Stockholms universitet

Analys av tidstrender av andelen som påbörjar universitets/högskole-studier efter gymnasiet

Daria Medvedeva

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2018:6
Matematisk statistik
Juni 2018

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Matematisk statistik
Stockholms universitet
Kandidatuppsats **2018:6**
http://www.math.su.se/matstat

Analys av tidstrender av andelen som påbörjar universitets/högskole-studier efter gymnasiet

Daria Medvedeva*

Juni 2018

Sammanfattning

Denna studie använder sig av data över andelen personer som gått vidare till universitet/högskola inom tre år efter avslutad gymnasial utbildning för examensåren 2000 till 2011. Arbetets syfte är att undersöka trender av andelen personer som påbörjat universitets/högskole-studier inom tre år i olika högskoleförberedande program. Vi vill även undersöka trender för kvinnor och män samt för personer med svensk och utländsk bakgund inom dessa program. Målet är att hitta modeller som bäst förklarar dessa trender, och vi använder oss av logistisk regressionsmodell med B-spline för detta ändamål.

I analysen kommer vi fram till att högskoleförberedande programmen har olika trender. Estetiska programmet har en kontinuerligt ökande trend mellan åren 2000 och 2009, medan de andra programmen inte uppvisar denna kontinuerliga ökning mellan samma år. Samtliga program har en ökande trend från år 2006 till år 2009 och en sjunkande trend från år 2009 till år 2011.

I vidare analys kommer vi fram till att andelen män ökar mer än andelen kvinnor inom Naturvetenskaps- och Samhällsvetenskapsprogrammet. För samtliga program, förutom International Baccalaureate där ingen signifikant skillnad mellan könen påvisades, är andelen kvinnor högre än andelen män. Denna trend gäller för alla år.

Vidare resultat visade att andelen med utländsk bakgrund ökar mer än med svensk bakgrund inom Estetiska- och Samhällsvetenskapsprogrammet. Andelen med utländsk bakgrund överstiger den med svensk bakgrund efter år 2007 inom Estetiska programmet, och inom Samhällsvetenskapsprogrammet är andelen med utländsk bakgrund högre för samtliga år. För International Baccalaureate gäller det motsatta gentemot Samhällsvetenskapsprogrammet och för Naturvetenskapsprogrammet påvisades ingen signifikant skillnad mellan bakgrunderna.

Slutligen visade resultat att andelen män med svensk bakgrund ökar mer än andelen kvinnor med svensk bakgrund inom Samhällsvetenskapsprogrammet. Inom Naturvetenskapsprogrammet är andelen utländska kvinnor högre än andelen svenska kvinnor, och denna trend gäller för alla år.

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige.
E-post: dasha.medv@gmail.com. Handledare: Jan-Olov Persson.

Förord

Detta arbete utgör mitt examensarbete motsvarande 15 högskolepoäng i matematisk statistik vid Matematiska intitutionen på Stockholms Universitet. Jag vill ge ett väldigt stort tack till min handledare Jan-Olov Persson för all hjälp, råd och feedback jag fått under arbetets gång.

Jag vill även tacka min studiekamrat och vän Magnus Pierrau för synpunkter, intressanta diskussioner, motivation och stöd samt tacka mina vänner Niklas Livchitz, Viveka Rangne och Åsa Mählquist Jonasson för peptalk och stöd.

Innehåll

1 Inledning **4**

2 Teori **6**
2.1 Generaliserad linjär modell 6
2.1.1 Slumpmässig komponent 6
2.1.2 Systematisk komponent 6
2.1.3 Länkfunktion 6
2.1.4 GLM för Binär och Binomial responsvariabel 7
2.2 Oddskvot 7
2.2.1 Konfidensintervall för oddskvot 8
2.3 Logistisk regression 8
2.3.1 Multipel logistisk regression 9
2.3.2 Oddskvot i logistisk regression 9
2.4 Likelihood 9
2.5 Likelihood-ratio test, LRT 10
2.6 Konfidensintervall 10
2.7 Splines 11
2.7.1 B-splines 11
2.8 Akaike information criterion, AIC 12

3 Databeskrivning **13**
3.1 Visualisering av data 14

4 Modellering och analys **17**
4.1 Logistisk regression 18
4.2 Analys av icke-linjär tidstrend 19
4.2.1 Analys av könens inverkan 23
4.2.2 Analys av bakgrundens inverkan 27
4.2.3 Analys av samspel mellan kön och bakgrund 31
4.3 Modell för ES 40
4.4 Modell för IB 40
4.5 Modell för NV 40
4.6 Modell för SP 41
4.7 Prediktion 41

5 Slutsats och diskussion **43**

6 Appendix **46**
6.A Tabeller 46
6.B Figurer 50

Referenser **53**

1 Inledning

Bakgrund

Högskolan i Sverige har utvecklats och expanderat och antalet studenter har haft en stor ökning under de senaste decennierna. Redan från slutet av 1980-talet började man se en ökning av övergångsfrekvensen från gymnasium till högskola.

I rapporten *Kvinnor och män i högskolan. Från gymnasium till forskarutbildning*, (1996), skriver Ewa Foss:

> Övergångsfrekvensen inom tre år från gymnasiet till högskolan har för avgångskullarna 1985/86–1991/92 ökat från 20 till 34 procent.

Även könsfördelningen har haft en förändring. I studien *Förkunskaper, social bakgrund och rekrytering till högskolan*, (2018), skriver Fredrik Svensson:

> Kvinnor var fram till 1970-talet underrepresenterade bland studenterna på högskolan. Antalet registrerade studenter var höstterminen 1945 drygt 14 000 och av dessa var 22 procent kvinnor. 1977 kom kvinnorna för första gången att vara i majoritet som en konsekvens av organisatoriska förändringar i högskolan.

För examensåret 2012 hade andelen som började studera vidare efter gymnasiet ökat från år 1992 och könsfördelningen inte förändrats. I rapporten *Kvinnor och män i högskolan*, (2016), skriver Magdalena Inkinen och Ingeborg Amnéus:

> Läsåret 2011/12 slutförde 97 700 elever gymnasieskolan och av dem började 42 procent studera vid högskolan inom tre år. Övergången till högskolan var högre bland kvinnor än bland män – 48 respektive 37 procent.

Fördelningen mellan personer med svensk och utländsk bakgrund har också genomgått en förändring. I studien *Förkunskaper, social bakgrund och rekrytering till högskolan*, (2018), skriver Fredrik Svensson:

> Före födelseåren runt mitten på 80-talet hade personer med utländsk bakgrund lägre övergång till högre studier. Men baserat på de senaste uppföljningsbara kohorterna har de med utländsk bakgrund passerat de med svensk bakgrund i övergången till högskolan.

Syfte

Med bakgrund i detta vill vi i denna studie undersöka och analysera hur trender av andelen som påbörjat studera vidare inom tre år efter gymnasiet ser ut för fyra högskoleförberedande program mellan examensåren 2000 och 2011. Vi vill även undersöka trender för kvinnor och män samt för personer med svensk och utländsk bakgrund inom dessa program. Har högskoleförberedande programmen olika tidstrender? Är tidstrender olika för kvinnor jämfört med män och är andelen kvinnor som påbörjat studier högre än andelen män? Är tidstrender olika för personer med svensk bakgrund jämfört med personer med utländsk bakgrund och blir andelen med utländsk bakgrund som påbörjat studier högre än andelen med svensk bakgrund någon gång mellan åren 2000 och 2011?

För att kunna besvara dessa frågor kommer analys och modellering göras där vi först undersöker hur tidstrender ser ut för programmen var för sig. Därefter undersöker vi hur de ser ut för kvinnor och män samt för personer med svensk och utländsk bakgrund var för sig inom programmen. Efter det kommer vi undersöka hur de ser ut för kvinnor respektive män med svensk respektive utländsk bakgrund inom programmen, vi kommer med andra ord undersöka om samspel mellan könen och bakgrunderna finns inom varje program. Till sist kommer en prediktion för examensåret 2012 göras.

2 Teori

2.1 Generaliserad linjär modell

För en Linjär regression måste antagandet om en normalfördelad responsvariabel vara uppfyllt. I det fallet då detta antagandet inte är uppfyllt kan istället en Generaliserad linjär modell, GLM, användas som är en generalisering av den Linjära modellen.

Den Generaliserade linjära modellen består av tre komponenter,
Agresti (2013, kapitel 4.1.1):

- En slumpmässig komponent
- En systematisk komponent
- En länkfunktion

2.1.1 Slumpmässig komponent

En responsvariabel Y med oberoende observationer $(y_1, y_2, ..., y_n)$ från en fördelning som tillhör en naturlig exponentialfamilj är den slumpmässiga komponenten av en GLM. Den naturliga exponentialfamiljen har täthetsfunktionen

$$f(y_i; \theta_i) = a(\theta_i)b(y_i)\exp\{y_i Q_i(\theta_i)\},\ i = 1, 2, ..., n \tag{1}$$

för någon funktion a, b och Q där $Q(\theta_i)$ kallas för den naturliga parametern.

2.1.2 Systematisk komponent

Om vi låter x_{ij}, $i = 1, 2, ..., n$, $j = 0, 1, 2...$ vara värden för observarioner i av förklaringsvariablerna j så har vi att den systematiska komponenten $\eta = (\eta_1, \eta_2, ..., \eta_n)$ av en GLM beskriver sambandet mellan förklaringsvariablerna och den slumpmässiga komponenten genom den linjära modellen

$$\eta_i = \alpha + \sum_j \beta_j x_{ij},\ i = 1, 2, ..., n;\ j = 1, 2, ... \tag{2}$$

där $\alpha = \beta_0$ är interceptet då $x_{i0} = 1$ för alla i. Denna linjärkombination kallas för linjär prediktor.

2.1.3 Länkfunktion

Länkfunktionen av en GLM kopplar ihop den slumpmässiga komponenten med den systematiska komponenten. Den kopplar ihop $\mu_i = \mathbb{E}[Y_i]$, $i = 1, 2, ..., n$, med η_i genom $\eta_i = g(\mu_i)$ där länkfunktionen g är monoton och differentierbar. Således har vi att g kopplar μ_i till förklaringsvariablerna genom

$$g(\mu_i) = \sum_j \beta_j x_{ij},\ i = 1, 2, ..., n;\ j = 0, 1, 2, ...\ . \tag{3}$$

Länkfunktionen $g(\mu_i) = Q(\theta_i)$ kallas för kanonisk-länk.

2.1.4 GLM för Binär och Binomial responsvariabel

I många fall där man använder sig av en GLM är responsvariabeln binär, den är då Bernoullifördelad med $\pi = \mathbb{P}(Y = 1) = 1 - \mathbb{P}(Y = 0)$ och $\mathbb{E}[Y] = \pi$. Sannolikhetsfunktionen för en Bernoullifördelad responsvariabel kan då skrivas som

$$p(y;\pi) = \pi^y(1-\pi)^{1-y} = (1-\pi)\{\pi/(1-\pi)\}^y = (1-\pi)\exp\left\{y\log\left(\frac{\pi}{1-\pi}\right)\right\}$$

där vi kan indentifiera den med täthetsfunktionen för den naturliga exponentialfamiljen (1) genom $\theta = \pi$, $a(\pi) = 1 - \pi$, $b(y) = 1$ och $Q(\pi) = \log\{\pi/(1-\pi)\}$. Den naturliga parametern $Q(\pi)$ av denna form heter logit av π (logaritmen av oddset).

I det fallet där responsvariabeln är summan av n oberoende och Bernoullifördelade variabler Y_i, $i = 1, ..., n$, så är den summan Binomialfördelad med parametrarna n och π, där $\mathbb{E}[Y] = n\pi$. Sannolikhetsfunktionen kan då skrivas om som

$$\begin{aligned} p(y;\pi) &= \binom{n}{y}\pi^y(1-\pi)^{n-y} = \binom{n}{y}\exp\{y\log(\pi) + (n-y)\log(1-\pi)\} = \\ &= \binom{n}{y}\exp\{y\log(\pi) - y\log(1-\pi)\}\exp\{n\log(1-\pi)\} = \\ &= \binom{n}{y}\exp\left\{y\log\left(\frac{\pi}{1-\pi}\right)\right\}\exp\{n\log(1-\pi)\} \end{aligned}$$

där vi ser att $a(\pi) = \exp\{n\log(1-\pi)\}$, $b(y) = \binom{n}{y}$ och $Q(\pi) = \log\{\pi/(1-\pi)\}$. Den naturliga parametern $Q(\pi)$ är även här $\text{logit}(\pi)$.

I det fallet där $Q(\theta) = \text{logit}(\theta) = \log\{\theta/(1-\theta)\}$ har vi att GLM är en *logistisk regressionmodell*, Agresti (2013, kapitel 4.1.2).

2.2 Oddskvot

För $\pi = \pi(x) = \mathbb{P}(Y = 1|X = x)$, där $0 < \pi < 1$, har vi att oddset är definierat som

$$\Omega = \pi/(1-\pi) \tag{4}$$

där Ω enligt definitionen aldrig är negativt. För $\Omega > 1$ är $Y = 1$ mer troligt än $Y = 0$ och för $\Omega < 1$ är $Y = 0$ mer troligt än $Y = 1$.

Om vi nu antar att vi har en 2×2-tabell, med två rader och två kolumner, så har vi att oddset i rad 1 är $\Omega_1 = \pi_1/(1-\pi_1)$ oddset i rad 2 är $\Omega_2 = \pi_2/(1-\pi_2)$. Kvoten mellan Ω_1 och Ω_2

$$\theta = \frac{\Omega_1}{\Omega_2} = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} \tag{5}$$

kallas för oddskvoten, Agresti (2013, kapitel 2.2.3). Om $\theta > 1$ kan det tolkas som att händelsen $Y = 1$ är mer trolig i rad 1 än i rad 2. Om $\theta < 1$ gäller det motsatta. För $\theta = 1$ är händelsen $Y = 1$ lika trolig i båda raderna, således är raderna och kolumnerna oberoende, Agresti (2013, kapitel 2.2.4).

2.2.1 Konfidensintervall för oddskvot

Vi antar vidare att vi har en 2×2-tabell med frekvenserna $n_{11}, n_{12}, n_{21}, n_{22}$, där n_{ij} är observerade värden i rad i kolumn j. Variansen för skattningen av $\log(\theta)$ ges av

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

Eftersom $\log(\theta)$ är approximativt normalfördelad för stora observationer så är ett $(1-\alpha) \cdot 100\%$ Wald-konfidensintervall för $\log(\theta)$ givet av

$$KI_{\log(\theta)} = \log(\hat{\theta}) \pm z_{\alpha/2} \sqrt{\widehat{\text{var}}(\log(\hat{\theta}))} = (I_l, I_u) \tag{6}$$

där $z_{\alpha/2}$ är kvantil för normalfördelning. Genom att transformera får vi konfidensintervall för θ enligt

$$KI_\theta = (\exp\{I_l\}, \exp\{I_u\}), \tag{7}$$

Agresti (2013, kapitel 3.1.1).

2.3 Logistisk regression

När vi har en binär responsvariabel Y med utfallen $Y = 0$ och $Y = 1$ kan logistisk regression vara lämplig att använda för modellering. Sannolikheten att $Y = 1$ givet en förklaringsvariabel X är formulerad som

$$\pi(x) = \mathbb{P}(Y = 1|X = x) = 1 - \mathbb{P}(Y = 0|X = x).$$

För en logistisk regressionsmodell är länkfunktionen den naturliga parametern $Q(\pi(x)) = \text{logit}(\pi(x))$ vilken skrivs som

$$\text{logit}(\pi(x)) = \log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \alpha + \beta x$$

Den logistiska regressionmodellen är då

$$\pi(x) = \frac{\exp\{\alpha + \beta x\}}{1 + \exp\{\alpha + \beta x\}}$$

där $\pi(x)$ minskar om $\beta < 0$, ökar om $\beta > 0$ och där Y är oberoende av X om $\beta = 0$, Agresti (2013, kapitel 5.1).

För modeller med kategoriska förklaringsvariabler används ofta indikatorvariabler för de olika kategorierna. För en förklaringsvariabel med K kategorier låter vi $x_i = 1$ för observarioner i kategorin i och $x_i = 0$ annars, $i = 2, ..., K$. Modellen ser då ut som

$$\text{logit}(\pi_i) = \alpha + \beta_1 x_2 + ... + \beta_{K-1} x_K$$

Agresti (2013, kapitel 5.3.2).

