Stockholms universitet

Regressionsanalys av lägenhetspriser i Stockholms Norrort

Ramtin Golrang

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2022:8
Matematisk statistik
Juni 2022

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Matematisk statistik
Stockholms universitet
Kandidatuppsats **2022:8**
http://www.math.su.se/matstat

Regressionsanalys av lägenhetspriser i Stockholms Norrort

Ramtin Golrang*

Juni 2022

Sammanfattning

Det huvudsakliga syftet med denna studie, är att undersöka vilka faktorer som faktiskt påverkar slutpriset av lägenheterna i Stockholms Norrort men även kunna prediktera framtida slutpriser. Metoden för att undersöka detta, har varit en linjär regressionsanalys där vi fått fram åtta olika modeller, som vi jämfört för att få fram den mest anpassade modellen med hög förklarings- och prediktionsförmåga. Modellerna togs fram med hjälp av transformationer och metoder som stegvis variabelselektion. Studien är begränsad till Stockholms Norrort på kommunal nivå och studien vänder sig till både säljare och köpare då den tar hänsyn till flertalet förklarande variabler.

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

Innehåll

1 Introduktion

Idag är det en stor efterfrågan på lägenheter och bostadsmarknaden har ökat drastiskt i pris. Ingen skulle väl tacka nej till en lägenhet men till vilket pris? När är det egentligen läge att köpa? När ska man sälja? Spelar det någon roll vilken mäklare man väljer? Det är några av många frågor denna studie besvarar och syftar till att hjälpa den nyfikne.

I denna studie behandlas lägenhetspriserna i Stockholms Norrort på kommunal nivå med hjälp av en regressionsanalys. Data är insamlad från Booli Search Technologies AB:s API mellan åren 2019-2022 och innehåller information om exempelvis vilken mäklare som sålde lägenheten, slutpriset och vilken säsong den såldes. Det huvudsakliga syftet med denna studie är att kunna använda den slutliga modellen för att kunna prediktera slutpriset på en lägenhet i Stockholms Norrort baserat på den information man har fått. Studien vänder sig till både säljare och köpare då den tar hänsyn till vilken säsong det är bäst att lägga ut sin lägenhet samt om det spelar någon roll vilken mäklarfirma man väljer. En köpare kan också finna det intressant att kontrollera vad ett rimligt pris är för en lägenhet som hen är intresserad av. Modellen kan också vara intressant för mäklare, dels för att få ett uppskattat pris, men även jämföra om det hade gjort någon skillnad om det var en annan mäklare.

Studien kommer att besvara vilka faktorer som faktiskt påverkar slutpriset på en lägenhet i Norrort samt ta fram en modell för att prediktera framtida slutpriser.

I studien kommer vi gå igenom teorier för att ge förståelse till hur en linjär regression fungerar och byggs. Vi kommer börja med en datahantering för att ta bort all ointressant data som exempelvis variabeln annons id för att få med det mest relevanta till modellen. Vi undersöker även kollinaritet och omvandlar variabler så att inte nödvändig information försvinner. Efter vi har kontrollerat att all data uppfyller kraven för linjära modeller så skapar vi vår första modell. Vi använder oss sedan av olika transformationer och metoder för att få ut en så bra modell som möjligt. När vi utfört alla stegen, jämför vi de modeller vi kommit fram till för att avgöra vilken modell som har den bästa predikterande samt den starkaste förklarande förmågan. Sedan undersöker vi i hur stor omfattning den utvalda modellen, uppfyller båda kriterium. Modellen vi kom fram till har en väldigt liten felmarginal i prediktionen och en hög förklaringsgrad.

Några frågor som läsaren kan tänka på ytterligare genom studien är:

- Är den starkaste förklarande modellen även den bäst predikterande modellen?
- Vilken förklarande variabel har störst påverkan på slutpriset?

2 Teoretisk bakgrund och metoder

I detta kapitel kommer vi gå igenom de grundläggande teorier och metoder som kommer användas genom hela arbetet.

2.1 Linjär regression

Linjär regression användas för att studera relationen mellan variabler, kontrollera sambanden mellan variablernas styrka och utföra prediktioner.

2.1.1 Multipel linjär regression

Med en modell från en multipel linjär regression kan man försöka prediktera ett värde på en responsvariabel (y) med hjälp av de linjära sambanden till de förklarande variablerna (x). Strukturen av modellen kan skrivas som:

$$Y = A\theta + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I_N) \tag{1}$$

Där

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix} \tag{2}$$

$$\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_N \end{pmatrix} \tag{3}$$

$$A = \begin{pmatrix} 1 & X \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & \cdots & x_{Nk} \end{pmatrix} \tag{4}$$

$$\theta = \begin{pmatrix} \alpha \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \tag{5}$$

varav Y är responsvektorn, ε residualvektorn, A designmatrisen, θ parametervektorn, α interceptet och $\beta_i, 1 \leq i \leq k$ motsvarar sambanden mellan responsvariabeln och de förklarande variablerna, k som motsvarar antalet förklarande variabler och slutligen N som motsvarar antalet responsvariabler (antalet linjära regressioner eller stickprovsmängden).[2]

För att få bygga en sådan modell så måste man uppfylla vissa antaganden:

- Linjäritet – Regressionsmodellens parametrar är linjära.
- Avsaknad av linjära kombinationer – Kolumnerna i X (de förklarande variablerna) är linjärt oberoende av varandra och det finns minst k stycken observationer.
- Homoskedasticitet – Residualernas varians är densamma oavsett värdet på de förklarande variablerna, $Var[\varepsilon_i] = \sigma^2$ för alla i.

- Ingen korrelation – Residualerna antas inte ha någon korrelation mellan varandra, $Cov(\varepsilon_i, \varepsilon_j) = 0$ för alla $i \neq j$.
- Normalitet – För att få göra en linjär regression antas det att residualerna är normalfördelade med väntevärde 0 och definieras som $\varepsilon_i \sim N(0, \sigma^2)$ vilket innebär att $E[\varepsilon_i] = 0$ för alla i och $\varepsilon|X = \varepsilon \sim N(0, \sigma^2 I_N)$.[1]

2.1.2 Enkel linjär regression

Enkel linjär regression är ett specialfall av multipel linjär regression där man bara har en förklarande variabel, strukturen blir då samma som innan i **ekvation (1)-(3)** men med $k = 1$ i **ekvation (4)-(5)**.[4]

2.1.3 Variabeltyper

En förklarande variabel x kan anta olika variabeltyper och man måste därför först bestämma vilken typ av data man har samlat in, för att man ska kunna avgöra vilken metod som behöver användas för att analysera datasetet. De förklarande variablerna är oftast en av följande variabeltyper:

- Kontinuerliga värden, som kan anta alla värden inom ett variationsområde, t.ex. koncentrationsmätningar.
- Räknedata, som bara kan anta heltal och uppstår när man räknar en slumpmässig händelse, t.ex. antal fjärilar på en plats.
- binära data, som bara kan anta två olika värden, t.ex finns/finns inte eller över/under ett gränsvärde.
- Kategorisk data, som kan anta ett fåtal olika värden, som t.ex. hög/medium/låg eller röd/grön/blå.

[6]

2.1.4 Sammanfoga kategorier

I vissa fall för linjära modeller, normalast i de kategoriska variablerna kan man stöta på kategorier med för lite data. Detta kan leda till att kategoriernas estimeringar inte är signifikanta och har därför en negativ påverkan på modellens riktighet. Lösningen i fallet med för lite data är att sammanfoga kategorierna till en ny kategori. Ett exempel på detta är om man har den kategoriska variabeln ålder och ett data som innehåller 100 personer med slumpvis ålder, då skulle det inte vara lämpligt att kategorisera den specifika åldern i data eftersom det lätt leder till många mindre kategorier. Istället skulle man kunna sammanfoga de mindre kategorierna till nya kategorier som tillexempel ”Ålder 1-25, 26-50, 51-75 och 75+”, vilket skulle leda till mer data i varje kategori och bättre estimeringar i modellerna.[5]

2.1.5 Parameterskattning (MK-Metoden)

För att kunna skatta parametrarna (interceptet och sambanden) till linjär regression, använder vi oss av minsta-kvadratmetoden (MK-metoden).

Metoden innebär att man vill skatta en parametervektor som ger så liten kvadratsumma av residualvektorn som möjligt, kvadratsumman kan skrivas som en skalärprodukt:

$$(Y - A\theta)^T(Y - A\theta) = ||(Y - A\theta)||^2 = ||\varepsilon||^2 \tag{6}$$

Man söker då parametervektorn som minimerar avståndet $||(Y - A\theta)||$.

Skattningen för parametervektorn ges av lösningen:

$$\hat{\theta} = (A^T A)^{-1} A^T Y = \begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = \begin{pmatrix} \hat{\alpha} \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{pmatrix} \tag{7}$$

Där $\hat{\theta} \sim N(\theta, \sigma^2 (A^T A)^{-1})$, det gäller även att $\hat{\alpha}$ och $\hat{\beta}_i$, $1 \leq i \leq k$ motsvarar skattningen av interceptet och sambanden.[2]

2.1.6 Hypotesprövning

Generellt för denna typ av studie är att man vill undersöka vilken av de förklarande variablerna som har en påverkan på responsvariabeln. Detta kan man göra via en hypotesprövning, vilket innebär att man har en nollhypotes H_0 som testas mot en alternativ hypotes H_1. För att ta reda på om de förklarande variablerna har någon effekt undersöker man hypotesen $H_0 : \theta_i = 0$ mot den alternativa hypotesen $H_1 : \theta_i \neq 0$.[1] Detta gör man genom att beräkna en statistika:

$$T = \frac{\hat{\theta}_i - \theta_i}{\hat{\sigma}\sqrt{(A^T A)^{-1}_{ii}}} \tag{8}$$

där $T \sim t(n-k)$ som motsvarar en t-fördelning med $n-k$ frihetsgrader och $\hat{\sigma} = \frac{1}{n}\sum_{j=1}^{n} \epsilon_j^2$ som motsvarar den skattade standardavvikelsen.[1][2] Hypotesen H_0 förkastas om statistikan är signifikant, detta testas via:

$$|T| = t_{\frac{\alpha}{2}}(n-k) \tag{9}$$

där α motsvarar signifikansnivån och nollhypotesen förkastas då $|T| \geq t_{\alpha/2}(n-k)$.[1]

2.1.6.1 P-värde metoden

Normalt använder man $\alpha = 0.05$ som en gräns för tester. Men α är egentligen ett P-värde, och kan beräknas via $p = P(förkasta \;\; H_0, \; givet \; att \; H_0 \; är \; sann)$, detta P-värde kan användas som ett direkt test för att undersöka hur stort motiv det finns att förkasta H_0. Om $p < \alpha$ så förkastas H_0, ju lägre P-värde desto större incitament till att förkasta H_0. P-värdet är sannolikheten att få det observerade värdet eller ett mer extremt.[7]

2.1.7 Korrelation

Korrelation anger inom statistiken styrkan och riktningen av ett samband mellan två eller flera variabler. Korrelationen anges ofta med en korrelationskoefficient (ρ) och har ett värde mellan 1 och -1, där 0 anger inget samband, 1 anger maximalt positivt samband och -1 anger maximalt negativt samband. Det finns många olika sätt att beräkna korrelationen men den vanligaste formen är Pearsons korrelationskoefficient och fås via:

$$\hat{\rho} = r = \frac{c_{xy}}{s_x s_y} \tag{10}$$

där $c_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$, $s_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}$ och $s_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \overline{y})^2}{n-1}}$.

VARNING: man ska inte tolka hög korrelationskoefficient som att det MÅSTE råda ett linjärt beroende mellan x och y. Även ett starkt olinjärt samband kan ge högt r-värde.[2]

2.1.8 Nästan-kollinearitet

Nästan-kollinearitet eller multikollinearitet som det också kallas, uppstår då det finns minst en korrelation mellan en och flera variabler. Det vill säga om informationen i en av de förklarande variablerna kan beskrivas med hjälp av några andra förklarande variabler, då har man multikollinearitet och informationen blir överflödig.[2]

2.1.8.1 Variance Inflation Factor

Ett sätt för att mäta multikollinearitet är Variance Inflation Factor (VIF-Faktorn) som fås av:

$$VIF_j = \frac{1}{1 - R_j^2}, 1 \leq j \leq k, \tag{11}$$

där R_j^2 står för den förklaringsgrad som anger hur mycket av variationen i x_j som förklaras av de andra förklarande variablerna och beskrivs nedanför i kapitel 2.2.3.1. Man strävar efter att alla $VIF_j < 5$ för att undvika multikollinearitet så gott det går, optimalt är om $VIF_j = 1$, då är x_j ortogonal mot de övriga variablerna.[2]

2.2 Modellval

Eftersom att man kan göra flera olika varianter (modeller) av linjära regressioner från samma dataset, så vill man jämföra dessa modeller med varandra och ta fram den som uppfyller villkoren bäst. Nedan kommer vi gå igenom metoder för att skapa olika modeller samt teorier för att jämföra dessa.

2.2.1 Outlier identifiering

När man skapar en linjär regression så kan det förekomma outliers, en outlier är en ovanlig observation som skiljer sig från andra observationer och kan förekomma av olika anledningar som till exempel felinmatning eller slumpen. En outlier har oftast stor påverkan på den linjära regressions modellen vilket ger modellen sämre förutsättningar.

För att identifiera en outlier kan man använda sig av olika outlier test, en av dem, som används i denna studie är kontrollen av Studentized residuals värdena. Studentized residuals beräknas som:

$$t_i = \frac{d_i}{s(d_i)} \tag{12}$$

där $s(d_i)$ är den skattade standarvavvikelsen av d_i och $d_i = y_i - \hat{y}_{(i)}$, y_i är det observade värdet för observation i, $\hat{y}_{(i)}$ är den predikterade värdet för observation i baserat på en modell där man tagit bort observation i.

Om $|t_i| > 3$ så kallas observationen en outlier och bör behandlas för att få en förbättrad modell.[8]

2.2.2 Transformationer

För att kunna förbättra modellerna bör man undersöka om det hjälper med transformationer, normalt kontrollerar man residualerna efter Hetroskedasticitet för att konstatera om en transformation behövs. Det är inget problem om respons- eller förklarande variabler skulle göras icke-lineära och det är därför fritt att transformera dem, men om parametrarna α och β inte är lineära så finns det inte en lineär modell och leder till andra problem. Detta problem går att lösa på olika sätt, ibland räcker det med en lämplig transformation för att överföra modellen till en lineär modell.

En typ av transformation som kommer användas genom denna studie är log transformation. En anpassningsmöjlighet för en modell kan vara en multiplikativ modell, som fås på formen:

$$Y = \alpha e^{\beta x} \varepsilon \tag{13}$$

En multiplikativ modell anses inte vara lineär men genom att logaritmera Y-värdena ovan kommer vi tillbaka till den lineära additiva modellen på formen:

$$ln(Y) = ln(\alpha) + \beta x + ln(\varepsilon) \tag{14}$$

[2]

2.2.3 Anpassningsmått

2.2.3.1 Förklaringsgraden (R^2)

Det vanligaste anpassningsmåttet för val av modell är förklaringsgraden (R^2). En modells förklaringsgrad kan beskrivas som andelen av den totala variationen modellen "förklarar" och ges av:

$$R^2 = \frac{Kvs(regression)}{Kvs(totalt)} = 1 - \frac{Kvs(residual)}{Kvs(totalt)}. \tag{15}$$

Där Kvs står för Kvadratsumma, $Kvs(regression) = \sum_{i=1}^{N}(\sum_{j=1}^{m} \hat{\beta}_j(x_{ij} - \bar{x}_{.j}))^2 = (X\hat{\beta})^T(X\hat{\beta})$, $Kvs(totalt) = \sum_{i=1}^{N}(y_i - \bar{y})^2$ och $Kvs(Residual) = Kvs(totalt) - Kvs(regression)$.

En nackdel med R^2 är att den ökar om man tillför fler förklarande variabler, det gäller även om det skulle vara en helt irrelevant variabel som slumpats fram. Därför blir det svårt att använda R^2 för att avgöra om en modell med en extra variabel förklarar mer än nuvarande modellen.

För att undkomma detta är det bättre att undersöka om σ^2 minskar, det kan tolkas som att det finns mindre slump kvar i modellen och måttet fås av *Adjusted* R^2 som ges av:

$$R^2_{adj} = 1 - \frac{MKvs(residual)}{MKvs(totalt)} = 1 - \frac{\frac{Kvs(residual)}{N-m-1}}{\frac{Kvs(total)}{N-1}} = 1 - (1 - R^2)\frac{N-1}{N-m-1}. \tag{16}$$

där N är antalet stickprov och m är antalet förklarande variabler. [2]

2.2.3.2 Residual Standard Error

Ett annat anpassningsmått är Residual Standard Error (RSE) som används för att se hur väl en modell passar ett dataset. RSE beräknas som:

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

Där y_i är det observerade värdet, $\hat{y}_i$ det skattade värdet och df är antalet frihetsgrader (antalet observationer - antalet modelparametrar)

Ju mindre värde på RSE desto bättre passar modellen datasetet.[10]

2.2.3.3 F-statistic och p-value

F-statistikan kan användas för att mäta sambandet mellan responsvariabeln och de förklarande variablerna, ju högre F-statistika, desto lägre p-värde. Vilket innebär att modellen blir mer signifikant och är bättre anpassad till datasetet.

F-statistikan fås genom formeln:

$$F = \frac{\frac{Kvs(regression)}{k-1}}{\frac{Kvs(residual)}{n-k}} \in F(k-1, n-k). \tag{18}$$

Och p-värdet fås av:

$$P(F(k-1, n-k) > \frac{\frac{Kvs(regression)}{k-1}}{\frac{Kvs(residual)}{n-k}}) = p. \tag{19}$$

[1]

2.2.4 Prediktionsmått

2.2.4.1 Mean Squared Error of Prediction

När man konstruerar en prediktor är det är viktigt att ha koll på prediktionsmåttet, en variant av prediktionsmått är det som kallas Mean Squared Error of Prediction (MSEP). För att beräkna MSEP så behöver man använda sig av en metod som kallas Korsvalidering. Korsvalidering innebär att man tillfälligt tar bort en observation i, från data och skattar en ny regression utan observation i, som man kallar $\hat{\mu}_{-i}$. Därefter använder man den för att prediktera värdet för den borttagna observationen $y_i^* = \hat{\mu}_{-i}(x_i)$. Då man vet y-värdet för observation i från data, så kan man beräkna prediktionsfelet $y_i - y_i^*$. Med hjälp av prediktionsfelet kan vi beräkna MSEP och får då:

$$MSEP = \frac{1}{N}\sum_{1}^{N}(y_i - y_i^*)^2. \tag{20}$$

Från MSEP finns de fler prediktionsmått som man kan använda, de två som används i studien är:

- RMSEP (Root Mean Squared Error of Prediction) som är kvadratroten ur MSEP.
- PRESS (PREdiction Sum of Squares) som är N*MSEP.

