Projekt inom linjär regression
FMSF80 Matematisk statistik, allmänn kurs

Linnea Munter, Edvin Gunnarsson och Casper Sjöström

November 2024

1 Inledning och frågeställning

Att kunna förutsäga fastighetsvärderingar med viss statistisk säkerhet är grundläggande för att kunna fatta välgrundade investreringsbeslut. I det här fallet undersöks värderingen av ett begränsat antal villor i Ulricht baserat på fyra olika parametrar.
Målet är att sortera ut och anpassa de variabler som har en tydlig koppling mot priset och skapa en linjär regressionsmodell. Sedan använda modellen för att ge en rimlig prisprediktion med prediktionsintervall för en villa med tomtarea på 150 kvm, boyta på 175 kvm, utan balkong byggd 2022. Dessutom ange sannolikheten att huset säljs för minst 875 000€.

2 Teori/ metod

Grunden i våran modell kommer vara en multipel regressionsmodell på formen:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 ... \beta_n x_n + \epsilon \tag{1}$$

Där α är en konstant, x_n är en variabel (ex. boyta), β_n är koffecienten och ϵ är residualen som beskriver hur mycket en mätning avviker från den skattade modellen. Detta kan också skrivas på matrisform:

$$Y = X\beta + \epsilon \tag{2}$$

Variablerna har samma betydelse men α har substituerats mot en extra kolumn ettor i X matrisen. Standardmetoden för att beräkna koeffcienten och interceptet görs genom minsta kvadratmetoden:

$$\hat{\beta} = (X^T X)^{-1} X^T y \tag{3}$$

$\hat{\beta}$ kommer då vara en kolumnvektor där $\hat{\beta_0}$ är interceptet och resterande $\hat{\beta_n}$ är koefficienterna tillhörande respektive variabel.

Först scatterplotas alla parametrar (boyta, tomtarea, år, balkong) mot priset för att visuellt avgöra vilka variabler som är releventa att ta med utifrån de mönster de uppvisar i datan. Linjära korrelationer kan direkt användas i regressionsmodellen. Data som uppvisar andra mönster behöver korrigeras utifrån pylonom eller andra godtyckliga funktioner. Sedan kan en mer rigorös kontroll utföras genom att scatterplotta residualerna mot x. Resultatet ska då framstå som slumpmässigt och inga mönster gå att urskilja. Residualerna ska också vara $\epsilon \in N(0, \sigma)$. Så länge något av dessa mått är dåliga så får man antingen transformera om datan alternativt förkasta den helt.

För att skapa själva regressionsmodellen delas husdatan upp i två delar. 80 procent används för att träna modellen resterande 20 används för att kontrollera den framtagna modellens korrekthet. När datan valts ut används frivilligt beräkningsprogram i det här fallet pythonbiblioteket Statsmodelför att skapa den multipla regressionsmodellen. För att avgöra vilka variabler som är signifikanta så kan koefficienterna hypotesprövas. Antingen antas H_0 alltså att $\hat{\beta} = 0$ eller H_n, vilket innebär: $\beta \hat{\neq} 0$. Det finns två sätt att hypotespröva (Statsmodel ger tre men endast två kommer tas upp här). Enligt det första sättet defineras en teststorhet t enligt nedan:

$$t_\beta = \frac{\hat{\beta}}{d(\hat{\beta})} \tag{4}$$

Där $\hat{\beta}$ är skattade koefficienten och $d(\hat{\beta})$ är standardfelet. Sedan beroende på vilka krav man har på modellen fastställer man ett α som avgör vilken noggranhet som uppnås. Om sannolikheten att finna $\mid t \mid$ är mindre än α förkastas nollhypotesen och variabeln anses ha signifikans.

Andra metoden tittar på konfidensintervallet för $\hat{\beta}$ istället för direkt på t kvantilen. Om 0 är inanför vårt valda konfidensintervall baserat på det α som väljs så förkastas variabeln. Matematiskt beskrivet så ges konfidensintervallet av:

$$I_\beta = [\hat{\beta} - t_{\alpha/2} \cdot d(\hat{\beta}), \hat{\beta} + t_{\alpha/2} \cdot d(\hat{\beta})] \tag{5}$$

Om ovan givna parametrar är inom acceptabla intervall så kan modellen användas för att predikera pris utifrån givna parametrar angivna i inledningen.

