Stockholms universitet

Prediktion av lägenhetspriser i Stockholm - en statistisk undersökning

Anna Flodström

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2009:7
Matematisk statistik
September 2009

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholms universitet

Matematisk statistik
Stockholms universitet
Kandidatuppsats **2009:7**,
http://www.math.su.se/matstat

Prediktion av lägenhetspriser i Stockholm - en statistisk undersökning

Anna Flodström*

September 2009

Sammanfattning

Den här uppsatsen har till syfte att undersöka prediktionsförmågan hos lägenhetspriser i Stockholms innerstad. Med hjälp av metoder inom Prediktion och Regressionsanalys ska vi konstruera en modell som på bästa sätt kan fungera utifrån vårt syfte. Redan innan studiens början förutspår vi boareans inverkan på priset. Vår analys riktar sig mot att undersöka vilka fler variabler, utöver boarea, och i vilken sammansättning av dessa som ger den mest användbara prediktionsmodellen. Våra vidare undersökningar exkluderar vissa variabler från fortsatt analys och lämnar oss med tio stycken förklarande variabler. Vi använder oss av metoder inom Regression och Stegvis Regressions analys för att ta fram ett antal modeller för vidare undersökningar. Modellerna applicerar vi dels på hela materialet och dels på delmaterial då vi delat upp lägenheterna efter antal rum. Vidare undersökningar leder oss till att skifta responsvariabel från slutpris till logaritmerat slutpris. För att avgöra vilken av modellerna som ger det mest tillfredställande resultatet, utifrån vårt syfte, använder vi metoder inom Cross Validation. Undersökningarna resulterar i olika modeller för de olika materialen. Dock kan vi dra slutsatsen att boarean är den variabel som i olika sammansättningar med andra variabler tjänar vårt syfte mest tillfredställande.

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige. E-post: anna@flodstrom.nu. Handledare: Maria Deijfen, Anders Björkström.

Prediktion av lägenhetspriser i Stockholm - en statistisk undersökning

Kandidatuppsats 15 högskolepoäng

Anna Flodström

16 september 2009

"Prediction is very difficult, especially if it´s about the future."

- Niels Bohr. *Nobel Prize in Physics, 1922.*

Sammanfattning

Den här uppsatsen har till syfte att undersöka prediktionsförmågan hos lägenhetspriser i Stockholms innerstad. Med hjälp av metoder inom Prediktion och Regressionsanalys ska vi konstruera en modell som på bästa sätt kan fungera utifrån vårt syfte. Redan innan studiens början förutspådde vi boareans inverkan på priset. Vår analys riktar sig mot att undersöka vilka fler variabler, utöver boarea, och i vilken sammansättning av dessa som ger den mest användbara prediktionsmodellen. Våra vidare undersökningar exkluderar vissa variabler från fortsatt analys och lämnar oss med tio stycken förklarande variabler. Vi använder oss av metoder inom Regression och Stegvis Regressionsanalys för att ta fram ett antal modeller för vidare undersökningar. Modellerna applicerar vi dels på hela materialet och dels på delmaterial då vi delat upp lägenheterna efter antal rum. Vidare undersökningar leder oss till att skifta responsvariabel från slutpris till logaritmerat slutpris. För att avgöra vilken av modellerna som ger det mest tillfredställande resultatet, utifrån vårt syfte, använder vi metoder inom Korsvalidering. Undersökningarna resulterar i olika modeller för de olika materialen. Dock kan vi dra slutsatsen att boarean är den variabel som i olika sammansättningar med andra variabler tjänar vårt syfte mest tillfredställande.

Abstract

This essay is intended to examine the ability to predict apartment prices in Stockholm City. Using techniques in Prediction and Regression, we shall construct a model that can best act on the basis of our purpose. The living areas impact on the price is predicted before the beginning of the investigation. The analysis aim is to examine which more variables than the living area and in which composition of these, that gives the most useful prediction model. Our further studies exclude certain variables from further analysis, leaving us with ten variables to use in our upcoming models. We use methods in Regression and Stepwise Regression to develop a number of models for further studies. The models we apply to both the entire data and also on parts of the data which we have split by the number of rooms of the apartments. Further investigations lead us to change the response variable from the final price to the logarithm of the final price. To determine which of the models that provides the most satisfying results according to our purpose, we use the method of Cross Validation. The different studies results in different models for the various data. However, our conclusion is that the living area in combination with other variables, in different compositions, serves our purpose most satisfying.

Förord och Tack

Denna uppsats utgör ett kandidatarbete på 15 högskolepoäng och resulterar i en kandidatexamen vid institutionen för Matematisk Statistik vid Stockholms Universitet.
Jag skulle vilja uttrycka min djupa och uppriktiga tacksamhet till mina två handledare, Forskningsingenjör Anders Björkström och Docent och forskarassistent Maria Deijfen, vid Stockholms Universitet avdelningen för Matematisk Statistik. Deras rådgivning har varit ovärderlig och jag tackar för deras vägledning och support vid arbetet med denna uppsats. Jag vill även uttrycka min stora tacksamhet till Fastighetsbyrån och Tomas Edlund för möjligheten att ta del av deras lägenhetsförsäljningar.
Ytterligare vill jag delge min tacksamhet till de personer i min närhet som uppmuntrat mig under arbetets gång.

Innehåll

1 Introduktion **1**
- 1.1 Inledning . . . 1
- 1.2 Syfte och metod . . . 1

2 Bakrundsfakta **2**
- 2.1 Fastighetsbyrån . . . 2
- 2.2 Materialet . . . 2

3 Metoder **5**
- 3.1 Regression . . . 5
  - 3.1.1 Enkel linjär regression . . . 5
  - 3.1.2 Multipel linjär regression . . . 5
- 3.2 Minsta kvadrat metoden . . . 6
- 3.3 Stegvisa regressions metoder . . . 6
  - 3.3.1 Stepwise Regression . . . 6
- 3.4 Korsvalidering . . . 7
- 3.5 Prediktion . . . 9

4 Viktiga begrepp som används **11**
- 4.1 Signifikant förklaringsvariabel . . . 11
- 4.2 Förklaringsgraden . . . 11

5 Resultat **12**
- 5.1 Datamaterialet . . . 12
  - 5.1.1 Undersökning av datamaterialet . . . 12
- 5.2 Test av hypoteser rörande datamaterialet . . . 17
- 5.3 Regressionsanalys . . . 24
  - 5.3.1 Test av normalfördelningsantagande . . . 28
- 5.4 Tillämpningar . . . 31
  - 5.4.1 Tillämpningar av Leave one out-metoden, Korsvalidering . . . 31
  - 5.4.2 Alternativ responsvariabel . . . 37

6 Slutsatser **44**

7 Diskussion **49**

Referenser **51**

1 Introduktion

1.1 Inledning

Priserna på bostadsmarknaden är idag ett hett ämne som diskuteras flitigt i media. Det handlar om att sälja i rätt läge för att göra så stor vinst som möjligt på sin bostad. Det skrivs otaliga artiklar om dagsläget på bostadsmarknaden och prognoser inför framtiden. Budskapet är att ett bostadsköp är en investering. Det nya sättet att se på bostadsmarknaden har väckt intresse för olika undersökningar på området. I artikeln "Nytt index ska ge bättre koll på bostadspriser" i nätupplagan av Dagens Industri, från den 25 november 2008, läses: "Idag använder sig såväl Riksbanken, som ska sätta reporäntan, som privatpersoner, som ska fatta köp- och säljbeslut, av bostadsprisstatistik från SCB:s småhusbarometer och Mäklarstatistik."
Forskare vid KTH framtar just nu ett prisindex som syftar till att göra det lättare att få en överblick över prisutvecklingen. Målet med framtagandet av detta index är att i framtiden kunna använda det som bas för fastighetsderivat. Men en intressant fråga som denna studie inte reflekterar över, är den som alla ställer sig på bostadsmarknaden: Hur mycket är min lägenhet värd? / Hur mycket får jag betala för en viss lägenhet?. Det var när jag befann mig i just denna situation och frågade mig just dessa frågor som jag fick idén om att undersöka prediktionsförmågan av dessa priser. Boarean hos en lägenhet har uppenbart en stor inverkan på slutpriset, men vilka andra variabler kan på bästa sätt hjälpa till att uppskatta ett framtida försäljningspris?

1.2 Syfte och metod

En undersökning av bostadspriserna och av mäklarstatistik kan syfta till många olika ändamål. Huvudsyftet med denna uppsats var att få fram en modell som på bästa möjliga sätt kan användas för att prediktera framtida försäljningspriser på lägenheter. Men även att med detta syfte i åtanke, utvidga mina kunskaper inom regressionsanalys.

För att utreda skillnaderna mellan olika modeller kommer vi ta hjälp av värden på Predicted Residual SS (PRESS), förklaringsgraden, signifikansen av enskilda variabler samt antalet variabler i modellen. Uteslutande genom arbetets gång har jag använt mig av programvaran SAS.

Datamaterialet som vi kommer att undersöka består av drygt 450 lägenhetsförsäljningar under perioden 2006-2009 på Östermalm, Gärdet och Hjorthagen i Stockholms innerstad. Datamaterialet kommer från Fastighetsbyrån, Swedbank, på Östermalm i Stockholm.

2 Bakrundsfakta

2.1 Fastighetsbyrån

Vi fick möjligheten att, till denna kandidatuppsats, använda oss av mäklarstatistik från Fastighetsbyrån. Fastighetsbyrån är en del av Swedbank-koncernen och fungerar idag som en franchise-kedja. Det har de gjort sen 1999, men de grundades redan 1966. Mäklarfirman är i dagsläget de som säljer flest bostäder inom Sverige och de har 230 kontor runt om i landet. Mäklarstatistiken som vi har använt oss av kommer från Fastighetsbyrån på Östermalm i Stockholm.

2.2 Materialet

Datamaterialet togs fram ur Fastighetsbyråns datasystem med hjälp av Tomas Edlund (fastighetsmäklare/franchisetagare). Materialet innefattar större delen av Fastighetsbyråns försäljningar från januari 2006 till Mars 2009 och är lägenheter belägna i huvudsak på Östermalm, Gärdet och i närbelägna områden i Stockholms innerstad. Vi opererade efter principen att ta fram så många möjliga variabler som möjligt för att sedan utesluta faktorer som inte visade sig vara signifikanta i senare undersökningar. Materialet från Fastighetsbyrån innehöll tillslut variablerna:

Slutpris
Slutpriset var det exakta beloppet som lägenheten har sålts för.

Startpris
Startpriset var det belopp som säljaren och mäklaren har kommit överrens om att budgivningen skulle starta på.

Avgift
Var den avgift som köparen är tvungen att betala i månaden till bostadsrättsföreningen.

Boarea
Lägenhetens beboliga area.

Antal rum
Antalet rum som lägenheten är uppdelad i.

Våningsplan
Vilket våningsplan i huset som lägenheten är belägen på.

Antal våningar
Hur många våningar som finns totalt i huset.

Balkong
Är en 0/1-variabel. Där 1 står för att lägenheten ifråga har balkong.

Hiss
Är en 0/1-variabel. Där 1 står för att huset har hiss.

Garage
Är en 0/1-variabel. Där 1 betyder att köparen har möjlighet till en garageplats.

Avtalsdag
Dagen/månad/år då kontrakten skrevs under. Alltså inte dagen då köparen får tillgång till lägenheten.

Byggnadsår
Året då huset byggdes.

Kön på mäklaren
Är en 0/1-variabel där 1 betyder att mäklaren som sålde lägenheten var en kvinna och därmed betyder 0 att mäklaren var man.

Nyproduktion
Är en 0/1-variabel där 1 hänvisade till att lägenheten var nybyggd. Det vill säga, ingen har bott i lägenheten innan.

Boareakälla
Boareakälla är en variabel som beskriver på vilket sätt som boarean hade blivit uppmätt. Den kunde blivit uppmätt på fem olika sätt.

Område
Denna kategori beskriver i vilket område av Stockholm som lägenheten var belägen. Det finns elva stycken dokumenterade områden: Östermalm, Östermalm/Vasastan, Nedre Gärdet, Hjorthagen, Gärdet, Djurgården, Ekhagen, Gärdet/Östermalm, Östermalm/Starrängsringen, Östermalm/Nedre Gärdet och Östermalm/Hjorthagen.

Adress
Vi har även tillgång till den exakta adressen som lägenheten var belägen på.

Vi utvidgade sedan detta datamaterial med två parametrar som båda hade till syfte att beskriva den rådande konjunkturen för att se vilken av dessa som lyckades bäst med denna uppgift.

CCI
CCI är en variabel ur Konjunkturinstitutets hushållsbarometer. Klas-Göran Warginger på Konjunkturinstitutet beskriver framtagandet av variabeln på följande sätt: ”CCI baseras på svaren på 5 av frågorna i enkätundersökningen Hushållsbarometern. Frågorna är 2 stycken som avser hur den egna ekonomiska situationen är just nu (jmf med för 12 månader sedan) samt hur denna kommer att vara om 12 månader (jmf med nuläget). Vidare 2 frågor om den ekonomiska situationen i Sverige, jämfört på samma sätt. Slutligen 1 fråga som avser om respondenten tycker det är lämpligt att köpa kapitalvaror i nuläget. Hanteringen av svaren innebär att man beräknar nettotal: dvs. den procentuella andel av de svarande som svarar ”mycket bättre”eller ”bättre” minus den procentuella andel som svarar ”mycket sämre”eller ”sämre”. Summan av dessa nettotal divideras med 5. Om CCI=0 innebär det att CCI är något under medelvärdet för CCI (som är 4,4 för perioden 1993 till mars 2009). Formell övre och undre gräns för CCI är +/- 100.”

PROK
PROK är en egen hopsatt variabel som är summan av två stycken procenttal från konjunkturinstitutets konjunkturbarometer hushåll. PROK är summan av procenttalen som svarat ”ja absolut”eller ”ja troligen” på frågan: ”Bygga eller köpa hus/lägenhet inom 12 mån”.

3 Metoder

3.1 Regression

Regressionsmodeller kännetecknas enligt Rolf Sundbergs kompendium i Tillämpad Matematisk Statistik av att en mätstorhet under slumpmässig osäkerhet beror genom ett linjärt funktionssamband av en eller flera precist kända variabler. Metodens syfte är att hitta en funktion som på bästa sätt passar observerade data. Uteslutande under hela arbetet har jag använt mig av regression som tillämpar minsta kvadrat metoden.

3.1.1 Enkel linjär regression

Enkel linjär regression definieras i Rolf Sundbergs kompendium i Tillämpad Matematisk Statistik enligt följande:

$$Y_i = \alpha + \beta * x_i + \epsilon_i$$

där Y_i är responsvariabeln (variabeln som påverkas), x_i är den förklarande variabeln (variabeln som påverkar), α och β är parametrar, ϵ_i står för de slumpmässiga variationerna och i=1,...,N där N är antalet gjorda mätningar. Om ϵ_i betraktas som oberoende och normalfördelade N(0,σ^2) har vi modellen enkel linjär regression. Alltså har modellen väntevärdesfunktionen:

$$\mu = \alpha + \beta * x$$

3.1.2 Multipel linjär regression

Multipel linjär regression används då man har misstanke om att en responsvariabel Y_i beror av två eller flera förklarande variabler. Denna form av linjär regression definieras enligt följande:

$$Y_i = \alpha + \beta_1 * x_{1i} + \beta_2 * x_{2i} + ... + \beta_n * x_{ni} + \epsilon_i$$

ϵ_i betraktas även här som oberoende och normalfördelade, N(0,σ^2), och i=1,..,N där N är antalet gjorda mätningar. Modellen har väntevärdesfunktionen:

$$\mu = \alpha + \beta_1 * x_1 + ... + \beta_n * x_n$$

3.2 Minsta kvadrat metoden

Minsta kvadrat metoden (Least square method) är en metod som först beskrevs av Carl Friedrich Gauss runt 1794 och som används i regressionsanalys för beräkning av en regressionslinje. Syftet med metoden är att minimera felet i en funktion som anpassas utifrån de observerade värdena. Det vill säga de observerade värdena ska ligga så nära den skattade linjen som möjligt. Metoden beräknar residualerna, det vill säga avståndet från varje observerat värde till den skattade linjen, och kvadrerar sedan detta avstånd för att komma ifrån problemet med positiva och negativa tal. e_i står för residualerna, y_i är den verkliga observationen och $\widehat{y}_i$ är det skattade funktionsvärdet.

$$e_i = y_i - \widehat{y}_i$$

Vi strävar efter att göra summan av kvadraterna på residualerna så liten som möjligt. Detta mått kallas RSS (residual sum of squares). Det minsta värdet på RSS ger enligt denna metod den bäst skattade regressionslinjen.

$$RSS = \sum e_i^2$$

3.3 Stegvisa regressions metoder

När vi använder oss av Regressionsanalys på datamaterial, med ett större antal möjliga förklarande variabler, så kan det vara svårt att välja ut vilka variabler som bör användas för att förklara materialet så tillfredställande som möjligt. För att underlätta det här problemet kan man använda sig av stegvisa regressionsmodeller.

