ATT PROGNOSTISERA DET FINANSIELLA RESULTATET I SVENSKA KOMMUNER

EN JÄMFÖRELSE MELLAN ETT ELASTISKT NÄT OCH EN BASELINE

SOFIA NILSSON

KANDIDATUPPSATS (STAH11 15 ECTS)

STATISTISKA INSTITUTIONEN

LUNDS UNIVERSITET

HANDLEDARE: BJÖRN HOLMQUIST

Innehållsförteckning

ABSTRACT:

Machine learning techniques have gained ground in macroeconomic forecasting in recent years. Estimations with the elastic net technique have yielded good results. There is currently not much research on how statistical techniques can be used to estimate the financial result in Swedish municipalities. This study adds to the current body of research by investigating how the elastic net technique can be used to forecast the financial result in Swedish municipalities.

Data is collected from several sources and is used to forecast the financial result in Swedish municipalities. Estimations are carried out using an elastic net and a baseline in the form of OLS and are then compared and discussed. The results show that the estimations using an elastic net give more accurate forecasts than OLS in Swedish municipalities between 2002 and 2018.

The most accurate forecasting method in this study was the Ridge regression. The study concludes by suggesting how the financial result in Swedish municipalities could be more thoroughly investigated using statistical methods.

KEY WORDS: Elastic net, Ridge regression, Lasso regression, OLS, finances of Swedish municipalities, prediction, forecasting

SAMMANFATTNING:

De senaste åren har maskininlärningstekniker börjat användas för att ta fram makroekonomiska prognoser. Framför allt har skattningar med elastiska nät givit goda resultat. I dagsläget finns det lite forskning på hur olika statistiska tekniker kan användas för att skatta det finansiella resultatet i svenska kommuner. Denna studie bygger på forskningen om ett elastiskt nät i prognosarbete och vänder blicken mot det finansiella resultatet i Sveriges kommuner.

Statistik samlas in från flera källor för att prognostisera resultatet i svenska kommuner. Skattningar tas fram och en jämförelse görs mellan ett elastiskt nät med en baseline i form av OLS. Resultaten visar att skattningar med ett elastiskt nät ger mer träffsäkra prognoser än OLS i Sveriges kommuner mellan 2002 och 2018.

Den mest träffsäkra prognosen i det här fallet togs fram med en Ridge regression. Studien konkluderar med att ge förslag på hur det finansiella resultatet i svenska kommuner skulle kunna genomlysas på en djupare nivå med hjälp av statistiska metoder.

NYCKELTERMER: Elastiskt nät, Ridge regression, Lasso regression, OLS, kommunalekonomi, prediktion, prognoser

1. PROBLEM FORMULERING

Statistisk analys tillämpas inom många områden. Bland annat används statistiska metoder för att genomlysa det rådande ekonomiska läget, analysera befolkningsutvecklingen och för att ta fram prognoser på olika ekonomiska nyckeltal så som BNP-tillväxt. På senare tid har statistiska metoder som ryms inom samlingstermen maskininlärning fått stor uppmärksamhet. Detta begrepp är brett och är inte tydligt definierat. Bland de mer uppmärksammande modellerna återfinns neurala nätverk och ”random forest” men även mer traditionella skattningsmodeller såsom OLS definieras ibland som maskininlärning.

Maskininlärning har fått allt mer uppmärksamhet i takt med att mängden information och statistik har ökat det senaste decenniet. Tekniker inom detta område kräver vanligtvis en stor mängd observationer. Inom analyser på mer aggregerad nivå inom exempelvis makroekonomi är antal observationer vanligtvis relativt få. Dock har maskininlärningstekniker börjat utforskas även inom detta område under senare år. Det finns hittills relativt lite forskning om tillämpningen av maskininlärning inom ekonomisk tillväxt. Basuchoudhary et al. (2017) tillämpar diverse maskininlärningstekniker för att prognostisera olika makroekonomiska variabler i olika länder med blandande resultat. Jung et al. (2018) undersöker olika maskininlärningstekniker för att ta fram makroekonomiska prognoser och fann att framförallt det elastiska nätet förbättrade prognosförmågan. Det elastiska nätet gav även goda resultatet i prognoser på Libanons tillväxt (Tiffin 2016). Den här studien bygger på tidigare forskning om maskininlärning inom ekonomiska prognoser genom att jämföra precisionen av en modernare skattningsmetodik med en konventionell skattningsmetodik – vilken vi benämner ”baseline” i prognoser av kommuners finansiella resultat i Sverige mellan 2002 och 2018. Enligt mina efterforskningar har detta område inte belysts i nuvarande forskning.

Sverige består av 290 kommuner som bedriver verksamhet för sina kommuninvånare. Varje år beräknas det finansiella resultatet för varje kommun. Om en kommun går med ett underskott måste antingen lån tas upp för att täcka underskottet eller måste kommunen i fråga öka sina intäkter genom att, som exempel höja skattesatsen. Det är därför viktigt för kommunerna att med viss säkerhet kunna förutspå resultatet ett givet år redan ett år eller tidigare, i förväg. Indikatorer från tidigare år på exempelvis arbetslöshet skulle kunna bidra till att prognostisera resultatet det kommande året. Statistiska tekniker skulle därmed kunna hjälpa kommuner att prognostisera resultatet det kommande året redan ett år före. I denna uppsats jämförs ett elastiskt nät som används av Jung et al. (2018) och Tiffin (2016) med en baseline. Baseline i denna uppsats är en OLS regression. Syftet med uppsatsen är att utforska om ett elastiskt nät tar fram en bättre prognos än en baseline (i detta fall en OLS regressionen) på kommunernas resultat mellan 2002 och 2018. Frågan som denna uppsats ämnar besvara är därmed:

Är ett elastiskt nät bättre än en baseline i form av OLS på att prognostisera resultatet i svenska kommuner mellan 2002 och 2018?

Uppsatsen är uppdelad i sju avsnitt. I avsnitt 2 ges en översikt över den teoretiska ansatsen. Det tredje avsnittet ger en översikt över datainsamlingen. Det fjärde avsnittet beskriver de två metoderna som jämförs. Det femte avsnittet redogör för resultaten. En diskussion av resultaten finns i avsnitt sex. Avsnitt sju konkluderar och ger förslag på vidareutveckling och förbättringar av studien.

2. TEORETISK ANSATS

2.1 PROGNOSTISERANDE MODELLER VS. FÖRKLARANDE MODELLER

De senaste åren har en del prognosmakare börjat rikta sin uppmärksamhet mot mer flexibla statistiska tekniker för att ta fram makroekonomiska prognoser. Att arbeta med prediktion är annorlunda från att arbeta med förklarande modeller. Prediktion handlar om att ta fram så träffsäkra prognoser som möjligt. Det innebär att fokus inte läggs på att förklara sambandet mellan de förklarande variablerna och den beroende variabeln (Shmueli 2010) utan på att ta fram så goda prognoser som möjligt.

Trots detta måste modellen ha förklarande variabler som ökar den prediktiva förmågan. Det är därför viktigt att få en förståelse för det finansiella resultatet i Sveriges kommuner och dess drivkrafter.

