Fertilitet och mortalitet i 1700- och 1800-talets Skellefteå

En modelleringsstudie

Lina Boberg

Lina Boberg
Vt 2010
Magisteruppsats i statistik, 15 hp
Statistikerprogrammet
Handledare: Erling Häggström Lundevaller

Fertility and mortality in 1700- and 1800's Skellefteå

A modeling study

Lina Boberg

Sammanfattning

Uppsatsen behandlar fertilitet och spädbarnsdödlighet på olika sätt under en period av Skellefteås historia. För att göra det användes data från folkbokföringen som tillhandahållits av Demografiska Databasen. Dels görs försök att modellera antal barn en moder får utifrån variabler om hennes levnadshistoria, med hjälp av olika varianter av generaliserade linjära modeller. Uppsatsen undersöker även om en familjs beslut att skaffa flera barn kan bero på huruvida man har en historia av spädbarnsdödlighet, detta med hjälp av logistisk regression. Slutsatserna blir att det går att modellera fertiliteten, men att ingen av de testade modellerna blir särskilt bra då de alla predikterar responsvariabeln dåligt. Inga av de logistiska modellerna får signifikanta resultat för skattningen av spädbarnsdödlighetsvariabeln. Detta gör att det i uppsatsen inte framkommer något samband mellan fortsatt barnskaffande och storleken på spädbarnsdödligheten inom familjen.

Abstract

This essay brings up fertility and infant mortality in various ways over a period of time in Skellefteå's history. To do this, data from the Swedish national registration was provided from the Demographic Data Base. The essay attempts to, by using different types of generalized linear models, model the number of children a mother gets by using variables from her life history. The essay also examines whether a family’s decision to have more children may depend on whether there is a history of infant mortality, by using logistic regression.
The conclusions became that it is possible to model fertility, but that none of the tested models are especially good when all of them predict the response variable poorly. No one of the logistic models get a significant result for the estimation of the variable for infant mortality. Because of this there are no indications in the essay that there is a connection between the size of infant mortality in the family and the decision of getting more children.

Innehållsförteckning

1. Inledning

1.1 Bakgrund

I vissa delar av Sverige har man sedan början på 1600-talet fört kyrkböcker och över hela landet kom det som krav 1686.[1] Kyrkböcker är förteckningar upptagna av en präst eller dylikt ur ens församling över medborgarna i Sverige och dess förehavanden vid olika tidpunkter såsom födelseort och datum, civiltillstånd, församling, när de flyttat samt dödsort och datum.

Dessa har digitaliserats på Demografiska databasen i Umeå[2] och kan nu användas för att göra statistik och undersöka fenomen med.[3] I denna uppsats används kyrkböcker från Skellefteåregionen från 1600-1900-talet, men mest fokus på perioden 1700- och 1800-talet. Regionen bestod då av ett fåtal församlingar runt själva staden. Befolkningen var i stor utsträckning bönder då det var ett bördigt landskap, förutom vid kusten där mindre industrier och även fiske förekom. Under regionens tid i registret skiljer sig sammansättningen åt ganska rejält. Dock fanns mellan 1721 och 1834 de församlingar som nu är Norsjö, Jörn, Skellefteå, Bureå och Byske med i regionen. [4] På grund av att många olika orsaker så som bränder, mänskliga faktorn och ovetskap är registren inte kompletta. Vissa av variablerna har många saknade värden och andra kan vara relativt kompletta, hur stor del av det använda materialet som är saknade värden beskrivs för vardera variabel i variabelkapitlet.

1.2 Frågeställning och syfte

Syftet med uppsatsen är att undersöka fertilitet och mortalitet i Skellefteå under 1700- och 1800-talet. Detta görs genom att undersöka om den totala fertiliteten kan modelleras med hjälp av variabler kring moderns levnadshistorik, att se efter hur mönstret för spädbarnsdödlighet och antal barn förändras över tid samt hur spädbarnsdödlighet påverkar barnafödslar inom familjen under vissa villkor. I dessa delar används olika varianter av generaliserade linjära modeller och data från Demografiska Databasen vid Umeå Universitet som är en inläsning av kyrkböcker från församlingar runt Skellefteå mellan 1615-1901.

Frågeställningarna i uppsatsen är alltså att

- Kan man modellera fertilitet med någon typ av räknedatafördelning?

[1] Skatteverket, hemsida
[2] Demografiska Databasen, hemsida
[3] Skatteverket, hemsida
[4] Demografiska Databasen, hemsida

- Finns det något gemensam mönster för antalet barn och spädbarnsdödligheten över tid?
- Påverkar historien av spädbarnsdödlighet inom en familj huruvida man skaffar flera barn?

I uppsatsen kommer kapitel 2 att ta upp variabler och hur de behandlas i uppsatsen, kapitel 3 att behandla de metoder och den teori som används under uppsatsens gång. Kapitel 4 beskriver resultatet av metoderna och kapitel 5 diskuterar detsamma. Kapitel 6 tar upp vilka referenser som använts och i kapitel 7 finns appendix.

2. Variabler och urval

Detta kapitel är uppdelat i två avsnitt. Det första beskriver det data som finns tillgängligt från början och det andra beskriver de transformationer och urval som har gjorts på vissa variabler för att göra om materialet till det som behövs för frågeställningarna.

2.1 Originalmaterial

Från början fanns ett stort antal variabler registrerade för varje individ. Dessa variabler visade saker inom olika kategorier med exempel inom parantes:

- Information om föräldrarna och dess historia (personnummer, yrke, socialgrupp)
- Vart barnet var fött (församling och exakt ort)
- Detaljer kring barnets födsel (flerbörd, placering i skaran, exakt datum, kön)
- Om barnet överlevde vissa tidpunkter och när det dog

Dessa olika variabler var indelade i flera olika dataset varav två användes under analysen. Det ena visar varje individ som bodde i Skellefteå under den aktuella tidsperioden med totalt 90 050 individer och den andra visar information om varje barn en kvinna fick. Ur det första datasetet fanns följande variabler som användes:

20år är en indikatorvariabel som visar om modern har bott i samma församling under minst 20 år, från 18 års ålder eller från att den flyttat in om den då var äldre än 18 år. Den har inga saknade värden då okända värden ses som obekräftade att de skulle ha bott i 20 år.

Församling är uppdelad i tre olika församlingar då dessa är de enda som hittills har blivit digitaliserade och därmed finns att tillgå i registret. I dataseten finns då Jörn, Norsjö och Skellefteå. Inte heller för denna variabel fanns några saknade värden då de är själva grunden för att hamna i kyrkboken. Totalt fanns den största andelen individer, cirka 80 procent, i Skellefteå. Norsjö och Jörn hade cirka 10 procent var.

Kön är uppdelad i tre kategorier. Man, kvinna och okänd. Denna variabel hade inte några saknade värden då även denna har kategorin okänd.

AntalBarn är precis vad det låter som, antalet barn varje individ fick under sin tid i Skellefteå. Denna variabel hade totalt cirka 50 procent saknade värden. Detta är inte orimligt då barnen endast är registrerade på modern och ej på fadern.

Socialgrupp är en variabel som visar vilken socialgrupp som personen maximalt uppnådde under den delen av sin livstid den bodde i Skellefteå. Socialgruppsvariabeln har inga saknade värden registrerade, då kategorin okänt samlar upp dessa individer.

1. Storföretagare, totalt 0,1 procent.
2. Högre tjänsteman, totalt 1,3 procent.

3. Bönder, totalt 49 procent och överlägset den största gruppen.
4. Småföretagare, totalt 0,7 procent.
5. Lägre tjänsteman, totalt 1,2 procent.
6. Kvalificerad arbetare, totalt 4,3 procent.
7. Agrar underklass, totalt 9,1 procent.
8. Okvalificerad arbetare, totalt 7,8 procent.
9. Före detta, totalt 0,8 procent.
10. Okänt, 26 procent den nästa största gruppen.

Det betyder att endast två av de tio grupperna har över 10 % av befolkningen och det är bönder och okänt.

Födelsedatum är en variabel som visar exakt vilket datum en person är född och är skrivet på formen ÅÅÅÅMMDD där Å är årtal, M månad och D datum. Variabeln *födelsedatum* har ungefär 0,3 % saknade värden. Dock kan det finnas problem i variabeln då värden ibland kan vara avrundade till ett nära ”jämnt” datum, t ex sista juni eller sista december då man inte vet när personen är född. Detta ger en osäkerhet i datat fastän det finns registrerade värden.

Dödsdatum är en variabel som visar exakt vilket datum en person är dog och är även den skrivet på formen ÅÅÅÅMMDD där Å är årtal, M månad och D datum. Den har cirka 60 % saknade värden.

Ur dataset två användes följande variabler:

För varje barn en mor fått har det registrerats om barnet har överlevt vissa hållpunkter. Dessa är bland andra födelsen, 8 dagar och 28 dagar. Dessa tre tillfällen har summerats ihop till den dikotoma variabel i registret som berättar om barnet dog vid något av dessa tillfällen. Variabeln i datasetet heter *Spädbarnsdöd* och används för samtliga av barnen, det betyder att totalt används 21 stycken *Spädbarnsdöd*-variabler från datasetet. Det fanns saknade värde på 17 % i den första delvariabeln, det vill säga det saknades registrering för 17 % av de förstfödda barnen.

Många av de variabler som fanns i dataseten berättar ungefär samma saker. T ex finns det förutom *20år* även en variabel som berättar om man bodde i församlingen från att man fyllde 18 år och lämnade maximal ett år. På grund av detta har det urvalet som gjorts sett som en reducering av dimensioner men inte särskilt stor informationsförlust.

2.2 Transformationer och urval

Variabeln *År* visar vilket år modern är född och är skapad genom att variabeln födelsedatum delats med 10 000 och sedan trunkerats för att få fram endast ett årtal.

Livslängd räknas ut genom att använda två variabler från originaldatat, *födelse-* och *dödsdatum*. Denna individuella livslängdsvariabel togs fram genom att subtrahera födelsedatumet från dödsdatumet, delad det värdet på 10 000 och sedan trunkera det värdet. Trunkering betyder att decimalerna från födelsedatumet tas bort så att det endast blir ett årtal kvar. För att sedan få ut en enklare variabel att arbeta med gjordes variabeln om till en gruppvariabel enligt följande princip:

Livslängd i år	≤40	41-50	51-60	61-70	71-80	81-90	90<
Grupp	1	2	3	4	5	6	7

Variabeln *Spädbarnsdödlighet* som använts är en summering av den dikotoma *Spädbarnsdöd* för varje barn för en mor. För att lätt kunna arbeta med variabeln *Spädbarnsdödlighet* delades den med variabeln *AntalBarn* och blev istället proportionen döda barn för varje moder. Den nya variabeln heter *Propdoda.*

För att få ut de individer som behövs för att bygga modeller och få de bästa prediktionerna gjordes ett urval inom variablerna enligt följande: *AntalBarn≥1, 20år=TRUE* och *Kön=kvinna.* Det betyder att de personer som ingick i urvalet var kvinnor som hade bott i någon av församlingarna i minst 20 år och fått minst ett barn. Anledningen till att det endast får vara kvinnor som bott minst 20 år i någon av församlingarna med är att för att få en helhetsbild av vardera personens barnafödslar. Med urvalet blev antalet kvinnor 7304 stycken.

För den tredje frågeställningen behövdes en ny dikotom variabel som visar huruvida modern skaffar flera barn efter ett visst antal barn. Variabeln heter *Flera* och skapas genom att koda om antalet barn en mor fått till antingen en etta eller noll beroende på hur många de är.