3.3.1 Stepwise Regression

Stepwise regression beskrivs i Rolf Sundbergs kompendium i Tillämpad Matematisk Statistik som den mest använda stegvisa proceduren för att välja variabler till ett regressionssamband. Metoden beskrivs som en mer avancerad metod av Forward selection. Metoden utgår från en modell som inte innefattar några förklarande variabler över huvudtaget, dvs. $\mu_i = \alpha$. Metoden inkluderar sedan stegvis en variabel i taget. Vilken variabel som inkluderas i modellen bestäms via test av hypotesen $\beta_h = 0$ för alla variabler x_h. Den variabel som visar sig vara mest signifikant är den som inkluderas i modellen. Till skillnad från Forward selection så testas efter varje steg

att variabler, som redan inkluderats i modellen fortfarande är signifikanta. Eventuella variabler, som inte längre visar sig vara signifikanta utesluts då ur modellen. Detta kan bero på att fler förklarande variabler tillsammans beskriver datamaterialet bra, medan de var för sig inte förklarar mer än en annan tredje variabel. Proceduren slutar då inga flera variabler var tillräckligt signifikanta på en förutbestämd signifikansnivå.

3.4 Korsvalidering

Korsvalidering är en teknik för att bedöma hur väl resultaten av en statistisk analys beskriver verkligheten. Metoden används främst inom prediktion där syftet med metoden är att uppskatta hur väl en predikterad modell kommer att fungera i praktiken.

Metoden delar upp materialet, som den har att arbeta med, i två undergrupper. Den ena delen blir ett utbildningsmaterial (training set) och den andra delen blir ett valideringsmaterial (validation set eller testing set). Metoden använder utbildningsmaterialet för att med hjälp av regressionsanalys få fram den mest tillfredställande modellen som passar dessa värden. Därefter appliceras denna modell på valideringsmaterialet för att undersöka hur bra modellen passar dessa värden. Det vill säga vi använder oss av Minsta kvadrat metoden och beräknar MSE (mean squared error). För att reducera variationerna så mycket som möjligt så kan man utföra denna metod flera gånger genom att dela upp materialet i utbildningsgrupp och valideringsgrupp på olika sätt. Det slutgiltiga resultatet blir då ett medelvärde över alla delresultaten.

Metoden vi använt oss av i vår undersökning är en form av Korsvalidering vid namn Leave One Out. Här använder vi bara en observation som valideringsmaterial och de övriga observationerna som utbildningsmaterial. Sedan reducerar vi variationen genom att låta alla observationer vara valideringsmaterial, en gång var. Det vill säga, vi utför denna undersökning lika många gånger som antalet observationer. Vi beräknar $\widehat{y}$ med hjälp av den skattade regressionslinjen och beräknar sedan residualskattningen med hjälp av det observerade värdet på y. Notationen (i) står för att observation i inte var med i utbildningsmaterialet som användes för att ta fram $\widehat{y}_{(i)}$.

$$\widehat{e_{(i)}} = y_i - \widehat{y}_{(i)}$$

Metoden leder fram till att vi får n stycken predikterade värden på residualerna, för våra n observationer. Dessa värden använder vi för att beräkna en statistika vid namn PRESS (predicted residual sum of squared) som

används som ett jämförelsemått av en modells prediktionsförmåga. PRESS är summan av de kvadrerade residualskattningarna där varje $\widehat{y}_{(i)}$ kommer från regressionen där dennas observation tillhörde valideringsmaterialet. Ett så lågt PRESS-värde som möjligt är önskvärt.

$$\text{PRESS} = \sum \widehat{e}^2_{(i)}$$

I våra undersökningar kommer vi använda oss av dels slutpriset och dels det logaritmerade slutpriset som responsvariabel. Det ligger i vårt intresse att skapa oss en bild av hur väl anpassad vår skattning är. Ett mått på detta skulle kunna vara prediktionsfelet. Vi kan aldrig beräkna prediktionsfelet exakt eftersom vi då aldrig skulle göra något fel. Men när responsvariabeln representeras av slutpriset kan vi, med hjälp av PRESS statistikan, beräkna det typiska prediktionsfelet i kronor. Vi bör dock komma ihåg att inte alla utfall är typiska. Vi använder oss av följande formel:

$$\text{Typiskt prediktionsfel i SEK} \approx \sqrt{\text{PRESS}/n}$$

När vi istället använder oss av det logaritmerade slutpriset som responsvariabel bör vi inte omvandla PRESS till ett prediktionsfel i kronor. Men vi kan beräkna en approximation av kvoten för det verkliga slutpriset och vår skattning. Vi bör även här komma ihåg att det är en approximation av den typiska kvoten. Skillnaden mellan vänster och högerled i formlen kan skilja sig så mycket att det inte är rimligt att använda approximationen. Med detta i åtanke, kan vi i denna uppsats använda kvoten för att skapa oss en bild av prediktionsfelets typiska storlek.

$$log(P) - log(\widehat{P}) \approx \sqrt{\text{PRESS}/n}$$

$$P \approx \widehat{P} * e^{\sqrt{\text{PRESS}/n}}$$

$$\frac{P}{\widehat{P}} \approx e^{\sqrt{\text{PRESS}/n}}$$

Vi får alltså en relation mellan skattningen av slutpriset och det verkliga värdet.

3.5 Prediktion

Prediktion är en metod som används för att uppskatta ett kommande utfall. I detta arbete kommer prediktion vara förknippat med regressionsmodeller. Metoden använder sig av tidigare kända värden för att med hjälp av dessa uppskatta kommande värden. Vi kan utgå från vilket värde vi vill på x-axeln och med hjälp av regressionslinjen (i enkel linjär regression) uppskatta ett värde på y-axeln. Samma princip gäller förstås med multipel regression. En prediktor för Y baserad på $\mathbf{X}$ definieras som en funktion d($\mathbf{X}$). Prediktorn är linjär ifall d($\mathbf{X}$) är linjär, det vill säga ifall

$$d(\mathbf{X}) = a + b_1X_1 + b_2X_2 + ... + b_nX_n$$

där a och b_1,...,b_n är parametrar. Prediktionsfelet ges utav:

$$Y - d(\mathbf{X})$$

eller med våran tidigare notation:

$$\widehat{e}_{(i)} = y_i - \widehat{y}_{(i)}$$

Det finns olika metoder för att avgöra vilken modell som är den bästa prediktorn. Vi kommer att använda oss av metoden Korsvalidering i just detta syfte. Vi kommer mer exakt att titta på PRESS-värdet. PRESS-värdet är kopplat till prediktionsfelet enligt formeln:

$$PRESS = \sum \widehat{e}^2_{(i)}$$

Den modell som ger det lägsta PRESS-värdet är också den modell som ger lägst RSS (summan av prediktionsfelen i kvadrat). Det vill säga även

den modell som är att föredra enligt minsta kvadrat metoden. Dock är det inte säkert att det är, den för vårt syfte, mest lämpade modellen att välja då vi bör ta hänsyn till andra faktorer, såsom antal variabler. En reducerad modell med en obetydlig ökning av PRESS-värdet är även den en bra kandidat.

Det vi bör reflektera över, innan vi använder oss av prediktion, är att de variabler som uppskattar kommande värden mest tillfredställande inte alls behöver vara de variabler som förklarar priset, det vill säga ligger till grund för det aktuella priset. Regression har i det här fallet gemensamma egenskaper med korrelation och det innebär att det inte behöver finnas ett orsakssamband mellan de olika variablerna.

4 Viktiga begrepp som används

4.1 Signifikant förklaringsvariabel

En förklaringsvariabel kallas signifikant om den vi test av hypotesen att motsvarande parameter är noll ger signifikant utslag. En nollhypotes förkastas om det observerade utfallet inträffar i mindre än 5 procent av fallen, givet nollhypotesen. Signifikansnivån , 5 procent, defineras som sannolikheten att förkasta en hypotes som egentligen är sann, på 95 procents-nivån.

4.2 Förklaringsgraden

Förklaringsgraden, R^2, är den del av variationen i Y som kan förklaras med hjälp av X. Den anges som ett mått i procent. R^2 är det vanligaste anpassningsmåttet i samband med linjära modeller enligt Rolf Sundbergs kompendium i Tillämpad Matematisk Statistik, och definieras enligt följande:

$$R^2 = \frac{KVS_{modell}}{KVS_{totalt}} = 1 - \frac{KVS_{residual}}{KVS_{totalt}}$$

R^2 är kvadraten på den så kallade multipla korrelationskoefficienten, som mäter korrelationen mellan $\mathbf{Y}$ och $\widehat{\mathbf{Y}}$. Vidare är:

$$KVS_{totalt} = \sum (y_i - \bar{y})^2$$

$$KVS_{modell} = \sum (\hat{y}_i - \bar{\hat{y}})^2$$

$$KVS_{totalt} = \sum (y_i - \hat{y})^2$$

5 Resultat

5.1 Datamaterialet

Datamaterialet innefattar information om 451 stycken sålda lägenheter. För att skapa oss en bild av materialet så börjar vi med att undersöka relationen mellan alla variabler genom att studera plottar över sambanden mellan dessa. Det främsta syftet med undersökningen av datamaterialet är att upptäcka om några av de förklarande variablerna korrelerar med varandra eller med andra ord om det finns variabler som samvarierar. Finns det flera variabler som samvarierar i materialet vill vi försöka reducera antalet till en. Nedan följer resultatet av denna undersökning samt resonemang kring plottarna.

5.1.1 Undersökning av datamaterialet

Vi börjar med att granska materialet för att se om data saknas för några av lägenheterna. Datamaterialet innehöll enbart statistik över variabeln startpris för halva materialet. Innan den 1 juli 2007 så bokfördes inte vilket pris som budgivningen startat på. Här får vi göra en avvägning om att antingen reducera materialet till hälften och använda oss av startpris eller helt utesluta denna variabel ur vårt material. En reducering av datamaterialet till hälften skulle resultera i ett material bestående av drygt 200 lägenheter. Vi väljer att prioritera data från så många lägenhetsförsäljningar som möjligt. Med detta i åtanke väljer vi därför att utesluta startpris ur vårt material och behålla alla observationerna.

En annan variabel som vi saknar information om, hos en större del av de sålda lägenheterna, var antalet våningar som finns i huset. Vi resonerar därför efter samma princip som ovan i det här fallet och väljer att ta bort variabeln ur materialet. Ytterligare en anledning till varför vi väljer att utesluta variabeln, antal våningar, är att vi har en variabel som anger på vilken våning som den aktuella lägenheten är belägen på. Båda dessa har som uppgift att representera våningseffekten och den sistnämnda är då mer relevant. Detta på grund av att en köpare antagligen är mer intresserad av på vilken våning den aktuella lägenheten ligger, än antalet våningar i huset totalt.

Boareakälla är en variabel som beskriver på vilket sätt som boarean har blivit uppmätt. Denna variabel behöver hanteras som en kategorivariabel i vår regressionsanalys. Då själva uppmätningsmetoden av boarea är betydelselös för vår undersökning väljer vi att utesluta denna ur vårt material.

Materialet innehåller även information om avtalsdagen. För oss är inte den specifika dagen av intresse utan vi väljer att koncentrera oss på månaden

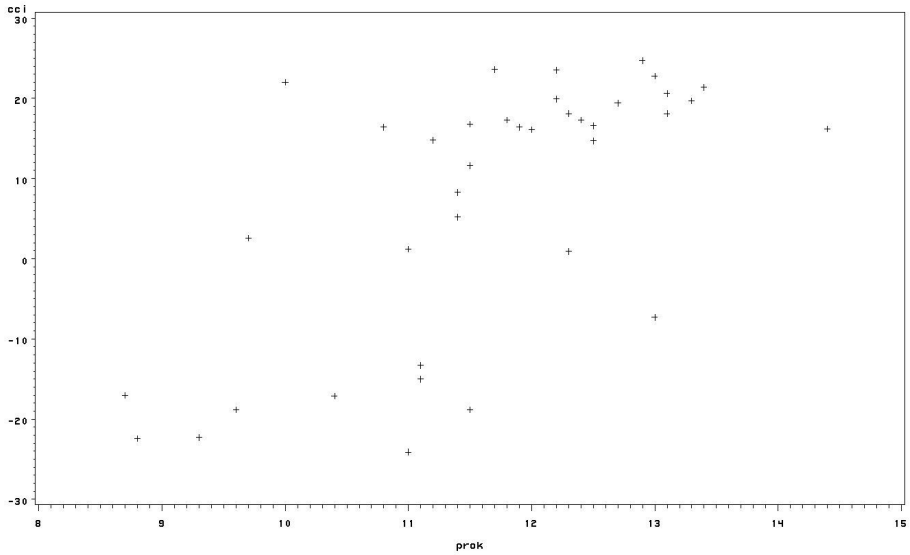
och året som lägenheten såldes. Vi väljer att konstruera variabeln som en siffra som löper från 1 till 39, där 1 motsvarar januari 2006 och 39 motsvarar mars 2009, och döper denna variabel till tidsvariabeln.

Det är inte oväntat att variablerna nyproduktion och byggår har ett samband då de båda ämnar till att beskriva samma effekt. Nyproduktion är en 0/1-variabel som antar värdet 1 ifall lägenheten ifråga är nyproducerad, det vill säga om ingen har bott i lägenheten innan. Det finns bara lägenheter byggda på senare delen av 2000-talet som anses vara nyproducerade, vilket är rimligt. Därför väljer vi att ta bort nyproduktion ur materialet eftersom variabeln byggår hjälper till att beskriva denna effekt.

Materialet innehöll specifik information gällande i vilket område varje lägenhet var belägen. Det skulle vara av intresse att använda områdesindelningen som en kategorivariabel och titta på skillnaderna mellan mindre områden i Stockholms innerstad. Men när vi börjar granska datamaterialet mer noggrant så ser vi att det är en väldigt diffus områdesindelning. Det finns inga tydliga gränser för var områdena börjar och tar slut. Det finns bland annat områden som beskrivs som: Gärdet, Nedre Gärdet, Gärdet/Östermalm, Nedre Gärdet/Östermalm. Lägenheter som ligger i husen bredvid varandra har bokförts på två olika områden. Detta kan bero på att olika mäklare har haft hand om de olika lägenhetsförsäljningarna och haft olika uppfattning om vilket område lägenheten hör till. Dessa fakta gör att vi väljer att inte använda oss av områdesindelningen som en variabel då den inte har en tillräckligt tillfredställande uppdelning.

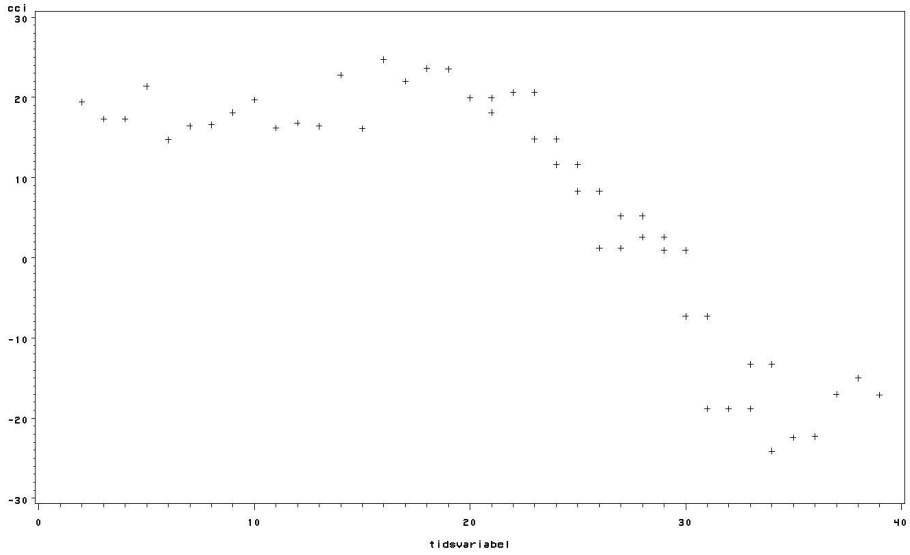
Datamaterialet bestod även av alla lägenheternas adresser, vilket är en variabel av stort intresse. Dock var det bara gatunamn och gatunummer som var dokumenterat och inte postkoden. För att ha möjlighet att använda postkoden som en variabel i undersökningen skulle ett alternativ vara att knappa in alla adresser i en sökmotor. Vi kom fram till att det skulle ta allt för stor del av tiden som vi har till vårt förfogande. Då alla lägenheter som vi undersöker ligger i samma del av Stockholm så har vi ändå ett relativt begränsat område som vi undersöker så därför valde vi att helt utesluta adressen som en variabel.

