Stockholms universitet

Regressionsanalys av nyproducerade bostadsrätts priser i Solna-Sundbybergs kommun

Ninorta Malki

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2022:2
Matematisk statistik
Januari 2022

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Matematisk statistik
Stockholms universitet
Kandidatuppsats **2022:2**
http://www.math.su.se/matstat

Regressionsanalys av nyproducerade bostadsrätts priser i Solna-Sundbybergs kommun

Ninorta Malki*

Januari 2022

Sammanfattning

Denna studie har som huvudsyfte att analysera vilka variabler som påverkar slutpriset av en nyproducerad lägenhet i Solna-Sundbybergs kommun. I studien har vi tagit fram linjära modeller som transformeras med logaritmer för att få fram en så förklarande modell som möjligt. Vi undersöker även modellernas predikterande förmåga då modellen kan användas för att prediktera framtida slutpriser. Den slutgiltiga modellen som redovisas med mest optimala förklarande förmåga, var en modell med log transformation på responsvariabeln och en av de förklarande variablerna.

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige.
E-post: Ninortamalki@gmail.com. Handledare: Pieter Trapman.

Innehåll

1 Introduktion **3**

2 Teori **4**
2.1 Linjär Regression . . . 4
2.1.1 Enkel linjär regression . . . 4
2.1.2 Multipel linjär regression . . . 4
2.1.3 Minsta-kvadratmetoden . . . 4
2.1.4 Hypotestestning . . . 5
2.1.5 Korrelation . . . 5
2.1.6 Multikollinearitet . . . 6
2.2 Modellval . . . 6
2.2.1 Studentized Residuals . . . 6
2.2.2 Transformation . . . 6
2.2.3 Förklaringsgraden R^2 . . . 7
2.2.4 Residual Standard Error, F-statistic och P-value . . . 7
2.2.5 Stegvis variabelselektion . . . 8
2.2.6 Akaikes Informationskriterium . . . 8
2.3 Prediktion . . . 8
2.3.1 "Mean Squared Error of Prediction" . . . 8

3 Data **9**
3.1 Original Data Set . . . 9
3.2 Variabel Transformation . . . 10
3.3 Slutgiltiga Datasetet . . . 10

4 Analys och Resultat **12**
4.1 Enkel linjär regression . . . 12
4.2 Multikollinearitet . . . 13
4.2.1 Korrelation före transformation . . . 13
4.2.2 Korrelation efter variabeltransformation . . . 14
4.3 Modeller . . . 15
4.3.1 Definition och överblick av alla modeller efter transformation . . . 17
4.3.2 Definition av alla modeller efter stegvis variabelselektion . . . 20
4.4 Prediktion . . . 23
4.4.1 MSEP och RMSEP . . . 24

5 Diskussion **25**
5.1 Resultat . . . 25
5.2 Andra analyser . . . 26

6 Referenser **27**

7 Appendix **28**

1 Introduktion

Bostadsmarknaden är idag en viktig del av människors liv. Oberoende av vilken del i livet de befinner sig i. Den har exempelvis en stor betydelse för en nystartad familj, en ung vuxen som för första gången ska flytta ut från sina föräldrars hem eller kanske en familj som har minskat i antal på grund av utflyttade barn. I vilket fall som än gäller, är bostadsmarknaden, specifikt priserna en väldigt intressant faktor i många fall.

Det här arbetet kommer behandla, precis som rubriken lyder, bostadspriserna men mer specifikt för nyproducerade lägenheter i området Solna-Sundbybergs kommun. Detta kommer göras med hjälp av en regressionsanalys. Det vill säga, undersöka ändringarna i slutpriserna av lägenheterna i Solna-Sundbybergs kommun. Detta gör vi genom att hitta en linjär regressionsmodell som förklarar relationen mellan slutpriset och de förklarande variabler vi har tillhanda, exempelvis variabeln *region* som tillåter oss gå in lite mer djupgående i analysen och visar oss vilka exakta områden som påverkar de resultat vi kommer fram till.
Vi börjar med att kontrollera vårt data och vilka variabler som är relevanta för arbetets gång. Detta med hjälp av metoder som *VIF* som visar eventuell *multikollinearitet* mellan de förklarande variabler vi valt ut för regressionsmodellen. Vi går sedan vidare och undersöker vad som möjligtvis kan förbättra vår modell. Det gör vi genom transformation av modellen och vidare med hjälp av stegvis variabelselektion. Arbetets slutliga skede behandlar även modellernas prediktionsförmåga för att visa vilken modell som mest optimalt skattar ett slutpris. Vi kommer slutligen fram till att den modell som vi under arbetets gång valt att kalla "Modell 4 Step" är den vinnande modellen baserat på att den mest optimalt visar vilka förklarande variabler som har mest påverkan på slutpriset och är även den modellen som har en väldigt god prediktionsförmåga.

Denna studie är relevant för de personer som är intresserade av att köpa eller sälja nyproducerade lägenheter i området Solna-Sundbybergs kommun. Studien ger ett underlag till mäklare och privatpersoner som är intresserade av slutpriserna i det nämnda området. Uppsatsen kommer ge en överblick av vilka faktorer som påverkar ett slutpris på en lägenhet.

Studien är uppdelad i fem huvudkapitel inklusive Referenser samt ett Appendix med utskrivna tabeller för värden av undersökningarna. Första huvudkapitlet är en *Introduktion* som går igenom syftet med arbetet samt en lätt genomgång av tillvägagångssättet i analysen. Därefter kommer andra huvudkapitlet *Teori* som innehåller den grundläggande teorin för studien, som behövs för att kunna följa analysen. All data som hämtats från företaget Booli Search Technologies AB:s API presenteras i det tredje huvudkapitlet *Data*. Där redovisas även alla ändringar som sker på data under analysens gång som leder till vårt slutgiltiga dataset. I det fjärde huvudkapitlet *Analys och Resultat* introduceras även våra undersökningar och utskrifter som resulterar i vår vinnande modell. I det sista kapitlet *Diskussion* går vi igenom resultatet och sammanfattar hur stor effekt varje förklarande variabel har på slutpriset.
Frågor som läsaren fortsättningsvis kan tänka på genom arbetets gång är:

- Vilka variabler påverkar ett slut pris på en nyproducerad lägenhet? Och vilka variabler kan bortses från?
- Hur mycket påverkar variablerna enskilt modellerna?
- Kan den förklarande modell som väljs ut även vara den mest optimala predikterande modellen?

2 Teori

I denna del går vi igenom nödvändig matematisk teori som kommer hänvisas till genom studien.

2.1 Linjär Regression

Det finns ett antal antaganden man måste göra när man handskas med linjära regressionsmodeller, dessa anges nedan

- Linjäritet : Regressionsmodellen är linjär i dess parametrar
- Normalitet : Residualerna i regressionsmodellen antas vara normalfördelade
- Homoskedacitet : Variansen för feltermerna är konstant, oavsett värdet för de förklarande variablerna ($Var(\epsilon) = \sigma^2$)
- Ingen multikollinearitet : Ingen korrelation mellan de förklarande variablerna [**1, 2.1, 2.2**][**2, 3.6.4, 3.6.5**]

2.1.1 Enkel linjär regression

Inom statistik är linjär regression en lämplig metod som man kan använda för att visa om det finns ett statistiskt samband mellan en eller flera förklarande variabler x och en responsvariabel y. De förklarande variablerna består oftast av kategoriska eller kontinuerliga variabler. En kategorisk variabel innebär att en variabel har två eller flera nivåer (grupper), där observationerna tillhör. En kontinuerlig variabel kan anta vilket värde som helst, exempelvis avstånd till vatten.[**1, 2.1**]

Modellen för enkel linjär regression definieras som

$$Y_j = \beta_0 + \beta_{x_j} + \epsilon_j, \qquad j = 1, ..., n \tag{1}$$

Där $\epsilon \sim N(0, \sigma^2)$ och oberoende.[**1, 2.1**]

2.1.2 Multipel linjär regression

Den allmänna formen för multipel linjär regression är

$$\begin{cases} y_1 & = & \beta_0 & + & \beta_1 x_{11} & + & \cdots & + & \beta_k x_{k1} & + & \epsilon_1 \\ y_2 & = & \beta_0 & + & \beta_1 x_{12} & + & \cdots & + & \beta_k x_{k2} & + & \epsilon_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \vdots \\ y_j & = & \beta_0 & + & \beta_1 x_{1j} & + & \cdots & + & \beta_k x_{kj} & + & \epsilon_j \end{cases}, \qquad \epsilon \sim N(0, \sigma^2) \tag{2}$$

Som kan skrivas om på enklare form till

$$Y_j = \beta_0 + \beta_1 x_{1j} + ... + \beta_k x_{kj} + \epsilon_j \tag{3}$$

där $j = 1, 2, \ldots, n$ och n är antalet datapunkter, ϵ_j är feltermen, vilket betyder att ϵ_j är skillnaden mellan datapunkterna y_n och dess motsvarande punkt på regressionslinjen. Detta innebär att ju mindre skillnad, eller ju mindre ϵ_j är, desto bättre anpassad är modellen till regressionslinjen. [**1, 2.1**][**2, 2.1.1**]
β_0 är interceptet för regressionslinjen, $\beta_1, \ldots, \beta_k$ är ett mått på hur mycket de förklarande variablerna $x_{1j}, \ldots, x_{kj}$ påverkar responsvariabeln Y_j. $\beta_1, \ldots, \beta_k$ är okända variabler som vanligtvis skattas med hjälp av minsta-kvadratmetoden. [**2, 2.1.1**]

2.1.3 Minsta-kvadratmetoden

För att skatta de okända variablerna $\beta_1, \ldots, \beta_k$ använder vi oss av minsta-kvadratmetoden. Som har fördelen av att den kan tillämpas även då man inte vet fördelningstypen av data man undersöker. Vi kan alltså inte utgå från att data är normalfördelad eller har någon annan specifik fördelningstyp. [**1, 2.3.1**]
ϵ_j är skillnaden mellan datapunkterna Y_j och motsvarande punkt på regressionslinjen $\beta_0 + \beta_1 x_{1j} + ... + \beta_k x_{kj}$. Självklart betyder detta att ju mindre skillnaden är desto bättre passar data till regressionsmodellen.