Variabeln uppdateras för varje ny modell och antagande enligt denna princip. Först exkluderas alla mödrar som fått färre barn än de som modellen berör, t ex fem barn, efter detta skapas variabeln genom att alla mödrar som fått fem barn får en nolla på variabeln *Flera* och alla som fått fler än fem barn en etta.

För denna del av analysen fanns det även ett behov av att göra om variabeln *Propdoda.* Här är det av intresse att se efter om andelen döda barn man hittills har fått påverkar om man skaffar flera barn. Alltså skapas en dödlighetsvariabel på samma sätt som variabeln *Propdoda* gjordes fast en för varje antal barn. Alltså gjordes 21 variabler benämnda som *Propdöda1-21* som visar hur hög proportion döda barn varje mor fått upp till ett visst antal barn.

3. Metod

Innehållet under denna rubrik kommer berätta mer om de regressionstekniker som används för olika typer av responsvariabler och generaliserade linjära modeller. Dessutom kommer även modellkontroll och funktioner i R som används i utförandet att beskrivas.

3.1 Regressionsanalys

Syftet med enkel linjär regression är att finna en linje som beskriver datat, det vill säga approximerar den sanna relationen mellan beroendevariabeln Y och den oberoende variabeln X.[5] Multipelregression kan ses som en extension av en enkla linjära med fler än en oberoende variabel enligt ekvationen nedan:

$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon$

där $\beta_0, \beta_1, \ldots, \beta_k$ är regressionskoefficienter som ska estimeras och $X_1,\ldots,X_k$ är oberoende variabler som är antingen en variabel ($X_k=X_k$) eller en funktion av andra variabler ($X_k=X_{k-1}$-X_{k-2}) och ε är felet.

För att estimera parametrarna används vanligtvis minsta kvadratmetoden som söker den bästa linjära anpassningen, det vill säga den estimation som skiljer sig minst från de observerade värdena. Minsta kvadrat metoden minimerar enligt denna formel:

$$\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{1i} - \cdots - \beta_k X_{ki})^2$$

Lösningen minsta kvadratmetoden ger är de estimerade värdena $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$.

3.2 Exponentialfamiljen

Exponentialfamiljen är en klass av fördelningar som innehåller specialfall av många välkända fördelningar som kan skrivas på formen:[6]

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

där a(.), b(.) och c(.) är någon funktion specifik för just denna fördelning. Θ är den kanoniska parametern som är en funktion av fördelningens lägesparameter och Ø är en skalparameter som för samtliga fördelningar som kommer att användas är känd. När formeln logaritmeras får man ut den generella formen av log-likelihoodfunktionen:

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

[5] Klienbaum, Kupper, Muller, Nizam, sidan 39
[6] Olsson, sidan 37

Vi vet att E(y)=μ=b´(θ) och att Var(y)= $a(\emptyset)$b´´(θ).[7]

3.3 Generaliserade linjära modeller

Generaliserade linjära modeller (GLM) är en generell klass av statistiska modeller som inkluderar många vanliga modeller som specialfall. De tar upp modellering med olika typer av linjär regression. I GLM används denna generella linjära modell $\boldsymbol{y = \beta X + e}$ där $\boldsymbol{\eta = \beta X}$ är den så kallade linjärprediktorn och visar effekten av de oberoende variablerna. GLM är en generalisering av vanliga linjära modeller av dessa anledningar:[8]

1. Vi kan tillåta fördelningen för **y** att vara vilken som helst av de som tillhör exponentialfamiljen och inte enbart normal som krävs i vanliga linjära modeller.
2. Den linjära funktionen av de oberoende variablerna, det vill säga linjärprediktorn, modelleras som en funktion av E(Y), g(**μ**) som även kallas länkfunktionen. Detta ger följande likhet: $\boldsymbol{\eta = \beta X}$= g(**μ**)

Länkfunktionen måste vara monoton och deriverbar och väljs utifrån vilket typ av data man har. För räknedata bör länkfunktionen begränsa μ så den endast kan anta positiva värden och för proportioner ska länkfunktionen se till att μ∈ [0,1]. Mer om olika länkfunktioner kan man läsa i Olsson sidan 40. En länkfunktion kan även vara en kanonisk sådan. Det innebär att den länken är mest naturlig för just den fördelningen då den gör om medelvärdet till en kanonisk lägesparameter från exponentialfamiljen, alltså så att η= g(μ)=θ. Bara för att en länkfunktion är kanonisk innebär inte att den är den bästa för modellen.

När modellen specificeras måste följande val göras: fördelning, länkfunktion och linjärprediktor.

För att få fram skattningar av parametrarna används oftast Maximum Likelihood Estimation (MLE), alltså de skattningar som maximerar likelihoodfunktionen. Detta görs i R med hjälp av Iteratively Reweighted Least Squares (IWLS).[9]

Nedan presenteras fyra olika varianter av GLM som kommer att användas, med fyra olika fördelningar som alla tillhör exponentialfamiljen.

3.3.1 Poissonfördelning

Poissonregressionsanalys är en teknik inom GLM som modellerar en beroendevariabel som beskriver räknedata.[10] Poissonfördelningen ser ut enligt följande:[11]

[7] Olsson, sidan 39
[8] Ibid, sidan 36
[9] Ibid, sidan 44
[10] Klienbaum, Kupper, Muller, Nizam, sidan 687

$f(y,\mu) = \frac{\mu^y e^{-\mu}}{y!}$ med E(y)=Var(y)= μ

När man använder poissonregression finns det några generella beaktanden man bör räkna med. Beroendevariabeln Y är vanligtvis en räknevariabel observerad en gång för varje vald undergrupp som är beskriven av ett antal prediktorer $X_1, X_2,\ldots,X_n$.

Y_i är det observerade antalet händelser i grupp *i*.

$\mathbf{X}_i=(X_{i1}, X_{i2},\ldots,X_{ik})$ är mängden prediktorer som är specifikt för just grupp *i*.

$\boldsymbol{\beta}= (\beta_0, \beta_1, \ldots, \beta_k)$ är en mängd okända parametrar och $\lambda(\mathbf{X}_i, \boldsymbol{\beta})$ är en specifik funktion för prediktorerna och parametrarna som representerar händelsekvoten för delgruppen, det förväntade antalet blir då:

$E(Y_i)= \mu_i= \lambda(\mathbf{X}_i, \boldsymbol{\beta})$ då $f(Y_i,\mu_i) = \frac{{\mu_i}^{Y_i} e^{-\mu_i}}{Y_i!}$

Ett problem som kan uppkomma när man använder poissonregression är att man kan ha en överspridning i datamaterialet.[12] Överspridning kan uppstå då variansen är större än förväntat och gör då så att modellen får en dålig anpassning. En annan effekt som kan uppstå är att standardavvikelserna för skattningarna blir för små, teststatistikan därmed blir för stor och att man då får för lätt att få signifikanta resultat.[13] Det finns två vanliga sätt att hantera överspridning på. Dels kan man använda sig av en skalparameter i variansfunktionen, $Var(Y)=\phi\mu$ där ϕ skattas genom *Devians/frihetsgraderna*, eller att byta fördelning till en Negativ Binomial. [14] När man väljer att lägga till en skalparameter använder man sig av en så kallad quasipoissonfördelning.

För poissonfördelningen i GLM är den kanoniska länkfunktionen en *log*länk, alltså ser den ut så här: $\log(\mu)= \eta$.

3.3.2 Trunkerad poissonfördelning

Trunkering härstammar från att värden över eller under en viss gräns inte finns med eller inte är tillgängliga i materialet. Det innebär alltså att när man har ett trunkerat dataset saknas värden under eller över ett visst värde på en variabel. I detta fall kan variabeln AntalBarn endast anta värden som minst ett, alltså måste man ha fått minst ett barn för att vara med i urvalet. Detta skulle kunna betyda att en trunkerad poissonfördelning som ej kan anta värdet noll är mer lämpad för modellanpassning. Fördelningen för en trunkerad poisson ser ut enligt följande:[15]

---

[11] Olsson, sidan 37
[12] Dobson, sidan 45
[13] Coxe, West, Aiken, sidan 130
[14] Dobson, sidan 167
[15] Lindsey, sidan 57

$$f(y,\mu) = \frac{\mu^y e^{-\mu}}{y!(1-e^{-\mu})}$$

Fördelningen tillhör fortfarande exponentialfamiljen och länkfunktionen som används kallas *loge* och ser såhär ut: log(μ)= η

3.3.3 Negativ binomialfördelning

När det finns överspridning i en poissonfördelning är ett av de bästa sätten att handskas med det att istället använda en negativ binomialfördelning. Det kan ses som en serie Bernoulliförsök tills *r* lyckade försök har uppnåtts. Fördelningen ser ut enligt följande:[16]

$f(y,p) = \binom{y-1}{r-1} p^r (1-p)^{y-r}$ där *p* är sannolikheten för att lyckas och E(*y*)=*r/p* och Var(*y*)=*r(1-p)/p*^{2}

För negativ binomialfördelningen är *log*länk, log(1-*r*/*p*)= η, den kanoniska länkfunktionen.

Skillnaden mellan fördelningarna för negativ binomial och poisson är att för samma medelvärde har negativ binomial större sannolikhet för att få nollresultat samt får en längre svans. En längre svans betyder då även en större varians och hjälp mot överspridning om sådan existerar. Att sannolikheten för nollresultat ökar med en negativ binomial jämfört med en poissonfördelning kan vara olämpligt då det inte existerar några nollor alls i responsvariabeln av intresse.

3.3.4 Logistisk regression

I logistisk regression används en binär variabel som responsvariabel, alltså beskriver variabeln som ska förklaras t ex om man får flera barn efter man har fått sitt femte och kan då bara anta värdet sant eller falskt. Det innebär att när värdet tas för varje moder kommer den binära variabeln att upprepas och summeras till en binomialfördelning:

$f(y,p) = \binom{n}{y} p^y (1-p)^{n-y} = \binom{n}{y} \left(\frac{p}{1-p}\right)^y (1-p)^n$ med E(*y*)=*np* och Var(*y*)= *np(1-p)*

Vid logistisk regression är *logit*länken den kanoniska länkfunktionen och ser ut så här: logit(*p*)=log$\left({}^{p}/_{1-p}\right)$= η. Kvoten som logaritmeras brukar kallas oddskvoten på grund av att den är oddsen för att lyckas, därav kallas logit ibland för logodds.

3.4 Modellkontroll

3.4.1 Waldtest

När man valt modell och vill testa om förklaringsvariablerna man har med i den bör behållas kan man använda ett Waldtest. Förklaringsvariablerna bör behållas om testet visar ett

[16] Olsson, sidan 134

signifikant resultat då det betyder att de förklarar en del av variationen i modellen och förstås att de är relaterade till responsvariabeln.[17] Man testar följande hypoteser:

$$H_0: \beta_1 = 0 \; och \; H_A: \beta_1 \neq 0$$

H_0 säger att den för stora stickprov MLE-skattade parametern inte har någon påverkan på modellen. Walds teststatistika kan definieras som:[18]

$$Wald = \frac{(\hat{\beta}_1 - \beta_1)}{\sqrt{\widehat{Var}\hat{\beta}_1}} \simeq z$$

Där *z* approximativt följer en standardnormalfördelning om nollhypotesen är sann, men kan också kvadreras och därmed följa en X^2-fördelning.

3.4.2 Likelihoodkvottest

Likelihoodkvottest används för att undersöka om en modell är bättre än en annan. L($\hat{\theta}_1$) är likelihoodfunktionen maximerad över det fulla parameterområdet och L($\hat{\theta}_0$) är den maximerade likelihoodfunktionen över parametrarna som hör till nollhypotesen.