Där N är antalet stickprov.[2]

2.2.4.2 Akaike's Information Criterion

Ett annat prediktionsmått som är ett likartat alternativ till MSEP är Akaike's Information Criterion (AIC) som är den likelihoodbaserade varianten för ett prediktionsmått och fås av [2]:

$$AIC = -2l(\hat{\theta}_{ML}) + 2p, \tag{21}$$

där p är antalet parametrar, $l(\hat{\theta}_{ML})$ är log-likelihood funktionen och $\hat{\theta}_{ML}$ maximum likelihood skattningen. Man söker en modell med så liten AIC som möjligt.[3]

2.2.5 Stegvis variabelselektion

Med multipel linjär regression vill man prediktera responsvariabeln (y), baserat på en kombination av de förklarande variablerna (x_i), detta kallas för en prediktionsmodell. Istället för att testa alla olika kombinationer av x_i, för att se vilken modell som är mest lämpad att prediktera med, så kan man använda sig av olika variabelselektions metoder. De mest välkända och användbara metoderna är Bakåt-metoden, Framåt-metoden och Stegvis-selektion.[1] Efter man utfört dessa metoder så kan man avgöra vilken av modellerna som är bäst anpassad, genom att kontrollera vilken modell som fått bästa resultat mot stoppkriteriet. Det finns olika stoppkriterier som man kan välja mellan, exempelvis förklaringsgraden (R^2) som mäter hur väl modellen beskriver data eller AIC som är besläktat till $RMSEP$ och som är ett bra kriterium till en prediktionsmodell.[2]

2.2.5.1 Bakåt-Metoden

Bakåt-metoden går till så att man börjar med en modell med alla förklarande variabler sedan, med hjälp av ett hypotestest undersöker man vilken av de förklarande variablerna, som är minst lämpad att vara kvar i modellen. Därefter tar man bort den förklarande variabeln som inte klarade hypotestestet och fick sämst resultat. Sedan gör man om proceduren tills man når sitt stoppkriterium.[1]

2.2.5.2 Framåt-Metoden

För Framåt-metoden börjar man i motsatt riktning, man har en tom modell och kontrollerar med hjälp av hypotestest, vilken förklarande variabel som är mest lämpad att tas med i modellen. Därefter skapar man en ny modell, som innehåller den mest lämpade förklarande variabeln, för att sedan använda med den nya modellen mot resterande förklarande variabler och se vilken som är mest lämpad att adderas till modellen. Detta gör man stegvis tills man når sitt stoppkriterium.[1]

2.2.5.3 Stegvis-selektion

Stegvis-selektion är ett tillägg för bakåt och framåt metoderna. Under varje steg som man utför någon av dessa metoder, så ska man dubbelkolla om något av de föregående stegen inte är signifikanta längre. Ett exempel skulle vara om man med Framåt-metoden byggt en modell med fyra förklarande variabler och precis lagt till en femte, då ska man kontrollera att de fyra man lagt till innan fortfarande blir signifikanta i hypotestestet.[1]

3 Data

I detta arbete har data givits via API från företaget Booli Search Technologies AB [9] som är en sökmotor för den svenska bostadsmarknaden. Datat innehöll 11288 lägenheter i Norrort som sålts under perioden 2019-01-01 till 2022-04-03. Genom arbetet har data modifierats och filtrerats via programspråket R.

3.1 Original data

Boolis original data innehöll information om dessa variabler:

Id: Ett annons-ID för boolis plattform.

url: Länk till original annonsens plattform.

Typ av försäljare: Vilken typ av försäljare, om det var en mäklare eller byggfirma.

Försäljarens namn: Vilket företag sålde lägenheten.

Rum: Antalet rum lägenheten innehåller.

ByggÅr: Året lägenheten byggdes.

SåldDatum: Datumet då lägenheten såldes.

Stadsdel: Vilken stadsdel i Norrort som lägenheten finns i.

Adress: Vilken adress som lägenheten ligger på.

Latitud: Vilken latitud som lägenheten ligger på.

Longitud: Vilken longitud som lägenheten ligger på.

Kommun: Vilken Kommun i Norrort som lägenheten finns i.

Län: Vilket län i Norrort som lägenheten finns i.

AvståndVattenMeter: Lägenhetens avstånd till närmsta sjö eller hav, mätt i m

Område: Vilket lokalt område i Norrort som lägenheten finns i.

BiArea: Storleken på bi-area om lägenheten har sådan, mätt i kvadratmeter.

Hyra: Hyra per månad för lägenheten.

BoArea: Storleken på lägenheten, mätt i kvadratmeter.

SåldPris Källa: Källan för vart det sålda priset hämtats ifrån.

ListPris: Mäklarens bedömning på vilket pris som lägenheten ska listas för.

SåldPris: Vilket pris som lägenheten såldes för.

Våning: Vilken våning som lägenheten ligger på.

3.2 Behandling av data

Datamängden innehöll en del överflödig och orelevant information som man kunde ta bort, viss information behövde göras om för att behålla informationen som man kunde få av data. Här kommer vi gå igenom vilken behandling vi valt att göra och varför. Det data vi valt att ta bort på grund av att det saknas relevant information är: **Id**, **url**, **Typ av försäljare** och **SåldPris Källa**. Vi tog även bort **BiArea** då den inte innehöll någon information eftersom lägenheterna inte hade någon bi-area. Det fanns även data som innehöll korrelation och behövde antingen omvandlas till nya kombinationer av data, för att behålla informationen eller tas bort från datamängden, då informationen redan beskrivs. Bland de geografiska data valde vi att behålla **Kommun** och ta bort **Stadsdel**, **Område**, **Adress**, **Latitud**, **Longitud** och **Län**. Anledningen till detta är för att informationen är korrelaterad till vilken kommun som lägenheten ligger i, vi är inte heller intresserade av att gå in mer djupgående i kommunerna då studien jämför alla Norrorts kommuner och nivån ovanför kommun (**Län**) ger ingen nyttig information då alla lägenheter i studien ligger i Stockholms län. Vi valde även att omvandla **SåldDatum** till **SåldSäsong** istället genom att kategorisera dem under kategorierna Vår (3-5), Sommar(6-8), Höst(9-11) och Vinter(12,1,2) baserat på vilken månad de såldes.

Försäljarens namn ändrade vi till variabeln **Mäklare** då båda typerna av säljare kan räknas som mäklare, då de också kan erbjuda hjälp med försäljningar. I denna kategori förekom det också flertal små mäklarfirmor, vi valde därför att sammanfoga alla mäklarfirmor som sålt under 100 lägenheter till kategorin "Mindre Mäklarfirmor" enligt teorin i kapitel 2.1.4. Enligt samma teori (kapitel 2.1.4) sammanfogade vi även kategorierna i **ByggÅr** och detsamma med kategorierna i **Våning**.

I analysen nedan (kapitel 4.2) påvisade vi även att variablerna Area, Listpris, Hyra och Rum är högt korrelaterade med varandra och till responsvariabeln **SåldPris**. Vi valde därför att göra om dessa till de nya variablerna **AreaPerRum**, **PrisPerKvm** och **HyraPerKvm** med hjälp av den kategoriska variabeln **Rum** som förblir likadan.

Slutligen delar vi upp datat i två delar, alla lägenheter som såldes år 2022 (1001 stycken) ska gå till test-data så att man kan testa modellens prediktionsförmåga, resterande lägenheter som såldes innan 2022 (10287 stycken) blir tränings-data som används till att bygga modellerna.

3.3 Slutgiltigt data

Det dataset som slutligen kommer att användas i våra modeller innehåller denna information:

Mäklare: Vilken mäklare som sålde lägenheten.

Rum: Antalet rum lägenheten innehåller, kategorierna som finns är 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5, 6, och 7 rum.

ByggÅr: Året lägenheten byggdes, organiserad i kategorierna Byggår innan 1890, 1891-1899, 1900-1960, 1961-1970, 1971-1980, 1981-1990, 1991-2010, 2011-2022 och Okänd.

Kommun: Vilken Kommun i Norrort som lägenheten finns i, kategorierna är Danderyd, Sollentuna, Täby, Upplands Väsby, Vallentuna, Österåker och Vaxholm.

AvståndVattenMeter: Lägenhetens avstånd till närmsta sjö eller hav, mätt i m, varierar mellan 15-10000m.

Våning: Vilken våning som lägenheten ligger på, organiserad i kategorierna Våning under 2, 2 till 6, 7 till 11, 12 till 15, Över 15 och Okänd.

SåldSäsong: Vilket säsong som lägenheten såldes i, kategorierna är Vår, Sommar, Höst och Vinter.

AreaPerRum: Kvoten mellan **Boarea** och **Rum**, varierar mellan 9.5-81.

PrisPerKvm: Kvoten mellan **Listpris** och **Boarea**, varierar mellan 4090-142353.

HyraPerKvm: Kvoten mellan **Hyra** och **Boarea**, varierar mellan 2.9-138.2.

SåldPris: Vilket pris lägenheten såldes för, varierar mellan 225000-18000000kr.

4 Analys

I detta kapitel kommer vi att använda oss av teorierna i kapitel 2. Vi kommer att gå igenom all analys av data som vi använde oss av, för att få fram det slutgiltiga datasetet som presenterades i kapitel 3.3 och därefter bygga olika modeller och jämföra dessa för att sedan presentera vilken modell som är bäst anpassad för prediktion samt den mest förklarande modellen i nästa kapitel.

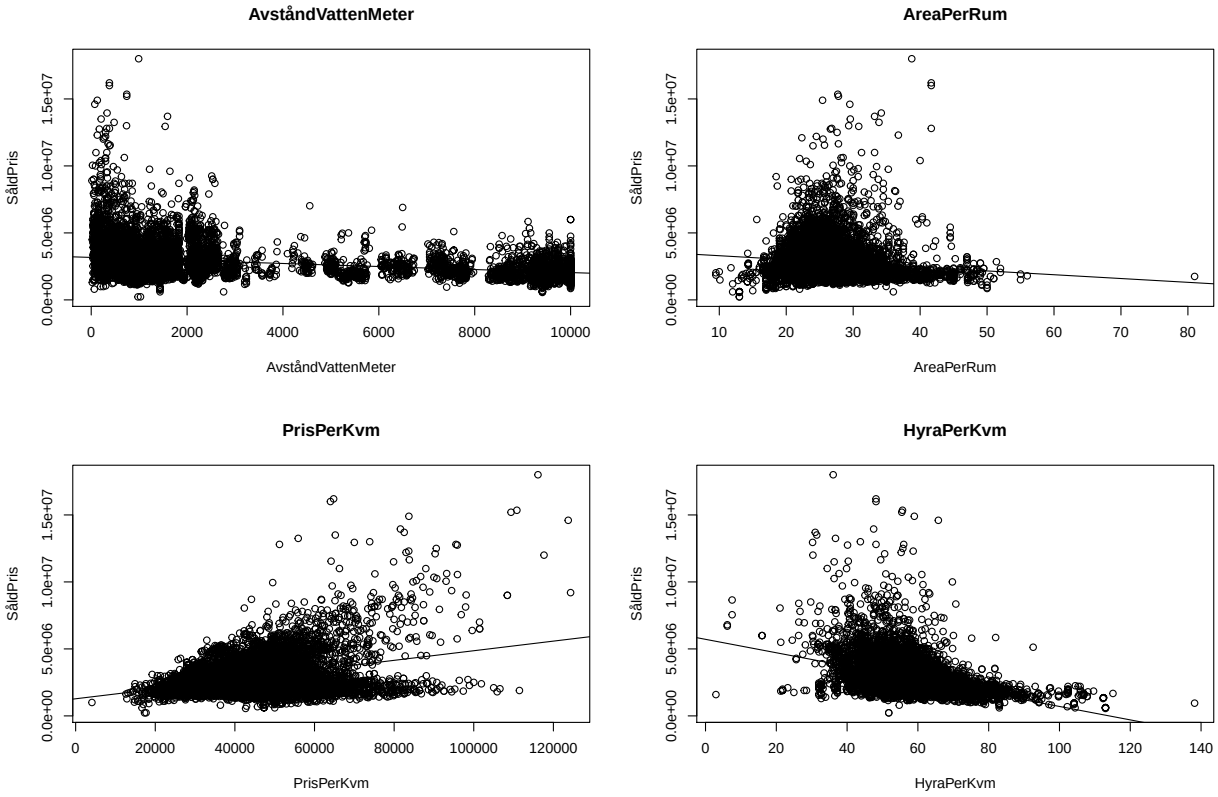
4.1 Enkel linjär regression

Vi börjar med att skapa enkla linjära modeller, för att undersöka de enskilda sambanden mellan respons och förklarande variablerna. Detta för att enklare kunna avgöra om regressionsmodellens parametrar är linjära, kontrollera om de är signifikanta och avgöra om det behövs en transformation till den multipla modellen.

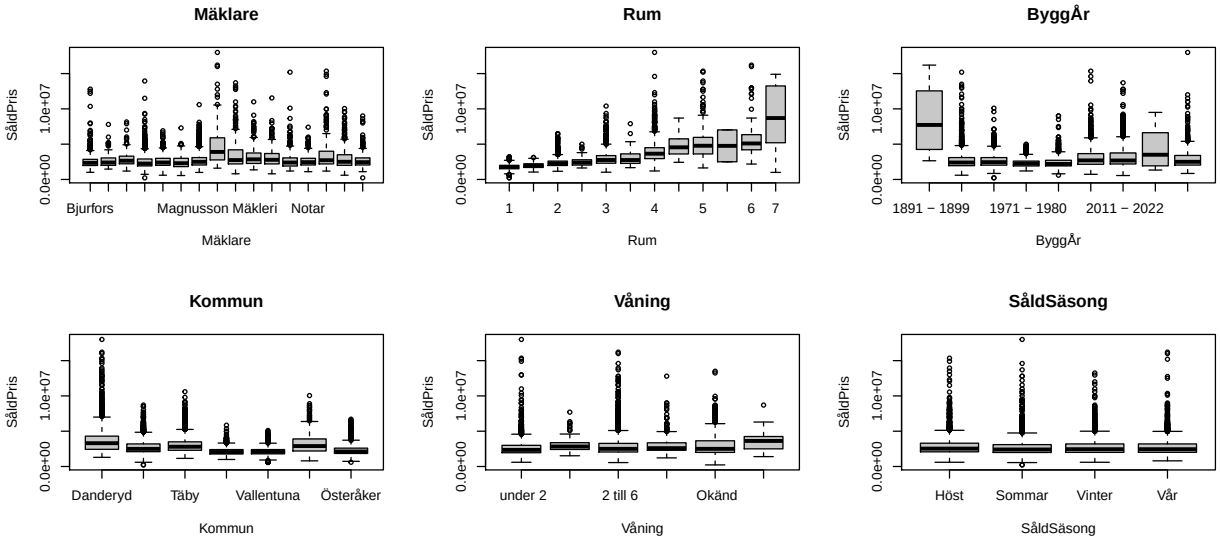
De fullständiga resultaten av våra enkla linjära modeller finns i Appendix under GitHub länken. I denna del går vi igenom Appendix kapitel 6.1. Där presenteras det att variablerna **Mäklare**, **Byggår** och **Våning** hade många icke signifikanta kategorier och med hjälp av sammanfogning (kapitel 2.1.4) så kunde man förbättra signifikansen i kategorierna avsevärt mycket. Här nedan sammanfattar vi modellerna för de slutgiltiga datasetet (kapitel 3.3) i en överskådlig tabell samt plottar ut modellerna:

Tabell 1: Sammanfattning av de slutgiltiga enkla linjära modellerna

Förklarande Variabler	R^2_{adj}	p-värde
Mäklare	0.07843	$< 2.2e-16$
Rum	0.4175	$< 2.2e-16$
ByggÅr	0.0705	$< 2.2e-16$
Kommun	0.1604	$< 2.2e-16$
AvståndVattenMeter	0.09269	$< 2.2e-16$
Våning	0.01073	$< 2.2e-16$
SåldSäsong	0.00146	0.0004349
AreaPerRum	0.01235	$< 2.2e-16$
PrisPerKvm	0.1387	$< 2.2e-16$
HyraPerKvm	0.171	$< 2.2e-16$



Figur 1: Scatterplottar av de slutliga kontinuerliga variablerna mot responsvariabeln



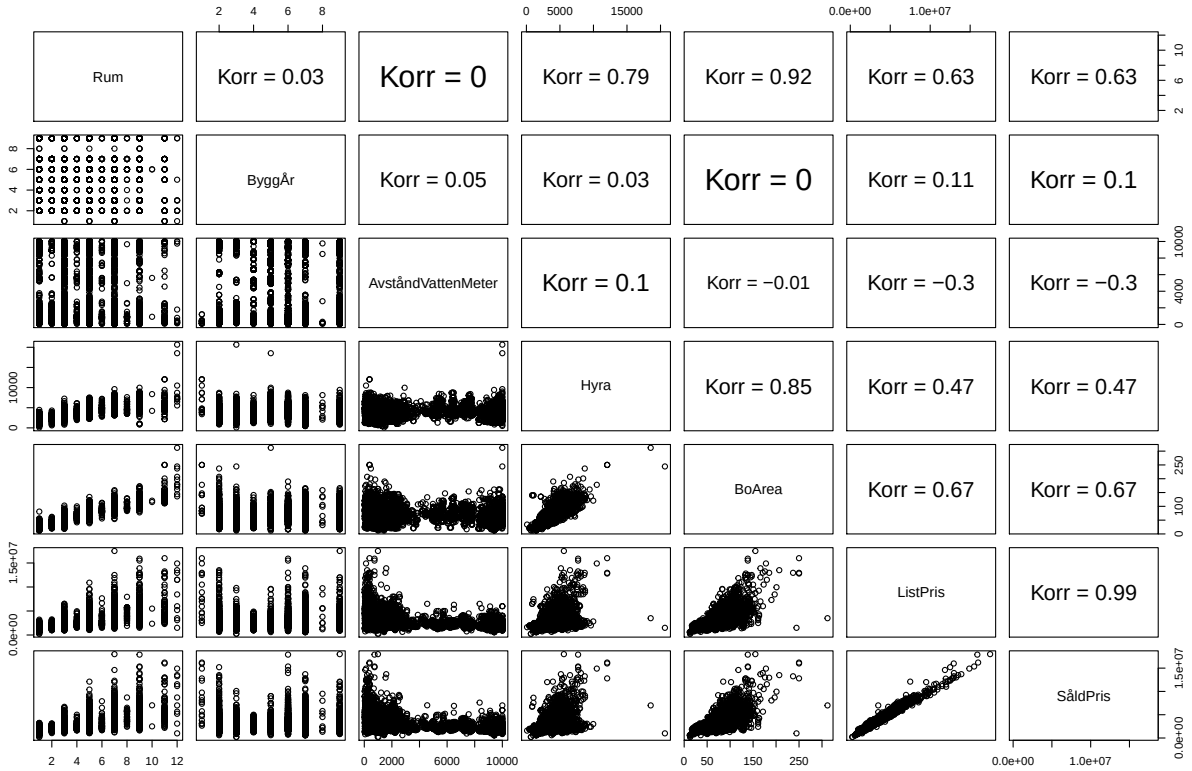
Figur 2: Boxplottar av de slutliga kategoriska variablerna mot resonsvariabeln

Från *Tabell 1* så kan man dra slutsatsen att **Rum** har störst påverkan av de förklarande variablerna om man jämför individuellt, därefter kommer **HyraPerKvm** och **PrisPerKvm**. Vi kan även notera från *Figur 1*, *Figur 2* samt residualplottarna i Appendix (kapitel 6.1.7 & GitHub) att parametrarna är linjära förutom **AreaPerRum**, **PrisPerKvm** samt **HyraPerKvm** som visar en konform i sin residualplott, vilket oftast innebär att en log transformation är lämplig att använda. Denna information tar vi med oss till konstruktionen av de multipla linjära modellerna då det finns en möjlighet att modellen bättras av en transformation.

