Stockholms universitet

En statistisk analys av antalet SD-röster

Jan Mikael Yousif

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2017:1
Matematisk statistik
Juni 2017

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholms universitet

Matematisk statistik
Stockholms universitet
Kandidatuppsats **2017:1**
http://www.math.su.se/matstat

En statistisk analys av antalet SD-röster

Jan Mikael Yousif*

Juni 2017

Sammanfattning

I denna uppsats så försöker vi skapa en modell som kan prediktera andelen röster ett missnöjes parti bör får ifall beroende på ett antal parametrar i samhället. Missnöjespartiet som vi har fokus på i detta fall bestämdes till Sverigedemokraterna, Att vi specifikt väljer Sverigedemokraterna beror på tillgängligheten av svenska samhällsdata och författarens intresse av partiets snabba tillväxt de senaste åren. Vi hoppas på att kunna skapa en modell som predikterar framtida tillväxt för partiet. Till förhoppningarnas spillo så lyckades vi i bästa fall hitta modeller vars prediktiva förmåga var väldigt dåliga och därmed bestämde vi oss för att att det inte fanns något större värde i att göra någon prediktion om inte man vill föra det av rent intresse. Vi lyckades dock hitta intressanta (och i vissa fall oväntade) korrelationssamband mellan variablerna i undersökningen.

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

Statistisk analys av andelen röster SD får i ett riksdagsval

Jan Yousif

8 februari 2017

Contents

Sammanfattning

I denna uppsats så försöker vi skapa en modell som kan prediktera andelen röster ett missnöjes parti bör får ifall beroende på ett antal parametrar i samhället. Missnöjespartiet som vi har fokus på i detta fall bestämdes till Sverigedemokraterna, Att vi specifikt väljer Sverigedemokraterna beror på tillgängligheten av svenska samhällsdata och författarens intresse av partiets snabba tillväxt de senaste åren. Vi hoppas på att kunna skapa en modell som predikterar framtida tillväxt för partiet. Till förhoppningarnas spillo så lyckades vi i bästa fall hitta modeller vars prediktiva förmåga var väldigt dåliga och därmed bestämde vi oss för att att det inte fanns något större värde i att göra någon prediktion om inte man vill föra det av rent intresse. Vi lyckades dock hitta intressanta (och i vissa fall oväntade) korrelationssamband mellan variablerna i undersökningen.

Abstract

The purpose of this report is to try and create a model to predict the growth of political parties of discontent. The specific party which was chosen in this report is the swedish party called “Sverigedemokraterna”, they were chosen because of the avalability of data regarding swedish societies and because of the writers interest of the growth of this perticular party these past years. Initially one hoped to be able to create a model with good predicting abilities but one was not able to do that. Instead the writer found some interesting correlations between the chosen variables in the report.

1 - Introduktion

Demokratiska partier som tjänar sina röster på missnöje är ingenting nytt i bland samhällen runtomkring i världen. Dessa typer av partier är väldigt får väldigt ofta stöd av invånarna i samhället då befolkningen inte är nöjda med arbetet som nuvarande lagstiftare gör. I denna uppsats kommer SD:s (Sverigedemokraterna) följare att analyseras med hjälp av ett antal variabler som bland annat representerar rubriker som tenderar att vara argument till partiets fördelar och vanliga stereotyper angående partiets sympatisörer. Syftet med denna uppsats är att med hjälp av de valda variablerna kunna prediktera antalet/andelen röster som SD kommer att få i riksdagsvalet år 2018.

I vardagligt tal så kan det vara väldigt lätt att fördöma de som röstar på SD till rasister (eller alternativt främlingsfientliga individer), men så behöver det inte alltid vara. En av många anledningar till att folk röstar på SD kan vara att befolkningen vara trötta på de befintliga partiernas lagstiftning och därför försöka markera att man vill ha en förändring, därav missnöjesparti. I denna uppsats så är författarens egna politiska åsikter och personliga relationer mot SD och dess väljare helt neutraliserade, syftet med denna uppsats är inte att fördöma eller dela in människor i fack. Vi vill helt enkelt av nyfikenhet se ifall vi kan hitta ett statistiskt samband mellan särskilda variabler i samhället och de drastiska ökningarna av SD röster.

Man kan diskutera vare sig SD har goda eller onda avsikter bakom sin politik och framför allt sin ställning mot invandring, och detta är egentligen helt oviktigt till denna rapport med tanke på att vi bara bryr oss om ifall vi kan hitta mönster bland variabler som exempelvis beskriver socioekonomiska indelningar eller mängden brottslighet i ett svenskt område. Men någonting som vi inte helt och hållet kan neka är SD:s hårda budskap mot den svenska immigrationspolitiken, och om det verkligen visar sig att SD enbart har en enda huvudfråga i sin politik (invandringen) så tror vi att man ska kunna ta reda på anledningarna till varför deras röster blir mer populära. Ett parti har som skyldighet att leva upp till dess följares önskemål och på så vis så kanske det inte finns något rätt eller fel inom politik utan istället så får man kanske lägga moraliska frågor på de individer som är med och röstar på hur riket ska föra sin politik. Men om ett enfrågeparti vinner majoriteten av folkets röster så kanske inte rikets framtid är den ljusaste. Av denna anledning så vill vi i denna uppsats undersöka vad som påverkar rösterna för SD.

I denna undersökning så kommer vi att bygga upp alla modeller för prediktion via logistisk regression på binomialfördelad data. Vi har valt att utföra 4 olika procedurer för variabelselektion för att få en utvidgad vision över vilka variabler som kan tänkas finnas i den bästa möjliga modellen. De selektionsmetoder som kommer att utföras är Meningsfull Selektion, korrelationsanalys, Framåt Selektion och Bakåt Eliminering. Innan någon analys påbörjas så kommer de två första kapitel i denna rapport att beskriva metoder, teoretiska element och variabler som vi har använt i undersökningen. Efter att kapitlet som undersökningen sker i så ämnar vi utrymme för diskussionsmoment angående resultaten vi har kommit fram till.

2 - Beskrivning av Data

2.1 - Inledande Ord

Eftersom att SD anses vara ett missnöjes parti så är även de förklarande variablerna för de kommande regressionsmodellerna valda efter vad som vi tror kan medföra missnöje runtom i landet. Missnöje behöver inte betyda att en individ har negativa syner mot specifikt invandring, missnöje är (åtminstone i denna uppsats) ett samlings ord för all typ av oro, besvär eller brist av tillit. Vi förutsätter alltså i denna rapport att en sjuk individ som känner att staten inte ger det stöd som individen behöver så kan denna individ på grund av sin situation vara benägen att rösta på SD.

Detta är självklart ett helt hypotetiskt exempel och därmed antyder vi inte att en person utan tvekan röstar på SD bara den blir sjuk eller hamnar i en olycklig situation, vi säger enbart att variablerna är valda under förutsättningen att ett missnöje som berör ens hälsa, trygghet, livskvalité etc. Kan vara en anledning till att en individ vill rösta på SD oavsett hur rimligt eller orimligt personens missnöje kan vara.

För att kunna få en bild över hur det ser ut över hela Sverige så är allt datamaterial baserad på Sveriges alla kommuner. Utöver detta så är även alla förklarande variabler dividerade med antalet invånare per kommun. Detta är gjort så att vi ska få en bättre bild av värdet av datamaterial per person. Undviker man detta så är det ganska lätt att man hittar mönster i sina modeller som inte existerar, exempelvis så är antalet invånare i Stockholms kommun mycket större än antalet invånare i Kungsör kommun och därmed är antalet brott som begås årligen större i Stockholm än i Kungsör. Men detta behöver inte nödvändigtvis betyda att Stockholm är farligare att leva i än i Kungsör. Undersöker vi istället andelen brott per person så kan vi kanske bättre avgöra vilken kommun som är säkrast att leva i.

Eftersom att datafilerna kommer från olika källor så har vi haft problem med att varje datafil har annorlunda ordning, namngivning och presentation av data. Därför har vi valt att formatera all data i R. Ett separat ark med kommenterad kod finns tillgänglig för läsning till den som undrar vilka modifieringar som har gjorts med datapunkterna.

2.2 Tabell över data

Variabel	Variabel namn	Variabeltyp	antagna värden	Värdetyp	Saknade	källa
Antal röster	Votes	Responsvariabel	$[0,\infty]$	Antal	0	[5]
Antal Icke-röster	Nonvotes	Responsvariabel	$[0,\infty]$	Antal	0	[5]
Våldsbrott	Assault	Förklarande variabel	[0,1]	Andel	0	[6]
Sexbrott	Sex	Förklarande variabel	[0,1]	Andel	0	[6]
Bachelorstudenter	Bachelor	Förklarande variabel	[0,1]	Andel	0	[5]
Mastersstudenter	Master	Förklarande variabel	[0,1]	Andel	0	[5]
Biståndstagare	Social	Förklarande variabel	[0,1]	Andel	12	[7]
Arbetslösa	Unem	Förklarande variabel	[0,1]	Andel	12	[7]
Utrikes födda	Immigrants	Förklarande variabel	[0,1]	Andel	0	[5]
Befolkning	bef	Förklarande variabel	[0,1]	Andel	0	[5]

2.3 - Antal röster SD fick år 2014

Datamaterialet avser antalet röster som SD fick år 2014 och sammanställt av SCB (Statistiska central byrån) där antalet röster per kommun ordnas efter kommunkoder i växande ordning. All data har blivit anpassad efter detta. Son kompletterande variabel till antalet röster så har vi framställt antalet icke-röster genom att ta fram en vektor som innehåller antalet röstberättigade personer per kommun och därefter subtrahera

antalet SD-röster från denna vektor. Anledningen till varför vi ville använde oss utav antalet röstberättigade människor och inte bara totala antalet röster i hela valet är för att vi vill även räkna med de människor som inte har röstat alls, vilket enligt SCB bestod utav 14.2 % av den svenska röstberättigade befolkningen år 2014. Då vi inte tycker att detta är en försumbar siffra så bestämde vi oss för att de med de personer som inte röstar alls i datamaterialet.

Antalet

2.4 - Antal våldsrelaterade och sexuellt relaterade brott per person

Inledningsvis var det tänkt att ordna data efter totala andelen brott som årligen sker per person men sedan så funderade vi ifall sexuellt relaterade brott och våldsrelaterade brott kanske kan ha annorlunda effekter på modellen, så vi bestämde oss för att dela upp variabeln för brottslighet i två separata variabler, en för våldsrelaterade brott och en för sexuellt relaterade brott. Data är sammanställt av BRÅ:s brottsregister där varje kolumn i datamaterialet avser antalet brott per kommun under år 2013.

2.5 - Andel universitetsutbildade personer indelat i mastersstudenter och Bachelor studenter

En vanlig stereotyp väljarna till missnöjespartier är att väljarna tenderar att vara outbildade eller okunniga, vi är intresserade av att undersöka ifall det finns någon grund till denna stereotyp eller ifall det bara är nonsens som icke-väljare står vid för att distansera sig ifrån partiet och dess väljare. Därför har vi valt att välja utbildning som en förklarande variabel. Data sammansatt av SCB och består utav andelen utbildade invånare per kommun på master- och bachelornivå i form av två separata förklarande variabler.

