Stockholms universitet

Analys och prediktion av lägenhetspriser med multipel linjär regression

Hiyab Munir

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2016:16
Matematisk statistik
Juni 2016

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholms universitet

Matematisk statistik
Stockholms universitet
Kandidatuppsats **2016:16**
http://www.math.su.se/matstat

Analys och prediktion av lägenhetspriser med multipel linjär regression

Hiyab Munir*

Juni 2016

Sammanfattning

Vår avsikt med arbetet är att undersöka vilka faktorer som kan tänkas förklara slutpriset för sålda lägenheter i Huddinge. Vi utför undersökningen med regressionsanalys. Datamaterial för 150 sålda lägenheter år 2015 samlades in. Halva datamaterialet användes för att förklara slutpriset, medan andra halvan av datamaterialet användes för att testa modellernas prediktionsförmåga. Vi kom fram till två potentiella modeller som förklarar slutpriset. En modell med utropspris och avgift per kvadratmeter visade sig ha bäst förklaringsgrad.

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

Innehåll

Introduktion

Lägenhetspriser i Stockholms innerstad har ökat de senaste åren, vilket leder till att fler söker sig utåt, till kommuner i de yttre delarna av Stockholm. Priserna har i överlag varit lägre än i innerstaden, men som konsekvens av prisökningar i Stockholms innerstad, har prisökningar även skett i många andra områden, bl.a. Huddinge kommun.

Huddinge kommun har ca 100 000 invånare. Det är, efter Stockholms stad, den största kommunen i Stockholms län. Vi vill undersöka vad som påverkar slutpriserna för lägenheter i Huddinge kommun. Vi gör det genom att utföra en regressionsanalys i syfte att finna en regressionsmodell som beskriver förhållandet mellan slutpriset på en lägenhet och ett antal förklarande variabler som vi har till förfogande. Modellen kommer även att testas i prediktionssyfte.

Frågeställningar

Syftet med arbetet är att undersöka vilka faktorer som kan tänkas förklara slutpriset av lägenhetsförsäljningar i Huddinge kommun. Vi vill även undersöka vad som förklarar prisskillnaderna mellan utropspriset och slutpriset.

Vi vill besvara dessa frågor i vår analys

- Vilka variabler förklarar slutpriset av en lägenhetsförsäljing?
- Vad förklarar prisskillnaden mellan utropspriset och slutpriset?
- Kan vi prediktera framtida slutpriser av lägenhetsförsäljningar?

Teori

Definitioner av begrepp och metoder i teoriavsnittet refereras till kompendiet ”Lineära Statistiska Modeller” av Rolf Sundberg (2015) . Definitioner av parameterskattningar refereras till kompendiet ”Notes in Econometrics” av Patrik Andersson & Joanna Tyrcha (2014).

Linjär regression

Linjär regression används för att förklara sambandet mellan en responsvariabel och en eller flera förklarande variabler. För fortsatt analys av en regressionsmodell så ska vissa grundläggande villkor vara antagna

- Linjäritet – Det är ett linjärt samband mellan responsvariabeln och de förklarande variablerna
- Normalfördelade slumpfel – Modellens slumpfel är oberoende och normalfördelade
- Homoskedasticitet- Variationen mellan slumpfelen ska vara konstant

Multipel linjär regression

Vi definierar vår multipla regressionsmodell på formen

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \ldots + \beta_m x_{mi} + \varepsilon_i \, , i = 1, \ldots, n. \; j = 1, \ldots, m.$$

där α betecknar parametern för regressionsmodellens intercept.
Parametern β_j beskriver de förklarande variablernas (x_{ji}) effekt på responsvariabeln. Parametern ε_i betecknar slumpfelet, avvikelsen mellan de observerade slutpriserna, och de slutpriser regressionsmodellen ger upphov till. $\varepsilon_i \sim N(0, \sigma^2)$ och antas vara oberoende och likafördelade (i.i.d).

Regressionsmodellen kan även skrivas i matrisform på följande sätt

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

där

$$\boldsymbol{Y} = (y_1, \ldots, y_n)^{\boldsymbol{T}}, \boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & \ldots & x_{m1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & \ldots & x_{mn} \end{pmatrix}, \boldsymbol{\beta} = (\alpha, \beta_1, .., \beta_m)^{\boldsymbol{T}}, \boldsymbol{\varepsilon_i} = (\boldsymbol{\varepsilon_1}, \ldots, \boldsymbol{\varepsilon_n})^{\boldsymbol{T}}$$

Hypotestest

Flera hypotestest kommer att utföras för att undersöka vilka förklarande variabler som ska inkluderas i modellen. Testet går ut på att undersöka om de förklarande variablerna har en signifikant inverkan, dvs. om deras motsvarande parameter är skild från noll, givet de redan inkluderade variablerna.
Hypoteserna kan se ut på följande vis

$$H_0: \beta_j = 0$$

$$H_1: \beta_j \neq 0 \; för \; något \; j = 1, \dots, m$$

Parameterskattning

Den skattningsmetod som är mest lämplig att använda är Minsta-kvadratmetoden (MK-metoden). Syftet med MK-metoden är att finna ett skattat värde för parametrarna då residualernas kvadratsummor är minimerat, så att modellen ger en god anpassning till data.

Låt $\boldsymbol{e_j}$ beteckna residualerna, Kvadratsumman kan i matrisform uttryckas på följande vis

$$\sum_j e_j^2 = \boldsymbol{e^T e} = (\boldsymbol{Y} - \boldsymbol{X\beta})^{\boldsymbol{T}}(\boldsymbol{Y} - \boldsymbol{X\beta})$$

MK-metoden ger oss följande skattningar av $\boldsymbol{\beta}$

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

där $\widehat{\boldsymbol{\beta}} \sim \boldsymbol{N}(\boldsymbol{\beta}, \boldsymbol{\sigma^2}(\boldsymbol{X^T X})^{-1})$

Multikollinearitet

I situationer då vi har flera potentiella förklarande variabler, så är det möjligt att linjära samband förekommer inom de förklarande variablerna. Signifikanta förklarande variabler med linjära samband kan ge missvisande resultat då de kan indikera att de inte har en signifikant påverkan på responsvariabeln när de inkluderas tillsammans i modellen. En möjlig åtgärd är att man reducerar modellen genom exkludering eller variabeltransformation.

Förklaringsgrad

Förklaringsgraden är ett mått på hur god anpassning regressionsmodellen har till data. Måttet betecknas R^2, och anger den förklarande andelen av modellens totala variation. Vi definierar förklaringsgraden på följande sätt

$$R^2 = \frac{KVS(Regression)}{KVS(Total)} = 1 - \frac{KVS(Residual)}{KVS(Total)}.$$

Där

$$KVS(Regression) = \sum_i (\hat{y}_i - \bar{y})^2$$

$$KVS(Residual) = \sum_i (y_i - \hat{y}_i)^2$$

$$KVS(Total) = \sum_i (y_i - \bar{y})^2$$

där $\bar{y}$ betecknar medelvärdet av responsvariabeln och $\hat{y}_i$ betecknar responsvariabelns skattade värden.

