Stockholms universitet

Vad är risken att dö i en trafikolycka? - En studie över hur kön och ålder hos en personbilsförare påverkar utfallet av en trafikolycka

Amanda Wiman

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2015:3
Matematisk statistik
Juni 2015

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholms universitet

Matematisk statistik
Stockholms universitet
Kandidatuppsats **2015:3**
http://www.math.su.se/matstat

Vad är risken att dö i en trafikolycka? - En studie över hur kön och ålder hos en personbilsförare påverkar utfallet av en trafikolycka

Amanda Wiman*

Juni 2015

Sammanfattning

Denna studie använder sig av trafikdata för åren 1999-2009 över personbilsförares trafikolyckor i Sverige. Med hjälp av en logistisk regressionsmodell ämnas att skatta en kurva som kan förklara sambandet mellan kön och ålder hos personbilsföraren med risken för dödligt utfall i en trafikolycka. För att skatta detta samband används B-splines, ett sätt att kombinera flera polynomfunktioner till en funktion. I studien testas olika antal polynomfunktioner och sedan väljs det antal som ger lägst AIC-värde för vår B-spline. Studien visar att de allra yngsta personbilsförarna är de som löper störst risk att omkomma om de är med i en trafikolycka. Risken är lägst för personer i 40- till 50-årsåldern och sedan ökar risken långsamt för högre åldrar. En annan slutsats är att män som är med i trafikolyckor i allmänhet löper större risk att omkomma än kvinnor som är med i trafikolyckor. Analysen visar även att det finns en viss osäkerhet att dra slutsatser om de allra yngsta kvinnornas påverkan på utfallet av en trafikolycka på grund av få observationer.

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige.
E-post: amandawiman12@gmail.com. Handledare: Martin Sköld och Tom Britton.

Förord

Detta arbete utgör mitt examensarbete i matematisk statistik om 15 högskolepoäng vid Stockholms Universitet. Jag vill ge ett stort tack till mina handledare Martin Sköld och Tom Britton för all hjälp och råd de gett under tidens gång. Jag vill även tacka studiekamraterna Sanna Kronman, Caroline Jernström, Martina Sandberg, Andrea Klemming och Sandra Brännstam för värdefull hjälp och feedback, samt Viktor Rutberg för motivation och stöd.

Innehåll

1 Inledning

Varje år dör flera hundra svenskar i trafikolyckor (Trafikanalys, 1999-2009). Flera studier har gjorts där man undersökt hur förarens kön och ålder inverkar på dödsfall vid trafikolyckor, bland annat en publicerad av Journal of Safety Research (Williams & Shabanova, 2003) och en av Accident Analysis & Prevention (Massie *et al.*, 1994), båda i USA.

I studien publicerad av Journal of Safety Research tittade man på trafikolyckor med dödligt utfall under en femårsperiod (1996-2000) där man avgjort vem av de inblandade som orsakade olyckan. Med det datamaterialet undersökte man vilket kön och vilken åldersgrupp som var ansvarig för flest dödsfall i trafiken. En av slutsatserna var att det är de yngsta och äldsta förarna som svarar för flest antal trafikolyckor med dödligt utfall. Man kommer även fram till att unga män orsakar fler trafikolyckor än unga kvinnor, samtidigt som kvinnor i 50-årsåldern orsakar fler trafikolyckor än män i samma ålder.

I studien publicerad Accident Analysis & Prevention studerades vilka kön och åldrar som oftast är inblandade i trafikolyckor, samt vilka kön och åldrar som oftast är inblandade i trafikolyckor med dödligt utfall. Där kom man fram till att de äldsta förarna är med i flest trafikolyckor med dödligt utfall, medan de yngre förarna har högst inblandning i trafikolyckor i allmänhet. Man kommer även fram till att män är med i fler trafikolyckor med dödligt utfall, medan kvinnor är med i flest trafikolyckor i allmänhet.

Båda dessa studier jämför olika åldersgrupper och kön med varandra. För att kunna göra detta behöver man ett mått på hur stor del av den dagliga trafiken som varje grupp utgör, då den inte är lika stor för alla grupper. Till exempel finns det troligtvis inte lika många kvinnor i 18-årsåldern i trafiken som det finns medelålders män. I studien publicerad av Journal of Safety Research har man använt antalet personer med körkort inom varje grupp som mått. Dock kan det bli ett problem att titta på hur många som har körkort i varje grupp då unga och äldre inte tenderar att köra lika mycket bil som till exempel personer i medelåldern (en grupp där det troligtvis finns fler som använder bilen på daglig basis). Studien från Accident Analysis & Prevention undviker detta genom att använda hur många mil det uppskattas att varje grupp kör, då de anser att mer tid i trafiken bidrar till större risk för att vara med i en trafikolycka. Dock är det väldigt svårt att uppskatta hur mycket bil

en viss grupp kör då detta är individuellt.

Med bakgrund av detta vill vi i denna studie med hjälp av svenskt trafikdata från Trafikanalys, en kunskapsmyndighet för transportpolitiken, undersöka hur ålder och kön på en personbilsförare påverkar utfallet av en trafikolycka. Med detta som mål behöver vi inte hitta ett mått på hur mycket varje grupp befinner sig i trafiken. Vi vill alltså undersöka givet att en personbilsförare med en viss ålder och av ett visst kön är inblandad i en trafikolycka, hur stor är risken att trafikolyckan har dödligt utfall? Och i och med detta kanske komma närmare att besvara frågan, vad är risken dö om du som förare är med i en trafikolycka?

I denna studie börjar vi med att gå igenom den nödvändiga teorin för att förstå och för att kunna utföra modelleringen som krävs för att komma fram till resultatet i denna studie. Vi beskriver sedan datamaterialet vi har till vårt förfogande och efter det går vi över till själva modelleringen. Där kommer vi först att utföra en analys där vi undersöker hur åldern påverkar utfallet av en trafikolycka. Därefter delar vi upp datamaterialet efter kön för att kunna se hur personbilsförarens kön påverkar utfallet av en trafikolycka. I denna del presenteras även resultatet av denna studie och vi kommer även att skapa konfidensintervall för att undersöka säkerheten i våra resultat. Sedan kommer diskussionen där vi försöker bryta ner resultatet och diskutera varför det blev som det blev.

2 Teori

Teorin om generaliserade linjära modeller, oddskvoter samt logistisk regression som följer nedan kommer från (Agresti, 2002), för djupare förståelse om dessa ämnen hänvisas att läsa kapitel 2, 4 och 5 i denna bok. Definitionen av splines och B-splines är tagen från (Maindonald & Braun, 2010), (Racine, 2014) samt (Ohlsson & Johansson, 2010). För mer om B-splines och dess egenskaper hänvisas att läsa Appendix B.2 i (Ohlsson & Johansson, 2010).

2.1 GLM

För att kunna använda multipel linjär regression krävs att responsvariabeln är normalfördelad, om detta inte är uppfyllt kan istället generaliserade linjära modeller (GLM) användas. En generaliserad linjär modell består av tre komponenter: en slumpmässig komponent, en systematisk komponent samt en länkfunktion.

2.1.1 Slumpmässig komponent

En slumpmässig komponent i den generaliserade linjära modellen består av en responsvariabel med oberoende observationer vars fördelning tillhör den naturliga exponentialfamiljen. En fördelning tillhör den naturliga exponentialfamiljen om dess sannolikhetsfunktion kan skrivas på formen

$$f(y_i;\theta_i) = a(\theta_i)b(y_i)exp\{y_i\mathcal{Q}(\theta_i)\}$$

där $\mathcal{Q}(\theta_i)$ kallas den naturliga parametern.

