Kunskapskontroll

Linjär modellering av priser på begagnade el- och hybridbilar



William Blennow

EC Utbildning

R-programmering DS23

202404

# Abstract

This report is based upon using linear regression modelling on a dataset of used car ads in Sweden. It researches whether there is a trend on newly registered cars in Sweden and discovers any recent trends based upon what type of fuel the car uses. It also outlines the usage of linear regression modelling and the method for ensuring that the data is of the assumed form that is needed for this type of modelling to draw conclusions. Thereafter it discusses the usage of the modelling and what type of variables that aren’t accounted for in the model that might have big effects on how usable it might be in the future.

Innehållsförteckning

[Abstract 2](#_Toc165035141)

[1 Inledning 1](#_Toc165035142)

[2 Teori 2](#_Toc165035143)

[2.1 Linjär regression 2](#_Toc165035144)

[2.1.1 Antaganden för linjär regression 2](#_Toc165035145)

[3 Metod 3](#_Toc165035146)

[3.1 Datainsamling 3](#_Toc165035147)

[3.1.1 Datainsamling från SCB 4](#_Toc165035148)

[3.2 Databehandling 4](#_Toc165035149)

[4 Resultat och Diskussion 5](#_Toc165035150)

[5 Slutsatser 7](#_Toc165035151)

[6 Teoretiska frågor 8](#_Toc165035152)

[7 Självutvärdering 11](#_Toc165035153)

[Appendix A 12](#_Toc165035154)

[Appendix B 13](#_Toc165035155)

[Källförteckning 15](#_Toc165035156)

# Inledning

Linjär regression är en av de mest pålitliga modellerna för både statistisk analys och statistisk inferens. I grund och botten så kan man genom linjär regression undersöka hur olika variabler är relaterade till varandra och hur de kan påverka exempelvis pris på en vara beroende på dess egenskaper.

För att förstå trender och tendenser så är det viktigt att kunna dra slutsatser från statistik och att kunna styrka upplevelser från ”magkänsla” med observationer och analys. Vi har en rad verktyg och modeller som kan hjälpa oss att förstå samband och generell utveckling.

Något som har belysts och som kommer att fortsätta spela en stor roll i samhället är miljön och klimatet, där vi ser en utveckling med mer och mer extrema väderförhållanden, naturkatastrofer och alarmerande rapporter om att vårt klimatavtryck inte är hållbart.

En stor del av detta handlar om användandet av fossila bränslen, flygresor, sjötransporter, industrier och vägtrafik har väldigt stora klimatavtryck och med modernare lösningar så hoppas man kunna stävja detta mönster.

Inom bilindustrin har det blivit tydligt att fordonen som trafikerar våra vägar inte nödvändigtvis måste vara drivna enbart på fossila bränslen, Tesla är ett varumärke som har blivit ett av världens högst värderade företag (Statista, 2023) baserat på deras produktion och försäljning av eldrivna bilar.

Denna rapport syftar till att undersöka huruvida det finns trender inom konsumentbeteenden i Sverige gällande val av drivmedel och samt en modellering kring vilka variabler som påverkar prissättningen på begagnade el och hybrid-bilar.

Rapporten syftar att besvara:

1. Har mönstret för vilken typ av fordon förändrats när man köper en bil över de senaste 5 åren?
2. Vilka variabler påverkar prissättningen för begagnade el och hybrid-bilar?

# Teori

## Linjär regression

Linjär regression är en statistisk metod som modellerar relationen mellan en beroende variabel (Y) och en eller flera oberoende variabler (X1, X2, ..) och modellen förutsätter att denna relation kan förklaras genom en linjär ekvation. En enkel linjär regression har enbart en beroende variabel och en oberoende variabel medans multipel linjär regression har flera oberoende variabler, modellen för denna är:

Y = β0 + β1X1 + β2X2 + ··· + βpXp + ϵ

Där Y som tidigare nämnt är den beroende variabeln, β0 är interceptet som modellen skapar, β1 och β2 är koefficienter för hur en ökning eller minskning av X1 resp X2 påverkar Y och ϵ modellens fel för det som inte går att förklara med modellen. (James, m.fl.)