Variablerna PROK och CCI är båda hämtade från konjunkturinstitutets databas och ämnar båda till att beskriva allmänhetens optimistiska investeringsvilja. Det vill säga båda ska beskriva den rådande konjunkturen. Dessa två variabler bör vara korrelerade med varandra och som vi ser i plotten så kan vi se ett positivt samband mellan dem. Ett högre värde på CCI motsvaras av ett högre värde på PROK.

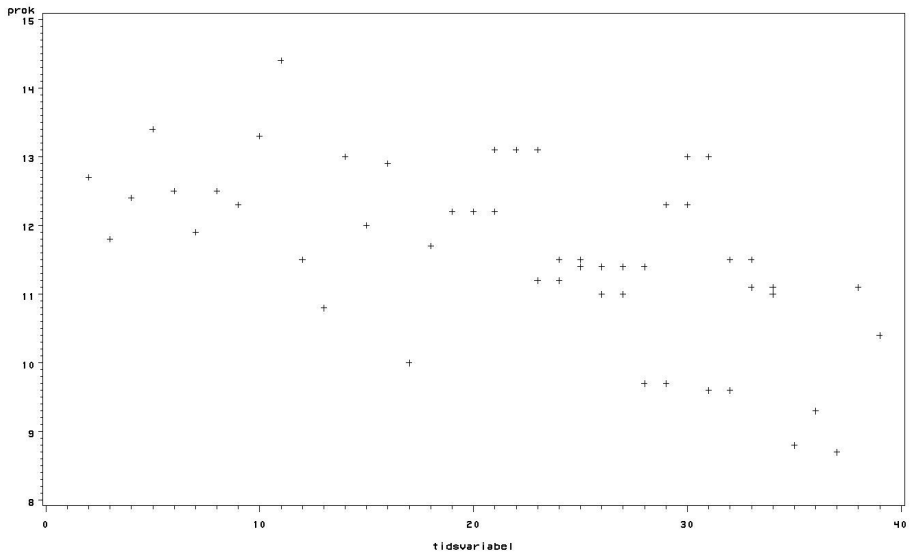


CCI i förhållande till PROK

För att avgöra vilken av dessa som vi anser mer lämpad för vår undersökning så plottar vi de båda variablerna mot tiden.



CCI i förhållande till tiden

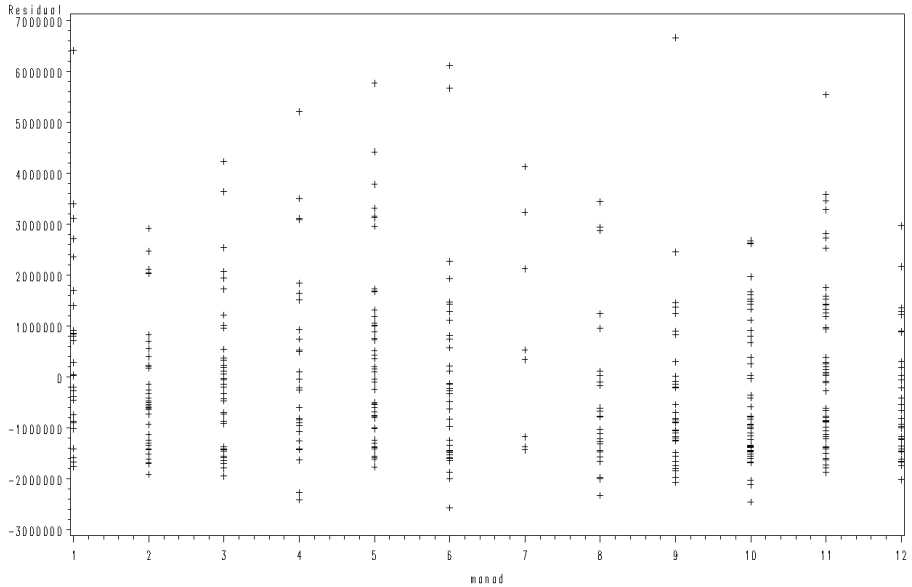


PROK i förhållande till tiden

Här ser vi att CCI ger en mer lättolkad bild av svängningarna i konjunkturen. Men för att undersöka variablerna ytterligare så utför vi en multipel regression på slutpriset med alla variabler som är kvar samt CCI och PROK. Det visar sig att CCI blir väldigt mer signifikant i regressionsanalysen. Vi väljer därför CCI till en bättre variabel för att beskriva konjunkturen i våran undersökning.

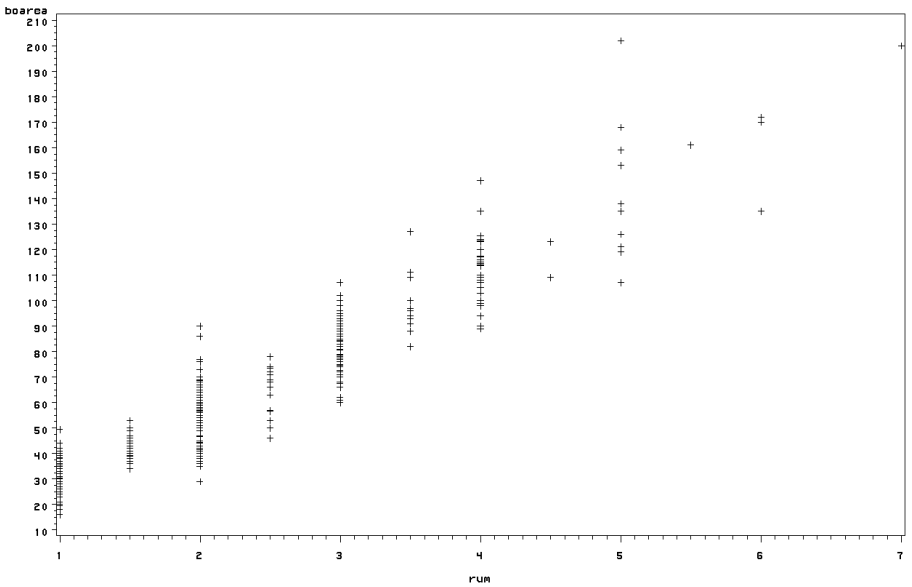
Ett problem som vi kommer att få hantera med CCI är att vi nu använder oss av CCI som motsvarar den månad då lägenheten såldes. Detta kommer bli ett problem vid prediktion. Av naturliga skäl kan inte konjunkturinstitutet släppa månadens CCI-siffra tidigare än i början av månaden efter. Vi diskuterade därför att skjuta CCI en månad fram i tiden så att januari månads CCI-siffra skulle vara med i modellen för lägenheter som säljs i februari månad. Men vårt resonemang var att det skulle bli snedvridet och inte statistiskt rätt med denna modifikation. Därför valde vi att ha kvar CCI för den månaden som siffran motsvarar. Men i en prediktionsmodell kan vi använda oss av den senaste siffran som vi har tillgång till eller så får vi använda oss av gammal statistik över CCI och information om dagsläget i ekonomin och på så sätt göra en kvalificerad gissning om vad CCI skulle vara den aktuella månaden.

Vi hade från början en misstanke om att det kunde finnas någon månadseffekt hos lägenhetsförsäljningarna. Vår teori är nämligen att de flesta gärna flyttar på sommaren. De borde i så fall säljas fler lägenheter under sommarmånaderna och det borde i sin tur påverka priset. Ifall detta vore fallet skulle det vara en idé att transformera tidsvariabeln på något lämpligt sätt. Men innan vi gör detta bör vi kontrollera ifall det finns något underlag för vår teori. Vi kontrollerar detta genom ett göra en enkel linjär regression med slutpris som responsvariabel och tidsvariabeln som den enda förklarande variabeln. Vi sparar residualerna från denna regression och plottar dem mot månad (en variabel som kan anta värdera 1-12, där 1 står för att lägenheten blev såld i januari och 12 för att lägenheten blev såld i december). I plotten kan vi se att enstaka försäljningar visar på högre slutpris i vissa spridda månader men vi kan inte konstatera någon månadseffekt. Dessa spridda högre slutpris (som kan ses i månad 1,4,5,6,9 och 11) kan förklaras med en konjunktureffekt. I januari 2007 befann sig Sverige inte i samma ekonomiska läge som vi gjorde i januari 2009. Denna effekt använder vi CCI för att beskriva. Därför väljer vi att behålla tidsvariabeln i dess ursprungliga skick.



Residualerna i förhållande till månad

Det ligger nära till hands att misstänka att boarea och antal rum är starkt korrelerade. Vi undersöker därför deras förhållande genom att plotta de båda variablerna mot varandra nedan.



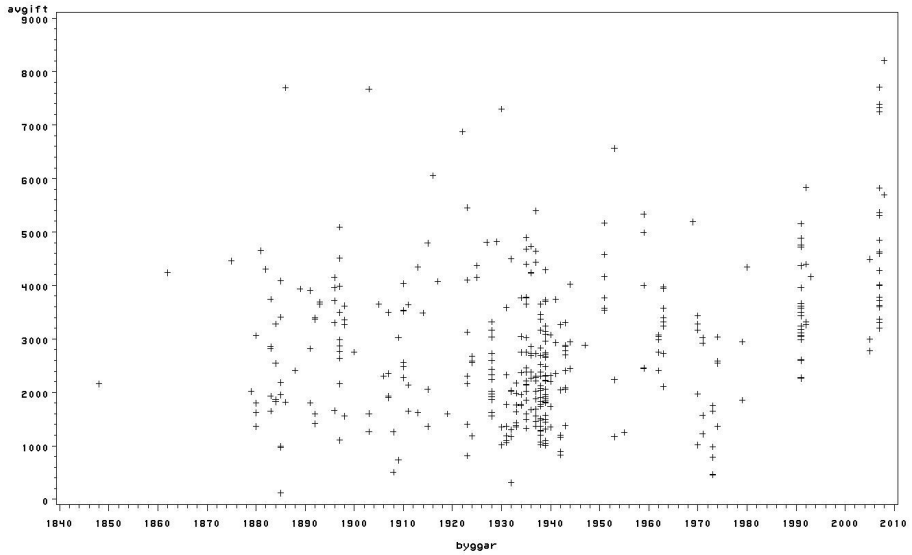
Boarean i förhållande till antalet rum

Vi ser ett tydligt samband mellan de båda variablerna. Vidare resonemang kring detta och deras eventuella korrelation med andra variabler i undersökningen, diskuteras i nästa avsnitt.

5.2 Test av hypoteser rörande datamaterialet

Vi fortsätter att undersöka datamaterialet men tar nu hjälp av Regression för våra undersökningar.

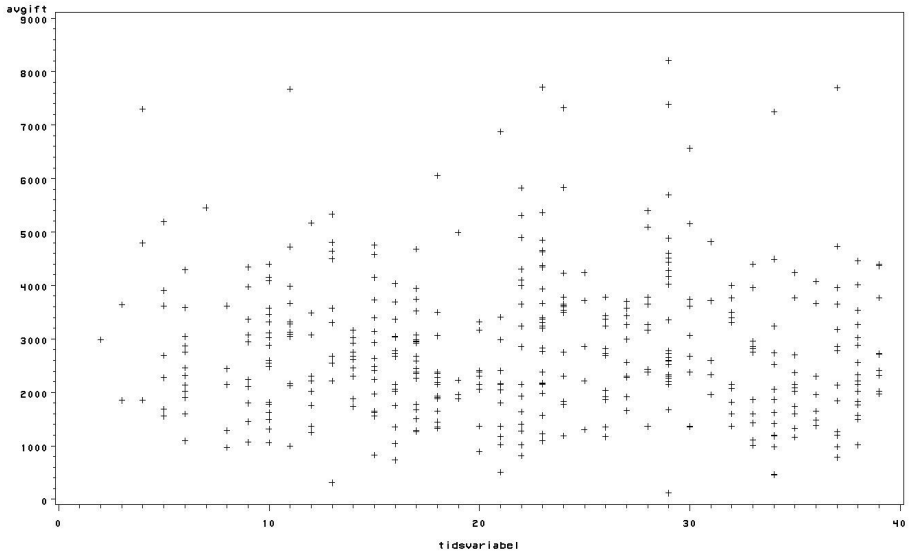
En hypotes som vi har är att avgiften bör vara högre i nyproducerade lägenheter än vad den är i lite äldre. Det vill säga lägenheter med ett sent byggår kan tänkas vara belastade med en högre avgift än de lägenheter som har ett tidigt byggår. För att se ifall detta var fallet studerar vi plotten nedan.



Avgiften i förhållande till byggår

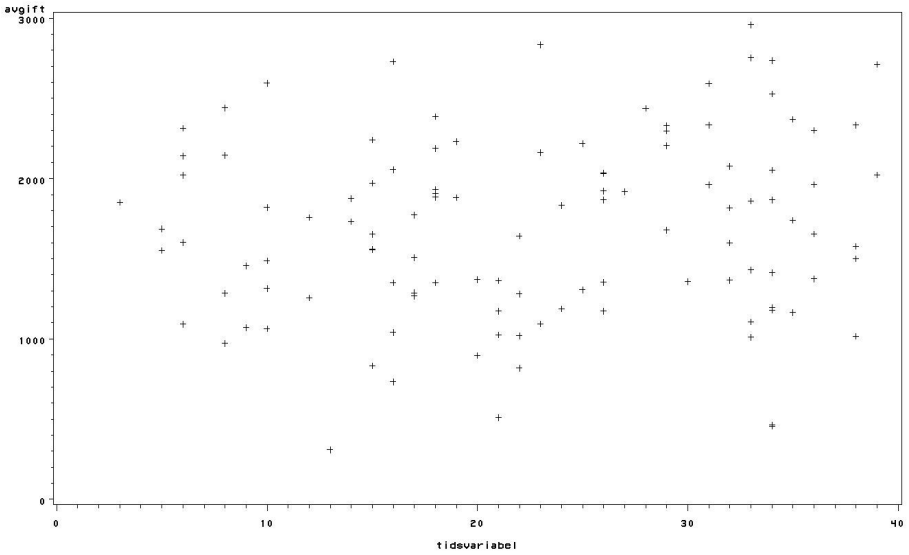
För att detta ska stämma bör vi se ett positivt samband mellan de båda variablerna. Man ser här en viss tendens till att ett senare byggår resulterar i att avgifterna ligger lite högre. Vi har alltså en viss korrelation mellan avgift och byggår. Vi får göra en avvägning av vad som är viktigast. Antingen informationen som försvinner genom att exkludera någon av dessa variabler eller att tvingas hantera den rådande korrelationen mellan variablerna. I vår undersökning är det svårt att hitta variabler som är helt oberoende av varandra. Vi väljer därför att prioritera informationen som de båda ger oss och att ta med båda variablerna i våra fortsatta undersökningar. Vi ska dock studera plotten lite ytterligare. Vi kan se att vi har lägenheter som byggdes 1980 men sedan är det ett stort hopp till 1990 och vidare ett stort hopp til 2005. Orsaken bakom detta kan spekuleras i. Bostäderna belägna på Östermalm är i huvudsak äldre byggnader så tidiga byggår är generellt inte konstigt. Men hoperna vid år 1990 och år 2005 är svårare att hitta en förklaring till. En möjlig teori kan vara att nyproduktioner av stora lägenhetsfastigheter i de aktuella områdena var färdiga för försäljning, under dessa årtal.

