Stockholms universitet

# Har faktorer olika inverkan på bostadspriser i olika städer?

Filip Axelsson

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2021:17
Matematisk statistik
Juni 2021

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholms universitet

Matematisk statistik
Stockholms universitet
Kandidatuppsats **2021:17**
http://www.math.su.se/matstat

# Har faktorer olika inverkan på bostadspriser i olika städer?

Filip Axelsson*

Juni 2021

**Sammanfattning**

Bostadsmarknaden är ett aktuellt område då den berör så pass många människor. Det finns en rad olika rapporter som beskriver och predikterar marknaden. En sak som är gemensam för dessa är att de oftast riktar in sig på större städer som Stockholm eller på hela Sverige. Det lyfts sällan fram rapporter kring hur bostadsmarknaden utvecklar sig i mindre städer, som exempelvis Östersund. Varför skrivs det färre rapporter och artiklar kring småstäder? En anledning skulle kunna vara att intresset för dessa rapporter är högre i stora städer, som exempelvis Stockholm. Möjligtvis kan behovet av att veta hur marknaden rör sig vara större där, eftersom bostäderna vanligtvis är dyrare i dessa städer och att det därmed leder till att bostadsköparna behöver ta större lån. När belåningsgraden är hög skulle det eventuellt kunna leda till ekonomiska konsekvenser om bostadsmarknaden skulle fallera. En annan anledning kan vara att bostadsmarknaden är densamma i små och stora städer. Detta leder till huvudfrågan, är faktorernas påverkan samma i Östersund som i Stockholm? Svaret är nej, de har inte samma påverkan i de två städerna. Detta kan visas genom att konstruera olika regressionsmodeller och sedan utföra hypotestester. Det kunde analyseras att utropspriset förklarade 98 % utav slutpriset, vilket ledde till misstankar om korrelation till responsvariabeln. Det fanns även misstankar kring multikollinearitet vilket resulterade till att utropspriset exkluderades som en förklarande variabel när linjära hypoteser utfördes. Det visade sig att åtminstone en utav faktorerna har en annan påverkan i Östersund och Stockholm. En modell med hjälp av stegvis variabelselektion anpassad för Östersund konstruerades också. Därefter kontrollerades ifall denna modell, applicerad i Stockholm, genererade att faktorerna hade samma påverkan. Hypotesen testades med hjälp av Chow's test of structural change och svaret blev även denna gång att faktorernas påverkan i Östersund och Stockholm var olika. Det finns dock olika tillvägagångssätt att konstruera dessa modeller som skulle kunna påverka resultatet, men det finns även andra faktorer som exempelvis tid och distrikt/områden i de två städerna som skulle kunna påverka resultatet. Detta med andra synpunkter tas upp i rapportens diskussion.

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige.
E-post: filip.axee@gmail.com. Handledare: Taras Bodnar, Tony Johansson och Måns Karlsson.

**Abstract**

The real estate market is currently a well discussed topic since it effects so many people. There are a number of different reports that describe and predict the market. One common thing to these reports is that they usually focus on larger cities such as Stockholm, or even in a bigger picture such the whole country Sweden. These reports rarely discuss on how the real estate market develops in Östersund. Why are there fewer reports and articles written about small towns such as Östersund? One reason could be that the interest in these is greater in larger cities such as Stockholm. Possibly there is a greater need to know how the market develops, this is because for the most part the apartments are more expensive in these cities. This usually leads to people have larger loans, which could possibly lead to financial consequences if the real estate market were to crash. Another reason would be that the real estate market is the same in Östersund and Stockholm. This leads to the main question, is the influence of the factors the same in Östersund as in Stockholm? The answer is no, they do not have the same effect in the two cities, this can be shown by constructing different regressions and then preforming hypothesis test. It could be analyzed that the asking price explained 98 % of the final price, which led to suspicions of correlation to the response variable. There were also suspicions about multicollinearity which resulted in the asking price being excluded as an explanatory variable linear hypotheses being performed. It turned out that at least one of the factors had different effect in Östersund and Stockholm. A model specified for Östersund was also constructed with help of a stepwise variable selection. The next step was to check if this model applied in Stockholm generated that the factors had the same effect. This hypothesis was tested by using Chow's test of structural change, the answer was also this time that the factors impact in Östersund and Stockholm was different. However, there are different ways of constructing these models that could affect the result, there are also other factors such as time and district/areas could have affected the result. This with other points of view is addressed in the report's discussion.

**Förord**

Jag vill rikta ett stort tack till mina tre handledare Taras Bodnar, Tony Johansson och Måns Karlsson för all vägledning och stöd som de har gett mig under mitt examensarbetet. Jag vill även tacka min familj, men framförallt min farmor och farfar för deras stora intresse för min skolgång.

# Innehåll

# 1 Introduktion

## 1.1 Inledning

Bostadsmarknaden är något som alltid har varit aktuellt och kommer förmodligen alltid förbli aktuellt då denna marknad berör så pass många människor. Media har alltid varit flitiga med att skriva artiklar och redovisat rapporter som ska visa hur marknaden har utvecklats samt prognoser inför framtiden. Att bostadspriserna generellt sett har gått upp, kan ha lett till att människor har tagit större lån eller utökat sina bolån för att finansiera en affär (Emanuelsson, Katinic & Spector, u. å.). Detta i sin tur skulle eventuellt kunna leda till ekonomiska konsekvenser vid en bostadskrasch. I dagsläget är Sverige inne i en pandemi vilket säkerligen har bidragit till osäkerhet kring bostadsmarknaden, vilket gör att ämnet är mer aktuellt än någonsin. En lämplig fråga som kan ställas är varför majoriteten utav dessa artiklar och rapporter riktar sig mot större städer eller mer generellt över riket. Det kan vara av många olika anledningar intresset utav dessa rapporter har större efterfrågan i de större städerna. En annan anledning skulle kunna vara att faktorerna har samma påverkan på bostadsmarknaden i alla städer. Utifrån den sista hypotesen och då det finns färre artiklar skrivna om mindre städer blir det mer intressant att ta fram information kring bostadsmarknaden för en utav Sveriges mindre städer, Östersund. Detta leder till huvudfrågan: har de förklarande variablerna i en modell för bostadspriser olika påverkan i Östersund och Stockholm? Gregory C. Chow är en ekonom som framförde att många linjära modeller ofta har använts vid tillämpningar inom ekonomin för att beskriva olika förhållanden (Chow, 1960). Chow beskriver utifrån detta att en lämplig fråga då blir om en linjär regression är och förblir samma i två olika tidsperioder. Denna fråga leder till hypotestestet The Chow test of structural change, vilket skapar en förutsättning till att kunna besvara huvudfrågan.

## 1.2 Syfte & metod

Många rapporter och artiklar har sitt huvudsyfte att förklara och prediktera lägenhetspriser i storstäder som Stockholm, dock finns det ytterst lite skrivet kring hur olika faktorer påverkar bostadspriserna i mindre städer som Östersund. Målet är att skapa en större förståelse angående om bostadsmarknaden skiljer sig från en stad till en annan. Genom att tillämpa en multipel linjär regressionsmodell kan det kontrolleras ifall faktorernas påverkan är samma i Östersund, som de är i Stockholm. Det är även av intresse att konstruera en modell som är anpassad för Östersund och sedan kontrollera om denna modells faktorer har samma påverkan i Stockholm. Målet är även att få en ökad förståelse och kunskap inom regressionsanalys.

I denna rapport kommer det genom användning av forward-selection, backward-elimination och stepwise-selection byggas en modell som beskriver hur olika faktorer påverkar bostadsmarknaden. The Chow test of structural change kommer att tillämpas för att kontrollera

om faktorerna är samma för de båda städerna. Genom hela arbetet kommer programmeringsspråket R att användas (R Core Team, 2020).

Datamängden som kommer användas för att utföra denna analys kommer hämtas ifrån Booli:s API. Det finns 498 observationer för Östersund som sträcker sig i en tidsperiod från 2020-01-01 till 2021-01-22 samt 6296 observationer för Stockholm, totalt 6794 observationer.

# 2 Teori

## 2.1 Linjär regression

Linjär regressionsanalys innebär att exempelvis modulera modeller. Det omfattar även att studera förhållandet mellan en responsvariabel och en eller flera förklarande variabler. I denna rapport framkommer två olika typer av linjär regressionsanalys, enkel och multipel.

### 2.1.1 Enkel linjär regression

Enkel linjär regression definieras som följande (Sundberg, 2020, Ch 3.1):

$$Y_i = \alpha + \beta x_i + \varepsilon_i \; i = 1, 2, \ldots N$$

$Y_i$ är responsvariabeln, $x_i$ är den förklarande variabeln, både $\alpha$ och $\beta$ räknas som parametrar, där $\alpha$ kallas för interceptet. Den sista termen $\varepsilon_i$ är en felterm, som är oberoende och normalfördelad med väntevärde 0 och varians $\sigma^2$ för alla $i$. Antalet observationer modellen anpassas på betecknas av $N$.

### 2.1.2 Multipel linjär regression

Om fallet skulle vara att flera förklarande variabler skulle kunna påverka responsvariabeln, så kan en multipel linjär regression tillämpas. Denna modell definieras enligt följande (Sundberg, 2020, Ch 3.2):

$$Y_i = \alpha + \beta_1 x_{1i} + \ldots + \beta_m x_{mi} + \varepsilon_i, \; i = 1, 2 \ldots N \tag{1}$$

Även som i fallet för enkel linjär regression så betraktas $Y_i$ som resposvariabel, skillnaden är antalet förklarande variabler. Både $\alpha$ och $\beta_m$ räknas som parametrar, $\alpha$ kallas även för intercept. Även i denna modell antas $\varepsilon_i$ vara oberoende och likafördelade $N(0, \sigma^2)$. Det som är nytt är att $m$ representerar antalet förklarande variabler. Genom följande matrisnotation kan modellen för multipel linjär regression (1) skrivas om (Andersson, Lindensjö & Tyrcha, 2019, Ch 2):

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

där

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

### 2.1.3 Parameterskattning

Genom att tillämpa minsta-kvadrat metoden på data kan den okända parametern $\boldsymbol{\beta}$ skattas. Detta innebär att minimera kvadratsumman av residualerna (Andersson m. fl., 2019, Ch 2).

$$\sum_{i=1}^{N} (Y_i - \alpha - \beta_1 x_{1i} - \beta_2 x_{2i} - \ldots - \beta_m x_{mi})^2$$

Det går att bevisa att ekvationen som minimerar uttrycket när $\boldsymbol{\beta}$ är okänd och ger parameterskattningen till $\boldsymbol{\beta}$ är följande:

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

där $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$ (Sundberg, 2020, Ch 3.2). Det är värt att poängtera att feltermer är något som används i teorin, med residualer är skillnaden mellan det faktiska värdet på $\mathbf{Y}$ och det skattade värdet $\hat{\mathbf{Y}}$, det vill säga (Andersson m. fl., 2019, Ch 2):

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}, \text{ där } \hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$$

Eftersom $\sigma^2$ är en okänd parameter skattas den med observationerna från data. Den väntevärdesriktiga skattning av $\sigma^2$ ges utav följande:

$$\hat{\sigma}^2 = \frac{\mathbf{e}^T\mathbf{e}}{N - m - 1}$$

$N - m - 1$ representerar antalet frihetsgrader i modellen.

