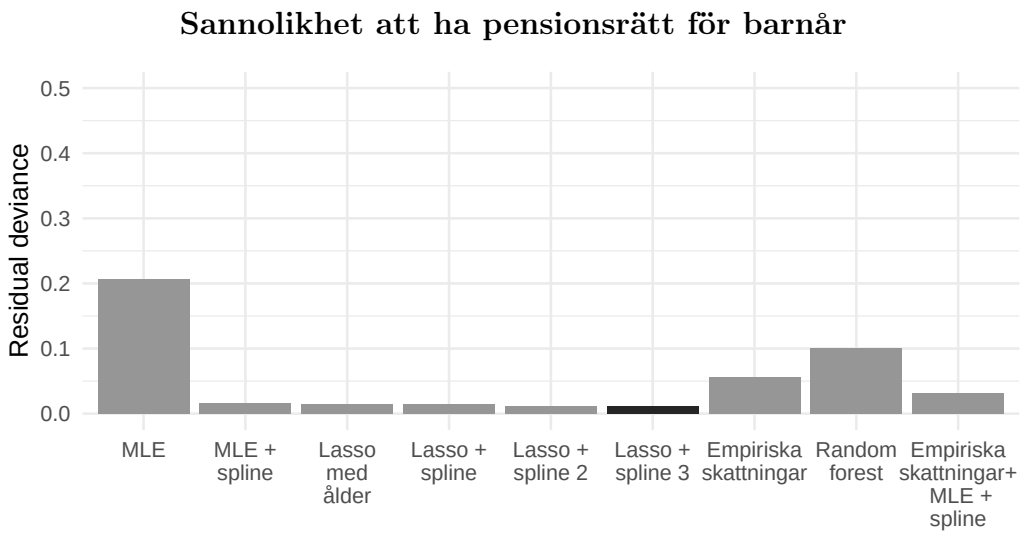
Figur 34:



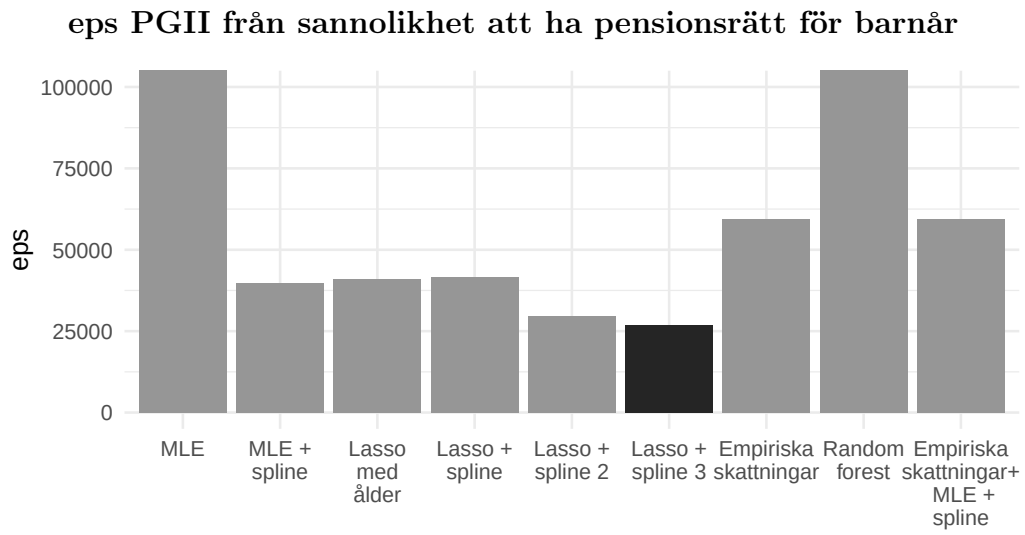
Figur 35:



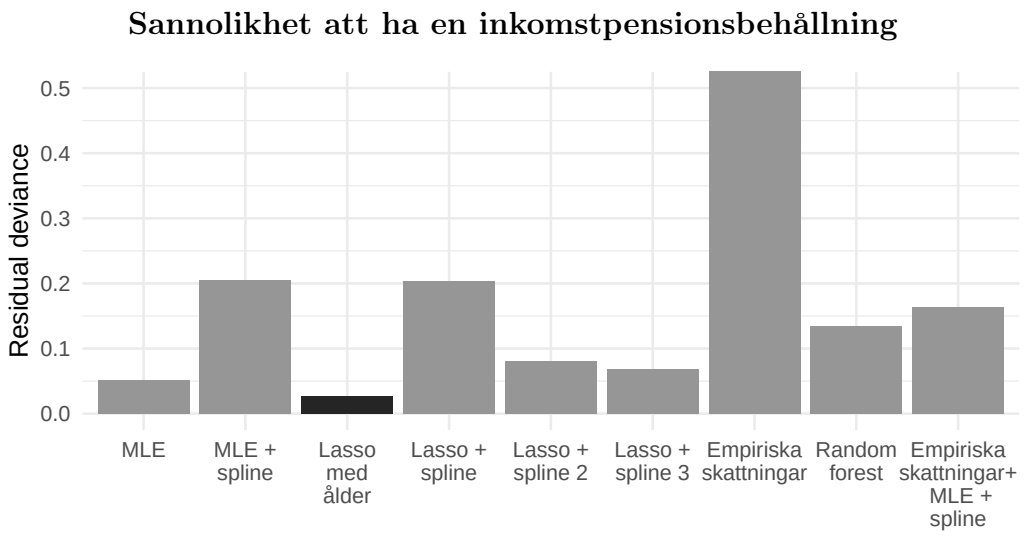
Figur 36:



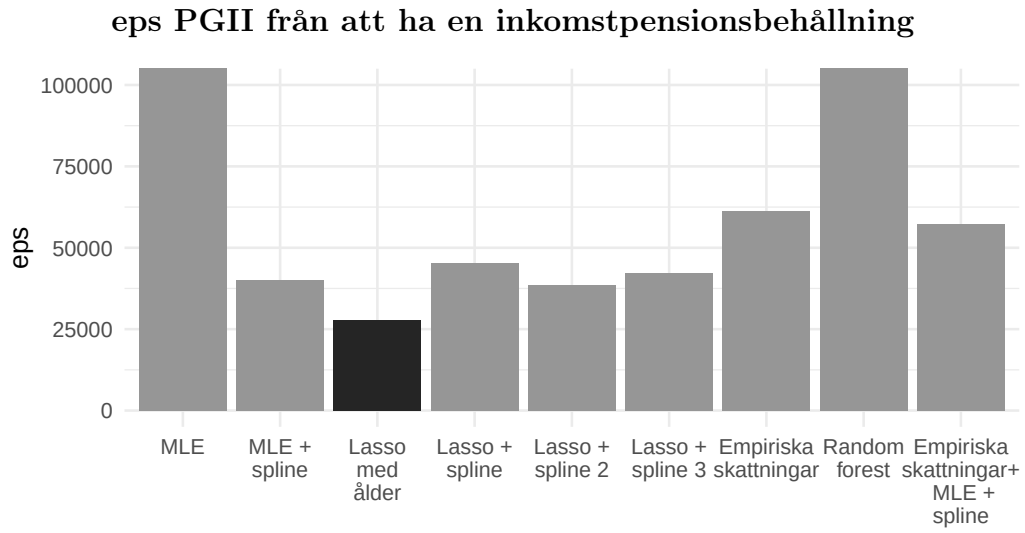
Figur 37:



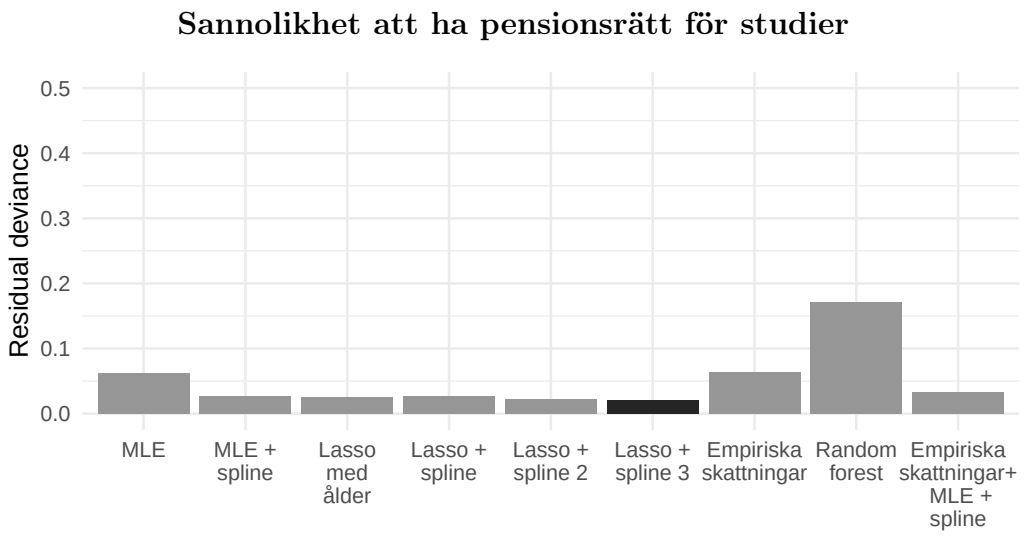
Figur 38:



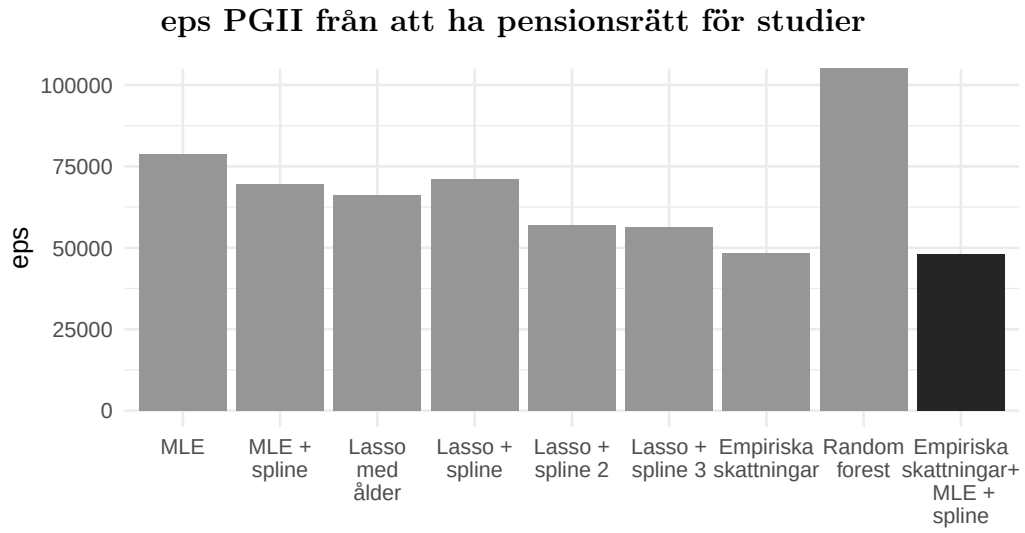
Figur 39:



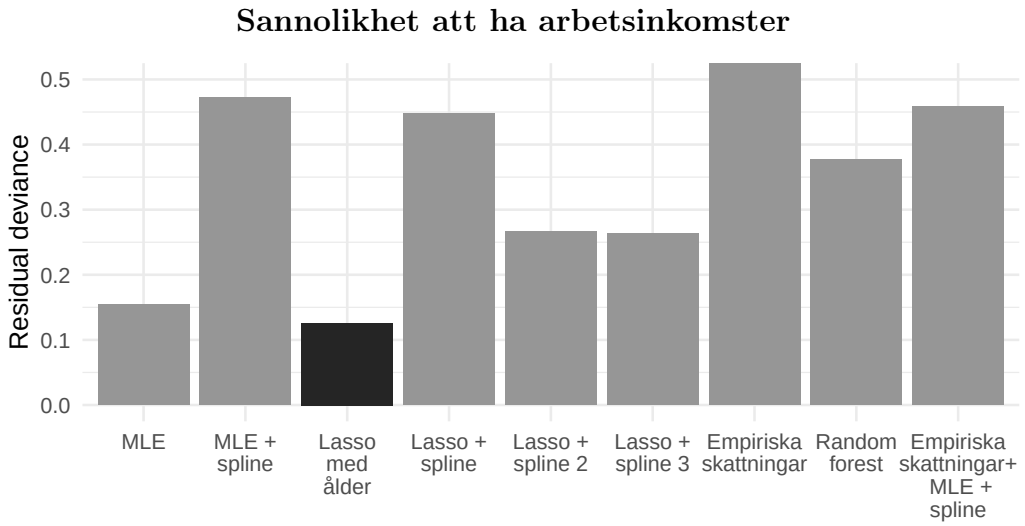
Figur 40:



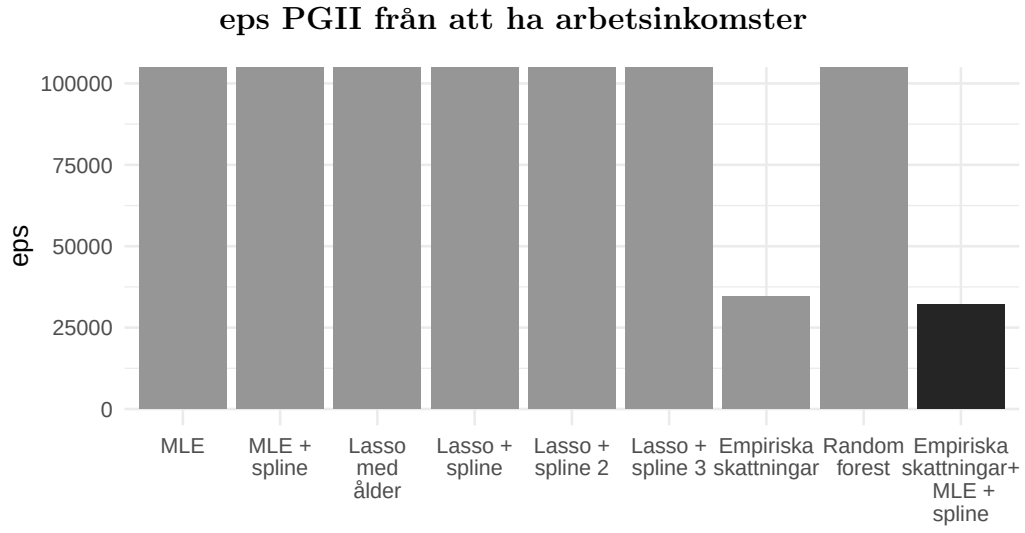
Figur 41:



Figur 42:

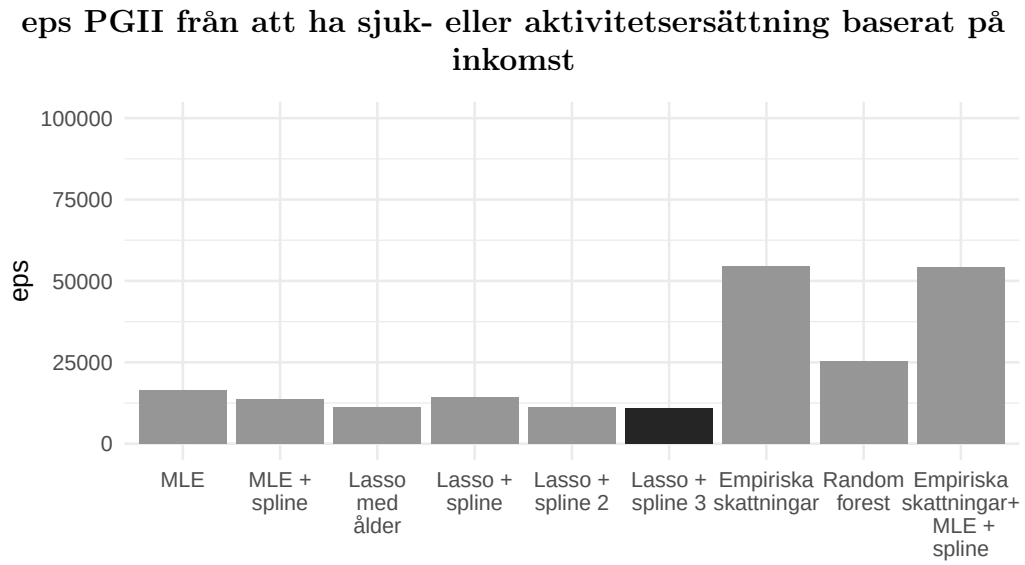


Figur 43:

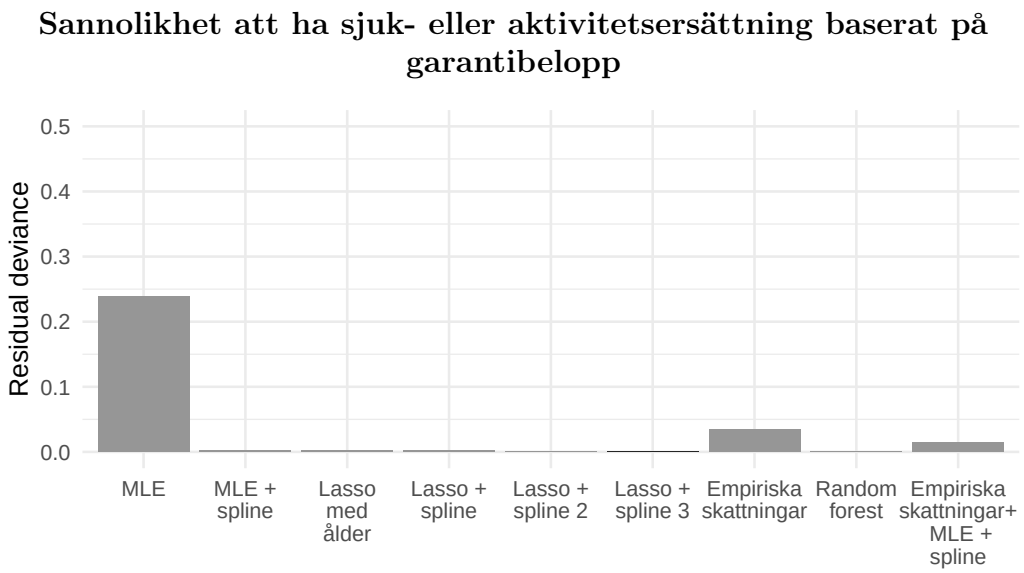


Figur 44:

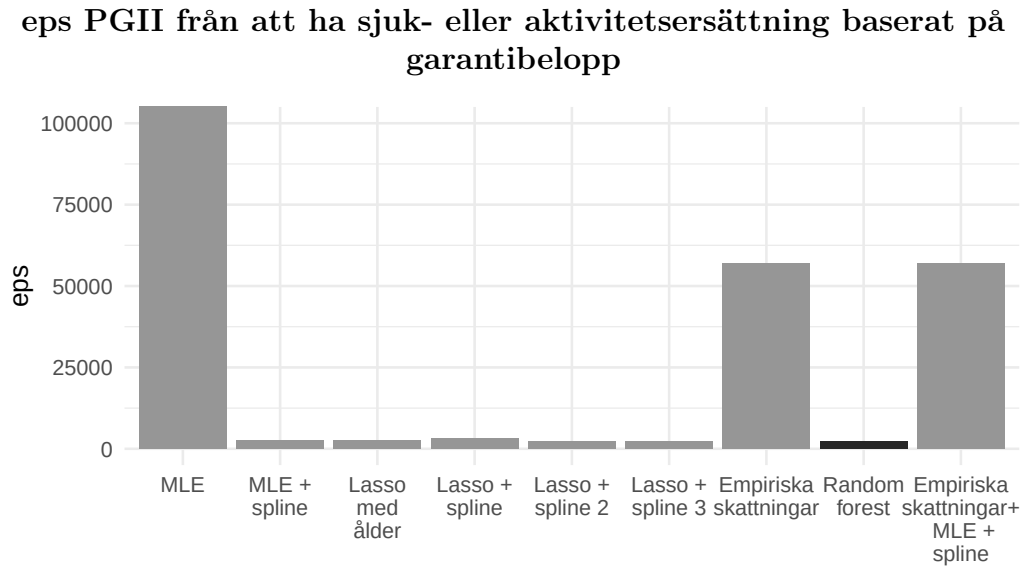
Figur 45:



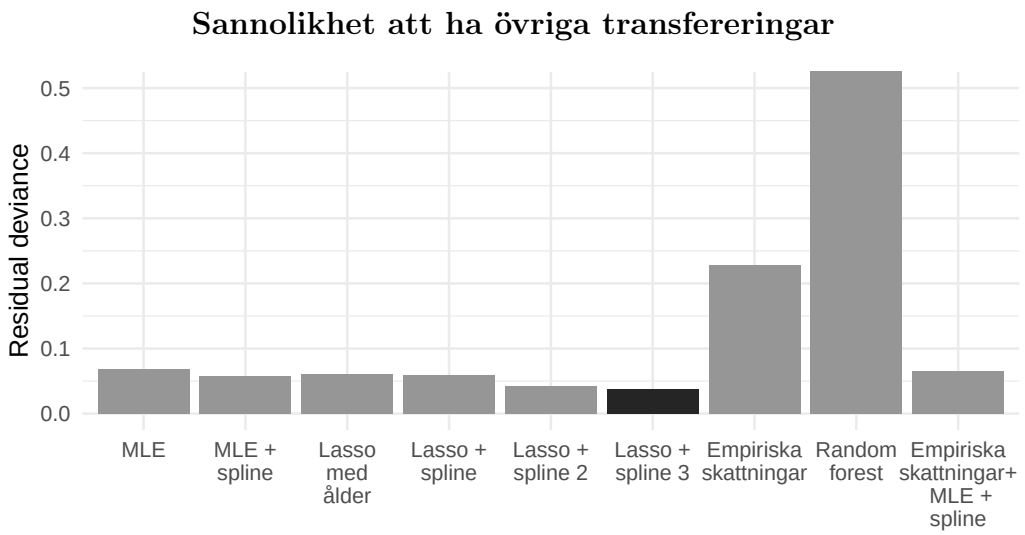
Figur 46:



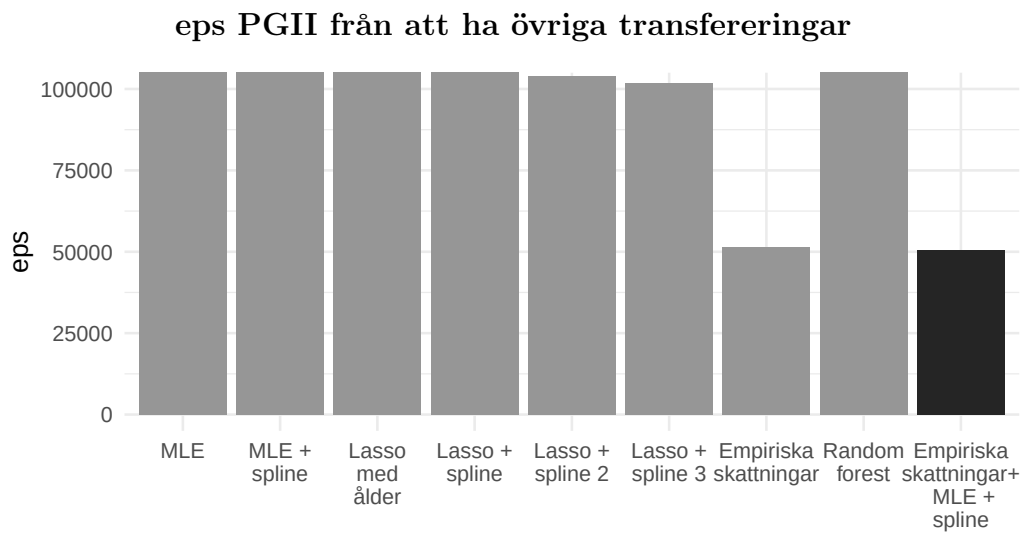
Figur 47:



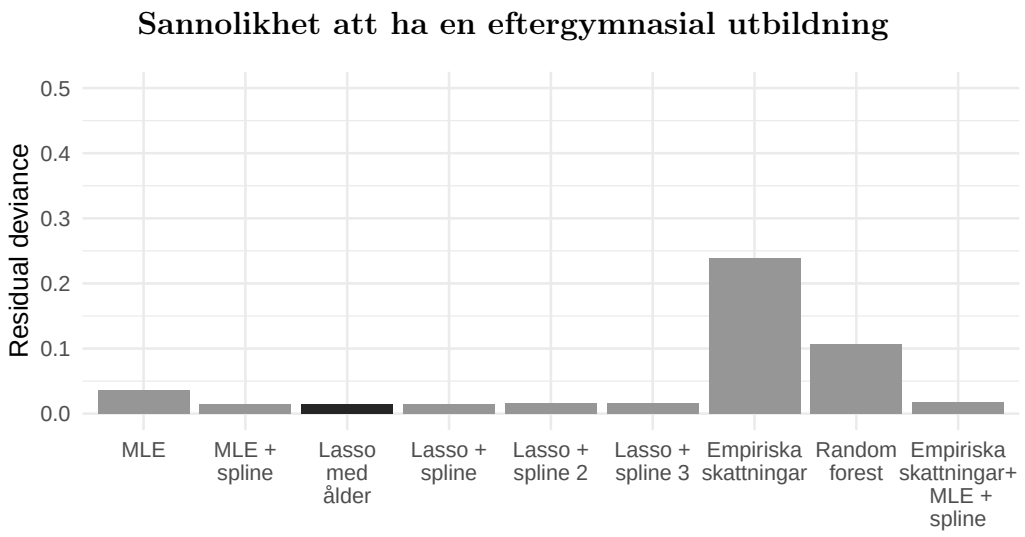
Figur 48:



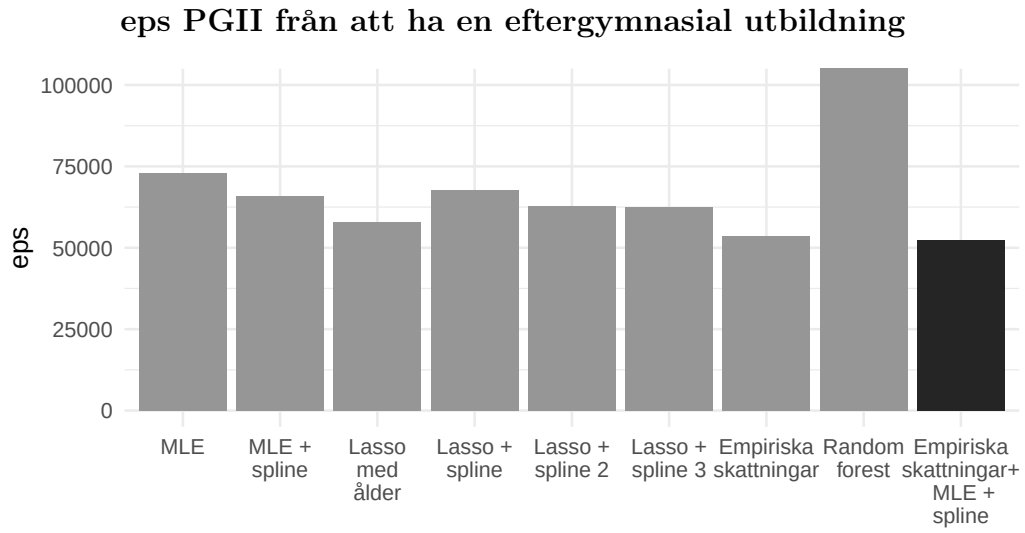
Figur 49:



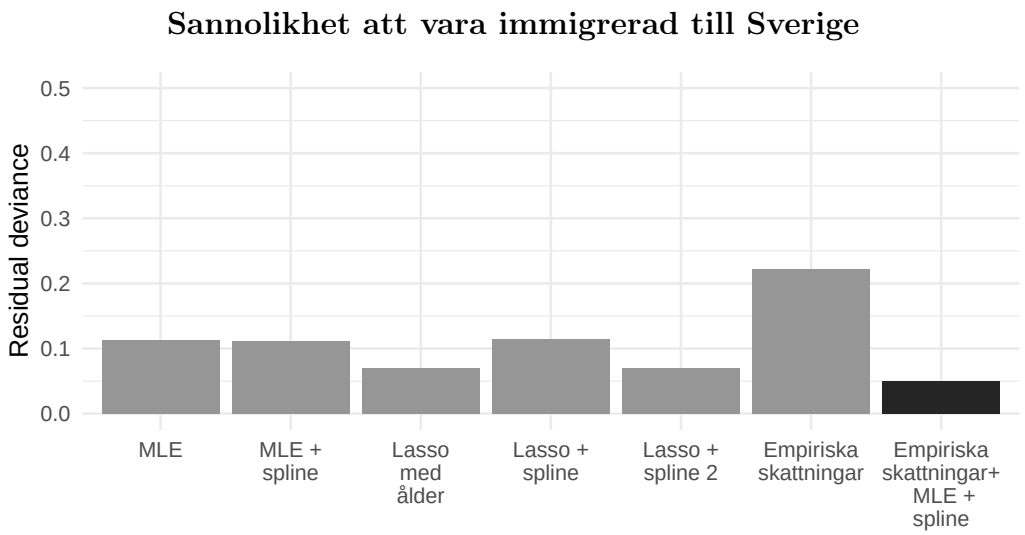
Figur 50:



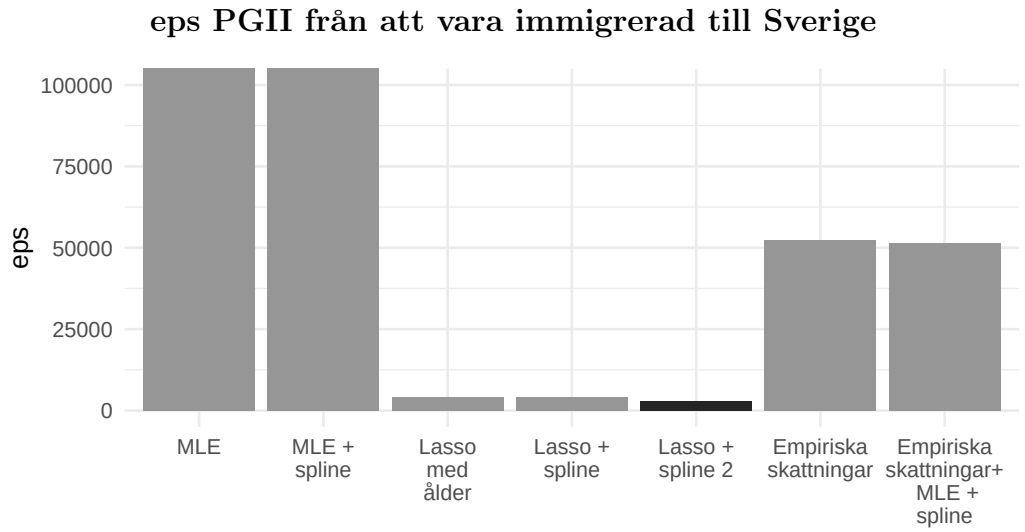
Figur 51:



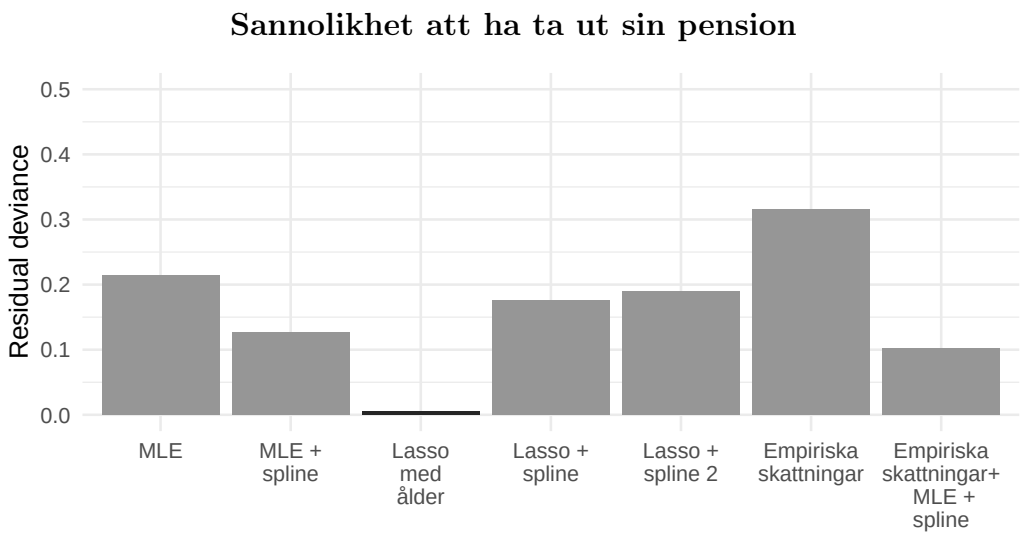
Figur 52:



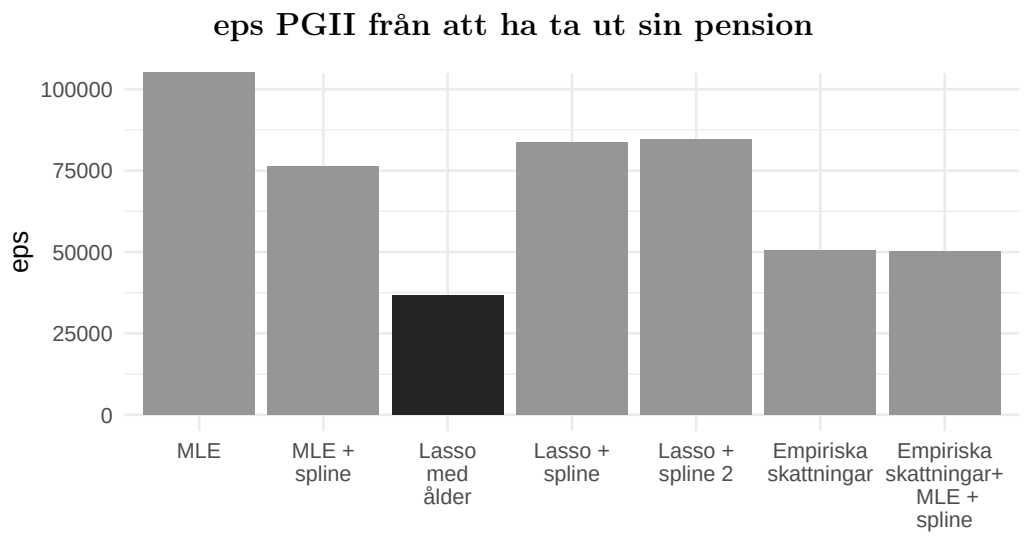
Figur 53:



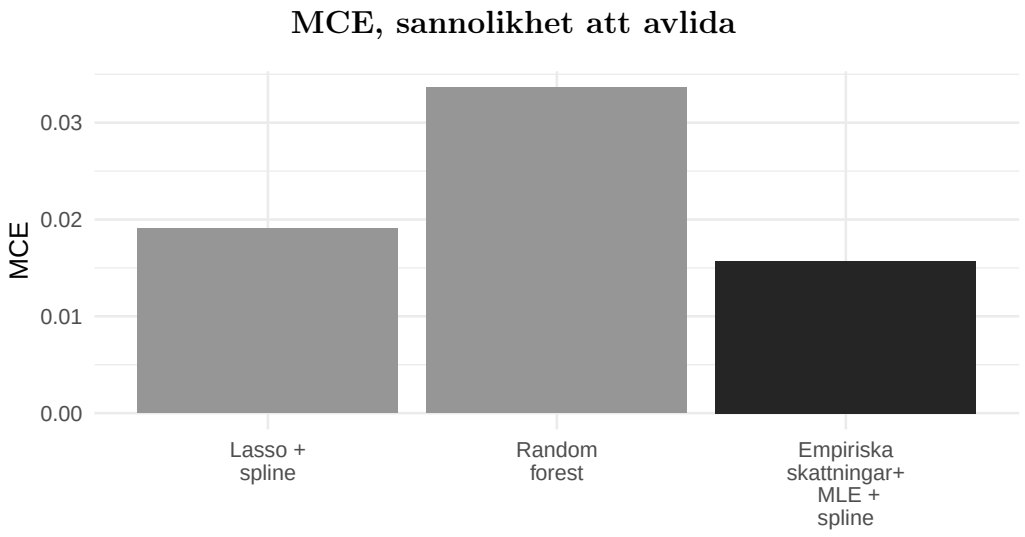
Figur 54:



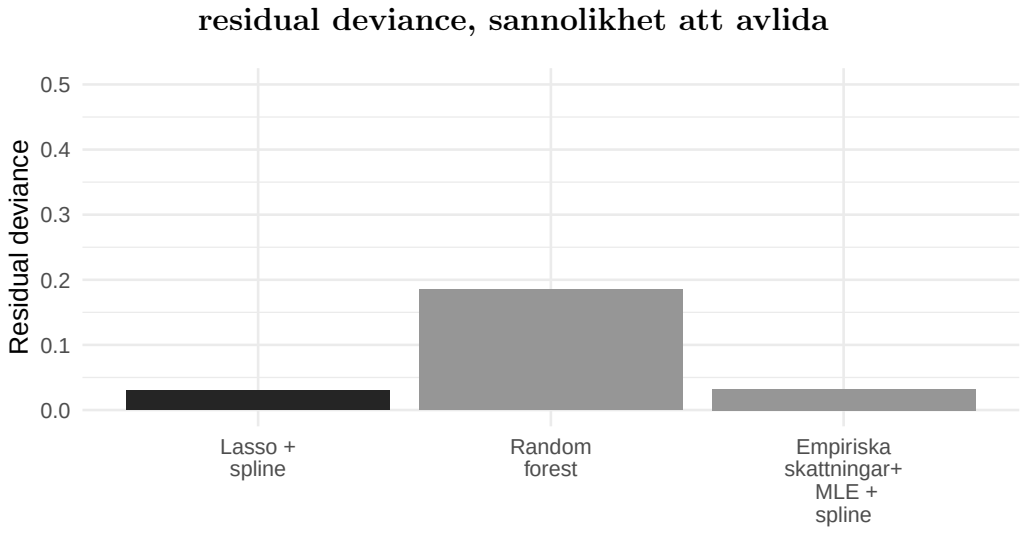
Figur 55: Vi ser att den logistiska modellen där ålder a är parametriserad som en kategorivariabel ger lägst fel. Detta kan bero på att det finns en hård gräns i regelsystemet vid 61 års ålder vilket t.ex. kubiska splinefunktioner har svårare att anpassa sig till.



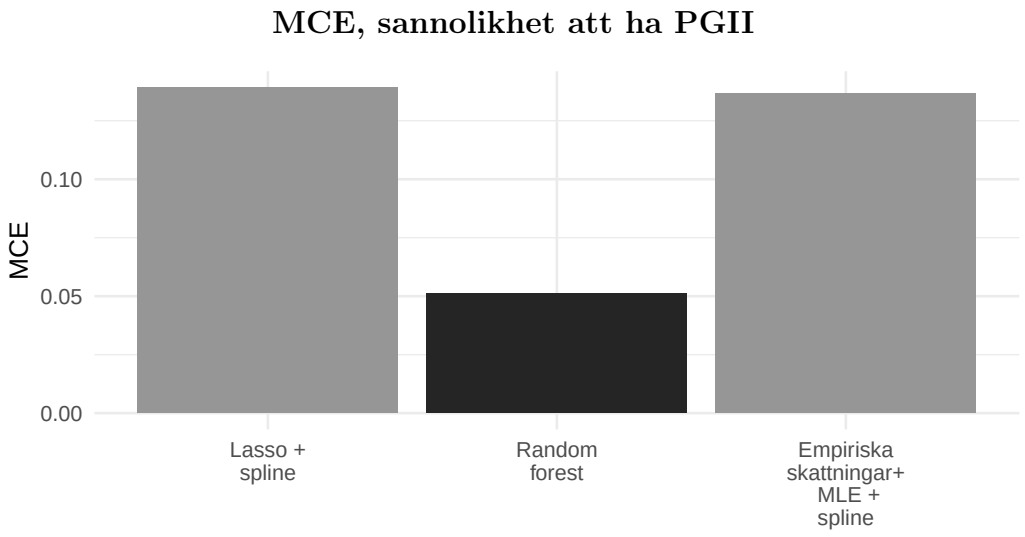
Figur 56: Vi ser att den logistiska modellen där ålder a är parametriserad som en kategorivariabel ger lägst fel. Detta kan bero på att det finns en hård gräns i regelsystemet vid 61 års ålder vilket t.ex. kubiska splinefunktioner har svårare att anpassa sig till.



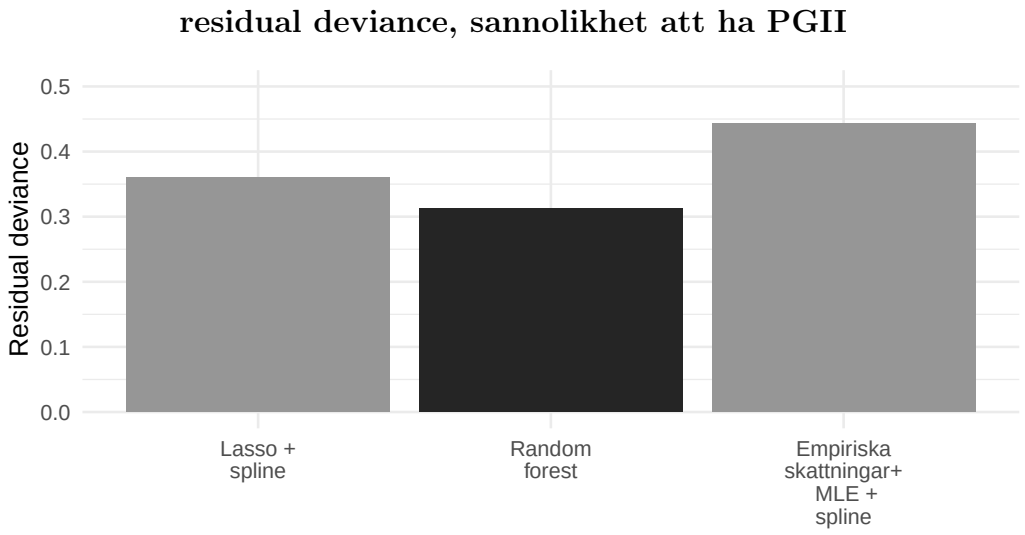
Figur 57: Ett urval på en tiondel av datamängden gjordes för att kunna beräkna Random Forest, då dessa beräkningar kraschade när hela datamängden användes. De andra modellerna i grafen är beräknade på samma urval