2.2 DET FINANSIELLA RESULTATET I SVENSKA KOMMUNER

Sverige består av 290 kommuner som bedriver verksamhet för sina invånare. Den främsta inkomstkällan för kommunerna utgörs av skatteintäkter från sina invånare. Kommunerna har även en del andra inkomster såsom utdelning från kommunala bolag, bidrag från staten (statsbidrag) och avgifter från de boende i kommunen. Dessa utgör en mindre andel. Utgifterna för kommuner består främst av utgifter för vård och omsorg samt pedagogisk verksamhet (se avsnitt 2.3 om drivkrafter bakom det finansiella resultatet i kommuner)

När en kommun går med ett underskott måste kommunen få in pengar för att täcka detta. Det kan ske genom att antingen ta upp lån eller, som alternativ, öka intäkterna genom att, som exempel höja skattesatserna. Kommunerna måste även följa kommunlagen. Det innebär att de måste uppfylla det så kallade balanskravet och ha en god ekonomisk hushållning. Balanskravet

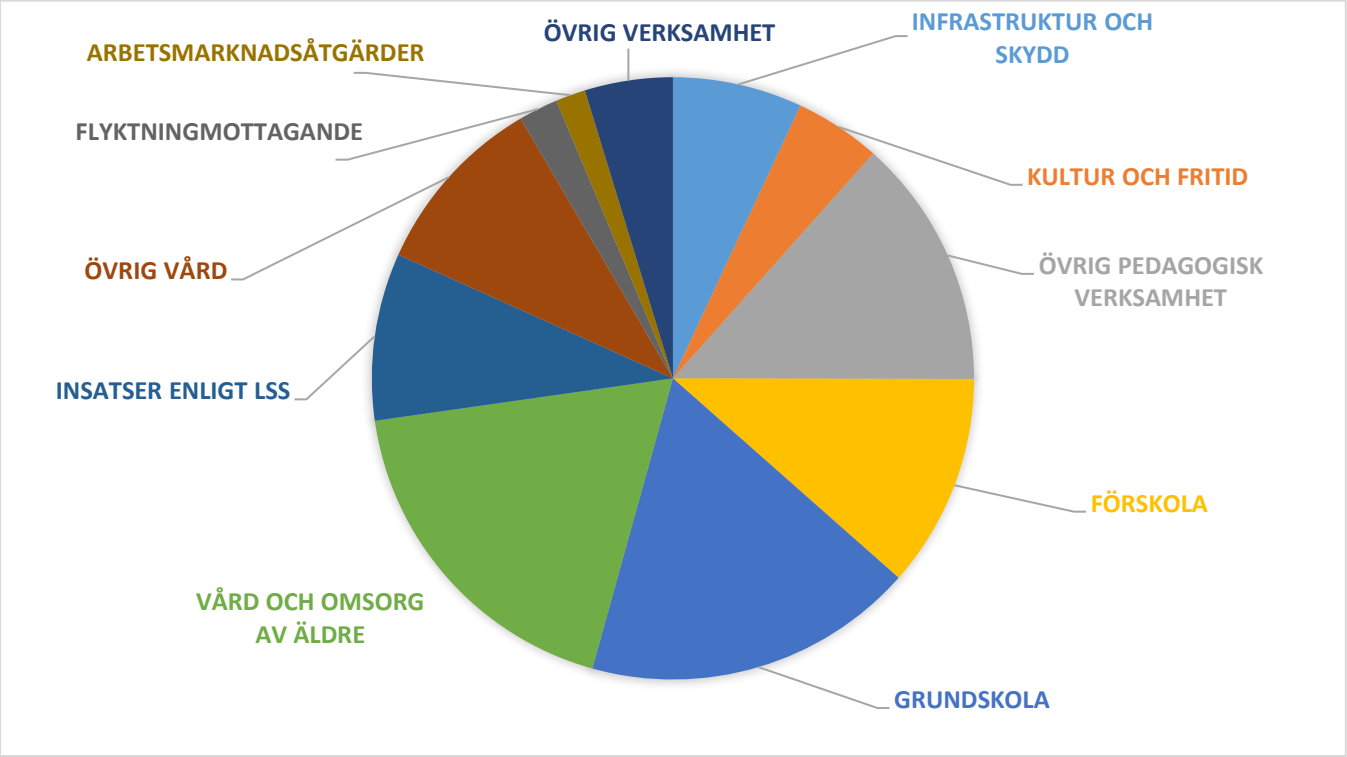
innebär att kommuner ska upprätta en budget för varje budgetår som inte innebär ett underskott. Kommunerna måste därmed upprätta en budget för varje budgetår så att intäkterna överskrider kostnaderna. (Konjunkturinstitutet, 2019). Om kostnaderna överstiger intäkter ett räkenskapsår ska detta negativa resultat tas igen inom tre år. God ekonomisk hushållning är inte lika väldefinierat och definieras av fullmäktige i varje kommun (Heed & Stark 2020).

På grund av ramverket skulle en träffsäker prognos på kommande budgetårs resultat kunna bidra till det ekonomiska arbetet i kommuner. En prognos på det finansiella resultatet av hög kvalitet skulle kunna vara ett komplement till de vanliga prognoserna som tas fram av kommunerna. Syftet med denna uppsats är inte att ta fram en så träffsäker prognos som möjligt utan att utforska om ett elastiskt nät producerar en bättre prognos än en baseline i form av OLS på kommuners finansiella resultat mellan 2002 och 2018. Om det är fallet skulle detta statistiska verktyg på sikt kunna vara till hjälp i budgetarbetet i kommuner.

2.3 DRIVKRAFTER BAKOM DET FINANSIELLA RESULTATET I SVENSKA KOMMUNER

I diagram 1 illustreras de största kostnadsområdena för kommunerna. Diagrammet visar tydligt att de största utgiftsposterna för kommuner består av vård och omsorg (till exempel vård och omsorg av äldre och insatser enligt LSS) samt pedagogisk verksamhet (till exempel grundskola och förskola). Bland de mindre utgiftsposterna finns arbetsmarknadsåtgärder och flyktingmottagande.

Diagram 1. Största kostnadsposter för kommuner 2019



Källa: Statistiska Centralbyrån

För att ta fram en prognos på det finansiella resultatet i en kommun med hjälp av ett elastiskt nät och OLS behöver data samlas in som representanter för de flesta kostnader som syns i diagram 1. Även mått på kommunernas intäkter behöver samlas samt några faktorer som varierar mellan kommunerna. I denna studie samlas skattebasen i kommunerna in samt skattesatsen. Mått på storleken på de olika kommuner och om de växer eller krymper kan också vara förklarande variabler. Folkmängd och befolkningsökning samlas därför in på kommunnivå.

2.4 FÖRKLARANDE VARIABLER

2.4.1 KOSTNADER

Indikatorer på de olika kostnaderna behöver samlas in. De största kostnaderna är för vård och omsorg av äldre samt pedagogisk verksamhet. För att representera dessa kostnadsposter samlas statistik in på antal äldre över 70 år samt antal personer mellan under 19 år från befolkningsstatistiken på SCB. Även medianåldern i samtliga kommuner samlas in och används som förklarande variabel. Denna variabel är en indikator på åldersstrukturen i svenska kommuner och behöver kontrolleras för.

Ett mått behövs även på antal personer som är i behov av LSS[1] i kommunerna. Kommunerna betalar ersättning för de första tjugo timmarna för personer som får statlig assistans. Försäkringskassan har statistik på antal personer fördelat på kommun som får mer än tjugo timmars assistans per månad. Detta mått är inte idealt då personer som får mindre än tjugo timmars assistans inte finns med. Denna statistik är den enda statistiken som finns på kommunnivå och samlas därför in för att jämföra de två skattningsmetoderna. Den är troligen en underskattning av det verkliga antalet som behöver får ersättning från kommunen för stöd genom LSS.