Då är det alltså likelihoodkvoten, L($\hat{\theta}_0$)/ L($\hat{\theta}$), som leder till teststatistikan för hypotesen, -2log(L($\hat{\theta}_0$)/ L($\hat{\theta}_1$)) ≥ $X^2(df_1\text{-}df_0)$.[19] Om kvoten är större än X^2-värdet och man får ett p-värde lika med eller under 0,05 är skillnaden mellan modellerna signifikant.

3.4.3 Deviansanalys

Devians är även det ett mått som används för att se hur bra en modell är.[20] Deviansen baseras dels på den modellen som har en perfekt anpassning till datamaterialet, en så kallad full modell, samt på den valda modellen. Den skalade deviansen är deviansen delad med skalparametern $a(\emptyset)$. Vi vet att för alla de fördelningar som ska användas gäller $a(\emptyset)$=1 och det betyder att den skalade är den samma som den vanliga. För att få fram deviansen används den logaritmerade likelihooden av maximum likelihood estimaten för både den valda modellen och för den fulla modellen enligt följande:

$D=2(l(\boldsymbol{y}, \emptyset; \boldsymbol{y})-l(\hat{\boldsymbol{\mu}}, \emptyset; \boldsymbol{y}))$

Modellen är bra om den skalade deviansen är liten, alltså att skillnaden mellan den fulla och valda modellen är liten. När man gör deviansanalys testar man om följande kvot, $D_{2-1} = D_2 - D_1 / df_2 - df_1$, är större än ett X^2-värde med samma frihetsgrader,

[17] Olsson, sidan 47
[18] Klienbaum, Kupper, Muller, Nizam, sidan 647
[19] Olsson, sidan 48
[20] Ibid, sidan 45

$D_{2-1} \geq X^2(df_2\text{-}df_1)$ där är D_1 den mindre modellen. Om kvoten är större än X^2-värdet och ett p-värde lika med eller under 0,05 erhålls är skillnaden mellan modellerna signifikant.

3.4.4 AIC

AIC är ett mått som vanligtvis används vid modellbyggnad för att välja den bästa modellen.[21] Som enskilt mått säger AIC ganska lite då den har en något godtycklig skala, dock är det ett bra mått att använda när man vill jämföra två olika modeller.

AIC=$2(k-l(\hat{\boldsymbol{\mu}}, \emptyset; \boldsymbol{y}))$

där k är antalet parametrar i modellen och $l(\hat{\boldsymbol{\mu}}, \emptyset; \boldsymbol{y})$ är maximumvärdet av likelihood-funktionen för den skattade modellen.

Ett bra AIC är ett lågt AIC, men en enkel modell kan automatiskt ge ett lågt AIC då det ”straffar” modeller med flera parametrar. Det betyder att det kan vara svårt att veta vad ett litet AIC är, men att man lätt kan jämföra värden mellan flera olika modeller.

3.4.5 Residualer

Residualerna för en GLM skiljer sig från dem från en vanlig linjär regression. I en vanlig linjär regression är residualerna skillnaden mellan det observerade och anpassade värdet, i en GLM är variansen av residualerna kopplad till de anpassade värdena. Man kan beskriva relationen med följande likhet: $\hat{\boldsymbol{y}}=\boldsymbol{Hy}$ där **H** är den såkallade hattmatrisen. Hattmatrisen används för att beskriva inflytandet varje observation har på sitt anpassade värde. För att kunna använda de råa residualerna till modellkontroll och liknande måste man välja att skala om dessa på något sätt.

Råa residualer och standardiserade deviansresidualer är de som kommer användas.[22] De råa residualerna skalas om till deviansresidualer med hjälp av att de multipliceras med roten ur d_i, det är den inverkan den i: te observationen har på deviansen för modellen.

$$e_{i,Devians} = sign(y_i - \hat{y}_i)\sqrt{d_i}$$

Termen sign betyder att den försäkrar att $e_{i,Devians}$ har samma tecken som andra typer av residualer.[23] Residualerna standardiseras med hjälp av diagonalelementen ur hattmatrisen:

$$e_{i,adj,D} = {e_{i,Devians}}\Big/{\sqrt{1-h_{ii}}}$$

[21] Olsson, sidan 46
[22] Ibid, sidan 56
[23] Dobson, sidan 127

3.4.6 Grafisk modellkontroll

Den grafiska modellkontrollen utförs med hjälp av residualer, anpassade värden och observerade värden. Dessa används för att se hur bra modellen passar datat och visas i två olika typer av grafer.

Dels plottas de råa residualerna eller deviansresidualerna mot deras anpassade värden. Här vill man att värden ska ligga nära nollinjen, då poängen är att residualerna ska sprida sig konstant kring medelvärde. Det betyder att det inte är önskvärt med någon typ av mönster i grafen då det tyder på att modellen inte tar med all påverkan.

En annan modellkontroll som kan göras är att plotta de observerade värdena mot de anpassade och där se efter om de av modellen anpassade värdena verkar träffa rätt. Målet här är att punkten ska ligga jämt mellan de olika axlarna, att det observerade värdet och det anpassade värdet för samma individ ska vara relativt lika varandra.

3.5 R funktioner

Funktionen glm används för att anpassa generaliserade linjära modeller, ge en förklaring av linjärprediktorn samt anta en fördelning och kan hittas i R-paketet stats. Funktionen glm.nb finns i paketet MASS och är en modifikation av glm, speciell för att kunna bygga generaliserade linjära modeller med fördelningen negativ binomial där man vill estimera en extra parameter. För att kunna använda mig av en trunkerad poissonfördelning som familj i generaliserade linjära modeller måste paketet VGAM och funktionen vglm användas. Trunkerad poissonfördelning heter *pospoisson* i paketet och här finns det tyvärr inte tillgång till att få ut andra residualer än de råa. För att testa överspridning användes paketet qcc och funktionen qcc.overdispersion.test där H_0= ingen överspridning. Den använder beräkningen

$$Varians_{Observerad} / (Varians_{Teoretisk} * (n-1))$$

och jämför sedan värdet med en $X^2(n-1)$ fördelning, vid högre värde och p-värde under 0,05 kan H_0 förkastas.

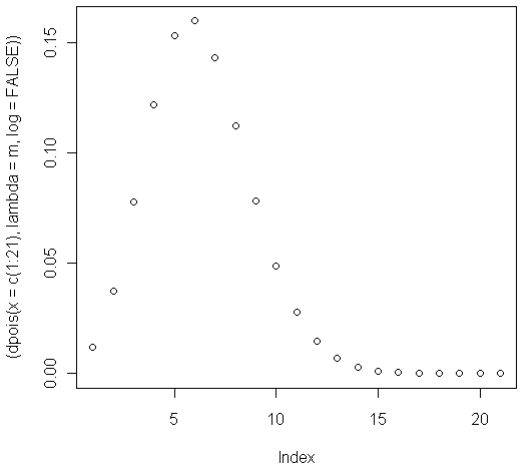
4. Resultat

4.1 Deskriptivt

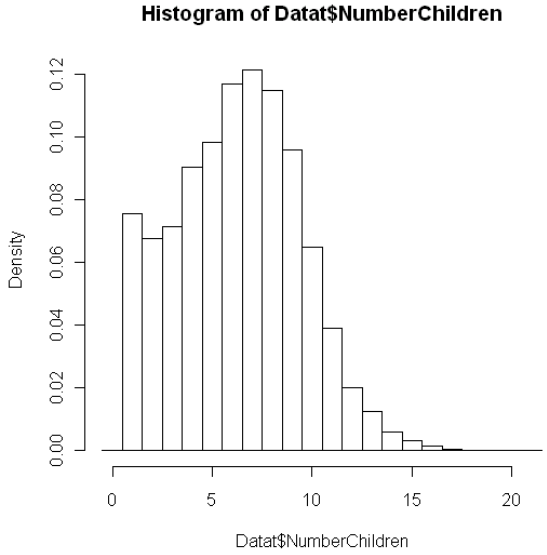
Tabell 1 visar en summering av variabeln *AntalBarn*

Minimum	Median	Medelvärde	Varians	Maximum
1	6	6,268	9,783	21

I tabell 1 kan man se att maxvärdet av antalet barn någon fått är 21 och att μ =6,268, vi kan även uppmärksamma att variansen är 9,783 det vill säga 1,5 gånger så stor som medelvärdet.



Figur 1 visar den sanna poissonfördelningen med ett μ som är medelvärdet av *AntalBarn*[1]



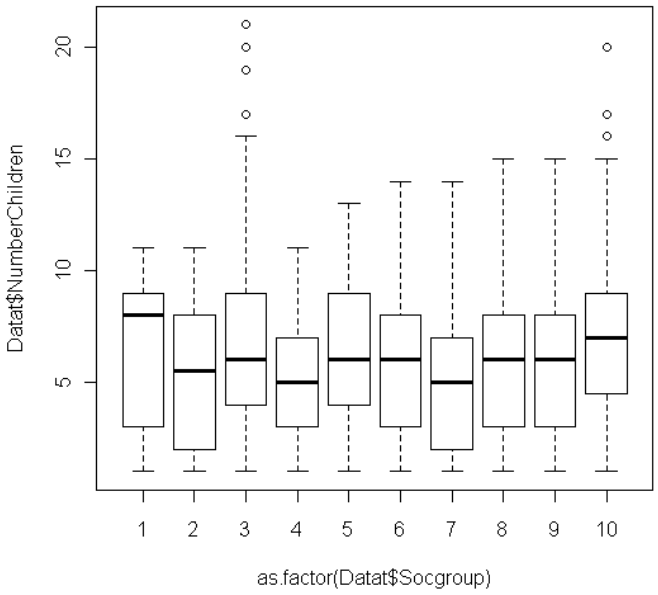
Figur 2 visar fördelningen av observationerna för variabeln *AntalBarn*.[1]

I figur 1 kan vi se den sanna fördelningen av antalet barn varje mor fått. Genom att jämföra den med figur 2 kan man även se att de två figurerna har nästan samma form, förutom att den observerade fördelningen har en fetare vänstersvans än den teoretiska. Detta leder till slutsatsen att en poissonfördelning skulle vara lämpligt för att modellera antalet barn.

