Projekt 1

Leon Voss Gustavsson, Erik Krook, David Helms

9/21/2019

Sammanfattning

Vi har arbetat med en sammanställning på hundratals lägenhetsförsäljningar som skedde mellan 2006-2009 i området Östermalm–Gärdet i Stockholm. Syftet med den insamlade datan var att studera de olika förekommande variablernas påverkan av slutpriset med regressions- och variansanalys och få fram en förklarande modell. Störst påverkan visade sig vara Boarea men även andrra förklarande variabler kom med, dessa var byggnadsår och våningsplan. Antal variabler vi till sist använde i modellen var alltså betydligt färre än de övriga 12 som fanns tillgänliga (därbland fanns antal rum, avgift etc). Något kortfattat var skälet att övriga hade hög grad av kolineritet (vilket resulterar i en mer svårtolkad modell) samt obetydlig inverkan på förklaringsgraden, mer om detta i rapporten.

Inledning

Priser på bostäder varierar och vi vill undersöka vilka variabler som påverkar prisskillnader mest mellan bostäderna, vad gör vissa bostäder mer attraktiva än andra?

Data och statistisk modellering

Vi har data från lägenhetsförsäljningar i Gärdet, där ingår det 429 observationer av 13 variabler. Vi exemplifierar med en rad nedan(observera att mäklar variabeln är 1 för en specifik mäklare och 2 för en annan).

```
A23[1, ]
```

```
     pris maklare boarea rum avgift  CCI startpris byggnadsar balkong garage
1 5450000       2    117   4   2985 19.4   5450000       1897       0      0
  hiss vaningsplan tid
1    0         0.5   2
```

Vårat mål är sedan att förklara priset med hjälp av de övriga variablerna. Metodiken är multipel linjär regression där priset är responsvariabeln och ett urval av de återstående variablerna blir förklarande.

Till en start så skapar vi multipel modell där vi inkluderar samtliga variabler.

```
full.model.A23 <- lm(pris ~., data = A23)
summary(full.model.A23)
```

```
Call:
lm(formula = pris ~ ., data = A23)

Residuals:
     Min       1Q   Median       3Q      Max
-1145881  -138949   -34281    68074  1227409
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.343e+06  9.341e+05   1.438  0.15131
maklare      8.248e+04  2.842e+04   2.902  0.00393 **
boarea       4.526e+03  2.033e+03   2.226  0.02665 *
rum          4.458e+04  3.439e+04   1.296  0.19575
avgift      -7.044e+00  1.717e+01  -0.410  0.68183
CCI          8.687e+03  1.499e+03   5.796 1.49e-08 ***
startpris    8.869e-01  2.719e-02  32.617  < 2e-16 ***
byggnadsar  -9.837e+02  4.912e+02  -2.002  0.04599 *
balkong      5.989e+03  3.185e+04   0.188  0.85097
garage      -6.944e+04  5.162e+04  -1.345  0.17937
hiss         8.220e+04  3.144e+04   2.614  0.00932 **
vaningsplan  3.002e+04  7.933e+03   3.785  0.00018 ***
tid          1.828e+04  2.548e+03   7.173 4.24e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 239300 on 358 degrees of freedom
  (58 observations deleted due to missingness)
Multiple R-squared:  0.9797,    Adjusted R-squared:  0.979
F-statistic:  1437 on 12 and 358 DF,  p-value: < 2.2e-16
```

