Stockholms universitet

Regressionsanalys av huspriser i Vaxholm

Rasmus Parkinson

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2015:19
Matematisk statistik
Juni 2015

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholms universitet

Matematisk statistik
Stockholms universitet
Kandidatuppsats **2015:19**
http://www.math.su.se/matstat

Regressionsanalys av huspriser i Vaxholm

Rasmus Parkinson*

Juni 2015

Sammanfattning

Det här arbetet har till avsikt att studera hur olika faktorer påverkar slutpriset vid försäljning av hus i Vaxholm, på öarna Vaxön och Resarö. Målet men den här uppsatsen är att hitta en modell som förklarar så mycket som möjligt av variationen i slutpriset. För att uppnå målet att hitta denna modell så används metoder i regressionsanalys. Grundmodellen bestod av nio stycken förklarande variabler. Flera modeller har testats och undersökts, bland annat med hjälp av den stegvisa variabelselektionen, stepwise regression. Utifrån den så har vi kunnat exkludera flera variabler i modellen. Av de nio förklarande variablerna från början återstår det tre stycken i slutmodellen som användes för det totala datamaterialet. Boarean visar sig vara den variabel med den största inverkan på slutpriset i denna modell. Slutpriset som var responsvariabeln i modellen har ändrats genom en transformation från slutpris till logaritmerat slutpris. Datamaterialet som användes i det här arbetet bestod av 87 hus sålda år 2013 respektive 2014 i Vaxholm på öarna Vaxö och Resarö.

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige.
E-post: rallep93@hotmail.com. Handledare: Maria Deijfen och Jan-Olov Persson .

Sammanfattning

Det här arbetet har till avsikt att studera hur olika faktorer påverkar slutpriset vid försäljning av hus i Vaxholm, på öarna Vaxön och Resarö. Målet men den här uppsatsen är att hitta en modell som förklarar så mycket som möjligt av variationen i slutpriset. För att uppnå målet att hitta denna modell så används metoder i regressionsanalys. Grundmodellen bestod av nio stycken förklarande variabler. Flera modeller har testats och undersökts, bland annat med hjälp av den stegvisa variabelselektionen, stepwise regression. Utifrån den så har vi kunnat exkludera flera variabler i modellen. Av de nio förklarande variablerna från början återstår det tre stycken i slutmodellen som användes för det totala datamaterialet. Boarean visar sig vara den variabel med den största inverkan på slutpriset i denna modell. Slutpriset som var responsvariabeln i modellen har ändrats genom en transformation från slutpris till logaritmerat slutpris.

Datamaterialet som användes i det här arbetet bestod av 87 hus sålda år 2013 respektive 2014 i Vaxholm på öarna Vaxö och Resarö.

Abstract

This work has the intention to investigate how much some variables influence the final price of the sale of the house in Vaxholm, on the islands Vaxön and Resarö. The goal with this essay is to find a model that explains as much as possible of the final price. To achieve the goal of finding this model, methods in regression analysis have been used. The basic model consisted of nine explanatory variables. Several models have been tested and investigated, including using the stepwise variable selection, stepwise regression. Based on this, we have been able to exclude several variables in the model. Of the nine explanatory variables we had in the beginning, it remains three in the final model. The living area turns out to have the biggest impact on the final price of this model. In the final model, the response variable has been modified by a logarithmic transformation.

The study used 87 houses sold between 2013-2014 in Vaxholm on the islands Vaxö and Resarö.

Förord med tack

Det här arbetet är en kandidatexamensuppsats på Stockholms Universitet. Arbetet omfattar 15 högskolepoäng på institutionen matematisk statistik.

Jag skulle först och främst vilja rikta ett stort tack till mina två handledare Maria Deijfen och Jan-Olov Persson vid Stockholms Universitet. För den stora rådgivningen som jag har fått och deras support och stöttande under arbetets gång. Det har betytt väldigt mycket. Jag skulle även vilja passa på att tacka Josefin Andersson Senko, student vid Stockholms Universitet, för hennes uppmuntran och hjälpsamhet.

Innehållsförteckning

1. Introduktion

1.1 Inledning

Ett av många aktuella samtalsämnen idag handlar om huspriser. Man kan se rubriker i olika tidningar så som 'huspriser slår rekord' skriver Aftonbladet [1], 'huspriserna fortsätter uppåt' från DN [2] och 'bostadspriserna i innerstan fördubblade på tio år' i Svenska Dagbladet [3]. Det verkar som priserna för hus ökar med åren. Vid försäljning av hus kan stora ekonomiska vinster göras men även en förlust. Det gäller att sälja och köpa vid rätt tillfälle för att kunna göra en så bra vinst som möjligt. En intressant fråga som man då kan ställa sig är, vilka faktorer är det som har betydelse vid försäljningen av ett hus? Är det husets storlek, husets läge eller när husets såldes som har störst betydelse?

1.2 Mål och syfte

Målet med det här arbetet är att undersöka hur ett visst val av faktorer inverkar på slutpriset av hus i Vaxholm, på öarna Vaxön och Resarö. Syftet med arbetet är att hitta en modell som förklarar slutpriset så bra som möjligt med hjälp av dessa variabler och även undersöka om det skiljer sig mellan försäljning av hus som såldes 2013 respektive 2014. Ytterligare ett syfte med arbetet är såklart att fördjupa mina kunskaper inom regression samt regressionsanalys.

Dessa frågor förväntas bli besvarade i detta arbete.

- Vilka av variablerna visar sig påverka slutpriset vid försäljning av ett hus?
- Skiljer sig sambanden mellan Vaxö och Resarö?
- Är det någon skillnad mellan slutpriserna för åren 2013 och 2014?

2. Metod

Definitionerna av de olika begreppen i detta kapitel hänvisas till Rolf Sundbergs kompendium 'Linjära Statistiska Modeller' där han förklarar dessa begrepp. Till hjälp för att definiera minstakvadratmetoden har även kompendiet 'Notes In Econometrics' skriven av Patrik Andersson och Johanna Tyrcha använts.

2.1 ***Regression***

För att beskriva ett samband mellan en responsvariabel och antingen en eller flera förklarande variabler används regression. Målet med regression är att hitta en funktion som till exempel kan bestå av en eller flera variabler som så bra som möjligt beskriver de observerade data.

2.1.1 Multipel linjär regression

Då vi har en responsvariabel (slutpriset) och flera förklarande variabler som kan tänkas inverka på responsvariabeln, kan man använda multipel linjär regression för att

undersöka om det finns något samband mellan de förklarande variablerna och responsvariabeln.

Detta skrivs på formen

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_n x_{ni} + \varepsilon_i, \quad där\ i = 1, \ldots, k, \quad k \in N.$$

där ε_i är oberoende och antags vara normalfördelad med väntevärdet noll och en konstant varians σ^2, $\varepsilon_i \sim N(0, \sigma^2)$. ε kallas ofta för försöksfelet. Parametern α är interceptet i modellen och de olika β_i i modellen är parametrarna som beskriver vilken typ av inverkan de olika förklarande variablerna x har på responsvariabeln. Dessa kommer skattas utifrån de data vi har.

Det är ofta smidigt att skriva modellen i matrisform, vilket blir

$$\boldsymbol{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \qquad \boldsymbol{\beta} = \begin{pmatrix} \alpha \\ \beta_1 \\ \vdots \\ \beta_n \end{pmatrix}, \qquad \boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix}, \qquad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$$\boldsymbol{Y} = \boldsymbol{\beta X} + \boldsymbol{\varepsilon}.$$

2.1.2 Minstakvadratmetoden

Minstakvadratmetoden (MK-metoden) är en metod som kan användas för att skatta parametrarna i en modell. Den går ut på att minimera det kvadratiska felet, det vill säga minimera avståndet mellan de observerade data och regressionslinjen. Residualerna är avståndet mellan den skattade linjen och varje observerat värde. I MK-metoden kvadreras avstånden för att hitta parameterskattningarna som ger en regressionslinje med minsta möjliga avstånd till observationerna. Residualen är e_i, den verkliga observationen är y_i och det skattade värdet är $\widehat{y_i}$. Avståndet mellan de observerade data och regressionslinjen som skall minimeras är alltså

$$e_i = y_i - \widehat{y_i}.$$

Summan av kvadraterna på residualerna strävas då alltid efter att göras så små som möjligt. Ett mått som mäter detta är RSS (residual sum of squares). Den bäst skattade regressionslinjen är den som ger det minsta värdet på RSS

$$RSS = \sum e_i^2.$$

Med hjälp av MK-metoden får man skattningarna av β som används i modellen

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X^T X})^{-1} \boldsymbol{X^T Y}.$$