2.3.1 Multipel logistisk regression

I fallet där det är $p > 1$ förklaringsvariabler, $\mathbf{x} = (x_1, ..., x_p)$, gäller det att

$$\text{logit}(\pi(\mathbf{x})) = \alpha + \beta_1 x_1 + ... + \beta_p x_p$$

och den logistiska regressionsmodellen är då

$$\pi(\mathbf{x}) = \frac{\exp\{\alpha + \beta_1 x_1 + ... + \beta_p x_p\}}{1 + \exp\{\alpha + \beta_1 x_1 + ... + \beta_p x_p\}}$$

där β_i är effekten av x_i i log-odds att $Y = 1$ när de övriga x_j, $j \neq i$, är fixa.

För kategoriska förklaringsvariabler används indikatorvariabler för kategorierna, som i enkel logistisk regression, Agresti (2013, kapitel 5.4).

2.3.2 Oddskvot i logistisk regression

Vi antar att vi ha en multipel logistisk regressionmodell med $\text{logit}(\pi(\mathbf{x})) = \alpha + \beta_1 x_1 + ... + \beta_p x_p$ där förklaringsvariabeln x_1 är kategorisk. Om vi håller fast värden för resten av förklaringsvariablerna får vi att logaritmen av oddskvoten mellan oddset för andra kategorin av x_1 och oddset för första kategorin av x_1, där vi låter $x_1 = 0$ för första kategin och $x_1 = 1$ för andra kategorin, är

$$\log(OR) = \log\left(\frac{\pi(\mathbf{x})_{x_1=1}/(1-\pi(\mathbf{x})_{x_1=1})}{\pi(\mathbf{x})_{x_1=0}/(1-\pi(\mathbf{x})_{x_1=0})}\right) = \text{logit}(\pi(\mathbf{x})_{x_1=1}) - \text{logit}(\pi(\mathbf{x})_{x_1=0}) =$$

$$= \alpha + \beta_1(1) + ... + \beta_p x_p - (\alpha + \beta_1(0) + ... + \beta_p x_p) = \beta_1.$$

Oddskvoten är då given som $OR = \exp\{\beta_1\}$, Agresti (2013, kapitel 5.4.1).

2.4 Likelihood

Den Generaliserade linjära modellen använder sig av maximum likelihood-metoden för att skatta parametrarna $\hat{\beta}_j$ i den systematiska komponenten (2) i modellen.

För logistisk regression med observerade data $\mathbf{y} = (y_1, ...y_n)$ från $\mathbf{Y} = (Y_1, ..., Y_n)$ där alla Y_i, $i = 1, ..., n$ är oberoende och likafördelade med sannolikhetsfunktion $f(y_i|\pi(\mathbf{x}))$ har vi att sannolikhetsfunktionen beror på $\boldsymbol{\beta}$ givet $\mathbf{x}$ på grund av länkfunktionen, så $f(y_i|\pi(\mathbf{x}))$ kan skrivas som $f(y_i|\boldsymbol{\beta})$ och likelihood fuktionen är då definierad som

$$L(\boldsymbol{\beta}) = f(\mathbf{y}|\boldsymbol{\beta}) = \prod_{i=1}^{n} f(y_i|\boldsymbol{\beta}) \tag{8}$$

och log-likelihood funktionen är given som

$$l(\boldsymbol{\beta}) = \log(f(\mathbf{y}|\boldsymbol{\beta})) = \sum_{i=1}^{n} \log(f(y_i|\boldsymbol{\beta})). \tag{9}$$

Maximum likelihood skattningen $\hat{\boldsymbol{\beta}}_{ML}$ av $\boldsymbol{\beta}$ fås genom att maximera likelihood eller log-likelihood funktionen. För att beräkna maximum likelihood skattningen

$\hat{\boldsymbol{\beta}}_{ML}$ kan man använda sig av Score funktionen som är första derivatan av log-likelihood funktionen och har formen

$$S(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\beta}} l_i(\boldsymbol{\beta}) = \sum_{i=1}^{n} S_i(\boldsymbol{\beta})$$

Genom att lösa Score ekvationerna $S(\boldsymbol{\beta}) = 0$ får man ut skattningen $\hat{\boldsymbol{\beta}}_{ML}$.

Den observerade informations matrisen är

$$H(\boldsymbol{\beta}) = -\frac{\partial^2}{\partial \boldsymbol{\beta}_i \partial \boldsymbol{\beta}_j} l(\boldsymbol{\beta}).$$

Den observerade och förväntade informationsmatrisen $F(\boldsymbol{\beta}) = \mathbb{E}(H(\boldsymbol{\beta}))$ är samma i den logistiska regressionmodellen, Fahrmeir, Kneib, Lang & Marx (2013, kapitel 5.1.2).

2.5 Likelihood-ratio test, LRT

Likelihood-ratio test används i hypotestes där man vill testa ifall en mindre modell håller givet att en större modell håller. Vi låter M_0 vara den mindre modellen och M_1 vara den större modellen.

Likelihood-ratio test statistikan är då given av

$$G^2(M_0|M_1) = -2 \log \left[\frac{L_0}{L_1} \right] = -2(l_0 - l_1) \tag{10}$$

där l_0 och l_1 är maximerade log-likelihoodfunktioner för modell M_0 respektive modell M_1, Agresti (2013, kapitel 5.4.3).

Likelihood-ratio test statistikan är under nollhypotesen asymptotiskt χ^2-fördelad med *df* som är lika med skillnaden i antalet parametrar mellan M_1 och M_0, Agresti (2013, kapitel 1.3.3).

2.6 Konfidensintervall

För en stor mängd n av observationer har ML-skattningen $\hat{\boldsymbol{\beta}}_{ML}$ av $\boldsymbol{\beta}$ en approximativ nomlafördelning

$$\hat{\boldsymbol{\beta}}_{ML} \sim \mathrm{N}(\boldsymbol{\beta}, F^{-1}(\hat{\boldsymbol{\beta}}_{ML}))$$

med den skattade kovariansmatrisen av $\hat{\boldsymbol{\beta}}_{ML}$ som är

$$\widehat{cov}(\hat{\boldsymbol{\beta}}_{ML}) = F^{-1}(\hat{\boldsymbol{\beta}}_{ML})$$

Diagonala elementen i kovariansmatrisen är skattningarna av varianserna för $\hat{\beta}_j$ i $\hat{\boldsymbol{\beta}}_{ML}$, Fahrmeir et al. (2013, kapitel 5.1.2).

Ett $(1-\alpha) \cdot 100\%$ Wald-konfidensintervall för $\hat{\beta}_j$ ges av

$$\hat{\beta}_j \pm z_{\alpha/2} \sqrt{\left[\widehat{cov}(\hat{\boldsymbol{\beta}}_{ML}) \right]_{jj}}$$

där $z_{\alpha/2}$ är kvantil för normalfördelningen.

För $\text{logit}(\hat{\pi}(\mathbf{x})) = \mathbf{x}^\top \hat{\boldsymbol{\beta}}_{ML}$ har vi att den skattade variansen är $\mathbf{x}^\top \widehat{cov}(\hat{\boldsymbol{\beta}}_{ML})\mathbf{x}$, så ett $(1-\alpha)\cdot 100\%$ Wald-konfidensintervall för $\text{logit}(\pi(\mathbf{x}))$ är

$$\text{logit}(\hat{\pi}(\mathbf{x})) \pm z_{\alpha/2}\sqrt{\mathbf{x}^\top \widehat{cov}(\hat{\boldsymbol{\beta}}_{ML})\mathbf{x}}$$

och genom att transformera med transformen $\pi = \exp\{\text{logit}\}/(1+\exp\{\text{logit}\})$ får vi konfidensintervallet för $\pi(\mathbf{x})$, Agresti (2013, kapitel 5.5.3).

2.7 Splines

Ibland när datapunkterna som man vill anpassa en Generaliserad linjär modell till inte har något bra linjärt samband, utan pendlar en del upp och ner, så kan man tänka sig att ett polynom av högre grad kan anpassa datapunkterna bättre. Ett sätt att göra detta på är att utnyttja splines som sätter ihop flera polynom till ett enda.

Definitionen av en spline är följande, Fahrmeir et al. (2013, kapitel 8.1.1):

> En funktion $f : [a,b] \to \mathbb{R}$ kallas för polynomspline av grad $p \geq 0$ med knutar $a = \kappa_1 < ... < \kappa_m = b$, om den uppföljer följande villkor:
>
> - $f(x)$ är $(p-1)$-gånger kontinuerligt deriverbar. Då $p = 1$ är $f(x)$ kontinuerlig men behöver ej vara deriverbar.
> - $f(x)$ är ett polynom av grad p på intervallen $[\kappa_j, \kappa_{j+1})$, $j = 1, ..., m-1$ som är definierade av knutarna.

Utifrån detta kan vi nu skriva en spline på följande sätt

$$f(x) = \alpha_0 P_0(x)\mathbb{1}\{\kappa_1 \leq x \leq \kappa_2\} + ... + \alpha_{m-1}P_{m-1}(x)\mathbb{1}\{\kappa_{m-1} \leq x \leq \kappa_m\} \quad (11)$$

där α_{j-1} är konstanter, $P_{j-1}(x)$ är polynom som kallas för basfunktioner och κ_j är knutar, $j = 1, ..., m-1$, Maindonald and Braun (2010, kapitel 7.5).

Vi ser att en spline sätter ihop m stycken basfunktioner, som är polynom, med en indikatorvariabel som antar värdet 1 i det intervallet där vi vill ha en viss funktion och 0 annars.

2.7.1 B-splines

En annan typ av spline är B-splines som är en linjärkombination av kontrollpunkter och basfunktioner. Vi har $N+2$ reella värden κ_i, $i = 0, ..., N+1$ som vi kallar för knutar, där $N \geq 2$ knutar kallas för inre knutar och 2 knutar κ_0 och κ_{N+1} kallas för yttre knutar, som uppfyller

$$\kappa_0 < \kappa_1 < ... < \kappa_{N+1}.$$

En B-splinefunktion av ordning p är definierad som

$$S(x) = \sum_{i=0}^{N+p} \beta_i B_{i,p}(x) \quad (12)$$

där β_i är kontrollpunkter (koefficienter) och $B_{i,p}(x)$ är basfunktioner av grad p.

Basfunktionerna $B_{i,p}(x)$ är definierade rekursivt som

$$B_{i,0}(x) = \begin{cases} 1 & \text{om } \kappa_i \leq x < \kappa_{i+1} \\ 0 & \text{annars.} \end{cases}$$

$$B_{i,j}(x) = \frac{x - \kappa_i}{\kappa_{i+p} - \kappa_i} B_{i,p-1}(x) + \frac{\kappa_{i+1+p} - x}{\kappa_{i+1+p} - \kappa_{i+1}} B_{i+1,p-1}(x)$$

Från definitionen av en B-splinefunktion ser vi att vi totalt kommer ha $K = N + p + 1$ basfunktioner i en B-splinefunktion av graden p med N stycken inre knutar. Basfunktionen $B_{0,p}(x)$ är oftast benämnd som "intercept",
Racine (2018, kapitel 3).

B-splines i R

En inbygd funktion för att anpassa B-splines finns i **R**. Vi kommer använda oss av standarden i **R**, där B-splines av grad 3 anpassas och knutar sätts ut vid kvantilerna som är jämnt fördelade, och bestäms av antalet frihetsgrader man väljer att ha. Antalet inre knutar bestäms av *frihetsgrader - polynomets grad*, så om vi vill ha 3 inre knutar så väljer vi att ha 6 *frihetsgrader*. Har vi 3 inre knutar så kommer de sättas ut på 25%, 50% och 75% kvantilerna. Totala antalet knutar är 5 om man har 3 inre knutar eftersom det alltid finns 2 yttre knutar, Maindonald et al. (2010, kapitel 7.5)

Standarden i **R**'s inbygda anpassning är att inte ha något intercept i B-splines, alltså att basfunktionen $B_{0,p}(x)$ inte är med, Racine (2018, kapitel 3). I detta fall kommer vi få $K = N + p$ basfunktioner och B-splinefuktionen kommer att se ut som $S(x) = \sum_{i=1}^{N+p} \beta_i B_{i,p}(x)$.

2.8 Akaike information criterion, AIC

Akaike information criterion (AIC), är ett värde på modellanpassning som används för att jämföra olika modeller som är baserade på samma data. Värt att notera är att AIC-värdet inte säger någonting om hur bra en enskiljd modell är anspassad, utan det säger hur bra en modell är i jämförelse med en annan. Definitionen av AIC är följande

$$\text{AIC} = -2(\log(L) - p) = -2\log(L) + 2p \tag{13}$$

där L är maximum värdet av likelihood funktionen för modellen och p är antalet parametrar i modellen. När vi jämför modeller med varandra för att välja den med bäst anpassning kommer vi att välja modellen som har lägst AIC-värde, Agresti (2013, kapitel 6.1.6).

3 Databeskrivning

I denna studie används data mellan åren 2000 och 2011 som innehåller information om antalet personer som påbörjat studier på högskola/universitet inom tre år efter avslutad gymnasial utbildning samt totala antalet personen som avslutat gymnasial utbildning på våren ett givet år. Dessa data är grupperade eftersom de innehåller information om antalet personer som är en summa. Data är hämtat från Skolverkets officiella statistik, för varje år mellan 2000 och 2011.

Antalet personer är vidare uppdelat på program, inom vilka personer tagit examen från gymnasiet, kön och personers bakgrund. Programmen som är med i denna studie är Estetiska programmet, International Baccalaureate, Naturvetenskapsprogrammet och Samhällsvetenskapsprogrammet. Dessa är högskoleförberedande program, alla personer har därmed behörighet till att söka vidare studier på högskola/universitet. Personers bakgrund är antingen svensk eller utländsk, där personer som är födda utomlands eller födda i Sverige med båda föräldrarna födda utomlands har utländsk bakgrund.

Tabell 1 nedan sammanfattar variablerna som är med i denna studie samt beteckningar som kommer användas för dessa.

Tabell 1: Variabler och deras beteckningar.

Variabel	Beteckning
Totalt	Antalet personer som avslutat gymnasiet inom ett program, N.
Påbörjat	Antalet personer som påbörjat studier, $Z = \sum_{i=1}^{N} Y_i$. $Y_i = 1$: påbörjat studier $Y_i = 0$: inte påbörjat studier
År	X: $x_i \in (2000, ..., 2011)$, $i = 1, ..., 12$.
Program	ES = Estetiska programmet IB = International Baccalaureate NV = Naturvetenskapsprogrammet SP = Samhällsvetenskapsprogrammet
Kön	K = Kvinna M = Man
Bakgrund	Sv = Svensk Utl = Utländsk

Vidare kommer analys och resultat presenteras av andelen personer som påbörjat vidare studier inom tre år av totala antalet som avslutade gymnasiet, för varje år mellan 2000 och 2011. Om personer avlutade gymnasiet våren 2000 så tittar vi på andelen personer som påbörjat vidare studier senast hösten 2003.

Tabell 2 nedan visar hur data kan se ut och innehåller frekvenser för examensåret 2000 uppdelat på program, kön och bakgrund. Totala antalet personer som tog examen är presenterat i kolumnen *Totalt*. Andelen personer som påbörjat studier inom tre år av totala antalet personer som tog examen är presenterat i kolumnen *Andel*.

Tabell 2: Frekvenser för examensåret 2000.

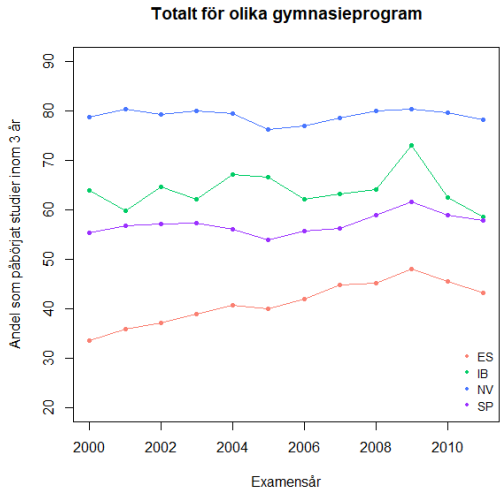
Program	Kön	Bakgrund	Påbörjat	Ej påbörjat	Totalt	Andel
ES	K	Sv	813	1426	2239	36.3
ES	K	Utl	68	153	221	30.8
ES	M	Sv	275	687	962	28.6
ES	M	Utl	23	75	98	23.5
IB	K	Sv	71	29	100	71
IB	K	Utl	13	17	30	43.3
IB	M	Sv	37	17	54	68.5
IB	M	Utl	10	11	21	47.6
NV	K	Sv	4893	1071	5964	82
NV	K	Utl	636	153	789	80.6
NV	M	Sv	6409	2001	8410	76.2
NV	M	Utl	832	238	1070	77.8
SP	K	Sv	6644	4518	11162	59.5
SP	K	Utl	899	585	1484	60.6
SP	M	Sv	2859	3174	6033	47.4
SP	M	Utl	408	437	845	48.3

3.1 Visualisering av data

I denna studie vill vi undersöka tidstrender av andelen som påbörjat högskole/universitets-studier för dem olika högskoleförberedande programmen, för män och kvinnor i dessa program samt för svensk och utländsk bakgrund i dessa program, mellan åren 2000 och 2011. Vi vill även undersöka om kvinnor samt män med olika bakgrund har olika tidstrender, det vill säga om samspel mellan kön, bakgrund och examensår finns inom dessa program, och försöka bygga upp en övergripande modell för tidstrenderna.