4.2 Undersökning av korrelation

Innan man gör en multipel linjär regression så måste man kontrollera villkoret om avsaknad av linjära kombinationer (kapitel 2.1.1). Korrelationen kan användas för att ta reda på om de förklarande variablerna är linjärt oberoende av varandra, beräkningarna fås fram med hjälp av teorin i kapitel 2.1.7 och 2.1.8.

Vi börjar med att undersöka data efter sammanfogning men innan omvandling av variablerna **Area**, **Listpris**, **Hyra** och **Rum**.



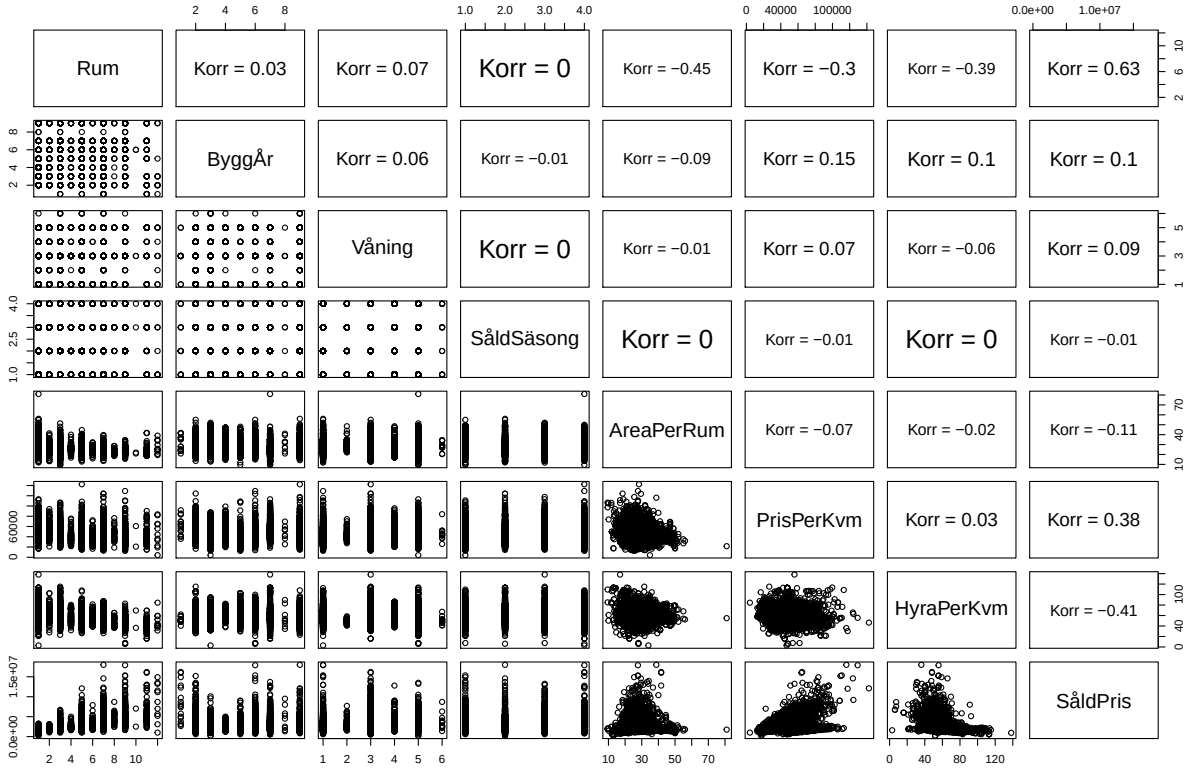
Figur 3: Korrelationsplott, Scatterplott mellan variablerna i nedre triangeln och korrelationskoefficienten i övre triangeln.

	VIF-värde
Rum	8.787999
ByggÅr	1.696282
AvståndVattenMeter	1.239940
Hyra	4.589016
BoArea	12.101674
ListPris	2.793636
Våning	1.152692
SåldSäsong	1.009978

Enligt korrelationsplotten i *Figur 3* samt VIF-värdena i tabellen ovan så kan man konstatera att variablerna **Area**, **Listpris**, **Hyra** och **Rum** är linjärt beroende av varandra. Detta leder oss till att omvandla dessa variabler då den unika informationen de innehåller är något vi vill behålla. Vi skapar därför variablerna

AreaPerRum, **PrisPerKvm** och **HyraPerKvm**, istället för **Area**, **Listpris**, **Hyra**. Vi behåller **Rum** som en "nyckel" för att kunna få fram informationen från de omvandlade variablerna.

Vi undersöker korrelationen på nytt med den nya omvandlingen av data.



Figur 4: Korrelationsplott efter omvandling, Scatterplott mellan variablerna i nedre triangeln och korrelationskoefficienten i övre triangeln.

	VIF-värde
Rum	3.120653
ByggÅr	1.638361
AvståndVattenMeter	1.329016
Våning	1.155942
SåldSäsong	1.011069
PrisPerKvm	2.089444
AreaPerRum	1.772787
HyraPerKvm	1.484015

Man kan enligt *Figur 4* och VIF-tabellen, konstatera att korrelationen minskat avsevärt mycket och det finns inga linjära beroenden mellan de förklarande variablerna. Då VIF-värdet blev lägre än 5 för de numeriska förklarande variablerna. Vilket innebär att man kan gå vidare med att skapa en multipel linjär modell.

4.3 Modeller

Här skapar vi den första modellen med hjälp av det slutgiltiga datasetet som presenterades i kapitel 3.3 och får då en modell på denna form,

Modell 1:

$$\begin{aligned} Såldpris_j = \alpha + \beta_1 Mäklare + \beta_2 Rum + \beta_3 ByggÅr + \beta_4 Kommun + \beta_5 AvståndVattenMeter + \\ \beta_6 Våning + \beta_7 SåldSäsong + \beta_8 AreaPerRum + \beta_9 PrisPerKvm + \beta_{10} HyraPerKvm + \epsilon_j \end{aligned} \tag{22}$$

Till en början verkar denna modell vara en bra modell, alla undersökningar hittas i Appendix kapitel 6.2.1 och GitHub. Enligt summary utskriften är modellen signifikant och visar ett högt R^2_{adj}-värde (0.9177) men när man undersöker nogrannare, så ser man på residualplotten och QQ-plotten att kraven för homoskedasticitet samt normalitet (kapitel 2.1.1), inte är uppfyllda. Vi kan även se att vi har ett flertal outliers, där studentized residuals > 3 (kapitel 2.2.1).

Vi valde att ta bort dessa outliers (172st), för att se om modellen skulle uppfylla villkoren. Resultaten finns i Appendix kapitel 6.2.2 och GitHub. Det visar sig att Modell 1, inte uppfyller kraven för homoskedasticitet men var på god väg att uppfylla kraven för normalitet.

Vi gick sedan vidare med att testa metoden att transformera modellen (kapitel 2.2.2) och tog därefter fram fyra ytterliggare modeller.

Modell 2:

$$\begin{aligned} log(Såldpris_j) = \alpha + \beta_1 Mäklare + \beta_2 Rum + \beta_3 ByggÅr + \beta_4 Kommun + \beta_5 AvståndVattenMeter + \\ \beta_6 Våning + \beta_7 SåldSäsong + \beta_8 AreaPerRum + \beta_9 PrisPerKvm + \beta_{10} HyraPerKvm + \epsilon_j \end{aligned} \tag{23}$$

Modell 3:

$$\begin{aligned} log(Såldpris_j) = \alpha + \beta_1 Mäklare + \beta_2 Rum + \beta_3 ByggÅr + \beta_4 Kommun + \beta_5 log(AvståndVattenMeter) + \\ \beta_6 Våning + \beta_7 SåldSäsong + \beta_8 log(AreaPerRum) + \beta_9 log(PrisPerKvm) + \beta_{10} log(HyraPerKvm) + \epsilon_j \end{aligned} \tag{24}$$

Modell 4:

$$\begin{aligned} log(Såldpris_j) = \alpha + \beta_1 Mäklare + \beta_2 Rum + \beta_3 ByggÅr + \beta_4 Kommun + \beta_5 AvståndVattenMeter + \\ \beta_6 Våning + \beta_7 SåldSäsong + \beta_8 log(AreaPerRum) + \beta_9 log(PrisPerKvm) + \beta_{10} log(HyraPerKvm) + \epsilon_j \end{aligned} \tag{25}$$

Modell 5:

$$\begin{aligned} log(Såldpris_j) = \alpha + \beta_1 Mäklare + \beta_2 Rum + \beta_3 ByggÅr + \beta_4 Kommun + \beta_5 AvståndVattenMeter + \\ \beta_6 Våning + \beta_7 SåldSäsong + \beta_8 log(AreaPerRum) + \beta_9 log(PrisPerKvm) + \beta_{10} HyraPerKvm + \epsilon_j \end{aligned} \tag{26}$$

Det kompletta resultatet för dessa modeller finns i GitHhub-länken i Appendix. Sammanfattningsvis så kan man säga att Modell 1 och 2 inte uppfyllde kraven för en multipel linjär regression, men Modell 3, 4 och 5 var väldigt starka modeller med marginella skillnader mellan varandra och uppfyllde alla krav i kapitel 2.1.1.

Vi applicerar metoden för stegvis variabelselektion på dessa modeller för att undersöka om de går att få ännu bättre modeller. Därefter har vi en översiktlig genomgång på modellerna.

4.4 Modeller efter stegvis variabelselektion

I denna sektion så utför vi stegvis variabelselektion (kapitel 2.2.5) på de tidigare modellerna Modell 1-5. Vi kommer använda oss av stoppkriteriumet *AIC* (kapitel 2.2.4.2), för att försöka få fram en så bra prediktionsmodell som möjligt. Både Bakåt-Metoden och Framåt-Metoden kommer användas med tillägget Stegvis-selektion och det bästa resultatet, utifrån stoppkriteriumet kommer plockas ut av varje modell. De nya modellerna definieras som:

Modell 1 Stepwise:

$$\begin{aligned} Såldpris_j = \alpha + \beta_2 Rum + \beta_3 ByggÅr + \beta_4 Kommun + \beta_6 Våning + \\ \beta_8 AreaPerRum + \beta_9 PrisPerKvm + \beta_{10} HyraPerKvm + \epsilon_j \end{aligned} \tag{27}$$

Modell 2 Stepwise:

$$\begin{aligned} log(Såldpris_j) = \alpha + \beta_1 Mäklare + \beta_2 Rum + \beta_3 ByggÅr + \beta_4 Kommun + \beta_5 AvståndVattenMeter + \\ \beta_6 Våning + \beta_8 AreaPerRum + \beta_9 PrisPerKvm + \beta_{10} HyraPerKvm + \epsilon_j \end{aligned} \tag{28}$$

Modell 3 Stepwise, **Modell 4 Stepwise** och **Modell 5 Stepwise** döps om till

Modell Stepwise:

$$\begin{aligned} log(Såldpris_j) = \alpha + \beta_2 Rum + \beta_3 ByggÅr + \beta_4 Kommun + \beta_6 Våning + \\ \beta_8 log(AreaPerRum) + \beta_9 log(PrisPerKvm) + \epsilon_j \end{aligned} \tag{29}$$

Det första man kan observera är att Modell 3, 4 och 5 blir samma modell efter stegvis variabelselektion, vi väljer därför att kalla den Modell Stepwise. En annan observation är att Modell 1 Stepwise och 2 Stepwise fortfarande inte uppfyller kraven för en multipel linjär regression men behålls för jämförelse. Summary och plottar finns i Appendix (Github) men vi väljer att sammanfatta summary output från alla våra modeller i tabellen nedan:

Tabell 4: Sammanfattning av summary output från alla modeller

Modell	Residual Standard Error	R^2	R^2_{adj}	F-statistic	p-värde
Modell 1	167800 on 9308	0.962	0.9619	5243 on 45 and 9308	$< 2.2e-16$
Modell 2	0.04784 on 9308	0.9755	0.9754	8230 on 45 and 9308	$< 2.2e-16$
Modell 3	9.912e-11 on 9308	1	1	1.965e+21 on 45 and 9308	$< 2.2e-16$
Modell 4	9.911e-11 on 9308	1	1	1.966e+21 on 45 and 9308	$< 2.2e-16$
Modell 5	9.911e-11 on 9308	1	1	1.966e+21 on 45 and 9308	$< 2.2e-16$
Modell 1 Stepwise	167900 on 9327	0.9619	0.9618	9062 on 26 and 9327	$< 2.2e-16$
Modell 2 Stepwise	0.04784 on 9311	0.9755	0.9754	8817 on 42 and 9311	$< 2.2e-16$
Modell Stepwise	9.911e-11 on 9328	1	1	3.538e+21 on 25 and 9328	$< 2.2e-16$

Från *Tabell 4* och anpassningsmåtten i kapitel 2.2.3 så kommer vi fram till att den mest förklarande modellen är Modell Stepwise med väldigt liten marginal från Modell 4 och 5, då det avgörande var att Model Stepwise hade högre värde på F-statistikan som i sin tur betyder mer signifikant modell (kapitel 2.2.3.3). Vidare vill vi undersöka modellernas predikterande förmåga.

4.5 Prediktion

För att undersöka modellernas predikterande förmåga så använder vi oss av test-data som består av 1001 stycken lägenheter som såldes år 2022 (kapitel 3.2). Vi sätter in test-data i modellerna och låter den sedan skatta vilket pris lägenheten såldes för, detta jämför vi med den observerade responsvariabeln **SåldPris** som vi har i datasetet. I Appendix (kapitel 6.4) finns jämförelsena mellan den observerade och den skattade responsvariabeln i form av en scatterplott och en plott av skillnaden mellan dem mot index. Vi använder oss av de predikterade och observerade värdena för att beräkna MSEP, RMSEP och PRESS enligt kapitel 2.2.4.1 för att kunna avgöra vilken modell som har bäst prediktionsförmåga. Nedan visas en tabell med de uträknade värdena:

Tabell 5: Prediktionsmåtten för att avgöra modell med starkast prediktionsförmåga.

	MSEP	RMSEP	PRESS (N = 1001)
Modell 1	1.786379e+11	4.226557e+05	1.756010e+14
Modell 2	1.151929e+12	1.073280e+06	1.132347e+15
Modell 3	1.000000e-07	3.655000e-04	1.313000e-04
Modell 4	1.000000e-07	3.651000e-04	1.310000e-04
Modell 5	1.000000e-07	3.647000e-04	1.308000e-04
Modell 1 Stepwise	1.793244e+11	4.234671e+05	1.762758e+14
Modell 2 Stepwise	1.155800e+12	1.075081e+06	1.136152e+15
Modell Stepwise	1.000000e-07	3.681000e-04	1.332000e-04

Från *Tabell 5* och kapitel 2.2.4.1 kan vi avgöra vilken modell som har starkast prediktionsförmåga genom att se vilken modell som har de minsta värdena. Vi kan se att Modell 5 har starkast predikterande förmåga med små marginaler från Modell 4 och 3, Modell Stepwise som hade störst förklarande förmåga var den sämre av de modellerna som uppfyller villkoren för linjära modeller. Det är normalt att sådant kan hända då man byter dataset, Modell Stepwise var bäst anpassad till dåvarande data men den predikterande förmågan testar egentligen förmågan i mer slumpad typ av data.

5 Diskussion

I denna sektion kommer vi att redovisa de resultat och besvara frågor samt syftet med denna studie. Vi kommer även att gå igenom de förklarande variablernas påverkan mot responsvariabeln, vilket ger en kännedom om vad man ska tänka på när man letar efter/säljer lägenheter.

5.1 Resultat

Från Analysen (kapitel 4) så kunde man dra slutsatsen av att Modell 5 hade bäst predikterande förmåga, Modell Stepwise hade bäst förklarande förmåga men sämre predikterande förmåga. Modell 5 har utöver starkast predikterande förmåga även en stark förklarande förmåga, den tar även hänsyn till fler förklarande variabler och bör därför anses va den mest lämpliga modellen i denna studie.