Nackdelen med R^2 är att inkludering av fler förklarande variabler alltid ökar förklaringsgraden. Därmed kan vi använda den justerade förklaringsgraden (R^2_{adj}) som mäter hur stor variationsminskningen blir i modellen när en variabel inkluderas. Vi definierar den justerade förklaringsgraden på följande vis

$$R_{adj}^2 = 1 - \frac{\widehat{\sigma^2}}{\widehat{\sigma_0}}$$

där $\widehat{\sigma_0^2}$ är variansskattningen när ingen förklarande variabel är inkluderad i modellen. Relationen mellan förklaringsgraden och den justerade förklaringsgraden kan vi definiera på följande sätt

$$1 - R_{adj}^2 = \frac{(1 - R^2)(N - 1)}{(N - m - 1)}$$

VIF-faktor

VIF-faktorn anger hur stor ökning variansen blir för den skattade parametern $\widehat{\beta_j}$ då vi inkluderar en variabel i modellen. Om VIF-faktorn antar ett värde större än fem, så kan vi stöta på problem med multikollinearitet .
VIF-faktorn kan uttryckas på följande vis

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

Där R_j^2 förklarar andelen av variationen i variabeln X_j som förklaras av resterande förklarade variabler.

Stegvis variabelselektion

Stegvis variabelselektion kan användas när man har tillgång till ett större antal potentiella förklarande variabler. Syftet med variabelselektion är att finna en modell med färre variabler, som kan ge en bättre förklaring för den totala variationen av responsvariabeln. Det är tre metoder som är vanligast att använda.

Forward selection

Här startar proceduren med en modell utan förklarande variabler. Stegvist inkluderas den variabel med lägst p-värde. Proceduren slutar när alla inkluderade variabler har en signifikant inverkan.

Backward Selection

Proceduren startas med alla förklarande variabler inkluderade. I varje steg exkluderas den variabel med högst p-värde. Proceduren slutar när alla inkluderade variabler är signifikanta.

Stepwise Selection

Stepwise proceduren är en kombination av forward- och backward procedurer. Proceduren startar med alla förklarande variabler exkluderade. Precis som Forward selection kommer den variabel med lägst p-värde att inkluderas i varje steg, men alla variabler som redan är inkluderade kommer att testas ifall de fortfarande är signifikanta. De variabler som inte längre är signifikanta exkluderas.

Vi kommer att använda oss av stepwise selection i arbetet.

Logtransformation

Logaritmtransformering är en åtgärd som kan komma till nytta om responsvariabeln visar icke-linjära samband med förklarande variabler.
Logtransformering kan även vara effektiv för att undvika heteroskedasticitet.

En multiplikativ modell kan uttryckas som

$$Y = \alpha * e^{\beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \ldots + \beta_m x_{mi}} * \varepsilon$$

där $\alpha > 0, \varepsilon > 0$.

Vi logaritmerar modellen och får

$$Y' = \ln Y = \alpha' + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \ldots + \beta_m x_{mi} + \varepsilon'$$

där $\alpha' = \ln \alpha$, $\varepsilon' = \ln \varepsilon$.

Vilket är en linjär modell.

Prediktion

För att testa hur väl vår modell är anpassad, så ska vi testa att prediktera slutpriser med ett nytt datamaterial som underlag. Vi kommer att använda oss av samma parametrar samt parameterskattningar som vår slutliga modell. Vi får följande modell i matrisform

$$\widehat{\boldsymbol{Y}} = \boldsymbol{X_{ny}} * \widehat{\boldsymbol{\beta}}$$

Vi kommer att jämföra de predikterade residualkvadratsummor av våra potentiella modeller, där ett lägre värde ger en modell med en bättre anpassad prediktionsförmåga.

Vi definierar de predikterade residualkvadratsumman på följande vis

$$MSEP = \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{n}$$

kvadratroten ur $MSEP$ ger oss ett predikterat värde av feltermen.

$$RMSEP = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n}}$$

Data

Data som används i detta arbete samlades manuellt från Booli. Data för 150 sålda lägenheter i Huddinge kommun år 2015 samlades in. Halva datamaterialet (lägenheter sålda mellan januari-juni 2015) kommer att användas för att försöka skapa en lämplig modell. Den resterande halvan (juli-december 2015) av datamaterialet kommer att användas för att testa modellens prediktionsförmåga.

Variabler

$Slutpris$ – Kontinuerlig variabel som anger det slutliga priset lägenheten såldes för. Varierar mellan 1280000-4550000kr

$Utropspris$ – Kontinuerlig variabel som anger priset när annonsen lades upp, det pris budgivningen startar på. Varierar mellan 895 000-3 695 000kr.

Yta – Kontinuerlig variabel som beskriver lägenhetens storlek i enheten m^2. Varierar mellan 23-118m^2.

Rum – Antalet rum lägenheten har. Varierar mellan 1-5 rum.

$Byggår$ – Kontinuerlig variabel som beskriver antalet år sedan lägenheten byggdes. Varierar mellan 8-66 år.

$Avgift$ – Kontinuerlig variabel som anger kostnaden per månad för att bo i lägenheten. Varierar mellan 1096-8570kr

$Våning$ – Kategorisk variabel som anger vilken våning lägenheten ligger i. Varierar mellan 1-7 våningar.

$Område$ – Dummyvariabel som antar värdet 1 om lägenheten ligger i västra delen av Huddinge Kommun, och värdet 0 om den ligger i östra delen.

$Datum$ – Kontinuerlig variabel som anger datumet lägenheten såldes. Varierar mellan 1-12 (januari-December).

Variabeltransformation

I Analys-delen kommer vi fram till att transformera vissa variabler för att undvika multikollineäritet.

$AvgiftYta$ – Kontinuerlig variabel som beskriver avgiften per kvadratmeter. Varierar mellan 36-101kr/m^2.

$Rumyta$ – Kontinuerlig variabel som beskriver antalet kvadratmeter per rum. Varierar mellan 18.5 - 53m^2/rum

$Priskvot$ – Kontinuerlig variabel som beskriver kvoten mellan slutpriset och utropspriset. Vi definierar den som $Priskvot = \frac{Slutpris}{Utropspris}$.
Varierar mellan 0.90 – 1.59.

Analys och Resultat

Multikollinjäritet

Vi börjar med att undersöka ifall en eventuell multikollinjäritet kan uppstå mellan våra potentiella förklarande variabler. Vi börjar med att undersöka korrelationer mellan prediktorerna. Utifrån korrelationsmatrisen kan vi läsa av starka korrelationer mellan $Yta - Rum$, samt $Yta - Avgift$.
Vi undersöker de förklarande variablernas VIF-värden när alla variabler är inkluderade i modellen. De får följande VIF-värden

Variabel	**VIF-Värde**
Rum	6.33720
Yta	8.56578
Våning	1.33189
Byggår	1.89907
Avgift	6.39281
Utropspris	3.13656
Område	1.57671
Datum	1.15747

Tabell 1. VIF-värden när alla potentiella förklarande variabler inkluderas i modell med slutpris som responsvariabel.

Yta, Rum och $Avgift$ har VIF-värden högre än fem, det ger starka indikationer på att multikollineäritet uppstår om de variablerna inkluderas tillsammans. För att åtgärda detta utför vi variabeltransformationerna $RumYta = Yta/Rum$ och $AvgiftYta = Avgift/Yta$ som ersätter variablerna Rum och Avgift. Variabeln $Utropspris$ har som väntat en stark korrelation med Slutpriset. $Utropspris$ förklarar enskilt 86.08 % av variationen för slutpriset. Vi försöker anpassa en modell där utropspriset är exkluderat från analysen.

