Stockholms universitet

Tidstrend inom självmord i Sverige - En jämförelse på kön och ålder

Heja Lindgren

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2017:5
Matematisk statistik
Juni 2017

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholms universitet

Matematisk statistik
Stockholms universitet
Kandidatuppsats **2017:5**
http://www.math.su.se/matstat

Tidstrend inom självmord i Sverige - En jämförelse på kön och ålder

Heja Lindgren*

Juni 2017

Sammanfattning

Arbetet har avsikt att svara på hur tidstrenden inom självmord i Sverige varierat mellan år 1969 och 2015, Vi gör en jämförelse mellan kön och åldersgrupperna 15–24 år, 25–44 år, 45–54 år, 55–64 år, 65–74 år och 75+ år. Sedan analyserar vi den enskilda åldersgruppen 15–24 år mellan år 2000 och 2014. Målet är att hitta en modell som bäst förklarar hur antalet självmord beror på tiden samt undersöker samspelet mellan kön och kalenderår. Vi ska även analysera samspelet mellan åldersgrupper och kalenderår. Till sist undersöker vi tidstrend för ungdomarna mellan år 2000 och 2014.

Det finns inte något signifikant samspel mellan kön och kalenderår. Oddset för att en man begår självmord är 2.46 gånger högre än oddset för att en kvinna begår självmord, däremot existerar ett signifikant samspel mellan åldersgrupper och kalenderår. När vi analyserar självmord för åldersgruppen 15-24 mellan år 2000 och 2014, visar det sig att antalet självmord har ökat för ungdomar i Sverige.

För att visuellt undersöka tidstrenden inom självmord för de sex åldersgrupperna, plottar vi B-spline för varje åldersgrupp. Vi ser att antalet självmord i åldersgrupperna 25–44 år, 55–64 år och 65–74 år avtar efter år 1985. Åldersgruppen 15–24 år skiljer sig från de andra åldersgrupperna. Denna grupp har lägst antal självmord genom åren.

Åldersgruppen 25–44 år har större antal självmord än åldersgruppen 15–24 år genom åren, men är mindre än de äldre åldersgrupperna. Åldersgruppen 45-54 år har högst antal självmord mellan år 1969 och 1986. Mellan år 1987 och 2005 har den äldsta åldersgruppen högst antal självmord.

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

Abstract

The work is intended to respond to how the time trend in suicide in Sweden varied between 1969 and 2015, We make a comparison between gender and age groups 15-24 years, 25-44 years, 45-54 years, 55-64 years, 65-74 Years and 75+ years. Then we analyze the individual age group 15-24 years between 2000 and 2014. The goal is to find a model that best explains how the number of suicides depends on time and examines the interaction between gender and calendar year. We will also analyze the interaction between age groups and calendar years. Finally, we investigate the time trend for young people between 2000 and 2014.

There is no significant interaction between gender and calendar year. The odds that a man commits suicide is 2.46 times higher than the odds of a woman committing suicide, on the other hand, there exists a significant interaction between age groups and calendar years. When analyzing suicide for the 15-24 age group between 2000 and 2014, it appears that the number of suicides has increased for young people in Sweden.

To visually examine the time trend in suicide for the six age groups, we plot B-spline for each age group. We see that the number of suicides in the age groups 25-44 years, 55-64 years and 65-74 years decreases after 1985. The age group 15-24 years differs from the other age groups. This group has the lowest number of suicides over the years.

The age group 25-44 years has a greater number of suicides than the age group 15-24 years over the years, but is less than the older age groups. The age group 45-54 years has the highest number of suicides between 1969 and 1986. Between 1987 and 2005, the oldest age group has the highest number of suicides.

Förord

Detta arbete utgör mitt examensarbete i matematik statistik om 15 högskolepoäng vid Stockholms Universitet. Jag vill ge ett stort tack till min handledare Martin Sköld för all hjälp och råd han gett under arbetets gång. Jag vill även tacka professor Ola G H Hössjer för värdefull hjälp och feedback.

Innehåll

1 Inledning **6**

2 Teori **7**

2.1 Multipel linjär regresionsmodell 7

2.1.1 Centrera förklarande variabler 7

2.2 GLM 8

2.2.1 Slumpmässig komponent 8

2.2.2 Systematisk komponent 8

2.2.3 Länkfunktion 9

2.3 Multipel logistisk regressionsmodell 9

2.4 Oddskvot 10

2.4.1 Konfidensintervall för oddskvot 10

2.5 Likelihood 11

2.6 Akaike information criterion 12

2.7 Konfidensintervall 12

3 Splines **14**

3.1 B-splines 14

4 Databeskrivning **15**

5 Analys av data **15**

6 Modeller och analys **16**

6.1 Linjär regression 17

6.2 Generaliserad linjär modell 18

6.3 Icke linjära tidstrender 20

6.3.1 Wald-konfidensintervall 23

6.3.2 Ökad tidstrend för ungdomar 24

6.3.3 Tidstrender för de sex åldersgrupperna 26

7 Diskussion **27**

8 Appendix **29**

8.1 Figurer 29

8.2 Tabller 29

9 Referenser **33**

1 Inledning

Självmord har diskuterats och studeras av människor i alla tider. Det är ett intressant ämne.

Danuta Wasserman, professor i psykiatri och suicidologi skriver i **Läkartidningen**, Nr 50, 2004:

> Ökat missbruk av alkohol och droger (inte sällan i syfte att lindra skadlig stress och dämpa oro och ångest) samt ökat antalet depressioner som debuterar allt längre ner i åldrarna har också framhållits som viktiga orsaker till de ökade självmorden bland unga. Pojkar liksom män är mer känsliga för psykosociala förändringar, och de har även sämre förmåga att be om och söka hjälp.

Författare Henrik Nordin som arbetar med statistiksamordning på socialstyrelsen, Skriver i en artikel i **välfärd**, Nr 4, 2009, titel för artikeln är "självmord är vanligast bland äldre män":

> Depression eller de kroppsliga sjukdomarna är de vanligaste orsakerna till äldres självmord enligt Nationell prevention av suicid och psykisk ohälsa, NASP, vid Karolinska Institutet.

Längre ner i samma artikel skriver Henrik Nordin:

> Det sker betydligt fler självmord bland män än bland kvinnor i alla åldrar, men skillnaden är störst bland de äldsta. Sammantaget är självmord mer än dubbelt så vanligt bland män som bland kvinnor.

Med bakgrund av detta vill vi i denna studie analysera hur tidstrend inom självmord för respektive kön och åldersgrupperna 15-24 år, 25-44 år, 45-54 år, 55-64 år, 65-74 år och 75+ år. Vi kommer att undersöka om det finns något signifikant samspel mellan kön och kalenderår samt samspel mellan åldersgrupper och kalenderår. Efter vi har hittat den lämpligaste B-splinefunktion för de sex åldersgrupperna, gör vi ett Wald-konfidensintervall för att undersöka hur säkert B-splines är. Sedan analyserar vi åldersgruppen 15–24 år mellan år 2000 och 2014. Till sist gör vi en plot där vi har alla B-splines för de sex åldersgrupperna för att visuellt analysera tidstrend inom självmord under perioden 1969-2015 för hela befolkningen i Sverige.

I denna rapport börjar vi med att gå genom teorin för bakomliggande modelleringen och analysen, sedan beskriver vi datamaterialet som vi har hämtat från Statistiska centralbyrån och Socialstyrelsen. Efter det analyserar vi vårt datamaterial. Vi kommer se att trend för båda kön ser snarlikt ut, dock är antalet självmord för män är högre än kvinnor. Vi går vidare till modeller och

analys, där vi försöker på bästa sätt hitta en modell som passar vår data. Vi prövar linjär regressionsmodell, generaliserad linjär modell och B-splines. Det visar sig att B-splines är den lämpligast modell som passar vår data. Vi analyserar den enskilda åldersgruppen 15–24 år mellan år 2000 och 2014. Här passar linjär regressionsmodell bra. Genom statistisk metod kommer vi fram till att det finns en signifikant ökad tidstrend för ungdomsgruppen under dessa år. Sist kommer diskussionen där vi diskuterar resultat och förbättringar.

2 Teori

2.1 Multipel linjär regresionsmodell

En multipel linjär modell där vår responsvariabel är W och vi har k förklaringsvariabler med n observationer för varje förklaringsvariabel. Vår ursprungsmodell är då

$$w_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki} + \epsilon_i,$$

där $i = 1, 2, 3...n$, och ϵ_i kallas för slumpterm och antas vara oberoende $N(0, \sigma^2)$.

