Stockholms universitet

Kan vi prediktera valresultat med logistisk regressionsanalys?

Agnes Wiiand

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2020:16
Matematisk statistik
Juni 2020

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Matematisk statistik
Stockholms universitet
Kandidatuppsats **2020:16**
http://www.math.su.se/matstat

Kan vi prediktera valresultat med logistisk regressionsanalys?

Agnes Wiiand*

Juni 2020

Sammanfattning

I denna uppsats sätter vi upp två olika logistiska regressionsmodeller med syftet att prediktera valresultaten, för de tre största riksdagspartierna, Socialdemokraterna, Moderaterna och Sverigedemokraterna. Vi anpassar de logistiska regressionsmodellerna genom att använda två variabelselektionsmetoder, stegvis variabelselektion och purposeful selection. Vi jämför sedan modellera partivis med hjälp av ROC/AUC analys, Hosmer-Lemeshow test och tränings- och testfel. Vi kommer fram till att ingen av modellerna är tillräckligt bra för att användas vid prediktion av valresultat, i synnerhet inte modellerna för Sverigedemokraterna.

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige.
E-post: agnes.wiiand@gmail.com. Handledare: Felix Wahl, Mathias Millberg Lindholm och Kristoffer Lindensjö.

Förord

Jag vill tacka mina handledare Felix Wahl, Mathias Millberg Lindholm och Kristoffer Lindensjö för all handledning och stöd genom uppsatsskrivandet.

Innehåll

1 Introduktion **5**

2 Beskrivning av data **5**

3 Teori **7**

3.1 Logistisk regression 7

3.2 Odds och oddskvot 8

3.3 Tränings- och testset 9

3.4 Antaganden 9

3.4.1 Multikolinjäritet och variansinflationsfaktor 10

3.4.2 Linjäritet i log-oddsen 10

3.5 Modellselektion 11

3.5.1 Waldtest 11

3.5.2 Likelihood-kvot test 12

3.5.3 Akaike's Information Criterion 12

3.5.4 Stegvis variabelselektion 13

3.5.5 Purposeful selection 13

3.6 Modellutvärdering 15

3.6.1 Receiver operating characteristic 15

3.6.2 Hosmer-Lemeshow test 16

3.6.3 Tränings- och testfel 16

4 Modellanpassning **16**

4.1 Multikolinjäritetsanalys 17

4.2 Med stegvis variabelselektion 17

4.2.1 Moderaterna 18

4.2.2 Sverigedemokraterna 18

4.2.3 Socialdemokraterna 19

4.3 Med purposeful selection 19

4.3.1 Moderaterna och Sverigedemokraterna 19
4.3.2 Socialdemokraterna 21

5 Modellutvärdering 22
5.1 ROC/AUC 22
5.2 Hosmer-Lemeshow test 23
5.3 Tränings- och testfel 24

6 Resultat 24

7 Diskussion 25

8 Referenslista 27
8.1 Datamaterial 27

A Appendix Figurer 29
A.1 Korrelationsmatris 29
A.2 Linjäritet i log-odds 30
A.2.1 Moderaterna 30
A.2.2 Sverigedemokraterna 31
A.2.3 Socialdemokraterna 32
A.3 ROC figurer 33

B Appendix Tabeller 34
B.1 Stegvis variabelselektion 34
B.2 Purposeful selection 37
B.2.1 Moderaterna 37
B.2.2 Sverigedemokraterna 39
B.2.3 Socialdemokraterna 41

1 Introduktion

Kan man på ett relativt enkelt sätt, med offentliga data, göra en prediktionsmodell som ger oss andelen röster ett parti får i nästa val? Är våra politiska åsikter så förutsägbara? Gissningsvis inte, men i denna uppsats gör vi ett försök. Vi kommer att skapa prediktionsmodeller för de tre största partierna i riksdagsvalet 2018, Socialdemokraterna, Moderaterna och Sverigedemokraterna. Detta gör vi genom att använda logistiska regressionsmodeller och valresultaten per kommun för riksdagsvalen som ägt rum under 2000-talet.

Vi kommer i avsnitt 2 beskriva den offentliga data vi kommer använda till modelleringarna i uppsatsen. Avsnitt 3 är teoridelen och i denna finns det flera delavsnitt. I delavsnitt 3.5 presenteras teorin för modellselektion och i delavsnitt 3.6 presenteras teorin för modellutvärdering. I avsnitt 4 gör vi två modellanpassningar per parti, en med stegvis variabelselektion och en med purposeful selection. I avsnitt 5 använder vi testerna presenterade i delavsnitt 3.6 för att utvärdera modellerna. Resultatet sammanställs i avsnitt 6 och en diskussion om modelleringarna och resultaten finns i avsnitt 7.

2 Beskrivning av data

Eftersom vi är intresserade av de svenska riksdagsvalen under 2000-talet har vi från statistiska centralbyrån, förkortas SCB, hämtat valresultaten för riksdagsvalen från år 2002 till 2018 per kommun. I datasetet finns antalet röster per riksdagsparti, antalet röster på övriga partier, antalet ogiltiga röster och antalet röstberättigade i kommunen som valde att inte rösta. I en logistisk regressionsmodell behöver vi ha antalet lyckade och antalet misslyckade utfall som vår responsvariabel, mer om logistisk regression i avsnitt 3.1. Vi får vår *responsvariabel* genom att antalet lyckade utfall blir antalet röster på det parti vi modellerar och antalet röster som lades på något annat parti eller inte alls blir antalet misslyckade utfall. Detta då varje förlorad röst, oavsett om den lades på något annat parti eller inte alls, är ett misslyckat resultat för partiet vi är intresserade av. Detta betyder mer konkret för modelleringen att varje rad, som representerar en kommun ett visst valår, kommer ha ett antal lyckade och misslyckade utfall.

De förklarande variablerna kommer från SCB, Arbetsförmedlingen och BRÅ och finns sammanställda i Tabell 1. Referenser till dataset och tabeller finns i referenslistan i avsnitt 8. För att sammanställa datasetet för denna uppsats har vi använt oss utav bland annat tidyverse i R.

Från SCB har vi fått de följande förklarande variablerna: *Medelålder*, som är medelåldern i kommunen vid givet valår. *Medelinkomst*, som är medelvärdet av

förvärvsinkomsten i kommunen uttryckt i tusen kronor vid givet valår. *Utrikes födda* är andelen utrikesfödda i en kommun. *Kvinnor* är andelen kvinnor i en kommun. *Total kommunal skattesats* är summan av den kommunala och regionala skatten. *Eftergymnasial utbildning 3 år+* är andelen i kommunen som studerat på högskola eller universitet i minst 3 år. *Forskarutbildning* är andelen i kommunen som har en forskarexamen. *Bruttoregionalprodukt*, förkortas till BRP, är den regionala motsvarigheten till bruttonationalprodukt, förkortat BNP. BRP är det sammanräknade värdet på alla varor och tjänster som producerats i regionen, i vårt fall länet, under ett år. De värden vi har med är BRP per capita, BRP värdet delat på antalet invånare i länet, uttryck i tusen kronor.

Från Arbetsförmedlingen har vi den förklarande variabeln *Inskrivet arbetslösa*, som är medelvärdet av andelen inskrivet arbetslösa under året i en kommun.

Från BRÅ har vi *Totalt anmälda brott*, vilket är antalet anmälda brott i en kommun per 100 000 invånare. Vi använder detta mått istället för det faktiska antalet för att kunna göra en mer rättvis jämförelse mellan kommuner av olika befolkningsstorlek. Vi har också variabeln *Totalt antal brott mot liv och hälsa*, som också är justerat för antalet 100 000 invånare kommunerna har. Brott mot liv och hälsa är enligt svensk lag (https://lagen.nu/1962:700#K3) de brott som uppsåtligen eller oaktsamt skadat en levande människa. Exempel på sådana brott är mord, vållande till annans död och misshandel.

Vi har med båda dessa brottsvariabler då Lag och Ordning var, enligt både VALU (https://www.svt.se/special/valu2018-valjarnas-viktigaste-fragor/) och Novus (https://novus.se/valjaropinionen/valjarforstaelse/viktigaste-politiska-fragan/), en av de viktigaste valfrågorna 2018 och var en fråga som diskuterades flitigt fram till covid-19 pandemin enligt Sveriges Radio (https://sverigesradio.se/sida/artikel.aspx?programid=83artikel=7459859).

Förklarande variabel	Källa
Totalt anmälda brott	BRÅ
Totalt antal brott mot liv och hälsa	BRÅ
Medelinkomst	SCB
BRP	SCB
Total kommunal skattesats	SCB
Kvinnor	SCB
Medelålder	SCB
Eftergymnasial utbildning 3år+	SCB
Forskarutbildning	SCB
Utrikes födda	SCB
Inskrivet arbetslösa	Arbetsförmedlingen

Tabell 1: Sammanställning av de förklarande variablerna

3 Teori

I detta avsnitt kommer teorin som används vid modelleringen och analysen att beskrivas. Avsnittet börjar med allmän teori som behövs för att kunna göra vår analys, så som teorin för logistisk regression i delavsnitt 3.1. I delavsnitt 3.4 presenteras de antaganden som måste undersökas för logistisk regression och sedan följer teorin för modellselektion i delavsnitt 3.5 och för modellutvärdering i delavsnitt 3.6.

3.1 Logistisk regression

Teorin för logistisk regression är främst från Agresti (2013, kapitel 5), men även Hosmer, Lemeshow och Sturdivan (2013, kapitel 1-2).

Logistisk regression är en regressionsmodell som beskriver relationen mellan en binär responsvariabel och en eller flera förklarande variabler. En binär responsvariabel innebär att utfallen enbart kan anta två värden, ett "lyckat" eller ett "misslyckat" utfall, vilka generellt betecknas med 1 och 0. Vi använder logistisk regression för att modellera sannolikheterna för ett "lyckat" utfall, givet de förklarande variablerna. I vår analys kommer ett lyckat utfall vara en röst på partiet av intresse och ett misslyckat utfall är en röst som ej lades på partiet av intresse.

Vi låter Y vara vår binära responsvariabel med möjliga värden 0 och 1. Vi har k stycken förklarande variabler som vi betecknar med $x_1, x_2, ..., x_k$. $\mathbf{X} = (x_1, x_2, ..., x_k)$ är då vår vektor med alla förklarande variabler. Då vi har fler än en förklarande variabel får vi en *multipel logistisk regressionsmodell*. Då logistisk regression används för att modellera sannolikheten att vi får ett "lyckat" utfall givet de förklarande variablerna i $\mathbf{X}$ får vi formeln presenterad i Ekvation (1). α står för interceptet i modellen och β_i är lutningskoefficienten för den förklarande variabeln x_i. Formeln för sannolikheten av ett "lyckat" utfall är

$$\pi(\mathbf{X}) = \mathbf{P}(Y = 1|X = \mathbf{X}) = 1 - \mathbf{P}(Y = 0|X = \mathbf{X}) = \frac{e^{\alpha+\beta_1 x_1+...+\beta_k x_k}}{1 + e^{\alpha+\beta_1 x_1+...+\beta_k x_k}}. \quad (1)$$

$\pi(\mathbf{X})$ kan enbart anta värden mellan 0 och 1, men om vi tar logiten av denna modell, presenterad i Ekvation (2), kan modellen anta alla reella tal och vi får en linjär relation mellan responsvariabeln och de förklarande variablerna. Formeln för logiten är

$$logit\,[\pi(\mathbf{X})] = log\left(\frac{\pi(\mathbf{X})}{1 - \pi(\mathbf{X})}\right) = \alpha + \beta_1 x_1 + ... + \beta_k x_k. \quad (2)$$

De värden våra β koefficienter får i logit-modellen representerar hur log-oddsen, som beskrivs i avsnitt 3.2, för ett lyckat utfall förändras om den förklarande

variabel ändras. Så om $\beta_i > 0$ betyder det att sannolikheten att vi får ett lyckat utfall ökar om x_i ökar och motsvarande betyder det att om $\beta_i < 0$ minskar sannolikheten för ett lyckat utfall om x_i ökar. Om $\beta_i = 0$ betyder det att log-oddsen, som beskrivs i avsnitt 3.2, för ett lyckat utfall inte förändras, oavsett om x_i ökar eller minskar. Sådana variabler kan vi utesluta för att få en bättre modell, mer om det i avsnitt 3.5.

