Stockholms universitet

Analys av lägenhetspriser i Hammarby Sjöstad med multipel linjär regression

Christian Aguirre

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2015:17
Matematisk statistik
Juni 2015

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Matematisk statistik
Stockholms universitet
Kandidatuppsats **2015:17**
http://www.math.su.se/matstat

Analys av lägenhetspriser i Hammarby Sjöstad med multipel linjär regression

Christian Aguirre*

Juni 2015

Sammanfattning

I det här examensarbetet undersöker vi vilka variabler som påverkar priset på en lägenhet i Hammarby Sjöstad. Vi använder oss av multipel linjär regression för att skapa en lämplig modell som hjälper oss med detta. Det visar sig dock att en linjär modell inte är lämplig för vårt syfte, däremot anses en multiplikativ modell vara lämpligare. Detta hanteras med hjälp av transformation av responsvariabeln som överför modellen till en linjär modell. Undersökningen börjar med 10 variabler som vi tror påverkar priset för att därefter, med hjälp av olika metoder för stegvis variabelselektion, successivt minska antalet variabler till 6 där vi behåller alla som visar sig vara signifikanta. Föga förvånande visar det sig att storleken på lägenheten har en stor påverkan på priset men även andra variabler såsom antal rum, våningsplan och område visar sig ha en viss påverkan. Vi kan även konstatera att priserna har stigit under perioden 03/09-2012 till 31/01-2015, som är alltså perioden lägenheterna i datamaterialet såldes.

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige.
E-post: chrissalva82@gmail.com. Handledare: Jan-Olov Persson, Maria Deijfen.

Innehåll

1 Introduktion

Bostadssituationen i Stockholm är något som ofta diskuteras och många verkar vara överens om att bostadsbristen är stor. Med långa bostadsköer ses ett köp av en lägenhet som ett alternativ. Priserna kan dock bli väldigt höga vilket gör det svårt för många att köpa sig en lägenhet. En intressant fråga som dyker upp är -vad bestämmer priset på en lägenhet? I det här examensarbetet ska vi koncentrera oss på att undersöka vad som påverkar priset på en lägenhet i Hammarby Sjöstad. Modellen vi ska använda oss av är multipel linjär regression, en statistiskt modell som används för att påvisa linjära samband mellan en responsvariabel och flera förklaringsvariabler. Vi kommer att analysera ett datamaterial bestående av 420 lägenhetsförsäljningar i Hammarby Sjöstad under perioden 03/09-2012 och 31/01-2015. I avsnitt 2 ska vi gå igenom den teori som är relevant för det här examensarbetet. Vidare ska vi i avsnitt 3 beskriva datamaterialet i detalj samt skapa och välja variabler vars påverkan på slutpriset ska analyseras närmare. I avsnitt 4 ska vi ta fram några olika modeller och välja den som vi anser vara mest lämplig för vårt syfte. Resultatet ska vi gå igenom i avsnitt 5 där vi ska titta närmare på alla signifikanta/icke-signifikanta variabler samt slutmodellen. I avsnitt 6 avslutar vi med lite diskussion och förslag på hur man kan förbättra den här undersökningen.

2 Teori: Multipel linjär regression[1]

Multipel linjär regression är en statistisk modell som används för att påvisa om en responsvariabel beror (eller beskrivs) linjärt av flera förklarande variabler. I det här examensarbetet ska vi använda oss av multipel linjär regression för att undersöka vilka variabler som påverkar eller beskriver slutpriset på en lägenhet i Hammarby Sjöstad.

2.1 Modellen

Den allmänna formen av en multipel linjär regressionsmodell definieras som:

$$Y_i = \beta_0 + x_{1i}\beta_1 + x_{2i}\beta_2 + ... + x_{mi}\beta_m + \varepsilon_i, \qquad i = 1, 2, ..., N \tag{2.1}$$

I uttrycket ovan är $\{Y_i,\ \ i = 1, ..., N\}$ en uppsättning av N oberoende stokastiska variabler vars observerade värden vi betecknar med $\{y_i,\ \ i = 1, ..., N\}$. Uppsättningen $\{\beta_j,\ \ j = 0, ..., m\}$ består av $k = m + 1$ okända parametrar och uppsättningen $\{x_{ji},\ \ j = 1, ..., m,\ i = 1, ..., N\}$ består av $m * N$ kända tal (värdet på den j:te förklarande variabeln för den i:te observationen). Variablerna $\{\varepsilon_i = Y_i - (\beta_0 + x_{1i}\beta_1 + x_{2i}\beta_2 + ... + x_{mi}\beta_m),\ \ i = 1, ..., N\}$ kallas för feltermer och är stokastiska variabler. I den här modellen antar vi att $\{\varepsilon_i,\ \ i = 1, ..., N\}$ är normalfördelade med väntevärde 0 och konstant varians σ^2.

Ett annat sätt att skriva (2.1) är att använda sig av matris- och vektornotation:

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

där

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{pmatrix}, \qquad \boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{m1} \\ 1 & x_{12} & \cdots & x_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1N} & \cdots & x_{mN} \end{pmatrix}, \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_m \end{pmatrix}, \qquad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{pmatrix}$$

Antagandet om fördelningen av $\{\varepsilon_i,\ \ i = 1, ..., N\}$ kan uttryckas som:

$$\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I}) \tag{2.2}$$

2.2 Parameterskattningar

I en multipel linjär regressionsmodell anger parametern β_j hur mycket den j-te förklarande variabeln, som vi betecknar med X_j, påverkar responsvariabeln. Då parametervektorn $\boldsymbol{\beta}$ är okänd måste vi skatta den ur data. För att skatta $\boldsymbol{\beta}$ använder vi oss av minsta-kvadratmetoden. Idén med minsta-kvadratmetoden är att hitta parametervektorn $\boldsymbol{\beta}$ som minimerar summan av den kvadratiska skillnaden mellan observationerna och deras väntevärden, dvs vi vill hitta $\{\beta_j\}$

som minimerar

$$\sum_{i=1}^{N} \big(y_i - \sum_{j=0}^{m} x_{ji}\beta_j\big)^2, \qquad \text{där } x_{0i} = 1 \text{ för alla } i \tag{2.3}$$

Vi kan skriva om (2.3) med hjälp av matris- och vektornotation:

$$(\boldsymbol{y} - \boldsymbol{X\beta})^T(\boldsymbol{y} - \boldsymbol{X\beta}) = \|\boldsymbol{y} - \boldsymbol{X\beta}\|^2 \tag{2.4}$$

Att minimera (2.4) är detsamma som att minimera $\|\boldsymbol{y} - \boldsymbol{X\beta}\|$. Det visar sig att parametervektorn som minimerar $\|\boldsymbol{y} - \boldsymbol{X\beta}\|$ är:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

Innan vi har observerat vektorn $\boldsymbol{y}$ ser vi dock $\hat{\boldsymbol{\beta}}$ som en stokastisk variabel då den beror på $\boldsymbol{Y}$. Vi har alltså:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} \tag{2.5}$$

Det är av intresse att veta vad (2.5) har för väntevärdesvektor och varians-kovariansmatris. Genom att utnyttja antagande (2.2) kan man visa följande:

$$E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta} \tag{2.6}$$

$$Var(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1} \tag{2.7}$$

Då variansen σ^2 är okänd skattar vi den ur data. En väntevärdesriktig skattning av σ^2 ges av:

$$\hat{\sigma}^2 = \frac{1}{N-k}\|\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2 \tag{2.8}$$

2.3 Hypotesprövning och p-värde

Då vi vill undersöka om variabeln X_j påverkar responsvariabeln, är det av intresse att testa hypotesen att β_j är skild från noll. Hypotesen kan uttryckas på följande sätt:

$$\begin{aligned} H_0 &: \beta_j = 0 \\ H_a &: \beta_j \neq 0 \end{aligned} \tag{2.9}$$

För att testa hypotes (2.9) använder vi oss av ett t-test. Om antagande (2.2) är uppfyllt följer det av (2.6) och (2.7) att $\hat{\beta}_j$ är normalfördelad med väntevärde β_j och varians $(\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})_{jj}$, där $(\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})_{jj}$ är det j-te diagonalelementet i matrisen (2.7). Eftersom σ^2 är okänd kan vi inte använda oss av normalfördelningen för att testa hypotesen, däremot kan vi använda oss av (2.8)

och t-fördelningen genom att utnyttja följande:

$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}\sqrt{((\boldsymbol{X}^T\boldsymbol{X})^{-1})_{jj}}} \sim t(N-k) \tag{2.10}$$

Med hjälp av (2.10) kan vi testa hypotes (2.9). Detta gör vi genom att utnyttja att teststatistikan $T_j = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{((\boldsymbol{X}^T\boldsymbol{X})^{-1})_{jj}}}$, under H_0 är t-fördelad med $(N-k)$ frihetsgrader. Eftersom t-fördelningen är symmetrisk kring 0 kommer stora och små värden på T_j att tala emot H_0, vilket leder till att vi förkastar H_0. Vilka värden som T_j måste anta för att H_0 ska förkastas beror på vilken signifikansnivå α vi väljer. I det här examensarbetet ska vi välja signifikansnivån α till 0.05, dvs vi kommer att förkasta H_0 om $|T_j| > t_{0.025}(N-k)$.