Minsta-kvadratmetoden använder summan av de kvadratiska residualerna, som ett generellt mått på avståndet mellan de observerade datapunkterna och regressionslinjen. [**2, 2.1.1**] Matematiskt formulerat ges uttrycket

$$\sum_{j=1}^{n} \epsilon_j^2 = \sum_{j=1}^{n} (Y_j - (\beta_0 + \beta_1 x_{1j} + ... + \beta_k x_{kj}))^2 \tag{4}$$

Detta kan skrivas om på vektorform, vilket ger oss $\epsilon = Y - X\beta$. Med hjälp av minsta-kvadratmetoden får vi nu skattningen för β [**2, 2.2.3**]

$$\hat{\beta} = (X^T X)^{-1} X^T Y \tag{5}$$

2.1.4 Hypotestestning

Vi vill veta hur mycket de förklarande variablerna $x_{1j}, \ldots, x_{kj}$ påverkar responsvariabeln Y_j. Detta innebär att vi måste formulera ett hypotestest, ett så kallat t-test för β_i där $i = 1, \ldots, k$ och k är antalet datapunkter. Detta gör vi för att testa om de förklarande variablerna har någon statistiskt signifikant påverkan på responsvariabeln. [**2, 3.3.2**]

$$\begin{cases} H_0 : \beta_i = 0 \\ H_1 : \beta_i \neq 0 \end{cases} \tag{6}$$

Vi har tidigare definierat $\epsilon \sim N(0, \sigma^2)$, detta leder till att $\hat{\beta} = \beta + (XX)^{-1} X^T \epsilon$. Vi kan nu dra slutsatsen av att $\hat{\beta} \sim N(\beta, \sigma^2 (X^T X)^{-1})$. [**2, 2.2.2, 3.2**]
Detta leder vidare till teststatistikan

$$T = \frac{\hat{\beta}_i - \beta_i}{\sqrt{V(\hat{\beta}_i)}} = \frac{\hat{\beta}_i - \beta_i}{\hat{\sigma}\sqrt{c_{ii}}} \tag{7}$$

c_{ii} är i:te diagonal elementet av $(X^T X)^{-1}$, vi har nu slutligen teststatistikan [**2, 3.3.2**]

$$T = \frac{\hat{\beta}_i - \beta_i}{\hat{\sigma}\sqrt{((X^T X)^{-1})_{ii}}} \sim t(n-k) \tag{8}$$

Då σ^2 sällan är känd använder vi därför det estimerade värdet $\hat{\sigma} = \frac{1}{n} \sum_{j=1}^{n} \epsilon_j^2$.[**2, 3.2**] På grund av nollhypotesen har vi ett tvåsidigt test, vi förkastar nollhypotesen då $|T| \geq t_{\alpha/2}(n-k)$ på signifikansnivån α och frihetsgraderna $n - k$. Ett mindre P-värde, desto starkare incitament finns för att förkasta noll hypotesen, vanligtvis sätts gränsen till att $\alpha \leq 0.05$. Detta innebär att P-värdet är sannolikheten att vi får ett minst lika extremt utfall som stickprovet vi har när nollhypotesen är sann. [**8, s.3**]

2.1.5 Korrelation

Både Pearsons korrelationskoefficient och enkel linjär regression kan användas för att påvisa linjärt samband mellan två variabler. Pearsons korrelationskoefficient är ett mått på både styrkan och riktningen av det linjära sambandet. Korrelationskoefficienten r är ett nummer som varierar mellan -1 och 1, där $r = -1$ beskriver perfekt negativt linjärt samband, $r = 1$ beskriver perfekt positivt linjärt samband och $r = 0$ betyder att inget linjärt samband finns.
Pearsons korrelationskoefficient beräknas genom

$$r = \frac{c_{xy}}{s_x s_y} = \frac{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1} \frac{\sum_{i=1}^{n} (y_i - \overline{y})^2}{n-1}}} \tag{9}$$

Där kovariansen definieras av $c_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$, stickprovets standardavvikelse för x definieras av $s_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{n-1}}$ och stickprovets standardavvikelse för y definieras av $s_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i-\overline{y})^2}{n-1}}$. Därpå $\overline{x}$ är medelvärdet av alla x värden, $\overline{y}$ är medelvärdet av alla y värden och n är antalet observationer. **[1, 3.3.4]**

2.1.6 Multikollinearitet

När man har ett stort antal förklarande variabler, är det vanligt att det finns parvisa korrelationer mellan dem. Detta kallas multikollinearitet. Idealt vill man att alla förklarande variabler endast har en korrelation till responsvariabeln, inte till varandra.
I regressionsanalys kan man ibland förvänta sig att vissa x-variabler ska vara statistiskt signifikanta. När de inte är det, kan det vara så att det beror på kollinearitetsproblem. Enkelt förklarat betyder det exempelvis, om två x-variabler är kollineära, så fungerar modellen ungefär lika bra med vilken av de två förklarande variablerna som helst. Skulle man ta in båda variablerna i modellen så kan vi se att båda då blir icke-signifikanta.
Ett användbart mått som ger indikation på att vi har multikollinearitet i modellen, är det statistiska måttet VIF, som står för "Variance Inflation Factor" (på svenska variansinflationsfaktor). VIF beräknas som

$$VIF = \frac{1}{1 - R_j^2} \tag{10}$$

Där R_j^2 (förklaringsgraden) som beskrivs nedanför i kapitel 2.2.3, anger hur mycket en förklaringsvariabel x_j beskrivs av variationen av de andra förklarande variablerna i en linjär kombination. Ett högre värde för $0 \leq R^2 \leq 1$ innebär ett högt VIF värde. Då VIF-värdet som vi tidigare nämnt förenklat uttrycker hur mycket varians minsta-kvadratskattningen $\hat{\beta}$ har i kombination av de andra regressionstermerna. Ett lämpligt högsta VIF-värde är vanligtvis någonstans mellan 5 och 10. **[1, 3.2.4]**

2.2 Modellval

Inom teorin för multipel linjär regression ingår valet av modell för det data set man granskar. Teorin för de metoderna vi använder i denna studie förklaras nedan.

2.2.1 Studentized Residuals

Ibland kan det vara svårt att identifiera en "outlier", då den potentiella "outliern" påverkar regressionsmodellen så pass mycket att den estimerade regressionsmodellen dras till "outliern". För att kunna handskas med detta problem kan man använda sig av *studentized residuals*, som fungerar genom att man handskas med en observation i taget. Formeln för *studentized residuals* beskrivs av:

$$t_i = \frac{d_i}{s(d_i)} \tag{11}$$

Där d_i är en raderad residual delad med sin skattade standardavvikelse. Allmänt funkar *studentized residuals* på sånt sätt att om observationen är större än 3 i absolut värde, kallas den för en "outlier" och man väljer då att behandla den.**[6]**

2.2.2 Transformation

Ibland kan modelltransformationer behövas för att förhindra överträdelser av modellantaganden såsom icke konstant felvarians (homoskedacitet). Det kan dessutom vara så att rådata man har, inte visar ett linjärt samband mellan responsvariabeln och de förklarande variablerna. I vissa fall krävs en transformation av antingen responsvariabeln eller de förklarande variablerna, men i vissa fall kan det behövas en transformation av både responsvariabeln och förklarande variablerna.
En vanlig metod som kan användas är log transformation. Om vi antar att vi har den multiplikativa modellen

$$Y = \alpha e^{\beta x}\epsilon \tag{12}$$

Kan vi genom log transformation, logaritmera modellen till den additiva modellen

$$\log Y = \log \alpha + \beta x + \log \epsilon \tag{13}$$

Där $\log \epsilon \sim N(0, \sigma^2)$. Den multiplikativa modellen kunde på grund av log transformationen omvandlas till en modell med ett linjärt samband mellan responsvariabeln och den förklarande variabeln. [**1, 3.3.2**]

2.2.3 Förklaringsgraden R^2

Förklaringsgraden R^2 är den procenthalt av den totala variationen i responsvariabeln som kan "förklaras" av de förklarande variablerna. Vilket innebär att förklaringsgraden berättar hur bra data setet passar till modellen ("goodness of fit"). Förhållandet beskrivs som

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} \tag{14}$$

Som kallas determinationskoefficienten. Där ESS = "Explained sum of squares", TSS = "Total sum of squares", RSS = "Residual sum of squares" och $TSS = ESS + RSS$, som har formeln

$$\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 + \sum_{i=1}^{n}(Y_i - \hat{Y})^2 \tag{15}$$

Där $\hat{Y} = \beta_0 + \beta_1 x_{1j} + ... + \beta_k x_{kj}$ är det predikterade värdet och $\overline{Y}$ är medelvärdet. En viktig egenskap R^2 har, är att den alltid ökar ju mer förklarande variabler man tillför till en modell. Det gäller även om den specifika förklarande variabeln inte är korrelerad till responsvariabeln. Av den orsaken, när man jämför olika modeller med samma responsvariabel, så måste man vara försiktig med att välja den modellen med högst R^2. För att se om det lönar sig att tillföra en ytterligare förklarande variabel kan det därför vara bra att använda R^2_{adj}, som tar hänsyn till antalet förklarande variabler som används i den linjära regressionen. R^2_{adj} är ett mått på hur mycket variansreduktion man uppnått med den valda modellen.[**2, 2.2.5**] Detta förhållande kan beskrivas som

$$R^2_{adj} = 1 - \frac{(1 - R^2)(N - 1)}{(N - m - 1)} \tag{16}$$

Där N = totala stickprovstorleken och m = antalet förklarande variabler. [**1, 3.2.2**]

2.2.4 Residual Standard Error, F-statistic och P-value

Residual standard error (RSE) är en metod som man kan använda för att mäta hur bra en linjär regression passar till data man har. Andra mått som kan användas för att bedöma om data är bra anpassad är R^2 eller R^2_{adj}, som tidigare nämnts. Residual standard error beräknas genom formeln

$$RSE = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{df}} \tag{17}$$

Där df är antalet observationer minus antalet parametrar man estimerar. Ju lägre värde RSE har desto mer anpassat är data setet till regressionsmodellen. [**5, 6.1**]
F-statistic är en indikation som beskriver om det finns ett samband mellan responsvariabeln och de förklarande variablerna. Ju högre siffra F-statistic är desto bättre är förhållandet mellan responsvariabeln och de förklarande variablerna. F-statistic har formeln [**2, 3.2.2**]

$$F = \frac{\frac{ESS}{k-1}}{\frac{RSS}{n-k}} \in F(k-1, n-k) \tag{18}$$

Där $F(k-1, n-k)$ är F-fördelningen med $k-1$ som är ESS frihetsgrader och där $n-k$ är RSS frihetsgrader. P-värdet beskriver tillsammans med F-statistic om modellen är signifikant, formeln ges av [**2, 3.3.2**]

$$P\left(F(k-1, n-k) \leq \frac{\frac{ESS}{k-1}}{\frac{RSS}{n-k}}\right) \tag{19}$$

2.2.5 Stegvis variabelselektion

När man handskas med multipel linjär regression, där man har identifierat en responsvariabel Y_j, som beskrivs av flera förklarande variabler x_j, är frågan hur många och vilka förklarande variabler man har med i sin regressionsmodell. Det finns ett flertal stegvisa metoder man kan använda, för att slutligen bestämma den slutgiltiga modellen. Valet av stoppkriterium som exempelvis AIC som förklaras nedanför och R^2 som har betydelse för det slutgiltiga resultatet. Metoderna som är mest vanliga att använda är "bakåt"-metoden ("Backward Elimination"), "framåt"-metoden ("Forward Selection") och stegvis variabelselektion.