2.2 Förklaringsgrad

Förklaringsgraden, R^2 är ett mått på hur bra modellen passar data. Hur stor del som förklaras av den totala variationen i modellen. Förklaringsgraden antar ett värde mellan noll och ett. Ett högt värde på förklaringsgraden kan tolkas som att anpassningen av den linjära modellen är bra medan ett lägre värde kan peka på att det inte är något linjärt samband mellan responsvariabeln och de förklarande variablerna. Ett lågt värde kan även antyda att den slumpmässiga variationen har en stor effekt. Förklaringsgraden definieras som

$$R^2 = \frac{Kvs(regression)}{Kvs(totalt)} = 1 - \frac{Kvs(residual)}{Kvs(totalt)}.$$

De olika kvadratsummorna definieras på följande sätt

$$Kvs(totalt) = \sum (y_i - \overline{y})^2$$

$$Kvs(residual) = \sum (y_i - \hat{y}_i)^2$$

$$Kvs(regression) = \sum (\hat{y}_i - \overline{y})^2.$$

2.3 Hypotestest

I de olika modellerna som kommer undersökas så kommer flera hypotestest att göras. Hypotestesten går ut på att undersöka om de olika parametrarna i modellen är signifikant skild från noll eller inte, det vill säga om motsvarande variabel har en inverkan eller inte.
Då datamaterialet ofta antags vara normalfördelad är skattningen av beta lika med $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X^T X})^{-1}\boldsymbol{X^T Y}$ som har fördelningen $N(\beta, \sigma^2(\boldsymbol{X^T X})^{-1})$.

2.4 Stegvis variabelselektion

När man har en stor mängd data och flera variabler som kan tänkas ha någon slags inverkan på responsvariabeln, kan det ofta vara intressant att se om variablerna är signifikanta eller inte. Om inte, kan de icke-signifikanta variablerna tas bort från modellen och modellen blir enklare med färre parametrar att skatta. Då man även har flera variabler som sinsemellan är korrelerade så kan dock märkliga resultat förekomma, till exempel osäkra skattningar av parametrarna. Exempel på tre stycken stegvisa variabelselektioner är backward elimination, forward selection och stepwise regression. Den metoden som vi kommer att använda oss av i detta arbete är stepwise regression. Den fungerar på så sätt att inga variabler är med från början i modellen, sen allteftersom i varje steg så utvidgas den med en ny variabel i taget. Den kontrollerar i varje steg om samtliga variabler fortfarande är signifikanta efter utökning av variabler. På detta sätt går den igenom samtliga variablerna i modellen, och slutar då det inte finns några fler signifikanta variabler.

2.4.1 Multikollinjäritet

Multikollinjäritet kan inträffa då man har flera förklarandevariabler som har ett linjärt samband sinsemellan. Detta kan bidra till att variablerna kan bli icke-signifikanta i modellen trots att de egentligen har en signifikant effekt på responsvariabeln. En förändring hos en variabel kan alltså bero på förändringen hos en annan variabel. Till exempel om man har två variabler som är korrelerade med varandra så kan det hända att båda visar sig icke-signifikanta medan de var för sig visar sig signifikanta, båda kan inte vara med i modellen samtidigt!

2.4.2 Heteroskedacitet och homoskedacitet

Problem som kan förekomma vid tolkning av resultatet vid regressionsanalys är heteroskedacitet. Ett av antagandena som görs i regressionsanalys är att det ska vara homoskedacitet, dvs. att residualerna är homogena hos variablerna annars råder heteroskedacitet. Då det råder heteroskedacitet kommer standardfelen för variablerna inte att bli korrekta, vilket kan leda till att värdena vid signifikansberäkning av koefficienterna inte blir korrekta. Ett bra sätt för att avgöra om heteroskedacitet råder eller inte är att plotta en scatterplot mot variablerna.

2.5 Logaritmtransformer

Då man har en multiplikativ modell som man vill göra om till en additiv (linjär) modell behöver man göra en transformation. Ett exempel på transformation som kan göras är logaritmering.

Exempel på multiplikativ modell

$$y = \alpha * e^{\beta x} * \varepsilon.$$

Efter logaritmeringen av denna modell fås

$$y' = \log(y) = \alpha' + \beta x + \varepsilon'$$

som är en linjär modell.

Ett annat syfte med att logtransformera är att få residualerna mer jämnt spridda längs regressionslinjen.

3. Material

3.1 Datamaterialet

Datamaterialet som används kommer från www.booli.se. Booli är en söktjänst som uppger sig vara Sveriges största söktjänst när det gäller att hitta hus till salu. På deras hemsida kan man bland annat hitta bostäder på marknaden oberoende av dem säljs av privatpersoner, mäklare med mera. Man kan se information om hur länge huset legat ute på marknaden, slutpriser på husen samt annan information om husen samt en karta över området. [4]

Materialet består av 87 sålda hus i Vaxholm på öarna Vaxö och Resarö under åren 2013 och 2014. Det är 43 respektive 44 hus på de olika öarna, så fördelningen av husen är jämn. Insamlingen av data har skett manuellt på www.booli.se. De 87 husen som är med i detta arbete är de som fanns tillgängliga på deras hemsida med alla förklarande variabler som presenteras nedan. Det fanns ytterligare några hus men där saknades det information om några av de förklarandevariablerna vilket gjorde att de inte kunde användas.

3.2 Variabler

En stor del av alla variabler som användes i det här arbetet kommer från www.booli.se hemsida, enda variabeln som inte gör det är variabeln trä. För att få fram information om den variabeln så åkte jag till de hus som det inte fanns någon bild på www.booli.se för att se om huset var byggt av trä eller inte.

3.2.1 Responsvariabeln

- **Pris** – slutpriset vid försäljning av huset.

Det dyraste huset i data såldes för 12 450 000kr och det billigaste för 2 750 000kr. Medelvärdet av alla huspriser är 5 813 176, 47 kr och medianen är 5 450 000 kr.

3.2.2 Förklarandevariabler

- **Boarea** – husets antal kvadratmeter.

Det huset med störst antal kvadratmeter som boarea låg på 281 kvm och det minsta huset på endast 30 kvm. Medelvärdet av boarean är 155,3 kvm och medianen är 148,5 kvm.

- **Tomt** – tomtens antal kvadratmeter.

Den största tomten i kvadratmeter låg på 5 760 kvm och den minsta tomten på 186 kvm. Medelvärdet över tomtarean ligger på 1 246,27 kvm och medianen är 949 kvm.

- **Byggår** – året då husets byggdes.

Det äldsta huset som var med i data byggdes år 1800 och det nyaste byggdes år 2012. Medelvärdet över då husen är byggda är år 1961,15 och medianen ligger på år 1970.

- **Hav** – avståndet till havet mätt i meter.

Denna variabel mäter avståndet (fågelvägen) mätt i meter till havet från huset. Huset som ligger närmast havet har ett avstånd på 20 meter och det huset med längst avstånd låg på 760 meter. Medelavståndet från husen till vattnet är 266,45 meter och medianen är 229 meter.

- **Rum** – antal rum huset har.

Mäter hur många rum huset har, det hus med mest antal rum hade elva rum och det hus med minst antal rum hade endast två rum. Medelvärdet av rummen är 5,89 rum per hus och medianen är 6 rum.

- **Sålt** – är en dummyvariabel, 1 står för att huset såldes 2014 och 0 för 2013.

Av de totala 87 husen så såldes 53 av dem år 2014 och de resterande 34 husen år 2013.

- **Hus** – är en dummyvariabel, 1 står för radhus/kedjehus och 0 för villa.

De allra flesta husen var en villa, 76 hus av de 87 totalt var en villa och endast 11 var ett radhus/kedjehus.

- **Ö** – är en dummyvariabel, 1 står för Vaxö och 0 för Resarö.

Denna variabel var jämn fördelad, 43 av husen är på Vaxö och 44 av husen är på Resarö.

- **Trä** – är en dummyvariabel, 1 står för om huset inte är av trä och 0 för att det är gjort av trä.

Av de totalt 87 husen så var 62 av dem byggda av trä och de andra 25 av tegel.

4. Genomförande

Regressionen i det här arbetet gjordes i programpaketet SAS, 'Statistical Analysis System'.

Då samtliga variabler ansågs relevanta i analysen så sorterades inga variabler bort innan första regressionen. Variablerna boarea, tomt, byggår, hav, rum, sålt, hus, ö och trä är inte så många och de alla kan tänkas ha en inverkan på slutpriset för ett hus.

Vi kommer använda tre olika uppdelningar av data, dels hela datamaterialet samt de två uppdelade datamaterialen utifrån om huset ligger på Vaxö eller Resarö. Uppdelningen kommer att göras då det kan tänkas att det är olika faktorer som inverkar olika mycket på de två öarna. På Resarö är det generellt sätt större hus med större tomter sett i kvadratmeter än på Vaxö. Vi kommer sedan att undersöka vilka tre modeller som beskriver slutpriset för dessa datamaterial på bästa sätt för att kunna jämföra dem sinsemellan.