När vi ska välja vilka variabler som ska ingå i vår modell ska vi välja variabler vars koefficienter är statistiskt signifikanta skilda från 0, dvs vi kommer att välja X_j om vi kan förkasta hypotesen $H_0 : \beta_j = 0$. Detta kommer vi att göra med hjälp av p-värdet för X_j. Om vi låter T_{j0} beteckna det observerade värdet på vår teststatistika T_j, definieras p-värdet för X_j som sannolikheten att $|T_j| \geq |T_{j0}|$ då H_0 är sann. Matematiskt kan detta utryckas på följande sätt:

$$p\text{-värde} := P_{H_0}(|T_j| \geq |T_{j0}|)$$

Variabeln X_j kommer att ingå i modellen om p-värdet $\leq \alpha = 0.05$.

2.4 Dummy-variabler

Dummy-variabler är variabler som endast kan anta värdet 0 eller 1. Dessa variabler tillåter kvalitativa variabler att ingå i regressionen då de kan användas för att ange förekomsten eller avsaknaden av ett attribut. Om en kvalitativ variabel endast kan anta två värden, t ex kön, kan vi skapa en dummy-variabel som antar värdet 1 om det är en kvinna och värdet 0 om det är en man. I fallet då en kvalitativ variabel kan anta fler än två värden kan vi använda oss av flera dummy-variabler för att ange värdet på den. Som exempel kan vi ta en kvalitativ variabel som kan anta tre olika värden, röd, gul och svart, då kan vi skapa två dummy-variabler för att ange värdet på den. Den första dummy-variabeln skulle kunna anta värdet 1 om färgen är röd och värdet 0 annars. Den andra dummy-variabeln skulle kunna anta värdet 1 om färgen är gul och 0 annars. Båda dummy-variabler antar värdet 0 om färgen är svart. I det här fallet använder vi färgen svart referens. Det går givetvis att använda en annan färg som referens. På samma sätt kan vi göra om vi har en kvalitativ variabel som kan anta fler än tre värden. Om en kvalitativ variabel kan anta n olika värden behöver vi $(n-1)$ dummy-variabler för att ange värdet på den.

2.5 R^2 och R^2_{adj}

Förklaringsgraden, som betecknas R^2, är ett mått på hur väl de förklarande variablerna i en linjär modell förklarar variationen av responsvariabeln och de-

finieras på följande sätt:

$$R^2 := \frac{Kvs(regression)}{Kvs(totalt)} = 1 - \frac{Kvs(residual)}{Kvas(Totalt)} = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

där $\hat{y}_i = \hat{\beta}_0 + x_{1i}\hat{\beta}_1 + x_{2i}\hat{\beta}_2 + ... + x_{mi}\hat{\beta}_m$ och $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$.

Förklaringsgraden kan användas för att jämföra olika modeller och antar värden mellan 0 och 1 där ett högre värde är att föredra. En nackdel med R^2 är att den alltid ökar då vi tillför fler variabler i modellen. I praktiken används därför ett annat mått, R^2_{adj}, som definieras på följande sätt:

$$R^2_{adj} := 1 - \frac{\hat{\sigma}^2}{\hat{\sigma}^2_0}$$

där $\hat{\sigma}^2_0$ är skattningen av variationen då vi inte har några förklarande variabler i modellen. R^2_{adj} mäter hur mycket variansen minskar i modellen jämfört med modellen då vi inte har några förklarande variabler.

2.6 Stegvis variabelselektion

När man försöker anpassa en multipel regressionsmodell är det vanligt att man samlar in en stor uppsättning av möjliga förklarande variabler. Dock är det sällan så att alla dessa variabler behövs för att få en tillfredställande modell. Det är därför viktigt att kunna välja ut en mindre uppsättning av möjliga förklarande variabler som ger oss den "bästa" modellen. Det finns flera metoder på hur man kan gå tillväga för att hitta just en sådan uppsättning. Vi ska titta närmare på tre olika metoder för stegvis variabelselektion som vi kommer att använda i det här examensarbetet.

Backward elimination
Den här proceduren gör en multipel regression med samtliga variabler och plockar sedan bort den variabeln med högst p-värde. Sedan fortsätter proceduren med resterande variabler. Proceduren slutar när alla kvarvarande variabler visar sig vara signifikanta.

Forward selection
Den här proceduren utgår från inga variabler och lägger sedan till en variabel i taget. I varje steg lägger proceduren till den variabeln med lägst p-värde vid inkludering i modellen. Proceduren fortsätter tills det inte finns fler signifikanta variabler att lägga till.

Stepwise selection
Den här proceduren är en blandning av föregående procedurer. Precis som forward selection lägger proceduren till en variabel i taget men med skillnad att den efter varje inkludering testar att alla variabler som redan inkluderats fortfarande är signifikanta. Om det visar sig att någon variabel inte längre är signifikant plockas den bort.

2.7 Transformation av variabler

När man hanterar verklig data stöter man ofta på problemet att sambandet mellan responsvariabel och de förklarande variablerna inte är linjär. Detta kan man i vissa fall hantera med hjälp av en lämplig transformation av variablerna och på så sätt få ett linjärt samband. En sådan transformation kan till exempel vara att logaritmera antingen responsvariabel (2.9), de förklarande variablerna (2.10) eller både responsvariabeln och de förklarande variablerna tillsammans (2.11).

$$log(Y_i) = \beta_0 + x_{1i}\beta_1 + ... + x_{mi}\beta_m + \varepsilon_i, \qquad i = 1, ..., N \tag{2.9}$$

$$Y_i = \beta_0 + log(x_{1i})\beta_1 + ... + log(x_{mi})\beta_m + \varepsilon_i, \qquad i = 1, ..., N \tag{2.10}$$

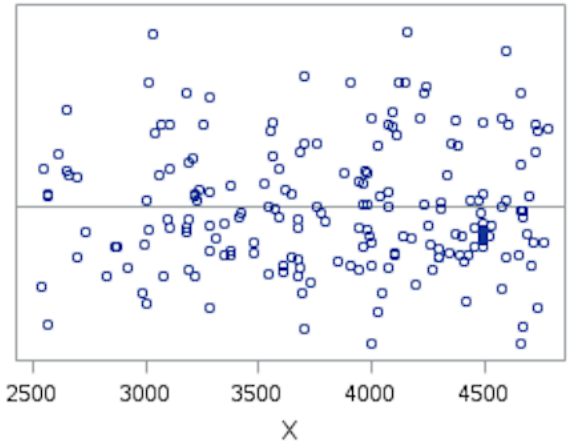
$$log(Y_i) = \beta_0 + log(x_{1i})\beta_1 + ... + log(x_{mi})\beta_m + \varepsilon_i, \qquad i = 1, ..., N \tag{2.11}$$

I (2.11) kan man även välja ut en delmängd av de förklarande variablerna som ska logaritmeras. Under modellframtagandet kommer vi att testa dessa transformationer för att ta fram en så bra modell som möjligt.

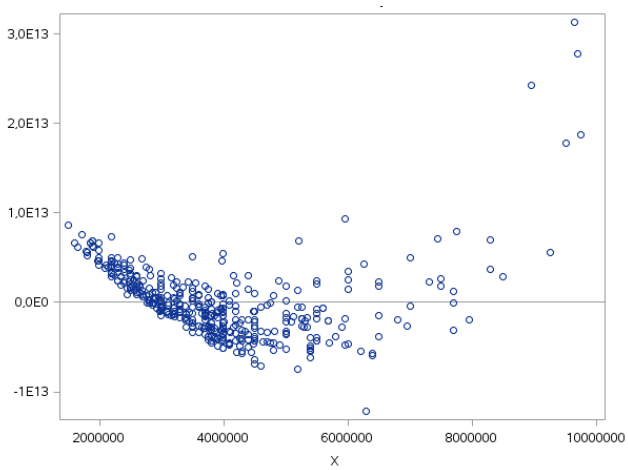
2.8 Modellvalidering

Det finns en mängd olika plottar man bör titta närmare på för att undersöka om den valda modellen är lämplig. I det här avsnittet ska vi gå igenom några plottar som är relevanta för det här examensarbetet.

Så som modell (2.1) är definierad är väntevärdesfunktionen av responsvariabeln, linjärt i parametrarna $\{\beta_j, \ \ j = 0, ..., m\}$. För att undersöka lineariteten ska vi titta närmare på residualerna plottade mot varje förklarande variabel. Om det finns ett linjärt samband mellan responsvariabeln och X_j, förväntar vi oss att residualerna ska vara jämnt utspridda runt 0. I Figur 1 kan vi se ett exempel på en plott som tyder på ett linjärt samband och i Figur 2 kan vi se ett exempel på en plott som tyder på ett icke-linjärt samband.



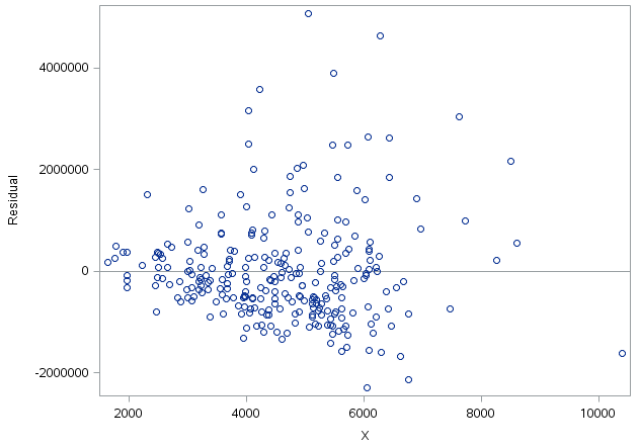
Figur 1: Plott av residualer mot X_j, linjärt samband.



Figur 2: Plott av residualer mot X_j, icke-linjärt samband.