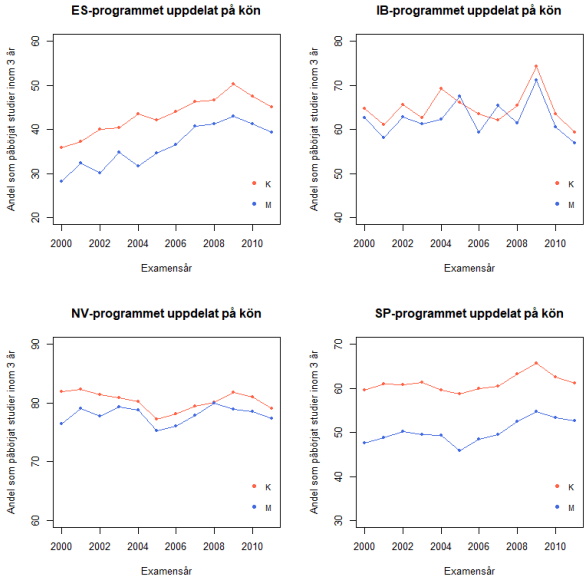
Innan en mer djupgående analys påbörjas kan det vara intressant att visualisera och undersöka det data vi har för att få en uppfattning om hur det är fördelat under åren 2000 till 2011 inom de olika programmen, samt för att se om några samspel kan tydas.

Figur 1 nedan visar andelen personer som påbörjat studier inom tre år för dem olika programmen. I figuren kan vi se att andelen inom ES ökar nästan varje år mellan åren 2000 och 2009, och ökningen ser ut att vara störst för detta program mellan dessa år jämfört med dem andra programmen. Andelen pendlar ganska mycket upp och ner mellan nästan alla år inom IB. Totalt sett har det skett en ökning av andelen mellan åren 2000 och 2009 inom alla program förutom NV och inom varje program sker en sänkning från år 2009, där denna sänkning är störst inom IB.



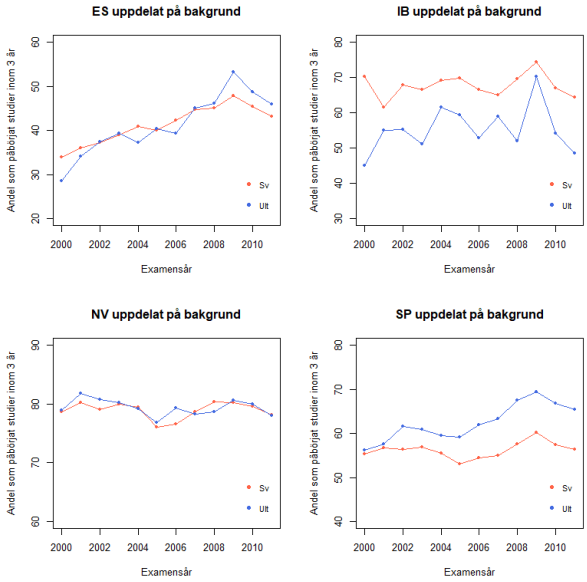
Figur 1: Andelen personer som påbörjat studier inom tre år efter avslutad gymnasial utbildning, för varje program.

I Figur 2 nedan visas andelen kvinnor respektive män i varje program som påbörjat studier inom tre år. Vi kan se direkt att andelen kvinnor som påbörjat studier är högre än andelen män inom alla program, dock ser skillnaden mellan könen ut att vara liten inom IB. Andelen kvinnor respektive män verkar ha olika förändringar med åren.



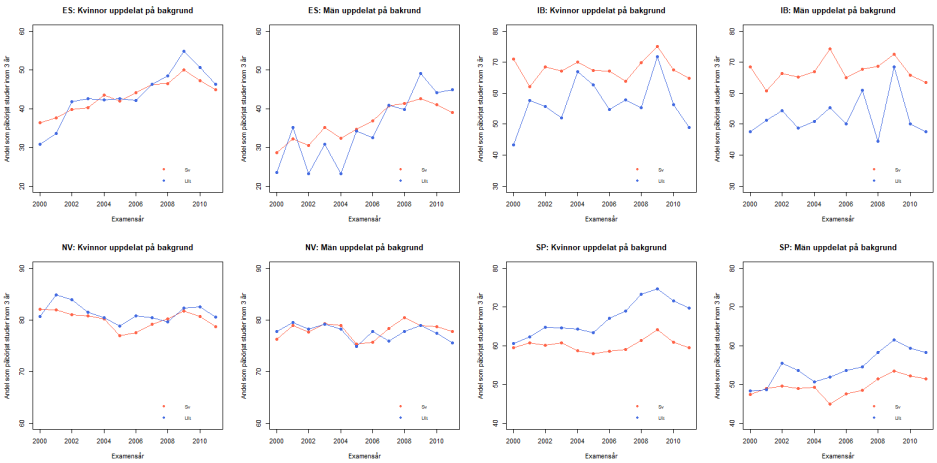
Figur 2: Andelen kvinnor respektive män som påbörjat studier inom tre år efter avslutad gymnasial utbildning.

I Figur 3 nedan visas andelen med svensk och utländsk bakgrund som påbörjat studier, inom varje program. Vi ser att andelen med utländsk bakgrund ökar mer med åren än med svensk bakgrund inom ES. Inom IB ser vi att andelen som påbörjat studier med svensk bakgrund är högre än andelen med utländsk bakgrund för alla år. Inom NV ser skillnaden ut att vara väldigt liten mellan bakgrunderna och förändringen med åren ser ut att vara rätt lika mellan bakgrunderna. Inom SP ser vi att andelen med utländsk bakgrund är högre än med svensk bakgrund för alla år och andelen med utländsk bakgrund ökar mer med åren.



Figur 3: Andelen med svensk respektive utländsk bakgrund som påbörjat studier inom tre år efter avslutad gymnasial utbildning.

Figur 4 nedan visar andelen kvinnor med svensk respektive utländsk bakgrund samt män med svensk respektive utländsk bakgrund, inom varje program. Inom ES ser vi att andelen kvinnor med utländsk bakgrund som påbörjat studier är högre än med svensk bakgrund för större delen av åren och andelen män med svensk bakgrund är högre än med utländsk bakgrund för större delen av åren. Inom NV ser vi ingen större skillnad mellan bakgrunderna för kvinnor samt män. Inom SP ser vi att andelen kvinnor med utländsk bakgrund ökar med åren varpå andelen kvinnor med svensk bakgrund verkar minska med åren, och andelen män med utländsk bakgrund ser ut att öka mer med åren än andelen män med svensk bakgrund.



Figur 4: Andelen kvinnor och män med svensk respektive utländsk bakrund som påbörjat studier inom tre år efter avslutad gymnasial utbildning.

4 Modellering och analys

Vi vill undersöka tidstrender av andelen som påbörjat högskole/universitetsstudier för olika gymnasieprogram (ES, IB, NV och SP), för män och kvinnor samt för personer med svensk och utländsk bakgrund, inom dessa program. För att kunna undersöka och analysera detta vill vi hitta en lämplig modell som kan tänkas beskriva vårt data på ett bra sätt.

Vi kommer till att börja med att anpassa en enkel logistisk regressionsmodell för varje program separat. Sedan kommer vi använda B-splines för att undersöka och modellera tidstrender mellan åren 2000 och 2011 för varje program separat. Detta gör vi för att se hur tidstrenden ser ut för varje program i helhet, och jämföra tidstrender mellan programmen.

Vi kommer därefter undersöka effekten av kön på tidstrenden inom varje program och se om kvinnor respektive män har olika tidstrender. Om könen har olika tidstrender kommer vi anpassa B-splines för könen separat. Detta gör vi för att kunna besvara frågorna om trenden är olika för kvinnor jämfört med män samt om andelen kvinnor som påbörjat vidare studier är högre än andelen män och om detta gäller för alla år mellan 2000 och 2011.

Därefter kommer vi utföra samma procedur för personers bakgrund som för könen. Detta för att kunna besvara frågorna om trenden är olika för personer med svensk bakgrund jämfört med utländsk bakgrund samt om andelen personer med svensk bakgrund som påbörjat vidare studier är högre än andelen med utländsk bakgrund eller om det mellan åren 2000 och 2011 sker en förändring där andelen med utländsk bakgrund blir högre än med svensk bakgrund.

Efter det kommer vi att undersöka om kvinnor och män har olika tidstren-

der beroende på om de har svensk eller utländsk bakgrund samt hur andelarna förhåller sig till varandra inom varje program mellan åren 2000 och 2011. Detta gör vi för att kunna få en större överblick och en helhetsbild av hur tidstrender inom varje program ser ut för könen och bakgrunderna beroende på varandra.

4.1 Logistisk regression

Eftersom en person som avslutat en gymnasial utbildnig antingen påbörjat vidare studier eller inte så har vi att $Y = 1$ om personen påbörjar studier inom tre år och $Y = 0$ om personen inte påbörjat studier inom tre år. Y är således Bernoullifördelad med sannolikheten π att en person påbörjat studier, $Y \sim Be(\pi)$.

Vårt data innehåller information om summan av Bernoullifördelade variabler Y_i där vi antar att variablerna är oberoende och likafördelade, då personer påbörjar respektive påbörjar ej studera vidare oberoende av varandra och med lika sannolikhet. Denna summa $Z = \sum_{i=1}^{n} Y_i$ är Binomialfördelad med parametrarna n och π, $Z \sim Bin(n, \pi)$, där n är antalet personer som avslutade gymnasial utbildning ett visst år.

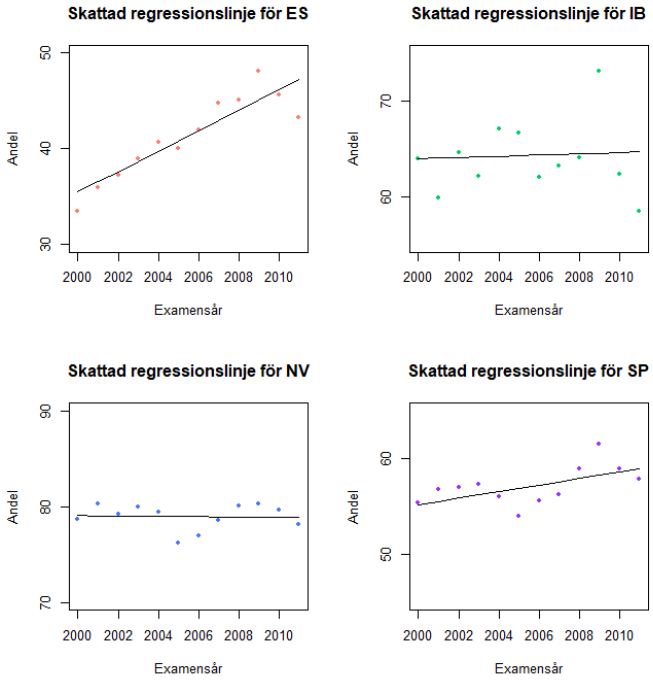
En Generaliserad linjär modell med Binomialfördelad responsvariabel och länkfunktion logit är en *logistisk regressionsmodell*, (2.1.4). Vår logistiska regressionsmodell för gymnasieprogrammen var för sig kan fomuleras som

$$\text{logit}(\pi(x_i)) = \alpha + \beta x_i, \ i = 1, 2, ..., 12$$

där x_i är examensår för åren 2000 till 2011, och $\pi(x_i) = P(Y = 1|X = x_i)$.

Eftersom vi i denna analys vill undersöka tidstrenden för dem olika programmen kommer vi till att börja med anpassa en enkel logistisk regressionmodell med examensår som förklarande variabel för varje program var för sig och plotta linjerna för att se om de passar datapunkterna väl. AIC-värden för modellerna kan ses i Appendix 6.A Tabell A.1.

I Figur 5 ser vi de skattade linjerna samt datapunkterna för varje program. Man kan se att linjerna inte passar datapunkterna väl, framförallt så ser man att linjerna inte anpsassar sänkningen från år 2009, för alla program. Punkterna för programmen har inte riktigt ett positivt linjärt samband mellan alla åren, de pendlar en del upp och ner. Man skulle därför behöva beskriva förändringen med tiden av andelarna med ett polynom av högre grad. Ett sätt att göra detta på är att anpassa B-splines som kan kombinera flera polynom av högre grad, vilket kan vara användbart eftersom våra punkter pendlar ganska ojämnt.



Figur 5: Skattade logistiska regressionslinjer av andelen som påbörjat studier inom tre år, för varje program.

4.2 Analys av icke-linjär tidstrend

Vi kommer nu anpassa B-splines för programmen var för sig och undersöka om modellerna passar datapunkterna bättre. Detta görs för se hur tidstrender ser ut för varje program i helhet och för att kunna göra en jämförelse av tidstrender mellan programmen. Till att börja med vill vi analysera ifall de olika programmen har olika tidstrender, det vill säga om det föreligger samspel mellan program och splinefunktion av examensår. Vi kommer vidare benämna splinefunktionen av examensår som tidstreden.

B-splines (2.7.1) kombinerar flera polynom av högre grad vid specifika punkter, så kallade knutar. Knutarna i vår analys kommer vara på examensåren och polynomen kommer att vara av grad 3. Vi kommer vidare testa olika antal knutar för varje anpassning och genom att notera AIC-värdet för olika antalet knutar kommer den modell som har lägst AIC värde att anses som den mest välanpassade. För att inte överanpassa kommer vi bara titta på modeller med upp till 6 knutar totalt, alltså med upp till fyra inre knutar plus två yttre knutar. Dessa yttre knutar kommer sitta på årtalen 2000 och 2011 i denna studie.

Vi vill alltså undersöka ifall det föreligger samspel mellan program och tidstrenden. Om samspel påvisas innebär det att de fyra programmen har olika tidstrender, modellerna kan ha olika antal knutar i splinefunktionerna. För att kunna jämföra två modeller, en med samspelsterm och en utan samspelsterm,

behöver modellerna ha lika antal knutar. Vi väljer antalet knutar genom att notera AIC-värden för båda modellerna och välja dem modeller där AIC är lägst, helst för båda modellerna. Efter att ha valt modellerna använder vi *Likelihood-ratio test* (2.5) för att se om det finns ett signifikant samspel. Tabell 3 nedan visar AIC-värden för modellerna med olika antal knutar.

Tabell 3: AIC för modeller utan och med samspel mellan program och tidstrend.

Antal knutar	AIC, utan samspel	AIC, med samspel
2	6487.85	6235.52
3	6163.63	5922.23
4	6192.43	5948.50
5	6144.61	5905.23
6	6133.15	5895.55

Från tabellen ser vi att vi borde välja modellerna med 6 knutar, lägst AIC. Modellen med samspelsterm ser ut på följande vis

$$\text{logit}(\pi(x_i)) = \alpha + \beta_1 \mathbb{1}\{\text{IB}\}_i + \beta_2 \mathbb{1}\{\text{NV}\}_i + \beta_3 \mathbb{1}\{\text{SP}\}_i + S_1(x_i) +$$

$$+ S_2(x_i)\mathbb{1}\{\text{IB}\}_i + S_3(x_i)\mathbb{1}\{\text{NV}\}_i + S_4(x_i)\mathbb{1}\{\text{SP}\}_i, \ i = 1, 2, \ldots, 12.$$

där $\mathbb{1}\{.\}$ är indikator för programmen som antar värdet 0 och 1, S_1 är splinefunktion för ES, $S_1 + S_2$ är splinefunktion för IB osv. Splinefunktionerna har följande form

$$S_k(x_i) = \beta_{k1} B_1(x_i) + \beta_{k2} B_2(x_i) + \ldots + \beta_{km} B_m(x_i), \ k = 1, 2, 3, 4$$

där B_n är basfunktioner för $n = 1, \ldots, 7$.