### 2.1.4 Antaganden

För att kunna applicera en linjär regressionsmodell krävs det att vissa antaganden är uppfyllda (Andersson m. fl., 2019, Ch 2, Ch 4):

- Linjäritet
  Detta innebär att ett linjärt samband finns mellan responsvariabeln och de förklarande variablerna. Skulle detta antaganden inte uppfyllas resulterat det i bland annat att parameterskattningarna inte är väntevärdesriktiga, det vill säga $E[\hat{\boldsymbol{\beta}}] \neq \boldsymbol{\beta}$. De skulle även sakna någon betydelse eftersom det inte skulle finnas något samband mellan responsvariabeln och de förklarande variablerna.

- Normalfördelade residualer Residualerna för modellen ska vara oberoende och normalfördelade med varians $\sigma^2$ och väntevärde 0. Detta antagande krävs för att kunna utföra konfidensintervall samt för att kunna konstruera hypotestester. Notera att detta antaganden endast krävs vid små dataset, vilket kan läsas i kapitel 2.3.4.

- Homoskedasticitet
  Variansen för residualerna är konstant oavsett värden för förklaringsvariablerna.

  $$V(\varepsilon_i|\mathbf{X}_i) = \sigma^2 \quad \forall i = 1, \ldots, N.$$

  Konsekvensen av att detta inte uppfylls är exempelvis att slutsatserna grundande på t-testerna blir felaktiga. Detta problem går dock att åtgärda genom transformation i hopp att konstant varians ska uppnås.

- Ej perfekt multikollinearitet Det ska inte finnas något linjärt beroende mellan de förklarande variabler som inkluderas i modellen. Variabler med hög multikollinearitet leder till missvisande parameterskattningar.

## 2.2 Anpassningsmått

### 2.2.1 Förklaringsgrad $R^2$

Förklaringsgrad beskrivs som ett utav de vanligaste anpassningsmåtten inom regression (Sundberg, 2020, Ch 3.2.2). Måttet är ett tal som ligger mellan 0 och 1 som beskriver den andel av variationen i responsen som förklaras av kovariaten. Förklaringsgraden definieras enligt:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}.$$

Notera att $\bar{Y}$ är medelvärdet av responsvektorn $\mathbf{Y}$.

### 2.2.2 Justerad Förklaringsgrad $R^2_{adj}$

Tillägg av en förklarande variabel kommer alltid att resultera i ett högre $R^2$ värde. Detta beror på att modellen får fler parametrar att arbeta med, vilket sin tur alltid ökar värdet på $R^2$. Notera att det även gäller om den förklarande variabeln skulle vara slumpgenererad (Sundberg, 2020, Ch 3.2.2). Denna egenskap hos $R^2$ gör att man kan blåsa upp värdet med insignifikanta kovariat. Därav introduceras en annat anpassningsmått, Justerad Förklaringsgrad $R^2_{adj}$ vilket tar hänsyn till antalet parametrar i modellen. Detta anpassningsmått mäter variansreduktionen som har uppnåtts i den aktuella modellen, det vill säga hur mycket variansen minskar jämfört med en modell där inga förklarande variabler är inkluderade. Den Justerade Förklaringsgraden definieras enligt följande (Sundberg,

2020, Ch 3.2.2) :

$$R^2_{adj} = 1 - \frac{\hat{\sigma^2}}{\hat{\sigma^2_0}} \Leftrightarrow R^2_{adj} = 1 - (1 - R^2)\frac{N-1}{N-m-1}.$$

Där $\hat{\sigma^2_0} = \frac{\sum(Y_i - \bar{Y})^2}{N-1}$, det vill säga den variansskattning där ingen förklaringsvariabel är inkluderar i modellen. $N$ definieras som antalet observationer och $m$ som antalet parametrar.

### 2.2.3 Akaike Information Criteria (AIC)

Det finns olika sätt att jämföra modeller i samma datamängd, AIC är ett alternativ. AIC definieras enligt följande (Held & Sabanés Bové, 2014, Ch 7)(Gordon, 2015, Ch 6.6):

$$AIC = -2\log(\hat{\theta}_{ML}) + 2(m+1).$$

Där $\hat{\theta}_{ML}$ representerar den skattade maximerade likelihooden och $m$ är antalet parametrar i modellen, det vill säga antalet förklarande variabler som är inkluderade i modellen. Vid jämförelse av modeller väljs den modellen med lägst AIC. Notera att man nödvändigtvis inte väljer en modell med lägst AIC-värde, om det finns en enklare modell med något högre AIC-värde. Då detta kriterium tar hänsyn förklaringsgraden samt antalet parametrar så tenderar AIC att straffa mer när en förklarande variabel läggs till än vad förklaringsgraden gör.

## 2.3 Hypotestest

En hypotes beskrivs som ett teoretiskt påstående eller som ett slags krav/antaganden. Målet med ett hypotestest är att ta ett beslut om vad man ska tro om verkligheten. En nollhypotes, $H_0$, ställs mot en alternativhypotes $H_1$, som tillsammans omfattar alla möjliga utfall. Man bestämmer sedan en signifikansnivå $\alpha$, som beskriver hur mycket data måste avvika från vad man förväntar sig under $H_0$ för att vi ska ändra uppfattning. Om data avviker tillräckligt mycket förkastar man $H_0$ till förmån för $H_1$.

### 2.3.1 Hypotesprövning

En utav de viktigaste hypotestesterna är att kontrollera om koefficienterna i $\boldsymbol{\beta}$ är signifikant skilda från noll, det vill säga följande hypotestest ställs upp:

$$H_0 : \beta_i = 0 \qquad H_1 : \beta_i \neq 0$$

Testet använder följande teststatistika som ges under nollhypotesen, $H_0$ (Andersson m. fl., 2019, Ch 3):

$$T_i = \frac{\hat{\beta}_i}{\hat{\sigma}\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{ii}}} \sim t(N-m-1) \tag{2}$$

Detta innebär att $T_i$ är t-fördelad med $(N-m-1)$ frihetsgrader, nollhypotesen $H_0$ förkastas vid signifikansnivån $\alpha$ ifall $|T_i| \geq t_{\alpha/2}(N-m-1)$. En vanlig signifikansnivån är $\alpha = 0.05$, värdet på $t_{\alpha/2}(N-m-1)$ hämtas från en tabell över t-fördelningens kvantiler.

Ett annat hypotestest som kommer utföras i denna rapport är att kontrollera ifall samtliga faktorer har samma effekt på repsonsvariabeln och kan ställas upp enligt följande (Andersson m. fl., 2019, Ch 3):

$$H_0 : \beta_i = \beta_j, i \neq j \qquad H_1 : \beta_i \neq \beta_j, i \neq j$$

Där teststatistika är under nollhypotesen:

$$F_{ij} = \frac{(\hat{\beta}_i - \hat{\beta}_j)^2}{\hat{\sigma}^2((\mathbf{X}^T\mathbf{X})^{-1}_{ii} + (\mathbf{X}^T\mathbf{X})^{-1}_{jj} - 2(\mathbf{X}^T\mathbf{X})^{-1}_{ij})} \sim F(1, N-m-1) \tag{3}$$

Ett ytterligare test som är utav intresse är att kontrollera ifall samtliga koefficienter är lika med noll, vilket kan utföras genom att tillämpa följande hypotes och teststatistiska:

$$H_0 : \beta_i = 0, \ \forall i \qquad H_1 : \beta_i \neq 0, \text{ för något } i$$

$$F = \frac{R^2/m}{(1-R^2)/(N-m-1)} \sim F(m, N-m-1) \tag{4}$$

### 2.3.2 P-värde

Ett tydligare sätt att se om en hypotes ska förkastas eller inte är att införa ett så kallat p-värde. Detta värde anger under $H_0$, hur stor sannolikheten är att den givna statistikan är större än det värde som hämtades från fördelningstabellen. Detta kan även beskrivas med matematiska termer och det definieras som $\mathbb{P}(|T_i| \geq t_{\alpha/2}(N-m-1)|H_0)$ (Andersson m. fl., 2019, Ch 3). P-värdet ska vara mindre än den givna signifikansnivån $\alpha$ för att genera att koefficienten $\beta_i$ är signifikant, det vill säga att nollhypotesen $H_0$ förkastas.

### 2.3.3 The Chow test of structural change

En viktigt antagande för linjära regressions modeller är att förklaringsvariablernas koefficienter är konstanta. Genom att applicera The Chow test of structural change, som är en mer generaliserad metod av The Chow forecast, så kan det undersökas ifall förklaringsvariablerna har olika påverkan i olika delpopulationer. Ett typiskt exempel är att jämföra ifall det är en skillnad i konsumtionen för ett land i krigstid och fredstid. I denna rapport kommer fokus vara på att kontrollera ifall det är någon skillnad mellan förklaringsvariablernas effekter i städerna Östersund och Stockholm. Vi presenterar nu the Chow test of structural change

(Andersson m. fl., 2019, Ch 4):
Genom den obegränsade modellen:

$$\begin{bmatrix}\mathbf{Y}_1\\ \mathbf{Y}_2\end{bmatrix} = \begin{bmatrix}\mathbf{X}_1 & 0\\ 0 & \mathbf{X}_2\end{bmatrix} \cdot \begin{bmatrix}\boldsymbol{\beta}_1\\ \boldsymbol{\beta}_2\end{bmatrix} + \boldsymbol{\varepsilon} \tag{5}$$

Där $\boldsymbol{\beta}_1$ och $\boldsymbol{\beta}_2$ är de $m$-vektorer av parametrar för respektive undergrupp (Östersund respektive Stockholm). Detta ger att nollhypotesen av ingen strukturskillnad blir följande:

$$H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \boldsymbol{\beta}$$

Den begränsade modellen ges av:

$$\begin{bmatrix}\mathbf{Y}_1\\ \mathbf{Y}_2\end{bmatrix} = \begin{bmatrix}\mathbf{X}_1\\ \mathbf{X}_2\end{bmatrix} \cdot \boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{6}$$

Detta ger följande teststatistika för nollhypotesen:

$$\begin{aligned} F &= \frac{(RSS_{restricted} - RSS_{unrestricted})/(m+1)}{RSS_{unrestricted}/(N-2(m+1))} \\ &= \frac{(\tilde{\boldsymbol{e}}^T\tilde{\boldsymbol{e}} - \boldsymbol{e}^T\boldsymbol{e})/(m+1)}{\boldsymbol{e}^T\boldsymbol{e}/(N-2(m+1))} \sim F(m+1, N-2(m+1)) \end{aligned} \tag{7}$$