Antagandet om feltermernas fördelning (2.2) är mycket viktigt i modell (2.1). De hypotesprövningar vi gör hänger till stor del på det här antagandet. Om antagandet är uppfyllt borde residualerna vara approximativt normalfördelade med väntevärde 0. För att undersöka detta ska vi plotta en normalfördelningsplott av residualerna där punkterna i plotten borde approximativt följa en rät linje.

Vi vill även undersöka om feltermerna har konstant varians. Detta kan vi göra genom att plotta residualerna (vi kan även välja att först standardisera residualerna genom att dela dessa med sina skattade standardavvikelse) mot varje förklarande variabel eller mot de skattade värdena. Precis som när vi testar lineariteten förväntar vi oss en plott som ser ut som i Figur 1. I Figur 3 kan vi se ett exempel på residualer som tyder på icke-konstant varians.

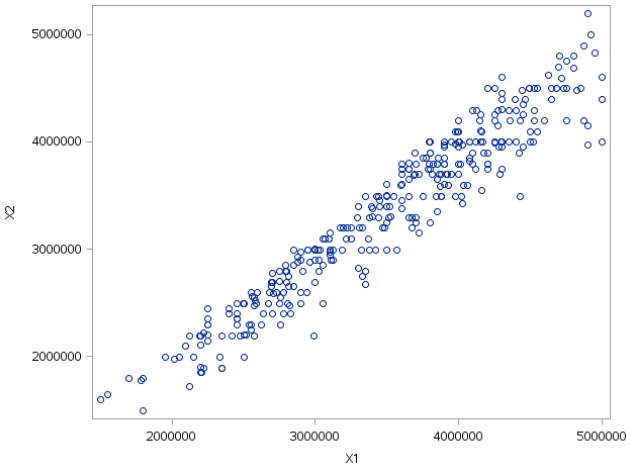


Figur 3: Exempel på residualer som tyder på icke-konstant varians.

En annan sak som är viktig att undersöka är om vi råkat ut för multikollinearitet. Multikollinearitet innebär att det finns approximativt linjära samband mellan förklarande variabler, dvs korrelation mellan variabler. Detta kan leda till att variabler som man förväntar sig bli signifikanta inte blir det, vilket inte är önskvärt.
Ett sätt att upptäcka multikollinearitet är att beräkna VIF-värdet för alla X_j. VIF står för *Variance Inflation Factor*. VIF-värdet för X_j beräknas på följande sätt:

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

där R_j^2 är förklaringsgraden i en multipel linjär regression med X_j som responsvariabel och och övriga X-variabler som förklarande variabler. VIF-värdet uttrycker hur mycket större variansen av $\hat{\beta}_j$ blir i en multipel linjär regression jämfört med variansen av $\hat{\beta}_j$ med endast X_j som förklarande variabel. Hur högt ett VIF-värde ska vara för att det ska anses vara ett problem är svårt att säga men enligt Rolf Sundbergs kompendium, *Linjära Statistiska Modeller* brukar en sådan gräns sättas vid 5 eller 10. I det här examensarbetet ska vi välja ett VIF-värde på 5 som gräns.
Ett annat sätt att upptäcka multikollinearitet är att plotta de olika förklarande variablerna mot varandra för att på så sätt upptäcka eventuella linjära samband. I Figur 4 kan vi se ett exempel på två variabler med ett tydligt linjärt samband.



Figur 4: Plott av två variabler med ett tydligt linjärt samband.

3 Datamaterial

Datamaterialet samlades in från *booli.se*[2] och består av försäljningar av lägenheter i Hammarby Sjöstad mellan perioden 03/09-2012 och 31/01-2015. Datamaterialet bestod från början av 619 observationer med 10 variabler som beskrivs närmare nedan. Värdet på variabeln *Trappa* saknades dock i flera observationer vilket gjorde att alla 619 observationer inte kunde användas i analysen. Med hjälp av data från *slutpris.se*[4] kunde värdet på *Trappa* i vissa fall kompletteras. Efter kompletteringen består datamaterialet av 420 observationer med följande variabler:

Slutpris. Priset lägenheten såldes för. Under den aktuella perioden såldes lägenheterna för mellan 1 500 000 och 10 350 000 kronor med ett medelvärde på ungefär 4 080 000 kronor.

Boarea. Storleken på lägenheten som i datamaterialet varierar mellan 27 och 164 m^2 med ett medelvärde på ungefär 76 m^2.

Trappa. Våningsplanet lägenheten är belägen i. Den här variabeln varierar mellan 1 och 11 där ungefär 56% av lägenheterna är belägna på våningsplan 3 eller under.

Byggår. Året då lägenheten byggdes. Byggåret för lägenheterna i datamaterialet varierar mellan 2000 och 2013.

Antal rum. Antal rum lägenheten är uppdelad i. Värdet på den här variabeln varierar mellan 1 och 7 där den största delen av lägenheterna, ungefär 78%, är uppdelade i 3 rum eller mindre.

Avgift. Månadsavgift som betalas till bostadsrättsföreningen. Månadsavgiften i datamaterialet ligger mellan 1 356 och 10 398 kronor med ett medelvärde på ungefär 4 570 kronor.

Mäklare. Namnet på mäklarfirman som sålde lägenheten. Totalt var det 22 olika mäklarfirmor som sålde lägenheter under den aktuella perioden. Trots många mäklarfirmor står 2 mäklarfirmor för över 50% av försäljningar av lägenheterna i datamaterialet.

Säljår. Året då lägenheten såldes, den här variabeln varierar mellan 2012 och 2015.

Adress. Adressen lägenheten är belägen i.

Period. Perioden mellan 03/09-2012 och 31/01-2015 är uppdelat i 10 olika perioder. Uppdelningen har gjorts årstidsvis där period 1 är hösten 2012, period 2 är vintern 2012/2103 osv till period 10 som är vintern 2014/2015.

3.1 booli.se och slutpris.se

booli.se är en internettjänst som tillhandahåller bland annat information om slutpriser på bostäder och beskriver sig som oberoende då de inte ägs av en bank eller mäklarfirma. Deras metod att samla in information beskriver de såhär[3]:

"Slutprisinformationen om bostadsrätter bygger på en strukturerad automatisk insamling av sista buden från öppna budgivningar på nätet. Det betyder att många slutpriser finns med men inte alla. Alla mäklarbyråer presenterar inte slutpriser öppet på sajten och även om en byrå brukar göra det så kan säljaren alltid välja att inte visa budgivningen".

Tyvärr kan slutpriser på vissa objekt vara felaktiga då informationen samlas in automatiskt, vilket kan vara ett problem. *booli.se* förklarar problemet på följande sätt[3]:

"Vi hämtar slutpriser för villor, tomter och lägenheter direkt från de öppna budgivningarna på mäklarens hemsida. Utslaget på en större mängd objekt är träffsäkerheten mycket hög, men på enskilda objekt kan indikationen skilja sig mer från det faktiska slutpriset. Vi är medvetna om att denna information inte alltid stämmer med det faktiska slutpriset. "

slutpris.se är också en internettjänst som tillhandahåller information om slutpriser på bostäder och beskriver sig också som oberoende utan kopplingar till branschorganisationer eller enskilda mäklarbyråer. Även *slutpris.se* har samma problem med felaktiga slutpriser och beskriver detta såhär[5]:

"Uppgifter om lägenheter på slutpris.se bygger på den information som mäklare tillhandahåller och publicerar i lägenhetsprospekt. slutpris.se kan därmed inte garantera att uppgifterna för varje enskilt objekt är korrekta. I undantagsfall kan t ex en lägenhet av misstag publiceras som såld på slutpris.se, eller visa ett pris som avviker från den slutliga köpeskillingen. Slutpris.se ansvarar inte för några följder av sådana eventuella fel eller avvikelser."

Att både *booli.se* och *slutpris.se* kan ha felaktiga uppgifter om slutpriser är ett problem, detta diskuteras vidare i avsnitt 6.

4 Statistisk modellering

4.1 Bearbetning av data

Ur datamaterialet ska vi skapa nya variabler i syfte att hitta en modell som på bästa sätt förklarar slutpriset på en lägenhet i Hammarby Sjöstad.

Ålder = (*Säljår*−*Byggår*). Den här variabeln ska undersöka om åldern på lägenhet påverkar slutpriset.

Kvadratmeter per rum=(*Boarea*/*Antal rum*). Vi misstänker att *Boarea* och *Antal rum* är korrelerade. För att hantera detta problem skapar vi den här variabeln.

Avgift per kvadratmeter = (*Avgift*/*Boarea*). Vi misstänker även att *Boarea* och *Avgift* är korrelerade och skapar därför även den här variabeln.

Mäklarfirmorna Fastighetsbyrån och Svenska Mäklarhuset står för 18.33% respektive 32.86% av alla försäljningarna i datamaterialet (tredje största mäklarfirman står för 10.24%). Tillsammans står de alltså för mer än hälften av alla försäljningar och det kan därför vara intressant att undersöka om slutpriset påverkas av att man väljer en av dessa mäklarfirmor. För att kunna undersöka detta skapar vi tre dummy-variabler.

Fastighetsbyrån = 1 om lägenheten har sålts av Fastighetsbyrån, 0 annars.

Svenska Mäklarhuset = 1 om lägenheten har sålts av Svenska Mäklarhuset, 0 annars.

Övriga Mäklarfirmor = 1 om lägenheten har sålts av andra mäklarfirmor än Fastighetsbyrån och Svenska Mäklarhuset, 0 annars.

Övriga Mäklarfirmor, kommer att fungera som referens.