I logistisk regression används *maximum likelihood-metoden* för att skatta koefficienterna. För den som vill läsa mer om det hänvisar vi till Hosmer et al. (2013, sid 8 – 10).

3.2 Odds och oddskvot

Teorin om odds och oddskvot kommer från Hosmer et al. (2013, kap 3) och Wiest, Lee och Carlin (2015).

Odds är sannolikheten att utfallet vi är intresserade av sker, sannolikheten betecknar vi med p, delat på sannolikheten att utfallet inte sker, $1 - p$. Om vi ersätter p med sannolikheterna från en logistisk regression ser vi att det logaritmerade oddset blir logit-modellen som vi definierade i avsnitt 3.1. Formeln för odds är

$$\text{odds} = \frac{p}{1-p} = \frac{\pi(x)}{1-\pi(x)}. \tag{3}$$

Oddskvot, förkortat **OR** efter engelskans *odds ratio*, används för att jämföra sannolikheten att det intressanta utfallet sker när en förklarande variabel förändras, ofta i termer utav en enhet. Exempel på en enhet i vårt fall är en procentenhet eller 1000kr mer i medelinkomst för en kommun. Oddskvoten för en univariat logistisk regression med förklarande variabel x som ökas med en enhet definieras som

$$\mathbf{OR} = \frac{odds(x+1)}{odds(x)} = \frac{\dfrac{\pi(x+1)}{1-\pi(x+1)}}{\dfrac{\pi(x)}{1-\pi(x)}} = \frac{e^{\alpha+\beta(x+1)}}{e^{\alpha+\beta x}} = e^{\beta}. \tag{4}$$

Om oddskvot värdet för till exempel x_i är 1.5 betyder det att oddset för att det intressanta utfallet sker ökar med en faktor på 1.5 om vi ökar x_i variabeln med en enhet, och i en multipel logistisk regression håller de andra förklarande variablerna konstanta. Oddskvoten kan anta alla positiva reella tal och 0. Om $0 \leq \mathbf{OR} < 1$, betyder det att sannolikheten för att det intressanta utfallet sker minskar om den förklarande variabeln ökar med en enhet. Om $\mathbf{OR} \geq 1$ ökar sannolikheten för utfallet. På grund av att $\mathbf{OR} \in [0, \infty)$ använder man sig oftast av *log-oddskvoten* som kan anta alla reella värden och är symmetrisk runt 0.

3.3 Tränings- och testset

Om man har tillräckligt med data bör man, enligt Hastie, Tibishirani och Friedman (2009, sid 219 - 223), dela upp sitt dataset i mindre delar, ett träningsset bestående av ungefär 80% av observationerna, och ett testset bestående av de resterande 20% av observationerna. Träningssetet används för att anpassa en eller flera modeller och testsetet används sedan för att kontrollera modellerna för bland annat överanpassning. Om en modell är överanpassad till det datamaterial vi använder när vi skapar modellen finns det risk för att modellen inte gör bra prediktioner när vi introducerar ny data vi vill prediktera. För att undersöka detta beräknar vi tränings- och testfel för våra modeller. Vi vill att båda dessa fel ska vara så små som möjligt, men en minskning i träningsfel kan kosta genom att modellen passar testsetet sämre, vilket ger oss ett högre testfel.

Träningsfelet beräknas som medelvärdet av prediktionsfelen som ges av vår modell. Formeln för träningsfelet finns i Ekvation (5). N_1 är antalet observationer i träningssetet, y_i är den observerade sannolikheten för ett positivt utfall och $\hat{y}_i$ är den predikterade sannolikheten för ett positivt utfall, som vi får från modellen vi anpassat med träningssetet. Formeln för träningsfel är

$$\mathrm{Err}_{\text{Träning}} = \frac{1}{N_1} \sum_{i=1}^{N_1} \left(y_i^{(\text{Träning})} - \hat{y}_i^{(\text{Träning})} \right)^2 . \tag{5}$$

Testfelet beräknas på samma sätt, men använder sig utav de N_2 observationerna från testsetet istället för träningssetet, men predikterar fortfarande med modellen vi anpassat med träningssetet. Detta betyder att vi kan se testsetet som "ny data" som modellen ska prediktera för. Formeln för testfelet är

$$\mathrm{Err}_{\text{Test}} = \frac{1}{N_2} \sum_{i=1}^{N_2} \left(y_i^{(\text{Test})} - \hat{y}_i^{(\text{Test})} \right)^2 . \tag{6}$$

I vår analys kommer träningssetet bestå av observationerna från valen mellan 2002 och 2014 och testsetet kommer vara observationerna från valet 2018. Vi gör denna indelning då detta är en naturlig uppdelning av vårt dataset och vi vill se hur väl våra modeller kommer kunna prediktera framtida val.

3.4 Antaganden

När vi gör en logistisk regression finns det några antaganden att undersöka och ta hänsyn till. Det får inte existera multikolinjäritet och de förklarande variablerna ska vara linjära i log-oddsen. Hur vi undersöker dessa antaganden presenteras i delavsnitten nedan.

3.4.1 Multikolinjäritet och variansinflationsfaktor

Variansinflationsfaktor, förkortas VIF, är enligt Sundberg (2016, sid 72-74, 89) ett statistiskt mått som används för att undersöka multikolinjäritet mellan de förklarande variablerna i en modell. Att förklarande variabler är multikolinjära betyder att minst en förklarande variabel är (nästan) en linjärkombination av de andra. Detta kan ställa till problem i modelleringen genom att ge (icke)signifikanta resultat där det inte borde och vi vill därför undersöka detta innan vi anpassar någon modell. Multikolinjärietet mellan två variabler kan man undersöka genom att beräkna korrelationskoefficienten. Om korrelationskoefficienten är stor mellan två förklarande variabler finns det multikolinjäritet. Dock kan vi inte använda oss utav korrelationanalys för att upptäcka multikolinjäritet mellan tre eller fler förklarande variabler, då använder vi oss utav VIF istället. Vi kommer använda oss utav den linjära varianten av VIF, där man enbart undersöker hur de förklarande variablerna förhåller sig till varandra, oberoende av hur responsvariabeln ser ut.

VIF-faktorn, definierad i Ekvation (7), anger hur mycket större varians vår koefficientskattning, för den förklarande variabeln j, får i den fulla modellen jämfört med variansen i den univariata modellen. R_j^2 är förklaringsgraden som anger hur mycket av variationen i x_j som förklaras av de andra förklarande variablerna. Formeln för VIF-faktorn är

$$\text{VIF} = \frac{1}{1 - R_j^2}. \tag{7}$$

VIF-faktorn kan anta värden större än eller lika med 1, där 1 betyder att den förklarande variabeln x_j är ortogonal mot de andra. Högre värden på VIF-faktorn indikerar att det existerar multikolinjäritet. Enligt Kutner, Nachtsheim, Neter och Li (2004, sid 408-409) vill vi ha VIF-faktorer under 10 för att kunna fortgå analysen med våra variabler.

3.4.2 Linjäritet i log-oddsen

Det finns några olika sätt att undersöka de kontinuerliga variablernas linjäritet i log-oddsen och vi väljer att göra det som beskrivet i Hosmer et al. (2013, sid 94 - 96), genom att kolla på kvartilplottar.

För att undersöka linjäriteten i log-oddsen transformerar vi de kontinuerliga förklarande variablerna, en och en, till kategoriska variabler bestående av kvartilerna, programmeras som dummyvariabler med fyra nivåer. Vi anpassar en multipel logistisk regressionsmodell där en av de kontinuerliga variablerna har bytts ut mot dess kategoriska motsvarighet, där den lägsta kvartilen blir referensgruppen. Vi plottar sedan våra koefficientskattningar i logit-modellen mot medelvärdena för kvartilerna med räta linjer mellan punkterna, första kvarti-

len får koefficientvärde 0. Om vi har linjäritet i log-oddsen ska denna plottade linje ha vara approximativt linjär och vi kan i sådana fall forsätta med vår kontinuerliga variabel. Om linjen inte är approximativt linjär kan vi ersätta den förklarande variabeln i modellen med den kategoriska motsvarigheten, bestående av kvartilerna.

3.5 Modellselektion

Enligt Agresti (2013, sid 207) finns det två mål när vi anpassar en modell. Modellen ska vara komplex nog att passa vår data bra, vilket vi kan undersöka genom goodness-of-fit tester. Dessa presenteras i avsnitt 3.6. Samtidigt bör modellen vara relativt enkel, inte för många förklarande variabler, så att vi inte överanpassar modellen till träningssetet. Detta är viktigt då vi vill kunna använda våra modeller för att prediktera nya data i framtiden. I detta delavsnitt kommer vi presentera teori för metoder att undersöka de förklarande variablerna med och hur vi anpassar modeller genom stegvis variabelselektion och purposeful selection.

3.5.1 Waldtest

Teorin om Waldtest och följande avsnitt om Likelihood-kvot test kommer från Agresti (2013, kap 1.3.3).

En bra modell har enbart med de förklarande variabler som har signifikanta koefficientskattningar. För att testa signifikanserna för β värdena kommer vi använda oss utav kvadratroten av *Wald-statistikan*, även kallat *z-statistikan*, som definieras i Ekvation (8).

Nollhypotesen i Wald- och z-test är $H_0 : \beta_i = \beta_0$ och under nollhypotesen är z-statistiskan approximativt standardnormalfördelad. Vi definierar vår maximum likelihood skattade koefficient för förklarande variabeln i med $\hat{\beta}_i$ och standard felet för skattningen med $\hat{SE}_i$. Z-satistikan har formeln

$$z = \frac{\hat{\beta}_i - \beta_0}{\hat{SE}_i} \sim N(0, 1). \tag{8}$$

Vi kommer testa för $\beta_0 = 0$, då vi är intresserade av att undersöka om de förklarande variablerna har koefficientskattningar som är signifikant skilda från noll. Vi kommer använda oss utav en signifikansnivå på 0.05 när vi undersöker våra förklarande variabler med z-test.

En viktig detalj enligt Hosmer et al. (2013, sid 41) är att om vi har en kategorisk variabel där någon nivå har ett p-värde större än 0.05 och vi väljer att utesluta denna måste alla nivåer för den kategoriska variabeln uteslutas.

3.5.2 Likelihood-kvot test

Med likelihood-kvot test prövar vi goodness-of-fit mellan en mer komplex modell, med k förklarande variabler, mot en enklare modell, med en delmängd av den komplexa modellens förklarande variabler, till exempel $k - m$. Detta test använder vi för att undersöka om den enklare modellen är tillräckligt bra för att ersätta den komplexa modellen. Nollhypotesen för likelihood-kvot testet är att den enklare modellen är tillräcklig, vilket betyder att de parametrarna som ingår i den mer komplexa modellen, men inte i den enklare modellen, är noll.

