Stockholms universitet

Bibelbältet, Brott och Aborter

Oskar Fryklöf

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2017:2
Matematisk statistik
Juni 2017

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholms universitet

Matematisk statistik
Stockholms universitet
Kandidatuppsats **2017:2**
http://www.math.su.se/matstat

Bibelbältet, Brott och Aborter

Oskar Fryklöf*

Juni 2017

Sammanfattning

Effekterna av abortförbud i länder som USA och Rumänien har visats ha ödesdigra följder för respektive samhälle. Genom att använda tre olika regressionanpassningar testas möjligheten att skapa en prediktionsmodell för att kunna avgöra hur många aborter som kommer genomföras i Sverige givet vissa samhällsomständigheter. Diagnostik för att hitta den bästa modellen tillämpas i stor grad för att sedan jämföra modellen med viss forskning på området. Två av modellerna som tillämpas tillhör specialfall av de allmänna modellantagandena där det lämpligaste modellvalet visar sig vara använda negativ binomial regression eftersom svarsvariabeln tillhör räknedata och datamaterialet är alldeles för spritt för att använda en Poissonmodell. Resultatet är en prediktionsmodell som bland annat visar att ett specifikt kluster av regioner bidrar negativt till aborttalet och att antal brott har en signifikant positiv effekt på antalet aborter. Modellen kan användas i sammanhang då antalet aborter för ett specifikt år, betingat av vissa omständigheter i samhället, ska skattas.

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

Bibelbältet, Brott och Aborter
En regressionsanalys av antalet aborter i Sverige

Oskar Fryklöf

4 juni 2017

Innehåll

1 Inledning **4**

2 Datakällor **4**

3 Teori **5**
3.1 Linjär regression 5
3.2 Förklaringsgrad, R^2 6
3.3 AIC 6
3.4 Allmäna modeller, GLM 7
3.5 Fördelningsantagande för GLM 7
3.6 IRLS - Iterated Reweighted Least Squares 8
3.7 Specialfall av GLM 9
3.7.1 Poisson regression 9
3.7.2 Poisson regression med offset 9
3.7.3 Överspridning 10
3.7.4 Negativ Binomial regression 11
3.7.5 LOOCV 11

4 Data **12**
4.1 Beskrivning av data 12
4.2 Varför ta hänsyn till antalet förlossningar 14

5 Resultat **16**
5.1 Explorativ dataanalys och visualisering 16
5.1.1 Korrelationsanalys 16
5.1.2 Yngre personer och graviditetslängder 17
5.2 Regressionsanalys 17
5.2.1 Enkel linjär regression 18
5.2.2 Multipel linjär regression 18

5.2.3 Poisson-regression 19
5.2.4 NB-regression 20
5.3 Analys 21
5.3.1 Diskussion 22

6 Slutsats **23**

7 Vidare **24**

8 Tabeller **25**

9 Grafer **27**

Sammanfattning

Effekterna av abortförbud i länder som USA och Rumänien har visats ha ödesdigra följder för respektive samhälle. Genom att använda tre olika regressionanpassningar testas möjligheten att skapa en prediktionsmodell för att kunna avgöra hur många aborter som kommer genomföras i Sverige givet vissa samhällsomständigheter. Diagnostik för att hitta den bästa modellen tillämpas i stor grad för att sedan jämföra modellen med viss forskning på området. Två av modellerna som tillämpas tillhör specialfall av de allmänna modellantagandena där det lämpligaste modellvalet visar sig vara använda negativ binomial regression eftersom svarsvariabeln tillhör räknedata och datamaterialet är alldeles för spritt för att använda en Poissonmodell. Resultatet är en prediktionsmodell som bland annat visar att ett specifikt kluster av regioner bidrar negativt till aborttalet och att antal brott har en signifikant positiv effekt på antalet aborter. Modellen kan användas i sammanhang då antalet aborter för ett specifikt år, betingat av vissa omständigheter i samhället, ska skattas.

Abstract

The effects of abortion bans in countries like the United States and Romania have been shown to have fatal consequences for the respective community. By using three different regression adjustments, the possibility of creating a prediction model is tested to determine how many abortions will be implemented in Sweden given certain social circumstances. Diagnostics to find the best model is largely used to compare the model with some research in the field. Two of the models apply to special cases of the general model assumptions where the most appropriate model selection appears to be a negative binomial regression because the response variable belongs to frequency data and the data is too widely spread to use a Poisson model. The result is a prediction model that shows, among other things, that a specific cluster of regions contributes negatively to the abortion rate and that the number of crimes has a significant positive effect on the number of abortions. The model can be used in policy making when the number of abortions, given some circumstances in the society, is to be estimated.

1 Inledning

1973 föll en avgörande dom i USA som sederemera blev känd som fallet 'Roe vs. Wade' (RvW). Effekten av domen innebar bland annat att kvinnor i USA fick bättre möjligheter att genomföra aborter. 16 år senare var kriminaliteten som högst i USA på länge. Experter kallade fenomenet "det kriminella monstret" och det påstods att kriminaliteten enbart skulle öka. Konsensus rådde bland experterna. De hade fel. Kriminialiteten började nämilgen avta 1992 och avtog som skarpast 1995. Ekonomiska forskare kunde senare koppla utgången av fallet RvW till den avtagande kriminaliteten [1]. Teorin är mycket simpel - i och med abortmöjligheten föddes färre barn i en ekonomisk misär vilket innebar färre kriminella sextonåringar. Andreas Madestam och Emilia Simionova skrev rapporten 'Children of the pill' där de socioekonomiska effekterna av subventionerade preventivmedel ledde till att kvinnorna i områden med subventionen bland annat hade högre utbildningsnivå och var mindre benägna att röka under graviditeten[2]. I och med utvecklingen inom medicin är det möjligt att utifrån fostret se vilka genetiska drag ett barn kommer att ha. Detta leder till etiska frågeställningar som att motiven bakom aborten inte enbart handlar om föräldraranas kapacitet att ta hand om barnet.

Den 1 oktober 2016 uppgav SvD att en av kandidaterna till nobelpriset i kemi var forskare som hittat en metod som bland annat upptäcker förekomsten av Down Syndrom i fosterstadiet [3]. Ponera att i framtiden mer sofistikerade metoder utvecklas vilka innebär att mer detaljerade drag hos individer kan förutspås hos barn. Detta skulle kunna finnas som beslutsunderlag för en kvinna att genomföra en abort. Med metoder som CRISPR är forskningen redan i det läge att arvsmassan går att förändra vilket innebär att människor kommer att kunna designas i framtiden[4]. Ska en abort få genomföras för att en parameter inte överensstämmer med föräldrarnas förväntan? Just dessa etiska frågor kommer inte denna rapport att beröra - enbart upplysa om. Abort är inte ett självklart ingrepp eftersom det, till skillnad från tex. en cancerbehandlning eller en hjärttransplantation, sällan är ett botemedel. Om en person drabbas av cancer är det självklara att försöka bota sjukdomen med hjälp av ett medicinskt ingrepp. En abort är ett medicinskt ingrepp men det genomförs inte med samma motiv som något annat medicinskt ingrepp som finns att tillgå idag.

I denna rapport kommer statistiska metoder att tillämpas för att se om vissa kovariater har en signifikant effekt på antalet aborter som genomförs i Sveriges 21 län.

2 Datakällor

Datan som använts i analysen är en sammanslagning av dataset från fyra olika källor. Dessa fyra set har slagits samman med avseende på år, län samt åldersintervall. Källor samt förklaring av respektive datakälla framgår i Tabell 5 i Avsnitt 8. Datan kommer från SCB, Socialstyrelsen, BRÅ och Ekonomifakta. Analysen görs med hjälp av mjukvaran R. Datan är indelad i åldersintervall

om fem år från $\leq$24 år till 35$\leq$ år. Eftersom analysen genomförs på paneldata över en 10-års period och ej specifika individer finns därför en sannolikhet att en person som genomför en abort och tillhör en åldersgrupp kan genomföra en abort vid ett senare tillfälle och därmed bidra till antalet aborter för en annan åldersgrupp. Denna företeelse beaktas ej i analysen eftersom specifika individer ej är av intresse. Detta innebär även att datan som är insamlad nödvändigtvis inte är samma individer eftersom datan kommer från olika källor. Vad som däremot är av intresse är de sociala förutsättningarna för individerna i varje åldersgrupp och län och se hur pass väl de sociala förutsättningarna kan prediktera antalet aborter. Alla koder samt filer för genomförandet av analysen förvaras på Github [5].

3 Teori

Förklaring av alla teorier som tillämpas. Källhänvisningar finns i respektive avsnitt.

3.1 Linjär regression

Anta n observationer, $x_i,\ y_i$ där y_i är en observation från den stokastiska variabeln Y_i där $\mu_i \equiv \mathbb{E}[Y_i]$. Anta vidare att en lämplig modell för förhållandet mellan x och y definieras som:

$$Y_i = \mu_i + \epsilon_i \text{ där } \mu_i = x_i\beta. \tag{1}$$

β är en okänd parameter i ekvationen (3.1) som skattas med hjälpa av Minstakvadratskattning och feltermerna, ϵ_i, antas vara oberoende likafördelade variabler med fördelning $\epsilon_i \sim N(0, \sigma^2)$ (s. 2 i [6]). Fem antaganden görs i huvudsak, listade i avtagande betydelse (s.45 i [7]):

a. Validitet: Rätt data används för att angripa problemet.

b. Linearitet: Förhållandet mellan kovariaterna och responsvariabeln antas vara linjärt, om antagandet bryts är det lämpligare att tex. logtransformera datan.

c. Oberoende: alla kovariater antas vara oberoende av varandra.

d. Homoskedasticitet: Alla ϵ antas ha konstant variants.

e. Normalitet: Alla $\epsilon_i \sim N(0, \sigma^2)$.

