LUNDS UNIVERSITET

EKONOMI-
HÖGSKOLAN

Politik i simplexrummet

Kompositionell dataanalys av valresultat

Författare: Theo Gleisner, Anhelina Lysobyk

Handledare: Jakob Bergman

Kandidatexamen i statistik höstterminen 2023

Abstract

In this thesis we discuss a potential problem of negative correlation arising when using ordinary statistical methods on compositional data. We discuss a less known statistical method, compositional data analysis, that stands on the pillars of the works of John Aitchison from the 1980-s and Pawlowsky-Glahn, Egozcue and Delgado from the 2015. Using results from the 2022 national parliament elections divided into 290 Swedish municipalities, we present how the method, together with multivariate regression, can be used to explore the effects of different social and economic aspects on how people tend to vote. We find compositional data analysis to be a powerful and underutilized tool for analyzing aggregate voter behavior, allowing an application of regression analysis to elections with results that are easy to visualize and present.

Inledning

Sveriges demokrati grundar sig på riksdagsval som hålls var fjärde år. Det har gjorts många undersökningar där olika socioekonomiska aspekter vägs in för att se vad som påverkar röstbeteende. Undersökningarna som genomförs använder ofta vanliga statistiska metoder som logistisk och linjär regressioner. Sådana studier fokuserar oftast på att bestämma faktorer som påverkar hur individer väljer, snarare än att betrakta ett helt valresultat som undersökningsvariabel. Problematiken med det senare uppstår i partiandelarnas korrelationsstruktur: då varje andel enbart kan ligga i ett spann mellan noll och ett, samt andelarna totalt sett måste summera till ett, så föreligger svårigheter med att tillämpa vanlig statistisk metodik på ett valresultat. Vi vill därför göra ett försök att använda en annan metod för att undersöka valdata. I denna kandidatuppsats kommer valdata i Sverige analyseras med hjälp av etablerad kompositionell dataanalys som metod, där kommuner betraktas som studieenheter. På så sätt kan vi anpassa en multivariat regression som förutspår sammansättningar av röstandelar.

Litteraturundersökning

Aspekter som påverkar hur man röstar är ett ämne många är intresserade av att undersöka. Typiskt sett så utgår sådana studier från opinionsundersökningar och analyserar partisympatisörers egenskaper, d.v.s. använder individer som studieenhet. Ett exempel är en artikel av Richard Örhvall och Mikaela Järnbert ”Partiernas sympatisörer” (2010). I denna artikel presenterar de analys av opinionsundersökningen från Statistiska Centralbyrå (SCB) där de undersöker röstbeteende baserat på kön, ålder och utbildning. Resultat de kommer fram till är bland annat att Vänsterpartiets sympatisörer tenderar att vara unga människor, medan Moderaternas sympatisörer är ofta i medelåldern och har högre inkomst. Miljöpartiets och Liberalernas väljare är ofta högutbildade och av Kristdemokraternas väljare finns det fler kvinnor än män.

Tillämpningar av kompositionell dataanalys på valresultat är desto färre; detta förefaller vara ett ganska outforskad ämne. En metod liknande den vi nedan presenterar används dock i Jonathan N. Katz och Gary Kings artikel *A Statistical Model for Multiparty Electoral Data* (1999), där de undersöker brittiska val för att utröna om det finns en tjänstgöringsfördel för sittande regeringspartier. Till skillnad från tidigare studier, som inte använt en kompositionell metodik, kan de påvisa en liten men tydlig fördel.

Bakgrund

Val i Sverige genomförs på så vis att svenska medborgare som är 18 år och äldre går till vallokaler och väljer personerna som ska representera de i riksdagen, regionen och kommunen, genom att fylla in en valsedel. Personerna som är medborgare i något land som ingår i EU, personerna har varit folkbokförda i Sverige i tre år och svenska medborgare har rätt att rösta i valet till region och kommun. Svenska medborgare har rätt att rösta i valet till region, kommun och riksdag. I denna uppsats kommer fokus ligga på val till riksdagen. På valdagens kväll räknas valsedlarna och det preliminära valresultatet annonseras. Cirka en vecka efter valdagen fastställs det slutliga valresultatet.

Perioden mellan valen kallas för mandatperiod. I riksdagen finns det 349 platser, som kallas för mandat. Mandat fördelas mellan partierna proportionellt, beroende på antal röster de får. För att ett parti ska kunna komma in i riksdagen och få mandat måste partiet få minst 4% i riksdagsvalet eller 12% i en valkrets. (Sveriges riksdag, 2023)

I valet till riksdagen delar man in landet i valkretsar. Sverige är indelat i 29 valkretsar. Till stor del motsvarar de län, men områden med mycket befolkning, som Skåne län, Stockholms län och Västra Götalands län är indelade i flera valkretsar. Eftersom valkretsar är uppdelade med avseende på befolkningen, täcker vissa valkretsar en mycket större geografisk yta än andra. För att kunna titta på olika socioekonomiska aspekter som påverkar valresultat i detalj och kunna se trender kommer kommuner undersökas istället för valkretsar. Röstfördelningen i en kommun betraktas som studieenheter.

Problemformulering

Platserna till riksdag är proportionellt fördelade mellan partierna. Detta innebär att partierna med största antal röster får också flest mandat i riksdagen. Därför pratar man ofta om andel röster ett visst parti får. Andelarna måste alltid summeras till ett, och om, till exempel, en andel ökar, måste minst en annan minska. Detta skapar negativ korrelation mellan andelarna, vilket kan leda till vilseledande resultat. Därför är vi intresserade av att undersöka kompositionell dataanalys, en statistisk metod som tillåter att analysera andelar av en helhet så att andelarna inte är negativt korrelerade.

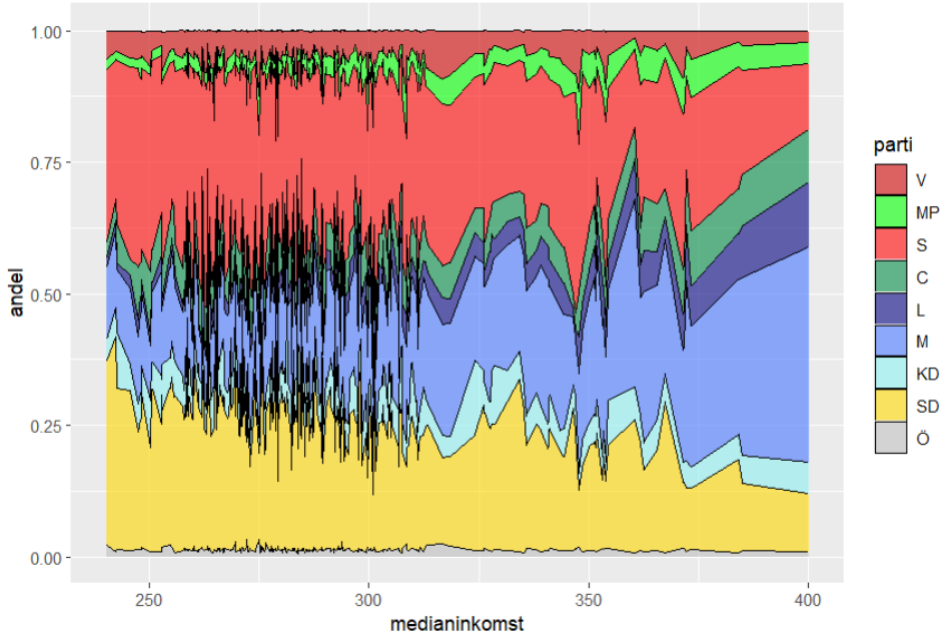
Syfte

Syftet med denna uppsats är att undersöka en metod som tillåter att prediktera röstfördelningen med regression. Val till riksdag är en viktig händelse och det finns därför intresse för trender om hur människor röstar och vilka aspekter påverkar deras val.