Likelihood-kvot statistikan, förkortas LR efter engelskans likelihood-ratio, är asymptotisk χ^2-fördelad under nollhypotesen. Antalet frihetsgrader fås genom att ta differensen av dimensionerna av parameterrummen för de två modellerna, i vårt exempel $k - (k - m) = m$. Statistikans formel finns i Ekvation (9). l står för maximum av log-likelihood funktionerna för våra modeller, l_E för den enklare modellen och l_K för den komplexa. Formeln för likelihood-kvoten är

$$\mathbf{LR} = -2 \cdot (l_E - l_K) \sim \chi^2(m). \tag{9}$$

3.5.3 Akaike's Information Criterion

Akaike's Information Criterion, förkortas AIC, är enligt Agresti (2013, kapitel 6.1.6) och Held och Sabané Bové (2014, kapitel 7.1.1) ett modellmått som vi kan använda för att jämföra modeller med varandra. AIC är ett typ av mått på balansen mellan goodness-of-fit och hur många förklarande variabler en modell har. Den modell som minimerar AIC är den modell kriteriet väljer som den bästa. Formeln för AIC finns i Ekvation (10) där l_{ML} är den maximerade log-likelihood funktionen för modellen vi undersöker och k är antalet förklarande variabler modellen har. Värdet vi får ut från AIC säger inget speciellt i sig utan värdet används för att jämföra modeller som har samma responsvariabel med varandra. AIC definieras som

$$\mathbf{AIC} = -2 \cdot (l_{ML} - k). \tag{10}$$

Enligt Agresti (2013, kap 6.1.6) är det inte enbart den modellen med lägst AIC som är intressant. Om två modeller har liknande AIC kan vi välja den modellen med något högre AIC om denna har färre förklarande variabler. Det finns inget som är mer korrekt än det andra i denna situation, utan varje fall måste analyseras och med hjälp av andra tester kan man bestämma vilken modell som är bäst för syftet.

3.5.4 Stegvis variabelselektion

Stegvis variabelselektion går ut på att hitta en bra modell som beskriver data med så få förklarande variabler som möjligt. Variabelselektion kan genomföras på flera sätt, vi kommer titta närmare på *Forward selection* och *Backward elimination* som beskrivet i Agresti (2013, kapitel 6.1.3) och Sundberg (2016, sid 71).

I *forward selection* utgår vi från en modell med enbart intercept som förklarande variabel. I varje steg adderas den förklarande variabel som ger den största förbättringen av modellen. Agresti (2013, sid 209) föreslår att man adderar den förklarande variabel som har lägst p-värde för z-testet i modellen, men man kan använda andra kriterier också. Vi kommer använda oss utav R's funktion *Step* när vi gör stegvis variabelselektion och detta mått utgår från förbättring av AIC värdet. Processen slutar när modellen inte längre minskar AIC värdet vid inkludering av någon variabel.

I *backward elimination* börjar vi istället med modellen med alla förklarande variabler och väljer, i varje steg, att exkludera den variabel som ger oss den största förbättringen av modellen. Som ovan beskrivit kommer vi använda oss utav AIC som mått på förbättring. Detta betyder att vi i varje steg exkluderar den förklarande variabel som ger oss den största förbättringen för AIC värdet. Processen slutar då modellen inte minskar AIC värdet vid exkludering av någon variabel.

Agresti (2013, kap 6.1.3, sid 210) lyfter fram att stegvis variabelselektion inte ska användas utan egen analys. Detta är speciellt viktigt om vi genom dessa metoder får fram två olika modeller.

3.5.5 Purposeful selection

Hosmer et al. (2013, sid 90 - 93) presenterar en alternativ variant av stegvis variabelselektion vid namn purposeful selection för logistisk regression. Denna variabelselektion har sju steg som presenteras nedan.

Steg 1: I detta steg undersöker vi om de förklarande variablerna överhuvudtaget är intressanta att ha med i vår modell. Vi undersöker detta genom att anpassa en univariat logistisk regressionsmodell för varje förklarande variabel och kontrollerar p-värdena för z-testerna. De variabler som har ett p-värde mindre än 0.25 fortsätter vi vår undersökning med.

Steg 2: Vi anpassar en multipel logistisk regressionsmodell med de förklarande variablerna som blev godkända i steg 1. Även i detta steg är vi intresserade av de förklarande variablernas p-värden för z-testen, men denna gång vill vi att p-värdena ska vara mindre än 0.05 för att variablerna ska få vara kvar i modellen. Om vi har förklarande variabler som har större p-värden än 0.05 anpassar vi

en ny multipel logistisk regressionsmodell utan dessa variabler och testar sedan denna enklare modell mot den fulla modellen med ett likelihood-kvot test. Om vi får ett p-värde större än 0.05 för testet är den enklare modellen tillräckligt bra enligt likelihood-kvot testet.

Steg 3: Om vi i steget ovan fick en enklare och en mer komplex modell måste vi göra en noggrannare analys innan vi kan bestämma oss för om den mindre modellen är tillräckligt bra för att kunna fortsätta analysen med. Vi jämför de estimerade koefficienterna mellan den mindre och den större modellen. Om någon av de förklarande variablerna får en drastisk ökning av sitt estimerade β värde i den mindre modellen indikerar det på att någon eller några av de förklarande variablerna som exkluderades i steg 2 är viktiga tillsammans med de resterande variablerna och bör då inkluderas i modellen igen. $\hat{\beta}_i$ är den estimerade koefficienten för den förklarande variabeln i i den komplexa modellen och $\hat{\theta}_i$ den estimerade koefficienten i den enklare modellen. Om $\Delta\hat{\beta}_i$, definierad i Ekvation (11), överstiger 0.2 inkluderar vi de förklarande variablerna vi uteslöt i steg 2, en i taget, och gör om steg 2 och 3 tills alla $\Delta\hat{\beta}_i$ är mindre än 0.2. Formulan för $\Delta\hat{\beta}_i$ är

$$\Delta\hat{\beta}_i = \frac{\hat{\theta}_i - \hat{\beta}_i}{\hat{\beta}_i}. \tag{11}$$

Steg 4: Om vi i steg 1 uteslöt någon förklarande variabel måste vi undersöka om dessa har signifikanta resultat tillsammans med de andra variablerna. Vi adderar de exkluderade variablerna en och en och upprepar steg 2 och 3 för varje variabel och tar med de variabler som har signifikanta p-värden. Den modell vi får efter detta steg kallas för den *preliminära huvudeffektsmodellen.*

Steg 5: För de kontinuerliga förklarande variablerna undersöks linjäritetsantagandet i log-oddsen i detta steg. Hur vi gör detta finns beskrivet i avsnitt 3.4.2. Modellen vi får efter detta steg kallas för *huvudeffektsmodellen.*

Steg 6: I detta steg undersöker vi relevanta samspel mellan de förklarande variablerna. Om en samspelseffekt ska vara med i modellen ska den ha ett p-värde mindre än 0.05, men även en logisk anledning. För att inte få för många samspelsvariabler har vi redan bestämt oss för vilka fem som skulle vara intressanta att undersöka. Dessa samspelsvariabler finns sammanställda i Tabell 2.

Vi adderar en samspelseffektsvariabel i taget till huvudeffektsmodellen och gör ett likelihood-kvot test för att undersöka om modellen blir bättre med samspelsvariabeln. Om modellen blir bättre får vi ett p-värde mindre än 0.05 för likelihood-kvot testet. När vi fått fram vilka samspel som förbättrar modellen anpassar vi en ny multipel logistisk regressionsmodell med variablerna från huvudeffektsmodellen och de signifikanta samspelen och gör om steg 2, men denna gång tas enbart eventuella samspelseffekter bort, inte de redan undersökta huvu-

deffekterna. Modellen vi får efter denna undersökning kallas för den *preliminärt färdiga modellen*.

Samspelsvariabler	
Totalt anmälda brott	Totalt antal brott mot liv och hälsa
Totalt anmälda brott	Utrikes födda
Eftergymnasial utbildning 3år+	Inskrivet arbetslösa
Eftergymnasial utbildning 3år+	Medelinkomst
Total kommunal skattesats	Medelinkomst

Tabell 2: Samspelsvariabler

Steg 7: Innan vi bestämmer oss för om modellen är bra måste vi testa modellens anpassning genom goodness-of-fit tester. Dessa finns beskrivna i Avsnitt 3.6.

3.6 Modellutvärdering

I detta avsnitt beskrivs teori för olika goodness-of-fit och prediktionstester vi kommer använda oss utav för att utvärdera och jämföra modellerna i avsnitt 5.

3.6.1 Receiver operating characteristic

När vi använder våra modeller för att prediktera utfall får vi sannolikheten för att den givna observationen kommer få ett lyckat utfall. Agresti (2013, sid 223) skriver att vi behöver bestämma ett gränsvärde, π_0, som hjälper oss klassificera våra predikterade observationer. Om den predikterade sannolikheten att få ett lyckat utfall är större än π_0 klassas utfallet som lyckat och för de observationer där sannolikheten att få ett lyckat utfall är mindre än π_0 klassas utfallen som misslyckade. För att se hur bra vår modell är på att prediktera rätt beräknar vi sensitiviteten och specificiteten för modellen. Sensitiviteten är sannolikheten att vi får ett lyckat utfall givet att det observerade utfallet är lyckat. Formeln är

$$\text{sensitivitet} = \mathbf{P}(\hat{y} = 1 | y = 1). \tag{12}$$

Specificiteten är sannolikheten att få ett misslyckat utfall, givet att det observerade utfallet är misslyckat. Formeln för specificiteten är

$$\text{specificitet} = \mathbf{P}(\hat{y} = 0 | y = 0). \tag{13}$$

Hur många lyckade utfall vi får för vår modell är beroende av vad vi väljer för π_0. Det är här *Receiver operating characteristic kurvan*, förkortas ROC, används. Enligt Agresti (2013, sid 223 - 224) är ROC kurvan en plot av sensitivitet mot

1 – specificitet för alla möjliga värden på π_0. Det vi är intresserade av är AUC, från engelskans Area under the curve, som kan ses som ett mått på modellens prediktionsförmåga. Hosmer et al. (2013, sid 178) skriver att om vi har ett AUC värde mellan 0.5 och 0.7 har vi en relativt dålig modell, mellan 0.7 och 0.8 har vi en godkänd modell och mellan 0.8 och 0.9 har vi en bra modell. Över 0.9 anses som en mycket bra modell.

3.6.2 Hosmer-Lemeshow test

Detta test finns beskrivet i Hosmer et al. (2013, sid 157 – 159). Vi delar in vårt datamaterial i G = 10 grupper, 10-percentilerna av observationernas predikterade sannolikheter. Den första gruppen innehåller de observationer med lägst predikterad sannolikhet för ett lyckat utfall och den sista gruppen innehåller de observationer med högst predikterad sannolikhet för ett lyckat utfall. Nollhypotesen är att de observerade och predikterade sannolikhetsproportionerna är densamma för grupperna. Hosmer-Lemeshow statistikan har formeln

$$\hat{C} = \sum_{g=1}^{G} \left(\frac{(O_{1,g} - \hat{e}_{1,g})^2}{\hat{e}_{1,g}} + \frac{(O_{0,g} - \hat{e}_{0,g})^2}{\hat{e}_{0,g}} \right). \tag{14}$$

$O_{1,g}$ står för antalet lyckade observationer i grupp g, medan $O_{0,g}$ står för antalet misslyckade observationer i grupp g. $\hat{e}_{1,g}$ är antalet predikterade lyckade observationer och $\hat{e}_{0,g}$ är antalet predikterade misslyckade observationer i grupp g. $\hat{C}$ följer en asymptotisk χ^2-fördelning med antalet frihetsgrader $G - 2 = 8$ i vårt fall. En viktig sak att anmärka är att om vi har många observationer kan mycket små skillnader i observerade och predikterade sannolikheter göra att vi förkastar nollhypotesen trots att prediktionerna inte är så dåliga.

3.6.3 Tränings- och testfel

För att utvärdera våra prediktionsmodeller kommer vi också beräkna tränings- och testfelen som definierades i avsnitt 3.3. Vi kommer ta kvadratroten ur formlerna som definierades för att få fram procentfelen, vilket är enklast att förstå då vi undersöker andelen röster. Detta test kallas även för RMSEP från engelskans Root Mean Squared Error of Prediction.

4 Modellanpassning

Vi kommer att göra två modellanpassningar för våra partier, en där vi använder oss utav stegvis variabelselektion, utan samspelsvariabler, och sedan en där vi använder purposeful selection och följer de sju stegen beskrivna i avsnitt 3.5.5.

I avsnitt 5 kommer vi sedan jämföra de modeller vi får fram i detta avsnitt partivis för att sedan bestämma den bättre modellen för varje parti.