Regressionsmodellen definieras som

$$y_i = \alpha + \sum_{j=1}^{k} \beta_j x_{ij} + \epsilon_i \ \text{ där } \ i \in \{1, \ldots, n\} \ \& \ \epsilon \sim \mathrm{N}(0, \sigma.$$

$j \in \{1, \ldots, k\}$, om $k = 1$ fås enkel linjär regression, och om $k > 1$ fås multipel linjär regression där j är ett index för antalet förklarande variabler i modellen.

3.2 Förklaringsgrad, R^2

R^2 är en skattning av den andelen av variationen i datan som förklaras av regression och definieras som

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \hat{\epsilon}_i^2/n}{\sum_{i=1}^{n} (y_i - \overline{y})/n}$$

där $\overline{y}$ är medelvärdet av alla observationer y_i, n är antalet observationer och $\hat{\epsilon}$ är feltermen från regressionen. Kvoten i detta fall är i stort sett en skattning av andelen av variansen som *ej* förklaras av regression. Förklaringsgraden är ett användbart mått för att jämföra styrkan mellan olika modeller. Ett högt R^2 är ofta associerat med att modellen är väl anpassad till modellen varför den modell med högst R^2 oftast är den som väljs.

3.3 AIC

Akaikes informationskriterium används för att jämföra modeller. I klassiska linjära modeller förväntas anpassningen att öka när en ytterligare parameter av rent brus läggs till. Läggs en prediktor till förväntas avvikelsen att minska med 1, på samma sätt väntas avvikelsen att minska med k om k prediktorer läggs till. Mer exakt innebär det att avvikelsen reduceras med en mängd motsvarande en χ^2-fördelning med k frihetsgrader. Om k prediktorer läggs till och avvikelsen reduceras med signifikant mer än k kan man dra slutsatsen att den observerade förbättringen i prediktiv förmåga är statistiskt signifikant. Således kan ekvationen

$$\text{justerade avvikelsen} = \text{avvikelsen} + \text{antal parametrar}$$

användas som ett justerat mått som approximativt tar hänsyn till ökningen i anpassningen som åstadkoms med att lägga till en prediktor i modellen - vilket kan jämföras med förklaringsgraden för enkel linjär regression.
Nästa steg, utöver att se om en förbättring är statistiskt signifikant, är att se om det är skattat att höja 'out-of-sample' prediktionsförmågan. I snitt måste en ytterligare prediktor reducera avvikelsen med två för att förbättra anpassningen till ny data. Akaikes Informations kriterium (AIC) är definierat som

$$AIC = \text{avvikelsen} + 2(\text{antal parametrar}) = \text{justerade avvikelsen} + \text{antal parametrar}.$$

I klassisk regression eller GLM färväntas en ny model att minska 'out-of-sample' prediktionsfelet om AIC minskar (s. 525 [7]). Baserat på log likelihooden fås således AIC till

$$AIC = -2\log(L(M)) + 2k(M)$$

där $L(M)$ är likelihooden för model M.

3.4 Allmäna modeller, GLM

För att kunna anpassa modeller där fördelningsantagandet för svarsvariablen är annan än normal och för att ha en nivå av icke-linearitet i modellen formuleras den funtamentala strukturen för en GLM som

$$g(\mu_i) = \mathbf{X}_i\boldsymbol{\beta}$$

där

a. $\mu_i \equiv \mathbb{E}[Y_i]$

b. g är en slät monotonisk 'länkfunktion'

c. $\mathbf{X}_i$ är den i'te raden i modelmatrisen $\mathbf{X}$

d. $\boldsymbol{\beta}$ är en vektor av okända parametrar.

Utöver ovanstående antas även att Y_i är oberoende och att

$$Y_i \sim \text{någon fördelning tillhörande exponentialfamiljen.}$$

GLM uttrycks i temrer av linjära prediktorer $\mathbf{X}\boldsymbol{\beta}$ vilket leder till att mycket av ramverket för linjär modellering överförs till GLM med skillnaden att en länk-funktion och fördelning måste väljas(s. 59 [6]).

3.5 Fördelningsantagande för GLM

Skattningar och inferens för GLM, är baserat på teori för ML-skattningar. Responsvariabeln i en GLM kan tillhöra en fördelning från exponentialfamiljen och kan skrivas som

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

där b, a och c är godtyckliga funktioner, ϕ en godtycklig skalär, och θ kännetecknas som 'kanonisk parameter' av fördelningen vilken även är känd som 'parametern av intresse'. Det är möjligt att erhålla generella uttryck för väntevärdet och variansen från fördelningar tillhörande exponentialfamiljen i termer av b, a och ϕ . Log-likelihooden av θ, givet något y, är $\log[f_\theta(y)]$. Dvs likelihood-kerneln fås som

$$l(\theta) = \frac{\{y\theta - b(\theta)\}}{a(\phi)}$$

s.a

$$\frac{\partial l}{\partial \theta} = \frac{\{y - b'(\theta)\}}{a(\phi)}.$$

Om l behandlas som en slumpmässig variabel och genom att byta ut y mot Y kan väntevärdet för $\frac{\partial l}{\partial \theta}$ räknas ut likt

$$\mathbb{E}\left(\frac{\partial l}{\partial \theta}\right) = \frac{\{\mathbb{E}[Y] - b'(\theta)\}}{a(\phi)}.$$

Genom att tillämpa det generella resultatet att $\mathbb{E}\left(\frac{\partial l}{\partial \theta}\right) = 0$ fås

$$\mathbb{E}[Y] = b'(\theta).$$

Dvs. medelvärdet, för någon slumpvariabel tillhörande en exponentialfamilj, ges av den första derivatan av b med avseende på θ (s.62 [6]). Om l deriveras en ytterligare gång fås

$$\frac{\partial^2 l}{\partial \theta^2} = -b''(\theta)/a(\phi),$$

tillämpning av $\mathbb{E}\left(\frac{\partial^2 l}{\partial \theta^2}\right) = -\mathbb{E}[\left(\frac{\partial l}{\partial \theta}\right)^2]$ ger

$$-b''(\theta)/a(\phi) = \mathbb{E}[\{\mathbb{E}[Y] - b'(\theta)\}^2]/a(\phi)^2,$$

vilket ger

$$var(Y) = b''(\theta)a(\phi).$$

På detta vis kan alltså väntevärde och varians räknas ut för alla slumpvariabler med en fördelning tillhörande exponentialfamiljen (s.62 [6]).

3.6 IRLS - Iterated Reweighted Least Squares

IRLS-algoritmen är en förenklad Maximum-Likelihood aligoritm och härledningen påminner mycket om härledningen för Newton Raphson-modeller. Likt ML-slkattningar är IRLS-metoden baserat på sannolikhetsfördelningar. Schemat för IRLS-alogritmen i generella fall ser ut likt

a. Inled med $\mathbb{E}[Y] = \mu$ & den linjära prediktorn η eller $g(\mu)$.

b. Räkna ut vikterna som $w^{-1} = Vg'(\mu)^2$ där V är variansen och $g'(\mu)$ är derivatan av länkfunktionen.

c. Räkna ut en lösningsrespons, en Taylorlinearisering av log-likelihoodfunktionen med standardform
$$z = \eta + (y - \mu)g'(\mu)$$

d. Tillämpa regression av z på prediktorerna $X_1, \ldots, X_n$ med vikter för att uppdatera paramerskattningarna β.

e. Räkna ut η baserat på estimaten.

f. Räkna ut μ som $g^{-1}(\mu)$.

g. Räkna ut avvikelsen eller log-likelihooden

h. Upprepa processen tills differensen mellan avvikelsen eller log-likelihooden är nära 0.

Denna algoritm används generellt av mjukvara då någon funktion för allmän regression tillämpas för att skatta koefficienterna i modellen då analys med hjälp av GLM önskas (s.26 [8]).

3.7 Specialfall av GLM

Vissa generella modeller beskrivs som specialfall där två av de olika modellerna samt analytiska verktyg definieras nedan.

3.7.1 Poisson regression

Poisson-fördelningen används för att modellera variationen i räknedata. Varje enhet i överensstämmer med en inställning (vanligtvis ett geografiskt område eller ett tidsintervall) för vilket y_i händelser observeras. Variationen i y kan förklaras med linjära prediktorer X. Modellen för vanlig Poisson-regression har formen

$$y_i \sim Poisson(\theta_i).$$

Där sannolikhetsfördelningen samt väntevärde och varians i det generella fallet med θ som intensitet är

$$f(k;\theta) = P(Y = k) = \frac{\theta^k e^{-\theta}}{k!}$$

$$\mathbb{E}[Y] = \mathrm{Var}[Y] = \theta.$$

Parametern θ_i måste vara positiv varför det är rimligt att anpassa en linjär regression på logaritmisk skala:

$$\theta_i = \exp(X_i\beta).$$

Koefficienterna β kan exponentieras och behandlas som multiplikativa effekter vilket leder till att modellen för k parametrar kan se ut som

$$y_i \sim Poisson(\exp(\alpha + \sum_{j=1}^{k} \beta_j X_{ji})),$$

där varje koefficient ger den förväntade förändringen i y på logaritmisk skala för varje värde X_{ij}. Förändringen av y är alltså $\exp(\beta_j X_{ij})$ för något $i \in (1, \ldots, n)$ och $j \in (1, \ldots, k)$(s. 111[7]).