Omskrivning av modellen i matris form:

$$\mathbf{W} = \mathbf{A}\ \boldsymbol{\theta} + \boldsymbol{\epsilon},$$

där $\mathbf{W} = (w_1, w_2, ...w_n)^T$,

$$\mathbf{A} = \begin{bmatrix} 1 & x_{11} & x_{21} & x_{31} & \dots & x_{k1} \\ 1 & x_{12} & x_{22} & x_{32} & \dots & x_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \\ 1 & x_{1n} & x_{2n} & x_{3n} & \dots & x_{kn} \end{bmatrix}$$

$\boldsymbol{\theta} = (\alpha, \beta_1, \beta_2, ..., \beta_k)^T$, där $(\beta_1, \beta_2, ..., \beta_k)$ är effekt parametrar, α är intercept och $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, ..., \epsilon_n)^T$ (Patrik Andersson och Joanna Tyrcha, 2014, s.26).

2.1.1 Centrera förklarande variabler

Om man är intresserad av att analysera en del av datamaterialet, och vill undersöka om en enskild koefficient är lika med noll, så ska man centrera de förklarande variablerna för att undvika korrelationer mellan koefficienterna.

Låt $W_i, i = 1, 2, ..., n$ vara oberoende och $N(\mu_i, \sigma^2)$ och vi har k olika förklaringsvariabler $x_1, x_2, ..., x_k$, μ_i kan skrivas på formen

$$\mu_i = \alpha + \beta_1(x_{1i} - \bar{x}_1) + \beta_2(x_{2i} - \bar{x}_2) + ... + \beta_k(x_{ki} - \bar{x}_k),$$

där $\bar{x}_1 = \sum_{i=1}^{n} x_{1i}/n$ och så vidare. Med minsta kvadratmetod får vi $\hat{\alpha} = \bar{w}$ (Rolf Sundberg, 2016,s.64-65).

2.2 GLM

I Generaliserad linjär modell behöver responsvariabeln inte vara normalfördelad. Den generaliserad linjär modell består av tre komponenter, De är

- ⋆ en slumpmässig komponent
- ⋆ en systematisk komponent
- ⋆ en länkfunktion(Alan Agresti,2002, s.116)

2.2.1 Slumpmässig komponent

Sannolikhetsfördelning för responsvariabeln Y med oberoende observationer $(y_1, y_2, ..., y_n)$ kommer från en naturlig exponentialfamilj. Den slumpmässiga komponenten av generaliserad linjär modell består av en sådan responsvariabel. Den naturliga exponentialfamiljen har sannolikhetsfunktionen

$$f(y_i; \theta_i) = A(\theta_i)B(y_i)exp(y_iQ(\theta_i)), i = 1, 2, ...n,$$

för någon funktion A, B och Q. Termen $Q(\theta)$ kallas för den naturliga parametern(Alan Agresti,2002, s.116).

2.2.2 Systematisk komponent

Låt $\mathbf{X_i} = (1, X_{i1}, X_{i2}, ...X_{ij}), i = 1, 2, ...I, j = 1, 2, ..., k$ vara prediktorsvariabler och β_j vara regressionsparametrar i modellen. Då ger den en systematisk komponent i en generaliserad linjär modell genom vektor $\eta = (\eta_1, \eta_2, ...\eta_n)$ enligt nedan.

$$\eta_i = \alpha + \sum_{j=1}^{k} \beta_j x_{ij}, i = 1, 2, ...I.$$

Denna linjära kombination av förklarande variabler kallas för linjär prediktor (Alan Agresti,2002,s.116).

2.2.3 Länkfunktion

Länkfunktion länkar ihop slumpmässiga komponenter och systematiska komponenter.

Modellen förbinder μ_i till η_i genom sätta $\eta_i = g(\mu_i)$, där länkfunktionen g är en monoton, differentierbar funktion. Således g länkar $E(\mathbf{Y}_i)$ till förklarande variabler genom formeln

$$g(\mu_i) = \sum_{j=1}^{n} \beta_j x_{ij}, i = 1, 2, ..., k.$$

Länkfunktion $g(\mu) = \mu$ kallas för identitetslänk som har $\eta_i = \mu_i$. Exempel för andra länkfunktioner är

- Log: $\eta = \log(\mu)$
- Logit: $\eta = logit(\mu)$
- Kanonisk: $\eta = Q(\theta)$

En generaliserad linjär modell använder maximum likelihood-metoden för att skatta parametrarna $\hat{\beta}_j$ för modellen. Eftersom vi i den generaliserade linjära modellen antar att responsvariabel inte är normalfördelad med konstant varians, så är minsta kvadratmetod uteslutet(Alan Agresti,2002, s.116-117).

2.3 Multipel logistisk regressionsmodell

Vi antar att vi har k prediktorsvariabler. $\mathbf{X} = (x_1, x_2, ...x_k)$. Prediktorsvariabler är intervallsvariabler, till exempel år, ålder, dos, och $Y \in \{0, 1\}$ är en binär responsvariabel. där $Y = 0$ betyder “Nej” och $Y = 1$ betyder ”Ja”.

Då gäller i den logistiska regressionen att

$$\pi(\mathbf{X}) = P(Y = 1|\mathbf{X} = (x_1, x_2, ...x_k)) = \frac{exp(\alpha + \sum_{i=1}^{k} \beta_i x_i)}{1 + exp(\alpha + \sum_{i=1}^{k} \beta_i x_i)}.$$

Vi får

$$logit(\pi(\mathbf{X})) = log(\frac{\pi(\mathbf{X})}{1-\pi(\mathbf{X})}) = \alpha + \sum_{i=1}^{k} \beta_i x_i,$$

där α är intercept parameter. β_i är multiplikative effekt i log-odds när x_i ökar med en enhet och de övriga x_j är fixt, där $i \neq j$ (Alan Agresti,2002, s.182-183).

2.4 Oddskvot

Låt $\pi = \pi(\mathbf{x})$ som ovan, sannolikheten för ett lyckad försök givet $\mathbf{X} = (x_1, x_2, ...x_k)$. Vi vet att $0 \leq \pi \leq 1$. då är definition för odds

$$\Omega = \pi/(1-\pi).$$

Enligt definitionen ser vi att oddset aldrig är negativt. Om $\Omega > 1$ betyder att lyckad är mer troligt än misslyckad. Om $\Omega < 1$ är misslyckad mer troligt än lyckad.

Antar att vi har två kolumner och två rader. $\pi_i = P(Y = 1|X = i)$ och $\pi_j = P(Y = 1|X = j)$, där $i = 1, 2, j = 1, 2$ och i $\neq$ j, då är oddskvoten mellan oddset för Ω_1 och Ω_2

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

Om oddskvoten är större än ett kan den tolkas som att oddset för händelse 1 är större än oddset för händelse 2. Om $0 \leq \Theta < 1$ tolkas oddskvoten som oddset för händelse 2 är större än händelse 1. I fallet $\Theta = 1$ är då oddset för att lyckas för de båda händelserna lika (Alan Agresti, 2002, s.44).

2.4.1 Konfidensintervall för oddskvot

Vi antar fortfarande att vi har 2×2 tabell, då är

$$\widehat{var}(\log \widehat{\Theta}) = \left(\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}\right),$$

där n_{ij} är det observerade värdet i rad i och kolumn j.

$\log \Theta$ är approximativ normalfördelad med "stora" observationer. Vi får därför $(1-\alpha) \cdot 100$ % konfidensintervall för $\log \Theta$

$$\text{KI}(\log\Theta) = \log\widehat{\Theta} \pm z_{\alpha/2} \cdot \sqrt{\widehat{var}(\log\widehat{\Theta})} = \big(I_o, I_u\big),$$

där $z_{\alpha/2}$ är kvantil för normalfördelning. I_o och I_u är övregräns respektive undregräns.

För att få konfidensintervall för Θ transformerar vi $\log\Theta$

$$\text{KI}\Big(\Theta\Big) = \Big(e^{I_o}, e^{T_u}\Big),$$

(Alan Agresti,2002,s.71).