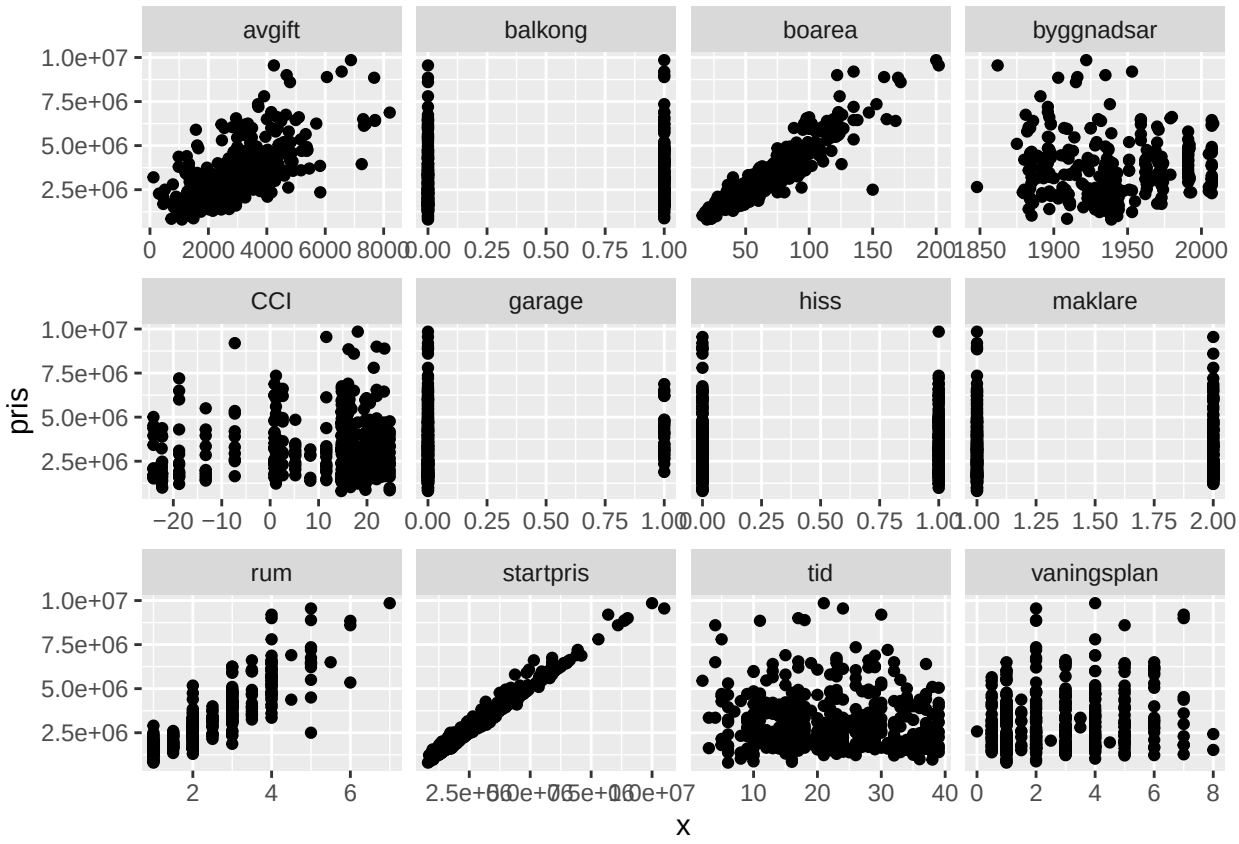
Vi ser direkt att den justerade förklaringsgraden $R^2 = 0.979$, alltså en rejält hög förklaringsgrad. Men det finns skäl att inte fira än. I en förklarande modell så vill vi gärna kunna tolka modellen, exempelvis kan vi vilja se hur en en specifik koefficient β_i i regressionen påverkar slutpriset. Är dom förklarande variablerna parvist beroende så kan vi ha fallet att dom var för sig är positiva (med ett signifkant p-värde som indikerar att båda har stark inverkan). Det medans dom tillsammans i samma modell kan få lutningskoefficienter som ändrat tecken. Variablen kan då helt plötsligt se ut att ha en negativ inverkan på priset, men hade samtidigt ensam förklarande variabel en positiv? Tråkigt nog blir det då svårt att tolka variablernas inverkan på slutpriset.

Då en viss grad av samverkan i praktiken alltid kommer att existera så ger för många förklarande variabler en svårtolkad modell. Det kan även vara så att en av dessa variabler i stort sett är densamma som den vi önskar att förklara. Datan vi undersöker har såväl startpris som pris vilket kan ses vara ett exempel av det. För mig låter det ungefär som att man skulle använda människors längd med skor för att förklara människors längd, alltså att variablen inte är förklarande utan snarare identisk, och därav intetsägande.

Ett tredje problem är att vissa av dom förklarande variablerna har oregistrerad data, exempelvis byggnadsår har flera rader saknande. I en multipel linjär regression där alla variabler påverkar varandra så raderas samtliga motsvarande observationer när endast observationen för en variabel saknas, den förklarande modellen byggs då på ett selektivt urval av grunddatan.

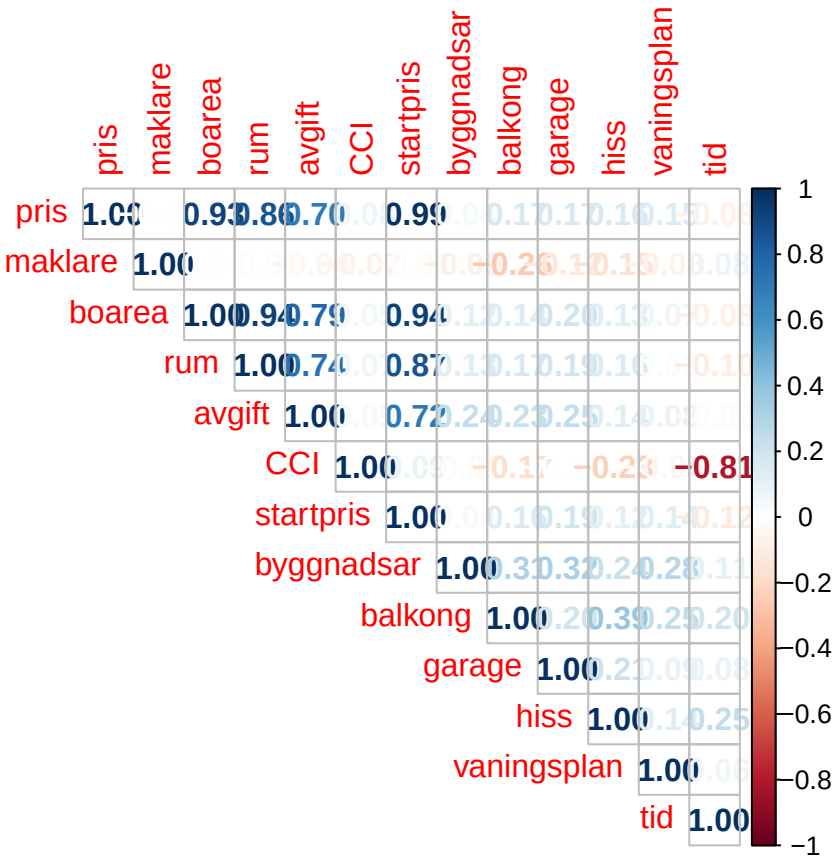
Målet då blir att förenkla modellen utan att tappa för mycket av förklaringsgraden. Plottande av datan blir ett bra första steg för att visuallisera inflytandet från våra möjliga responsvariabler.

```
A23 %>%
  pivot_longer(-pris, values_to = "x") %>%
  ggplot(aes(x = x, y = pris)) +
  geom_point() +
  facet_wrap(~name, scales = "free_x")
```