- "bakåt"-metoden: Startar med att inkludera alla variabler i regressionen, för att stegvis sedan eliminera en variabel i taget. Fram till att regressionen uppnår ett förvalt stoppkriterium.
- "framåt"-metoden: Startar med en tom modell, för att sedan utvidgas med en variabel i taget. Metoden inkluderar den mest signifikanta av variablerna som inte inkluderats i modellen än.
- Stegvis variabelselektion: Metoden är en förbättrad version av "bakåt"/"framåt"- metoderna. Förbättringen tillåter borttagning av en variabel som hamnat i modellen men i ett senare steg inte längre är signifikant. [**1, 3.2.3**]

2.2.6 Akaikes Informationskriterium

Akaikes Informationskriterium (AIC) mäter hur bra en modell passar till data setet man har, utan att addera för mycket förklarande variabler ("overfitting"). Den modellen med lägst AIC poäng, är den modell som är mest lämplig att använda. AIC beräknas genom formeln

$$AIC = -2l\left(\hat{\beta}_{ML}\right) + 2n \tag{20}$$

Där $l\left(\hat{\beta}_{ML}\right)$ = maximala värdet av log-likelihood-funktionen för modellen och n = antalet parametrar. [**3, 7.1.1**] Programspråket R använder AIC poäng för att välja ut den mest lämpade modellen när den genomför stegvis variabelselektion

2.3 Prediktion

När regressionsanalysen genomförts och man fått fram β skattningarna för att senare komma fram till en bra modell, är nästa steg att göra en prediktion. Detta innebär enkelt sagt att man använder den skattade regressionsmodellen man kommer fram till, för att predikera framtida värden. Vilken modell som är mest anpassad till data kan väljas med hjälp av $MSEP$.

2.3.1 "Mean Squared Error of Prediction"

"Mean squared error of Prediction" är prediktionsmåttet av storleken för det kvadratiska medelvärdet av feltermerna. Tillfälligtvis tas observation j ur data. Regressionen görs utan observation j, med det skattade prediktionsvärdet $\hat{y}_{-j}$. Detta för att i sin tur predikera värdet y_j. Prediktionsfelet beräknas som $y_j - \hat{y}_{-j}$. Prediktionsmåttet ges av

$$MSEP = \frac{1}{N}\sum_{1}^{N}\left(y_j - \hat{y}_{-j}\right)^2 \tag{21}$$

Ett ekvivalent prediktionsmått är $RMSEP$ ("Root Mean Squared Error of Prediktion"), som är kvadratroten ur MSEP. [**1, 3.2.2**]

3 Data

Data som används i detta examensarbete har samlats in från företaget Booli Search Technologies AB [**4**], som är en sökmotor baserad på den svenska bostadsmarknaden. Data laddades ner från webbplatsens egna API. Totalt samlades 231 dataobservationer in för områdena Solna-Sundbyberg. Datafilen som erhölls genom Boolis API var i JSON-format som omvandlades till CSV. Genom arbetet används programspråket R.

3.1 Original Data Set

Den ursprungliga datafilen som hämtades från Boolis API, innehöll dessa variabler:

Region : Områden i Solna-Sundbyberg

Stadsdel : Information om lägenheten befinner sig i Solna eller Sundbyberg

Avstånd Till Vatten : Avstånd till närmsta sjö eller hav, mätt i km

List Pris : Noteringspris för lägenheten

Avgift : Avgift per månad

Våning : Våning för den specifika lägenheten

Kvadratmeter : Storleken på lägenheten mätt i kvadratmeter

Rum : Antal rum inklusive vardagsrum

Byggår : Byggåret för lägenheten

Sälj Datum : Säljdatum för lägenheten

Sälj Pris : Slutpriset för lägenheten

Longitud/Latitud : Geografisk position

Mäklare : Vilken mäklarfirma som sålde lägenheten

3.2 Variabel Transformation

För vidare behandling av data finns ibland behovet av att ändra, ta bort eller lägga till variabler. Detta för att förhindra problem som kan uppkomma i analysen. Under arbetets gång kommer dessutom en prediktion göras på nyproducerade lägenhetspriser i området Solna-Sundbyberg, vilket innebär att variablerna *Longitud/Latitud* och *Mäklare* inte är relevanta för modellen som ska predikteras, på grund av frågeställningen vi har. Detta innebär att vi kan ta bort de variablerna från vårt data. För att behålla variabler som exempelvis har multikollinearitet, men vars information har en betydelse för modellens styrka, var vi tvungna att omdefiniera eller ta bort vissa variabler såsom *List Pris*, *Kvadratmeter*, *Avgift*, *Våning* och *Säljdatum*. Den variabeln vi tog bort var *List Pris* då den har en väldigt stor påverkan på modellerna och prediktionen. *List Pris* är dessutom väldigt korrelaterad till variabeln *Sälj Pris* som är vår responsvariabel i denna studie. Vidare definierades de nya variablerna till *Kvadratmeter/Rum*, *Avgift/Kvadratmeter* och till den kategoriska variabeln *Säsong* som delas upp i kategorierna Vår, Sommar, Höst och Vinter, vilket baserats på variabeln *Säljdatum*. Vi valde även att omgruppera variabeln *Våning* till våning $1-7$, $8-14$, 15, $16-18$, $19-20$ och våning "Okänd" då det var för lite data för varje enskild våning, vilket ger oss mer relevant information till den fortsatta studien.

3.3 Slutgiltiga Datasetet

Här presenteras det slutgiltiga datasetet som vidare kommer behandlas i vår analys.

Sälj Pris : Kontinuerlig responsvariabel som varierar mellan 1795000 och 6350000 (kr)

Region : Kategorisk variabel för områden i Solna-Sundbyberg

Stadsdel : Kategorisk variabel med information om lägenheten befinner sig i Solna/Sundbyberg

Avstånd Till Vatten : Kontinuerlig förklarande variabel, varierar mellan 1222 och 10000 (m)

Våning : Kategorisk variabel för de olika våningarna

Rum : Kategorisk variabel för antalet rum

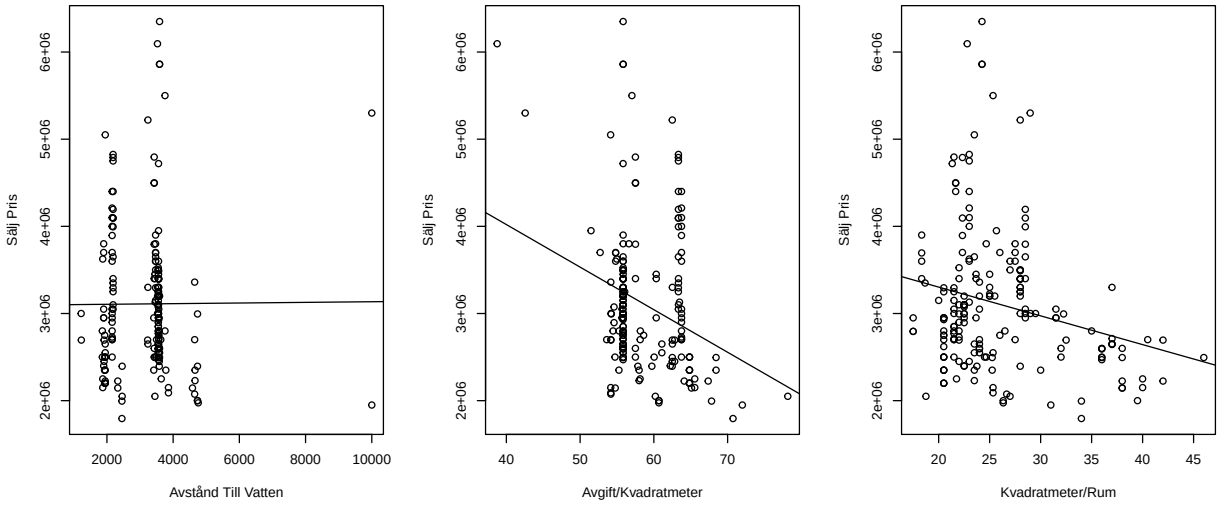
Byggår : Kategorisk variabel för byggåren 2018 till 2019

Säsong : Kategorisk variabel för de olika årstiderna

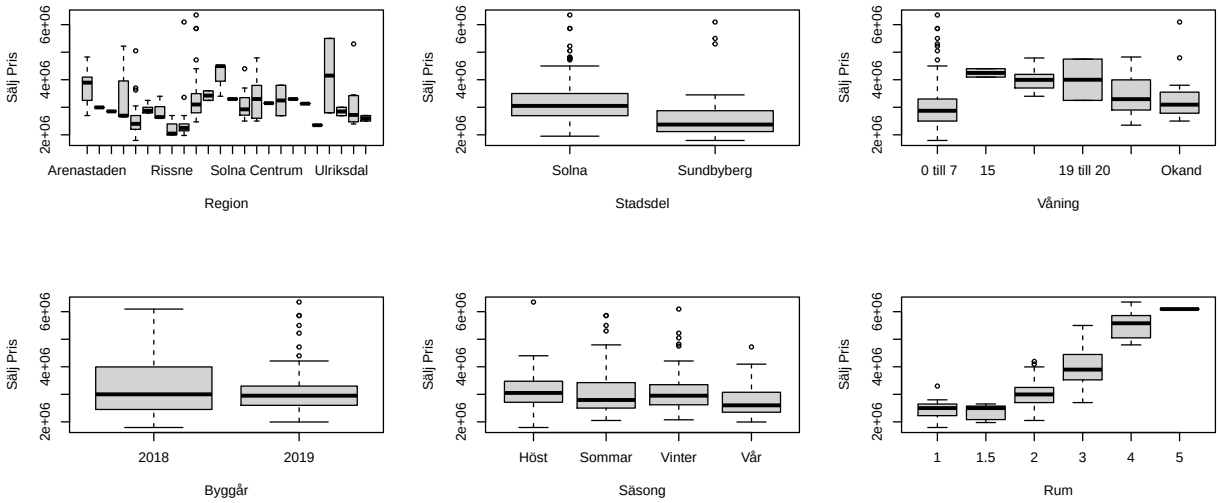
Kvadratmeter/Rum : Kontinuerlig förklarande variabel, varierar mellan 17.5 och 46 (m^2/rum)

Avgift/Kvadratmeter : Kontinuerlig förklarande variabel, varierar mellan 38.75 och 78.16 (kr/m^2)

Vi plottar responsvariabeln *Sälj Pris* mot varje enskild kontinuerlig variabel i Figur 1 och varje enskild kategorisk variabel i Figur 2, i ett rutnät av "scatterplots" och "boxplots" för enklare överblick.



Figur 1: Scatterplots av de kontinuerliga variablerna



Figur 2: Boxplots av de kategoriska variablerna

4 Analys och Resultat

Detta kapitel handlar om att undersöka ifall alla modellantaganden vi gör, gäller för de antagandena som råder för linjära regressioner, som nämns i kapitel 2.1 samt att presentera de resultat som även bekräknades i programspråket R. Kapitlet avslutas med en prediktion.

