Förutsägelse av försäljningspris för Volvobilar

En tillämpning av linjär regression och dataanalys

Mahad Elmi

EC Utbildning

Kunskapskontroll R

202504

Abstract

This project aimed to predict car prices listed on Blocket using linear regression modeling. The data was prepared by handling missing values, removing outliers, and creating relevant variables such as car age. Exploratory data analysis revealed important relationships between factors like mileage, horsepower, and price. The final regression model demonstrated good predictive performance with an R^2 of approximately 0.77, an RMSE of 0.191, and an MAE of 0.143 on the test set. The results highlight how statistical modeling can support accurate car price predictions on online marketplaces.

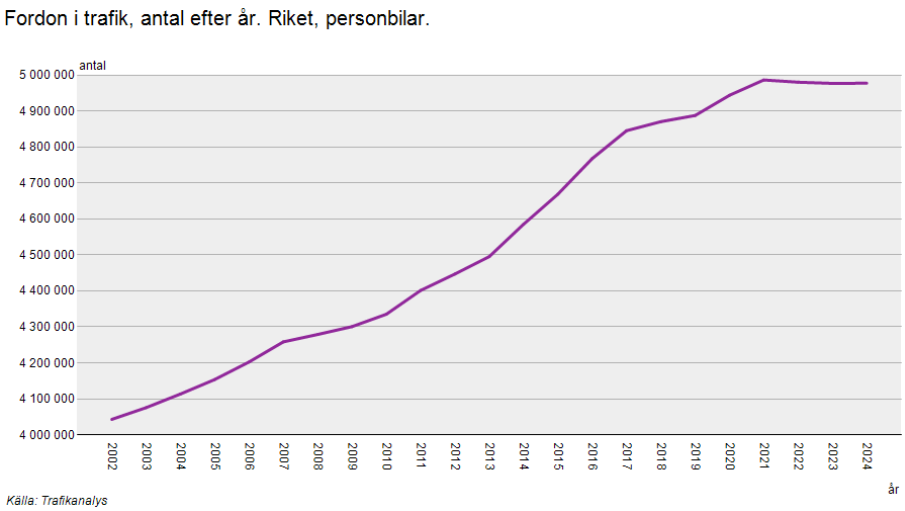
Innehåll

1 Inledning

Bilmarknaden är en av de största och mest dynamiska delarna av Sveriges ekonomi. Enligt statistik från SCB sker årligen ett omfattande antal bilaffärer, både av nya och begagnade fordon, vilket gör korrekt prissättning till en central fråga för både säljare och köpare. Priset på en bil påverkas av en rad faktorer såsom ålder, körsträcka, hästkrafter, bränsletyp och växellåda. Att kunna förstå och förutsäga dessa prisvariationer är därför av stort värde för privatpersoner, bilhandlare och andra aktörer på marknaden.

Med ökningen av digitala köp- och säljsajter såsom Blocket.se har tillgången till data om bilars egenskaper och försäljningspriser blivit större än någonsin. Detta öppnar möjligheter att använda statistiska modeller för att förutse bilars försäljningspris baserat på objektiva egenskaper, något som kan skapa mer transparens och effektivitet på marknaden.

Denna rapport fokuserar på att analysera ett insamlat datamaterial från Blocket.se med hjälp av statistisk modellering, specifikt en multipel linjär regressionsmodell, för att undersöka vilka faktorer som har störst påverkan på försäljningspriset, samt bygga en prediktiv modell som kan uppskatta priset på en bil givet dess egenskaper. Metoderna och principerna som tillämpas i detta arbete baseras på teorier och tekniker från *An Introduction to Statistical Learning* (James et al., 2021) samt *R for Data Science* (Wickham & Çetinkaya-Rundel, 2023).



Figur 1. Utvecklingen av antalet registrerade personbilar i trafik i Sverige från 2002 till 2024, vilket visar på en kontinuerlig tillväxt i fordonsbeståndet.

1.1 Syfte och frågeställning

Syftet med denna rapport är att utveckla en statistisk modell som kan prediktera bilars försäljningspris baserat på olika egenskaper såsom miltal, hästkrafter och bilens ålder.