Vi gör ett Likelihood-ratio test för att testa om programmen har olika tidstrender. Vår nollhypotes och alternativhypotes är

$$H_0 : \beta_{k1} = ... = \beta_{k7} = 0, \ k = 2, 3, 4$$
$$H_1 : \text{åtminstone en av } \beta_{kn} \neq 0$$

Vi väljer signifikansnivå på 5% och får p-värde $< 2.2 \cdot 10^{-16}$ för teststatistikan $G^2 = 279.6$ med $df = 21$. Vi förkastar nollhypotesen om att det inte finns något samspel mellan program och tidstrenden.

Vi vet nu att åtminstone en av S_2, S_3, S_4 inte är lika med noll. Vi fortsätter nu med att utföra samma test som ovan fast parvis mellan programmen, resultaten kan ses i Tabell 4 nedan. Alla tester har ett p-värde som är mindre än 0.05. Det finns alltså ett signifikant samspel mellan program och tidstrend och vi drar slutsatsen att programmen har olika tidstrender och kan ha olika antal knutar i sina splinefunktioner.

Tabell 4: Likelihood-ratio test av samspel mellan program och tidstrenden.

Program	G^2 (df)	p-värde
NV,SP	43.99 (7)	$2.14 \cdot 10^{-7}$
IB,NV	16.95 (6)	0.00946
IB,SP	21.42 (7)	0.003195
ES,IB	22.51 (4)	0.000159
ES,NV	251.34 (6)	$< 2.2 \cdot 10^{-16}$
ES,SP	162.82 (7)	$< 2.2 \cdot 10^{-16}$

Nu när vi har konstaterat att programmen har olika tidstrender vill vi anpassa B-splines för varje program separat. Vi testar olika antal knutar för varje program och väljer den modell för vilken AIC-värdet är lägst. Tabell 5 nedan sammanfattar AIC-värden för olika antal knutar för varje program.

Tabell 5: AIC-värden för modeller med B-spline.

ES		IB	
Antal knutar	AIC	Antal knutar	AIC
2	614.10	2	377.61
3	598.44	3	375.77
4	601.53	4	368.09
5	601.63	5	372.50
6	603.33	6	366.67
NV		**SP**	
Antal knutar	AIC	Antal knutar	AIC
2	734.55	2	4509.26
3	634.20	3	4313.82
4	648.35	4	4330.53
5	628.81	5	4302.28
6	639.76	6	4285.78

Vi ser från tabellen att ES har bäst anpassning med 3 knutar, IB har bäst anpassning med 6 knutar, NV har bäst anpassning med 5 knutar och SP har bäst anpassning med 6 knutar. Modellerna ser ut på följande vis

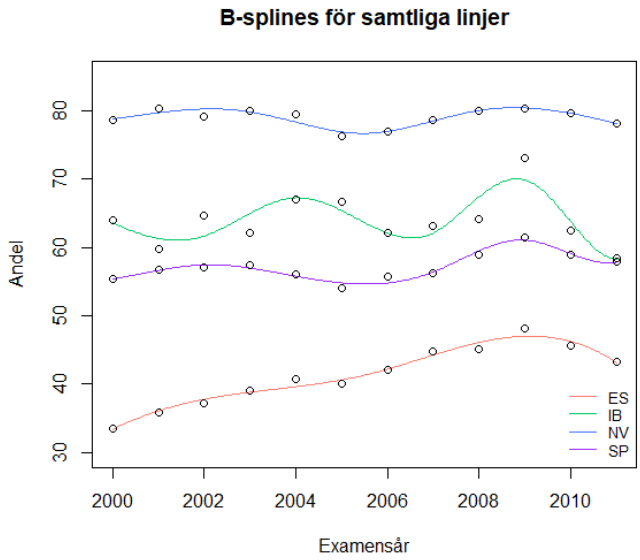
$$\text{ES: } \text{logit}(\pi(x_i)) = \alpha + S_{ES}(x_i);\ S_{ES}(x_i) = \sum_n \beta_{1n} B_n(x_i),\ n = 1, ..., 4$$

$$\text{IB: } \text{logit}(\pi(x_i)) = \alpha + S_{IB}(x_i);\ S_{IB}(x_i) = \sum_n \beta_{2n} B_n(x_i),\ n = 1, ..., 7$$

$$\text{NV: } \text{logit}(\pi(x_i)) = \alpha + S_{NV}(x_i);\ S_{NV}(x_i) = \sum_n \beta_{3n} B_n(x_i),\ n = 1, ..., 6$$

$$\text{SP: } \text{logit}(\pi(x_i)) = \alpha + S_{SP}(x_i);\ S_{SP}(x_i) = \sum_n \beta_{4n} B_n(x_i),\ n = 1, ..., 7$$

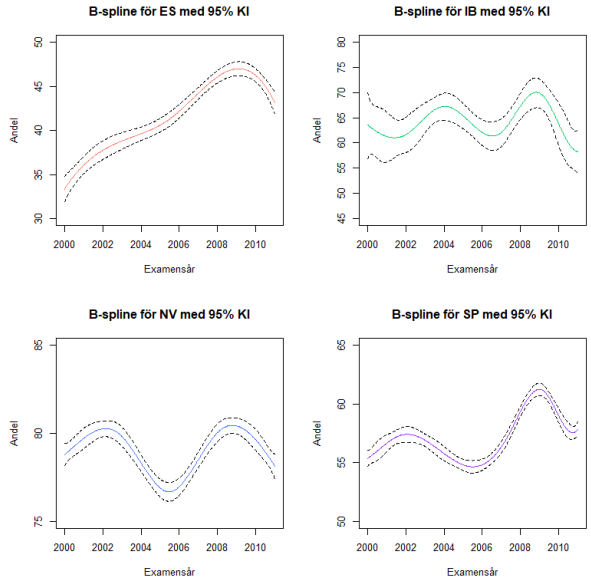
I Figur 6 nedan ser vi den anpassade tidstrenden för varje program. Linjerna ser ut att följa punkterna väl, mindre väl för IB då punkterna pendlar ganska mycket upp och ner mellan varje år. Vi ser att varje tidstrend anpassar den negativa lutningen från år 2009 väl, jämfört med dem logistiska regressionslinjerna från avsnitt 4.1. Ökningen med tiden mellan år 2000 och år 2009 är mest tydlig för ES där andelen ökar varje år.



Figur 6: Skattade tidstrender för andelen personer som påbörjat studier inom tre år, för varje program.

Vi vill vidare undersöka hur säkra våra skattningar är, och detta gör vi genom att skapa ett 95%-igt konfidensintervall för varje tidstrend. Först skapar vi ett *Wald-konfidensintervall* (2.6) för $\text{logit}(\pi(x_i))$ och transformerar sedan det för att få ett konfidensintervall för $\pi(x_i)$.

I Figur 7 ser vi tidstrenderna för varje program med 95%-iga konfidensintervall. För ES, NV och SP ser konfidensintervallen ganska smala ut vilket tyder på en god säkerhet i anpassningarna. För IB är konfidensintervallet bredare jämfört med de andra programmen samt mer ostabilt mellan åren 2000 och 2003, vilket tyder på mer osäkerhet i hur den anpassade tidstrenden förklarar andelen som påbörjat studier.



Figur 7: Skattade tidstrender med 95% konfidensintervall för andelen personer som påbörjat studier inom tre år.

4.2.1 Analys av könens inverkan

Vidare vill vi undersöka hur tidstrender för dem olika programmen ser ut beroende på kön, för att se om tidstrender är olika för kvinnor och män samt om andelen kvinnor som påbörjat studier är högre jämfört med andelen män och om så är fallet för alla år mellan 2000 och 2011.

För att göra detta vill vi först undersöka om det finns signifikant samspel mellan kön och tidstrenden inom varje program. Detta gör vi genom att jämföra två modeller, en utan samspelsterm mellan kön och tidstrenden och en med samspelsterm. Som tidigare i avsnitt 4.2 vill vi välja lika antal knutar för båda modellerna, och vi väljer antalet knutar där AIC-värdet är lägst. Tabell 6 innehåller AIC-värden för programmet IB. AIC-värden för de andra programmen kan ses i Appendix 6.A Tabell A.2, Tabell A.3 och Tabell A.4.

Tabell 6: AIC för modell utan och med samspel mellan kön och tidstrenden, inom IB.

Antal knutar	AIC, utan samspel	AIC, med samspel
2	376.43	382.15
3	374.60	382.10
4	366.85	375.99
5	371.35	382.66
6	365.47	378.26

I Tabell 6 ser vi att AIC-värdet är lägst med 6 knutar för modellen utan sam-

spel och lägst med 4 knutar för modellen med samspel. Eftersom AIC ökar mer från 4 till 6 knutar för modellen med samspel (378.26-375.99=2.27) jämfört med sänkningen i AIC för modell utan samspel (366.85-365.47=1.38) så väljer vi 4 knutar för båda modellerna. Från tabellerna i Appendix 6.A ser vi att bäst anpassning för ES är med 3 knutar för båda modellerna, för NV med 5 knutar och för SP med 6.

Modellerna med samspelsterm ser ut på följande vis

$$\text{logit}(\pi(x_i)) = \alpha + \beta \mathbb{1}\{\text{M}\}_i + S_1(x_1) + S_2(x_i)\mathbb{1}\{\text{M}\}_i, \ i = 1, ..., 12$$

där $\mathbb{1}\{\text{M}\}$ är indikator som antar värde 0 för kvinnor och 1 för män, S_1 är splinefunktion för kvinnor och $S_1 + S_2$ är splinefunktion för män. Splinefunktionerna har formen

$$S_1(x_i) = \beta_{11}B_1(x_i) + \beta_{12}B_2(x_i) + \ldots + \beta_{1m}B_m(x_i)$$

$$S_2(x_i) = \beta_{21}B_1(x_i) + \beta_{22}B_2(x_i) + \ldots + \beta_{2m}B_m(x_i)$$

där B_n är basfunktioner och $n = 1, ..., 4$ för ES, $n = 1, ..., 5$ för IB, $n = 1, ..., 6$ för NV och $n = 1, ..., 7$ för SP.

Vidare gör vi ett Likelihood-ratio test för att testa om kvinnor och män har olika tidstrender. Vår nollhypotes och alternativhypotes för varje program är då

$$H_0 : \beta_{21} = ... = \beta_{2m} = 0$$
$$H_1 : \text{åtminstone en av } \beta_{2n} \neq 0$$

Om samspelsterm är signifikant väljer vi modellen med samspel. Tidstrenden för kvinnor respektive män är då olika inom varje program och splinefunktionerna för könen kan ha olika antar knutar. Vi väljer signifikansnivå på 5% och Tabell 7 sammanfattar testen för de fyra programmen.

Tabell 7: Likelihood-ratio test av samspel mellan kön och tidstrenden.

Program	G^2 (df)	p-värde
ES	4.78 (4)	0.3102
IB	0.86 (5)	0.9732
NV	37.04 (6)	$1.727 \cdot 10^{-6}$
SP	28.21 (7)	0.0002

Från tabellen ser vi att p-värde för ES och IB är större än 0.05, så nollhypotesen för dessa program förkastas ej, kvinnor och män har inte olika tidstrend. För NV och SP är p-värdet mindre än 0.05, så nollhypotesen för dessa program förkastas, kvinnor och män har olika tidstrender.

Nu när vi vet om samspelstermen är signifikant eller inte vill vi anpassa B-splines för könen separat inom NV och SP. För ES och IB vill vi däremot undersöka huvudeffekten av kön, det vill säga hur andelen kvinnor som påbörjat vidare studier förhåller sig till andelen män.

ES och IB

Modellen för ES och IB ser nu ut på följande vis

$$\text{logit}(\pi(x_i)) = \alpha + \beta \mathbb{1}\{\text{M}\}_i + S(x_i),\ i = 1, ..., 12$$

där $S = S_{ES}$ för ES och $S = S_{IB}$ för IB som är de anpassade tidstrenderna från avsnitt 4.2 s.21.

Eftersom samspel mellan kön och tidstrenden inte är signifikat för båda programmen så har vi att oddskvoten mellan oddset för att en man påbörjar studier och oddset för att en kvinna påbörjar studier är konstant mellan åren 2000 och 2011. Om vi säger att OR är oddskvoten så har vi att $OR = \exp\{\beta\}$ (2.3.2) för båda programmen. Skattningen av β är $\hat{\beta} = -0.28558$ för ES, vilket ger

$$OR_{ES} = \exp\{-0.28558\} \approx 0.75$$

och skattningen av β är $\hat{\beta} = -0.1018$ för IB, vilket ger

$$OR_{IB} = \exp\{-0.1018\} \approx 0.90$$

Ett 95%-igt konfidensintervall (2.2.1) för OR_{ES} där standardavvikelsen för $\hat{\beta}$ är $\hat{\sigma}(\hat{\beta}) = 0.01922$, blir

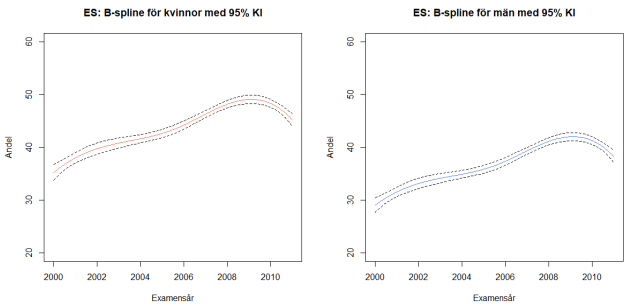
$$KI_{OR_{ES}} \approx (0.72, 0.78)$$

och ett 95%-igt konfidensintervall för OR_{IB} där standardavvikelsen för $\hat{\beta}$ är $\hat{\sigma}(\hat{\beta}) = 0.0565$, blir

$$KI_{OR_{IB}} \approx (0.809, 1.009).$$

Oddskvoten för ES är signifikant men för IB är den inte signifikant, kön har således inget signifikant huvudeffekt inom IB.

I Figur 8 ser vi den skattade tidstrenden för ES uppdelat på kvinnor och män med 95% konfidensintervall, där tidstrenden är samma för könen men med andelen kvinnor högre än män. I Appendix 6.B Figur A.1 kan vi se hur dessa tidstrender passar datapunkterna, och de ser ut att passa väl.



Figur 8: Skattad tidstrend för ES med 95% konfidensintervall uppdelat på kön.

NV och SP

Eftersom samspel mellan kön och tidstrend är signifikan inom NV och SP vill vi nu anpassa separata B-splines för kvinnor och män inom båda programmen. Vi gör detta på samma sätt som tidigare, genom att undersöka AIC- värdet för olika antal knutar och välja den modell där AIC-värdet är lägst.

Tabell A.5 och Tabell A.6 i Appendix 6.A innehåller AIC-värden av anpassning för kvinnor respektive män inom NV och SP. Modellen för kvinnor har bäst anpassning med 3 knutar (AIC=222.97) och modellen för män har bäst anpassning med 5 knutar (AIC=242.33) inom NV. Modellen för kvinnor har bäst anpassning med 6 knutar (AIC=886.07) och modellen för män har bäst anpassning med 5 knutar (AIC=413.22) inom SP. Tidstrenderna ser ut på följande sätt

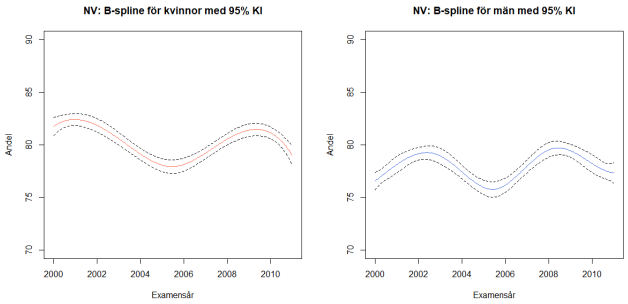
$$\text{K: logit}(\pi(x_i)) = \alpha + S_K(x_i);\ S_K(x_i) = \sum_n \beta_{Kn} B_n(x_i),$$

$$\text{M: logit}(\pi(x_i)) = \alpha + S_M(x_i);\ S_M(x_i) = \sum_n \beta_{Mn} B_n(x_i),$$

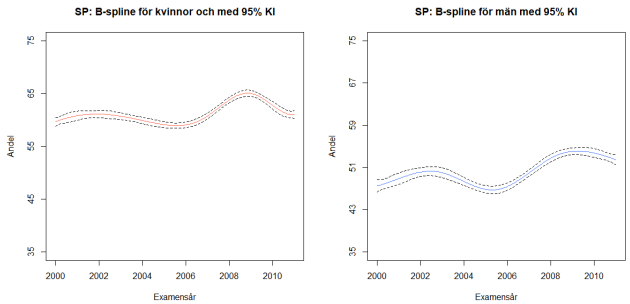
där $n = 1, ..., 4$ för kvinnor inom NV, $n = 1, ..., 7$ för kvinnor inom SP och $n = 1, ..., 6$ för män inom NV samt SP.