En annan hypotes, som vi vill undersöka är om avgiften korrelerar med tiden. Vår hypotes är att avgiften ökar med tiden, som bland annat kan bero på inflation. För att konstatera om vi ser ett sådant samband studerar vi plotten nedan.

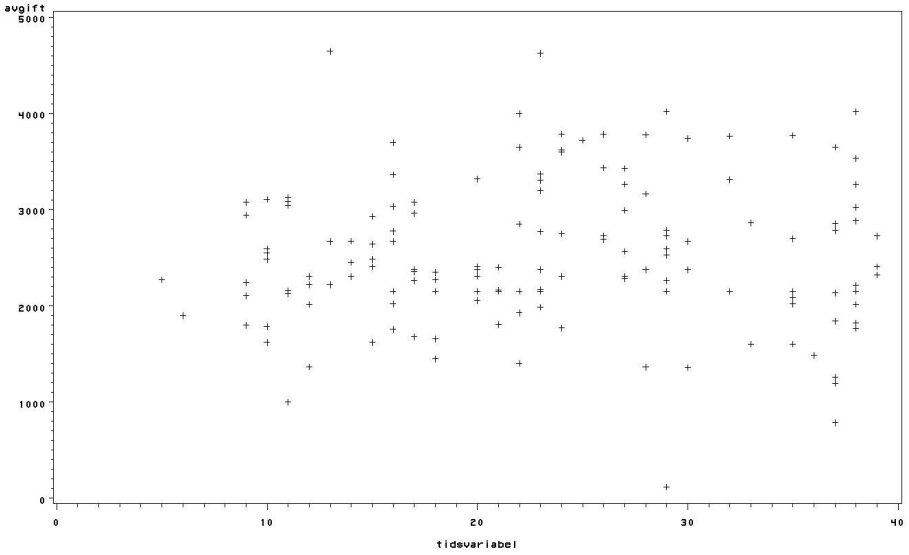


Avgiften i förhållande till tiden för hela materialet

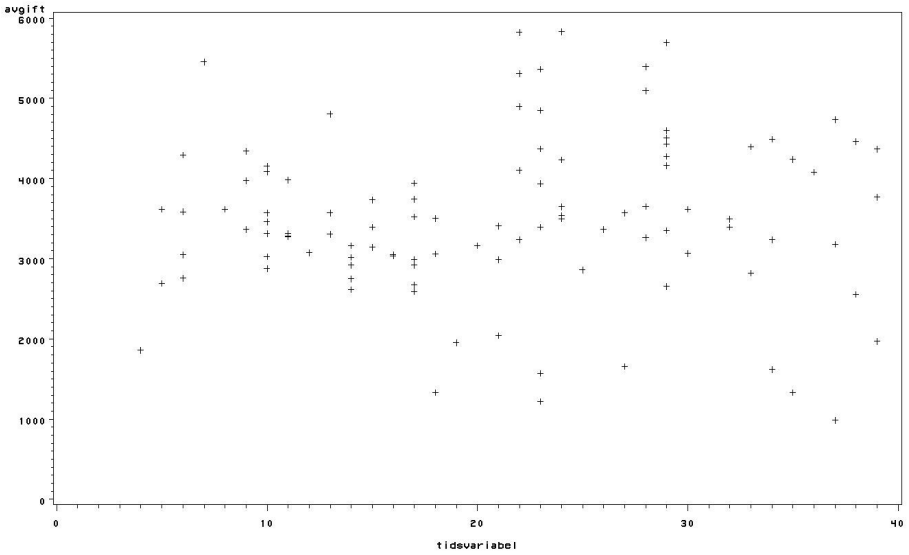
Vi kan inte se någon tydlig sådan trend. Vi studerar även samma samband fast för ettor, tvåor, treor och fyror var för sig. Vi väljer att dela upp materialet för att det finns en viss risk att ett samband mellan tid och avgift inte syns när vi studerar hela materialet. Det blir enklare att upptäcka om vi har en ökande avgift med tiden om vi studerar lägenheter av liknande storlek. Större lägenheter har högre avgift och vice versa. Vilket leder till att vårt hypotetiska samband kan ligga gömt, då spridningen på lägenheternas avgifter är stor redan från början.



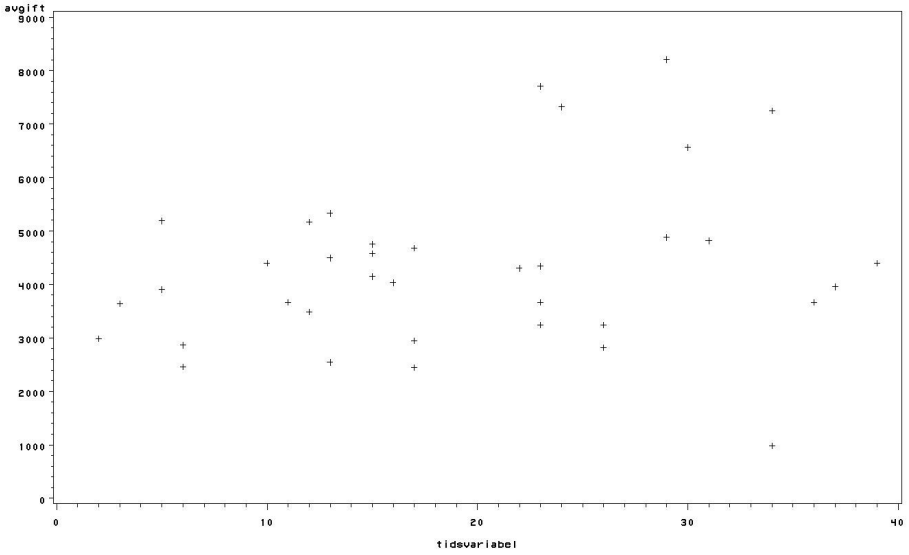
Avgiften i förhållande till tiden för ettor



Avgiften i förhållande till tiden för tvåor



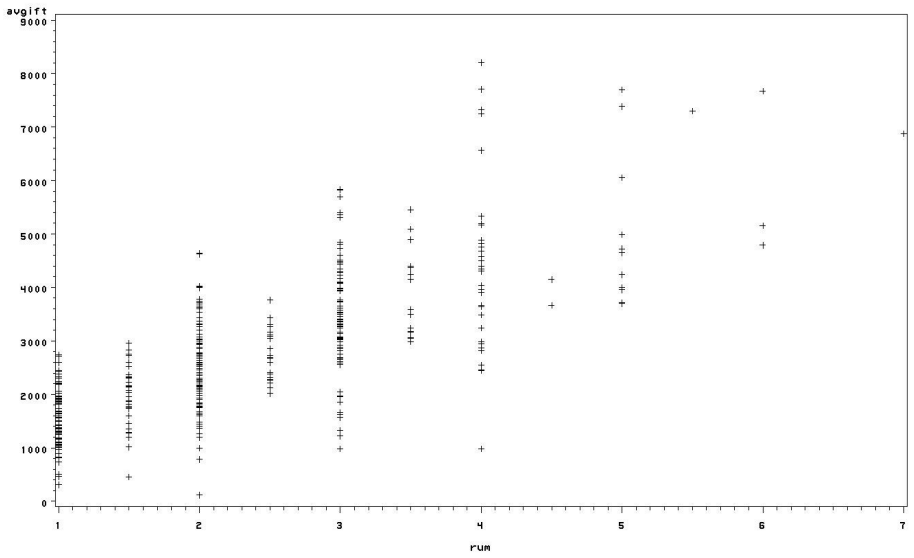
Avgiften i förhållande till tiden för treor



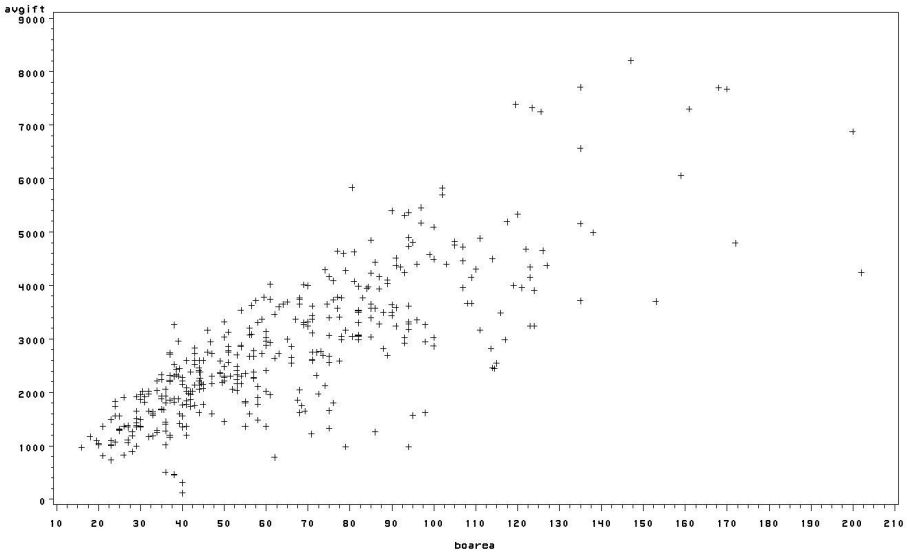
Avgiften i förhållande till tiden för fyror

Inte heller dessa plottar visar på ett samband mellan tid och avgift. Då vi studerar bostadsrätter är det inte säkert att avgiften höjs trots faktorer som inflation. En förklaring till detta kan vara att vi studerar bostadsrätter, i huvudsak belägna på Östermalm, vilket är äldre bostadsrättsföreningar med jämförelsevis små lån. Avgifterna kan då typiskt sett vara konstanta eller till och med sänkas då lån betalas av eller då räntor sjunker. Det kan även bero på att tiden vi studerar är relativt kort, enbart två år, och att den procentuella ökningen av avgiften äts upp utav variationen i avgiften hos de olika lägenheterna. Den kraftiga variationen beror på att alla observationer är på olika lägenheter det vill säga vi har inga lägenheter som vi har gjort flera mätningar på. Avgifterna kan alltså skilja kraftigt mellan lägenheterna.

Sedan tidigare har vi konstaterat den tydliga korrelationen mellan antal rum och boarean. Misstänkt är även att avgiften ökar med antalet rum och med boarean. Detta samband kan vi se tydligt i plottarna nedan.



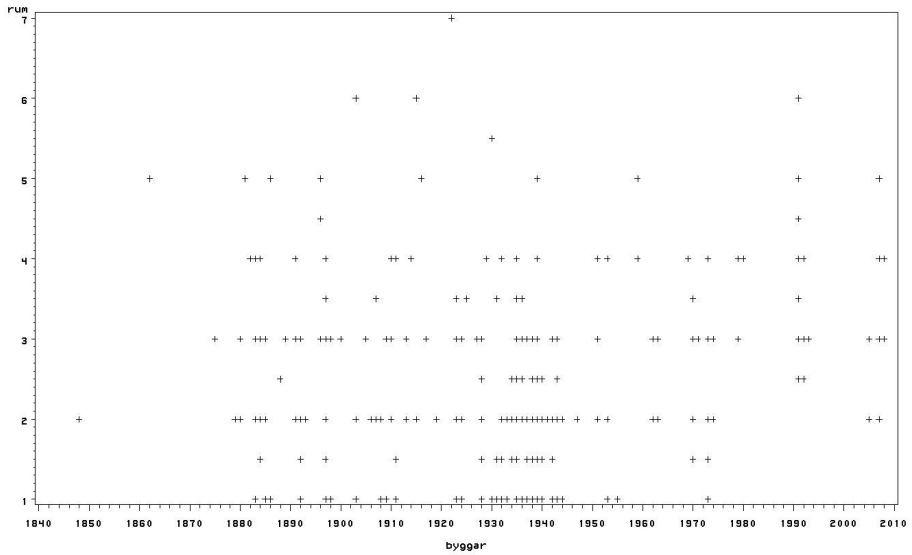
Avgiften i förhållande till antalet rum



Avgiften i förhållande till boarean

Alltså har vi en kraftig korrelation mellan boarea, antal rum och avgiften. Då är frågan hur bra det är att ta med alla dessa variabler i vår undersökning. Eftersom alla variablerna är ytterst intressanta för undersökningen och bidrar med viktig information så gör vi en större förlust genom att exkludera några av variablerna än att hantera korrelationen som de har med varandra.

Ytterligare en hypotes som vi vill undersöka är ifall antalet rum har något samband med tiden. Hypotesen som vi har är att det byggs lägenheter med mindre antal rum på senare år än förr i tiden. Men då inget sagt om boarean som vi inte tror följer det här sambandet. Det vill säga, hypotesen är att nybyggda lägenheter har en öppnare planlösning. För att se ifall vi kan dra dessa slutsatser studerar vi plotten nedan.



Antalet rum i förhållande till byggår

Ur denna plott kan vi inte se något tydligt samband som skulle kunna stärka vår hypotes. Vi väljer därför att bortse från denna.

5.3 Regressionsanalys

För att få en så bra bild som möjligt av datamaterialet så fortsätter vi med att undersöka hur bra vi skulle kunna förklara variationen i data med hjälp av de variabler som vi har valt ut. Som förklarande variabler har vi nu kvar:

Boraea
Avgift
Tidsvariabel
CCI
Antal Rum
Byggår
Våning
Balkong
Hiss
Garage
Kön

Därför väljer vi att börja med att göra en multipel regression på slutpriset med de övriga elva variablerna som förklarande variabler. ANOVA tabellen

över denna regression redovisas nedan.

Tabell 1

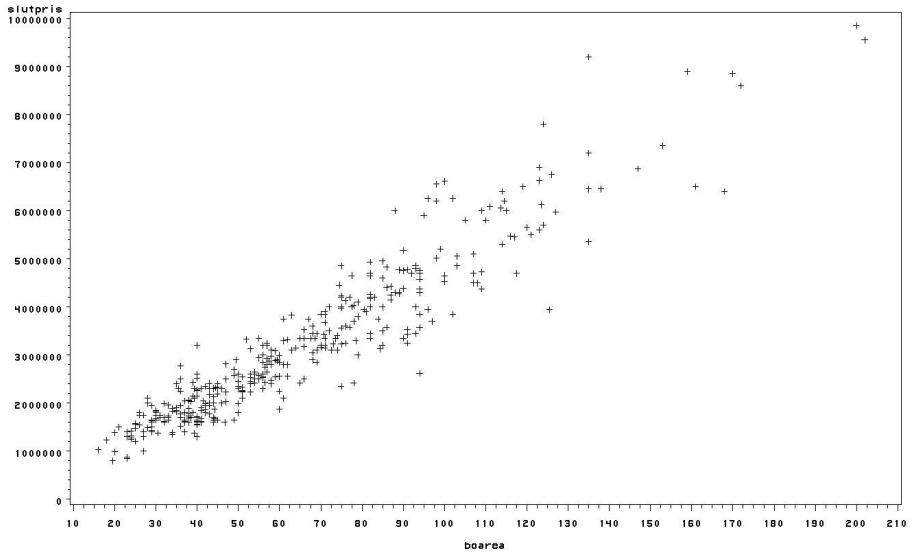
Källa	FRGR	KVS	MKVS	F-value	R^2
Model	11	$9.46532 * E^{14}$	$8.604836 * E^{13}$	388.06	0.9175
Error	384	$8.514908 * E^{13}$	$2.217424 * E^{11}$		
Corrected Total	395	$1.031681 * E^{15}$			

Vi ser här att vi kan förklara en stor del av variationen med hjälp av dessa elva variabler, 91.75 procent. Vi vill nu fortsätta att undersöka materialet med hjälp av regressions analys och se om vi kan reducera det ytterligare utan att förlora för mycket i förklaringsgrad. Vi väljer därför med att fortsätta analysen genom att använda oss av den stegvisa regressionsmetoden Stepwise Regression. Vi använder oss av samma förklarande variabler och fortfarande slutpris som responsvariabel. Resultatet visas nedan.