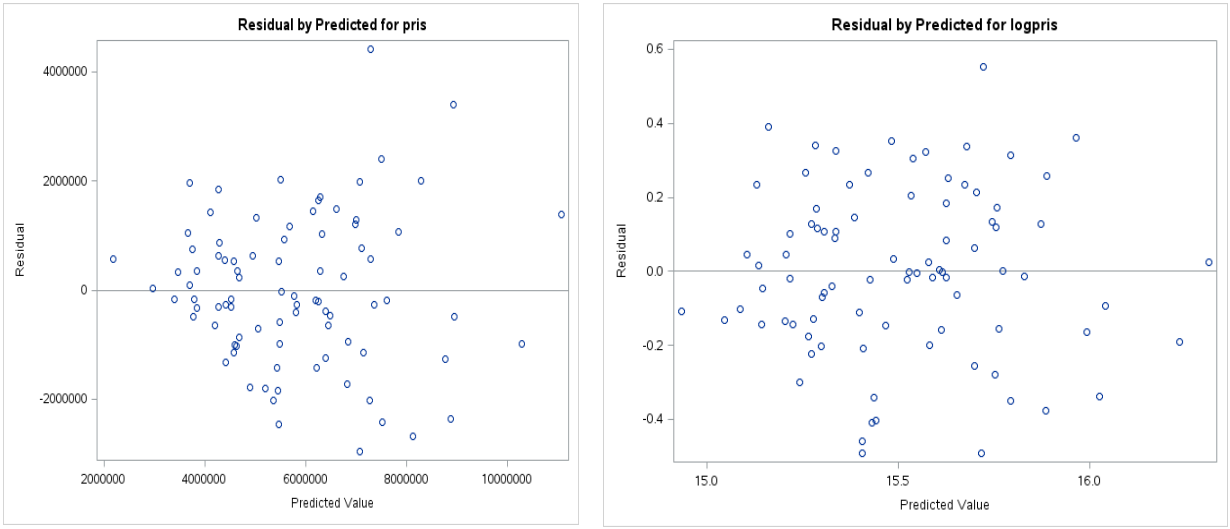
Innan regressionen gjordes var det av intresse att undersöka ifall några av variablerna korrelerade med varandra. Som förklarats tidigare så kan detta medföra att variabler som egentligen inte är signifikanta blir signifikanta eller att signifikanta variabler blir icke-signifikanta om så är fallet. För att motverka detta gjordes en scatterplot över samtliga variabler. Scatterplotten Figur 12 presenteras i appendix och man kan där se att det verkar finnas ett linjärt samband mellan variablerna boarea och rum. Det är alltså inte bra att ha med båda dessa variabler i modellen. Då boarea med största sannolikhet kommer att ha en väldigt stor inverkan på slutpriset vill vi göra något åt variabeln rum. För att minimera korrelationen mellan variablerna så ersattes variabeln rum med variabeln borum. Borum är definierat som boarea dividerat på antal rum i huset.

4.1 Totala datamaterialet

Första undersökningen av datamaterialet var då hela materialet användes. Första modellen som testades var då den innehöll alla variabler med rum ersatt av borum. Figur 13 och 14 visar olika residualplottar som presenteras i appendix. Normalfördelningsplotten av modellen var varken bra eller dålig. I de enskilda residualplottarna för variablerna kan man se att vissa enstaka punkter sticker ut,

tydligast visas för byggår då en punkt ligger vid 1800-talet då de andra ligger vid cirka 1900-talet och senare men även för variabeln boarea finns ett hus med mindre kvadratmeter än de andra. Ett av antagandena som gjorts är att det inte ska förekomma heteroskedacitet hos residualerna för variablerna. För antagandet ska vara uppfyllt fortsätter strävan efter att hitta en bättre modell.

Efter att ha studerat de olika residualplottarna samt de olika plottarna för de predikterade värdena mot priset så valdes att logaritmera responsvariabeln för att få residualerna mer jämnt spridda. I de två figurerna nedan presenteras hur de predikterade värdena såg ute före och efter logaritmeringen.



Figur 1 och 2. Den vänstra figuren visar residualerna mot de predikterade värdena innan responsvariabeln var logaritmerad och den högra figuren visar då responsvariabeln är logaritmerad.

Som syns ovan i figurerna så visar den högra figuren en mer homogen blandning av punkterna.

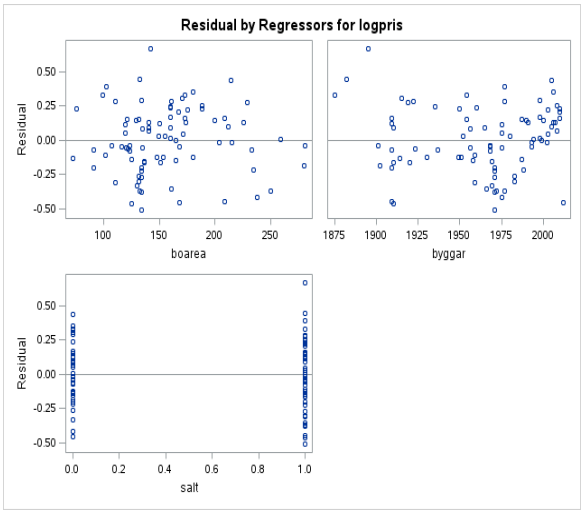
Än så länge har vi studerat modellen med samtliga variabler. Nästa steg är att undersöka om det är möjligt att finna en enklare modell, det vill säga om det går att exkludera några förklarande variabler från modellen utan att den försämras. Med att modellen försämras menas att modellen inte skulle förklara lika mycket av priset samt att residualerna inte försämras.

Till hjälp för att selektera bort variabler i modellen användes den stegvisa variabelselektionen stepwise regression. Efter att ha selekterat bort variabler som visat sig var icke-signifikanta på 5 % nivån återstod det tre stycken variabler i modellen, sex stycken variabler selekterades bort. Variablerna visade sig vara boarea, byggår och sålt. Modellen ser nu ut som följande

$$\log(pris) = \alpha + \beta_1 * boarea + \beta_2 * byggår + \beta_3 * sålt + \varepsilon.$$

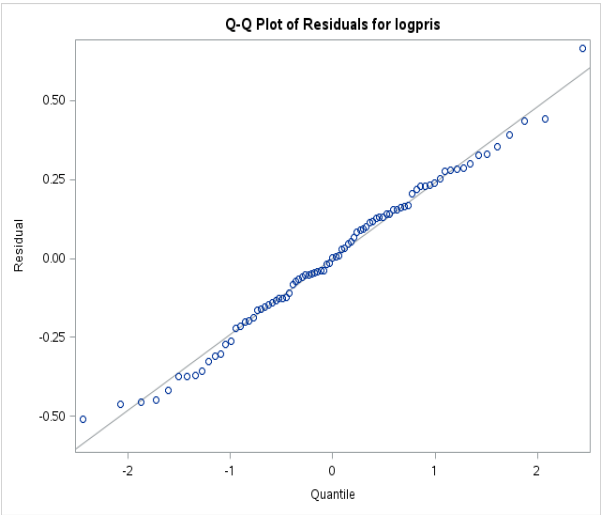
I normalfördelningsplotten till denna modell visar sig punkterna vara slutna kring den raka linjen i figuren. I Figur 15 i appendix presenteras den. Studerar man residualplottarna för de enskilda variablerna som presenteras i Figur 16 i appendix ser man att det fortfarande är vissa punkter som sticker ut från de övriga. Efter att ha tagit

bort dessa två punkter, huset som byggdes på 1800-talet och huset med lilla boarean gavs dessa residualer för de ensklida variablerna i modellen i Figur 3.



Figur 3. Residualplottar för de enskilda variablerna i modellen.

Jämförs denna figur med den då punkterna inte var borttagna ser man att dessa punkter är mer samlade. Punkterna anses vara jämnt fördelade, antagandet om homoskedacitet är uppfyllt. Normalfördelningsplotten för denna modell med de borttagna punkterna presenteras i Figur 4 nedan.



Figur 4. Normalfördelningsplot.