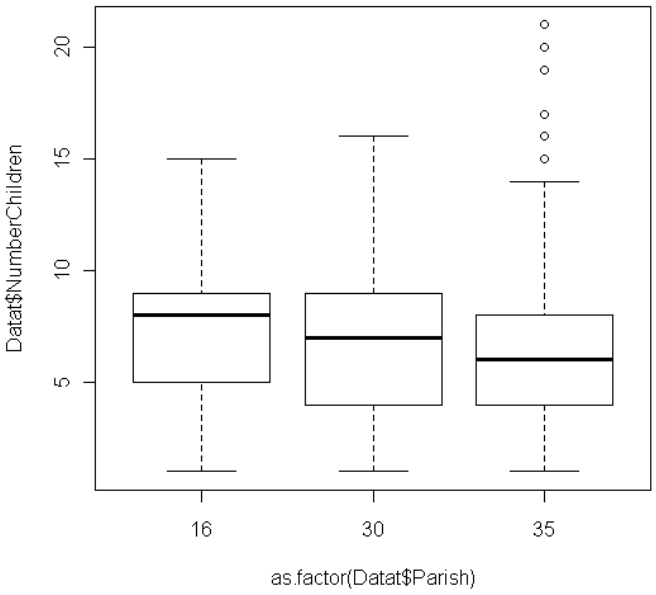
Tabell 2 visar resultatet av överspridningstestet

Overdispersion test poisson data		
Obs.Var/Theor.Var	Statistic	p-value
1,560784	11398,40	0

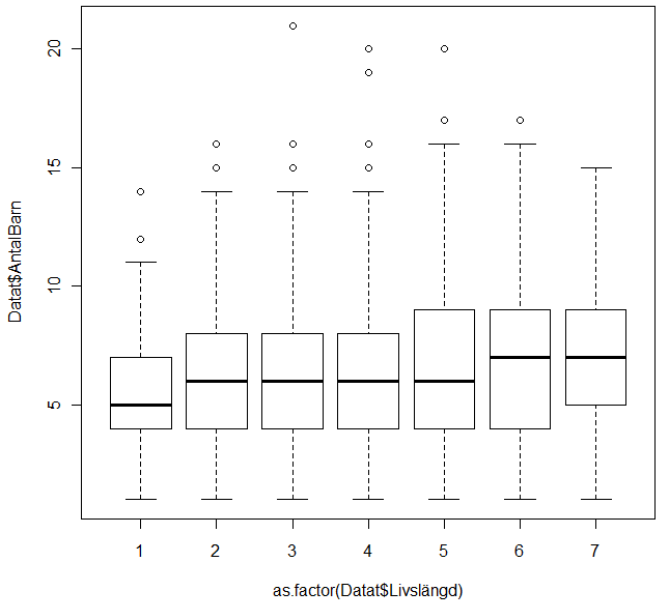
Överspridningstestet använder sig av den observerade och teoretiska variansen samt antalet observationer som finns i datasetet för att ta fram statistikan. Resultatet tyder på att H_0 kan förkastas och att det kan finnas överspridning i variabeln *AntalBarn* och därmed i materialet.



Figur 3 visar hur AntalBarn är fördelade över vilken socialgrupp deras mor tillhörde.



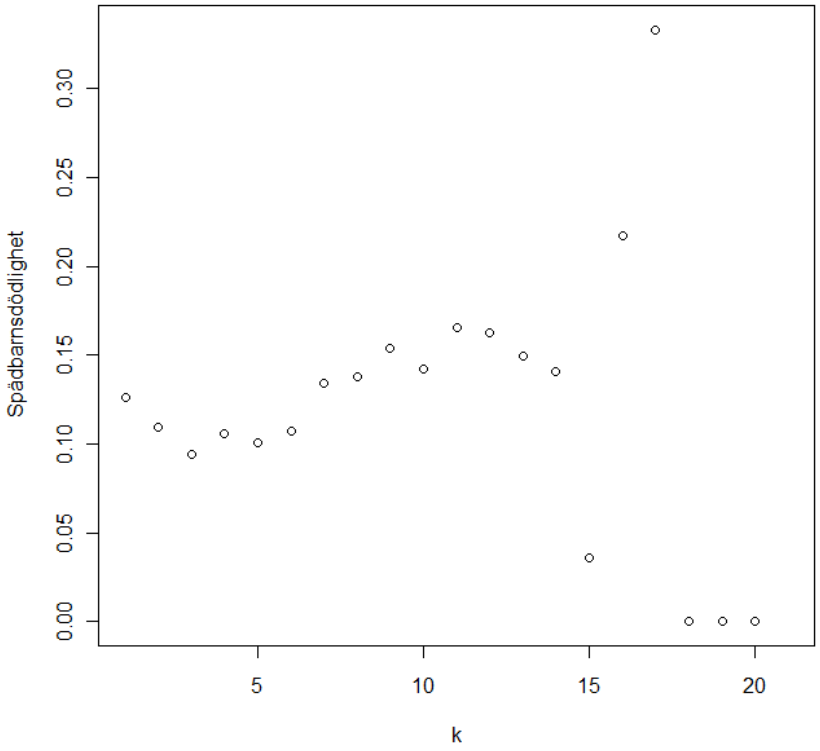
Figur 4 visar hur AntalBarn är fördelade över vilken församling deras mor bodde i där Jörn (=16), Norsjö (=30) och Skellefteå (=35).



Figur 5 visar hur AntalBarn är fördelade över vilken livslängd deras mödrar hade.

I figur 3-5 kan man se fördelningen av antalet barn för varje faktor av en viss variabel. Spridningen mellan grupperna skiljer sig mellan de olika variablerna. I figur 3 har grupp tre och tio den största spridningen av antalet barn. Grupp tre och tio är de som har flest individ i sina grupper. Grupp fyra är den grupp med klart minst spridning gällande både hela mängden men även mellan 25:te och 75:te percentilen. Även grupp ett och två har en liten spridning av hela sin mängd mödrar. Medianen är ganska jämn mellan grupperna, den ligger runt 5-7 barn, dock är grupp ett (Storföretagare) ett undantag med nästan 9 barn. Figur 4 visar de tre olika

församlingarna och att de har liten skillnad mellan sig. Församling 35 (Skellefteå) har den lägsta medianen och skillnaden mellan 25:te och 75:te percentilen men även den största spridningen mellan lägsta och högsta värde på variabeln, detta kan bero på att det är den klart största gruppen. Figur 5 visar att ju äldre mödrarna blev, desto högre blev median av antalet barn. Grupp 1 innehåller de mödrar som blev mellan 37-40 år och denna grupp har den minsta skillnaden mellan sin 25:te och 75:te percentil. Grupp 2, 3 och 4 är extremt lika varandra vilket gör att man kan misstänka att barnafödandet inte skiljde sig så mycket mellan åldersgrupperna 41-50, 51-60 och 61-70. Grupp 5 och 6 har ungefär samma spridning av antalet barn, men grupp 6 har högre median. Grupp 7 har högst median, men med ungefär samma spridning som grupp 2-4.



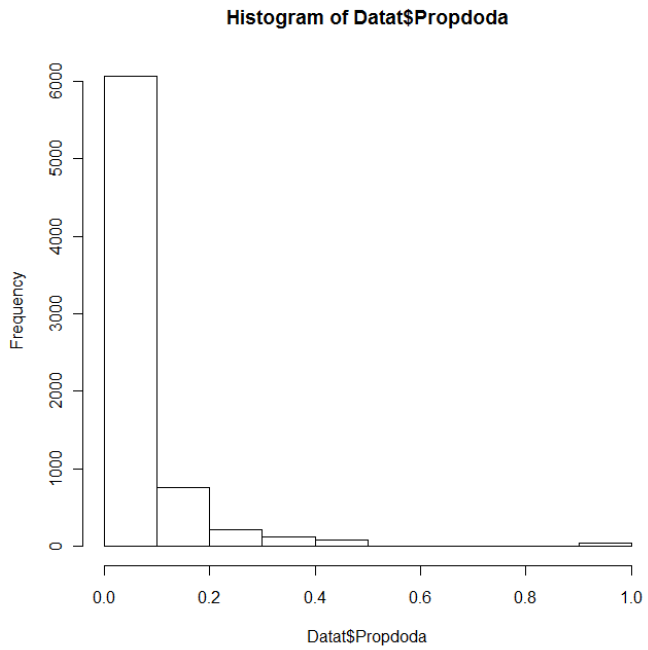
I figur 6 ser vi proportionen döda barn för varje barn i kullen

Tabell 3 visar numeriskt proportionen döda barn för varje barn i kullen.

Barn nummer	1	2	3	4	5	6	7	8	9	10	
Proportion döda	0,126	0,109	0,094	0,105	0,101	0,107	0,134	0,138	0,154	0,142	
	11	12	13	14	15	16	17	18	19	20	21
	0,165	0,163	0,149	0,141	0,036	0,217	0,333	0	0	0	NA

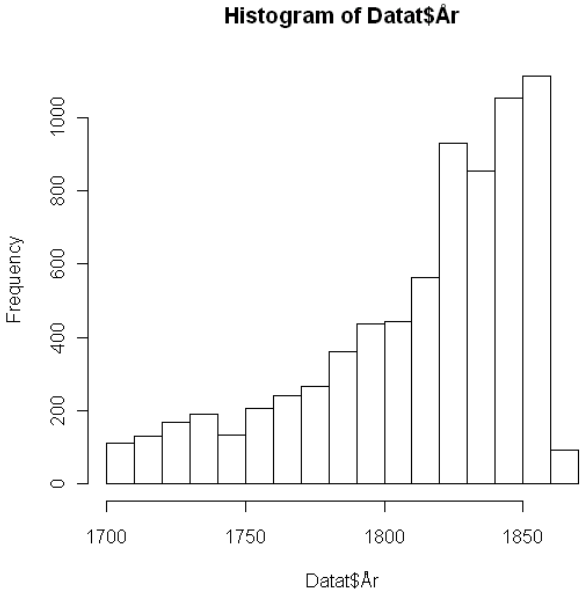
I figur 6 ovan syns proportionen döda barn för varje kull med barn som föddes. Den tillsammans med tabell 3 visar alltså hur stor andel av barnen som föddes som det första

barnet av sin moder som dog i spädbarnsdödlighet, ända upp till de som föddes som det 21:a barnet. I figuren ser vi att för barn 2 till 6 ligger andelen döda på runt 10 %. För barn 1 och 7 är proportionen snarare 13 %. Övriga barn har en proportion snarare upp mot 15 %. När man kommer upp mot barn 17 blir antalet så få att det är svårt att säga om siffrorna är helt rättvisande, dock dog en tredjedel av de barn som hade registrerade värden. I figur 6 saknas en punkt för den 21:a kullen barn. Detta beror på, precis som man kan se i tabell 3, att de barn som är det 21:a i sin kull inte har ett registrerat värde på den variabeln.



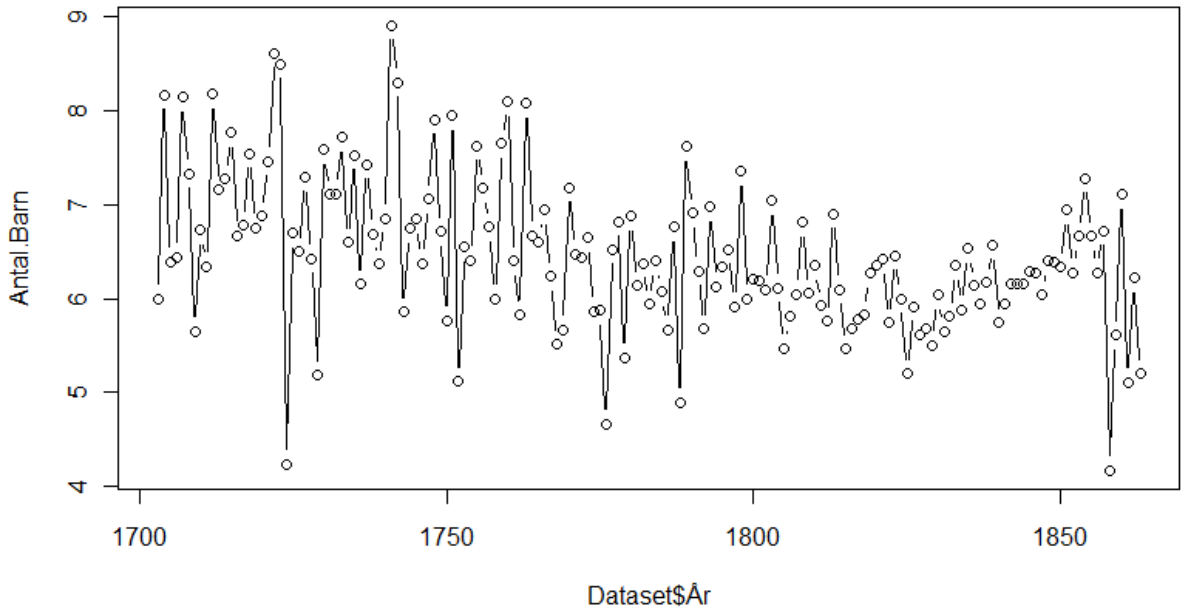
Figur 7 visar hur stor proportion döda barn alla mödrar fick.