Tabell 2

Källa	FRGR	KVS	MKVS	F-value	R^2
Model	8	$9.46032 * E^{14}$	$1.18254 * E^{14}$	534.32	0.9170
Error	387	$8.564904 * E^{13}$	$2.213153 * E^{11}$		
Corrected Total	395	$1.031681 * E^{15}$			

Stepwise Regression valde att utesluta de tre variablerna kön, garage och antal rum. Vi ser att trots detta så har förklaringsgraden enbart minskat med 0,05 procent. Vi fick nu alltså en modell med åtta stycken variabler. Här använder vi de förinställda kraven på signifikansnivån i programvaran SAS, vilket i det här fallet är 0,1500. Värt att observera från dessa beräckningar är att boarean bidrar med större delen av den totala förklaringsgraden. Boarea förklarar själv 89.04 procent av variationen. Det kan därför vara intressant att studera plotten mellan slutpriset och boarean för att se om vi även visuellt kan se denna samvariation. Plotten visar tydligt att detta är fallet.



Slutpriset i förhållande till boarean

För att få ytterligare förståelse för vårt datamaterial så valde vi att göra ett antal multipla regressionsmodeller med delar av de förklarande variablerna, för att se hur pass mycket förklaringsgraden påverkas samt vilka variabler som är signifikanta och inte. Vi kommer inte att redovisa alla dessa modeller här men det vi kan konstatera från dessa beräkningar är två saker. Det första är att den förklarande variabeln kön uteslutande är den variabel som är minst signifikant eller inte signifikant alls i alla olika modeller. På grund av detta tar vi beslutet att även utesluta denna variabel från fortsatta undersökningar. Det lämnar oss med tio stycken variabler att använda oss av i fortsättningen. Det andra som vi observerar är att boarean alltid bidrar med den större delen av den totala förklaringsgraden.

För att få en klar bild över den högsta förklaringsgraden vi kan få med de återstående tio variablerna så gör vi en multipel regression på slutpriset med dessa tio förklarande variabler. Resultatet följer nedan.

Tabell 3

Källa	FRGR	KVS	MKVS	F-value	R^2
Model	9	$9.461799 * E^{14}$	$9.461799 * E^{13}$	426.05	0.9171
Error	387	$8.550115 * E^{13}$	$2.220809 * E^{11}$		
Corrected Total	395	$1.031681 * E^{15}$			

Alltså är 91.71 procent den högsta förklaringsgrad som vi kan uppnå med dessa tio variabler. Eftersom boarean har en sådan kraftig inverkan på hur hög förklaringsgraden blir så väljer vi att fortsätta undersökningarna genom att först göra en enkel linjär regression med enbart boarea som förklarande variabel. Från denna regression sparar vi ner residualerna som en ny variabel. Vi fortsätter sedan med att göra multipel regression på residualerna med de övriga nio förklarande variablerna för att se hur mycket av variationen hos residualerna som dessa kan förklara. Detta gör vi dels för hela materialet men även uppdelat för både ettor, tvåor, treor och fyror, det vill säga fem stycken olika modeller. Vi har för få lägenheter representerade i materialet för att kunna göra detta för större lägenheter än fyror. Vi väljer att göra dessa multipla regressioner med Stepwise Regression. Tabell nummer 4 nedan visar resultatet för de enkla regressionerna med boarea och tabell nummer 5 visar resultatet av Stepwise Regression på residualerna för de fem olika datamaterialen.

Tabell 4 - Enkel linjär Regression med boarea

Material	Förklaringsgrad	Antal lägenheter
Hela materialet	0.8797	396
Ettor	0.5550	109
Tvåor	0.6573	138
Treor	0.4291	101
Fyror	0.3325	33

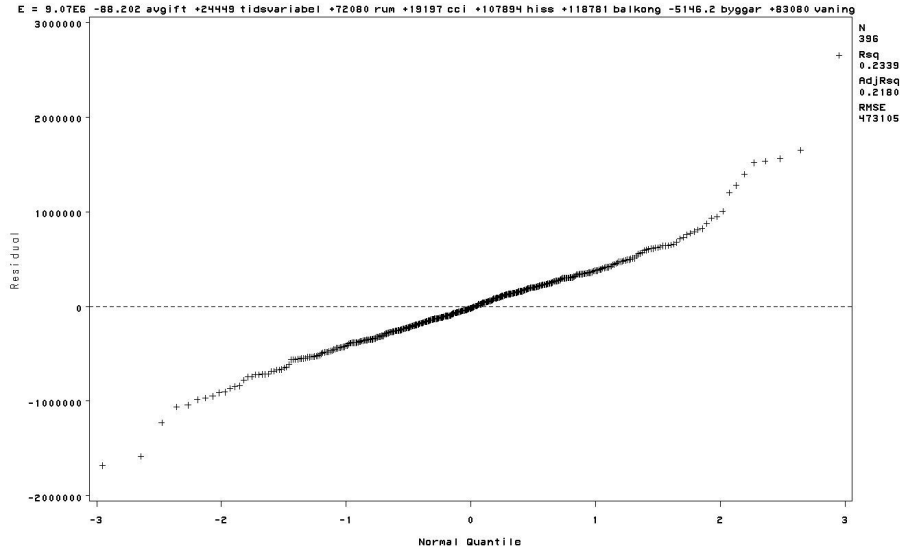
Tabell 5 - Stepwise Regression på residualerna

Material	Variabler i modellen vid Stepwise Regression	R^2
Hela materialet	våning, byggår, CCI, tid, hiss, avgift, rum och balkong	0.2438
Ettor	byggår, CCI, tid och våning	0.3056
Tvåor	garage, CCI, tid, byggår, våning, hiss och avgift	0.2948
Treor	byggår, våning, hiss, CCI, avgift, tid och balkong	0.3220
Fyror	våning, byggår, CCI och tid	0.6371

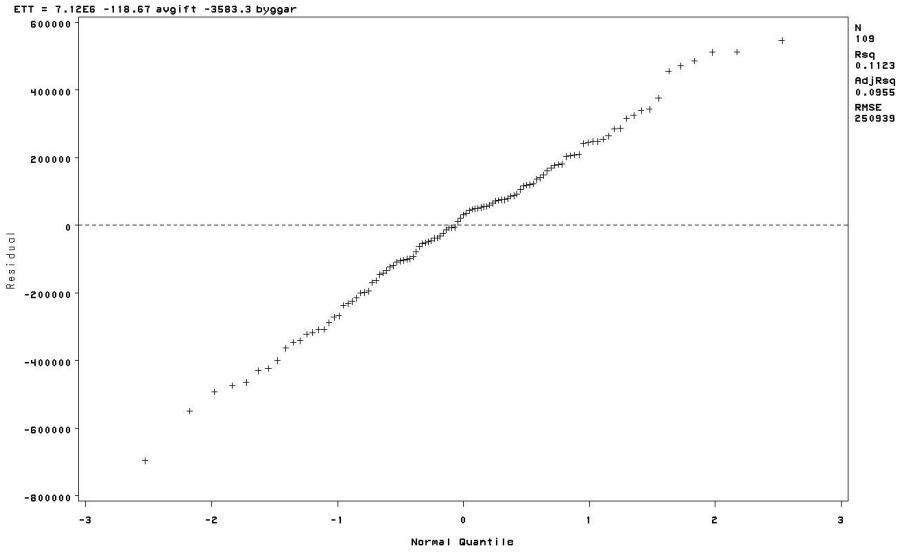
Som vi ser ovan så skiljer sig inte variablerna mellan de olika underlagen, efter Stepwise Regression, allt för mycket. Vi kan se att variablerna våning, byggår, CCI, och tid finns med i alla modellerna. När vi studerar hur variablerna påverkar slutpriset så är det i samma riktiningen som vi på förhand skulle kunna gissa. De enda variablerna som påvekar slutpriset negativt är avgiften och byggåret.

5.3.1 Test av normalfördelningsantagande

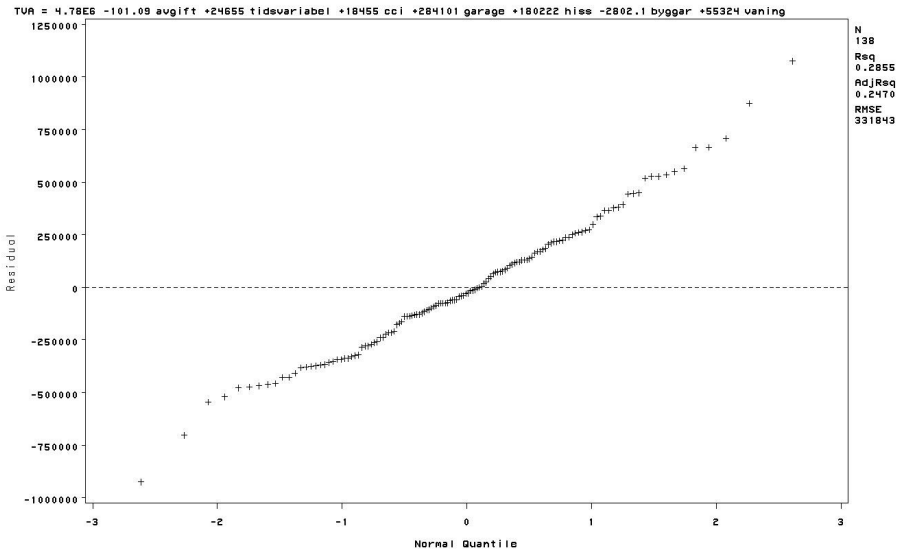
När vi använder oss av modellerna ovan så antar vi att residualerna är normalfördelade, N(0,σ^2). Vi vill därför undersöka om det är ett rimligt antagande. För att göra detta väljer vi att plotta residualerna från den enkla linjära regressionen med boarea mot normalfördelnings kvantilen för de fem olika datamaterialen, det vill säga för hela materialet, ettor, tvåor, treor och fyror. Det vi vill se är en rak linje, vilket skulle påvisa att vårt antagande är rimligt.



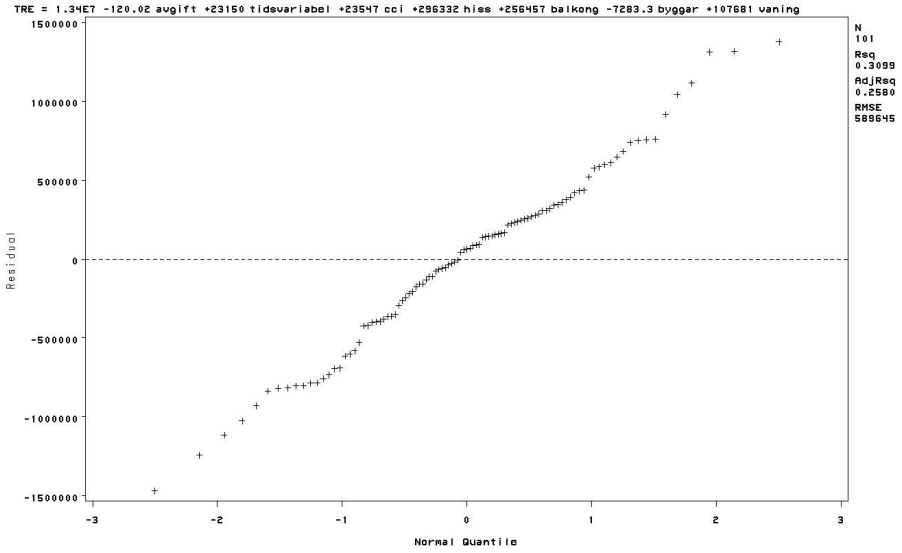
Residualerna i förhållande till normalfördelningskvantilen för hela materialet



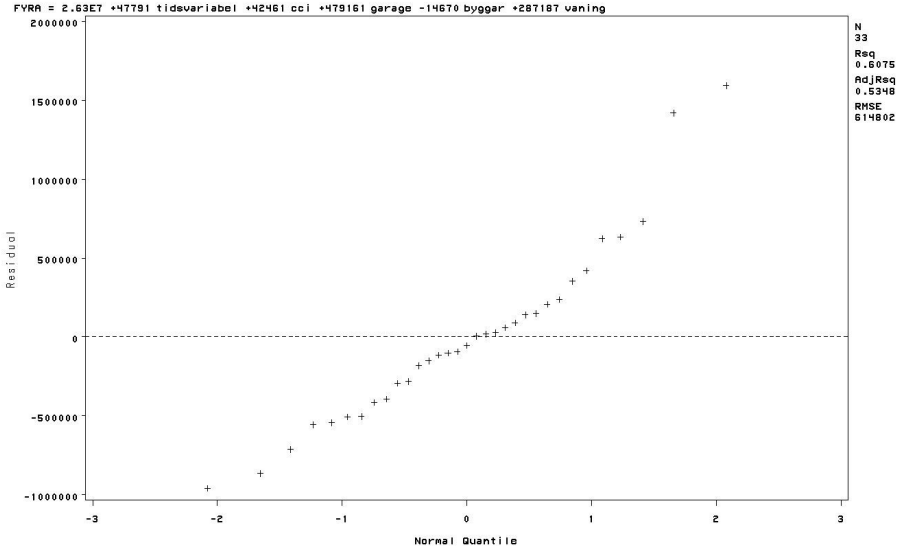
Residualerna i förhållande till normalfödelningskvantilen för ettor



Residualerna i förhållande tillcnormalfördelningskvantilen för tvåor



Residualerna i förhållande till normalfördelningskvantilen för treor



Residualerna i förhållande till normalfördelningskvantilen för fyror

Som vi ser i plottarna ovan så är linjerna i alla fem fallen tillräckligt raka för att vi ska kunna göra antagandet om residualernas normalfördelning i våra undersökningar. Dock kan vi se att de dyrare lägenheterna är dyrare och de

billiga lägenheterna billigare än om materialet vore helt normalfördelat.

5.4 Tillämpningar

5.4.1 Tillämpningar av Leave one out-metoden, Korsvalidering

Huvudsyftet med vår undersökning är att få fram en modell som på bästa sätt kan uppskatta ett framtida försäljningspris. En sådan modell behöver inte vara den modell som ger högst förklaringsgrad. Därför väljer vi att fortsätta genom att titta på PRESS-värdet istället, som är ett bättre jämförelsetal på hur bra en modell är i prediktionssyfte. I de här undersökningarna kommer vi använda oss av en variant av Korsvalidering, nämligen Leave One Out. Vi väljer nu att fortsätta med det ursprungliga materialet och inte materialet som bestod av residualerna.

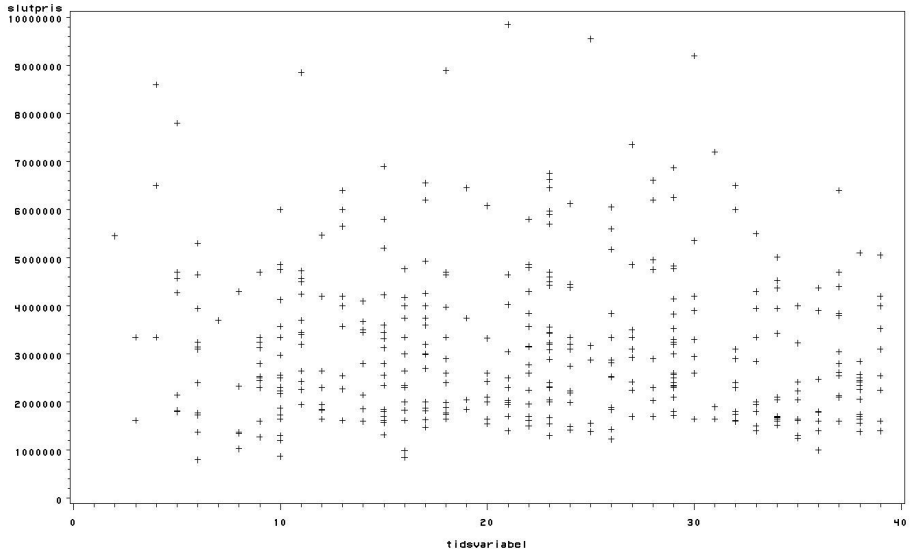
Vi har ingen möjlighet att testa alla möjliga modeller då vi har tio stycken förklarande variabler och därmed 2^{10}=1024 stycken olika modeller (varav en modell är helt utan förklarande variabler). Vi väljer att testa de modeller som vi anser troliga (baserat på våra tidigare observationer och teorier) för att se vilken av dessa modeller som kan ge det lägsta PRESS-värdet. Vi jämför även alla modeller för de fem olika materialen för att se ifall de skiljer sig mellan ettor, tvåor, treor och fyror. Resultatet av dessa undersökningar kan vi se i tabellerna nedan. Tabell nummer 7 visar resultatet för hela materialet och sedan följer för ettor, tvåor, treor och fyror.