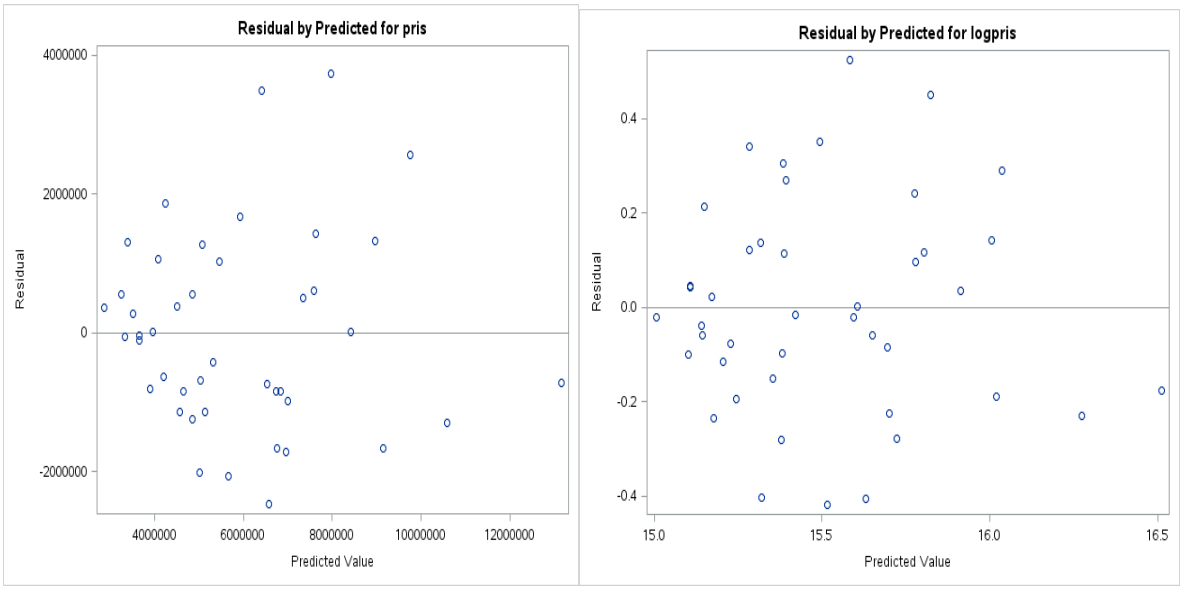
De tre parametrarnas skattning samt standardfel presenteras i Tabell 1 nedan.

Tabell 1. Presenterar de nya skattningarna samt standardfelen då två observationer är borttagna

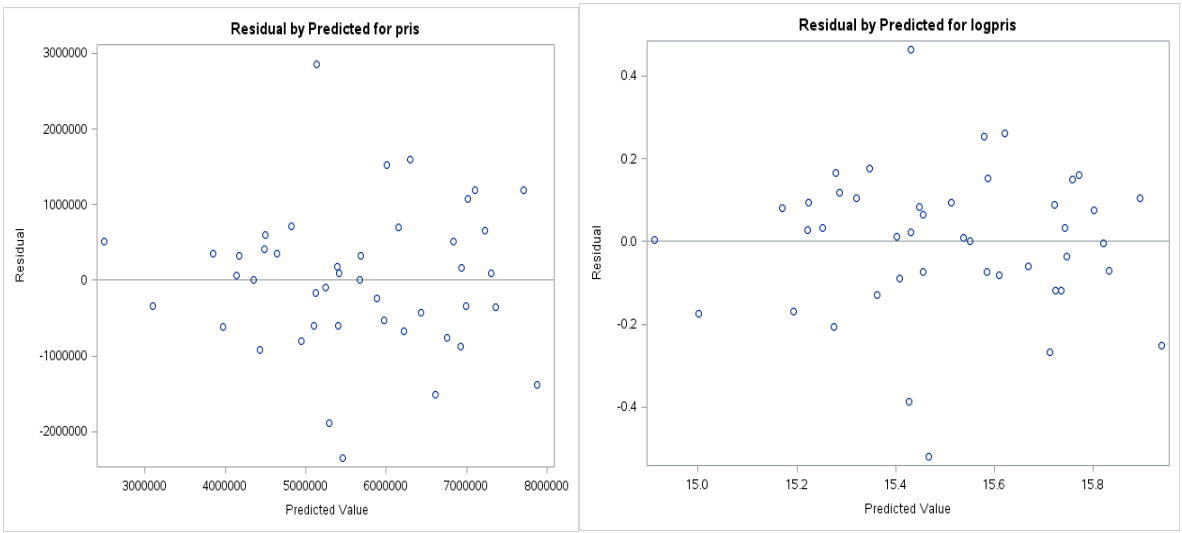
Variabel	Parameter-skattning	Standard-fel	p-värde
Intercept	18.21101	1.48227	
boarea	0.00558	0.00060820	<.0001
byggår	-0.00186	0.00075799	0.0161
sålt	0.13517	0.05474	0.0156

4.2 Uppdelning av datamaterialet

Därefter delades datamaterial upp utifrån om huset är beläget på Vaxö eller Resarö. I båda de fallen studerades även scatterplotten i Figur 12 för att se om några variabler korrelerade med varandra. Då samma variabler ingick i dess grundmodell fanns ett linjärt samband mellan variablerna boarea och rum, rum ersattes likaså med borum som definierats tidigare. Med samma motivering som tidigare logaritmerades även responsvariabeln för att få residualerna mer jämnt spridda.



Figur 5 och 6. Den vänstra figuren visar residualerna mot de predikterade värdena innan responsvariabeln var logaritmerad och den högra figuren visar då responsvariabeln är logaritmerad, datamaterialet med husen på Vaxö.



Figur 7 och 8. Den vänstra figuren visar residualerna mot de predikterade värdena innan responsvariabeln var logaritmerad och den högra figuren visar då responsvariabeln är logaritmerad, datamaterialet med husen på Resarö.

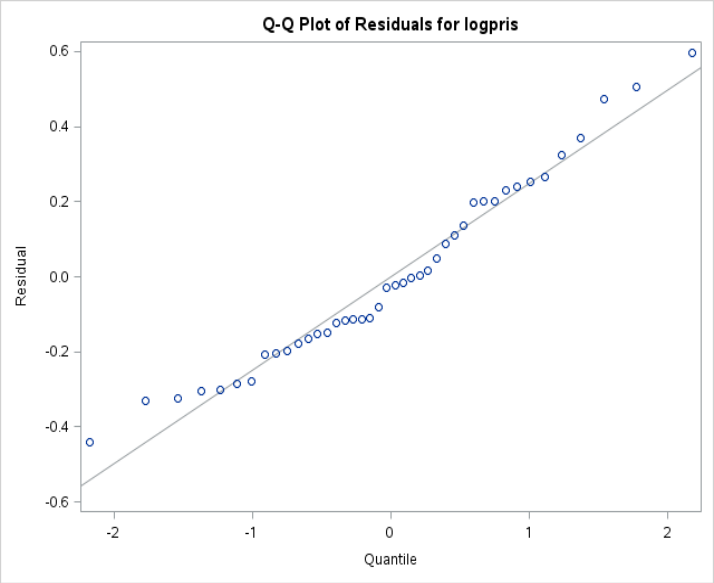
De fyra olika figurerna ovan illustrerar alla residualerna mot de predikterade värdena i modellerna. Det går inte lika tydligt att se en förbättring här som i det totala datamaterialet men studerar man noggrant så ser man att lägre värden för de

predikterade värdena ger ett smalare intervall för residualernas värde innan logaritmeringen. Det vill säga att ju större värde för det predikterade värdet ger en större variation i residualernas värde innan logartimeringen.

På samma sätt som gjordes i det totala datamaterialet undersöktes nu om det gick att exkludera några variabler i modellen. Till hjälp för att selektera variabler användes bland annat stepwise regression i SAS. För datamaterialen med husen på Vaxö återstod två stycken variabler signifikanta på 5 % nivån. De variablerna var boarea och byggår. Variabeln sålt visade sig inte vara signifikant på 5 % nivån men då man på senaste tiden har sett en tydlig prisökning av hus så beslöts det att ha med den variabeln i modellen ändå. Modellen blev följande

$$\log(pris) = \alpha + \beta_1 * boarea + \beta_2 * byggår + \beta_3 * sålt + \varepsilon.$$

För att undersöka huruvida denna modell uppfyller antagandena så studerades modellens residualplottar. Normalfördelningsplotten till denna modell visas i Figur 9 nedan. Residualplottarna för de enskilda variablerna hänvisas till appendix i Figur 17.



Figur 9. Normalfördelningsplot

Skattningarna för parametrarna i modellen och dess standardfel presenteras i tabellen nedan.

Tabell 2. Tabell över skattningarna samt standardfelen.