4.1 Enkel linjär regression

För att enklare se om det finns ett linjärt samband mellan varje enskild förklarande variabel presenteras alla värden, mer överskådligt i en tabell.

Tabell 1: Alla förklarande variabler mot responsvariabeln

Förklarande Variabler	Multiple R^2	P-värde
Region	0.2587	$7.036e-06$
Stadsdel	0.06426	$9.796e-05$
Avstånd till vatten	0.0001401	0.858
Våning	0.1035	0.0001566
Rum	0.667	$< 2.2e-16$
Byggår	0.004208	0.3263
Säsong	0.01771	0.2544
Kvadratmeter/Rum	0.04829	0.0007709
Avgift/Kvadratmeter	0.06806	$5.989e-05$

Vi kan se från Tabell 1, att variablerna *Avstånd till vatten*, *Byggår* och *Säsong*, inte är signifikanta vid enkel linjär regression. Den variabeln som har starkast inflytande på responsvariabeln *Sälj Pris* är variabeln *Rum*, med ett R^2-värde av 0.667. Näst på tur är variabeln *Region* med ett R^2-värde på 0.2587. Då det är endast enkel linjär regression så är det inte relevant att visa R_{adj} värdena.

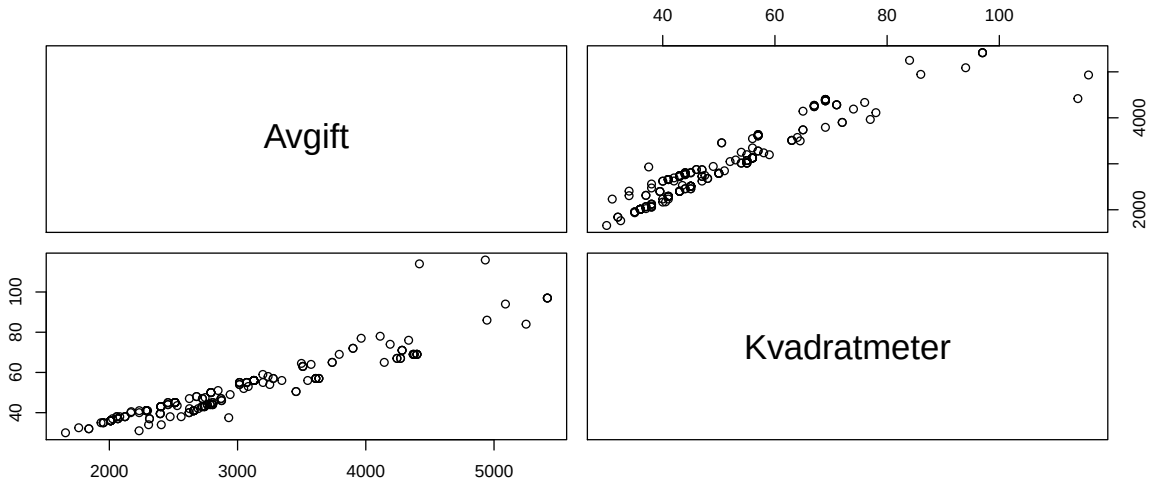
4.2 Multikollinearitet

4.2.1 Korrelation före transformation

Som vi tidigare nämnt i kapitel 3.2 så ändrades variablerna *Kvadratmeter* och *Avgift*. Detta eftersom enligt Pearsons korrelation, visade de förklarande variablerna *Avgift* och *Kvadratmeter* stark korrelation till varandra, där korrelationskoefficienten $r = 0.94$. Vi använde oss av den kategoriska variabeln *Rum* och skapade de nya förklarande variablerna *Kvadratmeter/Rum* och *Avgift/Kvadratmeter*. Transformationen behövs som det tidigare nämnts, då informationen som finns i variablerna inte ska gå förlorat, istället för att ta bort en av de variablerna. Tabell 3 visar att VIF värdet för variablerna *Kvadratmeter* och *Avgift* dessutom är över 5. Korrelationstabellen och korrelationsmatrisen presenteras i Tabell 2 och Figur 3 nedanför.

Tabell 2: Korrelationstabell mellan varje kontinuerlig variabel och responsvariabeln

	Avstånd Till Vatten	Avgift	Kvadratmeter	Sälj Pris
Avstånd Till Vatten	1			
Avgift	-0.14	1		
Kvadratmeter	0.03	0.94	1	
Sälj Pris	0	0.77	0.83	1



Figur 3: Korrelationsmatris mellan de mest korrelerade kontinuerliga variablerna

Tabell 3: VIF-värden för de kontinuerliga variablerna innan transformation

Avstånd Till Vatten	Kvadratmeter	Avgift
1.331182	11.2011	11.51766

4.2.2 Korrelation efter variabeltransformation

Nedanför presenteras resultaten i Tabell 4 och Tabell 5 genom en korrelationstabell och VIF-tabell efter att variablerna som tidigare hade hög korrelation transformerats. Vi kan se att de nya transformerade variablerna inte längre har någon stark korrelation. Dessutom ser vi att VIF-värdena nu är under 5.

Tabell 4: Korrelationstabell mellan de kontinuerliga variablerna efter transformation mot responsvariabeln

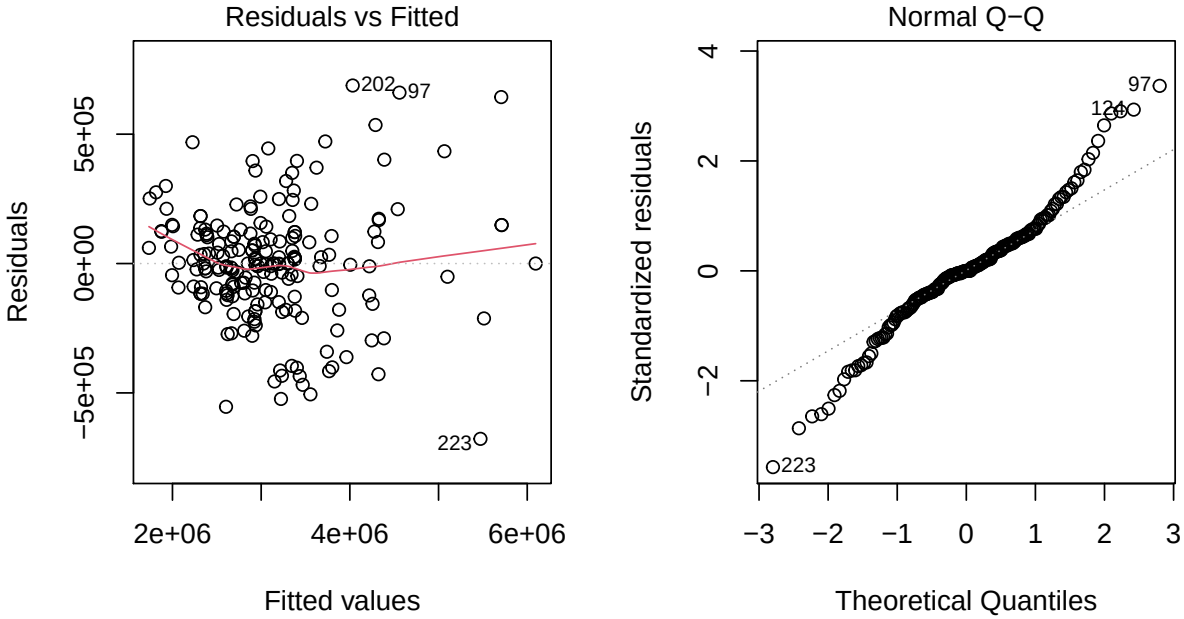
	Avstånd Till Vatten	Kvm/Rum	Avgift/Kvm	Sälj Pris
Avstånd Till Vatten	1			
Kvm/Rum	0.1	1		
Avgift/Kvm	-0.41	0	1	
Sälj Pris	0	-0.23	-0.28	1

Tabell 5: VIF-värden för de kontinuerliga variablerna efter transformation

Avstånd Till Vatten	Kvm/Rum	Avgift/Kvm
1.238755	1.016685	1.22663

4.3 Modeller

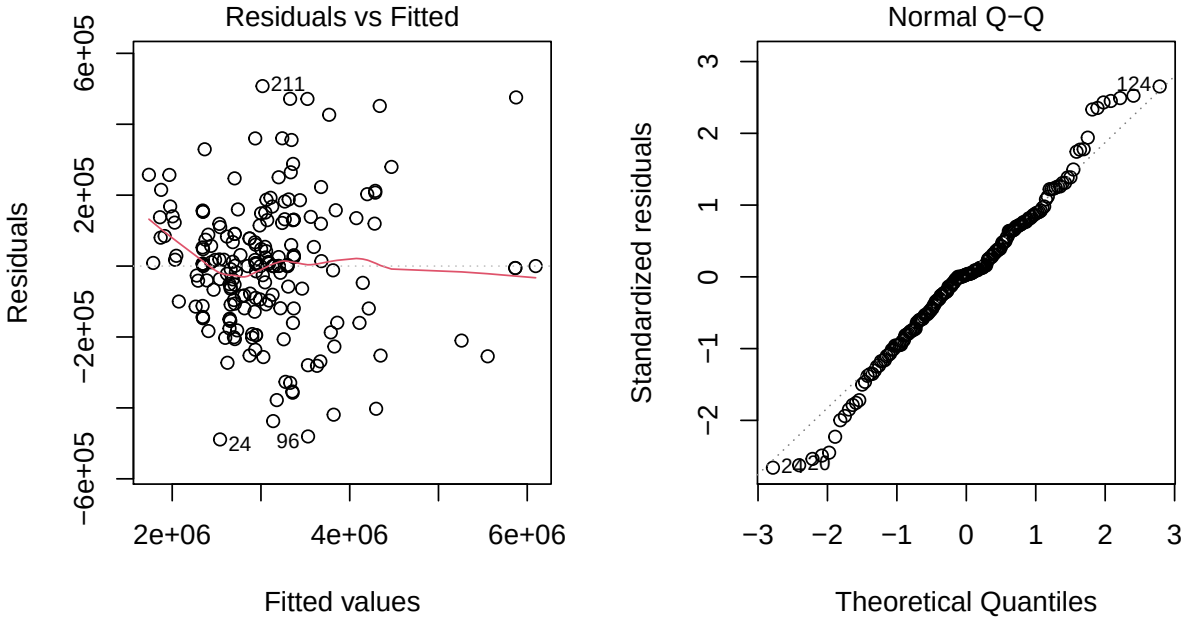
Eftersom VIF-värdena var lägre än 5 för alla variabler, väljer vi att behålla alla förklarande variabler i vad som nu blir den första modellen. Som vi kan se i Figur 4, kan det finnas eventuella outliers. Modellen har vi därför valt att namnge till "Modell 1 inkl outliers".