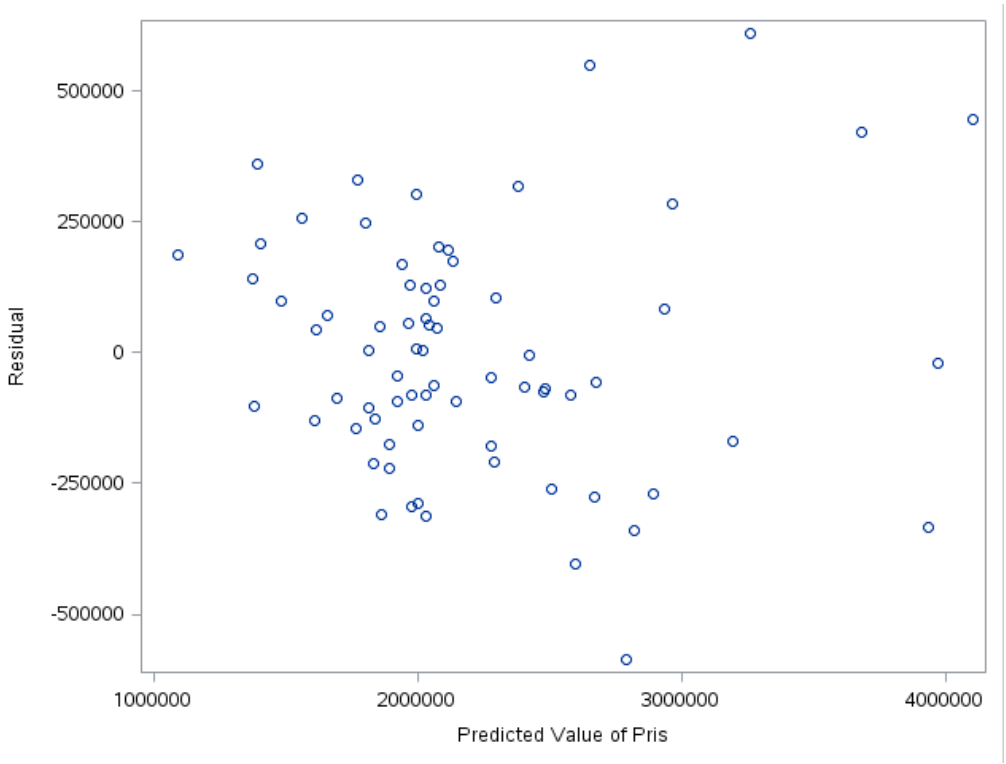
Multipel Linjär Regression

Första genomförandet

I vårt första genomförande inkluderar vi variabeln utropspris.
Vi anpassar en modell som innehåller alla förklarande variabler. Vi har då följande modell

$$\text{Slutpris}_i = \alpha + \beta_1 * Yta + \beta_2 * Utropspris + \beta_3 * RumYa + \beta_4 * AvgiftYta + \beta_5 * Byggår + \beta_6 * Våning + \beta_7 * Område + \beta_8 * Datum + \epsilon_i$$

Vi börjar med att undersöka plotten för residualerna mot predikterade värden för responsvariabeln Slutpris.



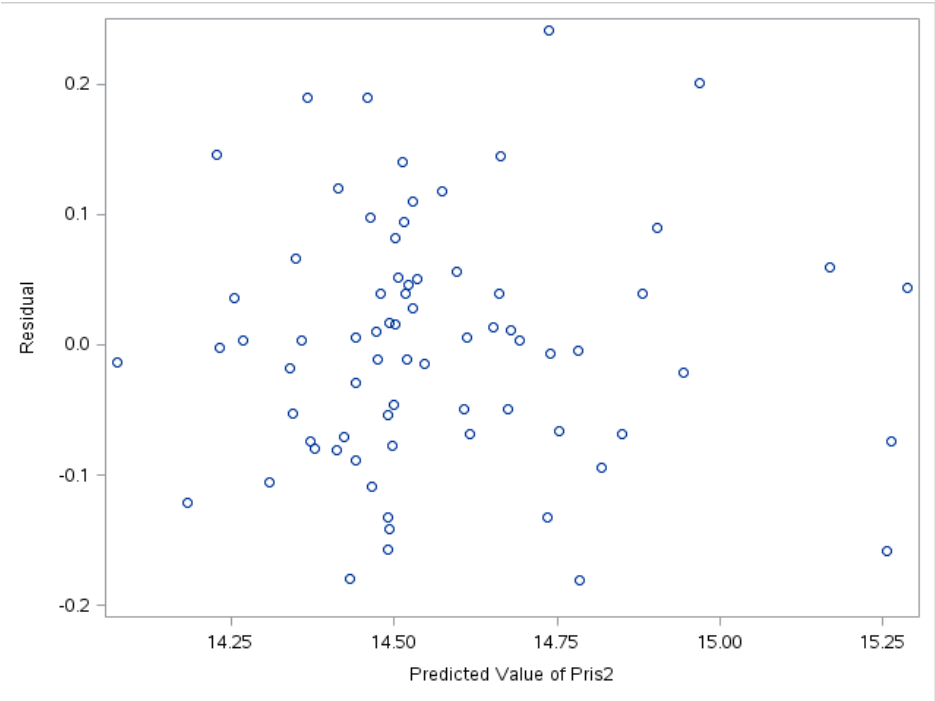
Figur 1. Plot som visar modellens residualer mot dess predikterade slutpris

Vi ser ovan att residualernas spridning tenderar att öka då Predikterade slutpriset ökar. Vi försöker åtgärda detta genom att utföra en logtransformation på responsvariabeln.

Vi har då följande additiva modell

$$\log(\text{Slutpris}_i) = \alpha + \beta_1 * Yta + \beta_2 * Utropspris + \beta_3 * RumYa + \beta_4 * AvgiftYta + \beta_5 * Byggår + \beta_6 * Våning + \beta_7 * Område + \beta_8 * Datum + \epsilon_i$$

Vi studerar om logaritmering av responsvariabeln förbättrar spridningen av modellens residualer.



Figur 2. Plot som visar modellens residualer mot dess predikterade slutpris efter logtransformering

Vi ser att spridningen av residualerna jämnade ut sig något. Logaritmering av responsvariabeln verkar vara en lämplig åtgärd i detta fall.

Stepwise selection

Vi har en additiv modell

$$\log(\text{Slutpris}_i) = \alpha + \beta_1 * Yta + \beta_2 * Utropspris + \beta_3 * RumYa + \beta_4 * AvgiftYta + \beta_5 * Byggår + \beta_6 * Våning + \beta_7 * Område + \beta_8 * Datum + \epsilon_i$$

där variablerna $Utropspris\ och\ AvgiftYta$ är signifikanta på 5 %-nivån när alla variabler är inkluderade i modellen. Vi använder oss utav Stepwise selection för att förenkla modellen, där vi sätter signifikansnivån på 5%.