2.6 - Andel biståndstagare

Biståndstagardata är sammansatt av Socialstyrelsens databas och representerar antalet personer per kommun som tar emot statliga bistånd. Inledningsvis så planerade vi inte att ta med denna kategori som förklarande variabel då statliga bistånd inte alltid handlar om att flyktingstöd eller liknande. Men vi kom att inse att all form av biståndstagande kan relateras till ett missnöje, så vi bestämde oss för att ta med denna data trots allt.

I datafilen som hämtades ifrån socialstyrelsens databas så fattades det 12 datapunkter. Dessa punkter ersattes med hjälp av medelvärdesimputering och vi har tagit detta som en möjlighet till någorlunda felaktiga resultat. Men vi förväntar oss inga signifikanta förändringar då datapunkterna som fattades tillhörde kommuner med små förändringar i antal biståndstagare per år.

2.7 - Andel Arbetslösa

Variabeln för andelen arbetslösa per kommun representerar andelen individer som tar emot monetärt stöd ifrån A-kassan skulle kunna ge en bra representation på hur många arbetslösa varje kommun har. Dock så går vi miste om några värden då man måste aktivt söka jobb i 3 månader innan man kan få stöd ifrån A-kassa. Detta innebär att någon som har blivit arbetslös i exempelvis november 2016 och söker bistånd ifrån A-kassa så blir personen en del av data ifrån 2017

Men med tanke på att ingen större nationalekonomisk händelse har skett i Sverige under perioden som datamaterialet avser så förväntar vi inte oss inte att datamaterialet går miste om väldigt många individer p.g.a. 90-dagars regeln.

Data är sammanställd av socialstyrelsen

2.8 - Andel Invandrare

Antal utrikes födda invånare per kommun är sammanställt av SCB och representerar andelen invånare med invandrarbakgrund per kommun. Data tar inte hänsyn till specifika ursprungen av dessa individer, alltså är exempelvis finska emigranter en del av data.

2.9 - befolkningsmändgen

För att omvandla all data till andelar så har vi även tagit med data ifrån SCB som talar om hur stor befolkningsmängden är per kommun.

3 - Teori

3.1 - Multipel Logistisk regression

När responsvariabeln i datamaterialet endast kan anta två värden (1 eller 0) så kan man använda logistisk regression för att ta fram modeller som predikterar sannolikheten att ett utfall blir lyckat. Eftersom att binomial logistisk regression förutsätter att modellens slumpmässiga variabel (responsvariabel) är binomialfördelad medan de förklarande variablerna kan vara både kontinuerliga eller kategorivariabler så blir detta ett perfekt verktyg för våran analys. Med tanke på att en individ endast kan rösta en gång så kan vi tolka responsvariabeln som 1 då en person röstar på SD eller 0 ifall personen avstår från att rösta.

Men innan vi går vidare med att definiera den logistiska regressionsmodellen så behöver vi först introducera uttrycket odds. Odds är en term som förklarar sannolikheten för att en en händelse ska vara ett lyckat utfall i förhållande med att händelsen ska vara ett misslyckat utfall. Man räknar ut oddset av en händelse genom $O = \frac{p}{1-p}$ och tolkas på så vis att alla värden av oddset då $O > 1$ säger att det är troligare att händelsen sker än att den inte sker och vice versa då oddset $O < 1$. Ju mer värdet av oddset avviker ifrån siffran 1 desto mer innebär det att händelsen troligen kommer ske eller inte ske (beroende på om oddset blir större än 1 eller mindre än 1).

Vi låter responsvariabeln Y_i vara en binär vektor vars innehåll representerar misslyckade eller misslyckade försök per observation. Låt dessutom π_i vara sannolikheten att en observation är ett lyckat försök. Då får säger vi att modellen för logistisk regression betecknas av

$$logit(\pi_i) = log(\frac{\pi_i}{1-\pi_i}) = \alpha + \sum_{j=1}^{n} \beta_j x_i j$$

,

Genom att slå ut logfunktionen i vänsterledet via en exponentialfunktionen så kan vi även visa att oddsfunktionen för π_i fås genom att ta exponentialfunktionen av regressionsmodellen.

$$log(\frac{\pi_i}{1-\pi_i}) = \alpha + \sum_{j=1}^{n} \beta_j x_i j \;\; <=> \;\; \frac{\pi_i}{1-\pi_i} = e^{\;\alpha + \sum_{j=1}^{n} \beta_j x_i j}$$

,

Via denna funktion så kan vi så kan vi alltså skriva om logit funktionen som sannolikheten att en given händelse är ett lyckat utfall. Vi får alltså

$$P(Y = 1|X) = \frac{e^{\;\alpha \; + \; \sum_{i=1}^{n} \beta_i \cdot x_i}}{1 + e^{\;\alpha \; + \; \sum_{i=1}^{n} \beta_i \cdot x_i}}$$

.

En viktig sak att nämna är att en logistisk regressionsmodell förutsätter att logit funktionen har ett linjärt samband med var och en av variablerna i modellen.

3.2 - Stegvis Variabelselektion

I en linjär modell så kan det ofta vara besvärande att bestämma vilka variabler man ska använda och till denna svårighet så kan det väldigt ofta vara ännu besvärligare att bestämma vilka variabler som ska vara kvar i modellen. I detta arbete så kommer vi att använda sig utav 4 huvudsakliga metoder för att avgöra vilka variabler som ska vara kvar i modellen. Dessa metoder kommer att vara

- Framåt Selektion
- Bakåt Eliminering
- Korrelationsanalys
- Meningsfull Selektion

3.2.1 - Framåt Selektion

När man utför en Framåt selektion för att avgöra vilken modell man ska använda så startar man med en modell utan variabler. För varje steg i den stegvisa proceduren så läggs en förklarande variabel till i modellen där man kontrollerar att variabeln är signifikant via Wald-statistikan med nollhypotesen nollhypotesen $H_0 : \ \beta_i = 0$. Det är av värde att nämna att det inte alltid behöver vara ett wald test som används vid prövning av hypotesen (exempelvis så använder man ofta T-statistikan vid prövning av vanliga linjära modeller) men vi nämnde specifikt detta test för att den kommer att användas i denna analys. [2, s.71]

3.2.2 - Bakåt Eliminering

till motsatsen av Framåt selektion så inleder man vid en Bakåt Eliminering istället med en modell som innehåller alla möjliga förklarande variabler. För varje steg i proceduren så prövar man nu de redan befintliga variablerna med ett Wald-test och kontrollerar att nollhypotesen $H_0 : \ \beta_i = 0$ stämmer. Ifall den ej stämmer så f;rkastas den testade variabeln, detta steg upprepas fram tills att det saknas insignifikanta variabler i modellen. [2, s.71]

3.3 - Meningsfull Selektion

Meningsfull selektion är en metod för variabelselektion som går baseras på 7 huvudsteg. Denna procedur är mer eller mindre anpassad för logistisk regression då ett utav stegen går ut på att kontrollera antaganden som specifikt gäller för logistisk regression. Man kan förmodligen ändra detta steg så att det passar ens egen regressionsmodell men då vi jobbar med logistisk regression så finns det ingen anledning att göra några justeringar.

Steg 1

Man inleder med att testa var och en av variablerna i ett likelihood ratio test. Vi sätter signifikansnivån på 0.25 för att undvika att förkasta variabler variabler som visar sig vara insignifikanta i en enkel modell medan de är signifikanta i en multivariat modell.

Steg 2

Skapa nu en multivariabel logistisk modell med de signifikanta variablerna ifrån steg 1 och exkludera de variabler som är insignifikanta under wald statistikan och en ytterligare modell utan att exkludera någon variabel. Testa därefter den mindre modellen mot den större i ett likelihood-ratio test för att avgöra vilken som passar data bäst.

Steg 3

Vi jämför nu lutningskoefficienterna β_i för respektive modell och söker efter de koefficienter som har förändrats märkvärdigt i jämförelse med den andra modellen och återlägg alla variabler vars lutningskoefficienter inte

har förändrats för mycket. Vi anser att en koefficient har förändrats för mycket då

$$\Delta\beta_i = \frac{\theta_i - \beta_i}{\beta_i} > 0.2$$

. Med andra ord så vill vi inte återlägga någon variabel vars lutningskoefficient har förändrats mer än 20%. Om inte koefficienten har gjort en förändring större än 20% så återläggs den till modellen.

Steg 4

I steg 4 så upprepar man steg 1 med de exkluderade variablerna där man prövar signifikansen av variablerna genom Wald statistikans p-värde.

Steg 5

Kontrollera att modellantagandet om linearitet mellan variablerna och logit funktionen som en funktion av modellens koefficienter. Detta kan göras genom att dela upp data i 4 kategorier som är baserade på datamaterialets kvartiler. Linearitеten kan kontrolleras via en plot.

Steg 6

Vi kontrollerar nu ifall det råder tvåvägssamspelseffekter mellan variablerna i den valda modellen. Ifall det råder samspel mellan två eller fler variabler så krävs det att samspelseffekten ska vara statistiskt signifikant för att få vara en del av modellen. Signifikansen testas med ett likelihood ratio test på signifikansnivån $\alpha = 0.05$.

Steg 7

I det sista steget så prövar vi modellens prediktiva förmåga. Detta kan göra med hjälp av ett Hosmer-Lemeshow test , AUC värden för ROC figurer och framförallt korsvalidering.

3.3 - Korrelationsanalys

I samband med alla dessa selektionsmetoder så har vi även valt att undersöka direkta korrelationssamband mellan förklarande variabler i syftet att finna kolineära samband mellan variablerna. Man kan göra denna analys genom att plotta variablerna mot varandra men vi föredrar att göra detta genom att skapa en korrelationsmatris och sedan jämföra vilka variabler som korrelerar med varandra. Detta är nödvändigt då man bör undersöka modellernas trovärdigheter gällande kolinearitet bland förklarande variabler. Vid fallet att det råder kolinearitet mellan två eller fler variabler så kan variationen för lutningskoefficienterna att visa sig vara större än vid utan kolinearitet vilket gör att modellen kan bli känslig mot små ändringar. Genom att exkludera variabler som korrelerar med varandra så kan man undvika problem som multikolinearitet.

3.4 - Akaike informationskriterium

Akaike informations kriterium (AIC - engelsk förkortning) är en konstant som används när man vill bedöma hur bra en modell passar det givna datamaterialet. AIC definieras via:

$$AIC = 2n - 2\log(\hat{\theta}_{ML})$$

,

där n är antalet variabler och $\hat{\theta}_{ML}$ den maximerade likelihoodskattningen av modellen. [4, s.224]

För övrigt så är AIC ett mått som saknar specifika parametrar som mäter ifall modellen passar den givna matamaterial bra eller ej, AIC anpassar sig efter den skapade modellen med avseende på antalet variabler och värdet av ML-skattningarna och tar fram ett värde som ska jämföras med andra modeller av samma typ. Generellt så vill man välja den modell med lägst AIC värde då tenderar AIC även att bestraffa överanpassning. Bestraffningen sker via uttrycket $2n$ i formeln för AIC. Då n är antalet variabler som finns med i modellen så innebär det att varje gång man lägger till en variabel i modellen så ökar alltså AIC värdet med 2.