3.7.2 Poisson regression med offset

I de flesta tillämpningar av Poissonregression kan antalen tolkas relativt någon basnivå eller "exponering" tex. befolkningsmängden i varje region. I den generella Poissonmodellen anses y_i som det antal fall i en process med intensitet θ_i och exponering u_i.

$$y_i \sim Poisson(u_i\theta_i),$$

där, som tidigare, $\theta_i = exp(X_i\beta)$. Logaritmen av exponeringen, $\log(u_i)$, kallas offset i GLM-terminologi. Inkludering av offset-variablen i modellen är likställt med att inkludera den som en regressionsprediktor men med dess koefficient fixerat till värdet 1. Ett alternativ är att inkludera den som en prediktor och låta dess koefficient skattas från datan (s.111 [7]).

3.7.3 Överspridning

Eftersom Poisson regression ej förser analysen med en oberoende variansparameter σ kan det resultera i att variansparametern är överspridd varför det undersöks ifall det föreligger överspridning bland de predikterade värderna. För $y_i \sim Po(u_i\theta_i)$ gäller att $E[y_i] = V[y_i] = u_i\theta_i$, dvs. $sd[y_i] = \sqrt{u_i\theta_i}$. De standardiserade residualerna kan definieras likt

$$\begin{aligned} z_i &= \frac{y_i - \hat{y}_i}{sd(\hat{y}_i)} \\ &= \frac{y_i - u_i\theta_i}{\sqrt{u_i\theta_i}} \end{aligned}$$

Om Poissonmodellen är sann ska alla z_i vara approximativt oberoende med väntevärde 0 och standardavvikelse 1. Om överspridning föreligger förväntas absolutbeloppen av alla z_i att vara större vilket reflekterar variationen som ligger bortom vad Poisson modeller kan prediktera. Överspridning kan testas i klassisk Poisson regression genom att räkna ut kvadratsummorna av de n standardiserade residualerna, $\sum_{i=1}^{n} z_i^2$, och jämföra det resultatet med en χ^2_{n-k}-fördelning vilket är väntat under modellantagandet eftersom

$$\begin{aligned} y_i &\sim \mathrm{N}(u_i\theta_i, u_i\theta_i) \\ \frac{y_i - u_i\theta_i}{\sqrt{u_i\theta_i}} &\sim \mathrm{N}(0,1) \end{aligned}$$

så att

$$z_i = \frac{y_i - u_i\theta_i}{\sqrt{u_i\theta_i}} \sim \mathrm{N}(0,1) \Rightarrow \sum_{i=1}^{n} z_i^2 \sim \chi^2_{n-k}. \tag{2}$$

$n-k$ istället för n frihetsgrader används för att ta hänsyn till de k regressionskoefficienterna. χ^2_{n-k}-fördelningen har ett genomsnitligt värde om $n-k$, så kvoten

$$\text{uppskattad överspridning} = \frac{1}{n-k}\sum_{i=1}^{n} z_i^2 \tag{3}$$

är en summering av överspridningen i datan jämfört med den anpassade modellen. Ett sätt att hantera överspridning är att justera inferensen genom att multiplicera alla regressionsfel med $\sqrt{\frac{1}{n-k}\sum_{i=1}^{n} z_i^2}$, ett enklare sätt är dock att anpassa en kvasi-Poission-modell alt. negativ binomial-modell (s.114 [7]).

3.7.4 Negativ Binomial regression

Negativ Binomial regression definieras av samma mängd som en Poissonregression. Den stora skillnaden är att NB-regression tillåter för överspridd data. På detta vis kan frekvensdata modelleras medan antagandet om att $E[Y] = V[Y]$ kan mildras. Detta är användbart för observerad data som är överspritt likt

$$Var[Y] = \delta E[Y], \text{ då } \delta >> 1.$$

Sannolikhetsfördelningen för en NB-fördelad variabel Y är

$$f(y; r, p) = P(Y = k) = \binom{k + r - 1}{r} p^k (1 - p)^r,$$

vilket som en del av exponentialfamiljen enligt avsnitt 3.5 skrivs

$$f(y; r, p) = \exp\left[y \ln(1 - p) + r(\ln(p)) + \ln \binom{k + r - 1}{r} \right].$$

Genom att dela in uttrycket likt metoden i avsnitt 3.5 räknas väntevärde och varians ut till

$$b'(\theta) = \frac{r(1 - p)}{p} = \mu$$

$$b''(\theta) = \frac{r(1 - p)}{p^2}$$

Dvs. $E[Y] = \mu$. IRLS-algoritmen tillämpas för att anpassa modellen till en NB-regression och eftersom IRLS-algoritmen normalt parametriseras i termer av den anpassade statistikan μ (s.84 [8]) parameteriseras p likt

$$\mu = \frac{r(1 - p)}{p} \iff \cdots \iff p = \frac{1}{1 + \frac{\mu}{r}}$$

Detta p sätts in i variansen $b''(\theta)$ så att

$$b''(\theta) = \frac{r(1 - p)}{p^2} = \cdots = \mu(1 + \frac{\mu}{r}).$$

Detta leder till att $E[Y] = \mu$ och $V[Y] = \mu(1 + \frac{\mu}{r})$ och eftersom länkfunktionen g är logaritmen formuleras modellen som

$$\log(\mu_i) = \alpha + \sum_{j=1}^{k} \beta_j x_{ji}.$$

3.7.5 LOOCV

LOOCV, eller Leave One Out Cross Validation, används för att mäta hur pass väl modellen predikterar den angivna datan. Processen ser ut som följer:

a. Ta bort observation i från datasetet.

b. Anpassa en model för den återstående datan innehållandes $n-1$ observationer.

c. Bilda $s_i^2 = (y_i - \hat{y}_i)^2$, där y_i är de observation som exkluderats och $\hat{y}_i$ är det anpassade värdet. s_i^2 är skattningen av felet.

d. Upprepa för n st observationer för att generera summan av CV_n.

Summan CV_n definieras som

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

CV_n är medelkvadratfelet varför den modell med lägst CV_n predikterar den observerade datan bäst.

4 Data

I denna sektion ges en förklaring till den ursprungliga idén för vilken data som skulle tillämpas och hur nya ideér kom till under arbetes gång. Svarsvariabeln har genom arbetet alltid varit antalet aborter.

4.1 Beskrivning av data

Det finns 1764 observationer från svarsvariabeln $Y_i :=$ "antalet aborter". Varje observation är uppdelad på region, år och åldersintervall och omfattar

a. En av graviditetslängderna, <8 veckor, 8 - 12 veckor eller 12 < veckor.

b. Uppskattat antal arbetslösa kvinnor, ej uppdelat på åldersintervall.

c. Antalet förlossningar för varje län.

d. Hur många registrerade patienter för preventivmedel, ej uppdelat på åldersintervall.

e. Hur många recept av vilken typ av preventivmedel som köpts - om det är långvarigt (till exempel hormonspiral och p-stav) eller kortvarigt (till exempel p-piller), ej uppdelat på åldersintervall.

Det bör uppmärksammas att 'Antalet patienter' $\leq$ 'Antal recept' och att årsintervallet går från 2006-2012. Det ska även uppmärksammas att variabeln Graviditetslängd (GL) ej är av intresse i regressionsanalysen. Anledeningen till detta är att det är uppenbart att antalet aborter avtar i och med att antalet graviditesveckor ökar. Detta leder till att den kortare graviditetslängden har en uppenbart signifikant effekt på antalet aborter. Graviditetslängden har även ett uppenbart förhållande till antalet aborter eftersom det från datan framgår att varje graviditetslängd leder till en abort. I Tabell 1 nedan framgår vad för typ

av data varje observation omfattar och i Tabell 2 nedan finns ett utdrag från datasetet som använts i R. För att ta hänsyn till demografin har även samtliga numeriska prediktionsvariabler dividerats med antalet kvinnor i respektive region för det relevanta åldersintervallet.

Variablen $\mathit{ÅR}$ kan hanteras som antingen en kategorisk variabel eller en intervallsvariabel. Om variabeln hanteras som en faktor kommer dolda effekter att synas i modellen vilket leder till spekulationer kring varför ett annat år var mer signfikant än det tidigare år och det blir svårare att prediktera från modellen. Sätts variabeln istället som linjär kommer de spekulativa aspekterna att utebli, vilket föredras, samt en prediktion i tiden kan genomföras. Variabeln sätts därför som linjär.