Tabell 7 - Hela Materialet

Nr	Modell	R^2	PRESS
1	Boarea, Avg, Tid, Rum, CCI, Garage, Hiss, Blkg, Byggår, Vån	0.9171	$9.264971 * E^{13}$
2	Boarea, Avg, Tid, CCI, Byggår, Vån, Hiss, Blkg	0.9070	$9.112817 * E^{13}$
3	Boarea, Avg, Tid, CCI, Hiss, Garage, Byggår, Vån	0.9159	$9.279109 * E^{13}$
4	Boarea, Tid, CCI, Byggår, Vån, Hiss, Garage, Blkg	0.9138	$9.452568 * E^{13}$
5	Boarea, Tid, Avg, CCI, Rum, Byggår, Garage	0.9064	$1.026861 * E^{14}$
6	Boarea, Tid, CCI, Byggår, Vån, Blkg	0.9130	$9.385767 * E^{13}$
7	Boarea, Tid, Avg, CCI, Rum, Vån	0.9079	$9.965326 * E^{13}$
8	Boarea, Tid, CCI, Byggår, Vån	0.9115	$9.495311 * E^{13}$
9	Boarea, Tid, CCI, Byggår	0.9030	$1.033456 * E^{14}$
10	Boarea	0.8904	$1.147226 * E^{14}$

Här ser vi att den andra modellen ger det lägsta värdet på PRESS statistikan. Det är en modell med åtta stycken variabler och vi får PRESS-värdet: $9.112817 * E^{13}$. Detta motsvaras av ett fel på 479 710 kronor. En annan intressant modell är modell nummer åtta. Den består utav fem variabler och har ett PRESS-värde på $9.495311 * E^{13}$. Detta resulterar i ett fel på 489 674 kronor. Dessa belopp låter skrämmande stora men när vi granskar plotten

nedan över alla lägenheternas slutpris ser vi att de flesta lägenheter har ett slutpris mellan två och sex miljoner.



Slutpriset i förhållande till tiden för hela materialet

Om vi generaliserar utifrån bilden ser det ut som att de flesta lägenheter såldes kring tre miljoner. Då motsvarar ett fel på 500 000 kronor 17 procent av försäljningspriset. Alltså är inte våra fel så skrämmande stora som det först verkar, men de är större än önskvärt.

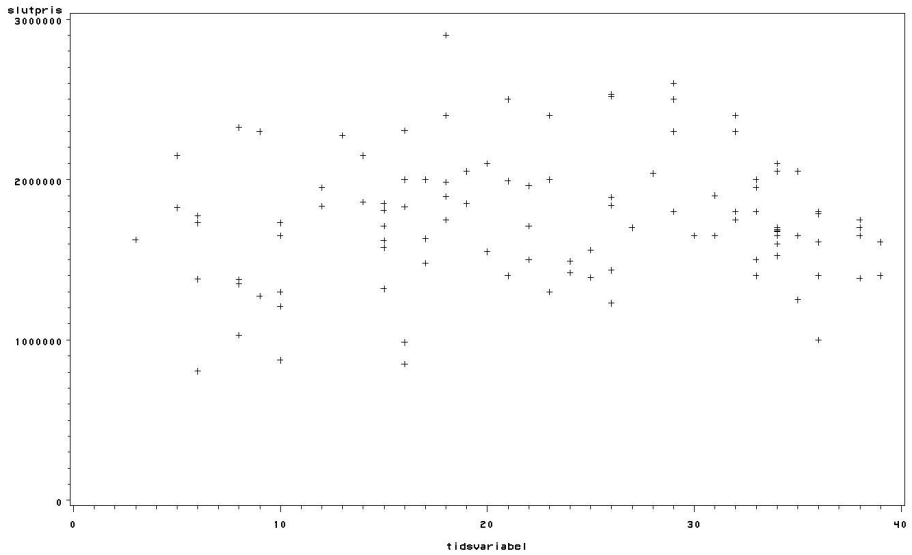
Tabell 8 - Ettor

Nr	Modell	R^2	PRESS
1	Boarea, Avg, Tid, Rum, CCI, Garage, Hiss, Blkg, Byggår, Vån	0.7145	$5.831517 * E^{12}$
2	Boarea, Avg, Tid, CCI, Byggår, Vån, Hiss, Blkg	0.6968	$6.091982 * E^{12}$
3	Boarea, Avg, Tid, CCI, Hiss, Garage, Byggår, Vån	0.6984	$5.959193 * E^{12}$
4	Boarea, Tid, CCI, Byggår, Vån, Hiss, Garage, Blkg	0.6568	$6.734752 * E^{12}$
5	Boarea, Tid, Avg, CCI, Rum, Byggår, Garage	0.7035	$5.70837 * E^{12}$
6	Boarea, Tid, CCI, Byggår, Vån, Blkg	0.6543	$6.659206 * E^{12}$
7	Boarea, Tid, Avg, CCI, Rum, Vån	0.6843	$6.094461 * E^{12}$
8	Boarea, Tid, CCI, Byggår, Vån	0.6491	$6.645024 * E^{12}$
9	Boarea, Tid, CCI, Byggår	0.6418	$6.642721 * E^{12}$
10	Boarea	0.5527	$7.815687 * E^{12}$

När vi delat upp så att vi endast undersöker ettor ser vi att det inte är sam-

ma modell, som för hela materialet, som ger det lägsta PRESS-värdet. Här är det modell nummer fem, med sju förklarande variabler, som ger det lägsta värdet på $5.70837 * E^{12}$, detta motsvaras av ett fel på 228 846 kronor. Här är även modell nio intressant med ett PRESS-värde på $6.642721 * E^{12}$ som motsvaras av 246 865 kr. Denna modell har endast fyra stycken förklarande variabler.

Vi granskar plotten över alla ettors slutpriser för att skapa oss en bild av hur bra modeller vi har hittat.



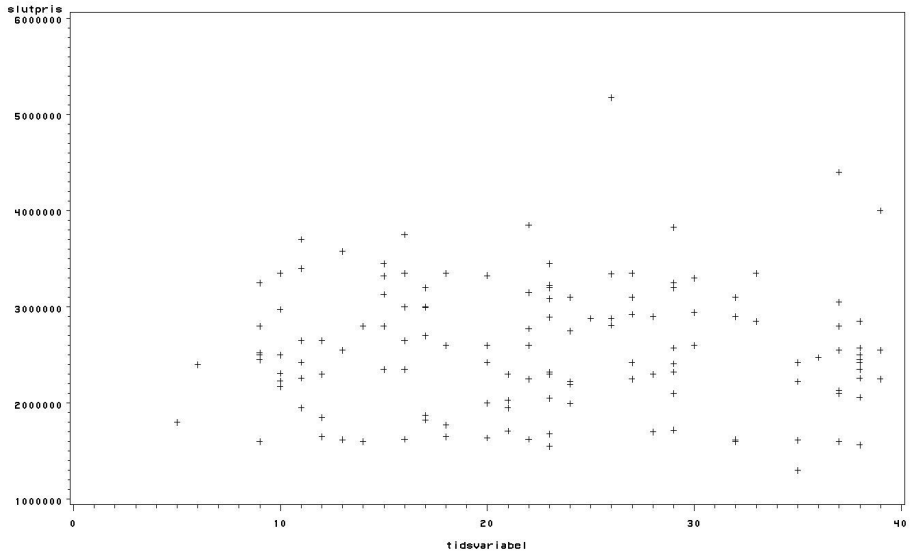
Slutpriset i förhållande till tiden för ettor

Här ser vi att ettorna har slutpris mellan 900 000 och 2 900 000 kronor. Om vi generaliserar och säger att de flesta lägenheterna ligger kring 1 800 000 kronor så motsvarar ett fel på 230 000 cirka 13 procent och ett fel på 250 000 ett fel på 14 procent.

Tabell 9 - Tvåor

Nr	Modell	R^2	PRESS
1	Boarea, Avg, Tid, Rum, CCI, Garage, Hiss, Blkg, Byggår, Vån	0.7748	$1.6271447 * E^{13}$
2	Boarea, Avg, Tid, CCI, Byggår, Vån, Hiss, Blkg	0.7655	$1.647563 * E^{13}$
3	Boarea, Avg, Tid, CCI, Hiss, Garage, Byggår, Vån	0.7740	$1.573928 * E^{13}$
4	Boarea, Tid, CCI, Byggår, Vån, Hiss, Garage, Blkg	0.7564	$1.690485 * E^{13}$
5	Boarea, Tid, Avg, CCI, Rum, Byggår, Garage	0.7432	$1.773478 * E^{13}$
6	Boarea, Tid, CCI, Byggår, Vån, Blkg	0.7375	$1.770102 * E^{13}$
7	Boarea, Tid, Avg, CCI, Rum, Vån	0.7399	$1.766332 * E^{13}$
8	Boarea, Tid, CCI, Byggår, Vån	0.7312	$1.78109 * E^{13}$
9	Boarea, Tid, CCI, Byggår	0.7159	$1.861962 * E^{13}$
10	Boarea	0.6681	$2.084148 * E^{13}$

Vi gör nu samma undersökningar för tvåor och kommer fram till att det är modell nummer tre, med åtta variabler, som ger det lägsta PRESS-värdet, nämligen $1.573928 * E^{13}$. Detta motsvaras av ett fel på 337 717 kr. Även modell nummer åtta är intressant då den har fem förklarande variabler och ger ett PRESS-värde på $1.78109 * E^{13}$, som motsvaras av ett fel på 359 255 kronor. Vi studerar plotten för tvåornas slutpris som funktion av tiden.



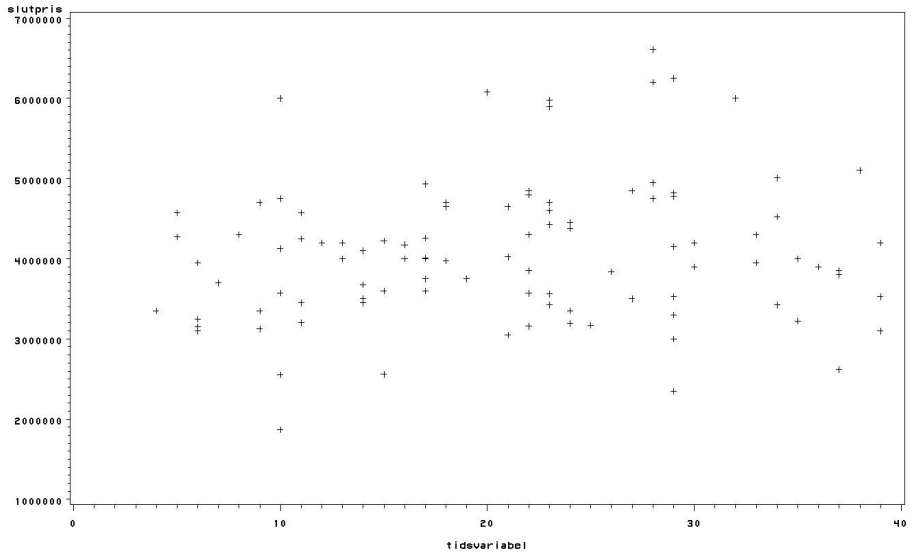
Slutpriset i förhållande till tiden för tvåor

Vi ser här att de flesta tvåorna har ett försäljningspris mellan 1 500 000 och 4 000 000 kronor. Vi generaliserar och säger att de flesta ligger kring 2 800 000 kronor. Då motsvarar ett fel på 340 000 kronor 12 procent och ett fel på 360 000 kronor 13 procent av slutpriset.

Tabell 10 - Treor

Nr	Modell	R^2	PRESS
1	Boarea, Avg, Tid, Rum, CCI, Garage, Hiss, Blkg, Byggår, Vån	0.6110	$4.107299 * E^{13}$
2	Boarea, Avg, Tid, CCI, Byggår, Vån, Hiss, Blkg	0.6085	$3.870603 * E^{13}$
3	Boarea, Avg, Tid, CCI, Hiss, Garage, Byggår, Vån	0.5932	$4.107491 * E^{13}$
4	Boarea, Tid, CCI, Byggår, Vån, Hiss, Garage, Blkg	0.5935	$4.094442 * E^{13}$
5	Boarea, Tid, Avg, CCI, Rum, Byggår, Garage	0.5201	$4.799656 * E^{13}$
6	Boarea, Tid, CCI, Byggår, Vån, Blkg	0.5702	$4.084802 * E^{13}$
7	Boarea, Tid, Avg, CCI, Rum, Vån	0.5049	$4.738185 * E^{13}$
8	Boarea, Tid, CCI, Byggår, Vån	0.5512	$4.174425 * E^{13}$
9	Boarea, Tid, CCI, Byggår	0.4979	$4.558346 * E^{13}$
10	Boarea	0.4284	$4.879072 * E^{13}$

När vi tittar på resultaten för treor kan vi ur tabellen utläsa att det är modell nummer två, med åtta förklarande variabler, som ger klart lägst PRESS-värde, nämligen $3.870603 * E^{13}$. Detta motsvaras av ett fel på 619 054 kronor. Precis som för tvåorna är modell åtta intressant även här. Den har fem förklarande variabler och ett PRESS-värde på $4.174425 * E^{13}$ som motsvarar ett fel på 642 891 kronor. Vi granskar plotten över treornas slutpris som funktion av tiden.



Slutpriset i förhållande till tiden för treor

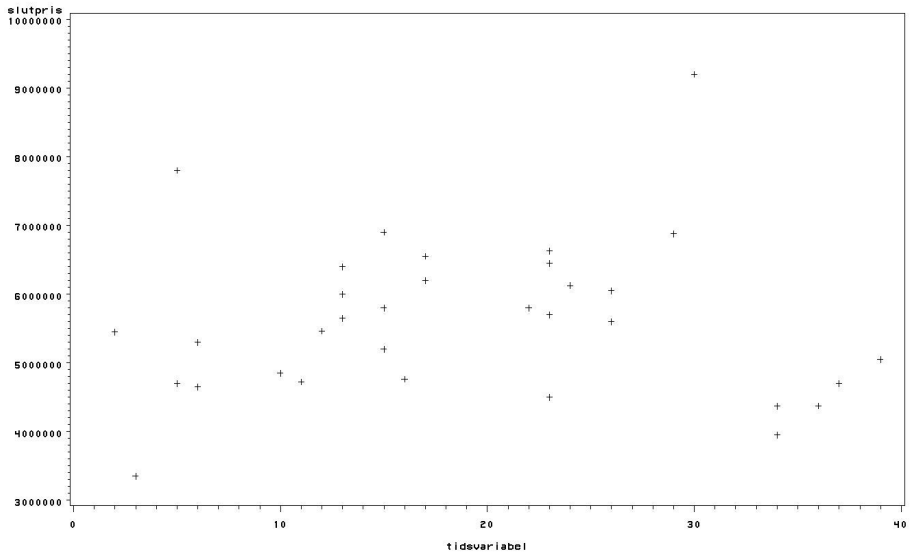
De flesta slutpriser ligger mellan 3 000 000 och 5 000 000 kronor. Om vi som i de övriga fallen först generaliserar genom att säga att de flesta ligger kring 4 000 000 kronor så motsvarar ett fel på 620 000 kr 15.5 procent av slutpriset och ett fel på 640 000 ett fel på 16 procent av slutpriset.

Tabell 11 - Fyror

Nr	Modell	R^2	PRESS
1	Boarea, Avg, Tid, Rum, CCI, Garage, Hiss, Blkg, Byggår, Vån	0.7724	$2.293438 * E^{13}$
2	Boarea, Avg, Tid, CCI, Byggår, Vån, Hiss, Blkg	0.7420	$2.163773 * E^{13}$
3	Boarea, Avg, Tid, CCI, Hiss, Garage, Byggår, Vån	0.7646	$1.956889 * E^{13}$
4	Boarea, Tid, CCI, Byggår, Vån, Hiss, Garage, Blkg	0.7638	$2.056 * E^{13}$
5	Boarea, Tid, Avg, CCI, Rum, Byggår, Garage	0.5498	$3.417128 * E^{13}$
6	Boarea, Tid, CCI, Byggår, Vån, Blkg	0.7407	$1.852998 * E^{13}$
7	Boarea, Tid, Avg, CCI, Rum, Vån	0.6220	$2.604913 * E^{13}$
8	Boarea, Tid, CCI, Byggår, Vån	0.7401	$1.722386 * E^{13}$
9	Boarea, Tid, CCI, Byggår	0.5411	$2.835219 * E^{13}$
10	Boarea	0.3906	$3.034549 * E^{13}$

Då vi studerar tabellen för fyrorna ser vi att modell nummer åtta ger helt klart lägst PRESS-värde på $1.722386 * E^{13}$. Då får vi ett fel på 722 451 kronor. Det finns inga andra modeller som har färre antal förklarande variabler och ett PRESS-värde som går att jämföra med modell nummer åttas. Vi

väljer att som i de övriga fallen granska plotten för slutpriset som funktion av tiden för alla lägenheter med fyra rum.

