Födas, åldras, låna

ett sökande efter samband mellan födslotal och skuldsättning

Jan Novotny

handledare: Jonas Wallin

Vårterminen 2018

Kandidatuppsats i statistik (15 hp)

Innehållsförteckning

Abstract

Household debts is interesting for a number of institutions in society. One reason for this is that it gives a picture of how severely a potential economic crisis might affect the country. If a model can be found which precisely explains the household debt levels, it can be used to predict dangerous situations at an earlier stage, so that resources to counteract the situation can be employed before it turns into a large-scale meltdown. This thesis tries to find such a model, by examining whether there is a relation between debt level (measured as the quota between total debt and disposable income) and the natural variation in the number of births each year over a longer period of time. The thesis also examines whether the number of births, when used as explanatory variable, can outperform a model where interest rate is used as explanatory variable instead, with regard to explanatory power. The models used are estimated via application of time series analysis (AR(p)-models) and LASSO regression. The author finds signs indicating that there might be a connection between the number of births and debt levels, but nothing indicates that number of births is a better explanatory variable than for instance interest rate levels. It is also shown that the time series analysis gives the largest contribution in explaining debt levels, and a more rigorous analysis of time series is proposed for potential future studies on the topic.

Key words: debt levels, number of births, times series analysis, LASSO regression

Sammanfattning

Hushållens skuldsättningsgrad är av intresse för en mängd instanser i samhället eftersom denna bland annat kan ge en bild av hur allvarligt en ekonomisk kris skulle drabba hushållen. Om man kan hitta en modell som väldigt väl förklarar hushållens skuldsättningsnivå kan man eventuellt också upptäcka potentiellt farliga situationer i ekonomin tidigare och då ha möjlighet att sätta in åtgärder mot dem innan de blommat ut i fullskalig kris. I denna uppsats görs ett försök att hitta en sådan modell, genom att undersöka ifall det finns ett samband mellan skuldsättningsnivån (mätt som skuldkvot mellan total skuld och disponibel inkomst) och den naturliga variationen i antalet födda barn varje år över en längre tidsperiod. Det undersöks även om antalet födda barn, då det används som förklarande variabel, ger en högre förklaringsgrad jämfört med en modell där låneränta används som förklarande variabel. De modeller som används tas fram genom tillämpning av tidsserieanalys (AR(p)-modeller) samt LASSO-regression. Författaren finner tecken på att det skulle kunna finnas samband mellan antalet födda barn och skuldkvoten, men inget som tyder på att antalet födda barn skulle vara en bättre förklarande variabel än exempelvis låneräntan. Det framkommer även att tidsserieanalysen ger det största bidraget i fråga om att förklara skuldsättningsnivån, och det föreslås en mer rigorös tidsserieanalys för eventuella framtida studier.

Nyckelord: skuldsättning, födslotal, tidsserieanalys, LASSO-regression

Inledning

Den som är satt i skuld är inte fri, brukar det heta. Tycka vad man vill om att låna pengar, men faktum är att det är helt avgörande för vårt sätt att leva. Enligt OECD (OECD, 2018) så lånade år 2015 det genomsnittliga svenska hushållet 1 krona och 78 öre för varje krona som det tjänade. Dessa lån kan vara av olika slag, som exempelvis huslån, studielån eller kreditkort.

I den bästa av världar hade man förstås varit så rik att man aldrig behövde låna pengar till någonting utan bara kunde köpa allt kontant. Sådan är dock inte världen för de allra flesta. De allra flesta måste låna pengar till sådant som hus och utbildning. Detta gör man mot en kostnad, kallad ränta, som helt enkelt är det pris man betalar för att kunna använda en viss mängd pengar idag istället för att behöva vänta tills man sparat ihop till så mycket själv.

Så, om räntan höjs, och det alltså blir dyrare att låna, verkar det rimligt att anta att folk kommer låna mindre, givet att inget annat förändras (vilket det förstås alltid gör, men det kan man försöka ta hänsyn till). Vill man bygga en statistisk modell där man förklarar skuldsättningen i ett land kan därför räntan förmodligen hjälpa till en hel del. Kanske skulle man emellertid även kunna förklara skuldsättningen på ett lite annorlunda sätt.

Vi har konstaterat att de *flesta* behöver låna pengar, men *alla* behöver inte låna pengar. Om du bor i hyreslägenhet, inte har någon högskoleutbildning, och inte handlar på kredit, ja då har du sannolikt inte behövt låna så mycket pengar i dina dagar. Likaså om du är barn. Är du pensionär har du möjligen lånat pengar i förfluten tid, men idag lär du inte behöva låna så mycket mer.

Om du å andra sidan nyss tagit studenten och vill börja studera, eller om du skaffat ett par barn och det börjar bli trångt i hyreslägenheten, eller om du nyss skaffat ditt första barn och pengarna inte alltid räcker till alla oväntade utgifter, ja då är nog sannolikheten att du behöver låna pengar högre.

Vad saker som högskolestudier och barnafödande har gemensamt är att de är starkt *åldersbundna*. De flesta studerar när de är unga, de flesta får barn när de är runt 30 år (SCB, 2013). Man kan således förmoda att om det ett givet år exempelvis finns många nittonåringar kommer volymen studielån öka det året (återigen givet att inget annat förändras).

Eftersom antalet barn som föds varje år varierar kommer också antalet potentiella låntagare ett visst antal år senare att variera, vilket i sin tur kan tänkas hänga ihop med antalet faktiska låntagare och därigenom volymen lån som tas i landet. Huruvida ett sådant samband mellan barnkullarnas storlek och skuldsättningen faktiskt går att finna är vad denna uppsats är ämnad att undersöka. Den ska även undersöka hur väl barnkullarnas storlek förklarar skuldsättningen i jämförelse med den mer rättframma modellen med räntenivå som beskrivits ovan.

Enligt dessa premisser upprättas alltså följande frågeställning:

- Vad finns det för samband mellan antalet födslar ett visst år och skuldsättningen som uppkommer en tid senare?
- Hur väl förklaras skuldsättningen av antalet födslar jämfört med hur väl den förklaras av räntenivån?

Data

Det ursprungliga syftet har inte varit att undersöka landspecifika data, men på grund av begränsningar i tid och tillgång till data har USA valts som land att studera. Eftersom olika länder skiljer sig åt i fråga om finansiella marknader, lagliga ramverk, politisk inriktning med mera bör detta betraktas som en studie av enbart de amerikanska förhållandena.

Insamlingen av data har skett via olika offentliga institutioners hemsidor. Exempel på dessa institutioner är *Federal Reserve*, *United States Census Bureau* och *Bureau of Labor Statistics*. Från de dataset som laddats ned har följande variabler använts för analys:

- Antalet födslar
- Disponibel inkomst
- Låneränta
- Utstående huslån
- Utstående övriga lån
- Utnyttjade krediter

Datan har efter behov bearbetats för att passa syftet. Antalet födslar (Centers for Disease Control and Prevention, 2017) har ursprungligen varit uppdelat på olika etniska grupper. Eftersom detta inte är relevant för uppsatsen har födslotalen istället lagts ihop för att göra det möjligt att studera det totala värdet. Det kan här tilläggas att det inte finns lika långtgående data för samtliga etniska grupper. Den enda gruppen där det finns observerade värden för hela den observerade perioden är ’*White*’. Detta är dock också den överlägset största gruppen, nästan fem gånger så stor som den näst största gruppen ’*Black*’. Övriga grupper har varit *’Asian/Pacific islanders*’ och *’American Indian/Alaska Native*’. I själva modellerna standardiseras antalet födslar. Mer om detta i avsnittet ”Metod”.

Vidare är antalet födslar den variabel, förutom utnyttjade krediter, med minst antal observerade värden, nämligen från år 1960 till 2013. För huslån och övriga lån finns det exempelvis data mellan åren 1952 och 2014. För att kunna undersöka hur födslotalet ett visst år påverkar skuldsättningen ett antal år senare kommer dock antalet observationer som faktiskt studeras bli ännu mindre. (För att se hur antalet trettioåringar ett givet år påverkar skuldkvoten kommer analysen behöva börja år 1990, alltså det år då det för första gången finns en barnkull som hunnit bli trettio år, de barn som föddes 1960. Detta innebär emellertid att data om skuldsättningsgrad från åren innan 1990 inte kan användas, eftersom det där inte finnas data för antalet födslar trettio år tidigare).

För variablerna huslån (Federal Reserve, 2017 b) och disponibel inkomst (U.S. Bureau of Economic Analysis, 2017) har värden från början varit givna för varje kvartal, och för övriga lån och utnyttjade krediter (Federal Reserve, 2017 a) har värdena angivits månadsvis. I samtliga fall används ett framräknat årsmedelvärde istället.

Låneräntan (International Monetary Fund, 2016) är inte räntenivån för någon särskild typ av lån utan istället ett genomsnitt för låneräntor för korta och medellånga lån för privata låntagare.

Datafilen för huslån har även varit uppdelad i många olika kategorier, där huvudkategorierna varit ”*typ av fastighet*” och ”*typ av låntagare*”. I huvudkategorin ”typ av låntagare” har underkategorin ”*individer och andra*” även valts ut, utöver de totala huslånen, för att studeras separat. Denna kategori består alltså främst av enskilda individers huslån och utgör cirka tio procent av de totala huslånen (resterande del av huslånen bärs av olika institutioner). Denna

variabel kommer i resten av texten kallas *”individers huslån”*, medan de totala huslånen helt enkelt kommer kallas *”totala huslån”*.

Vad gäller variablerna övriga lån och utnyttjade krediter kan även ett förtydligande vara på sin plats. Övriga lån är lån som på engelska kallas *’non-revolving credit*’, och rör sig främst om studielån, billån och liknande. Det är alltså den typen av lån där en *klumpsumma* betalas ut som sedan ska betalas tillbaka. Vad som här kallas ’utnyttjade krediter’ är vad som på engelska kallas *’revolving credit*’, och innefattar den typen av krediter som är gjorda för att användas vid upprepade tillfällen. Framförallt rör det sig här om kreditkort och liknande tjänster, där kunden kan utnyttja krediten upp till ett visst belopp, innan hen behöver betala tillbaka för att frigöra ny kredit att handla för.

För att skapa en bild av vilken storleksordning de olika variablerna har visas i *tabell 1* de olika variablernas värden år 2013.

Variabler	*Värden år 2013*
Antalet födslar	3.93 miljoner pers.
Disponibel inkomst	12 396 miljarder US $ *
Låneränta	3.25%
Utstående huslån	13 243 miljarder US $ *
Utstående övriga lån	2 167 miljarder US $ *
Utnyttjade krediter	845 miljarder US $ *

Tabell 1. Variabelvärden år 2013.
** 2013 års US $*
*** Basår 1967*

De variabler som anges i enheten US $ utgör det totala värdet för respektive variabel. De visar alltså med andra ord den sammanlagda summan av alla de enskilda hushållens disponibla inkomst, utstående huslån, och så vidare.

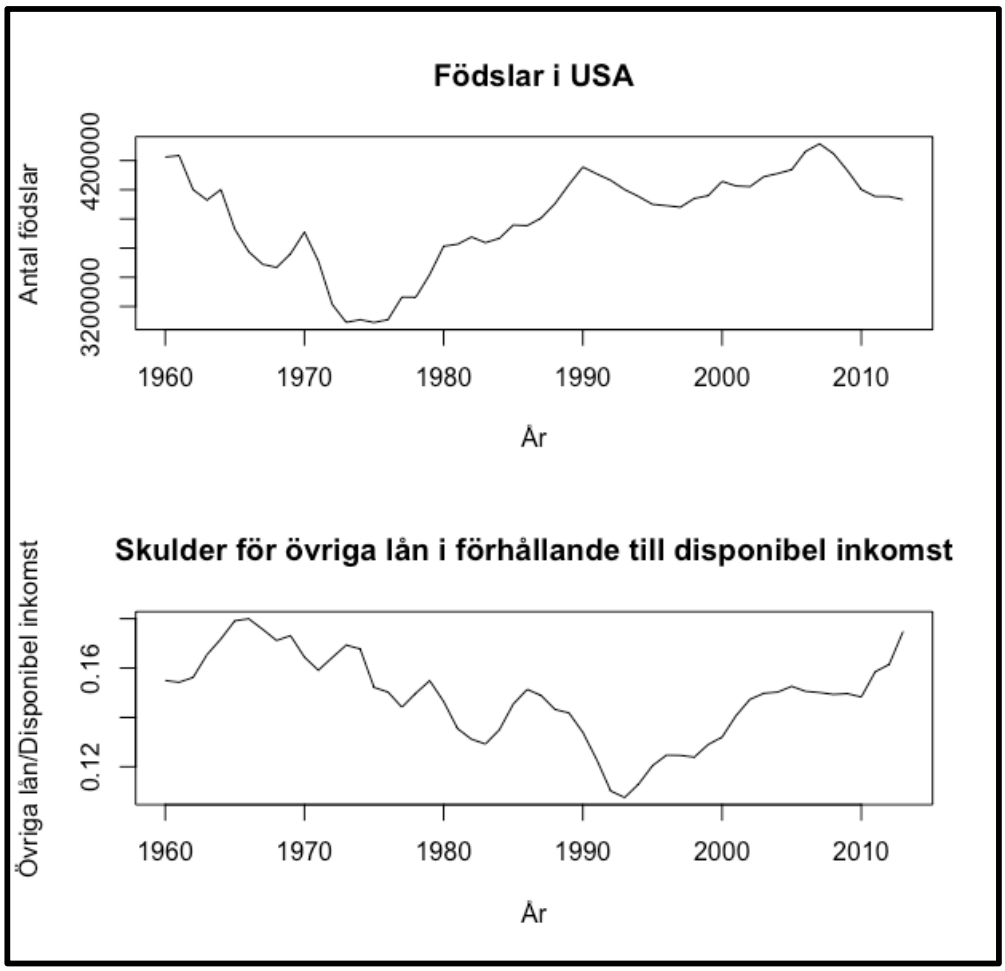
Eftersom totala mängden lån beror på saker som populationens storlek och inkomst är det olämpligt att studera den totala mängden lån. Istället divideras de olika lånebeloppen med den

totala inkomsten. Då används alltså datan för de olika lånetyperna tillsammans med datan för disponibel inkomst för att räkna ut vad som kallas för skuldkvot. Detta görs för alla de olika lånetyperna separat och för alla år mellan 1990-2013, och beräkningen ser ut enligt följande:

$$Skuldkvot = \frac{Utestående\ skuld}{Disponibel\ inkomst}.$$

Skuld är alltså detsamma som utstående lån eller utnyttjade krediter. Dessa tidsserier av skuldkvoter modifieras sedan ytterligare genom att värdena logaritmeras. Detta har gjorts i ett tidigt stadium för att eventuellt kunna analysera hur de förklarande variablerna påverkar skuldkvoten procentuellt.

