Stockholms universitet

Jämförelse av prediktionsförmågan mellan olika regressionsmodeller för icke-melanom hudcancer i Halland

Sebastian Bergström

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2016:14
Matematisk statistik
Juni 2016

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholms universitet

Matematisk statistik
Stockholms universitet
Kandidatuppsats **2016:14**
http://www.math.su.se/matstat

Jämförelse av prediktionsförmågan mellan olika regressionsmodeller för icke-melanom hudcancer i Halland

Sebastian Bergström*

Juni 2016

Sammanfattning

Antalet nya fall av icke-melanom hudcancer ökar i Sverige och Halland har visat en stor ökning sedan 1970. I den här uppsatsen har olika modeller jämförts för att se vilken som är lämpligast för prediktion av antalet nya cancerfall. De undersökta modellerna är Poisson- och negativ binomialregression samt splineregression baserat på dessa två. Under modelleringsarbetet prövades olika gruppindelningar av åldersgrupper samt beroenden mellan antalet nya hudcancerfall och antalet år efter 1970. Modellval gjordes baserat på AIC och devianceresidualplottar. Resultaten visade att en negativ binomial splineregression var lämpligast men led av bristen att den underskattar antalet nya hudcancerfall en aning. Tolkning av den slutgiltiga modellen visar att ökningen varit allvarligare för äldre åldersgrupper och att antalet nya hudcancerfall ökat något mer för kvinnor än för män över över tidsperioden 1970-2014.

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige.
E-post: sebastian.a.bergstrom@gmail.com. Handledare: Martin Sköld och Mikael Petersson.

Sammanfattning

Antalet nya fall av icke-melanom hudcance ökar i Sverige och Halland har visat en stor ökning sedan 1970. I den här uppsatsen har olika modeller jämförts för att se vilken som är lämpligast för prediktion av antalet nya cancerfall. De undersökta modellerna är Poisson- och negativ binomialregression samt splineregression baserat på någon av dessa två. Under modelleringsarbetet prövades olika gruppindelningar av åldersgrupper samt beroenden mellan antalet nya hudcancerfall och antalet år efter 1970. Modellval gjordes baserat på AIC och devianceresidualplottar. Resultaten visade att en negativ binomial splineregression var lämpligast men led av bristen att den underskattar antalet nya hudcancerfall en aning. Tolkning av den slutgiltiga modellen visar att ökningen varit allvarligare för äldre åldersgrupper och att antalet nya hudcancerfall ökat något mer för kvinnor än för män över över tidsperioden 1970-2014.

Abstract in English

The number of cases on non-melanoma skin cancer is increasing in Sweden and Halland has shown a large increase since 1970. This thesis has compared different regression models to find the most suitable one for predicting the amount of new skin cancer cases. The models compared are Poisson regression, negative binomial regression and spline regressions based on the two. During modeling, different age groupings and relationships between the number of new cancer cases and years after 1970 have been examined. Model selection was done on the basis of AIC and deviance residual plots. The results show that a spline regression based on the negative binomial case is the most suitable but that it suffers from underestimating the amount of new skin cancer cases. An interpretation of the results shows that skin cancer incidence has been most severe among the elderly and that women show a slightly larger increase than men over the time period 1970-2014.

Förord och tack

Jag skulle vilja tacka mina handledare Martin Sköld och Mikael Petersson för deras tålamod och rådgivning längs vägen av arbetet med den här uppsatsen. Den här uppsatsen utgör ett examensarbete på 15 högskolepoäng vid Matematiska institutionen vid Stockholms universitet som leder till en kandidatexamen i matematisk statistik.

Innehåll

1 Hudcancer

Hudcancer är en cancersjukdom som kan delas in i flera olika sorter. Termen ”icke-melanom hudcancer” syftar på cancertyper som långsamt utvecklas i huden. De vanligaste typerna är basalcellcancer och skivepitelcancer (NHSb, 2014). De flesta av dessa cancerfall tros orsakas av UVB-strålning och bruk av solarium ökar risken för icke-melanom hudcancer (NHS, 2014a).

2 Problemformulering och syfte

Antalet cancerfall ökar i Sverige och enligt Socialstyrelsen är hudcancer den cancerform som ökar mest. Ökningen sker främst i södra Sverige och Halland är särskilt drabbat (Socialstyrelsen, 2015). Det kan vara av intresse att modellera antalet nya cancerfall för prediktion ur budgetsyvinkel för landsting. I den här uppsatsen har olika regressionsmodeller jämförts för att se vilken som är lämpligast för att modellera antalet nya hudcancerfall.

3 Data och programvara

3.1 Socialstyrelens cancerregister

Datan om cancerfallen har hämtats från Socialstyrelsens cancerregister. Datan hämtades för åren 1970-2014 och innehöll antalet nya cancerfall för män och kvinnor inom olika åldersgrupper. Åldersgrupperna var 0-4 år, 5-9 år, ... , 80-84 år och 85+ år.

3.2 Statistiska centralbyrån - Folkmängd

Folkmängden hämtades från Statistiska centralbyråns statistikdatabas för samma årsperiod. Datafilen från Statistiska centralbyrån är uppdelad i kategorierna ”Ogifta”, ”Gifta”, ”Skilda” och ”Änkor/änklingar” för män och kvinnor under åren 1970-2014. Dessa kategorier har summerats för att ge den totala folkmängden i Halland varje år.

3.3 R

R är en statistisk programvara som kan användas för statistisk analys och modellering. R är ett s.k. ”open source project”, dvs vem som helst kan bidra till projektet (The R Project for Statistical Computing, 2016). Man kan använda diverse paket som finns på hemsidan, dessa kan hämtas från hemsidan CRAN (The Comprehensive R Archive Network). Paketen som använts är *MASS* och de paket som följer med installation av R. *MASS* har använts för modellering av negativ binomialregression, för Poisson- och splineregression följer paketen med installation av R.

4 Metoder

4.1 Generaliserade linjära modeller

Ifall en sannolikhetsfördelning tillhör en exponentiell spridningsfamilj kan sannolikhets-/täthetsfunktionen för en slumpvariabel Y med parameterar θ och ϕ (kallad spridningsparameter som antas vara känd) skrivas på formen som anges i Agresti (2013):

$$f(y;\theta,\phi) = exp(\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi))$$

Generaliserade linjära modeller (hädanefter kallade GLM:s) är modeller som är en generalisering av den allmänna linjära modellen. Den sistnämnda antar normalfördelade och oberoende observationer med feltermer som är normalfördelade med konstant varians. I GLM:s kan våra observationer vara från någon annan fördelningsfamilj tillhörandes en exponentiell spridningsfamilj så länge alla observationerna är från samma slags fördelning, t.ex. Poissonfördelningen och är oberoende (McCullagh & Nelder, 1989). Detta kan sedan användas för regressionsmodellering. En GLM bestäms av tre komponenter som anges nedan.

1. Slumpmässig komponent
2. Systematisk komponent
3. Länkfunktion

Den slumpmässiga komponenten består av slumpvariabler $Y_1, Y_2, ..., Y_n$, där n är antalet observationer. I en GLM tillåts dessa som sagt komma från en fördelning som tillhör den exponentiella spridningsfamiljen (McCullagh & Nelder, 1989). I regressionsarbete är dessa observationerna av vår responsvariabel. Den systematiska komponenten är en linjär prediktor η som är en funktion av våra förklarande variabler. Förhållandet anges nedan där p är antalet förklarande variabler och x_{ij} är värdet av den $j:e$ förklarande variabeln för observation i. Vi samlar alla våra η_i i vektorn $\eta = (\eta_1, ..., \eta_n)$, där n är antalet observationer (Agresti, 2013).

$$\eta_i = \sum_{j=1}^{p} \beta_j x_{ij}$$

Länkfunktionen $g(\mu)$ binder ihop dessa två komponenter via formen $\eta = g(\mu)$, där $g(\cdot)$ är en monoton och differentierbarfunktion och $E(Y) = \mu$ (McCullagh & Nelder, 1989). Ifall vi då har ett stickprov av slumpvariabler $Y_1, ..., Y_n$ kan varje enskild slumpvariabel Y_i med parametrar θ_i och ϕ enligt McCullagh & Nelder (1989) skrivas på formen

$$f(y_i;\theta_i,\phi) = exp(\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi))$$

Sambandet mellan θ_i och η_i kan ta flera former, t.ex. kan vi använda en kanonisk länk, då är $\theta_i = \eta_i$ i fallet för Poissonfördelningen med logaritmisk länk. Fördelningen av våra Y_i används i GLM på så sätt att vi gör regressionen m.h.a. vår länkfunktion. Nämnvärt är att vi inte längre antar homoskedasticitet i en GLM, vilket görs i en vanlig linjär modell. Anledning till detta är att $Var(Y_i) = b''(\theta_i)a(\phi)$, alltså är inte variansen konstant. Detta förhållande kan visas via beräkningar som för enkelhetens skull utelämnas i den här uppsatsen.