Figur 4: Modell 1 inkl outliers

Modell 1 inkl outliers :

$$\begin{aligned} SaljPris_j = \beta_0 + \beta_1 Region + \beta_2 Stadsdel + \beta_3 AvstandTillVatten + \beta_4 Vaning + \\ \beta_5 Rum + \beta_6 ByggAr + \beta_7 Sasong + \beta_8 KvadratmeterPerRum + \\ \beta_9 AvgiftPerKvadratmeter + \epsilon_j \end{aligned} \tag{22}$$



Figur 5: Modell 1 exkl outliers

Modell 1 exkl outliers :

$$\begin{aligned} SaljPris_j = \beta_0 + \beta_1 Region + \beta_2 Stadsdel + \beta_3 AvstandTillVatten + \beta_4 Vaning + \\ \beta_5 Rum + \beta_6 ByggAr + \beta_7 Sasong + \beta_8 KvadratmeterPerRum + \\ \beta_9 AvgiftPerKvadratmeter + \epsilon_j \end{aligned} \tag{23}$$

Vi kunde se i Figur 4 att modellen kanske har eventuella outliers. Därför var det lämpligt att använda ett outlier test som tar ut observationen med störst studentized residual värde. Vi valde att sätta gränsen för studentized residual värde på en outlier till ± 3 [**6**][**7**]. Vid en överblick över Tabell 6 nedanför, kan vi se att "Modell 1 exkl outliers" blir en lite starkare modell än "Modell 1 inkl outliers". Detta efter totalt sex outliers togs bort. Ett skevhetstest visade en förändring från -0.3811 till 0.0715. Vi kan däremot fortfarande se heteroskedacitet i Residual vs Fitted plotten, samt en konformig form både innan och efter borttagning av outliers. Detta innebär att en transformering av modellen kan vara relevant. Vidare väljer vi att gå vidare med "Modell 1 exkl outliers" och fortsättningsvis definiera den som "Modell 1". Vi presenterar även resterande modeller i stycke 4.3.1.

4.3.1 Definition och överblick av alla modeller efter transformation

Modell 1:

$$\begin{aligned} SaljPris_j = \beta_0 + \beta_1 Region + \beta_2 Stadsdel + \beta_3 AvstandTillVatten + \beta_4 Vaning + \\ \beta_5 Rum + \beta_6 ByggAr + \beta_7 Sasong + \beta_8 KvadratmeterPerRum + \\ \beta_9 AvgiftPerKvadratmeter + \epsilon_j \end{aligned} \tag{24}$$

Modell 2:

$$\begin{aligned} \log(SaljPris_j) = \beta_0 + \beta_1 Region + \beta_2 Stadsdel + \beta_3 AvstandTillVatten + \beta_4 Vaning + \\ \beta_5 Rum + \beta_6 ByggAr + \beta_7 Sasong + \beta_8 KvadratmeterPerRum + \\ \beta_9 AvgiftPerKvadratmeter + \epsilon_j \end{aligned} \tag{25}$$

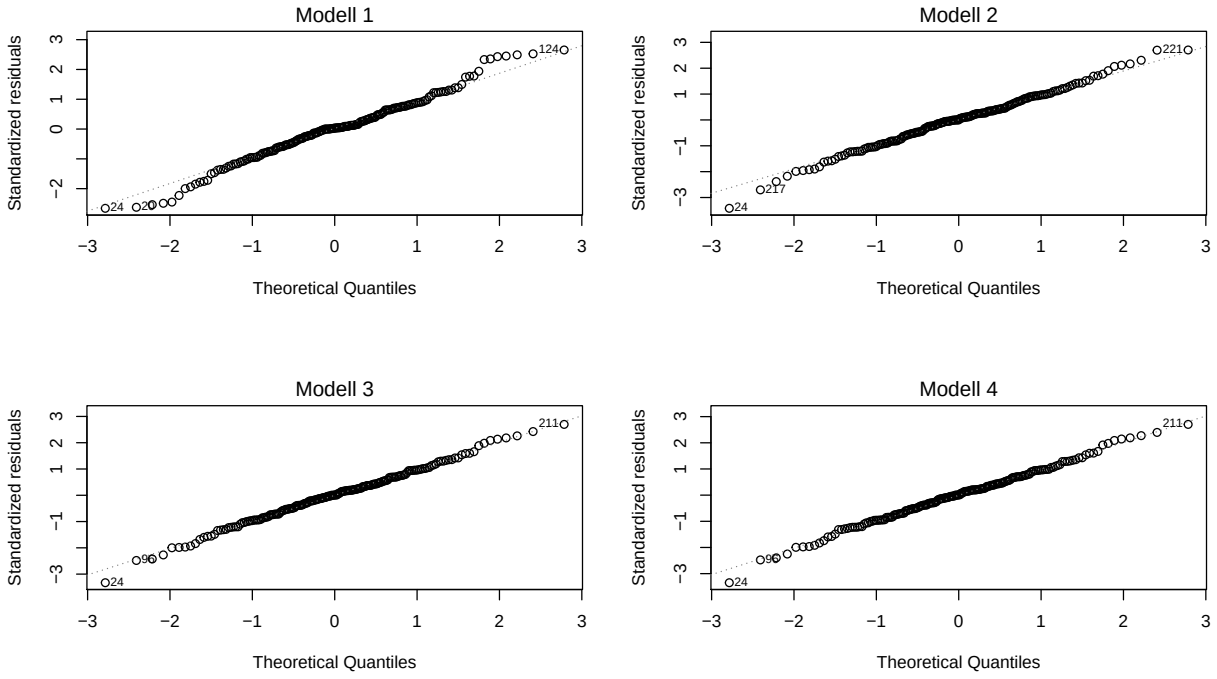
Modell 3:

$$\begin{aligned} \log(SaljPris_j) = \beta_0 + \beta_1 Region + \beta_2 Stadsdel + \beta_3 \log(AvstandTillVatten) + \beta_4 Vaning + \\ \beta_5 Rum + \beta_6 ByggAr + \beta_7 Sasong + \beta_8 \log(KvadratmeterPerRum) + \\ \beta_9 \log(AvgiftPerKvadratmeter) + \epsilon_j \end{aligned} \tag{26}$$

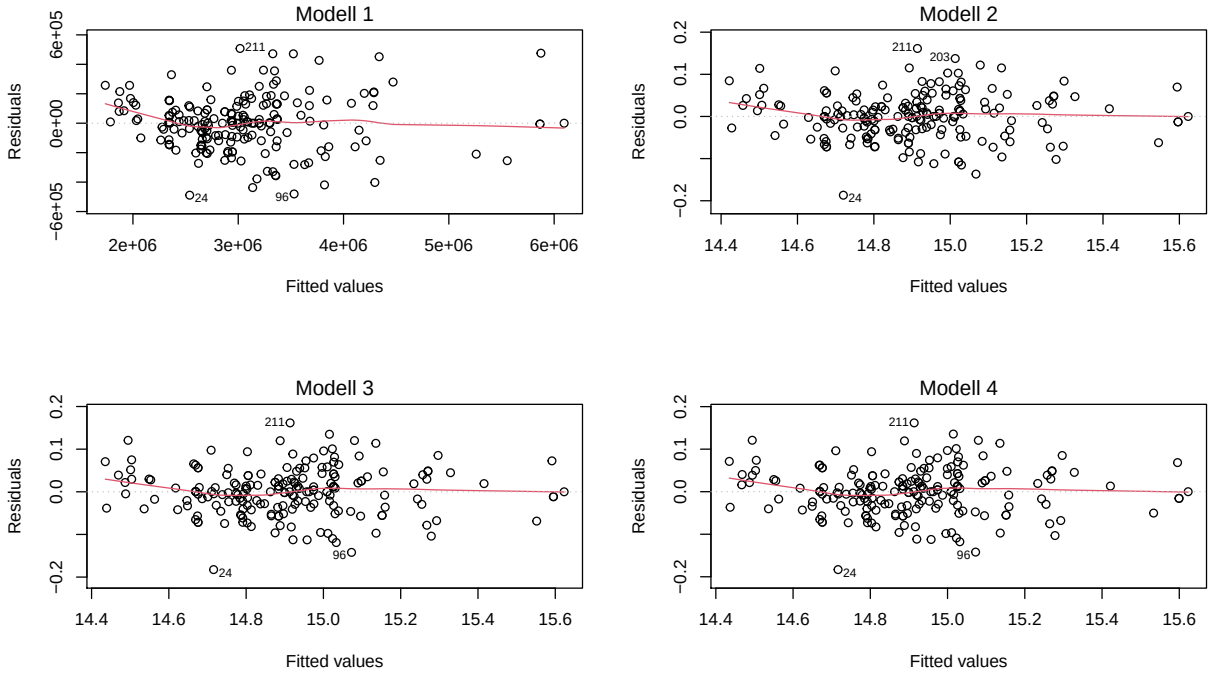
Modell 4:

$$\begin{aligned} \log(SaljPris_j) = \beta_0 + \beta_1 Region + \beta_2 Stadsdel + \beta_3 AvstandTillVatten + \beta_4 Vaning + \\ \beta_5 Rum + \beta_6 ByggAr + \beta_7 Sasong + \beta_8 \log(KvadratmeterPerRum) + \\ \beta_9 AvgiftPerKvadratmeter + \epsilon_j \end{aligned} \tag{27}$$

För en enklare överblick visar vi alla modellers plottar i ett rutnät. Det vi kan se från Figur 6 är att alla fyra modeller uppfyller normalitetsantagandet från kapitel 2.1. Dessutom visar modellerna, i Figur 7 relativt mycket homoskedacitet i residualerna.