Stepwise Selection exkluderade alla förklarande variabler utom $Utropspris$ och $Avgiftyta$, och vi får då följande additiva modell

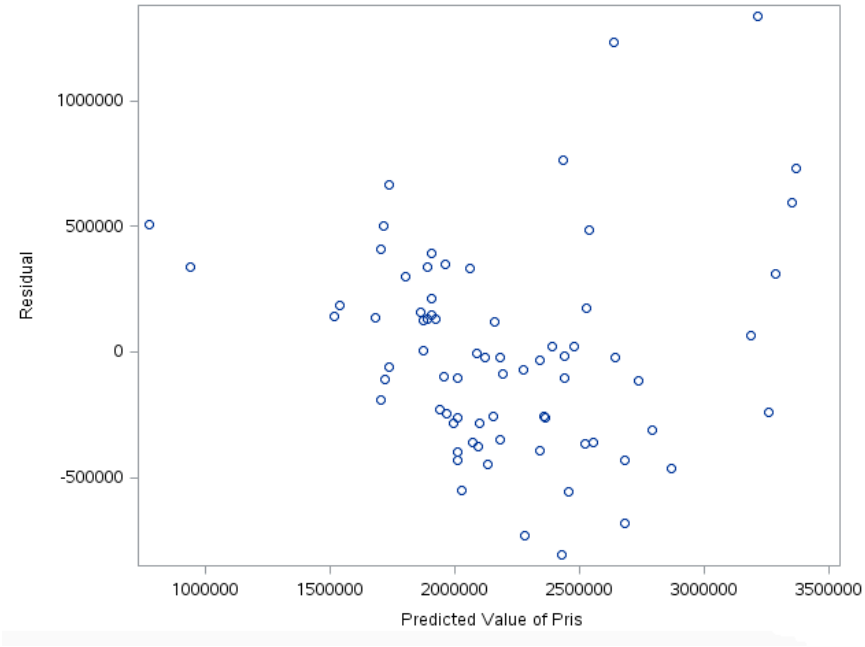
$$\log(Slutpris_i) = \alpha + \beta_1 * Utropspris + \beta_2 * Avgiftyta + \epsilon_i$$

Modellen har en justerad förklaringsgrad på 85.78%.

Andra genomförandet

Då $Utropspris$ förklarade en stor del av variationen för $Slutpris$, så ska vi anpassa en modell då $Utropspris$ exkluderas. Vi anpassar en modell med resterande förklarande variabler, och vi studerar residualerna mot de predikterade priserna.

$$\text{Slutpris}_\text{i} = \alpha + \beta_1 * Yta + \beta_2 * Rumyta + \beta_3 * Byggår + \beta_4 * AvgiftYta + \beta_5 * Område + \beta_6 * Våning + \beta_7 * Datum + \varepsilon_i$$

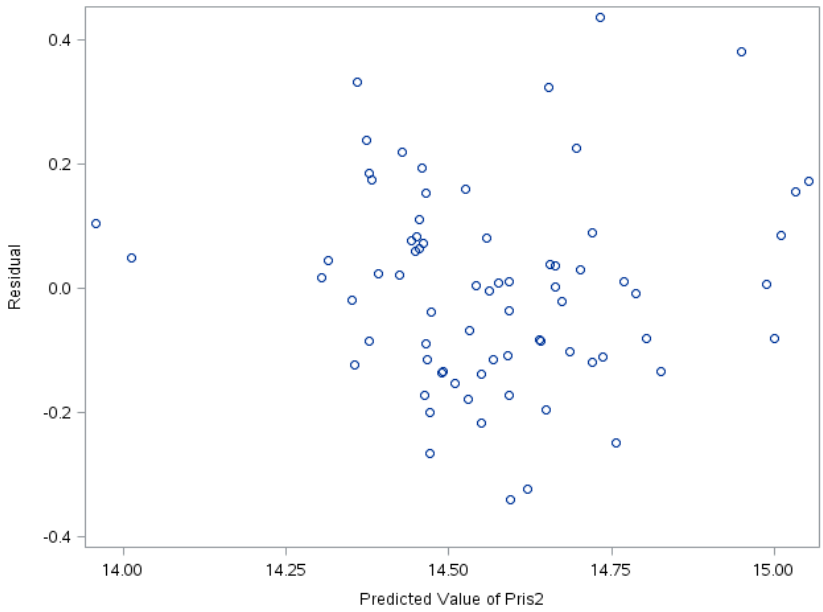


Figur 3. Plot som visar modellens residualer mot dess predikterade slutpris

Figuren ovan visar ojämnt spridda punkter. Som i första utförandet så utför vi en logtransformation av responsvariabeln. Det ger oss följande linjära modell

$$\log(\text{Slutpris}_i) = \alpha + \beta_1 * Yta + \beta_2 * Rumyta + \beta_3 * Byggår + \beta_4 * AvgiftYta + \beta_5 * Område + \beta_6 * Våning + \beta_7 * Datum + \varepsilon_i$$

Vi undersöker om spridningen förändras när vi har logaritmerat responsvariabeln



Figur 4. Plot som visar modellens residualer mot dess predikterade slutpris efter logaritmering

Vi har två observationer längst till vänster i figuren ovan vars predikterade värde avviker från resterande observationer. Bortsett från det så blev det en lite jämnare spridning efter log transformering av $Slutpris$.

Stepwise selection

Vi har följande linjära modell

$$\log(\text{Slutpris}_i) = \alpha + \beta_1 * Yta + \beta_2 * Rumyta + \beta_3 * Byggår + \beta_4 * AvgiftYta + \beta_5 * Område + \beta_6 * Våning + \beta_7 * Datum + \varepsilon_i$$

där alla variabler förutom $Datum$ är signifikanta på 5%-nivån. Stepwiseproceduren exkluderade endast variabeln $Datum$.

Vi får följande additiva modell

$$\log(\text{Slutpris}_i) = \alpha + \beta_1 * Yta + \beta_2 * Rumyta + \beta_3 * Byggår + \beta_4 * AvgiftYta + \beta_5 * Område + \beta_6 * Våning + \varepsilon_i$$

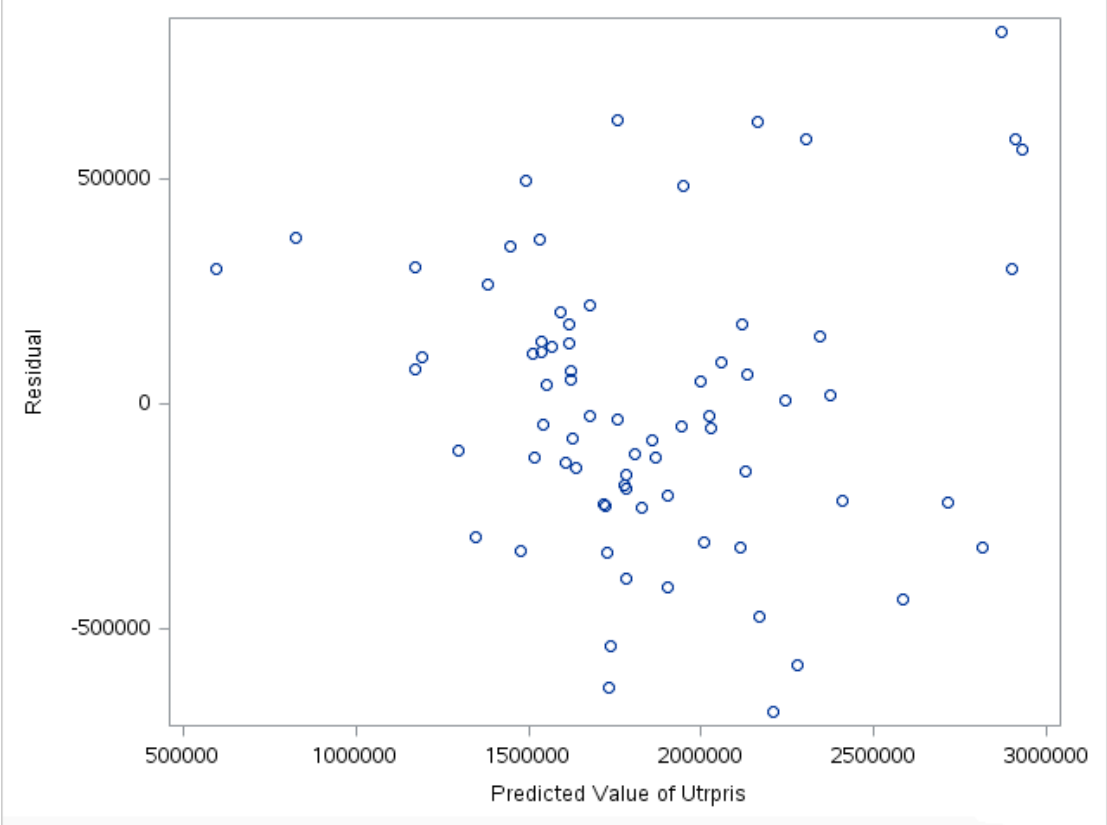
Modellen har en justerad förklaringsgrad på 59.45%.