4.2 Poissonregression

Poissonregression kan användas för att modellera antal av någon händelse, t.ex. en cancertyp i en viss region (Faraway, 2006). Ett exempel på detta är en rapport om incidensen av icke-melanom hudcancer i östra Skottland (Brewster et al., 2007) där författarna analyserade utvecklingen av icke-melanom hudcancer m.h.a. Poissonregression. Resultaten från denna artikel jämförs med dem från den här uppsatsen i diskussionsavsnittet 6.1. Låt Y vara en Poissonfördelad slumpvariabel med parameter $\mu > 0$, då har vi följande egenskaper för $y = 0, 1, 2, ...$:

$$P(Y = y) = \frac{1}{y!}\mu^y exp(-\mu)$$

$$E(Y) = Var(Y) = \mu$$

Via beräkningar kan man se att Poissonfördelningen tillhör en exponentiell spridningsfamilj, alltså kan vi använda det teoretiska ramverket för GLM för Poissonregression. För Poissonregression kommer vi ha att $\theta_i = \log(\mu_i)$, $a(\phi) = \phi$, $\phi = 1$ och $b'(\theta_i) = \mu_i$. Dessa kommer användas för att ta fram våra parameterskattningar i en regressionsmodell. Regressionen görs då på $\mu = (\mu_1, \mu_2, ..., \mu_n)$ och skrivs ut nedan. I ekvationen nedan betecknar x_{ij} värdet av den $j : e$ förklarande variabeln för observation i. Våra β_m är parametrarna i modellen (Agresti, 2013). För att skatta parametrarna använder vi oss av maximum likelihood-skattarna som fås av funktionen glm() i R m.h.a. metoden Iterated Reweighted Least Squares. Metoden beskrivs i avsnitt 4.6.

$$log(\mu_i) = \sum_{j=1}^{p} \beta_j x_{ij}$$

4.3 Negativ binomialregression

För en Poissonfördelning är $E(Y) = Var(Y)$. Det finns situationer när vi vill modellera antal men variansen verkar vara större än medelvärdet. I en sådan situation kan negativ binomialregression istället vara lämplig. Nämnvärt är att negativ binomialfördelningen endast kan ses som en GLM i specialfall

eftersom vi introducerar en okänd parameter α, se nedan (Hilbe, 2008b). Ett sätt att se på den negativa binomialfördelningen är att det är en Poissonmodell med gammaheterogenitet. Vad detta betyder är att ifall vi har en Poissonfördelad slumpvariabel med parameter λ så har λ en gammafördelning (Faraway, 2006). Denna s.k. "gammablandning" används när vi vill modellera antal av någon händelse och variansen är större än medelvärdet. Sannolikhetsfunktionen, väntevärdet och variansen för en negativ binomialfördelad slumpvariabel Y med parametrar μ och α kan uttryckas som följande (Hilbe, 2008b):

$$P(Y = y) = \frac{\Gamma(y + \frac{1}{\alpha})}{\Gamma(y+1)\Gamma(1\alpha)}(\frac{1}{1+\alpha\mu})^{\frac{1}{\alpha}}(\frac{\alpha\mu}{1+\alpha\mu})^{y}$$

$$E(Y) = \mu$$

$$Var(Y) = \mu + \alpha\mu^2$$

Det bör noteras att även om negativa binomialfördelningen inte nödvändigtvis är en GLM (α måste vara känd för att detta ska gälla) behandlas den som en sådan i R. För att skatta parametrar i en GLM använder R metoden "Iterated reweighted leastsquares" (beskrivs i avsnitt 4.6), men för negativ binomialregression använder R en uppdaterad funktion. Hilbe (2008b) beskriver den uppdaterade funktionen så att en Poissonmodell skattas för att sedan beräkna Pearsons spridningsstatistika, som definieras som

$$\chi^2_p = \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{Var(Y_i)}$$

Detta skattas och kallas ϕ, vilket är vår spridning. Man sätter $\phi = \alpha$ i vår negativ binomialmodell. Efter att denna negativ binomialmodell skattats beräknas en ny Pearson dispersionstatistika, denna gång multipliceras dock ϕ med vår redan erhållna spridning, vilket ger ett uppdaterat värde av ϕ. Processen upprepas tills konvergens för våra skattade parametrar nås, och då lagras värdet av α (Hilbe, 2008b). Nu en återkoppling till teorin om GLM. Vi antar att vi är i ett steg när α har skattats. Det går att visa (via tidskrävande beräkningar) att vi då får $\theta_i = -\log(\frac{1}{\alpha\mu_i+1})$, $b(\theta_i) = -\frac{\log(1+\alpha\mu_i)}{\alpha}$, $a(\phi) = 1$ och $b'(\theta_i) = \mu_i$. (Hilbe, 2008b).

4.4 Splineregression

Formen av beroendet mellan responsvariabeln och en eller flera av våra förklarande variabler kan vara svårbestämt och ett linjärt beroende kan verka orimligt. Ett möjligt alternativ kan vara att använda polynomregression och ett annat alternativ är att använda splines. Idén med splines är att vi har olika beroenden i olika intervall av vår förklarande variabel. I fallet för den här uppsatsen kan vi välja att ha olika polynom av årsvariabeln i vår regressionsfunktion beroende på vilken årsperiod vi betraktar. Exempelvis kan vi ha ett polynom mellan 1970 till 1979, ett annat mellan 1980 och 1989 osv. Punkterna där de olika intervallen

börjar och slutar kallas knutar (Fahrmeir et al., 2013).

I den här uppsatsen har B-splines använts, men dessa definieras diskuterar vi polynomregression och polynomsplines för att göra B-splines mer lättförståeliga. Informationen har hämtats från Fahrmeir et al . (2013). Vi börjar med att tänka oss en polynommodell $f(z_i) = \gamma_0 + \gamma_1 z_i + ... + \gamma_c z_i^c$ som beskriver effekten av den förklarande variabeln z m.h.a. ett polynom av grad c. Detta förhållande räcker inte alltid till för att ge en bra modell. En första idé är att titta på flera intervall av z_i och bestämma separata polynom för varje intervall. Ett problem här som exemplifieras i Fahrmeir et al. (2013) är att vi inte nödvändigtvis får en slät funktion som hänger ihop i början och sluten av våra intervall. Alltså kan vi introducera kravet att ett polynom i sin ändpunkt ska ha samma värde som det efterföljande polynomet i den punkten (vilket blir startpunkten för det efterföljande polynomet). Kravet blir då att vår funktion ska vara deriverbar $c - 1$ gånger i intervallgränserna. Detta krav gör att vi definierar polynomsplines, som görs i Fahrmeir et al (2013).

En funktion $f : [a, b] \rightarrow R$ är en polynomspline av grad $c \geq 0$ med knutar $a = k_1 < ... < k_m = b$ om följande villkor uppfylls:

1. $f(z)$ är deriverbar $c - 1$ gånger. Fallet $c = 1$ motsvarar att $f(z)$ är kontinuerlig men ej deriverbar, och för $c = 0$ sätter vi inga krav på släthet.

2. $f(z)$ är ett polynom av grad c på intervallen $[k_j, k_{j+1})$, och dessa intervall ges av våra knutar.

För en viss konfiguration av knutar och ett givet gradtal c behöver vi kunna representera vår mängd av polynom. Detta görs m.h.a. basfunktioner, som diskuteras nedan. I den här uppsatsen har som sagt B-splines använts. En kortfattad informell förklarning av B-splines är att vi vill ha en funktion $f(z)$ som definieras m.h.a. styckvisa polynom med kraven om deriverbarhet ovan. Våra basfunktioner skapas då m.h.a. dessa polynom som binds ihop vid knutarna. Detta görs genom att skapa basfunktioner som består av $c + 1$ stycken polynom av grad c som knyts ihop på ett sätt som är kontinuerligt deriverbart $c - 1$ gånger i knutarna. Ifall vi låter m vara antalet knutar, $d = m + c - 1$ och betecknar basfunktionerna som $B_i(z)$ kan vi då uttrycka vår splinefunktion $f(z)$ på följande sätt:

$$f(z) = \sum_{i=1}^{d} \gamma_i B_i(z)$$

Vi kan ange en mer precis definition av B-splines, vilket görs för specialfallen $c = 0$ och $c = 1$. Efter dessa anges definitionen för det generella fallet, vilket inkluderar en rekursiv funktion. Basfunktionen för gradtalet c betecknas som $B_i^c(z)$. I fallet för $c = 0$ använder vi en indikatorvariabel $I(k_j \leq z < k_{j+1})$ som antar värdet 1 om $k_j \leq z < k_{j+1}$ och 0 annars för $j = 1, ..., d - 1$. Indikator-

variabeln för $B_i^1(z)$ definieras på ett liknande sätt med lämpliga byten av index. Den generella formeln anges längst ned bland följande ekvationer (Fahrmeir et al., 2013).

$$B_i^0(z) = I(k_j \leq z < k_{j+1})$$

$$B_j^1(z) = \frac{z - k_{j-1}}{k_j - k_{j-1}} I(k_{j-1} \leq z < k_j) + \frac{k_{j+1} - z}{k_{j+1} - k_j} I(k_j \leq z < k_{j+1})$$

$$B_j^c(z) = \frac{z - k_{j-c}}{k_j - k_{j-c}} B_{j-1}^{c-1}(z) + \frac{k_{j+c} - z}{k_{j+1} - k_{j+1-c}} B_j^{c-1}(z)$$

Nu till en kommentar om knutarna. Hur man väljer att placera ut sina knutar kan variera från situation till situation. Man kan välja att placera ut knutarna likformigt över intervallet, välja knutar baserat på hur scatterplots ser ut eller använda kvantilbaserade plottar. De sistnämnda använder kvantilerna från våra observerade värden $z_1, z_2, ..., z_n$ (n=antalet observationer av z). Man bör även ha i åtanke att vi inte kan klara oss med endast våra valda knutar, för att kunna beräkna B-splines måste vi även definiera $2c$ stycken yttre knutar, där c som bekant är gradtalet i vårt polynom (Fahrmeir et al., 2013). Standardinställningen i R är att använda kvantilerna för vår variabel som inre knutar och randpunkterna av vårt intervall av variabeln som yttre knutar.