För att få en första bild av hur datan ser ut visas i *figur 1* tidsserien för antalet födslar i USA intill tidsserien för skuldkvoten för övriga lån i landet.



Figur 1. Födslar och skuldkvoten för övriga lån. Ett visst mönster kan anas; skuldkvoten tycks följa födslotalen, fast med en förskjutning på 18-20 år.

Mellan de lägsta värdena i den övre tidsserien och de lägsta värdena i den nedre tidsserien är det 18 år, och förskjuts den övre tidsserien ungefär 18 år framåt ser man att de båda tidsserierna tycks följa varandra relativt väl. Är detta ögats vilseledning eller ligger det något i det man ser?

Metod

Tidsseriemodellen

Eftersom de olika skuldkvoterna består av data insamlad över flera år är de att betrakta som tidsserier, alltså en serie stokastiska variabler där observationer samlas in över tid (även kallat en stokastisk process). Tidsserier har egenskapen att de är autokorrelerade, vilket innebär att föregående värden är korrelerade med nuvarande värde. Skuldkvoten för ett givet år kommer alltså högst sannolikt vara korrelerad med skuldkvoten för nästkommande år. Innan sambandet mellan skuldkvot och födslotal studeras behövs därför en analys av sambandet mellan skuldkvoternas värden vid olika tidpunkter.

En tidsserie där nuvarande värde påverkas av föregående värde kan beskrivas med formeln (Chan, Cryer, 2008, s. 66):

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + e_t \quad (I),$$

där värdet vid tiden *t* i tidsserien, Y_t, förklaras av en linjär modell innehållandes tidigare värden tillsammans med en s.k. ”innovationsterm”, e_t. Innovationstermen innefattar den slumpmässiga variationen vid tidpunkten *t*, alltså den variation som inte förklaras av föregående värden med tillhörande koefficienter ϕ_i.

AR-processer

Serien Y_t är en autoregressiv process, förkortat AR-process, eftersom den kan beskrivas som en regression på sina föregående värden. En AR-process av formen

$$Y_t = \phi_1 Y_{t-1} + e_t,$$

kallas för en AR(*1*)-process eftersom det nuvarande värdet endast antas bero på det föregående värdet plus innovationstermen. Det föregående värdet i tidsserien kallas även värdet vid *ett tidslagg*. Värdet innan det kallas värdet vid *två tidslagg*, och så vidare. En AR-process av formen som beskrivs i (*I*) är alltså en AR(*p*)-process, eller en AR-process av ordning *p*, eftersom den innefattar värdena vid ett tidslagg, två tidslagg, och så vidare upp till *p* tidslagg.

Vidare, om det gäller för varje sekvens av tidpunkter $t_1, t_2, \dots, t_n$ att den simultana täthetsfunktionen för $Y_{t_1}, Y_{t_2}, \dots, Y_{t_n}$ är densamma som för Y_{t_1-k}, $Y_{t_2-k}, \dots, Y_{t_n-k}$, så är tidsserien stationär, vilket antas för tidsserierna i denna uppsats. För stationära tidsserier kan korrelationen mellan Y_t och Y_{t-k} , kallad *autokorrelation* eftersom den beskriver hur tidsserien korrelerar med *sig själv* vid olika tidslagg, i en population skattas med hjälp av värdena i en observerad tidsserie, som då är att betrakta som ett stickprov. Den skattade autokorrelationen beskrivs då av formeln (Chan, Cryer, 2008, s. 46)

$$r_k = \frac{\sum_{t=k+1}^{n}(Y_t - \bar{Y})(Y_{t-k} - \bar{Y})}{\sum_{t=1}^{n}(Y_t - \bar{Y})^2} \qquad \text{för } k = 1,2, \dots$$

där r_k är den skattade autokorrelationen mellan variabler med tidslagg *k*. I och med att det inte finns någon tydlig ’cut-off’ i autokorrelationen då *k* ökar kan den inte användas som indikator för vilken ordning processen har (alltså vilket värde *p* i (*I)* antar) (Chan, Cryer, 2008, s. 112). För AR-processer kan man istället använda den *partiella* autokorrelationen vid tidslagg *k*, ϕ_{kk}, för att bestämma vilken modell som är mest lämplig. Den partiella autokorrelationen har nämligen egenskapen att den blir noll för alla tidslagg $k > p$. Denna korrelation kan ses som korrelationen mellan Y_t och Y_{t-k} då man kontrollerat för effekten av alla mellanliggande variabler $Y_{t-1}, Y_{t-2}, \dots, Y_{t-k+1}$. Ur en observerad tidsserie fås den skattade partiella autokorrelationen av formeln

$$\hat{\phi}_{kk} = \frac{r_k - \sum_{j=1}^{k-1} \hat{\phi}_{k-1,j} r_{k-j}}{1 - \sum_{j=1}^{k-1} \hat{\phi}_{k-1,j} r_j},$$

där lösningarna för termerna $\hat{\phi}_{k-1,j}$ fås via Yule-Walkers ekvationer (Chan, Cryer, 2008, s. 114-115), och där det kan antas att $\hat{\phi}_{11} = r_1$ (korrelationen mellan Y_t och Y_{t-1} är densamma

som den partiella autokorrelationen eftersom det inte finns några mellanliggande variabler mellan dessa två variabler).

Quenoulle (Chan, Cryer, 2008, s. 115) har visat att om en tidsserie kan beskrivas av en given AR(p)-modell gäller för alla tidslagg $k > p$ att den skattade partiella autokorrelationen är asymptotiskt normalfördelad med väntevärdet $E(\hat{\phi}_{kk}) = 0$ och variansen $Var(\hat{\phi}_{kk}) = \frac{1}{n}$. Detta innebär att om ett värde $|\hat{\phi}_{kk}|$, alltså den skattade partiella autokorrelationen vid tidslagg k, är större än $|\frac{1.96}{\sqrt{n}}|$ (signifikansnivå 5%), så tyder detta på att en AR-modell av *minst ordning k* kan vara lämplig att välja för att beskriva tidsserien i fråga. Omvänt, om $|\hat{\phi}_{kk}|$ är mindre än $|\frac{1.96}{\sqrt{n}}|$ så tyder det på att en AR-modell av *lägre ordning än k* är lämplig. Själva AR-modellerna skattas sedan med hjälp av maximum-likelihood-metoden (Chan, Cryer, 2008, s. 158-160).

Modellantaganden och residualanalys

Att *enbart* förlita sig på den skattade partiella autokorrelationen kan dock vara vanskligt, eftersom denna i praktiken inte garanterar att rätt modell väljs. Valet av modell kommer därför även avgöras av utseendet hos de residualer som uppstår då en viss modell används. Huruvida residualerna uppvisar ett slumpmässigt utseende då de plottas mot tiden, sig själv vid ett tidslagg och de skattade värdena (alltså att de är oberoende med konstant varians), och om de verkar vara normalfördelade, är saker som kommer undersökas.

Att residualerna uppvisar ett slumpmässigt utseende är viktigt för att modellen ska kunna anses vara välanpassad. Om modellen fångat upp den väsentliga variationen som finns i datamaterialet bör det inte finnas några ytterligare samband kvar att ta hänsyn till. Skulle det framkomma mönster i residualerna tyder detta på att de skattade värdena i någon utsträckning systematiskt avviker från de observerade värdena, vilket i sin tur talar för att det kan finnas en modell som beskriver det sanna sambandet mellan de olika variablerna bättre. Om residualerna istället tenderar att vara olika stora för olika nivåer på de skattade värdena är detta i så fall ett tecken på att de inte har konstant varians, och strider även det mot de antaganden som görs för modellen.

Normalfördelade residualer är ett annat viktigt modellantagande. Det kan visas att vid en korrekt anpassad modell bör residualerna vara normalfördelade med väntevärdet 0 och en viss varians σ_e^2 (Chan, Cryer, 2008, s. 42). Normalfördelning kontrolleras här med s.k. *Quantile-Quantile-plottar* (Q-Q-plottar), där värdena från den teoretiska *N(0,1)*-fördelningen plottas mot

de observerade residualerna. Ju närmare en rät linje dessa plottade punkter är, desto mer talar för att normalfördelningsantagandet är uppfyllt.

AIC och BIC

Viktigt att framhålla är dock att även modellens komplexitet är en faktor som vägs in i valet av modell. Åstadkoms endast marginella förbättringar hos residualerna med en mer komplex modell, alltså en AR(p)-modell av högre ordning, kan det vara önskvärt att använda en modell med lägre komplexitet. Anledning till detta är att utan restriktion på antalet skattade parametrar riskerar modellen att bli för stor för att kunna tolkas på ett begripligt sätt. För att kvantifiera denna önskan om att minimera komplexitet används därför det som kallas *Akaikes Information Criteria* (AIC) och det som kallas *Bayesian Information Criteria* (BIC). De matematiska formlerna för dessa mått skrivs följande sätt:

$$AIC = -2\log\left(\hat{L}\right) + 2K,$$

och

$$BIC = -2\log\left(\hat{L}\right) + \log(n) * K,$$

där $\hat{L}$ är den *likelihood* som fås av den skattade modellen givet de värden som observerats i stickprovet. K är här antalet skattade parametrar, alltså, riktningskoefficienter $\hat{\phi}_1, \dots, \hat{\phi}_p$, samt variansen $\hat{\sigma}^2$. Slutligen är n antalet observationer i stickprovet. Likelihood:en är alltså det som mäter hur välanpassad en modell är till en given uppsättning observationer medan termerna K respektive $\log(n) * K$ straffar modellens komplexitet.

Dessa båda mått fungerar rent praktiskt så att den modell som har lägst AIC eller- BIC-värde är mest lämplig (lågt värde motsvarar hög likelihood eftersom $\log\left(\hat{L}\right)$ multipliceras med -2). Skillnaden mellan dem är att i de flesta fallen straffar BIC komplexitet hårdare än vad AIC gör eftersom strafftermen i BIC är beroende av antalet observationer n (Sheather, 2009, s.230-233).

Regressionsmodellen

För att undersöka den eventuella kopplingen mellan hur många barn som föds olika år och olika skuldkvoter används regressionsanalys. Den allmänna formeln av en skattad regressionsmodell kan uttryckas på följande sätt (Sheather, 2009, s. 130-131):

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$$

där de olika koefficienterna $\hat{\beta}_j$ anger skattningen för hur mycket variabeln Yförändras då en förklarande variabel x_j förändras med en enhet samtidigt som övriga förklarande variabler hålls konstanta.

Minsta Kvadrat-metoden

För att skatta de olika koefficienterna $\hat{\beta}_j$ kan minsta-kvadratmetoden användas. Denna metod går ut på att minimera summan av de kvadrerade residualerna $e_i^2 = (y_i - \hat{y}_i)^2$ (även kallat medelkvadratfelet). Formeln för denna summa kan skrivas som (Sheather, 2009, s. 131)

$$RSS = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i} - \cdots - \hat{\beta}_p x_{pi})^2.$$

Skattningarna för de olika $\hat{\beta}_j$ fås sedan genom att derivera funktionen *RSS* med avseende på aktuellt $\hat{\beta}_j$ och sätta värdet för derivatan till 0, alltså $\frac{\partial RSS}{\partial \hat{\beta}_j} = 0$, och därefter lösa för aktuellt $\hat{\beta}_j$. För att sedan undersöka vilka koefficienter som är signifikanta, och således vilka förklarande variabler som kan antas ha en effekt på svarsvariabeln, kan ett *t-test* användas (Sheather, 2009, s. 135).