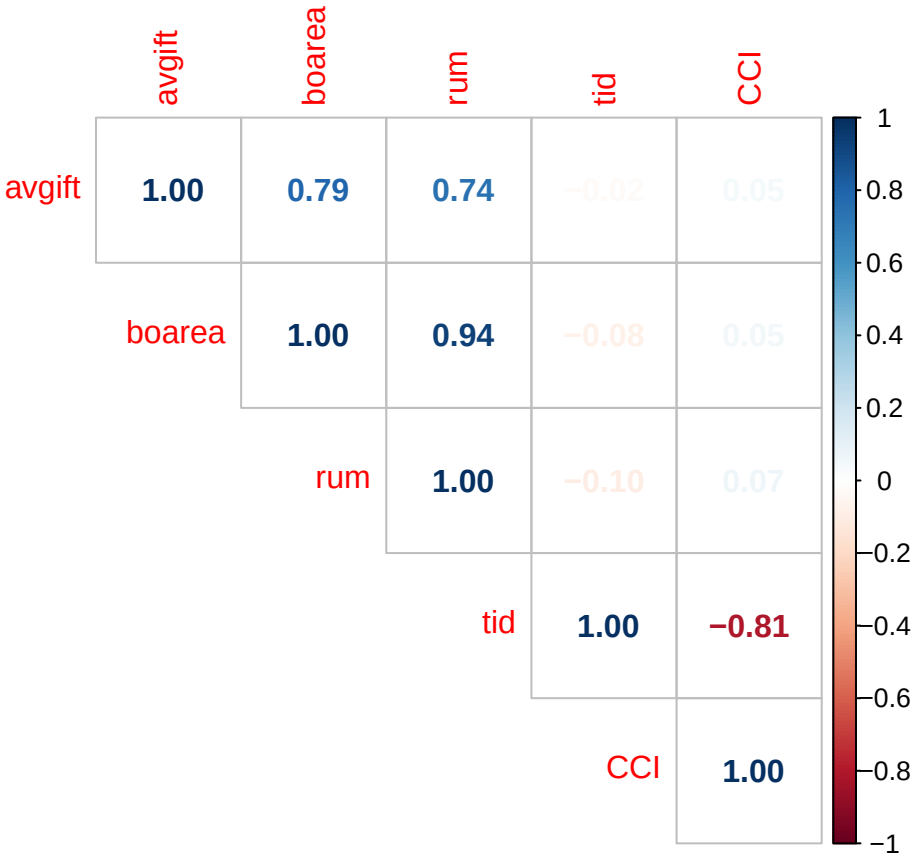
I plotten ovan så ser vi att variablerna mäklare, hiss, garage och balkong är kategoriska variabler som inte visar något tydligt samband med slutpris. Övrigt verkar boarea, avgift och rum ha absolut starkast samband med pris (bortsett från startpris, men den är som tidigare nämnt inte av så stort intresse just nu). Min misstanke är dock att dessa i sig är korrelerade vilket som tidigare nämnt skapar problem med modellen. Vi undersöker då korrealtionen.

```
corr_matrix <- cor(x = as.matrix(A23), use = "pairwise.complete.obs")

corrplot(corr_matrix, type = "upper", method = "number")
```



Plotten visar att vissa korrelationer sticker ut, vi plottar dessa närmare.

```
A23_subset<- A23 %>%
  select(avgift, boarea, rum, tid, CCI)
corr_matrix <- cor(x = as.matrix(A23_subset), use = "pairwise.complete.obs")
corrplot(corr_matrix, type = "upper", method = "number")
```



