Cardata analys i R



Ming Fondberg

EC Utbildning

R kunskapskontroll

2024-04

Abstract

Studien undersöker effektiviteten av en linjär regressionsmodell för att förutsäga bilpriser baserat på ett omfattande dataset med bilattribut som märke, årsmodell, bränsletyp, miltal, och växellåda. Modellen tränades och validerades med hjälp av metoder som korsvalidering och diagnostiska tester för att bedöma dess prediktionsprestanda och identifiera potentiella förbättringsområden. Resultaten visar att modellen har en rimlig förutsägelseförmåga med en genomsnittlig absolut fel (MAE) på 52,312 och en rot av medelvärdet av de kvadrerade felen (RMSE) på 74,306, vilket indikerar en god passform för många observationer men även betydande avvikelser för högre prisintervaller. Dessa avvikelser tyder på att modellen kan behöva ytterligare finjusteringar för att hantera dyrare bilar effektivare. Studien belyser också vikten av att inkludera relevanta variabler och använda avancerade modelleringsmetoder för att förbättra modellens precision och robusthet.

Innehåll

[Abstract 2](#_Toc165422657)

[1 Inledning 1](#_Toc165422658)

[2 Teori 2](#_Toc165422659)

[2.1Linjär Regression 2](#_Toc165422660)

[2.2 LOG-Transformation 2](#_Toc165422661)

[2.3 Modellens noggrannhet 2](#_Toc165422662)

[2.3.1 Modellens noggrannhet 2](#_Toc165422662)

[2.3.2 Outliers 2](#_Toc165422662)

[2.3.3 Hight leverage points 2](#_Toc165422662)

[2.3.4 R2 och justerad R2 2](#_Toc165422662)

[2.3.5 BIC 3](#_Toc165422662)

[3 Method 4](#_Toc165422664)

[3.1 Datainsamlingen 4](#_Toc165422665)

[3.2 Outliers 4](#_Toc165422666)

[3.3 Dummy- varialbler 6](#_Toc165422667)

[3.4 Korrelation 6](#_Toc165422668)

[3.5 Modellering 6](#_Toc165422668)

[4 API 9](#_Toc165422669)

[5 Resultat och diskussion 9](#_Toc165422669)

[6 Slutsatser 9](#_Toc165422670)

[7 Teoretiska frågor 10](#_Toc165422671)

[8 Självutvärdering 11](#_Toc165422672)

[9 Grupparbete 14](#_Toc165422673)

[Källförteckning 15](#_Toc165422674)

# Inledning

I dagens digitaliserade värld har dataanalys och prediktiv modellering blivit centrala verktyg för att fatta beslut inom många sektorer, inklusive bilindustrin. Att kunna förutsäga bilpriser exakt är av stor betydelse för både säljare och köpare. Prisprognoser kan hjälpa att göra bättre ekonomiska beslut medan säljare kan optimera sin prissättning. Denna studie fokuserar på att utveckla en robust omdell för att förutsäga bilpriser baserat på olika bilattribut .

Rapporten beskriver utvecklingen och utvärderingen av en linjär regressionsmodell för att förutsäga priser på begagnade bilar. Genom att använda ett datadrivet tillvägagångssätt syftar studien till att identifiera de mest signifikanta bilattributen som påverkar prisbildningen och att kvantifiera dessa effekter. Vi använder en omfattande dataset som består av tusentals biltransaktioner, där varje transaktion inkluderar detaljerad information om bilen.

Dataset med 6128 observationer och inkluderar variabler såsom märke, år, bränsletyp, körsträcka, växellåda och pris. jag ska använda linjär …..ta reda på

1. Kan jag uppnå en noggrannhet på 80%

2. Kan jag identifiera de mest signifikanta bilattributen som påverkar prisbildningen och att kvantifiera dessa effekter.

