Stockholms universitet

Prediktion av bostadrättspriser

Maximillian Alamgir

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2016:8
Matematisk statistik
Juni 2016

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholms universitet

Matematisk statistik
Stockholms universitet
Kandidatuppsats **2016:8**
http://www.math.su.se/matstat

Prediktion av bostadrättspriser

Maximillian Alamgir*

Juni 2016

Sammanfattning

Bostadsslutpriser beror på många olika faktorer och varierar kraftigt från område till område, år till år osv. Vi skall i detta arbete utreda vilka variabler som påverkar slutpriset på en lägenhet. Metoden vi ska använda oss av är multipel linjär regression. En faktor som påverkar bostadsslutpriset mer än andra variabler är utropspriset. Vi kommer därför att analysera två olika modeller en med utropspris och en utan utropspris samt se hur bra dessa två modeller predikterar slutpriset. Vi kommer att börja med att analysera en modell med slutpris som respons variabel och det kommer att visa sig att en linjär modell inte kommer att vara lämplig för vårt syfte. Detta kommer vi hantera med hjälp av transformation av slutpriset i form av en multiplikativ modell som överför modellen till en linjär modell.

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige.
E-post: Maximillianalamgir@gmail.com. Handledare: Jan-Olov Persson & Gudrun Brattström.

Innehållsförteckning

1. Introduktion

1.1 Introduktion

Bostadspriser är ett ämne som det alltid rapporteras om och det är ett ämne som berör de flesta någon gång förr eller senare i livet. Det skulle därför vara intressant att utreda olika faktorer som påverkar slutpriset för en lägenhet därför skall vi i detta arbete utreda olika utvalda faktorer som påverkar slutpriset för en bostad. Datan som är utvald består av 157 lägenheter runt om olika områden i Stockholm. De utvalda områdena är Östermalm, Södermalm, Solna, Älvsjö samt Hammarby Sjöstad. I avsnitt två går vi igenom den viktigaste teorin om multipel linjär regression för att sedan kunna göra en analys utifrån den teorin. I avsnitt tre beskriver vi vårt datamaterial detaljerat, i avsnitt fyra gör vi en analys och prediktion av vår data utifrån den teorin vi gått igenom i avsnitt två, i avsnitt fem beskriver vi resultatet av vår analys och vi avslutar sedan i avsnitt sex med en diskussion.

1.2 Syfte/Frågeställning

För att skapa en modell har vi olika förklarande variabler med i regressionen. Vissa variabler är mer inflytelserika i regressionen än andra variabler och en av dessa variabler är utropspris. Utropspriset är en mäklares värdering av lägenheten med hänsyn till de olika faktorerna som t.ex. yta, område, antal rum osv. Ibland så har man tillgång till utropspriset och ibland inte vi vill därför finna en lämplig modell som beskriver våra utvalda lägenheter om man har tillgång till mäklarens värdering av lägenheten och en modell där man inte har tillgång till mäklarens värdering av lägenheten. Vi vill med andra ord ta fram en modell utan utropspris som förklarande variabel och en annan modell med utropspris som förklarande variabel och därefter undersöka hur bra dessa modeller predikterar framtida försäljningspriser.

2. Teori Multipel Linjär Regression

2.1 Modell, parameterskattning och antaganden

2.1.1 Modell

Inom statistisk analys är det vanligt förekommande att en responsvariabel kan förklaras av flera olika förklarande variabler. Linjär regression är en statistisk modell där man studerar samband med mellan variabler. I modellen ingår en responsvariabel y och dess värde kan bero på en eller flera förklarande variabler x. Den allmänna modellen skrivs på följande sätt och typiskt för modellen är att den är linjär i sin parameteruppsättning[6]:

$$Y_i = \beta_0 + \sum_{j=1}^{m} \beta_j x_{ji} + \varepsilon_i, i = 1 \dots, n \quad (2.1)$$

I modellen ovan är Y_i en uppsättning av n oberoende stokastiska variabler. Uppsättningen β_j (j=0…,m) består av $k = m + 1$ okända parametrar och kallas även för regressionskoeffecient. β_j anger den genomsnittliga förändringen i y_i när x_{ji} ökar med en

enhet och övriga hålls konstanta. Och β_0 är det uppskattade värdet av y när alla förklarande variabler sätts lika med 0. Uppsättningen x_{ji} (j=1…,m, i=1…,n) består av m*n kända tal. Feltermen ε_i definieras som $\varepsilon_i = Y_i - (\beta_0 + \sum_{j=1}^{m} \beta_j x_{ji})$ där ε_i är stokastiska variabler. I denna modell antar vi att ε_i, $i = 1 \dots, n$ är oberoende och normalfördelade med väntevärde 0 och konstant varians σ^2.

Med vektornotation[4] kan vi istället skriva modellen 2.1 som:

Y=$X\beta$+ε, där

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ . \\ . \\ . \\ Y_n \end{pmatrix}, X = \begin{pmatrix} 1 & x_{11} & . & . & . & x_{m1} \\ 1 & x_{12} & . & . & . & x_{m2} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ 1 & x_{1n} & . & . & . & x_{mn} \end{pmatrix}, \begin{pmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ . \\ \beta_m \end{pmatrix} \text{ och } \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ . \\ \varepsilon_n \end{pmatrix}.$$

2.1.2 Parameterskattning

Som vi nämnde ovan anger β_j den genomsnittliga förändringen i y när x_j ökar med en enhet och övriga variabler hålls konstanta. Den förklarar med andra ord hur mycket den j:te variabeln X påverkar responsvariabeln Y. Parametervektorn β är okänd och måste därmed skattas med data. Tanken är att man vill minimera kvadratsumman av feltermerna $\sum_{i=1}^{n} \varepsilon_i^2$. Dvs. vi skall använda minsta kvadratmetoden för att skatta vår parametervektor.

Feltermen ε_i definierades som $\varepsilon_i = Y_i - (\beta_0 + \sum_{j=1}^{m} \beta_j x_{ji})$ enligt ovan. Eftersom vi kan skriva om detta på vektorform får vi alltså $\varepsilon = Y - X\beta$, vi vill alltså minimera kvadratsumman av uttrycket nedan[4]:

$|\varepsilon|^2 = (Y - X\beta)^T (Y - X\beta) = \|Y - X\beta\|^2$.

Man kan visa att den parametervektorn som minimerar uttrycket är:

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

Här betraktar vi $\hat{\beta}$ som stokastisk variabel, då parametervektorn beror på Y, när vi observerat Y sätter vi vi alltså y istället för Y.

Det går att visa att väntevärdesvektorn och kovariansmatrisen för $\hat{\beta}$ är

$E[\hat{\beta}] = \beta$ *samt* $\text{Var}[\hat{\beta}] = \sigma^2 (X^T X)^{-1}$.

Där en väntevärdesriktig skattning för σ^2 ges av följande formel:

$\widehat{\sigma^2} = \frac{1}{n-k}\left(Y - X\hat{\beta}\right)^2$.

2.1.3. Antaganden

Vi måste även ha med vissa antaganden när vi använder multipel linjär regression, dessa antaganden nämns i Joanna Tyrcha och Patrik Anderssons kompendium[1] och dessa lyder:

- $E[\varepsilon] = 0$
- $\text{Var}[\varepsilon] = \sigma^2$ dvs. konstant varians för varje kombination av $x_1, \ldots x_i$ har feltermen konstant varians(homoskedacitet)
- Normalfördelning för varje kombination av $x_1, \ldots x_i$ så följer värdena på feltermen en normalfördelning.
- Oberoende, alla feltermer är oberoende och normalfördelade med väntevärde 0 och varians σ^2.
- Inget $x_1, \ldots x_i$ ska kunnas skrivas som linjärkombination av de andra.

2.2 Hypotesprövning, P-värde

2.2.1 Hypotesprövning och p-värde

Inom linjär regression är det av stort intresse testa följande hypotes[1]:

$$H_0: \beta_j = 0$$
$$H_a: \beta_j \neq 0$$

Vi vill testa om vi kan förkasta H_0 eller inte, att en parameter är lika med noll, innebär med andra ord att motsvarande förklarande variabel x_j inte har påverkan på responsvariabeln.

Ovan konstaterade vi att kovariansmatrisen för $\hat{\beta}$ är $\sigma^2(X^TX)^{-1}$. Om vi nu vill ha variansen för en specifik parameter säg β_j, fås variansen alltså från denna matris på plats ”*jj*” dvs. $\sigma^2(X^TX)^{-1}{}_{jj}$. Vi testar ovanstående hypotes med teststatistikan

$$T_J = \frac{\widehat{\beta_J} - \beta_j}{\sqrt{\hat{\sigma}^2(X^TX)^{-1}{}_{jj}}} \sim t(n-k)$$

teststatistikan är alltså t-fördelad med *n-k* frihetsgrader under H_0. Vi förkastar nollhypotesen om $|T| > t_\alpha(n-k)$ där α är en lämplig vald felrisk. När man har en tvåsidig alternativ hypotes används t-kvantilen $t_{1-\alpha/2}(n-k)$.