Avgift, boarea, rum och avgift har alltså alla ett starkt beroende, något väntat. Mindre väntat för mig var kopplingen mellan tid och CCI(hushållens köpbenägenhet). Med all information vi hitills fått är det läge att se vilka förklarande variabler vi kan antingen ta bort eller göra lämplig transformation av. En sista bit innan vi går in på det specifikt är däremot att undersöka kateogriska variablernas inverkan separat då deras plottar inte gav något tydligt svar, svaret vi får är att ingen har ett R^2 värde större än 0.03. Därav ska vi undersöka om vi kan exkludera dem utan att minska R^2.

```
#OBS För att inte printa alla summaries kommenterar jag Adjusted R-squared bredvid modellen.
m_garage <- lm(pris ~ garage, data = A23) #Adjusted R-squared: 0.02759
m_hiss <- lm(pris ~ hiss, data = A23) #Adjusted R-squared:  0.0238
m_vaning <- lm(pris ~ vaningsplan, data = A23) #Adjusted R-squared:  0.01936
m_balkong <- lm(pris ~ balkong, data = A23) #Adjusted R-squared:  0.027
m_maklare <- lm(pris ~ maklare, data = A23) #Adjusted R-squared: -0.002294
```

```
A23_new <- subset(A23, select = -c(2, 9, 10, 11))
full.model.A23_new <- lm(pris ~., data = A23_new)
summary(full.model.A23_new)
```

```
Call:
lm(formula = pris ~ ., data = A23_new)

Residuals:
     Min       1Q   Median       3Q      Max
-1115191  -141540   -38859    82830  1259820

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)  1.487e+06  8.706e+05   1.708  0.08857 .
boarea       4.052e+03  2.006e+03   2.020  0.04408 *
rum          5.939e+04  3.414e+04   1.740  0.08277 .
avgift      -1.360e+01  1.718e+01  -0.792  0.42898
CCI          8.627e+03  1.517e+03   5.685 2.69e-08 ***
startpris    8.912e-01  2.710e-02  32.891  < 2e-16 ***
byggnadsar  -1.007e+03  4.607e+02  -2.186  0.02946 *
vaningsplan  2.994e+04  7.994e+03   3.745  0.00021 ***
tid          1.993e+04  2.520e+03   7.909 3.19e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 243400 on 362 degrees of freedom
  (58 observations deleted due to missingness)
Multiple R-squared:  0.9787,    Adjusted R-squared:  0.9783
F-statistic:  2083 on 8 and 362 DF,  p-value: < 2.2e-16
```

I koden ovan har vi exkluderat de kategoriska variablerna och ser att adjusted $R^2 = 0.9783$ medans det tidgare var 0.979, alltså en helt obetydlig skillnad. Tar vi däremot bort variablen startpris så ser vi att R^2 faller rejält till 0.9103, vi får även extremare fall av residualavstånd där den ökat till 2 300 000 från 1 300 000. Denna uppoffring anser jag däremot värt det av tidigare diskuterade skäl. Vi går nu vidare för att åtgärda kolineratiten.

```
A23_new <- subset(A23_new, select = -c(6))
full.model.A23_new <- lm(pris ~., data =  A23_new)
summary(full.model.A23_new) # R värde på 0.9103 .. Residual standard error: 495000
```

```
Call:
lm(formula = pris ~ ., data = A23_new)

Residuals:
     Min       1Q   Median       3Q      Max
-1380321  -318319    15774   279178  2348902

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 9191046.18 1655282.59   5.553 5.35e-08 ***
boarea        54814.33    2561.27  21.401  < 2e-16 ***
rum          -54411.48   67692.51  -0.804  0.42202
avgift         -112.71      33.54  -3.360  0.00086 ***
CCI           19461.57    2952.22   6.592 1.48e-10 ***
byggnadsar    -5279.26     872.74  -6.049 3.54e-09 ***
vaningsplan  110056.61   15269.09   7.208 3.17e-12 ***
tid           31449.01    5010.85   6.276 9.61e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 495000 on 374 degrees of freedom
  (47 observations deleted due to missingness)
Multiple R-squared:  0.912, Adjusted R-squared:  0.9103
F-statistic: 553.6 on 7 and 374 DF,  p-value: < 2.2e-16
```

```
model_without_rum <- lm(pris  ~ boarea + avgift + CCI + byggnadsar + vaningsplan + tid, data = A23_new)
model_without_boarea <- lm(pris  ~ rum + avgift + CCI + byggnadsar + vaningsplan + tid, data = A23_new)
```

```
#summary(model_without_rum) #Adjusted R-squared:   0.91 #Residual standard error: 494800
#summary(model_without_boarea) #Adjusted R-squared:  0.8011
```

När vi tog bort rum som förklarande variabel (då den hade en hög korrelation med boarea) så ser vi att förklaringsgraden endast sjunker minimalt(från 0.9103 till 0.91), inte bara det utan residualernas standardfel minskar till och med (från 495000 till 494800)! Vi får alltså en förenkling av modellen utan någon riktig uppoffring. I nästa steg tar vi bort resterande variabler som vi tidigare såg korrelera, dvs avgift.

```
simplified.model <- lm(pris  ~ boarea  + byggnadsar + vaningsplan + CCI + tid, data = A23_new)
# summary(simplified.model) fick Adjusted R-squared = 0.91.
```

Vi ser nu att adjusted $R^2 = 0.8878$, alltså en minimal minskning. Standard felet för residualerna ökar sedan från 494800 till 553100. Det är alltså rätt obetydliga förändringar. Däremot har min residualens avstånd ökat markant, nu är den så mycket som -4.5 miljoner. En modell som har en sådan skillnad mellan prediktion och uppmät värde är oroande, framför allt när priset rör sig mellan ungefär 1 till 10 miljoner kr. Något lugnande är däremot att första andra kvantilen har relativt låga värden samtidigt som medianen är nära noll. I helhet är ändå förenklingen att föredra.