3. kan jag lada ner data från URL i R med kod

# Teori

## 2.1Linjär Regression

Linjär regression är en grundläggande och omfattande använda teknik inom statistik och

maskininlärning som används för att förutsäga en kontinuerlig beroende variabel utifrån en eller flera oberoende variabel. Det primära syftet med linjär regression är att etablera en linjär relation (modell) mellan den beroende variabeln (Y) och en eller flera förklarande variabler(X). Modellen beskrivs allmänt med ekvationen:

där 𝛽0*β*0​ är skärningspunkten, 𝛽1,...,𝛽𝑛*β*1​,...,*βn*​ är regressionskoefficienter som representerar viktningen av varje oberoende variabel, och 𝜖*ϵ* är en felterm som fångar den slumpmässiga variationen i modellen.

I formel ovan är Y den beroende variabeln och är de oberoende eller förklarande variablerna. är okända konstanter som representerar effekt av på beroende variabel Y (James, Witten, Hastie & Tibshirani, 2014, p.72).

## 2.2 LOG-Transformation

Log- transformation är en vanlig statistisk teknik som används för att stabilisera varians, normalisera distributioner och göra datamängder mer hanterbara för analytiska modeller. I detta projekt tillämpades log-transformation på prisdatan för begagnade bilar. Detta steg var avgörande för att hantera problem med olika spridningar i subgrupper av data, vilka ofta förekommer i ekonomiska data.

## 2.3 Modellens noggrannhet

2.3.1 Outliers

Outliers är en observation som avviker betydligt från andra observationer i en datamängd. Utliggare kan uppstå av olika anledningar, såsom felaktig registrering av en observation under datainsamling. Utliggare kan snedvrida statistiska analyser och förvränga tolkningen av resultat, vilket gör det viktigt att identifiera och hantera dem på lämpligt sätt(James et al., 2023, p. 97).

#### 2.3.2 High leverage points

High leverage points är observationer med ovanligt extrema värden för prediktorvariabler jämfört med resten av datamängden. De kan påverka skattningen av regressionslinjen avsevärt och bör noggrant identifieras för att säkerställa korrekt modellering och tolkning

#### 2.3.3RSE och RMSE

RSE (Residual Standard Error) och RMSE (Root Mean Square Error) används för att utvärdera modellens noggrannhet. Modellen evalueras genom att undersöka dess prediktionsförmåga med hjälp av metriker som medelabsolutavvikelse (MAE) och rotat medelkvadratavvikelse (RMSE), vilket ger en uppfattning om modellens noggrannhet och användbarhet i praktiska scenarion. Lägre värde på RMSE, desto bättre passar modellen data.

#### och justerat

ger ett alternativt mått på passning och tar alltid värden mellan 0 och 1. visar vilken procent av variansen i datan kan förklaras med modell.

Justerat är ett mått som justerar träningsfelet för att kompensera för överanpassning i modellen. Det tar hänsyn till antalet parametrar och straffar överdriven komplexitet, vilket föredrar enklare modeller som ändå presterar väl. Genom att använda justerat kan vi jämföra modeller med olika antal parametrar utan att behöva använda olika data set för träning och utvärdering av modeller (James, Witten, Hastie & Tibshirani, 2014, p.235).

2.3.5 BIC

BIC är ett statistiskt mått som används för modellval, särskilt inom ramen för regressionsanalys. Det erbjuder ett sätt att jämföra olika modeller baserat på deras passform till datan samtidigt som det bestraffar modeller med fler parametrar, vilket hjälper till att förhindra överanpassning. Modeller med lägre BIC-värden föredras, vilket indikerar en bättre avvägning mellan passformens godhet och modellens komplexitet. (James, Witten, Hastie & Tibshirani, 2014, p.235).