Table 1: Uppskattad fördelning samt värdemängd av variablerna

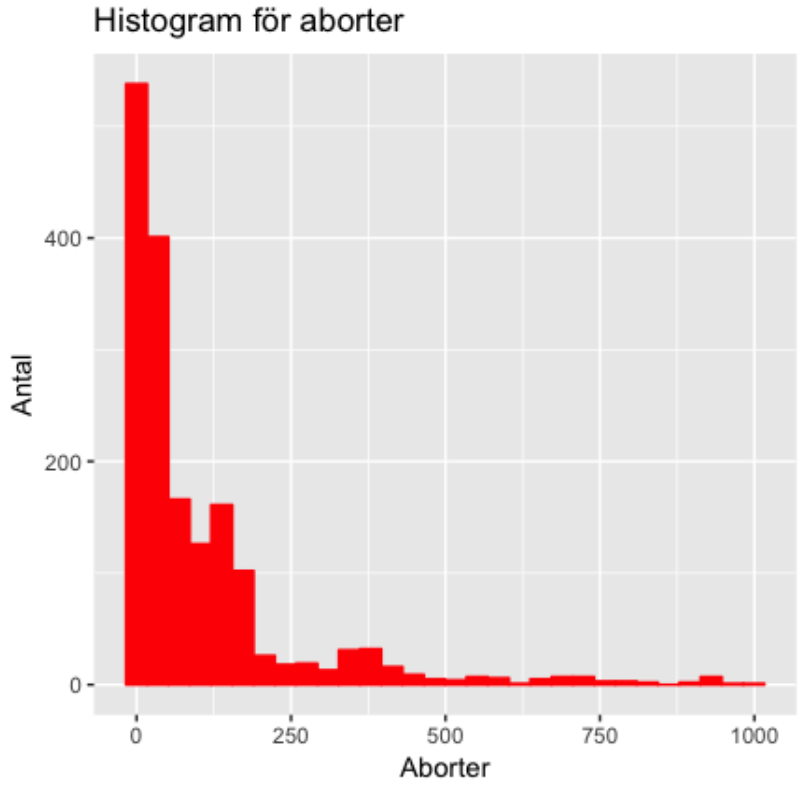
Variabel	**Förklaring**	**Variabeltyp**	**Fördelning**
RE	Region	Kategorisk	Sveriges 21 län
ÅR	Årtal	Intervall	(2006, 2012)
AGE	Ålder	Kategorisk	Åldersintervall om <24, 25-29, 30-34, 35<
GL	Graviditetslängd	Kategorisk	Antal graviditetsveckor <8, 8-11, 12< veckor
AB	Antalet aborter	Intervall	(0, 3072)
GRAV	Förlossningar	Intervall	(73, 10710)
ARB	Antalet arbetslösa	Intervall	(76.32, 16506)
PAK	Patienter till kortvarigt preventivmedel	Intervall	(0.1, 0.53)
RECK	Recept för kortvarigt preventivmedel	Intervall	(0.28, 1.44)
PAL	Patienter till långvarigt preventivmedel	Intervall	(0.01, 0.06)
RECL	Recept för långvarigt preventivmedel	Intervall	(0.01, 0.06)
BRO	Brott	Intervall	(1.01, 6.64)
POP	Population	Intervall	(1255, 234818)

Table 2: R-utskrift med de sex första raderna

	RE	ÅR	AGE	GL	AB	GRAV	PAK	RECK	PAL	RECL	ARB	BRO	POP
1	Blekinge	2006.00	<24	<8 veckor	128.00	220.00	0.48	1.07	0.03	0.03	564.37	1.89	8551.00
2	Blekinge	2006.00	<24	9-11 veckor	59.00	220.00	0.48	1.07	0.03	0.03	564.37	1.89	8551.00
3	Blekinge	2006.00	<24	>12 veckor	23.00	220.00	0.48	1.07	0.03	0.03	564.37	1.89	8551.00
4	Blekinge	2006.00	25-29	>12 veckor	6.00	510.00	0.44	1.00	0.03	0.03	244.13	4.36	3699.00
5	Blekinge	2006.00	25-29	<8 veckor	62.00	510.00	0.44	1.00	0.03	0.03	244.13	4.36	3699.00
6	Blekinge	2006.00	25-29	9-11 veckor	16.00	510.00	0.44	1.00	0.03	0.03	244.13	4.36	3699.00

Eftersom Y tillhör frekvensdata är det lämpligaste modellvalet någon variant av en GLM som tar hänsyn till svarsvariablens egenskaper - i sådana fall är de vanligaste modellvallen Poisson- eller Negativ Binomial-regression. I figur 1 nedan framgår spridningen av datan i Y.

Figure 1: Histogram för antalet aborter. Histogrammets x-axel är bruten vid 1000 av visuella skäl.



4.2 Varför ta hänsyn till antalet förlossningar

Eftersom datan är fördelad på regioner och antalet aborter ska predikteras är det uppenbart att alla kovariater tillhörande region Stockholm är signifikanta. Anledningen till detta är att Stockholm är den största regionen sett till antal invånare vilket under antagandet att *antalet aborter är konstanta i respektive region* leder till att Stockholm har flest aborter. Detta framgår också då medelvärdet för antalet aborter i Stockholm under perioden 2006 - 2012 är 836 medan snittet i Västra Götaland är 498 och lägsta snittet tillhör Gotland vilket är 21. För att få rättvisande resultat i analysen tas hänsyn till antalet kvinnor för respektive region. Dock visar det sig att det kan vara rimligare att ta hänsyn till antalet förlossningar istället för antalet kvinnor, alternativt en kombination av dem. En av dessa anledningar är trender som är typiska för åldersgrupper. En typisk sådan trend är att antalet förlossningar toppas av åldersgruppen 30-34. Detta leder till att en del av kohorten med kvinnor över 35 överhuvudtaget inte ingår i gruppen av kvinnor som blir gravida vilket i sin tur leder till att resultaten för den åldersgruppen ej är rättvisande om man tar hänsyn till an-

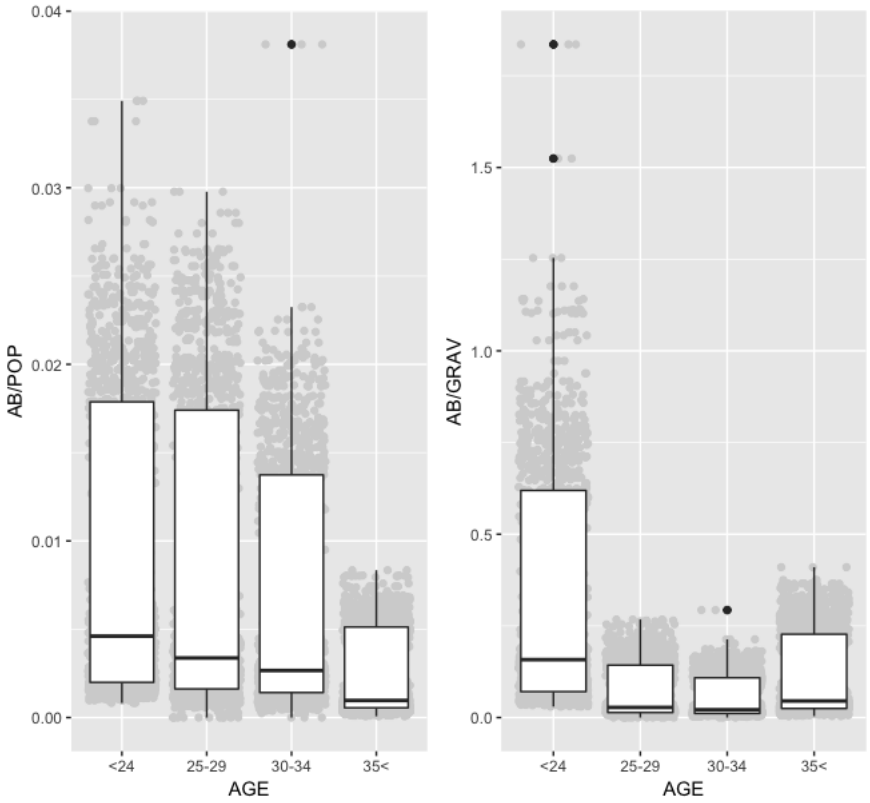
tal kvinnor. Data för antalet gravida finns inte, däremot finns data för antalet förlossningar. Således är det möjligt att anta att antalet gravida kan skattas av antalet förlossningar summerat med antalet aborter, men det avstås ifrån pga. två anledningar:

a. Variansen för variabeln GRAV är ca 3 gånger större än dito för AB

b. Aborterna gäller för de tre första månaderna, en tidsförskjutning uppstår i och med antalet aborter och förlossningar ej behöver höra till samma år.

I figur 4.2 nedan visas en visualisering av åldersfördelningen för de variabler som är rimliga kandidater att tillämpa som offset.

Figure 2: Lådagram där åldersförldelningen visualieras



I figur 4.2 framgår två separata trender. I det vänsta diagrammet har antalet aborter dividerats med antal kvinnor medan det högra visar antalet aborter dividerat med antalet förlossningar. Detta tillvägagångssätt leder även till att y-axeln i det högra diagrammet är ca 10 gånger större eftersom det existerar fler aborter per förlossning jämfört med aborter per invånare. Åldersgrupperna för datasetet tillhörande det högra diagrammet har slagits samman eftersom antalet

registrerade förlossningar börjar på <24 år och sträcker sig till 35< istället för 40<. Det framgår inte från Socialstyrelsen hur gammal den äldsta personen som genomförde en förlossning är, dock är det rimligt att av biologiska skäl anta att maxåldern är ca 50 år och att det är tillräckligt ovanligt förekommande för aborter över 40 år att datainsamlingen begränsats till intervallet 35<. Från graferna i 4.2 väljs således antalet aborter sett till antalet förlossningar, istället för antal kvinnor, eftersom det antas att resultaten kommer vara mer rättvisande och dessutom har lägre varians. Kostnaden för det valet är en analys med färre observationer eftersom datasetet för antal kvinnor är begränsat till sex istället för fyra åldersintervall.

5 Resultat

I detta avsnitt presenteras inledningsvis explorativ dataanalys följt av resultaten från de olika modellanpassningarna. Varje resultatet följs av en kort kommentar utan några större analytiska slutsatser vilka lyfts fram i ett senare avsnitt. Gränsen för felrisken att $H_0 : \beta = 0$ för samtliga kovariater är satt till 5%.

5.1 Explorativ dataanalys och visualisering

Resultatet från dataanalysen som presenteras är de resultat författaren uppfattar som intressanta. Datan styckas upp efter intresse och relevans.