Som vi tidigare har sett så är tid och CCI starkt korrelerade, ett sätt att komma tillrätta med det problemet men fortfarande låta båda faktorer vara med i modellen är genom att ta deras medelvärde, vi testar hur det påverkar förklaringsgraden.

```
A23_new <- A23_new %>%
  mutate(CCIplusTid = (CCI + tid)/2)

simplified.model_2 <- lm(pris  ~ boarea  + byggnadsar + vaningsplan + CCIplusTid, data = A23_new)
#summary(simplified.model_2) #Adjusted R-squared:  0.8837

#sum(is.na(A23$CCI)) Här får vi att 39 observationer saknas när det kommer till CCI
```

Vi ser då endast en minimal minskning då vi nu har att adjusted $R^2 = 0.8837$. Däremot så bidrar inte någon av dessa variabler till förklaringsgraden, varken enskilt eller tillsammans. Kombinationen av bristen på bidrag till förklaring, önskan av en så enkel modell som möjligt samt skälet att CCI innehåller flera saknade värden (39) gör att vi utesluter dessa från får slutgiltiga modell. Observera att saknade värden för en variabel gör att motsvarande observationer för övriga variabler raderas i modellen och vi vill bygga regressionen på så mycket data som möjligt

```
final.model <- lm(pris  ~ boarea + byggnadsar + vaningsplan, data = A23_new)
summary(final.model)
```

```
Call:
lm(formula = pris ~ boarea + byggnadsar + vaningsplan, data = A23_new)

Residuals:
     Min       1Q   Median       3Q      Max
-4595944  -320981    -9589   300900  2663039

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 9519069.1  1768243.5   5.383 1.22e-07 ***
boarea        47322.2      892.5  53.020  < 2e-16 ***
byggnadsar    -5012.3      921.0  -5.442 8.98e-08 ***
vaningsplan  106190.6    16920.1   6.276 8.72e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 576600 on 417 degrees of freedom
  (8 observations deleted due to missingness)
Multiple R-squared:  0.8737,    Adjusted R-squared:  0.8728
F-statistic: 961.6 on 3 and 417 DF,  p-value: < 2.2e-16
```

När vi nu har bestämt oss för dom förklarande variablerna vi önskar inkludera så är det dags att mäta deras kolinearitet med VIF-faktorn där $VIF = \frac{1}{1-R_j^2}$ där R_j^2 är ett numeriskt värde för hur mycket av variationen av den förklarande variabeln x_j som förklaras av de andra x-variablerna. Alltså så är ett VIF värde på 1 önskvärt.

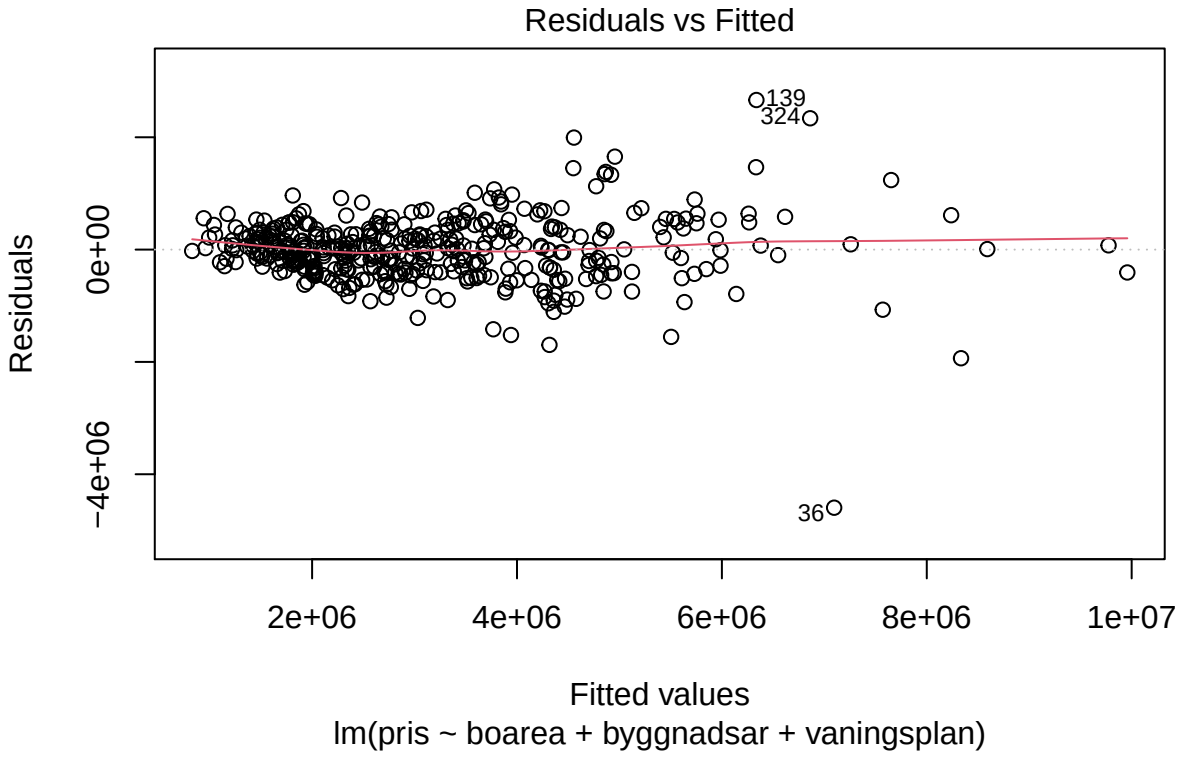
```
vif(final.model)
```

```
    boarea  byggnadsar vaningsplan
  1.015797    1.098436    1.087880
```

Tabellen ovan ger strålande resultat, alla förklarande variablerna är nästintill 1 så kolineratiten är till hög grad åtgärdad!