Tredje genomförandet

Som nämnts tidigare så finns det en stark korrelation mellan slutpriset och utropspriset. En möjlighet kan vara att våra variabler kan ge en bättre förklaring av utropspriset. Vi anpassar en modell med alla förklarande variabler inkluderade. Vår modell ser då ut på följande sätt

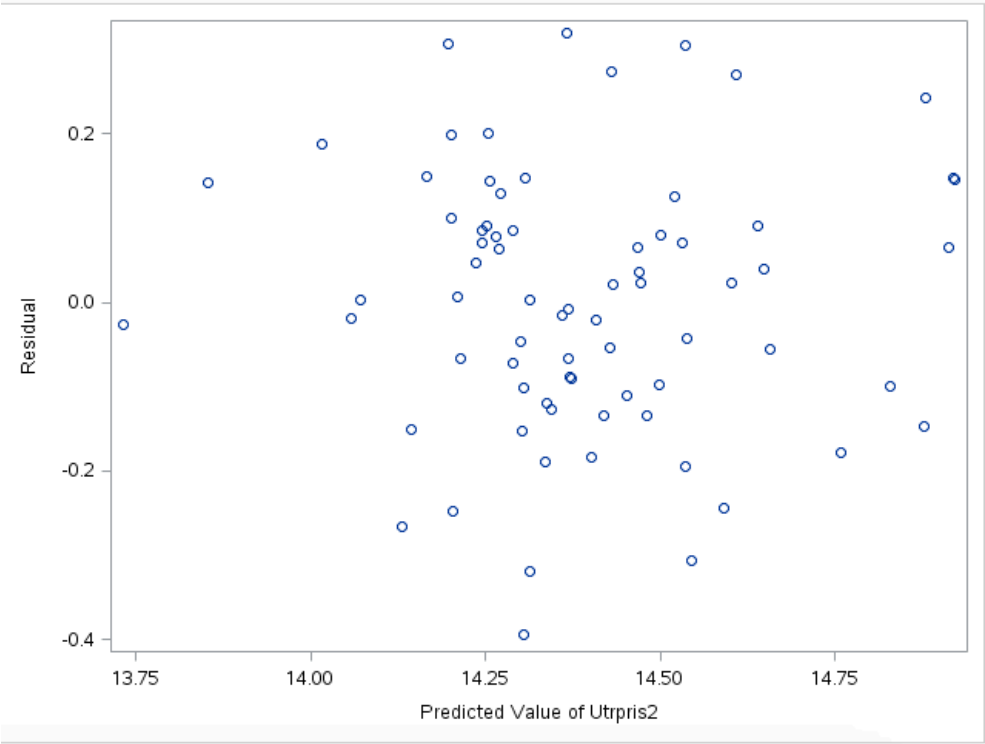
$$Utropspris_i = \alpha + \beta_1 * Yta + \beta_2 * RumYta + \beta_3 * Byggår + \beta_4 * Avgiftyta + \beta_5 * Våning + \beta_6 * Område + \beta_7 * Datum + \epsilon_i$$

När alla variabler är inkluderade förklarar de ungefär 64 % av den totala variationen för utropspriset. Vi tittar närmare på residualplottarna för att undersöka hur residualerna varierar.



Figur 5. Plot som visar modellens residualer mot dess predikterade utropspris

Residualernas spridning ser ut att öka när predikterat värde av utropspriset ökar. Vi undersöker om logaritmering av responsvariabeln jämnar ut spridningen av residualerna.



Figur 6. Plot som visar modellens residualer mot dess predikterade utropspris efter logtransformering

Vi ser ovan att logaritmering av utropspriset ger oss en jämnare variation av residualerna.

Stepwise selection

Vi fortsätter med att försöka förenkla modellen med stepwise selection. Vi väljer en signifikansnivå på 5 %.

En variabel visar sig vara icke-signifikant, det är variabeln Datum.

Vi har kvar följande additiva modell

$$\log(Utropspris_i) = \alpha + \beta_1 * Yta + \beta_2 * RumYta + \beta_3 * Byggår + \beta_4 * AvgiftYta + \beta_5 * Område + \beta_6 * Våning + \varepsilon_i$$

Modellen har en justerad förklaringsgrad på 65.63 %

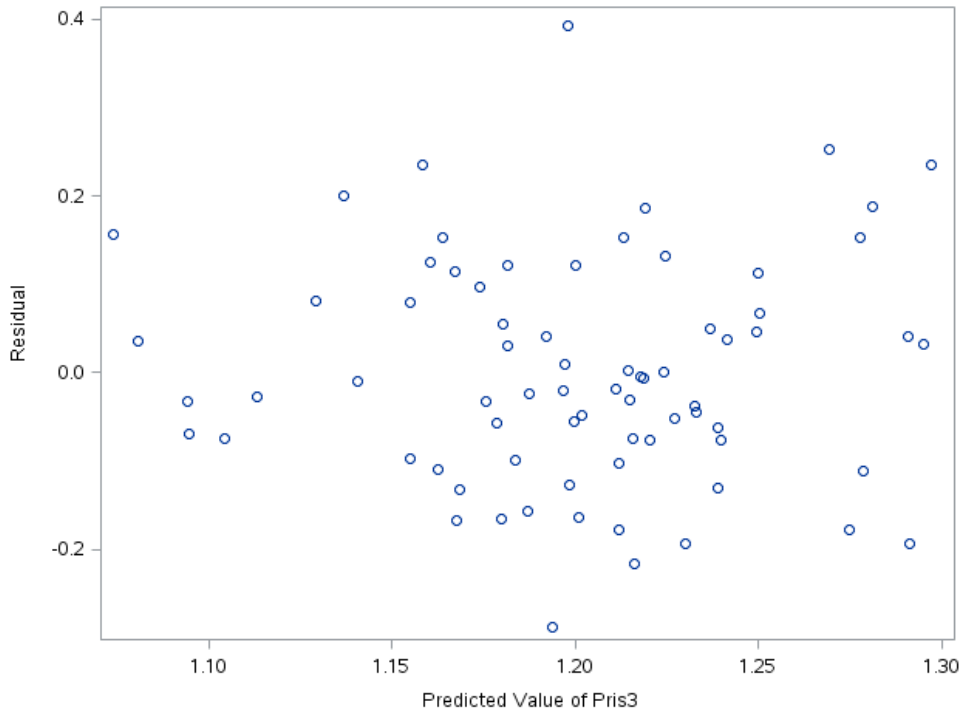
Fjärde genomförandet

Vi undersöker vad som förklarar den procentuella förändringen mellan utropspriset och slutpriset.