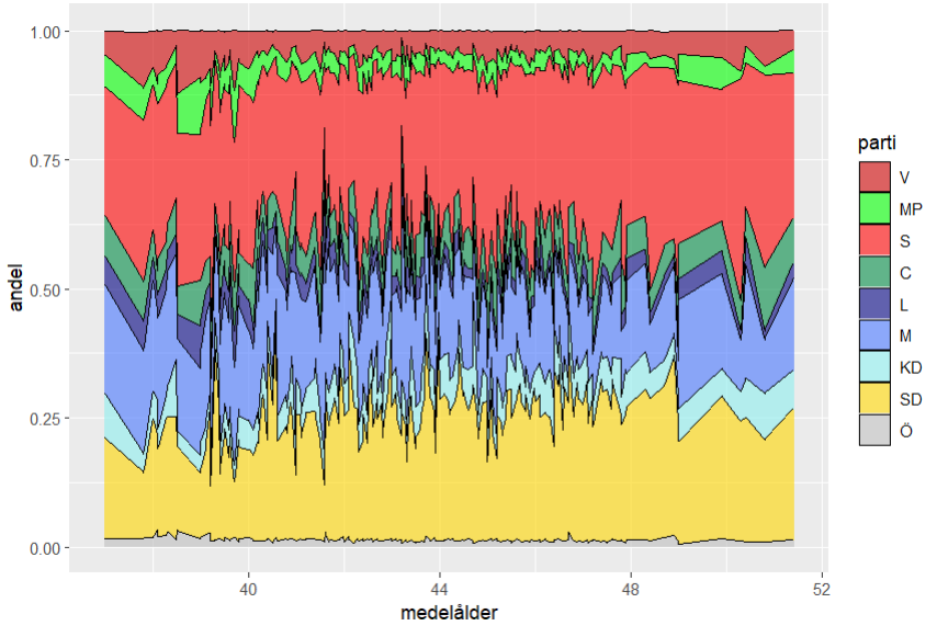
Data

Vi hämtar våra valdata från valmyndighetens webbsida (https://www.val.se/valresultat/riksdag-region-och-kommun/2022/radata-och-statistik.html#slutligtvalresultat). För varje kommun så finns röstandelar för de åtta riksdagspartierna, en kategori för övriga anmälda partier samt tre kategorier för ogiltiga röster. De röstandelar som används är riksdagspartierna samt övriga anmälda partier; vi har alltså 290 olika observationer (kommuner), där vardera består av nio andelar. Andelarna summerar i respektive kommun till ungefär 1, med små avvikelser som beror på att valmyndighetens data är avrundade till tre gällande siffror (detta utgör dock inte ett hinder för vår analys då vi enbart intresserar oss för andelarnas storlekar relativt varandra). Data om medelålder är hämtad från Statistiska Centralbyråns webbsida (https://www.statistikdatabasen.scb.se/pxweb/sv/ssd/START__BE__BE0101__BE0101B/BefolkMedianAlder/). Data om medianålder är hämtad från Statistiska Centralbyråns webbsida (https://www.statistikdatabasen.scb.se/pxweb/sv/ssd/START__HE__HE0110__HE0110A/SamForvInk1/).

Som en första visualisering av vårt datamaterial presenteras nedan två proportionella ytdiagram, med kommunens medianinkomst, figur 1, respektive medelålder, figur 2, på x-axlarna och de nio röstandelarna på y-axlarna. I figur 1 ser vi hur andel röster fördelas beroende på vilken inkomst man har. Diagrammen har en synlig störning mellan 250 och 310 tusen kronor på grund av att det finns väldigt många observationer, det vill säga kommuner, där medianinkomst ligger mellan dessa värden. I figur 2 ser vi hur andel röster fördelas beroende på ålder. Även här ser vi vissa störningar i diagrammet som beror på att det finns fler kommuner där medelålder är mellan 38 och 48 år än kommuner där medelålder är, exempelvis, 52 år. Dessa diagram ger även en visuell inblick i vad vi försöker simulera i de regressioner vi nedan kommer göra.



Figur 1. Empiriskt proportionellt ytdiagram. På y-axeln ser vi de olika partiernas röstandelar i riksdagsvalet kommunvis. X-axeln visar kommunens medianinkomst.

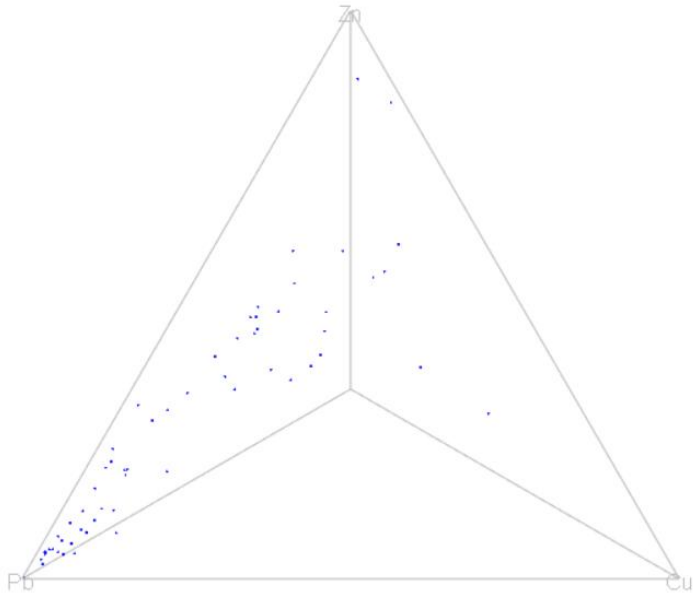


Figur 2. Proportionellt ytdiagram liknande figur 1. På y-axeln visas de olika partiernas röstandelar i riksdagsvalet kommunvis. X-axeln visar kommunens medelålder.

Kompositionell dataanalys

Röstfördelningen i en kommun är alltså ett exempel på en sammansättningsvariabel, eller en kompositionell variabel; andra exempel på sådana variabler är sammansättningen av mineraler i en sten eller djurarter i ett ekosystem. Metodiken i denna uppsats kretsar kring de särskilda egenskaper som sammansättningsvariabler har.

Andelar måste ligga mellan 0 och 1, och tillsammans måste andelarna i en sammansättning summera till 1. Geometriskt så betraktas ofta sammansättningar som en punkt på ett simplex, i dimensioner motsvarande antalet andelar minus 1. Sammansättningen av tre andelar kan till exempel representeras som en punkt på ett 2-simplex (en triangel), och sammansättningar av fyra andelar kan representeras som en punkt i ett 3-simplex (en tetraeder). En kommuns röstsammansättning, betraktat som en sammansättning av nio andelar (de åtta riksdagspartierna plus en andel för övriga partier) blir en punkt i ett 8-simplex. Eftersom det grafiskt sätt inte är möjligt att visualisera ett 8-diminsionellt simplex, visas nedan ett 3-dimensionellt simplex med simulerade data i figur 3. I figuren ser vi att många punkter ligger i nedre vänstra hörnet (Pb), vilket innebär att dessa observationer består av bly. Vidare ser vi att två observationer ligger i översta hörnet (Zn) och består till största del av zink. Observationer som ligger i mitten av tetraedern består av ungefär lika stora andelar av tre grundämnena.



Figur 3. Ett 3-simplex genererat i R där simulerade andelssammansättningar (i det här fallet grundämnen) plottas som punkter.

En vanlig ansats är att använda linjär regression med en andel, till exempel en procentsats, som beroende variabel. Linjär regression vilar emellertid på antagandet att den beroende variabelns väntevärde är normalfördelat kring de anpassade värdena, vilket är olämpligt för andelar; en sådan modell kan ge predikterade andelar utanför spannet [0, 1] med

konfidensintervall som skrider utanför dessa gränser. Att studera andelar en och en utnyttjar heller inte andelarnas relativa natur, d.v.s. att de konstant summerar till 1. En modell som förklarar röstsammansättningar bör ta hänsyn till att den beroende variabeln är kompositionell och förutspå röstssammansättningar som koordinater i ett simplex.

Den skotske statistikern John Aitchison har varit formativ för den moderna kompositionella dataanalysen. Han presenterar i sin artikel ”The Statistical Analysis of Compositional Data” (1982) nya metoder för att studera data bestående av sammansättningar och andelar (benämnda kompositioner respektive komponenter). Här introducerar han en stor del av den nu vedertagna terminologin för statistiskt arbete med sampling från populationer som är fördelade i ett simplexrum, och diskuterar problemen detta medför (till exempel svårigheten med att hitta fungerande beroendemått). Hans förslag är att införa transformationer från simplexrummet till ett mer lätthanterligt rum, det reella rummet, där vanlig multivariat analys kan tillämpas. Ett viktigt tillskott är den s.k. logkvotstransformationen, som omvandlar en komposition med D komponenter till en vektor med (D – 1) värden i det reella rummet. Hans metodik bygger på att komponenternas absoluta värden är irrelevanta då de enbart innehåller relativ information och att den relevanta informationen framgår i kvoter mellan komponenter.