Flyktingmottagande och arbetsmarknadsåtgärder är relativt små delar av de totala kostnaderna (se Diagram 1). Statistik på dessa samlas ändå in och tas med i skattningarna. Flyktingmottagande samlas in från Migrationsverket genom antal nyanlända på kommunnivå. En indikator på arbetsmarknadsåtgärder samlas in genom antal öppet arbetslösa på kommunnivå från Arbetsförmedlingen.

[1] Lagen om stöd och service till vissa funktionshindrade (LSS) innebär att personer med funktionshinder har rätt till stöd.

Antalet personer som får ekonomiskt bistånd samlas in från Socialstyrelsen. Denna variabel samlas in då den ger en indikator på den socioekonomiska situationen i kommunerna (se avsnitt 3 för mer information om datainsamlingen).

2.4.2 INTÄKTER

För att få information om intäkterna i kommunerna hämtas skatteunderlaget och skattesatsen in från SCB. Skatteunderlaget är summan av den beskattningsbara förvärvsinkomsten för personer kommuninvånare i kommunen föregående år. Det innebär att skatteunderlaget för exempelvis 2019 är summan av den beskattningsbara förvärvsinkomsten för 2018. Även skattesatsen samlas in från SCB då den kan ge en indikator på den ekonomiska situationen i kommunen. En hög skattesats kan tyda på en pressad ekonomisk situation och en lägre skattesats kan tyda på en mindre pressad situation.

2.4.3 ANDRA FÖRKLARANDE VARIABLER

Folkmängden samlas in från befolkningsstatistiken på SCB för att få en indikator på storleken på kommunen. Även befolkningsökningen samlas från SCB in då den ger en uppfattning om kommunen är en inflyttnings-eller en utflyttnings kommun.

Medianåldern i kommunen samlas in då en högre medianålder kan påverka en kommuns resultat då det kan ses på ett mått åldersstrukturen.

3. DATAINSAMLING

3.1 DATAINSAMLING

Det var ett omfattande arbete att sammanställa data för alla olika kommuner och år. Informationen hämtades från en mängd olika källor. Vissa kommuner exempelvis Heby bytte även län och kod under insamlingsperioden. Statistiken på antal nyanlända som tillhandahölls av Migrationsverket var i olika format för olika tidsperioder och sträckte sig inte långt tillbaka i tiden. Det saknades även observationer för flera kommuner och år.

När all statistik var insamlad visade det sig att analysperioden blev 2002 till 2018. Vissa källor fanns längre tillbaka i tiden men statistiken från Försäkringskassan fanns bara för 2002 och framåt. Analysperioden för att testa de två olika statistiska metoderna blev därmed årsdata från åren 2002 till 2018.

3.2 DATA ÖVERSIKT

Tabell 2 visar några nyckeltal för det finansiella resultatet i svenska kommuner från 2002 till 2018. Histogram för variablerna syns i bilaga 1. För flera av variablerna (bland annat det finansiella resultatet) är standardavvikelsen väsentligt större än medelvärdet, vilket innebär att det är betydande variation i dessa variabler i svenska kommuner under den här perioden.

En del observationer saknades för vissa variabler så som migration pga brister i datamaterialet (se avsnitt 3.1 för mer info). Ett elastiskt nät i R kan endast köras på kompletta dataset, vilket innebär att vissa observationer tas bort från datasetet. Det totala antalet observationer i skattningarna är 4 049 observationer trots att statistik på vissa variabler finns för samtliga 4 930 observationer.

Serien som används för att skatta fram resultaten testas för stationäritet med Dickey-Fuller testet. Stationäritet innebär att fördelningen är oberoende av tid. Medelvärdet och variansen är därmed konstanta över tid (Wooldridge 2012, s.381) Noll hypotesen om att tidserien är icke-stationär kan förkastas på 1% nivån för samtliga serier.

3.3 VARIABLER OCH KÄLLOR

Tabell 1 visar vilka variabler som samlades in och vilken lag som används i skattningarna. Samtliga förklarande variabler används med en lag på -1 i skattningarna. Det innebär att informationen från föregående år används i skattningarna för år t. Tabell 2 visar nyckeltal för variablerna som används i skattningarna. Histogram som visar fördelningen på storleken av variablerna finns i bilaga 1.

Tabell 1. Variabelförteckning

Variabel	Typ av variabel	Källa	Lag
Finansiellt resultat	Responsvariabel	Kommunräkenskaperna, SCB	Ingen
Skatteunderlag	Förklarande variabel	Kommunräkenskaperna, SCB	-1
Skattesats	Förklarande variabel	Kommunräkenskaperna, SCB	-1
Medianålder	Förklarande variabel	Kommunräkenskaperna, SCB	-1
Antal öppet arbetslösa	Förklarande variabel	Arbetsförmedlingen	-1
Folkmängd	Förklarande variabel	Befolkningsstatistiken, SCB	-1
Befolkningsökning	Förklarande variabel	Befolkningsstatistiken, SCB	-1
Antal personer över 70 år	Förklarande variabel	Befolkningsstatistiken, SCB	-1
Antal personer under 20 år	Förklarande variabel	Befolkningsstatistiken, SCB	-1
Antal nyanlända	Förklarande variabel	Migrationsverket	-1
Antal med LSS	Förklarande variabel	Försäkringskassan	-1
Antal med ekonomiskt bistånd	Förklarande variabel	Socialstyrelsen	-1

Tabell 2. Nyckeltal för variabler som används i studien för observationerna som används i skattningarna (n= 4 049)

Variabel	Medelvärde	Standardavvikelse	Minimum	Maximum
Finansiellt resultat	53 320	271 774	−799 916	9 594 969
Skatteunderlag	$6{,}068\times10^{9}$	13 537 349 896	$3{,}488\times10^{8}$	$2{,}460\times10^{11}$
Skattesats	21,48	1,38	17,12	33,60
Medelålder	42,74	2,54	36,10	49,80
Arbetslösa	781	1730	43	20 967
Folkmängd	35 892	69 577	2568	962 154
Befolkningsökning	314	1069	−524	19 297
Antal äldre	929	1691	72	22 535
Antal i skolålder (under 19 år)	8212	14 241	520	205 917
Antal nyanlända	117	242	0	5 401
Antal personer med LSS	56	95	3	1255
Antal personer med ekonomiskt bistånd	1074	2560	21	30 020

Möjligheten att använda dummies för kommuner i skattningarna undersöks. Eftersom statistiken som samlades in bedöms fånga merparten av variationen i kommuners resultat (se diagram 1) tas dummies inte med i regressionerna. Det är svårt att tänka ut någon specifik variabel som inte fångas av de insamlande variablerna men som en dummy skulle fånga upp. Ansatsen blir därmed en mer generell modell utan dummies.

4. METODER

Syftet med denna studie är att jämföra ett elastiskt nät med en baseline i att prognostisera det finansiella resultatet i svenska kommuner. I denna del av uppsatsen beskrivs först Ordinary Least Squares (OLS). Metoden baserad på ett elastiskt nät förklaras sedan genom att först gå igenom dess två beståndsdelar: Ridge och Lasso regressioner.

4.1 OLS – BASELINE

Baseline är en OLS regression i denna studie. OLS har funnits länge och beskrivs i en stor mängd läroböcker (James et al. 2013 s.59). Det är en relativt enkel modell och det betonas ofta att det är viktigt att förstå OLS innan man vänder blicken mot mer avancerade modeller (James et al. 2013, s.59).