Vi vill även kunna undersöka om området där lägenheten är belägen i påverkar slutpriset och skapar därför tre dummy-variabler som hjälper oss att undersöka just detta. Vi delar in Hammarby Sjöstad i 3 olika områden (se Figur 5). *Område 1* och *Område 2* är områden som ligger nära vattnet medan *Område 3* har lite längre avstånd till vattnet. Tyvärr blev inte uppdelningen optimalt då endast gatunamnet noterades under datainsamlingen men inte gatunumret vilket gjorde att vissa gator inte gick att dela på. Mer om detta i avsnitt 6.

Område 1 = 1 om lägenheten ligger i området 1, 0 annars.

Område 2 = 1 om lägenheten ligger i området 2, 0 annars.

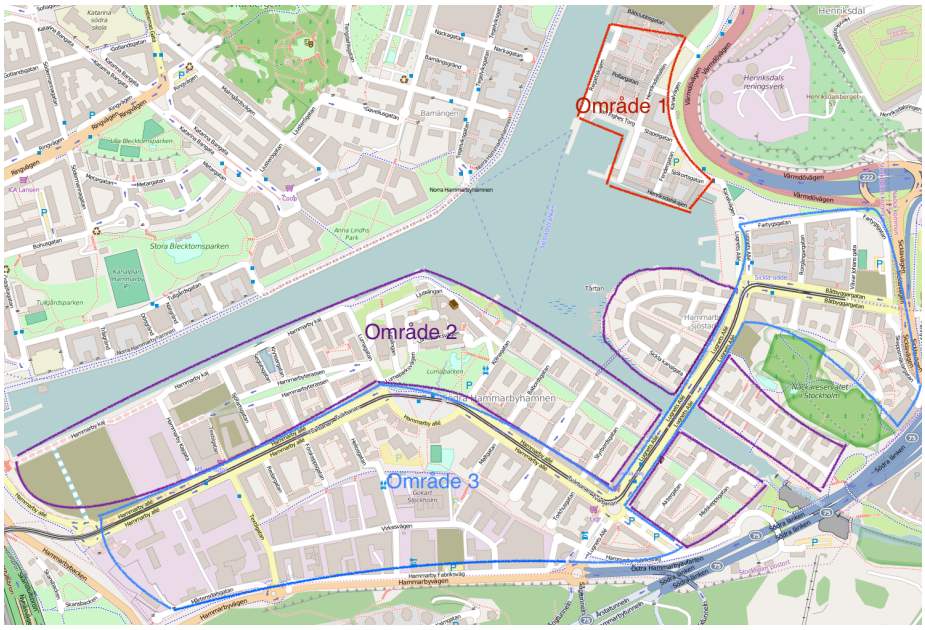
Område 3 = 1 om lägenheten ligger i området 3, 0 annars.

Område 3 kommer att fungera som referens.

Det kommer att visa sig att en modell med transformerade variabler är lämpligare för att beskriva sambandet mellan slutpris och våra förklarande variabler och skapar därför även följande variabler:

$logSlutpris = log(Slutpris)$.

$logBoarea = log(Boarea)$.



Figur 5: Karta[6] över Hammarby Sjöstad uppdelat i tre olika områden.

4.2 Val av variabler

Som vi tidigare nämnt är syftet med det här examensarbetet att undersöka vilka variabler som påverkar eller beskriver slutpriset på en lägenhet i Hammarby Sjöstad. Vi kommer att utgå från variablerna som beskrevs i avsnitt 3 och 4.1 och från dessa välja ut variablerna som vi har anledning att tro påverkar slutpriset. Nedan följer en lista på variablerna vi har valt ut.

Boarea. Det verkar rimligt att anta att storleken på lägenheten påverkar slutpriset och därför väljer vi att ha med den här variabeln i analysen.

Trappa. Det finns anledning att tro att våningsplanet lägenheten är belägen i påverkar slutpriset. Det kan till exempel vara så att en fin utsikt över området (som delvis beror på våningsplanet) kan vara något många är ute efter.

Antal rum, *Kvadratmeter per rum.* Det kan vara intressant att undersöka om antal rum påverkar slutpriset. Det kan vara så att man är benägen att betala mer för en lägenhet med fler rum, givet en viss boarea. Som vi nämnde tidigare

misstänker vi att variablerna *Boarea* och *Antal rum* är korrelerade och eftersom detta kan leda till att vi drar felaktiga slutsatser (se avsnitt 2.8) kommer vi att välja den variabeln som är minst korrelerad med *Boarea*.

Avgift, *Avgift per kvadratmeter*. Hur stor avgiften som betalas till bostadsrättsföreningen skulle kunna påverka slutpriset och därför tar vi med en av dessa variabler i analysen. Vilken av dessa variabler vi använder beror på korrelationen med *Boarea*.

Ålder. Lägenheterna i datamaterialet är upp till 15 år gamla, vi har därför ingen större anledning att tro att åldern påverkar slutpriset. Vi väljer dock att ha med den här variabeln för att övertyga oss om detta.

Fastighetsbyrån, *Svenska Mäklarhuset*. Som vi nämnde i avsnitt 4.1 står dessa två mäklarfirmor för över 50 % av alla lägenhetsförsäljningar i datamaterialet och därför är det intressant att undersöka om dessa mäklarfirmor påverkar slutpriset.

Period. Den här variabeln ska hjälpa oss att undersöka om priset på en lägenhet har påverkats över tid.

Område 1, *Område 2*. Till slut vill vi även undersöka om området lägenheten är belägen i påverkar slutpriset.

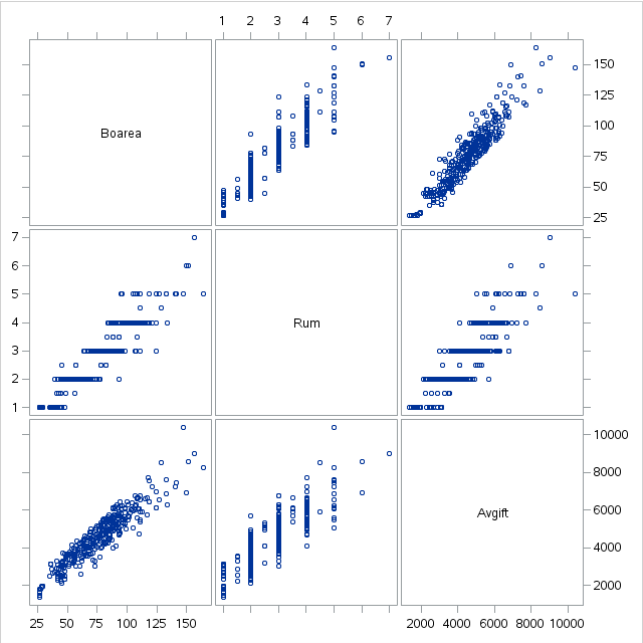
4.3 Analys av data

Det första vi ska göra är att undersöka hur det ligger till med korrelationen mellan våra förklarande variabler. Som vi nämnde i avsnitt 4.1 misstänker vi att det finns en korrelation mellan variablerna *Boarea*, *Antal rum*, och *Avgift*. Det kan även finnas korrelation mellan andra variabler och därför kommer vi att börja med att göra en multipel regression med *Boarea*, *Antal rum*, *Avgift*, *Trappa*, *Ålder*, *Fastighetsbyrån*, *Svenska Mäklarhuset*, *Period*, *Område 1* samt *Område 2* som förklarande variabler och titta närmare på VIF-värdet för dessa variabler. Resultatet sammanställs i Tabell 1.

Variabel	VIF**-värde**
Boarea	14.77316
Antal rum	6.10566
Avgift	9.07377
Trappa	1.13336
Ålder	1.66263
Fastighetsbyrån	1.14940
Svenska Mäklarhuset	1.25250
Period	1.09190
Område 1	1.68785
Område 2	1.29662

Tabell 1: VIF-värde för möjliga förklarande variabler.

Från Tabell 1 kan vi se att VIF-värdet för *Boarea*, *Antal rum* och *Avgift* är högre än 5 (gränsen vi satte i avsnitt 2.8). Vi undersöker korrelationen mellan dessa variabler närmare genom att plotta variablerna mot varandra. Resultatet visas i Figur 6.

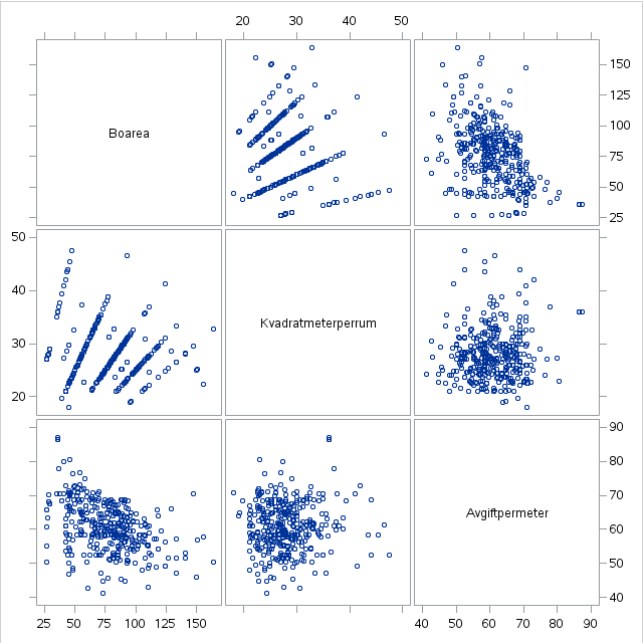


Figur 6: Variablerna *Boarea*, *Antal rum* samt *Avgift* plottade mot varandra.