Modell 5 fås av ekvationen:

$$\begin{aligned} log(Såldpris_j) = \alpha + \beta_1 Mäklare + \beta_2 Rum + \beta_3 ByggÅr + \beta_4 Kommun + \beta_5 AvståndVattenMeter + \\ \beta_6 Våning + \beta_7 SåldSäsong + \beta_8 log(AreaPerRum) + \beta_9 log(PrisPerKvm) + \beta_{10} HyraPerKvm + \epsilon_j \end{aligned} \tag{30}$$

För att undersöka påverkan på responsvariabeln vill vi bryta ut **SåldPris** och måste därför återgå till en multiplikativ modell och får då ekvationen:

$$\begin{aligned} Såldpris_j = e^{\alpha} * e^{\beta_1 Mäklare} * e^{\beta_2 Rum} * e^{\beta_3 ByggÅr} * e^{\beta_4 Kommun} * e^{\beta_5 AvståndVattenMeter} * \\ e^{\beta_6 Våning} * e^{\beta_7 SåldSäsong} * e^{\beta_8 log(AreaPerRum)} * e^{\beta_9 log(PrisPerKvm)} * e^{\beta_{10} HyraPerKvm} * e^{\epsilon_j} \end{aligned} \tag{31}$$

Med hjälp av Estimate kolumnen från summary som finns i kapitel 6.2.3 så kan vi få fram koefficienten framför den förklarande variabeln och på så sätt även avgöra vilken påverkan den förklarande variabeln har på responsvariabeln. Vi går igenom ekvationen stegvis och börjar med interceptet (α) och får då koefficienten $e^{\hat{\alpha}} = e^{1.802e-11} \approx 1.00000000001802$ vilket kan tolkas som bas av slutpriset när alla förklarande variabler har värdet noll, detta kommer dock aldrig kunna hända bland lägenheter då man inte kan sälja en lägenhet med t.ex 0 rum eller 0 i area.

Vi börjar nu undersöka de förklarande variablernas koefficienter:

Mäklare: Mäklarfirman Bjurfors har använts som en baslinje och har därför ingått i interceptet. Det innebär att den används som en standardpunkt och att procentuella ökningen/minskningen baseras på om man byter mäklarfirma. Eftersom **Mäklare** är kategorisk blir det enklare att visa i de olika $\hat{\beta}$ för **Mäklare** i en tabell:

Tabell 6: Mäklare, $\hat{\beta}$ och $e^{\hat{\beta}} - 1$ som den procentuella utvecklingen jämfört med Bjurfors

Mäklare	$\hat{\beta}$	%-utveckling
ERA	4.588e-12	4.588e-10
Erik Olsson Fastighetsförmedling	-1.461e-12	-1.461e-10
Fastighetsbyrån	4.058e-12	4.058e-10
HusmanHagberg	7.669e-12	7.669e-10
Jägholm Norrortsmäklarna	-1.274e-11	-1.274e-09
Länsförsäkringar Fastighetsförmedling	2.306e-12	2.306e-10
Magnusson Mäkleri	2.199e-12	2.199e-10
Mindre Mäklarfirmor	2.493e-12	2.493e-10
MOHV	8.472e-13	8.472e-11
Mäklarhuset	5.899e-12	5.899e-10
Mäklarringen	4.852e-12	4.852e-10
Notar	-4.758e-12	-4.758e-10
SkandiaMäklarna	-9.192e-12	-9.192e-10
Svensk Fastighetsförmedling	2.903e-12	2.903e-10
Svenska Mäklarhuset	-1.165e-12	-1.165e-10

Man kan tolka resultatet som att om man byter mäklare från Bjurfors till exempelvis ERA så säljs lägenheten för $(4.588 * 10^{-10})\%$ dyrare. Mäklare har inte en jätte stor påverkan men det finns skillnader, de har även en korrelation till **ListPris** vilket kan påverka responsvariabeln. Bästa mäklaren att välja enligt denna studie är HusmanHagberg som säljer lägenheten $(7.669 * 10^{-10})\%$ dyrare än baslinjen.

Rum: Rum1 har använts som en baslinje och har därför ingått i interceptet. Det innebär att den används som en standardpunkt och att procentuella ökningen/minskningen baseras på om man ökar antalet Rum. Eftersom **Rum** är kategorisk blir det enklare att visa i de olika $\hat{\beta}$ för **Rum** i en tabell:

Tabell 7: Rum, $\hat{\beta}$ och $e^{\hat{\beta}} - 1$ som den procentuella utvecklingen jämfört med Rum1

Rum	$\hat{\beta}$	%-utveckling
Rum1.5	0.4055	50.0052
Rum2	0.6931	99.9906
Rum2.5	0.9163	150.002
Rum3	1.099	200.116
Rum3.5	1.253	250.083
Rum4	1.386	299.882
Rum5	1.609	399.781

Vi noterar att för varje nivåökning på 0.5 i kategorierna **Rum** så ökar priset med ca 50%.

ByggÅr: kategorin "1900 - 1960" har använts som en baslinje och har därför ingått i interceptet. Det innebär att den används som en standardpunkt och att procentuella ökningen/minskningen baseras på om man köper nyare lägenheter. Eftersom **ByggÅr** är kategorisk blir det enklare att visa i de olika $\hat{\beta}$ för **ByggÅr** i en tabell:

Tabell 8: ByggÅr, $\hat{\beta}$ och $e^{\hat{\beta}} - 1$ som den procentuella utvecklingen jämfört med årsspannet 1900 - 1960

ByggÅr	$\hat{\beta}$	%-utveckling
ByggÅr 1961 - 1970	9.401e-12	9.401e-10
ByggÅr 1971 - 1980	8.782e-12	8.782e-10
ByggÅr 1981 - 1990	-3.528e-12	-3.528e-10
ByggÅr 1991 - 2010	8.545e-12	8.545e-10
ByggÅr 2011 - 2022	3.499e-12	3.499e-10
ByggÅr Okänd	2.503e-13	2.503e-11

Vi noterar att skillnaden i pris är inte så stor mellan årsspannen.

Kommun: kategorin Danderyd har använts som en baslinje och har därför ingått i interceptet. Det innebär att den används som en standardpunkt och att procentuella ökningen/minskningen baseras på om man ändrar kommun. Eftersom **Kommun** är kategorisk blir det enklare att visa i de olika $\hat{\beta}$ för **Kommun** i en tabell:

Tabell 9: Kommun, $\hat{\beta}$ och $e^{\hat{\beta}} - 1$ som den procentuella utvecklingen jämfört med Danderyds kommun

Kommun	$\hat{\beta}$	%-utveckling
Sollentuna	-1.723e-11	-1.723e-09
Täby	-1.873e-11	-1.873e-09
Upplands Väsby	-3.818e-12	-3.818e-10
Vallentuna	-5.438e-12	-5.438e-10
Vaxholm	-3.214e-11	-3.214e-09
Österåker	-1.914e-11	-1.914e-09

Vi noterar att skillnaden i pris är inte så stor mellan Kommunerna men däremot är alla nedåtgående.

AvståndVattenMeter: Den kontinuerliga variabeln **AvståndVattenMeter** har en koefficient som är $e^{-1.219e-15}$ den procentuella utvecklingen fås då av $(e^{-1.219e-15} - 1) * 100 = -1.219 * 10^{-13}\%$. Detta innebär att för varje meter som **AvståndVattenMeter** ökar så minskar lägenhets priset med $-1.219 * 10^{-13}\%$.

Våning: kategorin Våning under 2 har använts som en baslinje och har därför ingått i interceptet. Det innebär att den används som en standardpunkt och att procentuella ökningen/minskningen baseras på om man köper en lägenhet på en högre våning än 1. Eftersom **Våning** är kategorisk blir det enklare att visa i de olika $\hat{\beta}$ för **Våning** i en tabell:

Tabell 10: Våning, $\hat{\beta}$ och $e^{\hat{\beta}} - 1$ som den procentuella utvecklingen jämfört med Våningarna under 2

Våning	$\hat{\beta}$	%-utveckling
Våning 2 till 6	7.227e-12	7.227e-10
Våning 7 till 11	1.157e-11	1.157e-09
Våning 12 till 15	1.518e-11	1.518e-09
Våning Okänd	5.592e-12	5.592e-10

Vi noterar att skillnaden i pris är inte så stor mellan Våningarna men att priset är uppåtgående.

SåldSäsong: kategorin Höst har använts som en baslinje och har därför ingått i interceptet. Det innebär att den används som en standardpunkt och att procentuella ökningen/minskningen baseras på om man köper en lägenhet under en annan säsong än hösten. Eftersom **SåldSäsong** är kategorisk blir det enklare att visa i de olika $\hat{\beta}$ för **SåldSäsong** i en tabell:

Tabell 11: SåldSäsong, $\hat{\beta}$ och $e^{\hat{\beta}} - 1$ som den procentuella utvecklingen jämfört med köp under hösten

SåldSäsong	$\hat{\beta}$	%-utveckling
Sommar	-3.769e-12	-3.769e-10
Vinter	-8.264e-13	-8.264e-11
Vår	2.516e-12	2.516e-10

Vi noterar att skillnaden i pris är inte så stor mellan Säsongerna men resultet tyder på att det är mest givande att sälja under Våren och köpa under Sommaren.

log(AreaPerRum): Den kontinuerliga variabeln **log(AreaPerRum)** har en koefficient som är e^1 den procentuella utvecklingen fås då av $(e-1)*100 \approx 171.8282\%$. Detta innebär att för varje heltalsenhet som **log(AreaPerRum)** ökar så ökar lägenhets priset med ca 171.8282%. Observera att denna tolkning gäller då logaritmen av AreaPerRum ökar med ett heltal och inte AreaPerRum.

log(PrisPerKvm): Den kontinuerliga variabeln **log(PrisPerKvm)** har en koefficient som är e^1 den procentuella utvecklingen fås då av $(e-1)*100 \approx 171.8282\%$. Detta innebär att för varje heltalsenhet som **log(PrisPerKvm)** ökar så ökar lägenhets priset med ca 171.8282%. Observera att denna tolkning gäller då logaritmen av PrisPerKvm ökar med ett heltal och inte PrisPerKvm.

HyraPerKvm: Den kontinuerliga variabeln **HyraPerKvm** har en koefficient som är $e^{3.035*10^{-15}}$ den procentuella utvecklingen fås då av $(e^{3.035*10^{-15}} - 1)*100 = 3.035*10^{-13}\%$. Detta innebär att för varje heltalsenhet som **HyraPerKvm** ökar så ökar lägenhets priset med $3.035*10^{-13}\%$.

En slutsats av resultatet är att variabeln **Rum** har störst positiv påverkan på responsvariabeln **SåldPris**, utöver det, om man antar att de kontinuerliga variablerna ökar så gäller det att:

Variablerna **log(PrisPerKvm)**, **log(AreaPerRum)**, **HyraPerKvm** och **Våning** har en enbart positiv påverkan på **SåldPris**. Variablerna **Kommun** och **AvståndVattenMeter** har en negativ effekt på **SåldPris** och variablerna **Mäklare**, **ByggÅr** och **SåldSäsong** har en mixad påverkan.

5.2 Andra analyser och begränsningar

Booli erbjöd större mängder data men med mycket saknad av innehåll, vilket ledde till att man va tvungen att ta bort en hel del data som hade kunnat hjälpa till för en bättre modell. Även saknaden av data som skulle kunna påverka priset, exempelvis variablerna balkong och hiss skulle kunna bidra till en bättre modell och en mer överskådlig bild av vad mer som påverkar lägenhetspriserna. Det noterades även att enligt denna modell, så har en höjning av **HyraPerKvm** en positiv påverkan på responsvariabeln **SåldPris**, det må stämma enligt data, men i verkligheten är det inte attraktivt att antingen höja **Hyra** eller minska **BoArea**, eftersom det inte är en stor positiv påverkan och **HyraPerKvm** består av en kvot, så gissningsvis har det varit lite högre hyror på de lägenheter som sålts för mer pengar.

6 Appendix

För den intresserade så finns alla utskrifter för hela arbetet i Github på länken: https://github.com/RamtinGolrang/Regressionsanalys-av-l-genhetspriser-i-Stockholms-Norrort

6.1 Enkel linjär regression

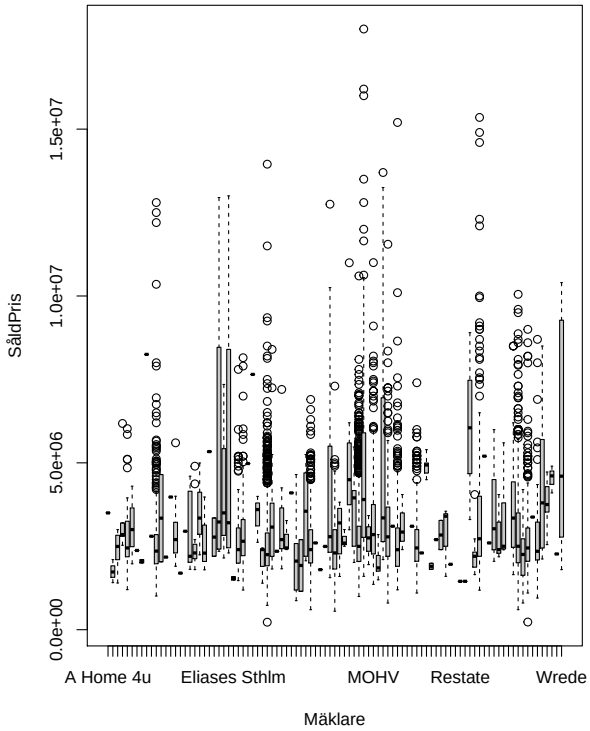
6.1.1 Mäklare innan sammanfogning:

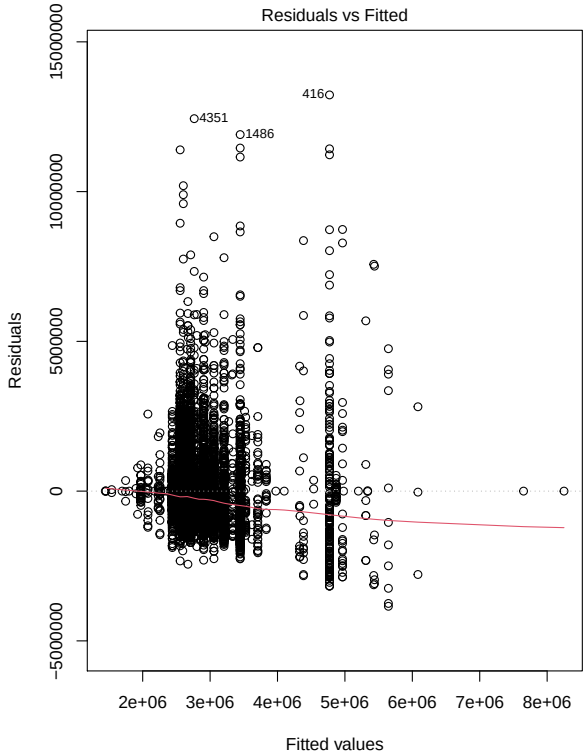
```
##
## Call:
## lm(formula = SåldPris ~ Mäklare, data = data_träning)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3845000  -712051  -262051   411878 13228417
##
## Coefficients:
##                                                Estimate Std. Error t value
## (Intercept)                                     3500000    1214651   2.881
## MäklareAFI FastighetsMäklare                   -1753333    1402558  -1.250
## MäklareAlertus Fastighetsbyrå AB               -1071875    1252034  -0.856
## MäklareAlexander White                            12000    1330584   0.009
## MäklareAlicia Edelman Fastighetsmäkleri         -634839    1234087  -0.514
## MäklareAmbassadör Fastighetsmäkleri             -408333    1402558  -0.291
## MäklareAndersson & Asplund Mäklarbyrå Stockholm -1125000    1717776  -0.655
## MäklareAsira                                   -1450000    1487638  -0.975
## MäklareBerggren Hörle                           4750000    1717776   2.765
## MäklareBest of Homes                            -700000    1717776  -0.408
## MäklareBjurfors                                 -897289    1215618  -0.738
## MäklareBlok                                     -157500    1487638  -0.106
## MäklareBlumenthalHoffman Fastighetsmäkleri     -1325000    1717776  -0.771
## MäklareBostadsrättsspecialisten                  475000    1717776   0.277
## MäklareBOSTHLM                                  -598095    1243235  -0.481
## MäklareBototal                                 -1805000    1717776  -1.051
## MäklareBremberg Fastighetsmäkleri               -550000    1717776  -0.320
## MäklareBrokr Fastighetsmäklare AB               -670000    1311973  -0.511
## MäklareBronze Fastighetsförmedling AB           -981750    1244647  -0.789
## MäklareBällstaudde Bostadsutveckling AB          -32105    1246206  -0.026
## MäklareCredentia AB                             -811667    1402558  -0.579
## MäklareDerome Bostad                            1837500    1487638   1.235
## MäklareDiplomat Fastighetsmäkleri AB            -725000    1487638  -0.487
## MäklareEdward & Partners Fastighetsmäklare AB   1938750    1358021   1.428
## MäklareEkenstam Fastighetsmäklare                833333    1402558   0.594
## MäklareEklund Stockholm New York                1927500    1358021   1.419
## MäklareEliases Sthlm                           -1960000    1487638  -1.318
## MäklareERA                                      -823288    1219500  -0.675
## MäklareErik Olsson Fastighetsförmedling         -572912    1218465  -0.470
## MäklareESSTATE AB                               1480000    1717776   0.862
## MäklareEstate Fastighetsbyrå AB                 4150000    1717776   2.416
## MäklareFabergé Fastighetsmäkleri                 -95000    1402558  -0.068
## MäklareFantastic Frank Fastighetsmäkleri       -1418333    1402558  -1.011
## MäklareFastighetsbyrån                          -944313    1214983  -0.777
```

```
## MäklareFastighetsmäklarna                        -163519    1236940  -0.132
## MäklareGardefalk & Co                           -1150000    1717776  -0.669
## MäklareGrand Fastighetsförmedling                -177500    1288332  -0.138
## MäklareGripsholms Fastighetsförmedling           -793333    1402558  -0.566
## MäklareHans Mörner Fastighetsförmedling           600000    1717776   0.349
## MäklareHemverket                                -1422663    1264249  -1.125
## MäklareHilmkil Fastighetsförmedling             -1575000    1487638  -1.059
## MäklareHistoriska Hem AB                          -26111    1280355  -0.020
## MäklareHusmanHagberg                             -905464    1215541  -0.745
## MäklareHägerstens mäklare                        -900000    1717776  -0.524
## MäklareHögalidsmäklarna                         -1700000    1717776  -0.990
## MäklareIndivida Fastighetsmäkleri               -1005000    1717776  -0.585
## MäklareInnerstadsspecialisten AB                  885357    1257283   0.704
## MäklareJägholm Norrortsmäklarna                 -1060646    1217148  -0.871
## MäklareKarlsson & Uddare                         -546250    1358021  -0.402
## MäklareLagerlöfs Fastighetsmäkleri               -801667    1402558  -0.572
## MäklareLe Grand Propriété                        1810000    1298517   1.394
## MäklareLiving Fastighetsmäkleri                   122000    1330584   0.092
## MäklareLänsförsäkringar Fastighetsförmedling     -787949    1214904  -0.649
## MäklareMagnusson Mäkleri                         1271583    1218020   1.044
## MäklareMats Holmgren Fastighetsbyrå              -793333    1402558  -0.566
## MäklareMOHV                                      -293381    1216356  -0.241
## MäklareMustonen & Hedlund                       -1527031    1233484  -1.238
## MäklareMäklarfirma Vincent Forssbeck AB          1464400    1238706   1.182
## MäklareMäklarhuset                               -443456    1215841  -0.365
## MäklareMäklarMäster                              -400000    1717776  -0.233
## MäklareMäklarringen                              -735426    1216490  -0.605
## MäklareNomad Mäkleri                             -425000    1358021  -0.313
## MäklareNoside Fastighetsförmedling               -400000    1717776  -0.233
## MäklareNotar                                     -839730    1215860  -0.691
## MäklareNybergs Hem                              -1200000    1717776  -0.699
## MäklareOBOS                                      1368921    1230530   1.112
## MäklareOlovsson - Stignäs AB                    -1600000    1487638  -1.076
## MäklarePeab Bostad                               -800000    1487638  -0.538
## MäklarePrivatmäklaren                            -662500    1487638  -0.445
## MäklareProperties & Partners Fastighetsmäklare   -650000    1402558  -0.463
## MäklareReal Vision Fastighetsmäklare            -1540000    1717776  -0.897
## MäklareRestate                                  -2050000    1487638  -1.378
## MäklareRiksmäklaren                             -2050000    1717776  -1.193
## MäklareSjönära Fastigheter AB                    2583333    1402558   1.842
## MäklareSjöös Fastighetsförmedling               -1269000    1254486  -1.012
## MäklareSkandiaMäklarna                           -53875    1216452  -0.044
## MäklareSkeppsholmen                              1700000    1717776   0.990
## MäklareSmart fastighetsförmedling                -900000    1717776  -0.524
## MäklareStadsvillan Fastigheter                     27907    1228694   0.023
## MäklareStefan Blomdin AB                         -621667    1402558  -0.443
## MäklareSthlmFast                                 -223571    1298517  -0.172
## MäklareSusanne Persson Fastighetsförmedling AB    208333    1229026   0.170
## MäklareSvensk Fastighetsförmedling               -595443    1215378  -0.490
## MäklareSvenska Mäklargruppen                    -1244000    1244647  -0.999
## MäklareSvenska Mäklarhuset                       -829166    1215794  -0.682
## MäklareTradition Fastighetsmäkleri AB            -125000    1717776  -0.073
## MäklareUnik Fastighetsförmedling                 -696857    1222315  -0.570
## MäklareURBAN by ESNY                              828636    1268662   0.653
```

```
## MäklareViktor Hanson                                 332895    1246206   0.267
## MäklareVision Fastighetsmäkleri AB                  1036667    1402558   0.739
## MäklareWiderlöv & Co                               -1230000    1717776  -0.716
## MäklareWrede                                        2145000    1268662   1.691
##                                                    Pr(>|t|)
## (Intercept)                                         0.00397 **
## MäklareAFI FastighetsMäklare                        0.21129
## MäklareAlertus Fastighetsbyrå AB                    0.39196
## MäklareAlexander White                              0.99280
## MäklareAlicia Edelman Fastighetsmäkleri             0.60697
## MäklareAmbassadör Fastighetsmäkleri                 0.77095
## MäklareAndersson & Asplund Mäklarbyrå Stockholm     0.51254
## MäklareAsira                                        0.32973
## MäklareBerggren Hörle                               0.00570 **
## MäklareBest of Homes                                0.68365
## MäklareBjurfors                                     0.46045
## MäklareBlok                                         0.91569
## MäklareBlumenthalHoffman Fastighetsmäkleri          0.44052
## MäklareBostadsrättsspecialisten                     0.78215
## MäklareBOSTHLM                                      0.63047
## MäklareBototal                                      0.29339
## MäklareBremberg Fastighetsmäkleri                   0.74884
## MäklareBrokr Fastighetsmäklare AB                   0.60959
## MäklareBronze Fastighetsförmedling AB               0.43026
## MäklareBällstaudde Bostadsutveckling AB             0.97945
## MäklareCredentia AB                                 0.56280
## MäklareDerome Bostad                                0.21679
## MäklareDiplomat Fastighetsmäkleri AB                0.62602
## MäklareEdward & Partners Fastighetsmäklare AB       0.15343
## MäklareEkenstam Fastighetsmäklare                   0.55242
## MäklareEklund Stockholm New York                    0.15583
## MäklareEliases Sthlm                                0.18769
## MäklareERA                                          0.49963
## MäklareErik Olsson Fastighetsförmedling             0.63823
## MäklareESSTATE AB                                   0.38894
## MäklareEstate Fastighetsbyrå AB                     0.01571 *
## MäklareFabergé Fastighetsmäkleri                    0.94600
## MäklareFantastic Frank Fastighetsmäkleri            0.31192
## MäklareFastighetsbyrån                              0.43704
## MäklareFastighetsmäklarna                           0.89483
## MäklareGardefalk & Co                               0.50321
## MäklareGrand Fastighetsförmedling                   0.89042
## MäklareGripsholms Fastighetsförmedling              0.57166
## MäklareHans Mörner Fastighetsförmedling             0.72688
## MäklareHemverket                                    0.26049
## MäklareHilmkil Fastighetsförmedling                 0.28975
## MäklareHistoriska Hem AB                            0.98373
## MäklareHusmanHagberg                                0.45635
## MäklareHägerstens mäklare                           0.60034
## MäklareHögalidsmäklarna                             0.32237
## MäklareIndivida Fastighetsmäkleri                   0.55852
## MäklareInnerstadsspecialisten AB                    0.48134
## MäklareJägholm Norrortsmäklarna                     0.38355
## MäklareKarlsson & Uddare                            0.68752
```

```
## MäklareLagerlöfs Fastighetsmäkleri                0.56762
## MäklareLe Grand Propriété                         0.16338
## MäklareLiving Fastighetsmäkleri                   0.92695
## MäklareLänsförsäkringar Fastighetsförmedling      0.51663
## MäklareMagnusson Mäkleri                          0.29652
## MäklareMats Holmgren Fastighetsbyrå               0.57166
## MäklareMOHV                                       0.80941
## MäklareMustonen & Hedlund                         0.21575
## MäklareMäklarfirma Vincent Forssbeck AB           0.23715
## MäklareMäklarhuset                                0.71532
## MäklareMäklarMäster                               0.81588
## MäklareMäklarringen                               0.54549
## MäklareNomad Mäkleri                              0.75432
## MäklareNoside Fastighetsförmedling                0.81588
## MäklareNotar                                      0.48980
## MäklareNybergs Hem                                0.48483
## MäklareOBOS                                       0.26596
## MäklareOlovsson - Stignäs AB                      0.28216
## MäklarePeab Bostad                                0.59075
## MäklarePrivatmäklaren                             0.65609
## MäklareProperties & Partners Fastighetsmäklare    0.64306
## MäklareReal Vision Fastighetsmäklare              0.37000
## MäklareRestate                                    0.16823
## MäklareRiksmäklaren                               0.23274
## MäklareSjönära Fastigheter AB                     0.06552 .
## MäklareSjöös Fastighetsförmedling                 0.31177
## MäklareSkandiaMäklarna                            0.96468
## MäklareSkeppsholmen                               0.32237
## MäklareSmart fastighetsförmedling                 0.60034
## MäklareStadsvillan Fastigheter                    0.98188
## MäklareStefan Blomdin AB                          0.65760
## MäklareSthlmFast                                  0.86330
## MäklareSusanne Persson Fastighetsförmedling AB    0.86540
## MäklareSvensk Fastighetsförmedling                0.62420
## MäklareSvenska Mäklargruppen                      0.31759
## MäklareSvenska Mäklarhuset                        0.49526
## MäklareTradition Fastighetsmäkleri AB             0.94199
## MäklareUnik Fastighetsförmedling                  0.56861
## MäklareURBAN by ESNY                              0.51367
## MäklareViktor Hanson                              0.78938
## MäklareVision Fastighetsmäkleri AB                0.45985
## MäklareWiderlöv & Co                              0.47398
## MäklareWrede                                      0.09091 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1215000 on 10194 degrees of freedom
## Multiple R-squared:  0.1179, Adjusted R-squared:   0.11
## F-statistic: 14.97 on 91 and 10194 DF,  p-value: < 2.2e-16
```

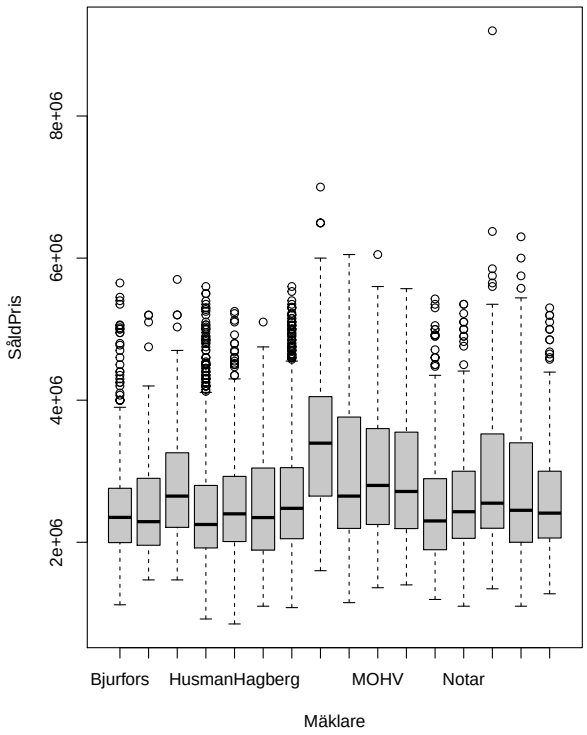


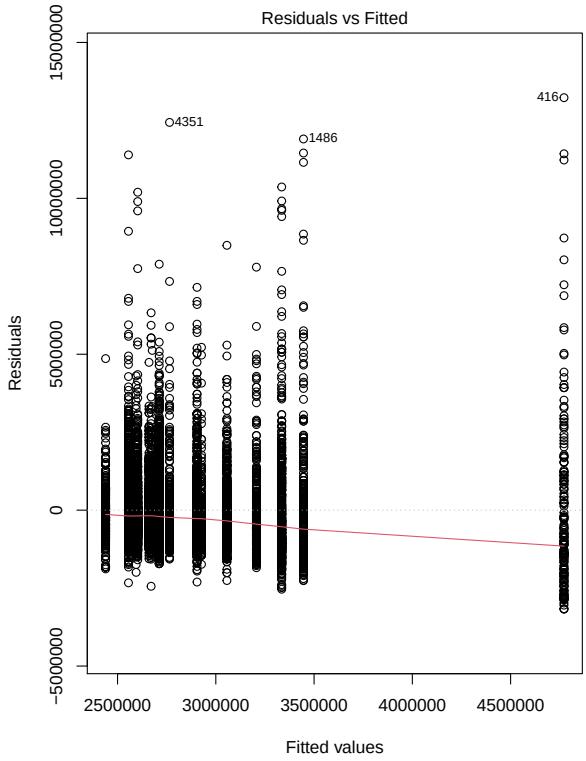


6.1.2 Mäklare efter sammanfogning:

```
##
## Call:
## lm(formula = SåldPris ~ Mäklare, data = data_träning)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3171583  -740587  -277400   437949 13228417
##
## Coefficients:
##                                          Estimate Std. Error t value
## (Intercept)                               2602711      49323  52.769
## MäklareERA                                  74001     121058   0.611
## MäklareErik Olsson Fastighetsförmedling    324377     109733   2.956
## MäklareFastighetsbyrån                     -47024      57167  -0.823
## MäklareHusmanHagberg                        -8175      68359  -0.120
## MäklareJägholm Norrortsmäklarna           -163357      93381  -1.749
## MäklareLänsförsäkringar Fastighetsförmedling  109340      55406   1.973
## MäklareMagnusson Mäkleri                  2168872     104501  20.755
## MäklareMindre Mäklarfirmor                 732578      69399  10.556
## MäklareMOHV                                603908      82002   7.365
## MäklareMäklarhuset                         453833      73678   6.160
## MäklareMäklarringen                        161863      84038   1.926
## MäklareNotar                                57559      74001   0.778
## MäklareSkandiaMäklarna                     843414      83464  10.105
```

```
## MäklareSvensk Fastighetsförmedling            301847      65287   4.623
## MäklareSvenska Mäklarhuset                     68123      72869   0.935
##                                             Pr(>|t|)
## (Intercept)                                  < 2e-16 ***
## MäklareERA                                   0.54102
## MäklareErik Olsson Fastighetsförmedling      0.00312 **
## MäklareFastighetsbyrån                       0.41077
## MäklareHusmanHagberg                         0.90481
## MäklareJägholm Norrortsmäklarna              0.08026 .
## MäklareLänsförsäkringar Fastighetsförmedling 0.04848 *
## MäklareMagnusson Mäkleri                     < 2e-16 ***
## MäklareMindre Mäklarfirmor                   < 2e-16 ***
## MäklareMOHV                                 1.91e-13 ***
## MäklareMäklarhuset                          7.56e-10 ***
## MäklareMäklarringen                          0.05412 .
## MäklareNotar                                 0.43670
## MäklareSkandiaMäklarna                       < 2e-16 ***
## MäklareSvensk Fastighetsförmedling          3.82e-06 ***
## MäklareSvenska Mäklarhuset                   0.34988
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1236000 on 10270 degrees of freedom
## Multiple R-squared:  0.07978,    Adjusted R-squared:  0.07843
## F-statistic: 59.35 on 15 and 10270 DF,  p-value: < 2.2e-16
```