Figur 6: Normal QQ Plott. Modell 1 till 4



Figur 7: Residuals vs Fitted. Modell 1 till 4

I ett försök att stärka Modell 1 till 4 testar vi att genomföra stegvis variabelselektion. Detta för att försöka se om vi kan hitta en modell som passar data setet vi har ännu bättre, än modellerna vi redan presenterat. Modellerna definieras nedanför i ekvation (28) till (31):

4.3.2 Definition av alla modeller efter stegvis variabelselektion

Modell 1 Steg :

$$\begin{aligned} SaljPris_j = \beta_0 + \beta_1 Region + \beta_2 Stadsdel + \beta_4 Vaning + \\ \beta_5 Rum + \beta_8 KvadratmeterPerRum + \epsilon_j \end{aligned} \tag{28}$$

Modell 2 Steg :

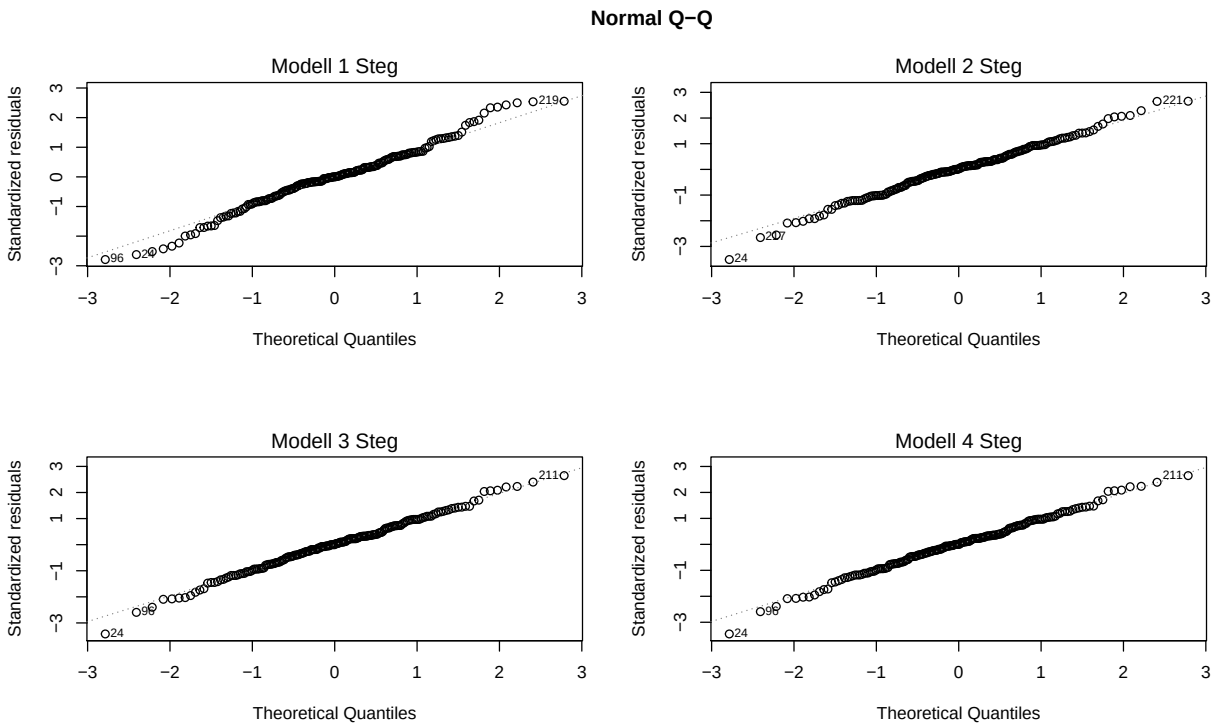
$$\begin{aligned} \log(SaljPris_j) = \beta_0 + \beta_1 Region + \beta_2 Stadsdel + \beta_4 Vaning + \\ \beta_5 Rum + \beta_8 KvadratmeterPerRum + \beta_9 AvgiftPerKvadratmeter + \epsilon_j \end{aligned} \tag{29}$$

Modell 3 Steg :

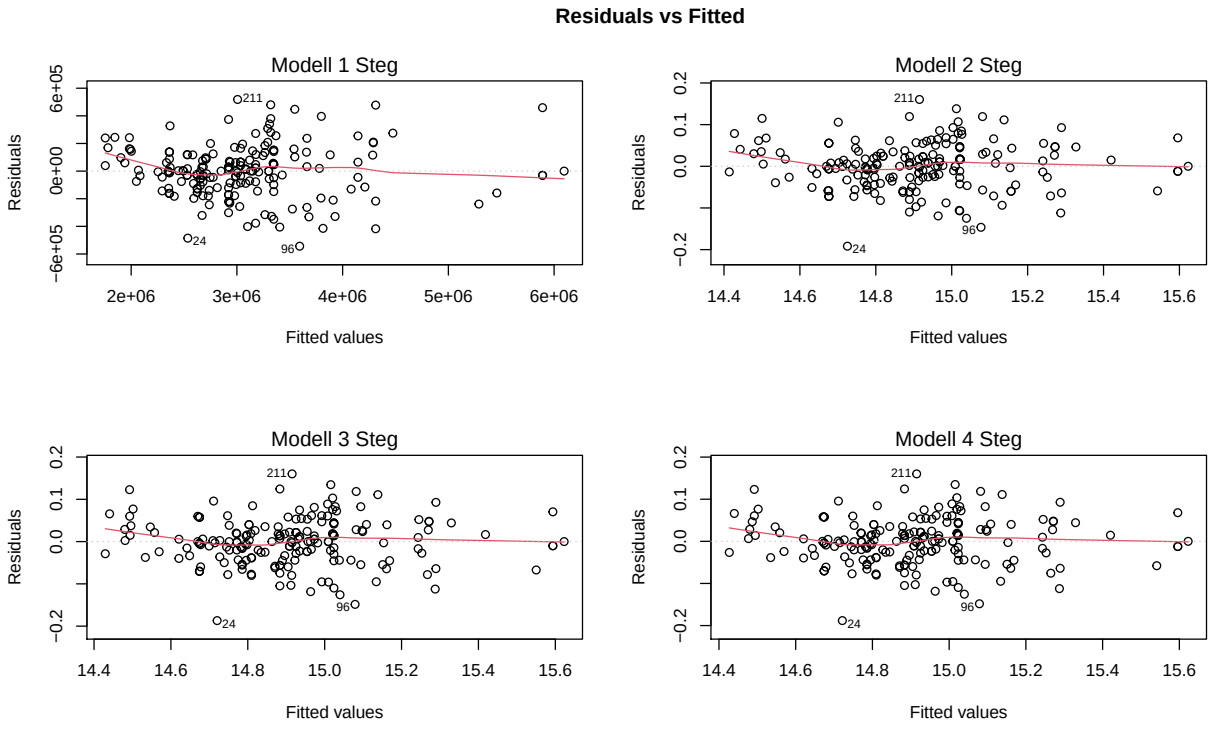
$$\begin{aligned} \log(SaljPris_j) = \beta_0 + \beta_1 Region + \beta_2 Stadsdel + \beta_4 Vaning + \\ \beta_5 Rum + \beta_8 \log(KvadratmeterPerRum) + \beta_9 \log(AvgiftPerKvadratmeter) + \epsilon_j \end{aligned} \tag{30}$$

Modell 4 Steg :

$$\begin{aligned} \log(SaljPris_j) = \beta_0 + \beta_1 Region + \beta_2 Stadsdel + \beta_4 Vaning + \\ \beta_5 Rum + \beta_8 \log(KvadratmeterPerRum) + \beta_9 AvgiftPerKvadratmeter + \epsilon_j \end{aligned} \tag{31}$$



Figur 8: Normal QQ Plott. Modell 1 Steg till 4 Steg



Figur 9: Residuals vs Fitted. Modell 1 Steg till 4 Steg

Tabell 6: Summary värden för alla modeller som presenterats

Modell	Residual Standard Error (df)	R^2	R^2_{adj}	F-statistic (df)	P-värde
Modell 1 inkl outliers	249200 on 156	0.9263	0.907	47.84 on 41 and 156	$< 2.2e-16$
Modell 1 exkl outliers	207600 on 151	0.9421	0.9267	61.41 on 40 and 151	$< 2.2e-16$
Modell 2	0.06175 on 151	0.9422	0.9269	61.53 on 40 and 151	$< 2.2e-16$
Modell 3	0.06189 on 151	0.9419	0.9266	61.25 on 40 and 151	$< 2.2e-16$
Modell 4	0.06181 on 151	0.9421	0.9268	61.41 on 40 and 151	$< 2.2e-16$
Modell 1 Steg	206500 on 157	0.9405	0.9276	72.93 on 34 and 157	$< 2.2e-16$
Modell 2 Steg	0.06122 on 156	0.9413	0.9281	71.49 on 35 and 156	$< 2.2e-16$
Modell 3 Steg	0.06114 on 156	0.9414	0.9283	71.67 on 35 and 156	$< 2.2e-16$
Modell 4 Steg	0.06111 on 156	0.9415	0.9284	71.74 on 35 and 156	$< 2.2e-16$

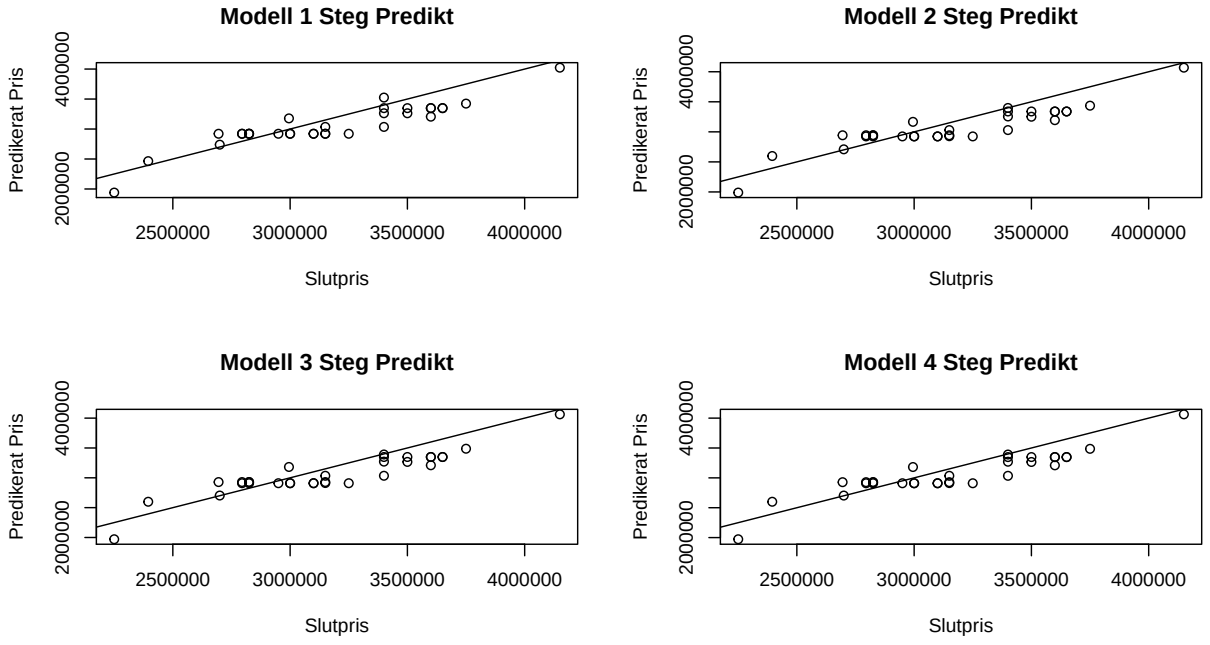
Det vi kan se i Tabell 6, är att "Modell 4", efter stegvis variabelselektion är den vinnande modellen med ett R_{adj}-värde på 0.9284, ett Residual Standard Error värde = 0.06111 på 156 frihetsgrader. Slutligen är även F-statistic värdet 71.74 på 35 och 156 frihetsgrader. "Modell 4 Steg" som är den vinnande modellen, definierades som

$$\log(SaljPris_j) = \beta_0 + \beta_1 Region + \beta_2 Stadsdel + \beta_4 Vaning +$$
$$\beta_5 Rum + \beta_8 \log(KvadratmeterPerRum) + \beta_9 AvgiftPerKvadratmeter + \epsilon_j$$

Vilket betyder att responsvariabeln $\log(Säljpris)$ för nyproducerade lägenheter i Solna-Sundbyberg beskrivs av variablerna *Region*, *Stadsdel*, *Våning*, *Rum*, $\log(KvadratmeterPerRum)$ och *AvgiftPerKvadratmeter*.