Från Figur 6 kan vi se ett tydligt linjärt samband mellan *Boarea*, *Antal rum* och *Avgift*. Då detta inte är önskvärt väljer vi att ersätta *Antal rum* med *Kvadratmeter per rum* och *Avgift* med *Avgift per kvadratmeter*. Innan vi går vidare måste vi dock undersöka hur det ligger till med korrelationen efter att vi ersatt med dessa variabler. Precis som tidigare gör vi en multipel regression och tittar närmare på VIF-värdet samt plottar. Resultatet visas i Tabell 2 och i Figur 7.

Variabel	VIF**-värde**
Boarea	1.31587
Kvadratmeter per rum	1.02996
Avgift per kvadratmeter	1.38946
Trappa	1.12941
Ålder	1.66082
Fastighetsbyrån	1.14876
Svenska Mäklarhuset	1.25249
Period	1.08898
Område 1	1.67791
Område 2	1.27781

Tabell 2: VIF-värde för möjliga förklarande variabler.

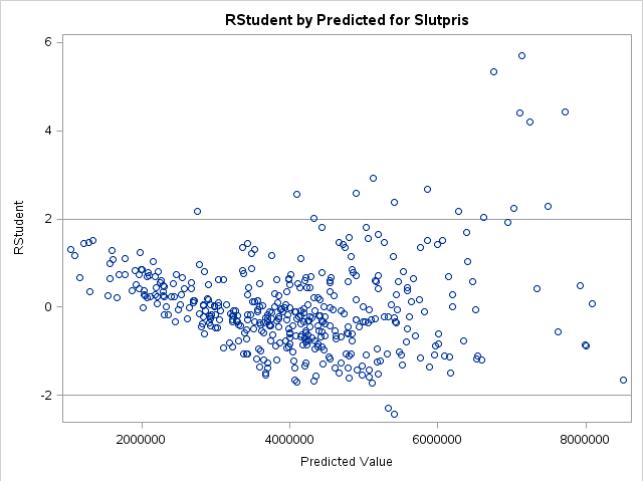


Figur 7: Variablerna *Boarea*, *Kvadratmeter per rum* och *Avgift per kvadratmeter* plottade mot varandra.

Från Tabell 2 och Figur 7 drar vi slutsatsen att korrelationen mellan våra förklarande variabler inte längre är ett problem. Vi kan nu gå vidare med analysen och utgår från variablerna i Tabell 2 som möjliga förklarande variabler och *Slutpris* som responsvariabel.

Under analysen ska vi använda oss av samtliga metoder för stegvis variabelselektion som beskrevs i avsnitt 2.6. Det visar sig att samtliga metoder ger samma resultat. Variablerna som visar sig vara statistiskt signifikanta (på 5% signifikansnivå) är *Boarea*, *Kvadratmeter per rum*, *Trappa*, *Period*, *Område 1* samt

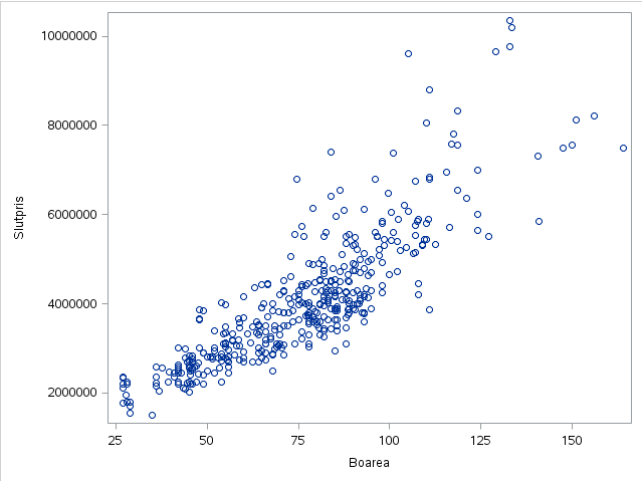
Område 2.
Innan vi tittar på resultatet av parameterskattningarna ska vi ta en närmare titt på residualerna som vi får ut, detta eftersom vi inte kan dra några korrekta slutsatser om modellen inte är lämplig.
I avsnitt 2.8 gav vi exempel på hur residualplottarna borde se ut för att modellen ska anses vara lämplig. Nedan visas en residualplott som vi ska titta närmare på.



Figur 8: Standardiserade residualer mot skattade värden. *Slutpris* som responsvariabel, *Boarea*, *Kvadratmeter per rum*, *Trappa*, *Period*, *Område 1* samt *Område 2* som förklarande variabler.

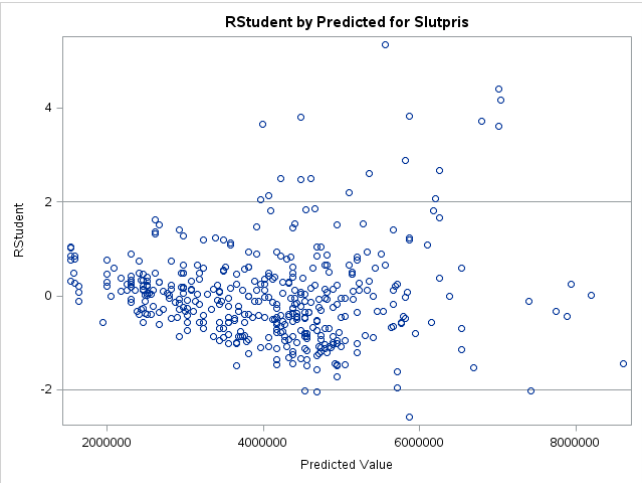
Residualerna i Figur 8 ser inte ut att vara jämnt utspridda runt 0. Först verkar residualerna följa ett nedåtgående mönster för att sedan bli väldigt utspridda. Dessutom ligger flera punkter väldigt långt ifrån 0. Detta kan bero på att vi inte har ett linjärt samband mellan *Slutpris* och våra förklarande variabler, eller att variansen inte är konstant. Detta är som vi tidigare nämt inte önskvärt då vi inte kan anse modellen vara lämplig.

För att hantera detta ska vi testa att transformera våra variabler (se avsnitt 2.7). Vilka variabler vi ska transformera är inte givet från början men en idé är att titta närmare på sambandet mellan *Slutpris* och den variabeln som förklarar ”mest” av slutpriset, dvs variabeln med högst förklaringsgrad vid en enkel regression. Med hjälp av *Forward selection* kan vi konstatera att *Boarea* ensamt förklarar mest av alla variabler i Tabell 2 med en förklaringsgrad på 0.7181. I Figur 9 kan vi se sambandet mellan *Slutpris* och *Boarea*.



Figur 9: Plott av *Slutpris* mot *Boarea*.

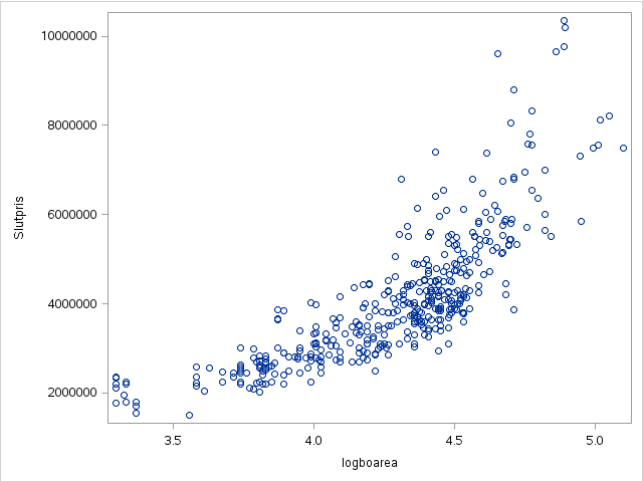
Från Figur 9 kan vi se att sambandet mellan *Slutpris* och *Boarea* inte verkar vara riktigt linjärt. För att få bättre förståelse över sambandet mellan dessa variabler gör vi nu en enkel regression med *Slutpris* som responsvariabel och *Boarea* som enda förklarande variabel och tittar närmare på plotten av standardiserade residualer mot skattade värden.



Figur 10: Standardiserade residualer mot skattade värden, *Slutpris* som responsvariabel och *Boarea* som förklarande variabel.

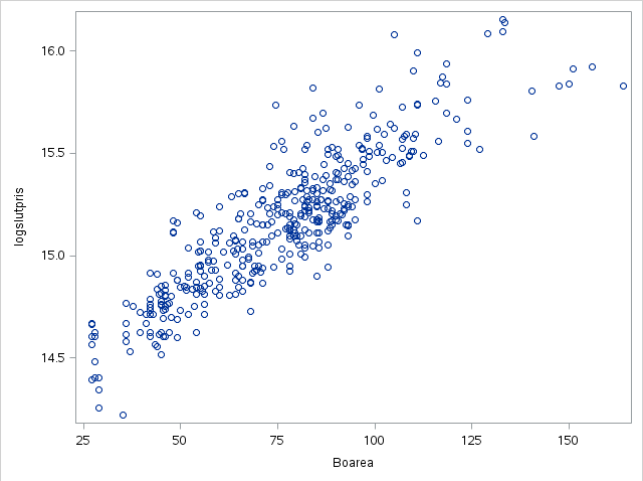
Vi ser en vis likhet mellan residualplottarna i Figur 8 och Figur 10, detta får oss att misstänka att en transformation (logaritmering) av *Slutpris*, *Boarea* eller båda samtidigt skulle resultera i en lämpligare modell. Vi tittar närmare på

sambandet mellan kombinationer av *logSlutpris*/*Slutpris* och *logBoarea*/*Boarea* för att upptäcka eventuella linjära samband och på så sätt bestämma vilka variabler vi väljer att transformera.