3.5 - Tester för modellval och prediktionsförmåga

3.5.1 Variations Inflationsfaktor

Syftet med korrelationsanalysen som görs i denna analys är att undersöka förekomsten av kolinearitet bland variablerna i modellerna och därmed skapa en modell med minst kolineära samband mellan de förklarande variablerna. Utöver att enbart läsa av korrelationsvärden mellan förklarande variabler så är VIF ett bra mått för att avgöra hur mycket variansen av en variabel påverkas pågrund av kolinearitet, fördelen med detta mått är att man kan för varje given modell kontrollera att ingen variabel har ett starkt kolineärt samband med någon annan. VIF räknas ut via:

$$VIF = \frac{1}{1 - R_i^2}$$

där R_i^2 är den förklaringsgrad som anger hur mycket av variationen i X_i kan förklaras av de andra variablerna i modellen. [2, s.73]

3.5.2 - Receiver operating characteristic

En metod som man kan använda sig utav för att få mäta hur väl en modell passar data utan att behöva köra ett flertal simuleringar är ROC kurvan (engelsk förkorning) ett bra alternativ. Det man gör när man sätter upp en ROC kurva är att man försöker prediktera datamaterialet som modellen baserar sig på och därefter räkna ut sensitiviteten och specificiteten för varje given brytpunkt. Sensitivitet definierar man som sannolikheten att den predikterade händelsen blir ett lyckat utfall givet att den sanna händelsen är ett lyckat utfall medan specificitet definieras som sannolikheten att en predikterad händelse är ett misslyckat utfall givet att den sanna händelsen också är ett misslyckat utfall. När man har räknat ut sensitiviteten och specificiteten så gäller det att man plottar de mot varandra och drar en rak linje från punkt [0,0] till punkt [0,1]. Kurvan kommer att ha ett konvext utseende. Arean av den yta som ligger mellan ROC kurvan och den raka linjen (Förkortas AUC - Area under curve) används därefter som en statistiska för att bedömma modellens prediktionsförmåga. Värden i tabell 2.1 beskriver vilka hurvida AUC värdet tolkar modellens prediktionsförmåga. [3, s. 149]

Tolkning	AUC
Dålig	$[0.5, 0.7]$
acceptabel	$[0.7, 0.8]$
utmärkt	$[0.8, 0.9]$
perfekt	$[0.9, 1]$

Ett problem vi ser med ROC kurvan är att den förklarar modellens prediktionsförmåga utifrån hurvida modellen predikterar den data som modellen är skapad på. Så som vi ser så tar inte ROC kurvan hänsyn till överanpassning, en nästan perfekt anpassning av alla datapunkter leder till att AUC värdet blir högre men bara för att en modell passar den data den är baserad på utan (eller med väldigt små) felmarginal så vet man egentligen inte vad modellen säger om ny data som den aldrig har bemött förr.

3.5.3 - Hosmer-Lemeshow Test

Ett Hosmer-Lemeshow test går ut på att testa hurvida en modell är bra anpassad till den data den baseras på. Testet görs genom att man delar upp data i g antal grupper (alltså g lika stora percentiler). Exempelvis om g = 10 så blir grupp 1 den grupp med de minsta percentilerna, grupp 2 med de näst minsta etc. Tanken med uppdelningen är att för varje percentil så så räknar testet ut medelprediktionsvärdet av sannolikheterna att man får ett lyckat och ett misslyckat fall. Därefter så skapar man en teststatistiska **K** enligt formeln

$$K = \sum_{i=1}^{g} \frac{(O_{1i} - E_{1i})^2}{E_{1i}} + \frac{(O_{0i} - E_{0i})^2}{E_{0i}}$$

,

Där O_{1i} & O_{0i} är de observerade fall då responsvariabeln $Y = 1$ respektive $Y = 0$ och E_{1i} & E_{0i} är de förväntade fallen då responsvariabeln $Y = 1$ respektive $Y = 0$. Denna statistika är chitvå fördelad med frihetsgrad $g - 2$ (alltså $K \in \chi^2(g-2)$) och testas alltså av ett chitvå test under nollhypotesen H_0 : "*Modellen anpassar data väl*"

Vi måste dock påpeka att ett högt P-värde inte behöver betyda att modellen passar data bättre än en annan modell med lägre p-värde. Med ett Hosmer-Lemeshow test kan man endast testa ifall modellen passar data väl eller inte. På så vis så bör man nog inte heller enbart utgå ifrån Hosmer-Lemeshow testet då det precis som ROC kurvan endast kontrollerar hur väl modellen passar det datamaterial den är skapad på. Visserligen så underlättar testet vid fallet att modellen inte kan prediktera det datamaterial den är baserad på men vid fallet att testet visar insignifikans så bör man använda sig av ytterligare metoder som kan mäta modellens prediktionskraft av ny data. [3, s.146]

3.5.6 - Lämna en ute-korsvalidering

Prediktionsförmågan för en modell anses vara bra eller dålig beroende på hur väl den predikterar data som den inte bygger på, man kan argumentera över över att en modell bör anses vara bra ifall den passar datamaterialet den är baserad på med små fel men om inte den kan prediktera data som ännu inte existerar så är den egentligen inte en bra modell. Då kan man allt undra hur man egentligen mäta en modells prediktionsförmåga ifall enda datamaterial man har är det som modellen bygger på. Tanken med korsvalidering är att man plockar bort en bit av det datamaterial man har och "låtsas" att det bortplockade datamaterialet är framtida data. Man skapar därefter en modell på det kvarstående datamaterial för att därefter försöka prediktera det borttagna datamaterialet.

Det finns många olika typer av korsvalidering men den typ som vi kommer att använda benämns på engelska som "leave one out - cross validation". Tanken med LOOCV är att man iterear genom datamaterialet och tar bort en rad data för varje iteration. I samma iteration så skapar man även en modell på det kvarstående datamaterielet för att därefter prediktera det datamaterial som man tog bort i början av iterationen. Därefter sparar man det predikterade värdet och jämför med det sanna värdet.

Men eftersom att våran modell är baserad på logistisk regrssion så kommer vi istället att prediktera fram sannolikheten att ett utfall blir lyckat. Så det vi kommer att göra istället för att jämföra prediktionsvärdet med det sanna värdet är att simulera data ifrån en binomialfördelning där varje simulerad händlse är baserad på de sannolikheter som räknas ut i korsvalideringen för att senare jämföra antalet lyckade/misslyckade utfall n antal gånger. Ifall jämförelsen visar att de simuerade värdena är godtyckligt nära de sanna värdena så accepterar vi modellens prediktionsförmåga.

Det är inte ovanligt att man skapar en statistiska av prediktionsfelets medelkvadratsumma $F = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}$ och kontrollerar värdet av den, i och med att medelkvadratsumman beräknar prediktionmedelsfelet i kvadrat så vill man alltså ha ett så lågt värde på F som möjligt. Men för vårat ändamål så nöjer vi oss istället simulera ny data som förklarat i föregående stycke då vi anser att det ger en bättre helhetsbild av antalet SD-röster.

4 - Analys

I detta kapitel så visar vi i närmre detalj på hur vi har använt oss utav selektionsmetoderna i kapitel 4 för att bygga upp modeller för pprediktion. Vi kommer att inleda med att bygga upp modeller via varje selektionsprocedur och därefter kommer man att testa modellernas prediktionsförmåga med Hosmer-Lemeshow goodness of fit test, AUC-värden för ROC-kurvor och framförallt korsvalidering. Målet är att bygga upp en modell med en tillräckligt godtycklig prediktionsförmåga för att i slutet av kapitlet prediktera antalet röster som SD kommer att få vid nästa riksdagsval.

4.2 - Konstruktion av modeller

Inledningsvis så började vi med att skapa en modell som inkluderade alla kovariater utan att ta hänsyn till outliers. En jämförelse av kovariaternas korrelationsnivåer, modellernas AIC värde, lutningskoefficienter samt p-värden av variabler visade ingen signifikant skillnad. I vissa fall handlade det om en skillnad i korrelationsnivåer på så lite som 2%. Så vi bestämde oss för att inte exkludera någon rad ifrån datamaterialet då vi inte fann något värde i att kasta iväg värdefull data.

4.2.1 - Korrelationsanalys

Vanligtvis kan brukar man utföra korrelationsanalyser genom att studera plottar och eventuella VIF-värden men vi föredrar att se numeriska korrelationsvärden och dra slutsatser i samband med plottar och VIF-värden istället för att enbart kika på plottar. Nödvändiga plottar kommer att hänvisas till appendix.

Tabell 4.1 - Matris med korrelationsvärden motsvarande varje variabel i modellen

Variabel	votes	nonvotes	assault	sex	bachelor	master	social	unem	immigrants
votes	1	-0.59	0.131	0.016	-0.46	-0.52	0.29	0.04	-0.08
nonvotes	-0.59	1	-0.17	-0.19	0.194	0.197	-0.12	0.08	-0.52
assault	0.08	0.05	1	0.31	-0.015	-0.05	0.032	-0.01	0.10
sex	0.40	0.33	0.31	1	0.14	0.02	-0.22	-0.16	0.41
bachelor	0.54	0.50	-0.015	0.14	1	0.90	-0.65	-0.39	-0.07
master	0.49	0.50	-0.05	0.02	0.90	1	-0.55	-0.34	-0.02
social	-0.46	-0.36	0.03	-0.22	-0.65	-0.55	1	0.44	0.07
unem	-0.30	-0.22	-0.01	-0.16	-0.39	-0.34	0.44	1	0.03
immigrants	0.14	0.14	0.10	0.41	-0.07	-0.02	0.07	0.03	1

Som vi kan se i Tabell 4.1 så finns det ingen större korrelationsnivå mellan mellan de förklarande variablerna förutom vid variablerna som motsvarar andelen studenter per kommun. Där ser vi en korrelationsnivå på 90 %. Resterade variabler korrelerar med mindre än 60 % (ifall vi bortser från bachelor mot social). Någonting som väckte intresse var antalet negativa korrelationsvärden som vi kan hitta bland variablerna för arbetslöshet och biståndstagande. Det var någorlunda förväntat att variablerna för utbildning skulle ha negativa värden då vi initialt trodde att utbildning skulle ha en dämpande eller negativ effekt till andelen biståndstagare och andelen arbetslösa, men någonting som helt gick emot våra förväntingar var de små korrelationsnivåer mellan arbetslöshet och SD-röster, vi förväntade oss verkligen ett större och framför allt positivt samband mellan SD-röster och arbetslöshet. Vi trodde även från första början att det skulle finnas ett signifikant missnöje med hur skattepengar distribueras till individer som inte jobbar eller individer som får någon form av

introduktionsbistånd vid emigration ifrån sitt hemland. Ett ytterligare resultat som inte stämde överens med förväntningarna var att andelen biståndstagare och arbetslösa även var negativt korrelerade med brottslighet. Vi förväntade oss inte en extremt hög korrelationsnivå mellan dessa variabler då vi gissade att den skulle hamna någonstans på intervallet $[0.15, 0.3]$ med tanke på att det inte finns samma typ av fattigdom i Sverige som i andra länder. Men negativa värden var helt oväntat. att andelen arbetslösa kan bero på att i Sverige så medför inte arbetslöshet samma typ av fattigdom som finns i andra länder med inferiöra stater. I Sverige (och andra länder i Skandinavien) så får man ett otroligt stöd vid fallet då att man har förlorat sitt jobb eller hamnat i någon annan olycksföljd situation därför är det inte sannolikt att en individ i en situation där hen behöver resultera sig till brott för att leva.