När nu det finns ett framresonerart antagande om vilka variabler vi önskar att inkludera så bör vi även undersöka hur bra vår ansatta modell uppfyller antagandena för en linjär regression. I linjär regression har vi att $Y = A\theta + \epsilon$ där $\epsilon \sim N(0, \sigma^2)$. Med andra ord så vill vi undersöka hurvida residualerna är normalfördelade med väntevärde 0 eller inte. Ett första steg är att att plotta residualerna för att undersöka heteroskedasticitet.

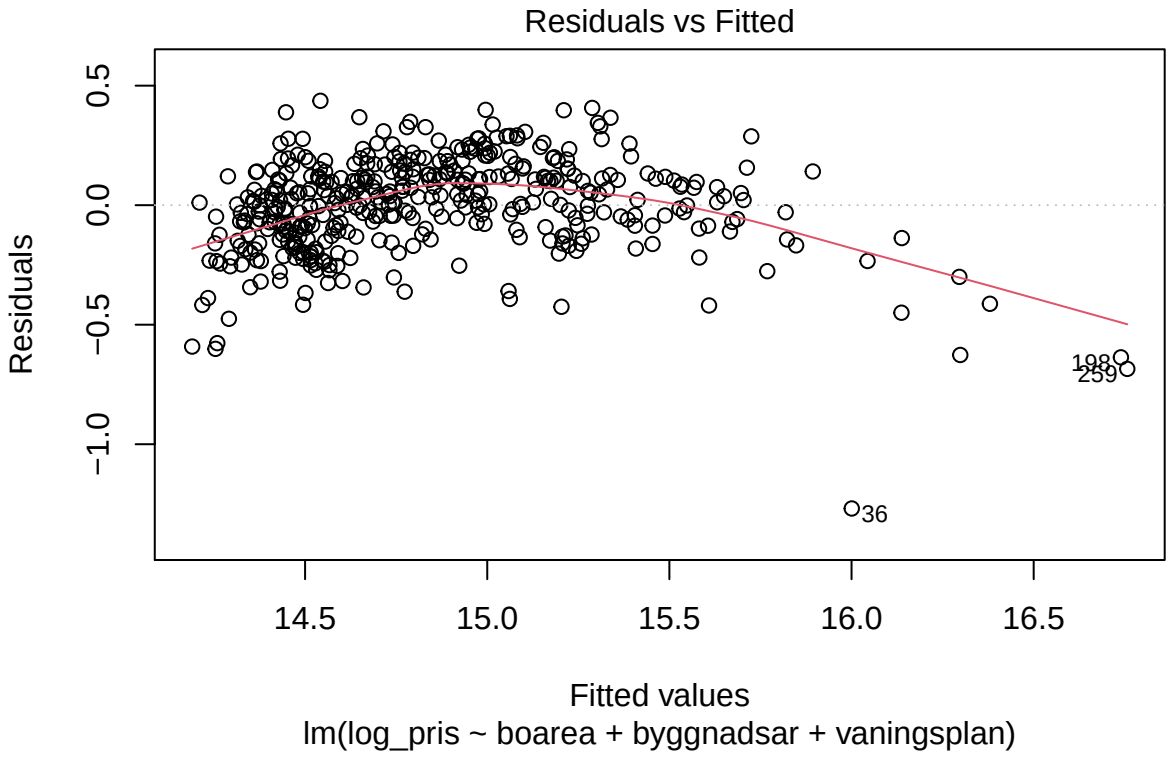
```
plot(final.model, 1)
```



Ovanstående graf visar att residualerna varierar mer desto större pris vi utgår från. Alltså är variansen inte konstant vilket vår tidigare antagande kräver. Åtgärden för en ökad variation för högre pris värden kan vara att logaritmera vår responsvariabel.

```
A23_new <- A23_new %>%
  mutate(log_pris = log(pris))
model_log <- lm(log_pris ~ boarea + byggnadsar + vaningsplan, data=A23_new)
```

```r
plot(model_log, 1)
```



