

Regressionsanalys av lägenhetspriser i Spånga

Mahamed Saeid Ali

Kandidatuppsats 2016:11
Matematisk statistik
Juni 2016

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Regressionsanalys av lägenhetspriser i Spånga

Mahamed Saeid Ali*

Juni 2016

Sammanfattning

Examensarbetet går ut på att få fram en lämplig modell som beskriver slutpriset av lägenheter i Spånga under perioden 2014-01-01 - 2015-12-31 genom att tillämpa multipel linjär regression. Vi har ett datamaterial som består av 221 observationer och av dem kommer vi att använda 158 observationer i vår undersökning där vi har 11 förklarande variabler som vi tror förklarar slutpriset, som är vår responsvariabel. Med hjälp av stegvis variabelselektion får vi en modell som består av sex variabler där bl.a. variabeln boarea förklarar mest av slutpriset vid enkel regression. Efter undersökning mellan flera olika modeller får vi fram två potentiella modeller och vi kommer fram till att modell 1, som är en multiplikativ modell där vi transformerar responsvariabeln slutpris med hjälp av logaritmering och som består av boarea, säljår, våning, ålder, avgiftarea och arearum som förklarande variabler, är den lämpligaste modellen som förklarar mest av slutpriset då modellen har högre förklaringsgrad, 0.7744.

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige.
E-post: mohamed940316@gmail.com. Handledare: Gudrun Brattström och Jan-Olov Persson.

Innehåll

1. Inledning.....	4
2. Teori.....	5
2.1 Multipel linjär regression.....	5
2.2 Minsta-kvadratmetoden.....	5
2.3 Hypotesprövning – tvåsidig test.....	6
2.4 Multikollinearitet –VIF.....	6
2.5 Stegvis variabelselektion.....	7
2.5.1 Backward elimination.....	7
2.5.2 Forward selection.....	7
2.5.3 Stepwise regression.....	7
2.6 Homoscedasticitet.....	7
2.7 Transformation.....	7
2.8 Anpassningsmått –förklaringsgrad.....	8
3. Beskrivning av data.....	9
3.1 Datamaterial.....	9
3.2 Statistisk modellering.....	10
4. Analys av data.....	11
4.1 Modellval.....	18
5. Resultat.....	19
6. Diskussion.....	20
7. Referenser.....	21

Sammanfattning

Examensarbetet går ut på att få fram en lämplig modell som beskriver slutpriset av lägenheter i Spånga under perioden 2014-01-01 - 2015-12-31 genom att tillämpa multipel linjär regression. Vi har ett datamaterial som består av 221 observationer och av dem kommer vi att använda 158 observationer i vår undersökning där vi har 11 förklarande variabler som vi tror förklarar slutpriset, som är vår responsvariabel. Med hjälp av stegvis variabelselektion får vi en modell som består av sex variabler där bl.a. variabeln *boarea* förklarar mest av slutpriset vid enkel regression. Efter undersökning mellan flera olika modeller får vi fram två potentiella modeller och vi kommer fram till att *modell 1*, som är en multiplikativ modell där vi transformerar responsvariabeln *slutpris* med hjälp av logaritmering och som består av *boarea*, *säljår*, *våning*, *ålder*, *avgiftarea* och *arearum* som förklarande variabler, är den lämpligaste modellen som förklarar mest av slutpriset då modellen har högre förklaringsgrad, 0.7744.

1. Inledning

När man köper en lägenhet undrar man vilka faktorer som påverkar priset på lägenheten. De vanligaste faktorer som man tänker på vid köp av en lägenhet är bl.a. hur stor lägenheten är och hur många rum lägenheten består av. I det här examensarbetet kommer vi undersöka vilka faktorer som påverkar mest på slutpriset och undersökningen kommer att baseras på 221 lägenheter som såldes i Spånga vid perioden 2014-01-01 - 2015-12-31. I kapitel 2 beskriver vi teorin som vi kommer att tillämpa i examensarbetet. I kapitel 3 beskriver vi datamaterialet och modellering av data. Kapitel 4 analyserar vi datamaterialet med hjälp av teorin i kapitel 2 och tar fram lämpliga modeller som vi därefter väljer en modell som vi anser är lämpligast. Vidare i kapitel 5 beskriver vi resultatet och tolkar parameterskattningarna och slutligen diskuterar vi resultatet som vi får fram i kapitel 6.

2. Teori

2.1 Multipel linjär regression

I det här examensarbetet kommer vi att tillämpa multipel linjär regression. Multipel linjär regression är en statistisk modell som används för att påvisa om det finns något linjärt samband mellan responsvariabeln och de förklarande variablerna.[1] I vårt fall kommer vi undersöka vilka variabler som beskriver slutpriset av lägenheter i Spånga. Den generella formen av en multipel linjär regression beskrivs på följande sätt:

$$Y_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \cdots \beta_k x_{kj} + \epsilon_j, \quad j = 1, 2, \dots, n \quad (1)$$

Där ϵ_j är feltermerna som är oberoende och normalfördelade med väntevärde noll och konstant varians, dvs.

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

Y_j är responsvariabeln (de observerade värdena y_j , $j = 1, 2, \dots, n$). Parametern β_0 tolkas som interceptet för regressionslinjen och $\beta_1, \beta_2, \dots, \beta_k$ beskriver på vilket sätt de k förklarande variablerna inverkar på responsvariabeln. Parametrarna, $\beta_0, \beta_1, \beta_2, \dots, \beta_k$, är okända vilket leder till att vi kommer att skatta dem ur vårt data med hjälp av minsta-kvadratmetoden. Vanligtvis brukar man uttrycka modellen i matris- och vektorform

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{k1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & \cdots & x_{kn} \end{pmatrix}, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

där

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}).$$

2.2 Minsta-kvadratmetoden

Som vi har nämnt i tidigare avsnitt så är $\boldsymbol{\beta}$ okända parametrar och de skattar vi med hjälp av minsta-kvadratmetoden. Metoden går ut på att minimera summan av kvadratiska skillnaden mellan observationerna y_j och väntevärden [1], dvs. minimera summan av de kvadratiska residualerna

$$\sum_{j=1}^n \left(y_j - (\beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \cdots \beta_k x_{kj}) \right)^2 = \sum_{j=1}^n e_j^2,$$