OLS är en linjär ekvation som estimerar $\beta_0, \beta_{1,\dots,}\beta_{p,}$ för att minimera:

$$RSS = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{i=1}^{p}\beta_j x_{ij}\right)^2 \quad \textbf{(1)}$$

RSS står för ”residual sum of squares”. OLS gör en rad antaganden som begränsar flexibiliteten i modellen. Det starkaste antagandet är att modellen är linjär. OLS kan också vara benägen till överanpassning (”overfitting”). Det innebär att OLS inte alltid ger träffsäker prediktion på ett nytt dataset eftersom den anpassar sig för mycket till datasetet som användes för att ta fram koefficienterna. Ett försök att göra OLS till en mer flexibel modell är att använda Ridge, Lasso och Elastiska nät skattningar.

4.2 ELASTISKT NÄT

4.2.1 ELASTISKT NÄT

Trevor och Hastie (2005) noterade att OLS ofta inte är en effektiv estimator för prediktion. I ett en studie där de utvecklar det elastiska nätet skrivs att det är välkänt att OLS *"performs poorly in both prediction and interpretation"* (Trevor & Hastie 2005, s. 301). Flera metoder har tagits fram för att göra OLS mer flexibel och därmed användbart för prognoser. Ridge regressionen togs fram av Hoerl och Kennard 1988 och Lasso regressionen arbetades fram av Tibshirani 1996. Ett elastiskt nät är en vidareutveckling av dessa två metoder (Trevor & Hastie, 2005). För att förstå det elastiska nätet beskriver vi först dess två komponenter och förklarar kortfattat hur dessa metoder fungerar.

4.2.2 RIDGE REGRESSION

I skattningar med Ridge, Lasso och ett elastiskt nät är det viktigt att variablerna är standardiserade. Det innebär att variablerna är på samma skala. Variablerna standardiseras med följande formel:

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_{ij} - \overline{x_j}\right)^2}} \quad \textbf{(2)}$$

I en OLS regression behöver koefficienterna inte standardiseras. Detta är eftersom koefficienterna är skal ekvivalenta (”scale equivalent”). Detta är inte fallet i skattningar med Ridge, Lasso eller elastiskt nät. Skattningarna varierar med skalan och behöver därmed standardiseras. Trots detta standardiseras ändå koefficienterna i OLS regressionen enligt (2) i denna studie.

En Ridge regression estimerar β_0, $\beta_{1,\dots,}\beta_{p,}$ för att minimera följande uttryck:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{i=1}^{p}\beta_j\tilde{x}_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2$$

$$= RSS + \lambda\sum_{j=1}^{p}\beta_j^2 \quad \textbf{(3)}$$

I uttrycket ovan är $\lambda \geq 0$ en ”tuning parameter”. RSS i uttrycket ovan tar fram koefficienter som passar statistiken väl och gör RSS så litet som möjligt. Den andra delen av uttrycket (3):

$$\sum_{j=1}^{p}\beta_j^2$$

kallas för ”shrinkage penality” och är liten när $\beta_{1,\dots,}\beta_p$ är nära noll. Det innebär att den andra delen av ekvation 1 krymper skattningarna av β_j mot noll. Parametern λ kontrollerar effekten av de olika delarna på skattningarna. När $\lambda = 0$ har den andra delen ingen effekt och Ridge skattningarna blir en vanlig OLS. När $\lambda \to \infty$ blir den andra delen av uttrycket starkare vilket innebär att koefficienterna närmar sig noll. Skattningar med Ridge regressionen generar olika set av koefficienter för varje värde av λ. Värdet av λ är därför viktigt och brukar tas fram genom kors-validering (cross-validation).

Kors-validering för att ta fram värdet på λ kan beskrivas på följande sätt:

Steg 1. Datasetet delas slumpmässigt in i ett antal lika stora grupper. I denna studie kommer 10 grupper att användas.

Steg 2. Observationerna i grupp ett, två fram till grupp nio delas in i så kallade träningsgrupper. Dessa grupper används för att skatta fram en regressionsmodell där λ sätts till ett litet tal nära noll.

Steg 3. Denna modell används för att skapa prediktioner för grupp tio. Denna grupp kallas för testgruppen och en medelkvadratsumma för residualerna (MSE) beräknas.

Steg 4. För samma värde på λ görs proceduren mellan steg 1–3 om där träningsgruppen och testgruppen skiftar. Alla grupper får vara testgrupper en gång var.

Steg 5. Medelvärdet för de framräknande MSE-värdena skattas fram.

Steg 6. Steg 1–5 upprepas med ett annat något högre λ-värde.

Steg 7. Steg 1–6 görs om tills λ-värdet är så pass högt att modeller med bara en koefficient erhålls (restriktionen här att alla andra koefficienter har satts till noll).

Steg 8. Värdet på λ med lägst genomsnittlig MSE (se steg 3) används för att skatta fram regressionsmodellen.

Det är värt att nämna att ”shrinkage penality” tillämpas på $\beta_{1,\dots,}\beta_p$ men inte på skärningen (intercept) β_0. Ridge regressionen minimerar RSS (se ekvation 1) med hänsyn till L2 normen. En Ridge regression tar med alla förklarande variabler i modellen och kan inte krympa en koefficient till noll, även om den kan krympa koefficienterna till nära noll.

4.2.3 LASSO REGRESSION

En Lasso regression bygger på Ridge regressionen men kan ta bort variabler från skattningarna (”subset selection”) genom att krympa vissa koefficienter till noll. En Lasso regression skattar fram $\beta_0,\ \beta_{1,\dots,}\beta_p,$ för att minimera följande uttryck:

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{i=1}^{p}\beta_j\tilde{x}_{ij}\right)^2+\ \lambda\sum_{j=1}^{p}|\beta_j| \\ =RSS+\ \lambda\sum_{j=1}^{p}|\beta_j| \qquad \textbf{(4)}$$

Uttrycket ovan är likt Ridge estimatorn men har en skillnad: β_j^2 i Ridge regressionen är ersatt med $|\beta_j|$ i Lasso estimatorn. Det innebär att en Lasso estimator använder en L1 normen istället för L2 i Ridge estimatorn.

Lasso estimatorn krymper koefficienterna mot noll men kan krympa koefficienterna till noll när ”tuning-parameter” λ är stor nog. Det innebär att en Lasso skattning kan välja ut variabler och en modell (”subset selection”).

4.2.4 ELASTISKT NÄT I FORMLER

Ett elastiskt nät kan ses som en generalisering av såväl Ridge regression som Lasso regression. Ett elastiskt nät skattar fram $\beta_0,\ \beta_{1,\dots,}\beta_{p,}$ för att minimera följande uttryck:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{i=1}^{p}\beta_j\tilde{x}_{ij}\right)^2 + \lambda\left((1-\alpha)\sum_{j=1}^{p}\beta_j^2 + \alpha\sum_{j=1}^{p}|\beta_j|\right)$$

$$= \text{RSS} + \lambda\left((1-\alpha)\sum_{j=1}^{p}\beta_j^2 + \alpha\sum_{j=1}^{p}|\beta_j|\right) \quad \textbf{(5)}$$

Ett elastiskt nät kombinerar därmed en Ridge regression och en Lasso regression. I uttrycket ovan är α en parameter som kan ändras för att kontrollera skattningen. När α = 0 så blir uttrycket ovan en Ridge regression och när α = 1 blir uttrycket en Lasso regression. När α ligger mellan 0 och 1 skattas en blandning mellan en Lasso och en Ridge regression. På samma sätt som i Lasso regressionen och Ridge regression skattas λ fram med kors-validering.[2] Värdet på α kan skruvas på för att undersöka olika modeller. Denna uppsats testar tio olika värden av α. Det går även att skatta fram α med kors-validering. Målet med denna uppsats är inte att hitta den optimala

[2] I paketet som används för skattningarna i denna studie (GLMNET i programvaran R) används kors-validering med 10 olika segment

modellen utan att jämföra prediktion med ett elastiskt nät med OLS. Det är därför intressant att skruva på värdet av α och jämföra det resulterande medelkvadratfelet från regressionen med medelkvadratfelet från en OLS skattning.