För att uppfylla syftet kommer följande frågeställningar att besvaras:

1. Vilka variabler har störst påverkan på bilars försäljningspris?

2. Hur väl kan en multipel regressionsmodell prediktera försäljningspriset på bilar baserat på tillgängliga egenskaper?

2 Teori

2.1 Explorativ Dataanalys (EDA)

Explorativ dataanalys används för att undersöka datastrukturen och identifiera mönster, avvikelser och samband mellan variabler. Visualiseringar som histogram, stapeldiagram och spridningsdiagram hjälper till att skapa en bättre förståelse för datamaterialet (Wickham & Grolemund, 2023).

2.2 Dataförberedelse

Korrekt hantering av saknade värden och kategoriska variabler är centralt innan en modell kan byggas. Saknade värden kan ersättas med exempelvis medianvärden eller kategorin "Okänd", och kategoriska variabler kodas som faktorer i R (Wickham & Grolemund, 2023).

2.3 Outlinerhantering

Outliers kan snedvrida analysen och modellens prediktioner. Ett vanligt sätt att identifiera outliers är genom interkvartilavstånd (IQR), där observationer som ligger utanför 1.5 gånger IQR betraktas som extrema (James et al., 2021). Multikollinearitet innebär att oberoende variabler är korrelerade med varandra, vilket kan leda till instabila koefficienter i regressionsmodeller. Variansinflationsfaktor (VIF) används för att mäta graden av multikollinearitet (James et al., 2021).

2.4 Multikollinearitet och Variansinflationsfaktor (VIF)

Multikollinearitet innebär att oberoende variabler är korrelerade med varandra, vilket kan leda till instabila koefficienter i regressionsmodeller. Variansinflationsfaktor (VIF) används för att mäta graden av multikollinearitet (James et al., 2021).

2.5 Linjär Regression

Linjär regression är en statistisk metod för att modellera sambandet mellan en beroende variabel och en eller flera oberoende variabler. Vid skattning används minsta kvadratmetoden för att hitta den bästa linjen genom datapunkterna (James et al., 2021).

2.6 Modellutvärdering

För att bedöma hur väl modellen presterar används flera mått:

- R^2 (R-squared): Andelen av variationen i den beroende variabeln som kan förklaras av modellen.
- RMSE (Root Mean Square Error): Mäter den genomsnittliga storleken på prediktionsfelen.
- MAE (Mean Absolute Error): Mäter det genomsnittliga absoluta felet mellan faktiska och predikterade värden (James et al., 2021; Wickham & Grolemund, 2023).

3 Metod

3.1 Datainsamling

Data som användes i detta arbete hämtades från Blocket.se, en svensk marknadsplats för försäljning av fordon. Urvalet bestod av bilar från märket Volvo, både nya och begagnade. Datan innehöll variabler som försäljningspris, miltal, motorstorlek, färg, modell, biltyp, drivning, växellåda, med flera. Data erhölls i form av en Excel-fil.

3.2 Dataförberedelse

För att säkerställa att analysen byggde på korrekt och användbar data genomfördes flera förberedande steg. Saknade värden i kategoriska variabler ersattes med kategorin "Okänd", medan numeriska saknade värden, såsom för hästkrafter, ersattes med medianvärdet. Vidare togs vissa irrelevanta kolumner, exempelvis "Motorstorlek" och "Datum_i_trafik", bort från datamängden för att fokusera analysen på relevanta variabler. Samtliga kategoriska variabler faktorerades för att möjliggöra korrekt behandling i regressionsmodellen. För att fånga in ålderns effekt på priset skapades en ny variabel, "Bilens ålder", baserad på modellår. Slutligen transformerades försäljningspriset genom logaritmering för att bättre uppfylla modellens antagande om normalfördelade residualer.

3.3 Explorativ Dataanalys

Visualiseringar i form av histogram, stapeldiagram, spridningsdiagram och boxplots användes för att undersöka variablers fördelning, identifiera eventuella avvikande värden och analysera samband mellan variabler.

3.4 Outlierhantering

Avvikande observationer hanterades med hjälp av interkvartilavstånd (IQR), där observationer utanför 1,5 gånger IQR för variablerna försäljningspris, miltal, bilens ålder och hästkrafter exkluderades från analysen.