När man väljer variabler som skall ingå i modellen väljer man variabler som är statistiskt signifikanta skilda från 0. Vi låter T_0 som det observerade värdet på vår teststatistikan T_J och definierar vi p-värdet för x_j som sannolikheten att $|T_J| \geq |T_0|$) då H_0 är sann. Matematiskt kan vi uttrycka detta på följande sätt:

$$p = P_{H_0}(|T_J| \geq |T_0|)$$

Om det gäller att $p \leq \alpha$, säger vi att variabeln är signifikant och skall därmed ingå i modellen.

2.3 Förklaringsgrad R^2 och justerad förklaringsgrad $\overline{R^2}$

2.3.1 Förklaringsgrad R^2

Förklaringsgraden är ett tal mellan 0 och 1(0 och 100 %) och är ett mycket vanligt anpassningsmått inom regression som Rolf Sundberg nämner i sitt kompendium[2]. Förklaringsgraden anger hur stor del av variationen i den beroende variabeln *y* som kan förklaras av den oberoende variabeln *x* under förutsättning att sambandet mellan *y* och *x* är linjärt. Vi definierar förklaringsgraden på följande sätt:

$$R^2 = \frac{kvs(regression)}{kvs(total)} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad eller$$

$$R^2 = 1 - \frac{kvs(residual)}{kvs(total)} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}.$$

2.3.2 Justerad Förklaringsgrad $\overline{R^2}$

Förklaringsgraden ökar ju fler *x* variabler vi lägger till i modellen, vilket i sin tur kan ge en missvisande förklaringsgrad därför finns justerad förklaringsgrad som tar hänsyn till detta. Den justerade förklaringsgraden mäter hur mycket variansen minskar i modellen jämfört med modellen när vi inte har några förklarande variabler.

$$\overline{R^2} = 1 - \frac{\widehat{\sigma^2}}{\hat{\sigma}_0^2} = 1 - \frac{kvs(residual)/n-k}{kvs(total)/n-1},$$

där vi betraktar $\hat{\sigma}_0^2$ som den variansskattning där ingen förklaringsvariabel ingår i modellen.

2.4 Dummyvariabler[1]

I många modeller kan det ingå kvalitativa och kvantitativa variabler. Det kan då vara lämpligt att införa en dummyvariabel. En dummyvariabel är en sådan variabel som antingen antar värdet 1 eller 0. Dummyvariabler som beskriver en specifik egenskap ska ha ett värde på 1 och 0 annars, exempelvis har vi att om en viss lägenhet har hiss kan vi skapa en

dummyvariabel sådan att variabel antar värdet 1 om lägenheten har hiss och 0 annars. På samma sätt kan man skapa dummyvariabler för t.ex. *n* olika kategorier antag att vi har tre kateogorier A, B samt C då skapas en dummyvariabel för dessa kategorier sådan att om kategorin är A så antar denna dummyvariabel värdet 1 om kategorin är A och 0 om kategorin är B eller C. På samma sätt skapas en till dummyvariabel för B, denna variabel antar värdet 1 om kategorin är B och värdet 0 om kategorin är A eller C. Kategori C används i detta fall som referens och det går lika bra att använda någon annan av dessa kategorier som referens. På motsvarande sätt kan vi skapa dummyvariabler för en kvalitativ variabel som kan anta *n* olika värden då behöver vi alltså skapa $(n-1)$ dummyvariabler.

2.5 Multikollineäritet och VIF

Multikollineäritet har vi när två eller flera förklarande variabler är linjärt beroende eller nästan linjärt beroende. Det är viktigt att undersöka detta eftersom när två förklarande variabler är högt korrelerade så kan man inte hålla isär effekterna av de två variablerna på den beroende variabeln. Multikollineäritet kan upptäckas genom att observera de skattade genom att exempelvis ta fram VIF värden.

En indikator för multikollineäritet som tas upp i Rolf Sundbergs kompendium[2] är variance inflation faktorn, VIF. En regression där vi sätter en av varje förklarande variabel som responsvariabel och erhåller alltså en viss förklaringsgrad för detta X_j när den används som responsvariabel och övriga variabler som förklarande variabler. Utgångsmodellen är $Y_i = \beta_0 + \sum_{j=1}^{m} \beta_j x_{ij} + \varepsilon_i, i = 1 \dots, n$ (2.1), via nedanstående formel kan vi beräkna VIF, där R^2 betraktas som förklaringsgraden för variabeln X_j när den används som responsvariabel i modellen:

$$VIF(x_j) = \frac{1}{1 - {R_j}^2}$$

Enligt Rolf Sundbergs bok är det vanligt att man sätter ett kritiskt värde på VIF faktorn mellan 5 och 10 och utfrån det avgör vi om variabeln skall ingå i modellen eller ej.

2.6 Stegvis variabelselektion

När vi anpassar en multipel linjär regressionsmodell är det vanligt att vi tar med många förklarande variabler. Däremot kan det vara så att alla dessa inte behövs för att förklara responsen. Vi kommer i detta arbete använda oss av tre metoder som hjälper oss att finna olika tillfredställande modeller med signifikanta variabler. I Rolf Sundbergs kompendium[2] beskrivs dessa på följande sätt:

2.6.1 Forward selection

I denna procedur utgår vi från att vi inte har några förklarande variabler. Vi lägger sedan till en variabel i taget med lägst p-värde. Proceduren upprepas sedan på detta sätt tills det inte finns några fler signifikanta variabler att inkludera i modellen.

2.6.2 Backward selection

Backward elimination är något av motsatsen till föregående procedur. Här börjar vi med att ha alla förklarande variabler i modellen. Sedan plockar vi bort den variabel med högst p-värde, proceduren stoppar då alla variabler i modellen är signifikanta.

2.6.3. Stepwise selection

Stepwise elimination är en kombination av stepwise och backward. Metoden börjar med endast responsvariabeln och utför inkludering enligt samma metod som forward selection. För att därefter utföra en backward elimination på en ny grupp av variabler. Men däremot så testar denna procedur till skillnad från ovanstående procedurer att de variabler man lagt till fortfarande är signifikanta. Om det visar sig att en variabel inte längre är signifikant plockas den bort.

2.7 Transformation av variabler

När man hanterar data är det inte alltid så att data visar på linjärt samband, då kan det vara en fördel att hantera data med transformationer som Rolf Sundberg[2] tar upp i sitt kompendium. En vanlig transformation är logtransformationen. Man kan antingen logaritmera responsvariabeln eller förklarande variablerna eller också både respons variabeln och förklarande variablerna. Exempel illustreras nedan. Antag att vi har grundmodellen $Y_i = \beta_0 + \beta_1 x_{1i} + \cdots, \beta_m x_{mi}$.

$$\ln(Y_i) = \beta_0 + \beta_1 x_{1i} + \cdots, \beta_m x_{mi},$$

$$Y_i = \beta_0 + \beta_1 \ln(x_{1i}) + \cdots, \beta_m \ln(x_{mi}),$$

$$\ln(Y_i) = \beta_0 + \beta_1 \ln(x_{1i}) + \cdots, \beta_m \ln(x_{mi}).$$

Om vi finner en bättre modell med hjälp av någon av ovanstående transformationen, är det ett alternativ till att arbeta med den transformerade datan istället.

2.8 Prediktion

2.8.1 Prediktion

Prediktion är en metod som används för att uppskatta ett kommande utfall. Metoden går ut på att man använder sig av kända förklarande variabler och med hjälp av dessa i en modell uppskatta ett kommande värde. En prediktor för **Y** baserad på **X** definierar vi som en funktion c(**X**) som alltså är linjär i parameteruppsättningen. I Anna Flodströms examensarbete[5] definerar hon prediktion på följande sätt:

$c(\mathbf{X})=b_0 + b_1X_{1i} + b_2X_{2i} + \cdots, +b_mX_{ji}$, $i=1\ldots,n$.

Där prediktionsfelet ges av

$\mathbf{Y}-c(\mathbf{X})$. Vi inför följande notation som prediktionsfel: $\widehat{\varepsilon_{(i)}} = y_i - \hat{y}_i$.