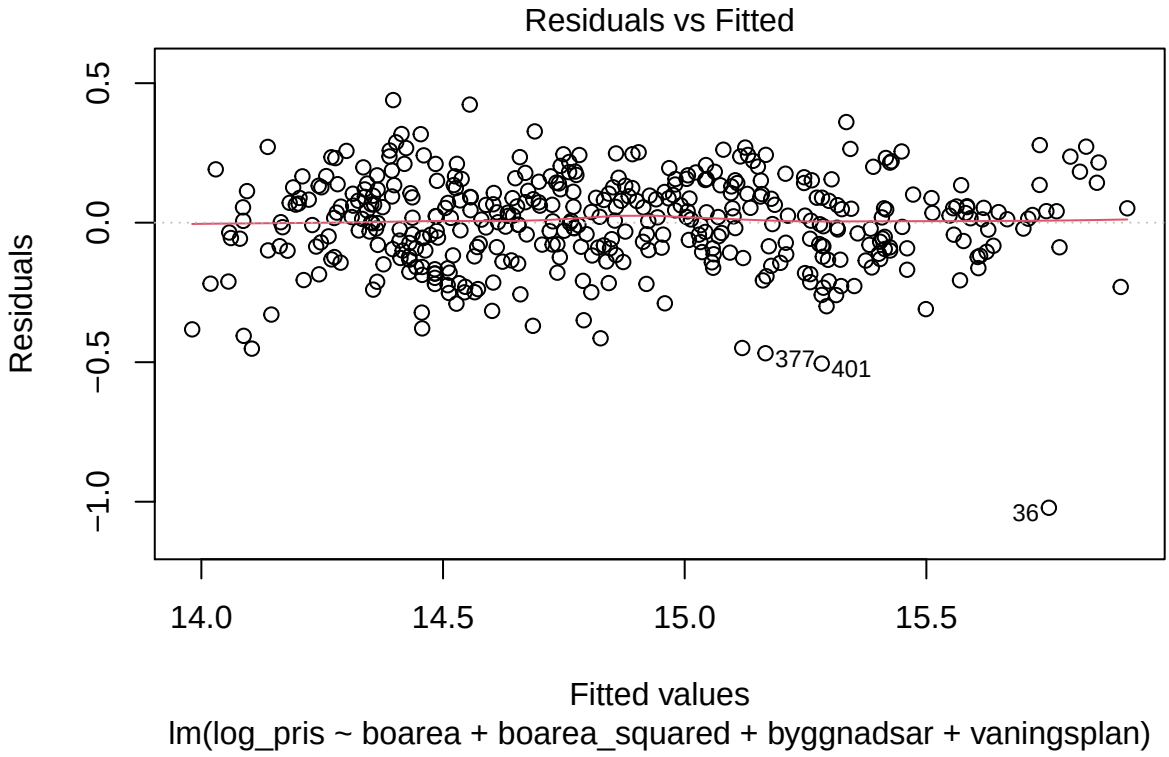
Nu ser vi att variansen inte längre ökar längst x-axeln, men variansen kan inte heller sägas vara konstant, den ser snarare ut att följa en kurva. Ett sätt att åtgärda det är genom att ansätta en av de föklarande variablerna med ett polynomt samband (istället för linjärt) med responsvariabeln. Tidigare har vi sett att boarea är den mest signifikanta av förklaringsvariablerna så vi ansätter ett polynom av den

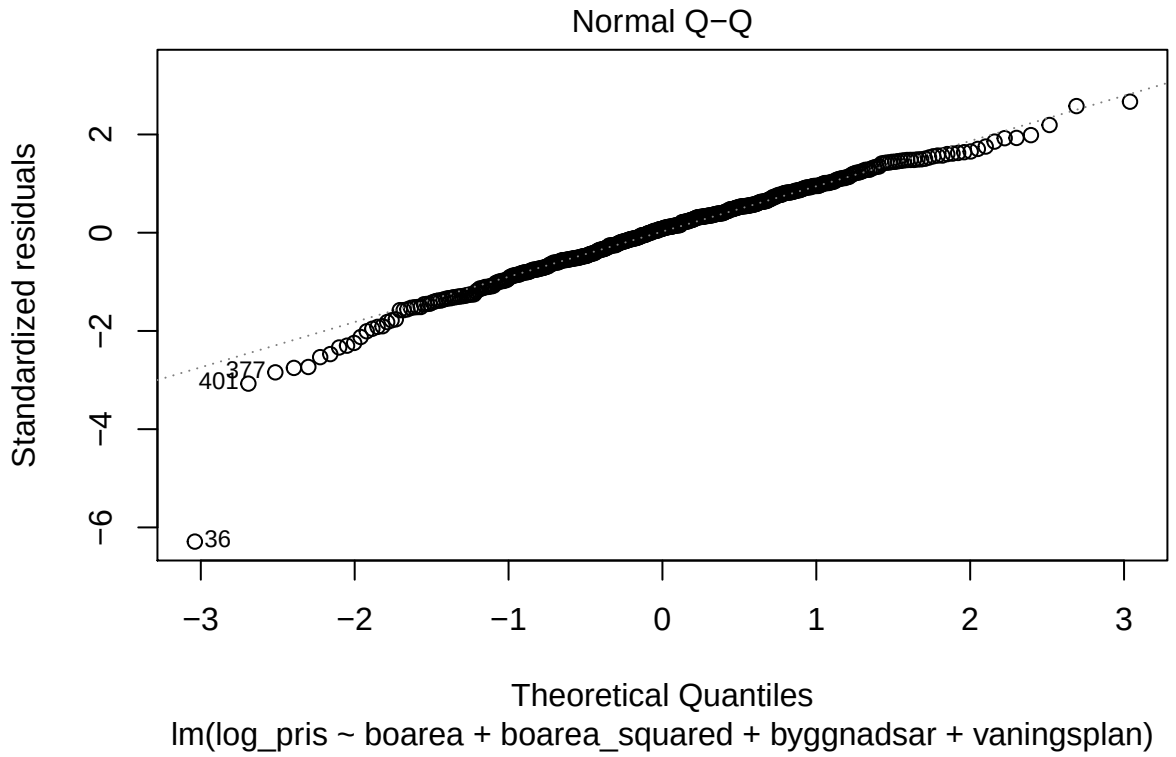
```r
A23_new <- A23_new %>%
  mutate(boarea_squared = boarea**2)

model_log_new <- lm(log_pris ~ boarea + boarea_squared + byggnadsar + vaningsplan, data=A23_new)
summary(model_log_new)
```

```
Call:
lm(formula = log_pris ~ boarea + boarea_squared + byggnadsar +
    vaningsplan, data = A23_new)

Residuals:
     Min       1Q   Median       3Q      Max
-1.02169 -0.09829  0.01348  0.10615  0.43933

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     1.627e+01  5.156e-01  31.556  < 2e-16 ***
boarea          2.609e-02  9.340e-04  27.937  < 2e-16 ***
boarea_squared -7.481e-05  5.434e-06 -13.767  < 2e-16 ***
byggnadsar     -1.443e-03  2.719e-04  -5.305 1.83e-07 ***
vaningsplan     2.639e-02  4.862e-03   5.428 9.71e-08 ***
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1651 on 416 degrees of freedom
  (8 observations deleted due to missingness)
Multiple R-squared:  0.8832,    Adjusted R-squared:  0.882
F-statistic: 786.1 on 4 and 416 DF,  p-value: < 2.2e-16
```

```r
plot(model_log_new, 1:2)
```