Vilket leder till att minsta-kvadratmetoden ger oss skattningen av $\boldsymbol{\beta}$

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

2.3 Hypotesprövning – tvåsidig test

Vi vill undersöka vilka förklarande variabler som påverkar responsvariabeln i modellen (1) genom att välja de β_i som blir statistisk signifikanta skilda från noll. Det gör vi genom att bilda en hypotestest att β_i är skilt från noll:

$$H_0: \beta_i = 0 \quad H_1: \beta_i \neq 0.$$

Vi vet från tidigare att feltermerna ϵ_j är oberoende normalfördelade med väntevärde noll och konstant σ^2 vilket leder till att fördelningen av skattningen $\boldsymbol{\beta}$ är

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}).$$

Vår test statistiska under noll hypotesen blir

$$T_i = \frac{\hat{\beta}_i}{\hat{\sigma} \sqrt{((\mathbf{X}^T \mathbf{X})^{-1})_{ii}}},$$

som är t –fördelad med $n - k - 1$ frihetsgrader. Eftersom

$$V(\hat{\boldsymbol{\beta}}) = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1},$$

så är $(\sigma^2 (\mathbf{X}^T \mathbf{X})^{-1})_{ii}$ den i :te diagonalelementen i kovariansmatrisen $V(\hat{\boldsymbol{\beta}})$. [1]

Vi förkastar vår nollhypotes vid signifikansnivå p om

$$|T_i| > t_{p/2}(n - k - 1).$$

2.4 Multikollinearitet – VIF

När vi har flera förklarande variabler kan det hända att det uppstår samvariation mellan variablerna, vilket betyder att vi har multikollinearitet. Det kan vara så att vissa förklarande variabler som man trodde skulle bli statistisk signifikanta inte är det pga. multikollinearitet eller exempelvis om kollineära variabler hamnar individuellt i modellen så är de statistisk signifikanta medan om variablerna tillsammans hamnar i samma modell får vi istället att båda blir icke-signifikanta.

För att få reda på om det råder multikollinearitet mellan de förklarande variablerna så använder vi en statistisk mått, VIF som står för *Variance Inflation Factor*. VIF-faktorn kan skrivas på följande sätt:

$$VIF = \frac{1}{1 - R_i^2},$$

Här står R_i^2 för den förklaringsgrad som anger hur mycket av variationen i x_i som förklaras av de andra x –variablerna. Höga värden på VIF indikerar att multikollinearitet existerar och man brukar sätta gränsen vid 5 eller 10. [2][3]

2.5 Stegvis variabelselektion

Det kan hända att mängden av de potentiella förklarande variablerna som vi använder i undersökningen inte är nödvändiga att ta med alla i modellen utan att man väljer ut de variabler som beskriver på bästa sätt responsvariabeln. Detta görs med hjälp av tre olika procedurer, backward elimination, forward selection och stepwise regression. De går ut på att eliminera eller inkludera en förklarande variabel i taget tills något stoppkriterium är uppfyllt, dvs. en risknivå som man har valt i förväg. Notera att de tre procedurer som vi har nämnt behöver inte ge oss liknande resultat. [2]

2.5.1 Backward elimination

Proceduren går ut på att man har samtliga förklarande variabler i modellen. I varje steg så utför man en hypotestest att parametrarna β_i är skilt från noll. Därefter elimineras i varje steg en variabel, med högst p-värde, i taget tills proceduren stoppar där variablerna som är kvar är signifikanta vid en förvald signifikansnivå. [2]

2.5.2 Forward selection

Man börjar med att modellen inte har några förklarande variabler alls. För att inkludera en variabel i taget så utför man ett hypotestest, liknande test som vi beskrev i backward elimination. Variabeln med högst p-värde inkluderas i modellen vid varje steg tills modellen består av variabler som är signifikanta vid en förväg bestämd risknivå. [2]

2.5.3 Stepwise regression

Denna procedur är lite mer avancerad än forward selection och den går ut på att för varje variabel som inkluderas i modellen så kontrollerar man om de tidigare variabler som inkluderades är signifikanta vid test av hypotesen att parametrarna är skilt från noll. Om en variabel inte är signifikanta vid en förvald risknivå så elimineras den ur modellen. [2]

2.6 Homoscedasticitet

Ett av antagandena om multipel linjär regression är att det råder konstant varians hos feltermerna, dvs.

$$V(\epsilon_j) = \sigma^2, \quad \text{för alla } j = 1, \dots, n.$$

Då har vi det som kallas för homoscedasticitet. Om variansen för feltermerna inte är konstant oavsett värdena på de förklarande variablerna, så råder det heteroscedasticitet. [1]

2.7 Transformation

Ibland kan det hända att det inte råder något linjärt samband mellan responsvariabeln och de förklarande variablerna vilket leder till att vi är tvungna att transformera modellen. Ett sätt att transformera modellen är att utnyttja logaritmen. Exempelvis om vi har den multiplikativa modellen

$$Y = \alpha e^{\beta x} \epsilon,$$

då blir den logarimerade modellen

$$\ln Y = \ln \alpha + \beta x + \ln \epsilon,$$

där $\ln \epsilon$ antas vara normalfördelad med väntevärde noll och varians σ^2 . Vi ser att vi får en linjär modell där responsvariabeln och förklarande variabeln har ett linjärt samband. [2]

2.8 Anpassningsmått – Förklaringsgrad

Förklaringsgraden, R^2 , är ett mått som används vid jämförelse av olika linjära modeller och multipla regressionsmodeller. Den visar hur väl de förklarande variablerna i en linjär modell förklarar den totala variationen och den antar värden mellan 0 och 1 där högre värden på R^2 visar att anpassningen av den linjära modellen är god. Förklaringsgraden definieras på följande sätt:

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

där kvadratsumman för regression är $\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$, kvadratsumman för residualerna $\sum_{i=1}^n (y_i - \hat{y}_i)^2$ och till slut kvadratsumman för den totala variationen $\sum_{i=1}^n (y_i - \bar{y})^2$.