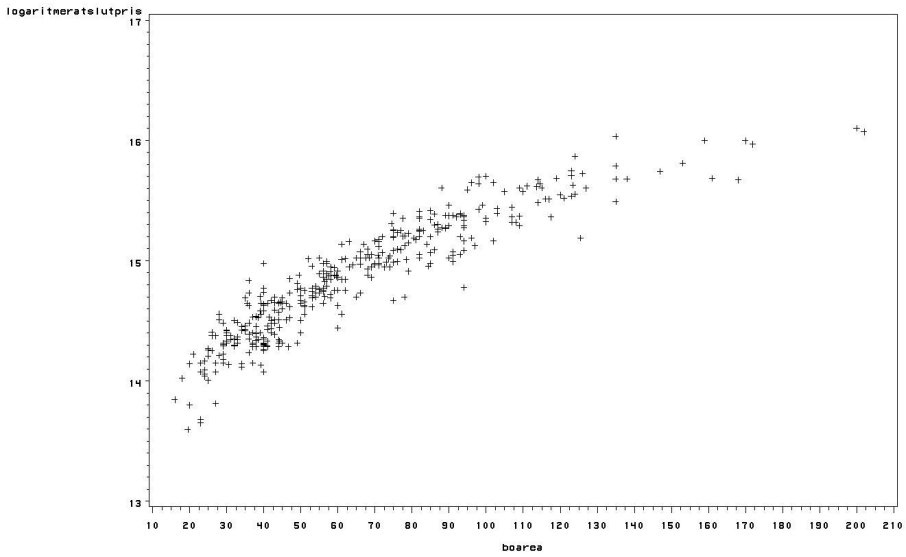


Slutpriset i förhållande till tiden för fyror

Vi ser att de flesta lägenheterna ligger mellan 4 500 000 och 7 000 000 kronor. Vi generaliserar och säger att de flesta ligger kring 6 000 000 kronor. Då motsvarar ett fel på 720 000 kronor 12 procent av slutpriset.

5.4.2 Alternativ responsvariabel

Innan vi utvärderar modellerna som vi har fått fram så vill vi testa att använda det logaritmerade slutpriset som responsvariabel istället för slutpriset. Då vi är medvetna om boareans betydelse väljer vi först att studera plotten mellan dessa två variabler. Vi har nu med hela materialet i undersökningarna.



Logaritmerat slutpris i förhållande till boarean för hela materialet

I plotten ser vi ett tydligt krökt samband. Vi ser att priset växer exponentiellt med boarean, men mattas av för de dyrare lägenheterna. Det här sambandet skulle kunna beskrivas som en andragradsfunktion av boarea. Eller ekvivalent, som ett regressionssammanband med två förklarande variabler, boarea och boarea i kvadrat. Vi får då följande ANOVA tabell.

Tabell 12

Källa	FRGR	KVS	MKVS	F-value	R^2
Model	2	81.16137	40.58069	1577.63	0.8892
Error	393	10.10894	0.02574		
Corrected Total	395	91.27031			

När vi sedan utför Leave One Out, Korsvalidering, på materialet så får vi ett PRESS värde på 10.28680. Nu ska vi dra till minnes att slutpriset var logaritmerat, alltså kan vi inte jämföra denna modell med de övriga modellerna. För att kunna göra en jämförelse får vi göra ytterligare en korsvalidering på dessa modeller men då vi har logaritmerat slutpris som responsvariabel istället. Resultaten av detta följer i tabell 13.

Tabell 13 - Hela materialet

Nr	Modell	R^2	PRESS
1	Boarea, Avg, Tid, Rum, CCI, Garage, Hiss, Blkg, Byggår, Vån	0.8685	13.02301
2	Boarea, Avg, Tid, CCI, Byggår, Vån, Hiss, Blkg	0.8675	12.90753
3	Boarea, Avg, Tid, CCI, Hiss, Garage, Byggår, Vån	0.8659	13.06399
4	Boarea, Tid, CCI, Byggår, Vån, Hiss, Garage, Blkg	0.8657	13.01212
5	Boarea, Tid, Avg, CCI, Rum, Byggår, Garage	0.8609	13.56025
6	Boarea, Tid, CCI, Byggår, Vån, Blkg	0.8643	13.01500
7	Boarea, Tid, Avg, CCI, Rum, Vån	0.8632	13.21247
8	Boarea, Tid, CCI, Byggår, Vån	0.8614	13.22643
9	Boarea, Tid, CCI, Byggår	0.8574	13.54099
10	Boarea	0.8451	14.40187
11	Boarea, Boarea^2	0.8892	10.28680

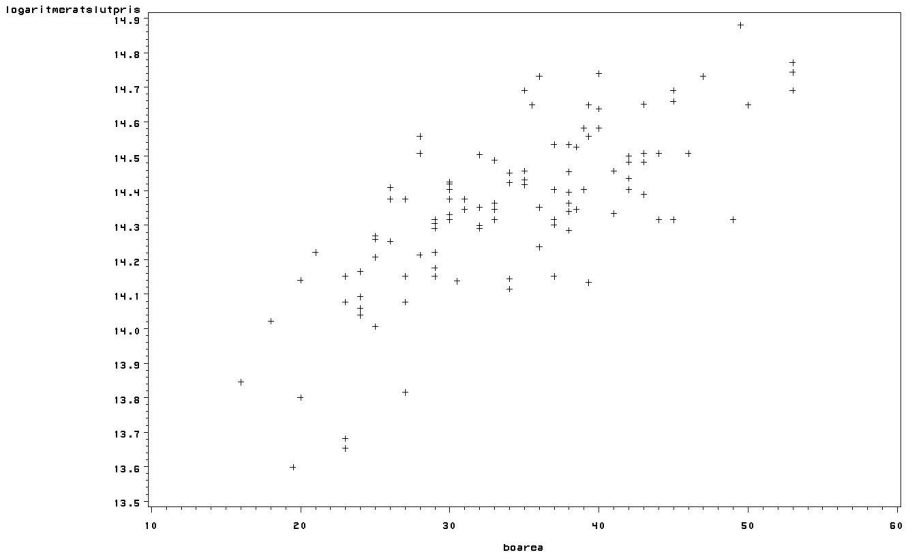
När vi använder det logaritemerade slutpriset som responsvariabel så bör vi inte omvandla resultaten till kronor. Förklaringsgraden skiljer sig nu från tidigare eftersom vi har bytt responsvariabel. Från tabellen kan vi utläsa två intressanta saker. Det första är att det är samma modell som ger det lägsta värdet på PRESS statistikan då vi har logaritmerat slutpris som responsvariabel istället för slutpris, nämligen modell nummer 2. Det andra är att alla modeller har ett högre värde på PRESS än vår nya modell, som finns representerad längst ner i tabellen, modell nummer elva.

Vi vill därför undersöka om detta är fallet även för det uppdelade materialet mellan ettor, tvåor, treor och fyror. Resultatet kan ses i tabellerna 14-17.

Tabell 14 - Ettor

Nr	Modell	R^2	PRESS
1	Boarea, Avg, Tid, Rum, CCI, Garage, Hiss, Blkg, Byggår, Vån	0.6813	2.36634
2	Boarea, Avg, Tid, CCI, Byggår, Vån, Hiss, Blkg	0.6602	2.48931
3	Boarea, Avg, Tid, CCI, Hiss, Garage, Byggår, Vån	0.6619	2.42861
4	Boarea, Tid, CCI, Byggår, Vån, Hiss, Garage, Blkg	0.6344	2.62610
5	Boarea, Tid, Avg, CCI, Rum, Byggår, Garage	0.6663	2.33282
6	Boarea, Tid, CCI, Byggår, Vån, Blkg	0.6316	2.59802
7	Boarea, Tid, Avg, CCI, Rum, Vån	0.6605	2.37818
8	Boarea, Tid, CCI, Byggår, Vån	0.6271	2.57970
9	Boarea, Tid, CCI, Byggår	0.6165	2.59750
10	Boarea	0.5409	2.92921
11	Boarea, Boarea^2	0.5642	2.83745

Här kan vi se att det precis som innan är modell nummer fem som ger det lägsta PRESS-värdet. Det som skiljer sig nu är att den nya modellen inte ger ett lägre PRESS-värde vilket den gjorde för hela materialet. Denna observation kan förklaras med att vi inte har ett lika tydligt krökt samband mellan det logaritmerade slutpriset och boarean, då vi endast undersöker ettor.



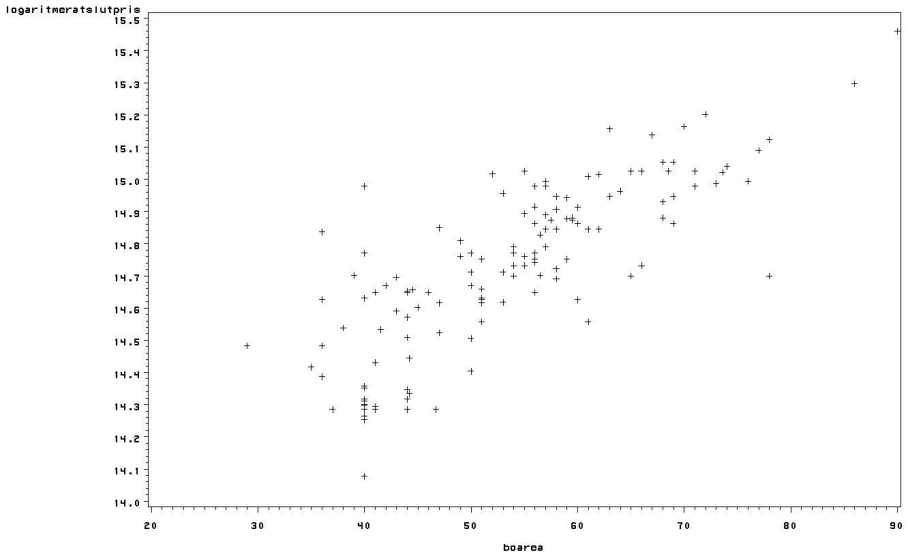
Logaritmerat slutpris i förhållande till boarean för ettor

Dock är skillnaden enbart marginell mellan PRESS-värdena och med tanke på antalet förklarande variabler i de båda modellerna, anses fortfarande den nya modellen vara ett bra alternativ. Vi gör nu samma undersökning fast för tvåor.

Tabell 15 - Tvåor

Nr	Modell	R^2	PRESS
1	Boarea, Avg, Tid, Rum, CCI, Garage, Hiss, Blkg, Byggår, Vån	0.7563	2.66694
2	Boarea, Avg, Tid, CCI, Byggår, Vån, Hiss, Blkg	0.7446	2.73642
3	Boarea, Avg, Tid, CCI, Hiss, Garage, Byggår, Vån	0.7523	2.62688
4	Boarea, Tid, CCI, Byggår, Vån, Hiss, Garage, Blkg	0.7430	2.70932
5	Boarea, Tid, Avg, CCI, Rum, Byggår, Garage	0.7172	2.95726
6	Boarea, Tid, CCI, Byggår, Vån, Blkg	0.7216	2.86297
7	Boarea, Tid, Avg, CCI, Rum, Vån	0.7114	2.97833
8	Boarea, Tid, CCI, Byggår, Vån	0.7072	2.96707
9	Boarea, Tid, CCI, Byggår	0.6898	3.10900
10	Boarea	0.6409	3.44739
11	Boarea, Boarea^2	0.6433	3.50243

Precis som för ettorna kan vi här se att det är samma modell som tidigare, modell tre, som ger det lägsta PRESS-värdet och att PRESS-värdet för den nya modellen inte heller blir lägst. Vilket har sin förklaring i att sambandet mellan logaritmerat slutpris och boarea inte heller här är lika tydligt.

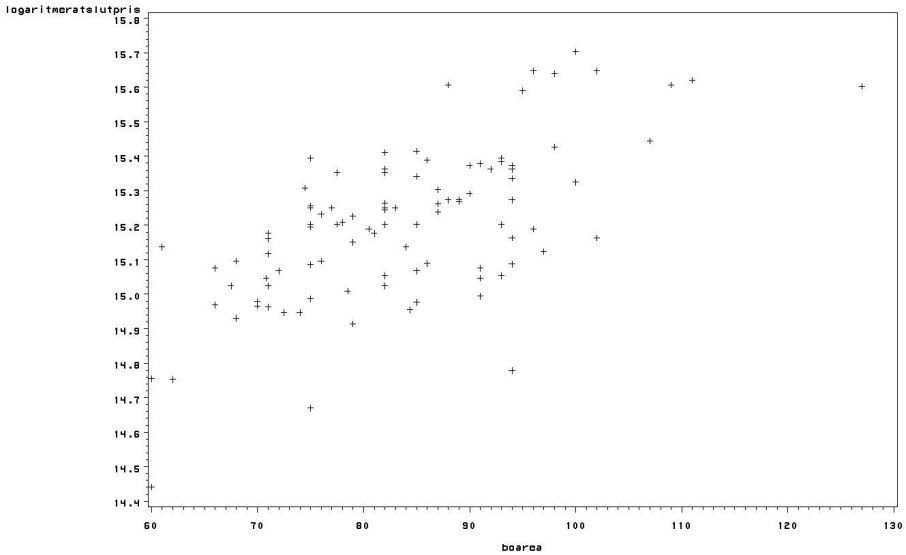


Logaritmerat slutpris i förhållande till boarean för tvåor

Vi kan dra samma slutsatser när vi studerar plottarna och tabellerna för treor och fyror. Här ger modell två det lägsta PRESS värdet för treorna och modell åtta det lägsta för fyrorna. Inte heller här ger den nya modellen den lägsta PRESS värdet. Vi ser inte heller samma krökta samband för något av materialen.