Slutligen är det viktigt att påpeka hur splinefunktionen binds ihop med själva regressionsmodellen. Vi föreställer oss ett enkelt exempel, då vi utför Poissonregression med två förklarande variabler X och Y utan samspel. Nedan anges en modell där vi använder splinefunktionen för X. Som vi ser har vi alltså ett linjärt beroende i våra baser i vår splinefunktion, vilket gör splinefunktioner väldigt användbara i vårt fall eftersom vi har en GLM.

$$log(\mu_{ij}) = \alpha + \sum_{k=1}^{d} \gamma_k B_k(X_i) + \beta Y_j$$

4.5 Modeller för frekvensdata

För vissa regressioner som räknar antal kan det vara lämpligt att introducera en offsetvariabel som kan användas för proportioner (McCullagh & Nelder, 1989). Det kan vara lämpligt eftersom bara de absoluta värdena kan vara vilseledande ibland. Ett exempel är ifall någon sitter och fiskar. Ifall personen ifråga fångar 3 fiskar på en timme en dag och 3 fiskar på åtta timmar en annan dag har personen ifråga visserligen fångat lika många fiskar båda dagarna men det gick ändå klart bättre den första dagen. I fiskeexempelt skulle antalet timmar vår fiskare spenderar på att fiska vara vår offsetvariabel. Vi kan skapa en regressionsmodell med responsvariabael antalet fångade fiskar per timme, snarare än bara antalet fångade fiskar med den här offsetvariabeln. Parametriseringen av för en modell med t.ex. Poissonfördelad data och offsetvariabel t och medföljande

regressionsmodell ges nedan. Indexen för de förklarande variablerna är samma som tidigare (Hilbe, 2008b). I den här uppsastsen har folkmängden i Halland ett visst år använts som offsetvariabel.

$$f(y_i;\mu_i) = \frac{1}{y_i!}(\frac{\mu_i}{t_i})^{y_i}\exp(\frac{\mu_i}{t_i})$$

$$log(\mu_i) = \beta_0 + \beta_1 X_{1i} + ... + \beta_p X_{pi} + log(t_i)$$

4.6 Iterated Reweighted Least Squares

Informationen om Iterated Reweighted Least Squares har hämtats från McCullagh & Nelder (1989). Maximum likelihood-skattarna i av våra parametrar $\beta_1, \beta_2, ..., \beta_p$ fås m.h.a. metoden Iterated reweighted least squares, hädanefter förkortad som IRLS. Idén är att vi inte utför en regression direkt på responsvariabeln Y utan snarare på Z. Z är en linjäriserad form av länfunktionen $g(\mu)$. En nämnvärd detalj om våra ursprungliga skattningar är att vi använder själva datan som ursprungliga skattning för $\hat{\mu_0}$. Metoden beskrivs nedan:

1. Låt $\hat{\eta_0}$ vara en ursprunglig skattare av vår linjära prediktor. Det motsvarande anpassade värdet i regressionsmodellen är $\hat{\mu_0}$

2. Vi bildar vår "nya responsvariabel" z_0 enligt följande:
$z_0 = \hat{\eta_0} + (y - \hat{\mu_0})(\frac{\partial \eta}{\partial \mu})_0$, där derivatan av länken evalueras i $\hat{\mu_0}$.

3. Vi skapar vikten $W_0^{-1} = (\frac{\partial \eta}{\partial \mu})_0^2 V_0$, där V_0 är variansfunktionen av Y evaluerad i $\hat{\mu_0}$

4. Vikten används sedan när vi gör en regression av z_0 på våra förklarande variabler $X_1, X_2, ..., X_p$ med vikten W_0. Detta ger en ny skattning av våra β_i som används för att ge en ny skattning av η, denna kallas $\hat{\eta_1}$.

5. Processen upprepas tills konvergens av parameterskattningarna nås.

4.7 Leave one out-cross validation

Informationen i det här avsnittet har hämtats från James et al. (2013). Korsvalidering är ett sätt att bedöma en modells prediktiva förmåga. I det här arbetet har "Leave one out-cross validation" (hädanefter kallad LOOCV) använts. Idén med LOOCV är att vi delar upp vårt givna dataset med observationer i två mängder, en som innehåller endast en enda observation och en del med resten av alla observationer. Det större datasetet används för att anpassa en modell och sedan göra en prediktion för det utelämnade värdet. Det empiriska värdet av den utelämnade observationen kallar vi y_i och det predikterade värdet från vår modell kallar vi $\hat{y_i}$. Vi bildar då $s_i^2 = (y_i - \hat{y}_i)^2$ som är en skattning av vårt

fel. Detta görs för alla våra observationer och kommer ge lika många skattade fel som observationer. Alltså, om vi har n stycken observationer i vårt dataset kommer vi då få $s_1^2, s_2^2, ..., s_n^2$. I det här skedet bildar vi en LOOCV-skattning som är ett medelvärde av alla våra s_i^2, kalla den CV_n som ges av

$$CV_n = \frac{1}{n} \sum_{i=1}^{n} s_i^2$$

Eftersom CV_n är en summa av kvadrerade feltermer vill vi välja den modell med lägst CV_n när vi jämför flera modeller m.a.p. LOOCV. Enligt Stone (1977) är modellval baserat på LOOCV asymptotiskt ekvivalent med modellval baserat på Akaikes Information Criterion som diskuteras i avsnitt 4.8.

4.8 Akaikes Information Criterion

Akaikes Information Criterion (förkortas hädanefter som AIC) är ett jämförelsemått mellan olika modeller för ett givet dataset som kan användas för att bestämma vilken modell som är lämpligast för att beskriva datan. För en given modell M definieras AIC som:

$$AIC(M) = -2L(M) + 2p(M)$$

Där $L(M)$ är den maximerade log-likelihoodfunktionen av modellens parametrar och $p(M)$ är antalet parametrar i modellen. Vi ser att AIC är en kompromiss mellan hur bra modellen passar datan och hur komplex den är. AIC är definierad så att ifall vi jämför två modeller är modellen med lägst AIC att föredra, dvs AIC straffar modeller med många parametrar. AIC kan användas mellan olika slags modeller, t.ex. en Poissonregression och en negativ binomialregression (Agresti, 2013). Det bör påpekas att AIC endast ger en bild av hur mycket bättre eller sämre en modell är jämfört med en annan, den ger inget absolut mått på hur bra modellen ifråga är. Som sagt är modellval baserat på AIC asymptotiskt ekvivalent med modellval baserat på LOOCV (Stone, 1977).

4.9 Devianceresidualer

Residualplottar används för att bedöma huruvida en modell är lämplig för att modellera vår givna data. I fallet för linjär regression plottar man residualerna mot modellens anpassade värden. Detta görs för att kontrollera antagandena om att residualerna har väntevärde 0 och uppvisar homoskedasticitet. För en GLM är situationen knepigare och det finns flera olika residualer man kan använda och flera olika alternativ för vad man kan plotta dessa mot. Man kan t.ex. plotta de valda residualerna mot de skattade värdena av vår linjära prediktor η eller mot modellens anpassade värden. I den här uppsatsen har devianceresidualerna plottats mot modellens anpassade värden. Enl. Faraway (2006) bör man plotta mot den linjära prediktorn, men han nämner även att detta kan vara problematiskt för Poissonregressionsmodeller med små värden på responsvariabeln, så som i vårt fall. Förklarningarna till detta illustreras i slutet av det här avsnittet

m.h.a. två residualplottar, se Figur 4.1 och Figur 4.2.