När man lägger till en x -variabel i modellen så ökar alltid R^2 , vilket kan uppfattas som en nackdel med R^2 . Istället använder man R_{adj}^2 som mäter hur mycket variansen minskar i modellen jämfört med modellen där vi inte har några förklarande variabler. R_{adj}^2 definieras på följande sätt:

$$R_{adj}^2 = 1 - \frac{\hat{\sigma}^2}{\hat{\sigma}_0^2},$$

Där $\hat{\sigma}^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 / (n - k - 1)$ och $\hat{\sigma}_0^2 = \sum_{i=1}^n (y_i - \bar{y})^2 / (n - 1)$ är σ^2 – skattningen när man inte har någon x -variabel i modellen. [2]

3. Beskrivning av data

3.1 Datamaterial

Datamaterialet är tagen från hemsidan *booli* som visar bl.a. försäljningar av lägenheter i Spånga vid perioden 2014-01-01 - 2015-12-31.[4] Totalt har vi 221 observationer men pga. brist på information om variabeln *driftkostnad* har vi tagit bort den och då får vi istället 158 observationer samt 11 förklarande variabler som vi tror förklarar slutpriset, som är responsvariabeln, av försäljning av lägenheter i Spånga vid den nämnda perioden. Slutligen, de 12 variabler beskriver vi på följande sätt:

- *Slutpris*: Det slutgiltiga priset som lägenheten såldes för.
- *Boarea*: Hur stor lägenheten är, i m^2 .
- *Antal rum*: Hur många rum lägenheten har.
- *Månadsavgift*: Avgiften som täcker fastighetsskötsel, reparationer m.m. och den betalar man till bostadsrättsföreningen som man tillhör.
- *Byggår*: Året lägenheten byggdes.
- *Säljår*: Året man sålde lägenheten.
- *Vatten*: Avståndet mellan lägenheten och den närmaste sjön, mätt i *km*.
- *Våning*: Vilken våning lägenheten ligger på.
- *Säsong*:
 - *Vår*: Försäljning av lägenheterna mellan perioderna 2014-03-01- 2014-05-31 och 2015-03-01- 2015-05-31.
 - *Sommar*: Försäljning av lägenheterna mellan perioderna 2014-06-01- 2014-08-31 och 2015-06-01- 2015-08-31.
 - *Höst*: Försäljning av lägenheterna mellan perioderna 2014-09-01- 2014-11-30 och 2015-09-01- 2015-11-30.
 - *Vinter*: Försäljning av lägenheterna mellan perioderna 2014-12-01- 2014-02-28 och 2015-12-01- 2015-02-28.

Följande tabell visar oss information om data för variablerna som vi har beskrivit:

	<i>Minimum:</i>	<i>Maximum:</i>	<i>Medelvärde, ca:</i>
<i>Slutpris</i>	900 000 <i>kr</i>	7 725 000 <i>kr</i>	2 242 244 <i>kr</i>
<i>Boarea</i>	18 m^2	195 m^2	68.85 m^2
<i>Antal rum</i>	1	8	2.6
<i>Månadsavgift</i>	1115 <i>kr</i>	7646 <i>kr</i>	4024 <i>kr</i>
<i>Säljår</i>	2014	2015	—
<i>Byggår</i>	1904	2008	1980.85
<i>Säljår</i>	2014	2015	—
<i>Vatten</i>	2.1 <i>km</i>	4.6 <i>km</i>	3.73 <i>km</i>
<i>Våning</i>	1	8	2.41

Tabell 1: Minimum, Maximum samt medelvärde av data.

3.2 Statistisk modellering

Innan vi går vidare till analys av datamaterialet skapar vi en ny förklarande variabel med hjälp av variablerna som vi har nämnt i förgående avsnitt 3.1, vilket leder till att vi har:

- Variabeln *Ålder* beskriver skillnaden mellan säljår och byggår.

Vi vill även undersöka om de fyra årstiden påverkar slutpriset och därför skapar vi tre dummyvariabler och *Vinter* blir vår referensvariabel.

- *Vår* = 1 om lägenheterna såldes mellan perioderna 2014-03-01- 2014-05-31 och 2015-03-01- 2015-05-31, 0 annars.
- *Sommar* = 1 om lägenheterna såldes mellan perioderna 2014-06-01- 2014-08-31 och 2015-06-01- 2015-08-31, 0 annars.
- *Höst* = 1 om lägenheterna såldes mellan perioderna 2014-09-01- 2014-11-30 och 2015-09-01- 2015-11-30, 0 annars.

Slutligen, de potentiella förklarande variablerna som vi väljer ut för att undersöka vilka som påverkar slutpriset, dvs. responsvariabeln, är:

- *Boarea*
- *Antal rum*
- *Månadsavgift*
- *Säljår*
- *Ålder*
- *Vatten*
- *Våning*
- *Årstid*
 - *Vår* – dummyvariabel
 - *Sommar* – dummyvariabel
 - *Höst* – dummyvariabel
 - *Vinter* – dummyvariabel, referens

4. Analys av data

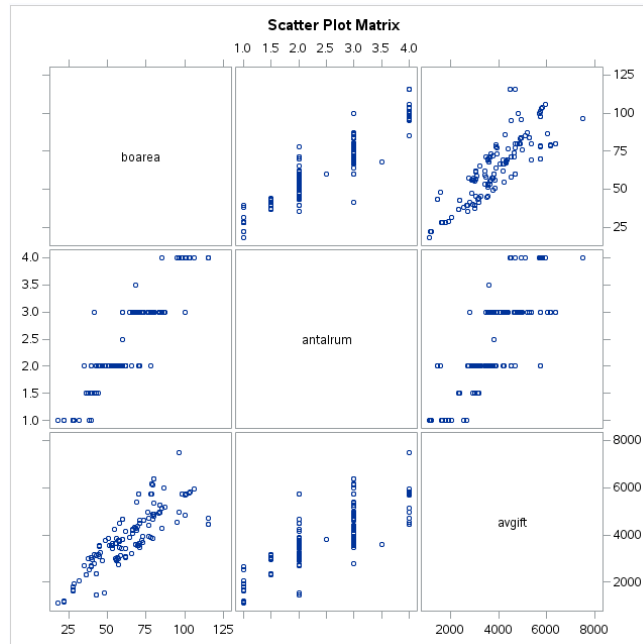
Under arbetet kommer vi att använda den statistiska programvaran SAS.