5.1.1 Korrelationsanalys

Korrelationer kan vara missvisande eftersom de kan bero på det logiska felslutet post hoc, trots detta är de användbara eftersom de ger viss information om sambanden mellan olika variabler. I graf 5 från avsnitt 9 framgår en visualisering av korrelationerna. Bland resultaten finns ett svagt negativt samband mellan antalet arbetslösa med brott och kortvariga recept. Korrelationen mellan antalet arbetslösa och antalet aborter är medelstarkt positiv samt antalet arbetslösa med långvarigt recept är svagt negativ. Antalet aborter har en svag negativ korrelation med antalet brott och antalet långvariga recept men en svagt positiv korrelation med antalet kortvariga recept. Korrelationen mellan antalet brott och kortvariga recept är svag. Korrelationen mellan aborter och arbetslöshet verkar ligga i linje med teorin som framgår i [1], speciellt om man antar att fler arbetslösa innebär lägre socioekonomisk säkerhet. Korrelationen mellan antalet arbetslösa och bruk av kortvarigt preventivmedel är intressant eftersom det kan styrkas av artikeln om subventionerade preventivmedel[2]. Ponera att priset för preventivmedel är konstant medan antalet arbetslösa ökar. Detta leder rimligtvis till sämre ekonomiska förutsättningar vilket leder till att konsumenter avstår från köp av preventivmedel. En väntad effekt är då att antalet aborter ökar.

5.1.2 Yngre personer och graviditetslängder

Ett tydligt samband verkar finnas i alla regioner, nämligen att kvoten $\frac{AB}{GRAV} > 0.5$ enbart omfattar åldersgruppen <24 och då graviditetslängden är <8 veckor. En ytterligare undersökning visar att kvoten $\frac{AB}{GRAV}$ är ca 6.5 gånger större för åldersgruppen < 24 år jämfört med åldersgruppen 30-34. Vid det lägre åldersintervallet befinner sig de flesta individer i någon form av ekonomisk osäkerhet varför det inte är ett rationellt beslut att föda ett barn. För det högre åldersintervallet är det rimligt att anta att den ekonomiska osäkerheten har avtagit och fler personer är lämpliga att bli föräldrar. DN skrev 2013 att Swedbank räknat ut att ett barn kostar ca 1.4 milj kronor fram till dagen barnet fyller 18 [9]. Swedbank rekommenderar även par att spara pengar en period för att vara solventa nog för att ha ett barn - en möjlighet som är liten för just den yngre åldersgruppen <24. Vidare är även svenska ungdomar enlig SvD sämst i norden på att skydda sig vid samlag där 21% av de tillfrågade personerna i åldrarna 15-20 år svarade att de inte skyddade sig alls [10]. Det är uppenbart att sämre användning av preventivmedel leder till att kvinnor i den åldersgruppen är mer benägna att bli gravida jämfört med andra åldersgrupper. Detta i kombination med socioekonomiska faktorer leder till att fler aborter genomförs.

5.2 Regressionsanalys

Tre typer av regressionsanalyser tillämpas för att anpassa prediktionsmodeller till datan. Varje typ av regressionsanalys inleds med att testa 'Den fulla modellen' vilken anpassas med alla variabler. Därefter undersöks diagnostiken för att se hur pass väl modellen är anpassad till datan. Eftersom kategoriska såväl som kontinuerliga variabler finns med i regressionen innehåller formeln för regressionen ett antal vektorer för dummy-variabler. De kontinuerliga variablera definieras som olika X_j medan de kategoriska variablerna bortsett från GRAV ges olika λ-beteckningar vilket förtydligas i tabellen endan

Table 3: Variabeldefinition för regressionsanalys

Variabel	**Ny notation**
AB	Y
ARB	X_1
RECK	X_2
RECL	X_3
BRO	X_4
PAK	X_5
PAL	X_6
RE	λ_k^{RE}
GL	λ_l^{GL}
AGE	λ_m^{AGE}
ÅR	λ_t^{AR}

På detta vis skrivs modellen för regressionsanalyser med fler än en variabel som

$$g(\mu_{iklmt}) = \alpha + \sum_{j=1}^{6} \beta_j X_{ij} + \lambda_k^{RE} + \lambda_l^{GL} + \lambda_m^{AGE} + \lambda_t^{\mathring{A}R} \tag{4}$$

där varje λ är en inställning för någon av nivåerna för vilken faktorvariabeln är inställd på. λ_k^{RE} representerar regionerna där $k = (0, \ldots, 21)$, λ_l^{GL} representerar graviditetslängd där $l = (0, 1, 2)$, λ_m^{AGE} representerar åldersgrupperna där $m = (0, 1, 2, 3)$, $\lambda_t^{\mathring{A}}$ representerar år där $t = (0, 1, 2, 3, 4, 5, 6)$ och en residual $\epsilon_i \sim N(0, 1)$. All diagnostik har fås m.h.a kommandot plot och summary i R.

5.2.1 Enkel linjär regression

Varje förklarande variabel testas enskilt mot responsvariabeln AB och resultaten från diagnostiken var dålig (se slutet av detta avsnitt). I och med att Blekinge är satt till basnivå är de län som har signfikant effekt Stockholm, Västra Götaland och Skåne vilket inte är oväntat eftersom de regionerna har flest invånare och därmed rimligtvis även högst antal aborter. Åldersgruppen '<24' år har en signifikant positiv effekt medan de resterande grupperna har en signifikant negativ effekt, detta framgår också från avsnitt 'yngre personer och graviditetslängd' om att kvoten $\frac{AB}{GRAV}$ är störst för just den åldersgruppen. Arbetslöshet har en signifikant positiv effekt på antalet aborter för varje ytterligare arbetslös kvinna. Antalet patienter med kortvariga recept samt antalet långvariga recept som inhandlas har var för sig en signifikant positiv effekt.

Problemet med alla dessa resultat är att de bryter mot fyra av de modellantaganden som finns i avsnitt 3.1. Residualerna är ej normalfördelade och ett märkligt beroende existerar i residualerna för alla förklarande variabler, ett axplock framgår i Grafen 4 i Avsnitt 9 då residualerna för modellen när brott är den förklarande variabeln plottas. De variabler som tillförde ett någorlunda R^2 var 'RE', 'RECK' och 'PAK' då $R^2_{RE} \approx 37\%$, $R^2_{ARB} \approx 26\%$ och $R^2_{POP} \approx 28\%$.

5.2.2 Multipel linjär regression

Eftersom svarsvariabeln 'AB' ser ut att vara exponentialfördelad på intervallet $(0, 1764)$ och de förklarande variablerna tillhör olika fördelningar blir denna form av analys märklig. Den fulla modellen LM_1 anpassas likt ekvation 5 bortsett från skillnaden att p_i byts ut mot en residualterm. Vidare exkluderas även variablerna X_5 & X_6 helt eftersom det från dataanalysen framgår tydligt att antal recept som konsumerats är fler än antalet patiener.

Modellen är alltså

$$\mu_{iklmt} = \alpha + \sum_{j=1}^{4} \beta_j X_{ij} + \lambda_k^{RE} + \lambda_l^{GL} + \lambda_m^{AGE} + \lambda_t^{\mathring{A}R} + \epsilon_i$$

Modellens förklaringsgrad är $R^2 = 59\%$ vilket tyder på att variationen i responsvariablen förklaras väl av kovariaterna. Alla åldersgrupper samt graviditetslängder är signifikanta medan bara vissa av regionerna är signifikanta. Antalet långvariga recept samt antalet brott per invånare är även dem signifikanta. Eftersom variabeln graviditetslängd är ointressant förefaller modellen som något oinformativ. Vidare framgår det tydligt från diagonstiken att modellantaganden som framgår i avsnitt 3.1 bryts till stor del. Det främsta problemet med denna modell är att den ej tar hänsyn till antalet förlossningar i varje region vilket intuitivt innebär att modellen borde ge relativt orättvisande resultat med anledning av argumenten i avsnitt 4.2. Diagnostikplottarna ger även konstiga resultat vilken framgår i Figur 6 i Avsnitt 9.

5.2.3 Poisson-regression

Tre uppenbara saker har observerats - svarsvariabeln tillhör räknedata, är exponentialfördelad; och kovariaterna tillhör olika fördelningar. Av de anledningarna tillämpas GLM-regressionen Poisson-regression. Poisson-regression kan ge resultat som inte överensstämmer med verkligheten, till exempel att ett antal aborter fås som ett rationellt tal. Vidare antas att $E[Y] = V[Y]$ vilket i fall av överspridning skapar en dålig anpassning. Då överspridning föreligger tillämpas modeller som tar hänsyn till överspridning i datan. Genom att lägga till en offset-variabel tar varje observation hänsyn till antalet förlossningar för respektive region. Därmed skapas en mer rättvisande modell eftersom områden med färre incidenter får en mer rättvisande effekt. Modellanpassningen för P_1 har formeln:

$$\log(\mu_{iklmt}) = \alpha + \sum_{j=1}^{4} \beta_j X_{ij} + \lambda_k^{RE} + \lambda_l^{GL} + \lambda_m^{AGE} + \lambda_t^{\mathring{A}R} + \log(p_i) \qquad (5)$$