2.1.2 Systematisk komponent

Den systematiska komponenten i en GLM ger en linjär koppling mellan den slumpmässiga komponenten och de förklarande variablerna. Om vi kallar parametrarna i modellen för β_i och har N observationer kan vi skriva vektorn $(\eta_1, ..., \eta_N)$ som

$$\eta_i = \sum_j \beta_j x_{ij}, \qquad i = 1, ...N.$$

Denna linjärkombination av de förklarande variablerna kallas för den linjära skattaren.

2.1.3 Länkfunktionen

Den systematiska faktorn visar alltså på att vi har en linjär funktion och vi behöver nu koppla ihop detta med vår slumpmässiga faktor, responsvariabeln. Länkfunktionen kopplar ihop den slumpmässiga och den systematiska

komponenten. Om vi låter $\mu_i = E[Y_i]$ där Y är responsvariabeln, så kommer $\eta_i = g(\mu_i)$, där g är länkfunktionen som är en monoton differentierbar funktion. Genom detta samband får vi formeln

$$g(\mu_i) = \sum_j \beta_j x_{ij}, \qquad i = 1, ...N.$$

Den länkfunktion som transformerar medelvärdet till den naturliga parametern kallas för den kanoniska länkfunktionen.

I en GLM skattas parametrarna med maximum likelihood-metoden, detta beror på att minsta kvadratmetod-skattningar bygger på en normalfördelning med konstant varians. Då detta inte är uppfyllt av en GLM är inte minsta kvadratmetoden längre optimal.

Teorin bakom GLM och dess tre komponenter kommer ifrån (Agresti, 2002) s.116-120.

2.2 Oddskvot

Oddset definieras som

$$\Omega = \frac{\pi}{1-\pi},$$

där π är sannolikheten för ett lyckat försök. Vi vet att Ω kommer ges av ett icke-negativt tal eftersom sannolikheten, π, ligger mellan noll och ett. Vi har två fall, då $\Omega > 1$ och då $0 \leq \Omega < 1$. Då Ω är större än ett är det troligare att få ett lyckat försök än ett misslyckat, om Ω ligger mellan noll och ett kommer ett misslyckat försök att vara mer troligt.

Om vi vill jämföra oddsen mellan två olika försök bildar vi en oddskvot som består av oddset för det ena försöket i täljaren och oddset för det andra försöket i nämnaren:

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

Här gäller det på samma sätt att om $\Theta > 1$ kommer oddset för händelse 1 att vara större än oddset för händelse 2 och om $0 \leq \Theta < 1$ kommer oddset för

händelse 1 att vara mindre än oddset för händelse 2. För $\Theta = 1$ vet vi att oddset för ett lyckat utfall hos de båda försöken är lika stort, s.44 (Agresti, 2002).

2.3 Logistisk regression

En logistisk regressionsmodell är en GLM med en binomialfördelad responsvariabel och en logit som länkfunktion, s.123 (Agresti, 2002).

Logistisk regression är lämpligt att använda då responsvariabeln är en kategorivariabel med två utfall, till exempel om responsvariabeln kan utfalla med 'ja' eller 'nej' med en viss sannolikhet p samt $1 - p$. Med logistisk regression kommer vi att få skattade sannolikheter som alltid ligger mellan noll och ett.

Om vi har en binär responsvariabel Y och en förklarande variabel X kan vi skriva $\pi(x) = P(Y = 1|X = x) = 1 - P(Y = 0|X = x)$. Den logistiska regressionsmodellen kommer då att ges av

$$\pi(x) = \frac{exp\{\alpha + \sum_{i=1}^{n} \beta_i x_i\}}{1 + exp\{\alpha + \sum_{1=1}^{n} \beta_i x_i\}}.$$

Vi kan definiera en logitfunktion som log av oddset och skrivs på följande sätt

$$logit(\pi(x)) = log\left(\frac{\pi(x)}{1 - \pi(x)}\right) = \alpha + \sum_{i=1}^{n} \beta_i x_i,$$

som alltså kommer att vara linjär i de förklarande variablerna, s.166 (Agresti, 2002).

2.4 Splines

Då vi använder oss av en generaliserad linjär modell är ofta syftet att hitta den modell som beskriver ett samband bra och alltså skatta en kurva som kan beskriva ens datapunkter. Om man har datapunkter som inte kan beskrivas av en rät linje kan polynom användas, där man alltså sätter in termer av högre ordning i modellen. För ett polynom av grad m kommer till exempel termer $x, x^2, ..., x^m$ krävas. För att skatta vissa kurvor bra kommer dock m kunna bli väldigt stort och för $m > 3$ kommer kurvan att bli svår att jobba

med då denna kommer att ha många upp- och nedgångar. Det är då inte längre lämpligt att jobba med ett polynom. Om detta sker kan man istället använda splines, som är ett sätt att kombinera två eller fler polynomkurvor av lägre ordning.

En spline är en funktion sammanställd av flera basfunktioner, dessa basfunktioner är polynom multiplicerade med indikatorvariabler, punkterna där basfunktionerna möts kallas knopar. Om vi betecknar knoparna med $u_0, u_1, ..., u_k$ kommer vi alltså att vilja använda basfunktionen $P_0(x)$ mellan knoppunkterna u_0 och u_1. Detta innebär att varje basfunktion multipliceras med en indikatorvariabel som antar värdet 1 för det intervallet där vi vill använda denna funktion och värdet 0 utanför det intervallet. Vi kan skriva funktionen på formen

$$F(x) = b_0 P_0(x)\mathbb{1}\{u_0 \leq x \leq u_1\} + b_1 P_1(x)\mathbb{1}\{u_1 < x \leq u_2\} + ... + b_k P_k(x)\mathbb{1}\{u_k < x \leq u_{k+1}\}.$$

där b_i är konstanter och $P_i(x)$ kallas för basfunktion och är alltså polynomfunktioner, (Maindonald & Braun, 2010).

Det man måste ha i åtanke när man jobbar med splines är att polynomfunktionerna måste mötas vid knoppunkterna och att detta bara kommer att hända för vissa uppsättningar av b:n. Detta innebär att $P_i(x)$ inte bara kommer att bero på x utan även på b_i. Det vill säga att om en koefficient ändras kommer även alla $P_i(x)$ behöva ändras. Istället kan en spline skrivas som en linjärkombination av en uppsättning basfunktioner. En sådan spline kallas för B-spline, s. 106 (Ohlsson & Johansson, 2010).

2.4.1 B-splines

Då vi använder B-spline kommer regressionsmodellen ha formen

$$F(x) = b_0 B_0(x) + b_1 B_1(x) + ... + b_k B_k(x) + \varepsilon$$

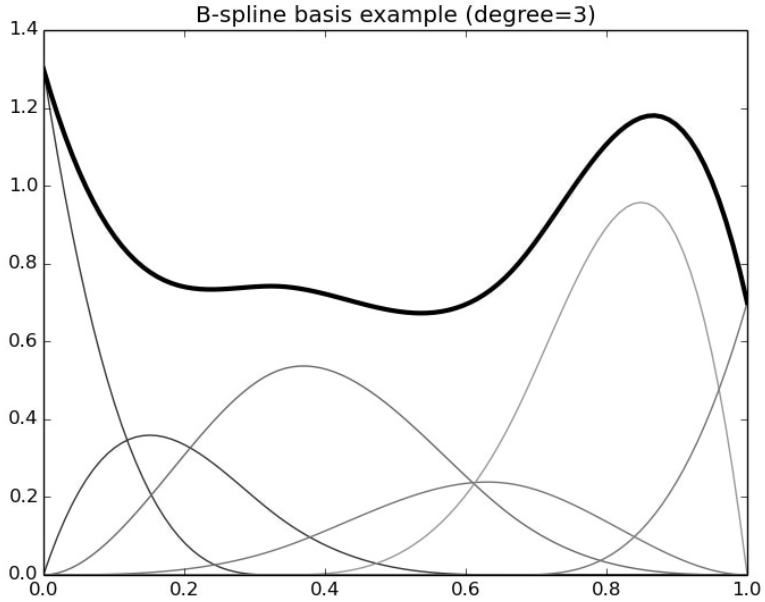
där ε är en felterm.