Vi börjar med att undersöka om det råder samvariation mellan x – variablerna genom ta en titt på VIF-värdet. Vi sätter gränsen för höga värden vid 5.[3] Med hjälp av SAS gör vi en multipel regression med samtliga variabler och vi får följande resultat:

<i>Förklarande variabel</i>	<i>VIF</i>
<i>Boarea</i>	8.83572
<i>Antal rum</i>	6.44064
<i>Månadsavgift</i>	4.61911
<i>Säljår</i>	1.05768
<i>Vatten</i>	1.74386
<i>Våning</i>	1.16779
<i>Ålder</i>	2.04327
<i>Vår</i>	1.95064
<i>Sommar</i>	1.76825
<i>Höst</i>	1.97270

Tabell 2: VIF-värden på x –variablerna.

Från tabell 2 ser vi att VIF-värdet för *boarea* och *antal rum* är högre än 5 och värdet på *månadsavgift* är ungefär nära gränsen, vilket vi tror beror på att den är korrelerad med *boarea* och *antal rum*. Vi undersöker korrelationen mellan dessa tre variabler genom att plotta mot varandra. Vi får följande scatter plot:



Figur 1: Scatter plot av *boarea*, *antal rum* och *avgift*.

Från figur 1 ser vi att det råder ett linjärt samband mellan *boarea*, *antal rum* och *avgift*. För att förhindra samvariationen mellan dessa variabler så transformerar vi till följande kvoter:

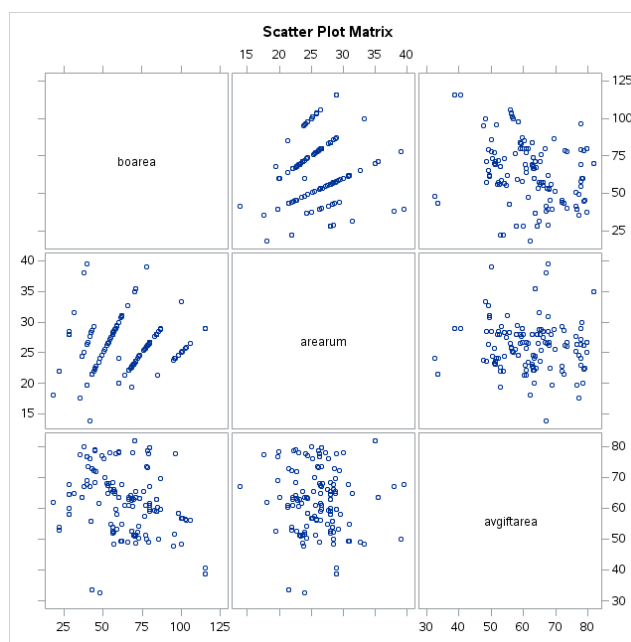
- *arearum*: kvoten mellan *boarea* och *antal rum*.
- *avgiftarea*: kvoten mellan *månadsavgift* och *boarea*.

Vi undersöker nu om det fortfarande råder korrelation mellan *arearum*, *avgiftarea* och de resterande förklarande variablerna, givet att *antal rum* och *avgift* byts ut mot *arearum* och *avgiftarea*, genom att köra en multipel regression i SAS. Vi får följande resultat:

Förklarande variabel	VIF
<i>Boarea</i>	1.43361
<i>Säljår</i>	1.05674
<i>Vatten</i>	1.72176
<i>Våning</i>	1.17114
<i>Ålder</i>	2.28430
<i>Vår</i>	1.95309
<i>Sommar</i>	1.78090
<i>Höst</i>	1.96546
<i>arearum</i>	1.02783
<i>avgiftarea</i>	1.71461

Tabell 3: VIF-värden av x –variablerna.

Från tabell 3 ser vi att alla värden på VIF är under 5 och för att säkerställa att det inte råder korrelation mellan *boarea*, *arearum* och *avgiftarea* plottar vi variablerna mot varandra. Vi får följande scatter plot:



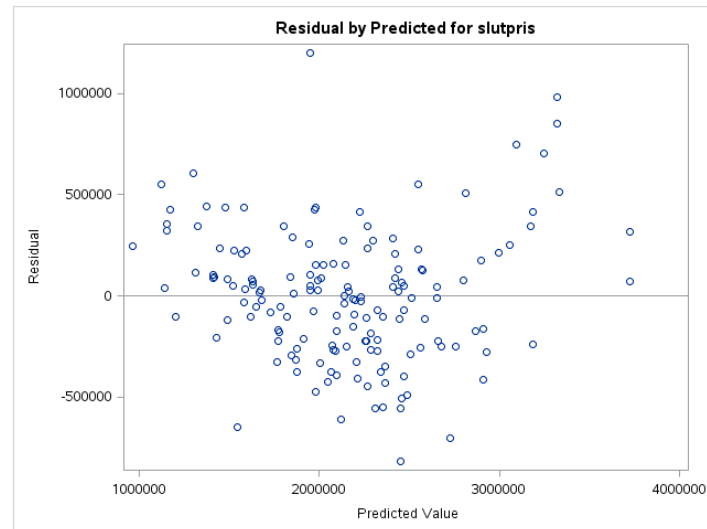
Figur 2: Scatter plot av *boarea*, *arearum* och *avgiftarea*.