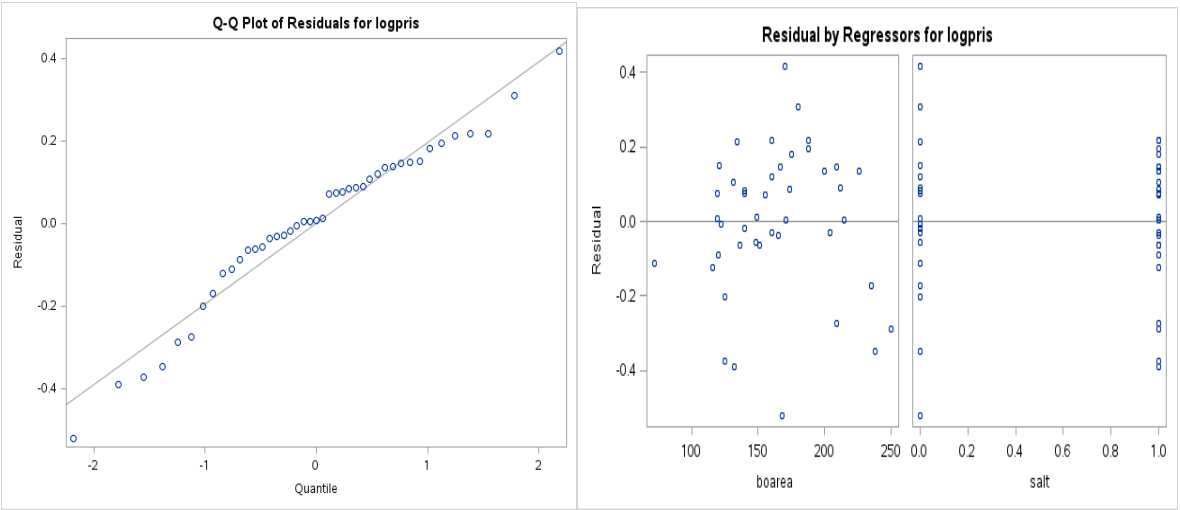
Variabel	Parameter-skattning	Standard-fel	P-värde
Intercept	24.10842	2.58604	
boarea	0.00549	0.00085719	<.0001
byggår	-0.00487	0.00131	0.0006
sålt	0.11649	0.08568	0.1820

För datamaterialen på Resarö återstod endast en variabel, boarea, efter att de icke-signifikanta variablerna hade selekterats bort. Efter att ha tagit bort en outlier så blev

även variabeln sålt signifikant på 5 % nivån. Observationen som togs bort var ett hus som endast hade 30 kvadratmeter i boarea på Resarö. Modellen ser ut som följande

$$logpris = \alpha + \beta_1 * logboarea + \beta_2 * sålt + \varepsilon.$$

Nedan presenteras normalfördelningsplotten och residualerna för de enskilda variablerna i denna modell.



Figur 10 och 11. Normalfördelningsplot presenteras i den vänstra figuren och residualerna över de enskilda variablerna i modellen presenteras i den högra figuren.

De olika parametrarnas skattningar i modellen samt deras standardfel visas i Tabell 3 nedan.

Tabell 3. Tabell med skattningarna samt standardfelen.

Variabel	Parameter-skattning	Standard-fel	P-värde
Intercept	14.69262	0.13344	
boarea	0.00462	0.00079418	<.0001
sålt	0.12744	0.06261	0.0485

5. Resultat

De tre modellerna som presenterades i avsnittet Genomförande är inte likadana. Resultatet blev alltså inte detsamma, olika variabler blev signifikanta för de olika datamaterialen. Den slutgiltiga modellen för de båda öarna tillsammans blev

$$\log(pris) = 18.21101 + 0{,}00558 * boarea - 0{,}00186 * byggår + 0{,}13517 * sålt + \varepsilon.$$

Det blev tre variabler som visade sig vara signifikanta på 5 % nivån. Modellen för enbart husen på Vaxön blev

$$\log(pris) = 24.10842 + 0.00549 * \ boarea - 0{,}00487 * byggår + 0.11649 * sålt + \varepsilon.$$

Modellen innehåller tre variabler. Den tredje modellen för husen på Resarö blev

$$\log(pris) = 14.69262 + 0.00462 * boarea + 0.12744 * sålt + \varepsilon.$$

De tre modellerna ovan är alla logaritmerade. Nu när parametrarna är skattade kan vi gå tillbaka till de multiplikativa modellerna från de här additiva modellerna. För att få tillbaka pris som responsvariabel måsta inversen av log tas på båda sidor, det vill säga e. Modellen för datamaterialet tillsammans ser ut som följande

$$pris = e^{18.21101+0.00558*boarea-0.00186*byggår+0.13517*sålt+\varepsilon}.$$

Modellen för datamaterialet med husen belägna på Vaxö

$$pris = e^{24.10842+0.00549*boarea-0.00522*byggår+0.11649*sålt+\varepsilon}.$$

Modellen för husen på Resarö ser ut som följande

$$pris = e^{14.69262+0.00462*boarea+0.12744*sålt+\varepsilon}.$$

Då de tre modellerna ovan har en multiplikativ inverkan mellan variablerna så kommer feltermerna ε att vara proportionell mot variablerna i modellen. Det vill säga att ju större värden modellen har desto större värden på felet kan det då bli. Men studerar man Figur 1,5 och 7 så var det anledningen till att en logaritmerad transformation gjordes då större värden i figurerna hade en större spridning bland residualerna.

Som man ser så är det inte lika många variabler i alla tre modeller. Modellen för Resarö och Vaxön tillsammans använder tre förklarandevariabler medan de var för sig har två respektive tre variabler. Variabeln boarea har visat sig i samtliga modeller vara den variabel som har störst inverkan på slutpriset för ett hus. Den variabeln är med i alla tre slutmodeller. En intressant iakttagelse är att när datamaterialet på Resarö användes så blev först bara variabeln boarea signifikant på 5 % nivån men när observationen då huset med endast 30 kvadratmeter togs bort blev även variabeln sålt signifikant.

Av de frågor som förväntades bli besvarade i arbetet så visade det sig alltså att olika variabler hade inverkan. De tre modellerna blev inte identiska. För det totala datamaterialet blev variabeln ö inte signifikant, vi kan då inte säkerställa att det skulle vara någon skillnad på de två öarna. Trots det visade det sig att modellerna för respektive ö skiljde sig när datamaterialet delades upp. En annan iakttagelse är att interceptet skiljer sig mycket åt mellan de två modellerna för Vaxö respektive Resarö. Orsaken till det är troligtvis då variabeln byggår blev signifikant för modellen för Vaxö men inte för modellen för Resarö. På frågan om det skiljde något mellan 2013 och 2014 visade det sig när det totala datamaterialet användes att variabeln sålt blev signifikant skild från noll på 5 % nivån, huspriserna stiger med åren.

Modellen för det totala datamaterialet och modellen för datamaterialet på Vaxön har samma förklarandevariabler med i modellen. Det enda som skiljer modellen för Resarö mot de andra är att variabeln byggår inte är med. De tre modellerna visade sig vara ganska lika varandra.

Ovan ser vi de tre olika modellerna samt deras skattningar för de olika parametrarna i varje modell. Något som är intressant att studera förutom att se vilka variabler som blev signifikanta är vilket tecken skattningen har, det vill säga om den är negativ eller positiv. I den första modellen i tabellen då det totala antalet data har använts, ser vi att vi har en negativ skattning β_2. Skattningen för β_2 hör ihop med variabeln byggår. Variabeln byggår anger vilket år huset byggdes. I den här modellen betyder det att äldre hus inte minskar slutpriset lika mycket som nyare hus gör. Medan de andra två variablerna i modellen boarea och sålt ökar slutpriset vid försäljning av ett hus. Det betyder i denna modell att ju större antalet kvadratmeter huset har och om det är sålt 2014 jämfört med 2013 ökar slutpriset av huset. Något som också är intressant utan att studera förutom vilka variabler som blev signifikanta är att se hur stora eller små skattningen av parametern blev, det vill säga om de har en stor påverkan eller inte. Något som är intressant att lägga märke till är att parametrarna för respektive variabel har ungefär samma skattning i de olika modellerna. Det betyder att variablerna ungefär har samma påverkan i de tre datamaterialen.
I den andra modellen för datamaterialet med husen på Vaxö ovan ser vi att även parameterskattningen för variabeln byggår är negativ. Annars är skattningarna positiva, vilket betyder att de ökar slutpriset vid försäljningen av huset.

Ett av antagandena som gjorts är att residualerna är normalfördelade i de olika modellerna. Det är därför av intresse att undersöka om antagandet verkar rimligt. I Figur 4, 9 och 10 så har normalfördelningsplottar gjorts över residualerna i de olika modellerna. Antagandet är uppfyllt då punkterna ligger längst den raka linjen i plotten, vilket de tycks göra på ett tillfredsställande sätt. I Figur 4 är punkterna mest slutna kring den raka linjen medan i Figur 9 och 10 ser avståndet från vissa punkter till linjen ut att vara större. I de två figurerna är det endast 42 respektive 43 observationer vilket inte är ett stort datamaterial. Några få avvikelser där ger större variation gentemot samma antal avvikelser jämfört med ett större datamaterial. I Figur 3, 11 och 17 visas residualplottarna för de enskilda variablerna i modellen, de anses visa en homoskedacitet, det vill säga en jämn fördelning bland punkterna.