För att beräkna sannolikheten att huset säljs för en viss summa tas ett $t_{\hat{y}}$ värde fram som beskrivs enligt följande formel:

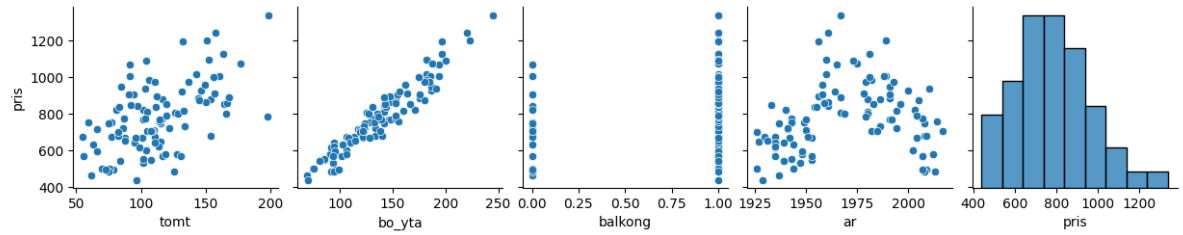
$$t_y = \frac{y^* - \hat{y}}{d(\hat{y})} \tag{6}$$

Där ekvationen tar skillnaden i modellens skattning med vårt valda värde dividerat med med medelfelet, vilket blir en standardiserad normalfördelning. Sedan används den standardiserade normalmfördelnigen för att beräkna sannolikheten enligt:

$$P(X \geq x) = 1 - \Phi(t_{\hat{y}}) \tag{7}$$

3 Analys

Först plottades priset enligt figur 1 mot de beroende variablerna och sig självt för att få en uppfattning om hur priset var fördelat.



Figur 1: Priset plottat mot sig själv i tusen euro och variablerna, vilka är: år, boyta m^2, tomtarea m^2 och, balkong

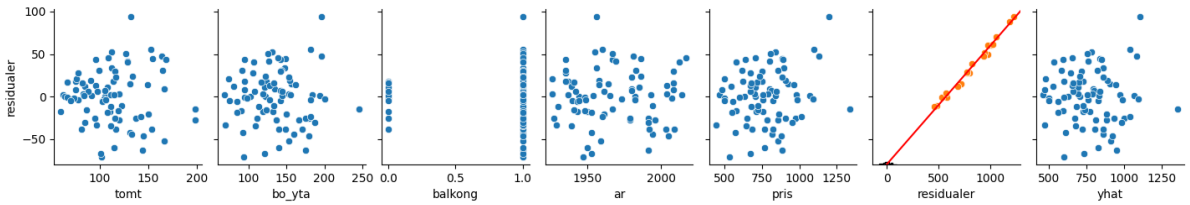
Från detta kan man tydligt se att priset över bo och tomtytan ger ett linjärt beroende. Där tomtytan har ett svagare samband eftersom värdena har större spridning. Modellen kommer därför inkludera både boyta och tomtstorlek utan att transformera dem.

När det kommer till åldern på huset så ökar priset först för att sedan avta. Därför testades ett andragrads polynom. Detta ledde till förbättring men en ännu bättre modell gavs av att lägga till en linjär komponent utöver den polynomiala. Transformationen blev alltså: $x^2 + x$.

Grafen för pris som funktion av balkong är diskret då huset antingen har balkong eller inte har det. Därför är det rimligt att anta att det finns en korrelation då det enligt figur 1 är de dyraste husen som har balkong.
För att undersöka variablernas signifikans görs en hypotesprövning, teorin togs upp i föregående kapitel. Detta ger P under den föredragna gränsen på $0,05$ för både tomtarea och boyta. Balkong däremot ger ett P värde på $0,895$ tydligt över 0.05. När balkong exkluderas från modellen ändras även P värdet för tomt från $0,042$ till $0,010$. Dess påverkan på tomtvariabeln tyder på korrelation mellan variablerna och gör att det blir svårare att isolera dem vilket i sin tur ger mindre precisa koefficienter. Eftersom balkongens korrelation med priset är lågt och den har negativ effekt på signifikansen för tomtytan så förkastas balkong helt som varibel. När residualerna plottas över våra olika variabler enligt figur 2 kan det för samtlig variabler, d.v.s yhat (predikterade värden), tomt, boyta och år, ses att punkterna är slumpmässigt fördelade kring 0, vilket enligt teorin ovan stärker lämpligheten för våra variabler.

Enligt teoridelen ska residualerna vara $N(0, \sigma)$. Statsmodell ger ett antal värden som beskriver hur väl detta stämmer. Dessa är Skewedness", Ömnibussöch "Kurtosis". Inga matematiska definitioner kommer ges utan endast beskrivning av deras funktion.