I figur 2 ser vi att de tre variablerna är nästan okorrelerade. Vidare så kan vi konstatera att de förklarande variablerna som är nämnda i tabell 3 är nästan okorrelerade, dvs. det råder inga samvariation mellan x –variablerna.

Eftersom vi vill ha en modell som beskriver slutpriset på bästa möjliga sätt så utnyttjar vi de tre olika procedurer av stegvis variabelselektion. Samtliga procedurer ger oss samma resultat vid 5 % signifikansnivå och de förklarande variablerna är:

- *Boarea*
- *Säljår*
- *Våning*
- *Ålder*
- *avgiftarea*
- *arearum*

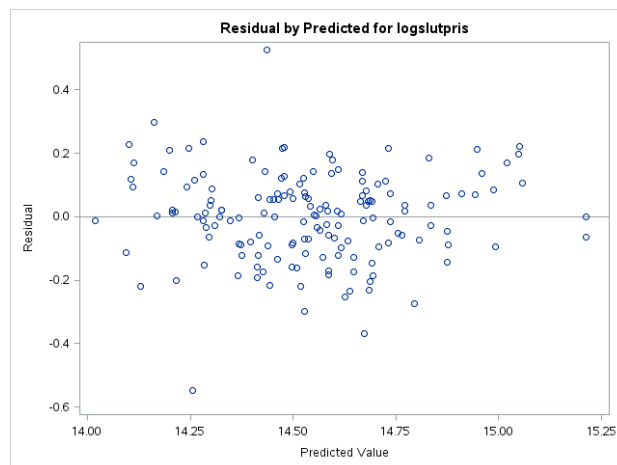
Vi undersöker om modellen, som vi fick fram från stegvis variabelselektion, är rimligt genom att titta på plotten av residualerna mot de predikterade värdena.



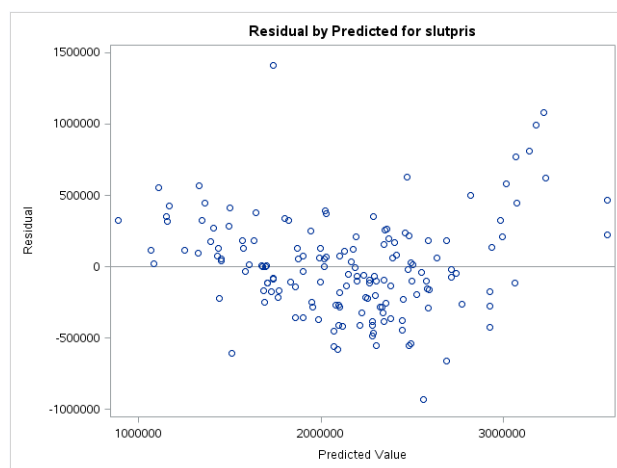
Figur 3: Residualplott. Regression av *slutpris* mot *boarea*, *säljår*, *våning*, *ålder*, *avgiftarea* och *arearum*.

Från figur 3 ser vi att vi har en modell som antar heteroscedasticitet, dvs. icke-konstant varians. För att undvika det kan vi transformera modellen med hjälp av logaritmering. Vi kan antingen logaritmera responsvariabeln eller båda responsvariabeln och de förklarande variablerna. Vi väljer att transformera responsvariabeln *slutpris* och förklarande variabeln *boarea* eftersom *boarea* förklarar mest av slutpriset vid enkel regression, dvs. med förklaringsgrad 0.4497. Vi undersöker plotten av residualerna mot predikterade värden av följande modeller när vi kör regression i SAS:

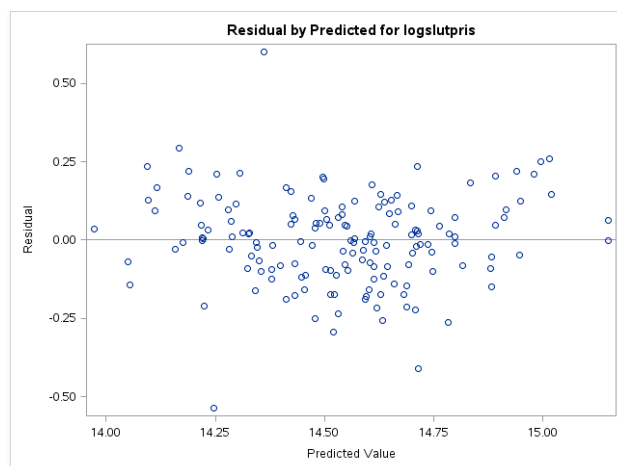
- Responsvariabeln *logslutpris* och *boarea*, *säljår*, *våning*, *ålder*, *avgiftarea* samt *arearum* som förklarande variabler.
- Responsvariabeln *slutpris* och *logboarea*, *säljår*, *våning*, *ålder*, *avgiftarea* samt *arearum* som förklarande variabler.
- Responsvariabeln *logslutpris* och *logboarea*, *säljår*, *våning*, *ålder*, *avgiftarea* samt *arearum* som förklarande variabler.



Figur 4: Plot av residualerna mot de predikterade värden då responsvariabeln är *logslutpris* och *boarea*, *säljår*, *våning*, *ålder*, *avgiftarea* samt *arearum* som förklarande variabler.

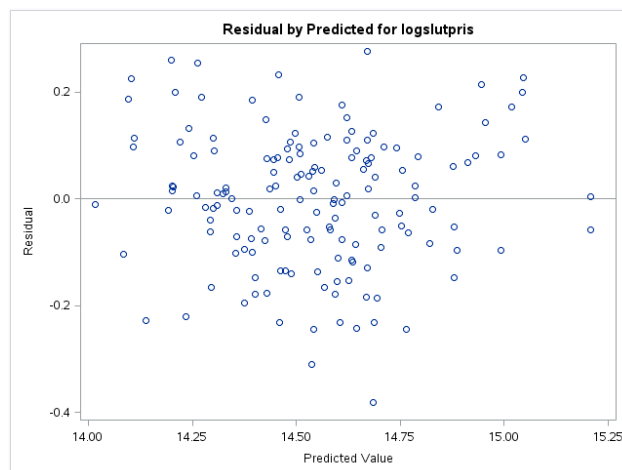


Figur 5: Plot av residualerna mot de predikterade värden då responsvariabeln är *slutpris* och *logboarea*, *säljår*, *våning*, *ålder*, *avgiftarea* samt *arearum* som förklarande variabler.