# Method

## 3.1 Datainsamlingen

Jag samma arbete med Nike, Mustafa och Parsan. Vi samlat in bil data från Blocket.se som säljas av företag i Göteborg. Efter jag tog översikt på datan

Brand Year Fuel Mileage Gearbox

Length:6128 Min. :1970 Length:6128 Min. : 0 Length:6128

Class :character 1st Qu.:2016 Class :character 1st Qu.: 3442 Class :character

Mode :character Median :2019 Mode :character Median : 7611 Mode :character

Mean :2018 Mean : 9027

3rd Qu.:2021 3rd Qu.:13206

Max. :2024 Max. :62000

Price

Min. : 11900

1st Qu.: 149900

Median : 239900

Mean : 303288

3rd Qu.: 379900

Max. :4999900

Var jag inte riktig nöjd med det. Från år 1970 tills 2024, samt priset från 11900 till 4999900kr. så begränsade jag

car\_data\_ny <- car\_data[car\_data$Price <= 1000000, ]

car\_data\_ny\_filtered <- car\_data\_ny[car\_data\_ny$Year >= 2010 & car\_data\_ny$Year <= 2024, ]

så vill jag koncentera på sista 14 år och tog jag bort stor del lyxbilar. Inför framtida datainsamlingar bör vi överväga att noggrant skilja på olika prisnivåer, speciellt mellan standard- och lyxbilar. Detta skulle inte bara förbättra modellens noggrannhet utan även ge en mer nyanserad förståelse av marknadsdynamiken.

Efter filtrering är datamängden fortfarande tillräckligt stor, d.v.s 5648 observationer för att ge robusta analyser av bilpriserna, körsträckan, tillverkningsåret, och andra bilattribut. Datamängden visar också en rimlig distribution av priser observationer.

### 3.2 Outliers

I nästa steg genomfördes en grundläggande kontroll av numerisk data i R med hjälp av pairs-funktionen. Från visualiseringen kunde en outlier observeras, diagram är särskilt användbara för att identifiera relationer och mönster som kan kräva ytterligare analys.En bild som visar text, skärmbild, diagram, Graf

Automatiskt genererad beskrivning

Genom att rensa bort outliers blir datan som används för analyser mer robust och mindre benägen att ledas vilse av atypiska värden. Detta är speciellt viktigt i situationer där outliers kan vara resultatet av felaktiga data eller ovanliga händelser som inte är relevanta för de generella trender och mönster man vill undersöka

En bild som visar text, skärmbild, diagram, design

Automatiskt genererad beskrivning

### 3.3 Dummy- varialbler

Vidare har jag skapa dummy-variabler. Att omvandla kategoriska data till dummy-variabler är nödvändigt för de flesta statistiska modeller och maskininlärningsalgoritmer som kräver numerisk input. när data är helt numerisk, kan den användas för att träna modeller som regression. Genom att ta bort den första dummy-variabeln undviker man problem relaterade till perfekt multicollinearitet, vilket är kritiskt för modeller som linjär regression.

En bild som visar text, skärmbild, Teckensnitt, nummer

Automatiskt genererad beskrivning

## 3.4 korrelation

Vid nästa steg undersökte jag korrelation mellan numeriska variabler. Korrelationskoefficienten är nära -1, vilket visar en stark negativ samband. Detta betyder att variablerna är invers relaterade; som väntat, ju nyare bilen är desto lägre är den ackumulerade körsträckan.

3.5 Modellering

Nästa steg delade jag upp datasetet i tränings- och testdata. Jag tränade en linjär regressionsmodell för att förutsäga bilpriser baserat på olika attribut som märke, år, miltal, drivmedel, och växellåda. Modellen inkluderade både numeriska och omvandlade kategoriska variabler. Jag genomförde diagnostiska tester för att utvärdera modellens anpassning, inklusive residualplots och leverage-analyser. Detta hjälpte till att identifiera data som hade stor påverkan på modellen och kunde potentiellt vara influerande observationspunkter. För att hantera icke-normalitet och heteroskedasticitet i residualerna, tillämpade du en log-transformation på pris. Detta resulterade i en förbättrad modell med bättre anpassning och residualfördelning.