LASSO-metoden

Det finns dock även andra sätt att avgöra vilka förklarande variabler som bör vara med i modellen. En metod är så kallad *LASSO-regression* (Least Absolute Shrinkage and Selection Operator). I regressionsmodeller där flera av de förklarande variablerna är relativt starkt korrelerade med varandra är det vanligt att deras koefficienter får väldigt stor varians (Hastie et. al., 2013, s. 441-449). Detta medför att de prediktioner modellen gör också får hög varians, och därför blir mindre tillförlitliga. LASSO-metoden försöker få bukt med detta problem genom att införa ett bivillkor för regressionskoefficienterna. *RSS* ska fortfarande minimeras, men nu

under bivillkoret att *summan av koefficienternas absolutbelopp* inte överstiger en viss konstant. I praktiken innebär detta att några av koefficienterna, de minst relevanta för regressionen (förhoppningsvis), sätts till noll och att deras respektive variabler således utesluts från modellen (eventuellt blir koefficienterna istället väldigt små i förhållande till övriga koefficienter, vilket då talar för att de relativt små koefficienternas motsvarande variabler är mindre viktiga för regressionen).

Genom att använda LASSO-metoden åstadkommer man på så sätt både skattning av koefficienter och eliminering av problematiska variabler på samma gång. Priset man betalar för att minska variansen på det här sättet är att koefficienterna får en bias, alltså blir icke-väntevärdesriktiga skattningar. Förhoppningen är dock att stickprovsvariansen ($\hat{\sigma}^2 = \frac{1}{N-K} RSS$, *K=antalet skattade parametrar*, (Sheather, 2009. s. 20)) för *nya* data blir mindre än om man använt den vanliga minsta-kvadrat-metoden, alltså att modellens träffsäkerhet för nya data förbättras med hjälp av LASSO-metoden.

Formeln för skattning av koefficienter enligt LASSO-metoden kan skrivas på följande sätt:

$$\hat{\beta}_j^{lasso} = argmin_{\beta_j} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i} - \cdots - \hat{\beta}_p x_{pi})^2 \; givet \; att \sum_{j=1}^{p} |\beta_j| \leq c.$$

Skattningen ovan går även att uttryckas enligt:

$$\hat{\beta}_j^{lasso} = argmin_{\beta_j} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i} - \cdots - \hat{\beta}_p x_{pi})^2 + \lambda \sum_{j=1}^{p} |\hat{\beta}_j|,$$

(Fahrmeir et. al., 2013, s. 208-209) där parametern $\lambda \geq 0$ har ett direkt samband med konstanten c i föregående formel och på motsvarande sätt justerar hur stor restriktionen, hur strängt bivillkoret, för koefficienterna β_j blir. Denna restriktion påverkar sedan i sin tur koefficienternas skattade värden. Är exempelvis $\lambda = 0$, allså om det inte finns någon restriktion

alls, blir koefficienterna samma som i den vanliga skattningen via minsta-kvadrat-metoden. Om man istället låter λ öka från 0 kommer fler och fler koefficienter tendera att bli mindre eller helt sättas till noll.

Korsvalidering för val av λ

Exakt vilket värde λ ska anta, med andra ord hur stor restriktionen på regressionskoeffcienterna ska vara, bestäms via så kallad *k-fold cross validation*, en typ av korsvalidering (Hastie et. al., 2013, s. 419). Datamaterialet delas då upp slumpmässigt in i ett godtyckligt antal lika stora grupper. I vårt fall har åtta grupper använts. Proceduren ser sedan ut som följande:

1. För observationerna i grupp ett, två osv. fram till och med grupp sju, då kallade *träningsgrupper*, anpassas en regressionsmodell med LASSO-metoden där λ sätts till ett litet tal nära noll.
2. Denna modell används för att skapa prediktioner för grupp åtta, då kallad *testgruppen*, varpå en *medelkvadratsumma för residualerna (MSE)* beräknas.
3. För samma λ-värde görs nu proceduren ovan om men där träningsgrupperna och testgruppen skiftar så att alla grupper får vara testgrupp *en gång var*.
4. *Medelvärdet* av de framräknade MSE-värdena beräknas.
5. Steg 1-4 upprepas, men nu med ett nytt, något högre λ-värde.
6. Steg 1-5 upprepas tills λ-värdet är så högt att modeller med endast en koefficient erhålls (alltså där restriktionen gjort att alla andra koefficienter satts till noll).
7. Det λ-värde som genererade *lägst genomsnittlig MSE* (se steg 4) väljs nu för den slutgiltiga regressionsmodellen.

På grund av den slumpmässiga indelningen av datamaterialet i grupper kommer λ-värdet variera om man utför korsvalideringen upprepade gånger. I och med detta kommer även de skattade koefficienterna att variera, vilket i sin tur kommer leda till att mått som R^2_{adj} (se nedan) kommer skilja sig från korsvalidering till korsvalidering.

Till skillnad från den konventionella minsta-kvadrat-metoden så är LASSO-metodens skattningar av riktningskoefficienterna beroende av vilken skala variablerna är i (t.ex. födslar i tiotusental, födslar i hundratusental). Eftersom alla förklarande variabler inte ursprungligen är i samma skala standardiseras de, vilket medför att modellen inte får något intercept (Hastie et. al., s. 439).

Justerade determinationskoefficienten R^2_{adj}

För att kunna jämföra resultatet mellan regressionsmodeller där olika förklarande variabler inkluderats används den justerade determinationskoefficienten, mer känt som R^2_{adj}. Denna beskrivs av formeln nedan (Sheather, 2009, s. 228):

$$R^2_{adj} = 1 - \frac{RSS/(n-p-1)}{SST/(n-1)},$$

där totala kvadratsumman $SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$ där $\bar{y}$ är stickprovsmedelvärdet. R^2_{adj} kan anta värden mellan noll och ett, och ju högre värde, desto högre förklaringskraft har den aktuella modellen. Anledningen till att R^2_{adj} används är för att den straffar modeller för att inkludera fler variabler, till skillnad från $R^2 = 1 - \frac{RSS}{SST}$ som alltid ökar ju flera variabler man använder. Eftersom AIC och BIC utgår från en maximum-likelihood-skattning av modellen används inte dessa här (Fahrmeir et. al., 2013, s. 664).

Tidsserier i regressionsmodeller

I regressionsmodellerna kommer som nämnts tidigare antalet födslar olika år användas som variabler för att förklara skuldkvoterna, men även skuldkvoterna själva vid olika tidslagg, alltså de AR-modeller som beskrivits tidigare i detta avsnitt, kan användas som förklarande variabler. Detta försvårar tolkningen av riktningskoefficienterna avsevärt.

Om AR-modellerna inte används blir tolkningen för en riktningskoefficient: ökningen i y (skuldkvot) då x_i (antalet födslar ett visst antal år tillbaka i tiden) ökar med en enhet givet att alla andra x_j hålls konstanta. Om istället laggade skuldkvoter används som förklarande variabler måste riktningskoefficienterna tolkas som hur mycket y ökar då x_i ökar med en enhet, *för en viss given nivå på föregående y-värde* (och övriga x_j hålls konstanta) (Hyndman, 2010).

Med anledning av denna tolkningssvårighet blir själva värdena på riktningskoefficienterna inte så intressanta då AR-modeller ingår bland de förklarande variablerna. Det som förblir intressant är istället riktningskoefficienternas tecken, alltså om variabeln har en positiv eller negativ inverkar på skuldkvoten, och riktningskoefficienternas storlek, eftersom det ger en fingervisning om hur viktig den tillhörande variabeln är i regressionen (se förklaringen av LASSO-metoden ovan).

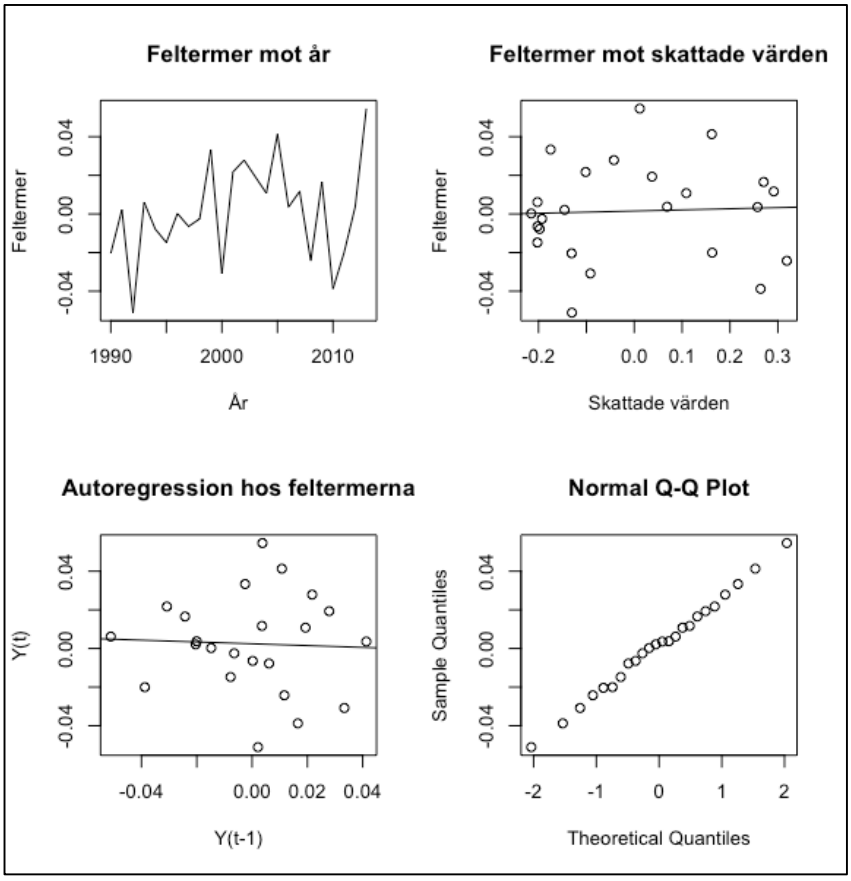
Resultat

Tidsserieanalys

Till en början presenteras och analyseras tidsserierna för skuldkvoterna. Som sagts i metodavsnittet är värdena för alla skuldkvoter logaritmerade. Resultatet från tidsserieanalysen används senare som en del av regressionsanalysen.

Huslån

När variabeln *totala huslån* analyseras tyder det mesta på att en AR(*2*)-modell är lämplig. Residualplottarna från denna modell kan ses i *figur 2*. (Residualplottar för båda AR(*1*) och-AR(*3*)-modellerna kan ses i *figur 5* och *6* i appendix, där även ACF och PACF kan ses i *figur 4*). PACF:en visar visserligen en signifikant partiell autokorrelation även vid tre tidslagg, men som ses i residualplottarna för AR(*3*)-modellen verkar denna sämre anpassad överlag, och från *tabell 2* framgår att både AIC och-BIC-värdena är högre för denna modell.



Figur 2. Residualplottar, totala huslån, AR(2)-modell

AR-modell	*AIC*	*BIC*	*Koefficienter*
AR(1)	-72.1381	-64.60394	0.96
AR(2)	-96.4932	-85.78099	1.74, -0.79
AR(3)	-94.94256	-81.0523	1.58, -0.46, -0.18

Tabell 2. Totala huslån; tabell över AIC, BIC och koefficienter för AR-modellerna

När *individers huslån* analyseras framgår att denna underkategori verkar passa in i samma modell som de totala huslånen, alltså en AR(*2*)-modell. Den partiella autokorrelationen är nämligen signifikant endast vid två tidslagg, och både AIC och BIC är lägst för AR(*2*)-modellen. Även olika residualplottar för de olika modellerna tyder på att AR(*2*)-modellen är bäst anpassad. PACF och residualer återfinns i *figur 7, 8* och *9* i appendix. Tabell över AIC och-BIC-värdena ses i *tabell 3* nedan.

AR-modell	*AIC*	*BIC*	*Koefficienter*
AR(1)	-49.15467	-41.62051	0.89
AR(2)	-76.3569	-65.64469	1.62, -0.88
AR(3)	-75.09401	-61.20374	1.81, -1.24, 0.22

Tabell 3. Individers huslån; tabell över AIC, BIC och koefficienter för AR-modellerna

Övriga lån

I fallet ”övriga lån” är det inte lika entydigt vilken modell som bäst beskriver skuldkvoterna. Enligt PACF:en (*figur 10* i appendix) är bara en av de partiella autokorrelationerna signifikant skild från noll, vilket skulle tala för en AR(*1*)-modell. Studerar man AR(*1*)-modellens residualplottar (*figur 11* i appendix) framkommer att denna är relativt välanpassad; residualerna har en viss autokorrelation, men den verkar främst bero på tre extremvärden som snedvrider sambandet. En viss regelbundenhet syns även då residualerna plottas mot skattade värden, men det är inget tydligt mönster som framträder. Normalfördelningsantagandet verkar någorlunda väl uppfyllt, bara några enstaka residualer till vänster om första kvartilen tycks avvika nämnvärt från en teoretisk normalfördelning.

Residualplottarna för AR(*2*)-modellen i *figur 12* visar förbättring på alla punkter (utom möjligen Q-Q-plotten där ingen tydlig förbättring eller försämring kan urskiljas). Vi kan även

se i *tabell 4* att AR(*2*)-modellen har lägre AIC och BIC-värden. Som nämnts ovan finns dock inget stöd i den partiella autokorrelationen för en AR(*2*)-modell. Därför kommer istället båda modellerna att prövas vid regressionsanalys.