I Figur 9 och Figur 10 nedan ser vi dem skattade tidstrenderna för kvinnor respektive män med 95% konfidensintervall inom NV och SP. Konfidensintervallen följer linjerna väl och är ganska smala för både NV och SP, skattningarna är således ganska säkra. I Appendix 6.B Figur A.2 och Figur A.3 kan vi se hur tidstrenderna passar datapunkterna inom NV och SP.

Vi kan se att andelen kvinnor verkar sjunkit från år 2000 till år 2011, medan andelen män verkar ökat lite från år 2000 till år 2011 inom NV. Inom SP kan vi se att andelen kvinnor och män verkar ökat lite från år 2000 till år 2011. Det sker en ökning mellan åren 2006 och 2009 samt en sänkning mellan åren 2009 och 2011 inom både NV och SP. Vi kan även se att andelen kvinnor som påbörjat studier är högre än andelen män för alla år mellan 2000 och 2011 inom både NV och SP.



Figur 9: Skattade tidstrender för kvinnor och män inom NV med 95% konfidensintervall.



Figur 10: Skattade tidstrender för kvinnor och män inom SP med 95% konfidensintervall.

4.2.2 Analys av bakgrundens inverkan

Näst vill vi undersöka hur tidstrenden för dem olika programmen ser ut beroende på personers bakgrund. Är tidstrenderna olika för personer med svensk bakgrund och för personer med utländsk bakgrund inom programmen och är andelen personer med svensk bakgrund som påbörjat studier högre än andelen personer med utländsk bakgrund eller blir andelen personer med utländsk bakgrund högre än andelen med svensk bakgrund någon gång mellan åren 2000 och 2011?

För detta vill vi först undersöka om det finns signifikant samspel mellan bakgrund och tidstrenden inom varje program. Vi gör det på samma sätt som i analysen av könets inverkan i avsnitt 4.2.1. Vi jämför modellen utan samspelsterm med modellen med samspelsterm, genom att välja lika antal knutar för modellerna där AIC-värdet är lägst.

Tabell 8 visar AIC-värden för programmet IB. AIC-värden för de andra programmen finns i Appendix 6.A Tabell A.7, Tabell A.8 och Tabell A.9.

Tabell 8: AIC för modell utan och med samspel mellan bakgrund och tidstrenden, inom IB.

Antal knutar	AIC, utan samspel	AIC, med samspel
2	304.08	308.76
3	302.89	308.58
4	295.25	302.99
5	299.70	308.03
6	294.48	303.19

Från Tabell 8 ser vi att modellen utan samspel har bäst anpassning med 6 knutar och modellen med samspel har bäst anpassning med 4 knutar. Vi kan se att sänkningen i AIC-värde från 4 till 6 knutar för modellen utan samspel är större än ökningen i AIC för modellen med samspel. Därför väljer vi 6 knutar för båda modellerna. Från tabellerna för ES, NV och SP i Appendix ser vi att bäst anpassning för ES är med 3 knutar, för NV med 5 knutar och för SP med 6 knutar.

Modellerna med samspelsterm ser ut på följande vis

$$\text{logit}(\pi(x_i)) = \alpha + \beta \mathbb{1}\{\text{Utl}\}_i + S_1(x_1) + S_2(x_i)\mathbb{1}\{\text{Utl}\}_i,\ i = 1, ..., 12$$

där $\mathbb{1}\{\text{Utl}\}$ är indikator som antar värde 0 för personer med svensk bakgrund och 1 för personer med utländsk bakgrund, S_1 är splinefunktion för svensk bakgrund och $S_1 + S_2$ är splinefunktion för utländsk bakgrund. Splinefunktionerna har formen

$$S_1(x_i) = \beta_{11}B_1(x_i) + \beta_{12}B_2(x_i) + \ldots + \beta_{1m}B_m(x_i)$$
$$S_2(x_i) = \beta_{21}B_1(x_i) + \beta_{22}B_2(x_i) + \ldots + \beta_{2m}B_m(x_i)$$

där B_n är basfunktioner och $n = 1, ..., 4$ för ES, $n = 1, ..., 7$ för IB, $n = 1, ..., 6$ för NV och $n = 1, ..., 7$ för SP.

Vidare gör vi ett Likelihood-ratio test för att testa om bakgrunderna har olika tidstrender inom varje program. Vår nollhypotes och alternativhypotes för varje program är då

$$H_0 : \beta_{21} = ... = \beta_{2m} = 0$$
$$H_1 : \text{åtminstone en av } \beta_{2n} \neq 0$$

Om samspelsterm är signifikant har vi att tidstrenden för personer med svensk och utländsk bakgrund är olika inom varje program, splinefunktionerna för bakgrunderna kan ha olika antar knutar.

Vi väljer signifikansnivå på 5% och Tabell 9 sammanfattar testen för dem fyra programmen.

Tabell 9: Likelihood ratio test av samspel mellan bakgrund och tidstrenden.

Program	G^2 (df)	p-värde
ES	11.66 (4)	0.0201
IB	5.30 (7)	0.6236
NV	8.99 (6)	0.1742
SP	118.33 (7)	$< 2.2 \cdot 10^{-16}$

Från tabellen ser vi att p-värdet för IB och NV är större än 0.05, nollhypotesen förkastas ej och personer med svensk respektive utländsk bakgrund har inte olika tidstrend. För ES och SP är p-värdet mindre än 0.05, nollhypotesen förkastas och personer med svensk respektive utländsk bakgrund har olika tidstrender.

Nu när vi vet om samspelstermen är signifikant eller inte vill vi anpassa B-splines för bakgrunderna var för sig inom ES och SP samt undersöka huvudeffekten av bakgrunderna inom IB och NV.

IB och NV

Modellen för IB och NV ser nu ut på följande vis

$$\text{logit}(\pi(x_i)) = \alpha + \beta \mathbb{1}\{\text{Utl}\}_i + S(x_1),\ i = 1, ..., 12$$

där $S = S_{IB}$ för IB och $S = S_{NV}$ som är de anpassade tidstrenderna från avsnitt 4.2 s.21.

Eftersom samspel mellan bakgrund och tidstrenden inte är signifikat så har vi att oddskvoten mellan oddset för att en person med utländsk bakgrund påbörjar studier och oddset för att en person med svensk bakgrund påbörjar studier är konstant över åren 2000 till 2011. Oddskvoten är $OR = \exp\{\beta\}$. Skattningen av β i modellen för IB är $\hat{\beta} = -0.49981$, vilket ger

$$OR_{IB} = \exp\{-0.49981\} \approx 0.61$$

och skattningen av β i modellen för NV är $\hat{\beta} = 0.02426$, vilket ger

$$OR_{NV} = \exp\{0.02426\} \approx 1.025.$$

Standardavvikelsen för $\hat{\beta}$ är $\hat{\sigma}(\hat{\beta}) = 0.05781$ för IB, vilket ger ett 95%-igt konfidensintervall för oddskvoten

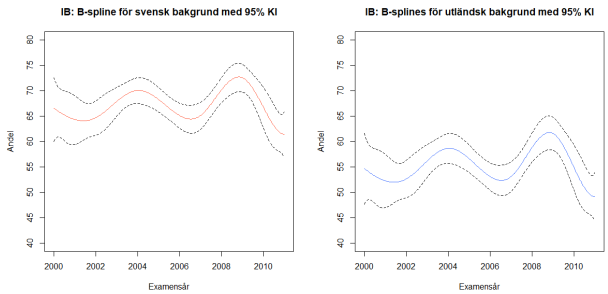
$$KI_{OR_{IB}} \approx (0.54, 0.68)$$

och standardavvikelsen för $\hat{\beta}$ är $\hat{\sigma}(\hat{\beta}) = 0.01729$ för NV, vilket ger ett 95%-igt konfidensintervall för oddskvoten

$$KI_{OR_{NV}} \approx (0.990, 1.060).$$

Oddskvoten för IB är signifikant men för NV är den inte signifikant.

I Figur 12 nedan ser vi den skattade tidstrenden för IB uppdelad på bakgrunderna med 95% konfidensintervall. Tidstrenden är samma för bakgrunderna men andelen med svensk bakgrund är högre än andelen med utländsk bakgrund för alla år mellan 2000 och 2011. I Appendix 6.B Figur A.5 kan vi även se hur tidstrenderna passar datapunkterna.



Figur 11: Skattad tidstrend för IB med 95% konfidensintervall uppdelad på bakgrund.

ES och SP

Eftersom samspel mellan bakgrund och tidstrend är signifikant för både ES och SP vill vi nu anpassa separata B-spline för personer med svensk och utländsk bakgrund inom båda programmen. Vi gör detta på samma sätt som tidigare, genom att undersöka AIC- värdet för olika antal knutar och välja den modell där AIC-värdet är minst.

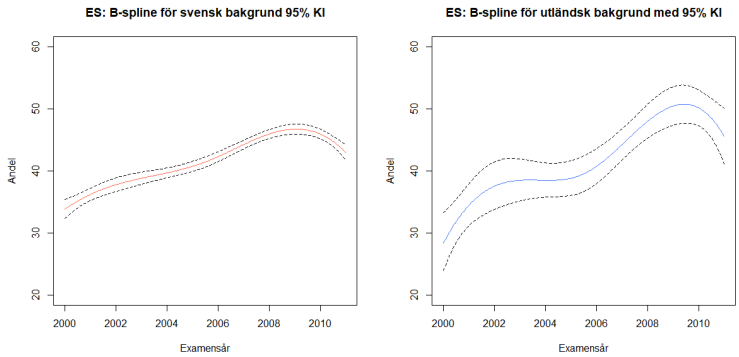
Tabell A.10 och Tabell A.11 i Appendix 6.A innehåller AIC-värden av anpassning för svensk respektive utländsk bakgrund inom ES respektive SP. Modellen för personer med svensk bakgrund har bäst anpassning med 3 knutar (AIC=423.44) och modellen för personer med utländsk bakgrund har bäst anpassning med 4 knutar (AIC=173.22) inom ES. Modellen för personer med svensk bakgrund har bäst anpassning med 6 knutar (AIC=2703.68) och modellen för personer med utländsk bakgrund har bäst anpassning med 6 knutar (AIC=842.33) inom SP. Tidstrenderna ser ut på följande vis

$$\text{Sv: logit}(\pi(x_i)) = \alpha + S_{Sv}(x_i);\ S_{Sv}(x_i) = \sum_n \beta_{Sn} B_n(x_i)$$

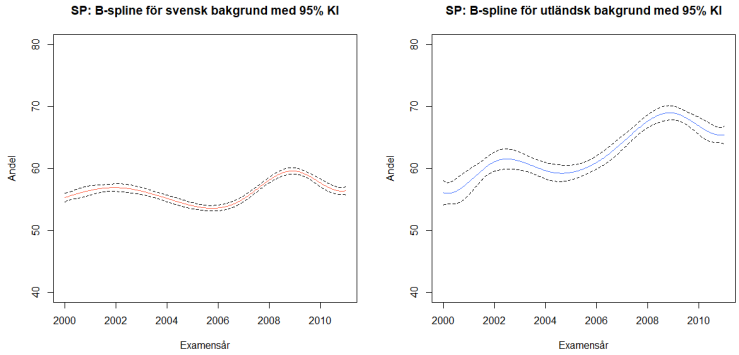
$$\text{Utl: logit}(\pi(x_i)) = \alpha + S_{Utl}(x_i);\ S_{Utl}(x_i) = \sum_n \beta_{Un} B_n(x_i)$$

där $n = 1, ..., 4$ för svensk bakgrund inom ES, $n = 1, ..., 5$ för utländsk bakgrund inom ES och $n = 1, ..., 7$ för både svensk och utländsk bakgrund inom SP.

I Figur 12 och Figur 13 nedan ser vi dem skattade tidstrenderna för båda bakgrunderna med 95% konfidensintervall för ES respektive SP. Konfidensintervallen följer linjerna väl, dock är det bredare konfidensintervall för utländsk bakgrund jämfört med svensk bakgrund inom både ES och SP. Skattningarna är alltså ganska säkra för svensk bakgrund och lite mindre säkra för utländsk bakgrund. I Appendix 6.B Figur A.4 och Figur A.6 kan vi även se hur tidstrenderna passar datapunkterna.



Figur 12: Skattade tidstrender med 95% konfidensintervall för svensk och utländsk bakgrund inom ES.



Figur 13: Skattade tidstrender med 95% konfidensintervall för svensk och utländsk bakgrund inom SP.

Inom ES och SP kan vi se att andelen med svensk och utländsk bakgrund har ökat från år 2000 till 2011, där ökningen är störst för personer med utländsk bakgrund jämfört med svensk bakgrund. Inom ES kan vi se att andelen med svensk bakgrund är högre än med utländsk bakgrund innan år 2007, och efter år 2007 är andelen med utländsk bakgrund högre än med svensk bakgrund. Inom SP kan vi se att andelen personer med utländsk bakgrund som påbörjat studier är högre än andelen med svensk bakgrund för alla år mellan 2000 och 2011.

4.2.3 Analys av samspel mellan kön och bakgrund

Nu när vi har undersökt hur kön och personers bakgrund inverkar på tidstrenden inom programmen så kan det vara intressant att undersöka om det föreligger samspel mellan kön och bakgrund inom varje program samt samspel mellan kön, bakgrund och tidstrenden. Vi kommer göra detta för varje program separat, som tidigare.

ES

För programmet ES finns det ett signifikant samspel mellan bakgrund och tidstrenden men inte mellan kön och tidstrenden. Kön har dock en signifikant huvudeffekt på tidstrenden. Det kan nu vara intressant att undersöka om det finns signifikant samspel mellan bakgrund och kön inom ES, för att se om effekten av kön är olika för personer med svensk bakgrund jämfört med utländsk bakgrund.

Som tidigare kommer vi jämföra två modeller, en modell utan samspelsterm och en med samspelsterm. I Tabell 10 ser vi AIC-värdena för båda modellerna och bäst anpassning är med 3 knutar där AIC-värdet är lägst.

Modellen med samspelsterm ser ut på följande vis

$$\text{logit}(\pi(x_i)) = \alpha + \beta_1 \mathbb{1}\{\text{M}\}_i + \beta_2 \mathbb{1}\{\text{Utl}\}_i + \beta_3 \mathbb{1}\{\text{M,Utl}\}_i +$$

$$+ S_1(x_i) + S_2(x_i)\mathbb{1}\{\text{Utl}\}_i, \; i = 1, ..., 12$$

där $\mathbb{1}\{\text{M,Utl}\}$ är indikatorvariabel som antar värdet 1 för män med utländsk bakgrund och 0 annars, S_1 är splinefunktionen för personer med svensk bakgrund och $S_1 + S_2$ är splinefunktionen för personer med utländsk bakgrund som vi anpassade i avsnitt 4.2.2 s.30.

Tabell 10: AIC för modell utan och med samspel mellan kön och bakgrund, inom ES.

Antal knutar	AIC, utan samspel	AIC, med samspel
2	392.26	393.88
3	375.53	377.18
4	380.43	382.07
5	382.84	384.48
6	385.96	387.61

Vi utför sedan ett Likelihood-ratio test för att se om huvudeffekten av kön är olika beroende på bakgrund. Vår nollhypotes och alternativhypotes är då

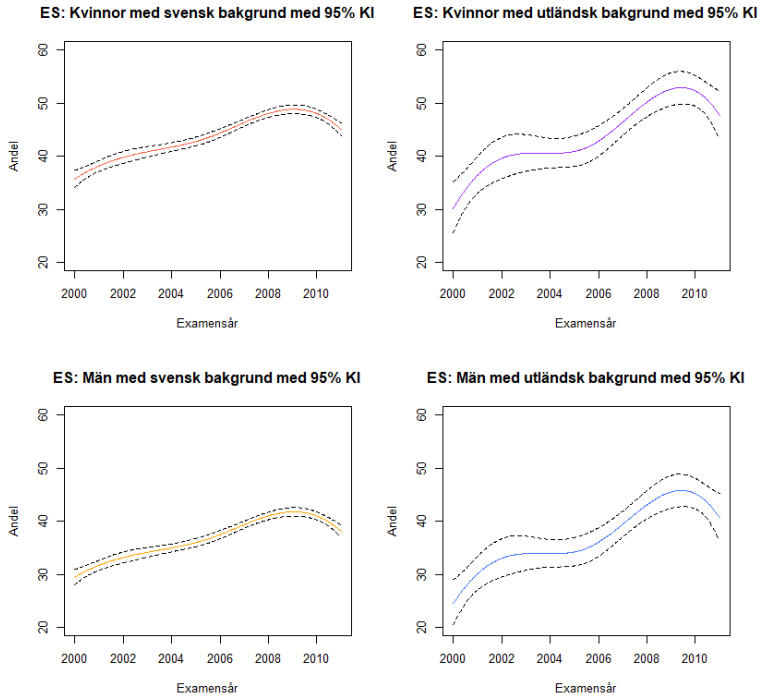
$$H_0 : \beta_3 = 0$$
$$H_1 : \beta_3 \neq 0$$

Vi sätter signifikansnivån till 5% och får p-värde = 0.552 för teststatistikan $G^2 = 0.354$ med $df = 1$. Vi förkastar inte nollhypotesen, huvudeffekten av kön är inte olika för dem två bakgrunderna.