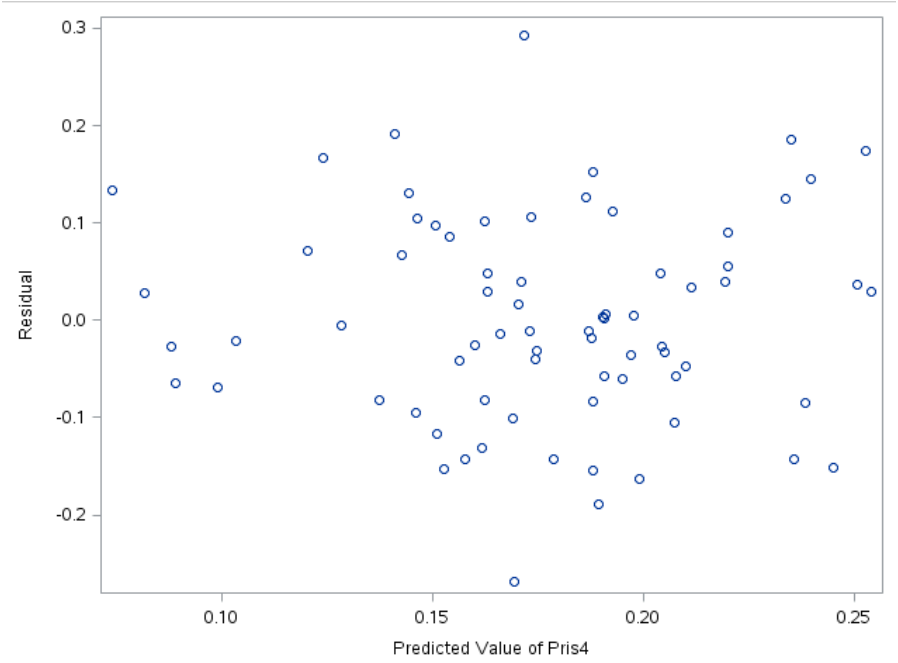
Vår modell ser ut på följande vis

$$Priskvot = \alpha + \beta_1 * Yta + \beta_2 * Rumyta + \beta_3 * Byggår + \beta_4 * Avgiftyta + \beta_5 * Område + \beta_6 * Våning + \varepsilon_i$$

När alla variabler inkluderas förklaras endast ca 4% av den totala variationen mellan priskvoten.



Ökningen av residualernas spridning försöker vi åtgärda genom att logaritmera responsvariabeln.



Logaritmering av responsvariabeln ger en mer homogen spridning av residualerna.

Stepwise selection

Vi fortsätter med att försöka förenkla modellen med stepwise selection. Vi väljer en signifikansnivå på 5 %.

En variabel visar sig vara signifikant, det är variabeln Yta.

Vi har kvar följande additiva modell

$$\log Priskvot_i = \alpha + \beta_1 * Yta + \varepsilon_i$$

Modellen har en justerad förklaringsgrad på 3,79 %.

Resultat

Våra genomföranden ger oss tre modeller. När vi har transformerat utropspriset får vi följande modell

$$\log(Slutpris_i) = \alpha + \beta_1 * Yta + \beta_2 * PrisYta + \epsilon_i$$

Där Yta och $PrisYta$ visar sig vara signifikanta på 5%-nivån. Modellen har en justerad förklaringsgrad på 78.02%.

Vi får följande parameterskattningar

Variabel	**Parameterskattning**	**Standardavvikelse**	**P-värde**	**95% Konfidensintervall**
Intercept	13.93979	0.07208	<.0001	(13.7960, 14.0835)
Utropspris	$4.06488 * 10^{-7}$	0.00083465	<.0001	$3.633 * 10^{-7}$
Avgiftyta	-0.00202	0.00000229	0.0106	(-0.0036, 0.00049)

Tabell 2. Skattningar, standardfel, P-värde samt konfidensintervall för modellens förklarande variabler.

När $Utropspris$ exkluderades från analysen var det fler variabler som visade sig vara signifikanta på 5% -nivån. Vi fick följande modell

$$\log(\text{Slutpris}_i) = \alpha + \beta_1 * Yta + \beta_2 * Rumyta + \beta_3 * Byggår + \beta_4 * AvgiftYta + \beta_5 * Område + \beta_6 * Våning + \varepsilon_i$$

Där $Yta, Rumyta, Byggår, AvgiftYta, Område$ och $Våning$ visade sig vara signifikanta på 5%-nivån. Modellen har en justerade förklaringsgrad på 59.45%, och har följande parameterskattningar

Variabel	**Parameterskattning**	**Standardavvikelse**	**P-värde**	**95% Konfidensintervall**
Intercept	15.26195	0.19087	<.0001	(14.8809, 15.6431)
Yta	0.00461	0.00108	<.0001	(0.0025, 0.0068)
Rumyta	-0.00976	0.00318	0.0031	(-0.0161, -0.0034)
Byggår	-0.00479	0.00149	0.0020	(-0.0078, -0.0018)
Avgiftyta	-0.01069	0.00156	<.0001	(-0.0138, -0.0076)
Våning	0.02069	0.01000	0.0424	(0.0007, 0.0407)
Område	0.13906	0.04921	0.0062	(0.0408, 0.2373)

Tabell 3. Skattningar, standardfel, P-värde samt konfidensintervall för modellens förklarande variabler när utropspriset exkluderas.

Vi har undersökt ifall slutpriset och utropspriset förklaras av samma variabler. Det visade sig att samma variabler var signifikanta när utropspriset användes som responsvariabel.

$$\log(Utropspris_i) = \alpha + \beta_1 * Yta + \beta_2 * RumYta + \beta_3 * Byggår + \beta_4 * AvgiftYta + \beta_5 * Område + \beta_6 * Våning + \varepsilon_i$$

Endast variabeln $Datum$ visade sig vara icke-signifikant på 5%-nivån. Modellen har en justerad förklaringsgrad på 65.63%.

Vi får fram följande parameterskattningar

Variabel	**Parameterskattning**	**P-värde**
Intercept	14.94333	<.0001
Yta	0.00645	<.0001
RumYta	-0.01129	0.0007
Byggår	-0.00447	0.0039
AvgiftYta	-0.01000	<.0001
Våning	0.02558	0.0131
Område	0.13623	0.0075

Tabell 4. Parameterskattningar och P-värden för de förklarande variabler med log(Utropspris) som responsvariabel.

När den procentuella förändringen mellan utropspriset och slutpriset undersöktes visade sig att endast variabeln Yta var signifikant på 5%-nivån. Vi fick följande modell

$$Priskvot = \alpha + \beta_1 * Yta + \varepsilon_i$$

Med följande parameterskattningar

Variabel	**Parameterskattning**	**P-värde**
Intercept	1.33821	<.0001
Yta	-0.00215	0.0073

Tabell 5. Parameterskattningar samt P-värden för modellen med priskvot som responsvariabel.