Pressvärdet definieras på följande sätt:

$PRESS = \sum_{i=1}^{n} \hat{\varepsilon}_{(i)}^2$,

Den modell som är att föredra enligt minsta kvadratmetoden är modellen med lägst pressvärde. Det är däremot inte säkertställt att det är den bästa modellen då vi även måste ta hänsyns till antal variabler osv. En modell med färre variabler med liten ökning av pressvärde kan också vara att föredra.

3. Data

3.1 Beskrivning av data

I detta examensarbete har vi valt ut 157 lägenheter inhämtade från booli.se[3] fördelade över fem olika områden. Områdena som valts ut är Östermalm, Södermalm, Solna, Älvsjö och Hammarbysjöstad, slutpriset är responsvariabeln och de förklarande variablerna beskrivs närmre här nedan. Antalet observationer är cirka 30 per område fördelade över 3 olika årtal dvs. 2013, 2014 samt 2015 med 10 observationer per år.

Slutpris- Priset lägenheten landade på när den såldes. Detta kommer att vara vår responsvariabel. Priset anges i enheten kr.

Boyta kvm- Storlek på lägenheten är en kontinuerlig variabel och varierar mellan 21 kvm och 202 kvm.

Antal rum- Antal rum som finns i lägenheten är en diskret variabel. I vårt datamaterial har vi 1 rum till 7 rum som mest.

Våning- Våningen lägenheten är belägen i dvs. en diskret variabel, den lägsta våningen är plan 1 och den högsta plan 19 i vårt datamaterial.

Avstånd till vatten km – Avstånd till vatten i kilometer en kontinuerlig variabel.

Utropspris- Utropspriset för lägenheten innan lägenheten såldes en kontinuerlig variabel som anges i enheten kronor.

Pris per kvm – Priset per kvadratmeter baserat på utropspriset för lägenheten och är en kontinuerlig variabel. Med andra ord utropspriset dividerat med antal kvadratmeter och vi får alltså priset i kr/kvm.

Månadsavgift – En avgift som betalas till bostadsrättsföreningen per månad. Desto ”bättre” förening desto lägre månadsavgift, varierar allt mellan 450 kr/mån upp till 5000 kr/mån alltså en kontinuerlig variabel.

Område – Kategorisk variabel, uppdelade över fem olika områden, Östermalm, Södermalm, Älvsjö, Solna samt Hammarbysjöstad.

Byggår – Året då lägenheten/fastigheten byggdes. Varierar från slutet av 1800-talet fram till 2013. Byggår är alltså en diskret variabel.

År- Året då lägenheten såldes, fördelad över åren 2013, 2014 samt 2015. År är alltså en diskret variabel.

Hiss – Om fastigheten har hiss eller inte.

3.2 Hantering av data

Innan vi börjar analysera data bör vi omformatera några variabler. Område är kategoriska variabler, vi gör om dessa till dummy variabler på följande sätt:

Område1(Östermalm) = 1 om lägenheten ligger i Östermalm och 0 annars

Område2(Södermalm) = 1 om lägenheten ligger på Södermalm och 0 annars

Område3(Solna) = 1 om lägenheten ligger i Solna och 0 annars

Område4(Hammarby Sjöstad) = 1 om lägenheten ligger i Hammarby sjöstad och 0 annars

Område 5(Älvsjö) kommer att fungera som referens.

Vi kommer även införa variablerna *yta per rum* samt *avgift per yta.*

Där *yta per rum* är *yta/rum* en kontinuerlig variabel angiven i enheten *kvadratmeter*.

Och *avgift per yta* är *avgift/yta* en kontinuerlig variabel som är angiven i enheten antal *kr/yta.*

År 2013, 2014, 2015 samt 2016 kommer vi i SAS att kalla för år 99, 100, 101 samt 102, där år 99 motsvarar 2013, år 100 år 2014 osv.

4. Analys av data

4.1 Analys av data

Vi skall nu analysera datan vi har och som vi nämnde ovan i avsnitt 1 så ville vi ta fram en modell med där utropspriset inte ingår som förklarande variabel och en modell där utropspriset ingår som förklarande variabel. Vi ska utifrån teorin vi gått igenom ovan ta fram två lämpliga modeller och därefter se hur bra dessa två modeller predikterar framtida lägenhetsslutpriser för år 2016.

Modell utan utropspris som förklarande variabel

Vi tar fram VIF-värdena för att upptäcka eventuell korrelation mellan variabler och som vi nämnde ovan satte vi gränsen för VIF-värde till 5. Vi erhåller följande tabell och som vi minns var formeln för beräkning av VIF-värde för en viss förklarande variabel följande: $\mathrm{VIF}(x_j) = \frac{1}{1-{R_j}^2}$.

Variabel	VIF-värde
Yta	11.37
Rum	8.74
Område 1	4.44
Område 2	5.68
Område 3	3.05
Område 4	4.64
Våning	1.11
Byggår	2.76
År	1.07
Hiss	1.45
Avgift	5.67
Vattenavstånd	3.94

Tabell 4.1: VIF-värden över olika möjliga förklarande variabler

Utifrån denna tabell ser vi att yta, rum, område 2 och avgift har VIF-värde som är större än 5. Vi tittar därför på korrelationsmatrisen med dessa mellan dessa variabler för att se hur korrelationen mellan dessa förhåller sig till varandra.

Pearson Correlation Coefficients, N = 157 Prob > |r| under H0: Rho=0				
	slutpris	yta	rum	avgift
slutpris	1.00000	0.72317 <.0001	0.70078 <.0001	0.39249 <.0001
yta	0.72317 <.0001	1.00000	0.92983 <.0001	0.80135 <.0001
rum	0.70078 <.0001	0.92983 <.0001	1.00000	0.73212 <.0001
avgift	0.39249 <.0001	0.80135 <.0001	0.73212 <.0001	1.00000

Tabell 4.2: Korrelationsmatris mellan variablerna slutpris, rum, yta och avgift.

Vi tittar på korrelationen mellan yta och rum och ser att korrelationskoeffecienten mellan dessa är 0.92983. Med andra ord förklarar dessa två variabler i princip samma sak. Tittar vi nu på korrelationen mellan slutpris och rum ser vi att den är 0. 70078 och 0. 72317 mellan slutpris och yta, vi väljer att plocka bort rum som förklarande variabel och vi inför en variabel ”*yta per rum*” samt ”*avgift per yta”*, som heller inte blir signifikanta och dessa variabler plockas alltså bort i den fortsatta analysen.

Vi kunde även se att område 2 hade högt VIF-värde, denna variabels VIF värde kommer inte påverka vår fortsatta analys eftersom det är en enskild dummy variabel som är negativt korrelerad med övriga områden, eftersom vi säkert vet att om en lägenhet ligger i område 1 så kan den omöjligt ligga i exempelvis område 3. Vi bortser alltså från att en enskild dummyvariabel har något högre VIF värde.

Vi utför en regression på modellen med de förklarande variablerna yta, område 1 till område 4, våning, byggår, år, hiss, avgift samt vattenavstånd och ser att variablerna yta, år, område 1, område 2, område 4, byggår och avgift blir signifkanta med andra ord blir dessa variablers p-värden mindre än 0.05 som vi satt vår signifikansnivå till och vi kan alltså förkasta hypotesen om att β_j=0 för dessa variabler.

När vi utför enkel linjär regression med de förklarande variablerna får vi att yta, område 1-område 4, byggår, år avgift samt vattenavstånd signifkanta. Eftersom hiss inte blir signifikant vid enkel som multipel linjär regression utesluter vi den variabeln i den fortsatta analysen.

Vi tar återigen fram VIF värden för de variabler vi valt att ha kvar i den fortsatta analysen:

Förklarande Variabel	*VIF-värde*
Yta	3.68
Område 1	4.32

Område 2	5.29
Område 3	2.95
Område 4	4.62
Avgift	4.67
Vattenavstånd	3.92
Byggår	2.27
År	1.02

Tabell 4.3 VIF-värden för förklarande variabler.

Vi kan se att VIF-värde längre inte är ett problem i den fortsatta analysen. Att område 2 har något högre VIF-värde än 5 är inget som påverkar den fortsatta analysen som vi nämnt tidigare eftersom det beror på att områdena är lineära kombinationer av varandra och alltså är korrelerade med varandra.