2.5 Likelihood

Givet observerad data $x_1, x_2, ..., x_n$ är ifrån ett slumpat stickprov $\boldsymbol{X} = (X_1, X_2, ..., X_n)$ där $X_i, i = 1, 2, ...n$ är oberoende och lika fördelad, kommer från en fördelning med okänd täthetsfunktion $p(x; \boldsymbol{\theta})$, där $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_n)^T \in \boldsymbol{\Theta}$ är okända vektor parametrar. Då definieras likelihood

$$L(\boldsymbol{\theta}; x_1, ..., x_n) = p(x_1, x_2, ..., x_n|\boldsymbol{\theta}) = \prod_{i=1}^{n} p(x_i|\boldsymbol{\theta}).$$

Log-likelihoodfunktionen är

$$l(\boldsymbol{\theta}; x_1, ..., x_n) = \log(L(\boldsymbol{\theta}; x_1, ..., x_n)) = \sum_{i=1}^{n} \log p(x_i|\boldsymbol{\theta}),$$

(Leonhard Held och Daniel Sabanes Bove, 2014, s.17-18).

Den maximum likelihood skattning $\hat{\boldsymbol{\theta}}$ av $\boldsymbol{\theta}$ fås genom maximera den likelihood funktionen.

$$\hat{\boldsymbol{\theta}}_{ML} = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} L(\boldsymbol{\theta}; x_1, ..., x_n),$$

(Leonhard Held och Daniel Sabanes Bove, 2014, s.27).

Score funktion av den okända vektor $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_n)^T$ är gradienten av log-likelihood funktion

$$S(\boldsymbol{\theta}) = \frac{\partial}{\partial\boldsymbol{\theta}} l(\boldsymbol{\theta}) = \left(\frac{\partial}{\partial\theta_1} l(\boldsymbol{\theta}), \frac{\partial}{\partial\theta_2} l(\boldsymbol{\theta}), ..., \frac{\partial}{\partial\theta_n} l(\boldsymbol{\theta})\right).$$

Man kan beräkna maximum likelihood skattningen för $\hat{\boldsymbol{\theta}}_{ML} = (\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_n)^T$ genom att lösa ut score ekvationen $S(\boldsymbol{\theta}) = \mathbf{0}$.

Fisher informationsmatris är den $n \times n$ symmetrisk Fisher informationsmatris $I(\boldsymbol{\theta})$, den fås genom

$$-\frac{\partial^2}{\partial\theta_i \partial\theta_j} l(\boldsymbol{\theta}), 1 \leq i, j \leq n.$$

$I(\hat{\boldsymbol{\theta}}_{ML})$ kallas för den observerade Fisher informationsmatris (Leonhard Held och Daniel Sabanes Bove, 2014, s.124-125).

Likelihoodkvot-statistikan är

$$G^2 = -2\left[\log \frac{L_0}{L_1}\right] \overset{appr}{\sim} \chi^2(df),$$

där L_0 är maximum likelihood funktion under nollhypotes, L_1 betecknar maximum likelihood funktion under alternativ hypotes och df är antalet parametrar i alternativ hypotes minus antalet parametrar i nollhypotes (Alan Agresti,2002,s.12).

2.6 Akaike information criterion

När vi jämför modeller använder oss av AIC-värde, Akaike information criterion. Definition för AIC är

$$AIC = -2\Big(\log(L) - p\Big) = -2\log(L) + 2p,$$

där L betecknar maximum värdet för likelihood funktion för modellen och p motsvarar antalet estimerade parametrar i modellen. Från formen ser vi att ju högre maximum värdet för likelihood funktion är och färre parametrar modellen har, desto lägre blir AIC-värdet, det vill säga om modellen är ”enkel”. Vi föredrar alltid enklare modeller som förklarar vår data väl, därför väljer vi den modellen som har lägst AIC-värdet (Alan Agresti, 2002,s.216).

2.7 Konfidensintervall

I generaliserad linjär modell skattar vi de okända parametrarna med maximum likelihood metod. När vi inte vet om fördelning för sådana parametrar kan vi använda oss en approximerad normalfördelning. Låt $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_n)^T$ vara parametervektor. Definition för Wald-statistikan är

$$\frac{\hat{\theta}_i - \theta_i}{se(\theta_i)} \overset{appr}{\sim} N(0,1), i = 1, 2, ...n,$$

där standardavvikelse för θ_i är definierad med hjälp av den förväntade Fisher informationsmatrisen

$$se(\hat{\theta}_i) = \sqrt{[I(\hat{\theta}_{ML})^{-1}]_{ii}},$$

Ett $(1-\alpha)\cdot 100$ % konfidensintervall för θ_i beräknas genom

$$\hat{\theta}_i \pm z_{\alpha/2} se(\hat{\theta}_i),$$

där $z_{\alpha/2}$ är kvantil för normalfördelningen (Leonhard Held och Daniel Sabanes Bove, 2014, s.128).

Enligt sats 3.1 (Allan Gut., 2009, s.120-121) om vi låter $\boldsymbol{Z}$ vara en slump $n \times 1$ vektor med vänte vektor $\boldsymbol{\mu}$ och kovarians matris $\boldsymbol{\Lambda}$ och $\boldsymbol{B}$ vara $m \times n$ matris. Med andra ord är $\boldsymbol{Z} \in N(\boldsymbol{\mu}, \boldsymbol{\Lambda})$. Vidare låter vi $\boldsymbol{Y} = \boldsymbol{BZ}$, då är $\boldsymbol{Y} \in N(\boldsymbol{B\mu}, \boldsymbol{B\Lambda B'})$. Nu antar vi

$$\boldsymbol{Z} = \begin{pmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \end{pmatrix} \overset{appr}{\sim} N\left(\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}, I^{-1}_{\hat{\theta}_1, \hat{\theta}_2} \right),$$

där $I_{\hat{\theta}_1, \hat{\theta}_2}$ är den förväntad Fisher informationsmatrisen

$$I_{\hat{\theta}_1, \hat{\theta}_2} = \begin{pmatrix} var(\hat{\theta}_1) & cov(\hat{\theta}_1, \hat{\theta}_2) \\ cov(\hat{\theta}_1, \hat{\theta}_2) & var(\hat{\theta}_2) \end{pmatrix}, \tag{1}$$

Då är

$$\boldsymbol{Y} = \hat{\theta}_1 + \hat{\theta}_2 x = \begin{pmatrix} 1 & x \end{pmatrix} \begin{pmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \end{pmatrix} \overset{appr}{\sim} N\left(\begin{pmatrix} 1 & x \end{pmatrix} \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}, \begin{pmatrix} 1 & x \end{pmatrix} I^{-1}_{\hat{\theta}_1, \hat{\theta}_2} \begin{pmatrix} 1 \\ x \end{pmatrix} \right), \tag{2}$$

Ett $(1-\alpha)\cdot 100\%$ konfidensintervall för $\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}$ är

$$\begin{pmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \end{pmatrix} \pm z_{\alpha/2} \left[\begin{pmatrix} 1 & x \end{pmatrix} I^{-1}_{\hat{\theta}_1, \hat{\theta}_2} \begin{pmatrix} 1 \\ x \end{pmatrix} \right]^{1/2}.$$

3 Splines

Vårt syfte är att hitta en modell som på bästa sätt passar vår data. Ibland har datapunkter inte något linjärt samband när vi använder oss av generaliserad linjär modell, då kan vi utnyttja spline som sätter ihop flera polynomen till ett.

En funktion $f : [a, b] \rightarrow$ kallas för polynom spline med grad $l \geq 1$ med knutar $a = k_1 < ... < k_m = b$, om den uppfyller följande villkor:

- $f(x)$ är $(l-1)$ gånger kontinuerlig differentialbar. Om $l = 1$ kräver vi enbart att $f(x)$ är kontinuerlig, behöver ej vara deriverbar.
- $f(x)$ är ett polynom med grad l i intervallet $[k_j, k_{j+1}], j = 1, 2, ..., m-1$.

(Ludwig Fahrmeir och Thomas Kneib och Stefan Lang och Brian Marx, 2013, s.418).

Vi kan skriva splinefunktionen på följande sätt:

$$F(x) = \alpha_1 B_1(x)\mathbb{1}\{k_1 \leq x \leq k_2\}+\alpha_2 B_2(x)\mathbb{1}\{k_2 < x \leq k_3\}+...+\alpha_m B_m(x)\mathbb{1}\{k_m < x \leq k_{m+1}\},$$

där α_j är konstanter och $B_j(x)$ betecknar basfunktioner, k_j är knutar där basfunktioner möts, $j = 1, 2, ...m$. (John Maindonald och W. John Braun, 2010).

Från splinefunktionen ser vi att den sätter ihop m-antal basfunktioner som består av polynom med hjälp av en indikator som är 1 i det intervallet där vi vill tillämpa denna funktion och 0 annars.