Vi byggde även upp modeller med alla variabler och undersökte modellens VIF-värden. Se tabell 4.2 för en fullständig utskrift av VIF-värden av variablerna i modellen. Genom att undersöka VIF-värden undersökning av VIF-värden i en modell med alla variabler gav dessutom resultaten

Tabell 4.2 - vif värden

variabel	VIF-värde
Assault	2.373640
Sex	1.204069
Bachelor	9.481792
Master	7.905667
Social	2.102070
unem	1.591979
Immigrants	1.841240

Bortsett ifrån kovariaterna för utbildning så verkar modellen inte lida av någon större kolinearitet mellan variablerna. Med tanke på att korrelationskoefficienterna var såpass låga för de flesta variablerna så fanns det initialt ingen anledning att misstänka en signifikant kolinearitet/multikolinearitet. vi gick vidare med att exkludera variabeln för bachelor och då kunde vi få ner alla vif-värden till under 2.

För att kontrollera ifall det råder samspel mellan variablerna i modellen så prövade vi även att sätta upp en modell med koeffecienter för samspel. Samspelskoeffecienten testas med wald statistikan precis som alla andra ariabler och exkluderas ifall $p > 0.05$. Modellen ifrån korrelationsanalysen visade att det råder samspel mellan assault/master, assault/immigrants och master/immigrants.

4.2.2 - Meningsfull Selektion

Steg 1

Vi inleder med att utföra ett likelihood-ratio test på varje enkel regression och kontrollerar att de är signifikanta på signifikansnivån $\alpha = 0.25$.

Tabell 4.3 - P-värden av alla variabler ifrån ett Likelihood-Ratio test

Variabel	P-värde
Assault	0.2572
Sex	$2.2 \cdot 10^{-16}$
Bachelor	$2.2 \cdot 10^{-16}$
Master	$2.2 \cdot 10^{-16}$
Social	$2.2 \cdot 10^{-16}$
Unem	$2.2 \cdot 10^{-16}$
Immigrants	$2.2 \cdot 10^{-16}$

Variabeln för våldsbrott visade sig vara insignifikant på signifikansnivån $\alpha = 0.25$ under ett likelihood ratio test så vi går alltså vidare till steg 2 i proceduren och skapar en en full modell och en modell utan denna variabel.

Steg 2

Tabell 4.4 - Utskrift av en multipel logistisk regressionsmodell utan variabeln för våldsrelaterade brott

Variabel	estimat	Avvikelse	P-värde
sex	-3.91866365	0.50252	0.186
bachelor	2.22406729	0.18058	$2 \cdot 10^{-16}$
Master	-6.06542889	0.06519	$2 \cdot 10^{-16}$
Social	4.50124196	0.27081	0.00207
Unem	-1.22024378	0.12096	$2 \cdot 10^{-16}$
Immigrants	0.06716663	0.05081	$6.29 \cdot 10^{-15}$

Tabell 4.5 - Utskrift av en multipel logistisk regressionsmodell utan insignifiaknta variabler

Variabel	estimat	Avvikelse	P-värde
Bachelor	1.9376636	8.191e-08	$2 \cdot 10^{-16}$
Master	-5.9346495	4.112e-02	$2 \cdot 10^{-16}$
Social	3.6347875	2.658e-01	$2 \cdot 10^{-16}$
Unem	-1.0970524	1.189e-01	$2 \cdot 10^{-16}$
Immigrants	0.1443933	5.157e-02	$3.78e-06$

vi ser att variabeln för sexuellt relaterade brott i tabell 4.4 är insignifikant under wald statistikan, så vi går vidare i detta steg och exkluderar den från modellen och gör jämför sedan den mindre modellen mot den större via ett likelihood-ratio test.

I ett likelihood-test där vi jämförde den mindre modellen mot den större modellen så visades insignifikanta resultat, vi går alltså vidare till steg 3 i proceduren under antagandet att den större modellen har en sämre prediktionsförmåga än den lilla.

Steg 3

variabel	modell 1	modell 2	procentuell förändring
Bachelor	1.9376636	2.22406729	13%
Master	-5.9346495	-6.06542889	2%
Social	3.6347875	4.50124196	19%
Unem	-1.0970524	-1.22024378	10%

en undersökning utav de procentuella skillnaderna mellan lutningskoefficienterna i de två modellerna visar att ingen utav förändringarna är större än den förbestämda signifikansnivån. Med en signifikansnivå på 20% så vi fortsätter till steg 4 i proceduren utan någon återläggning av förkastade variabler.

Steg 4

Återläggning av den enda LR-insignifikanta variabeln visade ett signifikant P-värde för alla variabler i modellen enligt Wald statistikan. Men återigen så finns det ingen större anledning att återlägga den i modellen då den fortfarande är insignifikant i ett LR-test och inte har någon större påverkan på lutningskoefficienterna för de resterande variablerna i modellen.

Vi går alltså vidare till steg 5 i proceduren med samma modell som vi kom fram till i steg 3.

Steg 5

Alla figurer till detta steg hänvisas till Figur 7.1 i appendix 7.2

Vi ser att ingen av plottarna har en perfekt linearitet, den som är mest ifrågasättbar utskriften för variabeln "Unem" där funktionen verkar ha någon tveksam böjning mellan första två punkterna. I övriga fall så ser man inte en fullständigt linjär tillväxt men vi anser inte att den avvikelse ifrån linearitеten är tillräckligt stor för att exkludera dessa variabler, så vi nöjer oss med att enbart exkludera variabeln för arbetslöshet. Vi går vidare till steg 5 i meningsfull selektion under antagandet att logit-funktionen för alla variabler förutom den för arbetslöshet uppfyller mer eller mindre ett linjärt samband med modellens koefficienter.

Steg 6

Med 5 stycken variabler i modellen så finns det alltså 10 möjligheter till signifikanta samspelseffekter mellan variablerna, därför inkluderar vi alla variabler för samspel i modellen och i syftet att undersöka vilka variabler som under wald statistikan kan samspela för en person som ska rösta i ett riksdagsval. Resultat av signifikanta (och insignifikanta) samspelseffeker presenteras i Tabell 4.6. Vid signifikans så presenteras ett ✓ och vid insignifikans så presenterasr ett ×.

Tabell 4.6 - Tabell med p-värden ifrån samspelseffekter mellan variablerna i modellen

variabler/p-värden	**Bachelor**	**Master**	**Social**	**Immigrants**	**unem**
Bachelor		×	×	×	×
Master			✓	✓	×
Social				✓	×
Immigrants					×

Vi fick alltså slutsligen fram en modell med variablerna för bachelorstudenter, mastersstudenter, biståndstagare, invandrare och arbetslöshet tillsammans med samspelsvariablerna som presenteras i tabell 4.6.

4.2.3 - Framåt Selektion & Bakåt Eliminering

Framåt Selektion & Bakåt Eliminering är de snabbaste variabelselektionsmetoderna av de 4 metoder som vi har använt oss utav i denna analys. Alla framställningar av modeller kommer här att hänvisas till appendix.

Framåt Selektion:

Vid utförandet av framåt selektion så kunde vi visa att variabeln för våldsrelaterade brott är insignifikant redan vid allra första steget av proceduren. Detta förväntades då variabeln även var insignifikant enligt ett LR test vid första steget i purposeful selection, så vi gick vidare i proceduren och exkluderade assault redan i steg 1. I de efterföljande stegen lyckades vi dessutom visa att variabeln för arbetslöshet också var insignifikant under Wald-statistikan. Alla andra variabler behöll ungefär samma P-värde igenom hela processen och behövde alltså inte exkluderas vid något steg. I tabell 4.7 så har vi samanställt varje steg i framåt selektionen. Inga P-värden framställs i tabellen, vid fallet att en variabel har exkluderats så har dens P-värde överskridit signifikansnivån $\alpha = 0.5$. vid fallet att en variabel visa sig vara insignifikant så kommer den att markeras med $\times$ och cellerna för den exkluderade variabeln ki kommande steg kommer att markeras med $-$, i annat fall så markeras den me ✓ och ifall variabeln ännu inte har introducerats i proceduren så kommer den betecknas av en tom cell.

Tabell 4.7 - En tabell som visar utfallen av varje steg i Framåt Selektion

Steg/Variabler	**Steg 1**	**Steg 2**	**Steg 3**	**Steg 4**	**Steg 5**	**Steg 6**	**Steg 7**	
Assault	×	−	−	−	−	−	−	−
Sex		✓	✓	✓	✓	✓	✓	✓
Bachelor			✓	✓	✓	✓	✓	✓
Master				✓	✓	✓	✓	✓
Social					✓	✓	✓	✓
unem							×	−
Immigrants								×

Vi utförde en prövning av samspelseffekter likt prövningen vi gjorde i kapitel 4.1. Resultaten gav oss en modell där samspelseffekter rådde mellan sex/master och sex/social. Ytterligare en undersökning av kolinearitet gav oss precis som vi korrelationsanalysen en modell där variablerna för utbildning har ett kolineärt samband, se Tabell 4.8. Detta ledde till att vi även exkluderade variebeln för bachelorstudenter ifrån modellen.

Tabell 4.8 - Vif värden av variabelrna inkluderade i modellen från framåt selektion

Variabel	Vif-värde
Sex	1.011114
bachelor	9.401003
master	7.574359
social	1.972192

Bakåt Eliminering

Bakåt Elimineringen gick lite snabbare att fullfölja än framåt selektionen. I denna procedur kunde vi enbart exkludera en variabel totalt vilket skedde i det första steget av selektionsmetoden. Denna variabel visade sig vara variabeln för sexuellt relaterade brott. Vi övervägde en stund att exkludera variabeln för våldsrelaterade brott då vi i framåt selektionen såg att den var insignifikant i en enkel logistisk modell, men vi bestämde oss för att ha kvar den ändå då den kanske kan bidra med någon stabiliserande effekt till modellen. Precis som vid famåt selektionen så vill vi även här skapa en tabell som beskriver varje delsteg i proceduren. Reglerna för hur vi markerar en vaiabel exkluderad ller inkluderad i Tabell 4.9 följer samma regler som Tabell 4.7 går efter.

Tabell 4.9 - En tabell som visar utfallen av varje steg i Bakåt Eliminering

Steg/Variabler	**Steg 1**	**Steg 2**
Assault	✓	✓
Sex	×	–
Bachelor	✓	✓
Master	✓	✓
Social	✓	✓
unem	✓	✓
Immigrants	✓	✓

Och precis som tidigare så prövade vi även modellen för samspel och fick fram att alla variabler förutom assault/master, Social/Unem och Assault/Sex hade ett signifikant samspel med varandra. Så dessa variabler fick inkluderas i modellen. Vi fick även i denna modell exkludera variabeln för bachelorstudenter p.g.a. Kolinearitet. Se vif värden i tabell 4.10

Tabell 4.10 - Vif-värden av variablerna inkluderade i modellen från Bakåt Eliminering

Variabel	Vif-värde
assault	1.056696
Bachelor	9.189875
Master	7.886336
social	1.768004
Immigrants	1.167187

Kapitel 5 - Resultat

Detta kapitel är en sammanfattning av analysen i kapitel 4. Här framställer vi resultaten av de tester, statistikor och andra analysmetoder i syftet att jämföra vilken modell som är den bästa. Efter presentation av dessa värden så är målet att vi ska kunna välja ut den bästa bästa modellen och därefter prediktera resultaten för SD under nästkommande riksdagsval. Det nya datamaterialet är sammanställt på exakt samma form som den som användes under konstruktionen av modellerna, den enda skillnaden är att det nya datamaterialet är sammanställd efter år 2014. Inga modell summaries kommer att framställas i detta kapitel då vi anser att det förstör läsligheten av kapitlet, vi hänvisar istället alla summaries och eventuella plottar till appendix.