John Aitchisons metoder har under de senaste decennierna vidareutvecklats. I den här uppsatsen har vi främst utgått från (Pawlowsky-Glahn, Egozcue & Tolosana-Delgado, 2015). Med utgångspunkt i John Aitchisons litteratur så formaliserar de ”The Aitchison Geometry” som en icke-euklidisk geometri med egna operatorer motsvarande addition och multiplikation, samt motsvarigheter till norm och avstånd, för statistiska tillämpningar i simplexrummet där kompositioner representeras som barycentriska koordinater. De beskriver tre olika logkvotstransformationer (varav de två första presenterades redan i (Aitchison, 1982)) som omvandlar kompositioner till vektorer i det reella rummet:

- Den additiva logkvotstransformationen (alr) uttrycks som $alr(x) = [log\frac{x_1}{x_D}, \dots, log\frac{x_{D-1}}{x_D}]$, där nämnaren $[x_1, x_2, \dots, x_D]$ är komponenterna i en komposition, och täljaren är en godtycklig komponent i kompositionen.
- Den centrerade logkvotstransformationen (clr) uttrycks som $clr(x) = [ln\frac{x_1}{g_m(x)}, \ ln\frac{x_2}{g_m(x)}, \dots, ln\frac{x_D}{g_m(x)}]$, där $[x_1, x_2, \dots, x_D]$ är komponenterna i en komposition och $g_m(x) = (\prod_{i=1}^{D} x_i)^{1/D}$ betecknar det geometriska medelvärdet.

- Den isometriska logkvotstransformationen (ilr) uttrycks som $ilr(x) = clr(x) * \Psi$ där Ψ är en s.k. kontrastmatris, vars kolumner utgör den ortonormala basen i simplexrummet (Pawlowsky-Glahn, Egozcue & Tolosana-Delgado, 2015). Sammanställning av kontrastmatrisen diskuteras senare i uppsatsen.

De ursprungliga komponenterna antas ligga i ett "Aitchison-simplex", och idén med transformationer är att avbilda kompositionernas barycentriska koordinater från simplex till kartesiska koordinater i det reella rummet, för att sedan kunna tillämpa de vanliga statistiska metoderna på data. De inversa transformationerna kan sedan användas för avbildningar tillbaka till simplexrummet. De inversa transformationer definieras på följande sätt:

$clr(x) = \ln(x)\,\Psi\Psi^{T} = \ln(x)\,I_{D-1}$ där I_{D-1} är en identitetsmatris, av detta följer

$ilr(x)^{-1} = C(\exp(\ln(x)\,I_{D-1}))$, där $C = [\frac{kx_1}{\sum_{i=1}^{D} x}, \dots, \frac{kx_D}{\sum_{i=1}^{D} x}]$, där 0 > k > 1 (Pawlowsky-Glahn, Egozcue & Tolosana-Delgado, 2015).

Den tredje transformationen (ilr) kommer att vara vårt tillvägagångssätt i den här uppsatsen. Denna har fördelen att transformationen från simplexrummet till det reella rummet är isometrisk, d.v.s. avstånd mellan observationer avbildas på ett ekvivalent sätt. Detta gör att ilr-transformationen lämpar sig väl för regressionsanalys. Tekniken bygger på att konstruera en ortonormal bas i simplexrummet, vilken representeras som rader i en kontrastmatris. Valet av ortonormal bas med tillhörande kontrastmatris kan göras på ett sådant sätt att de transformerade värdena blir tolkningsbara.

Transformationen från en sammansättning av D andelar till en vektor av D – 1 ilr-koordinater tillåter den regressionsanpassning som är utgångspunkten för vår uppsats. En multivariat linjär regression kan anpassas med ilr-koordinaterna som beroende variabel, varefter den inversa ilr-transformationen kan tillämpas på de skattade koefficienterna för att erhålla motsvarande koefficienter i simplexrummet.

Omvandling från andelar till balanser

Ett sätt att tillämpa ilr-transformationen som presenteras i (Pawlowsky-Glahn, Egozcue & Tolosana-Delgado, 2015) är genom att omvandla varje komposition (nio komponenter) till en vektor av åtta s.k. balanser. Denna metod har fördelen att de transformerade värdena kan göras förhållandevis lätta att tolka. Värdena i balansvektorn beskriver ett storleksförhållande mellan olika grupper av andelar inom sammansättningen, och liknas i litteraturen vid vikter i

ett besman. Nedan kommer vi först att göra en s.k. sekventiell binär uppdelning av partierna. Med utgångspunkt i denna uppdelning kan vi sedan konstruera en kontrastmatris, vilken slutligen kan användas för att transformera röstsammansättningarna till balanser (d.v.s. de ilr-koordinater som vi kommer att använda i vår regressionsanalys).

Vid beräkningen av balanser görs alltså först en sekventiell binär uppdelning av partierna, där partierna i varje steg tilldelas +1, –1 eller 0 som värde, beroende på vilka grupper av andelar man vill jämföra. Uppdelningen, som representeras i tabell 1 sker i åtta steg, d.v.s. tills varje parti i någon indelning har stått ensamt i jämförelse med ett annat parti eller grupp av partier. Vid varje steg räknas antalet partier i den positiva kategorin (r) och antalet partier i den negativa kategorin (s).

Vi börjar med att dela upp partierna i två grupper där övriga partier (Ö) tilldelas +1 och resterande partier tilldelas –1. Detta är för att se hur övriga partier förhåller sig till de andra partierna som ingår i riksdagen. Antal av de positiva och negativa ettorna beräknas och vi får r = 1, de positiva ettorna, och s = 8, de negativa ettorna. Fortsättningsvis sätter vi Ö till 0.

I andra ordningens uppdelning sätts partier som stödjer regeringen (M, KD, L, SD) till +1 och partier däremot (V, S, MP, C) till –1. Här blir alltså r = 4 och s = 4. I det tredje steget sätts vänsterblocket till 0 för att titta närmare på högerblocket. Vi sätter –1 på SD och +1 på resterande partier i högerblocket för att se hur SD förhåller sig till M, L och KD. I den fjärde ordningens uppdelning sätter vi SD till 0 och L till –1 för att se hur L förhåller sig till M och KD. I sista ordningen i högerblocket har vi M (+1) kontra KD (–1) kvar.

Därefter delar vi upp vänsterblocket på ett motsvarande sätt. I den sjätte ordningens uppdelning sätts V till –1 och MP, S respektive C sätts till +1. Sedan sätter vi –1 på C och +1 på MP och S. I sista ordningen tilldelas MP –1 och S +1.

Tabell 1. Sekventiell binär uppdelning av partierna. Partierna klassas i åtta steg (ordningar) i grupper betecknade med +1 och –1. Värdena r och s summerar antalet i respektive klass för varje ordning.

	V	**MP**	**S**	**C**	**L**	**M**	**KD**	**SD**	**Ö**	**r**	**s**
1	–1	–1	–1	–1	–1	–1	–1	–1	1	1	8
2	–1	–1	–1	–1	1	1	1	1	0	4	4
3	0	0	0	0	1	1	1	–1	0	3	1
4	0	0	0	0	–1	1	1	0	0	2	1
5	0	0	0	0	0	1	–1	0	0	1	1
6	–1	1	1	1	0	0	0	0	0	3	1
7	0	1	1	–1	0	0	0	0	0	2	1
8	0	–1	1	0	0	0	0	0	0	1	1

Nästa steg är att beräkna kontrastmatrisen. Kontrasterna beräknas radvis enligt formlerna $a_+ = \frac{1}{r}\sqrt{\frac{rs}{r+s}}$ och $a_- = -\frac{1}{s}\sqrt{\frac{rs}{r+s}}$ för partier som tilldelades +1 respektive –1. I figurerna nedan presenteras kontrastmatrisen i två delar (observera dock att den ska tolkas som en enda matris).

Tabell 2. Första delen av kontrastmatrisen där L, M, KD och SD är partier i det högra blocket och Ö är övriga partier.