3.1 B-splines

B-spline funktion består av en vektor av knutar $\mathbf{K} = (k_0 < k_1 < ... < k_m)$, en annan vektor $\mathbf{P} = (p_0, p_1, ..., p_m)$ vars värde är skattad och basfunktion som är polynom med grad l. Villkor för B-spline funktion är analogt splinefunktion att den ska vara kontinuerlig och $l-1$ gånger deriverbar. B-spline funktion av variabel X är definierad

$$S_l(x) = \sum_{j=0}^{m} p_j P_{j,l}(x),$$

(Esbjörn Ohlsson och Björn Johansson, 2010, s.108).

där $P_{j,l}(x)$ beräknas med Cox-de Boor rekursions former

$$P_{j,0} = \begin{cases} 1 & k_j \leq x < k_{j+1} \\ 0 & annars, \end{cases}$$

$$P_{j,l}(x) = \frac{x - k_j}{k_{j+l} - u_j} P_{j,l-1}(x) + \frac{k_{j+l+1} - x}{k_{j+l+1} - k_{j+1}} P_{j+1,l-1}(x),$$

(B-spline Basis Function: Definition).

Här får vi en konstant för varje basfunktion. I vår analys kommer vi att använda R:s standard med grad tre för polynom.

4 Databeskrivning

I rapporten presenteras antalet självmord (Avsiktligt självdestruktiv handling) i Sverige mellan år 1969 och 2015 i åldersgrupperna 15–24 år, 25–44 år, 45–54 år, 55–64 år, 65–74 år och 75+ år. För perioden 1997–2015 är data hämtad ifrån Socialstyrelsen. För perioden 1965–1996 kommer data ifrån Statistiska Centralbyrån, men Statistiska Centralbyrån har hämtat sin data ifrån Socialstyrelsen. Eftersom befolkning inte är konstant över tiden, så väljer vi representera antalet självmord per 100 000 invånare för att kunna lättare jämföra mellan olika åldersgrupper.

Statistikdatabasen behandlar endast avlidna personer som är folkbokförda i Sverige(Socialstyrelsen, dödsorsaker, 2015). Läkare skriver ut ett dödsbevis och ett dödsorsaksintyg för den avlidne. Dödsbeviset skickas till skattemyndigheten och dödsorsaksintyget sänds till Socialstyrelsen för registreringen (Socialstyrelsen, Dödsorsaksstatistik, 2010).

Tabell 1 visar variabelbeteckning för denna rapport.

Tabell 1: Variabelbeteckning

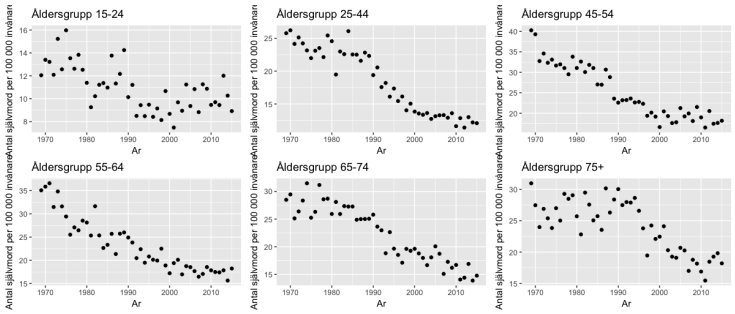
Variabel	**Beteckning**
Respons	$Y = 1$: begår självmord, $Y = 0$: begår inte självmord
År	X
Andel självmord	$P(Y = 1\|X = x_i) = \pi(x_i)$
Man	m
Kvinna	k

5 Analys av data

Vi vill i denna studie undersöka tidstrenden inom självmord under perioden 1969–2015 för åldersgrupperna 15–24 år, 25–44 år, 45–54 år, 55–64 år, 65–74 år

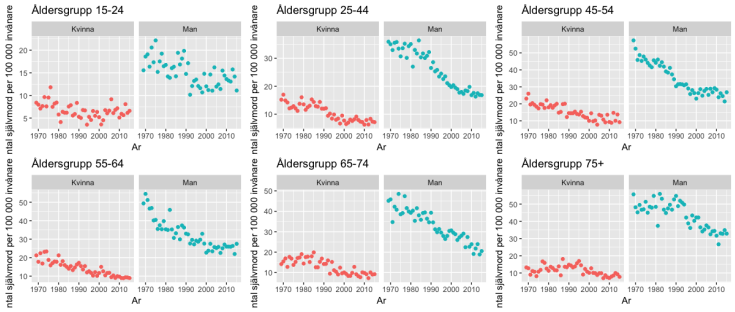
och 75+ år i Sverige. Vi är intresserade av samspel mellan kön och kalenderår samt samspel mellan de sex åldersgrupperna och kalenderår.

Vi undersöker hur vår data ser ut för den totala befolkningen. Figur 1 visar scatter-plottar för självmord mellan år 1969 och 2015 för åldersgrupperna 15–24 år, 25–44 år, 45–54 år, 55–64 år, 65–74 år och 75+ år i Sverige. Vi ser för åldersgruppen 15–24 minskar antalet självmord först, sedan ökar det igen. För de övriga åldersgrupperna verkar antalet självmord avta med tiden.



Figur 1: Scatter-plot av sex åldersgrupper för hela befolkning

Nu analyserar vi vidare hur vår data ser ut när vi separerar kön. Figur 2 visar scatter-plottar för självmord mellan år 1969 och 2015 för åldersgrupperna 15–24 år, 25–44 år, 45–54 år, 55–64 år, 65–74 år och 75+ år i Sverige för respektive kön. Vi ser tydligt att det är fler män som tar självmord än kvinnor, trenden ser ganska lika ut för båda kön i alla åldrar.



Figur 2: Scatter-plot av sex åldersgrupper för respektive kön, Röd är kvinnor, blå är män

6 Modeller och analys

Vi vill som sagt i denna studie undersöka tidstrenden inom självmord under år 1969 och 2015 för åldersgrupperna 15–24 år, 25–44 år, 45–54 år, 55–64 år, 65–74 år och 75+ år i Sverige, jämför sedan mellan kön och åldersgrupper.

För att undersöka detta behöver vi hitta en lämplig modell som kan beskriva vår data väl. Vi kommer att använda B-spline för att modellera hur självmordsfrekvens varierar under åren. Först vill vi undersöka om det finns signifikant samspel mellan kön och kalenderår, sedan vill vi analysera om samspel uppstår mellan åldersgrupper och kalenderår. Efter det vill vi skapa ett Waldkonfidensintervall för att se hur säkert våra B-splines är. I slutet av analysen vill vi testa om det finns signifikant ökad trend för ungdomsgruppen mellan år 2000 och 2014.

6.1 Linjär regression

Vi börjar med att undersöka hur den linjära modellen passar till vår data. Den linjära modellen har några modellförutsättningar som måste uppfyllas.

1 Linjäritet: Regressionsmodellen har linjärt samband

2 Ingen autokorrelation: Det finns inget tydligt samband mellan feltermen

3 Normalitet: Residualer ska vara normalfördelade med en konstant varians

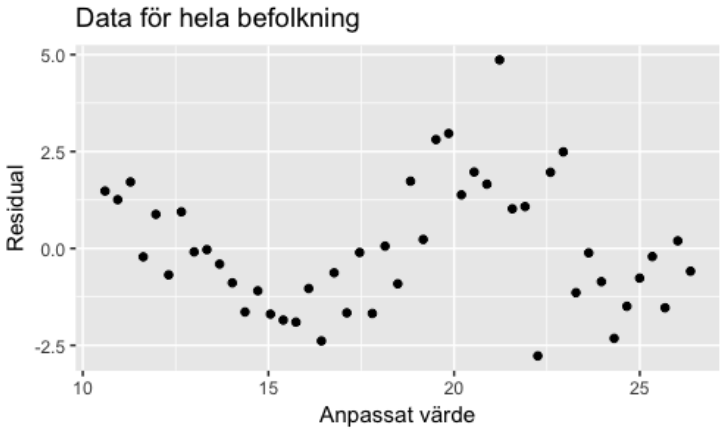
(Patrik Andersson och Joanna Tyrcha, 2014, s.23-25)

Låt W_i vara antalet självmord per 100 000 invånare, $i = 1, 2, ..., 47$. Vi definierar en linjär modell för vår data.

$$W_i = \alpha + \beta x_i + \epsilon_i, i = 1, 2, ...47,$$

där $x_i = 1969, 1970, ...2015$.