De allra flesta mödrarna, lite över 6 000, hade en proportion döda barn mellan 0 och 0,1, sedan hade cirka 800 en proportion mellan 0,1 och 0,2. Efter proportionen 0,2 är antalet inte alls lika stort utan planar ut ganska fort. Mellan 0,5 och 0,9 finns nästan inga observationer. Dock kommer en höjning mellan 0,9-1, vilket betyder att det finns en viss andel som förlorade nästan alla sina barn.

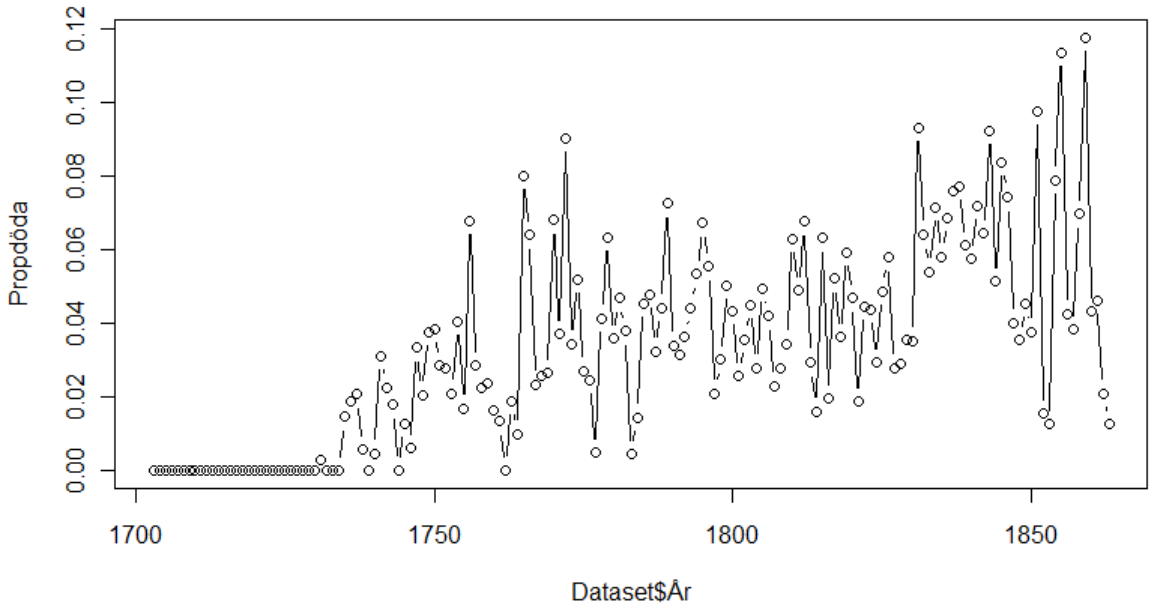


Figur 8 visar vilket år alla mödrar är födda med värden mellan 1703 och 1863.

I figur 8 kan man se att antalet mödrar i datasetet stiger ju längre tiden går. Åren fram till 1750 är det max 200 per årtionde, fram till 1800 max 400 och fram emot 1850 fanns det nästan 1000 mödrar med i urvalet. En av anledningarna till att den första och sista stapel när så små är att varje stapel representerar 10 år och det inte finns 10 födelseår registrerade för de staplarna.



Figur 9 visar medelvärdet för antal födda barn fördelat över år



Figur 10 visar medelvärdet för proportionen döda barn fördelat över år

För att titta närmre på delen rörande om det finns någon utveckling över tid för variablerna *Propdoda* och *AntalBarn* görs nämnda variabler om något. De summerades och det togs fram ett medelvärde för varje variabel per år. Dessa undersöktes sedan i figur 9 och 10. Om man tittar på de två olika figurerna ser man att de utvecklas åt olika håll. I figur 9 ser vi att antalet barn som kvinnorna födda ett visst år får minskar ju senare de är födda. Mellan 1800 och 1850 är variationen som minst mellan de olika åren, medan den svänger betydligt mer mellan 1700 och 1750. I figur 10 kan vi se att medelvärdet för proportionen döda barn för kvinnor födda ett visst år ökar när åren går. Variationen mellan de olika åren är svängig, dock ser den ut att vara mest stabil mellan ca 1780 till 1830. Det förekommer några ganska stora hopp mellan medelvärdena före 1780 och efter 1850.

4.2 Modellbyggnad med olika fördelningar

Modellbyggnaden är indelad i tre olika steg där alla utskrifter från modellerna finns i appendix A. Först används poisson som fördelning, sedan negativ binomial och sist trunkerad poisson. När man tittar på modellerna kan man se på signifikans på två olika sätt. Dels kan en variabel eller interaktion vara signifikant på 5%-ig nivå och kan då tolkas och anses bidraga till förklaringen av responsvariabeln. Men en tumregel säger att även en term som är signifikant på 25%-ig nivå kan vara med i modellen för förklaring, men att den inte kan tolkas i samma utsträckning.[24] De första två modellerna ser ut enligt följande:

[24] Olsson, sidan 26.

Barn.pois1 ← glm(AntalBarn ~ as.factor(Församling)+as.factor(Socgrupp)+as.factor(Livslängd) +Propdoda+År, data = Datat, family = poisson)

Den första modellen (tabell 8) innehåller samtliga förklaringsvariabler där variablerna *Församling, Socgrupp* och *Livslängd* används som faktorer. Den visar att ingen av faktorerna för *Socgrupp* är signifikanta på 5%-ig nivå. Endast två av dem (4 och 7) är signifikanta på 25%-ig nivå. Man kan även se att standardavvikelserna är väldigt stora för skattningarna av socialgrupperna. De är för flera av faktorerna mer än dubbelt så stora som själva skattningarna och medför att man inte kan veta om skattningarna har en positiv eller negativ påverkan på responsvariabel då skattningarna i sig är små. På skattningarna i modellen för församlingsfaktorerna kan man se att antalet barn minskar för Norsjö jämfört med Jörn och att detsamma gäller för Skellefteå jämfört med de andra orterna. Vi kan även se att med längre livslängd får man flera barn, alltså att skattningen blir större för varje grupp som går. Det intressanta i livslängdsvariabeln är att åldersgrupp 3 har en högre skattning än övriga, förutom den äldsta gruppen som är nummer 7. Alltså ligger grupp 2 och 4 på ungefär samma nivå medan grupp 3 har högre skattningar än både de två samt de för grupp 5 och 6. Totalt sett ligger skattningarna mellan 0,109 och 0,227. Man kan även se att proportionen döda barn har en positiv inverkan på antalet barn man får med en skattning på nästan 0,2. Vilket år man är född ger en negativ påverkan, alltså innebär det att ju senare man är född desto större negativ påverkan har det på antalet barn man skaffar. Dock är skattningen för variabeln inte särskilt stor så skillnaden bli liten mellan åren. AIC för modellen är 21 380.

Barn.pois2 ← glm(AntalBarn ~ as.factor(Församling)+as.factor(Livslängd)+År+Propdoda, data = Datat, family = poisson)

Den skillnad som utgör modell två som kan ses i tabell 9 är att variabeln *Socgrupp* plockats bort helt och hållet. Det gör att interceptet blir större, från 4,042 till 4,947. Den negativa påverkan som kom från vardera församlingsfaktorn blir större. Vidare förändras skattningarna för faktorerna för *Livslängd* nästan ingenting. För variabeln *År* blir inte den lilla skillnaden som uppkommer så betydande i verkligheten då skattningen i sig är så liten. Dock finns den en tydlig skillnad i skattningen för *Propdoda* från 0,188 till 0,148.

AIC är på 21 527. Det gör att skillnaden i AIC mellan modellerna är på cirka 150, i detta fall betyder det en ökning av AIC:t med ca 0,7 %.

Tabell 4 visar deviansanalysen för modell 1-2

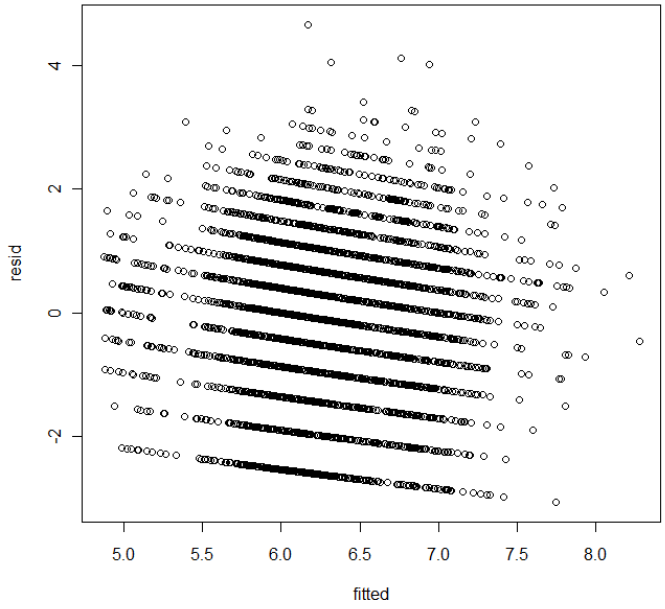
Deviansanalystabell					
Modell 1: AntalBarn~as.factor(Församling)+ as.factor(Livslängd) +Propdoda+År					
Modell 2: AntalBarn ~as.factor(Församling) +as.factor(Socgrupp)+ as.factor(Livslängd)+ Propdoda + År					
	Resid. Df	Resid. Dev	Df	Deviance	P(>|Chi|)

1	4051	7159,4			
2	4042	6993,9	9	165,5	5.297e-31

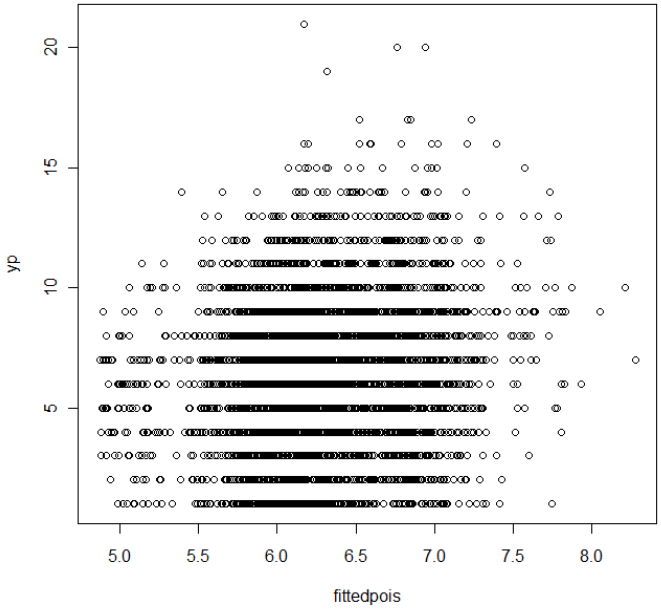
Modellerna lades in modellerna i analysen efter antalet parametrar som ingår i vardera modellen. I tabell 4 kan vi se att modell 2 är signifikant bättre än modell 1. Det betyder att modellen utan *Socgrupp* är signifikant bättre än den modellen med densamma.