Då vi använder B-splines kommer vi för alla uppsättningar av b_i få en B-spline och alltså beror inte $P_i(x)$ på de olika b_i:na vilket gör B-splines lättare

att jobba med.

En B-spline definieras av dess ordning n och dess inre knopar N, totala antalet knopar kommer att vara $N + 2$ då ändpunkterna även räknas som knopar. Polynomens grad kommer ges av ordningen för splinen minus ett, $n - 1$. En B-spline är den basfunktion som är maximalt deriverbar mellan knoparna (Racine, 2014).

I Figur 1 ses ett exempel på en B-spline (Smith, 2011-2013), de smala linjerna är basfunktioner av grad tre. Dessa tillsammans bildar B-splinen som ses högst upp i figuren (den tjocka linjen).



Figur 1: Exempel på en B-spline av grad 3 (Smith, 2011-2013).

Tittar man i Figur 1 kan man till exempel se att för $0 \leq x < 0.15$ (ungefär) används polynomet med den väldigt branta lutningen, för större värden på x går man istället över till ett polynom med en mindre brant lutning. Man ser även tydligt i slutet av B-splinen (för $0.7 < x \leq 1$ ungefär) att man använt sig av polynomet med en konkav topp.

2.5 AIC

Akaike informationskriterium (AIC) används för att jämföra olika modeller med varandra. AIC baseras på värdet av maximum log likelihood av modellen men tar även hänsyn till antal parametrar i modellen. Modellen med lägst AIC-värde är bäst. AIC definieras som, s.216 (Agresti, 2002)

AIC=-2(maximum log likelihood - antal parametrar i modellen).

Det är viktigt att ha i åtanke att AIC endast kan användas som ett mått på hur bra en modell är jämfört med en annan modell med samma responsvariabel baserat på samma data. Om en modell har lägre AIC-värde än en annan modell är den bättre men ett enskilt AIC-värde säger ingenting.

2.6 Konfidensintervall

För att skapa ett konfidensintervall för en GLM behöver vi fördelningen för våra parametrar. Vet vi inte fördelningen kan vi använda oss av definitionen om asymptotisk normalfördelning

$$\sqrt{n}(T_n - g(\theta)) \longrightarrow N_m(0, \Sigma(\theta)) \text{ för alla } \theta \, \epsilon \, \Theta.$$

Där T_n är en sekvens av parameterskattningar, $g(\theta)$ är en m-dimensionell parameter och $\Sigma(\theta)$ är en positivt definit kovariansmatris, s.114 (Liero & Zwanzig, 2011). Vi kan sätta kovariansmatrisen lika med den inversa Fisher informationsmatrisen, s.45 (Ohlsson & Johansson, 2010), och allmänt skriva ett konfidensintervall för θ som

$$\hat{\theta} \pm \lambda_{\alpha/2} I(\hat{\theta})^{-1/2}.$$

Detta kallas för ett wald-konfidensintervall.

3 Data

Datamaterialet i denna studie kommer från Trafikanalys (Trafikanalys, 1999-2009), där vi sammanställt elva tabeller från åren 1999-2009 till en tabell. Data innehåller information om hur många polisrapporterade trafikolyckor som personbilsförare i Sverige varit med i under dessa år. Data innehåller även information om vilka av dessa personbilsförare som var män respektive kvinnor samt vilken åldersgrupp de tillhörde. Åldersgrupperna i datamaterialet är följande: 18-19, 20-24, 25-34, 35-44, 45-54, 55-64, 65-74, 75+. Det är även angivet hur många av dess olyckor som hade ett dödligt utfall. Se Tabell 1 för data sammanställt i tabell. Se även Tabell 2 för beteckningar för variablerna vi kommer att använda i denna studie.

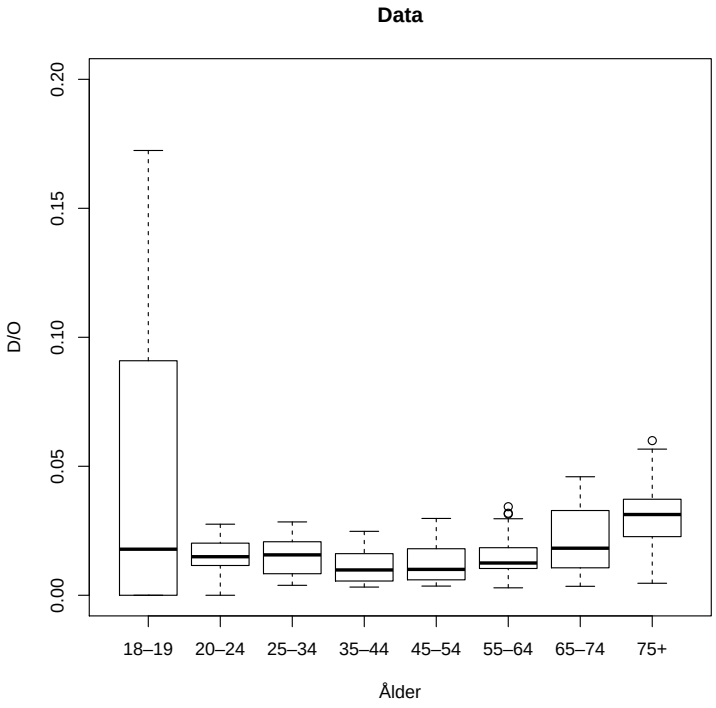
Tabell 1: Tabell över totala antalet olyckor fördelade mellan kön och ålder samt dödligt utfall eller ej.

Kön	Ålder	Dödligt utfall	
		Ja	Nej
Kvinna	18-19	2	72
	20-24	27	2772
	25-34	53	6011
	35-44	65	10871
	45-54	72	10804
	55-64	71	7812
	65-74	61	5592
	75+	50	2239
Man	18-19	24	296
	20-24	118	6016
	25-34	267	12667
	35-44	296	17303
	45-54	291	14988
	55-64	244	11215
	65-74	270	8274
	75+	193	4706

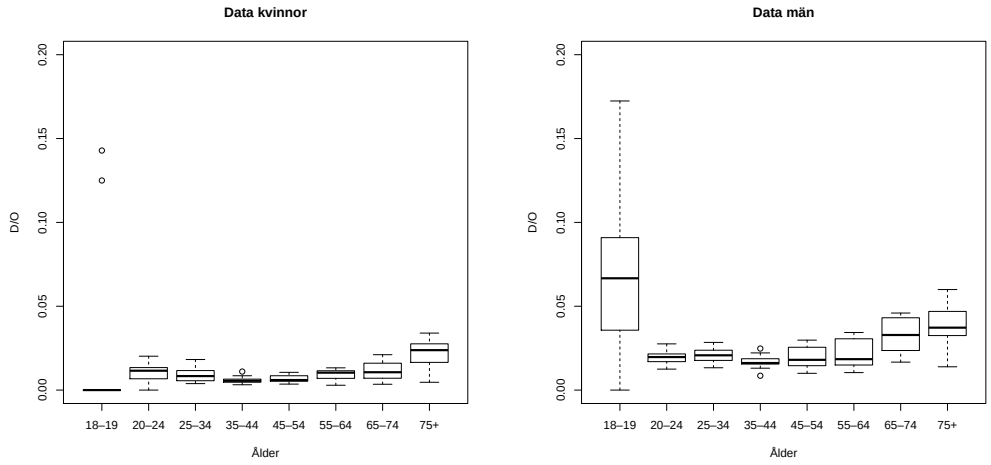
Tabell 2: Variabler i datamaterialet och dess beteckningar i denna studie.