En bild som visar text, diagram

Automatiskt genererad beskrivningEn bild som visar text, diagram

Automatiskt genererad beskrivning

En bild som visar text, skärmbild, diagram

Automatiskt genererad beskrivning

Histogrammet visar en mycket symmetrisk distribution av residualerna omkring noll, vilket är positivt eftersom det indikerar att medelvärdet av residualerna är centralt och att modellen inte systematiskt överskattar eller underskattar. Det finns en tydlig central topp, vilket kan tyda på en normalfördelning, men denna spetsighet kan också vara residualerna har tyngre eller lättare svansar än en normalfördelning. Shapiro-Wilk testet visade signifikant avvikelse från normalfördelning, ser både histogrammet och Q-Q plotten relativt bra ut. Detta kan ibland hända när storleken på datamängden är stor, vilket gör att även små avvikelser från normalitet blir statistiskt signifikanta.

En bild som visar text, diagram, Graf, skärmbild

Automatiskt genererad beskrivning

# API

Jag använda install.packages("httr") och install.packages("jsonlite") ladda ner data från SCB url <- "https://api.scb.se/OV0104/v1/doris/sv/ssd/BE/BE0101/BE0101A/BefolkningNy"

> print(data\_frame)

År Folkmängd

1 2010 9415570

2 2011 9482855

Så, sparade jag som Excel i min dator.

print(paste("Data har sparats som Excel-fil:", file\_path))

[1] "Data har sparats som Excel-fil: Folkmängd\_2010\_2011.xlsx"

Senare kund jag använda koden hämtade mer data

|  |
| --- |
| head(data\_df)  RegionKod År Folkmängd Folkökning RegionText  1 1 1968 1427236 NA Stockholm län  2 1 1969 1453754 26518 Stockholm län  3 1 1970 1477990 24236 Stockholm län  4 1 1971 1486648 8658 Stockholm län  5 1 1972 1485655 -993 Stockholm län  6 1 1973 1485587 -68 Stockholm län |
|  |
| |  | | --- | | > | |

Sen sparade jag i min dator som en excel fil.

En bild som visar text, programvara, Datorikon, Multimedieprogram

Automatiskt genererad beskrivningEn bild som visar text, skärmbild, programvara, skärm

Automatiskt genererad beskrivningEn bild som visar text, skärmbild, programvara, Datorikon

Automatiskt genererad beskrivning

# Resultat och Diskussion

Rapporten sammanfattar resultaten från en omfattande analys av begagnade bilars prissättning med hjälp av olika regressionsmodeller. Datamängden innehåller 6128 observationer av bilar som samlats från olika märken och modeller från 2010 till 2024. Genom datan rensades lyxbilar bort och endast bilar med priser upp till 1 000 000 kr inkluderades, vilket minskade urvalet till 5648 bilar.

Huvudanalysen involverade användning av linjär regression och log-linjär regression för att förstå vilka faktorer som mest påverkar bilpriser. Regressionsmodellerna analyserades noggrant med hjälp av diagnostiska plottar, inklusive Residuals vs Fitted, Q-Q plot, Scale-Location, och Residuals vs Leverage plottar, vilka tydligt illustrerar modellens passform och potentiella problem som hög leverage eller outlier.

Korrelationen mellan förutsagda och faktiska priser illustreras genom scatterplots, vilket visar att trots vissa avvikelser följer förutsägelserna generellt den förväntade trenden.

RMSE (Root Mean Square Error) och MAE (Mean Absolute Error) på testdata var höga, vilket indikerar att även om modellen presterar väl på träningsdata, finns det utrymme för förbättringar när det gäller generalisering till nya data.

Genom involverade i bilförsäljning inte bara förbättra sin prissättningsstrategi utan också erbjuda mer konkurrenskraftiga priser baserade på marknadsvärdet, vilket skulle kunna öka både kundnöjdhet och försäljning.

Utmanningar och möljgheter:

1. Även om modellerna förklarade en stor del av variansen i bilpriserna, indikerar diagnostiska tester som Shapiro-Wilk test och analys av residuals att det finns utrymme för förbättring. Specifikt kunde outliers och högleverage-punkter påverka modellens förmåga att förutsäga nya data korrekt.
2. RMSE och MAE på testdata visade att modellen kan ha svårigheter att generalisera till nya, oberoende data. Detta kan delvis bero på överanpassning eller att vissa variabler som påverkar pris inte fängslades av den nuvarande modellen.