Den bästa modellen för poissonfördelningen ser alltså ut enligt följande:

$$\log(\mu_i) = 4.9469 - 0.138\text{Norsjö} - 0.2711\text{Skellefteå} + 0.1086\text{Livslängd2} + 0.1544\text{L3} + 0.1101 * \text{L4} + 0.1213\text{L5} + 0.1408\text{L6} + 0.2232\text{L7} - 0.0017\text{Födelseår} + 0.1481\text{Pro. döda}$$



Figur 11 visar deviansresidualerna mot de anpassade värdena



Figur 12 visar de observerade värdena mot de anpassade värdena

För att kontrollera modellen grafiskt kan man se på figur 11 och 12. Figur 11 visar att residualerna för modell 2 är ganska jämt fördelade över nollinjen på y-axeln och att inget ickemodellerade mönster verkar finnas. I figur 12 kan vi se att modell 2:s anpassade värden inte sträcker sig över samma antal som de observerade gör. Anpassningen till datamaterialet är inte det bästa då så stor mängd av de observerade värdena inte ligger ens nära sitt anpassade värde.

När en rimlig modell med de parametrar som funnits att tillgå tagits fram samt att modellkontrollerna utförts blir nästa steg att gå vidare för att se efter om någon av de två andra fördelningarna kan ta bättre vara på underlaget i observationerna. Först ses det efter hur negativ binomialfördelningen kan passa. Överspridningstestet som tidigare gjordes visade att det skulle finnas överspridning i materialet, därav borde negativ binomial kunna passa.

Barn.negbin1 ← glm.nb(AntalBarn ~ as.factor(Församling)+as.factor(Socgrupp)+ as.factor(Livslängd)+ Propdoda+År, data = Datat)

I tabell 10 ser vi modell 3. Här ser vi att skattningarna för *Socgrupp* är ännu sämre än de var för poissonfördelningen och är verkligen inte signifikanta. Samma mönster återfinns för variabeln *Livslängd* som i poisson. De har nästan samma skattningar och nivå 3 och 7 står fortfarande ut på samma sätt som i modell 1. *Propdoda* ökar sin positiva påverkan jämfört med tidigare modeller medan skattningen för *År* ligger konstant på en väldigt låg nivå. Variabeln *Församling* har nästan exakt likadana skattningar på sina faktorer som i modell 1, dock blir inte faktorn för Norsjö signifikant på 5%-ig nivå. Eftersom den andra faktorn blir signifikant behålls ändå variabeln.

Här har vi ett AIC på 20 811, alltså cirka 500 enheter mindre än för modell 1.

Barn.negbin2←glm.nb(AntalBarn ~ as.factor(Församling)+ as.factor(Livslängd)+ Propdoda+ År, data = Datat)

Den andra modellen för negativ binomial är modell 4 i tabell 11, där återigen variabeln *Socgrupp* testats att tas bort. Det som skiljer modell 3 och 4 åt tydligast när *Socgrupp* plockats bort är precis som i modell 2 att skattningen för *Propdoda* minskar i sin påverkan samt att mer påverkan läggs i interceptet. Den positiva påverkan på antalet barn av *Livslängd* är fortfarande störst av faktor 3 och 7. Påverkan av faktorerna för *Församling* ökar, men dock inte lika mycket som de gjorde mellan modell 1 och 2. Även blir här faktorn för Norsjö insignifikant på 5%-ig men ej 25%-ig nivå, dock behålls den i modellen för uppbyggandens och de andra faktornivåernas skull. *År* förändras inte särskilt mycket. I övrigt stiger AIC:t till 20 893. Generellt är skattningarna väldigt lika de som fåtts i modell 2.

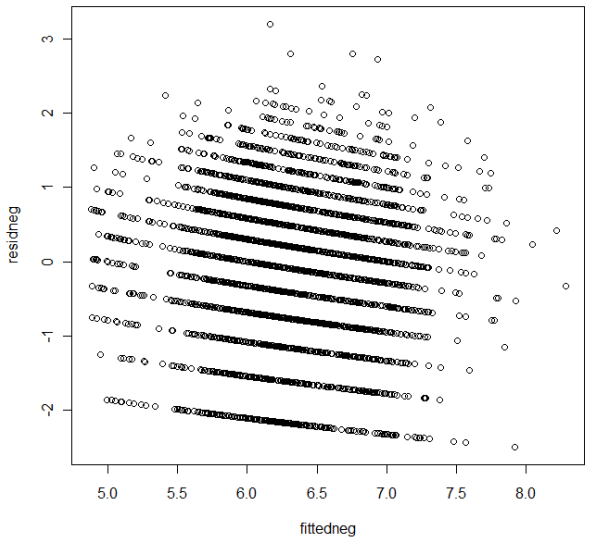
Tabell 5 visar likelihoodkvottest för modell 3-4

Likelihoodkvottest av negativ binomial modeller						
Respons: AntalBarn						
4 as.factor(Församling) + as.factor(Livslängd) + Propdoda + År						
3 as.factor(Församling) + as.factor(Socgrupp) + as.factor(Livslängd) + Propdoda + År						
Theta	Resid. Df	2 x log-lik.	Test	df	LR stat.	Pr(Chi)
8,588675	4051	-20868,77				
9,167183	4042	-20769,81	4 vs 3	9	99,78923	0,0000000

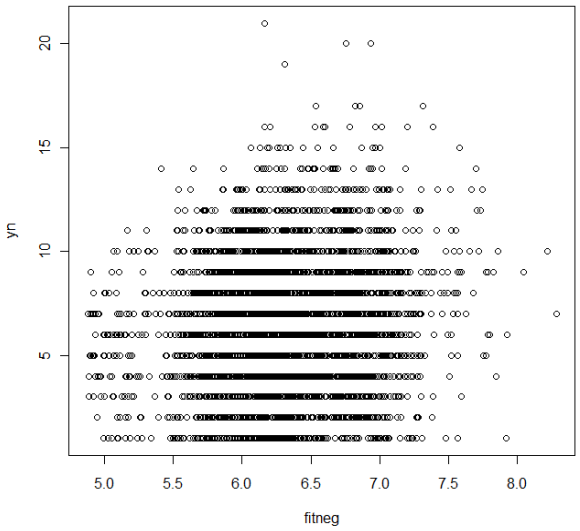
I tabell 5 återfinns likelihoodkvottestet för modellerna 3-4. Där kan vi se att skillnaden mellan modell 3 och 4 är signifikant, alltså är den mindre modellen bättre här. Det betyder att även här är modellen utan *Socgrupp* signifikant bättre än den andra.

Den bästa modellen för negativ binomialfördelningen ser alltså ut enligt följande:

$$\log(1 - r/p_i) = 4.9421 - 0.1349\text{Norsjö} - 0.2662\text{Skellefteå} + 0.1074\text{Livslängd2} + 0.1525\text{L3} + 0.1075\text{L4} + 0.119\text{L5} + 0.1385\text{L6} + 0.2212\text{L7} - 0.0017\text{Födelseår} + 0.1705\text{Pro. döda}$$



Figur 13 visar deviansresidualerna mot de anpassade värdena



Figur 14 visar de observerade värdena mot de anpassade värdena

Den grafiska modellkontrollen för modell 4 kan ses ovan i figur 13 och 14. Dessa två figurer visar samma sak som liknande figurerna för modellen med poissonfördelning. Residualerna är lite annorlunda fördelade över y-axeln, men mönstret dem emellan är sig likt. Figur 14 visar nästan exakt samma mönster som figur 12 visade.

Förutom att se efter om det finns överspridning i datat som kan modelleras kan det även testas om en bättre skattning kan fås genom att använda en positivt trunkerad poissonfördelning. Utgången är samma grundmodell som för de andra fördelningarna.

Barn.trunkpois1 ← vglm(AntalBarn ~ as.factor(Församling)+as.factor(Socgrupp)+ as.factor(Livslängd)+ Propdoda+År, data = Datat, family = pospoisson)

I modell 5 som kan ses i tabell 12 ser skattningarna väldigt lika ut som de i modell 1 och 3, dock mest som de i modell 1. Församlingsfaktorernas skattningar ger precis som tidigare ett negativt värde, socialgruppsfaktorerna blir inte signifikanta förutom två stycken, men då bara på 25%-ig nivå. Livslängdsfaktorerna är signifikanta, har en positiv påverkan samt att påverkan är högst för faktor 3 och 7. *Propdoda* ger en positiv påverkan på antalet barn på ungefär samma nivå som för modell 1 samt att variabeln *År* pekar på att ju senare man är född, desto färre barn får man. Modellen får ett AIC på 21 358,96. Det är alltså högre än för modellerna med negativ binomial som fördelning, men lägre för de med poissonfördelning.

Barn.trunkpois2 ← vglm(AntalBarn ~ as.factor(Församling)+as.factor(Livslängd)+Propdoda+År, data = Datat, family = pospoisson)

I den sista modellen som görs i detta avsnitt, modell 6 som kan ses i tabell 13, har *Socgrupp* testats att plockas bort. Här blir förändringen i interceptet väldigt stor, nästan en hel enhet. Precis som i tidigare jämförelser mellan modeller ökar storleken på skattningarna för församlingsfaktorerna samt för *År*. Samtidigt minskar de för *Propdoda* och är oförändrade för livslängdsfaktorerna. AIC:t som här är 21 509,9 skiljer sig ungefär 150 enheter från det värde fåtts för modell 5.

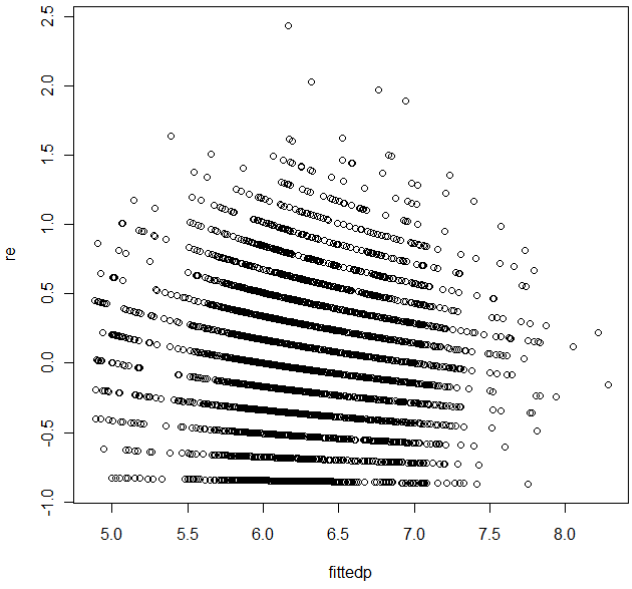
Tabell 6 visar deviansanalysen för modell 5 och 6

Deviansanalystabell				
Modell 5: AntalBarn~as.factor(Församling)+as.factor(Socgrupp)+as.factor(Livslängd)+Propdoda+År				
Modell 6: AntalBarn ~ as.factor(Församling) + as.factor(Livslängd) + År + Propdoda				
Resid. Df	Df	Devians	Chitvå	P(>|z|)
4051				
4042	9	18,77111	16,92	0

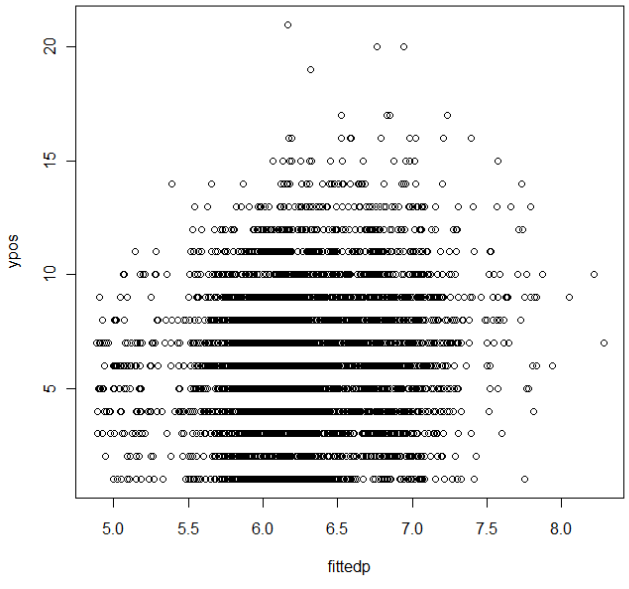
I tabell 6 syns deviansanalysen för modell 5 och 6. Den visar att modell 6 är signifikant bättre än 5. Alltså gäller det även här att modellen utan *Socgrupp* blir bättre. Det betyder att resultatet blir samstämmigt med det som framkommit i de övriga testen av modellerna.