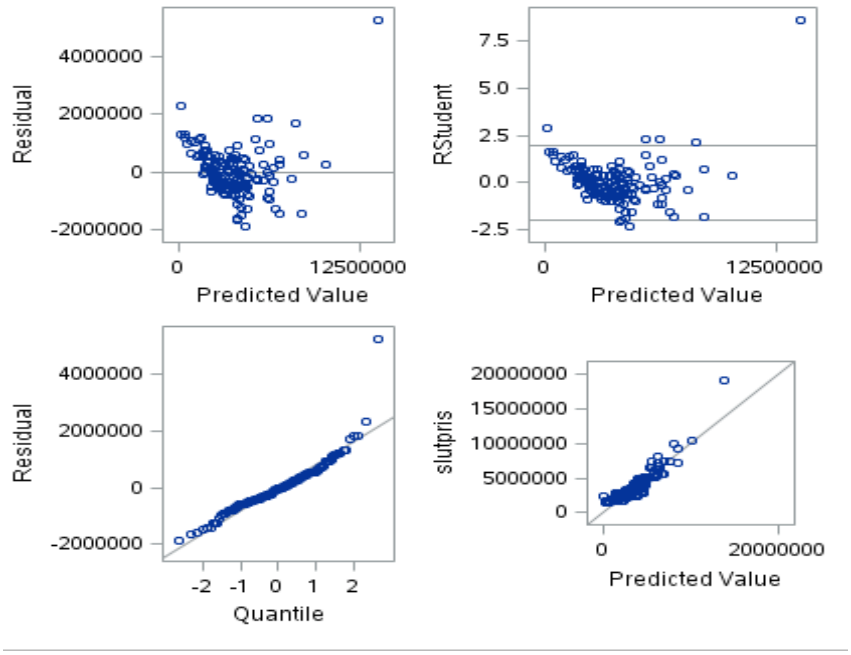
Vi ansätter en multipel linjär regressionsmodell med ovanstående förklarande och använder oss av samtliga variabel selektionsmetoder som finns i SAS. Resultatet illusteras i tabellen nedan:

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-46846202	8217519	2.26854E13	32.50	<.0001
yta	69899	4445.36476	1.725885E14	247.25	<.0001
omrade1	2842544	239085	9.867051E13	141.35	<.0001
omrade2	1978042	223746	5.455577E13	78.16	<.0001
omrade3	580363	210207	5.320907E12	7.62	0.0065
omrade4	1477471	213930	3.329463E13	47.70	<.0001
ar	456497	82152	2.155358E13	30.88	<.0001
avgift	-269.78832	95.28658	5.595798E12	8.02	0.0053

Tabell 4.4 Resultat av metoderna stepwise, forward samt backward.

Som vi kan se ovan ger alla tre metoder samma resultat. Förklaringsgraden för modellen blev 0.8470 och justerade förklaringsgraden 0.8398.

Vidare fortsätter vi titta på plottar för modellen vi fått fram:

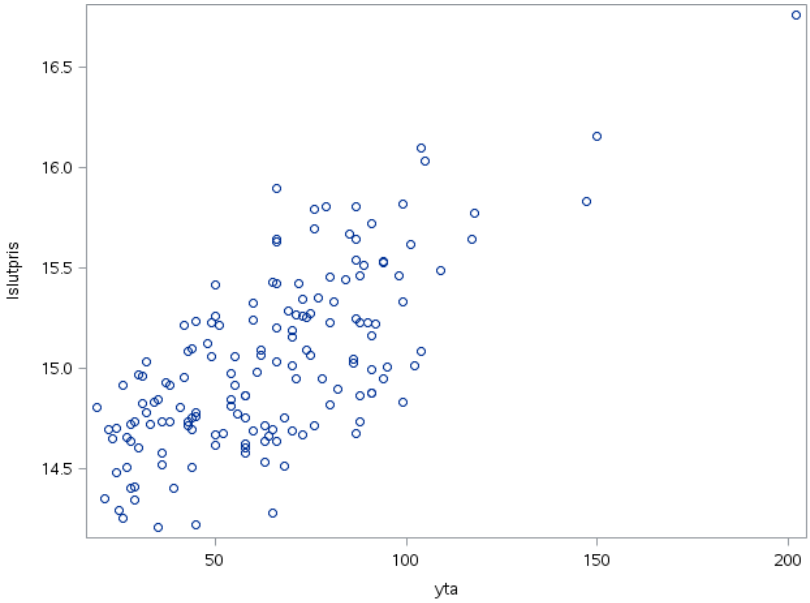


Figur 4.1 Plottar för Slutpris med förklarande variablerna Yta, avgift, vattenavstånd, år, samt område 1-4.

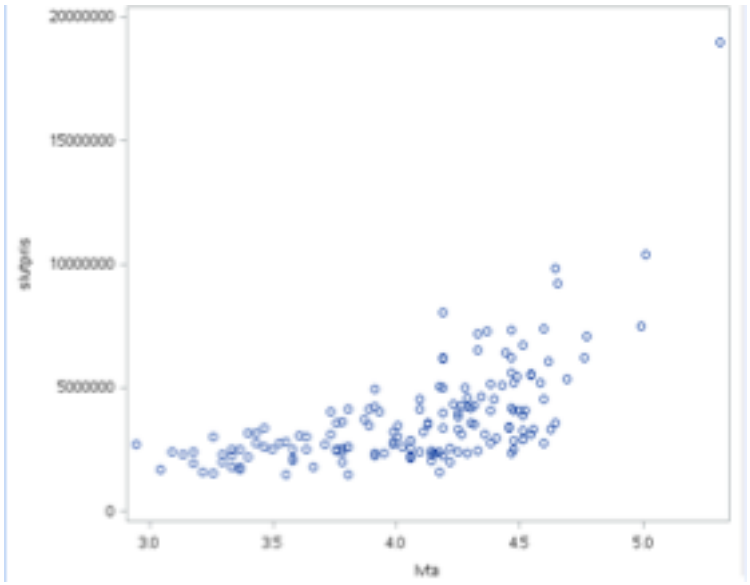
Längst upp till vänster har vi residualplot för regressorn mot predikterade slutpris, till höger om den har vi standardiserade residualer och nederst till vänster normalfördelningsplot för residualerna och till höger predikterade. Analyserar vi figuren till vänster längst upp ser vi att residualerna i figurerna inte är jämnt fördelade och symmetriska kring noll vilket inte är önskvärt, utan vi ser istället ett mönster som kan bero på att vi inte har ett linjärt samband mellan slutpris och våra förklarande variabler eller också att variansen inte är konstant. Tittar vi på normalfördelningsploten kan vi bl.a. se att en av observationerna avviker från övriga. Vi bör alltså testa någon form av transformation lämpligen att transformera slutpris eftersom residualplotten inte är jämnt utspridd och symmeterisk kring 0 är modellen med andra ord inte lämplig för vårt ändamål.

4.2 Transformerad modell

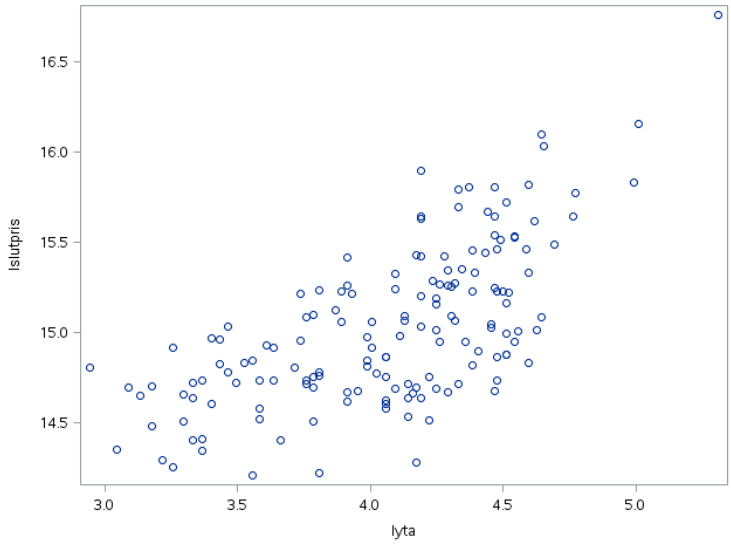
Yta är variabeln som förklarar mest i modellen vi plottar logaritmerat slutpris mot yta, slutpris plottad mot logaritmerat yta samt logaritmerat slutpris plottad mot logaritmerat yta.



Figur 4.2 Yta plottad mot logaritmerad slutpris.



Figur 4.3 slutpris plottad mot logaritmerad yta.



Figur 4.4 logaritmerad yta plottad mot logaritmerad slutpris.

Utifrån figurerna ovan ser vi från figur 4.2 samt 4.4 att logaritmering av slutpris är en lämplig transformation eftersom sambandet ser mest linjärt ut i dessa två figurer. Tittar vi istället på figur 4.3 så ser sambandet snarare ut att öka exponentiellt än linjärt. Vidare är det oklart om vi bör logaritmera yta eller inte vi fortsätter vidare i vår analys. Ovan hade vi variablerna yta per rum dvs. yta dividerat med rum, samt avgift per yta. Dessa variabler var inte signifikanta i analysen vi genomförde ovan däremot har vi nu ny responsvariabel och använder dessa i vår fortsatta analys och vi vet att avgift är korrelerad med yta och väljer därför att plocka bort variabeln avgift och använder istället variabeln avgift per yta.