4.1 Multikolinjäritetsanalys

Innan vi anpassar någon modell måste vi undersöka multikolinjäriteten för våra förklarande variabler. Som vi skrev i avsnitt 3.4.1 använder vi oss utav de linjära sambanden mellan de förklarande variablerna, vilket betyder att vi får samma multikolinjäritetsanalys för våra tre partier.

Vi börjar med att titta på korrelationskoefficienterna, som finns sammanställda i en korrelationsplot i appendix, Figur 1. Vi har två korrelationskoefficienter vars absolutvärde är 0.7 eller större. Korrelationen mellan brottsvariablerna *Totalt anmälda brott* och *Totalt antal brott mot liv och hälsa* är lika med 0.7, vilket inte är så konstigt då *Totalt antal brott mot liv och hälsa* är en del av *Totalt anmälda brott*. Korrelationen mellan *BRP* och *Medelinkomst* är lika med 0.74, vilket inte är så högt att vi direkt från denna analys kan säga att det existerar multikolinjäritet.

För att undersöka detta noggrannare beräknar vi VIF-faktorerna för våra förklarande variabler. Dessa finns sammanställda i Tabell 3. Vi ser att den största VIF-faktorn för våra förklarande variabler är för *Eftergymnasial utbildning 3år+* på 3.81. Detta är dock ett lågt tal, klart under gränsen på 10, vilket betyder att vi kan fortgå modelleringen med alla förklarande variabler.

Förklarande variabel	VIF
Eftergymnasial utbildning 3år+	3.81
Medelinkomst	3.59
Totalt anmälda brott	3.33
Totalt antal brott mot liv och hälsa	2.71
BRP	2.68
Medelålder	2.24
Forskarutbildning	2.17
Utrikes födda	2.14
Kvinnor	1.84
Inskrivet arbetslösa	1.79
Total kommunal skattesats	1.75

Tabell 3: VIF-faktorerna för de förklarande variablerna

4.2 Med stegvis variabelselektion

Vi börjar med att undersöka linjäritetsantagandet för partierna genom att utgå från den fulla modellen och byta ut en av de kontinuerliga variablerna i taget

mot den kategoriska motsvarigheten för att veta vilka variabler som ska ingå i modellen. Sedan anpassar vi en multipel logistisk regressionsmodell och gör som beskrivet i teoridelen.

4.2.1 Moderaterna

Vid undersökningen av linjäritetsantagandet ser vi att de förklarande variablerna *BRP*, *Utrikes födda* och *Inskrivet arbetslösa* inte ser ut att vara linjära i log-oddsen. Log-odds mot kvartilmedelvärdesplottarna för dessa variabler finns i appendix, Figur 2. Vi tar då med dessa variabler som kategoriska i vår resterande analys.

De estimerade koefficienterna, z-testen och p-värdena för den första modellen finns i appendix, Tabell 9. Vi ser att alla de förklarande variablerna är statistiskt signifikanta enligt z-testet, alla har ett p-värde mindre än 0.05. Detta innebär att vi i detta steg inte utesluter några förklarande variabler och kan gå vidare till stegvis variabelselektion.

Vid stegvis variabelselektion inkluderas alla variabler i forward selection och ingen exkluderas i backward elimination. Den bästa modellen enligt stegvis variabelselektion blir då den fulla modellen med kategoriska variabler för *Inskrivet arbetslösa*, *Utrikes födda* och *BRP*. AIC värdet för modellen är 367 273 och finns även med i den gemensamma sammanställningen i Tabell 5.

4.2.2 Sverigedemokraterna

När vi undersöker linjäritetsantagandet får vi att de förklarande variablerna *BRP*, *Medelålder*, *Total kommunal skattesats*, *Forskarutbildning* och *Inskrivet arbetslösa* inte ser ut att vara linjära i log-oddsen. Log-odds mot kvartilplottarna för dessa förklarande variabler finns i appendix Figur 3. Vi går vidare med analysen och för dessa förklarande variabler använder vi de kategoriska motsvarigheterna.

Modellens estimerade koefficienter, z-testen och p-värdena för den första modellen finns i appendix, Tabell 10. Vi får signifikanta z-test för alla variabler, vilket betyder att vi inte exkluderar några förklarande variabler. Vi går vidare till stegvis variabelselektion.

För vår modell utesluts inga variabler i backward elimination och alla variabler inkluderas i forward selection. Detta betyder att den bästa modellen för Sverigedemokraterna, enligt stegvis variabelselektion, är den fulla modellen med kategoriska variabler för *BRP*, *Medelålder*, *Total kommunal skattesats*, *Forskarutbildning* och *Inskrivet arbetslösa*. AIC värdet för modellen är 433 187 och finns även med i den gemensamma sammanställningen i Tabell 5.

4.2.3 Socialdemokraterna

När vi undersöker linjäritetsantagandet för Socialdemokraterna får vi flera variabler som inte ser ut att vara approximativt linjära i log-oddsen. Dessa är *Totalt anmälda brott*, *Totalt antal brott mot liv och hälsa*, *BRP*, *Kvinnor*, *Eftergymnasial utbildning 3år+* och *Utrikes födda*. Log-odds mot kvartilplottarna för dessa förklarande variabler finns i appendix Figur 4. Vi går vidare med modellanalysen med de ovan nämnda förklarande variablerna som kategoriska variabler.

Den anpassade modellens estimerade koefficienter, z-testen och p-värdena finns i appendix, Tabell 11. Vi ser att vi får signifikanta p-värden för alla förklarande variabler förutom för *Totalt anmälda brott kvantil 3*, som har ett p-värde lika med 0.21. Som nämnt i teoridelen måste alla nivåer av en kategorisk förklarande variabel vara inkluderade eller exkluderade ur modellen. För att undersöka om modellen blir bättre utan alla nivåer för den kategoriska variabeln gör vi ett likelihood-kvot test mellan den fulla modellen och en enklare modell utan *Totalt anmälda brott*. Vi får ett p-värde klart mindre än 0.05 för likelihood-kvot testet, vilket betyder att vi förkastar nollhypotesen att den enklare modellen är tillräckligt bra. Vi forsätter vår analys med den fulla modellen.

Vid stegvis variabelselektion får vi ett resultat där alla förklarande variabler inkluderas i forward selection och inga exkluderas i backward elimination. Den bästa modellen för Socialdemokraterna, enligt stegvis variabelselektion, blir då den fulla modellen med kategoriska variabler för *Totalt anmälda brott*, *Totalt antal brott mot liv och hälsa*, *BRP*, *Kvinnor*, *Eftergymnasial utbildning 3år+* och *Utrikes födda*. AIC värdet för modellen är 305 494 och finns också med i den gemensamma sammanställningen i Tabell 5.

4.3 Med purposeful selection

I detta avsnitt kommer vi göra de sex första stegen i purposeful selection för varje parti för att få fram de *preliminärt färdiga modellerna* för partierna. I avsnitt 5 gör vi steg 7 som är goodness-of-fit tester och drar slutsatser om hur bra modellerna är. Modelleringen för Moderaterna och Sverigedemokraterna är relativt lika, så vi presenterar modelleringen för dem tillsammans och Socialdemokraterna för sig.

4.3.1 Moderaterna och Sverigedemokraterna

Steg 1: För båda partierna har alla förklarande variabler p-värden mindre än 0.25 i de univariata logistiska regressionsmodellerna. Detta betyder att vi inte exkluderar någon förklarande variabel för något av partierna i detta steg.

Steg 2: Vi anpassar en multipel logistisk regressionsmodell för de båda partier-

na, estimerade koefficienter, z-test och p-värden för modellerna finns samlade i appendix i Tabell 12 för Moderaterna och i Tabell 14 för Sverigedemokraterna. P-värdena för samtliga förklarande variabler är mindre än 0.05 för båda partierna, vilket betyder att vi inte heller i detta steg inte utesluter någon förklarande variabel. Vi behöver inte göra något likelihood-kvot test.

Steg 3 och 4: Då vi inte uteslöt någon förklarande variabel i steg 2 för något av partierna, finns det inget att göra för våra modeller i steg 3. Detsamma gäller för steg 1, vilket betyder att steg 4 utesluts. Modellerna med alla förklarande variabler blir våra preliminära huvudeffektsmodeller för Moderaterna och Sverigedemokraterna.

Steg 5: Då vi i detta steg har samma modeller som när vi undersökte linjäritetsantagandet i stegvis variabelselektion får vi samma resultat som då. För Moderaterna är det de förklarande variablerna *BRP*, *Utrikes födda* och *Inskrivet arbetslösa* som inte uppvisar något approximativt linjärt samband i log-oddsen och vi ersätter då dessa med de kategoriska variabler med fyra nivåer. För Sverigedemokraterna är det de förklarande variablerna *BRP*, *Medelålder*, *Total kommunal skattesats*, *Forskarutbildning* och *Inskrivet arbetslösa* som inte uppvisar ett approximativt linjärt samband i log-oddsen. Dessa förklarande variabler byts ut mot deras kategoriska motsvarighet. Log-odds mot kvartilplottarna för de förklarande variabler är samma som i stegvis variabelselektion och finns i appendix, för Moderaterna i Figur 2 och för Sverigedemokraterna i Figur 3. Vår huvudeffektsmodell för Moderaterna blir då den fulla modellen med de förklarande variablerna *BRP*, *Utrikes födda* och *Inskrivet arbetslösa* som kategoriska. Huvudeffektsmodellen för Sverigedemokraterna blir den fulla modellen med de förklarande variablerna *BRP*, *Medelålder*, *Totalt kommunal skattesats*, *Forskarutbildning* och *Inskrivet arbetslösa* som kategoriska.

Steg 6: För båda partierna får vi p-värden mindre än 0.05 när vi utför likelihood-kvot testerna för samspelsvariablerna, vilket innebär att alla samspelsvariabler förbättrar modellerna. När vi anpassar en ny multipel logistisk regressionsmodell med de nya samspelsvariablerna får vi för Moderaterna p-värden mindre än 0.05 för alla samspelsvariabler, förutom för variabeln *Totalt anmälda brott* och *Utrikes födda nivå 4*. Då vi antingen måste utesluta alla tre samspelseffektsnivåerna eller ha kvar alla utför vi ett likelihood-kvot test för att se om den enklare modellen är tillräckligt bra. Vi får ett p-värde som är mindre än 0.05 för likelihood-kvot testet, vilket betyder att den komplexa modellen med alla förklarande variabler och de fem samspelseffektsvariablerna blir vår *preliminärt färdiga modell* för Moderaterna. När vi anpassar en ny multipel logistisk regressionsmodell med samspelsvariablerna för Sverigedemokraterna får alla samspelsvariabler signifikanta p-värden och denna modell blir den *preliminärt färdiga modell* för Sverigedemokraterna. Estimerade koefficienter, z-test och p-värden för dessa modeller finns i appendix i Tabell 13 för Moderaterna, och Tabell 15 för Sverigedemokraterna. Moderaternas modell har ett AIC värde på 352 096 och Sverigedemokraternas har ett AIC värde på 328 448. AIC värdena

finns även med i den sammanställda AIC tabellen, Tabell 5.

4.3.2 Socialdemokraterna

Steg 1: För alla våra univariata modeller får vi p-värden mindre än 0.25, så i detta steg utesluter vi inga förklarande variabler.

Steg 2: Vi anpassar en multipel logisitsk regressionsmodell med alla förklarande variabler och de estimerade koefficienterna, z-testen och p-värdena finns i appendix i Tabell 16. Vi får signifikanta resultat för alla förklarande variabler förutom för *Totalt anmälda brott* som har ett p-värde på 0.052. Vi anpassar en ny modell utan *Totalt anmälda brott* och gör ett likelihood-kvot test för att se om vi kan fortsätta med den enklare modellen. Vi får ett p-värde på 0.052 för vårt likelihood-kvot test, vilket betyder att vi kan gå vidare med den enklare modellen.