### Antaganden för linjär regression

För att en linjär regressionsmodell ska vara verksam så är ett viktigt antagande att det finns ett linjärt samband mellan den beroende och de oberoende variablerna, stämmer inte detta så kommer en förändring i variabeln göra en ickelinjär påverkan på den beroende variabeln.

Ett annat antagande för linjär regression är oberoende av residualerna, det vill säga att en residuals fel inte inte har ett samband med felet på en annan variabels residual.

Linjär regression förutsätter också homoskedasticitet, dvs att variansen i residualerna är konstant, om så inte är fallet så blir den statistiska inferensen som görs på modellen felaktig.

Ett annat antagande är normalfördelning av residualer, detta innebär att felen följer vårt antagande om att residualerna har en symmetrisk distribution vilket låter oss använda modellen för hypotesprövning och statistisk inferens.

Ingen kolinearitet, detta antagande innebär att inga variabler har linjära samband med varandra. Detta innebär att om så inte är fallet så går det inte att bedöma hur de enskilda variablerna påverkar modellen iom det linjära sambandet mellan dem.

# Metod

## Datainsamling

Datainsamlingen har skett genom att samla in information gällande begagnade bilar från en webbsida där annonser om bilar till salu finns samlade.

Enligt instruktioner så besvaras frågor kring datainsamlingen i punktform:

1. Vem du har arbetat i grupp med?

Gruppen har bestått av: Dan Heikenberg, Daniel Hemgren, Frida Kilby, Khaldoun Agha, Melike Koyuncu, Natalie Dobrovolska, Siarhei Thor Fedatsenka, William Blennow och Xiaoyong Yang.

2. Hur har ni i gruppen arbetat tillsammans?

Vi har, förutom att diskutera under lektionstider, samtalat i en gruppchatt och avsatt tid för att ha möten via microsoft teams där vi gått igenom vilken typ av data som är intressant, hur den ska bearbetas och hur den ska samlas in. Vi undersökte först genom manuell datainsamling för att senare hitta en teknisk lösning där vi skrapade blocket på information på tillgänglig information från annonserna. Sedan så bearbetades datan i Python och exporterades till CSV, där vi gemensamt gick igenom de olika stegen för datahanteringen.

3. Vad var bra i grupparbetet och vad kan utvecklas?

Vi löste uppgiften som grupp upplever jag och la ambitionen på en rimlig nivå, där vissa var beredda att lägga ner mer tid än andra på det. Jag har inga direkta tankar om utveckling då vi gemensamt enades om vad vi önskade att uppnå och vad resultatet blev.

4. Vad är dina styrkor och utvecklingsmöjligheter när du arbetar i grupp?

Mina styrkor i grupparbeten som dessa är att jag är bra på att strukturera upp arbeten och formulera en plan för genomförande. Jag är medveten om mina utvecklingsmöjligheter i grupparbeten när jag upplever att ”vi inte kommer vidare” att jag kan gå in och ta en ledande roll för att gemonförandet ska bli av. Detta kan innebära en risk att personers tankar och åsikter inte alltid får den plats som behövs, men jag upplever inte att så har varit fallet i detta.

5. Finns det något du hade gjort annorlunda? Vad i sådana fall?

Nej, jag upplever att vi har uppnått de mål vi satt ut och har diskuterat metod, data och tolkning av data på ett bra sätt i vår gemensamma grupp.

Hade det varit ett längre projekt så tror jag att det hade funnits mer facetter att reflektera över men som grupp kändes det som att vi lyckades väl med våra mål och gemensamt reflekterade kring datan och insamlingen av denna.