4.2.1 Modell med logaritmerat slutpris

Vi ansätter en multipel linjär regressionsmodell med de förklarande variabler vi har och får följande VIF-värden.

Förklarande variabel	VIF-värde
Yta	1.30
Yta per rum	1.12
Avgift per yta	1.69
År	1.06
Område 1	4.41
Område 2	5.75
Område 3	3.09
Område 4	4.85
Vattenavstånd	4.16
Hiss	1.48
Våning	1.12
Byggår	2.77

Tabell 4.5 Möjliga förklarande variabler med log slutpris som responsvariabel.

När vi utför regression med dessa variabler blir resultatet som ovan att hiss och våning blir inte signifikanta vid enkel som multipel linjär regression. Vi plockar bort dessa variabler som tidigare. Med samma resonemang som ovan så har område något högre VIF-värden då de beror av varandra. Vi kör nu alla tre variabel selektionsmetoder och får fram följande modell.

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.60909	1.53528	0.40	0.6921
yta	1	0.01191	0.00047240	25.21	<.0001
vattenavs	1	-0.05899	0.02328	-2.53	0.0123
ar	1	0.13493	0.01536	8.79	<.0001
omrade1	1	0.55169	0.06043	9.13	<.0001
omrade2	1	0.35993	0.07029	5.12	<.0001
omrade3	1	0.05627	0.05291	1.06	0.2893
omrade4	1	0.23987	0.06508	3.69	0.0003

Tabell 4.6 Resultat av Stepwise, Forward samt Backward med logslutpris som förklarande variabel.

Som vi ser ovan får vi en modell med de förklarande variablerna yta, vattenavstånd, område 1,2,4 samt år. Med förklaringsgrad på 0. 8773 och justerad förklaringsgrad 0. 8716 med standardavvikelse 0.15643.

Modellen är angiven på följande form:

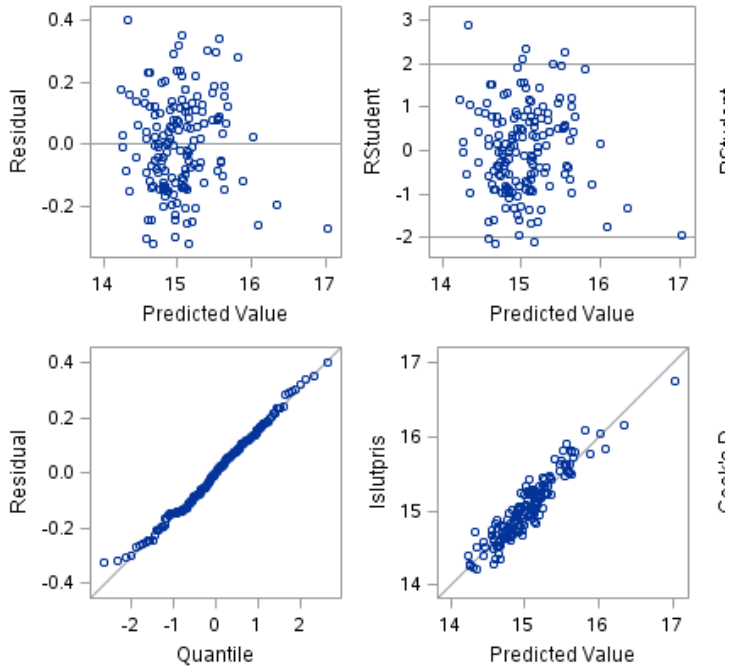
$$\ln(slutpris) = 0.6090 + 0.01191(yta) + (-0.05899)(vattenavstånd) + 0.1349(\text{år}) + 0.5516(område\ 1) + 0.3593(område\ 2) + 0.05627(område\ 3) + 0.2398(område\ 4) + \varepsilon$$

Dvs. med multiplikativa faktorer:

$$slutpris = e^{0.6090} * e^{0.01191(yta)} * e^{-0.05899(vattenavstånd)} * e^{0.1349(\text{år})} * e^{0.5516(område\ 1)} * e^{0.3599(område\ 2)} * e^{0.0562(område\ 3)} * e^{0.2398(område\ 4)} * e^{\varepsilon}.$$

Där e^{ε} är log normalfördelad med parameter 0 och 0.15643^2 eftersom $\hat{\sigma} = 0.15643$.

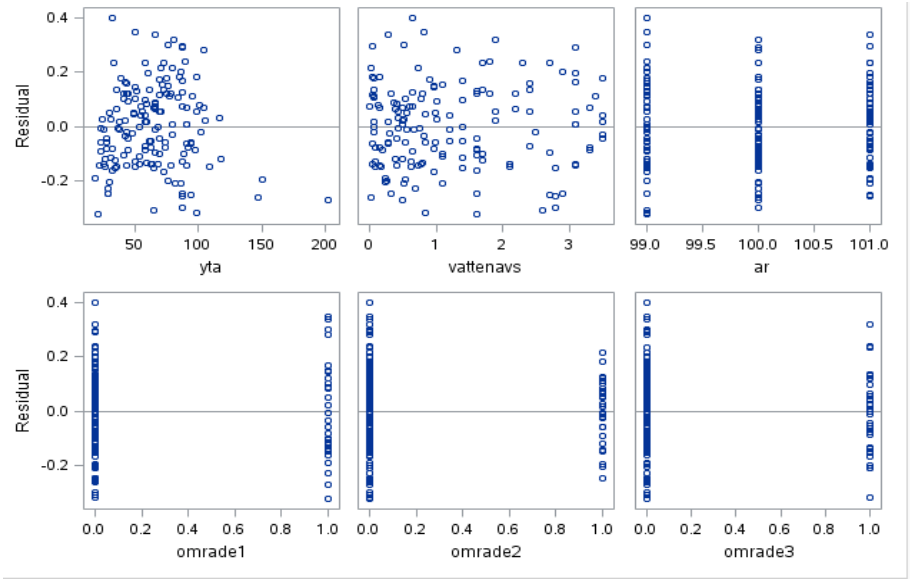
Modellen ovan tolkar vi på följande sätt. När *j:te* förklarande variabeln ökar med en enhet och övriga hålls konstanta så ger det en förändring i slutpris med en multiplikativ faktor på e^{β_j}. I avsnitt 5 går vi nämre in på hur vi tolkar varje variabel.

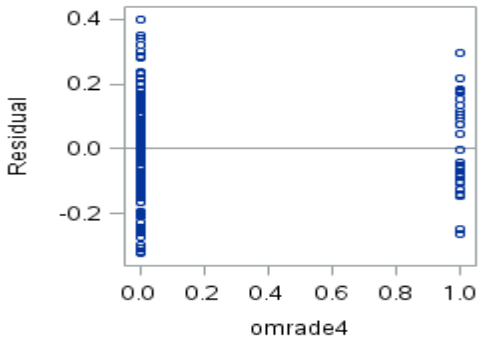


Figur 4.5 Residualplottar för logslutpris som responsvariabel med förklarande variablerna yta, vattenavstånd, år, område 1- 4.

Vi studerar plottarna ovan. Vi ser på plottarna ovan att residualerna ser jämnt utspridda ut och symmetriska kring noll. Tittar vi på de standardiserade residualerna dvs. figuren längst upp till höger ser vi att 7 av de standardiserade residualerna har absolutvärde något större än 2. Vidare på plottarna nedan ser vi att residualerna på figuren nederst till vänster anpassar sig bra till linjen och antagandet om feltermens varians verkar stämma. Tittar vi på de predikterade mot logaritmerad slutpris anpassar observationerna sig till linjen hyfsat väl.

Vi tittar vidare på residualplottar mot de förklarande variablerna.





Figur 4.6 Residualplottar för de olika förklarande variablerna

Vi studerar bilden ovan och ser att observationerna för varje förklarande variabel är jämnt utspridd kring noll. Spridningen säger oss att vi inte kan tala emot att sambandet skulle vara linjärt i parameteruppsättningen mot logslutpris som responsvariabel.

4.2.2 Logaritmerat slutpris och yta

Vidare var det oklart om vi skulle logaritmera yta eller inte eftersom både yta och logaritmerat yta visade sig vara linjära i plottarna ovan. Vi testar nu att göra det och låter precis som ovan SAS välja ut en lämplig modell åt oss med hjälp av de tre procedurerna dvs. stepwise, backward samt forward och får följande resultat.