Variabel	Beteckning
Antal olyckor	O
Antal döda	D
Män	M
Kvinnor	K

Då vi vill undersöka hur ålder och kön hos en personbilsförare påverkar risken för att få ett dödligt utfall givet att personen varit med i en trafikolycka kommer vi att använda kvoten D/O som responsvariabel (antalet olyckor med dödligt utfall delat med totala antalet olyckor). För att få en tydligare bild av hur datamaterialet ser ut, se Figur 2 och Figur 3. Där ser vi boxplottar för varje ålderskategori och varje box innehåller elva observationer, en för varje år.



Figur 2: Kvoten mellan antal olyckor med dödligt utfall och totala antalet olyckor plottat mot åldern.



(a) Kvinnor (b) Män

Figur 3: Kvoten mellan antal olyckor med dödligt utfall och totalt antal olyckor plottat mot åldern uppdelat för kön.

Varje olycka kommer att ha en bernoullifördelning då varje olycka antingen kan ha ett dödligt utfall eller ett ej dödligt utfall. Då vi har summan av flera sådana fördelningar i vårt dataset kommer vårt data vara binomialfördelat.

4 Modellering

Vi vill i denna studie undersöka hur ålder och kön hos en personbilsförare påverkar utfallet av en trafikolycka. För att undersöka detta vill vi först ställa upp en lämplig modell som kan beskriva vårt data. När vi hittat en lämplig modell vill vi undersöka hur ålder påverkar utfallet av en trafikolycka. Detta görs genom att skatta en B-spline som beskriver datapunkterna, detta och all annan modellering i denna studie sker i `R` (R Core Team, 2014). När vi hittat vår B-spline vill vi skapa ett konfidensintervall för denna för att kunna se hur säkerställd den kan antas vara. Vi delar sedan upp vårt data för män och kvinnor för att se hur de påverkar utfallet av en trafikolycka. På samma sätt kommer vi att konstruera B-splines för kvinnornas och männens data separat, samt konstruera konfidensintervall för dem båda.

4.1 Antagande om GLM

Eftersom vårt data inte är normalfördelat kan vi inte använda oss av multipel linjär regression, vi vill istället använda oss av en generaliserad linjär modell. Som vi beskrev i teoridelen vill vi då att vår sannolikhetsfördelning ska kunna skrivas på formen som en naturlig exponentialfamilj, alltså på formen

$$f(y_i;\theta_i) = a(\theta_i)b(y_i)exp\{y_i\mathcal{Q}(\theta_i)\}.$$

Vi har en binomialfördelning vars sannolikhetsfördelning ser ut på följande sätt

$$P(y|\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}, \quad n = 1, 2..., \quad y = 1, 2, ...n, \quad 0 \leq \theta \leq 1$$

där $\theta = \frac{D}{O}$ i vår modell. Vi kan skriva om sannolikhetsfördelningen som

$$\binom{n}{y}exp\{y\ln(\theta) + (n-y)\ln(1-\theta)\} =$$

$$= \binom{n}{y}exp\{y(\ln(\theta) - \ln(1-\theta))\}exp\{n\ln(1-\theta)\} =$$

$$= \binom{n}{y}exp\{y\ln\left(\frac{\theta}{1-\theta}\right)\}exp\{n\ln(1-\theta)\}.$$

Där vi kan skriva

$$a(\theta) = exp\{n\ln(1-\theta)\}$$

$$b(y) = \binom{n}{y}$$

$$\mathcal{Q}(\theta) = \ln\left(\frac{\theta}{1-\theta}\right) = logit(\theta).$$

Vår sannolikhetsfördelning tillhör alltså den naturliga exponentialfamiljen där vår kanoniska länkfunktion är en logitlänk.

4.2 Analys av ålderns inverkan

Första åtgärden blir, då vi inte är intresserade av att jämföra inverkan av ålder och kön för olika år, att lägga ihop datamaterialet för olika år. I denna stund lägger vi även ihop datat för olika kön då den inledande analysen endast undersöker vilken påverkan åldern hos en personbilsförare har.

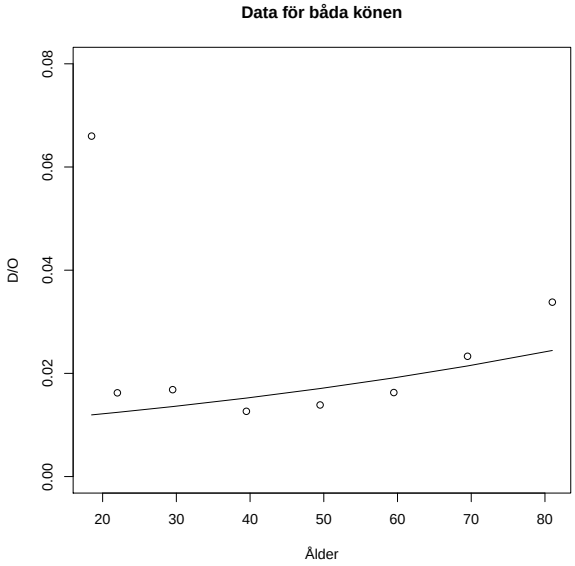
Då vi har en generaliserad linjär modell med en binomialfördelad responsvariabel och logit som länkfunktion har vi en logistisk regressionsmodell som vi kan skriva på formen

$$logit\left(\frac{D}{O}\right) = \alpha + \beta x_i \quad i = 1, ..., 8$$

där x_i anger vilken åldersgrupp personbilsföraren tillhör.

Nästa åtgärd blir att göra om ålder till en icke-kategorisk variabel och vi skapar därför vektorn (18.5, 22, 29.5, 39.5, 49.5, 59.5, 69.5, 81) som tar medelvärdet för varje ålderskategori. Valet av 81 år som medelvärdet för 75+ baseras på statistik från Statistiska Centralbyrån. Enligt Statistiska Centralbyrån levde 834 489 personer i Sverige som var 75 år eller äldre 2014, men antalet personer är inte likformigt fördelade för dessa åldrar. Vi kan beräkna att 439 006 av dessa personer var 81 år eller yngre vilket på ett ungefär är hälften av de som var 75 år eller äldre detta år. Åldersfördelningen i Sverige kan självklart ha ändrats sedan åren 1999-2009 men vi väljer att använda 81 år som ett approximerat medelvärde för de som är 75 år eller äldre.

Om vi plottar dessa åldrar mot kvoten D/O (antalet olyckor med dödligt utfall delat på totala antalet olyckor) får vi resultatet som ses i Figur 4. Detta ger oss en inblick i hur ålder påverkar utfallet av en trafikolycka. Dock bara för vissa åldrar, det vi nu vill göra för att kunna se ålderns inverkan är att skatta en linje som kan förklara detta samband för alla olika åldrar inom intervallet [18, 81].



Figur 4: Responsvariabeln antal döda delat på antal olyckor plottat mot åldern samt skattad regressionslinje.

Vi kan börja med att skatta en linje för vår logistiska regressionsmodell, vi anger vår modell i `R` med kommandot `glm` och skattar sedan en linje, kod för detta ses i appendix. Den skattade linjen vi får ses i Figur 4. Vi ser att denna linje inte passar datapunkterna särskilt bra. Skattningen kommer ha ett AIC-värde på 197.3929. Parameterskattningarna vi får ut i `R` för denna modell samt övriga modeller i detta arbete presenteras i appendix.

Att linjen inte passar särskilt bra är inte förvånande om vi tittar på datapunkterna i Figur 4. Datapunkterna har inte en tydlig "u-kurva", vi ser att vi har en väldigt brant negativ lutning mellan de två första datapunkterna för att sedan ha en positiv lutning, tillbaka till en negativ och sedan avsluta med en positiv lutning. Punkterna verkar vända fler gånger än vad som rimligt kan förklaras med endast ett polynom av låg grad. Vi vill därför använda oss av B-splines som kombinerar flera polynom, i denna studie av grad tre, till en kurva. Vi vill se om vi kan hitta en B-spline med lägre AIC-värde än 197.3929.