När man kör en multipel linjär regression så vill man inte att två eller flera variabler ska korrelera med varandra då deras skattningar kan bli osäkra. För att analysera hur flera variabler korrelerar med varandra är ett bra och smidigt sätt att plotta en scatterplot. Man får då en enkel överblick över hur de olika variablerna står i relation till varandra. Efter att det gjordes stod det klart att variablerna boarea och rum korrelerade med varandra. Variabeln rum ersattes då med variabeln borum.

Alla tre modeller har relativt låg förklaringsgrad, vilket säger att modellen inte förklarar så mycket som man egentligen skulle vilja att den gjorde. Att förklaringsgraden är så pass låg beror i största del på att det saknas variabler i modellerna som förklarar slutpriset vid försäljning av ett hus, exempel på variabler diskuteras i kapitlet Diskussion. En annan orsak till den låga förklaringsgraden kan bero på begränsningen av datamaterialet. I Tabell 7 nedan presenteras förklaringsgraderna för de tre olika modellerna.

Tabell 7. Visar förklaringsgraden för de tre olika modellerna.

Datamaterial för modellen	Förklaringsgrad
Vaxö och Resarö	0.5324
Vaxö	0.6420
Resarö	0.5117

6. Diskussion

Målet och syftet med det här arbetet var att hitta en modell av variabler som förklarar slutpriset vid försäljning av hus. För bedömning av vilken modell som är att föredra i detta syfte så har förklaringsgraden, de förklarande variablerna samt hur väl de olika residualplottarna uppfyller antagandena gjorts. Det finns även fler mått för att undersöka resultaten. Förklaringsgraden har prioriterats högt för att syftet var att undersöka vilka variabler som påverkar slutpriset, ger även ett bra mått variationen i modellen. Men det har även lagts en stor vikt på att studera antalet förklarande variabler i modellen. En enkel modell med färre variabler är att föredra om det är möjligt, det blir till exempel färre variabler att skatta i modellen då. Antalet variabler i en modell har en koppling med förklaringsgraden, ju fler variabler som tillförs till en modell ökar förklaringsgraden.

När vi nu har fått fram tre modeller kan det tänkas naturligt att ställa sig frågan om vilken av dem som bäst för detta syfte. Skulle man enbart studera förklaringsgraden så skulle modellen för det totala datamaterialet och modellen för Vaxön vara bäst och modellen för datamaterialet på Resarö vara sämst. Mellan modellerna för Vaxö och Resarö skiljer det drygt tio procentenheter. Enligt mig skulle modellen med tre stycken förklarande variabler ha valts dels då båda de modellerna har en högre förklaringsgrad samt att den ger mer information om huset.

Då variabeln byggår inte blev signifikant för datamaterialet för Resarö men för Vaxö, kan man ställa sig frågan om vad som skiljer sig åt mellan öarna. Vaxholm är en kommun som består av flera öar ute i skärgården. Vaxön är huvudön medan Resarö dels nu men framförallt för ett par år sedan bestod av flera fritidshus som många hade sommarställen på. Medianen för byggåret i datamaterialet är 1970, vilket betyder att hälften av husen är byggda redan innan då. Väldigt många hus användes nog då som fritidshus, vilket det inte alls har varit på samma sätt på Vaxön. Vilket kan vara en förklaring till varför byggår inte blev signifikant för datamaterialet på Resarö.

Alla modeller som testats och undersökt är en förenkling av verkligheten. Det är svårt om inte omöjligt att hitta en modell som skulle överensstämma helt med verkligheten då den är så invecklad. Däremot är en modell inte oanvändbar för det, utan den säger något om detta resultat. I det här arbetet har vi enbart studerat linjära modeller med additiva faktorer efter transformeringen. Givetvis finns det en hel del andra funktioner och alternativ som kan testas för att möjligtvis få en bättre modell för detta syfte. Med denna begränsning kan vi alltså inte garantera att det inte skulle finnas andra modeller som skulle visa sig vara bättre för detta ändamål.

Något som är viktigt att poängtera i detta arbete är att datamaterialet är litet, enbart 87 stycken observationer. Detta gör att resultaten som visat sig ska tas med en liten försiktighet. När man har ett mindre datamaterial att arbeta med är det lättare att till exempel skattningar av olika parametrar inte blir helt korrekta. Det är viktigt att ha i detta i åtanke.

För en framtida studie och eventuellt en större undersökning kan man tänka sig att ha fler observationer samt lägga till fler variabler. Det finns många intressanta variabler som inte är med i denna modell som kan tänkas ha en inverkan på priset. I grundmodellen finns det två variabler om husets lokalisering, variabeln ö och hav. Det skulle vara intressant att ha en mer exakt sådan, till exempel gatunamn, postkod med mera. Variabeln hav anger avståndet till havet men skulle även kunna tänkas intressant att ha med andra avstånd, till exempel avstånd till buss, centrum, matbutik med mera. En annan faktor som ökar priset är om det är sjöutsikt eller inte, vilket tyvärr var svårt att ta reda på. Husets skick är en annan intressant faktor. Detsamma gäller där att det är svårt att ta reda på hur gott skick ett hus är. Andra exempel på intressanta variabler kan vara vilken mäklare, typ av säljare, när på året försäljningen sker.

7. Referenslista/litteraturförteckning

[1] http://www.aftonbladet.se/minekonomi/sparasmart/article20423934.ab

[2] http://www.dn.se/ekonomi/huspriserna-fortsatter-uppat/

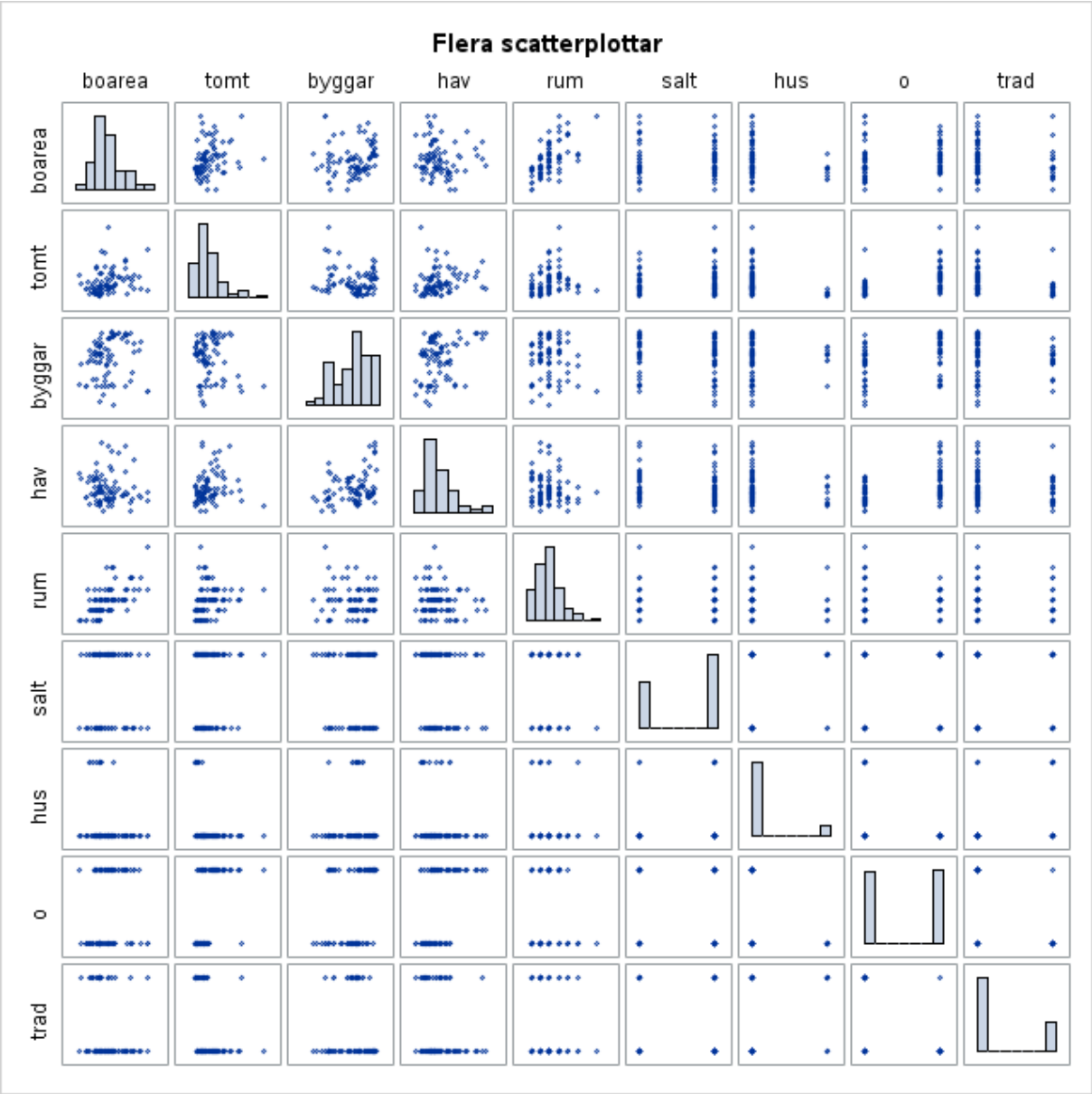
[3] http://www.svd.se/naringsliv/pengar/bostad/bostadspriserna-i-stockholms-innerstad-fordubblade-pa-tio-ar_4237035.svd