Variabeln $RSS$ (residual sum of squares), representerar kvadratsumman av residualerna. Det finns $RSS_{restricted} = \tilde{\mathbf{e}}^T\tilde{\mathbf{e}}$ som anger kvadratsumman för den begränsade regressionen (6), det vill säga när man har samma regressionskoefficienter för båda städerna. Den andra parametern $RSS_{unrestricted} = \mathbf{e}^T\mathbf{e}$ anger residualernas kvadratsumma utav den obegränsade regressionen (5), det vill säga när varje stad har en unik koefficientvektor. Vi förtydligar hur teststatistikan appliceras i fallet då två områden eller två tidsperioder testas mot varandra (Andersson m. fl., 2019, Ch 4):

$$F = \frac{(RSS - (RSS_1 + RSS_2))/(m+1)}{(RSS_1 + RSS_2)/((N_1+N_2) - 2(m+1))} \sim F(m+1, N-2(m+1)) \tag{8}$$

### 2.3.4 Icke-normalfördelningar och tester för stora datamängder

I föregående kapitel används t-fördelningen och F-fördelningen vid hypotestester, dessa teststatistiker bygger på att antagandet utav normalfördelade residualer uppfylls. I situationen utan detta antagende kommer den exakta fördelning av dessa teststatistiker bero på data och inte nödvändigtvis följa F- och t-fördelningar (Greene, 2002, Ch 6). Det visar sig dock att både den t- och F-fördelade teststatistikan fortfarande är användbara, i det mer generella fallet utan antagandet om normalfördelade residualer. De är istället uppfattade

som en approximation vars kvalité förbättras när antalet observationer ökar. Om data är godartad, ges den asymptotiska fördelningen för $\hat{\boldsymbol{\beta}}$ (Greene, 2002, Ch 6):

$$\hat{\boldsymbol{\beta}} \overset{a}{\sim} N[\boldsymbol{\beta}, \frac{\sigma^2}{N}\boldsymbol{Q}^{-1}] \text{ där } \boldsymbol{Q} = \text{plim}(\frac{\boldsymbol{X}^T\boldsymbol{X}}{N}).$$

Den kan vara värt att notera att den matematiska termen plim representerar sannolikhets gränsvärdet (probaility limit). Genom att anta följande tolkning att frånvaron utav normaliteten för residualerna, när antalet observationer $N$ växer, ger att normalfördelningen blir en allt bättre approximation till den sanna, men dock okända fördelningen av $\hat{\boldsymbol{\beta}}$. När $N$ växer, kommer fördelningen för $\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ att konvergera exakt till en normalfördelning. Resultatet grundar sig på centrala gränsvärdessatsen och beror inte på antagandet angående normalfördelade residualer. Det andra resultatet som behöver beaktas är det estimerade värdet utav $\sigma^2$ (Greene, 2002, Ch 6):

$$\text{plim } s^2 = \sigma^2, \text{ där } s^2 = \hat{\sigma}^2 = \frac{\boldsymbol{e}^T\boldsymbol{e}}{N - m - 1}.$$

Utifrån detta kan det nu erhållas några stora datamängds resultat för teststatistikor som visar hur processen ser ut i en ändlig observationsmängd med icke-normalfördelade residualer. Teststatistikan som används vid hypotesen att en utav koefficienterna $\beta_k$ är lika med ett specifikt värde $\beta_k^0$, kan skrivas på följande sätt (Greene, 2002, Ch 6):

$$t_k = \frac{\sqrt{N}(\hat{\beta}_k - \beta_k^0)}{\sqrt{s^2(\boldsymbol{X}^T\boldsymbol{X}/N)^{-1}_{kk}}}. \tag{9}$$

Under nollhypotesen, med normalfördelade residualer är $t_k$ exakt t-fördelad med $N - m$ frihetsgrader. I situationen då residualerna ej är normalfördelade är dock fördelningen okänd. Det går dock utifrån resultaten ovan (9), samt med mer komplicerade uträkningar komma fram till att nämnaren utav $t_k$ kommer att konvergera till $\sqrt{\sigma^2\boldsymbol{Q}^{-1}_{kk}}$ när antalet observationer går mot oändligheten. Detta leder till att den stora observationsmängdens fördelning för $t_k$ är samma som (Greene, 2002, Ch 6):

$$\tau_k = \frac{\sqrt{N}(\hat{\beta}_k - \beta_k^0)}{\sqrt{\sigma^2\boldsymbol{Q}^{-1}_{kk}}}.$$

Dock är $\tau_k = (\hat{\beta}_k - E[\hat{\beta}_k])/(\text{Asy. Var}[\hat{\beta}_k])^{1/2}$ från den asymptotiska normalfördelningen, vilket ger att $\tau_k$ har en standard normal asymptotiskt fördelning, vilket i sin tur ger den stora observationsmängdens fördelning för t-statiskern. Detta leder slutligen till att som approximation för stora observationsmängder kommer det att användas en standard normalfördelning för att approximera den sanna fördelningen för testsatiskern $t_k$ och även

använda de kritiska värdena från en standard normalfördelning vid hypotestester (Greene, 2002, Ch 6). Notera att detta resultat endast gäller då $N$ är stort, det vill säga när ett stort antal observationer används. Samma tillvägagångssätt kommer att användas för att analysera F-statistikan för hypotestestet för $m$ stycken linjära restriktioner. Första steget blir att visa att med normalfördelade residualer konvergerar $m \cdot F$ till en $\chi^2$-fördelning när observationer $N$ växer. Sedan kommer det visas att detta resultat faktiskt inte beror på att resiudalerna är normalfördelade utan det förlitar sig på den centrala gränsvärdessatsen. Sista och slutliga steget är som ovan, att ange de lämpliga kritiska värdena som kan användas för denna testsatistika, som också även endast gäller vid då stort antal observationer används (Greene, 2002, Ch 6).
Den F-statistika som används för att testa $J$ linjära restriktioner, $\boldsymbol{R\beta} - \boldsymbol{q} = \mathbf{0}$. Där $\boldsymbol{R}$ är en $r \times m$ matris som representerar de linjära restriktionerna och $\boldsymbol{q}$ är en $q \times 1$ matris där $q < m$ som representerar värdet restriktionen ska anta(Andersson m. fl., 2019, Ch 3) Utifrån detta med normalfördelade residualer under nollhypotesen, resulterar till att den exakta fördelningen utav denna statiska är $F(J, N - m - 1)$. En mer generell överblick av $F$ ges av (Greene, 2002, Ch 6):

$$F = \frac{(\boldsymbol{R\hat{\beta}} - \mathbf{q})^T(\mathbf{R}[\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}]\mathbf{R}^T)^{-1}(\boldsymbol{R\hat{\beta}} - \mathbf{q})}{J(s^2/\sigma^2)}.$$

Då plim $s^2 = \sigma^2$ och plim$(\mathbf{X}^T\mathbf{X}/N) = \mathbf{Q}$, ger det att nämnaren för $F$ kommer konvergera till $J$, och termen $[\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}]$ i täljaren kommer förhålla sig på samma sätt som $(\sigma^2/N)\mathbf{R}\mathbf{Q}^{-1}\mathbf{R}^T$. Oavsett vilken fördelningen detta är, och om $F$ har en begränsad fördelning kommer detta vara samma begränsade fördelning som för följande (Greene, 2002, Ch 6):

$$\begin{aligned} W^* =& \frac{1}{J}(\boldsymbol{R\hat{\beta}} - \mathbf{q})^T[\mathbf{R}(\sigma^2/N)\mathbf{Q}^{-1}\mathbf{R}^T]^{-1}(\boldsymbol{R\hat{\beta}} - \mathbf{q}) \\ &= \frac{1}{J}(\boldsymbol{R\hat{\beta}} - \mathbf{q})^T\{\text{Asy. Var}[\boldsymbol{R\hat{\beta}} - \boldsymbol{q}]\}^{-1}(\boldsymbol{R\hat{\beta}} - \mathbf{q}). \end{aligned}$$

Detta utryck är (1/J) gånger en Wald statistika, baserad på den asymptotiska fördelningen. Det stora observationsmängdens fördelningen för $W^*$ kommer därav att vara en (1/J) gånger en $\chi^2$-fördelning med $J$ frihetsgrader. När residualerna följer en normalfördelning kommer $J \cdot F$ att konvergera mot en $\chi^2$-fördelning med $J$ frihetsgrader. Det väsentliga resultatet här är att fördelningen för den Wald statistika är uppbyggd på fördelningen för $\hat{\boldsymbol{\beta}}$, som är asymptotiskt normal även i den situationen där resiudalerna inte är normalfördelade. Detta leder till att den lämpliga testsatiskan vid stora observationsmängder är $\chi^2$ fördelade (chi-squared = JF), även detta resultatet bygger centrala gränsvärdessatsen.
För att sammanfatta kort vad som har redovisats, är att även vid situationer då resiudalerna inte är normalfördelade kan fortfarande teststatistikor som ursprungligen följer antingen t-fördelningen eller F-fördelningen fortfarande användas, dock måste de kritiska värdena bytas ut och hämtas istället ifrån en standardnormalfördelning respektive en $\chi^2$-

fördelning med $J$ frihetsgrader. Notera att testen är tillförlitliga speciellt för stora observationsmängder.

## 2.4 Multikollinearitet

Vid många potentiella förklaringsvariabler kan ett linjärt beroende uppstå mellan dem, eller ett nästan linjärt beroende. Om detta skulle förekomma kan problem uppstå vilket även beskrevs i kapitel 2.1.4. Det kan till exempel leda till att variabler som förväntades vara signifikanta i modellen inte är det. Multikollinearitet kan till exempel upptäckas genom det statistiska måttet VIF.

### 2.4.1 Variansinflationsfaktor (VIF)

Detta statistiska mått beskrivs genom att först se på $Var(\hat{\beta}_j)$ för en enskild regressionskoefficient $\beta_j$ för att kunna få en tydligare bild (Sundberg, 2020, Ch 3.2.4). $Var(\hat{\beta}_j)$ uttrycks enligt följande:

$$Var(\hat{\beta}_j) = \sigma^2(\mathbf{S}^{-1})_{jj} = \frac{\sigma^2}{s_{jj}^2} \cdot VIF \tag{10}$$

Där den första faktorn $\frac{\sigma^2}{s_{jj}^2}$ i ekvation (19) är den varians som skulle ha erhållits vid enkel linjär regression av $y$ på $x_j$, eller om $x_j$ hade varit ortogonal mot resterande förklarande variabler. Den andra faktorn är den så kallade VIF-faktorn och kan beräknas enligt följande:

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

Där $R_j^2$ är förklaringsgraden för den j:te förklaringsvariabeln använd som responsvariabel och förklarad av resterande variabler. Det framgår även att den kritiska gränsen för höga värden för VIF antingen är 5 eller 10 (Sundberg, 2020, Ch 3.2.4). Det ska även noteras att variabler med höga VIF-värden inte direkt behövs tas bort, det är fler faktorer som även påverkar beslutet om vilken som ska tas bort.