5.1 - Tester och Värden

Nedan framställs ett hosmer-lemeshow test för goodness av fit på varje av de 4 modellerna som vi har framställt i kapitel 4, se tabell 5.1. Varje kolumn i Tabell 5.1 representerar en specifik modell medan raderna i tabellen representerar de förklarande variablernas lutningskoefficienter. Variabler som benämns "*Variabel1*:*Variabel2*" representerar variabeln för samspelseffekten mellan de två förklarande variablerna. Om en variabel inte finns med i modellen (p.g.a signifikans eller annan anledning) så kommer den att markeras med ett × tecken, vid Inkludering av variabeln så markeras variabeln med ✓. Se Summary 7.1 till Summary 7.4 i appendix 7.1 för exakta lutningskoefficienter och relevanta P-värden.

Tabell 5.1 - Tabell av De slutgiltiga modellerna från slektionsmetoderna i kapitel 4

Modell/koefficienter	**Korrelationsanalys**	**Meningsfull Selektion**	**FS**	**BE**
Assault	✓	×	×	✓
Sex	×	×	✓	×
Bachelor	×	✓	×	×
Master	✓	✓	✓	✓
Social	×	✓	✓	✓
Unem	×	×		×
Immigrants	✓	✓	×	✓
Samspelseffekter				
Assault:Sex	×	×	×	×
Assault:bachelor	×	×	×	×
Assault:master	✓	×	×	×
Assault:social	×	×	×	✓
Assault:unem	×	×	×	✓
Assault:immigrants	✓	×	×	✓
Sex:Bachelor	×	×	×	×
Sex:Master	×	×	✓	✓
Sex:social	×	×	✓	✓
sex:Unem	×	×	×	✓
bachelor:Master	×	×	×	×
Bachelor:social	×	×	×	×
Bachelor:Unem	×	×	×	×
Master:Social	×	✓	×	✓
Master:Unem	×	×	×	✓
Master:Immigrants	✓	✓	×	✓
Social:Unem	×	×	×	×
Social:Immigrants	×	✓	×	✓
Unem:Immigrants	×	×	×	✓

vi framställer även en tabell med nödvändiga värden och teststatistikor som användes vid modell-selektionerna. Dessa värden kommer nu att jämföras med varandra för att utse den bästa modellen mellan alla selektionsprocedurerna.

Tabell 5.2 - Tabell av jämförelsevärden för varje modell

Modell/Metod	**Korrelationsanalys**	**Meningsfull Selektion**	**FS**	**BE**
AIC	47116	47706	47803	46302
BIC	47141.62	47735.28	47825.17	46353.23
AUC	0.5810	0.5839	0.5809	0.5818
Hoslem	2.2e-16	2.2e-16	2.2e-16	2.2e-16
χ^2**-test**	2.2e-16	2.2e-16	2.2e-16	2.2e-16
CV-prediktionsfel	30%	26%	30%	27%

Som vi kan se så är det ingen av de valda modellerna som har en väldigt stark, eller för övrigt en bra prediktionsförmåga överhuvudtaget. Ifall vi ignorerar prediktionsfelen från korsvalideringen en stund och kikar på de andra testerna så ser vi att ROC areorna är inte av någon signifikans alls, i bästa fall så visar de att den givna modellen har en dålig prediktionsförmåga. Hosmer-Lemeshow testerna visar signifikanta resultat för alla tester vilket alltså förkastar nollhypotesen H_0 : " *Modellen passar data väl*". I praktiken så visade det sig att ingen av de slutgiltiga modellerna kunde visa någon bra prediktionsförmåga. Men för jämförelsens skull så går vi vidare med att säga att modellen som fås via Framåt Selektion är jämförelsevis den hittills sämsta modell som vi har kommit fram till. Modellens AIC värde är den modell med störst AIC och BIC värde samt så visar AUC är en av de minsta areorna bland modellerna. Generellt så ser vi att AUC värdet för modellen från meningsfull selektion är den bästa bland modellerna, den har dock inte lägst AIC värde men med tanke på att detta värde inte varierar så farligt mycket mellan modellerna om man bortser ifrån FS och BS modellerna så lägger man hellre vikt på aren under ROC kurvan istället för AIC värden.

Vi bekräftar nu att modellen från meningsfull selektion är den bästa genom att kika på prediktionsfelen för korsvalidering. Men någonting som är nämnvärt är att prediktionsfelen bör variera lite för varje gång som vi korsvaliderar modellen. Detta beror på att vi simulerar ny binomialfördelad data med avseende på de predikterade sannolikheterna, vilket innebär att varje ny simulering kommer variera lite i antalet lyckade/misslyckade utfall, men denna variation är inte någon signifikant variation så vi kan med gott samvete utgå ifrån att de värden som står i tabellen beskriver modellens prediktionsförmånga till en godtycklig nivå.

Kapitel 6 - Diskussion

Ursprungligen så hade vi hoppats på lite bättre resultat än dessa. Vi var medvetna om att vi inte nödvändigtvis skulle få en perfect fit eller för övrigt ens nära något perfect fit men vi förväntade oss åtminstone insignifikanta Hosmer-Lemeshow tester samt AUC värden sådana att $AUC \in [0.6, 0.8]$ vilket vore modeller med en hyfsat god eller godtycklig prediktionsförmåga. Men vi kunde redan vid korrelationsanalysen (Se tabell 4.1) se att ingen utav kovariaterna hade någon större korrelationsnivå med responsvariablerna och därmed kunde vi redan där ana att modellerna inte skulle bli särskilt bra. I bästa fall så hade kovariaterna för utbildning en korrelationsnivå på ca 50% med responsvariablerna.

Ett stort problem med analysen är att den data som modellerna byggs upp på inte är individdata. Då vi försöker modellera röster i ett riksdagsval så kan det vara svårt att få fram individuell information för var och en av ca 7.3 miljoner röstberättigade invånare. Därför har man i stort sett blivit tvungen till att generalisera individer till siffror som beskriver sannolikheten att en individ lever upp hamnar i en särskild kategori (sannolikheten att du är en brottsling, sannolikheten att du är utbildad etc.) istället för att ha exakt data på hur länge en individ har utbildat sig eller ifall hen har bidragit till brottsliga aktiviteter etc. Med andra ord så förklaras varje individ i en kommun med samma kovariat som alla andra personer i samma kommun vilket inte är ett rimligt antagande.

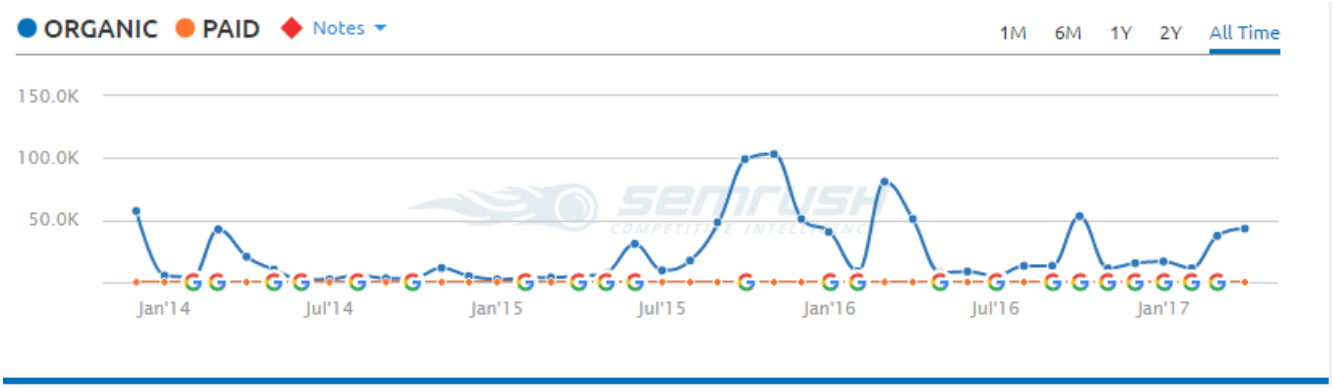
Med tanke på att lutningskoefficienterna i en logistisk regressionsmodell skattas via maximum-likelihood metoden så innebär det också att ML-skattningen utan tvekan kommer att visa annorlunda resultat med det datamaterial vi har än vid fallet då man har individdata. Att detta skulle ske visste man redan från första början men vi fortsatte med analysen med förhoppningar om att man kan generalisera individer i ett samhälle till siffror som som summerar samhället i helhet. Som vi kunde se i resultaten så hade vi fel. Efter analysen var färdig så fundrade vi ifall man kunde ha formulerat analysen på annorlunda sätt, vi hade kanske kunnat få bättre resultat ifall vi istället modellerade andelen röster per kommun i form av en vanlig multpel lnjär regression och på såvis försöka prediktera hur kommuner i helhet reagerar på de valda variablerna istället för att försöka mäta individer i form av logistisk regression.

Ett antagande som logistisk regression förutsätter är att log-oddset har ett linjärt samband mellan de enskilda variablerna i modellen, och som vi kunde se i steg 5 i kapitel 4.2 så var det långt ifrån alla variabler som visade det linjära sambandet som man vill ha. Ifall inte variablerna lever upp till detta antagande så kan man få problem med inkonsistenta estimat och avvikelser i skattningarna av deras standardavvikelser. Vill man tackla detta så kan man lösa problemet genom att transformera variabeln till en annan typ av funktion (exempelvis logaritmera, kvaderar etc.) och eventuellt inkludera samspelseffekter. Men trots försöket på att skapa de linjära samband man är ute efter så lämnades vi ändå med modeller som har en dåliog prediktionsförmåga, som man kunde se i Tabell 5.2 så fick vi fram modeller med ett prediktionsfel på upp till ca 33% enligt en LOO-korsvalidering. Detta gav oss egentligen inget annat val än att anta att den typ av data som användes i modelleringen inte kan generalisera svenska individer till vår datamaterial. I logistisk regression så förutsätter man även att man har oberoende observationer, men med tanke på att vi omvandlade i stort sett alla variabler till andelar innan någon analys gjordes så tror vi inte att detta antagande har blivit brutet.