[4] www.booli.se

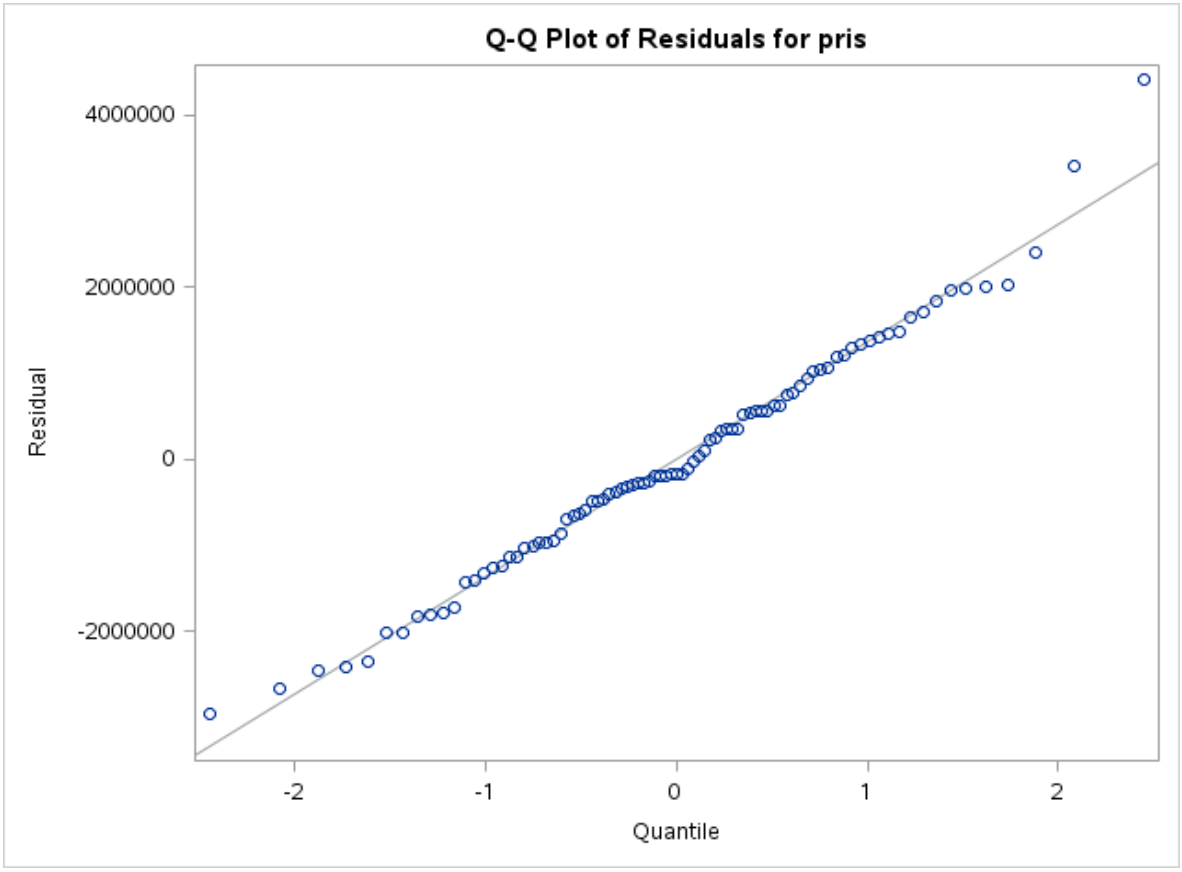
[5] Sundberg Rolf. Linjära Statistiska Modeller, 2014.

[6] Andersson Patrik och Tyrcha Johanna. Notes In Econometrics, 2014.