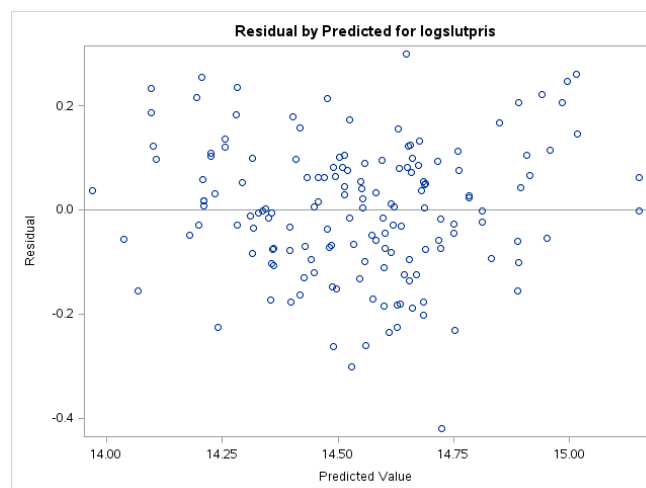


Figur 6: Plot av residualerna mot de predikterade värden då responsvariabeln är *logslutpris* och *logboarea*, *säljår*, *våning*, *ålder*, *avgiftarea* samt *arearum* som förklarande variabler.

Av figur 4 och 6 ser vi att residualerna är jämt utspridda kring noll vilket gör att båda figurerna antar homoscedasticitet jämfört med figur 5 där residualerna är inte jämnt utspridda kring värdet noll, dvs. figuren antar heteroscedasticitet, vilket inte är önskvärt. Vi ser att i figurerna 4 och 6 så har vi två punkter som avviker kraftigt från linjen, med andra ord så har vi två outliers. Den högsta outliern beror på att åldern på lägenheten är 110 år och den lägsta outliern beror på att den hade det lägsta priset vid försäljning av lägenheten dvs. 900 000 kr. Vi tar bort båda observationerna och vi får följande två plottar av residualerna mot predikterade värden.



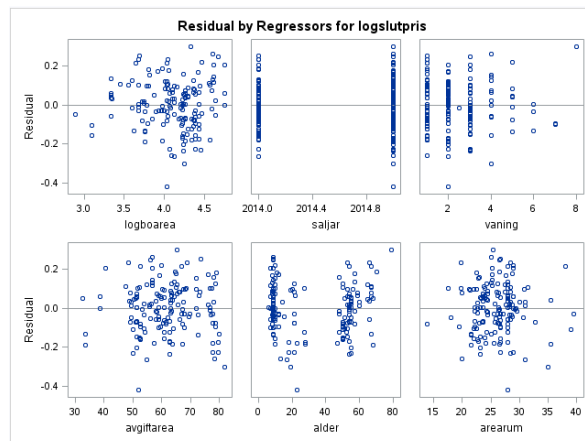
Figur 7: Plot av residualerna mot de predikterade värden då responsvariabeln är *logslutpris* och *boarea*, *säljår*, *våning*, *ålder*, *avgiftarea* samt *arearum* som förklarande variabler.



Figur 8: Plot av residualerna mot de predikterade värden då responsvariabeln är *logslutpris* och *logboarea*, *säljår*, *våning*, *ålder*, *avgiftarea* samt *arearum* som förklarande variabler.

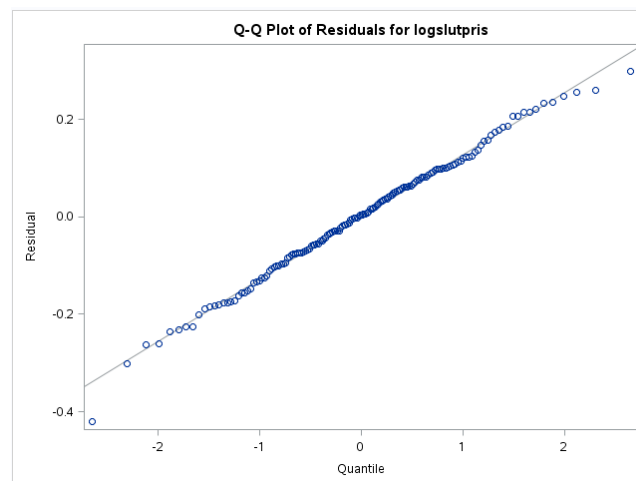
Vi ser att figurerna 7 och 8 ger oss bättre plottar än de vi fick tidigare, se figur 4 och 6, och därefter kan vi konstatera att det råder homoscedasticitet. Vi behöver inte undersöka modellen som består av *slutpris* som responsvariabel och *logboarea*, *säljår*, *våning*, *ålder*, *avgiftarea* samt *arearum* som förklarande variabler eftersom det ger oss liknande resultat som figur 5, dvs. att det råder heteroscedasticitet, vilket leder till att vi utesluter modellen.

Det leder till att vi får två olika modeller där den första modellen består av responsvariabeln *logslutpris* samt de förklarande variablerna *boarea*, *säljår*, *våning*, *ålder*, *avgiftarea* samt *arearum*. Den andra modellen är liknande som den första modellen men det som skiljer mellan de är att vi har logaritmerat *boarea*, dvs. vi får en ny förklarande variabel, *logboarea*. Vi börjar med att undersöka korrelationen mellan x-variablerna i den andra modellen med hjälp av *VIF*. Det ger oss att de förklarande variablerna är nästan okorrelerade eftersom alla variabler antar värden under 5. Nästa steg är att undersöka om residualplottarna av de förklarande variablerna är rimliga.