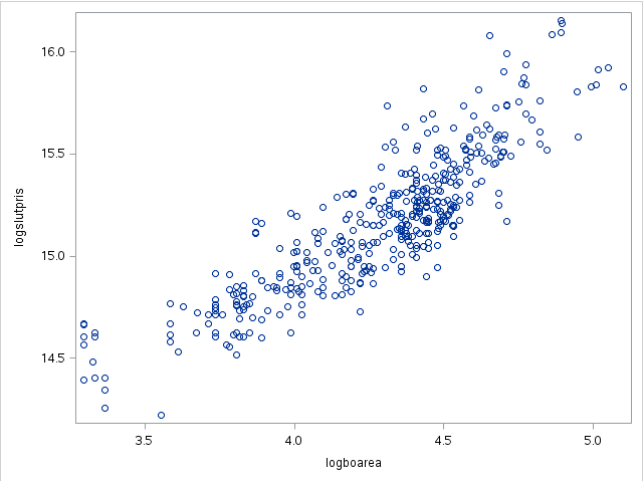


Figur 11: Plott av *Slutpris* mot *logBoarea*.

Från Figur 11 ser vi att sambandet mellan *Slutpris* och *logBoarea* inte verkar vara linjärt. Den här kombinationen av variabler väljer vi därför bort i den fortsatta analysen.



Figur 12: Plott av *logSlutpris* mot *Boarea*.



Figur 13: Plott av *logSlutpris* mot *logBoarea*.

Från Figur 12 och Figur 13 kan vi se ett tydligare linjärt samband mellan *logSlutpris* och *Boarea* respektive *logBoarea* och väljer därför att titta närmare på dessa kombinationer av variabler i den fortsätta analysen.

Vi ska alltså titta närmare på två olika modeller med *logSlutpris* som responsvariabel. Variablerna vi kommer att utgå ifrån i respektive modell visas i Tabell 3.

Modell 1	**Modell 2**
Boarea	*logBoarea*
Kvadratmeter per rum	*Kvadratmeter per rum*
Avgift per meter	*Avgift per meter*
Trappa	*Trappa*
Ålder	*Ålder*
Fastighetsbyrån	*Fastighetsbyrån*
Svenska Mäklarhuset	*Svenska Mäklarhuset*
Period	*Period*
Område 1	*Område 1*
Område 2	*Område 2*

Tabell 3: Möjliga förklarande variabler i Modell 1 respektive Modell 2.

4.3.1 Modell 1

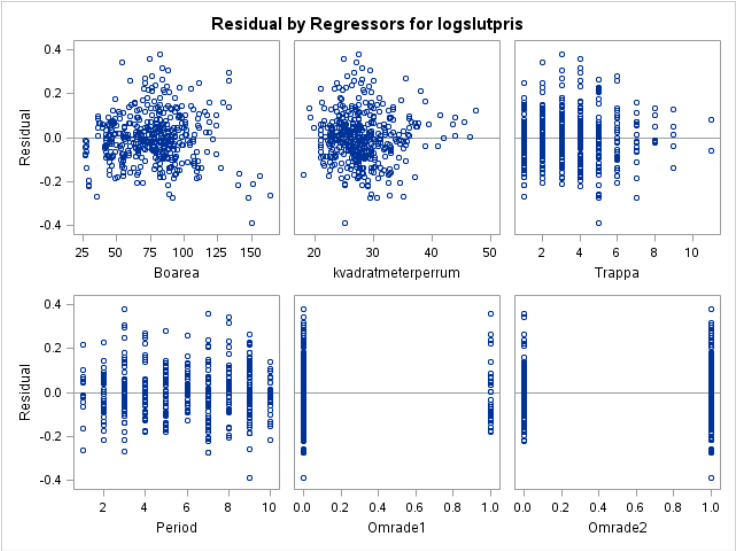
Vi börjar med Modell 1 där vi utgår från samma variabler som listades i Tabell 2. Vi har redan konstaterade att korrelationen mellan dessa variabler inte är ett problem och kan därför gå vidare utan att oroa oss för multikollinearitet. Även

här använder vi oss av samtliga metoder för stegvis variabelselektion som alla ger samma resultat av signifikanta variabler, detta listas i Tabell 4.

Variabler
Boarea
Kvadratmeter per rum
Trappa
Period
Område 1
Område 2

Tabell 4: Signifikanta variabler i Modell 1.

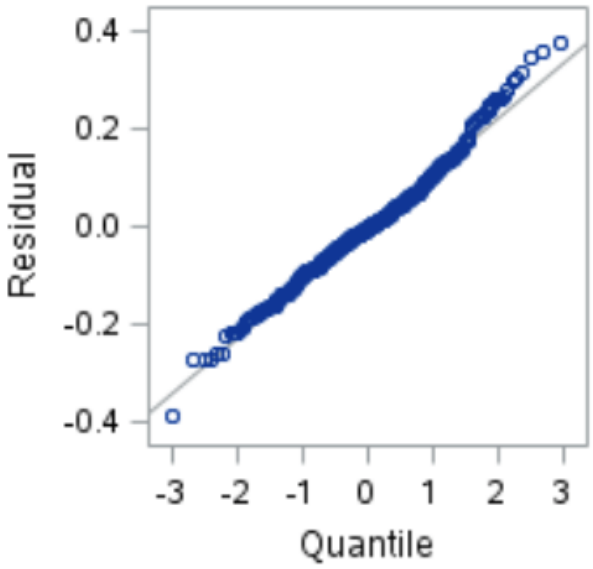
Nästa steg är att undersöka om modellen är lämplig. För att göra detta ska vi titta närmare på de residualplottar som vi får ut.



Figur 14: Residualplottar för samtliga variabler i Tabell 4.

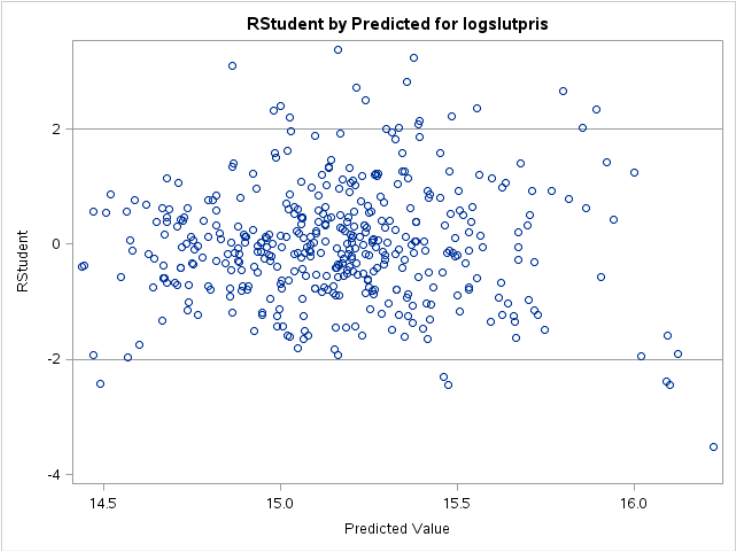
I Figur 14 kan vi se residualerna plottade mot varje förklarande variabel som visade sig vara signifikant i Modell 1. Vi kan inte se några tecken som tyder på ett icke-linjärt samband mellan *logSlutpris* och dessa variabler. Residualerna ser nämligen jämnt utspridda runt 0.

Figur 15 nedan ska hjälpa oss att undersöka antagandet om feltermernas fördelning (2.2). Vi ser att punkterna i Figur 15 ser ut att approximativt följa en rät linje, vilket är precis det vi önskar.



Figur 15: Normalfördelningsplott, Modell 1.

Slutligen ska vi undersöka om feltermerna har konstant varians.



Figur 16: Standardiserade residualer mot skattade värden, Modell 1.

Från Figur 16 ser vi inga tecken som tyder på icke-konstant varians. Vi kan alltså dra slutsatsen att variablerna i Tabell 4 ger oss en lämplig modell. Vi kan nu gå vidare och titta på bland annat parameterskattningar och förklaringsgrad

för Modell 1. Resultatet sammanställs i Tabell 5.

$\hat{\sigma}$=0.11336	R^2=0.8901	R^2_{adj}=0.8885
Variable	**Parameterskattning**	**p-värde**
Intercept	14.09838	< 0.001
Boarea	0.01201	< 0.001
Kvadratmeter per rum	-0.00676	< 0.001
Trappa	0.02500	< 0.001
Period	0.03095	< 0.001
Område 1	0.23315	< 0.001
Område 2	0.09053	< 0.001

Tabell 5: Parameterskattningar i Modell 1.

4.3.2 Modell 2

Skillnaden mellan Modell 1 och Modell 2 är att vi ersätter *Boarea* med *logBoarea* som möjlig förklarande variabel. Det som gör att vi är intresserade av att undersöka Modell 2, trots att vi redan hittat en lämplig modell, är att vi får ett mer intuitivt samband mellan *Slutpris* och *Boarea* när vi transformerar tillbaka variablerna, där priset går mot 0 när boarean går mot 0.
logBoarea är en variabel som vi hittills inte har använt och måste därför börja med att undersöka korrelationen med övriga variabler. Den här gången nöjer vi oss med att endast titta på VIF-värdet för att avgöra om korrelation mellan variablerna är ett problem. En multipel regression ger oss resultatet som visas i Tabell 6.