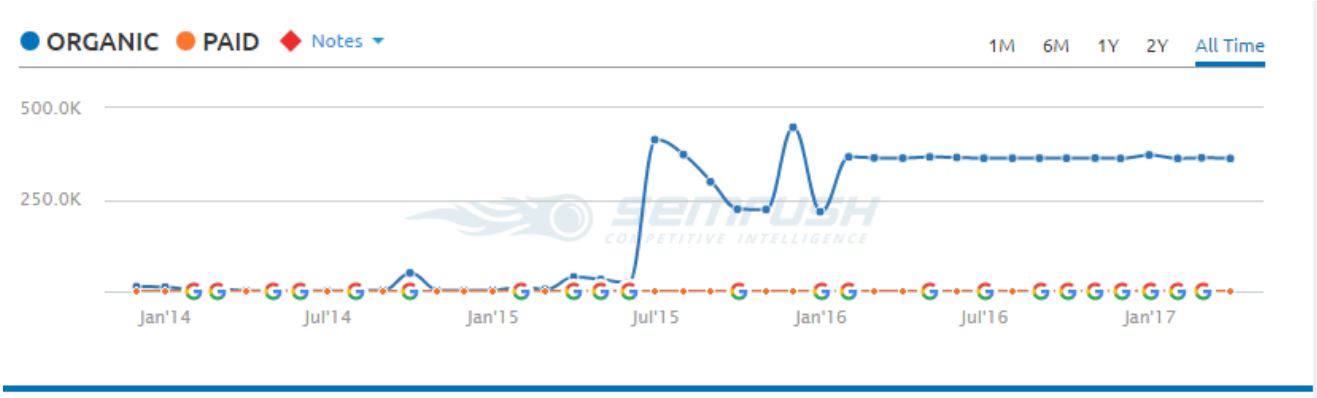
Vi hade mer än gärna velat göra en prediktion på nästa riksdagsval men som vi nämnde i kapitel 5 så finns det ingen större nytta i att göra någon prediktion då modellerna inte har någon pålitlig prediktionsförmåga. Hade man dock haft rätt typ av data och därmed eventuellt godtyckliga modeller så var planen från första början att sätta in nyare data (data sammanställd efter 2014) i den bästa modellen man har och därefter räkna ut sannolikheten för varje observation av att en individ röstar på SD enligt sannolikhetsfunktionen som beskrevs i kapitel 3.1. Efter att ha fått fram en vektor med sannolikheterna att en person röstar på SD så tänkte vi att simulera binomialfördelad data beroende på predikterade sannolikheterna, därefter kan man summera antalet lyckade utfall för att få en godtycklig prediktion av resultaten från nästkommande riksdagsval givet att modellen har en godtycklig preiktionsförmåga. Ett annat alternativ är också att simulera ett p-värde mellan $[0, 1]$ ifrån en likformig fördelninig där varje utfall blir 1 eller 0 beroende på om den simulerade refernspunkten är större eller mindre än den predikterade sannolikheten.

Generellt så tror vi fortfarande att man kan skapa en bra modell som predikterar SD-rösterna ifall man hade individdata. Detta håller vi oss till på grund av att SD gjorde ett otroligt hopp ifrån att inte ens vara ett

signifikant parti til att vara Sveriges trede största parti med stöd av ca 12.2% av befolkningen på bara två riksdagsval. Vi har svårt att tro att ett parti med den kontroversiella bakgrund som SD har kan samla på sig stödet från strax över en tiondel av av den svenska befolkningen utan någon anledning. Utöver detta så har man också kunnat se en ökning traffik på främlingsfientliga medier i högtid med valet 2014 (se Figurer 6.1 & 6.2) vilket vi tror kan vara en indikation på att det verkligen existerar ett missnöje på hur Sverige sköter sin invandringspolitik.



Figur 6.1 - En plot av antalet besökare på nyhetssidan www.friatider.se från 2014 fram till idag



Figur 6.2 - En plot av antalet besökare på nyhetssidan www.avpixlat.info från 2014 fram till idag

Någonting som vi dessutom var intresserade att undersöka var hur antalet SD-röster varierar bland svenska län och kommuner. Vi undrade ifall det kanske existerar några särskilda svenska områden där sympatin för Sverigedemokraernas politiska ideologi är lite större än andra områden. Vi plottade därför fram andelen röster per län och kommun (se Figur 6.3 och Figur 6.4) över en karta av Sverige för att visuellt analysera eventuella politiska trender i sverige och vi kunde se att i många fall så tillhörde de län/kommuner som röstade mest på SD den grupp av län/kommuner med minst invånare.

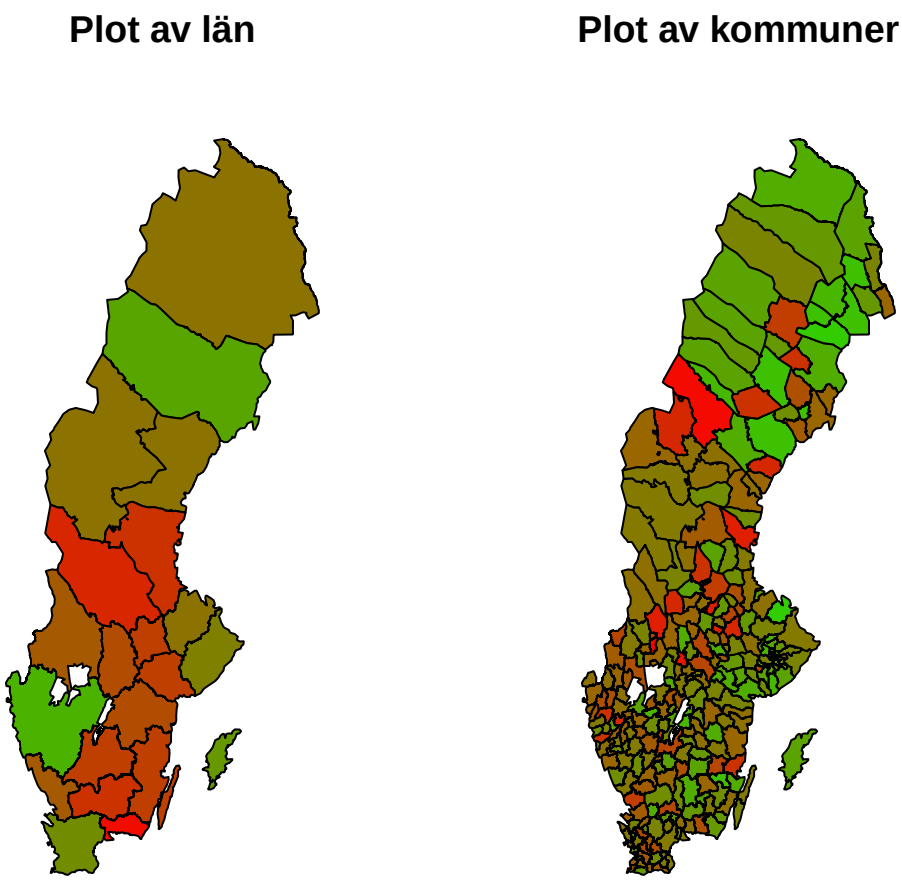


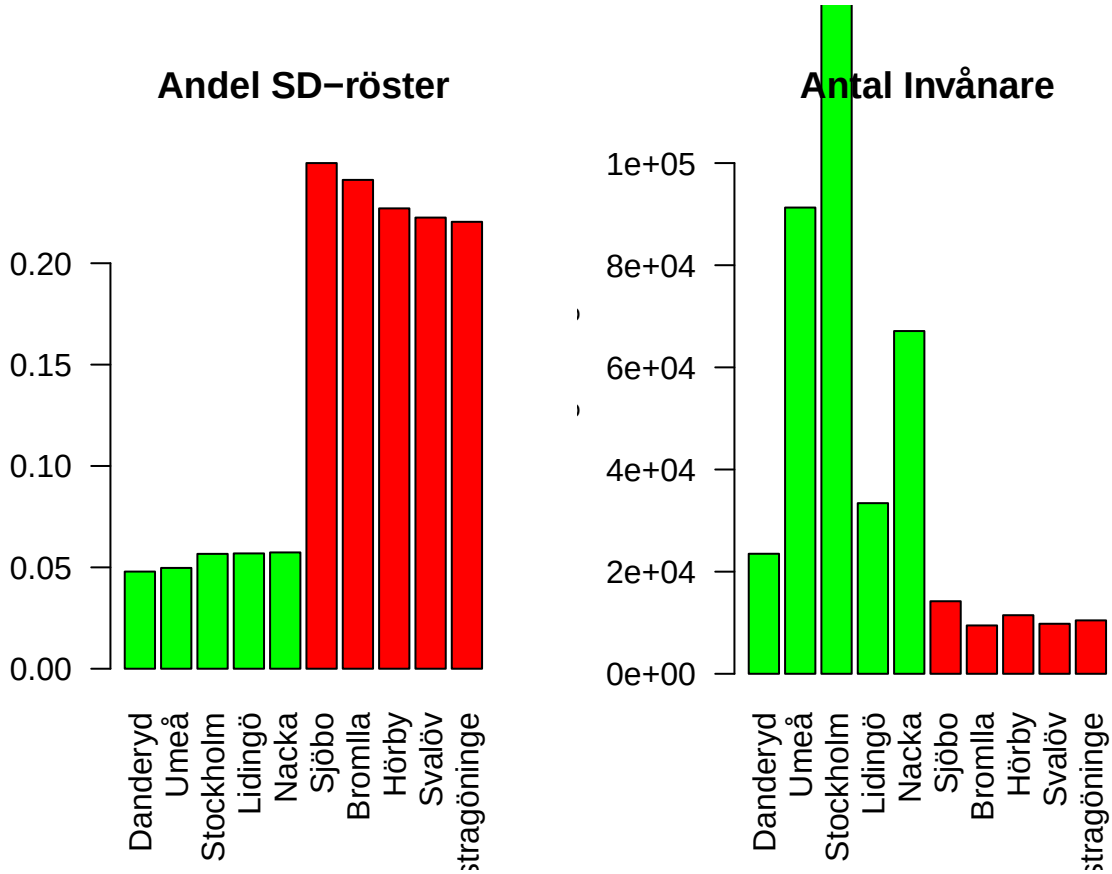
Figure 1: Figur 6.3 & Figur 6.4 - Plottar av andelar röster för varje län respektive kommun i sverige

Vi designade plotten sådan att ju grönare ett område är desto färre är andelen röster på det specifika området, därmed blir ett område mer och med rött då andelen röster ökar. För att resultatet ska vara mer lämpliga för presentation och för att man ska kunna se små ändringar mellan områdena så bestämde vi oss för att skära ner färgernas känslighet sådana att den rödaste färgen tillsattes till den kommun eller det län med störst anel SD röster och vice versa med de grönaste färgerna.

Generellt sett så ser vi att Stockholms län, Västra Götalands län, Skåne län, Gotlands län och Västerbottens län är de län vars andelar SD-röster ligger på intervallet $[0, 0.1]$. Resterande län verkar ligga på intervallet $[0, 0.2]$. Någonting märkvärdigt är att de län som står för majoriteten av Sveriges invånare (Stockholm, Västra Götaland och Skåne) tillsammans tillhör de län med minsta andel SD-röster medan de län med minst antal invånare bidrog med de största andelarna röster. När det gäller plotten för kommuner så var det dock lite svårare att se något mönster

vi har inte kommit fram till särskilt många anledningar till varför de mindre befolkade länen tenderar att rösta fram SD mer än de större länen men vi gissar på att man inte alltid får samma stora diversifiering av åsikter, diskussioner, protester och manifestationer i mindre samhällen som man kanske kan få i Stockholm. Vi säger inte nödvändigtvis att man är helt instängd ifrån diskussioner eller statlig fakta när man lever i mindre samhällen men man tror som sagt att det finns en möjlighet att man inte får samma variation av politiska diskussioner som vid storstäder. Och med tanke på att aktiviteten för främlingsfientliga medier har ökat enligt Figur 6.1 och Figur 6.2 så tror vi alltså att det finns en möjlighet att just dessa områden står för den största delen av ökningen. Dock så är detta påstående svårt att bevisa och bör då alltså inte tas på allvar.

För att förstora perspektivet lite till så prövade vi även med att jämföra andelen SD-röster mellan de största och de minsta kommunerna i Sverige och vad man såg var minst sagt intressant. Se Figur 6.5.