Vi ser ovan att de förklarande variablerna har en additiv effekt på den logaritmerade responsvariabeln. Det visar sig att de förklarande variablerna har en multiplikativ effekt på lägenhetspriserna. Vi får då följande modeller

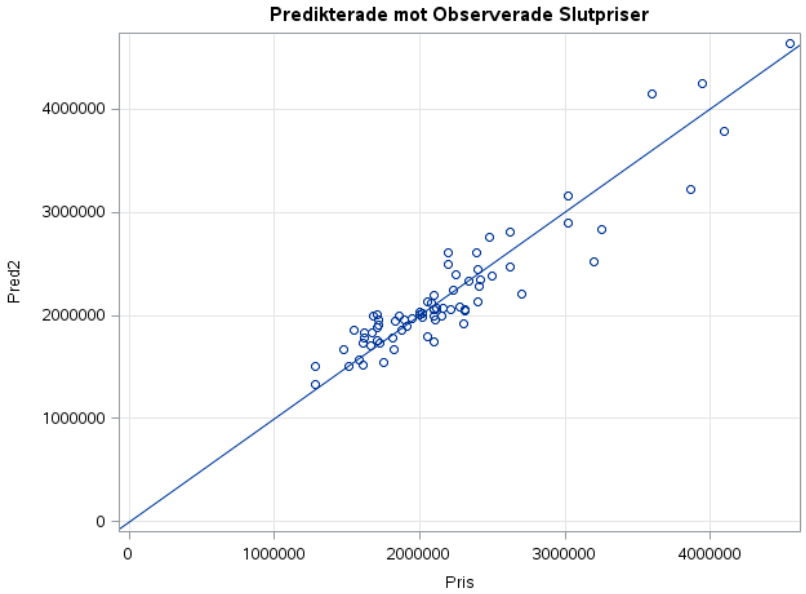
$$Slutpris_i = e^{\alpha} * e^{\beta_1 * Utropspris} * e^{\beta_2 * Avgiftyta} * e^{\varepsilon_i}$$

$$Slutpris_i = e^{\alpha} * e^{\beta_1 * Yta} * e^{\beta_2 * RumYta} * e^{\beta_3 * Byggår} * e^{\beta_4 * AvgiftYta} * e^{\beta_5 * Område} * e^{\beta_5 * Våning} * e^{\varepsilon_i}$$

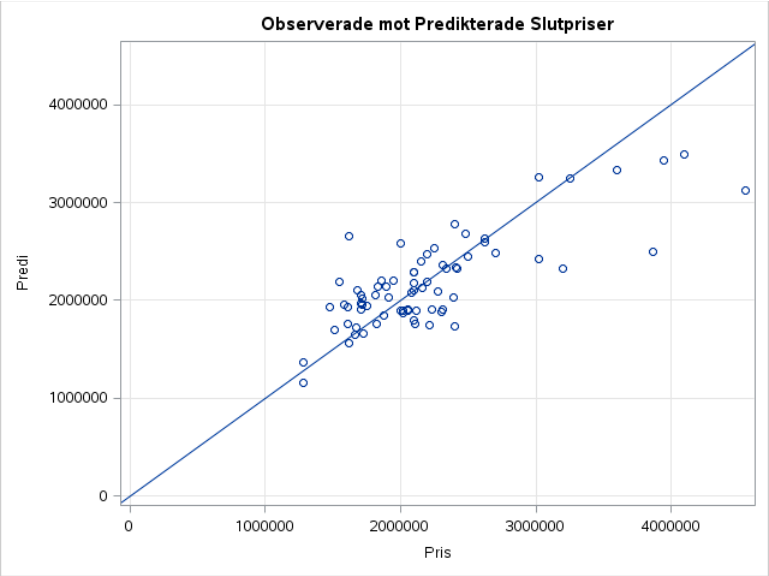
$$Utropspris_i = e^{\alpha} * e^{\beta_1 * Yta} * e^{\beta_2 * RumYta} * e^{\beta_3 * Byggår} * e^{\beta_4 * AvgiftYta} * e^{\beta_5 * Område} * e^{\beta_5 * Våning} * e^{\varepsilon_i}$$

Prediktion

Vi ska testa prediktionsförmågan för våra modeller. Vi tillämpar andra halvan av datamaterialet och predikterar nya slutpriser för lägenheterna. För att få en överblick över hur väl våra modeller har predikterat, undersöker vi plottar för det observerade slutpriset mot det predikterade. Vi kollar även på modellernas PRESS- samt RMSEP-värden.



Figur 7. Plot för predikterade mot observerade slutpriser när utropspris är inkluderat



Figur 8. Plot för predikterade mot observerade slutpriser när utropspris är exkluderat

Modellernas PRESS- och RMSEP-värden blev följande

Förklarande variabler	**MSEP**	**RMSEP**
Utropspris, Avgiftyta	50 026 830 991.2	223666.79
Yta, Rumyta, Byggår, Avgiftyta, Område, Våning	156,746,545,916.31	395912.30

Tabell 6. MSEP och RMSEP för modellerna när utropspris har inkluderas/exkluderats med slutpris som responsvariabel.

När utropspriset inkluderas ger det oss ett lägre RMSEP-värde.

Diskussion

Syftet med detta arbete var att undersöka vilka faktorer som kan förklara slutpriset av en såld lägenhet. Vi har fått fram två modeller som förklarar slutpriset av lägenheter. För den ena modellen har vi inkluderat utropspriset som förklarande variabel. För den andra modellen har vi exkluderat utropspriset helt. För att välja den bästa modellen mellan dem två så fokuserar vi på deras residualplottar samt deras justerade förklaringsgrad. Förklaringsgraden mellan dem skiljer sig med nästan 26 %. När vi inkluderar variabeln $Utropspris$ får vi en modell med en förklaringsgrad på 85.78%.vilket är den modell som förklarar slutpriset bäst. Dess residualplot (se figur 2) ser även ut att ha en mer homogen variation mellan residualerna, jämfört med den modell där utropspris inte är med (se figur 4).

Modellen med färre förklarande variabler har en positiv parameterskattning till variabeln $Utropspris$, och en negativ skattning till parametern $AvgiftYta$. En ökning av $Utropspris$ har en ökande effekt på slutpriset, medan en ökning av variabeln $AvgiftYta$ har en negativ effekt på slutpriset. Den sämre modellen har tre parameterskattningar som är negativa. De förklarande variablerna är $RumYta, Byggår$ och $AvgiftYta$. Vi kan tolka det som att dessa variabler har en negativ effekt på slutpriset.

Vi kom fram till att variabeln Yta var den enda variabeln som blev signifikant när vi undersökte den procentuella förändringen mellan utropspriset och slutpriset. Även den har en negativ parameterskattning. Vi kan tolka det som att en ökning av ytan hämmar den procentuella ökningen av priset.

Begränsningar

Datamaterialet vi har använt för arbetet består endast av 75 observationer. Ett datamaterial så litet kan påverka våra parameterskattningar, och kan öka osäkerheten av hur stor inverkan variablerna har. Den starka korrelationen mellan slutpriset och utropspriset ger som väntat en bättre förklaringsgrad när utropspriset inte exkluderas. När vi exkluderade utropspriset kom vi fram till en modell som minskade i förklaringsgrad. Det skulle vara intressant att undersöka om det finns fler variabler som kan tänkas ha en inverkan på slutpriset. Variabler som avstånd till centrum, hav, kommunala färdmedel. Andra variabler som hiss och mäklarfirma skulle också vara intressant att undersöka.