Figur 3 visar residualplot för linjär regressionsmodell för åldersgruppen 25–44 år. Här ser vi tydligt att residualerna verkar följa ett tredjegrads polynom och konstant varians existerar inte, därmed utesluter vi linjär modell för vår data. Residualplottar för de övriga grupperna finns i Appendix 8.1, figur 7.



Figur 3: Residualplot med Linjär regression för åldersgruppen 25-44 år

6.2 Generaliserad linjär modell

Nu när vi vet att linjär modell inte passar vår data vill vi pröva en generaliserad linjär modell. Vår responsvariabel Y är antingen 1, en person begår självmord eller 0, en person inte begår självmord, så $Y \sim Bern(\pi)$, där π är sannolikheten för en slumpmässig vald person begår självmord. Då är $Z = \sum_{i=1}^{n} Y_i \sim Bin(n, \pi)$, där alla Y_i är oberoende och n är befolkningsmängd i Sverige.

Vi notera att i denna modell har vi antagit antalet självmord per år är binomialfördelat, det vill säga vi antar att alla människor i befolkningen är lika fördelade. De har lika stor sannolikhet π att begå självmord ett givet år. Det vi vill ifrågasätta är det rimligt anta att människor beter sig oberoende och att alla människor har samma sannolikhet att ta livet av sig givet år? Självklart är det orimligt, till exempel självmordssannolikheten varierar väldigt mycket från person till person. En person som mår dåligt har väldigt stor sannolikhet att begå självmord medan en annan person mår bra, är nöjd med sig själv är sannolikheten att begå självmord väldigt liten. Vi behöver individdata med fler kovariater så som psykisk ohälsa, ekonomisk förutsättning, missbruk, familjeförhållande, utbildningsnivå och sysselsättning med mera. Eftersom vi inte har till gång till fler kovariater än år, så antar vi att antalet självmord per år är binomialfördelat i alla fall, även om det inte är realistiskt.

Nu vill vi skriva om sannolikhetsfördelning på formen

$$f(y_i; \theta_i) = A(\theta_i)B(y_i)exp(y_i Q(\theta_i)), i = 1, 2, ...n,$$

för någon funktion A, B och Q. Termen $Q(\theta)$ kallas för den naturliga parametern.

Sannolikhetsfördelning för Z är

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

där y är antalet personer som begår självmord, $\theta = \pi$. Vi kan skriva om sannolikhetsfördelning som

$$\binom{n}{y}\exp\{y\log(\theta) + (n-y)\log(1-\theta)\} =$$
$$= \binom{n}{y}\exp\{y(\log(\theta) - \log(1-\theta))\}\exp\{n\log(1-\theta)\} =$$
$$= \binom{n}{y}\exp\{y\log(\frac{\theta}{1-\theta})\}\exp\{n\log(1-\theta)\},$$

Vi får då

$$A(\theta) = \exp\{n\log(1-\theta)\},$$

$$B(y) = \binom{n}{y},$$

$$Q(\theta) = \log(\frac{\theta}{1-\theta}) = logit(\theta),$$

Vi kommer fram till att vår länkfunktion är logit.

Då vi har en generaliserad linjär modell med en binomialfördelad responsvariabel med en länkfunktion logit, så kan vi formulera vår logistiska regression på följande sätt

$$logit(\pi(x_i)) = \alpha + \beta x_i, \text{ där } i = 1, 2, \ldots\ldots, 47,$$

där x_i är kalenderår, vilket är från och med år 1969 till och med år 2015.

I Appendix 8.1, figur 8 visas datapunkter och en skattad logistiska regressionslinje för de sex åldersgrupperna, vi ser att linjerna inte passar våra datapunkter. AIC-värden för modellerna finns i Appendix 8.2, Tabell 4.

6.3 Icke linjära tidstrender

Vi bestämmer oss för att pröva B-splines och undersöker om den modellen passar bättre för våra datapunkter. Det finns många sätt att konstruera B-splines. Här väljer vi ett av dem. Vi bestämmer hur många knutar vi ska ha och var vi ska placera dem så att vi får ett så litet AIC-värde som möjligt.

Vi använder R:s inbyggda funktion. R använder grad tre som standard, df är frihetsgrad. Antal inre knutar blir df minus frihetsgrader. Sedan finns det två stycken yttre knutar som sitter, i vårt fall är år 1969 den första punkten och år 2015 den sista punkten. Efter man har bestämt antalet inre knutar beräknar R kvantiler för att bestämma var brytpunkter ska vara. Till exempel om vi sätter df till 6, då får vi $6 - 3 = 3$ inre knutar, R sätter kvantilerna till 25%, 50% och 75%. Vi får då inre knutar $1980.5, 1992, 2003.5$.

I tidigare analys av data såg vi att antalet självmord var fler för män än kvinnor och trenden såg ganska lika ut för båda könen, därmed vill vi analysera om det finns något samspel mellan kön och kalenderår. Vi har två modeller, en med samspel, en utan samspel. För att kunna jämföra två modeller tillsammans måste vi välja lika många knutar till båda modellerna. Först väljer vi ut antal knutar med hjälp av AIC-värde. minst antal knutar är två, viket är startpunkten och slutpunkten. I vårt fall är år 1969 och år 2015. Sedan använder vi likelihood-test för att avgöra om det finns något signifikant samspel. Tabell 2 visar AIC-värde för de två modellerna med olika antal knutar.

Tabell 2: AIC B-spline för samspel och utan samspel mellan kön och kalenderår

Antal knutar	AIC utan samspel	AIC med samspel
2	9427.768	9430.275
3	9414.126	9417.784
4	9387.987	9393.769
5	9394.055	9401.517
6	9389.108	9396.139
7	9389.405	9399.761
8	9388.756	9401.829

Enligt Tabell 2 ska vi välja modellen med fyra knutar, eftersom den modellen har lägst AIC-värde. Nu ska vi göra likelihood-test.

Vår modell är

$$logit(\pi(x_i)) = \alpha + \beta_1 \mathbb{1}\{m\}_i + S_1(x_i) + S_2(x_i)\mathbb{1}\{m\}_i,$$

där $\mathbb{1}$ är indikator, m är man, x_i är kalenderår och $S_1 + S_2$ är splinefunktion för män.

Vidare är

$$S_1(x_i) = \beta_{11}b_1(x_i) + \beta_{12}b_2(x_i) + ... + \beta_{1k}b_k(x_i),$$

och

$$S_2(x_i) = \beta_{21}b_1(x_i) + \beta_{22}b_2(x_i) + ... + \beta_{2k}b_k(x_i),$$

där $b_j, j = 1, 2, ..., k = 5$ är basfunktioner.

Vår nollhypotes

$$H_0 : \beta_{21} = \beta_{22} = \beta_{23} = \beta_{24} = \beta_{25} = 0,$$

det vill säga $S_2 = 0$ och splinefunktionen är samma för män och kvinnor.

Alternativ hypotes

$$H_a : \beta_{2j} \neq 0 \text{ för åtminstone en av } \beta_{2j}.$$

Vi sätter signifikansnivå på 5%. P-värde blir 0.5185 (utskrift av testet finns i Appendix 8.2, Tabell 6), så vi ska inte förkasta nollhypotesen, vilket medför att vi behåller modellen som inte har samspel.

Nu när vi har konstaterat att kön och kalenderår inte har något signifikant samspel, så är oddskvoten mellan oddset för att en man begår självmord och oddset för att en kvinna begår självmord konstant över åren. För kvinnor ser logistisk regressionsmodell med splinefunktionen ut på följande sätt

$$logit(\pi(x_i)) = \alpha + S(x_i),$$

och logistisk regressionsmodell med splinefunktionen för män ser ut på detta sätt

$$logit(\pi(x_i)) = \alpha + \beta + S(x_i),$$

där $S(x_i)$ är splinefunktion för kalenderår.

Låt $\pi(x_i)_m$ vara $\pi(x_i)$ för män, $\pi(x_i)_k$ vara $\pi(x_i)$ för kvinnor och OR vara oddskvot mellan oddset för en man begår självmord och oddset för en kvinna begår självmord, då får vi log oddskvot

$$log(OR) = log\big(\frac{\pi(x_i)_m/(1-\pi(x_i)_m)}{\pi(x_i)_k/(1-\pi(x_i)_k)}\big) = logit(\pi(x_i)_m) - logit(\pi(x_i)_k) =$$

$$= \alpha + \beta + S(x_i) - \alpha - S(x_i) = \beta \Longrightarrow OR = e^{\beta},$$

Från summary av modellen (Appendix 8.2, Tabell 7) får vi $\hat{\beta} = 0.900233 \Longrightarrow OR = e^{0.900233} = 2.46$.