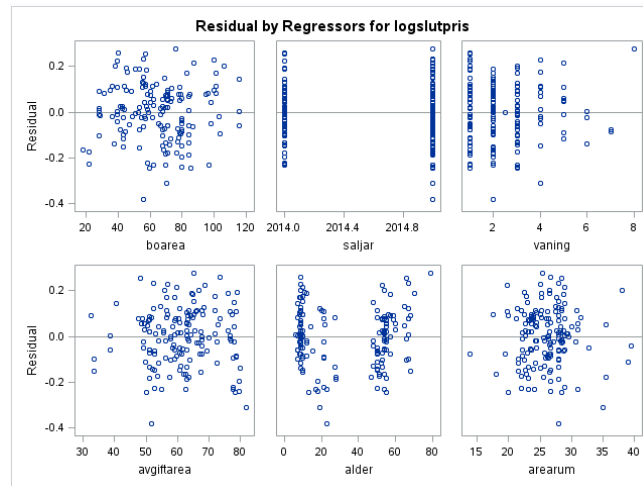
Figur 9: Residualplot av de förklarade variablerna.

Av figur 9 kan vi se att residualerna är jämt fördelade kring linjen vilket tyder på att vi har ett linjärt samband mellan responsvariabeln *logslutpris* och de förklarande variablerna *logboarea*, *säljår*, *våning*, *ålder*, *avgiftarea* samt *arearum*. Vi ska dessutom undersöka om feltermerna antar sin fördelning, dvs. normalfördelning, genom att plotta en Q-Q plot.



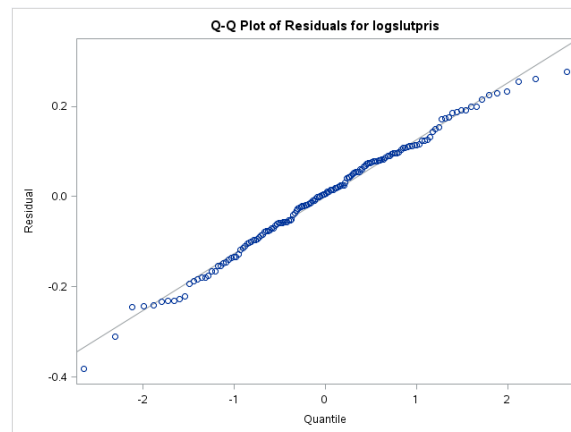
Figur 10: Q-Q plot.

För att feltermerna ska vara normalfördelade ska punkterna ungefär följa den rätta linjen, vilket vi kan se i figur 10 att den gör. Vi utför en liknande procedur för modellen med responsvariabeln *logslutpris* och de förklarande variablerna *boarea*, *säljår*, *våning*, *ålder*, *avgiftarea* samt *arearum*. Som vi tidigare har nämnt så är de förklarande variablerna nästan okorrelerade vilket leder till att vi inte behöver tänka på multikolaritet. Vi undersöker residualplottarna av de förklarande variablerna för att se om det råder linjärt samband mellan responsvariabeln och de förklarande variablerna.



Figur 11: Residualplot av de förklarande variablerna.

I figur 11 ser vi att residualerna är jämnt fördelade kring värdet noll vilket visar oss att det råder linjärt samband mellan *logslutpris* och *boarea*, *säljår*, *våning*, *ålder*, *avgiftarea* samt *arearum*. Vi undersöker om modellens feltermar antar normalfördelning genom att plotta en normalfördelningsplott, Q-Q plot.



Figur 12: Q-Q plot.

Av figur 12 ser vi att punkterna följer ungefär den rätta linjen, vilket vi konstaterar att feltermerna antar normalfördelning.

4.1 Modellval

Till slut har vi två lämpliga modeller med följande parameterskattningar, standardfel och p -värde.

- *Modell 1:*

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	Intercept	1	-460.94838	42.78949	-10.77	<.0001
boarea	boarea	1	0.00580	0.00061044	9.50	<.0001
saljar	saljar	1	0.23623	0.02124	11.12	<.0001
vaning	vaning	1	0.03435	0.00774	4.44	<.0001
avgiftarea		1	-0.00877	0.00132	-6.66	<.0001
alder		1	-0.00425	0.00060545	-7.02	<.0001
arearum		1	-0.00730	0.00277	-2.64	0.0092

Tabell 4: Parameterskattningar av modell 1, *logslutpris* som responsvariabel.

- *Modell 2:*

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	Intercept	1	-470.23380	43.18086	-10.89	<.0001
logboarea		1	0.31419	0.03401	9.24	<.0001
saljar	saljar	1	0.24043	0.02144	11.21	<.0001
vaning	vaning	1	0.03403	0.00782	4.35	<.0001
avgiftarea		1	-0.00991	0.00128	-7.75	<.0001
alder		1	-0.00472	0.00059006	-8.00	<.0001
arearum		1	-0.00786	0.00280	-2.81	0.0056

Tabell 5: Parameterskattningar av modell 2, *logslutpris* som responsvariabel.

För att bestämma vilken modell som förklarar *logslutpriset* på bästa sätt får vi genom att välja den modell som har högt värde på förklaringsgraden. Vi kör en regression i SAS för varje modell och vi får följande förklaringsgrad.

- *Modell 1:* $R^2 = 0.7744$
- *Modell 2:* $R^2 = 0.7697$

Vi ser att *modell 1* har högre förklaringsgrad vilket leder till att *modell 1* förklarar variationen i *logslutpris* på bästa sätt.

5. Resultat

Till slut har vi kommit fram till att modell 1 är den lämpligaste modellen och med hjälp av parameterskattningarna i tabell 4 får vi ett samband mellan *logslutpris* och de förklarande variablerna på följande sätt:

$$\logslutpris = -460.94838 + 0.0058 \cdot boarea + 0.23623 \cdot säljår + 0.03435 \cdot våning \\ - 0.00877 \cdot avgiftarea - 0.00425 \cdot ålder - 0.00730 \cdot arearum + \epsilon,$$

Där $\epsilon \sim N(0, \sigma^2)$.