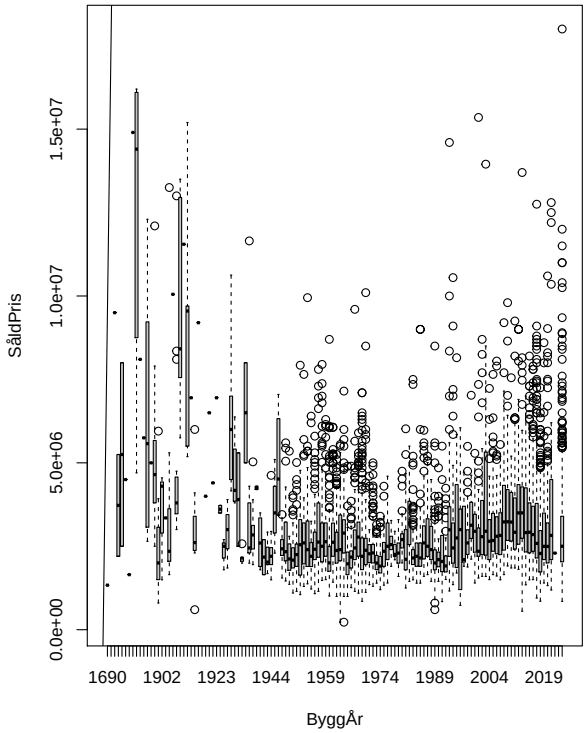


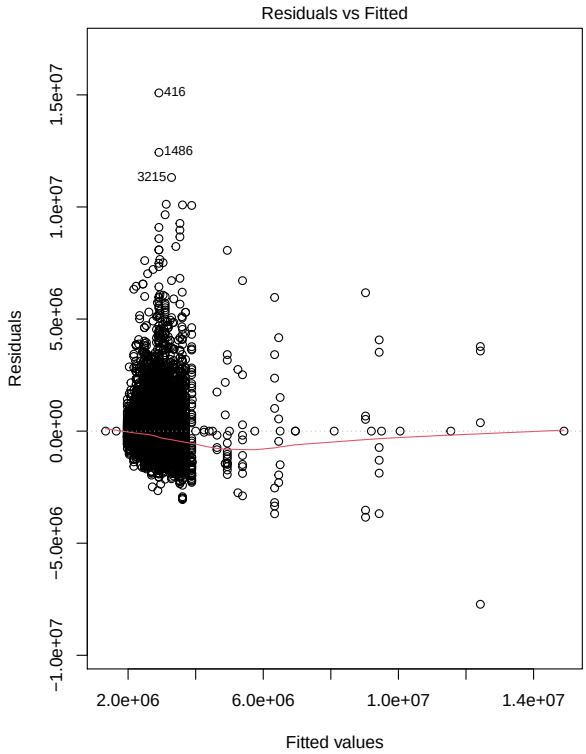
6.1.3 ByggÅr innan sammanfogning:

```
##
## Call:
## lm(formula = SåldPris ~ ByggÅr, data = data_träning)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -7725000  -687281  -214011   394362 15087722
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1335000    1164236   1.147 0.251543
## ByggÅr1820   8165000    1646478   4.959 7.20e-07 ***
## ByggÅr1830   2390000    1425892   1.676 0.093741 .
## ByggÅr1870   3915000    1425892   2.746 0.006050 **
## ByggÅr1880   3165000    1646478   1.922 0.054598 .
## ByggÅr1889    315000    1646478   0.191 0.848281
## ByggÅr1892  13565000    1646478   8.239  < 2e-16 ***
## ByggÅr1893  11090000    1301655   8.520  < 2e-16 ***
## ByggÅr1895   6765000    1646478   4.109 4.01e-05 ***
## ByggÅr1897   4415000    1646478   2.681 0.007342 **
## ByggÅr1898   5001875    1234859   4.051 5.15e-05 ***
## ByggÅr1899   3665000    1646478   2.226 0.026038 *
## ByggÅr1900   4051500    1221061   3.318 0.000910 ***
## ByggÅr1901   1085385    1208185   0.898 0.369015
## ByggÅr1902   2113333    1344344   1.572 0.115977
## ByggÅr1903   2015000    1646478   1.224 0.221047
## ByggÅr1904   1794194    1182865   1.517 0.129343
## ByggÅr1905   8715000    1646478   5.293 1.23e-07 ***
## ByggÅr1906   3601000    1202418   2.995 0.002753 **
## ByggÅr1907   8095833    1257519   6.438 1.27e-10 ***
## ByggÅr1908  10215000    1646478   6.204 5.71e-10 ***
## ByggÅr1909   7692000    1275357   6.031 1.68e-09 ***
## ByggÅr1911   5615000    1646478   3.410 0.000651 ***
## ByggÅr1912   1630714    1244621   1.310 0.190154
## ByggÅr1914   7865000    1646478   4.777 1.81e-06 ***
## ByggÅr1917   2665000    1646478   1.619 0.105563
## ByggÅr1920   5165000    1646478   3.137 0.001712 **
## ByggÅr1922   3065000    1646478   1.862 0.062696 .
## ByggÅr1923   5615000    1646478   3.410 0.000651 ***
## ByggÅr1924   2275000    1425892   1.595 0.110633
## ByggÅr1925    996667    1344344   0.741 0.458482
## ByggÅr1927   1581667    1344344   1.177 0.239409
## ByggÅr1929   5122000    1275357   4.016 5.96e-05 ***
## ByggÅr1930   3296250    1301655   2.532 0.011345 *
## ByggÅr1934   2565000    1425892   1.799 0.072068 .
## ByggÅr1935    851000    1275357   0.667 0.504618
## ByggÅr1936   5165000    1425892   3.622 0.000293 ***
## ByggÅr1938   2079427    1197989   1.736 0.082636 .
## ByggÅr1939   1530484    1182865   1.294 0.195737
## ByggÅr1940   2915000    1425892   2.044 0.040946 *
## ByggÅr1941   1287500    1301655   0.989 0.322625
## ByggÅr1942    840000    1425892   0.589 0.555804
```

```
## ByggÅr1943    869091    1216005   0.715 0.474805
## ByggÅr1944    929423    1186414   0.783 0.433417
## ByggÅr1945   2296667    1344344   1.708 0.087594 .
## ByggÅr1946   3541250    1301655   2.721 0.006528 **
## ByggÅr1947   1135652    1176823   0.965 0.334560
## ByggÅr1948   1415521    1176301   1.203 0.228863
## ByggÅr1949    890217    1189276   0.749 0.454154
## ByggÅr1950    908942    1175377   0.773 0.439351
## ByggÅr1951    996974    1170348   0.852 0.394311
## ByggÅr1952   1279922    1173296   1.091 0.275353
## ByggÅr1953   1403750    1173296   1.196 0.231562
## ByggÅr1954   1404355    1170479   1.200 0.230240
## ByggÅr1955    979081    1170908   0.836 0.403078
## ByggÅr1956   1316250    1170284   1.125 0.260731
## ByggÅr1957   1693509    1169331   1.448 0.147572
## ByggÅr1958   1505557    1167891   1.289 0.197384
## ByggÅr1959   1467423    1168328   1.256 0.209144
## ByggÅr1960    903036    1166313   0.774 0.438792
## ByggÅr1961   1454841    1166248   1.247 0.212260
## ByggÅr1962   1204728    1165993   1.033 0.301525
## ByggÅr1963   1382282    1166175   1.185 0.235922
## ByggÅr1964   1546647    1166544   1.326 0.184923
## ByggÅr1965    699947    1170348   0.598 0.549808
## ByggÅr1966    890268    1174585   0.758 0.448503
## ByggÅr1967   1244229    1166610   1.067 0.286208
## ByggÅr1968   1477318    1166263   1.267 0.205288
## ByggÅr1969   1148755    1166431   0.985 0.324723
## ByggÅr1970   1157632    1167635   0.991 0.321498
## ByggÅr1971    959479    1176301   0.816 0.414705
## ByggÅr1972   1085127    1167187   0.930 0.352552
## ByggÅr1973    699145    1168059   0.599 0.549485
## ByggÅr1974    643125    1170832   0.549 0.582819
## ByggÅr1975    984620    1171581   0.840 0.400693
## ByggÅr1976   1378304    1170546   1.177 0.239028
## ByggÅr1977   1247200    1187292   1.050 0.293533
## ByggÅr1978   1046875    1234859   0.848 0.396587
## ByggÅr1979    978182    1190402   0.822 0.411253
## ByggÅr1980   1440444    1177101   1.224 0.221085
## ByggÅr1981   1140313    1169422   0.975 0.329530
## ByggÅr1982   1847509    1174772   1.573 0.115830
## ByggÅr1983   1008234    1167717   0.863 0.387925
## ByggÅr1984    923234    1171772   0.788 0.430776
## ByggÅr1985   1104717    1169287   0.945 0.344794
## ByggÅr1986   1439460    1169469   1.231 0.218401
## ByggÅr1987   1442009    1169664   1.233 0.217664
## ByggÅr1988   1222805    1178348   1.038 0.299421
## ByggÅr1989    835090    1167312   0.715 0.474381
## ByggÅr1990    919083    1173898   0.783 0.433685
## ByggÅr1991   1040652    1170546   0.889 0.374007
## ByggÅr1992    749557    1171581   0.640 0.522329
## ByggÅr1993   1950193    1171228   1.665 0.095927 .
## ByggÅr1994   1700965    1174404   1.448 0.147546
## ByggÅr1995   1983250    1192987   1.662 0.096459 .
## ByggÅr1996   1357632    1194481   1.137 0.255738
```

```
## ByggÅr1997   1546429    1244621   1.242 0.214085
## ByggÅr1998   1493429    1180751   1.265 0.205967
## ByggÅr1999   2031667    1185600   1.714 0.086629 .
## ByggÅr2000   1657000    1180751   1.403 0.160545
## ByggÅr2001   1579481    1171772   1.348 0.177707
## ByggÅr2002   2258750    1184844   1.906 0.056630 .
## ByggÅr2003   2547460    1173440   2.171 0.029959 *
## ByggÅr2004   1586548    1178014   1.347 0.178075
## ByggÅr2005   1717407    1169614   1.468 0.142039
## ByggÅr2006   1831628    1175595   1.558 0.119254
## ByggÅr2007   1672222    1177101   1.421 0.155456
## ByggÅr2008   2200302    1169243   1.882 0.059889 .
## ByggÅr2009   2267540    1168847   1.940 0.052410 .
## ByggÅr2010   2152708    1176301   1.830 0.067269 .
## ByggÅr2011   2016282    1169201   1.724 0.084649 .
## ByggÅr2012   2365238    1169985   2.022 0.043244 *
## ByggÅr2013   2275433    1168572   1.947 0.051539 .
## ByggÅr2014   2011750    1168921   1.721 0.085276 .
## ByggÅr2015   1913926    1167005   1.640 0.101029
## ByggÅr2016   1851519    1166077   1.588 0.112358
## ByggÅr2017   1761690    1165570   1.511 0.130707
## ByggÅr2018   1205488    1165791   1.034 0.301137
## ByggÅr2019   1532259    1167577   1.312 0.189435
## ByggÅr2020   1606883    1168010   1.376 0.168931
## ByggÅr2021   2198584    1169376   1.880 0.060118 .
## ByggÅr2022    960000    1646478   0.583 0.559864
## ByggÅrOkänd  1577278    1164729   1.354 0.175702
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1164000 on 10163 degrees of freedom
## Multiple R-squared:  0.1921, Adjusted R-squared:  0.1824
## F-statistic: 19.81 on 122 and 10163 DF,  p-value: < 2.2e-16
```

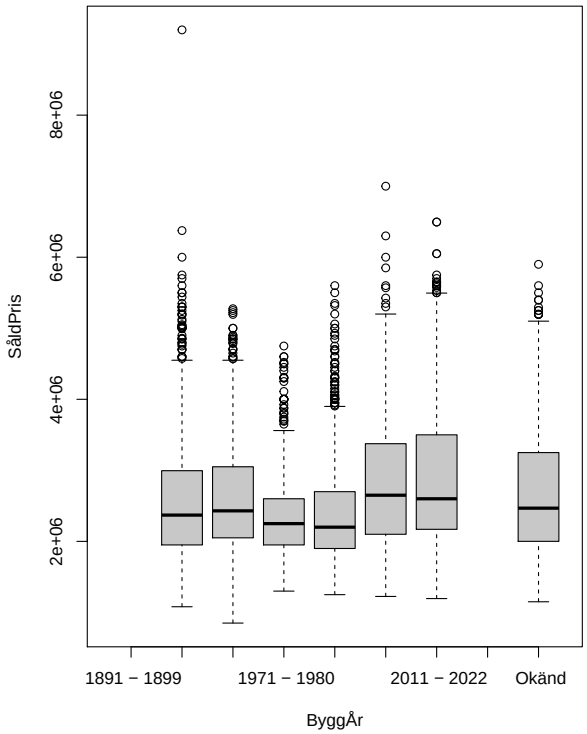


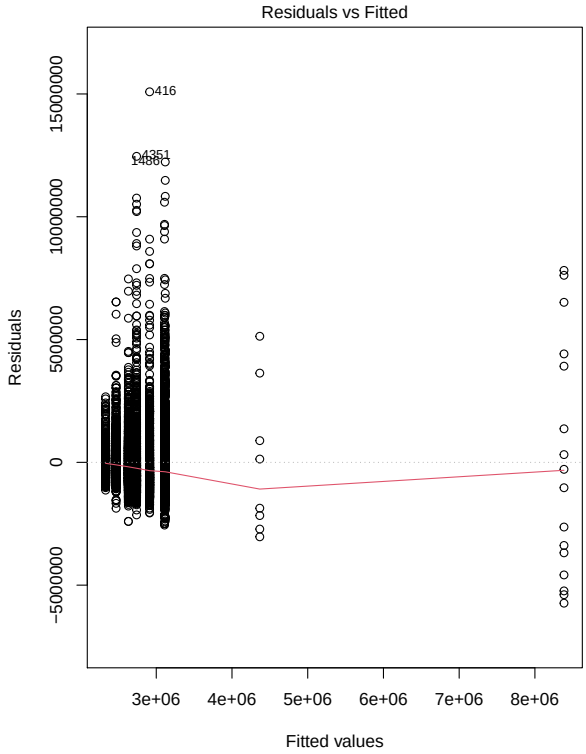


6.1.4 ByggÅr efter sammanfogning:

```
##
## Call:
## lm(formula = SåldPris ~ ByggÅr, data = data_träning)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -5734062  -740051  -281903   442574 15087722
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          8384063     310336  27.016  < 2e-16 ***
## ByggÅr 1900 - 1960 -5644012     311895 -18.096  < 2e-16 ***
## ByggÅr 1961 - 1970 -5752160     311417 -18.471  < 2e-16 ***
## ByggÅr 1971 - 1980 -6052801     313602 -19.301  < 2e-16 ***
## ByggÅr 1981 - 1990 -5914612     312727 -18.913  < 2e-16 ***
## ByggÅr 1991 - 2010 -5265307     312472 -16.850  < 2e-16 ***
## ByggÅr 2011 - 2022 -5275067     311435 -16.938  < 2e-16 ***
## ByggÅr innan 1890  -4017188     537517  -7.474 8.44e-14 ***
## ByggÅr Okänd       -5471784     312431 -17.514  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1241000 on 10277 degrees of freedom
## Multiple R-squared:  0.07122,    Adjusted R-squared:  0.0705
```

```
## F-statistic: 98.51 on 8 and 10277 DF,  p-value: < 2.2e-16
```

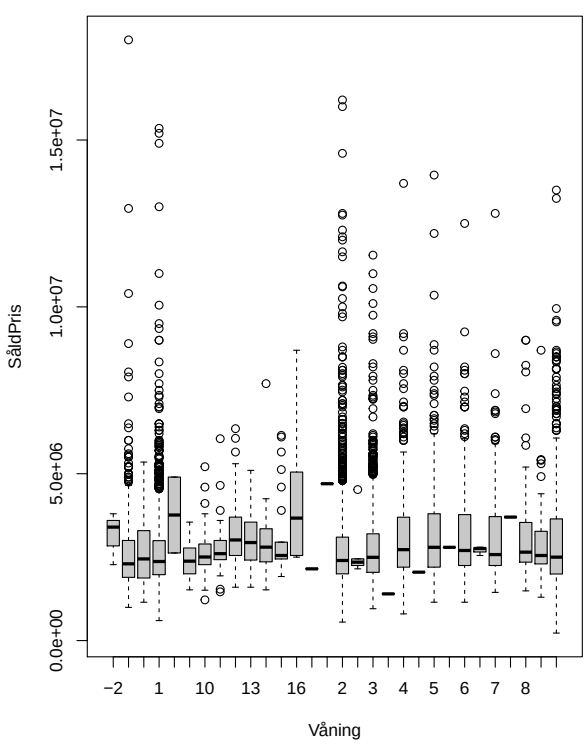


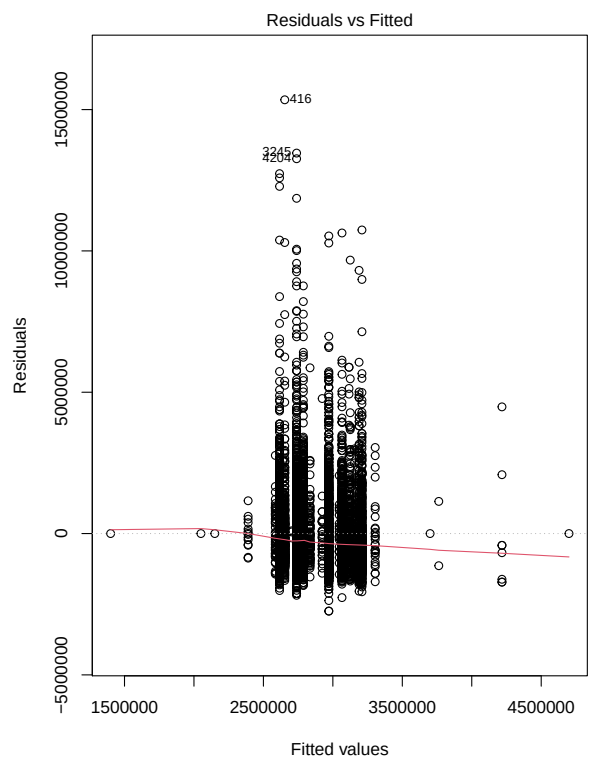


6.1.5 Våning innan sammanfogning:

```
##
## Call:
## lm(formula = SåldPris ~ Våning, data = data_träning)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2745921  -770921  -321272   412471 15346388
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3158333     735967   4.291 1.79e-05 ***
## Våning0      -504721     738116  -0.684    0.494
## Våning0.5    -572471     754761  -0.758    0.448
## Våning1      -542061     736413  -0.736    0.462
## Våning1.2     604167    1163666   0.519    0.604
## Våning1.5    -768387     810996  -0.947    0.343
## Våning10     -519105     755086  -0.687    0.492
## Våning11     -378191     766860  -0.493    0.622
## Våning12      146488     774391   0.189    0.850
## Våning13     -112583     789237  -0.143    0.887
## Våning14     -235167     771889  -0.305    0.761
## Våning15       -3533     778874  -0.005    0.996
```

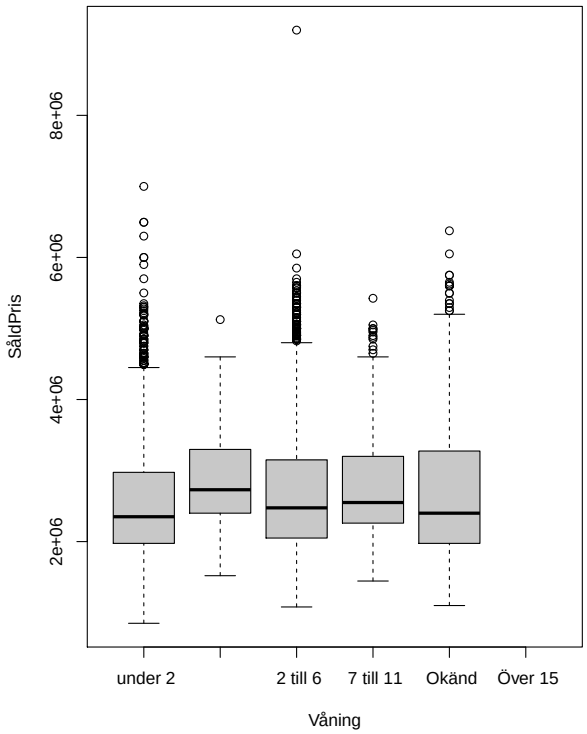
```
## Våning16      1059167     862998   1.227    0.220
## Våning16.5   -1008333    1471934  -0.685    0.493
## Våning17.5    1541667    1471934   1.047    0.295
## Våning2       -419358     736423  -0.569    0.569
## Våning2.5     -413333     930933  -0.444    0.657
## Våning3       -370804     736706  -0.503    0.615
## Våning30.5   -1758333    1471934  -1.195    0.232
## Våning4        -92551     737533  -0.125    0.900
## Våning4.5    -1108333    1471934  -0.753    0.451
## Våning5         51045     738466   0.069    0.945
## Våning5.5     -363333    1471934  -0.247    0.805
## Våning6         30467     739764   0.041    0.967
## Våning6.5     -455000    1040815  -0.437    0.662
## Våning7        -33261     741255  -0.045    0.964
## Våning75.5     541667    1471934   0.368    0.713
## Våning8        -41901     743867  -0.056    0.955
## Våning9       -322726     746213  -0.432    0.665
## VåningOkänd   -187412     736883  -0.254    0.799
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1275000 on 10256 degrees of freedom
## Multiple R-squared:  0.02258,    Adjusted R-squared:  0.01982
## F-statistic: 8.172 on 29 and 10256 DF,  p-value: < 2.2e-16
```