## 2.5 Stegvis variabelselektion

Vid en stor uppsättning av potentiella förklaringsvariabler behövs det en metod för att kunna avgöra vilka som ska inkluderas i modellen. Det finns flera olika metoder som kan användas, dock följer samtliga samma princip att det läggs till eller tar bort en variabel. Detta upprepas tills dess att någon form av stoppkriterium uppfylls, som till exempel att vi når det lägsta AIC-värdet eller det högsta $R^2$-värdet. Notera att valet av stoppkriterium och selektionsmetod har stor betydelse för resultatet. I denna rapport kommer tre olika metoder användas för att hitta olika modeller. Dessa metoder kan beskrivas på följande sätt (Sundberg, 2020, Ch 3.2.3)

### 2.5.1 Backward elimination

I denna metod utgår vi ifrån samtliga förklarande variabler, sedan elimineras den variabel med högst p-värde. Denna procedur upprepas sedan tills att alla variabler är signifikanta. Notera om fallet skulle vara att flera variabler inte är signifikanta, så elimineras den variabel med högst p-värde.

### 2.5.2 Forward selection

Denna metod är motsatsen till föregående metod, Backward elimination. Man börjar med en modell som inte innehåller någon förklaringsvariabel och sedan adderar man en variabel i taget till modellen. I varje steg adderas endast den variabel som är mest signifikant i den nuvarande modellen, det vill säga ger lägst p-värde. Detta upprepas tills att det inte längre finns någon signifikant variabel att inkludera i modellen.

### 2.5.3 Stepwise regression

Stepwise regression kan ses som en kombination av de ovanstående metoderna, eller som en mer avancerad metod av Forward selection. Man utgår från en modell utan någon förklaringsvariabel och använder sedan samma procedur som i Forward selection, det vill säga adderar den variabel som är mest signifikant. Skillnaden är att efter varje steg kontrolleras om alla inkluderade variabler fortfarande är signifikanta. I situationen där en nuvarande variabel i modellen har övergått från att ha varit signifikant till att vara icke-signifikant efter adderat en annan variabel plockas den nuvarande variabeln bort från modellen, detta gäller även om flera variabler har övergått till att vara icke-signifikant. Notera att om en variabel har plockats bort, kan den aldrig inkluderas i modellen igen.

## 2.6 Transformation

### 2.6.1 Dummyvariabel

Kategoriska variabler är något som vanligt förekommer, för att kunna inkludera dessa variabler i modeller införs så kallade dummyvariabler. Dessa variabler kan endast anta värdet 1 eller 0 (Andersson m. fl., 2019, Ch 4). Ett exempel där det endast finns två värden att anta för den kategoriska variabeln är kön, man och kvinna. Genom att då skapa en dummyvariabel som antar värdet 1 om det är en man och 0 annars, det vill säga om det är en kvinna. Detta ger möjligheten att inkludera kön som en förklarande variabel i modellen. Om det skulle finnas $n$ stycken värden kategorin kan anta, skapas $n-1$ dummyvariabler. Genom att använda en utav de $n$ värden som referens så krävs endast $n-1$ dummyvariabler. Notera att en referens ej behövs användas vid konstruktion utav dummyvariabler. En annan sak som är viktigt att notera är att i denna rapport behandlas endast kategoriska variabler utan någon ordning.

### 2.6.2 Logtransformering

Logtransformering kan vara ett alternativ för att lösa problemet ifall sambandet mellan responsvariabeln och de förklarande variablerna visar att vara icke-linjärt eller om variansen ökar med responsvärdet. I Alices rapport (Vijitraphongsa, 2018) beskrivs det att genom logtransformering av responsvariabeln fås följande modell:

$$\log(y_i) = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_m x_{mi} + \epsilon_i \tag{11}$$

Det förklaras även att en eller flera förklarande variabler kan logtransformeras, samt att även både förklarande variabler och responsvariabeln kan logtransformeras samtidigt.

$$y_i = \alpha + \beta_1 \log(x_{1i}) + \beta_2 \log(x_{2i}) + \ldots + \beta_m \log(x_{mi}) + \epsilon_i \tag{12}$$

$$\log(y_i) = \alpha + \beta_1 \log(x_{1i}) + \beta_2 \log(x_{2i}) + \ldots + \beta_m \log(x_{mi}) + \epsilon_i \tag{13}$$

En multiplikativ modell kan fås genom att använda (11) och sedan transformera tillbaka till en icke-logaritmerad responvariabeln, vilket ger följande:

$$y_i = e^{\alpha} \cdot e^{\beta_1 x_{1i}} \cdot e^{\beta_2 x_{2i}} \cdot \ldots \cdot e^{\beta_m x_{mi}} \cdot \epsilon'$$

Där $\epsilon' = e^{\epsilon}$ samt är log-normalfördelad med väntevärde $e^{\sigma^2/2}$ och varians $(e^{\sigma^2} - 1)e^{\sigma^2}$.

# 3 Datamaterial

Datamaterialet är hämtat ifrån Booli's API 2021-01-23 (*Booli API*, 2021), som är ett svenskt företag som samlar in data angående bostadsmarknaden i Sverige. Denna rapport kommer främst att fokusera på lägenhetspriser i två olika städer Östersund och Stockholm, med huvudfokus på den mindre staden Östersund. För att kunna få tillång till detta material krävs en API-nyckel som ges utav Booli, sedan krävs en viss kunskap inom programmering. Datamaterialet består av totalt 6794 observationer varav 498 är ifrån Östersund respektive 6296 observationer är ifrån Stockholm. Den period som valdes är för att få en så rättvis bild av bostadsmarknaden som möjligt. Viruset Covid-19 kom till Sverige runt början av året 2020. Denna pandemi kan ha påverkat bostadsmarknaden, både negativt och positivt. Människor kan ha blivit mer intresserade av större bostäder då till exempel ett hemmakontor blivit mer aktuellt men det kan även ha skapats en viss rädsla över att ekonomin skulle rasa vilket skulle leda till nedgång av bostadspriser. Utifrån detta och för att utesluta skillnader av bostadsmarknaden före pandemin men även för att få en så aktuell bild av bostadsmarknaden valdes tidsperioden 2020-01-01 till 2021-01-22. Datamaterialet innehåller totalt 8 stycken olika variabler varav en utav dessa är en responsvariabel.

### 3.1 Beskrivning av variabler

- *soldPrice*
  Detta är responsvariabeln och anger det pris en lägenhet såldes för i enheten svenska kronor.

- *listPrice*
  Detta är en kontinuerlig variabel som anger utropspriset för en lägenhet mätt i enheten svenska kronor.

- *rent*
  Anger den månatliga avgiften som betalas till bostadsrättsföreningen, mätt i enheten svenska kronor. Denna variabel är kontinuerlig.

- *floor*
  Våningen lägenheten befinner sig på, detta är en diskret variabel.

- *livingArea*
  Denna variabel anger bostadsytan på lägenheten, mäts i enheten kvadratmeter $m^2$, och är en kontinuerlig variabel.

- *rooms*
  Anger antalet rum som lägenheten innefattar, detta är en diskret variabel.

- *constructionYear*
  Denna variabel är diskret och anger vilket år lägenheten/fastigheten byggdes.

- *Area*
  Detta är en kategorisk variabel som anger vilken stad lägenheten tillhör, det vill säga antingen Östersund eller Stockholm.

### 3.2 Transformation av data

Genom att transformera vissa variabler i datamängden kan de enklare anpassas och tolkas vid modellanpassningen. De variabler som kommer att transformeras är följande:

- *Area*
  För att kunna implentera denna variabel, kommer den transformeras till så kallad dummyvariaber. Detta innebär att om lägenheten ligger inom ett specifikt område kommer den dummyvariabeln anta värdet 1, medan de resterande antar värdet 0. För att kunna utföra vissa hypotestester kommer det inte användas en referens vid konstruktion av variabeln. Följande kommer introduceras i modellen

$$\begin{cases} D_i^{\ddot{O}} = 1 & \text{Om lägenheten ligger i staden Östersund.} \\ D_i^{\ddot{O}} = 0 & \text{Annars.} \end{cases}$$

$$D_i^S = 1 - D_i^{\ddot{O}} \text{ är dummyvariabeln för staden Stockholm}$$

- *constructionYear*
  Genom att istället ange hur länge sedan lägenheten byggdes relativt till nutiden (2021) kan en kontinuerlig variabel användas, vilket kan leda till tydligare resultat vid analys. Det vill säga $2021 - constructionYear$ kommer användas istället.

Ett utdrag på data efter transformationen kan visualiseras i *Tabell 1*

| listPrice | rent | floor | livingArea | rooms | soldPrice | constructionYear | Area |
|---|---|---|---|---|---|---|---|
| 825000 | 4501 | 2 | 53.5 | 2 | 875000 | 77 | Östersund |
| 925000 | 2694 | 3 | 45.0 | 1 | 900000 | 44 | Östersund |
| 1400000 | 3592 | 1 | 74.5 | 3 | 1650000 | 55 | Östersund |
| 1325000 | 3071 | 2 | 57.5 | 2 | 1500000 | 61 | Östersund |
| 1100000 | 3276 | 3 | 62.0 | 2 | 1380000 | 59 | Östersund |

Tabell 1: Exemepl på strukturen för data efter transformation.

# 4 Statistisk analys

## 4.1 Enkel linjär regression

Enkel linjär regression kommer att tillämpas för att skapa en initial bild av de förklarande variablernas korrelation med responsen.

| Varibel | Estimat | P-värde | $R^2$ |
|---|---|---|---|
| listPrice | 1.042 | <2e-16 | 0.9812 |
| floor | 203215.0105 | <2e-16 | 0.0238 |
| livingArea | 83870.0789 | <2e-16 | 0.4924 |
| rooms | 1852371.8228 | <2e-16 | 0.3349 |
| constructionYear | 19043.1621 | <2e-16 | 0.0576 |
| rent | 762.4804 | <2e-16 | 0.0975 |
| Area | -3657554.4261 | <2e-16 | 0.0942 |

Tabell 2: Enkel linjär regression för respektive förklarande variabel.