Figur 6.5 & Figur 6.6 - Plottar av andelen SD-röster per plottad kommun och invånare per plottad kommun

Som man kan se så tillhör de kommuner med störst andel SD-röster (ca 24%) nästan exklusivt en grupp

med ca 10 000 röstberättigade invånare per kommun medan de kommuner med minsta andel röstar varierar mellan 23 000 - 67 000 invånare (om vi bortser från Stockholm kommun som har 677144 röstberättigade invånare). Nog är detta inte ensamt tillräckligt för att påstå att funderingarna i föregående stycken är utan tvivel korrekta men det väcker en och annan frågeställning kring varför det kan vara sådana skillnader mellan stora och små kommuner.

Kapitel 7 - Appendix

7.1 - Referenser

[1] Alan Agresti - Categorical Data Analysis, second edition, 2013, Wiley

[2] Rolf Sundberg - Lineaära Statistiska Modeller, 2016, Stockholms Universitet.

[3] David Hosmer & Stanley Lemeshow - Applied Logistic Regression, Second Edition, 2000,

[4] Leonhard Held & Daniel Sabanés Bové - Applied Statistical Inference, Likelihood and Bayes, 2014, Springer.

[5] Statistiska Centralbyrån, www.scb.se

[6] BRÅ - http://statistik.bra.se/solwebb/action/index

[7] Socialstyrelsen - http://www.socialstyrelsen.se/statistik/statistikdatabas

[8] Inspektionen för arbetslöshetsförsäkringen - http://www.iaf.se/Statistik/Statistikdatabasen/

7.2 - Summaries

```
summary(glm(cbind(votes,nonvotes)~assault+master+immigrants+assault:master+assault:immigrants+master:im
```

```
##
## Call:
## glm(formula = cbind(votes, nonvotes) ~ assault + master + immigrants +
##     assault:master + assault:immigrants + master:immigrants,
##     family = "binomial")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -44.045   -8.021   -0.798    5.446   47.163
##
## Coefficients:
##                       Estimate Std. Error  z value Pr(>|z|)
## (Intercept)        -1.624e+00  1.195e-02 -135.859  < 2e-16 ***
## assault             2.511e+02  6.461e+00   38.866  < 2e-16 ***
## master             -4.486e+00  8.592e-02  -52.218  < 2e-16 ***
## immigrants          1.075e+00  1.572e-01    6.836 8.15e-12 ***
## assault:master     -1.195e+03  4.812e+01  -24.838  < 2e-16 ***
## assault:immigrants -1.649e+03  7.208e+01  -22.875  < 2e-16 ***
## master:immigrants   1.145e+01  9.868e-01   11.606  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 106741  on 289  degrees of freedom
## Residual deviance:  44440  on 283  degrees of freedom
## AIC: 47116
##
## Number of Fisher Scoring iterations: 4
```

```
# summary(glm(votes3~assault3+master3+immigrants3+assault3:master3+assault3:immigrants3+master3:immigra
```

Summary 7.1 - Summary av slutgiltiga modellen ifrån korrelationsanalys

```
summary(glm(cbind(votes,nonvotes)~bachelor + master+social+immigrants+master:social+master:immigrants+s
```

```
##
## Call:
## glm(formula = cbind(votes, nonvotes) ~ bachelor + master + social +
##     immigrants + master:social + master:immigrants + social:immigrants,
##     family = "binomial")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -43.502   -7.887   -1.631    5.786   47.513
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept)        -1.34542    0.01531  -87.88   <2e-16 ***
## bachelor            3.46548    0.19319   17.94   <2e-16 ***
## master             -8.44846    0.14869  -56.82   <2e-16 ***
## social            -24.01305    1.06979  -22.45   <2e-16 ***
## immigrants         -2.66291    0.18102  -14.71   <2e-16 ***
## master:social     259.23115    8.67789   29.87   <2e-16 ***
## master:immigrants  19.63011    1.24418   15.78   <2e-16 ***
## social:immigrants 114.19350    8.69757   13.13   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 106741  on 289  degrees of freedom
## Residual deviance:  45028  on 282  degrees of freedom
## AIC: 47706
##
## Number of Fisher Scoring iterations: 4
```

```
# summary(glm(votes3~bachelor3 + master3+social3+immigrants3+master3:social3+master3:immigrants3+social
```

summary 7.2 - Summary av den slutgiltiga modellen ifrån Meningsfull Selektion

```
summary(glm(cbind(votes,nonvotes)~sex+master+social+sex:master+sex:social,family="binomial"))
```

```
##
## Call:
## glm(formula = cbind(votes, nonvotes) ~ sex + master + social +
##     sex:master + sex:social, family = "binomial")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -42.701   -7.797   -1.008    4.970   47.285
##
## Coefficients:
##               Estimate Std. Error  z value Pr(>|z|)
## (Intercept)   -1.58288    0.01427 -110.959  < 2e-16 ***
## sex           16.58673    1.33812   12.396  < 2e-16 ***
## master        -3.63701    0.07932  -45.854  < 2e-16 ***
## social        -2.86057    0.79136   -3.615 0.000301 ***
## sex:master  -140.80509    6.93960  -20.290  < 2e-16 ***
## sex:social   787.86982   86.16339    9.144  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 106741  on 289  degrees of freedom
## Residual deviance:  45129  on 284  degrees of freedom
## AIC: 47803
##
## Number of Fisher Scoring iterations: 4
```

```
# summary(glm(votes3~sex3+bachelor3+master3+social3+sex3:master3+sex3:social3,family="binomial"))
```

summary 7.3 - Summary av den slutgiltiga modellen ifrån Frammåt Selektion

```
summary(glm(cbind(votes,nonvotes)~assault+master+social+unem+immigrants+assault:social+assault:unem+assa
```

```
##
## Call:
## glm(formula = cbind(votes, nonvotes) ~ assault + master + social +
##     unem + immigrants + assault:social + assault:unem + assault:immigrants +
##     master:social + master:unem + master:immigrants + social:immigrants +
##     unem:immigrants, family = "binomial")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -43.324   -7.600   -1.090    4.982   41.470
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -1.451e+00  1.663e-02 -87.264  < 2e-16 ***
## assault             1.655e+02  5.402e+00  30.647  < 2e-16 ***
## master             -6.670e+00  1.065e-01 -62.608  < 2e-16 ***
## social             -1.926e+01  1.362e+00 -14.147  < 2e-16 ***
## unem               -2.121e+00  3.345e-01  -6.343 2.26e-10 ***
## immigrants          5.345e-01  2.091e-01   2.557   0.0106 *
## assault:social     -1.191e+03  1.759e+02  -6.772 1.27e-11 ***
## assault:unem       -3.704e+02  5.056e+01  -7.327 2.36e-13 ***
## assault:immigrants -1.523e+03  7.370e+01 -20.665  < 2e-16 ***
## master:social       2.064e+02  1.068e+01  19.328  < 2e-16 ***
## master:unem         3.573e+01  3.017e+00  11.843  < 2e-16 ***
## master:immigrants   1.120e+01  1.158e+00   9.675  < 2e-16 ***
## social:immigrants   1.245e+02  1.225e+01  10.170  < 2e-16 ***
## unem:immigrants    -1.501e+01  2.856e+00  -5.257 1.47e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 106741  on 289  degrees of freedom
## Residual deviance:  43611  on 276  degrees of freedom
## AIC: 46302
##
## Number of Fisher Scoring iterations: 4
```

```
# summary(glm(votes3~assault3+bachelor3+master3+social3+unem3+immigrants3+assault3:bachelor3+assault3:s
```

summary 7.4 - Summary av den slutgiltiga modellen ifrån bakåt eliminering

```
summary(glm(cbind(votes,nonvotes)~assault+master+immigrants,family="binomial"))
```

```
##
## Call:
## glm(formula = cbind(votes, nonvotes) ~ assault + master + immigrants,
##     family = "binomial")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -44.418   -7.965   -0.892    5.370   43.977
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.380572   0.003978 -347.01   <2e-16 ***
## assault     37.694513   1.403121   26.86   <2e-16 ***
## master      -5.461632   0.024311 -224.66   <2e-16 ***
## immigrants  -0.444955   0.040415  -11.01   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 106741  on 289  degrees of freedom
## Residual deviance:  45572  on 286  degrees of freedom
## AIC: 48242
##
## Number of Fisher Scoring iterations: 4
```

```
# summary(glm(votes3~assault3+master3+immigrants3,family="binomial"))
```

summary 7.5 - Summary av en modell jämförd i korrelationsanalysen

```
summary(glm(cbind(votes,nonvotes)~sex+master+social+unem+immigrants,family="binomial"))
```

```
##
## Call:
## glm(formula = cbind(votes, nonvotes) ~ sex + master + social +
##     unem + immigrants, family = "binomial")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -44.131   -7.711   -1.272    5.184   47.598
##
## Coefficients:
##              Estimate Std. Error  z value Pr(>|z|)
## (Intercept) -1.335528   0.005664 -235.807  < 2e-16 ***
## sex         -3.791967   0.500644   -7.574 3.61e-14 ***
## master      -5.350966   0.029119 -183.761  < 2e-16 ***
## social       3.400053   0.253008   13.439  < 2e-16 ***
## unem        -0.959753   0.071249  -13.470  < 2e-16 ***
## immigrants   0.004940   0.050493    0.098    0.922
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 106741  on 289  degrees of freedom
## Residual deviance:  45905  on 284  degrees of freedom
## AIC: 48579
##
## Number of Fisher Scoring iterations: 4
```

```
# summary(glm(votes3~sex3+master3+social3+unem3+immigrants3,family="binomial"))
```

summary 7.6 - Summary av en modell jämförd i korrelationsanalysen

```
summary(glm(cbind(votes,nonvotes)~sex+master+social+unem,family="binomial"))
```

```
##
## Call:
## glm(formula = cbind(votes, nonvotes) ~ sex + master + social +
##     unem, family = "binomial")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -44.136   -7.706   -1.275    5.184   47.594
##
## Coefficients:
##             Estimate Std. Error  z value Pr(>|z|)
## (Intercept) -1.335565   0.005651 -236.326   <2e-16 ***
## sex         -3.761202   0.389544   -9.655   <2e-16 ***
## master      -5.350575   0.028844 -185.498   <2e-16 ***
## social       3.404142   0.249530   13.642   <2e-16 ***
## unem        -0.959579   0.071226  -13.472   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 106741  on 289  degrees of freedom
## Residual deviance:  45905  on 285  degrees of freedom
## AIC: 48577
##
## Number of Fisher Scoring iterations: 4
```

```
# summary(glm(votes3~sex3+master3+social3+unem3,family="binomial"))
```

summary 7.7 - Summary av en modell jämförd i korrelationsanalysen

```
summary(glm(cbind(votes,nonvotes)~sex+bachelor+unem+immigrants,family="binomial"))
```

```
##
## Call:
## glm(formula = cbind(votes, nonvotes) ~ sex + bachelor + unem +
##     immigrants, family = "binomial")
##
## Deviance Residuals:
```