3.5 Modellering

En multipel linjär regressionsmodell tränades på 70 % av datan och testades på de återstående 30 %. Den beroende variabeln var den logaritmerade försäljningspriset ("LogPris"), medan oberoende variabler inkluderade miltal, hästkrafter, bilens ålder samt flera kategoriska variabler som växellåda, säljare och biltyp.

3.6 Modellutvärdering

För att bedöma modellens prestanda användes måtten R^2 (förklaringsgrad), RMSE (Root Mean Square Error) och MAE (Mean Absolute Error). Modellens antaganden undersöktes vidare genom analys av residualerna, inklusive QQ-plot, histogram och Cook's Distance för att identifiera eventuella avvikelser eller inflytelserika observationer.

4 Resultat och Diskussion

Den multipla linjära regressionsmodellen som tränades på den förberedda datan visade en god förklaringsgrad. Modellen uppnådde ett R^2-värde på 0,829 på testdatan, vilket innebär att cirka 82,9 % av variationen i bilarnas logaritmerade försäljningspris kunde förklaras av de valda prediktorerna. Även felmåtten RMSE och MAE var relativt låga, vilket indikerar att modellen har en god träffsäkerhet i sina prediktioner.

För att förbättra modellens antaganden om normalfördelning av residualer transformerades försäljningspriset med en naturlig logaritm. Saknade värden hanterades genom att ersätta kategoriska variabler med "Okänd" och numeriska variabler med medianvärdet. Denna förenklade metod säkerställde att alla observationer kunde användas i modellen, men kan samtidigt ha introducerat viss osäkerhet, eftersom "Okänd" kan representera flera olika verkliga förhållanden.

Outliers identifierades och togs bort med hjälp av en IQR-baserad metod. Detta bidrog till en stabilare modell, men innebar också att extrema men legitima datapunkter riskerade att uteslutas. Detta kan begränsa modellens generaliserbarhet till ovanliga biltyper eller prisklasser.

Trots modellens goda resultat finns det flera förbättringsmöjligheter. En viktig aspekt är modellens antagande om en linjär relation mellan prediktorer och målvariabel. I verkligheten kan sambanden vara mer komplexa och icke-linjära. Användning av mer avancerade metoder, såsom random forests, gradient boosting eller neurala nätverk, hade kunnat fånga dessa komplexa mönster och potentiellt förbättra prediktionerna.

En annan utvecklingsmöjlighet vore att använda en mer avancerad metod för hantering av saknade värden, exempelvis multipel imputering, vilket kan minska risken för bias. Dessutom hade en djupare analys av interaktionseffekter mellan variabler (t.ex. mellan bilens ålder och drivtyp) kunnat ge ytterligare förbättringar av modellens förklaringskraft.

Sammanfattningsvis visar arbetet att en väl genomförd multipel linjär regressionsmodell, med noggrann dataförberedelse, kan ge goda prediktioner av bilpriser på en marknadsplats som Blocket. För framtida förbättringar rekommenderas användning av mer flexibla modeller och mer sofistikerad hantering av datans egenskaper.

5 Slutsatser

Syftet med denna studie var att utveckla en prediktionsmodell som kan förutse bilars försäljningspris baserat på information från Blocket.se. Genom att använda en multipel linjär regressionsmodell på en rensad och förberedd datamängd lyckades modellen uppnå en träffsäkerhet (R^2) på 82,9 % på testdatan. Detta indikerar att modellen fångar en stor del av variationen i bilpriserna.

Arbetet visade vikten av noggrann dataförberedelse, inklusive hantering av saknade värden, borttagning av outliers och transformation av variabler. Dessutom visade resultaten att ålder, miltal och hästkrafter var viktiga faktorer för bilens värde.

Även om modellen presterade bra, finns möjligheter till förbättring, till exempel genom att testa mer avancerade modeller som random forest eller gradient boosting, eller genom att inkludera ytterligare beskrivande variabler (exempelvis bilens skick eller utrustningsnivå).

Sammanfattningsvis visar arbetet att linjär regression är ett kraftfullt verktyg för prisförutsägelser inom andrahandsmarknaden, men också att noggrann förberedelse och val av variabler är avgörande för modellens framgång.