Nu vill vi konstruera ett 95% konfidensintervall för oddskvoten. Vi vet att

$$\hat{\beta} \overset{appr.}{\sim} N(\beta, var(\hat{\beta})) = N(\beta, 0.00855^2),$$

Standardavvikelse för $\hat{\beta}$ kommer från summary (Appendix 8.2, Tabell 7). Ett 95% konfidensintervall för β ges av

$$(\hat{\beta} \pm 1.96\sqrt{var(\hat{\beta})}) = (0.900233 \pm 1.96 \cdot 0.008555) = (0.8834652, 0.9170008),$$

Genom transformation får vi ett 95% konfidensintervall för oddskvoten

$$\text{KI(OR)}=(e^{0.8834652}, e^{0.9170008}) = (2.419268, 2.501776).$$

Detta kan tolkas som att oddset för att en man begår självmord är 2.46 gånger högre än oddset för att en kvinna begår självmord. Ett 95% intervall för oddskvoten är $(2.42, 2.50)$. Eftersom konfidensintervall för oddskvoten är en bit ifrån 1, så är oddskvoten stark.

Eftersom kön och kalenderår inte har något samspel, så analyserar vi vidare hur tidstrenden inom självmord ser ut för hela befolkningen. Vi undersöker först om det finns något samspel mellan åldersgrupper och kalenderår. Modellen där alla åldersgrupper är med kan beskrivas

$$logit(\pi(x_i)) = \alpha + \beta_2 \mathbb{1}\{aldrsgrupp2\} + \beta_3 \mathbb{1}\{aldrsgrupp3\} + ... + \beta_6 \mathbb{1}\{aldrsgrupp6\} + S_3(x_i) +$$

$$+ S_{21}(x_i)\mathbb{1}\{aldrsgrupp2\} + S_{31}(x_i)\mathbb{1}\{aldrsgrupp3\} + ... + S_{61}(x_i)\mathbb{1}\{aldrsgrupp6\},$$

där $\mathbb{1}$ är indikator. S_3 är splinefunktion för åldersgrupp ett.

H_0 : alla $S_{n1} = 0$,

H_a : åtminstone ett av $S_{n1} \neq 0$, där $n = 2, 3, ..., 6$.

Vi väljer lika många knutar för båda modellerna enligt AIC-värde. Likelihood-test ger p-värde som är mindre än $2.2 \cdot 10^{-16}$, vilket medför att vi ska förkasta nollhypotesen på alla signifikansnivåer. Alltså det finns signifikant samspel mellan åldersgrupper och kalenderår.

Vi tar sedan bort åldersgrupp ett, och gör samma test. Vi får samma p-värde som ovan. Vi testar vidare med att reducera åldersgrupperna successivt. Tabell 3 nedan visar resultatet. Inget p-värde är signifikant på 1% signifikansnivå. Vi drar slutsatsen om att det inte finns något signifikant samspel mellan åldersgrupper och kalenderår, därmed är splinefunktion för olika åldersgrupper inte samma. Vi vill här poängtera att åldersgrupper tre och fem hade p-värde på 0.8526%, vilket är ganska nära 1%.

Tabell 3: Likelihood-test B-spline för samspel mellan åldersgrupper och kalenderår

Aldersgrupper	P-varde
1,2,3,4,5,6	< 0.0001
2,3,4,5,6	< 0.0001
3,4,5,6	< 0.0001
3,4,5	< 0.0001
4,5	< 0.0001
3,5	0.008526

Eftersom samspel uppstår mellan åldersgrupper och kalenderår, medför detta att splinefunktioner för de olika åldersgrupper är olika, det vill säga de kan ha olika antal knutar i sina splinefunktioner.

Nu vill vi göra B-splines för varje åldersgrupp för hela befolkningen. Vi väljer den modellen som har lägst AIC-värde. Tabell 8, 9, 10, 11, 12 och 13 i Appendix 8.2 visar alla AIC-värde för de sex åldersgrupperna. För åldersgruppen 15–24 är AIC-värdet ganska högt för de mindre antalet knutar, därför tas dem inte med i tabellen. Enligt tabellerna är antalet knutar för åldersgruppen 15–24 år 11, för åldersgruppen 25–44 år 4, för åldersgruppen 45–54 år 5, för åldersgruppen 55–64 år 3, för åldersgruppen 65–74 år 2 och för åldersgruppen 75+ år 3.

6.3.1 Wald-konfidensintervall

Nu när vi har valt ut vår modell är det bra att undersöka med hjälp av Wald-konfidensintervall hur säker vår skattning är. Ett 95% konfidensintervall för $logit(\pi(x_i))$ är

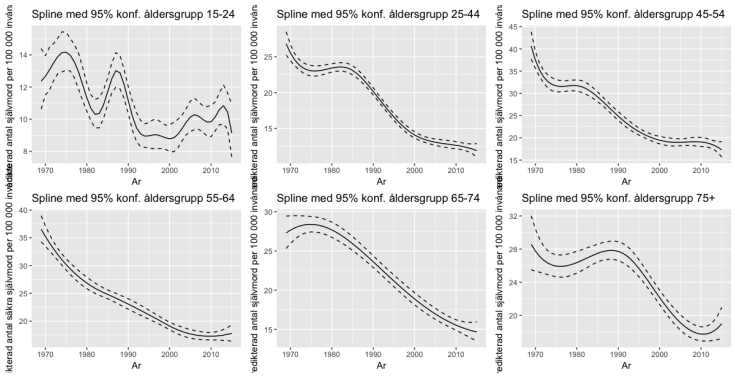
$$logit(\pi(x_i)) = \hat{\alpha} + \hat{\beta}x_i \pm z_{0.05/2} * \left[\begin{pmatrix} 1 & x_i \end{pmatrix} I^{-1}_{\hat{\alpha},\hat{\beta}} \begin{pmatrix} 1 \\ x_i \end{pmatrix} \right]^{1/2} = (\xi_u, \xi_l),$$

där $I^{-1}_{\hat{\alpha},\hat{\beta}}$ är den inverterade Fisher informationsmatrisen, ξ_u är övregräns och ξ_l är undregräns.

Härifrån kan vi transformera detta konfidensintervall för att få ett 95% konfidensintervall för $\pi(x_i)$.

$$KI(\pi(x_i)) = \left(\frac{e^{\xi_u}}{1 + e^{\xi_u}}, \frac{e^{\xi_l}}{1 + e^{\xi_l}} \right).$$

Figur 5 visar B-splines med 95% konfidensintervall för de sex åldersgrupperna för hela befolkningen. Vi ser att konfidensintervall för åldersgrupper 25–44 år, 45–54 år och 55–64 år ser ganska "bra" ut. Det är inte för brett eller för smalt någonstans. Vi säger att B-spline förklarar bra på hur självmord beror på år för de åldersgrupperna. För den allra yngsta gruppen är konfidensintervallet något större mellan år 1969 och 1976 samt mellan år 1994 och 2015. För åldersgruppen 65–74 år är bandet större för perioden 1969–1975 och för perioden 2013–2915, Åldersgruppen 75+ år är bandet större för perioden 1969-1990 samt de senaste två åren. Att bandet är något större betyder att det är mindre säkert hur vår B-splines förklarar antal självmord under dem åren.