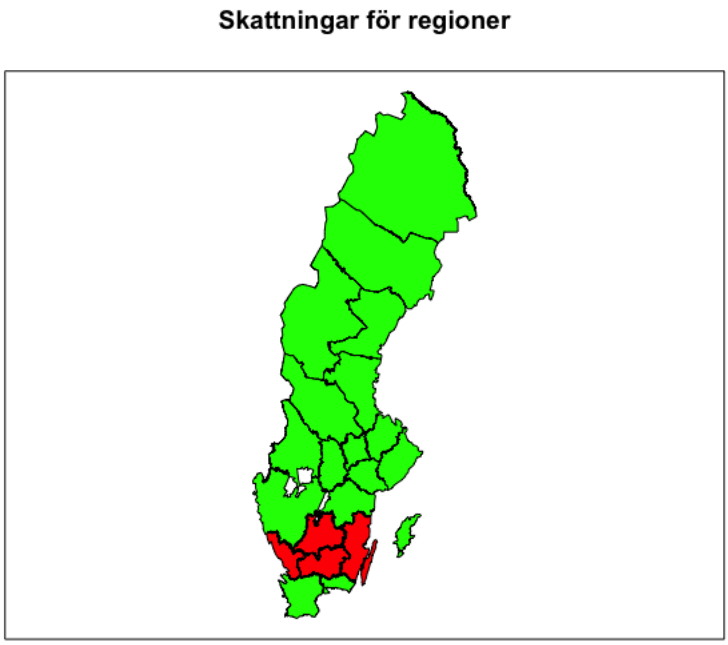
där varje λ är ett mått på hur många procent det logaritmerade väntevärdet av μ_i ökar eller minskar.
Basnivån för regressionsmodellen är då graviditetslängden är '<8 veckor', regionen Blekinge och åldersgruppen '<24'. AIC för modellen är 18215 och LOOCV ger ett CV_n om 1926 vilket ska jämföras med modeller som dyker upp senare i resultaten. Från denna modell framgår bland annat att alla åldersgrupperna, bortsett från '<24', har en signifikant negativ effekt på antalet aborter medan alla regioner bortsett från Kronoberg, Halland och Jönköping har en signifikant positiv effekt. Den långvariga formen av preventivmedel har en signifikant negativ effekt vilket är väntat - fler recept borde leda till färre aborter givet antagandet att recepten används. Ytterligare intressanta aspekter framgår men undersökningen ifall överspridning föreligger samt en kort diagnostisk undersökning av residualerna, vilka framgår i Figur 7 i Avsnitt 9, leder till att ett annat modellantagande testas. Överspridning i P_1 testas genom att tillämpa ekvationerna (2) och (3) från avsnitt 3.7.3 i R. Resultatet från (2) jämförs med en χ^2-fördelning med $1764 - 8 = 1756$ frihetsgrader (eftersom 8 förklarande

variabler inkluderas) vilket ger ett p-värde om 1, dvs. H_0 : "Överspridning föreligger ej" kan förkastas på alla rimliga signifikansnivåer och med hjälp av (3) framgår att en överspridningskvot om ca 5 föreligger. Ett ytterligare beslutsunderlag för att testa en ny modell är en undersökning av diagnostiken som ser förhållandevis bra ut bortsett från att antagandet om normalitet i residualerna inte är tillräckligt bra, se Figur 7 i Avsnitt 9.

5.2.4 NB-regression

Eftersom överspridning föreligger i Poisson-modellen genomförs en analys med Negativ Binomial regression. NB-modeller används för att anpassa frekvensdata och innehåller en parameter som används för att justera variansen oberoende av medelvärdet. På samma sätt som i Poisson-regressionen används en offset-variabel varför den fulla modellen har samma struktur som P_1, definierat i ekvation (5). Från modellsummeringen framgår att samtliga faktornivåer för graviditetslängd samt åldersintervall har en signfifikant negativ effekt på antalet aborter. Samtliga regioner bortsett från Kalmar, Västerbotten, Västmanland samt Örebro var signifikanta. I figur 3 framgår ifall skattningarna för regionerna är positiva, markerat i grönt, eller negativa, markerat i rött. Långvarig receptbehandling har en negativ effekt medan även år har en signifikant negativ effekt. AIC är 14008, CV_n är 4352 och residualplottarna ser ut att uppfylla de krav som sätts i avsnitt 3.1 vilket framgår i Figur 8 i Avsnitt 9. Vidare har brott en signifikant positiv effekt medan antalet arbetslösa har en signifikant negativ effekt.

Figure 3: Skattningarna för varje region, grönt är positiv skattning, rött är negativ skattning.



5.3 Analys

Förstahandsvalet vid regressionsanalyser är att nyttja linjära modeller på grund av dessas enkelhet. Problemet med linjär regression är att varje variabel antas tillhöra samma typ av fördelning. I de fall då olika fördelningar föreligger i datamaterialet är det därför lämpligast att tillämpa GLM. Från resultaten framgick att vanliga linjära modeller ej är lämpliga antaganden till detta dataset. Parameterskattningarnas signifikansnivåer kombinerat med den höga förklaringsgraden kan absolut leda till att modellen ger en bra prediktion av datan, men vid närmare undersökning framgår tydligt ur residualplottarna i frigur 6 från appendix 9 att inga av de fyra sistnämnda antagandena för en linjär modell uppfylls. Analysen går naturligt över till att testa andra modeller vilket leder till nyttjandet av specialfall av GLM. Den första modellen som anpassats m.h.a Poisson-regression ser inledningsvis bra ut men, precis som i de linjära modellerna, tappar kvalité i anpassningsförmåga då diagnostiken undersöks. Residualplottarna ser ut att uppfylla kraven men modellen faller då överspridning testas. Med hjälp av (3) från avsnitt 3.7.3 räknas överspridningen ut till att

motsvara ca fem gånger väntevärdet vilket bryter mot antagandet om Poissonfördelningar att $E[X] = V[X]$. Sannolikheten att datan ska anta ett värde som modellen skattade var 0 varför NB-modellen testas. I NB-modellen är större delen av kovariaterna och faktornivåer signifikanta. Residualplottarna ser ut att uppfylla de krav som ställs på deras utseende för en god prediktionsförmåga hos en modell. I tabellen 4 nedan jämförs AIC och CV för respektive modell.

Table 4: Tabell för AIC och CV

Modell	**AIC**	**CV**
P1	18215	1926
NB1	14008	4352
Δ	4207	-2406

Från jämförelsen framgår att AIC för NB-modellen är 4207 enheter lägre än Poisson-modellen. Dock har Poisson-modellen enligt CV en bättre prediktiv förmåga med ett CV som är 2426 enheter lägre. Detta är märkligt eftersom NB-modellen är bättre anpassad till datan och residualerna från diagnostiken har en lägre spridning jämfört med diagnostiken för Poisson-modellen. Detta kan förklaras med att överanpassning föreligger och eftersom NB-regression är mer flexibel än Poisson-regression kan det därmed även vara så att NB-modellen tar hänsyn till någon extrem outlier som Poisson-modellen möjligtvis tryckt ner i och med antagandet om att $E[Y] = Var[Y]$.

Det finns således två modeller att välja mellan:

a. Poisson-modellen: Sämre anpassad till datan än NB-modellen samt överspridning föreligger.

b. NB-modellen: Tveksamt sämre precision jämfört med Poisson-modellen, dock bättre anpassad till datan samt modellen tar hänsyn till överspridning i datan.

Informationen från a och b talar för att en NB-regression ska tillämpas för att kunna anpassa en modell till ett dataset som är överspritt till den aktuella nivån. I tidigare avsnitt påpekas att variablen GL ej är intressant att ha med som kovariat i regressionen, men eftersom AIC höjdes med 5192 enheter i samband med att GL exkluderas föredras alltså att inkludera GL för att få en bättre anpassning till datan. Skattningarna för NB-modellen framgår i Tabell 6 i Avsnitt 8.

5.3.1 Diskussion

I artikeln [2] och boken [1] påpekas att socioekonomiska faktorer har en effekt på antalet födda barn samt dessa barns förutsättningar senare i livet. Hög arbetslöshet och ett högt antal brott är tydliga indikatorer på ett samhälle i förfall. Den delen av analysen kan gå djupare då dessa två faktorer bland annat kan kopplas till värderingar bland vissa grupper eller samhällsstrukturer som

leder till en ökning av de faktorerna. På grund av omfattningen av detta arbete kommer den djupare analysen att exkluderas i största möjliga grad varför vissa antaganden är spekulationer.

I det fall en responsvariabel antar ett numeriskt reelt värde är oftast en multipel linjär regression det enklare valet. I detta fall ser responsvariabeln ut att vara exponentialfördelad vilket framgår ur figur 1 från Avsnitt 5. På grund av att responsvariabelns fördelning ser ut att tillhöra exponentialfamiljen ska regression på det logaritmerade väntevärdet av responsvariabeln genomföras. Ett enket fall hade varit om de förklarande variablerna var numeriska, men i denna analys förekommer en blandning av fördelningar för de förklarande variablerna. Parameterskattningarna i den slutgiltiga modellen tolkas som den procentuella effekten varje skattning har på responsvariabeln.

Ett problem med denna analys är att variabeln 'Brott' räknar alla brott från varje län där vissa av dessa brott ej är konsekvenser av miserabla förhållanden i någon specifik region. Det förväntades även att antalet aborter skulle minska ju mer preventivmedel som säljs i ett län enligt [2], vilket det finns tecken på eftersom de långvariga recepten har en signifikant negativ effekt. Anledningen till det är slutsatsen att ju mer preventivmedel som säljs desto färre graviditeter vilket leder till färre aborter.

En mycket intressant del i analysen är de regioner som har negativa estimat i NB-modellen. Det framgår tydligt från figur 3 att de regioner med en negativ skattning ligger i samma område i Sverige som Sveriges Radio refererat till som 'Bibelbältet' eller 'Smålands Jerusalem' [11].