Den bästa modellen för trunkerad poissonfördelningen ser alltså ut enligt följande:

$$\log(\mu_i) = 4.9805 - 0.1392\text{Norsjö} - 0.274\text{Skellefteå} + 0.1109\text{Livslängd2} + 0.1573\text{L3} + 0.1125 * \text{L4} + 0.1239\text{L5} + 0.1435\text{L6} + 0.2265\text{L7} - 0.0017\text{Födelseår} + 0.1499\text{Pro. döda}$$



Figur 15 visar de råa residualerna mot de anpassade värdena



Figur 16 visar de observerade värdena mot de anpassade värdena

Precis som i de tidigare modellkontrollgraferna visar dessa för modell 6 att det inte verkar finns något tydligt ickemodellerat mönster i figur 15, men dock att de tydligt markerade

linjerna är något mer krökta här. I denna figur används de råa residualerna istället för deviansresidualerna. Figur 16 visar ingen ändring i observerade värden mot anpassade.

Jämför man modellernas AIC ser vi i tabellen nedan att av de valda modellerna har modell 4 med negativ binomial som fördelning har det minsta och därmed bästa AIC:t.

Tabell 7 visar ACI för de utvalda modellerna

Modell	AIC
Barn.pois2	21527
Barn.negbin2	20893
Barn.trunkpois2	21 509,90

4.3 Logistiska modeller

För att kunna svara på den andra delen av syftet användes logistisk regression. Syftet där är att se över huruvida man kan se om en familj skaffar fler barn beroende på hur många barn de fått och hur många av dem som har avlidit vid tidig ålder. För att kunna göra detta har variabeln *Propdöda* använts som förklaringsvariabel och variabeln *Flera* använts som responsvariabel enligt nedan:

GLM.i ← glm(Flerai ~ Propdödai, family=binomial(logit), data=Datat) i=1,…,21

Sedan har modeller skapats för varje kull barn, från ett fram till 21. Efter femton barn har inte flera modeller tagits med då antalet barn inom varje kull inte är lika stora längre. I tabell 14 i appendix B kan man se skattningarna och signifikansnivåerna för varje modell. Tabellen visar att ingen av de femton modellerna får några signifikanta resultat för skattningarna för variabeln *Propdöda.* På endast 25%-ig nivå blir variabelskattningen signifikant för 2, 4 och 7 barn. För samtliga av de tio första intercepten gäller att de blir signifikanta på 5%-ig nivå och detsamma för det tolfte. Inget av intercepten blev signifikant på enbart 25%-ig nivå.

Detta tyder alltså på att ingen av modellerna kan säga helt klart att proportionen döda barn en kvinna fått påverkar om hon skaffar flera barn. Dock har proportionen en viss påverkan efter barn 2, 4 och 7. Men då detta bara är på 25%-ig nivå kan dessa skattningar inte tolkas. Vad som dock kan sägas om dem är att skattningarna av standardavvikelserna är ganska stora jämfört med skattningarna av parametern, de är mellan 70-85% av storleken på skattningarna av parametrarna. Det betyder att även om skattningarna kan erbjuda en viss förklaring till om man skaffar fler barn eller inte, så kan de variera mellan att vara positiva och negativa på grund av att standardavvikelsens storlek gör att de kan gå över nollgränsen.

5. Diskussion

I figur 9 kan man se att antalet mödrar som är med i urvalet ökar med åren som går och då minskas även påverkan av enskilda väldigt höga eller låga observationer. Detta kan vara orsaken till att variationen av medelvärdet mellan åren minskar med tiden. I figur 10 framgår det att proportionen döda barn var noll ända fram till 1735, förutom några enskilda hopp. Detta känns underligt att inga barn skulle ha dött i början av registreringen, även om vi sett att antalet mödrar är färre då. Här kan man misstänka att något fel i datamaterialet uppkommit.

När man jämför skattningarna för de olika modellerna i tabell 8-13 kan man se att generellt har negativ binomial något lägre värden (eller högre för negativa värden) än de för poissonfördelningarna. Detta varierar något, men för de flesta variablerna stämmer det då det egentligen bara är *Propdoda* som är större i påverkan för negativ binomialfördelningen. Vad som dock är mer tydligt när man tittar på modell 1-6 är att standardavvikelserna är större för skattningarna i Barn.negbin1 och Barn.negbin2 än för de övriga modellerna. Standardavvikelserna är högre för samtliga variabler och det borde kunna bero på att negativ binomialfördelningen förutsätter längre svansar än poissonfördelningen.

I jämförelsen av de olika modellerna Barn.pois2, Barn.negbin2 och Barn.trunkpois2 kan man dra några slutsatser. Barn.pois2 är grundmodellen som har den fördelning som söks efter, de andra två är modeller som har fått anpassas efter det ”brister” och avvikelser från grunden som existerar i datamaterialet. De värden som är lättast att jämföra är AIC som i tabell 7 visar att Barn.negbin2 med familjen negativ binomial är den med lägst värde. Dock är värdena alla ganska stora vilket gör att skillnaden mellan dem inte blir så stor i sammanhanget, då skillnaden är cirka 2 % av storleken på AIC-värdena.

I den grafiska modellkontrollen syns det inte att någon av modellerna ska vara sämre eller bättre än någon annan. Barn.trunkpois2 har residualerna med minst spridning, dock säger inte det inte jättemycket om hur bra själva modellen. De streckmönster som syns i samtliga grafer har troligtvis uppkommit på grund av att det handlar om räknedata, då de observerade värdena inte kan sprida sig mellan heltalen. Men helt säker på att streckmönstren enbart beror på hur de observerade värdena ser ut kan man inte vara. Även om de tre residualgraferna inte uppvisar ett tydligt ickemodellerat mönster tyder graferna för observerade värden mot anpassade värden på en dålig modellanpassning. Nästan inga värden under fem eller över åtta har predikterats av modellerna som ansågs vara bäst och de har alla fått snarlika anpassade värden. Detta tillsammans med vetskapen om att parameterskattningarna för de olika modellerna även de var snarlika tyder på att det finns delar av variationen för responsvariabeln *AntalBarn* inte förklaras av de variabler som finns i modellerna. Dock kan

en misstänkt poissonfördelad variabel ha en ganska stor varians, så att modellerna inte predikterar varje observation rätt är inte helt oväntat.

Då Barn.negbin2 och Barn.trunkpois2 är anpassade efter datat för att kunna finna en bättre anpassning än Barn.pois2 kunde erbjuda, och ändå bara presterar på samma nivå som Barn.pois2, kan den modellen ändå anses vara den bästa av dem. Detta tros att testet som gjordes visade på överspridning och att nollorna för variabeln *AntalBarn* faktiskt saknas. Dock är ingen av dem, som konstaterat, en särskilt bra modell för datamaterialet i sig.

Frågeställningarna som väcktes i början av uppsatsen var totalt tre stycken. Den första berörde huruvida den totala fertiliteten kan modelleras med hjälp av variabler kring moderns levnadshistorik. På den frågan är svaret att det går att modellera, tyvärr blir dock inte modellerna så väldigt bra. En vidaretagning av problemet här vore att undersöka flera variabler och se efter vad det är som gör att modellen predikterar dåligt.

Mönstret för spädbarnsdödlighet och antalet barn över tid handlade frågeställning två om. Dessa undersöktes endast med hjälp av två figurer 8-9. I dem visade det sig att de två variablerna hade olika utveckling över tid, men misstanke om att det fanns något fel i materialet eller insamlingen då vissa värden var underliga finns. Här vore det intressant att gå vidare rent djupgående för att se hur och om de två variablerna påverkar varandra och påverkas gemensamt av tiden.

Den sista av frågeställningarna tar upp huruvida historien av spädbarnsdödlighet inom en familj kan påverka om de skaffar flera barn. För att undersöka detta gjordes logistiska modeller som inte gav något utslag alls, inte för någon av barnkullarna. Utifrån modellerna i del ett kan man se att spädbarnsdödlighet har en påverkan på antalet barn då variabeln blir signifikant i samtliga modeller, men det var svårare att hitta ett samband mellan enbart antal barn och proportionen döda barn. Själva logistiska modellerna som är skapade är tyvärr svåra att tolka in mer i. Dels är det endast tre stycken som har en signifikant skattning på något utom interceptet, men även för att de parameterskattningarna hade ganska stora standardavvikelser som kan skapa problem vid eventuell tolkning.

Även om det är ett stort material som tillhandahållits kan ibland urvalet kännas lite smått. I figur 6 och tabell 3 ser vi proportionen döda barn för varje kull barn som föddes i materialet. Här minskar antalet barn i varje kull och det är endast en person av totalt 7403 stycken som fått 21 barn. Detta betyder att i vissa lägen, som i ovan nämnda figur, blir ett utslag för stort då det finns väldigt få observationer och resultat då endast överspeglar dem. Till exempel dör en tredjedel av alla barn som är det 17:e i sin familj, dock är det endast nio familjer som får så många barn och därmed är det tre barn som dör.

6. Referenslista

Tryckta källor

Coxe, S., West, S., Aiken L. (2009). The Analysis of Count Data: A Gentle Introduction to Poisson Regression and Its Alternatives. *Journal of Personality Assessments* 91(2), 121-136

Dobson, A., *An Introduction To Generalized Linear Models,* Chapman & Hall/CRC 2002

Klienbaum, D., Kupper, L., Muller, K., Nizam, A., *Applied Regression Analysis and Other Multivariate Methods,* Duxbury Press 1998

Lindsey, James K., *Applying Generalized Linear Models,* Springer-Verlag 1997

Olsson, U., *Generalized Linear Models,* Studentlitteratur AB 2007

McCulloch, C., Searle, S., Neuhaus, J., *Generalized, Linear, and Mixed Models,* Wiley 2008

Internetkällor

Demografiska Databasen,

http://www.ddb.umu.se, 5/4 2010

Demografiska Databasen, *Kyrkböcker – församlingsinformation och registreringsperioder* http://www.ddb.umu.se/material/kyrkbok 26/4 2010

Skatteverket,

http://www.skatteverket.se/privat/folkbokforing/omfolkbokforing/folkbokforingenigaridag/folkbokforingenshistoria.4.18e1b10334ebe8bc80003006.html, 10/3 2010

Skatteverket,

http://www.skatteverket.se/privat/folkbokforing/omfolkbokforing/folkbokforingenigaridag/sverigesforsamlingargenomtiderna.4.18e1b10334ebe8bc80003817.html, 10/3 2010

7. Appendix

7.1 Appendix A

Tabell 8 visar modell 1: Barn.pois1 ← glm(AntalBarn ~ as.factor(Församling)+ as.factor(Socgrupp)+ as.factor(Livslängd)+Propdoda+År, data = Datat, family = poisson)