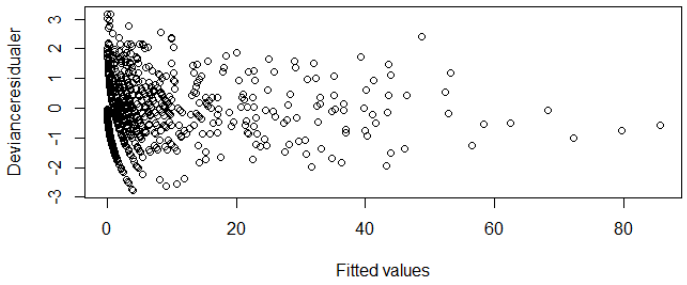
Innan devianceresidualer definieras är det bäst att definiera deviance och illustrera med ett enkelt exempel. Antag att vi har en slumpvariabel Y med sannolikhets-/täthetsfunktion $f(y;\theta) = l(\theta;y)$, där $l(\theta;y)$ är likelihoodfunktionen och $L(\theta;y) = \log(l(\theta;y))$ är log-likelihoodfunktionen. Då definieras deviance som $D = 2(L(\theta;y) - L(\hat{\theta};y))$, där $\hat{\theta}$ är våra parameterskattningar från den valda modellen. Deviance för modellen är alltså ett mått på hur mycket modellen ifråga avviker från den mättade modellen som tilldelar en parameter till varje enskild observation. Definitionen av devianceresidualen r_D och förhållanden den uppfyller anges nedan.

$$r_D = sign(y_i - \hat{\mu}_i)\sqrt{d_i}$$

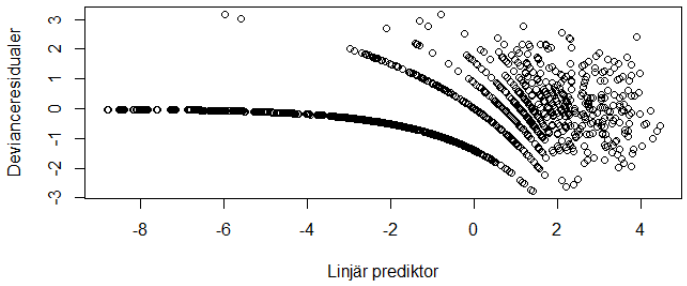
$$\sum r_D^2 = \sum_i d_i = D$$

Alltså är d_i ett mått på hur mycket en enskild observation bidrar till vår deviance. Med hjälp av definitionen av deviance ovan kan vi då få fram utseendet för individuella d_i via lite tidskrävande beräkningar vars utseende kommer bero på vilken fördelning vi har. För t.ex. en Poissonmodell blir då våra d_i på formen $2y_i \log(y_i \cdot \frac{1}{\hat{\mu}_i}) - (y_i - \hat{\mu}_i)$ (McCullagh & Nelder, 1989).

Som förklarning till devianceresidualer plottats mot anpassade värden snarare än den linjära prediktorn kan man betrakta Figur 4.1 och Figur 4.2 som fås av den slutgiltigt valda modellen.



Figur 4.1: Anpassade värden mot devianceresidualer



Figur 4.2: Linjär prediktor mot devianceresidualer

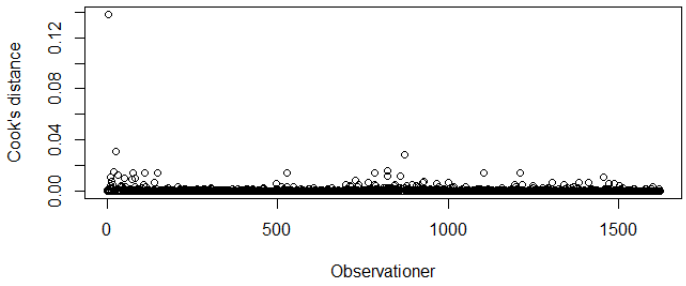
I Figur 4.1 har devianceresidualer plottats mot modellens anpassade värden, i Figur 4.2 har de plottats mot den linjära prediktorn. Den övre plotten visar att residualerna är ojämnt utspridda, vilket försvårar tolkningen. Dock är den nedre också svårtolkad och för alla undersökta modeller var läget detsamma. Under arbetets gång användes figurer av anpassade värden mot devianceresidualer. Det önskade utseendet för en residualplott är detsamma som när vi har en linjär modell, dvs residualplotten för en bra modell ska visa att residualerna är koncentrerade kring värdet 0 för residualerna samt uppvisa homoskedasticitet (Faraway, 2006).

4.10 Cook's Distance

Cook's Distance är ett mått på hur inflytelserik en observation är och kan användas för att identifiera outliers. Definitionen av Cook's distance för en enskild observation ges av

$$D_i = \frac{(\hat{\beta_{(i)}} - \hat{\beta})(X^T X)(\hat{\beta_{(i)}} - \hat{\beta})}{ps^2}$$

där p är antalet parametrar i modellen, $\hat{\beta_{(i)}}$ är våra skattningar för β utan den $i : e$ observationen och s^2 är en skattning av medelvärdet av våra fel i regressionsmodellen (McCullagh & Nelder, 1989). Notera att β här är en parametervekor och X är matrisen innehållandes värdena av våra förklarande variabler. Vad som är en godtagbar storlek går att diskutera och beror på hur situationen ser ut, tröskeln som använts i uppsatsen är värdet 1 som föreslogs av Cook & Weisberg (1982). Ifall man bedömer att en observation är en outlier bör man undersöka den observationen närmare (Faraway, 2006) och undersöka ifall den bör strykas ur modellen. Bland de modeller som undersökts har inga observationer gett ett så stort värde på Cook's distance, därför har inga observationer uteslutits. Ett exempel på en plot av Cook's distance finns i Figur 4.3. Plotten kom från den första negativa binomialmodellen som undersöktes.



Figur 4.3: Cook's distance för negativ binomialmodell

5 Modellering

5.1 Variabler och deras notation

X: år efter 1970, X=0,...,44.

Y: kön, Y=1 för kvinnor och Y=0 för män. Y inkluderas som en kategorisk variabel i regressionen.

Z: åldersgrupp, varje nivå betecknas med medelvärdet av den yngsta och äldsta åldern i varje åldersgrupp. Exempelvis motsvarar beteckningen 2 gruppen 0-4 år, 7 motsvarar 5-9 år osv. Den äldsta åldersgruppen betecknas som 85. Notationen för åldersgrupper ändras under modelleringarbetets gång p.g.a. ihopslagning av vissa åldersgrupper. Z är en kategorisk variabel och värdena 2, 7 osv. är till för att ge en idé om vilka åldersgrupper som inkluderas i en given nivå av variabeln.

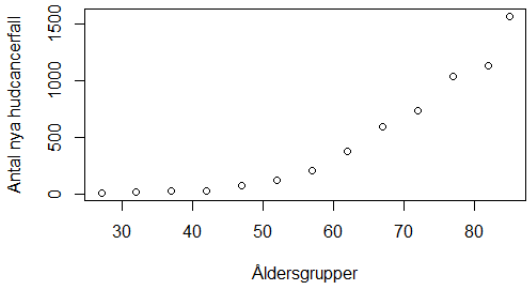
Responsvaribaeln antalet nya hudcancerfall för ett givet årtal, kön och åldersgrupp betecknas som N_{ijk}, där $X = i, Y = j, Z = k$ och $E(N_{ijk}) = \mu_{ijk}$. Nedan ges en modell som endast innehåller huvudeffekterna med offsetvariabel t_i för varje år. λ_k^Z är ett mått på hur mycket större det logaritmerade värdet av responsvariabelns väntevärde är när vi jämför nivå k av variabeln Z med referensnivån och de andra två variablerna hålls konstanta. I det här fallet är det ett mått på hur mycket större eller mindre vi förväntar oss att det logaritmerade väntevärdet av antalet hudcancerfall i åldersgrupp k är jämfört med den yngsta åldersgruppen inom ett kön för ett givet år. För att undvika överparametrisering har parametern med lägst index för åldersgrupper tilldelats värdet 0. Detsamma gäller för alla samspel innehållandes denna åldersgrupp. Ifall vi har en modell där den yngsta åldersgruppen 2 är referensnivån blir då $\lambda_2^Z = \lambda_{i2}^{XZ} = \lambda_{j2}^{YZ} = \lambda_{ij2}^{XYZ} = 0$. För variabeln Y är $\lambda_0^Y = 0$.

$$log(\mu_{ijk}) = \alpha + \beta_1 X_i + \lambda_j^Y + \lambda_k^Z + \log(t_i)$$

5.2 Gemensamma metoder och bedömningsmått

Målet med uppsatsen är som sagt att bestämma vilken regressionsmodell som är lämpligast ur prediktionssyfte. De tre måttstockarna som använts för att jämföra modeller är devianceresidualplottar, AIC och enkelhet. Anledningen till att residualplottarna är med är för att de agerar som en diagnostisk kontroll av modellen. Som sagt visade Stone (1977) att modellval baserat på AIC är asymptotiskt ekvivalent med modellval baserat på LOOCV, så detta ger ett mått av den prediktiva förmågan hos en modell samtidigt som den straffar komplexitet i en modell. Modelleringsarbetet följde samma modell för Poisson- och negativ binomialregression, för splineregression tillämpas andra resonemang som diskuteras i avsnitt 5.5. Av att titta på datan kunde man se att det var väldigt få cancerfall i vissa åldersgrupper (se Figur 5.1), så provades att slås ihop

till en stor grupp i Poisson- och negativ binomialregression. I splineregressionerna valdes den färdiga gruppindelningen från de slutgiltiga Poisson- respektive negativ binomialmodellerna. Fler möjliga gruppihopslagningar jämfördes m.a.p. AIC och residualplottarnas utseende. Idéerna till dessa gruppihopslagningar var detsamma som för den stora gruppen, dvs att vi undersöker om vi kan slå ihop grupper där antalet cancerfall verkar vara ungefär lika stora. Tabeller med dessa AIC-värden finns i Appendix A och ett exempel av en residualplott finns i Figur 4.1. Detta gjordes för modellen innehållandes huvudeffekter, tvåfaktorsamspel och trefaktorsamspel. En detalj ang. notationen av ihopslagna grupper: Ifall åldersgrupperna m upp till k slås ihop (där $k > m$) till en enda grupp får denna nya grupp indexet $\frac{m+k}{2}$.