6 Slutsats

Den valda modellen är en negativ binomial regression vilken valdes i huvudsak för att det modellantagandet var bäst anpassat till datan sett till AIC och överpsridningen i datan. Prediktionsförmågan var något sämre än alternativmodellen men det är rimligare att välja en modell utifrån ett uppfyllande av grundläggande mått. Vissa mycket intressanta aspekter från modellen bör uppmärksammas. Samtliga åldersgrupper är signifikanta och har negativa skattningar vilket pekar åt att den åldersgrupp som är mest benägen att genomföra en abort är ungdomar tillhörande åldersgruppen "<24" år. Detta är mycket rimligt utifrån vissa argument som tas upp i avsnitt 5.1.2, nämligen att yngre personer generellt ej befinner sig i en tillräckligt stabil livssituation för att kunna föda upp ett barn.

Att arbetslösheten har en negativ effekt är något svårare att tolka. Datan för arbetslösheten är hämtad från Ekonomifakta och deras definition av en arbetslös person är någon som kan börja jobba inom 14 dagar och aktivt har sökt arbete de senaste fyra veckorna [12]. En tolkning är att det är lättare för en arbetslös kvinna att ta beslutet att föda ett barn eftersom hon på grund av arbetslöshet kan lägga mer tid på barnet. Någonting ytterligare intressant är det kluster av regioner med negativa skattningar som framgår i figur 3. Djupt religiösa grupper i världen tenderar att ha en negativ inställning till abort som ingrepp varför det

är mycket intressant att det område i Sverige som även är känt som 'Bibelbältet' är de enda regionerna som har just negativa effekter. Det är mycket svårt att tolka det som en slump. Variabeln $\mathit{ÅR}$ betraktas som en linjär variabel och har en signifikant negativ effekt på antalet aborter. Detta tolkas som att världen blir säkrare och framför allt, enligt bland annat tidsskriften 'The Economist', är Stockholm världens säkraste stad [13] vilket leder till att personer rimligtvis finner det lämpligare att föda barnet eftersom förutsättningarna blir bättre med tiden. I teorin från [1] väntades att antalet brott skulle ha en positiv effekt på antalet aborter - vilket det även har. Ju fler brott som begås desto lägre är säkerheten i samhället varför fler väljer att genomföra en abort eftersom förutsättningarna inte är bra. Antal långvariga recept är signifikant negativt vilket är väntat eftersom fler personer som använder långvarigt preventivmedel leder till färre graviditeter vilket leder till färre aborter. Utifrån [2] kan därmed slutsatsen att högre priser för preventivmedel leder till att färre konsumerar preventivmedel vilket leder till fler aborter.

Modellen är bra ur en prediktiv ståndpunkt. Det framgår att färre aborter genomförs med tiden dock kommer osäkerheter som är resultat av ökad brottslighet att leda till fler aborter. Modellen kan användas som ett verktyg för beslutsfattande i samhället för att kunna uppskatta hur många fler aborter som kan tänkas genomföras och på så vis tex. förbereda myndigheter i det specifika området.

7 Vidare

Något av ett juridiskt problem är datan gällande aborter från Socialstyrelsen. Fram till 2012 är datan för aborter fördelat över region vilket ger utrymme för att hitta vilka personer som genomfört en abort. Denna företeelse går på tvärs mot anonymiteten varför utvärderingen kom fram till att datan ej ska vara indelad på länsnivå utan istället visas för hela riket. Resultatet av denna begränsning är bland annat att framtida analyser kommer bli svårare - nästan icke-existerande alternativt kräva ett forskningsanslag för att finansiera utvinningen av den datan som krävs. Om samma årtal som i denna analys tillämpas borde brotten delas in med större varsamhet - till exempel är finansiella brottslighet sällan konsekvensen av socioekonmisk misär. Den regressionsmodell som valts i denna analys är nödvändigtvis inte optimal. En modell där bland annat samspel mellan vissa variabler sätts som kovariater vore intressant att se, dock är teorin på området för samspelseffekter i NB-modeller mycket skral varför sådana metoder exkluderas från analysen i detta arbete.

8 Tabeller

Källorna för respektive dataset finns i nedan tabell, datan är indelad efter län (region), år och åldersintervall där åldersintervall är indelade på femårsintervall från <24 till 35<.

Table 5: Datakällor

Data	Förklaring	URL-källa
SCB		
Invånare	Antal invånare	Tabell Demografi
Socialstyrelsen		
Aborter	Antal aborter samt graviditetslängd tills utförande	Tabell Aborter
Preventivmedel	Vilken typ av preventivmedel, lång/kort-varigt, samt antal recept och patienter	Tabell Läkemedel
Förlossningar	Antal förlossningar	Tabell Förlossningar
BRÅ		
Brott	Antal brott	Tabell Brott
Ekonomifakta		
Arbetslöshet	Andelen arbetslöshet	Tabell Arbetslöshet

Table 6: R-utskrift för NB-modell

	Estimate	Std. Error	z value	Pr(> | z |)
(Intercept)	29.7627	6.9882	4.26	0.0000
ARB	-0.0000	0.0000	-7.61	0.0000
GL>12 veckor	-2.5355	0.0154	-164.95	0.0000
GL9-11 veckor	-1.6713	0.0134	-124.60	0.0000
AGE25-29	-1.8414	0.0548	-33.60	0.0000
AGE30-34	-2.0782	0.0542	-38.33	0.0000
AGE35<	-1.0276	0.0411	-25.03	0.0000
REDalarna	0.1941	0.0435	4.47	0.0000
REGotland	0.2790	0.0572	4.88	0.0000
REGävleborg	0.2260	0.0429	5.27	0.0000
REHalland	-0.0967	0.0415	-2.33	0.0198
REJämtland	0.2219	0.0451	4.92	0.0000
REJönköping	-0.1875	0.0429	-4.37	0.0000
REKalmar	-0.0024	0.0429	-0.06	0.9545
REKronoberg	-0.1000	0.0449	-2.23	0.0259
RENorrbotten	0.2563	0.0417	6.15	0.0000
RESkåne	0.3305	0.0496	6.67	0.0000
REStockholm	0.5800	0.0596	9.73	0.0000
RESödermanland	0.1232	0.0449	2.74	0.0061
REUppsala	0.1547	0.0447	3.46	0.0005
REVärmland	0.3242	0.0416	7.80	0.0000
REVästerbotten	0.0717	0.0458	1.57	0.1173
REVästernorrland	0.1780	0.0437	4.07	0.0000
REVästmanland	0.0820	0.0439	1.87	0.0619
REVästra Götaland	0.3503	0.0489	7.17	0.0000
REÖrebro	0.0536	0.0428	1.25	0.2104
REÖstergötland	0.1104	0.0425	2.60	0.0095
RECK	-0.0584	0.0634	-0.92	0.3575
RECL	-2.8790	1.4297	-2.01	0.0440
ÅR	-0.0150	0.0035	-4.28	0.0000
BRO	0.0650	0.0191	3.41	0.0007

9 Grafer

Figure 4: Residualer för GRAV

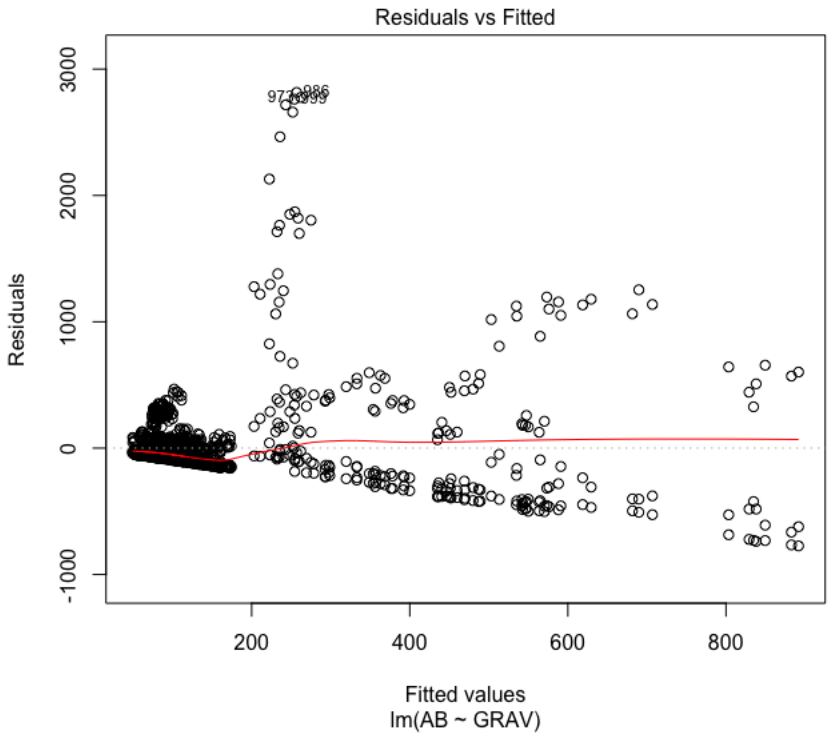


Figure 5: Visualisering av korrelationer mellan de variabler som framgår i avsnitt 'korrelationsanalys'