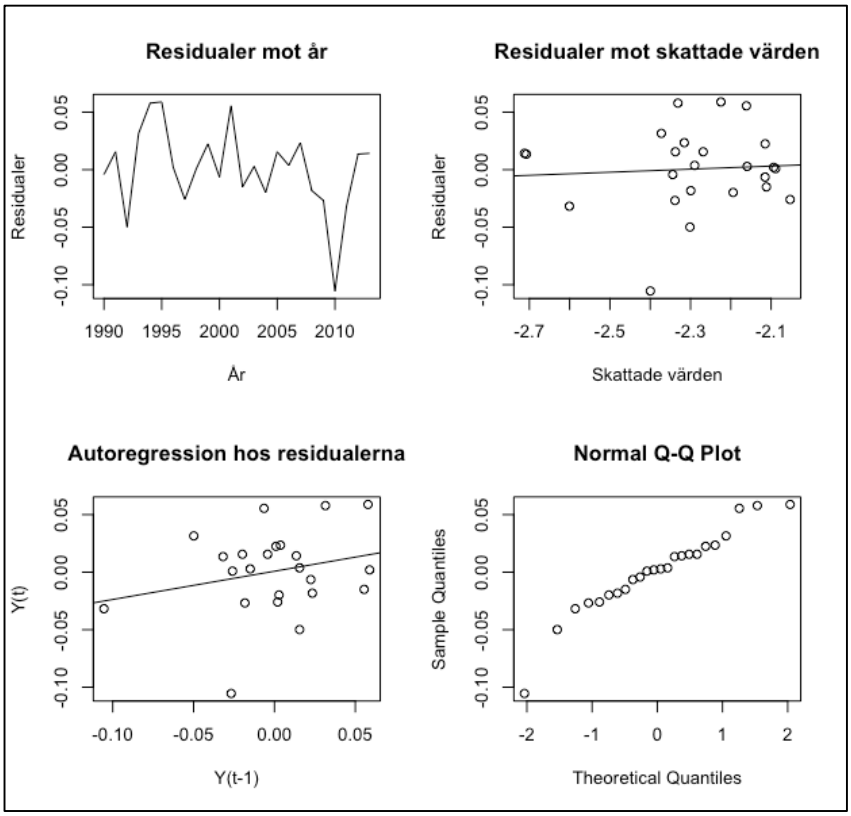
AR-modell	*AIC*	*BIC*	*Koefficienter*
AR(1)	-72.42424	-64.89008	0.95
AR(2)	-81.52867	-70.81645	1.61, -0.69

Tabell 4. Övriga lån; tabell över AIC, BIC och koefficienter för AR-modellerna

Krediter

Slutligen analyseras även skuldkvoten för krediter. ACF och PACF i *figur 13* visar att endast den skattade partiella autokorrelationen vid tidslagg 1 är signifikant skild från noll, vilket återigen tyder på en AR(*1*)-modell. När denna modell skapas syns dock tecken på att den är missanpassad. Av dess residualplottar i *figur 14* framgår att residualerna verkar autokorrelerade. De tycks även följa någon form av svag trend då de plottas mot årtalen.

Då en AR(*2*)-modell istället anpassas till materialet försvinner en betydande del av autokorrelationen som fanns i AR(*1*)-modellen (se *figur 3*). Den svaga trenden som kunde urskönjas då residualerna plottades mot årtalen verkar här helt obefintliga. Ett lite udda mönster syns fortfarande då residualerna plottas mot de skattade värdena, men det är inte mer udda än i föregående modell. Även vad gäller Q-Q-plotten tycks den ha blivit något bättre här. Tillsammans med att AIC och BIC är lägre för AR(*2*)-modellen (*tabell 5*) leder detta till att AR(2)-modellen väljs för regressionsanalys.



Figur 3. Residualplottar, krediter, AR(2)-modell

AR-modell	*AIC*	*BIC*	*Koefficienter*
AR(1)	-61.34219	-53.80803	0.97
AR(2)	-80.08749	-69.37528	1.71, -0.79

Tabell 5. Krediter; tabell över AIC, BIC och koefficienter för AR-modellerna

Därmed har ett antal tidsseriemodeller fastställts. Dessa används nu vidare i regressionsanalysen tillsammans med antalet födslar i en mer omfattande regressionsmodell för de olika skuldkvoterna.

Regressionsanalys

För denna del redovisas tabeller med riktningskoefficienterna för de förklarande variabler som tagits med av de olika LASSO-modellerna (vissa variabler kommer ju som konstaterats att elimineras, alltså deras koefficienter sättas till noll), de erhållna R^2_{adj}-värdena samt de

logaritmerade λ-värdena som valts. Alla dessa tabeller utom den första har lagts i appendix för att underlätta läsningen. I *figurer 18-22* i appendix ses även plottarna för de olika λ-värden som testats och vilken genomsnittlig medelkvadratsumma för residualerna (MSE) de genererat i de olika modellerna, samt hur många koefficienter som tagits med av de olika modellerna.

Huslån

Med hjälp av LASSO-modellen som beskrivits i metod-avsnittet skapas nu en funktion där skuldkvoten för totala huslån (kategorin där även de huslån som bärs av olika institutioner ingår) antas bero på antalet födslar för det aktuella året, alla 30 tidigare år samt föregående års värden två år tillbaka i tiden (eftersom AR(*2*)-modell fastslagits för skuldkvoten, se avsnittet om tidsserieanalyserna ovan). Dessutom anpassas även en modell till den del av huslånen som bärs av hushållen eller individerna själva, kallat ’individers huslån’.

För de totala huslånen utfaller följande bild av modellen (*tabell 6*).

Förklarande variabel	**Riktningskoefficienter**
Födslotal, t	0.0278
Födslotal, t-1	0.0121
Födslotal, t-14	0.0203
Födslotal, t-17	0.0010
Födslotal, t-19	0.0010
Födslotal, t-20	0.0192
Födslotal, t-24	0.0004
Födslotal, t-25	0.0050
Födslotal, t-30	−0.0253
Skuldkvot, t-1	0.1134
R^2_{adj}	0.940
Log(λ)	-6.075

Tabell 6. Riktningskoefficienter, modell för totala huslån

Här syftar alltså exempelvis *Födslotal, t-1* på variabeln födslotal med ett års laggade värden, alltså födslotalet för året som föregick det aktuella året. Alla koefficienter utom den för

födslotal, t-30 är positiva. Residualplot samt Q-Q-plot över residualerna för denna modell finns i appendix i *figur 15*. Inget tydligt mönster går att urskönja i residualplotten som skulle tyda på varierande varians eller beroende mellan resiudalerna, och normalfördelningsantagandet ser ut att vara någorlunda väl uppfyllt.

I *tabell 8* i appendix ges resultatet från analysen av huslån som bärs av individer specifikt. Hälften av koefficienterna är positiva. Residualerna för denna modell syns också i *figur 15*. Krav på konstant varians och oberoende residualer tycks relativt väl uppfyllda. Däremot tyder Q-Q-plotten på större avvikande från normalfördelningen än vad som fanns i modellen för totala huslån.

Övriga lån

För Övriga lån används LASSO-metoden för att skatta två funktioner; en där skuldkvoten antas bero på de laggade födslotalen och föregående års skuldkvot (AR(*1*)-modellen), och en där man använder skuldkvoten både för föregående år och året innan (AR(*2*)-modellen) tillsammans med de laggade födslotalen som förklarande variabler. Det visar sig dock att även då vi använder AR(*2*)-modellen för skuldkvoterna så kommer bara den första termen, alltså skuldkvoten för föregående år, med i LASSO-modellen. Den andra termen, skuldkvoten två år tillbaka, elimineras med andra ord på grund av restriktionen vi satt på riktningskoefficienterna. I övrigt blir riktningskoefficienterna för de båda skattade funktionerna så gott som identiska i fråga om vilka som elimineras och deras värde. R^2_{adj} skiljer sig dessutom med mindre än en tusendels procentenhet. Därför redovisas bara en tabell (*tabell 9* i appendix), den för resultatet då AR(*1*)-modellen använts (R^2_{adj} för AR(*2*)-modellen kommer redovisas senare i avsnittet, se *tabell 7*):

Residualplot tillsammans med Q-Q-plot över residualerna för både modellen med AR(*1*)-termerna och AR(*2*)-termerna kan ses i *figur 16*. Residualerna har en något märklig spridning, (dock inte uppenbart beroende) och dessutom viss skevhet i svansarna, vilket gäller för båda fallen då deras residualer är nästan identiska.

Krediter

Då en LASSO-modell anpassas för skuldkvoten för krediter fås det som ses i *tabell 10* i appendix. Här blir nästan alla riktningskoefficienter negativa. Residualplot samt Q-Q-plot syns i *figur 17*. Residualerna har en aningen udda spridning, som inte nödvändigtvis tyder på vare

sig beroende eller icke-konstant varians, men som ändå bör noteras. Q-Q-plotten tyder på att residualerna verkar avvika något från normalfördelningen i svansarna.

Jämförelse mellan modeller

Som nämnts i inledningen skapas även modeller där låneräntan används som förklarande variabel för skuldkvoterna. Någon vidare utläggning om dessa görs inte då deras syfte endast är att agera referenspunkt för modellerna där födslotal använts som förklarande variabler. Därför redovisas här bara deras R^2_{adj} för att just kunna jämföra de olika modellerna i fråga om förklaringsgrad.

En LASSO-modell har även skapats där endast AR-termerna, alltså skuldkvoterna vid ett och två tidslagg, använts som förklarande variabler (förutom i övriga lån AR(1), markerad med * i *tabell 10* nedan, där endast skuldkvoterna vid ett tidslagg finns med. Där har istället en regressionsmodell enligt minsta-kvadrat-metoden anpassats eftersom det inte är möjligt att använda LASSO-metoden för endast variabel). Detta görs för att kunna jämföra hur mycket bättre, eller sämre, förklaringsgraden blir av att faktiskt lägga till förklarande termer utöver AR-termerna.

Lånetyp	**R^2_{adj}, Låneränte-modell**	**R^2_{adj}, Födslotals-modell**	**R^2_{adj}, ren AR-modell**
Totala huslån	0.976	0.940	0.975
Individers huslån	0.922	0.842	0.923
Övriga lån, AR(1)	0.857	0.888	0.866*
Övriga lån, AR(2)	0.920	0.888	0.895
Krediter	0.962	0.939	0.954

Tabell 7. R^2_{adj} för låneränte-modeller, födslotals-modeller och renodlade AR-modeller

Diskussion

Ovan resultat kan säga oss något om förhållandet mellan födslotal, låneränta och skuldkvoter (i USA åren 1990-2013). Innan vidare diskussion om detta ska bara konstateras att tillräcklig kunskap om de aktuella variablerna saknas för att kunna utreda resultatet på detaljnivå. Varför exempelvis födslotalet vid år *t* eller *t-1* tagits med i nästan alla modellerna går förstås att spekulera i, men om man vetat vad exakt som faktiskt ingick i de olika lånetyper hade analysen förmodligen kunnat göras mer precis. Anledningen till att denna information saknas är att den inte funnits på de hemsidor där data hämtats och det inte funnits tid att göra närmare efterforskningar eftersom denna typ av detaljfakta inte varit prioriterat i uppsatsarbetet.

Med det sagt kan det vara intressant att börja med att titta på just vilka variabler som verkar ha någon betydelse. *Figur 1* i slutet av data-avsnittet visade exempelvis att skuldkvoten för Övriga lån (där studielån ingår) tycktes följa födslotalen 18-20 år tillbaka relativt väl. Detta skulle alltså antyda att antalet födda för ungefär 18 år sedan hade en effekt på skuldkvoten idag, och motiveras, som gjorts i inledningen, med att det är runt den åldern många börjar studera och eventuellt ta studielån. I den regression som gjorts för Övriga lån har också födslotalen vid *t-19* och *t-20* behållits av LASSO-modellen, och de har ett positivt samband med skuldkvoten. Detta är förstås inget slutgiltigt bevis, men talar åtminstone inte emot att det finns ett samband mellan antalet födda ett år och skuldkvoten för Övriga lån 18-20 år senare. Utöver detta går det förstås att diskutera mycket i resultatet för Övriga lån. Varför har exempelvis födslotalen vid *t, t-2* och *t-3* fått negativa koefficienter, eller varför har födslotalet vid *t-13* tagits med och vad det kan betyda? Eftersom det som nämnts ovan dock inte är uppsatsens huvudsakliga syfte att reda ut de olika sambanden på detaljnivå lämnas många av dessa diskussioner därhän.

Regressionsmodellerna för de andra lånetyperna ger ganska varierande resultat. För Totala huslån är nästan alla riktningskoefficienter positiva, medan ungefär hälften av koefficienterna är positiva för Individers huslån. Födslotalen vid *t-14, t-24, t-25* och *t-30* tas med som variabler i båda modellerna. Att *t-24* och *t-25* tas med och får positiva riktningskoefficienter i båda modellerna kan tyckas märkligt; hur många 25-åringar äger sin bostad? Likaså verkar det aningen kontraintuitivt att *t-30* får negativ riktningskoefficient i båda modellerna, med tanke på att det förmodligen är en vanligare ålder att köpa sin bostad vid.

Slutligen har vi lånetypen Krediter, där alla födslotalsvariabler förutom *t-1* får negativa värden. Att just *t-1* är positiv kanske kan förklaras med att man får mer oväntade utgifter första året eller åren efter att man fått barn, varför krediter blir en lösning. Frågan om varför de övriga födslotalsvariablerna får negativa värden låter sig inte förklaras lika intuitivt, och överlåts till läsaren själv att fundera på.