Tabell 16-Treor

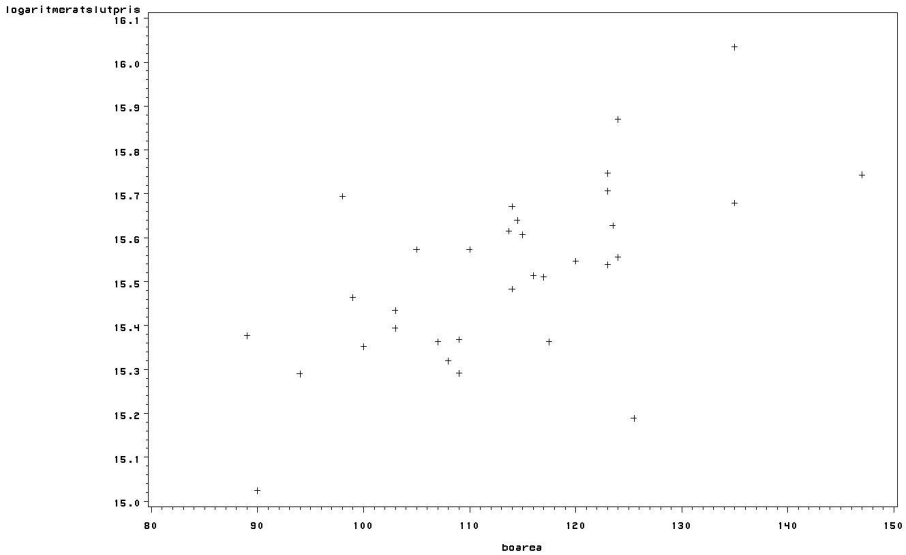
Nr	Modell	R^2	PRESS
1	Boarea, Avg, Tid, Rum, CCI, Garage, Hiss, Blkg, Byggår, Vån	0.5872	2.59495
2	Boarea, Avg, Tid, CCI, Byggår, Vån, Hiss, Blkg	0.5851	2.45842
3	Boarea, Avg, Tid, CCI, Hiss, Garage, Byggår, Vån	0.5700	2.60248
4	Boarea, Tid, CCI, Byggår, Vån, Hiss, Garage, Blkg	0.5682	2.61001
5	Boarea, Tid, Avg, CCI, Rum, Byggår, Garage	0.5067	2.92854
6	Boarea, Tid, CCI, Byggår, Vån, Blkg	0.5541	2.53438
7	Boarea, Tid, Avg, CCI, Rum, Vån	0.4943	2.87658
8	Boarea, Tid, CCI, Byggår, Vån	0.5383	2.57020
9	Boarea, Tid, CCI, Byggår	0.4831	2.81204
10	Boarea	0.4128	3.01029
11	Boarea, $Boarea^2$	0.4181	3.03283



Logaritmerat slutpris i förhållande till boarea för treor

Tabell 17 - Fyror

Nr	Modell	R^2	PRESS
1	Boarea, Avg, Tid, Rum, CCI, Garage, Hiss, Blkg, Byggår, Vån	0.7950	0.59794
2	Boarea, Avg, Tid, CCI, Byggår, Vån, Hiss, Blkg	0.7694	0.56604
3	Boarea, Avg, Tid, CCI, Hiss, Garage, Byggår, Vån	0.7885	0.53045
4	Boarea, Tid, CCI, Byggår, Vån, Hiss, Garage, Blkg	0.7888	0.53376
5	Boarea, Tid, Avg, CCI, Rum, Byggår, Garage	0.5936	0.94869
6	Boarea, Tid, CCI, Byggår, Vån, Blkg	0.7646	0.50081
7	Boarea, Tid, Avg, CCI, Rum, Vån	0.6733	0.68095
8	Boarea, Tid, CCI, Byggår, Vån	0.7646	0.46812
9	Boarea, Tid, CCI, Byggår	0.5792	0.80416
10	Boarea	0.3883	0.92543
11	Boarea, $Boarea^2$	0.3892	1.01268



Logaritmerat slutpris i förhållande till boarea för fyror

6 Slutsatser

Syftet med uppsatsen är att hitta en modell som vi kan använda för att prediktera framtida försäljningspriser på lägenheter i Stockholms innerstad. I bedömningen av vilken modell som är den mest lämpade i detta syfte så använder vi oss av fakta kring modellens PRESS-värde, förklaringsgrad och antalet förklarande variabler. Frågan är då vid vilken av dessa faktorer vi ska lägga störst vikt. Vid prediktion är inte förklaringsgraden av lika stort intressant då vi inte är ute efter att undersöka vilka variabler som påverkar priset eller i vilken utsträckning de kan förklara variationen. Därför väljer vi att lägga minst vikt vid förklaringsgraden. I våra undersökningar har vi naturligt valt att lägga störst vikt vid PRESS-värdet. Men den modell som ger lägst PRESS-värde kommer inte självfallet vara den vi anser mest lämpad till vårt syfte. Vi kommer även att ta hänsyn till antalet förklarande variabler. Då en modell har obetydligt högre PRESS-värde men betydligt färre förklarande variabler så är denna modell en mer lämpad kandidat. När vi i framtiden vill tillämpa modellen så önskar vi en så lätthanterlig modell som möjligt. Det är naturligt lättare att söka upp värden på färre antal variabler. Hur stor ökning i PRESS-värdet man kan tänka sig för att bli av med ytterligare en variabel är "individuellt". Det finns inga riktlinjer och bedömningen är olika från fall till fall. Därför har vi valt modeller utifrån våra egna preferenser och det finns möjlighet att dra annorlunda slutsatser.

Då vi ska välja modell som representerar hela materialet så blir valet inte så svårt. Modellen med lägst PRESS-värde är även den med minst antal förklarande variabler, då vi studerar tabell nummer 13. Det var när vi ändrade responsvariabel till logaritmerat slutpris som vi hittade sambandet mellan boarean och det logaritmerade slutpriset. Vi fick en modell med en andragradsfunktion av boarean, modell nummer elva i tabell nummer 13. Modellen gav lägst PRESS-värde och innehöll enbart en förklarande variabel. I våra analyser är den här modellen den helt klar mest lämpade för att använda som prediktionsmodell då alla sorters lägenheter är inkluderade. Nedan visas ANOVA tabellen för denna modell samt parameterskattningarna.

Tabell 18 - Hela materialet

Källa	FRGR	KVS	MKVS	F-value	R^2
Model	2	81.16137	40.58069	1577.63	0.8892
Error	393	10.10894	0.02572		
Corrected Total	395	91.27031			

Parameterskattningarna

Variabel	FRGR	Parameter estimate	Standard error	t-value	Pr< $\|t\|$
Intercept	1	13.59799	0.03393	400.79	<0.0001
Boarea	1	0.02461	0.00089731	27.42	<0.0001
$Boarea^2$	1	-0.00006514	0.00000520	-12.51	<0.0001

I avsnittet för Korsvalidering beskriver vi hur vi kan med hjälp av PRESS-värdet beräkna en approximativ kvot mellan det verkliga priset och skattning av priset från den analyserade modellen. Vi använder formeln:

$$\frac{P}{\widehat{P}} \approx e^{\sqrt{PRESS/n}}$$

Applicerar vi det på modell nummer elva, med PRESS-värdet 10.28680 och n=396, får vi en kvot på 1,1749. Detta innebär att om det verkliga slutpriset blir P SEK kommer vår skattning hamna i intervallet $P/1.1749 \leq \widehat{P} \leq 1.1749 * P$. Det vill säga en felmarginal på cirka 17.5 procent.

För ettorna är valet inte lika självklart. Vi studerar nu tabell nummer 8 och 14. Modell nummer fem ger lägst PRESS-värde både då vi har slutpris respektive logaritmerat slutpris som responsvariabel. Dock har denna modell sju stycken förklarande variabler. Om vi tittar på skillnaden i PRESS-värde mellan de olika modellerna då logaritmerat slutpris är responsvariabel ser vi att modell elva, andragradsfunktionen, ger en acceptabel ökning av PRESS-värdet. Vi väljer därför att även denna gång använda denna modell. Ty en minskning av antalet förklarande variabler från sju till en är att föredra framför den marginella ökningen av PRESS-värdet. I tabellerna nedan redovisas ANOVA tabellen och parameterskattningarna.

Tabell 19 - Ettor

Källa	FRGR	KVS	MKVS	F-value	R^2
Model	2	3.45506	1.72753	68.61	0.5642
Error	106	2.66879	0.02518		
Corrected Total	108	6.12385			

Parameterskattningarna

Variabel	FRGR	Parameter estimate	Standard error	t-value	Pr< $\|t\|$
Intercept	1	13.11152	0.22251	58.93	<0.0001
Boarea	1	-0.05229	0.01308	4.0	0.0001
Boarea^2	1	-0.00044294	0.00018625	-2.38	0.0192

För att beräkna kvoten $P/\widehat{P}$ använder vi PRESS-värdet 2.83745 från modell nummer elva samt n=109. Då blir kvoten 1.1751. Alltså har vi en skattning inom intervallet $P/1.1751 \leq \widehat{P} \leq 1.1751 * P$ och även här en felmarginal på cirka 17.5 procent.

När vi studerar tabell nummer 9 och 15, då vi enbart undersöker tvåor, ser vi att modell tre ger det lägsta PRESS-värdet för båda responsvariablerna. Här ser vi dock att modell nummer elva har ett betydligt högre PRESS-värde jämfört med modell tre, när vi studerar tabell nummer 15. Tidigare resonemang om att välja denna modell ty vinsten i färre variabler kan enligt våran mening ej rättfärdigas i det här fallet. Dock har modell nummer åtta ett acceptabelt PRESS-värde och använder sig av fem förklarande variabler vilket är tre färre än modell nummer tre. Alltså anser vi att modell nummer åtta är den mest lämpade modellen för ändamålet. Nedan följer ANOVA tabell och parameterskattningar för denna modell.

Tabell 20 - Tvåor

Källa	FRGR	KVS	MKVS	F-value	R^2
Model	5	6.57167	1.31433	63.77	0.7072
Error	132	2.72069	0.02061		
Corrected Total	137	9.29236			

Parameterskattningarna

Variabel	FRGR	Parameter estimate	Standard error	t-value	Pr< $\|t\|$
Intercept	1	15.55299	0.76544	20.32	<0.0001
Boarea	1	0.01877	0.00112	16.79	<0.0001
Tid	1	0.01131	0.00293	3.86	0.0002
CCI	1	0.00819	0.00179	4.59	<0.0001
Byggår	1	-0.00114	0.00041340	-2.76	0.0067
Våning	1	0.02146	0.00766	2.80	0.0059

Kvoten $P/\widehat{P}$ beräknas med PRESS-värdet 2.96707 från modell nummer 8 och n=138. Resultatet blir 1.1579. Vår skattning ligger alltså inom intervallet $P/1.1579 \leq \widehat{P} \leq 1.1579 * P$ som resulterar i en felmarginal på cirka 16

procent.

Då vi studerar PRESS-värden för treorna, i tabell nummer 10 och 16, så är det modell nummer två som ger lägst PRESS-värde för båda responsvariablerna. Dock kan vi konstatera att denna modell innehåller hela åtta förklarande variabler. När vi studerar tabell nummer 16 ytterligare, ser vi att modell nummer tio med enbart boarea som förklarande variabel ger ett acceptabelt PRESS-värde och sänker antalet förklarande variabler från åtta till en. Vinsten i färre variabler anser vi större än förlusten i ökat värde på PRESS. Som modell för treorna väljer vi därför en enkel regressionsmodell med boarea. ANOVA tabellen och parameterskattningarna följer nedan.

Tabell 21 - Treor

Källa	FRGR	KVS	MKVS	F-value	R^2
Model	1	2.02583	2.02583	69.59	0.4128
Error	99	2.88216	0.02911		
Corrected Total	100	4.90799			

Parameterskattningarna

Variabel	FRGR	Parameter estimate	Standard error	t-value	Pr< $\|t\|$
Intercept	1	14.19523	0.12166	116.68	<0.0001
Boarea	1	0.01199	0.00144	8.34	<0.0001

För att beräkna kvoten $P/\widehat{P}$ använder vi oss av PRESS-värdet 3.01029 från modell nummer tio och n=101. Vi får en kvot på 1.1884. Intervallet för vår skattning blir därför $P/1.1884 \leq \widehat{P} \leq 1.1884 * P$ och en felmarginal på cirka 19 procent.

Vid samma analys fast för fyrorna ser vi i tabell 11 och 17 att det är modell nummer åtta med fem förklarande variabler som ger det lägsta PRESS-värdet för de båda responsvariablerna. Vi ser även att den enkla regressionsmodellen med enbart boarea, modell nummer tio, ger ett acceptabelt PRESS-värde men ändå betydligt högre. Här är avvägningen svår mellan de båda modellerna. Men vi anser att det är en för stor ökning av PRESS-värdet i jämförelse med vinsten av färre variabler. Vi väljer därför modell nummer åtta. Nedan redovisas ANOVA tabell och parameterskattningar för denna modell.

Tabell 22 - Fyror

Källa	FRGR	KVS	MKVS	F-value	R^2
Model	5	0.99546	0.19909	17.54	0.7646
Error	27	0.30647	0.01135		
Corrected Total	35	1.30193			

Parameterskattningarna

Variabel	FRGR	Parameter estimate	Standard error	t-value	Pr< $\|t\|$
Intercept	1	18.28579	1.03644	17.64	<0.0001
Boarea	1	0.00898	0.00149	6.01	<0.0001
Tid	1	0.01074	0.00338	3.18	0.0037
CCI	1	0.00934	0.00229	4.08	0.0004
Byggår	1	-0.00216	0.00054335	-3.98	0.0005
Våning	1	0.04662	0.01011	4.61	<0.0001

Vi använder PRESS-värdet 0.46812 från modell nummer åtta, i tabell nummer 17, och n=33 för att beräkna kvoten. Resultatet blir 1.1265 vilket innebär att vi får intervallet $P/1.1265 \leq \widehat{P} \leq 1.1265 * P$ för vår skattning och en felmarginal på cirka 13 procent.

Nedan följer en tabell över de beräknade felmarginalerna för de olika materialen.

Tabell 23-Felmarginal

Material	Felmarginal
Hela	17.5 procent
Ettor	17.5 procent
Tvåor	16.0 procent
Treor	19.0 procent
Fyror	13.0 procent

7 Diskussion

Statistiska modeller är ofta en enklare version av verkligheten och mer komplexa fenomen. Därför är det ett misstag att anta att en modell beskriver den exakta formationen av det bakomliggande fenomenet. Men detta innebär inte att modellen inte är användbar i vissa syften, i vårt fall prediktion. Syftet med modellvalet är alltså att hitta en modell som på bästa sätt tjänar vårt syfte med undersökningen, oavsett hur väl modellen beskriver de verkliga sambanden.

Vi valde att begränsa uppsatsen till att undersöka linjära modeller. Utan denna begränsning finns det oändligt med funktioner som på ett eller annat sätt kan användas som prediktionsmodell. Alltså kan vi inte utesluta att det kan finnas modeller som tjänar vårt syfte mer tillfredställande inom andra modellantaganden. Då vi inte har möjlighet att undersöka alla modeller inom vårt modellantagande (linjära modeller) utan tar ut modellerna utifrån våra preferenser och tidigare undersökningar, finns det även en risk att vi missat en mer tillfredställande modell bland de 1024 stycken möjliga.

Boareans signifikanta effekt på slutpriset förutspådde vi i början av denna undersökning. Våra dragna slutsatser visar tydligt att detta antagande är väl motiverat. Det går inte att bortse från att boarean är den variabel som onekligen har störst inverkan på priset. I tre av fem olika modeller för våra material är det endast boarean som används som förklarande variabel, dock ser modellerna inte identiska ut. Men två modeller använder även andra förklarande variabler. En frågeställning i början av arbetet var vilka fler variabler som kunde användas för att prediktera ett så korrekt slutpris som möjligt. Det visade sig att svaret på denna fråga därför beror på vilket material vi undersöker. Lägenheter av olika storlek attraherar olika klientgrupper och därmed blir även olika variabler signifikanta. Enligt våra slutsatser visar det sig vara motiverat att även undersöka värden på variablerna: CCI, Tid, Byggår och Våning.

Felmarginalerna i vår analys hamnar mellan 13-19 procent för de olika materialen. Intressenter vid en lägenhetsvisning eller en säljare eftertraktar säkerligen en mer precis modell men för en mäklare kan modellen vara användbar. En mäklarfirma skulle tillämpa modellen på ett stort antal försäljningar, och i det långa loppet skulle över- och underskattningar ta ut varandra. Vi har inte haft startpriset som en variabel i vår analys och en mäklare kan därför anse modellen behjälplig vid beslut om utgångspris vid försäljning.

Vidare finns det flera sätt att utveckla dessa undersökningar. En möjlig fortsättning är att undersöka en modell som tar hänsyn till den exakta

lokaliseringen av lägenheterna. En intressant variabel i en sådan undersökning skulle kunna vara statsdel eller postkoden, som begränsar området ytterligare.

Vi valde att dela upp materialet på antal rum. Vi kan även tänka oss att dela upp det på antal kvadratmeter. Det vill säga göra intervall av olika kvadratmeterstorlekar och dela in lägenheterna i olika grupper efter dessa preferenser.

Vid en större undersökning finns det även möjlighet att utöka både antalet variabler och antalet sålda lägenheter. Förslag till intressanta variabler för en liknande undersökning i Stockholms innerstad skulle kunna vara meter till tunnelbanan, meter till mataffär, meter till grönområde, öppen spis, renoverings behov, typ av säljare (familj, ensamstående, ung, äldre osv.) samt ålder på mäklare.

Referenser

[1] Freund, R. J. and Littell, R.C. SAS System for Regression, SAS Institute Inc., Cary, NC, USA, 1986.

[2] www.fastighetsbyran.se

[3] Geisser Seymour. Predictive Inference: An Introduction.

[4] Gut Allan. An Intermediate Course in probability, 2005.

[5] www.ida.liu.se/ 732G17/labs/lab1.pdf

[6] www.konjunkturinstitutet.se

[7] www.ltrr.arizona.edu/ dmeko/notes 12.pdf

[8] Sundberg Rolf. Kompendium i Tillämpad Matematisk Statistik, 1997.

[9] www.wikipedia.org (Least squares method, Coefficient of determination)

[10] Weisberg, S. Applied Linear Regression, 2nd ed., John Wiley, New York, 1985.