Steg 3: Vi beräknar $\Delta\hat{\beta}_i$ för koefficienterna i den enklare modellen, dessa finns i Tabell 4. Alla absolutvärden för $\Delta\hat{\beta}_i$ är mindre än 0.2, det högsta är 0.12, vilket gör att vi kan forsätta vår analys med den enklare modellen.

Förklarande variabel	$\Delta\beta_i$
Totalt antal brott mot liv och hälsa	0.1201
Eftergymnasial utbildning 3år+	0.0085
Medelålder	0.0024
Medelinkomst	0.002
Forskarutbildning	0.0004
Inskrivet arbetslösa	-0.001
Kvinnor	-0.0043
BRP	-0.0049
Total kommunal skattesats	-0.0076
Utrikes födda	-0.0087

Tabell 4: $\Delta\hat{\beta}_i$ värdena för Socialdemokraterna i purposeful selection

Steg 4: Då vi inte uteslöt någon variabel i steg 1 finns det ingen analys att göra i detta steg. Vår preliminära huvudeffektsmodell blir modellen utan variabeln *Totalt anmälda brott*.

Steg 5: Vi undersöker linjäritetsantagandet för våra förklarande variabler och ser att variablerna *Totalt antal brott mot liv*, *BRP*, *Kvinnor*, *Eftergymnasial utbildning 3år+* och *Utrikes födda* inte är approximativt linjära i log-oddsen. Plottarna finns i appendix, Figur 4. Vår huvudeffektsmodell blir då modellen med de ovan nämnda variablerna som kategoriska och utan *Totalt anmälda brott*.

Steg 6: Eftersom vi i steg 2 uteslöt *Totalt anmälda brott* tar vi inte med de samspelsvariablerna där *Totalt anmälda brott* är delaktig. Vi undersöker de övriga tre samspelen och får signifikanta resultat i likelihood-kvot testerna och vår *preliminärt färdiga modell* blir som huvudeffektsmodellen och de tre samspelsvariablerna. Tabell med koefficienter, z-test och p-värden finns i appendix Tabell 17. AIC värdet för modellen är 294 369 och som de med de andra finns AIC värdet med i den sammanställda tabellen nedan, Tabell 5.

Parti	Modell	AIC
Moderaterna	Stegvis	367 273
	Purposeful	352 096
Sverigedemokraterna	Stegvis	433 187
	Purposeful	328 448
Socialdemokraterna	Stegvis	305 494
	Purposeful	294 369

Tabell 5: AIC värden för våra anpassade modeller

5 Modellutvärdering

För att kunna beräkna de prediktionsmått vi presenterade i avsnitt 3.6 med hjälp av paket och funktioner i R krävs data på individnivå. På grund av begränsat datorminne har detta inte varit möjligt att göra. Manuella beräkningar hade varit möjliga att göra, men inom tidsramen för detta arbete har vi valt att göra på följande sätt: En kommun klassificeras som ett lyckat utfall om andelen röster på partiet av intresse är större för kommunen än medianandelen för partiet i Sverige. Om andelen är mindre än medianandelen klassificeras utfallet som misslyckat.

För alla utvärderingsmått använder vi de anpassade modellerna från avsnitt 4 och våra testset för partierna, för att se hur bra modellerna fungerar på okänd data.

5.1 ROC/AUC

Vi beräknar sensitivitet och specificitet för våra modeller för olika värden på π_0 och plottar ROC kurvorna. Dessa finns i appendix, ROC kurvorna för Moderaterna i Figur 5, Sverigedemokraterna i Figur 6 och Socialdemokraterna i Figur 7. AUC värderna för modellerna finns också sammanställda i Tabell 6.

Det vi ser är att båda modellerna för Moderaterna klassificeras som bra mo-

deller enligt Hosmer et al.. Socialdemokraterna har värden precis under 0.8, vilket gör att modellerna klassas som godkända. Sverigedemokraternas modeller är dock knappt bättre än gissningar och klassas som relativt dåliga modeller. Det vi ser är att för både Moderaterna och Socialdemokraterna har modellen som anpassats med stegvis variabelselektion något högre AUC värden, vilket indikerar på att dessa är något bättre än de modeller valda genom purposeful selection. För Sverigedemokraterna kan vi dra slutsatsen att båda modellerna är dåliga enligt ROC/AUC.

Parti	Modell	AUC
Moderaterna	Stegvis	0.83
	Purposeful	0.821
Sverigedemokraterna	Stegvis	0.528
	Purposeful	0.529
Socialdemokraterna	Stegvis	0.798
	Purposeful	0.783

Tabell 6: AUC värden för våra modeller

5.2 Hosmer-Lemeshow test

I Tabell 7 finns våra resultat från Hosmer-Lemeshow testen för våra modeller. Vi ser att vi får p-värden klart mindre än 0.05 för alla modeller, vilket gör att vi kan förkasta nollhypotesen att de observerade och predikterade sannolikheterna är desamma för alla grupper. Detta kan, som vi nämnde i teroidelen, bero på att vi har ett stort antal observationer, vilket gör att mycket små skillnader mellan de predikterade och observerade sannolikheterna fångas upp. Vi kan inte säga att modellerna är dåliga enligt detta test, men de ser inte heller bra ut.

Parti	Modell	X^2	P-värde
Moderaterna	Stegvis	212.11	$<$ 2e-16
	Purposeful	192.31	$<$ 2e-16
Sverigedemokraterna	Stegvis	447.39	$<$ 2e-16
	Purposeful	209-09	$<$ 2e-16
Socialdemokraterna	Stegvis	140.46	$<$ 2e-16
	Purposeful	98.94	$<$ 2e-16

Tabell 7: Hosmer-Lemeshow test resultat

5.3 Tränings- och testfel

I Tabell 8 har vi sammanställt tränings- och testfelen för våra modeller. Överlag kan vi se att tränings- och testfelen är stora. Prediktionsfel större än 4% ger oss stora prediktionsintervall. Vi går igenom resultaten parti för parti.

För Moderaterna är träningsfelen för båda modellerna väldigt lika, medan testfelet för modellen vald av stegvis variabelselektion är 0.5 procentenheter lägre. Detta indikerar att modellen vald med purposeful selection kan vara överanpassad. Utifrån detta test skulle vi säga att modellen anpassad med stegvis variabelselektion är bättre än den anpassad med purposeful selection.

Sverigedemokraterna har relativt låga träningsfel för båda modellerna, under 4%, men extremt höga testfel på ungefär 10%. Detta betyder att båda modellerna troligen är överanpassade till träningssetet och ingen av dessa modeller bör användas i prediktionssyfte.

Träningsfelen för Socialdemokraterna är lika för båda modellerna, under 6%. Vi gör samma observation för Socialdemokraterna som för Moderaterna, testfelet för purposeful selection modellen är klart större än för stegvis variabelselektionsmodellen. Detta tyder på att vi, även för Socialdemokraterna, överanpassat purposeful selection modellen.

Parti	Modell	Typ av fel	Fel i %
Moderaterna	Stegvis	Träning	4.71%
		Test	6.98%
Moderaterna	Purposeful	Träning	4.73%
		Test	7.53%
Sverigedemokraterna	Stegvis	Träning	3.81%
		Test	9.53%
Sverigedemokraterna	Purposeful	Träning	3.32%
		Test	10.98%
Socialdemokraterna	Stegvis	Träning	5.99%
		Test	5.44%
Socialdemokraterna	Purposeful	Träning	5.73%
		Test	7.7%

Tabell 8: RMSEP för våra modeller

6 Resultat

Från testerna vi genomförde i avsnitt 5 kan vi dra slutsatsen att ingen av modellerna är bra på att prediktera valresultaten per kommun för något av par-

tierna. Modellerna för Moderaterna hade bra AUC värden, men både Hosmer-Lemeshow testerna och tränings- och testfelen visar på sämre resultat. Om vi ska bestämma vilken av modellerna som är den bättre för Moderaterna, utav den anpassad med stegvis variabelselektion eller purpouseful selection, väljer vi stegvis variabelselektions modellen. Den har förvisso ett högre AIC-värde än purposeful selectionmodellen, men också något högre AUC värde och lägre testfel.

För Socialdemokraterna hade vi godkända AUC värden, men höga tränings- och testfel och små p-värden för Hosmer-Lemeshow testen. Men om vi, som för Moderaterna, ska välja den bättre modellen av den anpassad med stegvis variabelselektion och den anpassad med purposeful selection väljer vi, trots högre AIC-värde, den modell som anpassats med stegvis variabelselektion. Den gav något högre AUC värde och mer än två procentenheter lägre testfel.

Då båda modellerna för Sverigedemokraterna gav dåliga AUC värden, mycket små p-värden för Hosmer-Lemeshow testen och mycket höga testfel är det inte intressant att bestämma vilken av modellerna som är den bättre modellen av de två. Vi nöjer oss med slutsatsen att båda modellerna är dåliga prediktionsmodeller.

7 Diskussion

Som vi skrev i avsnitt 6 kan vi dra slutsatsen att våra modeller inte är bra på att prediktera valresultaten per kommun för något av partierna, speciellt inte för Sverigedemorkraterna. Att prediktionsförmågan för Sverigedemokraternas modeller är sämre än för Moderaterna och Socialdemokraterna kan beor på att datamaterialet vi använde när vi anpassat modellerna är valresultaten från riksdagsvalen 2002 - 2014. Under denna tidsperiod gick Sverigedemokraterna från att inte ingå i Riksdagen till att få 12% av rösterna vid Riksdagsvalet 2014 och under de nästkommande fyra åren, fram till valet 2018, ökade Sverigedemokraternas stöd ytterligare till 17%. Detta är en enorm framväxt som tydligen inte kunde fångas upp av våra modeller. Socialdemokraterna och Moderaterna har varit mer stabila under 2000-talet, vilket kan förklara att dessa modeller var något bättre på att prediktera testseten.

Det hade varit mer intressant att analysera de manuella ROC/AUC beräkningar på individnivå än den analysen vi har i denna uppsats, men på grund av begränsat datorminne och tidsramen för kandidatarbetet var det inte möjligt för detta arbete. Det hade också varit intressant att ta med förklarande variabler som hade kunnat kvantifiera andra viktiga valfrågor. Enligt VALU (https://www.svt.se/special/valu2018-valjarnas-viktigaste-fragor/) har sjukvården och äldreomsorgen varit viktiga frågor för väljarna under hela 2000-talet och hade varit intressanta att ha med i analysen. Det som är svårt är att kvantifiera

dessa på ett rättvist sätt mellan kommuner och län.

Vår slutsats av detta arbete är att vi inte kunde göra en bra prediktionsmodell på den offentliga data vi valde att analysera. Moderaterna och Socialdemokraterna fick något bättre resultat än Sverigedemokraterna, men ingen av modellerna rekommenderas att använda för att prediktera valresultat i framtiden.

8 Referenslista

Agresti, A. (2013): *Categorical data analysis*. 3rd edition. Hoboken: John Wiley & Sons.

Hastie, T., Tibshirani, R., Friedman, J. (2009): *The Elements of Statistical Learning*. 2nd edition. New York: Springer New York. DOI: 10.1007/b94608.

Held, L., Sabanés Bové, D. (2014): *Applied Statistical Inference*. Springer-Verlag Berlin Heidelberg.

Hosmer Jr, D.W., Lemeshow, S., Sturdivant, R.X. (2013): *Applied Logistic Regression*. 3rd edition. Hoboken: John Wiley & Sons.

Kutner, M. H., Nachtsheim, C. J., Neter, J., Li, W. (2004): *Applied Linear Statistical Models*. 5th edition. New York: McGraw-Hill/Irwin

Sundberg, R. (2016): *Lineära Statistiska Modeller*. Opubliserat manuskript. Stockholms universitet.