För att skapa en B-spline behöver vi bestämma hur många knopar vi ska ha och vart dessa ska sitta, alltså från vilken punkt vi vill byta från ett poly-

nom till ett annat. För att välja den B-spline som passar datat bäst testar vi olika antal knopar vid olika punkter och jämför deras AIC-värde och väljer sedan den modell med lägst AIC.

Då vi inte har alltför många datapunkter kan vi börja med att studera Figur 4 för att se om vi kan avgöra hur många inre knopar som kan anses lämpligt. Till exempel kan vi se att mellan åldrarna 20 och 30 ser det ut att vara lämpligt att byta polynom då vi först har en väldigt hög negativ lutning och behöver sedan gå över till en mindre negativ lutning, på samma sätt som vi såg som exempel i Figur 1. Dock är det svårt att med blotta ögat avgöra exakt hur många inre knopar som är lämpligt. Vi kan se att fler än tre inre knopar antagligen är överflödigt och att antalet inre knopar borde ligga mellan ett och tre. Men detta är något vi behöver testa vidare.

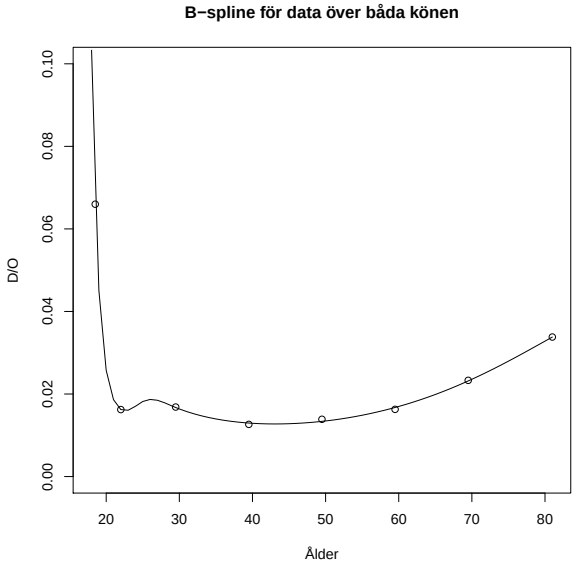
För att hitta den B-spline med lägst AIC-värde sätter vi de inre knoparna lika med vektorn av alla ålderskategorier förutom den lägsta och den högsta åldern då dessa kommer att vara våra yttre knopar. Knoparna behöver inte ligga vid datapunkterna men vi börjar med dem för att få en känsla av hur många inre knopar som krävs. När vi vet detta kan vi sedan flytta runt knoparna för att se om detta kan ge ett lägre AIC-värde.

Då vi sätter knoparna lika med vektorn av ålderskategorierna (22, 29.5, 39.5, 49.5, 59.5, 69.5) kommer vi att få en B-spline som sticker iväg väldigt mycket mellan datapunkterna, det verkar lämpligt med färre inre knopar. Vi testar oss fram genom att utesluta olika knopar och ser att om vi utesluter 59.5 och 69.5 kommer AIC-värdet att vara oförändrat, detta tyder på att dessa knopar är överflödiga. Vi går vidare genom att utesluta en knop, lägga tillbaka denna och sedan utesluta nästa. Vi undersöker vilken av dessa som ger den största minskningen i AIC och tar sedan bort den. Första omgången ger uteslutning av 49.5 störst minskning i AIC och vi tar bort den som inre knop. Tar vi bort ytterligare en knop kommer AIC fortfarande att minska vilket tyder på att det är lämpligt med färre än tre knopar, störst minskning sker när vi tar bort 39.5 och denna utesluts nu helt. Tar vi bort ytterligare en knop kommer AIC att öka, vi verkar alltså få den bästa passningen när vi har två inre knopar. Se Tabell 3 för några av de olika testade knopar och dess AIC-värden.

Tabell 3: Olika knopar och dess AIC värden.

Knopar	*AIC*
22, 29.5, 39.5, 49.5, 59.5, 69.5	73.34522
22, 29.5, 39.5, 49.5	73.34522
22, 29.5, 39.5	71.99861
22, 29.5	70.34309
25, 29	70.34309
29.5	94.8877
22	75.98131

Nu återstår alltså knoparna (22, 29.5) med ett AIC-värde på 70.34309, dock finns det som sagt inget som säger att knoparna måste ligga vid datapunkter. Med dessa punkter som utgångspunkter kan vi därför flytta runt dessa lite för att se om vi kan nå ett lägre AIC-värde, dock verkar inte AIC-värdet bli lägre. Däremot kommer knoparna (22, 29.5) ge en ganska hackig kurva mellan åldrarna 20 och 30. Det går att få en mjukare kurva genom att flytta knoparna lite. Vi testar oss fram och får att med knoparna (25, 29) kommer vi att få en B-spline som också har ett AIC-värde på 70.34309 men en mjukare kurva mellan åldrarna 20 och 30 som kan anses mer rimligt. Vi testar även att lägga till ytterligare en knop vid olika ställen för att se om vi kan få ett lägre AIC-värde, men vi får inte ett lägre AIC-värde vilket styrker antagandet om att två knopar är lämpligt för att skapa en B-spline som förklarar datat. Vi väljer alltså B-splinen med knopar=(25, 29) som kan ses i Figur 5. För att få ut vår B-spline använder vi oss av funktionen `bs` i `R` som anger att det är en B-spline. Kod för hur man anger en B-spline samt inre knopar i `R` ses i appendix.



Figur 5: B-spline med inre knopar=(25, 29).

Som vi ser i Figur 5 kommer vår B-spline att följa datapunkterna mycket bättre än när vi skattade en linje med linjär regression. Vi har även sett att B-splinen har ett AIC-värde på 70.34309 som är betydligt lägre än 197.3929.

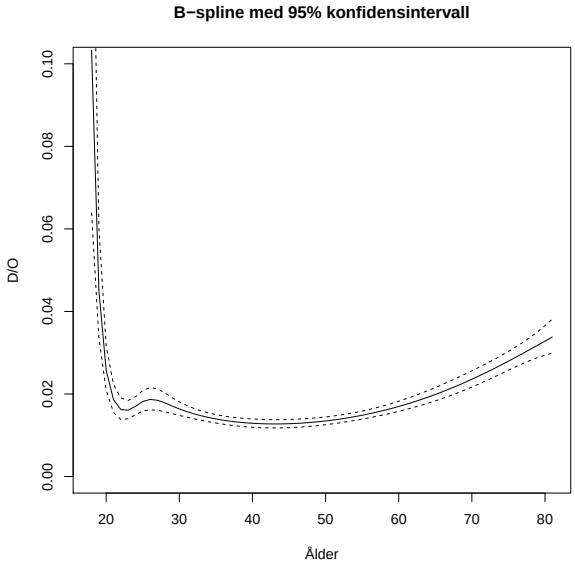
Nu när vi hittat vår B-spline är nästa steg att på något sätt undersöka hur säker denna skattning är. Vi väljer att göra detta genom att skapa ett 95%-igt konfidensintervall. Vi kan börja med att skapa ett konfidensintervall för $logit(\frac{D}{O})$ för att sedan transformera detta och då få ett konfidensintervall för $\frac{D}{O}$. Vi använder oss av ett wald-konfidensintervall

$$\text{KI}(\text{logit}(\tfrac{D}{O})) : \hat{\alpha} + \hat{\beta}_1 x_i \pm 1.96 * I^{-1/2} = (\gamma_1, \gamma_2)$$

Vi transponerar sedan detta konfidensintervall för att få ett 95%-igt konfidensintervall för $\frac{D}{O}$, detta ges av

$$\text{KI}(\tfrac{D}{O}) = \left(\tfrac{exp\{\gamma_1\}}{1+exp\{\gamma_1\}}, \tfrac{exp\{\gamma_2\}}{1+exp\{\gamma_2\}}\right).$$

Konfidensintervallet beräknat för vår B-spline ses i Figur 6.