Figur 5.1: Antal cancerfall i de olika åldersgrupperna

Under arbetets gång framgick det att residualplottarna inte skiljde sig nämnvärt mellan olika modeller. I Figur 3.1 redovisas residualplotten för den slutgiltigt valda modellen och i avsnitt 6.2 diskuteras huruvida den ser ut som vi hade önskat. Ingen av Cook's distance-plottarna visade observationer med ett Cook's distance-värde nära 1, så inga observationer undersöktes närmare. På grund av detta redovisas endast en sådan plot i uppsatsen (se Figur 4.3).

I både Poisson- och negativ binomialregression visade det sig att ifall åldersgrupper 2-22 slogs ihop fick vi lägst AIC när vi ville skapa en grupp där hudcancer är ovanligt. Denna grupp tilldelades kallades då grupp 12. Vi såg även att om grupperna 27-37 slogs ihop sänktes AIC i båda fallen. Den resulterande gruppen kallas då 32. Där slutar dock likheterna, då det visade sig att i Poissonregression gav fler ihopslagningar ett högre AIC medan i negativ binomialregression sänktes AIC av att slå ihop grupp 77 och 82 till en kategori som kallas 79.5. Dessa gruppindelningar användes då i en modell innehållandes huvudeffekter, alla tvåfaktorsamspel och även trefaktorsamspelet. Alltså hade modellerna utseendet

$$\log(\mu_{ijk}) = \alpha + \beta_1 X_i + \lambda_j^Y + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ} + log(t_i)$$

där $i = 0, ..., 44, j = 0, 1$ och $k = 12, 32, 42, ..., 85$ (i fallet för Poisson) och $k = 12, 32, 42, ..., 79.5, 85$ (i fallet för negativ binomial). $\log(t_i)$ är det logaritmerade värdet av folkmängden i Halland år i, vilket är vår offsetvariabel. En tabell innehållandes AIC-värden av olika gruppindelningar för att skapa vår referensgrupp (den där hudcancer är ovanligt) redovisas i avsnittet för Poissonregression (se Tabell 5.1), resterande tabeller finns i Appendix A.

Efter gruppindelningen undersöktes icke-linjära sampand mellan vår responsvariabel och variabeln X, vilket var antalet år efter 1970. Detta gjordes på två olika sätt. För Poisson- och negativ binomialregression adderades en term av X med stigande potens åt gången, dvs för X^2, därefter X^3, X^4 osv. I båda fallen minimerades AIC när den högsta potensen var 7. Ett alternativ till denna metod är att använda sig av regressionsplines i Poisson- och negativ binomialmodeller, och för dessa ändrade vi helt enkelt antalet knutar i splinefunktionen för X. Efter beroendet mellan responsvariabeln och X hade valts undersöktes ifall samspel kunde strykas ur modellen. För att välja den bästa modellen av de olika kandidaterna betraktas AIC och residualplottarna för de undersökta modellerna. Tabeller innehållandes AIC-värden för olika potenser finns i Appendix A. För residualplotten av den bästa modellen, se Figur 4.1. Som sagt presenteras endast den residualplotten eftersom de andra liknade den såpass mycket att man inte kan urskilja en klar vinnare bland modellerna baserat på plottarna.

Både för Poisson- och negativ binomialmodellerna undersökte vi alltså olika modeller med utseende som anges nedan och samma index som ovan. l är här ett heltal som uppfyller $l \geq 1$ och $p_l(X)$ är ett polynom med grad l av variabeln X. Varje potens av X_i har en tillhörande koefficient som kan skattas. Under ekvationen för Poisson-/negativ binomialmodellen finner vi det generella utseendet för splinemodellerna, där m är antalet knutar och $s_m(X)$ är splinefunktionen av X med m knutar. Anledningen till att dessa inte innehåller något trefaktorsamspel diskuteras i avsnitt 5.5.

$$\log(\mu_{ijk}) = \alpha + p_l(X_i) + \lambda_j^Y + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ} + \log(t_i)$$

$$\log(\mu_{ijk}) = \alpha + s_m(X_i) + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \log(t_i)$$

5.3 Poissonregression och resultat

Som sagt visade det sig att den gruppindelning som gav lägst AIC för Poissonregression blev ihopslagningen grupp 2-22 tillsammans som kallas grupp 12, grupp 27-37 tillsammans som kallades grupp 32 och resten separata. I tabell 5.1 redovisas AIC för olika modeller som jämfördes för att skapa en grupp där hudcancer är ovanligt, resterande tabeller finns i Appendix A. För att testa högre

ordningens termer av X visade det sig att modellen med högsta potens 7 gav lägst AIC (se tabell 7.1), så denna valdes som modell med slutgiltigt årsberoende för Poissonregression. Att utesluta parametrar höjde AIC-värdet, så den slutgiltiga modellen blev alltså den med gruppbeskrviningen ovan, högsta potens 7 för årsvariabeln och samtliga samspel. Modellens utseende presenteras senare i stycket. I tabell 5.1 visas som sagt AIC-värden för den första ihopslagningen medan övriga tabeller presenteras i Appendix A. Samma slags tabeller finns för negativ binomialregression i Appendix A.

Tabell 5.1: AIC och ihopslagning av grupper för att ge en grupp innehållandes de åldersgrupper där hudcancer är ovanligt

Gruppindelning	AIC
Alla separata	3791.3
2-7 ihop	3784.7
2-12 ihop	3782.6
2-17 ihop	3777.7
2-22 ihop	3772.1
2-27 ihop	3778.6
2-32 ihop	3798.3
2-37 ihop	3815.1
2-42 ihop	3840.7

Som vi ser minimeras AIC av att slå ihop grupperna 2-22 till en stor grupp där hudcancer är ovanligt. Som sagt visar det sig att AIC sänks ytterligare av att slå ihop 27-37 till en grupp. När vi provar att lägga till termer av X med högre potens än 1 och jämför dessa med avseende på AIC märker vi att modellen vars term med högsta potens är 7 minimerar AIC-värdet, vilket är 3657.6. Härnäst vill vi undersöka ifall någon av samspelstermerna kan strykas från modellen. Samtliga modeller med ett eller flera uteslutna samspel höjer AIC, t.ex. höjs AIC till 3674.3 ifall trefaktorsamspelet utesluts. En tabell innehållandes AIC-värden i modeller där ett visst samspel utesluts finns i Appendix A.

Som sagt var residualplottarna för de olika modellerna väldigt lika varandra och det går inte att säga att en modell skulle vara märkbart bättre än en annan baserat på dessa. Ifall vi betecknar ett polynom av X med grad l som $p_l(X)$ har vår slutgiltiga modell har alltså utseendet

$$log(\mu_{ijk}) = \alpha + p_7(X_i) + \lambda_j^Y + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ} + log(t_i)$$

för $i = 0, ..., 44, j = 1, 0$ och $k = 12, 32, 42, ..., 85$. Varje potens av X_i har en tillhörande koefficient som kan skattas. Det visade sig att alla huvudeffekter var signifikatna på 5%-nivån (här inkluderas alla potenser av årsvaribeln X).

Tvåfaktorsamspelen visade sig i regel inte vara signifikanta på den nivån, det enda undantaget var samspelet mellan grupp 47 och kön. Samtliga trefaktorsamspel förutom de två för grupp 42 och grupp 85 var signifkanta, dock hade dessa två p-värden mycket nära 0.05. Parameterskattningarna tolkas ej eftersom en annan modell valdes som bästa modell.

5.4 Negativ binomialregression och resultat

För negativ binomialregression blev resultaten lite annorlunda. Precis som för Poissonregression gavs lägst AIC av att slå ihop grupperna 2 till 22 som grupp där hudcancer är ovanligt. Den resulterande gruppen kallas samma sak som i Poissonfallet, dvs den kallas 12. Sedan visade det sig att ifall grupp 27-37 slogs ihop sänktes AIC ytterligare (även här kallas den resulterande gruppen 32), men här slutar likheterna för gruppindelning. Ifall vi slår ihop grupp 77 och 82 sänktes AIC lite mer (gruppen kallas 79.5), alltså blev den slutgiltiga gruppindelningen lite annorlunda. Som sagt följde modelleringsmetodiken samma idé som för Poissonregression, så tabellerna presenteras ej i texten utan i Appendix A.