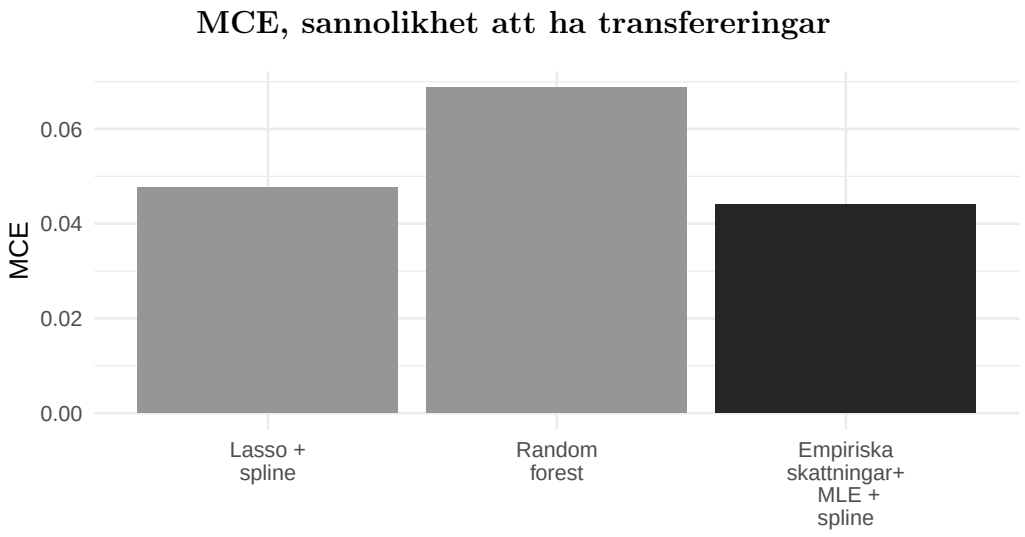
Figur 58: Ett urval på en tiondel av datamängden gjordes för att kunna beräkna Random Forest, då dessa beräkningar kraschade när hela datamängden användes. De andra modellerna i grafen är beräknade på samma urval



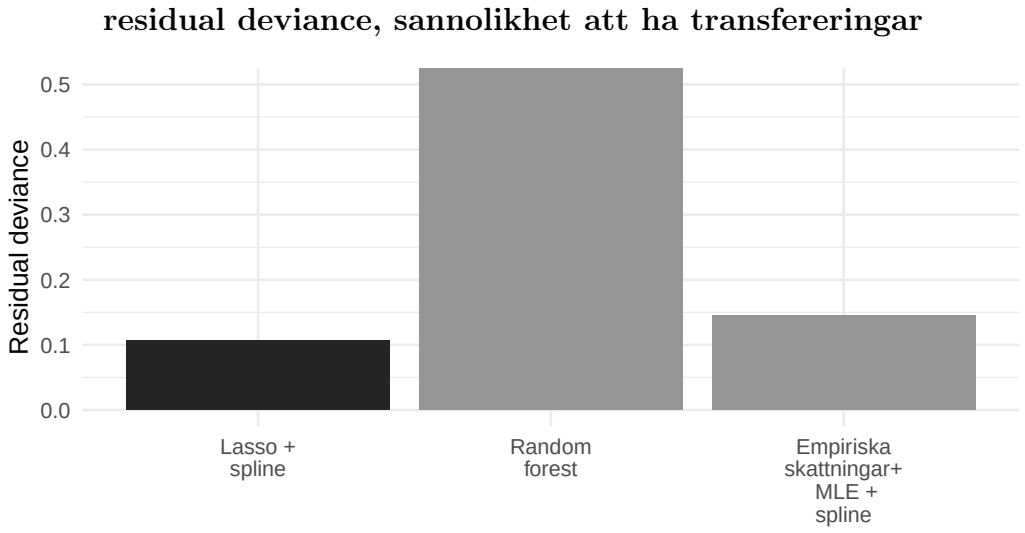
Figur 59: Ett urval på en tiondel av datamängden gjordes för att kunna beräkna Random Forest, då dessa beräkningar kraschade när hela datamängden användes. De andra modellerna i grafen är beräknade på samma urval



Figur 60: Ett urval på en tiondel av datamängden gjordes för att kunna beräkna Random Forest, då dessa beräkningar kraschade när hela datamängden användes. De andra modellerna i grafen är beräknade på samma urval



Figur 61: Ett urval på en tiondel av datamängden gjordes för att kunna beräkna Random Forest, då dessa beräkningar kraschade när hela datamängden användes. De andra modellerna i grafen är beräknade på samma urval



Figur 62: Ett urval på en tiondel av datamängden gjordes för att kunna beräkna Random Forest, då dessa beräkningar kraschade när hela datamängden användes. De andra modellerna i grafen är beräknade på samma urval

9.2 Omskrivning av parameterisering logit/multinomial

$$P(X_1 = x_1, X_2 = x_2) = P(X_1 = x_1) \cdot P(X_2 = x_2 | X_1 = x_1) \tag{17}$$

Detta kan visas vara en multinomial fördelning med fyra nivåer. En parameterisering av en multinomial fördelning har följande utseende:

$$\log\Big(\frac{\pi_j}{\pi_J}\Big) = \alpha_j + \sum_i \beta_{ij} x_i \tag{18}$$

Om man sätter parametrarna för modellen för X_1 och $X_2|X_1$ i högerledet i ekvation (17) får man:

$$P(X_1 = x_1) \cdot P(X_2 = x_2 | X_1 = x_1) = \frac{\exp(\alpha_1 x_1)}{1 + \exp(\alpha_1)} \cdot \frac{\exp(\alpha_2 x_2 + \beta x_1 x_2)}{1 + \exp(\alpha_2 + \beta x_1)} =$$

$$\frac{\exp(\alpha_1 x_1 + \alpha_2 x_2 + \beta x_1 x_2)}{1 + \exp(\alpha_1) + \exp(\alpha_2 + \beta x_1) + \exp(\alpha_1 + \alpha_2 + \beta x_1)}$$

I det här läget har vi en variabel x_1 i nämnaren, vilket vi måste hantera om vi vill skriva detta uttryck på formen som följer av parameteriseringen i (18). Vi ansätter en ny parameter Δ_{x_1} som logaritmen av kvoten mellan de två möjliga nämnarna:

$$\Delta_{x_1} := \log\Big(\frac{1 + \exp(\alpha_1) + \exp(\alpha_2 + \beta) + \exp(\alpha_1 + \alpha_2 + \beta)}{1 + \exp(\alpha_1) + \exp(\alpha_2) + \exp(\alpha_1 + \alpha_2)}\Big)$$