L	M	KD	SD	Ö
$-\frac{1}{8}\sqrt{\frac{1\cdot 8}{1+8}}$	$-\frac{1}{8}\sqrt{\frac{1\cdot 8}{1+8}}$	$-\frac{1}{8}\sqrt{\frac{1\cdot 8}{1+8}}$	$-\frac{1}{8}\sqrt{\frac{1\cdot 8}{1+8}}$	$\frac{1}{1}\sqrt{\frac{1\cdot 8}{1+8}}$
$\frac{1}{4}\sqrt{\frac{4\cdot 4}{4+4}}$	$\frac{1}{4}\sqrt{\frac{4\cdot 4}{4+4}}$	$\frac{1}{4}\sqrt{\frac{4\cdot 4}{4+4}}$	$\frac{1}{4}\sqrt{\frac{4\cdot 4}{4+4}}$	0
$\frac{1}{3}\sqrt{\frac{3\cdot 1}{3+1}}$	$\frac{1}{3}\sqrt{\frac{3\cdot 1}{3+1}}$	$\frac{1}{3}\sqrt{\frac{3\cdot 1}{3+1}}$	$-\frac{1}{1}\sqrt{\frac{3\cdot 1}{3+1}}$	0
$-\frac{1}{1}\sqrt{\frac{2\cdot 1}{2+1}}$	$\frac{1}{2}\sqrt{\frac{2\cdot 1}{2+1}}$	$\frac{1}{2}\sqrt{\frac{2\cdot 1}{2+1}}$	0	0
0	$\frac{1}{1}\sqrt{\frac{1\cdot 1}{1+1}}$	$-\frac{1}{1}\sqrt{\frac{1\cdot 1}{1+1}}$	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0

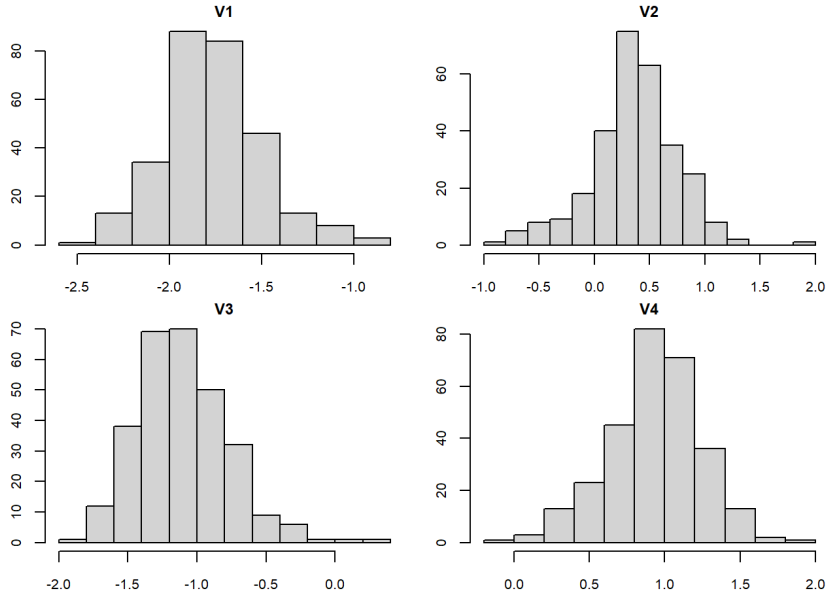
Tabell 3. Andra delen av kontrastmatrisen där V, MP, S och C är partier i det vänstra blocket.

V	MP	S	C
$-\frac{1}{8}\sqrt{\frac{1\cdot 8}{1+8}}$	$-\frac{1}{8}\sqrt{\frac{1\cdot 8}{1+8}}$	$-\frac{1}{8}\sqrt{\frac{1\cdot 8}{1+8}}$	$-\frac{1}{8}\sqrt{\frac{1\cdot 8}{1+8}}$
$-\frac{1}{4}\sqrt{\frac{4\cdot 4}{4+4}}$	$-\frac{1}{4}\sqrt{\frac{4\cdot 4}{4+4}}$	$-\frac{1}{4}\sqrt{\frac{4\cdot 4}{4+4}}$	$-\frac{1}{4}\sqrt{\frac{4\cdot 4}{4+4}}$
0	0	0	0
0	0	0	0
0	0	0	0
$-\frac{1}{1}\sqrt{\frac{3\cdot 1}{3+1}}$	$\frac{1}{3}\sqrt{\frac{3\cdot 1}{3+1}}$	$\frac{1}{3}\sqrt{\frac{3\cdot 1}{3+1}}$	$\frac{1}{3}\sqrt{\frac{3\cdot 1}{3+1}}$
0	$\frac{1}{2}\sqrt{\frac{2\cdot 1}{2+1}}$	$\frac{1}{2}\sqrt{\frac{2\cdot 1}{2+1}}$	$-\frac{1}{1}\sqrt{\frac{2\cdot 1}{2+1}}$
0	$-\frac{1}{1}\sqrt{\frac{1\cdot 1}{1+1}}$	$\frac{1}{1}\sqrt{\frac{1\cdot 1}{1+1}}$	0

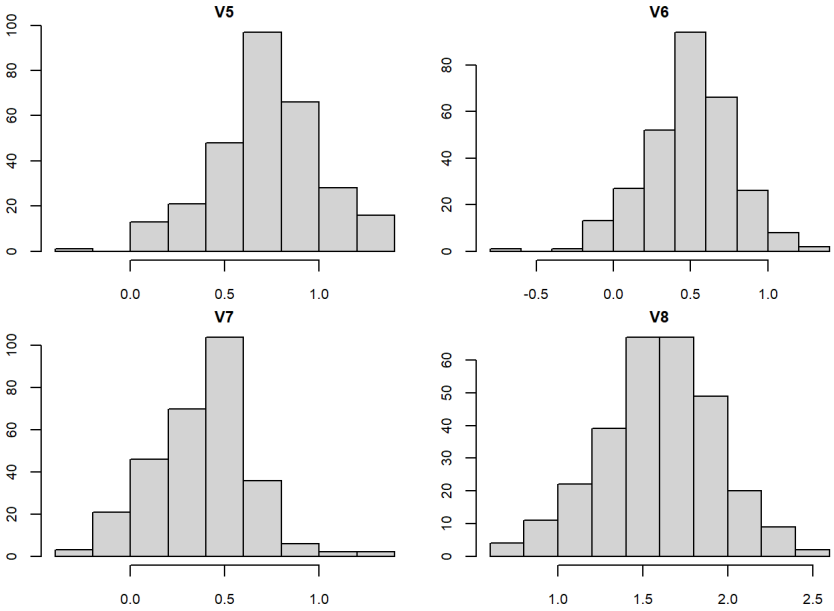
Balanser definieras som den normaliserade logkvoten mellan de geometriska medelvärdena för två grupper av andelar:

$$b = \ ln \frac{(x_1, \dots, x_r)^{a_+}}{(x_{r+1}, \dots, x_D)^{a_-}}$$

Där a_+ respektive a_- är de värden som beräknades i kontrastmatrisen. Slutligen har vi nu för varje kommun en vektor av åtta balanser, eller ilr-koordinater, som kan undersökas vidare. I figur 4 och 5 redovisas histogram för balanserna (V1, V2, …, V8) för att få en inblick i hur balanserna är fördelade.



Figur 4. Histogram för de fyra första balanserna.



Figur 5. Histogram för de fyra sista balanserna.

Den första balansen (V1) beskriver förhållandet mellan å ena sidan övriga partier (+) och å andra sidan samtliga riksdagspartier (–). Som väntat så antar dessa balanser negativa värden, vilket innebär att riksdagspartierna har större "vikt" - en term som används i (Pawlowsky-Glahn, Egozcue & Tolosana-Delgado, 2015) - i sammansättningen, eller större normerat geometriskt medelvärde. Partierna i nämnaren och täljaren är i snitt ungefär lika stora om balansen är 0. Om partierna i täljaren är större än partierna i nämnaren, är balansen positiv, och om partierna i täljaren är mindre än partierna i nämnaren, är balansen negativ.

Den andra balansen (V2) jämför partier i högerblock som ingår i eller stöder regeringen (+) och partier i vänsterblock, som ingår i oppositionen (–). Här är förhållandet jämnare, med något större vikt på regeringssidan.