Det ska även nämnas att skuldkvoterna kan påverkas av födslotal som inte tagits med i LASSO-regressionen. Exempelvis skulle antalet födda på 40-talet, och som därmed börjat komma i pensionsåldern runt år 2013, mycket väl kunna påverka skuldkvoterna det året. Om födslotalen under 40-talet dessutom är korrelerade med någon eller några av de födslotalsvariabler som tagits med i regressionsmodellen, vilket är rimligt att anta, kan detta vara ytterligare något som påverkar riktningskoefficienterna i modellen.

Oavsett hur de olika riktningskoefficienterna ska tolkas så finns det en grundläggande svaghet med födslotals-modellerna, nämligen att de inte presterar så bra som sina konkurrenter. För alla lånetyper, med undantag av övriga lån där enbart AR(*1*)-termen inkluderats, så får de modeller där låneränta används som förklarande variabel högre R^2_{adj}. Även om det i de flesta fallen bara rör sig om några enstaka procentenheter talar detta just för att man förlorar på att använda alla dessa födslotal som förklarande variabler. En simplare modell där man endast inkluderat låneränta (och AR-termerna) ger likvärdig eller bättre förklaring, och är samtidigt mindre komplex.

Låneränte-modellerna är dessvärre inte heller helt problemfria. Deras marginella bidrag är relativt litet vid jämförelse med de renodlade AR-modellerna. I de flesta fall skiljer sig dessa två modeller minimalt i förklaringsgrad, och i två fall är till och med de rena AR-modellerna bättre. Detta talar visserligen för att låneräntan kan göra vår modell åtminstone en aning bättre, men det ger också upphov till frågan om inte mer avancerad tidsserieanalys hade kunnat vara ännu bättre. Det är trots allt i AR-termerna som den absolut största delen av förklaringsgraden tycks ligga.

Det kan förstås förekomma situationer där man vill göra prediktioner långt fram i tiden, exempelvis hur skuldsättningen kommer se ut om 20 år. I en sådan situation har man ju varken tillgång skuldkvoten året innan, eller vilken nivå räntan ligger på det året. Kanske skulle födslotalen för nuvarande år (som då i fallet med prediktioner 20 år fram i tiden motsvarar

variabeln *födslotal*, *t-20*) tillsammans med födslotalen för ett visst antal år tillbaka (*t-21, t-22,* etc.) då kunna vara till hjälp? Sådana modeller har testats men inte tagits med i resultatet eftersom det inte ingår i det huvudsakliga syftet, som ju rätt och slätt varit att hitta potentiella samband mellan skuldkvoterna och *alla* tänkbara födslotalsvariabler. Även om dessa modeller inte ger riktigt lika hög förklaringsgrad, eller om de inte blir lika välanpassade i fråga om de antaganden som görs för regressionsanalys, så kan de ändå vara intressanta och förtjäna vidare studier just eftersom de går att använda i fall där de modeller som presenterats i resultatet inte är till någon hjälp. Samtidigt kan man förstås fråga sig om inte låneränta eller skuldkvot med 20 års tidslagg skulle ge likvärdiga eller bättre resultat om de användes som förklarande variabler i en sådan situation.

Viktigt att poängtera är även det lilla antalet observationer. Detta är ett grundläggande problem som gör alla skattningar tämligen osäkra. Exempelvis gör det att valet av λ-värde kommer variera märkbart från korsvalidering till korsvalidering, och därmed påverka sådant som riktningskoefficienternas storlek och därigenom även R^2_{adj}. Detta gör de jämförelser som gjorts mellan modeller mindre pålitliga och alla potentiella slutsatser mer vaga. Ett annat fenomen som påverkar möjligheten att dra några tydliga slutsatser är att många av de LASSO-modeller som anpassats avviker från modellantagandena om framförallt normalfördelning. Eftersom detta antagande bör vara uppfyllt för att modellen egentligen ska kunna anses välanpassad och lämplig att använda för framtida prediktioner är avvikelser från normalfördelningen problematiska.

Slutsats

Som framkommit i diskussionen ovan är det svårt att dra några alltför stora slutsatser av arbetet. Det kan påvisas att det verkar finnas en del betydande samband mellan födslotal och skuldkvot, vilket var en del av syftet med denna uppsats. Detta kan vara värdefull information att ha med sig om man vill prediktera skuldkvoter långt fram i tiden. Å andra sidan, om man jämför födslotalens förklaringskraft med låneräntans så verkar den senare komma vinnande ur striden. Dessutom, oavsett om man använder födslotal eller låneränta i sin modell, så är den enskilt största förklaringsfaktorn till skuldkvoten ett givet år skuldkvoten det föregående året. Kanske skulle en ännu mer välanpassad tidsseriemodell, som en cyklisk tidsseriemodell eller exempelvis någon form av GARCH-modell eller spektralanalys, vara det bästa att satsa på i eventuella framtida studier.

Litteraturlista

Centers for Disease Control and Prevention, (2017), NCHS - Natality Measures for Females by Race and Hispanic Origin: United States. Tillgängligt via: https://catalog.data.gov/dataset/births-birth-rates-and-fertility-rates-by-race-of-mother-united-states-1960-2013 [2018-01-19]

Chan, Kung-Sik, Cryer, J.D., (2008), *Time Series Analysis with applications in R*, 2:nd ed., New York: Springer Science+Business Media

Fahrmeir, L., Kneib, T., Lang, S., Marx, B., (2013), *Regression: Models, Methods and Applications*, Berlin, Heidelberg: Springer-Verlag

Federal Reserve, (2017 a), Consumer Credit – G.19. Tillgängligt via: https://www.federalreserve.gov/releases/g19/current/default.htm [2018-01-19]

Federal Reserve, (2017 b), Mortgage Debt Outstanding. Tillgängligt via: https://www.federalreserve.gov/data/mortoutstand/current.htm [2018-01-19]

Hastie, T., James, G., Tibshirani, R., Witten, D., (2013), *An Introduction to Statistical Learning with applications in R,* New York: Springer Science+Business Media

Hyndman, R.J., (2010), The ARIMAX model muddle, *Rob J Hyndman* [Blogg], 4 okt 2010. Tillgängligt via: https://robjhyndman.com/hyndsight/arimax/ [2018-01-19]

International Monetary Fund, (2016), Lending interest rate. Tillgängligt via: https://data.worldbank.org/indicator/FR.INR.LEND?locations=US&view=chart [2018-01-19]

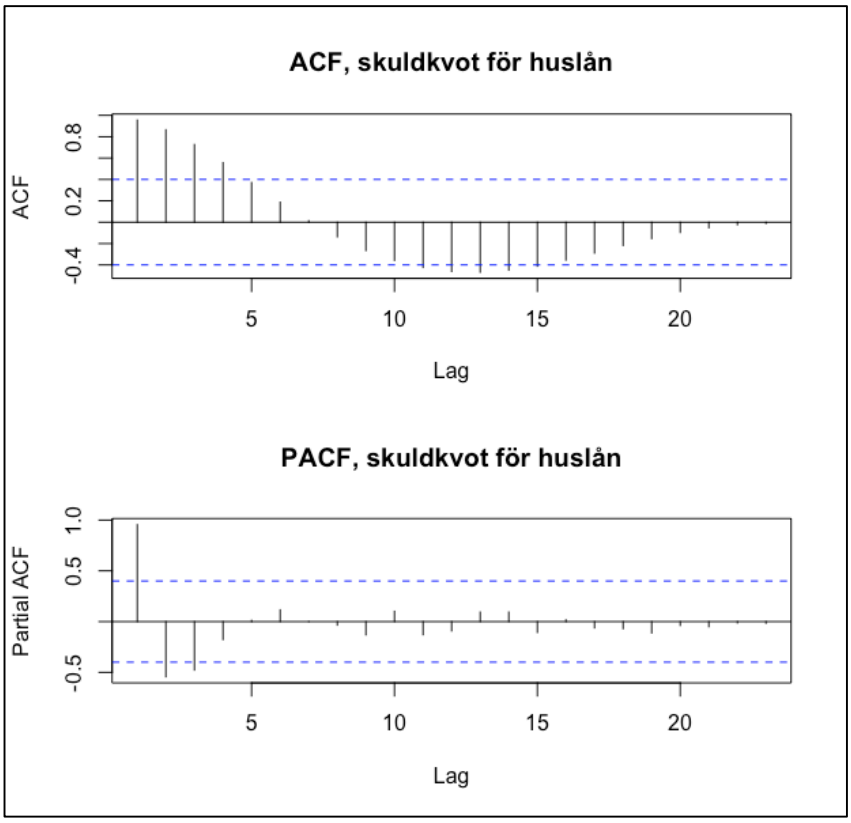
OECD, (2018), Household debt (indicator). doi: 10.1787/f03b6469-en. Tillgängligt via: https://data.oecd.org/hha/household-debt.htm [2018-01-19]

SCB, (2013), Förstagångspappor äldre än förstagångsmammor. Tillgängligt via: https://www.scb.se/sv_/Hitta-statistik/Artiklar/Forstagangspappor-aldre-an-forstagangsmammor/ [2018-01-19]

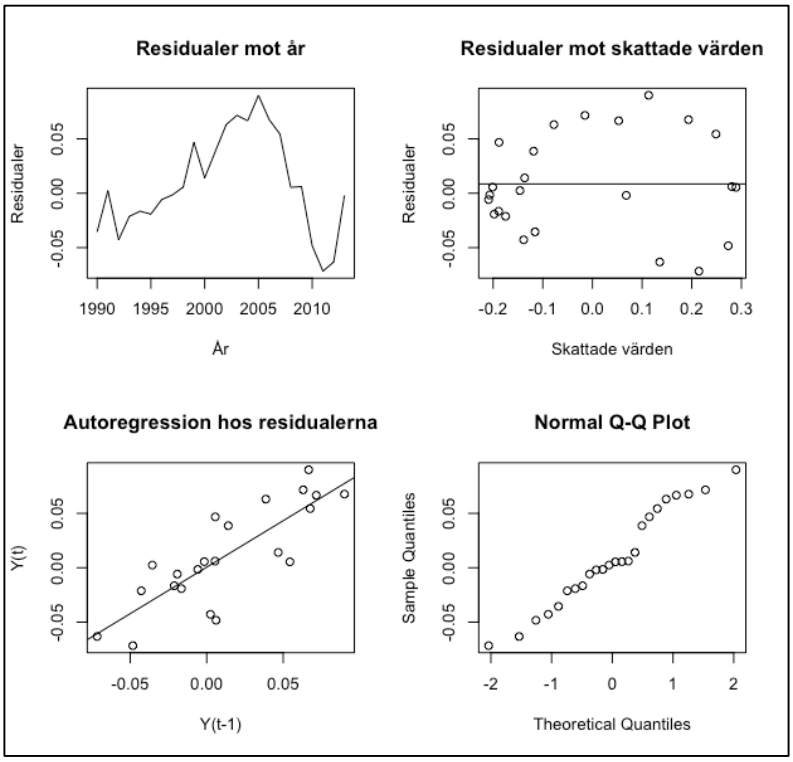
Sheather, S.J., (2009), *A Modern Approach to Regression with R*, New York: Springer Science+Business Media

U.S. Bureau of Economic Analysis (2017), Disposable Personal Income [DPI]. Tillgängligt via: https://fred.stlouisfed.org/series/DPI