Således har vi att huvudeffekten av kön på tidstrenden för svensk bakgrund är samma som på tidstrenden för utländsk bakgrund. Oddskvoten mellan oddset för att en man påbörjar studier och oddset för att en kvinna påbörjar studier är samma som oddskvoten från avsnitt 4.2.1 s.25

$$OR = \exp\{\hat{\beta}_1\} \approx 0.75.$$

I Figur 14 ser vi de skattade tidstrenderna för båda bakgrunderna uppdelade på män och kvinnor med 95% konfidensintervall. För personer med svensk bakgrund är konfidensintervallen smala och följer linjerna bra, skattningarna är ganska säkra. För personer med utländsk bakgrund är konfidensintervallen bredare så skattningarna här är mindre säkra. Andelen kvinnor är högre än andelen män för både svensk och utländsk bakgrund. I Appendix 6.B Figur A.7 kan vi se hur de skattade tidstrenderna passar datapunkterna.



Figur 14: Skattade tidstrender med 95% konfidensintervall för andelen med svensk och utländsk bakgrund inom ES uppdelat på kön.

IB

För programmet IB finns det inget signifikant samspel mellan kön och tidstrenden samt mellan bakgrund och tidstrenden, men bakgrund har dock en signifikant huvudeffekt på tidstrenden. Det kan nu vara intressant att undersöka om det finns signifikant samspel mellan bakgrund och kön eftersom huvudeffekten av kön kan vara signifikant i samspel med bakgrund. Vi vill alltså undersöka om huvudeffekterna av kön och bakgrund är olika beroende på varandra.

Som tidigare kommer vi jämföra två modeller, en modell utan samspelsterm och en med samspelsterm. I Tabell 11 ser vi AIC-värdena för båda modellerna och bäst anpassning är med 6 knutar där AIC-värdet är lägst.

Modellen med samspelsterm ser ut på följande vis

$$\text{logit}(\pi(x_i)) = \alpha + \beta_1 \mathbb{1}\{\text{M}\}_i + \beta_2 \mathbb{1}\{\text{Utl}\}_i + \beta_3 \mathbb{1}\{\text{M,Utl}\}_i +$$

$$+S(x_i),\ i = 1, ..., 12$$

där $\mathbb{1}\{\text{M,Utl}\}$ är indikatorvariabel som antar värdet 1 för män med utländsk bakgrund och 0 annars och $S = S_{IB}$ är tidstrenden som anpassades i avsnitt 4.2 s.21.

Tabell 11: AIC för modell utan och med samspel mellan kön och bakgrund, inom IB.

Antal knutar	AIC, utan samspel	AIC, med samspel
2	304.15	304.49
3	302.97	303.38
4	295.28	295.74
5	299.79	300.27
6	294.53	294.94

Vidare gör vi ett Likelihood-ratio test för att se om huvudeffekten av kön och bakgrund är olika beroende på varandra. Vår nollhypotes och alternativhypotes är då

$$H_0 : \beta_3 = 0$$
$$H_1 : \beta_3 \neq 0$$

Vi sätter signifikansnivån till 5% och får p-värde=0.2076 för teststatistikan $G^2 = 1.5879$ med $df = 1$. Vi förkastar inte nollhypotesen, huvudeffekten av kön och bakgrund är inte olika beroende på varandra.

I avsnitt 4.2.1 s.25 kom vi fram till att oddskvoten mellan oddset för att en man påbörjar studier och oddset för att en kvinna påbörjar studier inte är signifikant inom IB. På grund av detta är huvudeffekten av kön på tidstrenden inte signifikant.

Oddskvoten mellan oddset för att en person med utländsk bakgrund påbörjar studier och oddset för att en person med utländsk bakgrund påbörjar studier är samma som oddskvoten från avsnitt 4.2.2 s.29

$$OR = \exp\{\hat{\beta}\} \approx 0.61.$$

NV

För programmet NV finns det ett signifikant samspel mellan kön och tidstrenden men inte mellan bakgrund och tidstrenden. Bakgrund har även ingen signifikant huvudeffekt på tidstrenden. Det kan nu vara intressant att undersöka om det finns signifikant samspel mellan bakgrund och kön eftersom huvudeffekten av bakgrund kan vara signifikant i samspel med kön. Vi vill alltså undersöka om huvudeffekten av bakgrund på tidstrendern för kvinnor respektive män är olika.

Som tidigare kommer vi jämföra två modeller, en modell utan samspelsterm och en med samspelsterm. I Tabell 12 ser vi AIC-värdena för båda modellerna och bäst anpassning är med 5 knutar där AIC-värdet är lägst.

Modellen med samspelsterm ser ut på följande vis

$$\text{logit}(\pi(x_i)) = \alpha + \beta_1 \mathbb{1}\{\text{M}\}_i + \beta_2 \mathbb{1}\{\text{Utl}\}_i + \beta_3 \mathbb{1}\{\text{M,Utl}\}_i +$$
$$+ S_1(x_i) + S_2(x_i)\mathbb{1}\{\text{M}\}_i, \ i = 1, ..., 12$$

där $\mathbb{1}\{\text{M,Utl}\}$ är indikatorvariabel som antar värdet 1 för män med utländsk bakgrund och 0 annars, S_1 är splinefunktionen för kvinnor och S_1+S_2 är splinefunktionen för män som vi anpassade i avsnitt 4.2.1 s.26.

Tabell 12: AIC för modell utan och med samspel mellan kön och bakgrund, inom NV.

Antal knutar	AIC, utan samspel	AIC, med samspel
2	571.50	563.39
3	472.84	464.73
4	487.75	479.66
5	468.12	459.99
6	481.60	473.45

Vidare utför vi ett Likelihood-ratio test för att undersöka om huvudeffekten av bakgrund är olika för könen. Vår nollhypotes och alternativhypotes är då

$$H_0 : \beta_3 = 0$$
$$H_1 : \beta_3 \neq 0$$

Vi sätter signifikansnivån till 5% och får p-värde=0.00146 för teststatistikan $G^2 = 10.127$ med $df = 1$. Vi förkastar alltså nollhypotesen, huvudeffekten av bakgrund är olika för könen. Således har vi att huvudeffekten av bakgrund på tidstrenden för kvinnor respektive män är olika.

Oddskvoten mellan oddset för att en kvinna med utländsk bakgrund påbörjar studier och oddset för att en kvinna med svensk bakgrund påbörjar studier är given av $OR_K = \exp\{\beta_2\}$. Skattningen av β_2 i modellen är $\hat{\beta}_2 = 0.08148$ vilket ger oddskvoten

$$OR_K = \exp\{0.08148\} \approx 1.08$$

med ett 95%-igt konfidensintervall för oddskvoten där standardavvikelsen för $\hat{\beta}_2$ är $\hat{\sigma}(\hat{\beta}_2) = 0.02589$

$$KI_{OR_K} = (1.03, 1.14).$$

Oddkvoten mellan oddset för att en man med utländsk bakgrund påbörjar studier och oddset för att en man med svensk bakgrund påbörjar studier är given av $OR_M = \exp\{\beta_2+\beta_3\}$. Skattningen av $\beta_2+\beta_3$ i modellen är $\hat{\beta}_2+\hat{\beta}_3 = -0.029101$ vilket ger oddskvoten

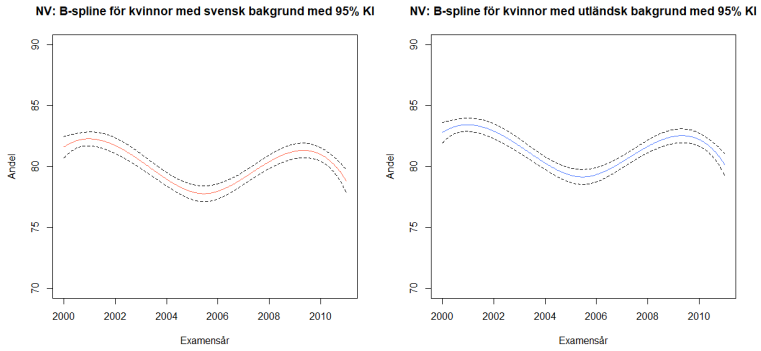
$$OR_M = \exp\{-0.029101\} \approx 0.97$$

med ett 95%-igt konfidensintervall för oddskvoten där standardavvikelsen för $\hat{\beta}_2 + \hat{\beta}_3$ är given av $\sqrt{\hat{\sigma}^2(\hat{\beta}_2) + \hat{\sigma}^2(\hat{\beta}_3) + 2\widehat{cov}(\hat{\beta}_2, \hat{\beta}_3)} = 0.023287$

$$KI_{OR_M} = (0.93, 1.02).$$

Oddskvoten OR_K är signifikant men oddkvoten OR_M är inte det. På grund av detta är huvudeffekten av bakgrund på tidstrenden för män inte signifikant, andelen män med utländsk bakgrund som påbörjar studier är således samma som andelen män med svensk bakgrund för alla år mellan 2000 och 2011.

I Figur 15 ser vi den skattade tidstrenden för kvinnor uppdelad på bakgrunderna med 95% konfidensintervall. Tidstrenden för kvinnor med svensk bakgrund är samma som med utländsk bakgrund, men andelen kvinnor med utländsk bakgrund är högre än andelen med svensk bakgrund för alla år mellan 2000 och 2011. I Appendix 6.B Figur A.8 kan vi se hur de skattade tidstrenderna passar datapunkterna, där den skattade tidstrenden för män också är med.



Figur 15: Skattad tidstrend med 95% konfidensintervall för andelen kvinnor uppdelad svensk och utländsk bakgrund, inom IB.

SP

För programmet SP finns det signifikant samspel mellan kön och tidstrenden samt mellan personers bakgrund och tidstrenden. Här kan det nu vara intressant att undersöka om det finns ett signifikant tredjegradssamspel mellan kön, personers bakgrund och tidstrend. Om ett sådant samspel är signifikant så innebär det att kvinnor med svensk bakgrund, kvinnor med utländsk bakgrund, män med svensk bakgrund och män med utländsk bakgrund kan alla ha olika tidstrender.

Vi vill jämföra modellen utan tredjegradssamspel med modellen med tredjegradssamspel. I Tabell 13 kan man se AIC-värdena för båda modellerna. Som vi ser så är bästa anpassningen för båda modellerna är med 6 knutar, lägst AIC.

Modellen med trejdegradssamspel ser ut på följande vis

$$\text{logit}(\pi(x_i)) = \alpha + \beta_1 \mathbb{1}\{\text{M}\}_i + \beta_2 \mathbb{1}\{\text{Utl}\}_i + \beta_3 \mathbb{1}\{\text{M,Utl}\}_i +$$

$$+S_1(x_i) + S_2(x_i)\mathbb{1}\{\text{M}\}_i + S_3(x_i)\mathbb{1}\{\text{Utl}\}_i + S_4(x_i)\mathbb{1}\{\text{M,Utl}\}_i, \;\; i = 1, ..., 12$$

Tabell 13: AIC för modell utan och med tredjegradssamspel mellan kön, bakgrund och tidstrenden.

Antal knutar	AIC, utan samspel	AIC, med samspel
2	753.15	724.44
3	556.04	527.88
4	575.98	549.41
5	551.77	526.39
6	534.76	511.70

Vi utför ett Likelihood-ratio test av nollhypotesen och alternativhypotesen

$$H_0 : \beta_{41} = ... = \beta_{47} = 0$$
$$H_1 : \text{åtminstone en av } \beta_{4n} \neq 0,\ n = 1, ..., 7$$

med signifikansnivå 5% och får p-värde= $4.787 \cdot 10^{-6}$. Nollhypotesen förkastas, det finns ett signifikant tredjegradssamspel mellan kön, personers bakgrund och tidstrenden. Åtminstone två av dem fyra tidstrenderna som är nämnda ovan är olika (har olika splinefunktioner).

Vi vill nu vidare undersöka om det finns ett signifikant samspel mellan tidstrenden för kvinnor och personers bakgrund, tidstrenden för män och personers bakgrund, tidstrenden för personer med svensk bakgrund och kön samt tidstrenden för personer med utländsk bakgrund och kön. Detta är på grund av att vi inte säker vet om alla fyra tidstrender som är nämnda ovan är olika, vi vet bara att åtminstone två av dem är det.

Vi kommer nu utföra fyra stycken Likelihood-ratio test, där vi kommer jämför en modell utan samspel och en med samspel för kvinnor, män, personer med svensk och utländsk bakgrund. Tabell 14 visar AIC-värden för modellerna utan och med samspel mellan kön och tidstrenden för personer med utländsk bakgrund. Resterande tre tabellerna med AIC-värden finns i Appendix 6.A Tabell A.12, Tabell A.13 och Tabell A.14.

Tabell 14: AIC för modell utan och med samspel mellan kön och tidstrenden för personer med utländsk bakgrund inom SP.

Knutar	AIC, utan samspel	AIC, med samspel
2	258.27	258.54
3	220.23	220.84
4	225.32	227.83
5	219.45	223.18
6	216.68	221.90

Från Tabell 14 ser vi att bäst anpassning för modellen utan samspel är med 6 knutar och för modellen med samspel är med 3 knutar. Eftersom sänkningen i AIC-värde från 3 till 6 knutar för modellen utan samspel är större än ökningen

för modellen med samspel så väljer vi 6 knutar för båda modellerna. Från tabellerna i Appendix 6.A ser vi att bäst anpassning för kvinnor är med 6 knutar, för män med 5 knutar och för personer med svensk bakgrund med 6 knutar.

Vi utför Likelihood-ratio test av nollhypotesen att samspel inte finns mot alternativhypotesen att samspel finns. Vi sätter signifikansnivån på 5%, och sammanfattning av testen kan ses i Tabell 15.

Tabell 15: Likelihood-ratio test av samspel mellan kvinnor, män, svensk bakgrund och utländsk bakgrund.

Samspel	G^2 (df)	p-värde
Kvinnor och bakgrund	117.11 (7)	$< 2.2 \cdot 10^{-16}$
Män och bakgrund	20 (6)	0.00277
Svensk bakgrund och kön	32.20 (7)	$3.731 \cdot 10^{-5}$
Utländsk bakgrund och kön	8.78 (7)	0.2686

Från tabellen ser vi att kvinnor med svensk och utländsk bakgrund har olika tidstrender, män med svensk och utländsk bakgrund har olika tidstrender samt kvinnor och män med svensk bakgrund har olika tidstrender. Kvinnor och män med utländsk bakgrund har dock samma tidstrend då p-värdet är större än 0.05, nollhypotesen att inget samspel finns mellan utländsk bakgrund och kön förkastas ej.