Wiest, M.M., Lee, K.J., Carlin, J.B. (2015): Statistics for clinicians: An introduction to logistic regression. *Journal of Paediatrics and Child Health*. vol. 51, ss. 670 - 673. DOI: 10.1111/jpc.12895

8.1 Datamaterial

Arbetsförmedlingen:
Förklarande variabel: *Inskrivet arbetslösa*.
Tabell: *Arbetssökande 1996 - 2019*. Hämtad: 2020-04-19.
Hemsida: https://arbetsformedlingen.se/om-oss/statistik-och-analyser/statistik/tidigare-statistik

BRÅ:
Förklarande variabler: *Totalt anmälda brott* och *Totalt antal brott mot liv och hälsa*.
Dataset: *Årsvis - Kommun och storstädernas stadsdelar 1996-*.
Hämtad: 2020-04-20.
Hemsida: http://statistik.bra.se/solwebb/action/index

Statistiska Centralbyrån (SCB):
Hemsida: http://www.statistikdatabasen.scb.se/pxweb/sv/ssd/.

Förklarande variabler: *Eftergymnasial utbildning 3år+* och *Forskarutbildning*.
Tabell: *Befolkning 16-74 år efter kön, ålder, nationell bakgrund, utbildningsnivå*

och utbildningsinriktning SUN 2000. År 2000 - 2018. Hämtad: 2020-03-19.

Förklarande variabel: *Medelålder*.
Tabell: *Befolkningens medelålder efter region och kön. År 1998 - 2019*. Hämtad: 2020-04-19.

Förklarande variabel: *BRP*.
Tabell: *Bruttoregionprodukt (BRP), sysselsatta och löner (ENS2010) efter region (län, riksområde). År 2000 - 2018*. Hämtad: 2020-04-19.

Förklarande variabel: *Kvinnor*.
Tabell *Folkmängden efter region, civilstånd, ålder och kön. År 1968 - 2019*. Hämtad: 2020-02-17.

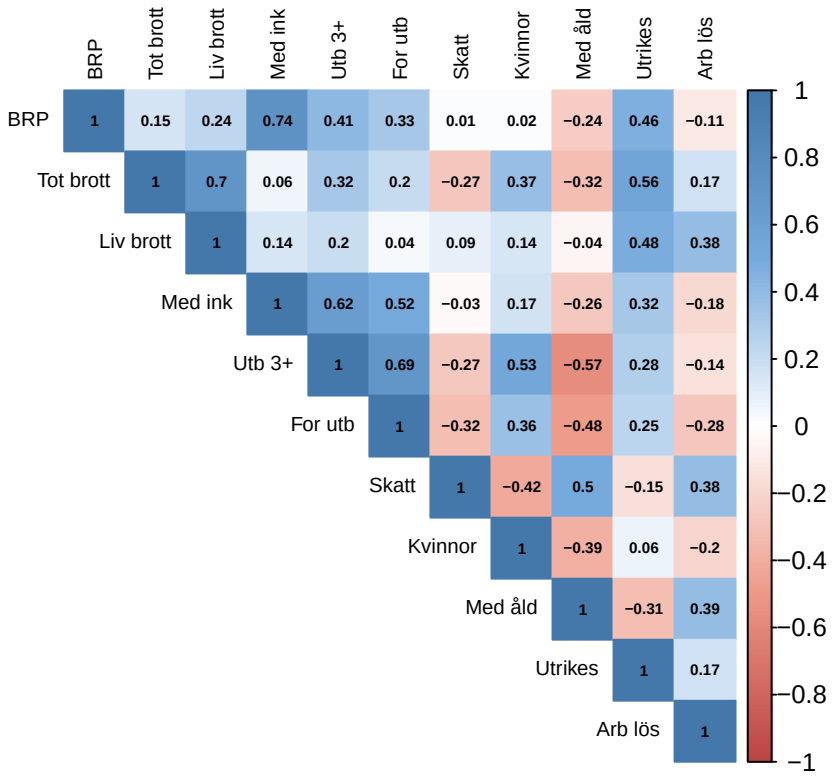
Förklarande variabel: *Utrikes födda*.
Tabell: *Inrikes och utrikes födda efter region, ålder och kön. År 2000 - 2019*. Hämtad: 2020-04-19.

Förklarande variabel: *Total kommunal skattesats*.
Tabell: *Kommunala skattesatser. År 2000 - 2020*. Hämtad: 2020-04-19.

Förklarande variabel: *Medelinkomst*.
Tabell: *Sammanräknad förvärvsinkomst för boende i Sverige hela året efter region, kön, ålder och inkomstklass. År 1999 - 2018*. Hämtad: 2020-04-19.

A Appendix Figurer

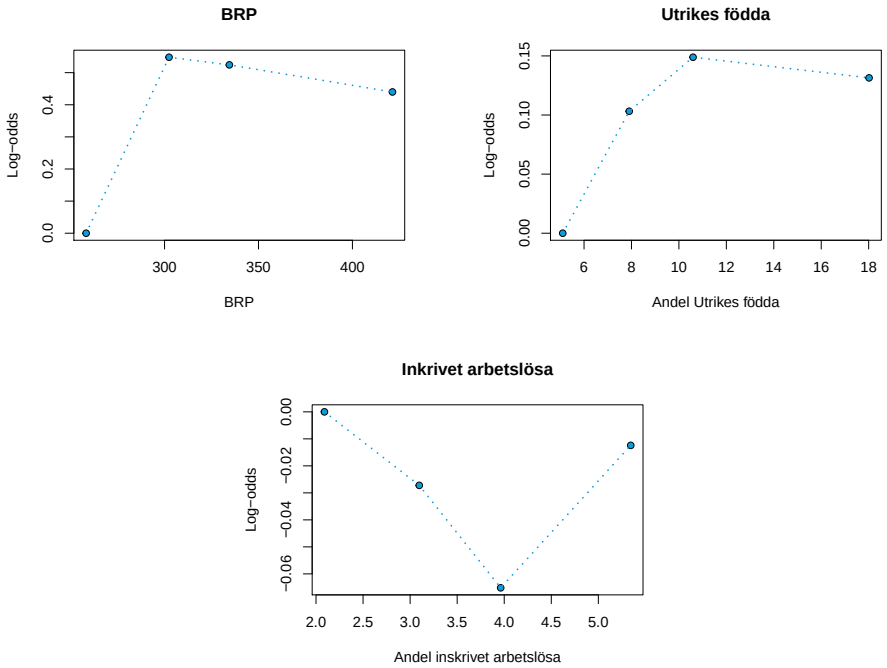
A.1 Korrelationsmatris



Figur 1: Korrelationsmatris för våra förklarande variabler

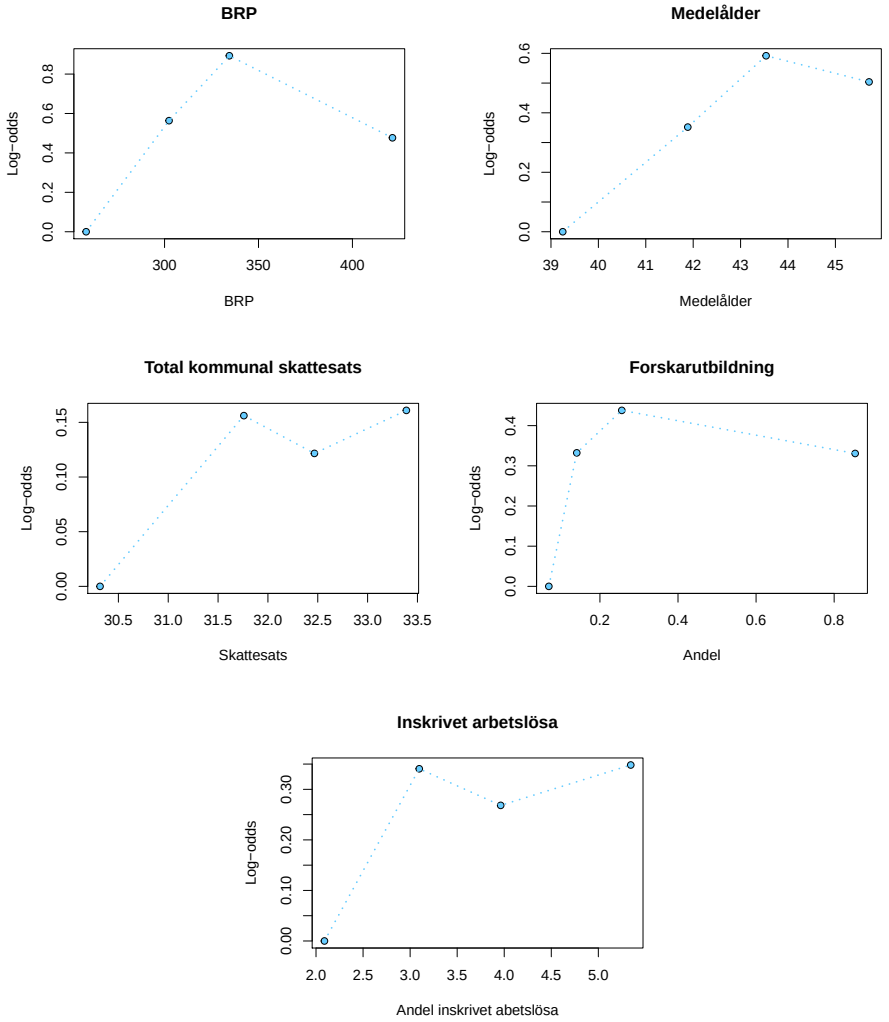
A.2 Linjäritet i log-odds

A.2.1 Moderaterna



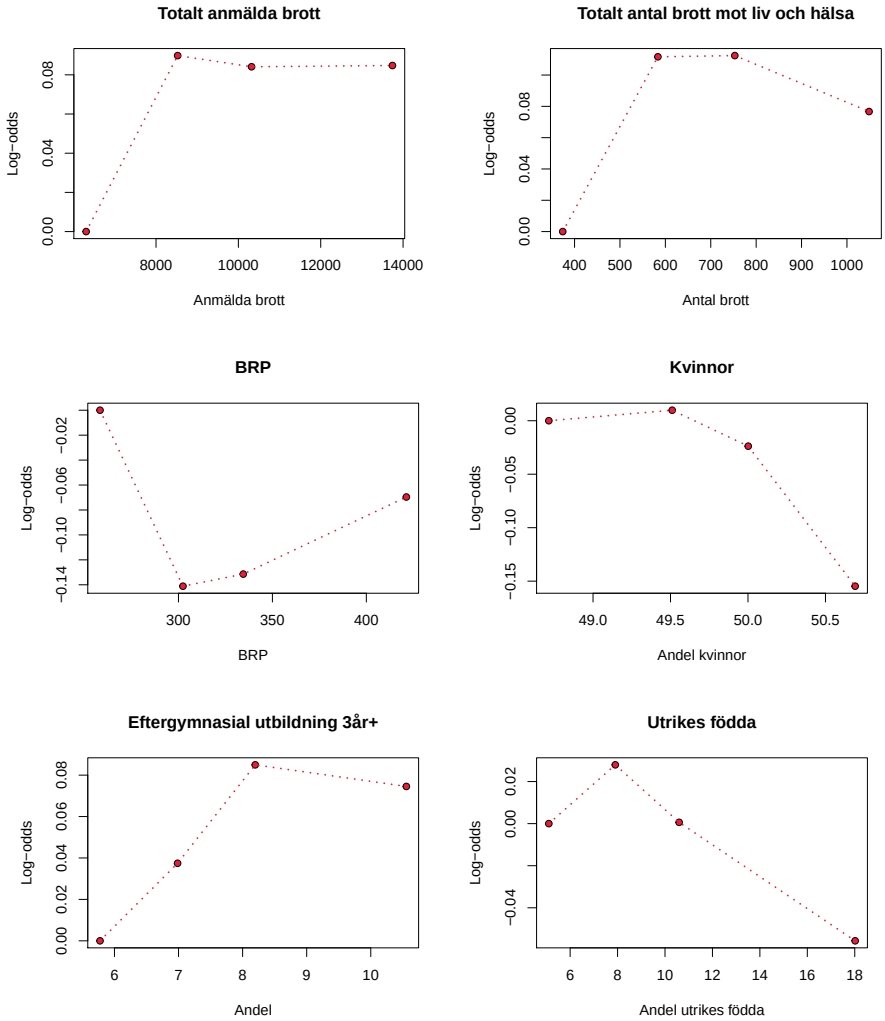
Figur 2: Log-odds mot kvantilerplottar för Moderaterna