Vi har även fortsatt diskuterat uppgiften, kanske att dessa diskussioner hade kunnat vara fler och mer uttömmande men jag upplever inte att någon hade de önskemålen.

### Datainsamling från SCB

För att samla in data från Statistiska Centralbyrån (SCB) har R-biblioteket PXweb använts för att genom API:er (Application Program Interface) samla in data och sedan behandlas i R Studio. Samt har exporterats och behandlats i excel.

## Databehandling

Förutom datarensning i Python och Excel så har all modellering och visualisering skett i programmeringsspråket R och i konsolen R studio.

Datan har först undersökts för att sedan skapa en ny data frame som består enbart av el- och hybridbilar där data finns om antalet hästkrafter. Därefter har en första undersökning av outliers på pris (den beroende variabeln) gjorts.

Nästa steg är att filtrera ut värden som kan indikera att annonsen är för en ny bil, eller att bilen är en leasingbil samt filtrera ut de outliers som identifierats. Därefter har det skapats dummies n-1 på vårda variabler som inte har numeriska värden för att vi ska kunna utföra en linjär regression på vårt dataset. En model skapas och från modellen så plottar vi för att undersöka de antaganden som krävs för linjär regressionsmodellering.

Därefter undersöks huruvida vår beroende variabel antas vara normalfördelad genom shapiro wilkes test, sedan transformeras den med den naturliga logaritmen och ett nytt transformerat dataset används för vår andra modell.

För den modellen som visar tecken på att det finns outliers som påverkar vår modell så vi använder cooks distance för att identifiera outliers, inspektera dem och sen bedöma om de kan exkluderas från vårt dataset.

En till modell skapas utan dessa outliers och sedan undersöks multikollinearitet genom VIF. Då vi har en stor mängd variabler genom våra skapade dummies så använder vi forward stepwise selection för att se vad vi kan göra för antagande om träningsfelet i vår modell.

# Resultat och Diskussion

Insamlingen av data från SCB visar att nyregistreringen av el- och hybridbilar har ökat markant över de 5 senaste åren, från en ganska obetydlig andel 2018 till en klar majoritet 2023.

En bild som visar text, skärmbild, Teckensnitt, nummer

Automatiskt genererad beskrivning

Figur 1: Antal nyregistrerade bilar per drivmedelstyp och år (SCB)

En bild som visar text, diagram, linje, Graf

Automatiskt genererad beskrivning

Figur 2: Visualisering av tabell 1

Den linjära regressionsmodellen visar på ett adjusted R squared på 0.8978 och ett F-värde som indikerar hög signifikans. Samt även ser ut att uppnå de antaganden som krävs för linjär regressionsmodellering.

En bild som visar text, Teckensnitt, vit

Automatiskt genererad beskrivning

En bild som visar text, Teckensnitt, vit

Automatiskt genererad beskrivning

Figur 3 : Värden för regressionsmodellen

En bild som visar text, diagram, skiss

Automatiskt genererad beskrivning

Figur 4: plottar för antaganden

# Slutsatser

För att besvara den första frågan så kan vi tydligt se genom vår datainsamling att ja, mönstret har tydligt förändrats de senaste 5 åren och om man ska tolka indikationerna så ser det ut att fortsätta på inslagen takt där fossildrivna bilar ersätts av el- och hybridbilar.

Den andra frågeställningen kring variabler visar på att vi i vår modell har oerhört många signifikanta variabler (appendix 1). Begagnade elbilar tenderar att vara dyrare än hybridbilar. Kausaliteten till detta behöver undersökas vidare, det kan ju vara så att elbilar (som säljs i högre grad det senaste året) prissätts utifrån årsmodell, vilket är en annan variabel som påverkar priset. En annan påverkande faktor till detta kan ju också vara att, som vi har identifierat, människor i Sverige tenderar att välja en elbil om de köper en ny bil, vilket skulle kunna förklara ett högre pris på marknaden för begagnade bilar.