Den tredje balansen (V3) jämför å ena sidan MD, KD och L (+) och å andra sidan SD (–). Maktbalansen mellan SD och övriga partier i högerblocket har stor politisk betydelse då SD i viss mån agerar som regeringens högeropposition. Här antar balanserna negativa värden, d.v.s. vikten är större på SD i de flesta kommuner.

Den fjärde balansen (V4) beskriver förhållandet mellan å enda sidan regeringspartierna M och KD (+) och å andra sidan L (–). Här antar balanserna som väntat nästan enbart positiva värden, d.v.s. vikten är större på M och KD i nästan alla kommuner.

Den femte balansen (V5) jämför M (+) och KD (–). Detta är den sista uppdelningen inom högerblocket och ger balansen mellan de två regeringspartierna. Här antar balanserna nästan enbart positiva värden, d.v.s. M är det större partiet i nästan alla kommuner.

Efter att vi har tittat på alla partier i högerblocket övergår vi till att titta på balanserna i det vänstra blocket. Den sjätte balansen (V6) beskriver förhållandet mellan å ena sidan MP, S och C (+) och å andra sidan V (–). Vänsterpartiet, V, har en traditionell roll som vänsteropposition till S-ledda regeringar och har även idag en delvis oppositionell roll till regeringsunderlaget MP-S-C. Här antar balanserna både positiva och negativa värden, med något större vikt på MP, S och C i de flesta kommuner.

Den sjunde balansen (V7) beskriver förhållandet mellan å ena sidan MP och S (+) och å andra sidan C (–). Denna jämförelse är politiskt viktig på grund av spänningen mellan C:s liberala politik kontra S och MP:s vänsterpolitik. Här antar balanserna främst positiva värden, d.v.s. vikten är större på MP och S.

Den åttonde och sista balansen (V8) beskriver förhållandet mellan S (+) och MP (–). Detta är den sista uppdelningen inom vänsterblocket. Här antar balanserna enbart positiva värden, d.v.s. S är det större partiet i samtliga kommuner.

Multivariat regression

För att undersöka förhållandet mellan en responsvariabel och en förklarande variabel använder man sig av en linjär regression. Ifall man har flera förklarande variabler tillämpas det en multipel linjär regression. I vårt fall vill vi bestämma hur en sammansättning av andelar, d.v.s. en matris av responsvariabler, beror på en förklarande variabel (medianinkomst respektive ålder) i två olika regressioner. Vi använder oss därför av en multivariat regression och sammanställer två regressioner. I den första regressionen är vi intresserade av att undersöka hur inkomst påverkar valbeteende. I den andra regressionen undersöker vi hur ålder påverkar valbeteende. Den multivariata regressionen kan visas på ett följande sätt:

$$Y = Z\beta + \varepsilon$$

där Y är en matris av responsvariabler, Z är en matris av förklarande variabler, β är en matris av koefficienter d, som är associerade med responsvariabel d, och är en vektor av feltermer. Feltermerna antas vara oberoende, normalfördelade med väntevärde 0 (Johnson, Wichern, 1998).

I fallet då vi jobbar med regressioner i simplexrummet, är det inte möjligt att multiplicera och addera element som man gör i det reella rummet. Vi använder oss av operatorer som definieras i (Pawlowsky-Glahn, Egozcue & Tolosana-Delgado, 2015) för beräkningar i ett simplexrum. Perturbation är definierat som ekvivalent till addition i det reella rummet och powering, (potens på svenska) är definierat som ekvivalent till multiplikation i det reella rummet:

- $x \oplus y = C[x_1 y_1, \dots, x_D y_D]$, där x och y är vektorer och $C = [\frac{kx_1}{\sum_{i=1}^{D} x}, \dots, \frac{kx_D}{\sum_{i=1}^{D} x}]$, där 0 > k > 1
- $a \odot y = C[y_1^a, \dots, y_D^a]$, där a är en konstant, y är en vektor och $C = [\frac{kx_1}{\sum_{i=1}^{D} x}, \dots, \frac{kx_D}{\sum_{i=1}^{D} x}]$, där k <= 1

Den multivariata regressionen i simplexrummet ser ut på ett följande sätt:

$$Y = Z \odot \beta \oplus \varepsilon$$

där $\odot$ – betecknar powering och $\oplus$ – betecknar perturbation.

Vidare är det viktigt att undersöka signifikansen av förklarande variabler. För att kunna göra det utför vi ett Pillais spårtest, som är en del av MANOVA. Pillais spår kan variera i sitt värde mellan 0 och 1, där värde nära 0 betyder att förklarande variabler inte alls är signifikanta och värde nära 1 betyder att förklarande variabler är mycket signifikanta.

Medianinkomst och röstsammansättning

Den första regressionen beskriver det multivariata linjära sambandet mellan de åtta balanserna och medianinkomsten i en given kommun. I tabell 4 visas de anpassade koefficienterna med tillhörande intercept, samt t-testvärden för koefficienterna.

Tabell 4. De anpassade koefficienterna med intercept, t-värden och signifikansnivå.

Balans	**V1**	**V2**	**V3**	**V4**	**V5**	**V6**	**V7**	**V8**
(intercept)	–0.9022	–0.0920	–3.4915	2.7722	–0.7897	–0.3993	0.2858	3.6476
Medianinkomst	–0.0029	0.0016	0.0082	–0.0063	0.0052	0.0030	0.0003	–0.0070
t-värde	–5.775	1.985	17.9	–12.63	10.982	5.632	0.602	–12.81
signifikansnivå	***	*	***	***	***	***	Ej signifikant	***

På balans 1 (övriga partier kontra riksdagspartier) ser vi att medianinkomst har en signifikant negativ inverkan. Balans 2 (högerblocket kontra vänsterblocket) och balans 3 (MD, KD och L kontra SD) har båda en positiv koefficient och t-värde som visar signifikans. Balans 4 (M och KD kontra L) har ett tydligt signifikant samband och en negativ koefficient. Balans 5 och 6 (M kontra KD, respektive MP, C och S kontra V) har båda en positiv koefficient och ett t-värde som visar på signifikans. Balans 7 (S och MP kontra C) har en positiv koefficient; här är dock t-värdet lågt, och ingen signifikans kan påvisas. Balans 8 (S kontra MP) har en negativ koefficient och ett t-värde som visar på signifikans. Pillais spårtest ger ett testvärde

på 0.64562 vilket innebär att modellen som helhet är kraftigt signifikant (med ett försvinnande litet p-värde).

När vi analyserar residualerna finner vi tecken på heteroskedasticitet kring de anpassade andelarna. När t-värdena justeras med en robust (White-justerad) variansmatris så förändras dock inte signifikansnivåerna nämnvärt: de koefficienter som tidigare klassats som statistiskt säkerställda har fortfarande en hög signifikansnivå, och vice versa. Vi har dessvärre inte kunnat hitta eller konstruera en robust motsvarighet till Pillais spårtest i R. Heteroskedasticiteten i residualerna påverkar dock inte själva skattningarna, vilket är det väsentliga.