För årstermerna provades samma metodik som för Poissonregressionen, dvs termer av variabeln X med ökande potens lades till i modellen. AIC minskade upp till potensen 7, därefter höjdes den för de två nästföljande. Det slutgiltiga årsberoendet blev alltså ett polynom av grad 7 med AIC=3633.3. Härnäst vill vi undersöka ifall någon av samspelstermerna kan strykas från modellen. Precis som i fallet för Poissonregression höjs AIC när en eller flera samspel utesluts. Att utesluta trefaktorsamspelet höjde AIC till 3643.6, att stryka samspelet mellan X och Y gav AIC=3700.8, ifall samspelet mellan Y och Z höjde AIC till 3682.6 och ifall samspelet mellan X och Z utesluts höjs AIC till 3682.6. Den slutgiltiga modellen hade AIC-värdet 3633.3 och såg ut som följande:

$$log(\mu_{ijk}) = \alpha + p_7(X_i) + \lambda_j^Y + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ} + log(t_i)$$

Såsom i fallet för Poissonregressionen var inte residualplottarna för de olika modellerna märkbart annorlunda, så dessa inkluderas inte. Modellens slutgiltiga utseende anges ovan. Även här var alla huvudeffekter signifikanta på 5%-nivån, inkluderande de olika potenserna av årsvariabeln X. Inte heller här verkar tvåfaktorsamspelen ha någon signifikant effekt, dock finns enstaka undantag. Situationen med trefaktorsamspelens signifikans liknar den i Poissonregressionen, nu var dock trefaktorsamspelet innehållandes den äldsta åldersgruppen signifikant. En nogrannare tolkning av resultaten görs ej då en annan modell väljs som slutgiltig.

5.5 Splineregression och resultat

I den här uppsatsen används splineregression i både Poisson- och negativ binomialmodeller. I båda fallen valdes startpunkten där gruppindelningen var klar, därefter varieras antalet knutar i splinefunktionen för X. Denna betecknas som $s_m(X)$ när vi har m knutar. Ursprungsmodellernas utseende diskuterades i avsnitt 5.2.

En nogrannare igenomgång för Poisson- och negativ binomialfallen anges i följande stycken. För att börja experimentera med splines användes en knut till att börja med i båda ursprungsmodellerna. Senare i modelleringsarbetet provades att öka antalet knutar. Knutselektion gjordes via R, som använder kvantilerna av X som knutar. För exempelvis en knut används då 50%-kvantilen som knut, men för 9 knutar används 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80% och 90%- kvantilerna av årtalen som knutar. Vid det här laget kan det vara instruktivt att minnas splinefunktionernas utseende, vilket diskuterades i avsnitt 4.4. Standardinställningen i R är att skapa polynom av grad 3. Beroende på antal knutar kommer våra knutar vara annorlunda. Beräkning av basfunktionernas exakta utseende kan göras via tidskrävande uträkningar som använder formlerma ovan och de valda knutarna för en viss modell.

Vi börjar med Poissonregression. Ursprungsmodellen led av problemet att vissa anpassade värden av responsvariabeln blev 0, så en modell utan trefaktorsamspel undersöktes. Denna hade AIC-värdet 3736.8 utan det ovan nämnda problemet. Därefter provades att ändra antalet knutar i huvudeffekten, och det visade sig att det optimala antalet knutar var 8, vilket gav AIC=3661.6. En tabell innehållandes AIC-värdena för olika antal knutar presenteras i Appendix A. Efter antalet knutar valts undersökte vi ifall några av tvåfaktorsamspelen kunde uteslutas ur modellen. Det visade sig att modellen blev väsentligt sämre av att utesluta samspel. Att ta bort samspelet mellan X och Y höjde AIC till 3724.5, att ta bort samspelet mellan X och Z höjde AIC till 3685.7 och att ta bort samspelet mellan Y och Z höjde AIC till 3720.5. Den slutgiltiga modellen för Poissonbaserad splineregression blev alltså den med splinefunktionen $s_8(X)$, AIC=3661.6 och utseendet

$$log(\mu_{ijk}) = \alpha + s_8(X) + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + log(t_i)$$

Ursprungsmodellen för negativ binomialfallet hade samma problem som i Poissonfallet. Återigen uteslöts trefaktorsamspelet vilket åtgärdade problemet och gav AIC=3684.4. Att öka antalet knutar i huvudeffekten sänkte AIC och precis som i Poissonfallet visade det sig att det optimala antalet knutar var 8, vilket gav AIC=3634.9. En tabell innehållandes AIC-värdena för olika antal knutar anges i Appendix A. Därefter provade samspelstermer att uteslutas. Återigen visade det sig att modellen blev sämre av detta. Att ta bort samspelet mellan X och Y höjde AIC till 3692.4, att ta bort samspelet mellan X och Z höjde AIC

till 3641.7 och att ta bort samspelet mellan Y och Z höjde AIC till 3674.5. Den slutgiltiga modellen hade alltså AIC=3634.9, 8 stycken knutar och utseendet

$$log(\mu_{ijk}) = \alpha + s_8(X) + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + log(t_i)$$

5.6 Jämförelse och slutgiltig modell

En tabell innehållande AIC-värdena för de fyra slutgiltiga modellerna från varje regressionstyp presenteras i Tabell 5.2.

Tabell 5.2: AIC och slutgiltiga modeller

Modell	AIC
Poisson	3657.6
Poisson-spline	3661.6
Negativ binomial	3633.3
Negativ binomial - spline	3634.9

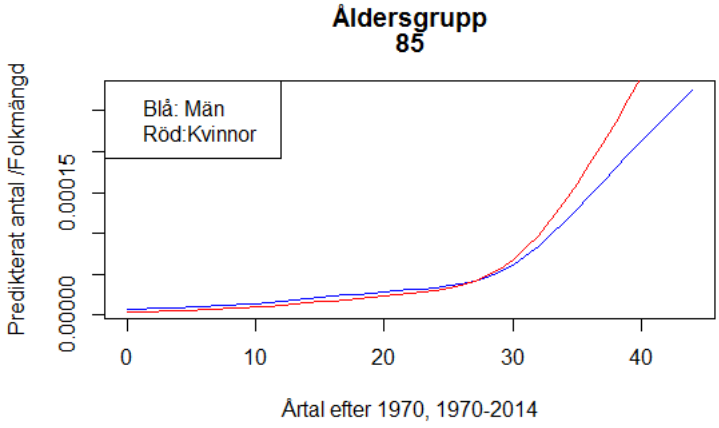
Vi kan se att negativ binomialmodellerna är bättre än Poissonmodellerna och att negativ binomial utan splinefunktionen ger snäppet lägre AIC än den med. Skillnaden är dock bara 1.6 och en tumregel som används är att ifall skillnaden i AIC är mindre än 2 kan modellerna anses vara likvärdiga (Cavanaugh, 2012). Dock är splinemodellen lite enklare eftersom den saknar trefaktorsamspel. Även om vi främst var intresserad av prediktion är enkelhet en önskvärd egenskap hos en modell. Det bör dock påpekas att även denna modell är svårtolkad och man rekommenderas inte försöka tolka tvåfaktorsamspelens parameterskattningar (Hilbe, 2008a). Hur vi väljer att försöka tolka modellen diskuteras i diskussionsavsnittet. I och med att residualplottarna som sagt har ungefär samma utseende väljs då negativ binomialregressionen med splinefunktion som slutgiltigt val. Resultaten från denna modell jämförs med existerande forskning i diskussionsavsnittet.