6 Teoretiska frågor

1. Vad är en Quantile-Quantile (QQ) plot?

   En QQ plot är ett verktyg för att grafisk jämföra fördelningen av ett dataset mot en teoretisk fördelning, oftast normalfördelningen. Om datapunkterna i QQ plottet ligger nära en rak linje tyder det på att datan följer den förväntade fördelningen. Avvikelser från linjen indikerar att datan avviker från den teoretiska fördelningen.

2. Vad menas med att Maskininlärning fokuserar på prediktion medan regressionsanalys även fokuserar på inferens?

   Maskininlärning har primärt som mål att skapa modeller som kan ge så bra prediktioner som möjligt på ny data, ofta utan att tolka sambanden mellan variablerna. Statistisk regressionsanalys fokuserar både på prediktion och på att dra slutsatser om relationerna mellan variabler, t.ex. om en viss faktor har en signifikant effekt på utfallet. Ett exempel på inferens är att använda p-värden för att bedöma om en viss variabel är viktig.

3. Vad är skillnaden på konfidensintervall och prediktionsintervall för predikterade värden?

   Ett konfidensintervall beskriver osäkerheten i skattningen av det genomsnittliga utfallet för en given kombination av prediktioner. Ett prediktionsintervall är bredare och inkluderar både osäkerheten i skattningen och den naturliga variationen hos enskilda observationer. Med andra ord: konfidensintervall för medelvärdet, prediktionsintervall för en ny observation.

4. Hur tolkas beta parametrarna i en multipel linjär regressionsmodell?

   I en multipel linjär regressionsmodell representerar varje beta-koefficient (β_1, β_2, ..., β_p) den genomsnittliga förändringen i responsvariabeln Y för en enhets förändring i den tillhörande prediktorn X, när alla andra prediktorer hålls konstanta.
   β_0 (interceptet) representerar det förväntade värdet på Y när alla prediktorer är noll.

5. Behöver man använda tränings, validerings, och test set om man använder BIC?

   Om man använder modellurvalskriterier som BIC (Bayesian Information Criterion) kan man i teorin välja en modell utan att behöva dela upp datan i träning, validering och test.
   BIC tar hänsyn till både modellens anpassning (RSS) och komplexitet (antal parametrar).
   Logiken är att BIC straffar överanpassade modeller, vilket minskar risken för överfitting.

Dock kan det i praktiken ändå vara fördelaktigt att använda en separat testmängd för att få en objektiv utvärdering av modellens prediktionsförmåga.

6. Förklara algoritmen för ”Best subset selection”

   Best subset selection innebär att man testar alla möjliga kombinationer av prediktorer för varje antal prediktorer k = 1, 2, ..., p.
   För varje k väljs den modell som har lägst RSS eller högst R^2.
   Sedan väljer man den bästa modellen totalt sett med hjälp av kriterier som valideringsfel, C_p, AIC, BIC eller adjusted R^2.
   Detta gör att man kan identifiera den mest lovande modellen utan att behöva testa alla på ny data.

7. Förklara citatet ”All models are wrong, some are useful”

   Citatet av George Box betyder att alla statistiska modeller är förenklingar av verkligheten och kan aldrig beskriva den exakt — därför är de tekniskt sett alltid "fel".
   Trots detta kan modeller ändå vara mycket användbara om de fångar de viktigaste mönstren eller sambanden i datan.
   En modell behöver alltså inte vara perfekt för att vara praktiskt användbar, exempelvis för prediktion, beslutsstöd eller förståelse av samband.

7 Självutvärdering

1. Utmaningar du haft under arbetet samt hur du hanterat dem.
   En utmaning var att hantera saknade värden och outliers korrekt. Jag använde medianimputering och borttagning med IQR-metoden. Jag hann dock inte genomföra alla moment som krävs för högsta betyg.

Källförteckning

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An Introduction to Statistical Learning (2nd ed.). Springer. Tillgänglig på: https://www.statlearning.com/

Wickham, H., & Çetinkaya-Rundel, M. (2023). R for Data Science (2nd ed.). Tillgänglig på: https://r4ds.hadley.nz/

Statistiska centralbyrån (SCB). (2024). Antal registrerade personbilar i trafik 2002-2024. Hämtad från: https://www.scb.se/

YouTube, StatQuest with Josh Starmer (2016). Quantile-Quantile (QQ) Plots Explained. Tillgänglig på: https://www.youtube.com/watch?v=X9_ISJ0YpGw