Det är mycket som vår modell inte kan förklara, utan den kan bara berätta om ögonblicksbilden på den begagnade marknaden på webbhandelsplatsen i fråga och predicera vilket pris en bil skulle listas till. Vi kan också göra ett antagande att fler och fler elbilar kommer dyka upp på andrahandsmarknaden då de säljs i långt högre utsträckning de senaste åren, däremot så kan vi inte predicera hur detta kommer påverka prissättningen för andrahandsmarknaden.

I stort så har vi en modell som har utformats för att lösa predicering av prissättning just i detta nu, men tillämpningen av modellen och dess användbarhet bedömer jag vara väldigt begränsad.

Däremot kan vi se att vi skapar en bättre modell genom att tillämpa strukturerad granskning och analys av vår träningsdata på modellen, detta är lärdomar och insikter som vi kan med oss vidare och utvecklingen på de olika modellernas plottar för antagande av datan (appendix b) talar sitt tydliga språk.

# 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.*

En QQ-plot visar hur vår data (i fallet av en regressionsmodell så visar den våra standardiserade residualer) följer förväntningarna om normalfördelning. Linjen mellan y och x visar den förväntade fördelningen av datan och om den avviker så finns det anledningar att undersöka varför avvikelser uppstår och om datan behöver transformeras.

*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?*

Dels så kan ju linjär regression ingå i machine learning som en övervakad metod för att skapa prediktioner, men precis som Karin nämner så kan vi utföra statistisk inferens med regressionsanalys. Detta innebär att vi kan undersöka samples från en stor population för att dra slutsatser om populationen genom vårt dataset. T.ex så kan vi predicera konfidensintervall för att en urvalsgrupp skulle rösta på ett givet politiskt parti utifrån utbildningsnivå, eller om vi vill testa andra nollhypoteser kring befolkningen givet olika variabler.

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

Ett konfidensintervall utgår från medelvärdet, medan prediktionsintervallet är vad vi predicerar för en framtida observation. Ett exempel för att särskilja detta är att vi slår en sexsidig tärning. Vårt prediktionsintervall på detta slaget är 1/6 för varje sida av tärningen medan vårt konfidensintervall på slaget utgår från medelvärdet som kommer vara 3,5 då summan av sidorna delat på sidorna = 3,5

*4. Den multipla linjära regressionsmodellen kan skrivas som:*

*𝑌 = 𝛽0 + 𝛽1𝑥1 + 𝛽1𝑥2+ . . . + 𝛽𝑝𝑥𝑝 + 𝜀 .*

*Hur tolkas beta parametrarna?*

Betaparametrarna visar koeffecienten som en förändring i variablerna påverkar Y. beta 0 innebär modellens intercept, tittar man vidare så ser det ut som att x1 och x2 använder samma koefficient. Det vedertagna sättet att modellera skulle vara 𝑌 = 𝛽0 + 𝛽1𝑥1 + 𝛽2𝑥2+ . . . + 𝛽𝑝𝑥𝑝 + 𝜀 .

Ett exempel på detta är att om koeffecienten för b1 är 0,8 och x1 är 2 så kommer det enligt modellen påverka Y med 1.6

*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?*

Logiken bakom detta är att desto fler variabler du har i din modell så kommer den prestera bättre, det finns ett naturligt samband i detta där mängden variabler ökar modellens R2. Då löper vi absolut risk för att overfit, dvs att vår modell presterar extremt bra på datan den lärt sig av men inte nödvändigtvis på ytterligare observationer. Dvs att en modell kan prestera jättebra på träningsdata men inte testdata. Detta kan absolut kontrolleras mot validering, där man önskar att felet på träningsdata är i paritet med felet på testdata, då har vi en användbar modell.