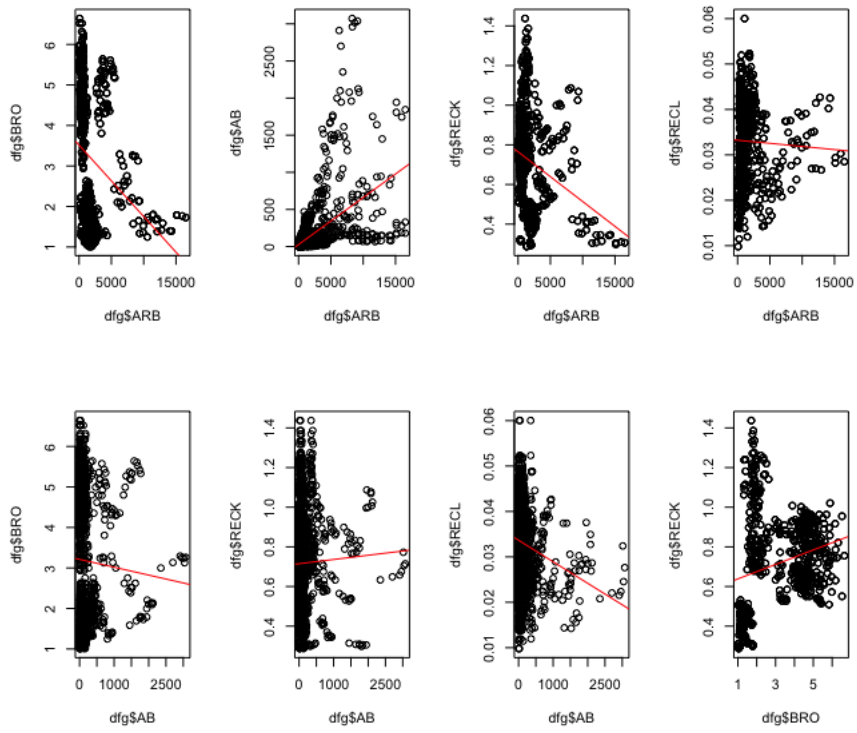


Figure 6: Diagnostik för LM1

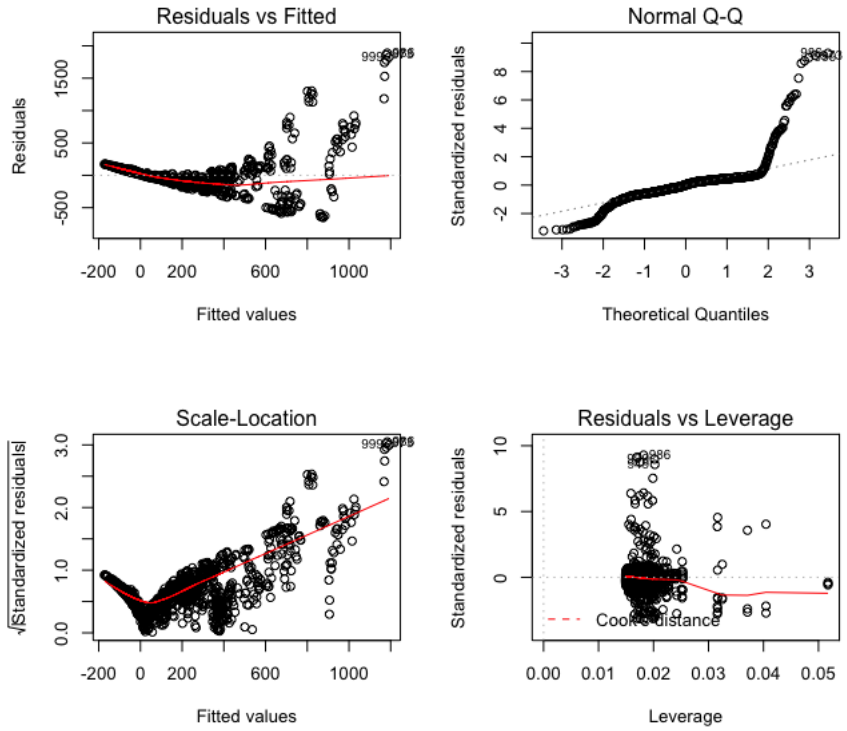


Figure 7: Residualer för P1

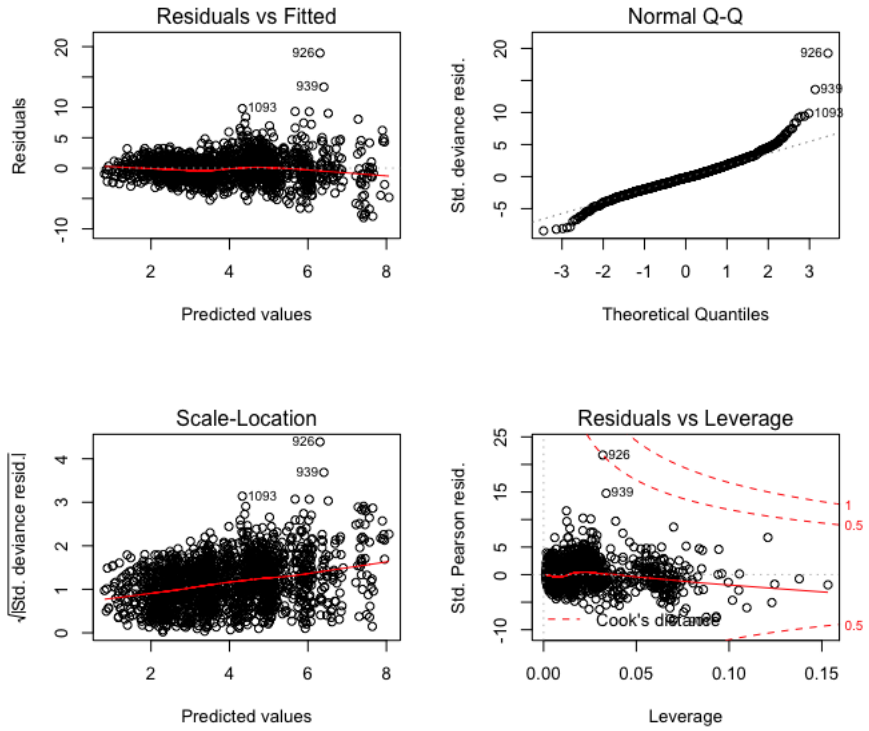
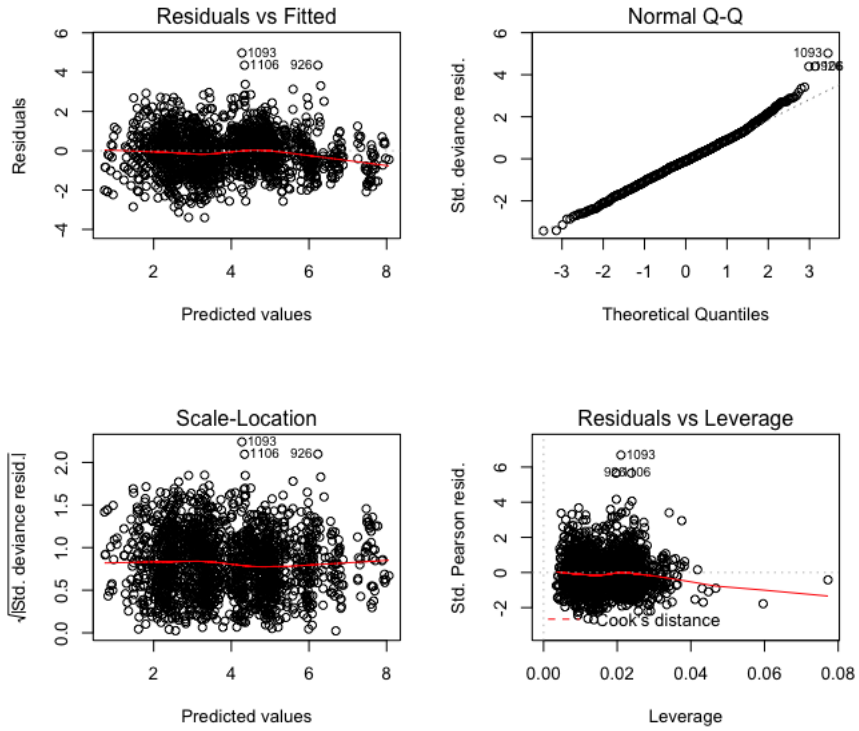


Figure 8: Residualer för NB1



References

[1] S. D. Levitt and S. J. Dubner, *Freakonomics*, vol. 61. Sperling & Kupfer editori, 2014.

[2] A. Madestam and E. Simeonova, "Children of the pill: The effect of subsidizing oral contraceptives on children's health and wellbeing," in *2012 AAPPAM Fall Research Conference*, pp. 8–10, 2012.

[3] F. Svensson, "Lista, här är hetaste kandidaterna i kemi," in *www.svd.se*, 2016.

[4] M. Nyman, "Forskare skapar genetiska förändringar i befruktade humana ägg," in *www.genteknik.se*, 2015.

[5] BABOUCHEE, "B.sc-mathematical-statistics," in *github.com/BABOUCHEE/B.sc-Mathematical-Statistics*, 2017.

[6] S. Wood, *Generalized additive models: an introduction with R*. CRC press, 2006.

[7] A. Gelman and J. Hill, *Data analysis using regression and multilevelhierarchical models*, vol. 1. Cambridge University Press New York, NY, USA, 2007.

[8] J. M. Hilbe, *Negative binomial regression*. Cambridge University Press, 2011.

[9] C. Englund, "Så mycket kostar det att ha barn," in *Dagens Nyheter*, 2013.

[10] A. Engström, "Unga svenskar sämst på skyddat sex," in *Svenska Dagbladet*, 2009.

[11] O. Mattisson, "Då kan bibelbältet flyttas från jönköpings län," in *sverigesradio.se*, 2016.

[12] C. Holmström, "Arbetslöshet," in *Ekonomifakta.se*, 2017.

[13] S. Murray, "The safe cities index 2015," in *safecities.economist.com*, 2015.