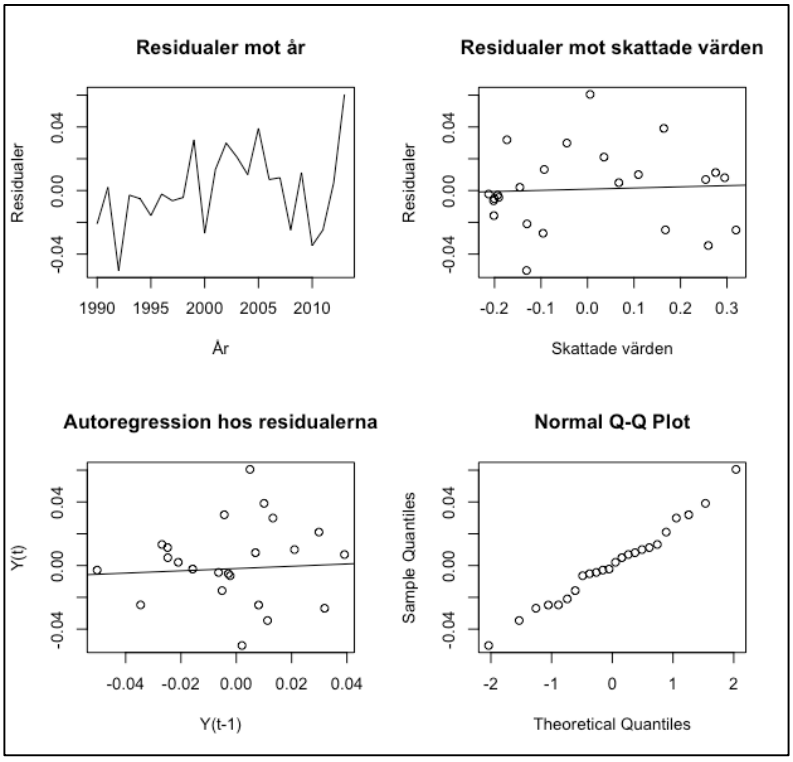
Appendix



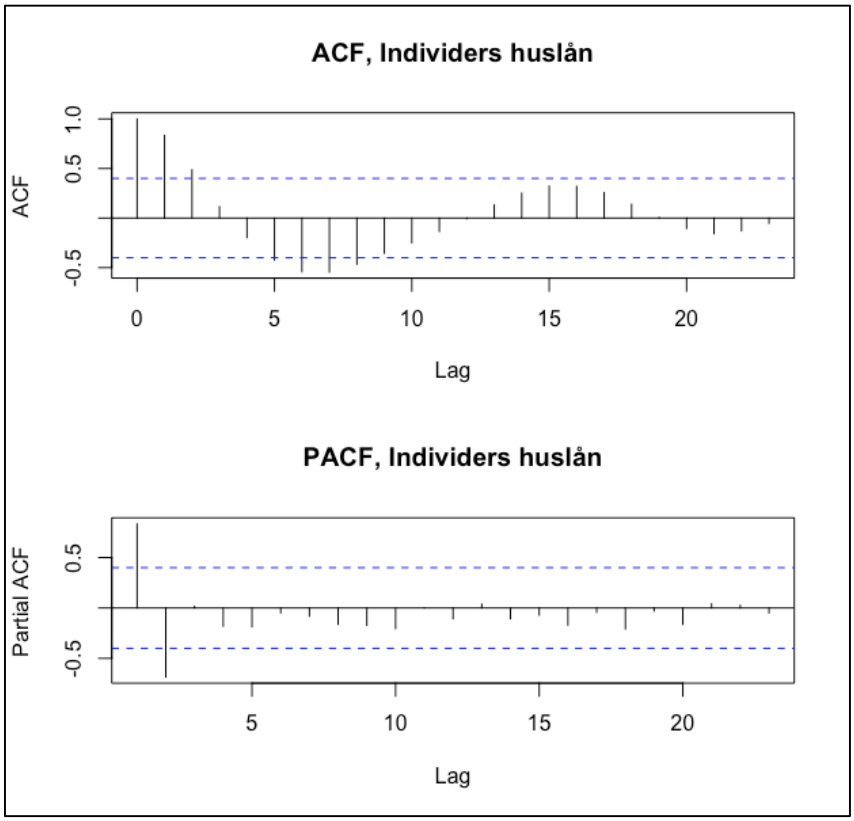
Figur 4. ACF och PACF, ***Totala huslån***



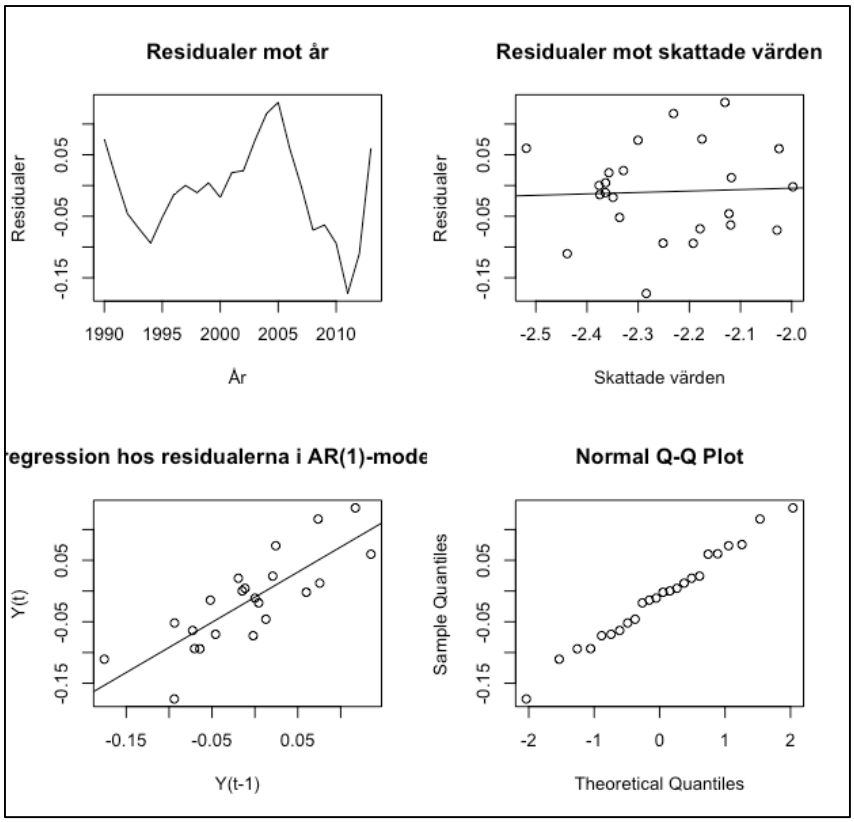
*Figur 5. Residualplottar, **Totala huslån**, AR(1)-modell*



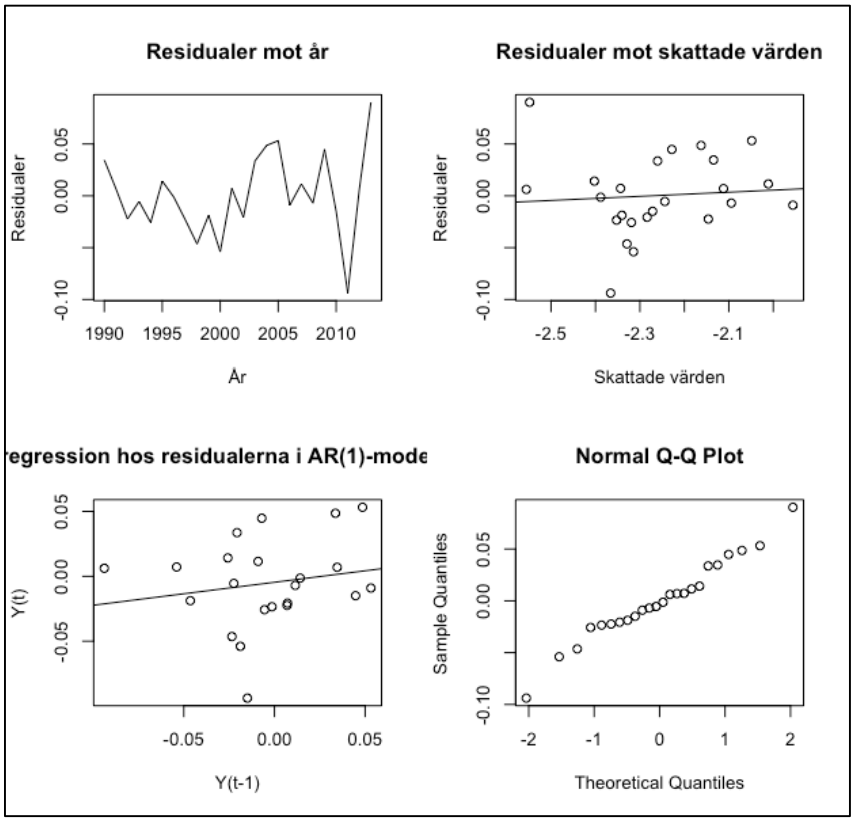
*Figur 6. Residualplottar, **Totala huslån**, AR(3)-modell*



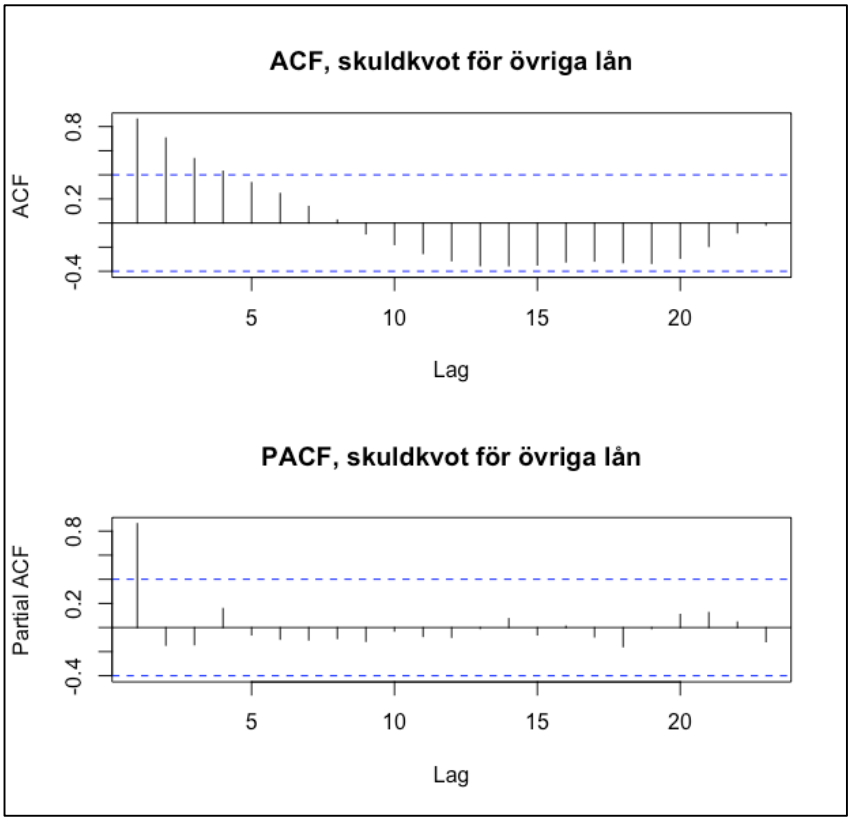
Figur 7. ACF och PACF, ***Individers huslån***



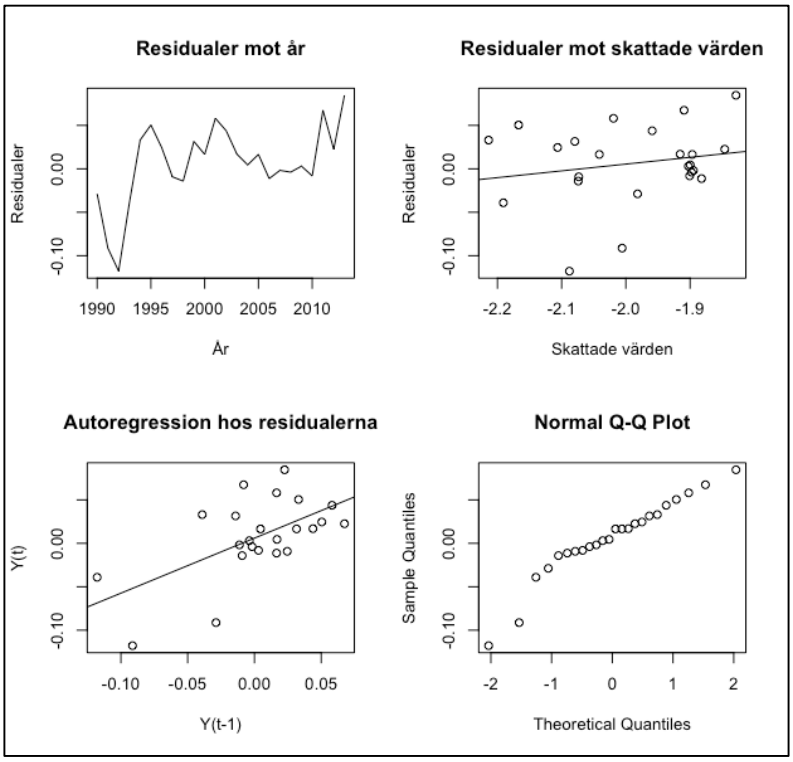
Figur 8. Residualplottar, ***Individers huslån****, AR(1)-modell*



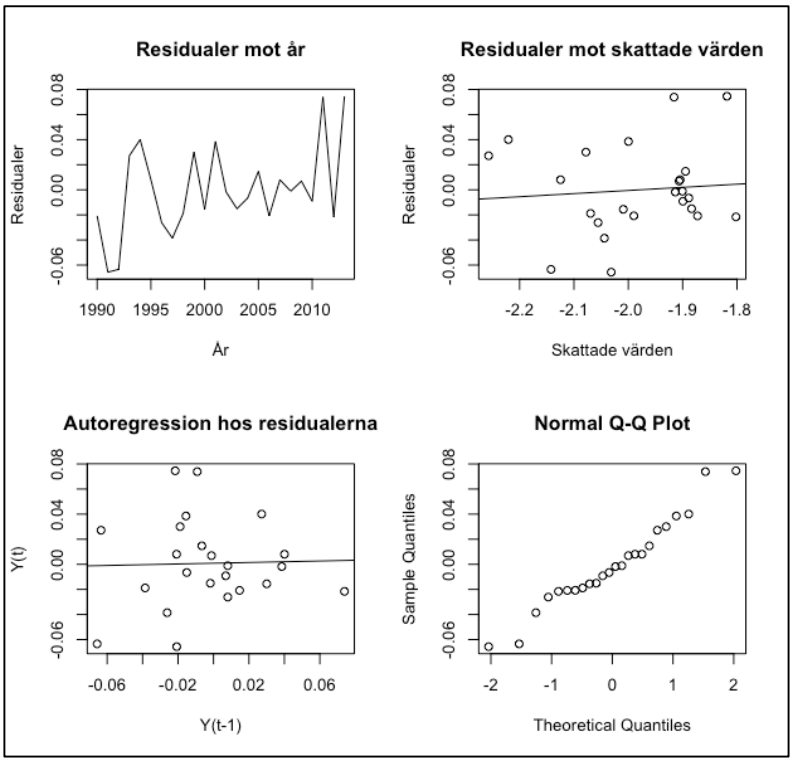
*Figur 9. Residualplottar, **Individers huslån**, AR(2)-modell*