Figur 4: Spline med 95% konfidensintervall för sex åldersgrupper för hela befolkningen

6.3.2 Ökad tidstrend för ungdomar

Danuta Wasserman, professor i psykiatri och suicidologi skriver i **Läkartidningen**, Nr 50, 2004 att självmord har ökat för ungdomar i Sverige. Vi vill här och med undersöka med statistiska metoder om detta är ett tillfällighet eller om det finns en signifikant ökad tidstrend för ungdomar

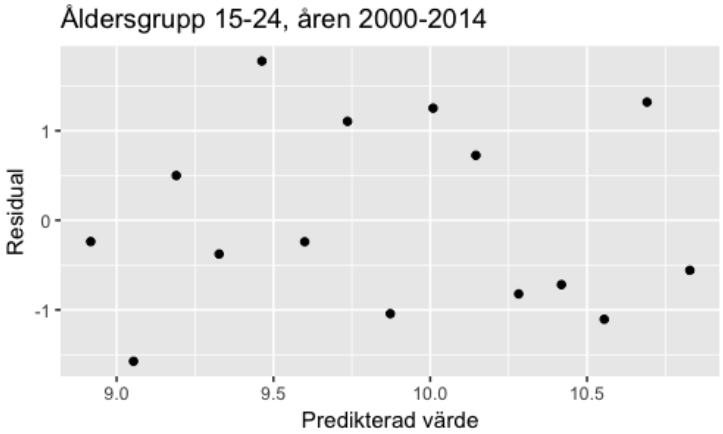
Vi analyserar åldersgruppen 15–24 mellan år 2000 och 2014. Vi antar att linjär modell passar bra till vår data. Eftersom vi tar ut en del av datan, och kommer att testa om en kofficient är lika med noll, så ska vi centrera vår förklarande variabel för att undvika korrelation mellan α och β. Låt W_j vara antalet självmord

per 100 000 invånare, $j = 1, 2, ..., 15$. Den linjära modellen kan formuleras

$$W_j = \alpha + \beta(x_j - \bar{x}) + \epsilon_j,$$

där $j = 1, 2, ...15, x_j = 2000, 2001, ...2014$ och $\bar{x}$ är 2007.

Figur 5 visar residualplot för modellen. Det ser ut som att residualer inte har något tydligt samband. Vi säger att modellförutsättningarna för linjär modell är uppfyllda.



Figur 5: Residualplot för linjär modell åldersgruppen 15-24, åren 2000-2014

Nu vill vi testa om det finns någon signifikant ökad tidstrend.

$$H_0 : \beta = 0,$$
$$H_a : \beta > 0.$$

Från summary av modellen (Appendix 8.2, tabell 14) får vi $\hat{\beta} = 0.13646$ och $\sqrt{v(\hat{\beta})} = 0.06419$.

Vi vet att

$$\hat{\beta} \sim N(\beta, var(\hat{\beta})) = N(\beta, 0.06419^2),$$

Vi får då teststorhet

$$T = \frac{\hat{\beta} - 0}{\sqrt{v(\hat{\beta})}} \sim N(0, 1),$$

Så

$$T = \frac{\hat{\beta} - 0}{\sqrt{v(\hat{\beta})}} = \frac{0.13646}{0.06419} = 2.13 > z_{0.05} = 1.64.$$

Vi förkastar nollhypotesen med 5% signifikansnivå till förmån för alternativa hypotesen, det vill säga tid har inverkat på självmorden. Ju högre årtalet är desto större är självmordstalet.

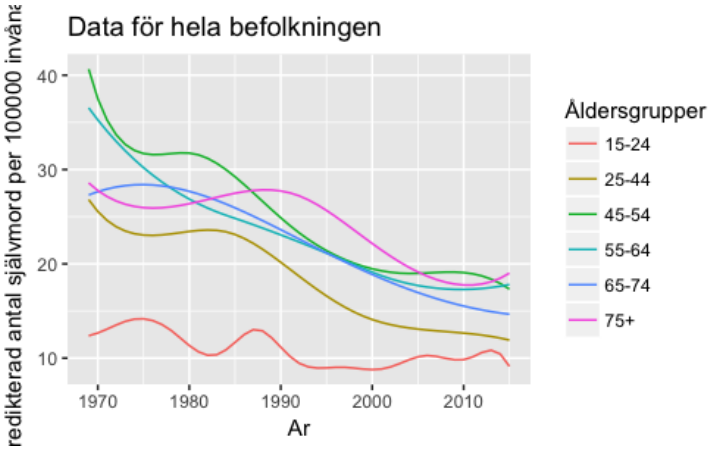
6.3.3 Tidstrender för de sex åldersgrupperna

Nu är vi intresserad av att plotta alla sex B-splines tillsammans för att visuellt analysera hur tidstrenderna inom självmord är för de olika åldersgrupperna.

Figur 6 visar B-splines för alla sex grupper för hela befolkningen. Det visar sig att åldersgruppen 15–24 år har lägst antal självmord genom tiden jämfört med de övriga åldersgrupperna. Självmordstalet är lägst mellan år 1993 och 2001 för denna åldersgrupp. Antalet självmord för åldersgruppen 25–44 år är högre än åldersgruppen 15–24, men lägre än resterade åldersgrupperna genom tiden.

För åldersgrupper 25–44 år, 45–54 år, 55–64 år och 65–74 år verkar antalet självmord avta sedan år 1985 och fortsätter så fram tills år 2005. Åldersgrupper 25–44 år och 65–74 år fortsätter avta efter år 2005. Mellan år 2005 och 2011 ändras inte antalet självmord för åldersgruppen 55–64 år, efter år 2011 ökar antalet självmord något fram tills år 2015 för denna åldersgrupp. Självmordstalet ökar något för åldersgruppen 45–54 år mellan år 2006 och 2011, sedan avtar det succesivt fram tills år 2015.

Åldersgruppen 45–54 har högsta självmordstalet mellan år 1969 och 1986 samt mellan år 2005 och 2012. Mellan år 1987 och 2005 samt år 2013 och 2015 har den allra äldsta gruppen högst antal självmord. Antalet självmord är som högst för åldersgruppen 75+ år 1990, sedan minskar det fram tills år 2010, efter detta året ökar det igen ända fram tills år 2015. Eftersom konfidensintervallet är brett mellan år 2010 och 2015 för denna åldersgrupp, kan vi med mindre säkerhet dra slutsats om att antalet självmord ökar mellan dessa år och antalet självmord är högst under de senaste två åren.



Figur 6: B-spline för alla sex åldersgrupper

7 Diskussion

I analysen kommer vi fram till att kön har inget signifikant samspel med kalenderår. Oddset för att en man begår självmord är 2.46 gånger högre än oddset för att en kvinna begår självmord. Ett 95% konfidensintervall för oddskvoten är (2.42, 2.50), det vill säga tidstrenden är samma för män som för kvinnor, men antalet självmord är större för män än kvinnor. Det överensstämmer med vad författare Henrik Nordin skriver i **välfärd**, Nr 4, 2009.

Vi kommer även fram till att det finns ett signifikant samspel mellan åldersgrupper och kalenderår för hela befolkningen. Vi konstruerar en B-spline för varje åldersgrupp. Ett 95 % Wald-konfidensintervall för splinefuktioner visar att konfidensintervallet för åldersgruppen 15–24 något större mellan år 1969 och 1976 samt mellan år 1994 och 2015. För åldersgruppen 65–74 år är intervallet större mellan år 1994 och 2015. Hos den allra äldsta åldersgruppen 75+ år är bandet något större mellan år 1968 och 1990, samt de två senaste åren. Konfidensintervallet ser "bra" ut för åldersgrupper 25–44 år, 45–54 år och 55–64 år.

I inledningen nämner vi att Danuta Wasserman, professor i psykiatri och suicidologi är orolig för att självmordsfrekvensen bland unga har ökat. Vi analyserar åldersgruppen 15-24 år mellan år 2000 och 2014. Det visar sig att det finns signifikant ökad tidstrend för ungdomar under dessa år.

Vi analyserar tidstrenden för de sex åldersgrupperna genom att plotta de sex splines i en och samma figur. Där ser vi att åldersgruppen 15–24 år har lägst antal självmord genom tiden jämfört med de övriga åldersgrupperna. Självmordstalet är lägst mellan år 1993 och 2001 för denna åldersgrupp. Antalet självmord för åldersgruppen 25–44 år är högre än åldersgruppen 15–24 år, men lägre än resterade åldersgrupperna genom tiden.

För åldersgrupper 25–44 år och 65–74 år avtar antalet självmord sedan år 1983. Självmordstalet avtar mellan år 1969 och 2010 för åldersgruppen 55–64 år, de senaste fem år har det inte ändrats.

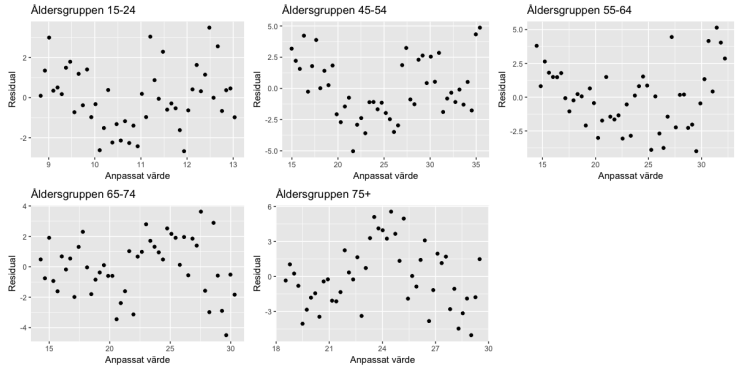
Åldersgruppen 45–54 år har högst självmordstal mellan år 1969 och 1986 samt mellan år 2005 och 2012. Mellan år 1987 och 2005 samt mellan år 2013 och 2015 har den allra äldsta gruppen högst antal självmord.