A.2.2 Sverigedemokraterna



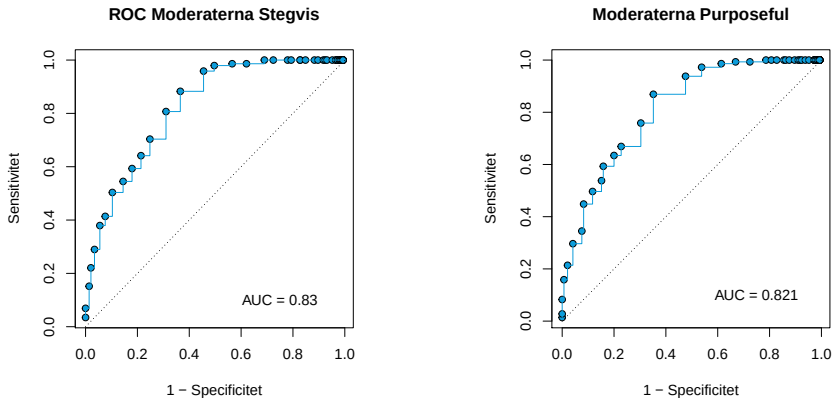
Figur 3: Log-odds mot kvantilerplottar för Sverigedemokraterna

A.2.3 Socialdemokraterna

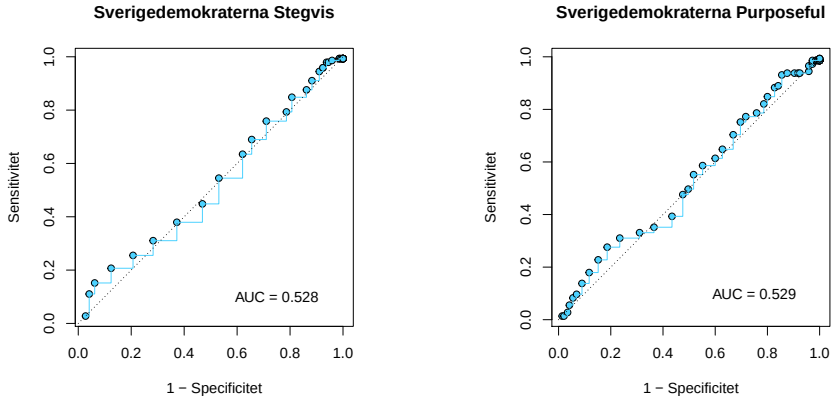


Figur 4: Log-odds mot kvantilerplottar för Socialdemokraterna

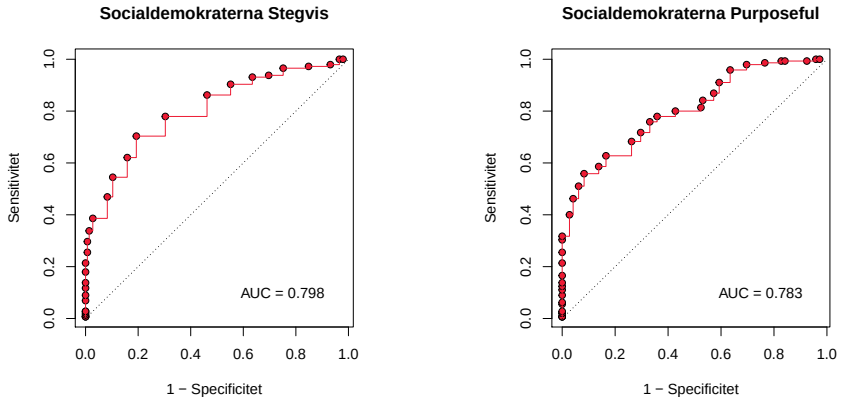
A.3 ROC figurer



Figur 5: ROC kurvor och AUC värden för modellerna för Moderaterna



Figur 6: ROC kurvor och AUC värden för modellerna för Sverigedemokraterna



Figur 7: ROC kurvor och AUC värden för modellerna för Socialdemokraterna

B Appendix Tabeller

B.1 Stegvis variabelselektion

Förklarande variabel	Estimerade koeff	z-värden	Pr(> |z|)
Intercept	-6.4	-119.76	$< 2 \cdot 10^{-16}$
Totalt anmälda brott	-0.000026	-94.21	$< 2 \cdot 10^{-16}$
Totalt antal brott mot liv och hälsa	0.00021	58.8	$< 2 \cdot 10^{-16}$
Medelinkomst	0.061	284.91	$< 2 \cdot 10^{-16}$
BRP kvantil 2	0.55	280.37	$< 2 \cdot 10^{-16}$
BRP kvantil 3	0.51	245.69	$< 2 \cdot 10^{-16}$
BRP kvantil 4	0.45	201.45	$< 2 \cdot 10^{-16}$
Total kommunal skattesats	-0.053	-104.1	$< 2 \cdot 10^{-16}$
Kvinnor	0.14	146.51	$< 2 \cdot 10^{-16}$
Medelålder	-0.046	-136	$< 2 \cdot 10^{-16}$
Eftergymnasial utbildning 3år+	-0.0098	-19.71	$< 2 \cdot 10^{-16}$
Forskarutbildning	-0.0089	-8.45	$< 2 \cdot 10^{-16}$
Utrikes födda kvantil 2	0.055	27.85	$< 2 \cdot 10^{-16}$
Utrikes födda kvantil 3	0.078	39.93	$< 2 \cdot 10^{-16}$
Utrikes födda kvantil 4	0.025	11.52	$< 2 \cdot 10^{-16}$
Inskrivet arbetslösa kvantil 2	-0.0095	-5.86	$\mathbf{4.7 \cdot 10^{-9}}$
Inskrivet arbetslösa kvantil 3	-0.042	-22.71	$< 2 \cdot 10^{-16}$
Inskrivet arbetslösa kvantil 4	-0.023	-11.06	$< 2 \cdot 10^{-16}$

Tabell 9: Första modellen för Moderaterna i stegvis variabelselektion

Förklarande variabel	Estimerade koeff	z-värden	Pr(> |z|)
Intercept	6.01	70.88	$< 2 \cdot 10^{-16}$
Totalt anmälda brott	-0.00002	-38.22	$< 2 \cdot 10^{-16}$
Totalt antal brott mot liv och hälsa	-0.00011	-16.8	$< 2 \cdot 10^{-16}$
Medelinkomst	0.14	345.56	$< 2 \cdot 10^{-16}$
BRP kvantil 2	0.62	133.97	$< 2 \cdot 10^{-16}$
BRP kvantil 3	0.92	196.43	$< 2 \cdot 10^{-16}$
BRP kvantil 4	0.51	100.83	$< 2 \cdot 10^{-16}$
Total kommunal skattesats kvantil 2	0.12	43.74	$< 2 \cdot 10^{-16}$
Total kommunal skattesats kvantil 3	0.18	60.72	$< 2 \cdot 10^{-16}$
Total kommunal skattesats kvantil 4	0.23	71.92	$< 2 \cdot 10^{-16}$
Kvinnor	-0.26	-149.63	$< 2 \cdot 10^{-16}$
Medelålder kvantil 2	0.41	138.49	$< 2 \cdot 10^{-16}$
Medelålder kvantil 3	0.48	123.75	$< 2 \cdot 10^{-16}$
Medelålder kvantil 4	0.43	97.58	$< 2 \cdot 10^{-16}$
Eftergymnasial utbildning 3år+	-0.032	-35.9	$< 2 \cdot 10^{-16}$
Forskarutbildning kvantil 2	0.25	60.82	$< 2 \cdot 10^{-16}$
Forskarutbildning kvantil 3	0.31	76	$< 2 \cdot 10^{-16}$
Forskarutbildning kvantil 4	0.17	33.49	$< 2 \cdot 10^{-16}$
Utrikes födda	0.039	181	$< 2 \cdot 10^{-16}$
Inskrivet arbetslösa kvantil 2	0.38	113.3	$< 2 \cdot 10^{-16}$
Inskrivet arbetslösa kvantil 3	0.45	125.14	$< 2 \cdot 10^{-16}$
Inskrivet arbetslösa kvantil 4	0.59	154.51	$< 2 \cdot 10^{-16}$

Tabell 10: Första modellen för Sverigedemokraterna i stegvis variabelselektion

Förklarande variabel	Estimerade koeff	z-värden	Pr($>\|z\|$)
Intercept	-2.21	-124.14	$< 2 \cdot 10^{-16}$
Totalt anmälda brott kvantil 2	0.031	16.27	$< 2 \cdot 10^{-16}$
Totalt anmälda brott kvantil 3	-0.0027	-1.25	**0.21**
Totalt anmälda brott kvantil 4	-0.015	-5.96	$2.5 \cdot 10^{-9}$
Totalt antal brott mot liv och hälsa kvantil 2	0.13	71.1	$< 2 \cdot 10^{-16}$
Totalt antal brott mot liv och hälsa kvantil 3	0.14	70.47	$< 2 \cdot 10^{-16}$
Totalt antal brott mot liv och hälsa kvantil 4	0.11	47.71	$< 2 \cdot 10^{-16}$
Medelinkomst	-0.05	-211.87	$< 2 \cdot 10^{-16}$
BRP kvantil 2	-0.16	-101.74	$< 2 \cdot 10^{-16}$
BRP kvantil 3	-0.16	-91.11	$< 2 \cdot 10^{-16}$
BRP kvantil 4	-0.12	-64.14	$< 2 \cdot 10^{-16}$
Total kommunal skattesats	0.042	99.32	$< 2 \cdot 10^{-16}$
Kvinnor kvantil 2	0.017	9.26	$< 2 \cdot 10^{-16}$
Kvinnor kvantil 3	-0.027	-14.47	$< 2 \cdot 10^{-16}$
Kvinnor kvantil 4	-0.15	-75.1	$< 2 \cdot 10^{-16}$
Medelålder	0.02	65.55	$< 2 \cdot 10^{-16}$
Eftergymnasial utbildning kvantil 2	0.027	13.95	$< 2 \cdot 10^{-16}$
Eftergymnasial utbildning kvantil 3	0.083	41.61	$< 2 \cdot 10^{-16}$
Eftergymnasial utbildning kvantil 4	0.087	37.93	$< 2 \cdot 10^{-16}$
Forskarutbildning	-0.047	-55.55	$< 2 \cdot 10^{-16}$
Utrikes födda kvantil 2	0.011	6.81	$9.5 \cdot 10^{-12}$
Utrikes födda kvantil 3	-0.02	-11.77	$< 2 \cdot 10^{-16}$
Utrikes födda kvantil 4	-0.067	-36.06	$< 2 \cdot 10^{-16}$
Inskrivet arbetslösa	0.053	112.53	$< 2 \cdot 10^{-16}$

Tabell 11: Första modellen för Socialdemokraterna i stegvis variabelselektion

B.2 Purposeful selection

B.2.1 Moderaterna

Förklarande variabel	Estimerade koeff	z-värden	Pr(> |z|)
Intercept	-6.77	-123.83	$< 2 \cdot 10^{-16}$
Totalt anmälda brott	-0.000054	-197.83	$< 2 \cdot 10^{-16}$
Totalt antal brott mot liv och hälsa	0.00061	176.19	$< 2 \cdot 10^{-16}$
Medelinkomst	0.07	265.97	$< 2 \cdot 10^{-16}$
BRP	0.00031	28.32	$< 2 \cdot 10^{-16}$
Total kommunal skattesats	-0.042	-82.21	$< 2 \cdot 10^{-16}$
Kvinnor	0.12	127.52	$< 2 \cdot 10^{-16}$
Medelålder	-0.023	-65.4	$< 2 \cdot 10^{-16}$
Eftergymnasial utbildning 3år+	0.012	24.8	$< 2 \cdot 10^{-16}$
Forskarutbildning	-0.0094	-9.18	$< 2 \cdot 10^{-16}$
Utrikes födda	0.0049	36.3	$< 2 \cdot 10^{-16}$
Inskrivet arbetslösa	-0.026	-42.24	$< 2 \cdot 10^{-16}$