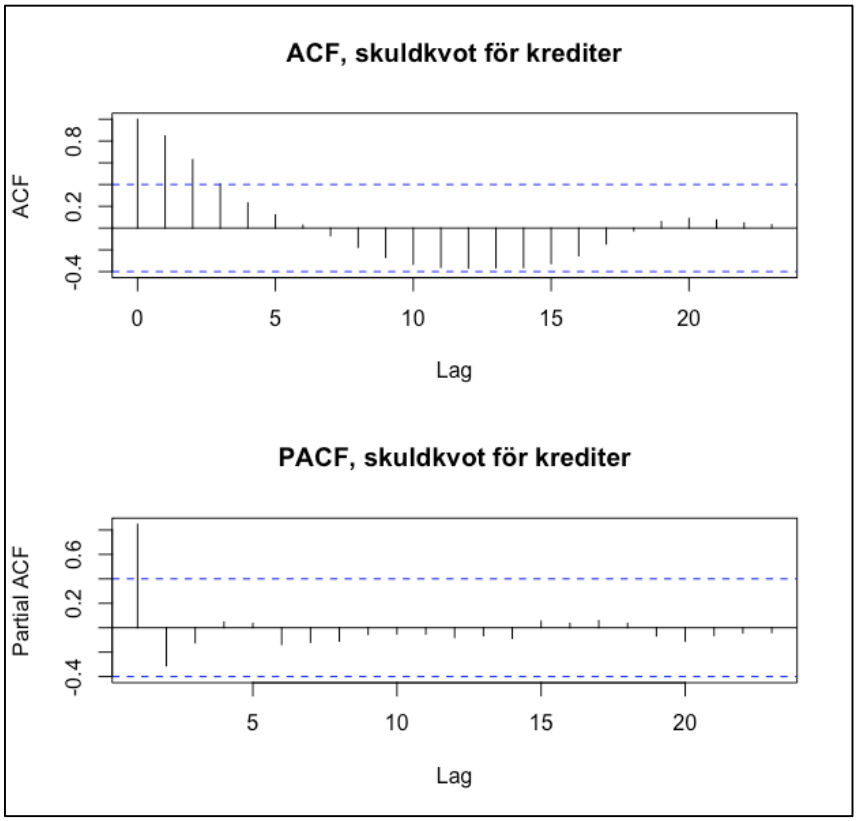
*Figur 10. ACF och PACF, **Övriga lån***



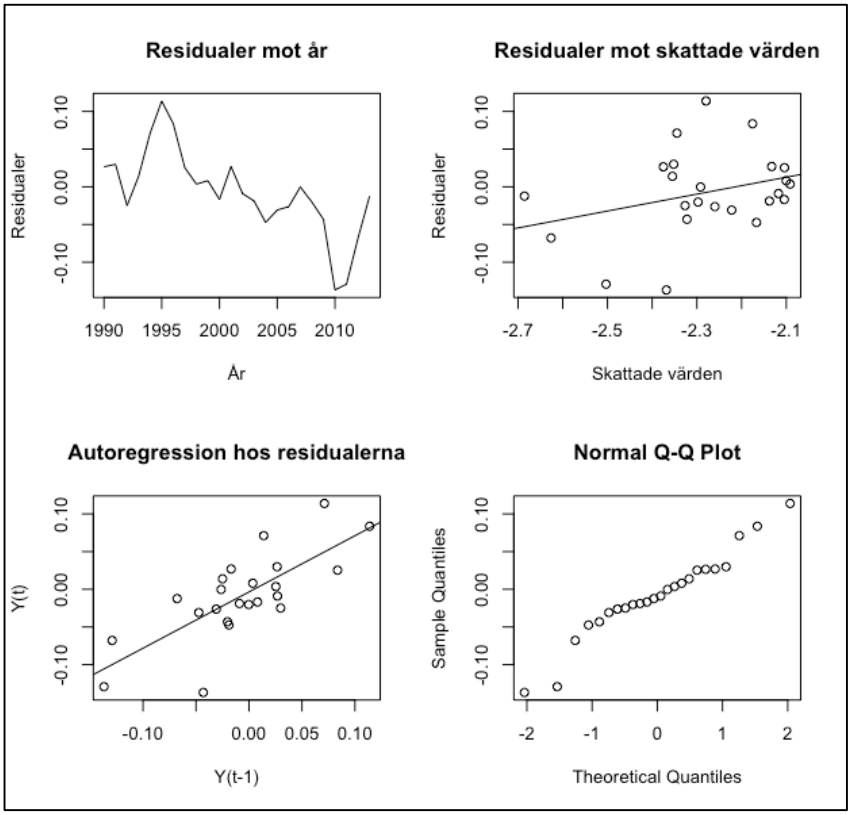
*Figur 11. Residualplottar, **Övriga lån**, AR(1)-modell*



*Figur 12. Residualplottar, **Övriga lån**, AR(2)-modell*



*Figur 13. ACF och PACF för **Krediter***



*Figur 14. Residualplottar, **Krediter**, AR(1)-modell*

Förklarande variabel	**Riktningskoefficienter**
Födslotal, t-4	-0.0620
Födslotal, t-5	-0.0255
Födslotal, t-8	-0.0183
Födslotal, t-10	-0.0262
Födslotal, t-11	-0.0118
Födslotal, t-14	0.0353
Födslotal, t-15	0.0584
Födslotal, t-24	0.0143
Födslotal, t-25	0.0056
Födslotal, t-27	0.0060
Födslotal, t-28	0.0004
Födslotal, t-30	-0.0103
Skuldkvot, t-1	0.1041
Skuldkvot, t-2	-0.0055
R^2_{adj}	0.842
Log(λ)	-6.474

*Tabell 8. Riktningskoefficienter, modell för **Individers huslån**. AR(2)-termerna används i denna modell. Variablerna i vänstra kolumnen är de som tagits med av LASSO-regression, alltså vars riktningskoefficienter inte satts till noll. R^2_{adj} anger förklaringsgrad på ca. 84%.*

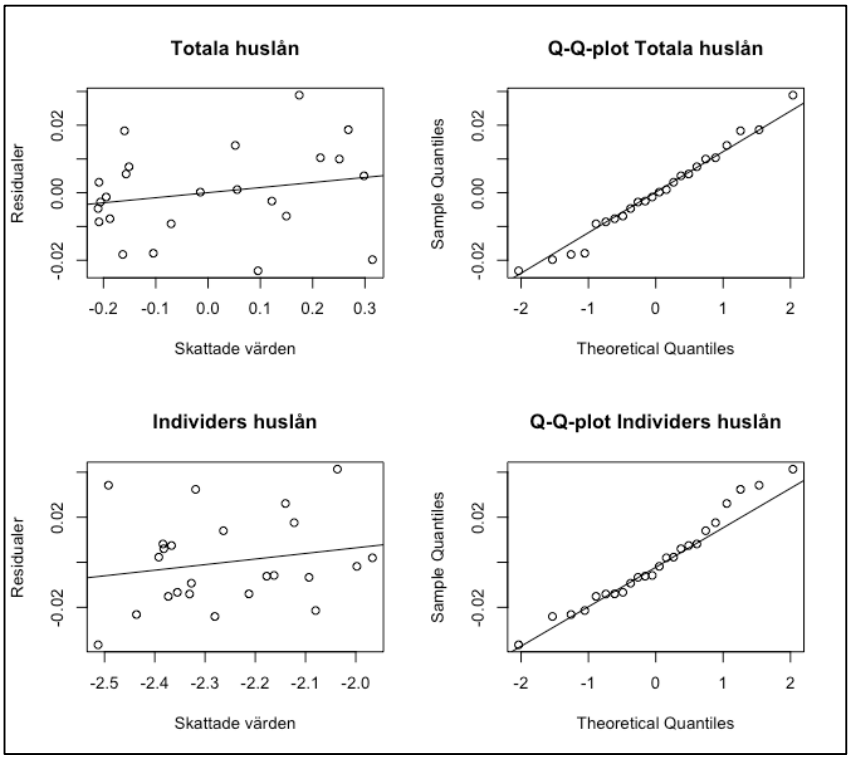
Förklarande variabel	**Riktningskoefficienter**
Födslotal, t	-0.0029
Födslotal, t-2	-0.0111
Födslotal, t-3	-0.0175
Födslotal, t-5	0.0069
Födslotal, t-9	0.0016
Födslotal, t-11	0.0131
Födslotal, t-13	0.0273
Födslotal, t-19	0.0203
Födslotal, t-20	0.0239

Födslotal, t-25	0.0237
Födslotal, t-26	0.0056
Födslotal, t-28	−0.0085
Födslotal, t-29	−0.0034
Skuldkvot, t-1	0.0485
R^2_{adj}	0.888
Log(λ)	-6.646

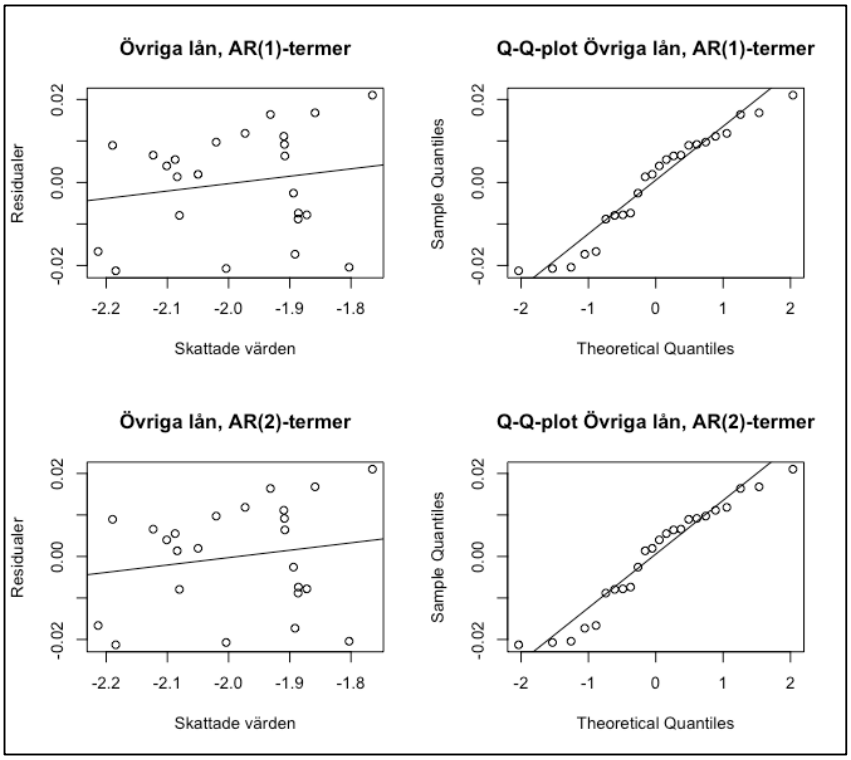
*Tabell 9. Riktningskoefficienter, modell för **Övriga lån**, i fallet då AR(2)-termerna använts. Variablerna i vänstra kolumnen är de som tagits med av LASSO-regression, alltså vars riktningskoefficienter inte satts till noll. R^2_{adj} anger förklaringsgrad på ca. 89%. R^2_{adj} då endast AR(1)-termen använts redovisas i tabell 7, och eftersom skattningarna i den modellen är näst intill identiska med dem som ses här i tabell 9 skapas ingen ytterligare tabell för de skattningarna. .*

Förklarande variabel	**Riktningskoefficienter**
Födslotal, t-1	0.0156
Födslotal, t-4	−0.0235
Födslotal, t-20	−0.0032
Födslotal, t-21	−0.0329
Födslotal, t-22	−0.0348
Födslotal, t-23	−0.0302
Födslotal, t-28	−0.0088
Födslotal, t-29	−0.0248
Födslotal, t-30	−0.0024
Skuldkvot, t-1	0.0781
R^2_{adj}	0.939
Log(λ)	-6.254

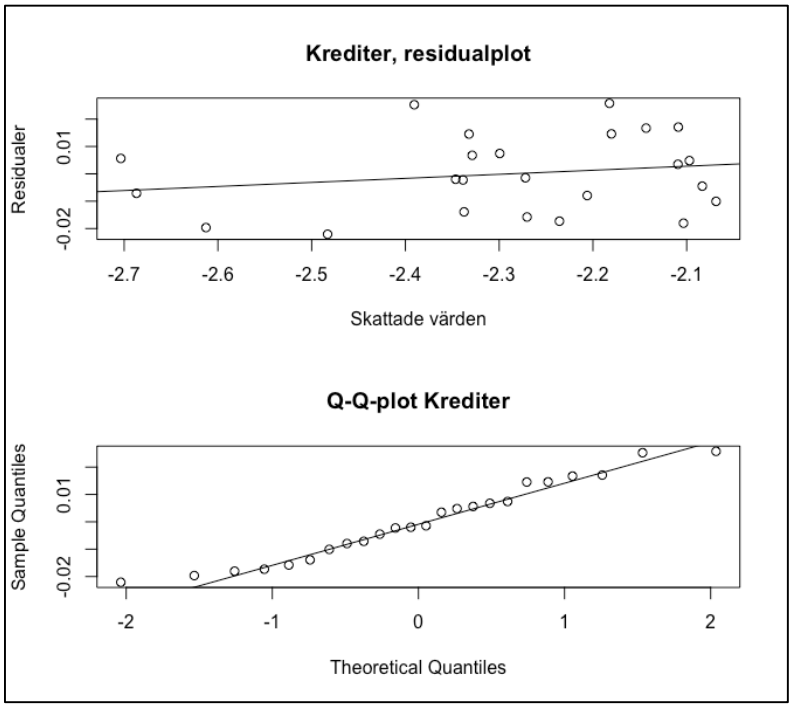
*Tabell 10. Riktningskoefficienter, modell för **Krediter**. AR(2)-termerna används i denna modell. Variablerna i vänstra kolumnen är de som tagits med av LASSO-regression, alltså vars riktningskoefficienter inte satts till noll. R^2_{adj} anger förklaringsgrad på ca. 94%.*