Variable	**VIF-värde**
logBoarea	1.31716
Kvadratmeter per rum	1.02999
Avgift per kvadratmeter	1.39032
Trappa	1.12869
Ålder	1.67117
Fastighetsbyrån	1.15050
Svenska Mäklarhuset	1.24960
Period	1.08879
Område 1	1.67758
Område 2	1.27827

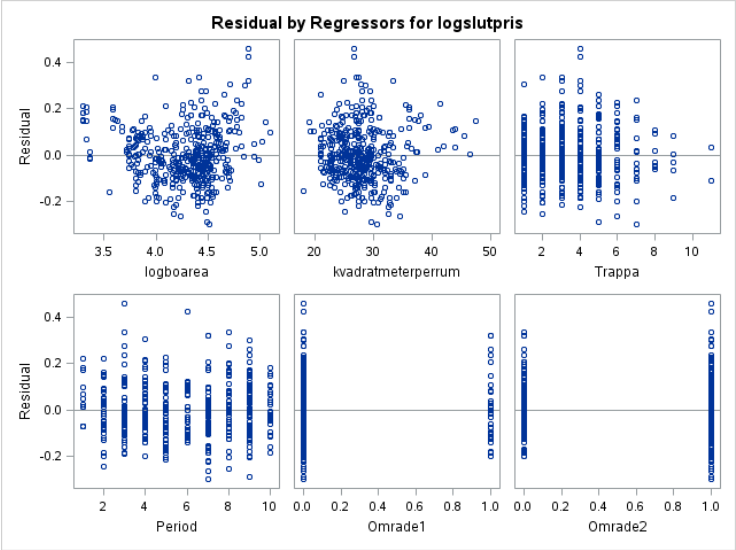
Tabell 6: VIF-värdet för möjliga förklarande variabler i Modell 2.

Från Tabell 6 kan vi konstatera att VIF-värdena är så pass låga att vi även här drar slutsatsen att korrelation mellan variablerna inte är ett problem. På sedvanligt sätt ska vi nu avgöra vilka variabler som visar sig vara signifikanta. Resultatet visas i Tabell 7.

Variable
logBoarea
Kvadratmeter per rum
Trappa
Period
Område 1
Område 2

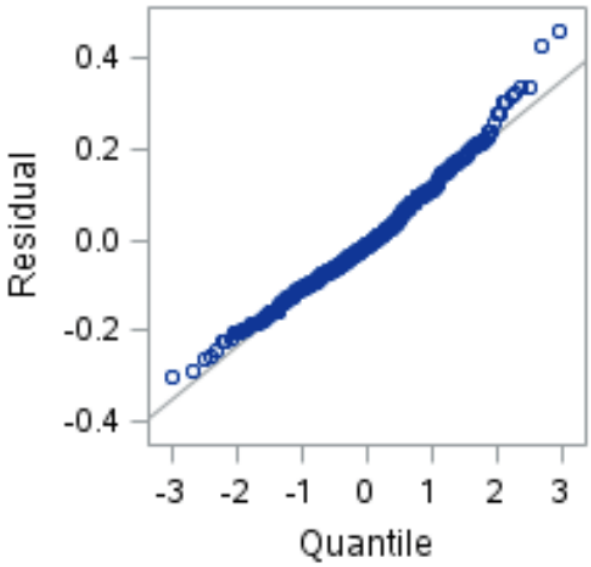
Tabell 7: Signifikanta variabler i Modell 2.

Även här måste vi undersöka om modellen anses vara lämplig. Vi gör på samma sätt som tidigare och tittar närmare på residualplottarna vi får ut.



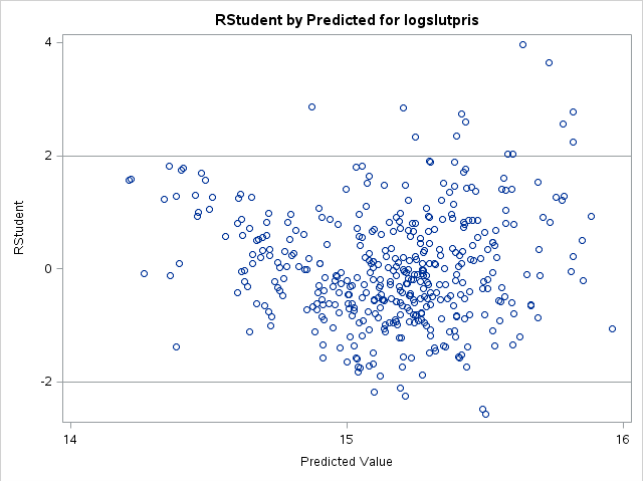
Figur 17: Residualplottar för samtliga variabler i Tabell 7.

Från Figur 17 kan vi vi inte se några tecken på icke-linjärt samband mellan *logSlutpris* och variablerna i Tabell 7. Därmed kan vi konstatera att *logSlutpris* är linjär i parametrarna. Vidare ska vi undersöka hur det ligger till med fördelningen av feltermerna.



Figur 18: Normalfördelningsplott, Modell 2.

Från Figur 18 ser vi att punkterna i normalfördelningsplotten verkar approximativt följa en rät linje. Vi kan därför dra slutsatsen att antagandet om feltermernas fördelning är uppfyllt. Vidare ska vi undersöka om feltermerna har konstant varians.



Figur 19: Standardiserade residualer mot skattade värden, Modell 2.

Från Figur 19 ser vi att residualerna är utspridda runt 0 dock med några ensta-

ka punkter med långt avstånd från 0. Speciellt ser vi två punkter väldigt nära 4. Då vi tittar på standardiserade residualer är punkter nära 4 väldigt osannolika men då vi har 420 punkter drar vi ändå slutsatsen att även Modell 2 är lämplig. I Tabell 8 tittar vi närmare på bland annat parameterskattningar och förklaringsgrad.

$\hat{\sigma}$=0.11830	R^2=0.8803	R^2_{adj}=0.8785
Variable	**Parameterskattning**	***p*-värde**
Intercept	11.41630	< 0.001
logBoarea	0.83814	< 0.001
Kvadratmeter per rum	-0.00689	< 0.001
Trappa	0.02654	< 0.001
Period	0.03088	< 0.001
Område 1	0.23511	< 0.001
Område 2	0.10699	< 0.001

Tabell 8: Parameterskattningar i Modell 2.

4.3.3 Val av modell

Vi har nu landat i två olika modeller som beskriver variationen av *logSlutpris* på ett tillfredställande sätt. Frågan vi ställer oss nu är vilken av dessa modeller vi ska välja. Det enda som skiljer modellerna åt är att vi använder variablerna *Boarea* och *logBoarea* i respektive modell. Som vi nämnde i avsnitt 4.3.2 så har Modell 2 ett mer intuitivt samband mellan *Slutpris* och *Boarea* då slutpriset går mot 0 då boarean går mot 0, vilket inte är fallet i Modell 1. Då vi strävar efter att hitta modellen som på bästa sätt förklarar variationen i *Slutpris* är ett högt värde på förklaringsgraden, R^2, väldigt viktigt när vi ska välja modell.
Vi börjar med att undersöka hur mycket respektive variabel (*Boarea* och *logBoarea*) ensamt förklarar *logSlutpris*. Vi gör därför en enkel regression med *Boarea* respektive *logBoarea* som enda förklarande variabel. Resultatet visas i Tabell 9.

Variable	R^2
Boarea	0.772
logBoarea	0.758

Tabell 9: Förklaringsgrad för modellerna med *Boarea* respektive *logBoarea* som enda förklarande variabel och *logSlutpris* som responsvariabel.

Vi ser att förklaringsgraden är högre för modellen med *Boarea*. Från Tabell 5 och Tabell 8 ser vi även att modellen med *Boarea* ihop med övriga variabler har en högre förklaringsgrad (0.8901 mot 0.8803) än modellen med *logBoarea* ihop övriga variabler. Trots att Modell 1 inte ger oss ett samband där *Slutpris* går mot 0 då *Boarea* går mot 0 väljer vi Modell 1 då vi anser Modell 1 vara giltig i intervallet där vi har observationer.

5 Resultat

I det här avsnittet ska vi titta närmare på resultatet av våra parameterskattningar. Vi ska bland annat undersöka rimligheten i våra parameterskattningar. I avsnitt 4 kom vi fram till följande Resultat:

$\hat{\sigma}$=0.11336	R^2=0.8901	R^2_{adj}=0.8885
Variable	**Parameterskattning**	**p-värde**
Intercept	14.09838	< 0.001
Boarea	0.01201	< 0.001
Kvadratmeter per rum	-0.00676	< 0.001
Trappa	0.02500	< 0.001
Period	0.03095	< 0.001
Område 1	0.23315	< 0.001
Område 2	0.09053	< 0.001

Tabell 10: Parameterskattningar i Modell 1.