4.3 MÅLVARIABEL

För att jämföra vilken metod som tar fram de mest träffsäkra prognoserna på det finansiella resultatet i kommunerna behövs ett mått för att jämföra de olika metoderna. I detta fall är medelkvadratfelet eller ”mean-square error” (MSE) ett bra mått på vilken metod som är mest träffsäker.

Medelkvadratfelet kan definieras på följande sätt där N är antalet observationer:

(6)

$$\frac{1}{N}\sum_{i=1}^{N}\left(Y_i^{prognos} - Y_i\right)^2$$

I denna studie är N=810 då medelkvadratfelet inte beräknas på hela datasetet utan endast på den del av datasetet som används för att prognostisera det finansiella resultatet (se avsnitt 5.2). 20 procent av datasetet används därmed för att testa prognosernas träffsäkerhet.

Detta mått ger en bild av hur långt det prognostiserade värdet ligger från det riktiga värdet. Uttrycket inom parentesen $Y_i^{prognos} - Y_i$ kvadreras för att ta hänsyn till att det ibland sker underskattning och ibland överskattningar. Varje prognos jämförs därmed med det verkliga värdet och kvadreras. Medelkvadratfelen summeras sedan och delas med antalet observationer för att få fram ett mått hur bra en skattningsmetod är i genomsnitt.

Denna målvariabel skattas fram för både OLS och elastiska nät med olika värden på α. Syftet med den här uppsatsen är att jämföra OLS med ett elastiskt nät. Därför prövas olika värden av α för att ta fram elva elastiska nät

(med andra ord en Ridge regression, en Lasso regression och nio elastiska nät) och jämföra dessa med en OLS skattning.

5. RESULTAT

5.1 SKATTNINGAR

Programvaran R används för att göra skattningarna. För att ta fram skattningarna för de elastiska näten används paketet glmnet. Ett annat paket som används för att ta fram skattningarna är dplyr som används för att lagga variablerna. Andra paket som används är bland annat lmtest.

Syftet med den här uppsatsen är inte att förklara sambandet mellan den beroende variabeln (det finansiella resultatet) och de förklarande variablerna utan att jämföra två skattningsmetoder. Detta samband kan vara komplext. Skattningarna från OLS redovisas även i uppsatsen även om dessa inte analyseras vidare. Detta ligger utanför syftet med den här studien.

5.2 TESTDATA OCH TRÄNINGSDATA I PREDIKTION

För att ta fram prognoser görs först skattningar som sedan testas på en ny del av datasetet. Delen av datasetet som används för att ta fram skattningar som sedan testas kallas för träningsdata. Dessa skattningar testas sedan på en mindre del av datasetet som ibland kallas för ett testdata. Det finns inga tydliga riktlinjer på storleken på dessa dataset (Rossi & Inoue, 2012).

Hastie et al. (2008) skriver att minst hälften av datasetet ska utgöra träningsdatamängden. Denna studie använder 80 procent av datasetet för att ta fram skattningar och 20 procent för att testa skattningarna. Observationerna tilldelas till de olika dataseten slumpmässigt.

5.3 RESULTAT FRÅN SKATTNINGAR

Medelkvadratfelet i de olika skattningarna syns i tabell 3.

Tabell 3. Resultat från olika skattningar

Skattningsmetod	Värde på α	Medelkvadratfel
OLS	Inget, då skattningsmetoden inte är ett elastiskt nät	12 153 223 048
Ridge (elastiskt nät med α = 0)	0	8 626 794 455
Elastiskt nät	0,1	9 651 991 006
Elastiskt nät	0,2	11 581 856 497
Elastiskt nät	0,3	10 309 774 905
Elastiskt nät	0,4	9 926 760 910
Elastiskt nät	0,5	10 519 474 830
Elastiskt nät	0,6	11 676 944 837
Elastiskt nät	0,7	11 978 223 997
Elastiskt nät	0,8	10 675 478 739
Elastiskt nät	0,9	10 586 870 526
Lasso (elastiskt nät med α = 1)	1,0	12 051 872 495

Även om fokus i uppsatsen inte är att förklara sambandet mellan det finansiella resultatet och de förklarande variablerna redovisas parametrarna för OLS skattningarna i tabell 4. Variablerna är standardiserade enligt ekvation 2 (se avsnitt 4.2)

Tabell 4. Parametrar från OLS skattning på det kompletta datasetet med standardiserade variabler, 4 049 observationer

Variabel	Koefficient
Intercept	-1,53***
Skatteunderlag$_{-1}$	1,12 **
Skattesats$_{-1}$	0,013
Medianålder$_{-1}$	0,077***
Antal öppet arbetslösa$_{-1}$	0,68***
Folkmängd$_{-1}$	-1,21*
Antal med ekonomiskt bistånd $_{-1}$	-0,15
Befolkningsökning$_{-1}$	0,033
Antal med LSS $_{-1}$	0,012
Antal över 70 år$_{-1}$	-0,68***
Antal i skolålder $_{-1}$	0,80**
Antal nyanlända$_{-1}$	-0,017
Justerad R^2	0,34

* $p \leq 0,05$
** $p \leq 0,01$
*** $p \leq 0,001$

Heteroskedasticitet testades med Breush-Pagan test i R och det visade sig att heteroskedasticitet var närvarande då vi kunde förkasta noll hypotesen på 1% nivån. Det är viktigt att testa för heteroskedasticitet då OLS modellen bygger på att variansen av standardfelet (”error term”) är konstant. När heteroskedasticitet är närvarande innebär det att variansen av standardfelet

inte är konstant. I praktiken kan då variansen av standardfelet öka med värdet på variabeln. För att rätta till detta estimeras standardfelen med en annan teknik. I denna studie användes robusta standardfel enligt White (Wooldridge 2012, s.271). I praktiken innebär detta att det går att se vilka variabler som är statististiskt signifikanta.

Tabell 5. Parametrar från Ridge skattning på det kompletta datasetet med standardiserade variabler, 4 049 observationer

Variabel	Koefficient
Intercept	−0,78
Skatteunderlag$_{-1}$	0,31
Skattesats$_{-1}$	0,011
Medianålder$_{-1}$	0,033
Antal öppet arbetslösa$_{-1}$	0,23
Folkmängd$_{-1}$	0,063
Antal med ekonomiskt bistånd $_{-1}$	−0,15
Befolkningsökning$_{-1}$	0,071
Antal med LSS $_{-1}$	0,043
Antal över 70 år$_{-1}$	−0,21
Antal i skolålder $_{-1}$	0,16
Antal nyanlända$_{-1}$	0,045

5.4 JÄMFÖRELSE MELLAN OLS OCH ELASTISKT NÄT

Tabell 3 visar att ett elastiskt nät ger ett lägre medelkvadratfel jämfört med baseline OLS. Skattningarna med OLS ger ett högre medelkvadratfel jämfört med samtliga elastiska nät. Medelkvadratfelet med OLS blir 12 153 223 048. Det innebär att i genomsnitt hamnar en prognos med OLS runt 110 000 kronor från det riktiga finansiella resultatet.