	Estimat	Std.avvikelse	z-värde	Pr(>|z|)
Intercept	4,042	0,413	9,777	<2e-16
as.factor(Församling)[T.30]	-0,120	0,063	-1,905	0,057
as.factor(Församling)[T.35]	-0,233	0,052	-4,520	0,000
as.factor(Socgrupp)[T.2]	-0,100	0,197	-0,509	0,611
as.factor(Socgrupp)[T.3]	0,021	0,180	0,117	0,907
as.factor(Socgrupp)[T.4]	-0,321	0,220	-1,457	0,145
as.factor(Socgrupp)[T.5]	-0,057	0,192	-0,298	0,766
as.factor(Socgrupp)[T.6]	-0,104	0,183	-0,569	0,570
as.factor(Socgrupp)[T.7]	-0,290	0,182	-1,593	0,111
as.factor(Socgrupp)[T.8]	-0,095	0,186	-0,51	0,610
as.factor(Socgrupp)[T.9]	-0,221	0,197	-1,119	0,263
as.factor(Socgrupp)[T.10]	0,102	0,181	0,56	0,575
as.factor(Livslängd)[T.2]	0,109	0,033	3,308	0,001
as.factor(Livslängd)[T.3]	0,152	0,033	4,642	0,000
as.factor(Livslängd)[T.4]	0,109	0,032	3,399	0,001
as.factor(Livslängd)[T.5]	0,121	0,032	3,733	0,000
as.factor(Livslängd)[T.6]	0,140	0,034	4,065	0,000
as.factor(Livslängd)[T.7]	0,227	0,054	4,176	0,000
Propdoda	0,188	0,059	3,207	0,001
År	-0,0012	0,000	-5,985	0,000
Null deviance: 7315.3 on 4061 degrees of freedom				
Residual deviance: 6993.9 on 4042 degrees of freedom				
Number of Fisher Scoring iterations: 5	AIC: 21380			

Tabell 9 visar modell 2: Barn.pois2←glm(AntalBarn~as.factor(Församling)+as.factor(Livslängd)+År+Propdoda,data = Datat, family = poisson)

	Estimat	Std.avvikelse	z-värde	Pr(>|z|)
Intercept	4,947	0,344	14,396	<2e-16
as.factor(Församling)[T.30]	-0,138	0,0631	-2,187	0,029
as.factor(Församling)[T.35]	-0,271	0,0514	-5,276	0,000
as.factor(Livslängd)[T.2]	0,109	0,033	3,286	0,001
as.factor(Livslängd)[T.3]	0,154	0,0327	4,722	0,000
as.factor(Livslängd)[T.4]	0,110	0,032	3,435	0,001
as.factor(Livslängd)[T.5]	0,121	0,0322	3,765	0,000
as.factor(Livslängd)[T.6]	0,141	0,034	4,100	0,000
as.factor(Livslängd)[T.7]	0,223	0,054	4,117	0,000
Propdoda	0,148	0,0578	2,564	0,010
År	-0,0017	0,000183	-9,054	<2e-16
Null deviance: 7315.3 on 4061 degrees of freedom				
Residual deviance: 7159.4 on 4051 degrees of freedom				
Number of Fisher Scoring iterations: 4	AIC: 21527			

Tabell 10 visar modell 3: Barn.negbin1 ← glm.nb(AntalBarn ~ as.factor(Församling) +as.factor(Socgrupp) +as.factor(Livslängd)+ Propdoda +År, data = Datat)

	Estimat	Stdavvikelse	z-värde	Pr(>|z|)
Intercept	4,038	0,539	7,494	0,000
as.factor(Församling)[T.30]	-0,120	0,084	-1,421	0,155
as.factor(Församling)[T.35]	-0,229	0,069	-3,313	0,001
as.factor(Socgrupp)[T.2]	-0,100	0,254	-0,396	0,692
as.factor(Socgrupp)[T.3]	0,017	0,233	0,075	0,940
as.factor(Socgrupp)[T.4]	-0,325	0,279	-1,164	0,244
as.factor(Socgrupp)[T.5]	-0,063	0,248	-0,253	0,801
as.factor(Socgrupp)[T.6]	-0,108	0,236	-0,458	0,647
as.factor(Socgrupp)[T.7]	-0,292	0,235	-1,243	0,214
as.factor(Socgrupp)[T.8]	-0,098	0,241	-0,407	0,684
as.factor(Socgrupp)[T.9]	-0,223	0,253	-0,88	0,379
as.factor(Socgrupp)[T.10]	0,099	0,235	0,423	0,672
as.factor(Livslängd)[T.2]	0,107	0,042	2,556	0,011
as.factor(Livslängd)[T.3]	0,151	0,042	3,64	0,000
as.factor(Livslängd)[T.4]	0,106	0,041	2,602	0,009
as.factor(Livslängd)[T.5]	0,117	0,041	2,849	0,004
as.factor(Livslängd)[T.6]	0,136	0,044	3,098	0,002
as.factor(Livslängd)[T.7]	0,224	0,071	3,138	0,002
Propdoda	0,202	0,076	2,651	0,008
År	-0,0012	0,000	-4,572	0,000
Null deviance: 4562.8 on 4062 degrees of freedom				
Residual deviance: 4369.5 on 4042 degrees of freedom				
Number of Fisher Scoring iterations: 1				
AIC: 20811	θ-est: 9,167			

Tabell 11 visar modell 4: Barn.negbin2 ← glm.nb(AntalBarn ~ as.factor(Församling)+as.factor(Livslängd) +Propdoda+År, data = Datat)

	Estimat	Std.avvikelse	z-värde	Pr(>|z|)
Intercept	4,942	0,456	10,848	<2e-16
as.factor(Församling)[T.30]	-0,135	0,0854	-1,579	0,114
as.factor(Församling)[T.35]	-0,266	0,0700	-3,803	0,000
as.factor(Livslängd)[T.2]	0,107	0,0424	2,531	0,011
as.factor(Livslängd)[T.3]	0,152	0,042	3,628	0,000
as.factor(Livslängd)[T.4]	0,108	0,0411	2,614	0,009
as.factor(Livslängd)[T.5]	0,119	0,0414	2,874	0,004
as.factor(Livslängd)[T.6]	0,138	0,0443	3,124	0,002
as.factor(Livslängd)[T.7]	0,221	0,0723	3,060	0,002
Propdoda	0,171	0,0764	2,233	0,026
År	-0,0017	0,000243	-6,830	0,000
Null deviance: 4453,8 on 4061 degrees of freedom				
Residual deviance: 4364,1 on 4051 degrees of freedom				
Number of Fisher Scoring iterations: 1				
AIC: 20893	θ-est: 8,589			

Tabell 12 visar modell 5: Barn.trunkpois1 ← vglm(AntalBarn ~ as.factor(Församling) +as.factor(Socgrupp)+ as.factor(Livslängd)+Propdoda+År, data = Datat, family = pospoisson)

	Estimat	Std.avvikelse	t-värde	Pr(>|t|)
Intercept	4,066	0,416	9,777	<2e-16
as.factor(Församling)[T.30]	-0,121	0,0634	-1,915	0,056
as.factor(Församling)[T.35]	-0,235	0,0517	-4,555	0,000
As.factor(Socgrupp)[T.2]	-0,102	0,198	-0,514	0,607
As.factor(Socgrupp)[T.3]	0,0214	0,181	0,118	0,906
As.factor(Socgrupp)[T.4]	-0,331	0,223	-1,485	0,138
As.factor(Socgrupp)[T.5]	-0,0579	0,193	-0,300	0,764
As.factor(Socgrupp)[T.6]	-0,106	0,184	-0,574	0,566
As.factor(Socgrupp)[T.7]	-0,299	0,183	-1,630	0,103
As.factor(Socgrupp)[T.8]	-0,0968	0,188	-0,516	0,606
As.factor(Socgrupp)[T.9]	-0,225	0,199	-1,132	0,258
As.factor(Socgrupp)[T.10]	0,102	0,183	0,561	0,575
As.factor(Livslängd)[T.2]	0,112	0,0335	3,341	0,001
As.factor(Livslängd)[T.3]	0,155	0,0331	4,679	0,000
As.factor(Livslängd)[T.4]	0,112	0,0325	3,434	0,001
As.factor(Livslängd)[T.5]	0,123	0,0327	3,77	0,000
As.factor(Livslängd)[T.6]	0,143	0,0348	4,102	0,000
As.factor(Livslängd)[T.7]	0,230	0,0547	4,216	0,000
Propdoda	0,191	0,059	3,240	0,001
År	-0,0012	0,000198	-6,03	<2e-16
Log-likelihood: -10659,428 on 4042 degrees of freedom				
Number of iterations: 3	AIC: 21358,96			

Tabell 13 visar modell 6: Barn.trunkpois2← vglm(AntalBarn ~ as.factor(Församling) +as.factor(Livslängd)+Propdoda +År, data = Datat, family = pospoisson)

	Estimat	Std.avvikelse	t-värde	Pr(>|t|)
Intercept	4,981	0,345	14,4142	<2e-16
as.factor(Församling)[T.30]	-0,139	0,0633	-2,1991	0,028
as.factor(Församling)[T.35]	-0,274	0,0515	-5,3176	0,000
as.factor(Livslängd)[T.2]	0,111	0,0334	3,3170	0,001
as.factor(Livslängd)[T.3]	0,157	0,0331	4,7575	0,000
as.factor(Livslängd)[T.4]	0,113	0,0324	3,4691	0,001
as.factor(Livslängd)[T.5]	0,124	0,0326	3,8000	0,000
as.factor(Livslängd)[T.6]	0,144	0,0347	4,1368	0,000
as.factor(Livslängd)[T.7]	0,227	0,0545	4,1548	0,000
Propdoda	0,150	0,0581	2,5822	0,010
År	-0,0017	0,000184	-9,1105	<2e-16
Log-likelihood: -10743,95 on 4051 degrees of freedom				
Number of iterations: 3	AIC: 21509,90			

7.2 Appendix B

Tabell 14 visar summering av vardera logistisk modell

Antal barn	Koefficient	Estimat	Std.avvikelse	z-värde	Pr(>|z|)
1	(Intercept)	2,514	0,046	55,147	<2e-16
	Propdöda1	-0,158	0,193	-0,819	0,413
2	(Intercept)	2,558	0,049	52,247	<2e-16
	Propdöda2	-0,396	0,282	-1,404	0,160
3	(Intercept)	2,383	0,048	49,845	<2e-16
	Propdöda3	0,439	0,403	1,089	0,276
4	(Intercept)	2,060	0,044	46,525	<2e-16
	Propdöda4	-0,428	0,362	-1,184	0,236
5	(Intercept)	1,799	0,043	41,639	<2e-16
	Propdöda5	0,141	0,416	0,337	0,736
6	(Intercept)	1,408	0,041	34,003	<2e-16
	Propdöda6	0,167	0,422	0,397	0,691
7	(Intercept)	1,064	0,043	24,921	<2e-16
	Propdöda7	0,554	0,457	1,214	0,225
8	(Intercept)	0,757	0,047	16,193	<2e-16
	Propdöda8	-0,052	0,482	-0,109	0,913
9	(Intercept)	0,423	0,055	7,706	0,000
	Propdöda9	0,372	0,574	0,648	0,517
10	(Intercept)	0,221	0,070	3,167	0,002
	Propdöda10	0,690	0,714	0,966	0,334
11	(Intercept)	0,103	0,094	1,097	0,273
	Propdöda11	0,481	0,891	0,539	0,590
12	(Intercept)	0,251	0,129	1,942	0,052
	Propdöda12	-0,840	1,160	-0,725	0,469
13	(Intercept)	0,015	0,169	0,088	0,930
	Propdöda13	-0,892	1,477	-0,604	0,546
14	(Intercept)	-0,030	0,244	-0,122	0,903
	Propdöda14	0,585	2,342	0,250	0,803
15	(Intercept)	-0,044	0,347	-0,127	0,899
	Propdöda15	-0,913	3,369	-0,271	0,786