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-1.27820	1.47661	-0.87	0.3881
lyta	1	0.75227	0.02842	26.47	<.0001
omrade1	1	0.62161	0.05869	10.59	<.0001
omrade2	1	0.40635	0.06800	5.98	<.0001
omrade3	1	0.05134	0.05077	1.01	0.3136
omrade4	1	0.24191	0.06253	3.87	0.0002
ar	1	0.13072	0.01476	8.85	<.0001
vattenavs	1	-0.06480	0.02231	-2.90	0.0042

Tabell 4.7 Resultat av stepwise, forward och backward med logslutpris som responsvariabel och logyta, vattenavstånd, område 1-4 och år som förklarande variabler.

Modellen vi får fram har samma förklarande variabler som modellen utan logaritmerat yta dock med andra skattningar. Modellen har förklaringsgrad på 0.8868 och justerad förklaringsgrad på 0.8815 med standardavvikelse 0.1503.

Vi får följande med modell:

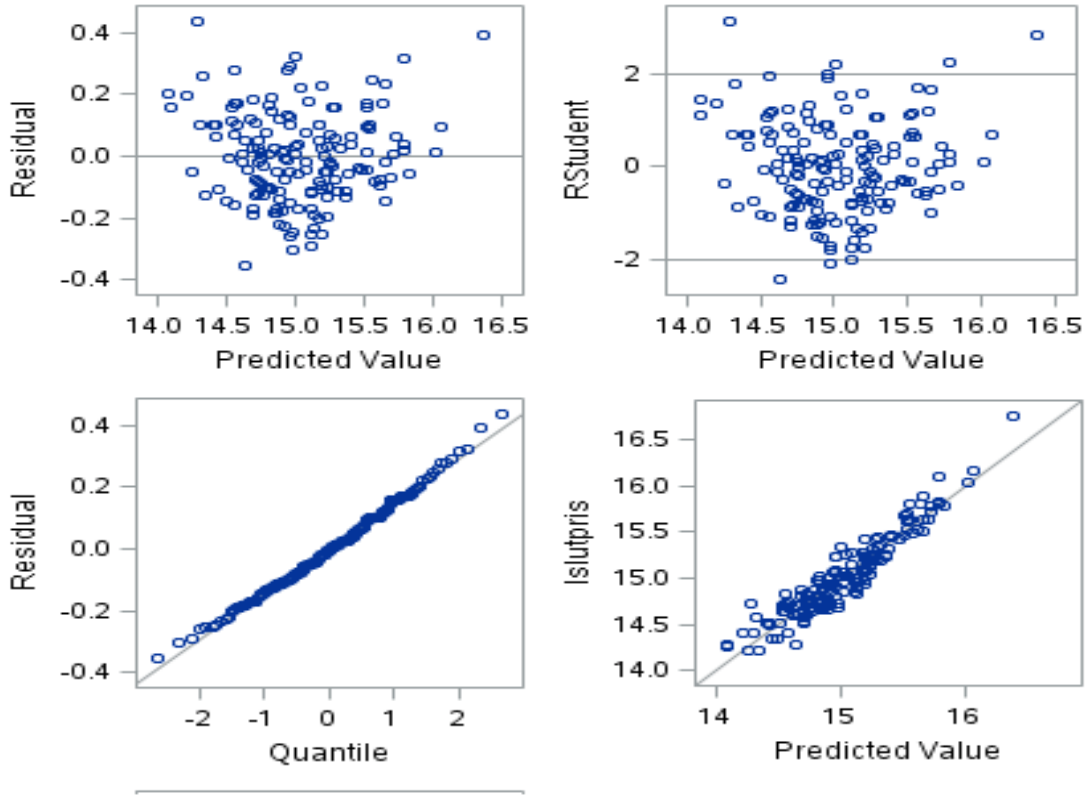
$$\ln(slutpris) = \\ -1.27 + 0.7522\ln\big((yta)\big) + 0.1307(\text{år}) + (-0.06480)(vattenavstånd) + \\ 0.6216(område\ 1) + 0.4063(område\ 2) + 0.0513(område\ 3) + 0.2419(område\ 4) + \varepsilon$$

Med multiplikativa faktorer.

$$slutpris = e^{-1.27} * e^{0.7522(logaritmerad\ yta)} * e^{0.13072(\text{år})} * e^{(-0.06480)(vattenavstånd)} \\ * e^{0.6216(område\ 1)} * e^{0.4063(område\ 2)} * e^{0.0513(område\ 3)} * e^{0.2419(område\ 4)} * e^{\varepsilon}$$

På samma sätt tolkar vi parameterna i modellen som ovan utom β_1 som alltså är logaritmerade ytan så säg att för varje enhet yta ökar slutpris med en faktor av den logaritmerade ytan multiplicerat med 0.744. Och e^{ε} är lognormal fördelat med parameter 0 och 0.1488^2=0.02214.

Vi studerar nu plottarna på samma sätt som vi gjort ovan.

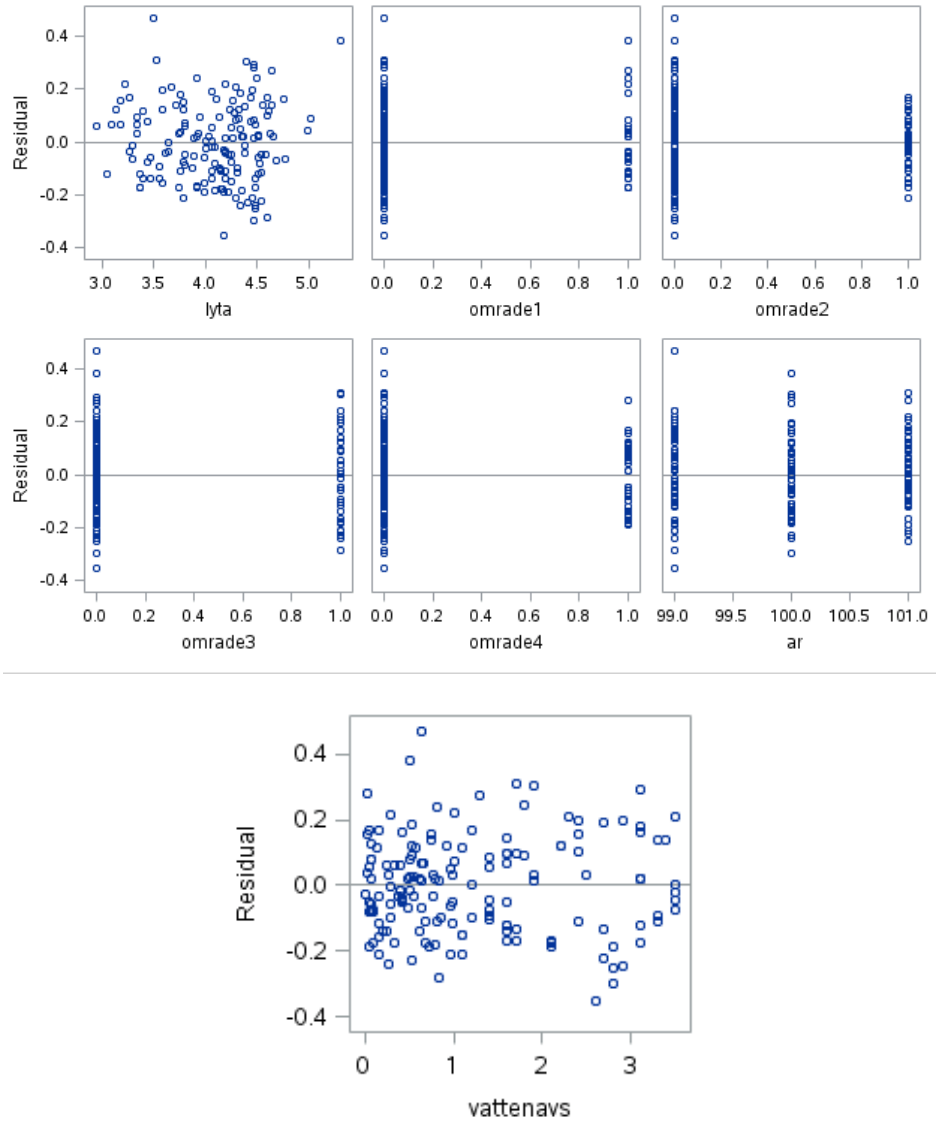


Figur 4.7 Residualplottar för modell med logslutpris som respons och logaritmerad yta, område 1,2,3,4, år och vattenavstånd som förklarande variabler.

Studerar vi plotten ovan till vänster så ser vi att den ser bra ut tittar vi på plotten för de standardiserade residualerna ovan till höger ser vi att åtminstonde fem av observationerna har

något större absolutbelopp än två. Och studerar vi nu hur residualerna anpassar sig till linjen så anpassar de sig bra och antagandet om feltermens varians stämmer alltså.