Vidare anpassar vi B-splines för kvinnor med svensk bakgrund och män med svensk bakgrund. Tabell A.15 i Appendix 6.A innehåller AIC-värden för olika antal knutar för anpassning av tidstrenderna. Kvinnor med svensk bakgrund har bäst anpassning med 6 knutar (AIC= 152.35) och män med svensk bakgrund har bäst anpassning med 5 knutar (AIC=140.60). Tidstrenderna ser ut på följande vis

$$\text{K,Sv: logit}(\pi(x_i)) = \alpha + S_{KS}(x_i);\ S_{KS}(x_i) = \sum_n \beta_{KSn} B_n(x_i),\ n = 1, ..., 7$$

$$\text{M,Sv: logit}(\pi(x_i)) = \alpha + S_{MS}(x_i);\ S_{MS}(x_i) = \sum_n \beta_{MSn} B_n(x_i),\ n = 1, ..., 6$$

För personer med utländsk bakgrund har vi bäst anpassning med 6 knutar (AIC=842.33) som vi även tog fram i avsnitt 4.2.2 s.30. Oddskvoten mellan oddset att en man påbörjar studier och att en kvinna påbörjar studier är konstant mellan åren 2000 och 2011 för personer med utländsk bakgrund. Oddskvoten är given som $OR = \exp\{\beta_1 + \beta_3\}$. Skattningen av $\beta_1 + \beta_3$ i modellen är $\hat{\beta}_1 + \hat{\beta}_3 = -0.54888$ vilket ger oddskvoten

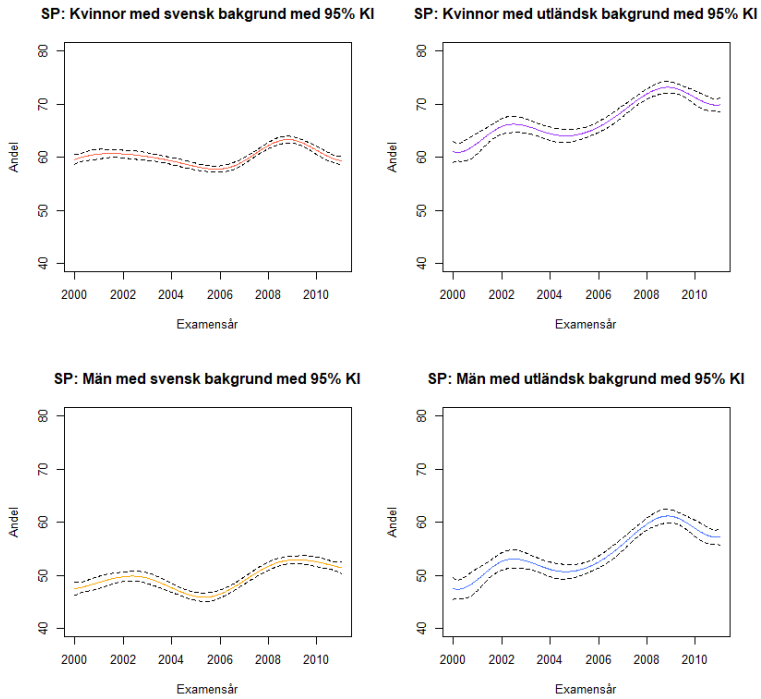
$$OR = \exp\{-0.54888\} = 0.58$$

och ett 95%-igt konfidensintervall för oddskvoten, där standardavvikelsen för $\hat{\beta}_1 + \hat{\beta}_3$ är $\hat{\sigma}(\hat{\beta}_1 + \hat{\beta}_3) = 0.02192$, är följande

$$KI_{OR} = (0.55, 0.60)$$

Oddskvoten är således signifikant.

I Figur 16 ser vi de skattade tidstrenderna med 95% konfidensintervall för kvinnor och män med svensk och utländsk bakgrund. Konfidenintervallen är smala och följer linjerna väl, skattningarna är således ganska säkra för dessa modeller. I Appendix 6.B Figur A.9 kan vi se hur väl de skattade tidstrenderna följer punkterna och vi kan säga att tidstrenderna följer punkterna väl.



Figur 16: Skattade tidstrender med 95% konfidensintervall för kvinnor med svensk och utländsk bakgrund samt för män men svensk och utländsk bakgrund, inom SP.

Andelen personer med utländsk bakgrund som påbörjat studier är högre än andelen personer med svensk bakgrund för kvinnor mellan åren 2000 och 2011 och för män mellan åren 2001 och 2011. Andelen kvinnor som påbörjat studier är högre än andelen män för både svensk och utländsk bakgrund. Andelen ökar mest för personer med utländsk bakgrund och andelen män med svensk bakgrund har ökat mer än andelen kvinnor med svensk bakgrund från år 2000 till år 2011.

4.3 Modell för ES

Det finns signifikant samspel mellan bakgrund och tidstrend men inte mellan kön och tidstrend inom ES. Kön har däremot en signifikant huvudeffekt på tidstrenderna. Slutgiltiga modellen för ES blev

$$\text{logit}(\pi(x_i)) = \alpha + \beta_1 \mathbb{1}\{\text{M}\}_i + \beta_2 \mathbb{1}\{\text{Utl}\}_i + S_1(x_i) + S_2(x_i)\mathbb{1}\{\text{Utl}\}_i,\ i = 1, ..., 12$$

där S_1 är splinefunktionen för personer med svensk bakgrund och $S_1 + S_2$ är för personer med utländsk bakgrund som anpassades i avsnitt 4.2.2 s.30.

Oddskvoten $OR = \exp\{\beta_1\}$ är 0.75. Vi kan tolka värdet på oddkvoten som att oddset för att en man påbörjar studier är 0.75 av oddset för att en kvinna påbörjar studier.

4.4 Modell för IB

Inom IB finns inget signifikant samspel med mellan kön och tidstrend, och kön har ingen signifikant huvudeffekt. Detta innebär att män och kvinnor har samma tidstrend och andelen kvinnor och män som påbörjat vidare studier är samma. Inget signifikant samspel mellan personers bakgrund och tidstrenden finns men personers bakgrund har signifikant huvudeffekt på tidstrenden inom IB. Slutgiltiga modellen blev

$$\text{logit}(\pi(x_i)) = \alpha + \beta \mathbb{1}\{\text{Utl}\}_i + S(x_i),\ i = 1, ..., 12$$

där S är splinefunktionen för IB som anpassades i avsnitt 4.2 s.21.

Oddskvoten $OR = \exp\{\beta\}$ är 0.61. Vi kan tolka värdet på oddskvoten som att oddset för att en person med utländsk bakgrund påbörjar studier är 0.61 av oddset för att en person med svensk bakgrund påbörjar studier.

4.5 Modell för NV

NV har signifikant samspel mellan kön och tidstrend samt inget signifikant samspel mellan personers bakgrund och tidstrend. Signifikant samspel mellan kön och bakgrund finns. Slutgiltiga modellen blev

$$\text{logit}(\pi(x_i)) = \alpha + \beta_1 \mathbb{1}\{\text{M}\}_i + \beta_2 \mathbb{1}\{\text{Utl}\}_i + \beta_3 \mathbb{1}\{\text{M,Utl}\}_i +$$

$$+ S_1(x_i) + S_2(x_i)\mathbb{1}\{\text{M}\}_i,\ i = 1, ..., 12$$

där $\beta_3 = -\beta_2$ eftersom andelen män med utländsk bakgrund som påbörjat studier är samma som andelen män med svensk bakgrund som påbörjat studier. S_1 är splinefunktionen för kvinnor och $S_1 + S_2$ är splinefunktionen för män som anpassades i avsnitt 4.2.1 s.26.

Oddskvoten $OR = \exp\{\beta_2\}$ är 1.08. Vi kan tolka värdet på oddskvoten som att oddset för att en kvinna med utländsk bakgrund påbörjar studier är 1.08 av oddset för att en kvinna med svensk bakgrund påbörjar studier.

4.6 Modell för SP

Inom SP finns ett signifikant tredjegradssamspel där kvinnor med svensk bakgrund, män med svensk bakgrund samt personer med utländsk bakgrund har olika tidstrender. Slutgiltiga modellen blev

$$\text{logit}(\pi(x_i)) = \alpha + \beta_1 \mathbb{1}\{\text{M}\}_i + \beta_2 \mathbb{1}\{\text{Utl}\}_i + \beta_3 \mathbb{1}\{\text{M,Utl}\}_i +$$

$$+S_1(x_i) + S_2(x_i)\mathbb{1}\{\text{M}\}_i + S_3(x_i)\mathbb{1}\{\text{Utl}\}_i + S_4(x_i)\mathbb{1}\{\text{M,Utl}\}_i, \ i = 1, ..., 12$$

där $S_4 = -S_2$ eftersom tidstrenden för kvinnor och män med utländsk bakgrund är samma. S_1 är splinefunktionen för kvinnor med svensk bakgrund, $S_1 + S_2$ är för män med svensk bakgrund som anpassades i avsnitt 4.2.3 s.36 och $S_1 + S_3$ är splinefunktionen för personer med utländsk bakgrund som anpassades i avsnitt 4.2.2 s.30.

Oddskvoten $OR = \exp\{\beta_1 + \beta_3\}$ är 0.58 och kan tolkas som att oddset för att en man med utländsk bakgrund påbörjar studier är 0.58 av oddset för att en kvinna med utländsk bakgrund påbörjar studier.

4.7 Prediktion

Nu när vi har undersökt och fått fram modeller som kan tänka sig beskriva tidstrenden av andelen som påbörjat studier inom tre år efter avslutad gymnasial utbildning för de olika programmen så kan det vara intressant att se hur väl dessa modeller predikerar framtiden. Här väljer vi att predikera andelen som påbörjar studier för examensår 2012. Från Statistiska centralbyrån har vi hämtat information om andelen som påbörjat studier inom tre år för de fyra programmen samt för kvinnor och män inom dessa program. Information om personers bakgrund fanns tyvärr inte med.

Vi börjar med att predikera andelen för varje program utan uppdelning på kön. Modellerna för programmen är logistiska regressionmodeller med en B-splinefunktion av examensår där ES är med 3 knutar, IB är med 6 knutar, NV är med 5 knutar och SP är med 6 knutar. Yttre knutarna sitter på åren 2000 och 2011 och för att predikera värden för året 2012 behöver den övre yttre knuten sättas på det året. När vi sätter den yttre knuten på år 2012 och tar bort knuten på 2011 förändras inte AIC-värdet för modellerna, således är knuten vid 2011 inte nödvändig. Tabell 16 nedan visar predikerade värden med 95% konfidensintervall från våra modeller samt värden hämtade från SCB.

Tabell 16: Prediktion och sanna värden av andelen som påbörjar studier inom tre år efter avslutad gymnasial utbildning våren 2012.

Program	Prediktion %, (KI)	SCB %
ES	37.1, (34.4, 40.0)	44.0
IB	58.7, (29.0, 83.2)	60.8
NV	76.5, (72.7, 79.8)	84.8
SP	60.1, (56.0, 64.0)	59.7

Som vi kan se så stämmer prediktioner från våra modeller bra för IB och SP och mindre bra för ES och NV.

Vidare predikerar vi andelen kvinnor och män inom varje program. Eftersom IB inte hade någon signifikant effekt av kön kommer vi använda samma predikerade värde för kvinnor och män som det predikerade värdet för IB ovan. För ES är oddskvoten mellan män och kvinnor 0.75 och B-splinefunktionen samma som ovan. Modellerna för NV är med 3 knutar för kvinnor och 5 knutar för män. Modellerna för SP är med 6 knutar för kvinnor och med 5 knutar för män. Vi predikerar värden som ovan genom att sätta en yttre knut vid år 2012 och ta bort knuten på år 2011. Tabell 17 nedan visar predikerade värden med 95% konfidensintervall samt värden från SCB för kvinnor och män i varje program.

Tabell 17: Prediktion och sanna värden av andelen kvinnor och män som påbörjar studier inom tre år efter avslutad gymnasial utbildning våren 2012.

Program	Kön	Prediktion %, (KI)	SCB (%)
ES	Kvinnor	39.1, (36.3, 42.0)	45.5
	Män	32.6, (30.0, 35.2)	41.2
IB	Kvinnor	58.7, (29.0, 83.2)	60.2
	Män	58.7, (29.0, 83.2)	62.0
NV	Kvinnor	73.7, (70.9, 76.3)	84.5
	Män	78.4, (73.5, 82.6)	85.1
SP	Kvinnor	63.0, (57.9, 67.9)	62.6
	Män	50.9, (46.2, 55.5)	54.8

Vi ser att prediktionerna från våra modeller stämmer bra för IB och SP. Prediktionerna stämmer mindre bra för ES och NV som ovan, dock så stämmer prediktionerna av att andelen kvinnor är större än andelen män inom ES och att andelen män är större än andelen kvinnor inom NV.

5 Slutsats och diskussion

I analysen kom vi fram till att högskoleförberedande programmen har olika tidstrender. Vi anpassade tidstrender med B-splines med 95% konfidensintervall för varje program och kom fram till att andelen personer som påbörjat vidare studier inom tre år, för alla examenår mellan 2000 och 2011, är störst inom NV, följt av IB, SP och ES. Tidstrenderna ser olika ut mellan programmen, där tidstrenden för IB pendlar mer än för de andra programmen.

Att NV-programmet i vår analys har störst andel personer som påbörjat vidare studier jämfört med dem andra programmen stämmer överens med vad Inkinen et al. skriver i rapporten *Universitet och högskolor: Årsrapport 2013*, (2013):

> Naturvetenskapsprogrammet är det gymnasieprogram som har i särklass högst övergångsfrekvens.

En anledning till varför NV-programmet har störst andel som påbörjat vidare studier kan vara att personer som går detta program redan från början har inställningen att i framtiden påbörja vidare studier eftersom programmet ger en bred grundläggande utbildning med stor valmöjlighet av högskole/universitetsutbildningar. Detta är dock bara spekulationer.

Att IB-programmet har en mer pendlande tidstrend kan bero på att denna grupp är betydligt mindre varpå andelen personer som påbörjat vidare studier påverkas mer av en ökning eller säkning av antalet som påbörjar vidare studier jämfört med dem andra programmen. Konfidensintervallet är för IB bredare än för dem andra programmen, vilket också kan vara på grund av att gruppen är mindre än dem andra, och på grund av detta är det mer osäkert hur trenden egentligen förändras mellan åren 2000 och 2011.

Det visade sig även att andelen personer från programmet ES som påbörjat vidare studier har en kontinuerligt ökande trend mellan åren 2000 och 2009, medan trenden inte ändras så mycket för NV jämfört med dem andra programmen. Alla program har en ökande trend mellan åren 2006 och 2009 och minskande mellan åren 2009 och 2011. Tidstrendens högsta värde är året 2009 för alla program.

Att trenden kontinuerligt ökar från år 2006 till 2009 kan tänkas ha en förklaring i att konjukturen under dessa år försämrades med finanskrisen 2008, vilket Niklas Karlsson skriver i analysen *Nybörjare i högskolan 2015/16*, (2016):

> Från läsåret 2006/07 till läsåret 2009/10 ökade antalet högskolenybörjare kraftigt. Under denna period ökade efterfrågan på högskoleutbildning i samband med en försämring i konjunkturen.

Arbetsmarknaden för unga var ganska svag och arbetslösheten steg under den perioden varpå fler yngre valde att börja studera vidare, och i den statistiska analysen *Samband mellan konjunktur och högskolestudier?*, (2009), av Torbjörn Lindqvist har det påvisats att fler yngre, framförallt fler 19-åringar, påbörjat studier i och med att arbetslösheten steg. Anledningen för minskning i trenden från år 2009 kan även här vara arbetsmarkanaden, som då blev starkare.

Vidare kom vi fram till att inget signifikant samspel finns mellan kön och tidstrend inom ES och IB och att signifikant samspel finns inom NV och SP. Kvinnor och män har samma tidstrend inom ES, men andelen kvinnor är högre än andelen män som påbörjat vidare studier inom tre år för alla examensår mellan 2000 och 2011. Inom IB kom vi fram till att kvinnor och män har samma tidstrend samt att ingen signifikant skillnad av andelen som påbörjat vidare studier mellan kvinnor och män finns.

Inom programmen NV och SP kom vi fram till att kvinnor och män har olika tidstrender, och vi anpassade dessa med B-splines med 95% konfidensintervall. För både kvinnor och män inom NV pendlar andelen mellan åren 2000 och 2011, där andelen minskade från år 2000 till år 2011 för kvinnor och ökade lite för män. Tidstrenden för kvinnor inom SP är mer stabil mellan åren 2000 och 2006 än för män, där den pendlar upp och ner igen fram till år 2006. Även här har både kvinnor och män inom NV samt SP en ökning av andelen från år 2006 till 2009 och en sänkning från år 2009 till 2011. Andelen kvinnor som påbörjat vidare studier är högre än andelen män för alla år mellan 2000 och 2011 inom både NV och SP.

En icke-spekulativ förklaring till skillnaden mellan könen i andelen som påbörjat vidare studier är väldigt svår att ge eftersom skillnaden kan bero på många andra socionomiska faktorer som inte ingår i denna studie.

Vi kom sedan fram till att inget signifikant samspel mellan bakgrund och tidstrend finns inom IB och NV och att signifikant samspel finns inom ES och SP. Personer med svensk och utländsk bakgrund har samma tidstrend inom IB, men andelen med svensk bakgrund som påbörjat vidare studier är högre än andelen med utländsk bakgrund. Inom NV har personer med svensk och utländsk bakgrund samma tidstrend och ingen signifikant skillnad av andelen som påbörjat vidare studier finns mellan bakgrunderna.