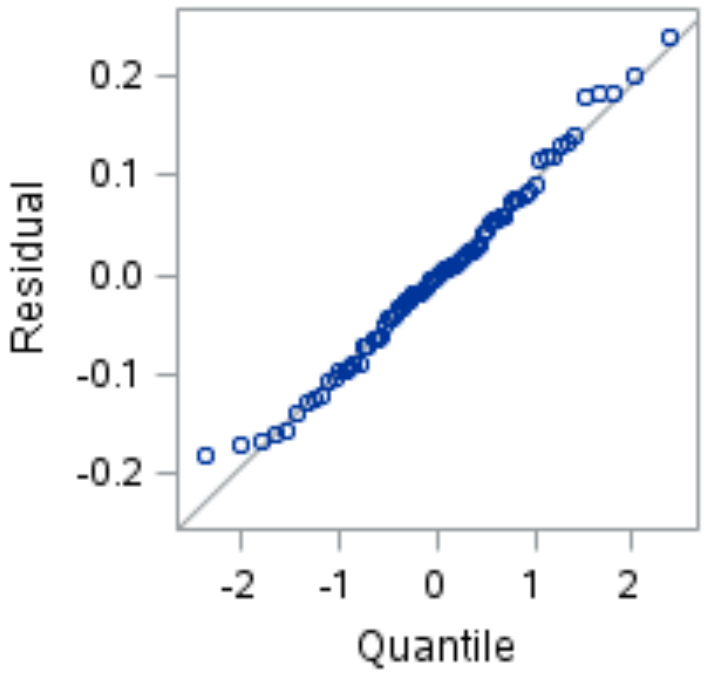
Variabeln $Datum$ visade sig inte vara signifikant i någon modell. Datamaterialet består av observationer för 150 lägenheter under 2015. Regressionen utfördes på 75 observationer, där datumen lägenheterna såldes varierar mellan januari-Juni 2015. Priserna för lägenheterna har ökat dem senaste åren, och skulle vi utföra analysen med data i ett större tidsintervall så tror jag att datum skulle ha en signifikant inverkan

Referenser

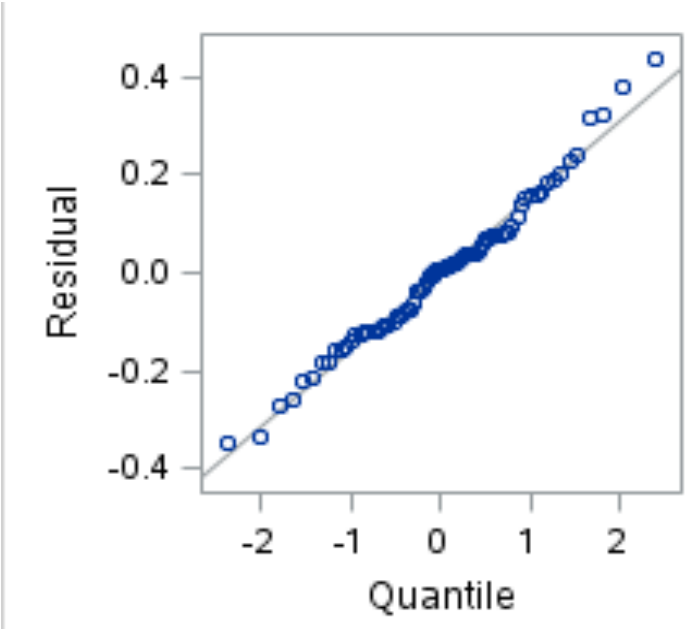
[1] Rolf Sundberg*, Lineära Statistiska Modeller*, 2015

[2] Patrik Andersson & Joanna Tyrcha, *Notes in Econometrics*, 2014

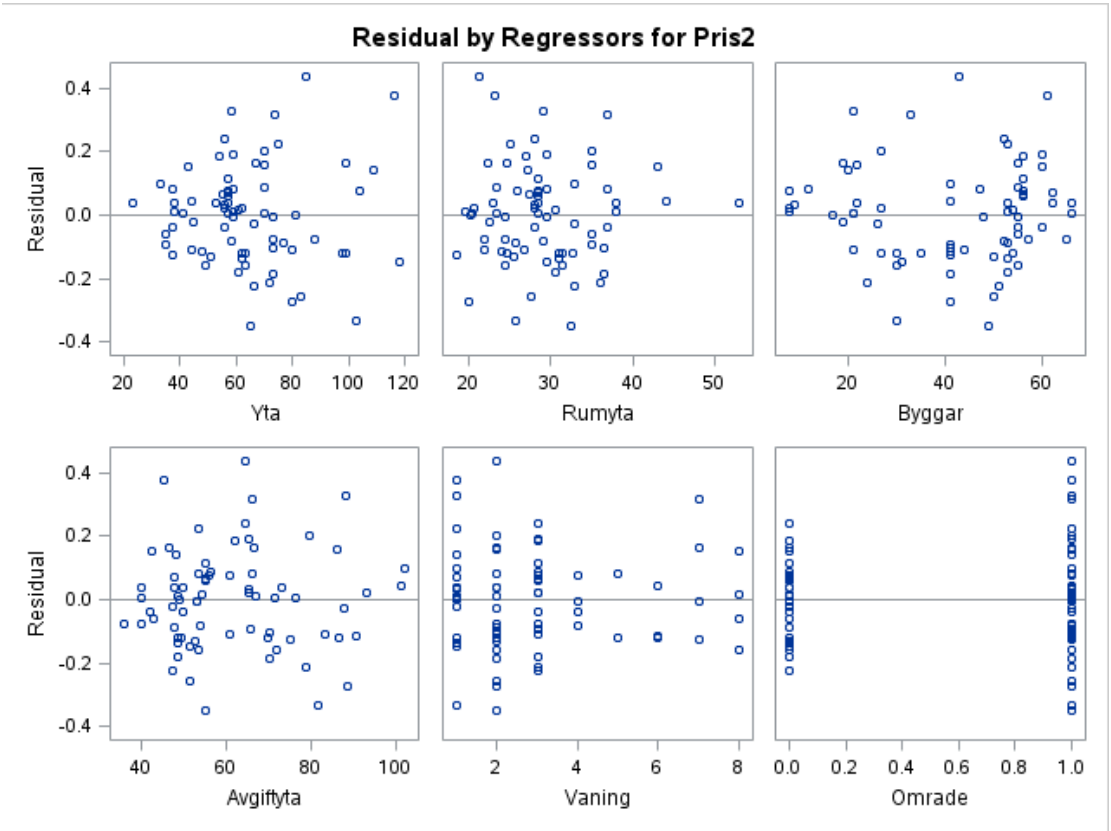
Appendix



FIgur 9. Normalfördelningsplot för modell med utropspriset inkluderat.



Figur 10. Normalfördelningsplot för modell med utropspriset exkluderat.



Figur 11. Residualplottar för de signifikanta förklarande variablerna I modellen med utropspris exkluderat.

Figur 12. Residualplottar för de signifikanta variablerna I modellen med utropspris inkluderat.