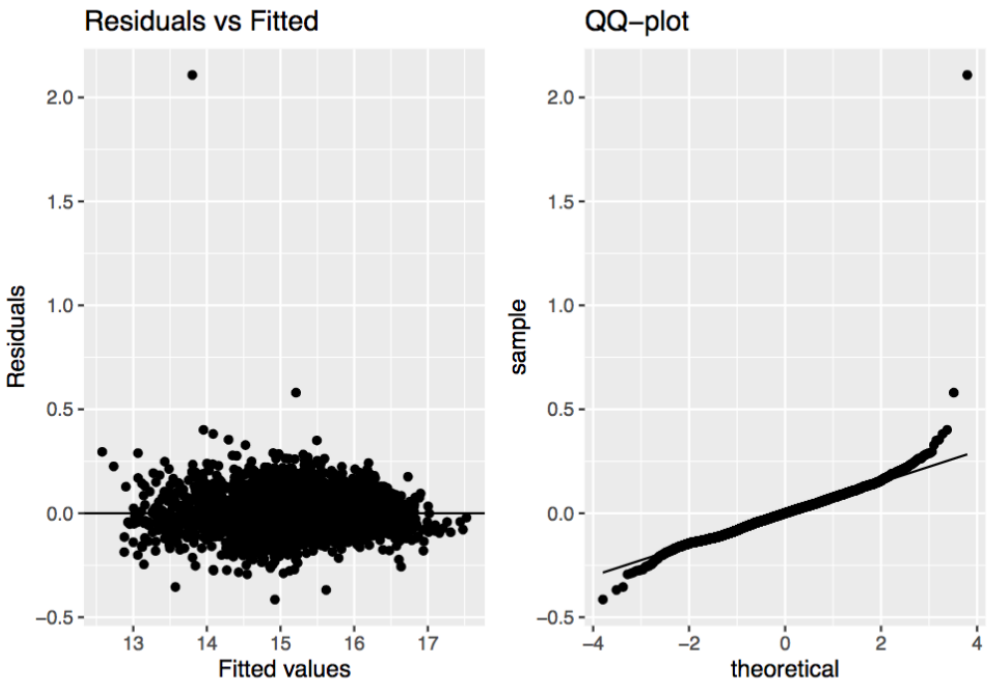
Genom att analysera *Tabell 2* fås informationen att samtliga variabler har en signifikant effekt på responsvariabeln. Utropspriset (*listPrice*) förklarar slutpriset (*soldPrice*) bäst med $R^2 = 0.98$ vilket är en väldigt hög förklaringsgrad.

## 4.2 Utropspriset inkluderas

Första steget i denna analys kommer vara att konstruera en multipel linjär regression. Modellen kommer tillämpas med samtliga variabler, detta på grund utav kunna avgöra ifall de olika kovariaten har samma påverkan på bostadspriserna i de olika städerna.

$$\begin{aligned}\log(\text{soldPrice}_i) = {} & \text{Area}_{\ddot{\text{O}}} + \text{Area}_S + \beta_1^{\ddot{\text{O}}} \cdot \log(\text{listPrice}) \cdot D_i^{\ddot{\text{O}}} + \beta_1^{\text{S}} \cdot \log(\text{listPrice}) \cdot D_i^{\text{S}} \\ & + \beta_2^{\ddot{\text{O}}} \cdot \text{floor} \cdot D_i^{\ddot{\text{O}}} + \beta_2^{\text{S}} \cdot \text{floor} \cdot D_i^{\text{S}} + \beta_3^{\ddot{\text{O}}} \cdot \text{livingArea} \cdot D_i^{\ddot{\text{O}}} \\ & + \beta_3^{\text{S}} \cdot \text{livingArea} \cdot D_i^{\text{S}} + \beta_4^{\ddot{\text{O}}} \cdot \text{rooms} \cdot D_i^{\ddot{\text{O}}} + \beta_4^{\text{S}} \cdot \text{rooms} \cdot D_i^{\text{S}} \\ & + \beta_5^{\ddot{\text{O}}} \cdot \text{constructionYear} \cdot D_i^{\ddot{\text{O}}} + \beta_5^{\text{S}} \cdot \text{constructionYear} \cdot D_i^{\text{S}} \\ & + \beta_6^{\ddot{\text{O}}} \cdot \text{rent} \cdot D_i^{\ddot{\text{O}}} + \beta_6^{\text{S}} \cdot \text{rent} \cdot D_i^{\text{S}} + \varepsilon_i\end{aligned} \tag{14}$$

Modellen som inkluderar utropspriset och som ges utav (14) är den modell som bäst uppfyller antagandena för linjär regression för denna situation. Detta kan analyseras genom studera *Figur 1* där den vänstra bilden visar en residualplott som indikerar ifall residualerna följer ett mönster och därav kan de ej tolkas som slumpmässiga runt 0. I detta fall ser man att punktmolnet är smalare ut till höger i residualplotten och detta skulle kunna bero på att färre observationer i det responsintervallet. Antagandet om konstant varians är helt, men förmodligen tillräckligt uppfyllt. Den högra bilden, visar en QQ-plot som indikerar ifall residualerna följer en normalfördelning eller inte, detta analyseras genom att kontrollera hur punkterna förhåller sig till linjen, om punkterna följer linjen indikerar detta att residualerna är normalfördelade. I detta framgår det att residualerna inte kan antas vara normalfördelade, detta på grund av att svansarna har för stor avvikelse. Det kan även analyseras att det finns outliers, det vill säga observationer som avviker väldigt mycket från de resterande observationerna. Dessa outliers kommer inte tas bort, utan kommer endast vara i åtanke under analysen. Denna modell ger en förklaringsgrad, $R^2 = 0.99$, vilket även är samma förklaringsgrad som generas när utropspriset exkluderas i modellen, vilket framkommer i kapitel 4.3. Om en modell generar samma förklaringsgrad med eller utan utropspriset som ett kovariat ger indikationer på att de resterande kovariaten kan förklara utropspriset, vilket i sin tur leder till att utropspriset bör exkluderas i modellen. Detta är för att undvika eventuell multikollinearitet.

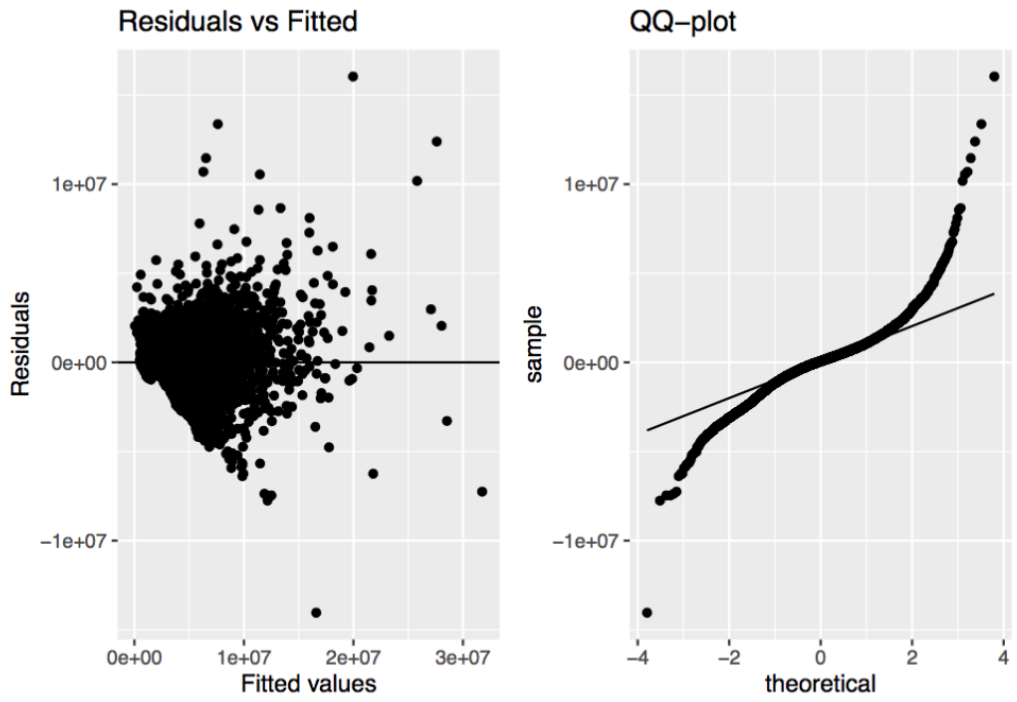



Figur 1: Kontroll av homoskedasticitet och normalitet när samtliga variabler används samt att responsvariabeln och utropspriset är logaritmerad

### 4.3 Multipel linjär regression, Utropspriset exkluderad

Utropspriset hade en hög förklaringsgrad vid enkel linjär regression vilket är på grund av mycket hög korrelation med slutpriset. Emellertid vill vi förstå vilka egenskaper hos lägenheten som påverkar slutpriset, och därför konstruerar vi nu:

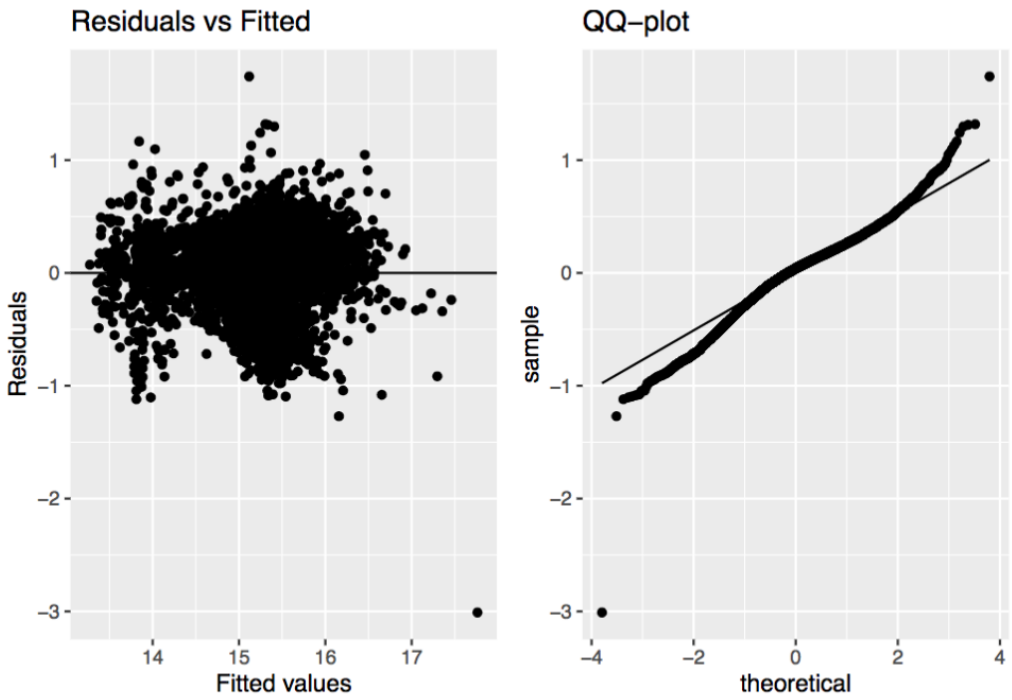
$$\begin{aligned}\text{soldPrice}_i = {} & \text{Area}_{\ddot{\text{O}}} + \text{Area}_S + \beta_1^{\ddot{\text{O}}} \cdot \text{floor} \cdot D_i^{\ddot{\text{O}}} + \beta_1^{\text{S}} \cdot \text{floor} \cdot D_i^{\text{S}} + \beta_2^{\ddot{\text{O}}} \cdot \text{livingArea} \cdot D_i^{\ddot{\text{O}}} \\ & + \beta_2^{\text{S}} \cdot \text{livingArea} \cdot D_i^{\text{S}} + \beta_3^{\ddot{\text{O}}} \cdot \text{rooms} \cdot D_i^{\ddot{\text{O}}} + \beta_3^{\text{S}} \cdot \text{rooms} \cdot D_i^{\text{S}} \\ & + \beta_4^{\ddot{\text{O}}} \cdot \text{constructionYear} \cdot D_i^{\ddot{\text{O}}} + \beta_4^{\text{S}} \cdot \text{constructionYear} \cdot D_i^{\text{S}} \\ & + \beta_5^{\ddot{\text{O}}} \cdot \text{rent} \cdot D_i^{\ddot{\text{O}}} + \beta_5^{\text{S}} \cdot \text{rent} \cdot D_i^{\text{S}} + \varepsilon_i\end{aligned} \tag{15}$$