Tabell 12: Första modellen för Moderaterna i purposeful selection

Förklarande variabel	Estimerade koeff	z-värden	Pr(> |z|)
Intercept	-9.49	-91.95	$< 2 \cdot 10^{-16}$
Totalt anmälda brott	-0.000046	-60.58	$< 2 \cdot 10^{-16}$
Totalt antal brott mot liv och hälsa	-0.000019	-2.27	**0.023**
Medelinkomst	0.27	58.5	$< 2 \cdot 10^{-16}$
BRP kvantil 2	0.52	257.05	$< 2 \cdot 10^{-16}$
BRP kvantil 3	0.48	217.29	$< 2 \cdot 10^{-16}$
BRP kvantil 4	0.43	185.09	$< 2 \cdot 10^{-16}$
Total kommunal skattesats	0.014	5.034	$4.81 \cdot 10^{-7}$
Kvinnor	0.14	133.82	$< 2 \cdot 10^{-16}$
Medelålder	-0.039	-113.15	$< 2 \cdot 10^{-16}$
Eftergymnasial utbildning 3år+	0.1	55.03	$< 2 \cdot 10^{-16}$
Forskarutbildning	0.043	36.54	$< 2 \cdot 10^{-16}$
Utrikes födda kvantil 2	-0.16	-18.7	$< 2 \cdot 10^{-16}$
Utrikes födda kvantil 3	-0.12	-14.58	$< 2 \cdot 10^{-16}$
Utrikes födda kvantil 4	0.03	3.61	**0.00031**
Inskrivet arbetslösa kvantil 2	-0.083	-11.64	$< 2 \cdot 10^{-16}$
Inskrivet arbetslösa kvantil 3	-0.29	-38.74	$< 2 \cdot 10^{-16}$
Inskrivet arbetslösa kvantil 4	-0.63	-76.76	$< 2 \cdot 10^{-16}$
Anmälda brott:Brott mot liv	2e-08	30.83	$< 2 \cdot 10^{-16}$
Anmälda brott:Utrikes födda kvantil 2	0.000021	23.61	$< 2 \cdot 10^{-16}$
Anmälda brott:Utrikes födda kvantil 3	0.000019	22.52	$< 2 \cdot 10^{-16}$
Anmälda brott:Utrikes födda kvantil 4	0.0000016	1.9	**0.057**
Utbildning 3+:Arbetslösa kvantil 2	0.0088	12.24	$< 2 \cdot 10^{-16}$
Utbildning 3+:Arbetslösa kvantil 3	0.027	35.32	$< 2 \cdot 10^{-16}$
Utbildning 3+:Arbetslösa kvantil 4	0.066	76.13	$< 2 \cdot 10^{-16}$
Utbildning 3+:Medelinkomst	-0.008	-80.85	$< 2 \cdot 10^{-16}$
Kommunalskatt:Medelinkomst	-0.004	-28.82	$< 2 \cdot 10^{-16}$

Tabell 13: Perliminärt färdiga modellen för Moderaterna i purposeful selection

B.2.2 Sverigedemokraterna

Förklarande variabel	Estimerade koeff	z-värden	Pr(> |z|)
Intercept	2.18	21.6	$< 2 \cdot 10^{-16}$
Totalt anmälda brott	-0.000037	-73.2	$< 2 \cdot 10^{-16}$
Totalt antal brott mot liv och hälsa	0.00051	85.61	$< 2 \cdot 10^{-16}$
Medelinkomst	0.22	480.89	$< 2 \cdot 10^{-16}$
BRP	-0.0043	-232.59	$< 2 \cdot 10^{-16}$
Total kommunal skattesats	0.1	104.26	$< 2 \cdot 10^{-16}$
Kvinnor	-0.33	-193.56	$< 2 \cdot 10^{-16}$
Medelålder	0.11	173.85	$< 2 \cdot 10^{-16}$
Eftergymnasial utbildning 3år+	0.05	52.56	$< 2 \cdot 10^{-16}$
Forskarutbildning	-0.28	-121.77	$< 2 \cdot 10^{-16}$
Utrikes födda	0.052	232.44	$< 2 \cdot 10^{-16}$
Inskrivet arbetslösa	0.0756	73.94	$< 2 \cdot 10^{-16}$

Tabell 14: Första modellen för Sverigedemokraterna i purposeful selection

Förklarande variabel	Estimerade koeff	z-värden	Pr(> |z|)
Intercept	0.56	5.85	$4.8 \cdot 10^{-9}$
Totalt anmälda brott	0.00006	56.19	$< 2 \cdot 10^{-16}$
Totalt antal brott mot liv och hälsa	0.00055	34.9	**0.023**
Medelinkomst	0.26	117.21	$< 2 \cdot 10^{-16}$
BRP kvantil 2	0.53	111.29	$< 2 \cdot 10^{-16}$
BRP kvantil 3	0.75	154.07	$< 2 \cdot 10^{-16}$
BRP kvantil 4	0.33	62.75	$< 2 \cdot 10^{-16}$
Kommunal skattesats kvantil 2	-2.48	-164.83	$< 2 \cdot 10^{-16}$
Kommunal skattesats kvantil 3	-3.82	-176.41	$< 2 \cdot 10^{-16}$
Kommunal skattesats kvantil 4	-2.96	-114.65	$< 2 \cdot 10^{-16}$
Kvinnor	-0.18	-97.76	$< 2 \cdot 10^{-16}$
Medelålder kvantil 2	0.26	86.95	$< 2 \cdot 10^{-16}$
Medelålder kvantil 3	0.47	119.58	$< 2 \cdot 10^{-16}$
Medelålder kvantil 4	0.44	100.48	$< 2 \cdot 10^{-16}$
Eftergymnasial utbildning 3år+	0.12	28.93	$< 2 \cdot 10^{-16}$
Forskarutbildning kvantil 2	-0.01	-2.33	**0.02**
Forskarutbildning kvantil 3	-0.063	-14.32	$< 2 \cdot 10^{-16}$
Forskarutbildning kvantil 4	-0.27	-51.38	$< 2 \cdot 10^{-16}$
Utrikes födda	0.031	42.13	$< 2 \cdot 10^{-16}$
Inskrivet arbetslösa kvantil 2	-0.2	-13.15	$< 2 \cdot 10^{-16}$
Inskrivet arbetslösa kvantil 3	-0.13	-8.6	$< 2 \cdot 10^{-16}$
Inskrivet arbetslösa kvantil 4	-0.54	-33.49	$< 2 \cdot 10^{-16}$
Anmälda brott:Brott mot liv	-6e-08	-55.16	$< 2 \cdot 10^{-16}$
Anmälda brott:Utrikes födda	1.5e-07	2.75	**0.0059**
Utbildning 3+:Arbetslösa kvantil 2	0.068	42.84	$< 2 \cdot 10^{-16}$
Utbildning 3+:Arbetslösa kvantil 3	0.065	40.18	$< 2 \cdot 10^{-16}$
Utbildning 3+:Arbetslösa kvantil 4	0.13	76.06	$< 2 \cdot 10^{-16}$
Utbildning 3+:Medelinkomst	-0.013	-66.5	$< 2 \cdot 10^{-16}$
Kommunalskatt kvantil 2:Medelinkomst	0.12	166.53	$< 2 \cdot 10^{-16}$
Kommunalskatt kvantil 3:Medelinkomst	0.19	181	$< 2 \cdot 10^{-16}$
Kommunalskatt kvantil 4:Medelinkomst	0.15	118.07	$< 2 \cdot 10^{-16}$

Tabell 15: Perliminärt färdiga modellen för Sverigedemokraterna i purposeful selection

B.2.3 Socialdemokraterna

Förklarande variabel	Estimerade koeff	z-värden	Pr(> |z|)
Intercept	3.044	63.74	$< 2 \cdot 10^{-16}$
Totalt anmälda brott	5e-07	1.95	**0.052**
Totalt antal brott mot liv och hälsa	0.00003	9.83	$< 2 \cdot 10^{-16}$
Medelinkomst	-0.056	-215.85	$< 2 \cdot 10^{-16}$
BRP	-0.000074	-7.36	$1.9 \cdot 10^{-13}$
Total kommunal skattesats	0.033	72.6	$< 2 \cdot 10^{-16}$
Kvinnor	-0.089	-103.7	$< 2 \cdot 10^{-16}$
Medelålder	0.0095	30.87	$< 2 \cdot 10^{-16}$
Eftergymnasial utbildning 3år+	0.02	44.72	$< 2 \cdot 10^{-16}$
Forskarutbildning	-0.085	-86.95	$< 2 \cdot 10^{-16}$
Utrikes födda	-0.01	-88.31	$< 2 \cdot 10^{-16}$
Inskrivet arbetslösa	0.051	99.2	$< 2 \cdot 10^{-16}$

Tabell 16: Första modellen för Socialdemokraterna i purposeful selection

Förklarande variabel	Estimerade koeff	z-värden	Pr(> |z|)
Intercept	0.25	3.24	**0.0012**
Totalt brott mot liv 2	0.1	59.65	$< 2 \cdot 10^{-16}$
Totalt brott mot liv 3	0.1	58.61	$< 2 \cdot 10^{-16}$
Totalt brott mot liv 4	0.08	41.34	$< 2 \cdot 10^{-16}$
Medelinkomst	-0.16	-39.39	$< 2 \cdot 10^{-16}$
BRP kvantil 2	-0.17	-113.54	$< 2 \cdot 10^{-16}$
BRP kvantil 3	-0.18	-104.8	$< 2 \cdot 10^{-16}$
BRP kvantil 4	-0.14	-74.75	$< 2 \cdot 10^{-16}$
Total kommunal skattesats	-0.068	-27.78	$< 2 \cdot 10^{-16}$
Kvinnor kvantil 2	0.039	20.98	$< 2 \cdot 10^{-16}$
Kvinnor kvantil 3	0.0064	3.36	**0.00078**
Kvinnor kvantil 4	-0.11	-56.48	$< 2 \cdot 10^{-16}$
Medelålder	0.018	59.56	$< 2 \cdot 10^{-16}$
Utbildning 3+ kvantil 2	0.46	26.72	$< 2 \cdot 10^{-16}$
Utbildning 3+ kvantil 3	1.1	65.71	$< 2 \cdot 10^{-16}$
Utbildning 3+ kvantil 4	1.22	79.96	$< 2 \cdot 10^{-16}$
Forskarutbildning	-0.05	-57.07	$< 2 \cdot 10^{-16}$
Utrikes födda kvantil 2	-0.0064	-4.016	$5.9 \cdot 10^{-5}$
Utrikes födda kvantil 3	-0.055	-32.98	$< 2 \cdot 10^{-16}$
Utrikes födda kvantil 4	-0.1	-54.78	$< 2 \cdot 10^{-16}$
Inskrivet arbetslösa	0.06	50.42	$< 2 \cdot 10^{-16}$
Utbildning 3+ kvantil 2:Arbetslösa	0.013	8.17	$3 \cdot 10^{-16}$
Utbildning 3+ kvantil 3:Arbetslösa	0.015	10.77	$< 2 \cdot 10^{-16}$
Utbildning 3+ kvantil 4:Arbetslösa	-0.021	-16.43	$< 2 \cdot 10^{-16}$
Utbildning 3+ kvantil 2:Medelinkomst	-0.034	-30.76	$< 2 \cdot 10^{-16}$
Utbildning 3+ kvantil 3:Medelinkomst	-0.07	-67.08	$< 2 \cdot 10^{-16}$
Utbildning 3+ kvantil 4:Medelinkomst	-0.069	-70.56	$< 2 \cdot 10^{-16}$
Kommunal skattesats:Medelinkomst	0.0057	45.68	$< 2 \cdot 10^{-16}$

Tabell 17: Perliminärt färdiga modellen för Socialdemokraterna i purposeful selection