Ett elastiskt nät ger en mer träffsäker prognos. Den bästa prognosen skattas i detta fall fram av ett elastiskt nät med $\alpha = 0$ eller med andra ord en Ridge regression. I detta fall blir medelkvadratfelet 8 626 794 455. Det innebär att i genomsnitt hamnar en prognos med ett elastiskt nät runt 93 000 kr från det verkliga resultatet i svenska kommuner mellan 2002 och 2018. Det kan sättas i förhållande till standardavvikelsen som ligger på ungefär 270 000 kr (se tabell 2). Att standardavvikelsen är relativt hög i förhållande till medelvärdet innebär att spridningen i det finansiella resultatet mellan kommuner och tid är ganska stor. Eftersom standardavvikelsen är så pass hög tyder det på att det elastiska nätet ger en relativt god prognos över det finansiella resultatet i svenska kommuner.

I det här fallet ger $\alpha = 0$ det lägsta medelkvadratfelet. Det innebär att en Ridge regression skattar fram den bästa prognosen i det här fallet. Det innebär att alla förklarande variabler tas med i skattningen och ingen ”subset-selection” äger rum (dvs. inga förklarande variabler plockas bort). Detta är trots att alla variabler inte är statistiskt signifikanta i OLS regressionen (se tabell 4). Den minst träffsäkra prognosen av de elastiska näten är när $\alpha = 1$ eller i andra ord en Lasso regression. Denna prognos är ändå mer träffsäker än resultat med OLS.

6. DISKUSSION

En Ridge regression eller med andra ord ett elastiskt nät med $\alpha = 0$ ger det lägsta medelkvadratfelet. Denna metod är bättre på att förutspå det finansiella resultatet i svenska kommuner mellan 2002 och 2018 jämfört än en baseline i form av skattningar med OLS. Samtliga elastiska nät som användes i skattningarna gav lägre medelkvadratfel än baseline (skattningar med OLS) i denna studie. Det innebär att flexibiliteten som ett elastiskt nät ger innebär mer träffsäkra prognoser jämfört med OLS i det här fallet.

Alla förklarande variabler i OLS skattningarna är inte signifikanta (se tabell 4). Trots detta ger en Ridge regression den bästa skattningen. En Ridge regression tar med samtliga förklarande variabler i regressionen. Detta tyder på att samtliga förklarande variabler bidrar till prediktionen och har förklaringsvärde trots att de inte är statistiskt signifikanta.

En kritisk synpunkt är att samtliga modeller i denna studie inte är så träffsäkra. Avsikten med den här studien har endast varit att undersöka om det går att åstadkomma någonting bättre genom ett använda ett elastiskt nät jämfört med OLS. Resultaten visar att det elastiska nätet ger mer träffsäkra skattningar även om dessa inte kan betraktas som tillfredsställande.

Det hade nog varit möjligt att få en avsevärt mycket bättre prediktion genom att göra en djupare analys av hur de olika förklarande variablerna påverkar resultatet. Denna studie har antagit en linjär påverkan. Histogrammen i bilaga 1 visar en del variabler har många små värden och ett fåtal stora värden. Det innebär att effekten av skattningen då blir beroende av resultaten vid de stora värden på den förklarande variabeln. En vidareutveckling av denna studie skulle därför vara att titta närmare på effekten av de olika variablerna på det finansiella resultatet och använda denna analys för att ta fram en bättre prognos.

7. SLUTSATSER OCH MÖJLIGA FÖRBÄTTRINGAR

Denna studie visar att elastiskt nät skattar mer träffsäkra prognoser av det finansiella resultatet i svenska kommuner mellan 2002 och 2018 än en baseline. Den bästa predikatorn av det finansiella resultatet enligt denna studie är Ridge regressionen (ett elastiskt nät med $\alpha = 0$). Detta är i linje med vad Hastie et al. (2005) föreslog när de utvecklade det elastiska nätet. Ett elastiskt nät är i regel mer träffsäker i prediktion jämfört med OLS.

Denna studie visar att ett elastiskt nät ger en indikator av det finansiella resultatet i svenska kommuner genom att använda värden på förklarande variabler från föregående år. Studien kan vidareutvecklas och förbättras på flera sätt.

Skulle man vilja använda resultaten i praktiken bör uppdelningen i de två dataseten (testdata och träningsdata) genomlysas. Resultaten är känsliga för den här uppdelningen. Mer data skulle kunna ge fler möjligheter att testa resultaten. Om man väntar några år blir datasetet större och modellen kan vidareutvecklas.

En intressant ny studie skulle kunna göras om man delade upp kommuner i olika grupper baserade på exempelvis landsbygd och storstad. Man skulle sedan kunna testa metoderna som analyserades i denna studie på det stratifierade datasetet och på så sätt utreda om metoderna ger bättre resultat på en viss del av datasetet.

Det skulle vara intressant att jämföra andra modeller i att ta fram en prognos på det finansiella resultatet i svenska kommuner. Exempelvis skulle principalkomponent analys kunna ge intressanta resultat. Det skulle även vara intressant att testa icke-parametriska modeller i framtagandet i prognoser.

Dessa modeller skulle troligen kräva en större datamassa men skulle kunna ge intressanta resultat.

Det skulle även vara intressant att genomföra en analys av sambandet mellan de förklarande variablerna och det finansiella resultatet i kommunerna mer utförligt. Detta ligger utanför denna uppsats men skulle troligen höja prognosernas träffsäkerhet markant.

REFERENSER

Arbetsförmedlingen (2019), ”Statistik”, Tillgänglig online [2019-12-29] https://arbetsformedlingen.se/om-oss/statistik-och-analyser/statistik

Basuchoudhary, Atin. Bang, James. & Sen, Tinni. (2018). Machine-learning Techniques in Economics – New Tools for Predicting Economic Growth, 1:a upplagan, New York: Springer

Försäkringskassan (2020), ”Statistik på antal assistansberättigade på kommunnivå”, Tillgänglig online [2020-01-09] https://www.forsakringskassan.se/statistik/funktions-nedsattning/assistansersattning

Hastie, Trevor. Tibshirani, Robert. & Friedman, Jerome. (2009). The Elements of Statistical Learning: Data mining, Inference, and Prediction, 12:e upplagan. New York: Springer

Heed, Robert. & Stark, Hans. (2020). Regler och principer för god ekonomisk hushållning och RUR. Sveriges Kommuner och Regioner, Tillgänglig online [2020-03-09]:

https://skr.se/ekonomijuridikstatistik/ekonomi/godekonomiskhushallning/reglerochprinciperforgodekonomiskhushallningochrur.14862.html4

James, Gareth. Witten, Daniela. Hastie, Trevor. & Tibshirani, Robert. (2013). An Introduction to Statistical Learning – with Applications in R, 8:e New York: Springer

Jung, Jin-Kyu. Patnam, Manasa. & Martirosvan, Anna-Ter. (2018). An Algorithmic Crystal Ball: Forecasts based on Machine Learning, *IMF Working Paper*, Tillgänglig online [2020-11-24]:

https://www.imf.org/en/Publications/WP/Issues/2018/11/01/An-Algorithmic-Crystal-Ball-Forecasts-based-on-Machine-Learning-46288

Konjunkturinstitutet. (2019). Konjunkturläget december 2019. Stockholm: Konjunkturinstitutet. Tillgänglig online [2020-06-28]: https://www.konj.se/download/18.4a42c8be16f1a7f992c34cf/1576744727462/KLDec2019.pdf

Migrationsverket (2020), "Anvisning till kommuner och bosättning" Tillgänglig online [2020-01-07] https://www.migrationsverket.se/Om-Migrationsverket/Statistik/Anvisning-till-kommuner-och-bosattning.html

Rossi, Barbara. & Atsushi, Inoue. (2012). Out-of-Sample Forecast Tests Robust to the Choice of Window Size, *Journal of Business and Economic Statistics,* Vol. 30, No. 3, s. 432-453.