I vårt fall har vi nu en additiv modell som är logariterad och för att få ett samband med *slutpris* som responsvariabel får vi istället en multiplikativ modell:

$$slutpris = e^{-460.94838} \cdot e^{0.0058 \cdot boarea} \cdot e^{0.23623 \cdot säljår} \cdot e^{0.03435 \cdot våning} \cdot e^{-0.00877 \cdot avgiftarea} \\ \cdot e^{-0.00425 \cdot ålder} \cdot e^{-0.00730 \cdot arearum} \cdot \epsilon',$$

Där $\epsilon' = e^\epsilon$ är lognormalfördelad. Utifrån modellen kan vi tolka det som att om vi ökar en enhet av exempelvis den förklarande variabeln *boarea*, då förändras responsvariabeln *slutpris* med en multiplikativ faktor på $e^{0.0058}$. Med andra ord, responsvariabeln förändras med en multiplikativ faktor e^{β_i} vid ökning av en enhet av den *i*-te förklarande variabeln.

Som vi har tidigare skrivit så förändras *slutpriset* med $e^{0.0058} = 1.0058$ vid ökning av en enhet av den förklarande variabeln *boarea*, i det här fallet ökning av $1 m^2$. Eftersom parameterskattningen är positiv tolkar vi det som att *boarea* har en positiv påverkan på *slutpris* vilket är rimligt då många köpare är villiga att betala mer ju större boarea lägenheten har. Vidare så kan vi se från den multiplikativa modellen att både *säljår* och *våning* har även en positiv påverkan på slutpriset med en multiplikativ faktor på $e^{0.23623} = 1.2665$ respektive $e^{0.03435} = 1.0349$ vilket kan tolkas som att slutpriset ökas om köparen köper lägenheten nästa år samt att ju högre våning lägenheten ligger på så ökar slutpriset. *Avgiftarea*, *ålder* och *arearum* har negativa parameterskattningar vilket betyder att de har negativ påverkan på slutpriset med en multiplikativ faktor på $e^{-0.00877} = 0.9913$, $e^{-0.00425} = 0.9958$ och $e^{-0.0073} = 0.9927$ vid ökning av $1 kr/m^2$, år samt m^2/rum .

6. Diskussion

Vår uppgift var att undersöka vilka faktorer som påverkar mest slutpriset vid försäljningar av lägenheter i Spånga. Vi har kommit fram till att *modell 1*, se avsnitt 4.1, är den lämpligaste modellen som förklarar slutpriset. Om vi undersöker lite närmare på resultatet som vi har fått fram enligt *modell 1* så ser vi att variabeln *boarea* har en positiv effekt på slutpriset på så sätt att priset ökas med 0.58% vid ökning av 1 m^2 . Det kan förstås låta rimligt eftersom många hellre skulle betala mer för lägenheten utifrån hur stor kvadratmeter lägenheten har men i detta fall så är ökningen lite lägre än förväntad då vi trodde att *boarea* skulle leda till att slutpriset skulle öka betydligt mer. En anledning kan vara att lägenheterna ligger en bit bort från innerstaden där *boarea* kan vara en viktigare faktor för köparen som vill köpa en lägenhet där än att köpa en lägenhet i Spånga. En annan variabel som har en positiv påverkan på slutpriset är *våning*. Av den variabel kunde vi se att priset ökas med 3.49% vid ökning av 1 våning, dvs. ju högre våning lägenheten ligger på ökas slutpriset. Eftersom majoriteten av lägenheterna i Spånga som vi undersöker inte har en utsikt över sjön, den närmaste utsikten över en sjö för en lägenhet i Spånga ligger 2.1 km bort, tror vi istället att säkerheten spelar en viktig roll. Ju högre våningsplan lägenheten ligger på så är sannolikheten att få inbrott i lägenheten mindre jämfört med om lägenheten ligger på bottenvåningen då de flesta inbrott sker just i bottenvåningen.[5] Variabeln *säljår* kunde vi se att den hade en stor påverkan på slutpriset då ökning av ett år ger oss 26 % ökning på slutpriset vilket vi ser i vårt data att slutpriset ökas från år 2014 till 2015. Till slut har vi de tre variablerna *avgiftarea*, *ålder* och *arearum* som har en negativ effekt på slutpriset vilket leder till att priset minskas med ungefär 1 % då en av de tre variablerna ökas med en enhet. I förhand visste vi redan att dessa variabler inte skulle ha någon positiv effekt på slutpriset vilket stämmer med hjälp av resultatet.

Eftersom vi var tvungna att ta bort variabeln *driftkostnad* i datamaterialet, se avsnitt 3.1, fick vi mindre observationer att undersöka i vårt arbete. Det kan leda till att skattningarna på parametrarna inte blir rimliga och på så sätt får vi ett osäkert resultat. För att undvika det kan vi lägga till flera observationer samt flera variabler bl.a. om lägenheten har hiss, balkong, garage för bilparkering eller avstånd till centrum. Men informationen om variablerna ska vara tillgängliga för samtliga observationer annars går det inte att ta med de i undersökningen vilket leder till att man får arbeta med få observationer. Därför bör man ha ett datamaterial som består av många observationer som är korrekt med rätt information om alla variabler för varje observation och det kan man få bl.a. från mäklarfirmorna eftersom om man gör det manuellt för varje lägenhet slösar man flera timmar åt att samla in information om variablerna, vilket inte är önskvärt.

7. Referenser

- [1] *Notes in Econometrics*. Patrik Andersson, Joanna Tyrcha, juli 2014.
- [2] *Linjär Statistiska Modeller*. Rolf Sundberg, oktober 2015.
- [3] *Linjär Statistiska Modeller*, sid. 74. Rolf Sundberg, oktober 2015.
- [4] www.booli.se/slutpriser/spanga/393538/?objectType=l%C3%A4genhet&minSoldDate=2014-01-01&maxSoldDate=2015-12-31&sort=soldDate&direction=desc
- [5] <http://media.folksam.se/sv/2013/09/27/flest-inbrott-under-fjarde-kvartalet/>