Till sist har vi då residualer med konstant varians. I Normal Q-Q plotten så kan vi även se att reisudalerna ser ut att passa väl längst linjen, linjen som innebär normalfördeling för residualerna. Vi kan även observera att observation 36 har ett kraftigt avstånd till den tänkta linjen och bidrar antagligen till att flera andra punkter underskattas. Behövligt är då en undersökning av observationen.

```
A23_new[36, ]
```

```
      pris boarea rum avgift  CCI byggnadsar vaningsplan tid CCIplusTid
36 2500000    150   5     NA 18.1       1942           2   9      13.55
   log_pris boarea_squared
36  14.7318          22500
```

Här ser vi att vi har en observation som sticker ut, priset är "endast" 2.5 miljoner för 150 kvm, kan det förklaras genom avgift? Det är en fråga jag hade velat ha svar på men eftersom informationen saknas så är detta omöjligt. Här får vi även en förklaring till varför reisudalfelen förstorades när vi inte inkluderade avgift, det var helt enkelt så att observationen ströks då. Frestande är att utesluta observationen, men samtidigt vet vi inte upphovet till varför priset blivit så högt (även om vi gissar att det har att göra med avgift) så det finns inget bra skäl till exklusion.

Vi har då vår slutgiltiga modell.

```
summary(model_log_new)
```

```
Call:
lm(formula = log_pris ~ boarea + boarea_squared + byggnadsar +
    vaningsplan, data = A23_new)

Residuals:
     Min       1Q   Median       3Q      Max
-1.02169 -0.09829  0.01348  0.10615  0.43933
```

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     1.627e+01  5.156e-01  31.556  < 2e-16 ***
boarea          2.609e-02  9.340e-04  27.937  < 2e-16 ***
boarea_squared -7.481e-05  5.434e-06 -13.767  < 2e-16 ***
byggnadsar     -1.443e-03  2.719e-04  -5.305 1.83e-07 ***
vaningsplan     2.639e-02  4.862e-03   5.428 9.71e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1651 on 416 degrees of freedom
  (8 observations deleted due to missingness)
Multiple R-squared:  0.8832,    Adjusted R-squared:  0.882
F-statistic: 786.1 on 4 and 416 DF,  p-value: < 2.2e-16
```

Till slut har vi då fått adjusted R^2 på ungefär 0.9, en hög förklaringsgrad! Vi kan även notera att vi har utmärkta p-värden för att motsägga nollhypotesen att lutningskoefficienterna $\beta_i = 0$.

Sista transformationen vi utförde har försvårat tolkbarheten något då vi nu har boarea som påverkande koefficient två gånger om, vi måste även självklart inse att koefficienterna från ovan inbegriper ett logaritmerat pris. Hur som helst så kan vi läsa av att det generellt gäller att äldre lägenheter är dyrare. Något mer väntat vi ser är även att desto högre våningsplan, desto högre pris, men det behövde man knappast någon undersökning för att bekräfta.

Utöver tidigare diskussioner om samband mellan variabler och andra brister i datan så kan vi till sist även lägga till att den förklarande modellen vi använder här är baserat på ett specifikt datamaterial. Resultaten från den här regressionen är inte till för att förstå sig på lägenhetspriser under större tidperioder från flera områden utan ska bäst förklara data med liknande/samma bakgrund som denna. Vi kan även dra slutsattser från de stora residualfelen som dök upp, när vi undersökte en sådan saknade information som kunde förklara ett orimligt billigt pris. Var det ett renoveringsprojekt? Var det hög hyra? Fler variabler än de vi inkludera kan behövas för att framförallt förklara extremfallen, alltså kan modellen ge en generell förklaring men människan som tolkar den bör vara medveten om den stora skillnaden (från modellens predikton) som kan uppstå när en icke registrerad variabel påverkar priset.

Resultat

Genom analys av hur observationerna påverkar varandra och deras på verkan på modellen har kunnat komma fram till en enklare multipel linjär regressionsmodell som väl förklarar slutpriset av lägenheterna.