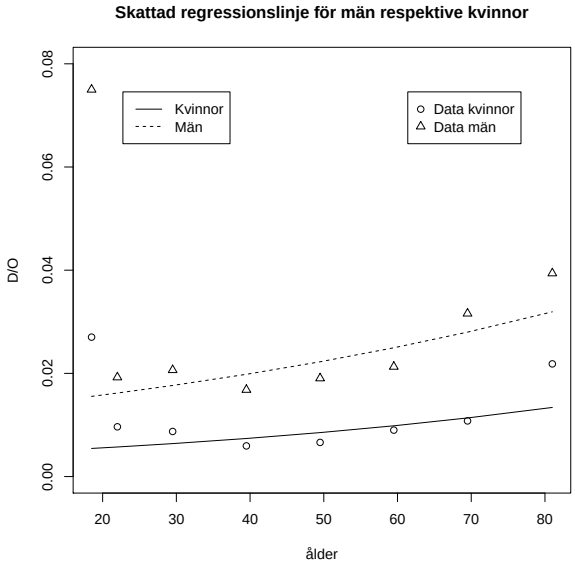


Figur 6: B-spline med 95%-igt konfidensintervall.

Konfidensbanden tyder på att B-splinen på ett bra sätt förklarar hur dödsolyckor beror på ålder. B-splinen verkar vara aningen mer osäker för personer i åldrarna 20-30 samt för de allra högsta åldrarna, men i allmänhet verkar konfidensbanden följa mönstret av vår B-spline.

4.3 Analys av könets inverkan

Nu delar vi upp datasetet för män och kvinnor för att se deras inverkan på dödsfall hos personbilsförare i trafikolyckor. På samma sätt som för hela datamaterialet kan vi plotta responsvariabeln, D/O, mot ålder uppdelat för de båda könen och skatta linjer, detta ses i Figur 7.



Figur 7: Skattad regressionslinje för män respektive kvinnor samt deras datapunkter.

Vi ser i Figur 7 att vi har två linjer som inte följer datapunkterna särskilt bra. AIC-värdet för linjen skattad för kvinnor är 81.09217 och för männen 143.0656. Vi vill även här skatta B-splines för de båda datamaterialen för att kunna få en bättre anpassning.

Vi vill nu hitta en B-spline som kan förklara datapunkterna för män och en B-spline som kan förklara datapunkterna för kvinnor. På samma sätt som för det gemensamma datamaterialet letar vi efter de knoparna som ger oss B-splinen med lägst AIC-värde. Vi vill helst ha samma inre knopar för männen som för kvinnor, detta för att det underlättar om man sedan vill jämföra kurvorna genom deras uppsättning av skattade koefficienter.

För att hitta knoparna börjar vi med att lägga till en knop, ta bort denna och lägga till en annan. Vi gör detta med alla de sex inre åldrarna (22, 29.5, 39.5, 49.5, 59.5, 69.5) och ser vilken av dem som ger lägst AIC-värde, denna behåller vi. Vi lägger sedan till ytterligare en knop, en i taget av de fem återstående värdena, den som ger oss lägst AIC-värde behåller vi. Vi gör detta tills en ytterligare knop inte ger ett lägre AIC-värde.

För kvinnor kommer 22 som knop ge lägst AIC-värde i första steget och addering av ytterligare en av de fem knoparna kommer inte att ge ett lägre AIC-värde. För männen kommer även 22 som knop ge lägsta AIC-värdet i första steget och här får vi även lägre AIC om vi inkluderar ännu en knop. Lägst AIC får vi när vi inkluderar 29.5, inkludering av en tredje knop kommer inte att ge ett lägre AIC-värde. För kvinnorna har vi alltså en inre knop=(22) och för männen har vi två inre knopar=(22, 29.5). Vi kan notera att de inre knoparna som ger lägst AIC för männen, (22, 29.5), även är samma knopar som gav lägst AIC för hela datamaterialet.

Som vi nämnde tidigare vill vi helst ha samma knopar för de båda datamaterialen och vi vill därför bestämma och vi ska använda en eller två inre knopar för båda datamaterialen. Vi kan beräkna att om vi lägger till knopen 29.5 till datat för kvinnorna kommer detta att ge en ökning på ca 2.6 procentenheter AIC. Om vi istället tar bort 29.5 som knop för männen kommer detta att ge en ökning på ca 5 procentenheter AIC. B-splinen för männen ”lider“ alltså mer av att gå över till knoparna för kvinnorna än tvärtom och vi väljer därför att använda knoparna=(22, 29.5) för både män och kvinnor.

På samma sätt som för det gemensamma datamaterialet gäller det att de inre knoparna inte nödvändigtvis måste ligga vid datapunkterna. Då datapunkterna för männen på många sätt liknar de gemensamma datapunkterna kommer vi även här få en mjukare B-spline om vi byter till knoparna=(25, 29). Detta kommer inte att ändra AIC-värdet varken för männen eller kvinnorna. För några olika knopar och AIC-värden, se Tabell 4. För B-splines för män och kvinnor med knopar=(25, 29) se Figur 8.

Tabell 4: Olika knopar och dess AIC-värden.

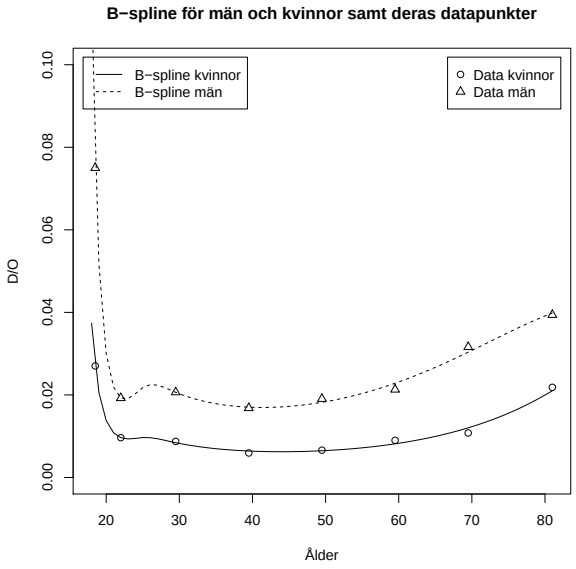
(a) Kvinnor

Knopar	*AIC*
22	55.94231
22, 29.5	57.4049
25, 29	57.4049

(b) Män

Knopar	*AIC*
22	73.81953
22, 29.5	70.07225
25, 29	70.07225

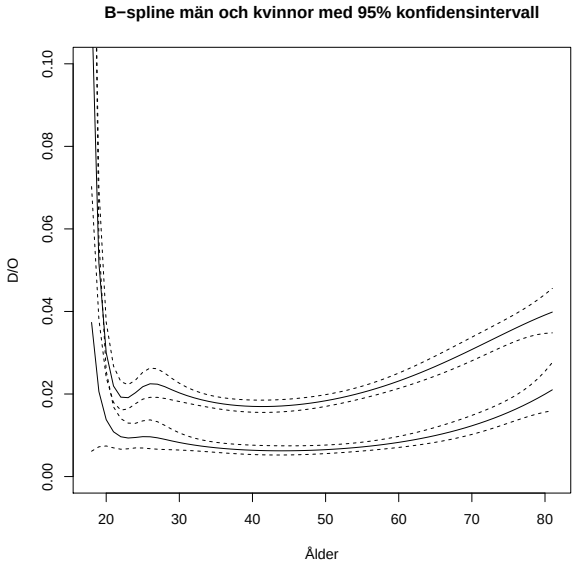
När vi tittar i Tabell 4 måste vi ha i åtanke att vi inte kan jämföra AIC-värden mellan män och kvinnor. Att det är lägre AIC-värden för kvinnor än för män säger inget om att kvinnornas B-spline passar bättre än B-splinen för män då dessa värden är baserade på olika datamaterial. Dock kan vi jämföra dessa AIC-värden med dem vi fick ut när vi utförde linjär regression. Vi ser att B-splinen både för män och kvinnor kommer att ge lägre AIC-värden än med linjär regression.