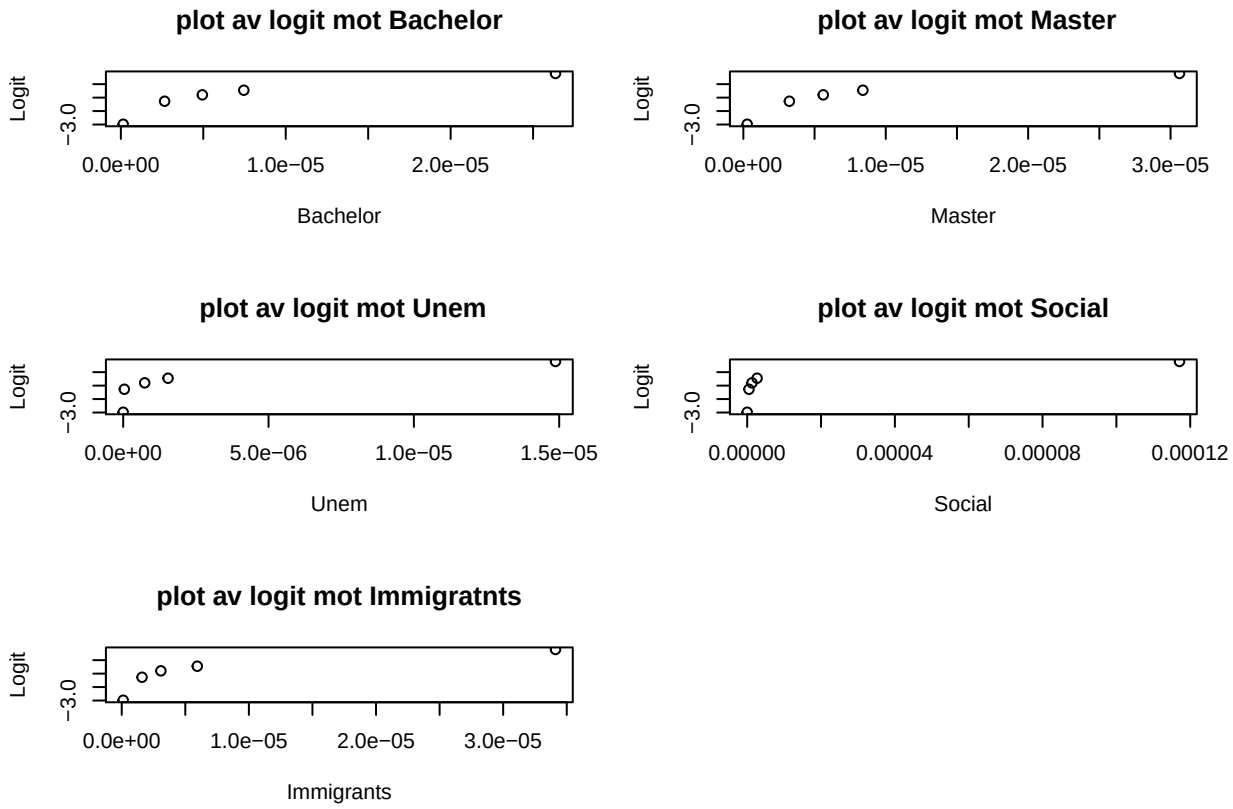
```
##     Min       1Q   Median       3Q      Max
## -61.265   -8.472   -1.188    5.471   50.382
##
## Coefficients:
##              Estimate Std. Error  z value Pr(>|z|)
## (Intercept)  -0.68440    0.00755  -90.646   <2e-16 ***
## sex          -4.49899    0.47489   -9.474   <2e-16 ***
## bachelor    -13.01951    0.07165 -181.699   <2e-16 ***
## unem         -1.31923    0.06936  -19.019   <2e-16 ***
## immigrants   -0.67853    0.04968  -13.658   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 106741  on 289  degrees of freedom
## Residual deviance:  54710  on 285  degrees of freedom
## AIC: 57382
##
## Number of Fisher Scoring iterations: 4
# summary(glm(votes3~sex3+bachelor3+unem3+immigrants3,family="binomial"))
```

summary 7.8 - Summary av en modell jämförd i korrelationsanalysen

```
##                  votes2    nonvotes2      assault2        sex2   bachelor2
## votes2       1.00000000  0.845986430  0.0554512712  0.30995708  0.58907057
## nonvotes2    0.84598643  1.000000000 -0.0020676029  0.24762702  0.71934997
## assault2     0.05545127 -0.002067603  1.0000000000  0.31355499 -0.03518335
## sex2         0.30995708  0.247627024  0.3135549892  1.00000000  0.03874670
## bachelor2    0.58907057  0.719349968 -0.0351833501  0.03874670  1.00000000
## master2      0.43234287  0.629354507 -0.0838042281 -0.09549908  0.88728083
## social2     -0.71000731 -0.673080009  0.0485585165 -0.15935929 -0.61855784
## unem2       -0.45754874 -0.400063374 -0.0002036576 -0.11655409 -0.35977427
## immigrants2  0.02508132  0.056315498  0.0999322467  0.39814962 -0.15399004
##                 master2     social2         unem2 immigrants2
## votes2       0.43234287 -0.71000731 -0.4575487400  0.02508132
## nonvotes2    0.62935451 -0.67308001 -0.4000633743  0.05631550
## assault2    -0.08380423  0.04855852 -0.0002036576  0.09993225
## sex2        -0.09549908 -0.15935929 -0.1165540916  0.39814962
## bachelor2    0.88728083 -0.61855784 -0.3597742720 -0.15399004
## master2      1.00000000 -0.52677567 -0.3049120440 -0.09760243
## social2     -0.52677567  1.00000000  0.4072236298  0.10900771
## unem2       -0.30491204  0.40722363  1.0000000000  0.05267675
## immigrants2 -0.09760243  0.10900771  0.0526767490  1.00000000
```

Summary 7.9 - En korrelationsmatris av data med avseende på outliers

7.2 - plottar



Figur 7.1 - En figur av logit plottar för alla variabler i modellen

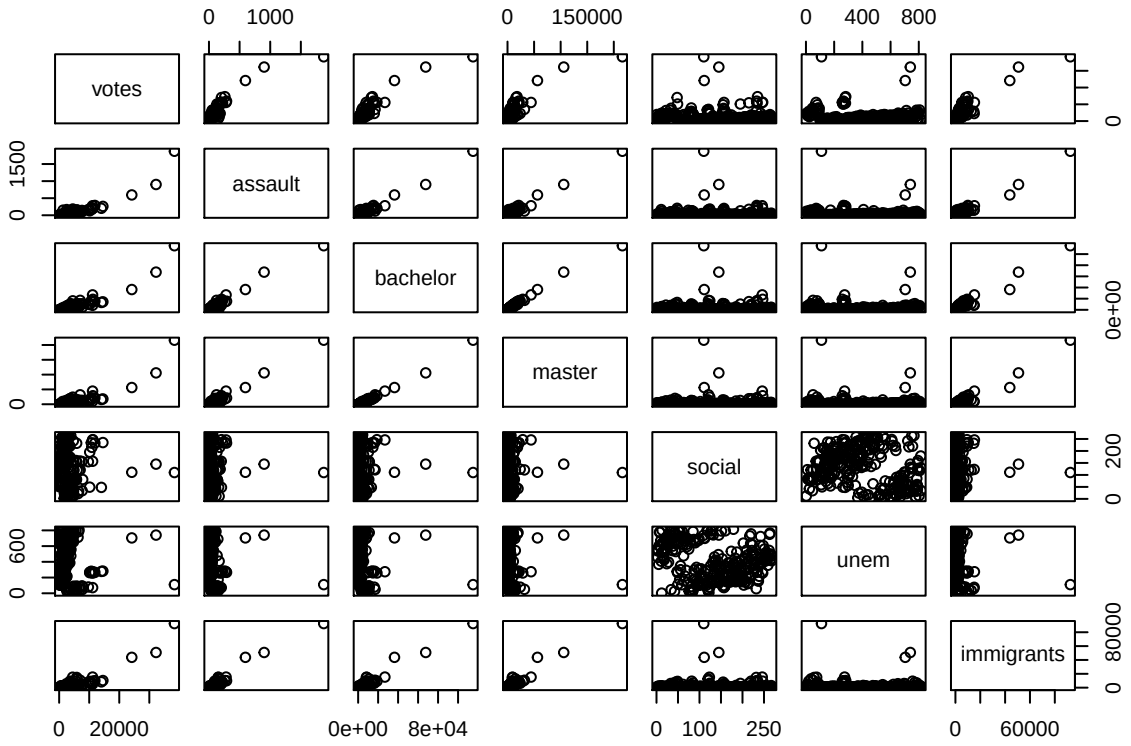


Figure 2: Figur 7.2 - En Pairs-plot över alla kovariater i form av antal som har varit med i undersökningen

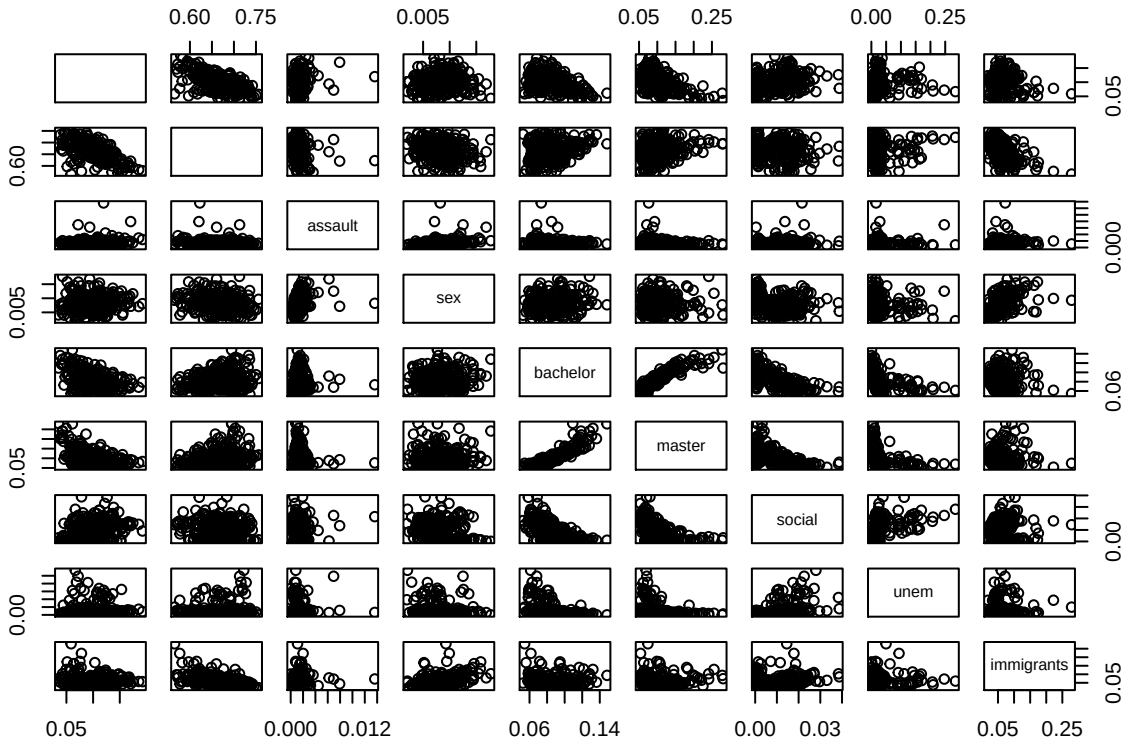


Figure 3: Figur 7.3 - En Pairs-plot över alla kovariater i form av andelar som har varit med i undersökningen