6 Diskussion

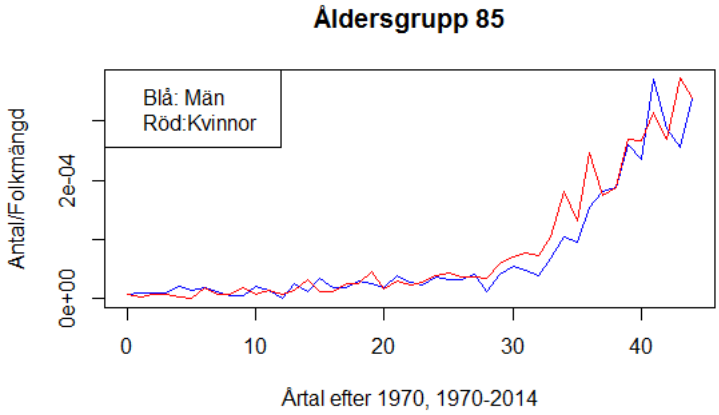
6.1 Tolkning av slutgiltig modell

I negativ binomialregression gör vi som sagt regression för logartimen av responsvariabelns väntevärde. I ett enkelt fall med t.ex. endast en numerisk förklarande variabel X med parameter β får vi då veta hur mycket det logaritmerade väntevärdet av antalet hudcancerfall ändras när X ökar med en enhet. I en modell med t.ex. två stycken kategoriska variabler utan samspel är tolkningen lite annorlunda. En parameterskattning säger då hur mycket större det logaritmerade väntevärdet är för ena kategorinivån än referensnivån givet att alla andra förklarande variabler hålls konstanta. I vårt fall kommer inte parameterskattningarna för den slutgiltiga modellen tolkas av två anledningar. För att förstå hur t.ex. λ_j^Y ska tolkas jämför vi skillnaden mellan män och kvinnor givet att X och Z antar värdet 0. I vårt fall får vi alltså endast information om skillnader mellan män och kvinnor det första året i den yngsta åldersgruppen, vilket inte är särskilt informativt. En liknande situation råder för våra λ_k^Z. Som sagt är även tvåfaktorsamspelens parameterskattningar svårtolkade eftersom vi har med samtliga möjliga tvåfaktorsamspel och man rekommenderas inte att försöka tolka dessa enligt Hilbe (2008a). Ett försök att förstå modellen görs istället grafiskt.

För att förstå effekten av kön kan vi studera Figur 6.1 som visar vår släta splinefunktion. Vi kan se att enligt den ökar antalet nya hudcancerfall varje år mer för kvinnor med tiden än för män i den äldsta åldersgruppen. Samma mönster upprepade sig i de flesta andra åldersgrupper, om än något svagare. I Figur 6.2 har de empiriska värdena plottats för att illustrera en svaghet med modellen, nämligen att den underskattar det totala fallet hudcancerfall. Den svagheten syns även i samma slags plottar för andra, men inte alla, åldersgrupper. Endast denna presenteras för att illustrera detta. Vad detta kan bero på diskuteras i avsnitt 6.2.



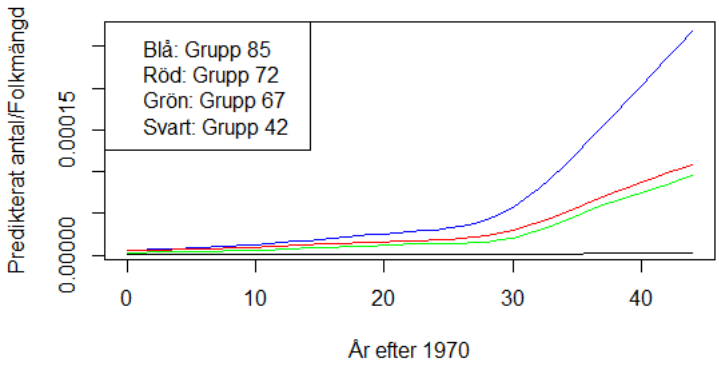
Figur 6.1: Anpassade värden av antalet nya hudcancerfall för män och kvinnor i åldersgrupp 85



Figur 6.2: Empiriska värden av antalet nya hudcancerfall för män och kvinnor i åldersgrupp 85

För att få en idé om åldersgruppernas påverkan vänder vi oss till Figur 6.3 som visar att läget är allvarligare för de äldre åldersgrupperna. I grupp 42 ser vi knappt någon ökning, vilket även gäller för grupp 12 och 32. Bland åldersgrupperna som är äldre än dessa ser vi dock en klar ökning med tiden, som sagt är allvarligare för de äldre åldersgrupperna. För övriga åldersgrupper som inte inkluderas i figuren är trenden densamma. Vi kan alltså se att antalet nya hudcancerfall ökar med tiden i vår slutgiltiga modell, vilket gäller för både

män och kvinnor och de flesta åldersgrupper.



Figur 6.3: Anpassade värden av antalet nya hudcancerfall åldersgrupperna 85, 72, 67 och 42

Resultaten från analysen kan jämföras med forskning som finns inom området, då man bl.a. analyserat hur antalet nya hudcancerfall har utvecklats i Skottland m.h.a. Poissonregression (Brewster et al., 2007). En artikel från ett land som ligger nära Sverige har valts så klimatskillnader inte kan tänkas vara alltför stora. I uppsatsen visar det sig att hudcancer ökat som mest i de äldre åldersgrupperna, vilket stämmer överens med resultaten från vår slutgiltigt valda modell. Även Brewster et al. (2013) observerade en något större ökning i tiden för kvinnor jämfört med män, vilket som sagt illustreras i Figur 6.1.

6.2 Begränsningar och möjliga förbättringar

En möjlig svaghet med analysen är att den allra yngsta åldersgruppen som endast innehöll nollor inkluderats. Som sagt kan det stora antalet nollor dämpa p-värden märkbart (Hilbe, 2008b) och leda till att vissa förklarande variabler verkar ha en signifikant effekt trots att de egentligen inte är signifikanta. I och med att det är såpass ovanligt med hudcancer i de yngre åldersgrupperna kan det diskuteras huruvida dessa nollor är slumpmässiga eller systematiska, dvs att denna åldersgrupp aldrig kommer innehålla några fall av icke-melanom hudcancer. Eftersom de flesta län inte hade data för denna åldersgrupp är det vågat att dra slutsatsen att dessa var systematiska, så de inkluderades. Ifall man undersökt fler regioner och inte observerar några cancerfall i den åldersgruppen någonstans kan det dock vara lämpligt att pröva modeller som har systematiska nollor i åtanke, t.ex. ”zero inflated poisson regression” och ”zero inflated nega-

tive binomial regression".

För att få en uppfattning om hur bra modellen är kan vi jämföra de anpassade med de empiriska värden. För att få en bild av modellens prediktiva förmåga kan man utesluta vissa observationer t.ex. de sista åren i vårt dataset, anpassa en modell på det mindre datasetet och sedan predikterar värden för dessa år. Ifall de predikterade värdena stämmer väl överens med de empiriska har modellen god prediktionsförmåga. I vårt fall behöver vi inte ens göra det för att upptäcka brister i modellen, som illustreras av att jämföra Figur 6.1 med Figur 6.2. Här jämförs endast figurer för den äldsta åldersgruppen. Vi kan se att modellen verkar underskatta antalet nya hudcancerfall. Underskattningen syns även i vissa andra åldersgrupper. En möjlig förklaring till detta är att datamängden modellen anpassades efter innehöll väldigt många nollor och små responsvärden. Eftersom modellen inte kan prediktera den empiriska datan som den baserades på drar vi slutsatsen att även om den var bäst bland de undersökta är den ändå bristfällig. En möjlig anledning till underskattningen är att vi som sagt inkluderat den yngsta åldersgruppen som bekant endast innehöll värdet 0 för responsvariabeln. Som sagt är modellen även svårtolkad m.a.p. parameterskattningar. Ett annat sätt att välja modeller hade kunnat göra att vi slapp det problemet. Även om AIC straffar komplexitet finns andra mått som straffar komplexitet mer, t.ex. BIC (Bayesian Information Criterion). En möjlig förändring kan var att göra modellval baserat på BIC kombinerat med LOOCV.

En möjlig vidareutveckling av arbetet skulle kunna vara att inte endast undersöka Halland utan även närliggande län för att få en bild av cancerutveckling i sydvästra Sverige. I en sådan uppsats kan det vara möjligt att undersöka hur cancerincidensen i Halland skiljer sig från andra län ifall man introducerar en kategorisk variabel för län.

6.3 Slutsats

Som sagt var negativ binomialmodellerna bättre än Poissonmodellerna och splinemodellens AIC-värde var inte särskilt mycket högre än den utan splines och bristen av trefaktorsamspel talar för splinemodellen i negativ binomialfallet. Ifall vi undersöker residualplottarna för de slutgiltiga modellerna ser vi att de liknar varandra, så baserat på dessa är det svårt att säga att en av dem skulle vara den klart bästa. Överlag var alltså splinemodellen baserad på negativ binomialregression den bästa. Det bör dock påpekas att residualplotten inte visar det utseende vi helst hade kunnat önska oss. Att tolka dessa plottar kan dock som sagt vara svårt eftersom att vi har många observationer med få värden (Faraway, 2006), men överlag ser inte modellen ut att passa datan så bra som vi hade önskat. Samtidigt verkar modellen underskatta antalet nya hudcancerfall en aning, vilket är en nackdel för prediktion. Slutsatsen är att även om en negativ binomialmodell med splines var bäst bland de undersökta modellerna har den sina brister.