Figur 8: Datapunkter samt B-splines uppdelat för män och kvinnor.

Jämför vi de två B-splines vi ser i Figur 8 med den för hela datamaterialet (Figur 5) kan vi se att B-splinen för män har samma form, även den för kvinnor är snarlik bortsett från att den har en mindre böjd kurva mellan åldrarna 20-30 samt inte lika brant lutning för de högsta och lägsta åldrarna.

Vi vill även för dessa B-splines undersöka hur bra de kan anses vara och vi kommer på samma sätt som för det gemensamma datamaterialet att skapa ett varsitt 95%-igt konfidensintervall. Dessa konfidensintervall för män respektive kvinnor ses i Figur 9. Den övre linjen är B-splinen för män och den undre linjen är B-splinen för kvinnor.



Figur 9: B-spline uppdelat för män och kvinnor med 95%-iga konfidensintervall.

Konfidensintervallet för B-splinen skattad för att förklara männens påverkan på utfall av olyckor liknar i många avseenden Figur 6, som visar konfidensintervallet för hela datasetet. Där splinen verkar passa bra för att beskriva de allra lägsta åldrarna samt åldrarna 30-70 år och för åldrarna 20-30 år samt för de allra äldsta åldrarna ser resultatet aningen osäkrare ut.

Konfidensintervallet för B-splinen för kvinnor är lite annorlunda jämfört med de vi sett tidigare. Detta konfidensintervall är som bredast för de allra lägsta åldrarna. Det är så pass brett att för dessa åldrar verkar vi inte kunna dra några slutsatser alls om hur dessa åldrar påverkar utfallet av en trafikolycka. I övrigt liknar konfidensintervallet de tidigare, det är som smalast för åldrarna 30-70 och bredare för de lägre och högre åldrarna.

För de allra lägsta åldrarna kommer konfidensintervallen för män och kvinnor att korsa varandra. Detta tyder på att för dessa åldrar kan vi inte säkerställa på 95%-nivån att risken för en personbilsförare att omkomma i en trafikolycka skiljer sig för de olika könen.

5 Diskussion

Genom vår analys har vi kommit fram till att 18-åriga personbilsförare är de som löper absolut störst risk att dö givet en trafikolycka. Risken verkar sedan minska drastiskt fram tills åldern 22. Risken ser sedan ut att öka en aning fram tills 28-årsålder där den sedan minska långsamt fram till 40-årsåldern, där risken ser ut att vara som lägst. För högre åldrar kommer risken att öka långsamt.

När vi skapar ett 95%-igt konfidensintervall för B-splinen för det gemensamma datamaterialet kan vi se att konfidensintervallet följer B-splinen relativt bra. Konfidensintervallet är aningen större mellan åldrarna 20-30 år, samt för personer äldre än 70 år. Mellan åldrarna 20-30 år kan vi alltså inte säkerställa att risken går upp, då konfidensintervallet är så pass brett att vi skulle kunna ha en konstant minskning från 18 till 40 år. För personer över 70 år kan vi säkerställa på 95%-nivån att risken fortsätter att öka men inte riktigt hur mycket.

När vi delar upp datat på kön och skattar B-splines för varje kön separat kan vi se att dessa till stor del följer samma form som B-splinen för det gemensamma datamaterialet. B-splinen för respektive kön skiljer sig dock åt på formen mellan åldrarna 20-30 år, männen har samma uppåtgående trend som det gemensamma datamaterialet medan kvinnorna har en mer plan kurva. Konfidensintervallet för det gemensamma datamaterialet pekar på precis detta, det är svårt att säkerställa om risken går upp eller inte mellan dessa åldrar.

Vi ser också att B-splinen för män generellt sett ligger högre än B-splinen för kvinnor. Detta tyder på att män i allmänhet löper större risk att omkomma vid en trafikolycka, oavsett ålder.

Om man tittar på konfidensintervallet för männens data kan vi se att det till en början är väldigt snävt, men redan för åldrarna 20-30 år blir det bredare. Där kan vi inte längre avgöra om det finns en reell ökning i risk eller om kurvan planar ut. Högre upp i åldrarna blir konfidensintervallet snävare, för att sedan bli bredare igen vid 65 års ålder. Där ser vi att risken ökar, men inte med hur mycket. I mångt och mycket är konfidensintervallet för männen likt konfidensintervallet för det gemensamma datamaterialet.

Konfidensintervallet för kvinnornas data är på vissa ställen bredare jämfört med männens konfidensintervall. För 18-åriga kvinnor är konfidensintervallet så pass brett att vi inte alls kan avgöra om det finns en negativ eller positiv lutning. Vi ser även att för 18-åringar kommer konfidensintervallen för kvinnor och män att korsa varandra, vilket betyder att för 18-åringar kan vi på 95%-nivån inte säkerställa att kön har en inverkan på utfallet av en trafikolycka.

I likhet med männens konfidensintervall kommer även kvinnornas konfidensintervall att vara lite bredare mellan åldrarna 20-30 år, det är därför svårt att avgöra hur risken förändras mellan dessa åldrar. Vi ser, också i likhet med männens konfidensintervall, att det finns en viss osäkerhet för de högsta åldrarna. Det finns en positiv lutning, men det är svårt att avgöra hur stor lutningen är.

I allmänhet liknar datamaterialet för männen det gemensamma datamaterialet mer än det för kvinnor. En anledning till detta kan vara att det finns fler observationer för männen. Om det beror på att män är med i fler trafikolyckor eller att män spenderar mer tid i trafiken är något vi inte kan säkerställa i denna studie.

En slutsats är att det hade krävts mer data för en bättre analys, detta på gott och ont. Mer data hade gett ett säkrare resultat i vår analys samtidigt som att man får vara lättad över att det inte sker fler trafikolyckor än vad det gör i Sverige. En förbättring hade varit att ha tillgång till mer historisk data (äldre än 1999 och senare än 2009), vilket hade lett till fler observationer.

Vårt resultat tyder på att 18-åringar löper absolut störst risk att omkomma givet att de varit med i en trafikolycka. Bortsett från denna ålder ser dock äldre personer, som varit med i en trafikolycka, ut att löpa en förhöjd risk att omkomma jämfört med lägre åldrar. Detta skulle kunna förklaras av en av slutsatserna som de kom fram till i (Williams & Shabanova, 2003). Där skrev de att unga personer som var ansvariga för trafikolyckor med dödligt utfall ofta själva inte omkom, medan äldre som var ansvariga för trafikolyckor med dödligt utfall ofta själva omkom. Då vi inte tittar på vem som är ansvarig i olyckan utan alla inblandade förare skulle detta kunna stämma även i vårt fall. Om äldre själva omkommer i olyckorna de är inblandade

i och yngre förare är med i väldigt många olyckor men ej omkommer själva kommer kvoten D/O att bli mindre för yngre förare och större för äldre förare. Dock är resultatet i (Williams & Shabanova, 2003) baserat på trafikdata från USA men det är inte omöjligt att dessa slutsatser även skulle kunna dras från svenskt trafikdata. Detta är något vi inte kan säkerställa i denna studie men något som skulle vara intressant att undersöka i framtiden.