Eftersom vi använder *logSlutpris* som responsvariabel får vi följande samband mellan *logSlutpris* och våra förklarande variabler:

$$\begin{aligned} logSlutpris = {} & 14.09838 + 0.01201(Boarea) - 0.00676(Kvadratmeter\ per\ rum) + \\ & + 0.02500(Trappa) + 0.03095(Period) + 0.23315(Område\ 1) + \\ & + 0.09053(Område\ 2) + \varepsilon \end{aligned} \tag{5.1}$$

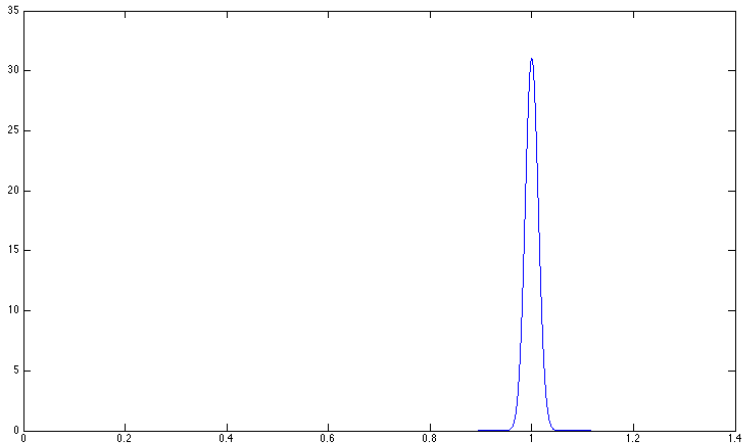
där $\varepsilon \sim N(0, \sigma^2)$.

(5.1) ger oss följande multiplikativa samband mellan *Slutpris* och våra förklarande variabler:

$$\begin{aligned} Slutpris = {} & e^{14.09838} * e^{0.01201(Boarea)} * e^{-0.00676(Kvadratmeter\ per\ rum)} * \\ & * e^{0.02500(Trappa)} * e^{0.03095(Period)} * e^{0.23315(Område\ 1)} * \\ & * e^{0.09053(Område\ 2)} * \varepsilon' \end{aligned} \tag{5.2}$$

där ε' är lognormal fördelat med parametrar 0 och σ^2.

Från Tabell 10 ser vi att den skattade standardavvikelsen $\hat{\sigma}$ är lika med 0.11336, vilket ger oss den skattade variansen $\hat{\sigma}^2 = 0.0128504896$. I Figur 20 ser vi täthetsfunktionen för en lognormalfördelad variabel med parametrar 0 och 0.0128504896.



Figur 20: Täthetsfunktion för en lognormalfördelad variabel med parametrar 0 och 0.0128504896.

Samband (5.2) säger att en ökning på en enhet av den j-te förklarande variabeln ger en förändring i *Slutpris* med en multiplikativ faktor på e^{β_j}, där β_j är parameterskattningen av den j-te förklarande variabeln.

5.1 Signifikanta variabler

Boarea. Att den här variabeln blev signifikant är inget som förvånar, inte heller det faktum att den ensam förklarar en stor del av slutpriset. Som vi såg av samband (5.2) ger en ökning på en m^2 av *Boarea* en ökning av slutpriset med en multiplikativ faktor på $e^{0.01201} = 1.01208$.

Kvadratmeter per rum. Med den här variabeln ville vi undersöka om antalet rum, givet en viss boarea, påverkar priset. Det visar sig att den här variabeln har en negativ effekt på priset då $e^{-0.00676} = 0.99326$. Det kan vara svårt att förstå hur den här variabeln egentligen påverkar priset och därför följer ett exempel nedan.
Låt oss säga att vi har två lägenheter med lika stor boarea på 100 m^2. Låt vidare lägenheterna ha 4 respektive 5 rum vardera. Variabeln *Kvadratmeter per rum* kommer att vara $\frac{100}{4} = 25$ respektive $\frac{100}{5} = 20$, dvs slutpriset på lägenheterna kommer att ändras med en multiplikativ faktor på $e^{-0.00676*25} = e^{-0.00676*(20+5)} = e^{-0.00676*20} * e^{-0.00676*5}$ respektive $e^{-0.00676*20}$. Slutpriset på en lägenhet med 4 rum kommer alltså att skilja sig från slutpriset på en lägenhet med 5 rum med en multiplikativ faktor på $e^{-0.00676*5} = 0.96676$. Den här variabeln säger alltså att man betalar mer för fler antal rum, givet en viss boarea.

Trappa. Redan tidigare misstänkte vi att den här variabeln skulle ha en positiv effekt på slutpriset. En hypotes var att utsikten, som till stor del beror på våningsplanet, ökar priset på en lägenhet. Resultatet i Tabell 10 bekräftar

våra misstankar då varje ökning av *Trappa* ökar slutpriset med en multiplikativ faktor på $e^{0.03095} = 1.03143$

Period. Vi ser att tiden har haft en positiv effekt på slutpriset då parameterskattningen blev positiv, $e^{0.03095} = 1.03143$. Exakt vad det beror på kan vi tyvärr inte svara på med hjälp av våra resultat, vi kan bara konstatera att priserna på bostadsrätterna i datamaterialet ökat med tiden.

Område 1, *Område 2.* Låt oss komma ihåg att vi använder *Område 3* som referens vilket innebär att parameterskattningar av *Område 1* och *Område 2* ska ses relativt till *Område 3.* Parameterskattningarna för dessa variabler blev positiva, 0.23315 respektive 0.09053. Detta innebär att slutpriset är högre med en multiplikativ faktor på $e^{0.23315} = 1.262457$ för område 1 respektive $e^{0.09053} = 1.094754$ för område 2 jämfört med slutpriset i område 3.

5.2 Icke-signifikanta variabler

Avgift per kvadratmeter. Med ett p-värde på 0.1982 blev den här variabeln inte signifikant. Resultatet visar att den här variabeln inte påverkar slutpriset. Lite förvånande då den här variabeln påverkar den totala månadskostnaden för lägenheten.

Ålder. Redan tidigare misstänkte vi att åldern på lägenheten inte skulle påverka slutpriset vilket resultatet bekräftade. Den här variabeln hade ett p-värde på 0.2941. Då åldern på lägenheterna i datamaterialet är upp till 15 år gamla förväntade vi oss inget annat resultat.

Fastighetsbyrån, *Svenska Mäklarhuset.* Med dessa variabler ville vi undersöka om det var någon skillnad mellan de två mäklarfirmorna som sålde mest jämfört med resten av mäklarfirmorna. Det visade sig att det inte fanns någon signifikant skillnad mellan mäklarfirmor då dessa variabler hade p-värden på 0.4323 respektive 0.3402.

6 Diskussion

I det här examensarbete har vi undersökt vilka variabler som påverkar priset på en lägenhet i Hammarby Sjöstad. I resultatet kom vi bland annat fram till att Boarea förklarar en stor del av slutpriset. Detta verkar väldigt rimligt då de flesta vill ha ett rymligt boende och betalar gärna mer för en större lägenhet. Vidare kom vi fram till att fler antal rum, givet en viss boarea, påverkar priset i positiv riktning. Detta verkar också rimligt då priserna på lägenheter i Hammarby Sjöstad i regel är rätt höga och med fler antal rum kan fler personer bo i en lägenhet. Våningsplan visade sig också ha en positiv påverkan på priset. Detta trodde vi redan före undersökningen där vi hade som hypotes att man gärna betalar för en bra utsikt, vilket i sin tur till stor del påverkas av våningsplanet. Området lägenheten är belägen i hade också en stor påverkan på priset. Vi såg att både område 2 och område 3 ökade priset på en lägenhet. Detta är dock inte så förvånande då område 3 inte har så många lägenheter nära vattnet, vilket område 2 och område 3 har. Att område 2 och område 3 påverkade olika mycket kan bero på saker som, utbud av restauranger i närheten av områdena, närhet till skola, närhet till kollektivtrafik, osv. Vi kom även fram till att priserna har ökat under perioden som lägenheterna i datamaterialet såldes. Följer man nyhetsrapporteringen om bostadsmarknaden så säger den samma sak. Tyvärr kan inte i den här undersökningen säga mer om varför priserna har stigit bara konstatera att resultatet visar oss en trend av stigande lägenhetspriser i Hammarby Sjöstad. En variabel som visade sig vara icke-signifikant var avgiften. Detta är förvånade då avgiften påverkar den totala månadskostnaden. En annan variabel som inte visade sig påverka priset var åldern på lägenheten. Som vi nämnde tidigare så är lägenheterna som undersöktes upp till 15 år gamla vilket gör resultatet rimligt då alla lägenheter fortfarande kan anses vara nya. Vi har även kommit fram till att mäklarfirmor inte skiljer sig när det gäller priser på lägenheter. Detta är egentligen inte så förvånansvärd då det verkar rimligt att köparen ser en viss lägenhet som lika mycket värd oavsett vem som säljer den.

Som vi tidigare nämnt kan vissa observationer i vårt datamaterial vara felaktiga. Detta är såklart ett problem som skapar en osäkerhet i vår modell. Osäkerheten är inte så stor om eventuella felaktiga observationer inte är så kallade outliers då dessa har en stor påverkan på resultatet. Detta är tyvärr något vi inte kan kontrollera och rätta till. För att undvika det här problemet skulle en framtida undersökning kunna hämta data direkt från mäklarfirmorna och på så sätt minimera antalet felaktiga observationer. Dock kan den här lösningen leda till att man begränsar sig till färre mäklarfirmor vilket kan leda till färre observationer. En annan sak som kan förbättras är uppdelnings av område. När datamaterialet samlades in noterades endast gatunamn men inte gatunummer vilket gjorde att långa gator inte gick att dela på. Detta resulterade till exempel i att delar som passade bättre i område 2 hamnade i stället i område 3.

7 Referenser

[1] Rolf Sundberg, Lineära Statistiska Modeller, 2014

[2] http://www.booli.se

[3] http://support.booli.se

[4] http://www.slutpris.se

[5] https://slutpris.se/om-oss

[6] http://www.openstreetmap.org/, http://www.openstreetmap.org/copyright, © OpenStreetMaps bidragsgivare