Vi fortsätter titta på residualer mot varje förklarande variabel.



Figur 4.8 Residualplottar för förklarande variablerna i modell med logslutpris som responsvariabel.

Studerar vi plottarna ovan, finns inget som talar emot linjärt samband mellan logaritmerat slutpris och de förklarande faktorerna då de ser jämnt utspridda ut kring noll.

4.3 Modell med utropspris som förklarande variabel

4.3.1 Analys med utropspris i modellen

På samma sätt som vi gjort analysen ovan gör vi nu med utropspris ingående i modellen och kommer fram till följande modell:

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-2.03932	0.73765	0.04316	7.64	0.0064
lutrpris	0.92723	0.01353	26.50848	4694.69	<.0001
ar	0.03248	0.00757	0.10400	18.42	<.0001

Tabell 4.8 förklarande variabler tillsammans med logaritmerat utropspris som förklarande variabel och logaritmerad slutpris som responsvariabel.

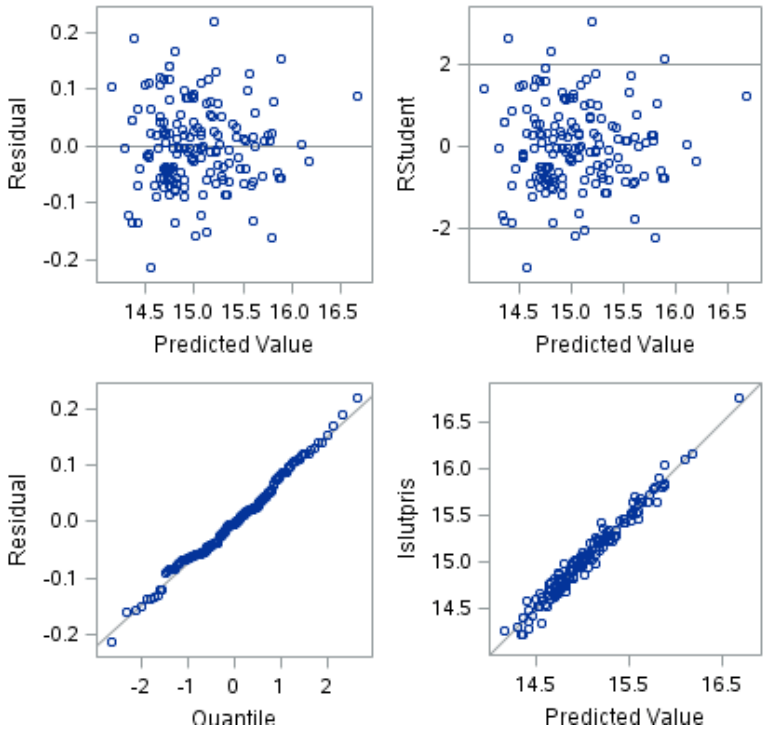
Modellen är alltså på följande form

$$\ln(slutpris) = -2.03 + 0.92723 * ln(utropspris) + 0.03248(\text{År}).$$

Med multiplikativa faktorer:

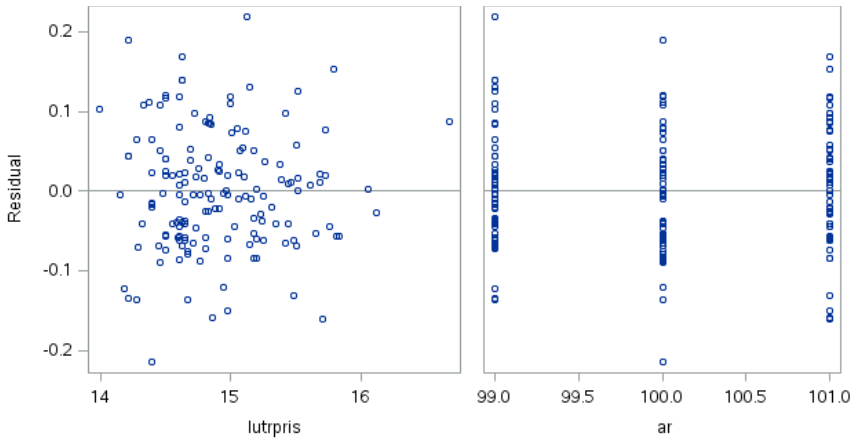
$$slutpris = e^{-2.03} * 0.92723(logaritmerat\ utropspris) * e^{0.03248(\text{år})}$$

Och för denna modell får vi följande plottar:



Figur 4.9 Residualplottar för modell med logaritmerat utropspris och år som förklarande variabler och logaritmerad slutpris som responsvariabel.

Plottarna ovan är tillfredställande enligt resonemangen vi förde ovan, för att dra slutsatsen om att modellen är godtagbar. Vidare tittar vi på varje förklarande variabels residualplottar.



Figur 4.10 Residualplottar för de förklarande variablerna logaritmerat utropspris och år

Vi kan konstatera att plottarna följer linjärt samband. Modellen har förklaringsgrad på 0.9707 och justerad förklaringsgrad på 0.9704 med standardavvikelse 0.005.

4.2 Modellval

Vi repeterar kort vad vi gjort i analysen ovan. Vi började med att göra en analys på olika modeller utan utropspris som förklarande variabel och kom fram till två lämpliga modeller. Därefter genomförde vi alltså samma analys men med utropspris som förklarande variabel och fick fram en lämplig modell.

När vi genomförde analysen ovan utan utropspris ingående som förklarande variabel kom vi fram till följande två modeller:

$R^2 = 0.8773$
$\overline{R^2} = 0.8716$
std = 0.15643

Parameter	**Skattning**
Intercept	0.6090
Yta	0.01191
Vattenavstånd	−0,0589
År	0.1349
Område 1	0.5516
Område 2	0.3599
Område 3	0.0562
Område 4	0.2398

R^2=0.8868
$\overline{R^2} = 0.8815$
std=0.1503

Parameter	Skattning
Intercept	−1.27
Log yta	0.7522
Vattenavstånd	−0.06480
År	0.1307
Område 1	0.6216
Område 2	0.4063
Område 3	0.0513
Område 4	0.2419

Tabell 4.9 Till vänster har vi modell med logaritmerat slutpris som responsvariabel och yta som förklarande variabel. Till höger har vi modell med logaritmerat slutpris samt logaritmerat yta som förklarande variabel.

Vi står nu inför ett modellval mellan dessa två modeller som båda i sig är lämpliga, men vi väljer modellen till höger om figuren eftersom förklaringsgraden är högre samt att standardavvikelsen är lägre än modellen som inte har logaritmerat yta som förklarande variabel.

Slutligen har vi alltså landat på två olika modeller för vidare analys av prediktion dvs. modellen ovan med logarimerat yta som förklarande variabel samt modellen med logaritmerat utropspris som förklarande variabel som vi illustrerar nedan.

R^2 =0.9707
$\overline{R^2}$ = 0.9704
std=0.005

Variabel	*Skattning*
Intercept	-2.03
Log utropspris	0.9273
År	0.03248

Tabell 4.10 Modell med logarimterat slutpris som respons variabel samt logaritmerat utropspris och år som förklarande variabler.

4.3 Prediktion med slutgiltiga modeller

Till detta avsnitt har vi hittat slutpriser för lägenheter år 2016 i de olika områdena. Vi har nu två modeller att utvärdera och det vi ska utvärdera är hur bra dessa två modeller predikterar det verkliga slutpriset på en lägenhet. De två modeller vi har att arbeta med illusteras nedan:

$$slutpris = e^{-1.27} * e^{0.7522(logaritmerat\ yta)} * e^{0.1307(\text{år})} * e^{(-0.06480)(vattenavstånd)} * e^{0.6216(område\ 1)} * e^{0.4063(område\ 2)} * e^{0.0513(område\ 3)} * e^{0.2419(område\ 4)} * e^{\varepsilon}$$

samt

$$slutpris = e^{-2.03} * 0.92723\beta_1(logaritmerat\ utropspris) * e^{0.03248(\text{år})}$$

Vi skall nu sätta in våra observerade värden(x) i modellerna ovan och får då följande två tabeller:

Slutpris (y_i)	*Skattad slutpris* ($\widehat{y_i}$)	*slutpris − skattad slutpris* ($y_i - \widehat{y_i}$)
650	732.9648	−82.9648
525	612.2787	−87.2787
291	357.8378	-66.8378
760.5	761.1605	−0.6605
475	442.0097	28.4902
340	337.0188	2.9811
206	298.2233	−92.2233
330	352.6211	−22.6211
530	452.0808	77.9191
460	441.5032	18.4967

Slutpris(y_i)	*Skattad slutpris*($\widehat{y_i}$)	*slutpris − skattad slutpris* ($y_i - \widehat{y_i}$)
650	694.6274	−44.6274
525	603.3529	−78.3529
291	308.1329	−17.1329
760.5	801.4472	−40.9472
475	498.8594	−28.3594
340	364.9991	−24.9991
206	238.7646	−32.7646
330	364.9991	−34.9991
530	531.5847	−1.8547
460	531.8547	−71.8547

Tabell 4.11 Slutpriser, skattad slutpris med modell utan utropspris som förklarande variabel samt differensen mellan slutpris och skattad slutpris. Priset är angivet i tiotusentalskronor.