Framåtblickande Strategier

1. Fortsatt arbete för att säkerställa datakvalitet kommer att vara avgörande. Detta inkluderar noggrann rengöring av data, behandling av saknade värden och korrekt hantering av outliers.
2. Det finns en potential att utforska mer komplexa modeller som kan hantera icke-linjära samband och interaktioner mellan variabler mer effektivt. Maskininlärningstekniker som slumpmässiga skogar eller gradientboosting kan erbjuda förbättringar i prediktionsnoggrannhet och robusthet.

Feature Engineering

Utveckla ytterligare features baserat på befintlig data, såsom bilmodellens livslängd (beräknad från skillnaden mellan nuvarande år och tillverkningsår) eller en kombinerad variabel av miltal och ålder, kan ge djupare insikter och förbättra modellens prestanda.

# Slutsatser

Jag har använt linjär och log-linjär regressionsanalys för att förstå och förutsäga begagnade bilars prissättning baserat på olika attribut som år, miltal, bränsletyp, växellåda och märke. Analysen har gett viktiga insikter och identifierat några nyckelfaktorer som påverkar begagnade bilars priser. Nedan följer de huvudsakliga slutsatserna från studien:

Prediktiva Faktorer

1. År och miltal : De är starka prediktorer för bilpriser. Nyare bilar och bilar med lägre miltal tenderar att säljas till högre priser.

2. Bränsletyp: Hybrid- och elbilar tenderar att ha högre priser, vilket återspeglar marknadens växande preferens för miljövänliga alternativ.

3. Växellådan och märke: Växellåda och specifika bilmarken är också betydelsefulla faktorer som påverkar prissättningen, vilket indikerar att konsumentpreferenser kan variera beroende på tekniska specifikationer och varumärkesimage.

Baserat på resultaten och observationerna från denna studie, här är några riktade rekommendationer för både bilåterförsäljare och forskare som arbetar med prediktiva modeller för prisbestämning av begagnade bilar:

Datadriven prissättning

**Återförsäljare** bör använda datadrivna metoder för att anpassa priserna baserat på faktorer som bilens ålder, miltal, och bränsletyp. Detta kan hjälpa till att optimera försäljningsstrategier och maximera lönsamheten.

Implementera dynamiska prissättningsmodeller som kontinuerligt kan uppdateras med nya data för att återspegla aktuella marknadstrender.

Kundfokuserad strategi

Integrera kundfeedback och preferensdata i prissättningsstrategin för att bättre möta marknadens behov och förväntningar.

Uttveckla segmenterade marknadsföringsstrategier baserade på analys av kundbeteende och preferenser, vilket kan leda till mer målinriktade och effektiva försäljningskampanjer.

Innan jag började vill jag nå nogrannhet på 80%, jag når inte mitt mål, borde ju göra mer.

|  |  |  |
| --- | --- | --- |
| MAE | RMSE | Felprocent |
| 276140kr | 78318kr | 28% |

det betyder att jag kan göra det bättre. Jag har identifierat de mest signifikanta bilattributen som påverkar prisbildningen och att kvantifiera dessa effekter. Sist lyckades jag hämtade data från URL PÅ SCB

# Teoretiska frågor

1.Kolla på följande video: <https://www.youtube.com/watch?v=X9_ISJ0YpGw&t=290s> , beskriv kortfattat vad en Quantile-Quantile (QQ) plot är.

QQ plot är en grafisk metod som används för att bedöma om en datamängd följer en specifik teoretisk fördelning, ofta normalfördelningen. Ploten jämför kvantilerna i den observerade datan med kvantilerna från den förväntade fördelningen. Om datapunkterna på ploten följer en linje, indikerar det att datamängden ungefär överensstämmer med den teoretiska fördelning

2. Din kollega Karin frågar dig följande: *”Jag har hört att i Maskininlärning så är fokus på prediktioner medan man i statistisk regressionsanalys kan göra såväl prediktioner som statistisk inferens. Vad menas med det, kan du ge några exempel?”* Vad svarar du Karin?