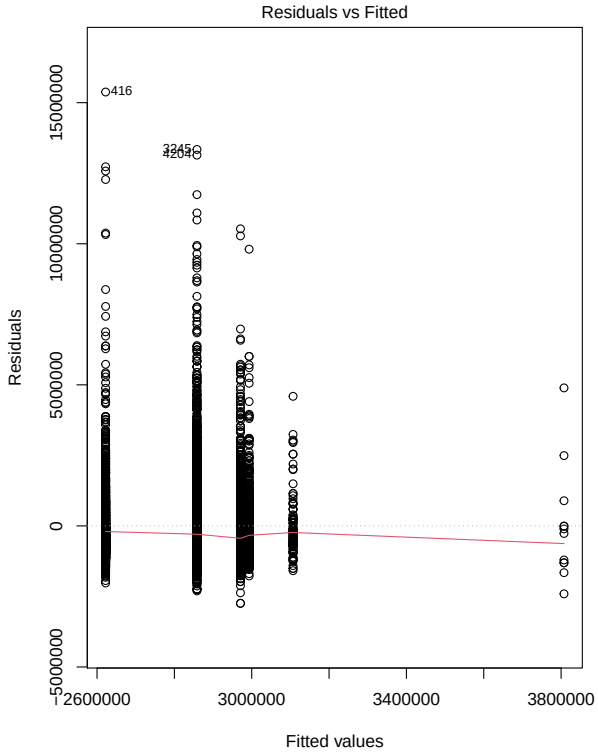




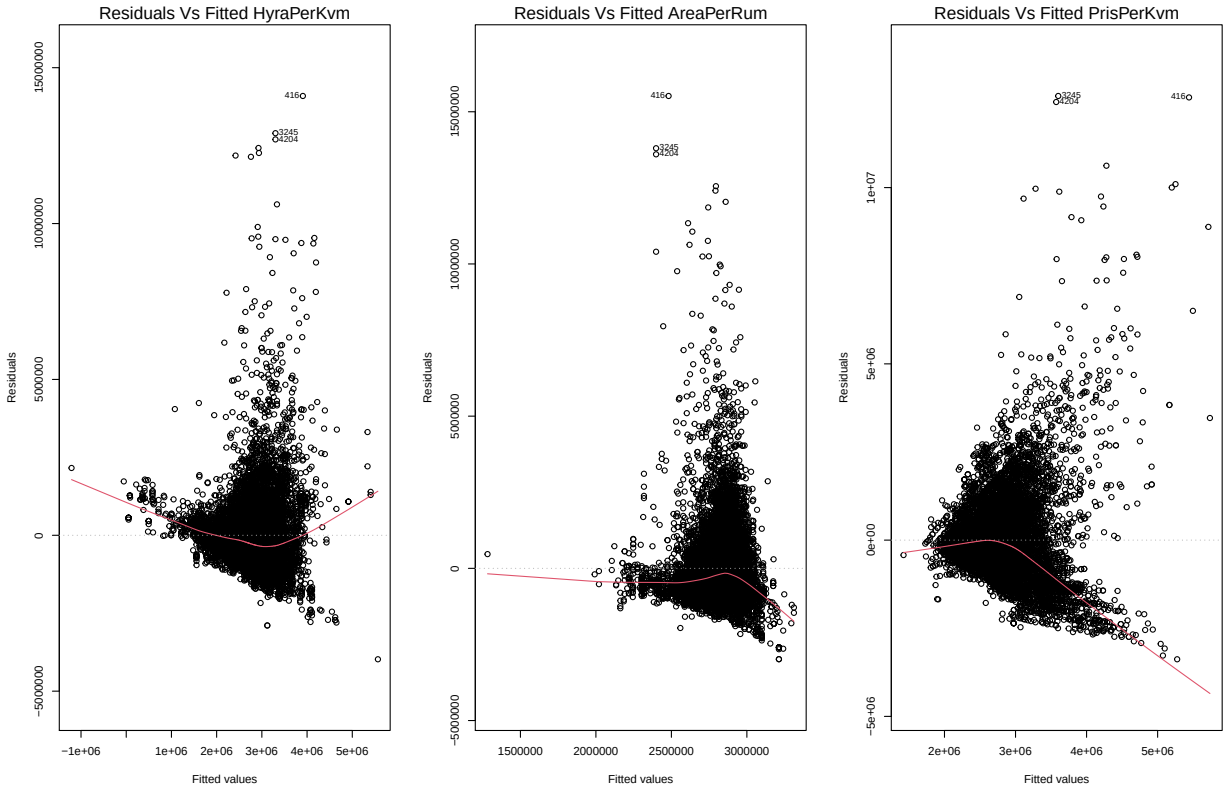
6.1.6 Våning efter sammanfogning:

```
##
## Call:
## lm(formula = SåldPris ~ Våning, data = data_träning)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2745921  -770921  -343315   407523 15377803
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         2622197      23143 113.304  < 2e-16 ***
## Våning 12 till 15    484745     128289   3.779 0.000159 ***
## Våning 2 till 6      236188      29012   8.141 4.37e-16 ***
## Våning 7 till 11     371118      59492   6.238 4.60e-10 ***
## Våning Okänd         348724      43550   8.007 1.30e-15 ***
## Våning Över 15      1185303     370411   3.200 0.001379 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1281000 on 10280 degrees of freedom
## Multiple R-squared:  0.01121,    Adjusted R-squared:  0.01073
## F-statistic:  23.3 on 5 and 10280 DF,  p-value: < 2.2e-16
```





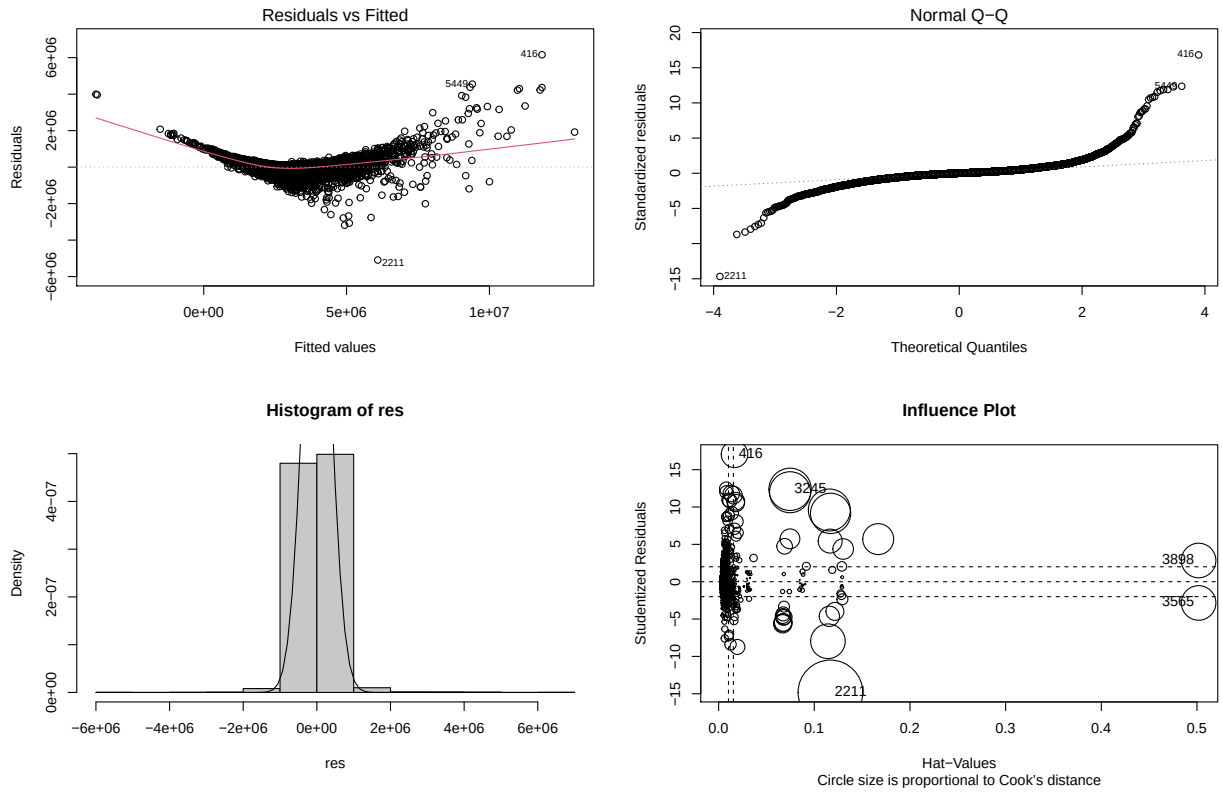
6.1.7 Relevanta residual plottar:



6.2 Modeller:

6.2.1 Modell 1, innan borttagning av outliers:

```
##       rstudent unadjusted p-value Bonferroni p
## 416   17.05724         2.3842e-64   2.4524e-60
## 2211 -14.82397         3.3167e-49   3.4115e-45
## 5449  12.45967         2.2358e-35   2.2997e-31
## 3245  12.37089         6.6688e-35   6.8595e-31
## 1693  11.98350         7.1935e-33   7.3992e-29
## 4204  11.98017         7.4838e-33   7.6979e-29
## 1486  11.80492         5.9312e-32   6.1008e-28
## 4351  11.58624         7.5361e-31   7.7517e-27
## 2871  10.94496         1.0006e-27   1.0292e-23
## 2734  10.84276         3.0351e-27   3.1219e-23
```

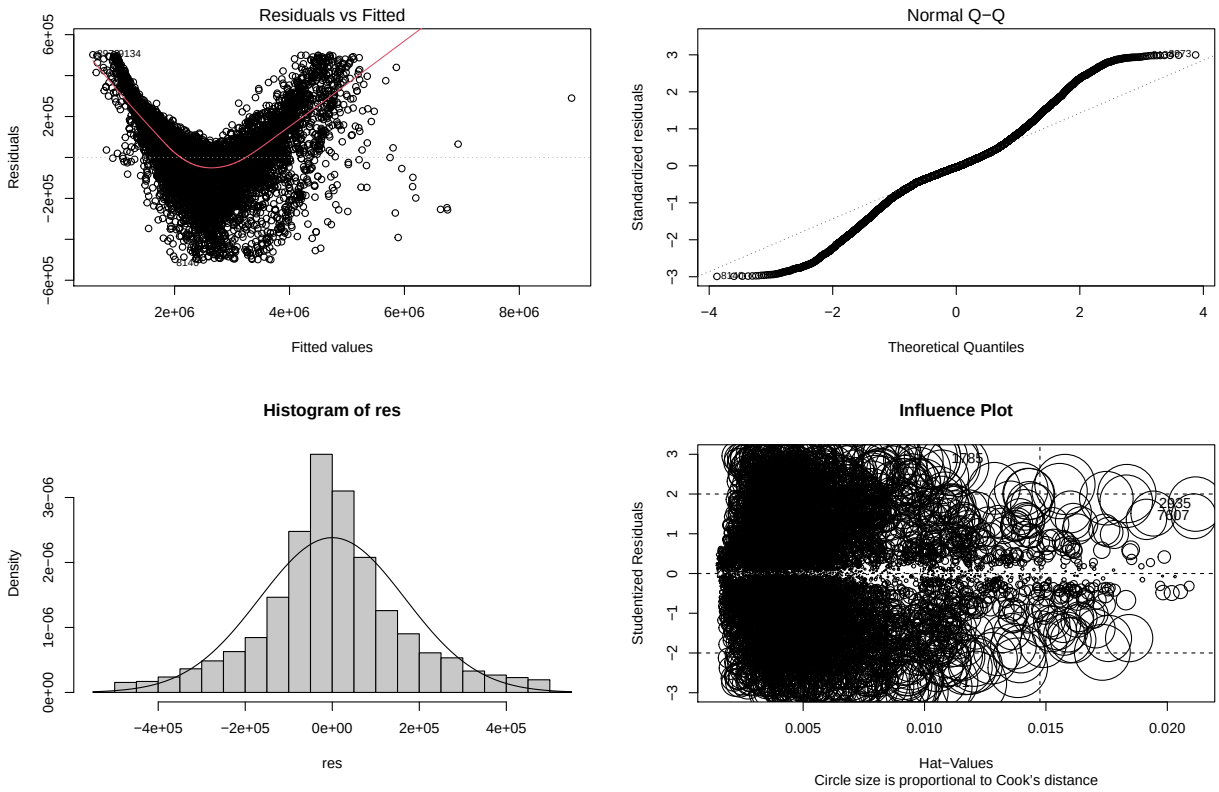



```
##         StudRes        Hat      CookD
## 416   17.057240 0.01662838 0.09026916
## 2211 -14.823968 0.11686017 0.53716052
## 3245  12.370892 0.07468370 0.22964480
## 3565  -2.840779 0.50169797 0.15319643
## 3898   2.840779 0.50169797 0.15319643
```

skewness
2.118497

6.2.2 Modell 1, efter borttagning av outliers:

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##      rstudent unadjusted p-value Bonferroni p
## 8973  2.99939          0.0027124           NA
```




```
##        StudRes         Hat        CookD
## 1785  2.878508 0.010926683 0.0019883697
## 2935  1.746395 0.021164967 0.0014333088
## 7607  1.441010 0.021088918 0.0009723829
## 8140 -2.992545 0.003551962 0.0006933720
## 8201 -2.766727 0.012046150 0.0020275710
## 8973  2.999390 0.004828952 0.0009481791
```

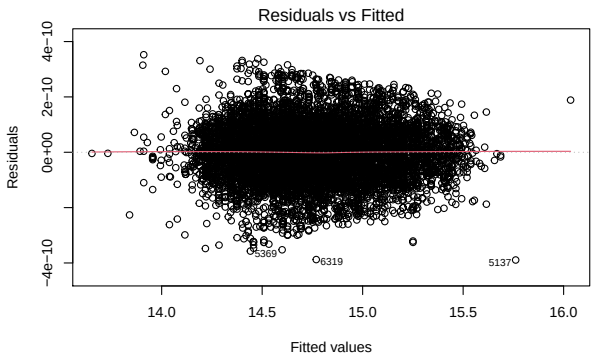
skewness
0.1071339

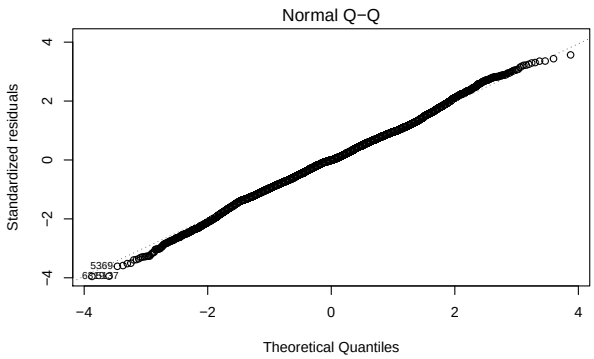
6.2.3 Modell 5, log-transformation enligt scatterplott i enkel linjär regression:

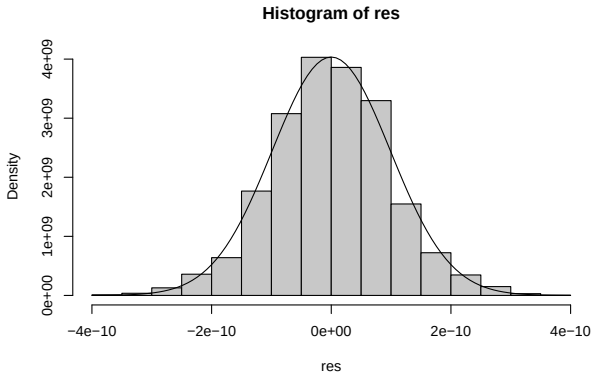
```
##
## Call:
## lm(formula = log(SåldPris) ~ Mäklare + Rum + ByggÅr + Kommun +
##     AvståndVattenMeter + Våning + SåldSäsong + log(AreaPerRum) +
##     log(PrisPerKvm) + HyraPerKvm, data = data_träning)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -3.893e-10 -6.579e-11 -3.900e-13  6.575e-11  3.525e-10
##
## Coefficients:
```

```
##                                            Estimate Std. Error    t value
## (Intercept)                                1.802e-11  1.000e-10  1.800e-01
## MäklareERA                                 4.588e-12  1.038e-11  4.420e-01
## MäklareErik Olsson Fastighetsförmedling   -1.461e-12  9.485e-12 -1.540e-01
## MäklareFastighetsbyrån                     4.058e-12  4.967e-12  8.170e-01
## MäklareHusmanHagberg                       7.669e-12  5.877e-12  1.305e+00
## MäklareJägholm Norrortsmäklarna           -1.274e-11  8.248e-12 -1.544e+00
## MäklareLänsförsäkringar Fastighetsförmedling  2.306e-12  4.736e-12  4.870e-01
## MäklareMagnusson Mäkleri                   2.199e-12  1.028e-11  2.140e-01
## MäklareMindre Mäklarfirmor                 2.493e-12  6.170e-12  4.040e-01
## MäklareMOHV                                8.472e-13  7.273e-12  1.160e-01
## MäklareMäklarhuset                         5.899e-12  6.992e-12  8.440e-01
## MäklareMäklarringen                        4.852e-12  7.339e-12  6.610e-01
## MäklareNotar                              -4.758e-12  6.363e-12 -7.480e-01
## MäklareSkandiaMäklarna                    -9.192e-12  7.300e-12 -1.259e+00
## MäklareSvensk Fastighetsförmedling         2.903e-12  5.814e-12  4.990e-01
## MäklareSvenska Mäklarhuset                -1.165e-12  6.376e-12 -1.830e-01
## Rum1.5                                     4.055e-01  7.509e-12  5.400e+10
## Rum2                                       6.931e-01  4.703e-12  1.474e+11
## Rum2.5                                     9.163e-01  1.046e-11  8.757e+10
## Rum3                                       1.099e+00  5.779e-12  1.901e+11
## Rum3.5                                     1.253e+00  1.285e-11  9.753e+10
## Rum4                                       1.386e+00  6.893e-12  2.011e+11
## Rum5                                       1.609e+00  1.015e-11  1.585e+11
## ByggÅr 1961 - 1970                         9.401e-12  3.546e-12  2.651e+00
## ByggÅr 1971 - 1980                         8.782e-12  4.777e-12  1.838e+00
## ByggÅr 1981 - 1990                        -3.528e-12  4.415e-12 -7.990e-01
## ByggÅr 1991 - 2010                         8.545e-12  4.411e-12  1.937e+00
## ByggÅr 2011 - 2022                         3.499e-12  3.935e-12  8.890e-01
## ByggÅr Okänd                               2.503e-13  4.195e-12  6.000e-02
## KommunSollentuna                          -1.723e-11  5.140e-12 -3.352e+00
## KommunTäby                                -1.873e-11  4.866e-12 -3.850e+00
## KommunUpplands Väsby                      -3.818e-12  9.678e-12 -3.940e-01
## KommunVallentuna                          -5.438e-12  1.011e-11 -5.380e-01
## KommunVaxholm                             -3.214e-11  7.616e-12 -4.221e+00
## KommunÖsteråker                           -1.914e-11  6.495e-12 -2.947e+00
## AvståndVattenMeter                        -1.219e-15  9.606e-16 -1.269e+00
## Våning 12 till 15                          1.518e-11  1.073e-11  1.414e+00
## Våning 2 till 6                            7.227e-12  2.371e-12  3.048e+00
## Våning 7 till 11                           1.157e-11  5.103e-12  2.267e+00
## Våning Okänd                               5.592e-12  3.784e-12  1.478e+00
## SåldSäsongSommar                          -3.769e-12  2.852e-12 -1.322e+00
## SåldSäsongVinter                          -8.264e-13  2.986e-12 -2.770e-01
## SåldSäsongVår                              2.516e-12  2.697e-12  9.330e-01
## log(AreaPerRum)                            1.000e+00  8.790e-12  1.138e+11
## log(PrisPerKvm)                            1.000e+00  7.072e-12  1.414e+11
## HyraPerKvm                                 3.035e-15  1.266e-13  2.400e-02
##                                           Pr(>|t|)
## (Intercept)                               0.857046
## MäklareERA                                0.658546
## MäklareErik Olsson Fastighetsförmedling   0.877610
## MäklareFastighetsbyrån                    0.413968
## MäklareHusmanHagberg                      0.191901
## MäklareJägholm Norrortsmäklarna           0.122607
```