Tabellen ovan till vänster är predikterade slutpriser för modellen utan utropspris som förklarande variabel och tabellen till höger modellen med utropspris ingående som förklarande variabel, priserna är angivna i tiotusentalskronor.

Pressvärdet för modellen utan utropspris som förklarande variabel blev $3.52*10^{12}$ och för modellen med utropspris ingående som förklarande variabel är press värdet $1.89*10^{12}$. I predktionssyfte är alltså modellen med utropspris den bästa och vi drar slutsatsen att utropspris är en viktig förklarande variabel då man vill prediktea ett framtida försäljningspris.

5 Resultat

Vi har kommit fram till två lämpliga modeller. En modell med utropspris och en utan.

5.1 Modell utan utropspris

Variabel	*Skattning*	*P-värde*
Intercept	-1.27	0.38
Logaritmerat yta	0.7522	<0.001
Vattenavstånd	-0.06480	0.0042
Område 1	0.6216	<0.001
Område 2	0.4063	<0.001
Område 3	0.0513	0.31
Område 4	0.2419	0.0002
År	0.1307	0.0491

Tabell 5.1 Modell med logaritmerat slutpris som responsvariabel och logaritmerat yta, vattenavstånd område 1-4 samt år som förklarande variabler.

Förklaringsgraden R^2 för denna modell är 0,8868 och justerade förklaringsgraden är 0.8815. Standardavvikelsen för modellen är 0.1503 som vi nämnt ovan.

Modellen är på följande form:

$$\ln(slutpris) = \\ -1.27 + 0.7522\ln\big((yta)\big) + 0.1307(\text{år}) + (-0.06480)(vattenavstånd) + \\ 0.6216(område\ 1) + 0.4063(område\ 2) + 0.0513(område\ 3) + 0.2419(område\ 4) + \varepsilon$$

Med multiplikativa faktorer.

$$slutpris = e^{-1.27} * e^{0.7522(logaritmerad\ yta)} * e^{0.13072(\text{år})} * e^{(-0.06480)(vattenavstånd)} \\ * e^{0.6216(område\ 1)} * e^{0.4063(område\ 2)} * e^{0.0513(område\ 3)} * e^{0.2419(område\ 4)} * e^{\varepsilon}$$

Med e^{ε} som är lognormalfördelad med väntevärde 0 och varians $0{,}1503^2$.

Vi skall nu tolka varje skattning i denna modell:

Logaritmerat yta skattningen för denna variabel är $e^{0.7522}$=2.12 dvs. om ytan ändras med faktorn *k* så ändras priset med en faktor $e^{0.7522}$=2.12.

År skattningen för variabeln är $e^{0.13072}$=1.13, variabeln ger alltså förändring i slutpris med en multiplikativ faktor på 1.13 för varje år.

Vattenavstånd skattningen för denna variabel är $e^{-0.06480}$=0.93. Skattningen har alltså negativ effekt på slutpriset och säger att för varje km längre bort vi kommer från vattnet ger det en sänkning i slutpris med multiplikativ faktor på 0.93.

Område 1 dvs. lägenheter i Östermalm, skattningen för denna dummy variabel är $e^{0.6216} = 1.86$ och *område 2* dvs. lägenheter på Södermalm skattningen för denna dummy variabel är $e^{0.4063} = 1.50$ och för *område 3* har vi skattningen $e^{0.0513} = 1.05$ och för *område 4* är skattningen $e^{0.2419}$=1.27. När vi nu tolkar dessa skattningar skall vi tolka de relativt område 5 eftersom vi satte område 5 till referens. Vi tolkar alltså skattningen för exempelvis område 1 som vi fick till 1.86, skattningen innebär alltså att slutpriset är med en multiplikativ faktor på 1.86 högre i område 1 jämfört med lägenheter i område 5 dvs. lägenheter i Älvsjö.

5.2 Modell med utropspris

Den andra modellen med utropspris som förklarande variabel så fick vi följande modell:

Variabel	*Skattning*	*P-värde*
Intercept	-2.03	0.0064
Logaritmerat utropspris	0.92723	<0.001
År	0.03248	<0.001

Tabell 5.2 Modell med logaritmerat slutpris som responsvariabel och utropspris ingående som förklarande variabel.

Modellen har förklaringsgrad på 0.9707 och justerad förklaringsgrad på 0.9704 med standardavvikelse 0.005 som vi nämnt ovan i analysdelen.

$$\ln(slutpris) = -2.03 + 0.92723 * ln(utropspris) + 0.00119(\text{År})$$

Med multiplikativa faktorer:

$$slutpris = e^{-2.03} * e^{0.92723(logaritmerat\ utropspris)} * e^{0.00119(\text{år})}$$

Där e^{ϵ} är lognormalfördelad med väntevärde 0 och varians 0.0056^2.

Vi tolkar varje skattning på samma sätt som ovan för denna modell.

Logaritmerat utropspris skattningen för denna variabel är 0, 92723 dvs. om utropspriset förändras med en enhet *k* så ändras slutpriset med en faktor $e^{0.92723}$=2.52.

År skattningen för denna variabel är $e^{0.00119} = 1.0011$ dvs. för varje år ökar lägenhetslutpriset med en faktor på 1.0011.

När det gäller resultatet av prediktionen så var den modellen med utropspris bäst på att prediktera framtida bostadsrättsslutpriset vilket också var väntat.

6 Diskussion

I detta arbete ville vi finna en bra modell som beskrev lägenhetspriser i fem olika områden runt om i Stockholm. Vi hade med ett antal förklarande variabler som vi trodde påverkade slutpriset på lägenheterna, de variablerna vi valde att ha med i arbetet var boarean, område, avstånd till vatten, byggår, året då lägenheten såldes, utropspriset, hiss, rum och våning. Variabler som inte blev signifikanta vid regressionen var byggår, hiss och våning. Men vi måste komma ihåg att dessa inte blev signifikanta vid regression i vårt datamaterial, logiskt sätt bör t.ex. våning ha en positiv påverkan på slutpriset men i och med att vi hade lägenheter som inte var belägna så högt upp så blev denna variabel inte signifikant. Variabler som blev signifikanta var då boarean som har stor inverkan på slutpriset och påverkade slutpriset positivt vilket i sin tur är väldigt rimligt då folk är beredda på att betala mer desto större lägenheten är.

Att finna en bra modell och en modell som är bra på att prediktera framtida bostadspriser är egentligen en svår uppgift vi började med att ta fram två lämpliga modeller som beskrev lägenhetslutpriset i vårt material därefter hittade vi tio nya observationer 2016 och predikterade slutpriset med de två modeller vi valt ut. Det finns alltså många fler förklarande variabler som man bör ha med i regressionen för att finna en modell som predikterar lägenhetspriser bättre än de vi hade med i vår regression exempelvis bör man ta hänsyn till ekonomin dvs. inflation, deflation, befolkningsmängd, ränta hos banken, lönenivåer osv. Dessutom hade vi endast 157 lägenheter att finna en bra modell med så få observationer är inte heller att föredra då man vill finna en bra modell och en modell som dessutom predikterar framtida slutpriser. Däremot kan vi dra slutsatsen att modellen håller för att beskriva lägenhetspriserna i vårt utvalda datamaterial som bestod av de 157 lägenheterna.

7. Referenser

[1] Tyrcha. J & Andersson. P. *Econometrics*. Kompendium Stockholms universitet, 2014.

[2] Sundberg. R. *Lineära statistiska modeller*. Kompendium Stockholms universitet, 2015.

[3] https://booli.se

[4] Aguirre. C. *Analys av lägenhetspriser i Hammarby Sjöstad med multipel linjär regression*. Självständigt arbete Stockholms universitet, 2015.

[5] Flodström. A. *Prediktion av lägenhetspriser i Stockholm- En statistisk undersökning*. Självständigt arbete Stockholms universitet, 2009.

[6] Björck. Å. *Numerical methods for least squares problems*. Philadelphia: SIAM, 1996.