Vi använder den inversa ilr-transformationen och erhåller koefficienternas motsvarighet i simplexrummet (d.v.s. sambandet som berör de ursprungliga andelarna snarare än de beräknade balanserna). Vi erhåller då följande koefficienter:

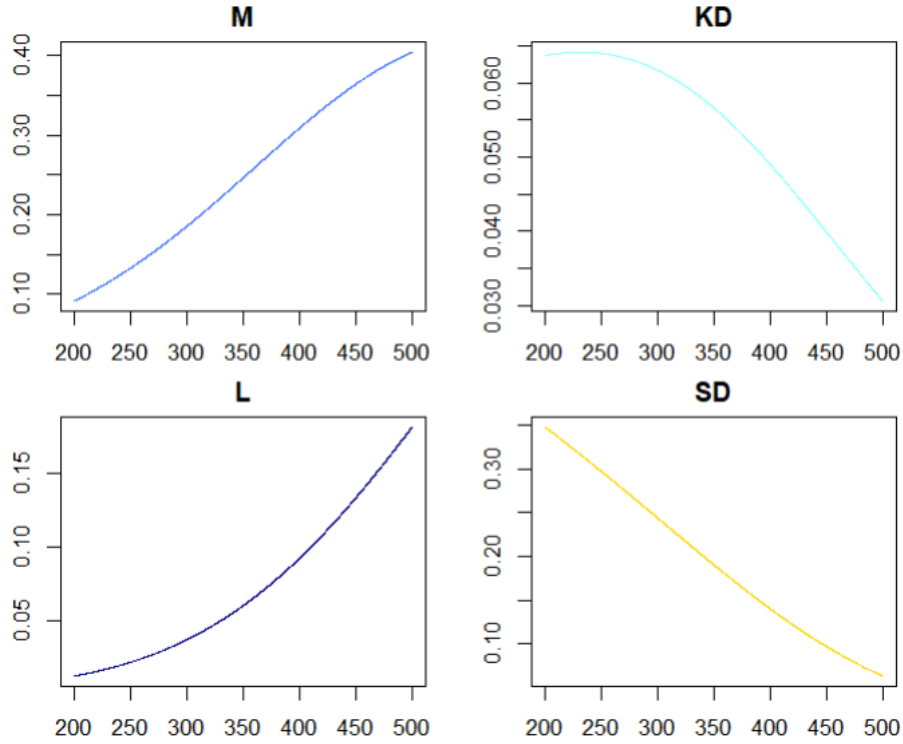
Tabell 5. Koefficienter som erhölls som resultat av ilr-transformation.

Andel	V	MP	S	C	L	M	KD	SD	Ö
(Intercept)	0.0376	0.0020	0.3518	0.0188	0.0009	0.0161	0.0494	0.5134	0.0099
medianinkomst	0.1108	0.1117	0.1106	0.1112	0.1121	0.1116	0.1107	0.1104	0.1108

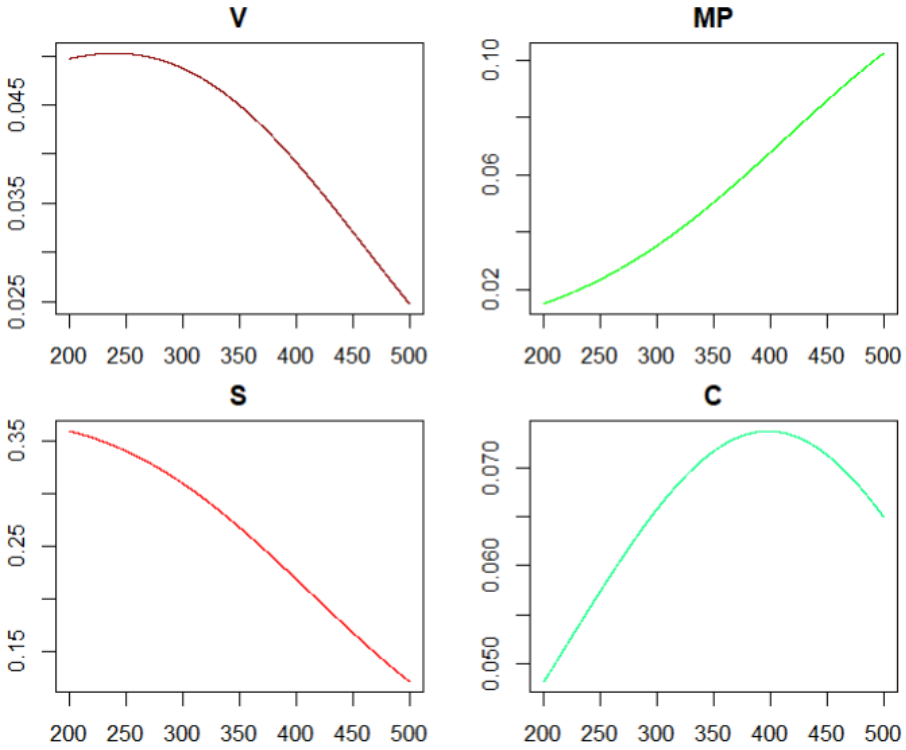
Dessa värden kan vid en första anblick verka underliga då samtliga koefficienter är positiva. Det ska dock hållas i åtanke att dessa koefficienter verkar i ett barycentriskt koordinatsystem där mittpunkten (motsvarigheten till origo) är den punkt där alla 9 komponenter har värdet 1/9. Annorlunda uttryckt: om koefficienten är lägre än $1/9 \approx 0.111$, så krymper den tillhörande predikterade andelen när medianinkomsten ökar. Vänsterpartiet, Socialdemokraterna, Kristdemokraterna, Sverigedemokraterna samt övriga partier-kategorin har koefficienter som är lägre än $1/9 \approx 0.111$, medan Miljöpartiet, Centerpartiet, Liberalerna och Moderaterna har koefficienter som är högre än $1/9 \approx 0.111$.

Följande ska här anmärkas: den regression som anpassats med avseende på balanserna är linjär, men den transformerade regression som beskriver andelarnas samband med medianinkomst har inte ett linjärt utseende. Koefficienterna i simplexrummet är något svårtolkade. I figurer 6 och 7 nedan presenteras anpassade värden för de olika partiernas

andelar grafiskt. Vi beräknar dessa genom att anpassa värden på balanserna och sedan använda den inversa ilr-transformationen för att återfå predikterade andelar, vilket är ekvivalent med att använda de transformerade koefficienterna direkt i simplexrummet.

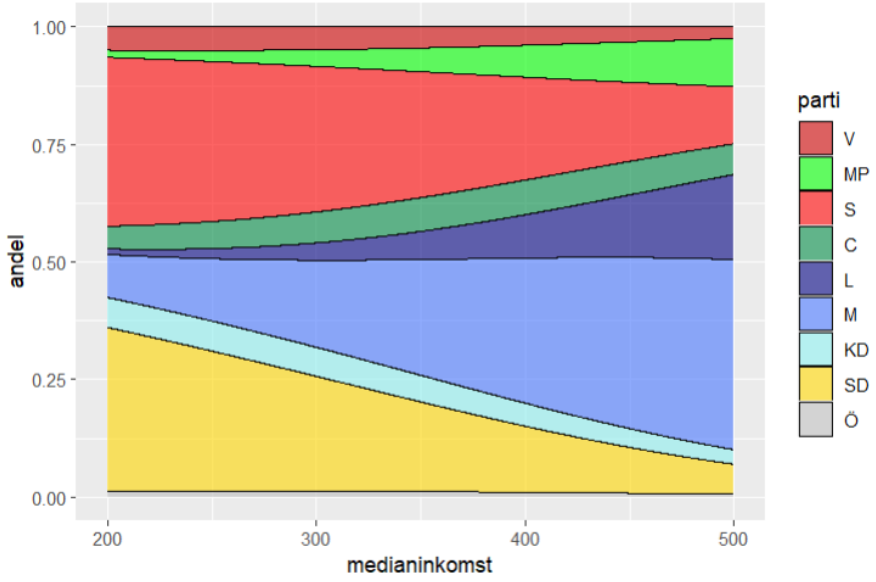


Figur 6. Anpassade kompositionella regressioner för M, KD, L och SD. Y-axeln visar den predikterade andelen röster i riksdagsvalet i en kommun med en medianinkomst given av x-axeln.



Figur 7. Anpassade kompositionella regressioner för V, MP, S och C.

Ett mer elegant sätt att presentera sambandet mellan andelssammansättningen och medianinkomst är i ett staplat områdesdiagram. I figur 8 presenteras varje parti som en färg, vars andel av y-axelns längd varierar längs x-axeln. Utifrån figur 8 kan vi se olika tendenser i hur befolkningen tenderar att rösta baserad på dess medianinkomst. Andel röster Vänsterpartiet (V) får sjunker långsamt, medan andel röster Miljöpartiet (MP) får blir betydligt större när väljarna har högre inkomst. Socialdemokraternas (S) och Sverigedemokraternas (SD) andelar sjunker betydligt, medan Liberalernas (L), och Moderaternas (M) andelar tenderar att växa när väljarna har högre medianinkomst. Centralpartiets (C) och Kristdemokraternas (KD) andelar ändras ganska lite när väljarna är höginkomsttagande.



Figur 8. Proportionellt ytdiagram. På y-axeln visas de olika partiernas predikterade röstandelar i riksdagsvalet i en kommun med en medianinkomst given av x-axeln.