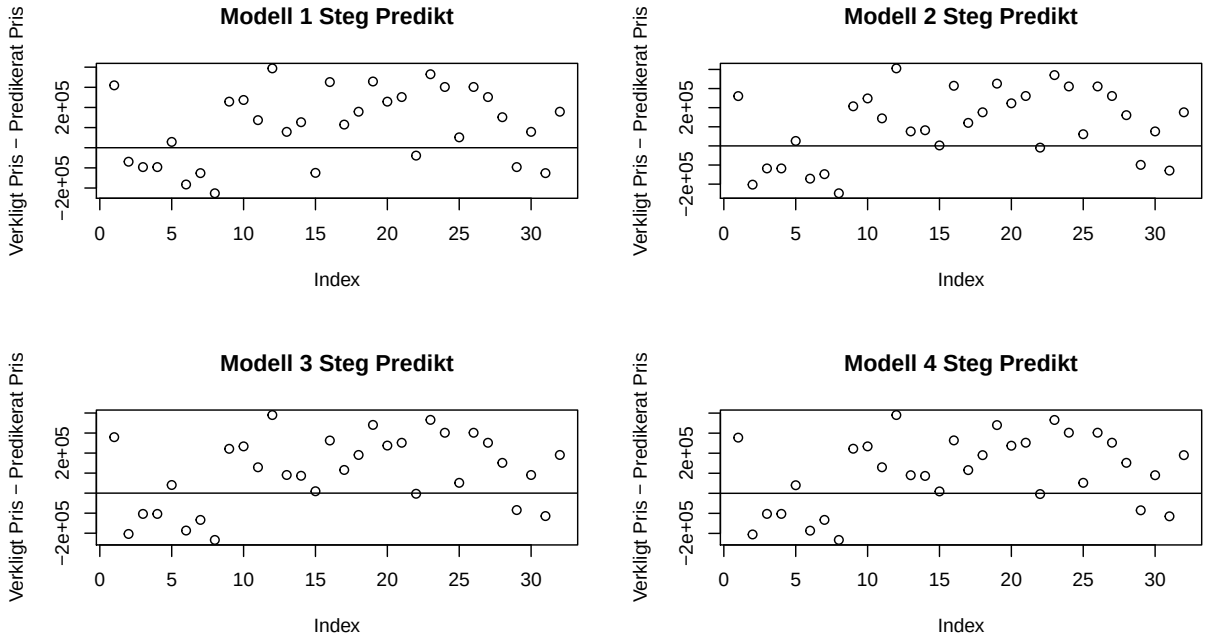
4.4 Prediktion

Modell (28) till (31) används vidare för att göra en prediktion av framtida slutpriser. Detta genom att använda all data för år 2020 som sparades för framtida värden. För att kunna se hur stark modellen är jämför vi det predikterade slutpriset mot det verkliga slutpriset. Målet är att värdena hamnar på eller så nära regressions linjen som möjligt. Resultaten presenteras nedanför i Figur 10 och 11.



Figur 10: Predikerat Pris mot Slutpris

Från Figur 10 kan vi dra slutsatsen av att "Modell 3 Steg Predikt" och "Modell 4 Steg Predikt" verkar vara de modeller med starkast predikterande förmåga. Det är även dessa modeller som var starkast i kapitel 4.3.2, gällande vilken modell som har bäst beskrivande förmåga av responsvariabeln *SäljPris*.



Figur 11: Skillnaden mellan Verkligt Pris och Predikerat Pris

Även här ser vi att skillnaden mellan det verkliga priset och det predikterade priset verkar vara minst för "Modell 3 Steg Predikt" och "Modell 4 Steg Predikt", detta tydliggörs även nedanför med hjälp av MSEP och RMSEP värdena, vilket antyder att dessa är de starkaste modellerna.

4.4.1 MSEP och RMSEP

Tabell 7: MSEP & RMSEP värden för de predikterade modellerna

	MSEP-värde	RMSEP-värde
Modell 1 Steg Predikt	43283103782	208045.9
Modell 2 Steg Predikt	44141406991	210098.6
Modell 3 Steg Predikt	42539220291	206250.4
Modell 4 Steg Predikt	42601095888	206400.3

Det vi kan se i Tabell 7 är alla MSEP och RMSEP värden för de fyra olika modellerna. Här kan vi se att "Modell 3 Steg Predikt" och "Modell 4 Steg Predikt" har de lägsta MSEP och RMSEP värdena. Vilket innebär att dessa modeller har starkast predikterande förmåga.

5 Diskussion

5.1 Resultat

Vi kan dra slutsatsen av att "Modell 4 Steg" är den vinnande modellen i vår analys. Detta då den använder sig av mindre transformationer samt har lite bättre R^2_{adj} värde än "Modell 3 steg". Tabell 14 visar alla värdena i utskriften från regressionen. Vi hade dessutom tidigare definierat "Modell 4 Steg" i ekvation (31), vilket därför slutligen ger den multiplikativa modellen:

$$SaljPris_j = e^{\beta_0} * e^{\beta_1 Region} * e^{\beta_2 Stadsdel} * e^{\beta_4 Vaning} * e^{\beta_5 Rum} * e^{\beta_8 \log(KvadratmeterPerRum)} \\ * e^{\beta_9 AvgiftPerKvadratmeter} * \hat{\epsilon_j} \tag{32}$$

Om vi analyserar modellen mer djupgående, så kan vi få fram de förklarande variablernas påverkan mot responsvariabeln *SäljPris*. Detta genom att vi väljer att sätta in våra estimerade β-värden från Tabell 14 i varje individuell förklarande variabel. Då det finns flera möjliga värden för de kategoriska förklarande variablerna, leder det till flera modellalternativ. Vi börjar med att kolla på interceptet, $e^{13.4253493} \approx 676947.7$, där det kan tolkas som baspriset men är omöjligt att uppnå eftersom det exempelvis skulle betyda att det inte finns någon kvadratmeter till en lägenhet. Vi analyserar vidare och får:

Region : Som baslinje har regionen Arenastaden använts, vilket innebär att den ingår i interceptet och att procentuella ökningen/minskningen baseras på om nyproduktionen *inte* är i Arenastaden. Eftersom *Region* är en kategorisk variabel får vi olika β beroende på vilken *Region* lägenheten ligger i, detta är lättare att visa i tabellform:

Tabell 8: Regions betaskattning och %-utveckling jämfört med Arenastaden

Region	$\hat{\beta}$	%-utveckling
Duvbo	-0.0860326	-8.24356
Entréterrass	-0.0432998	-4.23758
Gamla Råsunda	0.0006756	0.06758
Järvastaden	-0.1701975	-15.65018
Näckrosen	-0.0100285	-0.99784
Näckrosen-Råsunda	-0.0048373	-0.48256
Rissne	-0.2149445	-19.34138
Rissne/Ursvik	-0.1604216	-14.82154
Råsunda	-0.0011626	-0.11619
Råsunda Solna	0.0556797	5.7259
Skytteholm	0.1087519	11.48858
Solna-Näckrosen-Råsunda	0.0354704	3.6107
Solna Arenastaden	-0.1186083	-11.18444
Solna Centrum	0.0827923	8.63162
Solna Gamla Råsunda	0.0892614	9.33664
Solna Järvastaden	-0.0810937	-7.78927
Solna Råsunda	-0.0248958	-2.45885
Stora Ursvik	-0.1091447	-10.33994
Sundbyberg Centrum	0.0597189	6.15381
Ulriksdal	-0.0462027	-4.51517
Ursvik	-0.0984286	-9.37396
Västra Skogen	-0.0541826	-5.27409

Tabell 8 visar den procentuella utvecklingen som beräknas med hjälp av $e^{\hat{\beta}} - 1$ och kan tolkas som: En lägenhet i Duvbo får *SäljPris* att minska med 8.24% mot om lägenheten skulle vara i Arenastaden.

Stadsdel : En kategorisk variabel med baslinje Solna, vi kan se att $\hat{\beta}$ för Sundbyberg är $e^{-0.0972329} = 0.9073447$. Detta innebär att en lägenhet i Sundbyberg minskar *SäljPris* med 9.27% $(0.9073447 - 1 = -0.0926553)$ i jämförelse med om den skulle varit i Solna.

Våning : En kategorisk variabel med baslinje *Våning* $1 - 7$, jämför man med de resterande skattningarna så kan man se att *SäljPris* ökar för alla grupper, detta presenteras nedanför i Tabell 9:

Tabell 9: Vånings betaskattning och %-utveckling jämfört med Våning 1-7

Våning	$\hat{\beta}$	%-utveckling
8 till 14	0.0826798	8.6194
15	0.2534432	28.84542
16 till 18	0.1389763	14.90969
19 till 20	0.1658543	18.04011
Okänd	0.0093035	0.934691

Rum : En kategorisk variabel med baslinje *Rum* 1, jämför vi de resterande skattningarna så kan vi se att *SäljPris* ökar för varje grupp, detta presenteras nedanför i Tabell 10:

Tabell 10: Rums betaskattning och %-utveckling jämfört med Rum 1

Rum	$\hat{\beta}$	%-utveckling
1.5	0.1679386	18.2864
2	0.3511094	42.06427
3	0.6118369	84.38152
4	0.9879503	168.5724
5	1.2132777	236.44944

log(KvadratmeterPerRum) : En kontinuerlig variabel, vilket innebär att *SäljPris* ändras kontinuerligt beroende på variabelns värde. Vi får att $e^{0.4498372} = 1.568057$ som visar att för varje heltalsökning i $\log(KvadratmeterPerRum)$ så ökar *SäljPris* från interceptet med 56.81%.

AvgiftPerKvadratmeter : En kontinuerlig variabel vilket innebär att *SäljPris* ändras kontinuerligt beroende på variabelns värde. Vi får att $e^{-0.0044854} = 0.99552464$ som visar att för varje heltalsökning i *AvgiftPerKvadratmeter* så minskar *SäljPris* från interceptet med 0.4475%.

En överblick visar att de variablerna som har en positiv effekt på *SäljPris* är *Våning*, *Rum* och *KvadratmeterPerRum*. De variablerna som har en negativ effekt på *SäljPris* är *Stadsdel* och *AvgiftPerKvadratmeter*. Vi kan se att variabeln *Region* har blandad effekt, men visar mer negativa utfall än positiva. Från det slutgiltiga datasetet i kapitel 3.3 kan vi även se att variablerna *Avstånd Till Vatten*, *Byggår* och *Säsong* inte haft någon signifikant påverkan på *SäljPris* och därför tagits bort ur modellen under analysen i kapitel 4.