Shmueli, Galit. (2010). To Explain or to Predict, *Statistical Science*, vol.25. No. 3, s. 290-310

Socialstyrelsen (2019), "Statistik om ekonomiskt bistånd", Tillgänglig online [2020-01-08] https://www.socialstyrelsen.se/statistik-och-data/statistik/statistikamnen/ekonomiskt-bistand/

Statistiska centralbyrån (2020), " Folkmängden efter region, civilstånd, ålder och kön. År 1968 - 2019", Tillgänglig online [2020-01-03] http://www.statistikdatabasen.scb.se/pxweb/sv/ssd/START__BE__BE0101__BE0101A/BefolkningNy

Statistiska centralbyrån (2020), " Kommunala skattesatser. År 2000 - 2020", Tillgänglig online [2020-01-02]

http://www.statistikdatabasen.scb.se/pxweb/sv/ssd/START__OE__OE0101/Kommunalskatter2000/

Statistiska centralbyrån (2020), ” Kostnader och intäkter för kommuner efter region och verksamhetsområde. År 2011 - 2019”,

Tillgänglig online [2020-01-02] https://www.statistikdatabasen.scb.se/pxweb/sv/ssd/START__OE__OE0107__OE0107B/KostnDR/

Statistiska centralbyrån (2020), ” Resultaträkning för kommuner efter region och resultaträkningsposter. År 1998 - 2018”, Tillgänglig online [2020-01-02]http://www.statistikdatabasen.scb.se/pxweb/sv/ssd/START__OE__OE0107__OE0107A/ResultKn/

Statistiska centralbyrån (2020), ” Skatteunderlag och skattekraft. År 1995 - 2020”, Tillgänglig online [2020-01-02] http://www.statistikdatabasen.scb.se/pxweb/sv/ssd/START__OE__OE0101/SkatteKraft/

Tibshirani, Robert. (1996). Regression Selection and Shrinkage via the Lasso, *Journal of the Royal Statistical Society*. Series B (Methodological), Vol. 58, No. 1, s. 267–288.

Tiffin, Andrew. (2016). Seeing in the Dark: A Machine-Learning Approach to Nowcasting in Lebanon, *IMF Working Paper*, Tillgänglig online [2020-11-24]: https://www.imf.org/external/pubs/ft/wp/2016/wp1656.pdf

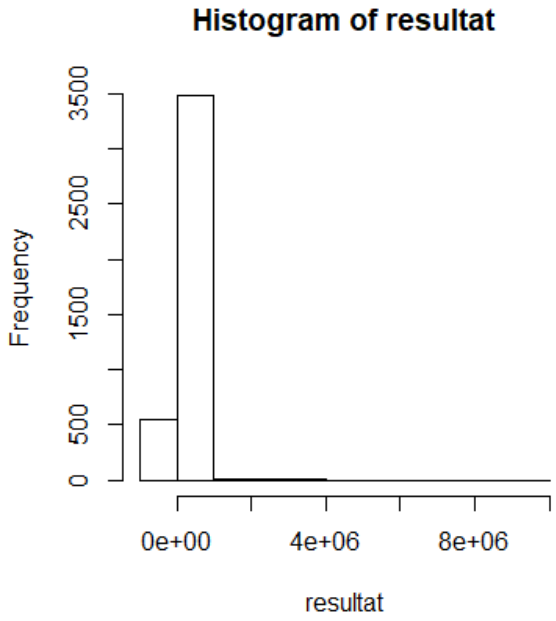
Varian, Hal R. (2014). Big data: New tricks for econometrics, *The Journal of Economic Perspectives*, No. June 2013, s. 1–36.

Wooldridge, Jeffrey M. Introductory Econometrics: A Modern Approach, 5:e, Boston: Cengage Learning

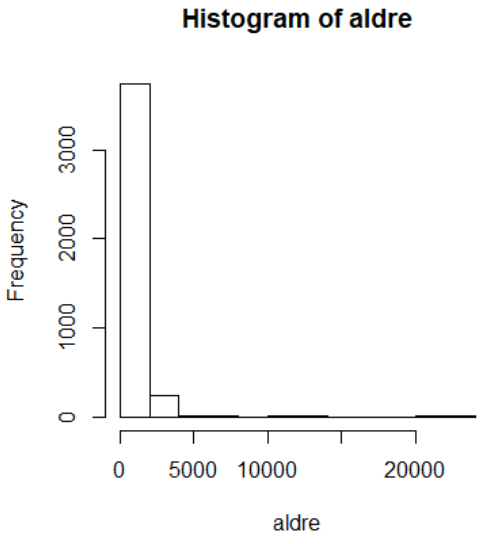
Zou, Hui. & Hastie, Trevor. (2005). Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society*, Vol. 67, No. 2, s. 301–320.

Bilaga 1.

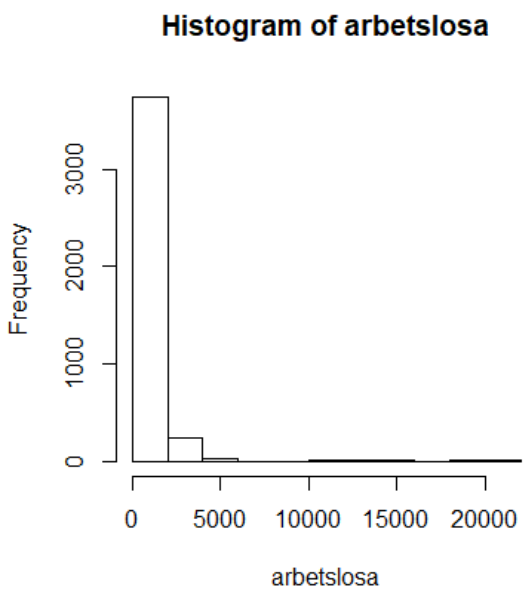
Histogram på det finansiella resultatet i kommuner, n= 4 049



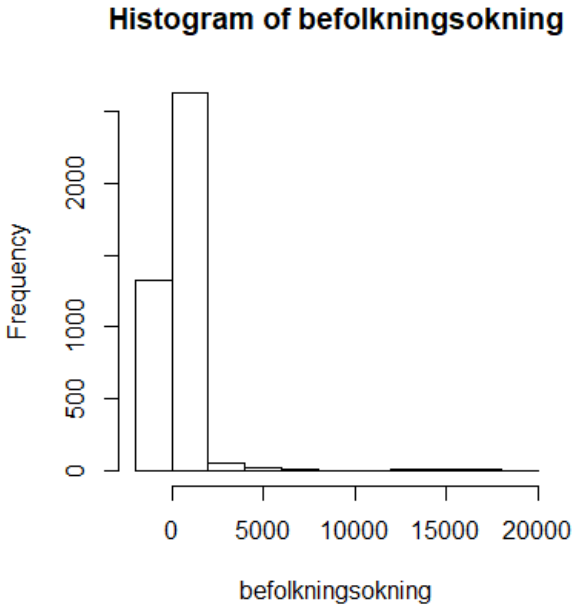
Histogram på antal personer över 70 år i kommuner, n=4 049