I (Massie *et al.*, 1994) kom man bland annat fram till att män är med i flest trafikolyckor med dödligt utfall, medan kvinnor är med i fler trafikolyckor i allmänhet. Med samma resonemang som i föregående stycke skulle detta kunna förklara varför kvinnornas B-spline för alla åldrar ligger lägre än männens. I (Massie *et al.*, 1994) kom man även fram till att äldre personer är med i fler trafikolyckor med dödligt utfall, medan yngre personer är med i fler trafikolyckor i allmänhet. Samma slutsats skulle även kunna dras i denna studie (bortsett från 18-åringarna), men för att säkerställa detta behövs ett mått på hur mycket varje åldersgrupp befinner sig i trafiken, ett mått vi i denna studie har velat undvika.

Nu när vi kommit fram till detta resultat är den naturliga frågan: varför? Hur kommer det sig att män löper större risk att omkomma i en trafikolycka än kvinnor, och varför löper yngre och äldre personer oavsett kön en större risk att omkomma än personer i medelåldern? Det finns flera olika teorier som skulle vara intressanta att undersöka. Det kan vara bristande erfarenhet bland yngre förare och långsamma reaktionsförmågor hos äldre. Kvinnor kanske tenderar att köra stora och tunga familjebilar medan män i större utsträckning kör lättare sportbilar. Män, eller yngre förare i allmänhet, kanske tenderar att överskatta sin egen eller bilens förmåga. Kanske utsätts yngre förare för mer mörkerkörning än andra, och för äldre personer kanske inte synen är vad den en gång var. Detta är frågor som skulle vara intressanta att studera vidare i framtiden men detta ligger utanför denna studie.

6 Appendix

6.1 Kod

Den logistiska regressionsmodellen skrivs in i `R` på följande sätt

`model = glm(cbind(D, O-D) ~ ålder, family=binomial(logit), data=data).`

För att ange att vi vill använda oss av en B-spline med inre knopar = (25, 29) skriver vi

```
library(splines)
spline <- glm(cbind(D, O-D) ~ bs(ålder,knots=c(25,29)),
family=binomial(link=logit), data=data).
```

6.2 Parameterskattningar

```
Deviance Residuals:
   Min      1Q  Median      3Q     Max
-4.107  -3.131   2.303   4.041   6.888

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.630050   0.071540  -64.72   <2e-16 ***
ålder        0.011643   0.001343    8.67   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 210.46  on 7  degrees of freedom
Residual deviance: 136.05  on 6  degrees of freedom
AIC: 197.39

Number of Fisher Scoring iterations: 4
```

Figur 10: Parameterskattningar för vår linjära modell.

```
Deviance Residuals:
      1       2       3       4       5       6       7       8
 0.0000  0.0000  0.1268 -0.4694  0.7115 -0.4916  0.1166  0.0021

Coefficients:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -2.6500     0.2029 -13.059  < 2e-16 ***
bs(ålder, knots = c(25, 29))1 -1.9280     0.2997  -6.433 1.25e-10 ***
bs(ålder, knots = c(25, 29))2 -1.1553     0.2301  -5.021 5.13e-07 ***
bs(ålder, knots = c(25, 29))3 -2.2622     0.2640  -8.567  < 2e-16 ***
bs(ålder, knots = c(25, 29))4 -1.2281     0.2528  -4.857 1.19e-06 ***
bs(ålder, knots = c(25, 29))5 -0.7029     0.2128  -3.303 0.000956 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 210.46489  on 7  degrees of freedom
Residual deviance:   0.99787  on 2  degrees of freedom
AIC: 70.343

Number of Fisher Scoring iterations: 3
```

Figur 11: Parameterskattningar för vår B-spline.

```
Deviance Residuals:
      1       2       3       4       5       6       7       8
 0.0000  0.0000  0.2638 -0.6179  0.1904  0.8261 -0.8692  0.2748

Coefficients:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -3.5835     0.7169  -4.999 5.77e-07 ***
bs(ålder, knots = c(25, 29))1 -1.3220     0.9478  -1.395   0.1631
bs(ålder, knots = c(25, 29))2 -0.9160     0.7634  -1.200   0.2302
bs(ålder, knots = c(25, 29))3 -2.0156     0.8143  -2.475   0.0133 *
bs(ålder, knots = c(25, 29))4 -1.1606     0.7923  -1.465   0.1429
bs(ålder, knots = c(25, 29))5 -0.2578     0.7310  -0.353   0.7243
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 53.9067  on 7  degrees of freedom
Residual deviance:  2.0011  on 2  degrees of freedom
AIC: 57.405

Number of Fisher Scoring iterations: 4
```

Figur 12: Parameterskattningar för vår B-spline för kvinnor.

```
Deviance Residuals:
       9        10        11        12        13        14        15        16
 0.00000   0.00000   0.01336  -0.21935   0.76463  -1.11439   0.66769  -0.16311

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                    -2.5123     0.2122 -11.837  < 2e-16 ***
bs(ålder, knots = c(25, 29))1  -1.8888     0.3202  -5.900 3.65e-09 ***
bs(ålder, knots = c(25, 29))2  -1.1207     0.2435  -4.603 4.17e-06 ***
bs(ålder, knots = c(25, 29))3  -2.0634     0.2838  -7.270 3.60e-13 ***
bs(ålder, knots = c(25, 29))4  -0.9790     0.2713  -3.608 0.000308 ***
bs(ålder, knots = c(25, 29))5  -0.6696     0.2240  -2.990 0.002794 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 148.1744  on 7  degrees of freedom
Residual deviance:   2.3472  on 2  degrees of freedom
AIC: 70.072

Number of Fisher Scoring iterations: 3
```

Figur 13: Parameterskattningar för vår B-spline för män.

Referenser

Agresti, Alan. 2002. *Categorical Data Analysis*. 2 edn. John Wiley Sons.

Liero, Hannelore, & Zwanzig, Silvelyn. 2011. *Introduction to the theory of statistical inference*. Taylor & Francis Ltd.

Maindonald, John, & Braun, John. 2010. *Data Analysis and Graphics Using R, Chapter 7*. 3 edn. Cambridge University Press.

Massie, Dawn L, Campbell, Kenneth L, & Williams, Allan F. 1994. *Traffic accident involvement rates by driver age and gender*. 2015-02-03. Accident Analysis & Prevention, Vol. 27, No. 1. http://ac.els-cdn.com/000145759400050V/1-s2.0-000145759400050V-main.pdf?_tid=aca75d4e-f498-11e4-82c0-00000aab0f27&acdnat=1430989909_0804f6ac4b87ef8326503062ee78e5a4.

Ohlsson, Esbjörn, & Johansson, Björn. 2010. *Non-life insurance pricing with generalized linear models*. Springer-verlag Gmbh.

R Core Team. 2014. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Racine, Jeffrey S. 2014. *A primer on regression splines*. 2015-02-12. http://cran.r-project.org/web/packages/crs/vignettes/spline_primer.pdf.

Smith, Nathaniel J. 2011-2013. *Spline regression*. 2015-04-02. http://patsy.readthedocs.org/en/latest/spline-regression.html.

Trafikanalys. 1999-2009. *Vägtrafikskador*. 2015-02-20. http://www.trafa.se/sv/Soksida/#query=vÃďgtrafikskador.

Williams, Allan F, & Shabanova, Veronika I. 2003. *Responsibility of drivers, by age and gender, for motor-vehicle crash deaths*. 2015-02-03. Journal of Safety Research. http://ac.els-cdn.com/S0022437503000732/1-s2.0-S0022437503000732-main.pdf?_tid=d8c87322-f498-11e4-adf6-00000aacb35f&acdnat=1430989983_3e6c2a569e2dac945a29932b93a2a611.