I vår analys är självmord vanligast för 75+ år mellan år 1987 och 2005 samt mellan år 2013 och 2015. Så vår analys överensstämmer inte med det författaren Henrik Nordin skriver i **välfärd**, Nr 4, 2009 att självmord är vanligaste bland äldre män. Detta kan bero på att Henrik Nordin använder data med säkert självmord och osäkert självmord (skadehändelser med oklar avsikt) och han undersöker om åldersgrupperna 65+ och 85+ i sin analys.

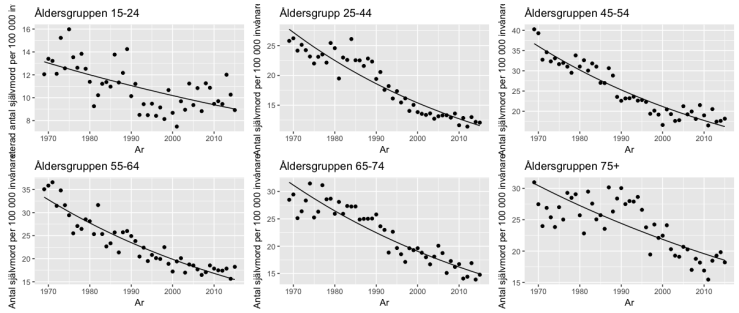
Det hade varit intressant om vi hade haft flera förklaringsvariabler så som psykisk ohälsa, missbruk av alkohol och droger, utbildningsnivå, ekonomisk förutsättning, familjförhållande samt sysselsättning med mera, då skulle vår analys bli bättre. Man kan ju ifrågasätta varför självmordstalet är högre för män än kvinnor? Några anledningar till detta kan vara män är sämre än kvinnor på att söka hjälp när de mår psykisk dåligt. De är våldsammare när de tar livet av sig själva. På tal på detta skulle det vara intressant att i framtiden även ta med antalet självmordsförsök, för att analysera hur antalet självmordsförsök och antalet självmord relaterad till respektive kön. Det skulle också vara intressant att analysera tidstrend inom självmord geografiskt.

8 Appendix

8.1 Figurer



Figur 7: Residualplott med linjär regressionsmodell i olika åldersgrupper för hela befolkningen



Figur 8: Logistisk regressions linje i olika åldersgrupper för hela befolkningen

8.2 Tabller

Tabell 4: AIC logistisk modell för hela befolkning

Åldersgrupper	AIC
15-24	425.04
25-44	539.71
45-54	442.48
55-64	421.57
65-74	398.49
75+	423.83

Tabell 5: AIC B-spline för hela befolkning

Åldersgruppen	AIC
15-24	385.04
25-44	442.11
45-54	409.21
55-64	406.09
65-74	379.99
75+	377.67

Tabell 6: likelihoodtest B-spline för kön och kalenderår

```
Model 1: cbind(antal, befolkning - antal) ~ Kon + bs(Ar, df = 5, degree = 3,
    intercept = FALSE)
Model 2: cbind(antal, befolkning - antal) ~ Kon * bs(Ar, df = 5, degree = 3,
    intercept = FALSE)
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       557     5766.8
2       552     5762.6  5   4.2176   0.5185
```

Tabell 7: Summary B-spline kön och kalenderår utan samspel

```
Coefficients:
                                            Estimate Std. Error  z value
(Intercept)                                -8.721448   0.018174 -479.887
KonMan                                      0.900233   0.008555  105.230
bs(Ar, df = 5, degree = 3, intercept = FALSE)1 -0.199117   0.034034   -5.850
bs(Ar, df = 5, degree = 3, intercept = FALSE)2 -0.045412   0.025558   -1.777
bs(Ar, df = 5, degree = 3, intercept = FALSE)3 -0.641474   0.034314  -18.694
bs(Ar, df = 5, degree = 3, intercept = FALSE)4 -0.616154   0.028749  -21.432
bs(Ar, df = 5, degree = 3, intercept = FALSE)5 -0.676033   0.027543  -24.545
                                           Pr(>|z|)
(Intercept)                                 < 2e-16 ***
KonMan                                      < 2e-16 ***
bs(Ar, df = 5, degree = 3, intercept = FALSE)1  4.9e-09 ***
bs(Ar, df = 5, degree = 3, intercept = FALSE)2   0.0756 .
bs(Ar, df = 5, degree = 3, intercept = FALSE)3  < 2e-16 ***
bs(Ar, df = 5, degree = 3, intercept = FALSE)4  < 2e-16 ***
bs(Ar, df = 5, degree = 3, intercept = FALSE)5  < 2e-16 ***
```

Tabell 8: AIC B-spline åldersgrupp 15-24 for hela befolkningen

Antal knutar	AIC
9	386.96
10	390.35
11	385.04
12	385.83
13	387.37

Tabell 9: AIC B-spline åldersgrupp 25-44 for hela befolkningen

Antal knutar	AIC
2	480.34
3	468.41
4	442.11
5	448.13
6	443.56

Tabell 10: AIC B-spline åldersgrupp 45-54 for hela befolkningen

Antal knutar	AIC
2	426.78
3	428.56
4	409.51
5	409.21
6	411.89
7	412.99

Tabell 11: AIC B-spline åldersgrupp 55-64 for hela befolkningen

Antal knutar	AIC
2	408.02
3	406.09
4	408.20
5	409.21
6	410.30

Tabell 12: AIC B-spline åldersgrupp 65-74 for hela befolkningen

Antal knutar	AIC
2	379.99
3	381.70
4	381.39
5	383.51
6	382.26

Tabell 13: AIC B-spline åldersgrupp 75+ for hela befolkningen

Antal knutar	AIC
2	395.11
3	377.67
4	381.36
5	381.64
6	384.25

Tabell 14: Summary linjär modell för åldersgruppen 15-24 mellan år 2000-2014

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.87267    0.27732  35.600 2.41e-14 ***
Ar           0.13646    0.06419   2.126   0.0532 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.074 on 13 degrees of freedom
Multiple R-squared:  0.258,     Adjusted R-squared:  0.2009
F-statistic:  4.52 on 1 and 13 DF,  p-value: 0.05324
```

9 Referenser

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

Patrik Andersson och Joanna Tyrcha, 2014, *Notes in Econometrics,* Matematiska Institutionen, Stockholms universitet

Ludwig Fahrmeir och Thomas Kneib och Stefan Lang och Brian Marx, 2013, *Regression, Models, Methods and Applications, chapter 8,* Springer Heidelberg New York Dordrecht London

Leonhard Held och Daniel Sabanes Bove, 2014, *Applied Statistical Inference,* Springer Heidelberg New York Dordrecht London

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

Esbjörn Ohlsson och Björn Johansson, 2010, *Non-Life Insurance Pricing with Generalized Linear Model, Chapter 5,* Springer Heidelberg Dordrecht London New York

Allan Gut, 2009, *An Intermediate Course in Probability, Second Edition.* Springer Dordrecht Heidelberg London New York

Rolf Sundberg, 2016, *Lineära Statistiska Modeller,* Matematiska Institutionen, Stockholms universitet

Socialstyrelsen-folkhälsorapport, 2009, *Ungdomars hälsa*, Hämtad 2017-05-24

Danuta Wasserman, 2004, *Självmord bland unga ökar i Sverige.* Läkartidningen, Hämtad 2017-05-24

Henrik Nordin, 2009, *Sjävmord är vanligaste bland äldre män*, Välfärd, Hämtad 2017-05-24

Socialstyrelsen, dödsorsaker, 2015, http://www.socialstyrelsen.se/statistik/statistikdatabas/hjalp/dodsorsaker, Hämtad 2017-05-24

Socialstyrelsen, Dödsorsaksstatistik, 2010, *Sveriges officiella statistik, Hälso- och sjukvård, Sidan 7*, Hämtad 2017-05-24

B-spline Basis Function: Definition, https://www.cs.mtu.edu/ shene/COURSES/cs3621/NOTES/spline/B-spline/bspline-basis.html, Hämtad 2017-05-24