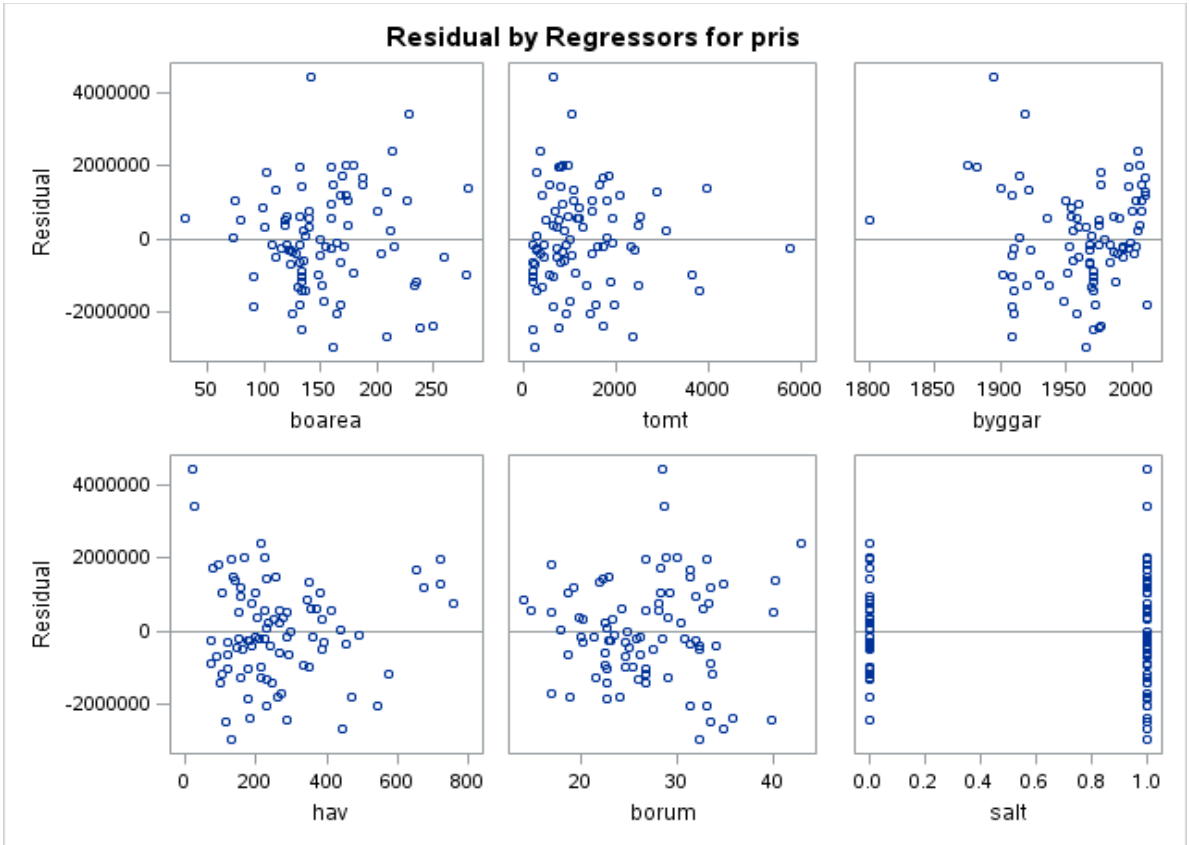
8. Appendix



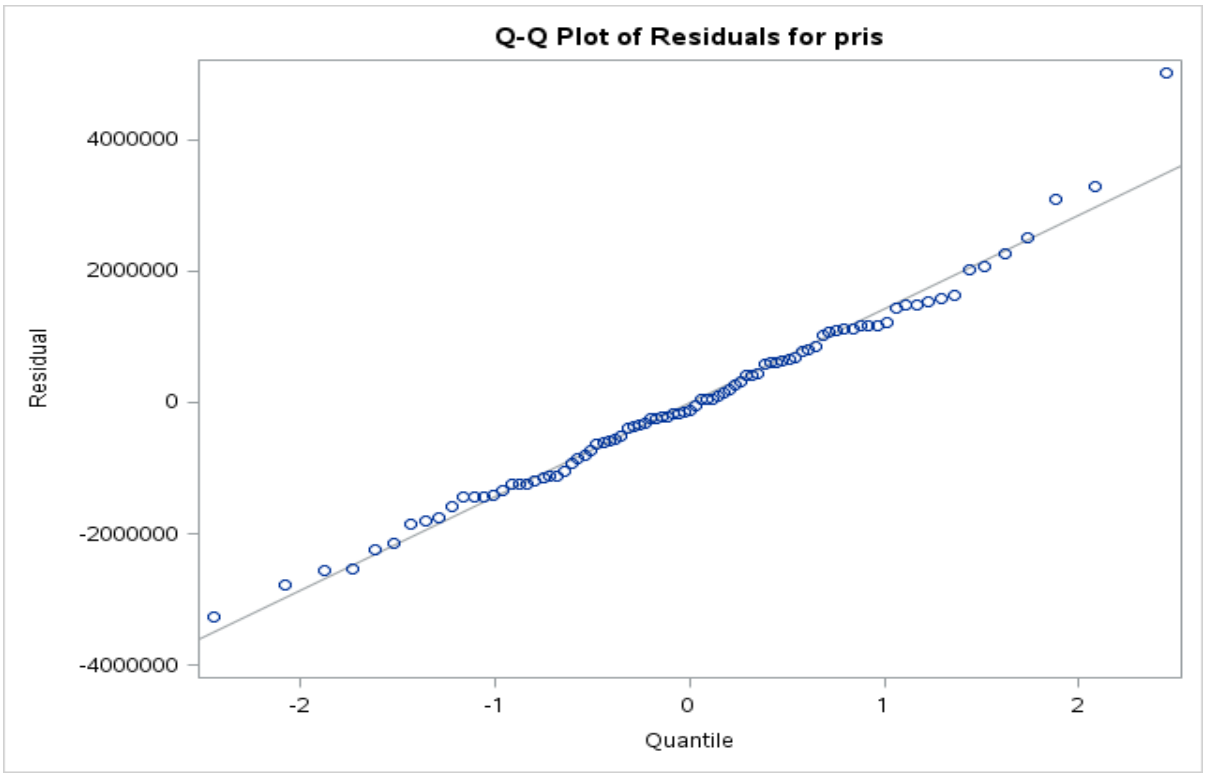
Figur 12. Scatterplot över samtliga variabler.



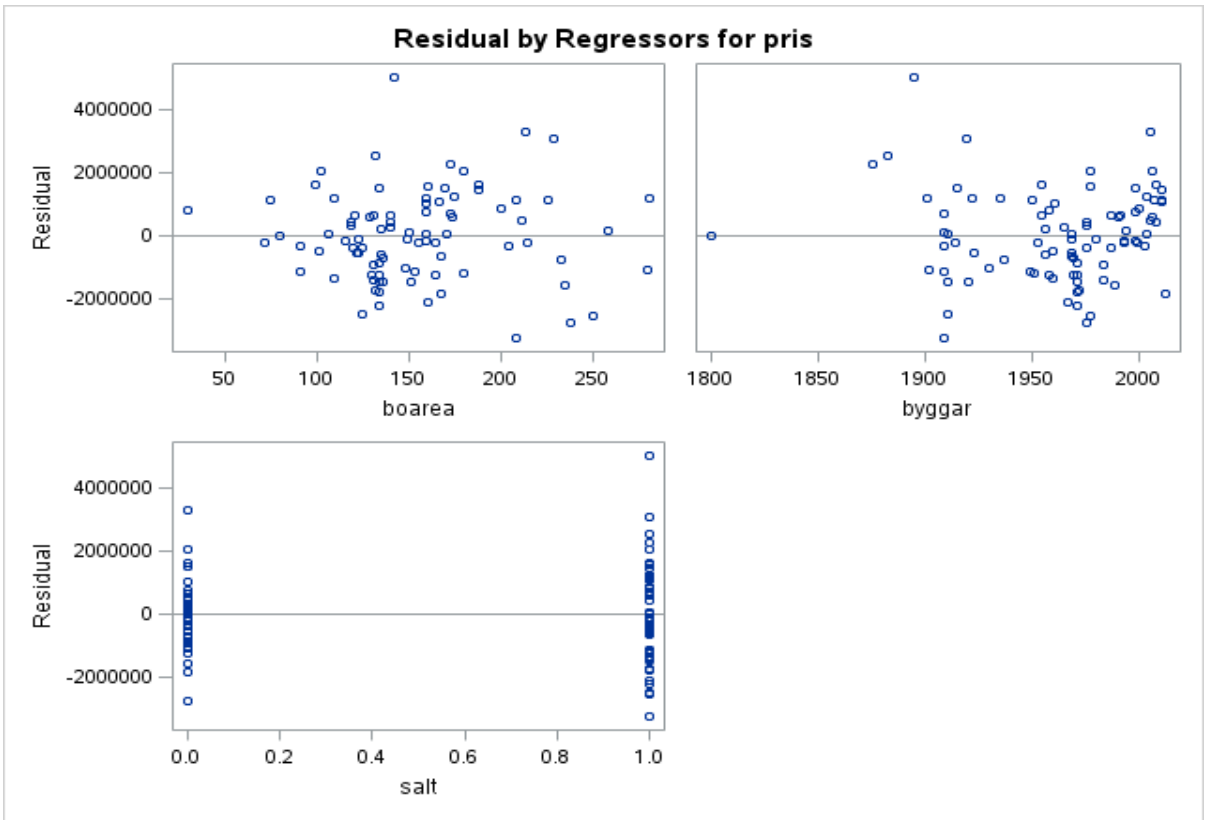
Figur 13. Normalfördelningsplot för alla variabler med rum ersatt av borum.



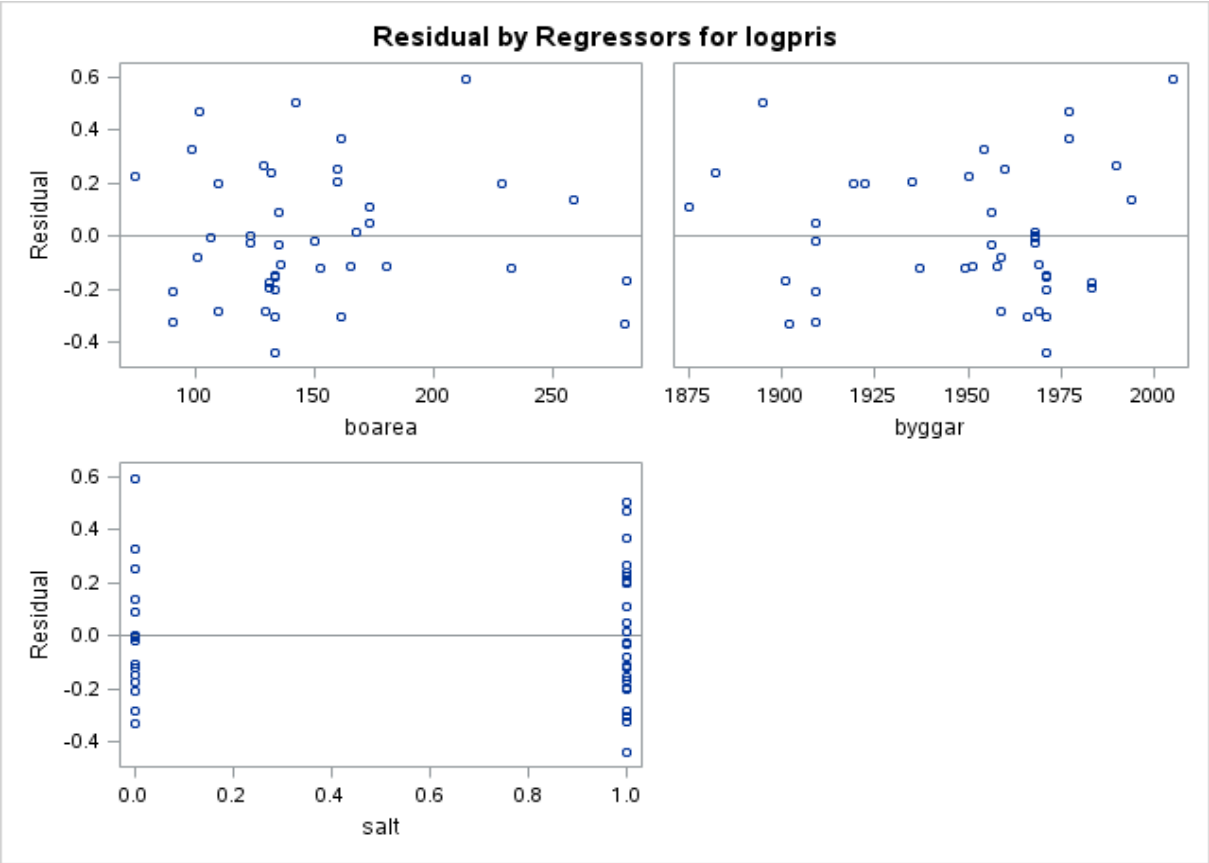
Figur 14. Residualplot för vissa av de enskilda variablerna i modellen.



Figur 15. Normalfördelningsplot efter variabler har exkluderats från grundmodellen.



Figur 16. Residualplot för de enskilda variablerna i modellen.



Figur 17.Residualpot för de enskilda variablerna i modellen.