Figur 2: Kontroll av homoskedasticitet och normalitet när samtliga variabler används

Enligt *Figur 2* kan det inte antas att residualerna följer en normalfördelning, det påvisar inte heller att homoskedasticitet existerar. En logtransformering på slutpriset och boarean utförs för att eventuellt förbättra situationen, det vill säga följande modell tillsätts:

$$\begin{aligned}\log(\text{soldPrice}_i) = {} & \text{Area}_{\ddot{\text{O}}} + \text{Area}_S + \beta_1^{\ddot{\text{O}}} \cdot \text{floor} \cdot D_i^{\ddot{\text{O}}} + \beta_1^{\text{S}} \cdot \text{floor} \cdot D_i^{\text{S}} + \beta_2^{\ddot{\text{O}}} \cdot \log(\text{livingArea}) \cdot D_i^{\ddot{\text{O}}} \\ & + \beta_2^{\text{S}} \cdot \log(\text{livingArea}) \cdot D_i^{\text{S}} + \beta_3^{\ddot{\text{O}}} \cdot \text{rooms} \cdot D_i^{\ddot{\text{O}}} + \beta_3^{\text{S}} \cdot \text{rooms} \cdot D_i^{\text{S}} \\ & + \beta_4^{\ddot{\text{O}}} \cdot \text{constructionYear} \cdot D_i^{\ddot{\text{O}}} + \beta_4^{\text{S}} \cdot \text{constructionYear} \cdot D_i^{\text{S}} \\ & + \beta_5^{\ddot{\text{O}}} \cdot \text{rent} \cdot D_i^{\ddot{\text{O}}} + \beta_5^{\text{S}} \cdot \text{rent} \cdot D_i^{\text{S}} + \varepsilon_i\end{aligned} \tag{16}$$




Figur 3: Kontroll av homoskedasticitet och normalitet när samtliga variabler används

Utifrån *Figur 3* visar det sig att problemet med heteroskedasticitet har förbättrats samt att residualerna följer en normalfördelning bättre, detta leder till att modellen som beskrivs i ekvation (16) kommer användas vid hypotesprövningar, denna modell generade även en förklaringsgrad $R^2 = 0.99$. Normalitet var ett utav de antaganden som behövdes uppfyllas för kunna utföra hypotestester, i detta fall kunde inte de antagandet uppfyllas. Detta kommer dock inte ge några problem vid hypotesprövning då endast det som är skrivit i kapitel 2.3.4 kan tillämpas och därav fortsätta utan vidare transformationer eller andra åtgärder. Det vill säga de kritiska värdena kommer hämtas från en standard normalfördelning samt en $\chi^2$-fördelning.

### 4.3.1 Hypotesprövning

Det försat hypotestestet som utförs är att kontrollera ifall någon variabel är skild från noll, det vill säga $H_0 : \beta_i = 0,\ \forall i$ vs $H_1 : \beta_i \neq 0,$ för något $i$. Detta utförs genom att använda den teststatistiska som beskrivs i ekvation (4).

$$F = \frac{R^2/m}{(1-R^2)/(N-m-1)} = \frac{0.99/(14)}{(1-0.99)/(6794-14-1)} = 47937.21$$

Detta ger att nollhypotesen kan förkastas med oavsett signifikansnivå $\alpha$ på grund av den oerhört stora teststatiskan, det vill säga det går med statistisk säkerhet sägas att utifrån samtliga variabler är åtminstone en variabel skild från noll.
Genom att använda den teststatistika (2) som beskrivs i kapitel 2.3.1 kan det kontrolleras ifall varje enskild förklaringsvariabel är signifikant skild från noll.

| | Estimat | Std.Error | t värde | Pr(N(0,1)>\|t\|) |
|---|---|---|---|---|
| AreaStockholm | 11.1080 | 0.0645 | 172.1967 | 0.0000 |
| AreaÖstersund | 11.7457 | 0.3534 | 33.2368 | 0.0000 |
| rent:AreaStockholm | -0.0002 | 0.0000 | -31.2269 | 0.0000 |
| rent:AreaÖstersund | -0.0002 | 0.0000 | -6.3326 | 0.0000 |
| AreaStockholm:floor | 0.0300 | 0.0016 | 18.4901 | 0.0000 |
| AreaÖstersund:floor | 0.0025 | 0.0122 | 0.2079 | 0.4177 |
| AreaStockholm:log(livingArea) | 1.0137 | 0.0208 | 48.7185 | 0.0000 |
| AreaÖstersund:log(livingArea) | 0.4858 | 0.1060 | 4.5815 | 0.0000 |
| AreaStockholm:rooms | 0.1020 | 0.0084 | 12.1421 | 0.0000 |
| AreaÖstersund:rooms | 0.2579 | 0.0347 | 7.4298 | 0.0000 |
| AreaStockholm:constructionYear | 0.0036 | 0.0001 | 34.4691 | 0.0000 |
| AreaÖstersund:constructionYear | 0.0020 | 0.0007 | 2.8226 | 0.0024 |

Tabell 3: Hypotestest $H_0 : \beta_i = 0$ vs $H_1 : \beta_i \neq 0$ för respektive förklaringsvariabel

Utifrån *Tabell 3* kan det konstateras att alla variabler förutom den variabel som när indikerar vilken våning lägenheten befinner sig på för Östersund är signifikanta. Denna modell generade en förklaringsgrad, $R^2 = 0.99$ vilket är samma förklaringsgrad som med utropspriset inkluderad. Det kommer inte genomföras någon sorts utav variabelselektion på grund av målet är att kontrollera ifall samtliga variabler har samma effekt i de två olika städerna. Utifrån detta kan det nu kontrolleras ifall variablerna har samma påverkan i de två olika städerna, detta görs genom att använda teststatistika som anges i ekvation (3). Följande hypotes ska därav prövas:

$$H_0 : \beta_i^{\text{Ö}} = \beta_i^{\text{S}} \ \forall i \quad H_1 : \beta_i^{\text{Ö}} \neq \beta_i^{\text{S}} \text{ för något } i$$

```
## Linear hypothesis test
##
## Hypothesis:
## AreaStockholm - AreaÖstersund = 0
## rent:AreaStockholm - rent:AreaÖstersund = 0
## AreaStockholm:floor - AreaÖstersund:floor = 0
## AreaStockholm:log(livingArea) - AreaÖstersund:log(livingArea) = 0
## AreaStockholm:rooms - AreaÖstersund:rooms = 0
## AreaStockholm:constructionYear - AreaÖstersund:constructionYear = 0
##
## Model 1: restricted model
## Model 2: log(soldPrice) ~ rent %in% Area + floor %in% Area + log(livingArea) %in%
##     Area + rooms %in% Area + constructionYear %in% Area + Area -
##     1
##
##   Res.Df     RSS Df Sum of Sq  Chisq Pr(>Chisq)
## 1   6788 1386.75
## 2   6782  628.95  6     757.8 8171.4  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Tabell 4: R (paket: "car", funktion: linearHypothesis") utskrift som utför hypotestestet $H_0 : \beta_i^{\text{Ö}} = \beta_i^{\text{S}} \; \forall i \quad H_1 : \beta_i^{\text{Ö}} \neq \beta_i^{\text{S}}$ för något $i$

Efter att ha analyserat *Tabell 4* kan det med statistisk säkerhet sägas att åtminstone en variabel har en annan inverkan i Östersund än i Stockholm, med andra ord nollhypotesen förkastas.

### 4.3.2 Modellval för Östersund

De olika faktorerna visade sig inte ha samma påverkan i de två städerna, detta leder till att det finns ett intresse konstruera en förklarande modell anpassad för Östersund. Genom att först tillämpa enkel linjär regression kan en bättre uppfattning kring hur variablerna förhåller sig till responsvariabeln, resultatet visas nedanför i *Tabell 5* där det kan konstateras att boarean har högst förklaringsgrad.

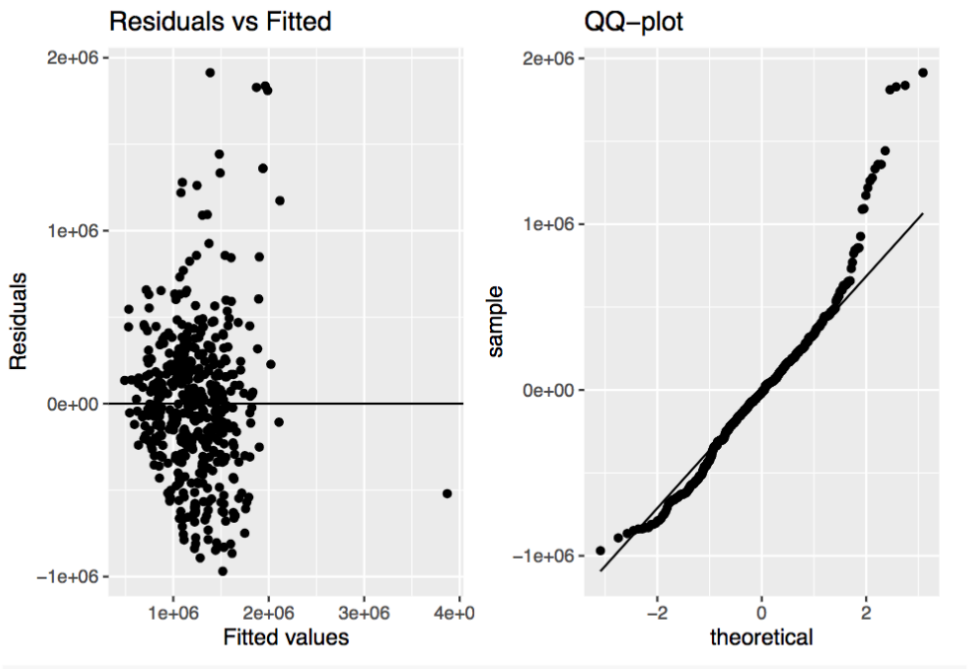
| Varibel | Estimat | P-värde | $R^2$ |
|---|---|---|---|
| listPrice | 1.0097 | <2e-16 | 0.9475 |
| floor | -2149.8353 | 0.92 | 0 |
| livingArea | 15569.7 | <2e-16 | 0.3063 |
| rooms | 327656.4766 | <2e-16 | 0.2762 |
| constructionYear | -2266.9926 | 0.0612 | 0.007 |
| rent | 183.7259 | 1.37e-14 | 0.1128 |

Tabell 5: Enkel linjär regression tillämpad på Östersunds observationer

Det är nu i intresse att tillämpa en multipel linjär regression på datasetet, detta görs genom att utgå från följande modell:

Modell 1

$$\text{soldPrice}_i = \alpha + \beta_1 \cdot \text{floor} + \beta_2 \cdot \text{livingArea} + \beta_3 \cdot \text{rooms} + \beta_4 \cdot \text{constructionYear} + \beta_5 \cdot \text{rent} + \varepsilon_i$$

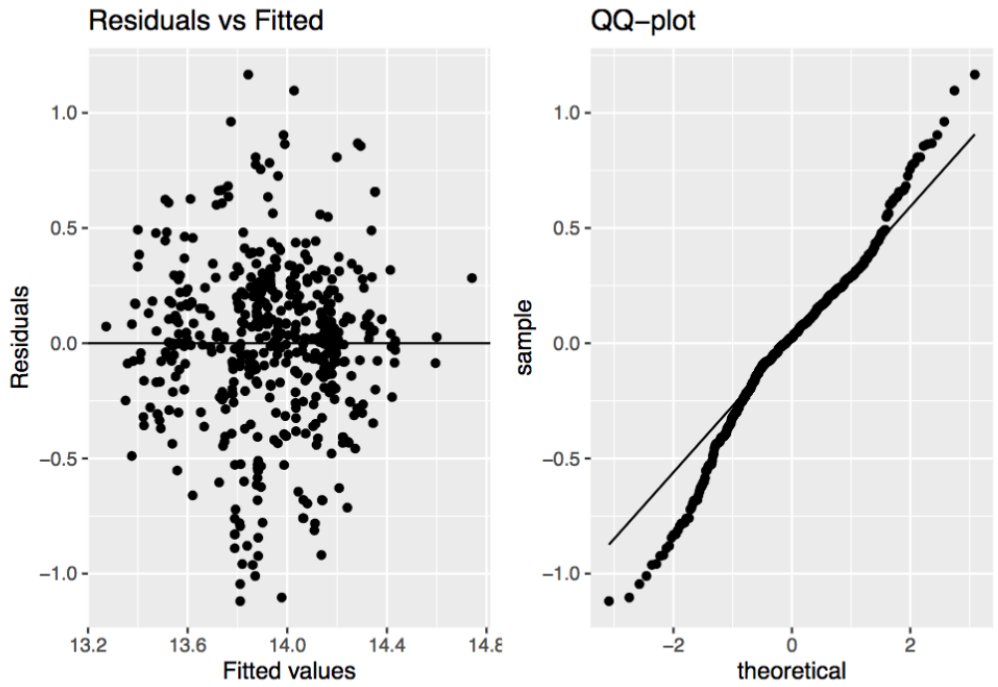



Figur 4: Kontroll av homoskedasticitet och normalitet för Modell 1

Vid kontroll utav homoskedasticitet och normalitet visar det sig ännu en gång vara ett problem, detta resultat fås genom att analysera *Figur 4*. Detta leder till att även här utförs en logtransformering, dock i detta fall logaritmeras slutpriset och boarean, vilket leder till följande modell:

Modell 2

$$\log(\text{soldPrice}_i) = \alpha + \beta_1 \cdot \text{floor} + \beta_2 \cdot \log(\text{livingArea}) + \beta_3 \cdot \text{rooms} + \beta_4 \cdot \text{constructionYear} + \beta_5 \cdot \text{rent} + \varepsilon_i$$




Figur 5: Kontroll av homoskedasticitet och normalitet för Modell 2

Genom att studera *Figur 5* kan slutsatsen att Modell 2 är bättre anpassad till datamaterialet och uppfyller de antaganden bättre än Modell 1. Detta leder till att Modell 2 kommer användas som grundmodell. Dock kan inte antagandet om normalfördelade residualer antas, detta på grund av att det är för stor avvikelse i svansarna vilket kan observeras i QQ-plotten i *Figur 5*, därav kommer det som har skrivits i kapitel 2.3.4 tillämpas vid hypotesprövning, vilket leder till att de kritiska värdena kommer hämtas från en standard normalfördelning samt en $\chi^2$-fördelning. Nästa steg nu blir att utföra stegvis variabelselektion, där metoderna Backard elimination, Forward selection och Stepwise regression. Samtliga metoder generade samma modell, vilket blev att variabeln som angav vilken våning lägenheten befanns sig på togs bort. Detta är även något som var väntat, eftersom enligt *Tabell 5* hade den variabeln en förklarningsgrad, $R^2 = 0$. Den modell som togs fram utav selektiv variabelselektion blev därav följande modell:

Modell 3

$$\log(\text{soldPrice}_i) = \alpha + \beta_1 \cdot \log(\text{livingArea}) + \beta_2 \cdot \text{rooms} + \beta_3 \cdot \text{constructionYear} + \beta_4 \cdot \text{rent} + \varepsilon_i$$

Genom att studera *Tabell 6* kan det uteslutas att multikollinearitet existerar i modellen. Vidare visualiseras samtliga variablers estimerade värden i *Tabell 7*, det kan avläsas att

samtliga variabler även är signifikanta. Förklaringsgraden och AIC presenteras i *Tabell 8* vilken endast generade 0.33 respektive 394.5. Dessa värden är visar att Modell 3 är sämre än Modell 1 där endast en logaritmerad version utav utropspriset användes.

| Variabel | VIF |
|---|---|
| log(livingArea) | 6 |
| rooms | 5 |
| constructionYear | 1 |
| rent | 3 |

Tabell 6: Kontroll av multikollinearitet för Modell 3

| | Estimat | Std.Error | t värde | Pr(N(0,1)>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 11.7484 | 0.4141 | 28.3742 | 0.0000 |
| log(livingArea) | 0.4867 | 0.1242 | 3.9181 | 0.0000 |
| rooms | 0.2573 | 0.0405 | 6.3448 | 0.0000 |
| constructionYear | 0.0020 | 0.0008 | 2.4013 | 0.0082 |
| rent | -0.0002 | 0.0000 | -5.3987 | 0.0000 |

Tabell 7: Parameterskattningar och P-värde för Modell 3

| $R^2$ | $R^2_{adj}$ | AIC |
|---|---|---|
| 0.3333 | 0.3279 | 394.4564 |

Tabell 8: Förklaringsgrad, Justerad förklaringsgrad och AIC för Modell 3

### 4.3.3 The Chow test of structural change

Genom att använda Modell 3 kan det nu testas ifall den framtagna modellens faktorer även generar samma påverkan vid tillfället modellen tillämpas i Stockholm. Detta kontrolleras genom att använda The Chow test of structural change, samma hypotes som i kapitel 4.2.3

ställs upp igen och samma statistika används.

$$F = \frac{(RSS_{restricted} - RSS_{unrestricted})/m + 1}{RSS_{unrestricted}/(N - 2(m+1))} = \frac{(\tilde{\boldsymbol{e}}^T\tilde{\boldsymbol{e}} - \boldsymbol{e}^T\boldsymbol{e})/(m+1)}{\boldsymbol{e}^T\boldsymbol{e}/(N-2(m+1))}$$
$$= \frac{(RSS - (RSS_{\ddot{\text{O}}} + RSS_S))/(m+1)}{(RSS_{\ddot{\text{O}}} + RSS_S)/((N_{\ddot{\text{O}}} + N_S) - 2\cdot(m+1))}$$

Utifrån ovanstående ekvation kan endast respektive värden stoppas in och därav fås följande:

$$F = \frac{(1465.07 - (62.847 + 597.81))/(4+1)}{(62.847 + 597.81)/((498 + 6296) - 2\cdot(4+1))} = 1652.03$$

Även denna gång kan nollhypotesen förkastas oavsett val på signifikantsnivå. Vilket leder att modellen konstruerad för Östersund tillämpad på Stockholm ger att åtminstone en faktor inte har samma inverkan i de två olika städerna.

# 5 Resultat

Utropspriset exkluderades som en förklaringsvariabel på grund av att det fanns indikationer på multikollinearitet, men även för att förstå vilka egenskaper hos en lägenhet som påverkar priset och om dessa påverkar på samma sätt i Östersund och Stockholm. Åtgärden blev att använda den modell som beskrivs i ekvation (16), det vill säga följande modell:

$$\begin{aligned}\log(\text{soldPrice}_i) = {} & \text{Area}_{\ddot{\text{O}}} + \text{Area}_S + \beta_1^{\ddot{\text{O}}}\cdot\text{floor}\cdot D_i^{\ddot{\text{O}}} + \beta_1^{\text{S}}\cdot\text{floor}\cdot D_i^{\text{S}} + \beta_2^{\ddot{\text{O}}}\cdot\log(\text{livingArea})\cdot D_i^{\ddot{\text{O}}} \\ & + \beta_2^{\text{S}}\cdot\log(\text{livingArea})\cdot D_i^{\text{S}} + \beta_3^{\ddot{\text{O}}}\cdot\text{rooms}\cdot D_i^{\ddot{\text{O}}} + \beta_3^{\text{S}}\cdot\text{rooms}\cdot D_i^{\text{S}} \\ & + \beta_4^{\ddot{\text{O}}}\cdot\text{constructionYear}\cdot D_i^{\ddot{\text{O}}} + \beta_4^{\text{S}}\cdot\text{constructionYear}\cdot D_i^{\text{S}} \\ & + \beta_5^{\ddot{\text{O}}}\cdot\text{rent}\cdot D_i^{\ddot{\text{O}}} + \beta_5^{\text{S}}\cdot\text{rent}\cdot D_i^{\text{S}} + \varepsilon_i.\end{aligned}$$

Utifrån detta kontrollerades ifall faktorerna hade samma påverkan i Östersund och Stockholm. Detta gjordes med hjälp av ett hypotestest och därav kunde det konstateras att i detta fall hade åtminstone en faktor olika inverkan i de två städerna. En modell som är anpassad för Östersund konstruerades även och efter selektiv variabelselektion togs Modell 3 fram som den optimala.

$$\log(\text{soldPrice}_i) = 11.75 + 0.49\cdot\log(\text{livingArea}) + 0.26\cdot\text{rooms} + 0.002\cdot\text{constructionYear} - 0.0002\cdot\text{rent} + \varepsilon_i$$

Denna modell kan även skrivas med en icke-logaritmerad responsvariabel

$$soldPrice_i = e^{11.75}\cdot e^{0.49\cdot\log(\text{livingArea})}\cdot e^{0.26\cdot\text{rooms}}\cdot e^{0.002\cdot\text{constructionYear}}\cdot e^{-0.0002\cdot\text{rent}}\cdot\epsilon'$$

Detta visar att boarea, antal rum och hur länge sedan lägenheten byggdes har en positiv inverkan på slutpriset, medan avgiften har en negativ inverkan. Denna modell generade en