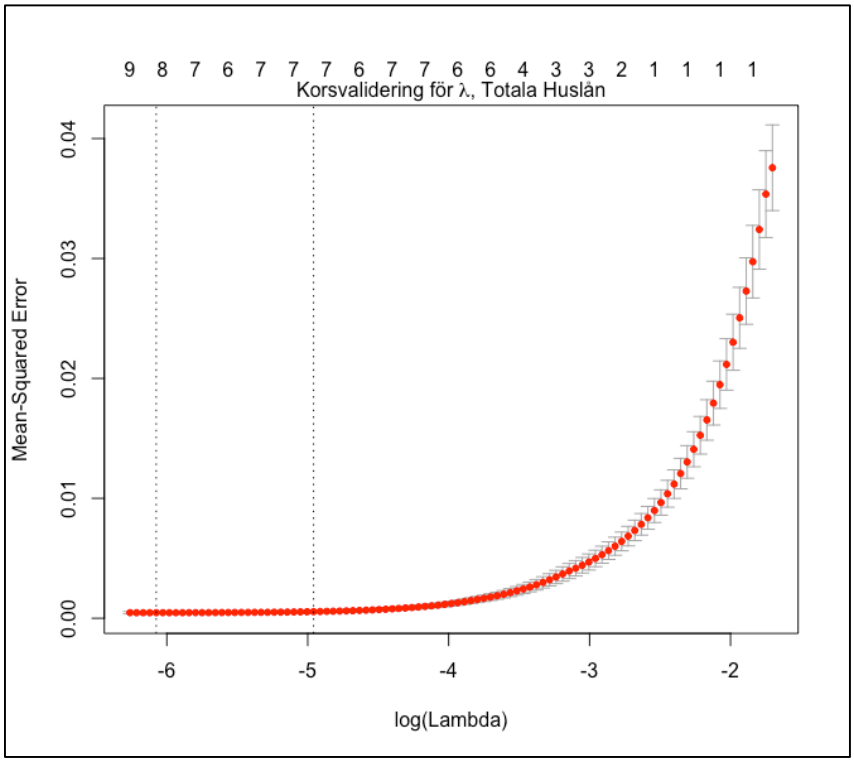
*Figur 15. Residualplottar och Q-Q-plot för residualer i LASSO-modell för **Huslån (totala och individers)***



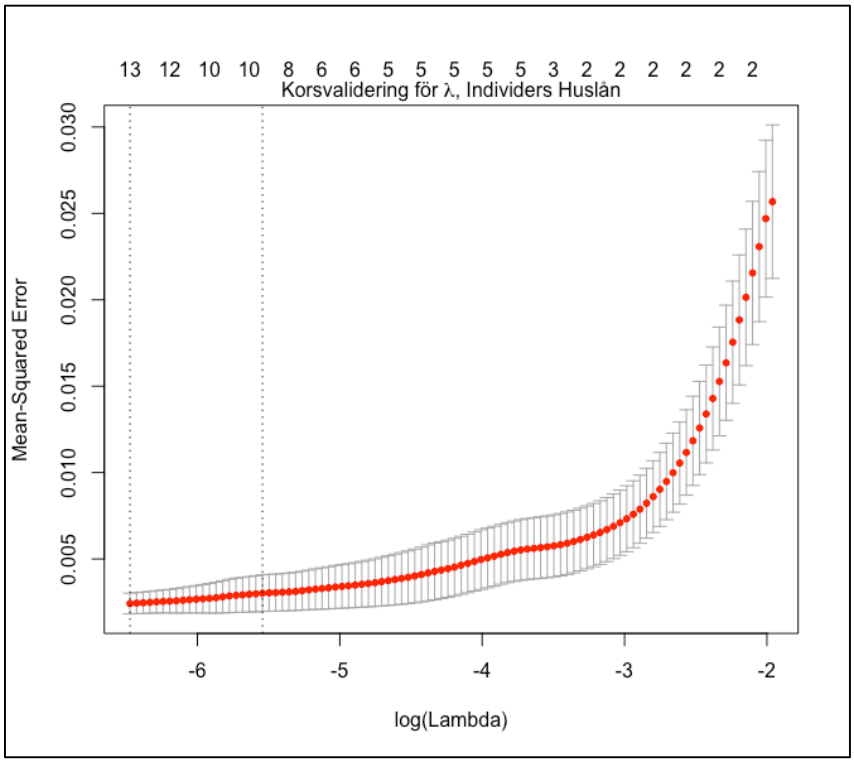
*Figur 16. Residualplott och Q-Q-plot för residualer, LASSO-modell för **Övriga lån***



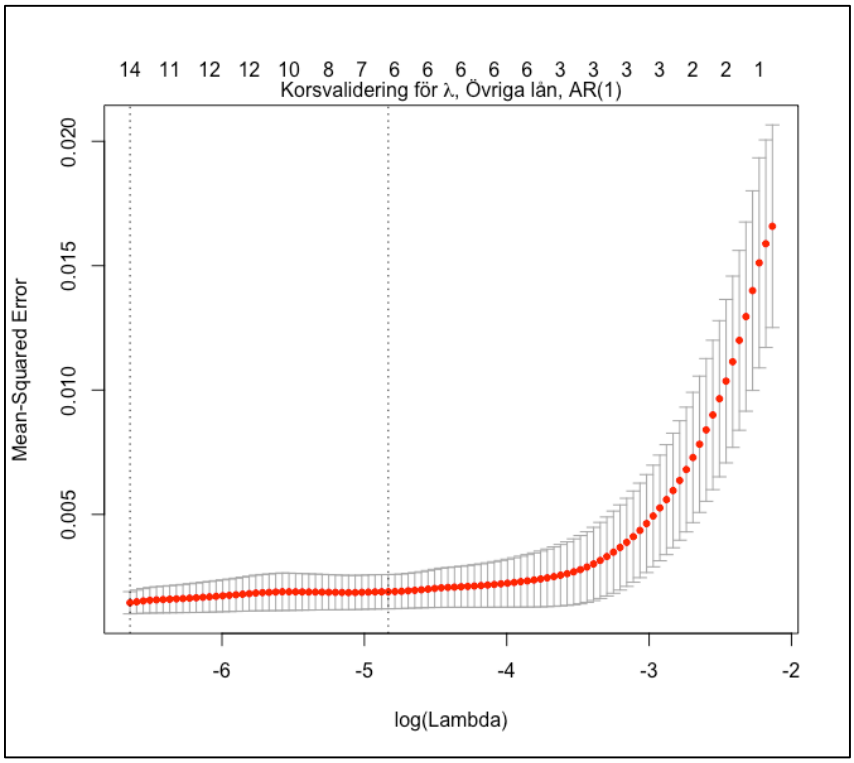
*Figur 17. Residualplot och Q-Q-plot för residualer, LASSO-modell för **Krediter***



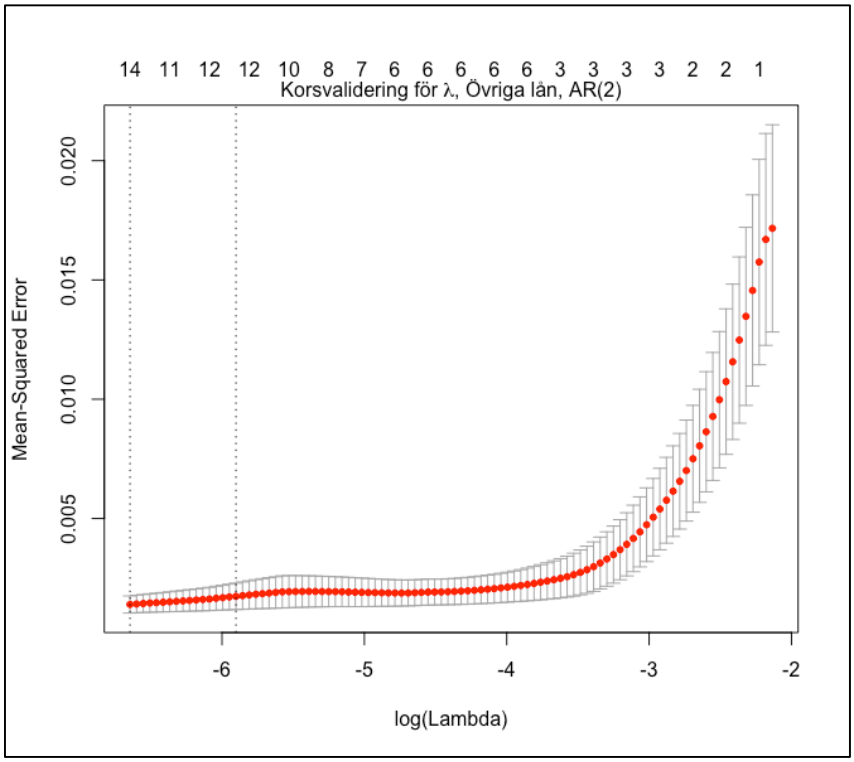
*Figur 18. Plott över korsvalidering för λ, **Totala huslån**. Y-axel: MSE. Övre x-axel: antal variabler som tas med i en viss LASSO-regressionsmodell. Nedra x-axel: Logaritmerat λ för viss LASSO-regressionsmodell (logaritmering görs av läsvänlighetsskäl).*



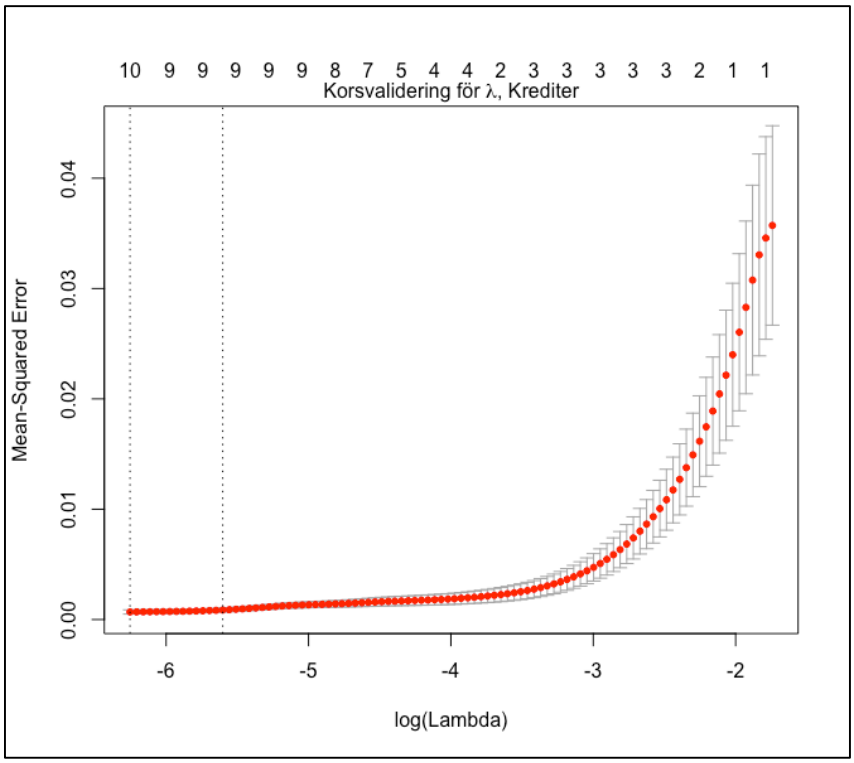
*Figur 19. Plott över korsvalidering för λ, **Individers huslån**. Y-axel: MSE. Övre x-axel: antal variabler som tas med i en viss LASSO-regressionsmodell. Nedra x-axel: Logaritmerat λ för viss LASSO-regressionsmodell (logaritmering görs av läsvänlighetsskäl).*



*Figur 20. Plott över korsvalidering för λ, **Övriga lån, AR(1**). Y-axel: MSE. Övre x-axel: antal variabler som tas med i en viss LASSO-regressionsmodell. Nedra x-axel: Logaritmerat λ för viss LASSO-regressionsmodell (logaritmering görs av läsvänlighetsskäl).*



*Figur 21. Plott över korsvalidering för λ, **Övriga lån, AR(2)**. Y-axel: MSE. Övre x-axel: antal variabler som tas med i en viss LASSO-regressionsmodell. Nedra x-axel: Logaritmerat λ för viss LASSO-regressionsmodell (logaritmering görs av läsvänlighetsskäl).*



*Figur 22. Plott över korsvalidering för λ, **Krediter**. Y-axel: MSE. Övre x-axel: antal variabler som tas med i en viss LASSO-regressionsmodell. Nedra x-axel: Logaritmerat λ för viss LASSO-regressionsmodell (logaritmering görs av läsvänlighetsskäl).*