I maskininlärning ligger huvudfokus verkligen på att göra prediktioner, vilket innebär att modeller tränas för att förutsäga utfall baserade på ingångsdata. Det handlar främst om modellens förmåga att ge korrekta förutsägelser, oavsett hur eller varför dessa samband existerar.

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

Konfidensintervallet uppskatta var en populationsparameter ligger med viss säkerhet medan prediktionsintervallet försöker uppskatta var en enskild observation kan ligga med samma säkerhet.

4.Den multipla linjära regressionsmodellen kan skrivas som:

𝑌 = 𝛽0 + 𝛽1𝑥1 + 𝛽1𝑥2+ . . . + 𝛽𝑝𝑥𝑝 + 𝜀 . Hur tolkas beta parametrarna?

Beta-parametrarna (𝛽) i en multipel linjär regressionsmodell representerar effekten av varje oberoende variabel (𝑥𝑖) på den beroende variabeln (𝑌). Varje 𝛽𝑖 visar förändringen i den beroende variabeln för en enhets ökning i den oberoende variabeln, medan alla andra oberoende variabler hålls konstanta.

5.Din kollega Hassan frågar dig följande: ”Stämmer det att man i statistisk regressionsmodellering inte behöver använda träning, validering och test set om man nyttjar mått såsom BIC? Vad är logiken bakom detta?” Vad svarar du Hassan?

BIC kan hjälpa till att välja den bästa modellen. BIC inte behovet av att dela upp datan i träning, validering och test set. Detta beror på att BIC primärt fokuserar på teoretisk modellvalidering inom samma dataset, och ger inte information om modellens prestanda på helt ny, osett data.

6.Förklara algoritmen nedan för ”Best subset selection”

En bild som visar text, skärmbild, Teckensnitt, nummer

Automatiskt genererad beskrivning

Best subset selection är en metod inom regressionsanalys där man syftar till att hitta den mest effektiva undergruppen av prediktorer (oberoende variabler) för en given datamängd. Best subset selection ger alltid den bästa kombinationen av prediktorer men kan vara enormt påfrestande för datorn, speciellt vid en stor mängd prediktorer.

7.Ett citat från statistikern George Box är: “All models are wrong, some are useful.” Förklara vad som menas med det citatet.

Citatet betonar att varje modell är en förenkling eller approximation av verkligheten. Detta innebär att alla modeller nödvändigtvis kommer att ha begränsningar och fel, men det betyder inte att de är oanvändbara. Tvärtom, många modeller är extremt värdefulla trots sina brister. De kan hjälpa oss att förstå komplexa system, förutsäga framtida händelser och fatta informerade beslut.

# Självutvärdering

1. Utmaningar du haft under arbetet samt hur du hanterat dem.

Jag började tycker att R är enkelt, för att jag hittade bra vidior på kinesiska, när vi studerade djupare, då förstod jag att det är mycket att jag behöver lära mig. Jag tycker att R är enkelt använda, så som install package, tar information när jag behöver hjälp.

1. Vilket betyg du anser att du skall ha och varför.

Jag tycker att jag kan bli nöjd av G

# Grupparbete

Jag jobbade med Nike, mustafa och Parsan. Vi var bra överans att samla in bildata från blocket.se om bilar som företag säljer i Göteborg. Jag rekommenderade att samla begränsar från år 2010, men kom inte fram, när jag började arbeta, känsla för mig är att jag vill koncentra senast tiden ändå. Samt anser jag att lyxbilar ska tas bort också. Jag tycker att nästa gång om vi samla in data kan vi göra det mycket bättre för att vi vet nu vilka data ge bättre analys resultat.

# Källförteckning

<https://www.youtube.com/playlist?app=desktop&list=PLGy-tIPhuAbY_ttheQ6GEbgAGX8Mqdv2k(på> kinesiska)

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2023). An introduction to statistical learning: with Applications in R. Springer.

Hadley Wickham, Romain François, Lionel Henry, Kirill Müller, Davis Vaughan. (2023). dplyr: A Grammar of Data Manipulation. <https://cran.r-project.org/web/packages/dplyr/index.html>