Medelålder och röstsammansättning

Vi gör en andra regression för att undersöka sambandet mellan kommuners medelålder och röstsammansättning i riksdagsvalet. Liksom i fallet med medianinkomst så anpassar vi först en regression m.a.p. de åtta balanserna.

Tabell 6. Anpassade koefficienter från regression 2, samt intercept, t-värden och signifikansnivå.

	V	MP	S	C	L	M	KD	SD	Ö
(Intercept)	0.0288	0.1635	0.0415	0.0222	0.2601	0.4458	0.0069	0.0177	0.0135
medelålder	0.1113	0.1059	0.1152	0.1127	0.1049	0.1076	0.1156	0.1169	0.1099

På balans 1 (övriga partier kontra riksdagspartier) finnes att en högre medelålder har en negativ inverkan, dock med ett p-värde strax över 0,05. På balans 2 (högerblocket kontra vänsterblocket) hittar vi en negativ koefficient, men t-testet ger ingen signifikans här. På balans 3 (M, L och KD kontra SD) ser vi en negativ koefficient och tydlig signifikans. På balans 4 (M och KD kontra L) har medelålder en positiv koefficient, med trestjärnig signifikans. På balans 5 (S, C och MP kontra V) ser vi en negativ och signifikant koefficient. Balans 6 har inte ett signifikant samband med medelålder. Däremot har balans 7 (S och MP kontra C) och 8 (S kontra MP) ett negativt respektive positivt samband med medelålder, där t-testet visar signifikans. Pillais spårtest ger ett testvärde på 0.46057, vilket visar på att regressionen sammantaget har ett högt förklaringsvärde.

Liksom i regressionen med medianinkomst som förklarande variabel utför vi här en analys av residualerna och finner tecken på heteroskedasticitet. Även här gäller dock att t-värden beräknade med en White-justerad variansmatris pekar mot att de koefficienter som tidigare klassats som statistiskt säkerställda fortsätter att ha en hög signifikansnivå och vice versa.

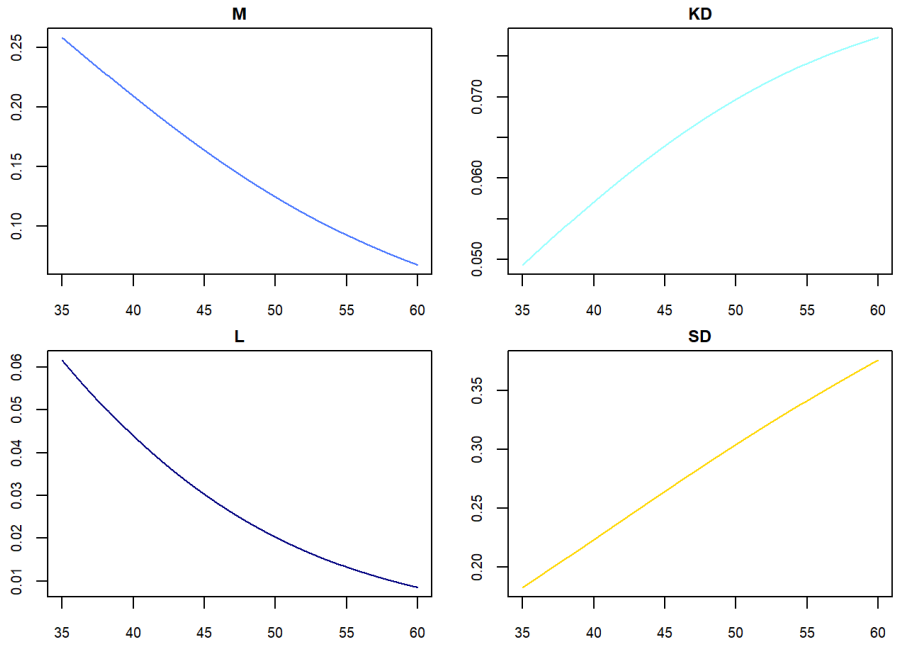
Vi använder återigen den inversa ilr-transformationen för att hitta motsvarande koefficienter för en kompositionell regression i simplexrummet.

Tabell 7. Resulterande koefficienter när regression 2 transformeras tillbaka till simplexrummet.

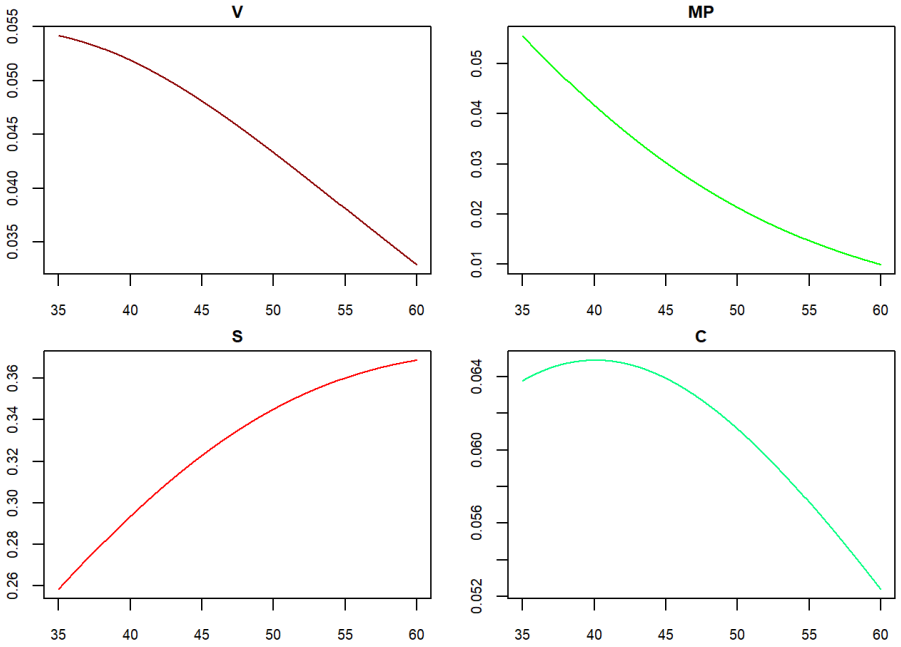
	V1	V2	V3	V4	V5	V6	V7	V8
(Intercept)	–1.2889	0.4197	1.4393	–1.2592	2.9431	0.5324	1.0711	–0.9689
medelålder	–0.0109	–0.0013	–0.0581	0.0501	–0.0506	–0.0007	–0.01589	0.0587

Liksom i den första regressionen gäller här att de predikterade andelarna växer när medelåldern ökar om koefficienten är högre än $1/9 \approx 0.111$, och krymper om koefficienten är lägre än $1/9 \approx 0.111$. Vi ser på så sätt att MP, L, M samt övriga partier förväntas ha lägre röstandelar när medelåldern är hög, medan det motsatta gäller för V, S, C, KD och SD.

Nedan följer grafiska presentationer av den anpassade kompositionella regressionen. Figur 9 och 10 visar riksdagspartiernas skattade samband med medelålder ett parti i taget, medan figur 11 visar ett proportionellt ytdiagram av samma slag som i det förra avsnittet.

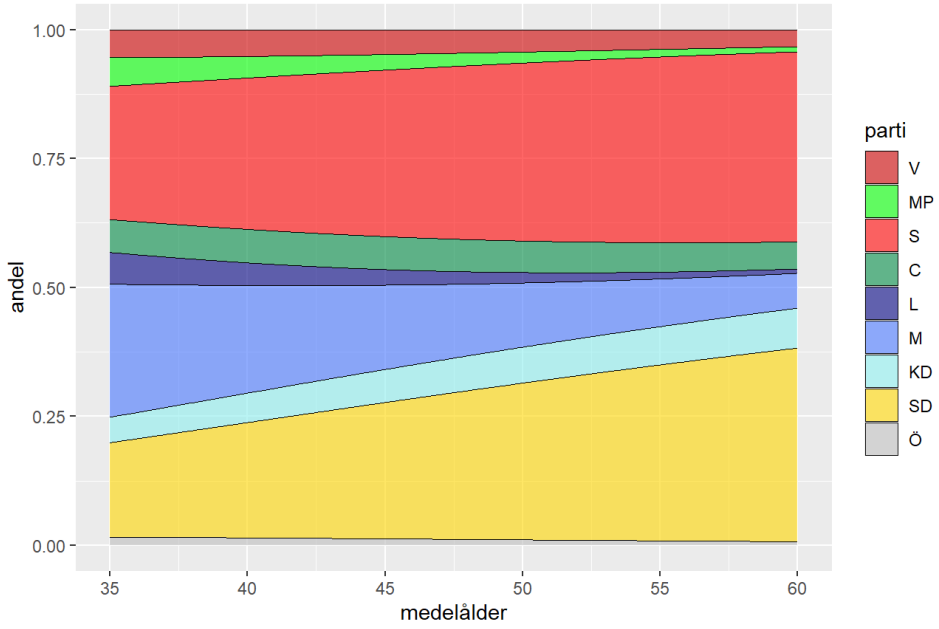


Figur 9. Andelar för M, KD, L och SD i den anpassade kompositionella regressionen. På y-axeln visas den predikterade röstandelen i en kommun med en medelålder givet av x-axeln.