```
## MäklareLänsförsäkringar Fastighetsförmedling 0.626378
## MäklareMagnusson Mäkleri                     0.830683
## MäklareMindre Mäklarfirmor                   0.686182
## MäklareMOHV                                  0.907277
## MäklareMäklarhuset                           0.398831
## MäklareMäklarringen                          0.508504
## MäklareNotar                                 0.454654
## MäklareSkandiaMäklarna                       0.207979
## MäklareSvensk Fastighetsförmedling           0.617612
## MäklareSvenska Mäklarhuset                   0.855011
## Rum1.5                                        < 2e-16 ***
## Rum2                                          < 2e-16 ***
## Rum2.5                                        < 2e-16 ***
## Rum3                                          < 2e-16 ***
## Rum3.5                                        < 2e-16 ***
## Rum4                                          < 2e-16 ***
## Rum5                                          < 2e-16 ***
## ByggÅr 1961 - 1970                           0.008040 **
## ByggÅr 1971 - 1980                           0.066043 .
## ByggÅr 1981 - 1990                           0.424239
## ByggÅr 1991 - 2010                           0.052772 .
## ByggÅr 2011 - 2022                           0.373957
## ByggÅr Okänd                                 0.952425
## KommunSollentuna                             0.000806 ***
## KommunTäby                                   0.000119 ***
## KommunUpplands Väsby                         0.693244
## KommunVallentuna                             0.590704
## KommunVaxholm                                2.46e-05 ***
## KommunÖsteråker                              0.003220 **
## AvståndVattenMeter                           0.204391
## Våning 12 till 15                            0.157293
## Våning 2 till 6                              0.002308 **
## Våning 7 till 11                             0.023434 *
## Våning Okänd                                 0.139471
## SåldSäsongSommar                             0.186312
## SåldSäsongVinter                             0.781987
## SåldSäsongVår                                0.350857
## log(AreaPerRum)                               < 2e-16 ***
## log(PrisPerKvm)                               < 2e-16 ***
## HyraPerKvm                                   0.980879
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.911e-11 on 9308 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 1.966e+21 on 45 and 9308 DF,  p-value: < 2.2e-16
```

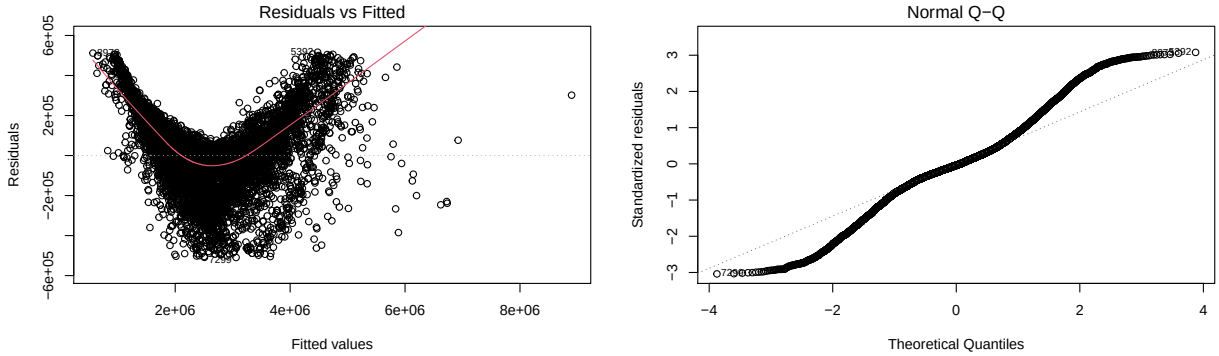

Residuals vs Fitted
Residuals
Fitted values
4e−10
2e−10
0e+00
−4e−10
14.0
14.5
15.0
15.5
16.0
5369
6319
5137

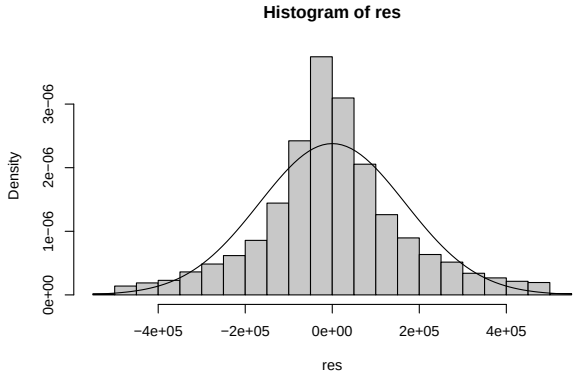

Normal Q−Q
Standardized residuals
Theoretical Quantiles
4
2
0
−2
−4
5369


Histogram of res
Density
res
4e+09
3e+09
2e+09
1e+09
0e+00
−4e−10
−2e−10
0e+00
2e−10
4e−10

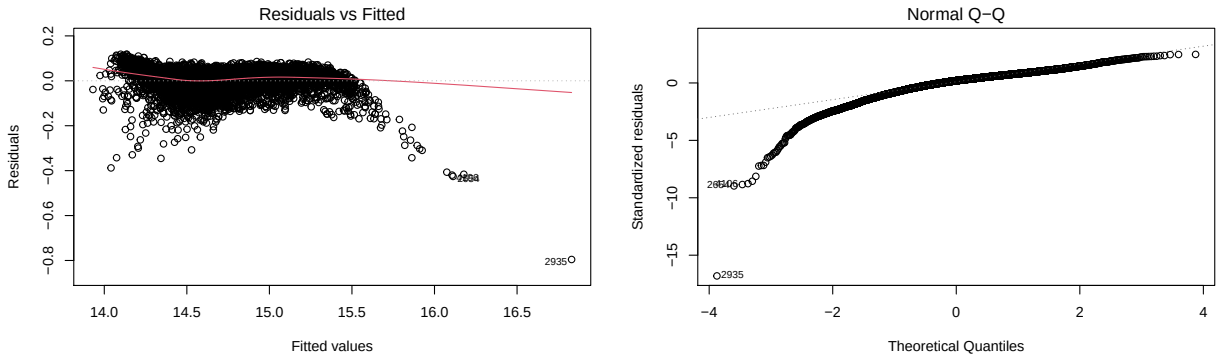
6.3 Modeller efter stegvis variabelselektion:

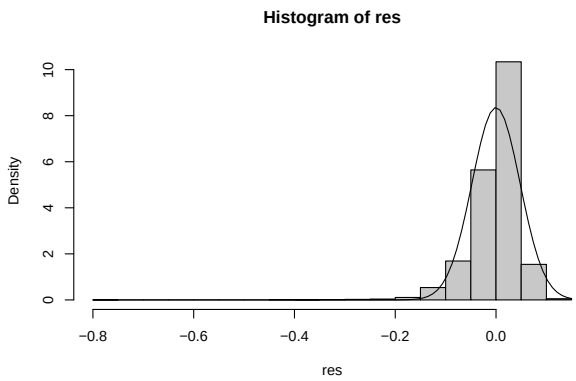
6.3.1 Modell 1 Stepwise:



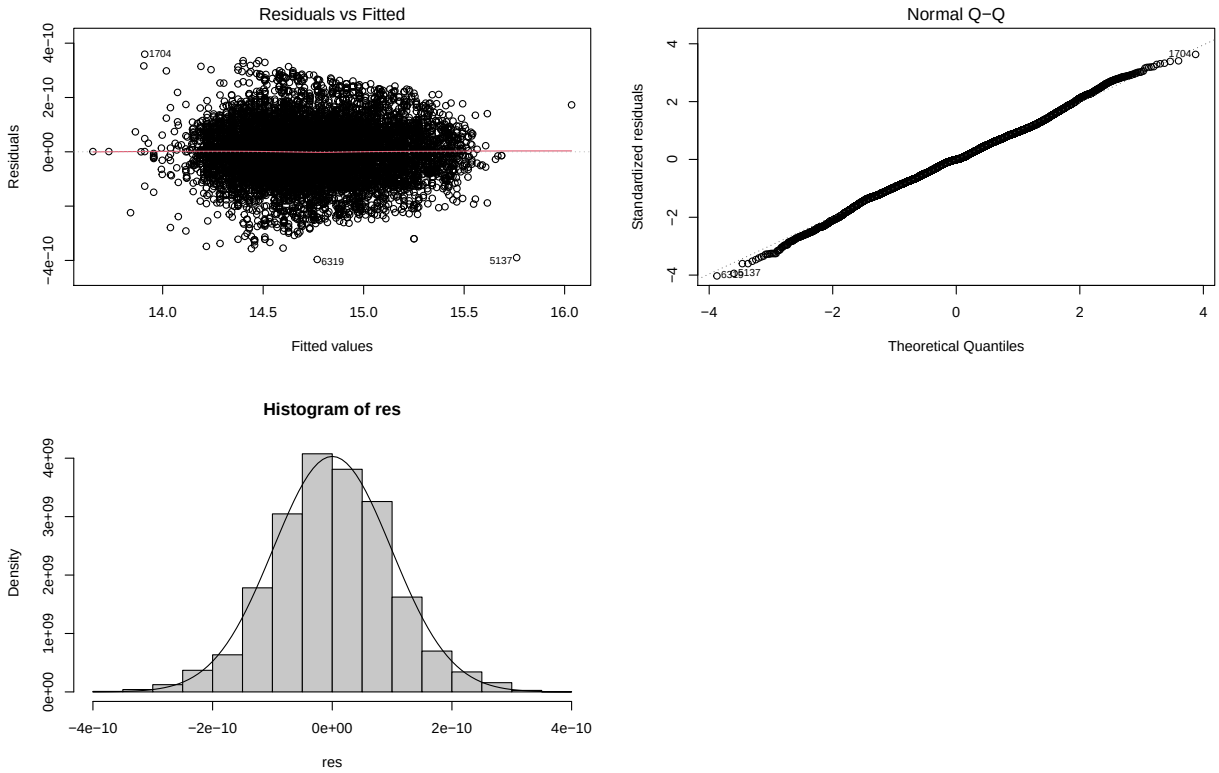


6.3.2 Modell 2 Stepwise:

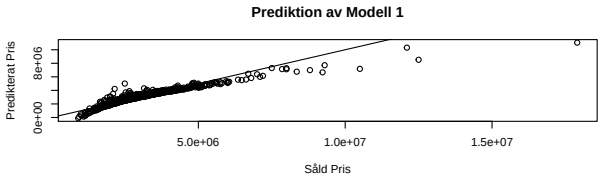


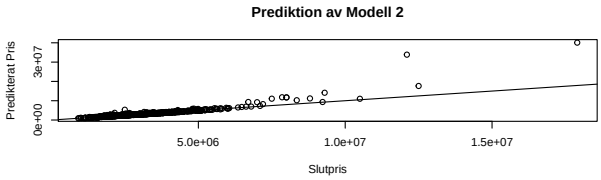


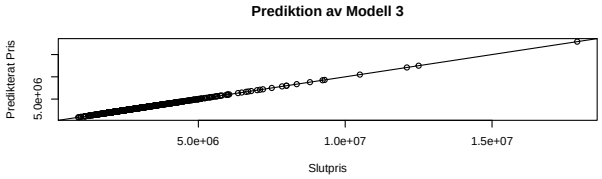
6.3.3 Modell Stepwise:

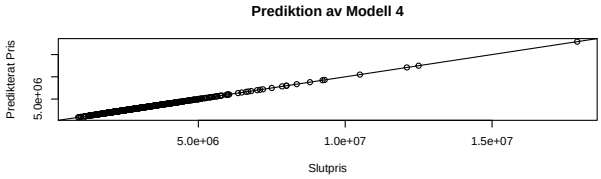


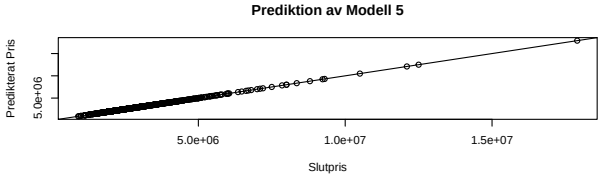
6.4 Prediktion:

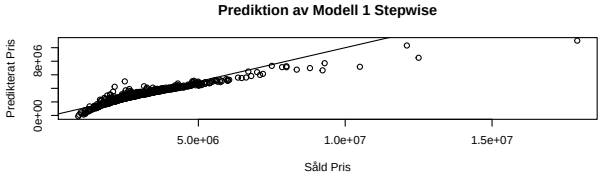


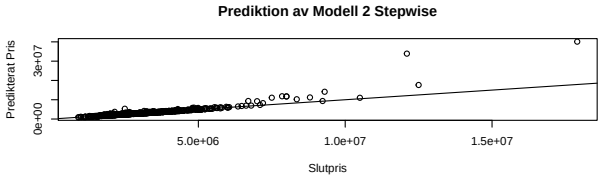


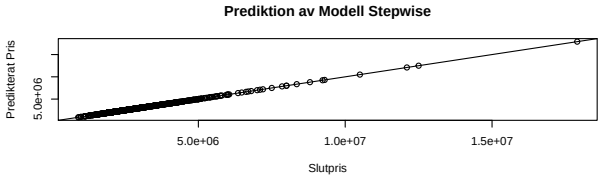


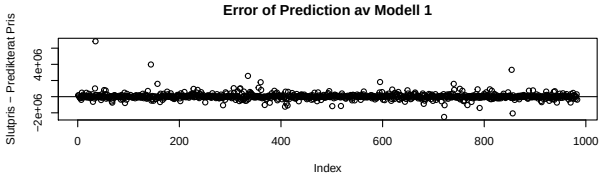


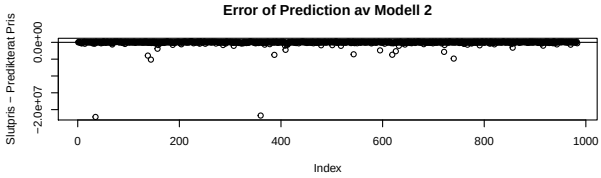


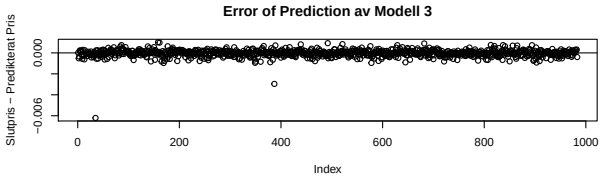


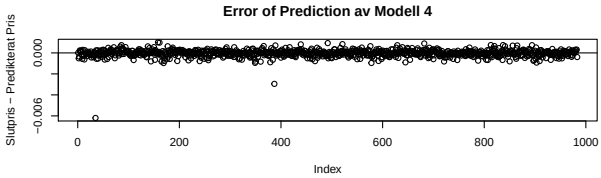


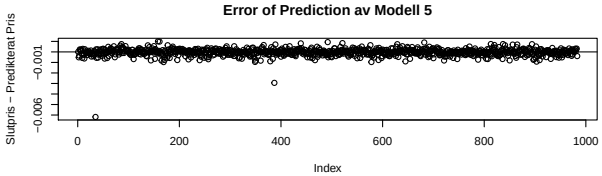


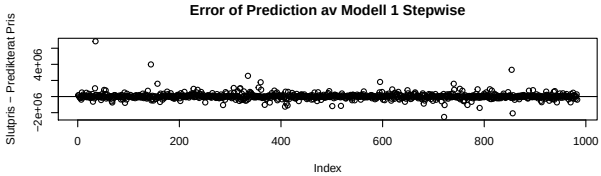


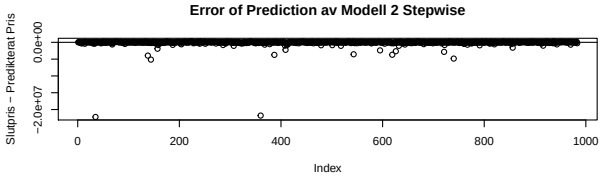


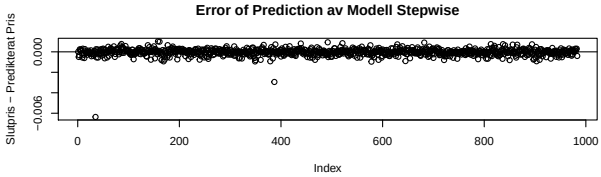












7 Referenser

[1] Tyrcha J, Andersson P, 2016, Notes in Econometrics, Department of Mathematics, Stockholm University

[2] Sundberg R, 2020, Lineära Statistiska Modeller, Department of Mathematics, Stockholm University

[3] Held L, Sabanés Bové D, 2014, Applied Statistical Inference Likelihood and Bayes, Springer

[4] Grandell J, Koski T, 07.03.2016, SF1901: Sannolikhetslära och statistik, föreläsning 15.
Enkel linjär regression, https://www.math.kth.se/matstat/gru/sf1901/F/fo15sf1901ny.pdf

[5] Ohlsson E, Johansson B, 2010, Non-Life Insurance Pricing with Generalized Linear Models, Springer

[6] Naturvårdsverket, Hämtad: 2022-05-01, http://www.miljostatistik.se/datatyper.html

[7] Olsson F, Matematikcentrum, Matematisk statistik Lunds universitet, October 13, 2019, FMSF30 - Hypotesprövning (kapitel 9),
https://www.lth.se/fileadmin/matematik_lth_hbg/kurshemsida_matstat_filer/hyptest.pdf

[8] The Pennsylvania State University, STAT 462, 2018, Hämtad: 2022-05-01, https://online.stat.psu.edu/stat462/node/247/

[9] Booli Search Technologies AB, Booli API, Hämtad: 2022-04-03, https://www.booli.se/p/api

[10] How to Interpret Residual Standard Error, Hämtad: 2022-05-02,
https://www.statology.org/how-to-interpret-residual-standard-error/#