5.2 Andra analyser

Vi valde att undersöka om resultatet på modellerna skulle kunna stärkas om variabeln *Rum*, behandlades som en kontinuerlig variabel istället för en kategorisk variabel. Det vi kunde komma fram till var att i fallet då variabeln behandlats som kontinuerlig, så vann "Modell 4 Steg" precis som när variabeln behandlats som kategorisk. Det vi däremot kunde observera var att i fallet då *Rum* behandlats som en kategorisk variabel, resulterade det i en starkare modell. Jämförelsen presenteras i Tabell 15.

6 Referenser

[1] Sundberg, R. (2020). Lineära Statistiska Modeller. Stockholms Universitet.

[2] Andersson, P. Tyrcha, J. (2016). Notes in Econometrics. Stockholms Universitet.

[3] Held, L. Sabanés Bové, D. (2014). Applied Statistical Inference. Springer

[4] Booli.se. Booli API. Hämtad 5 november 2020. https://www.booli.se/p/api.

[5] Arnholt, T. (2020). Correlation and Regression. Hämtad 30 december 2020. https://stat-ata-asu.github.io/CorrelationRegression/.

[6] The Pennsylvania State University. Hämtad 30 december 2020. https://online.stat.psu.edu/stat462/node/247/.

[7] The Pennsylvania State University. Hämtad 30 december 2020. https://online.stat.psu.edu/stat501/lesson/11/11.3.

[8] Thim, J. (2020). Linköpnings Universitet. Föreläsning 5: Hypotesprövningar. Hämtad 7 januari 2022 https://users.mai.liu.se/johth11/tams24_fo05.pdf.

7 Appendix

Tabell 11: Model 1 Step Summary

$\beta-$koefficient	$\hat{\beta}$	P-värde
(Intercept)	528132.279	0.0145373
RegionDuvbo	-352980.032	0.1515710
RegionEntréterrass	-82016.294	0.7115116
RegionGamla Råsunda	-105916.545	0.5428240
RegionJärvastaden	-513337.989	0.0000000
RegionNäckrosen	3564.475	0.9755500
RegionNäckrosen-Råsunda	40888.669	0.7806823
RegionRissne	-641726.497	0.0001517
RegionRissne/Ursvik	-494972.282	0.0004201
RegionRåsunda	46098.181	0.5688021
RegionRåsunda Solna	178671.551	0.2922050
RegionSkytteholm	483728.229	0.0012756
RegionSolna-Näckrosen-Råsunda	168591.430	0.4468891
RegionSolna Arenastaden	-401956.045	0.0000390
RegionSolna Centrum	327290.801	0.0137487
RegionSolna Gamla Råsunda	303437.539	0.1735497
RegionSolna Järvastaden	-245726.311	0.1593543
RegionSolna Råsunda	-50853.055	0.8326412
RegionStora Ursvik	-556010.797	0.0259108
RegionSundbyberg Centrum	309246.461	0.2071799
RegionUlriksdal	-60469.439	0.7281482
RegionUrsvik	-458680.213	0.0021104
RegionVästra Skogen	-164948.151	0.3556704
StadsdelSundbyberg	-196455.398	0.0446117
Vaning 15	879823.983	0.0000002
Vaning 16 till 18	467951.240	0.0000843
Vaning 19 till 20	592758.756	0.0005545
Vaning 8 till 14	216747.700	0.0009194
Vaning Okand	48536.819	0.5931474
Rum1.5	712600.631	0.0000000
Rum2	1179045.743	0.0000000
Rum3	2044041.812	0.0000000
Rum4	3935109.369	0.0000000
Rum5	4910860.323	0.0000000
KvadratmeterPerRum	56969.222	0.0000000

Tabell 12: Model 2 Step Summary

$\beta-$koefficient	$\hat{\beta}$	P-värde
(Intercept)	14.4016829	0.0000000
RegionDuvbo	-0.1058393	0.1509449
RegionEntréterrass	-0.0455039	0.4933579
RegionGamla Råsunda	-0.0111854	0.8284545
RegionJärvastaden	-0.1725122	0.0000000
RegionNäckrosen	-0.0184471	0.6065184
RegionNäckrosen-Råsunda	-0.0006700	0.9879506
RegionRissne	-0.2271294	0.0000098
RegionRissne/Ursvik	-0.1700654	0.0000603
RegionRåsunda	-0.0021520	0.9314833
RegionRåsunda Solna	0.0440188	0.3891776
RegionSkytteholm	0.1084599	0.0145473
RegionSolna-Näckrosen-Råsunda	0.0392009	0.5530996
RegionSolna Arenastaden	-0.1187365	0.0000419
RegionSolna Centrum	0.0967801	0.0156387
RegionSolna Gamla Råsunda	0.0810563	0.2252301
RegionSolna Järvastaden	-0.0812026	0.1170251
RegionSolna Råsunda	-0.0280869	0.6942094
RegionStora Ursvik	-0.1154063	0.1174364
RegionSundbyberg Centrum	0.0822825	0.2731879
RegionUlriksdal	-0.0375925	0.4746135
RegionUrsvik	-0.1132600	0.0103470
RegionVästra Skogen	-0.0668951	0.2128460
StadsdelSundbyberg	-0.0876600	0.0030330
Vaning 15	0.2545473	0.0000005
Vaning 16 till 18	0.1406756	0.0000798
Vaning 19 till 20	0.1670163	0.0010482
Vaning 8 till 14	0.0839398	0.0000592
Vaning Okand	0.0141343	0.6008958
Rum1.5	0.2079932	0.0000000
Rum2	0.3860406	0.0000000
Rum3	0.6463227	0.0000000
Rum4	1.0269147	0.0000000
Rum5	1.2418278	0.0000000
KvadratmeterPerRum	0.0176890	0.0000000
AvgiftPerKvadratmeter	-0.0046511	0.0037787

Tabell 13: Model 3 Step Summary

$\beta-$koefficient	$\hat{\beta}$	P-värde
(Intercept)	14.2428354	0.0000000
RegionDuvbo	-0.0862913	0.2395867
RegionEntréterrass	-0.0432142	0.5150087
RegionGamla Råsunda	0.0010531	0.9836726
RegionJärvastaden	-0.1705129	0.0000000
RegionNäckrosen	-0.0099946	0.7809588
RegionNäckrosen-Råsunda	-0.0047113	0.9153290
RegionRissne	-0.2144592	0.0000277
RegionRissne/Ursvik	-0.1607984	0.0001455
RegionRåsunda	-0.0011267	0.9640866
RegionRåsunda Solna	0.0558408	0.2762183
RegionSkytteholm	0.1089891	0.0139605
RegionSolna-Näckrosen-Råsunda	0.0353812	0.5919883
RegionSolna Arenastaden	-0.1186848	0.0000414
RegionSolna Centrum	0.0833989	0.0359524
RegionSolna Gamla Råsunda	0.0894277	0.1813444
RegionSolna Järvastaden	-0.0810584	0.1172339
RegionSolna Råsunda	-0.0251832	0.7240289
RegionStora Ursvik	-0.1086610	0.1397776
RegionSundbyberg Centrum	0.0607726	0.4142228
RegionUlriksdal	-0.0467539	0.3739927
RegionUrsvik	-0.0991654	0.0242819
RegionVästra Skogen	-0.0548619	0.3062238
StadsdelSundbyberg	-0.0979373	0.0009736
Vaning 15	0.2532250	0.0000006
Vaning 16 till 18	0.1388089	0.0000969
Vaning 19 till 20	0.1657892	0.0011237
Vaning 8 till 14	0.0824642	0.0000780
Vaning Okand	0.0093642	0.7279509
Rum1.5	0.1689333	0.0000001
Rum2	0.3519402	0.0000000
Rum3	0.6128169	0.0000000
Rum4	0.9863285	0.0000000
Rum5	1.1944043	0.0000000
log(KvadratmeterPerRum)	0.4510017	0.0000000
log(AvgiftPerKvadratmeter)	-0.2666115	0.0057063

Tabell 14: Model 4 Step Summary

$\beta-$koefficient	$\hat{\beta}$	P-värde
(Intercept)	13.4253493	0.0000000
RegionDuvbo	-0.0860326	0.2405759
RegionEntréterrass	-0.0432998	0.5139176
RegionGamla Råsunda	0.0006756	0.9895191
RegionJärvastaden	-0.1701975	0.0000000
RegionNäckrosen	-0.0100285	0.7799904
RegionNäckrosen-Råsunda	-0.0048373	0.9130179
RegionRissne	-0.2149445	0.0000266
RegionRissne/Ursvik	-0.1604216	0.0001477
RegionRåsunda	-0.0011626	0.9629016
RegionRåsunda Solna	0.0556797	0.2772996
RegionSkytteholm	0.1087519	0.0141283
RegionSolna-Näckrosen-Råsunda	0.0354704	0.5908106
RegionSolna Arenastaden	-0.1186083	0.0000415
RegionSolna Centrum	0.0827923	0.0373333
RegionSolna Gamla Råsunda	0.0892614	0.1818965
RegionSolna Järvastaden	-0.0810937	0.1168971
RegionSolna Råsunda	-0.0248958	0.7268966
RegionStora Ursvik	-0.1091447	0.1378297
RegionSundbyberg Centrum	0.0597189	0.4224216
RegionUlriksdal	-0.0462027	0.3789068
RegionUrsvik	-0.0984286	0.0251768
RegionVästra Skogen	-0.0541826	0.3114890
StadsdelSundbyberg	-0.0972329	0.0010636
Vaning 15	0.2534432	0.0000005
Vaning 16 till 18	0.1389763	0.0000945
Vaning 19 till 20	0.1658543	0.0011124
Vaning 8 till 14	0.0826798	0.0000744
Vaning Okand	0.0093035	0.7295199
Rum1.5	0.1679386	0.0000001
Rum2	0.3511094	0.0000000
Rum3	0.6118369	0.0000000
Rum4	0.9879503	0.0000000
Rum5	1.2132777	0.0000000
log(KvadratmeterPerRum)	0.4498372	0.0000000
AvgiftPerKvadratmeter	-0.0044854	0.0052431

Tabell 15: Jämförelse mellan variabeln Rum som kategorisk och Rum som kontinuerlig

Modell	Residual Standard Error (df)	R^2	R^2_{adj}	F-statistic (df)	P-värde	RMSEP-värde
Modell 4 Steg kontinuerlig	0.06471 on 161	0.9329	0.9196	69.97 on 32 and 161	$< 2.2e-16$	224677.3
Modell 4 Steg kategorisk	0.06111 on 156	0.9415	0.9284	71.74 on 35 and 156	$< 2.2e-16$	206400.3