förklaringsgrad på 0.33 och en justerad förklaringsgrad på 0.327 samt en *AIC* på 394. Det resulterade även att denna modell tillämpad på Stockholm inte generade att faktorerna hade samma påverkan i båda städerna. Denna hypotes kontrollerades med hjälp utav The Chow test of structural change. I *Tabell 9* visas de estimerade värden både för Östersund och Stockholm när Modell 3 är tillämpad. De största skillnader mellan de estimerade värdena är för boarean och antal rum. Det går även att i *Tabell 10* se att Modell 3 ger en högre förklaringsgrad och justerad förklaringsgrad i Stockholm än i Östersund. Dock generas en lägre AIC i Östersund. Tabellen visar att koefficienten för den logaritmerade boarean över 1 i Stockholm medan strax under 0.5 i Östersund. Slutsatsen blir därmed att boarean har över 100 % större påverkan för bostadspriserna i Stockholm jämfört med Östersund. På motsvarande sätt konstateras att antal rum har över 150 % större påverkan för bostadspriserna i Östersund jämfört med Stockholm. Att den logaritmerade boarean har en koefficient över 1 i Stockholm medan strax under 0.5 i Östersund innebär att boarean har över 100 % större påverkan i Stockholm. Det går även att konstatera att antal rum har över 150 % större påverkan i Östersund jämfört med i Stockholm. Variabeln som representerar lägenhetens ålder har också en skillnad mellan de olika städerna. I Stockholm påverkar variabeln bostadspriserna ungefär 60 % mer än i Östersund. De resterande estimerade parametrarna, interceptet och avgiften har ungefär samma koefficienter i båda städerna. Den stora skillnaden för de estimerade parametrarna skapar en markant skillnad i slutpriset i de olika städerna. Genom att till exempel estimera slutpriset på en stor etta med hjälp av den modell med en icke-logaritmerad responsvariabel ger följande resultat:

Östersund

$$e^{11.7484} \cdot e^{0.4867 \cdot \log(45)} \cdot e^{0.2573 \cdot 1} \cdot e^{0.002 \cdot 1} \cdot e^{-0.0002 \cdot 1500} = 774833$$

Stockholm

$$e^{11.16} \cdot e^{1.0306 \cdot \log(45)} \cdot e^{0.1004 \cdot 1} \cdot e^{0.0032 \cdot 1} \cdot e^{-0.0002 \cdot 1500} = 2918980$$

Detta leder till att det estimerade slutpriset i Stockholm är ungefär 278 % högre än i Östersund för en lägenhet med samma egenskaper. Detta visar tydligt att åtminstone en faktorer har olika inverkan på bostadspriset i de två städerna. Notera att detta gäller för det beskrivna exemplet.

| | Östersund | Stockholm |
|---|---|---|
| (Intercept) | 11.7484 | 11.1600 |
| log(livingArea) | 0.4867 | 1.0306 |
| rooms | 0.2573 | 0.1004 |
| constructionYear | 0.0020 | 0.0032 |
| rent | -0.0002 | -0.0002 |

Tabell 9: De estimerade värden vid tillämpning utav Modell 3 på respektive stad

| | Östersund | Stockholm |
|---|---|---|
| $R^2$ | 0.3333 | 0.6224 |
| $R^2_{adj}$ | 0.3279 | 0.6221 |
| AIC | 394.4564 | 3056.0028 |

Tabell 10: AIC, Förklaringsgrad och Justerad förklaringsgrad för respektive stad vid tillämpning utav Modell 3

# 6 Diskussion

Rapporten visade att utropspriset hade en väldigt hög förklaringsgrad vid tillämpning av enkel linjär regression. Detta resultat var inget som var oväntat på grund av att utropspriset oftast är en indikation på vad bostadens värde. Utifrån detta kunde misstankar finnas om en hög korrelation mellan utropspriset (förklaringsvariabel) och slutpriset (responsvariabel). Detta i samband med att förklaringsgraden blev densamma vid multipel linjär regression oavsett om utropspriset var inkluderad i modellen. Utifrån detta valde jag att exkludera utropspriset helt och hållet som en förklarande variabel. Jag ansåg att detta skulle ge en tydligare bild och lättare förstå om faktorerna påverkade bostadspriserna på samma sätt i de två olika städerna, men även eftersom utropspriset inte beskrev någon utav lägenhetens egenskaper. Det noterades att det fanns en stor prisskillnad på lägenheterna i Östersund jämfört med Stockholm. Denna prisskillnad skulle kunna ha tagits hänsyn till genom att standardisera priserna i de två städerna. Åtgärden skulle eventuellt kunna leda till nya resultat.
Jag valde att ta med de två förklarande variablerna som angav boarean och den som angav antal rum i modellen, som beskrivs i kapitel 4.3. Anledning till att ta med dessa variabler var på grund av att det blir vanligare med öppen planlösning, med detta menas att det inte är lika säkert att en lägenhet på 70 kvm har 3 rum. Det ska dock noteras att en lägenhet som är 25 kvm troligtvis inte har 3 rum. Att jag tog med två variabler som anger någon form utav indikation på lägenhetens storlek kan resultera i multikollinearitet. Ett sätt att motverka denna korrelation skulle kunna ha varit genom att transformera till exempel antal rum till antal rum per kvadratmeter. Denna transformation skulle även kunnat ha genomförts på avgiften eftersom avgiften oftast brukar tendera att bli större i samband med större lägenheter. Med andra ord hade variabeln istället blivit avgift per kvadratmeter.
Analysen delades upp i två steg, ett där samtliga variabler användes och ett där en modell anpassades till Östersund, sedan genomfördes variabelselektion. Att redan från första början utföra variabelselektion skulle kunna medföra att ett annat resultat uppnås. I resultatet visade det sig att boarean hade större påverkan på slutpriset i Stockholm än i Östersund medan antal rum hade större påverkan i Östersund än i Stockholm. Den skillnad som uppstod skulle kunna vara att det finns olika preferenser i de två städerna. I Stockholm

är kanske köparen endast ute efter en lägenhet med stor boarea medan i Östersund är det viktigt med antal rum i lägenheten.
Egentligen finns många fler faktorer som skulle kunna påverka slutpriset. Antalet budgivare skulle även kunna vara en faktor på grund av att Stockholm har en så mycket större population. Det skulle kunna leda till att fler spekulanter finns per lägenhet vilket i sin tur skulle genera fler budgivare och därmed leda till att priserna blir högre. Detta var inget som räknades med som en faktor, vilket dock bör tas hänsyn till om en likande studie ska genomföras igen.
Det resultat som togs fram, det vill säga att åtminstone en variabel hade en annan påverkan i de två städerna var ändå ganska väntat. Att bostadsmarknaden skulle vara lika i Östersund och Stockholm är ganska osannolikt på grund av det är två skilda miljöer. Vi vet att områden har en stark påverkan på priset i Stockholm, till exempel kommer en lägenhet på Östermalm att vara dyrare än en lägenhet i samma standard i till exempel Solna. Detta samband var något som inte existerade i samma utsträckning i Östersund, inte heller samma struktur på uppdelningen på områden. Utifrån detta valde jag att inte inkludera områden som en förklarande variabel. Dock kan detta skapa problem eftersom det finns så stora prisskillnader i Stockholm beroende på område, vilket skulle kunna resultera i att vissa parametervärden blir uppblåsta. En annat sätt att hantera detta problem hade varit att endast tagit med ett område från Stockholm vilket också skulle kunna leda till ett nytt resultat.
I kapitel 4.3 tas modellen som beskrivs i (16) fram och visade sig ha en förklaringsgrad på 0.99. Denna modell används för hela datasetet, samtidigt som en modell (Modell 3) anpassad till Östersund och tillämpad i Stockholm separat generar en mycket lägre förklaringsgrad. En orsak skulle kunna vara att antalet parametrar i respektive modell. I den modell som beskrivs i (16) finns det totalt 14 stycken parametrar medan i Modell 3 endast har 5 stycken parametrar. Det visade sig även att vid användning av en referensgrupp det vill säga ett intercept i (16) resulterade i en förklaringsgrad på 0.75. Orsaken till detta skulle kunna bero på hur funktionen som används vid konstruktion av de linjära modellerna i R (paket: *stats*, funktion: *lm*) hanterar en modell med intercept och utan. Dock skulle detta problem kunna åtgärdas genom att använda en referens i variabeln Area, vilket skulle genera ett intercept i modellen. Sedan skulle det endast kontrolleras ifall de resterande variablerna i modellen är lika. Denna lösning skulle även kunna ge en tydligare bild ifall en lägenhets egenskaper påverkar slutpriset på samma sätt i de två städerna.
Som tidigare nämnts finns det olika sätt gå tillväga för att konstruera modellerna. Ska det införas dummyvariabels och på vilket sätt ska de tillämpas, vilka förklarande variabler ska tas med från början och vilka ska inte? Utifrån den information som jag har kunnat ta del av, anser jag ändå att modelleringen blev lyckad.
Det ska förstås även poängteras att antagandet om normalfördelade residualer inte uppfylldes. Detta var dock inget problem utan istället tillämpades icke-normalfördelningar, som kunde utföras på grund av att det fanns många observationer. Ett annat alternativ för att eventuellt lösa detta problem hade varit att slumpmässigt plocka ut $X$ antal observationer

från respektive stad.

# Referenser


Andersson, P., Lindensjö, K. & Tyrcha, J. (2019). Notes in econometrics. *Matematiska institutionen. Stockholms universitet*.

*Booli API.* (2021). `https://www.booli.se/p/api/`. (Online; hämtat 2021-01-23)

Chow, G. C. (1960). Tests of equality between sets of coefficients in two linear regressions. *Econometrica: Journal of the Econometric Society*, 591–605.

Emanuelsson, R., Katinic, G. & Spector, E. (u. å.). Utvecklingen på bostadsmarknaden och dess bidrag till hushållens skulder.

Gordon, R. A. (2015). *Regression analysis for the social sciences*. Routledge.

Greene, W. H. (2002). Econometric analysis 5th edition. *International edition, New Jersey: Prentice Hall*.

Held, L. & Sabanés Bové, D. (2014). Applied statistical inference. *Springer, Berlin Heidelberg, doi*, *10*(978-3), 16.

R Core Team. (2020). R: A language and environment for statistical computing [handbok till mjukvara]. Vienna, Austria. Hämtad från `https://www.R-project.org/`

Sundberg, R. (2020). Lineära statistiska modeller. *Matematiska institution, Stockholms universitet*.

Vijitraphongsa, A. (2018). Regressionanalys av priser på mobiltelefoner. *Matematiska institutionen. Stockholms universitet*. (Kandidatuppsats i matematik statisik)