Skewedness"är ett mått på hur centrerad runt noll normalfördelning av residualerna är. Ett värde så nära noll som möjligt är alltså önskvärt. Kurtosis är ett mått på extremvärdena i normalfördelningen om Kurtosis = 3 så är den identisk med normalfördelning om Kurtosis är större än 3 är det en indikation



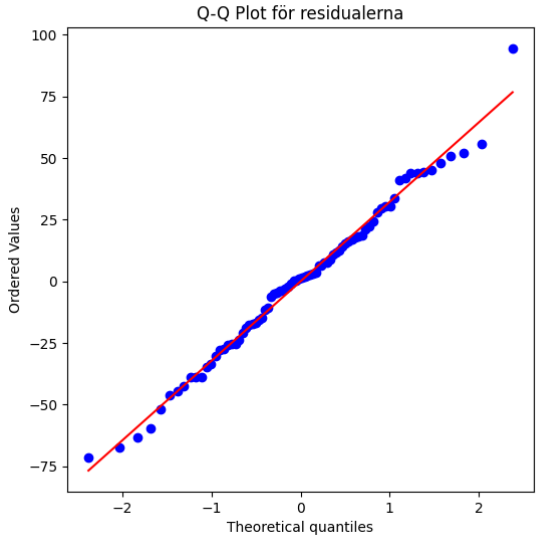
Figur 2: Residual spridningen plottade för varje enskild variabel

på att det finns värden som avviker mycket och att modellen ej är optimal. Omnibus tar både Skewedness och Kurtosis och ger ett sammanvägt värde av hur väl residualerna är normalfördelade. Här är ett värde $p > 0.05$ önskvärt. Statsmodell ger också ett värde kallat R^2. Det är ett mått på hur väl modellen förklara huspriset. om $R^2 = 1$ betyder det att modellen beskriver huspriset perfekt utan avvikelser. Ett $R^2 = 0$ betyder i sin tur att modellen inte alls lyckas korrelera variblerna med huspriset. Som kan ses i tabell 1 så är parametrar inom önskade interval med ett R^2 nära 1.

Skew	Kurtosis	Omnibus	R^2
0.06	3.049	0.255	0.967

Tabell 1: Uträknade värden från modellen

Modellen kontrollerades ytterligare en gång nu grafiskt för att helt säkerställa normalfördelningen hos residualerna. Först plottades värdena på en QQ-plot som är normerad så att linjen x = y följer en normalfördelning, se figur 3.



Figur 3: QQ-plot som visar hur väl residualerna är normalfördelade

Om punkterna i QQ-plotten ligger nära den diagonala linjen (x = y), tyder det på att residualerna är ungefär normalt fördelade. Om punkterna avviker från linjen eller uppvisar något annat mönster exempelvis ett polynom så indikerar det att residualerna inte är normalfördelade och att modellen inte uppfyller kraven. Vid granskning av QQ-ploten i figur x framgår det att punkterna ligger nära linjen.

När konfidensen hos vår modell är säkerställs kan frågeställningen från inledningen besvaras: Statsmodel ger ett uppskattat pris på 822 000€ . Med en sannolikhet på 7.6% att huset säljs för 875 000€ eller mer.

4 Slutsatser och sammanfattning

Vi har utvecklat en regressionsmodell för att förklara och förutsäga huspriser som en funktion av boyta, tomtarea och byggår. Vi identifierade att boyta och tomtarea har linjära samband med huspriset och direkt kunde tas in i modellen. Byggår visade sig ha ett icke-linjärt samband och modellerades med en polynomiell komponent i kombination med en linjär komponent för att bättre fånga mönstret. Variabeln balkong valde vi att förkasta då den visade låg påverkan på priset samtidigt som den hade negativ påverkan genom korrelation till andra variabler, särskilt tomtarean. Vi utförde därefter en residualanalys som påvisade att vår modell uppfyllde kravet på normalfördelning. Vi kollade värden såsom "Omnibuss", Skewness", "Kurtosis" samt R^2 och såg att dessa låg på rimliga och lämpliga värden vilket påvisade att vår modell var lämpad till det den var designad för. Slutligen matade vi in datan och fick det förväntade huspriset till 822 000€ med ett prediktionsintervall mellan 748 162€ och 897 302€ samt en sannolikhet på 7,6% att priset är 875 000 € eller högre. Förbättringar till nästa gång skulle kunna inkludera ett nytt försök att implementera balkongen men nu med någon typ av transformation av variabeln för att få datan att passa in. Andra förbättringar skulle kunna innebära ett större mängd data som tar hänsyn till fler parametrar. Exempelvis en variabel som tar i beaktande tillståndet på bostadsmarknaden.

5 Använding av AI

Ai har inte använts för att producera någon text eller grafer. Ai har däremot utnyttjatts för att förklara statistiska koncept och ekvationer. Men dessa har sedan kontrollerats mot trovärdiga källor såsom läroböcker och föreläsningsanteckningar.

6 Författarbidrag

Härledningar: Edvin Gunnarsson
Analys: Casper Sjöström, Edvin Gunnarsson och Linnea Munter
Programmering: Linnea Munter
Visualisering : Casper Sjöström och Linnea Munter
Skrivande — Originalutkast: Casper Sjöström
Skrivande — Revision o editering: Edvin Gunnarsson och Linnea Munter
Projektledning: Linnea Munter