Inom ES och SP har personer med svensk och utländsk bakgrund olika tidstrender, som vi anpassade med B-splines med 95% konfidensintervall. Inom ES ökar andelen med svensk bakgrund stadigt mellan åren 2000 och 2009 varpå andelen med utländsk bakgrund också ökar under dessa år men med ett stabilt värde mellan åren 2003 och 2005. Mellan åren 2000 och 2007 är andelen som påbörjat vidare studier med svensk bakgrund högre än med utländsk bakgrund och efter år 2007 är andelen med utländsk bakgrund högre än andelen med svensk bakgrund. Konfidensintervallet för utländsk bakgrund är dock bredare än för svensk bakgrund, således är det mer osäkert hur trenden egentligen förändras för personer med utländsk bakgrund mellan åren 2000 och 2011. Tidstrenden för svensk bakgrund liknar den för kvinnor inom SP, den är mer stabil mellan åren 2000 och 2006 än för utländsk bakgrund där den ökar och sjunker lite mellan dessa år. Andelen personer som påbörjat vidare studier ökar mer från år 2000 till år 2011 för utländsk bakgrund än för svensk bakgrund och andelen med utländsk bakgrund är högre än andelen med svensk bakgrund för alla år mellan 2000 och 2011.

En förklaring till att andelen med utländsk bakgrund är högre respektive blir högre än andelen med svensk bakgrund som påbörjat studier, inom SP och ES, kan vara att man som nyligen invandrad vill fortsätta studera för att etablera sig i det nya landet, samt att föräldrar kan ha andra förväntningar på grund av en annan kultur. Att detta gäller för ES och SP och inte för NV kan kanske bero på att de som är födda utomland eller nyligen invandrade går på programmen ES och SP eftersom dessa är enklare att komma in på jämfört med NV, där det i NV kanske snarare går de som är födda i Sverige men har föräldrar födda utomlands. Anledningen till att andelen med svensk bakgrund är högre än andelen med utländsk bakgrund inom IB kan vara att personer med utländsk bakgrund går just detta program på grund av att de eller deras föräldrar inte planerar att stanna i Sverige. Detta är dock inget vi vet eftersom dessa faktorer samt uppdelningen inom utländsk bakgrund inte ingår i denna studie.

I vidare analys kom vi fram till att inget samspel finns mellan kön och bakgrund inom programmen ES och IB. Inom ES är andelen kvinnor med svensk och utländsk bakgrund som påbörjat studier högre än andelen män med svensk och utländsk bakgrund. För både män och kvinnor inom ES blir andelen med utländsk bakgrund högre än andelen med svensk bakgrund efter år 2007. Samspel mellan kön och bakgrund finns inom NV, där vi kom fram till att andelen kvinnor med utländsk bakgrund är högre än andelen med svensk bakgrund, för alla år mellan 2000 och 2011, och att andelen män med utländsk respektive svensk bakgrund inte har någon signifikant skillnad. Inom SP finns samspel mellan kön, bakgrund och tidstrend, där vi kom fram till att kvinnor och män med svensk bakgrund har olika tidstrender samt att kvinnor och män med utländsk bakgrund har samma tidstrend. Andelen kvinnor är högre än andelen män som påbörjat studier för bägge bakgrunderna och denna trend gäller för alla år mellan 2000 och 2011. Andelen med utländsk bakgrund är högre än andelen med svensk bakgrund för bägge könen och denna trend gäller för alla år mellan 2001 och 2011.

I denna analys hade vi inte med så många faktorer som skulle kunna förklara varför tidstrenderna ser ut som de gör. Vi hade även inte ett så stort tidsspann, det är därför svårt att säga hur trender sett ut tidigare och famförallt hur de skulle kunna se ut i framtiden. För framtida udersökningar hade det varit intressant att ha med fler faktorer så som föräldrarnas utbildningsnivå, uppdelning inom utländsk bakgrund samt geografisk uppdelning. Något som även hade varit intressant att undersöka är hur tidstrenden beror på andelen som börjat studera direkt efter gymasiet. Det hade även varit bra att ha ett större tidsspann, för en bättre prediktion av framtiden.

6 Appendix

6.A Tabeller

Tabell A.1: AIC för logistisk regressionlinje för program.

Program	AIC
ES	666.19
IB	380.29
NV	739.91
SP	4551.10

Tabell A.2: AIC för modell utan och med samspel mellan kön och tidstrenden, inom ES.

Knutar	AIC, utan samspel	AIC, med samspel
2	393.37	394.28
3	377.24	380.46
4	380.27	382.25
5	380.62	385.09
6	382.02	387.76

Tabell A.3: AIC för modell utan och med samspel mellan kön och tidstrenden, inom NV.

Knutar	AIC, utan samspel	AIC, med samspel
2	600.13	570.75
3	497.69	472.31
4	511.94	487.20
5	492.62	467.58
6	503.44	481.05

Tabell A.4: AIC för modell utan och med samspel mellan kön och tidstrenden, inom SP.

Knutar	AIC, utan samspel	AIC, med samspel
2	1531.95	1511.93
3	1337.55	1319.74
4	1353.43	1337.69
5	1328.31	1313.76
6	1312.12	1297.91

Tabell A.5: AIC tidstrenden för kvinnor och män inom NV.

Knutar	AIC kvinnor	AIC män
2	268.76	301.98
3	222.97	249.34
4	227.37	259.83
5	225.25	242.33
6	228.75	250.66

Tabell A.6: AIC tidstrenden för kvinnor och män inom SP.

Knutar	AIC kvinnor	AIC män
2	1014.96	496.97
3	902.50	417.24
4	910.48	427.22
5	900.54	413.22
6	886.07	413.56

Tabell A.7: AIC för modell utan och med samspel mellan bakgrund och tidstrenden, inom ES.

Knutar	AIC, utan samspel	AIC, med samspel
2	616.00	613.10
3	600.32	596.66
4	603.41	601.63
5	603.51	603.78
6	605.21	607.15

Tabell A.8: AIC för modell utan och med samspel mellan bakgrund och tidstrenden, inom NV.

Knutar	AIC, utan samspel	AIC, med samspel
2	734.84	738.92
3	634.23	639.12
4	648.41	655.05
5	628.84	631.85
6	639.80	646.81

Tabell A.9: AIC för modell utan och med samspel mellan bakgrund och tidstrenden, inom SP.

Knutar	AIC, utan samspel	AIC, med samspel
2	3878.63	3772.44
3	3677.47	3571.95
4	3694.93	3590.26
5	3665.26	3561.84
6	3648.33	3544.01

Tabell A.10: AIC tidstrenden för personer med svensk och utländsk bakgrund inom ES.

Knutar	AIC svensk	AIC utländsk
2	434.23	178.87
3	423.44	173.22
4	426.31	175.32
5	426.61	177.18
6	428.60	178.65

Tabell A.11: AIC tidstrenden för personer med svensk och utländsk bakgrund inom SP.

Knutar	AIC svensk	AIC utländsk
2	2887.52	884.92
3	2725.36	846.59
4	2738.54	851.72
5	2716.77	845.07
6	2703.68	842.33

Tabell A.12: AIC för modell utan och med samspel mellan bakgrund och tidstrenden, för kvinnor inom SP.

Antal knutar	AIC, utan samspel	AIC, med samspel
2	498.42	393.43
3	381.97	276.15
4	390.26	285.48
5	379.62	275.25
6	364.44	261.33

Tabell A.13: AIC för modell utan och med samspel mellan bakgrund och tidstrenden, för män inom SP.

Knutar	AIC, utan samspel	AIC, med samspel
2	345.00	331.01
3	263.46	251.72
4	273.69	263.92
5	259.14	251.14
6	259.51	252.31

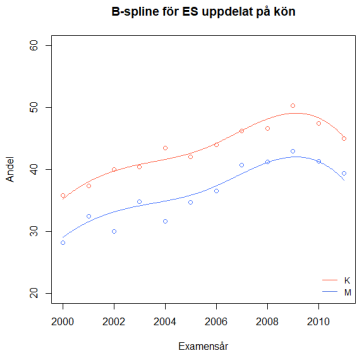
Tabell A.14: AIC för modell utan och med samspel mellan kön och tidstrenden, för personer med svensk bakgrund inom SP.

Knutar	AIC, utan samspel	AIC, med samspel
2	490.33	465.90
3	329.05	307.04
4	341.98	321.57
5	322.49	303.21
6	309.94	291.75

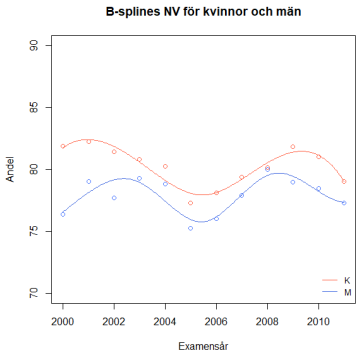
Tabell A.15: AIC tidstrenden för kvinnor och män med svensk bakgrund inom SP.

Knutar	AIC kvinnor	AIC män
2	255.46	210.44
3	162.99	144.05
4	168.08	153.50
5	162.61	140.60
6	152.35	142.79

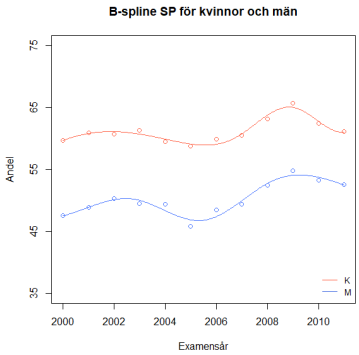
6.B Figurer



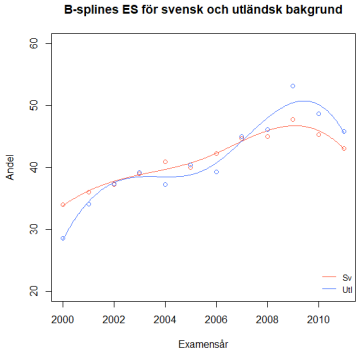
Figur A.1: Skattad tidstrend för ES uppdelad på kvinnor och män.



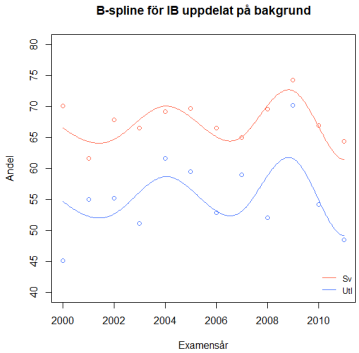
Figur A.2: Skattade tidstrender för kvinnor och män, NV.



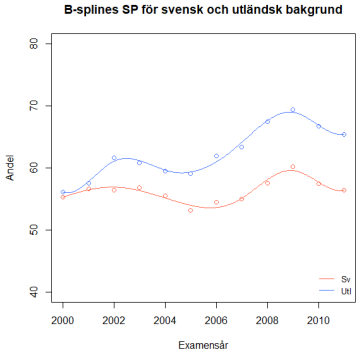
Figur A.3: Skattade tidstrender för kvinnor och män, SP.



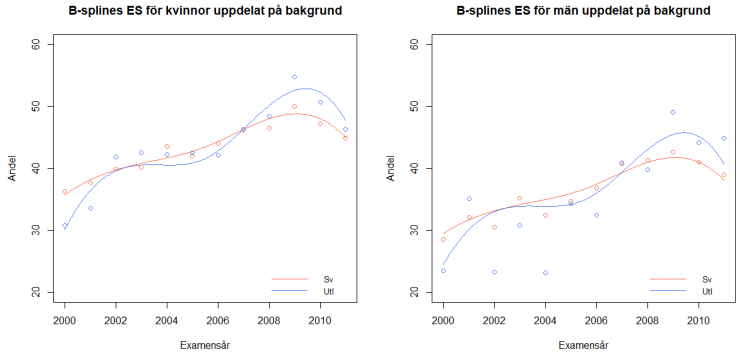
Figur A.4: Skattade tidstrender för personer med svensk och utländsk bakgrund inom ES.



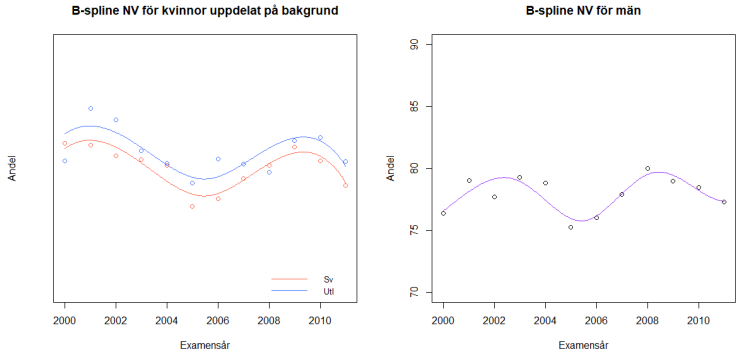
Figur A.5: Skattad tidstrend för IB uppdelad på personer med svensk och utländsk bakgrund.



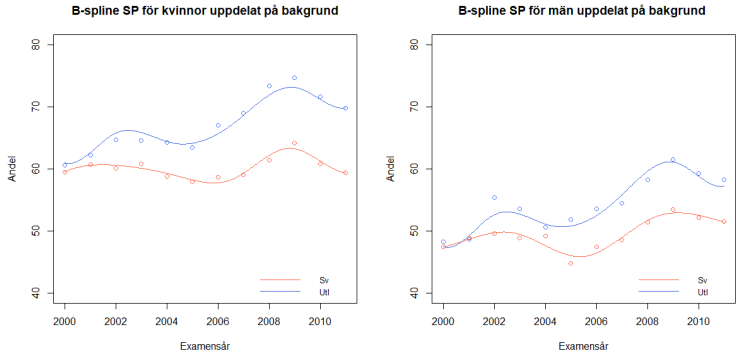
Figur A.6: Skattade tidstrender för personer med svensk och utländsk bakgrund inom SP.



Figur A.7: Skattade tidstrender för personer med svensk och utländsk bakgrund inom ES uppdelade på kön.



Figur A.8: Skattade tidstrender för kvinnor med svensk och utländsk bakgrund samt män inom NV.



Figur A.9: Skattade tidstrender för kvinnor och män med svensk och utländsk bakgrund inom SP.

Referenser

[1] Agresti, Alan (2013): *Cathegorical Data Analysis*, 3rd ed., John Wiley & Sons.

[2] Fahrmeir, Ludwig och Kneib, Thomas och Lang, Stefan och Marx, Brian (2013): *Regression: Models, Methods and Applications, Kapitel 5 och 8*, Berlin, Heidelberg: Springer Berlin Heidelberg.

[3] Foss, Ewa (1996): *Kvinnor och män i högskolan. Från gymnasium till forskarutbildning, Kapitel Övergångsfrekvens inom tre år*, SCB, i samarbete med Högskoleverket.
`http://www.uka.se/download/18.12f25798156a345894e29ab/1487841906405/9613R.pdf`

[4] Inkinen, Magdalena och Amft, Andrea och Callerholm, Carin och Furusten, Thomas och Gillström, Per och Lindqvist, Torbjörn och Westman, Per (2013): *Universitet och högskolor: Årsrapport 2013, Kapitel Utbildning på grundnivå och avancerad nivå*, UKÄ.
`http://www.uka.se/download/18.12f25798156a345894e4be7/1487841872014/Arsrapport-2013.pdf`

[5] Inkinen, Magdalena och Amnéus, Ingeborg (2016): *Kvinnor och män i högskolan, Kapitel Utbildning på grundnivå och avancerad nivå*, UKÄ.
`http://www.uka.se/download/18.5d85793915901d205f9a852/1487841854015/rapport-2016-10-14-kvinnor-och-man-i-hogskolan.pdf`

[6] Karlsson, Niklas (2016): *Nybörjare i högskolan 2015/16: Färre svenska högskolenybörjare, men fler inresande högskolenybörjare*, UKÄ.
`http://www.uka.se/download/18.5d85793915901d205f9a838/1487841856290/statistisk-analys-2016-12-06-nyborjare-i-hogskolan.pdf`

[7] Lindqvist, Torbjörn (2009): *Statistisk analys: Samband mellan konjunktur och högskolestudier?*, Högskoleverket.
`http://www.uka.se/download/18.12f25798156a345894e1f02/1487841916718/Statistisk-analys-2009-9.pdf`

[8] Maindonald, John och Braun, W. John (2010): *Data Analysis and Graphics Using R: An Example-Based Approach, Kapitel 7*, 3rd ed., Cambridge: Cambridge University Press.

[9] Racine, S. Jeffrey (2018): *A primer on Regression splines*, `https://cran.r-project.org/web/packages/crs/vignettes/spline_primer.pdf`.

[10] Svensson, Fredrik (2018): *Förkunskaper, social bakgrund och rekrytering till högskolan*, UKÄ.
`http://www.uka.se/download/18.2158bbb51621ecd5a963850/1522142679612/statistisk-analys-2018-03-27-forkunskaper-social-bakgrund-och-rekrytering-till-hogskolan.pdf`