7 Appendix - Tabeller för regressioner

7.1 Poissonregression

Tabell 7.1: AIC för olika potenser av variabeln X

Högsta potens	AIC
2	3728.8
3	3726.1
4	3684.9
5	3671.1
6	3670.4
7	3657.6
8	3659.3

Tabell 7.2: AIC för ihopslagning av flera åldersgrupper

Gruppindelning	AIC
27-57 ihop	4208.8
27-52 ihop	3969.8
27-47 ihop	3852.8
27-42 ihop	3769.2
27-37 ihop	3768.2
27-32 ihop	3772.1
42-47 ihop	3794.3
47-52 ihop	3771.7
52-57 ihop	3786.9
77-82 ihop	3773

Tabell 7.3: Modeller där samspel stryks ett åt gången

Utesluten parameter	AIC
λ_{ijk}^{XYZ}	3674.3
λ_{ij}^{XY}	3737.2
λ_{ik}^{XZ}	3733.2
λ_{jk}^{YZ}	3698.3

7.2 Negativ binomialregression

Tabell 7.4: AIC för ihopslagning av grupper för att ge en grupp innehållandes de åldersgrupper där hudcancer är ovanligt

Gruppindelning	AIC
Alla separata	3737.4
2-7 ihop	3730.7
2-12 ihop	3728.6
2-17 ihop	3723.7
2-22 ihop	3718.1
2-27 ihop	3724.6
2-32 ihop	3744.2
2-37 ihop	3760.7
2-42 ihop	3786

Tabell 7.5: AIC för ihopslagning av flera åldersgrupper

Gruppindelning	AIC
27-57 ihop	4107.8
27-52 ihop	3902.8
27-47 ihop	3794.3
27-42 ihop	3714.9
27-37 ihop	3714.1
26-32 ihop	3716.7
42-47 ihop	3737.8
47-52 ihop	3716.9
52-57 ihop	3729
77-82 ihop	3714

Tabell 7.6: AIC för olika potenser av variabeln X

Högsta potens	AIC
2	3681.3
3	3682.6
4	3656.3
5	3641.9
6	3643.4
7	3633.3
8	3635.2
9	3633.7

Tabell 7.7: AIC och uteslutna samspel

Utesluten parameter	AIC
λ_{ijk}^{XYZ}	3643.6
λ_{ij}^{XY}	3700.8
λ_{ik}^{XZ}	3682.6
λ_{jk}^{YZ}	3649.8

7.3 Splinegressioner

Tabell 7.8: AIC och antal knutar i Poissonspline

Antal knutar i $s_i(X)$	AIC
1	3736.8
2	3740.1
3	3716.9
4	3670.8
5	3669.3
6	3675.5
7	3664.9
8	3661.6
9	3667.9
10	3664.8

Tabell 7.9: AIC och antal knutar i negativ binomialspline

Antal knutar i $s_i(X)$	AIC
1	3684.4
2	3685.9
3	3675.5
4	3641.4
5	3638.9
6	3644.8
7	3637.5
8	3634.9
9	3640.5
10	3638.4

8 Appendix B - Modellutskrift

Signifikansnivå 0.001 betecknas ***, 0.01 ** och 0.05-nivån av *. Övriga är osignifikanta. Tabell 8.1 ges på nästa sida.

Tabell 8.1: Utskrift av summary() från R för den slutgiltiga modellen

Parameter	Skattning	Standardfel
α	-19.324^{***}	1.744
λ^{Z}_{32}	3.837^{**}	1.797
λ^{Z}_{42}	4.736^{***}	1.815
λ^{Z}_{47}	4.427^{**}	1.800
λ^{Z}_{52}	5.393^{***}	1.768
λ^{Z}_{57}	5.685^{***}	1.757
λ^{Z}_{62}	6.589^{***}	1.746
λ^{Z}_{67}	7.059^{***}	1.742
λ^{Z}_{72}	7.651^{***}	1.740
$\lambda^{Z}_{79.5}$	7.917^{***}	1.737
λ^{Z}_{85}	7.570^{***}	1.739
X, knut 0	0.813	0.503
X, knut 1	1.661^{**}	0.722
X, knut 2	1.898^{**}	0.931
X, knut 3	2.521^{**}	1.153
X, knut 4	2.433^{*}	1.374
X, knut 5	3.978^{**}	1.597
X, knut 6	4.218^{**}	1.752
X, knut 7	5.231^{**}	2.193
X, knut 8	4.523^{**}	1.912
λ^{Y}_{1}	-1.976^{**}	0.827
$\lambda^{YZ}_{1,32}$	1.658	0.882
$\lambda^{YZ}_{1,42}$	1.714	0.909
$\lambda^{YZ}_{1,47}$	1.113	0.853
$\lambda^{YZ}_{1,52}$	1.098	0.842
$\lambda^{YZ}_{1,57}$	1.041	0.833
$\lambda^{YZ}_{1,62}$	0.669	0.828
$\lambda^{YZ}_{1,67}$	0.953	0.825
$\lambda^{YZ}_{1,72}$	0.685	0.824
$\lambda^{YZ}_{1,79.5}$	0.813	0.821
$\lambda^{YZ}_{1,85}$	1.339	0.822
$\lambda^{XY}_{i,1}$	0.024^{***}	0.003
$\lambda^{XZ}_{i,32}$	-0.072	0.048
$\lambda^{XZ}_{i,42}$	-0.079	0.049
$\lambda^{XZ}_{i,47}$	-0.027	0.048
$\lambda^{XZ}_{i,52}$	-0.042	0.047
$\lambda^{XZ}_{i,57}$	-0.034	0.047
$\lambda^{XZ}_{i,62}$	-0.038	0.047
$\lambda^{XZ}_{i,67}$	-0.042	0.047
$\lambda^{XZ}_{i,72}$	-0.050	0.046
$\lambda^{XZ}_{i,79.5}$	-0.048	0.046
$\lambda^{XZ}_{i,85}$	-0.035	0.046

9 Referenser

Agresti, A. 2013. *Categorical data analysis*. New Jersey: John Wiley & Sons.

Brewster,.H., Bhatti, L.A., Inglis, J.H.C., Nairn, E.R. och Doherty, V.R. 2007. Recent trends in incidence of nonmelanoma skin cancers in the Eastof Scotland, 1992–2003, *British Journal of Dermatology*. 156,pp:1295-1300.

Cavanaugh, J.E., 2012. *171:290 Model Selection Lecture XIV: The Application of Model Selection Criteria*[pdf].<`http://www.myweb.uiowa.edu/cavaaugh/ms_lec_14_ho.pdf`>[Användes 5/4-16]

Cook, D.R. och Weisberg, S. 1982.*Residuals and Influence in Regression*. London: Chapman and Hall.

Fahrmeir, L., Kneib, T., Lang, S. och Marx, B. 2013. *Regression - Models, Methods and Applications*. Berlin Heidelberg: Springer-Verlag.

Faraway, J.J. 2006. *Extending the Linear Model with R*. Boca Raton:Chapman & Hall\CRC.

Gareth, J., Witten, D., Hastie, T. och Tibshirani, R. 2013. *An Introduction to Statistical Learning with Applications in R*. Berlin Heidelberg: Springer-Verlag.

Hilbe, J.M. Cambridge University. 2008a. *Brief overivew on interpreting count model risk ratios*[pdf].<`http://courses.statistics.com/count/HILBE_NBR_OVERVIEW_ON_INTERPRETING_RISK_RATIOS.pdf`>[Användes 17/5-16]
Hilbe, J.M., 2008b. *Negative Binomial Regression*. New York: Cambridge University Press

McCullagh, P och Nelder, J.A., 1989. *Generalized Linear Models*. Boca Raton:Chapman & Hall\CRC.

National Health Service, 2014a. *Skin cancer (non-melanoma) - Causes*[online] Tillgänglig vid <`http://www.nhs.uk/Conditions/Cancer-of-the-skin/Pages/Causes.aspx`>[Användes 4/5-2016]

National Health Service, 2014b. *Skin cancer (non-melanoma)*[online] Tillgänglig vid <`http://www.nhs.uk/conditions/cancer-of-the-skin/pages/introduction.aspx`>[Användes 4/5-2016]

Ripley, B., Venables, B., Bates, Douglas M., Hornik, K., Gebhardt, A. och Firth, David. Comprehensive R Archive Network. 2015. *Package 'MASS'*[pdf].<`https://cran.r-project.org`>[Användes 14/2-2016]

The R Project for Statistical Computing. 2016. *R Project*[online].<https://www.r-project.org/about.html>[Användes 19/3-16]

Socialstyrelsen, 2015. *Cancerincidens i Sverige 2014 – Nya diagnosticerade cancerfall år 2014*.[pdf] Tillgänglig vid <https://www.socialstyrelsen.se/Lists/Artikelkatalog/Attachments/20008/2015-12-26.pdf>

Socialstyrelen, 2016. *Statistikdatabas för cancer*[online]<http://www.socialstyrelsen.se/statistik/statistikdatabas/cancer>[Användes 26/1-2016]

Statistiska centralbyrån, 2016-02-22. *Folkmängden efter region, civilstånd, ålder och kön. År 1968 - 2015*[online] <http://www.statistikdatabasen.scb.se/pxweb/sv/ssd/START__BE__BE0101__BE0101A/BefolkningNy/?rxid=e661b55a-\e226-4208-ada1-1ad3536c44cd>[Användes 30/1-2016]

Stone, M. An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike's Criterion. *Journal of the Royal Statistical Society. Series B (Methodological)*.Vol. 39, No. 1., pp:44-47