Detta gör att vi kan skriva

$$\frac{\exp(\alpha_1 x_1 + \alpha_2 x_2 + \beta x_1 x_2)}{1 + \exp(\alpha_1) + \exp(\alpha_2 + \beta x_1) + \exp(\alpha_1 + \alpha_2 + \beta x_1)} =$$

$$\frac{\exp(\alpha_1 x_1 + \alpha_2 x_2 + \beta x_1 x_2)/\exp(\Delta_{x_1} x_1)}{(1 + \exp(\alpha_1) + \exp(\alpha_2 + \beta x_1) + \exp(\alpha_1 + \alpha_2 + \beta x_1))/\exp(\Delta_{x_1} x_1)} =$$

$$\frac{\exp(\alpha_1 x_1 + \alpha_2 x_2 + \beta x_1 x_2)/\exp(\Delta_{x_1} x_1)}{1 + \exp(\alpha_1) + \exp(\alpha_2) + \exp(\alpha_1 + \alpha_2)} =$$

$$\frac{\exp(\alpha_1 x_1 + \alpha_2 x_2 + \beta x_1 x_2 - \Delta_{x_1} x_1)}{1 + \exp(\alpha_1) + \exp(\alpha_2) + \exp(\alpha_1 + \alpha_2)} =$$

$$\frac{\exp((\alpha_1 - \Delta_{x_1}) x_1 + \alpha_2 x_2 + \beta x_1 x_2)}{1 + \exp(\alpha_1) + \exp(\alpha_2) + \exp(\alpha_1 + \alpha_2)}$$

Vi ansätter en ny parameter $\tilde{\alpha}_1 := \alpha_1 - \Delta_{x_1}$ och kan då skriva:

$$P(X_1 = x_1, X_2 = x_2) = \frac{\exp(\tilde{\alpha}_1 x_1 + \alpha_2 x_2 + \beta x_1 x_2)}{1 + \exp(\alpha_1) + \exp(\alpha_2) + \exp(\alpha_1 + \alpha_2)}$$

Vidare, eftersom vi söker en form så att:

$$\sum_{x_{1i}} \sum_{x_{2j}} P(X_1 = x_{1i}, X_2 = x_{2j}) = 1$$

kan vi alltså skriva:

$$P(X_1 = x_1, X_2 = x_2) = \frac{\exp(\tilde{\alpha}_1 x_1 + \alpha_2 x_2 + \beta x_1 x_2)}{\sum_{x_{1i}} \sum_{x_{2j}} \exp(\tilde{\alpha}_1 x_{1i} + \alpha_2 x_{2j} + \beta x_{1i} x_{2j})}$$

Om vi introducerar ytterligare en variabel X_3 och söker den simultana sannolikheten för:

$$P(X_1 = x_1, X_2 = x_2, X_3 = x_3) =$$

$$P(X_1 = x_1) \cdot P(X_2 = x_2 | X_1 = x_1) \cdot P(X_3 = x_3 | X_2 = x_2, X_1 = x_1)$$

och vi använder ordinär logistisk regression för sannolikheten $P(X_3 = x_3 | X_2 = x_2, X_1 = x_1)$ får vi:

$$P(X_1 = x_1, X_2 = x_2) \cdot P(X_3 = x_3 | X_2 = x_2, X_1 = x_1) =$$

$$\frac{\exp(\tilde{\alpha}_1 x_1 + \alpha_2 x_2 + \beta x_1 x_2)}{\sum_{x_{1i}} \sum_{x_{2j}} \exp(\tilde{\alpha}_1 x_{1i} + \alpha_2 x_{2j} + \beta x_{1i} x_{2j})} \cdot \frac{\exp(\alpha_3 x_3 + \beta_{31} x_3 x_1 + \beta_{32} x_3 x_2)}{1 + \exp(\alpha_3 + \beta_{31} x_1 + \beta_{32} x_2)}$$

$$\frac{\exp(\tilde{\alpha}_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \beta_{21} x_1 + \beta_{31} x_3 x_1 + \beta_{32} x_3 x_2)}{h(x_1, x_2)}$$

Där

$$h(x_1, x_2) := \Big(\sum_{x_{1i}} \sum_{x_{2j}} \exp(\tilde{\alpha}_1 x_{1i} + \alpha_2 x_{2j} + \beta x_{1i} x_{2j}) \Big) \cdot (1 + \exp(\alpha_3 + \beta_{31} x_1 + \beta_{32} x_2))$$

På analogt sätt som tidigare:

$$\Delta_{x_1}^{x_3} := \log\Big(\frac{h(1, x_2)}{h(0, x_2)}\Big)$$

$$\Delta_{x_2}^{x_3} := \log\Big(\frac{h(0, 1)}{h(0, 0)}\Big)$$

Då kan vi återigen omparametrisera intercepten och skriva:

$$\frac{\exp(\hat{\alpha}_1 x_1 + \tilde{\alpha}_2 x_2 + \alpha_3 x_3 + \beta_2 x_1 x_2 + \beta_{31} x_1 x_3 + \beta_{32} x_2 x_3)}{\sum_{x_{1i}} \sum_{x_{2j}} \sum_{x_{3k}} \exp(\hat{\alpha}_1 x_{1i} + \tilde{\alpha}_2 x_{2j} + \alpha_3 x_3 + \beta_2 x_{1i} x_{2j} + \beta_{31} x_{1i} x_{3k} + \beta_{32} x_{2j} x_{3k})}$$

Där

$$\hat{\alpha}_1 := \tilde{\alpha}_1 - \Delta_{x_1}^{x_3}$$

och

$$\tilde{\alpha}_2 := \alpha_2 - \Delta_{x_2}^{x_3}$$

Vi får alltså paramterisera om $\tilde{a}_1$ ännu en gång. Slutsatsen är alltså att om vi har en kedja av logitmodeller som inte innehåller några interaktionstermer, är β-parametrarna i någon mening absoluta, d.v.s. det spelar ingen roll i vilken ordning som kedjan går i, β-parametrarna kommer ändå bli desamma. Däremot kommer intercepten i respektive logit-modell påverkas av i vilken ordning man gör betingningarna.

Referenser

[1] Esbjörn Ohlsson, Björn Johansson, *Non-Life Insurance Pricing with Generalized Linear Models* , Springer, 2010.

[2] Robert Tibshirani, *Regression shrinkage and selection via the lasso*, Journal of the Royal Statistical Society. Series B (Methodological) Vol. 58, No. 1 (1996), pp. 267-288.

[3] Trevor Hastie, Robert Tibshirani, Jerome Friedman *The Elements of Statistical Learning*, Springer Series in Statistics, 2008

[4] Pensionsmyndigheten, *Orange Rapport 2018*, 2019.

[5] Pensionsmyndigheten, *Pensionsmodellen, användarmanual*, 2015.

[6] Pensionsmodellens hemsida, `https://www.pensionsmyndigheten.se/statistik-och-rapporter/pensionsmodellen/pensionsmodellen`