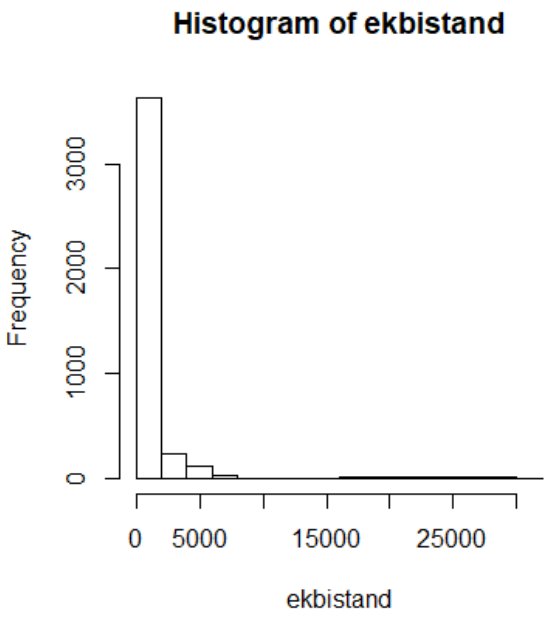
Histogram på antal öppet arbetslösa i kommuner, n= 4 049



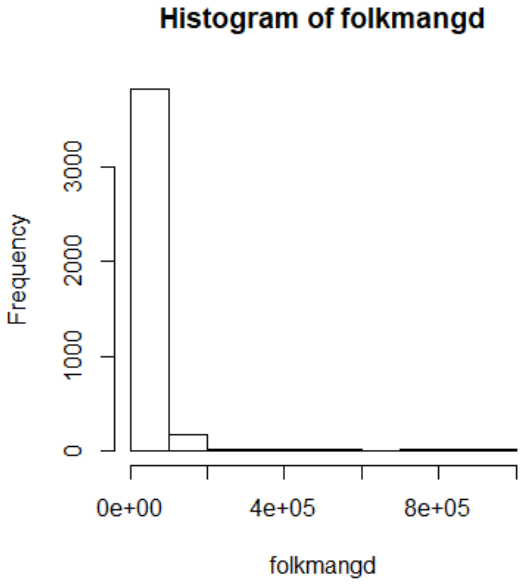
Histogram på befolkningsökningen i kommuner, n=4 049



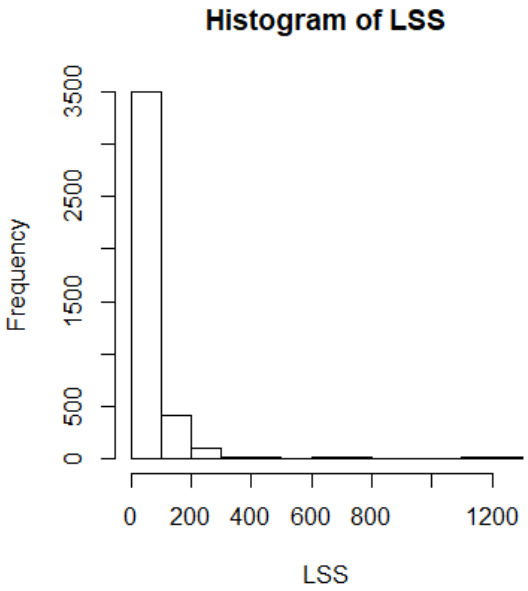
Histogram på antal personer som får ekonomiskt bistånd i kommuner, n=4 049



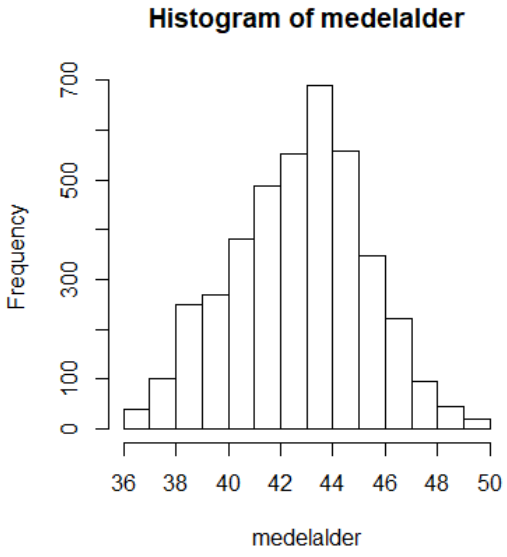
Histogram på folkmängd i kommuner, n=4 049



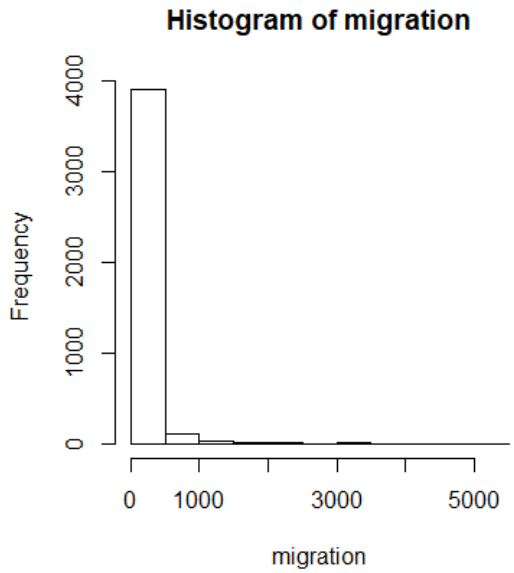
Histogram på antal personer med LSS i kommuner, n=4 049



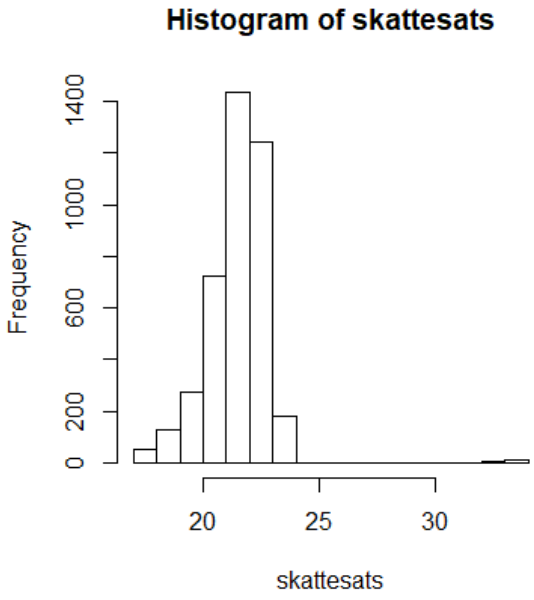
Histogram på medianåldern i kommuner, n=4 049



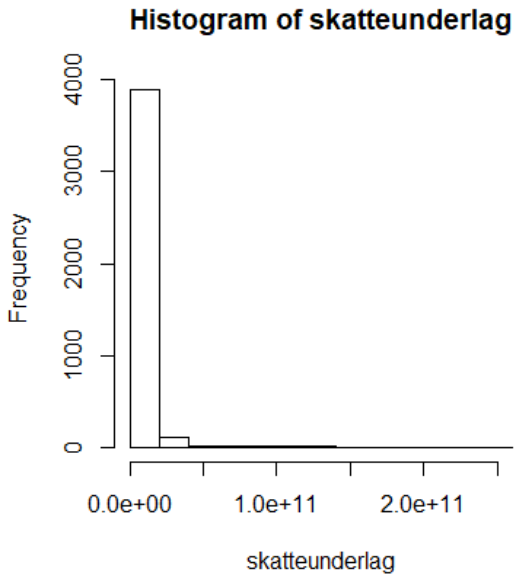
Histogram på antal nyanlända i kommuner, n=4 049



Histogram på skattesats i kommuner, n=4 049



Histogram på skatteunderlaget i kommuner, n=4 049



Histogram på antal personer under 19 år i kommuner, n=4 049