Figur 10. Andelar för V, MP, S och C i den kompositionella regressionen.

I figur 11 ser vi hur befolkningen tenderar att rösta baserat på deras medelålder. Vi kan avläsa att andel röster Socialdemokraterna (S) och Sverigedemokraterna (SD) växer ju äldre väljarna är, medan andel Miljöpartiet (MP), Liberalerna (L) och Moderaterna (M) sjunker betydligt ju äldre väljarna är. Andel röster Centralpartiet (C) och Kristdemokraterna (KD) ändras väldigt lite med väljarnas ålder.



Figur 11. Proportionellt ytdiagram för den kompositionella regressionen. På y-axeln visas den predikterade sammansättningen av röstandelar i en kommun med en medelålder givet av x-axeln.

Diskussion

Vi har utfört den isometriska logkvotstransformationen på data och skapat två separata regressioner för att kunna analysera data. Som förklarande variabler i den multivariata regressionen använde vi oss av data för medianinkomst och medelålder. Regressionerna gav, till stor del, signifikanta resultat, vilket var förväntat. Det vi här har utfört bör främst betraktas som ett “proof-of-concept" vad gäller kompositionell dataanalys och valresultat - en mer vederhäftig prediktionsmodell hade behövt använda en multipel regressionsmodell som tar hänsyn till fler socioekonomiska och geografiska faktorer (utbildningsnivå, län och dylikt). Med det sagt så visar sig kompositionell regression vara ett kraftfullt redskap för att förstå valresultat; de proportionella ytdiagrammen vi kunde generera ger en tydlig inblick i hur de förklarande variablerna påverkar röstandelarna, och är lättbegripliga även för en lekman. Kompositionell dataanalys är ett underutnyttjat redskap för politisk forskning.

I vår analys finner vi vissa samband som stämmer överens med våra förutfattade meningar om saken; till exempel att Moderaterna, Liberalerna och Miljöpartiets andelar tilltar när inkomsten ökar, medan Sverigedemokraternas, Socialdemokraternas och Vänsterpartiets minskar. Många resultat vad gäller kommunens medelålder stämmer också överens med allmänna uppfattningar, till exempel att Socialdemokraterna och Sverigedemokraterna har högre andelar när medelåldern är hög och att Miljöpartiet har ett negativt samband med medelålder. Det finns dock även överraskande resultat. Vi förväntade oss till exempel inte att Liberalerna krymper när medelåldern ökar, eller att Kristdemokraterna har högre andelar i kommuner med låg medianinkomst. Vi finner vissa andra anmärkningsvärda resultat, till exempel hur snarlika Socialdemokraternas och Sverigedemokraternas röstandelars samband med medianinkomst är. Intressant är även att Centerpartiet är den enda andel vars kurva har en lutning som byter riktning - både i medianinkomst och i medelålder har partiet en andel som ökar, når en topp, och sedan sjunker.

Som fördjupning i ämnet hade man kunnat utföra fler tester samt utföra en multipel regression där den simultana påverkan av medianinkomst och medelålder skulle undersökas. Det skulle även vara fördelaktigt att använda en större regressionsanalytisk verktygslåda; främst tänker vi här på ett multivariat F-test med en robust variansmatris för att korrigera för heteroskedasticitet i residualerna. Kompositionell dataanalys kan tillämpas i många fall där data består av andelar.

Slutsats

I denna kandidatuppsats undersökte vi hur medianinkomst och ålder påverkar beteende vid val till riksdag med hjälp av kompositionell dataanalys. Mandat till riksdagen är proportionellt fördelade mellan partierna vilket innebär att man ofta pratar om andelar när man pratar om valresultat. Om man använder sig av vanliga statistiska metoder, som linjär eller logistisk regressioner, kan det skapa vilseledande resultat då andelarna är negativt korrelerade mellan varandra. Vi har därför valt att testa en ny metod, kompositionell dataanalys, för att avgöra ifall den kan tillämpas på undersökning av valbeteende och valresultat. Kompositionell dataanalys, som bygger på att andelarna av en helhet ligger i ett annat rum än det reella rummet, simplex, tillåter att göra regressioner på data där andelarna inte är negativt korrelerade. För att kunna arbeta med data har vi skapat balanser och transformerat data med hjälp av den isometriska logkvotstransformationen, ilr, för att vidare kunna skapa regressioner på data. Vi gjorde en enkel ansats och undersökte påverkan av förklarande variabler, medianinkomst och ålder, i två separata multivariata regressioner. Resultat av regressioner presenterades med hjälp av diagram och signifikans av koefficienter undersöktes, i sin helhet, med hjälp av Pillais spårtest, som är en del av MANOVA och var för sig med hjälp av t-test.

Kompositionell dataanalys är en underskattad metod som inte används vid undersökning av valresultat. I denna uppsats visar vi att kompositionell dataanalys är ett kraftfullt verktyg för att analysera valdata vilket respekterar utfallsrummets inneboende restriktioner.

Referenser

Aitchison, J. (1982) The Statistical Analysis of Compositional Data, Journal of the Royal Statistical Society, Vol. 44 No. 2, 1982, Tillgänglig online via: http://leg.est.ufpr.br/lib/exe/fetch.php/pessoais:abtmartins:thestatisticalanalysisofcompositionaldata.pdf [hämtad 3 oktober 2023]

Johnson A. R., Wichern W. D., (1998) Applied Multivariate Statistical Analysis, Upper Saddle River, Prentice-Hall, Inc.

Katz, Jonathan N. & King, Gary (1999). A Statistical Model for Multiparty Electoral Data, The American Political Science Review, Vol. 93 No. 1. Tillgänglig online via: https://www.jstor.org/stable/2585758?seq=16 [hämtad 18 januari 2024]

Pawlowsky-Glahn V., Egozcue J. J., Tolosana-Delgado R. (2015) Modeling and Analysis of Compositional Data, Chichester, John Wiley & Sons, Ltd

Statistiska centralbyrå (SCB) (2022). Befolkningens medelålder och medianålder efter kön. År 1968 - 2022, Tillgänglig online via: https://www.statistikdatabasen.scb.se/pxweb/sv/ssd/START__BE__BE0101__BE0101B/BefolkMedianAlder/ [hämtad 23 oktober 2023]

Statistiska centralbyrå (SCB) (2021). Sammanräknad förvärvsinkomst för boende i Sverige hela året efter region, kön, ålder och inkomstklass. År 1999 – 2021, Tillgänglig online via: https://www.statistikdatabasen.scb.se/pxweb/sv/ssd/START__HE__HE0110__HE0110A/SamForvInk1/ [hämtad 22 oktober 2023]

Sveriges riksdag (2023). Val till riksdagen, Tillgänglig online via: https://www.riksdagen.se/sv/teckensprak/sa-fungerar-riksdagen/val-till-riksdagen/ [hämtad 19 oktober 2023]

Valmyndighet (2023). Rådata och statistik, Tillgänglig online via: https://www.val.se/valresultat/riksdag-region-och-kommun/2022/radata-och-statistik.html#slutligtvalresultat [hämtad 23 oktober 2023]

Örhvall R., Järnbert M. (2010) Partiernas sympatisörer, Richard Örhvall, Tillgänglig online via: https://richardohrvall.rbind.io/sv/publication/2010_2_partiernas_sympatis%C3%B6rer/ [hämtad 4 januari 2024]