Däremot för linjär regression så kan vi använda flera olika metoder för att uppskatta träningsfelet, bla mallows C, BIC och adjusted R2 som undersöker vilka variabler som skapar den bästa modellen, genom att undersöka dessa så kan vi identifiera vilken av våra linjära regressionsmodeller som kommer att antas prestera bäst på testdata. Detta görs generellt genom att bestraffa RSS och låta oss identifiera det optimala antalet variabler i modellen.

*6. Förklara algoritmen nedan för ”Best subset selection”*

En bild som visar text, skärmbild, Teckensnitt, nummer

Automatiskt genererad beskrivning

Vi börjar med en modell utan några prediktorer, sedan i steg 2 så skapar vi modeller för antalet variabler vi har som jämför RSS mellan variablerna. Låt oss utgå från att vi har 4 oberoende variabler i vår data så kommer vi skapa 4 olika modeller (Mk) och en Null-model. För m1 så inkluderar vi enbart en variabel och jämför således vilken av dem som ger oss lägst RSS, för M2 så inkluderar vi alla kombinationer av 2 variabler vilket då skulle vara 6 stycken. För M3 har vi 4 kombinationer och för M4 1 kombination.

Sedan så väljer vi den bästa av dessa 5 modeller baserat på vårt uppskattade genomsnittsfel baserat på validering eller en av de metoder som nämns i svaret på fråga 5.

Genom att göra detta så har vi hittat det bästa subsetet för vår modell

*7. Ett citat från statistikern George Box är: “All models are wrong, some are useful.”*

*Förklara vad som menas med det citatet.*

Citatet belyser det faktum att vi aldrig kommer kunna skapa en perfekt modell på någonting. Det kommer alltid finnas felkällor eller outliers eller en rad olika saker som inte går att förklara genom modellering. Däremot så belyser citatet att det centrala i modellerandet är att kunna applicera modellen. Så det vi behöver sträva efter är att skapa modeller som har ett tillämpningsområde, snarare än att sträva efter ”perfekta” modeller.

# Självutvärdering

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

Jag har haft lite svårt att hitta ett konkret syfte med modellen, vilket jag diskuterar i mina slutsatser, för mig så har den mest intressanta delen av kursen varit hur man ska hantera data för att kunna modellera på den, och jag hade gärna att vi spenderat mer tid på lektionstillfällena med detta, både i from av föreläsningar och kodningslabbar, jag känner att jag vill lära mig mer på den fronten för jag upplever att det är en central och viktig del av hela utbildningen som jag gärna hade fokuserat mer på.

Jämför jag denna uppgiften med den förra kursen så upplever jag att det var mer konkret och enklare att förhålla sig till i Machine Learning än i denna och därför har även min upplevelse av detta arbete känns lite spretigt och trevande.

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

Som alltid en svår fråga att besvara, då jag studerar programmet för att lära mig så mycket som möjligt snarare än att få så höga betyg som möjligt så min vana trogen refererar jag till betygskriterierna och resonerar kring dem:

Jag upplever att jag med råge har nått samtliga läranderesultat för kursen så då är det minst G.

Tittar vi på VG:

*• I en skriftlig rapport lösa ett problem genom att implementera metoder och modeller från regressionsanalys på ett fördjupat sätt med hög säkerhet*

Jag upplever nödvändigtvis inte att jag har löst ett problem, men jag upplever att jag har följt instruktionerna från kunskapskontrollen och använt mig av API. Vidare så har jag visat på en förståelse och anpassning av variabler och tillgodosett de antaganden som krävs för linjära regressioner, jag vet inte exakt hur fördjupat sättet är eller hur hög säkerheten är, men jag upplever att jag har en större förståelse för det än mina kurskamrater när vi har diskuterat dessa frågor och jag hoppas att detta framgår genom metodbeskrivningen och min R-kod

• Redogöra för och kritiskt diskutera modellval, modellanpassning och modellutvärdering

Detta stämmer överens med mitt överstående svar, jag tycker att appendix B och mitt r-script visar på en tydlig röd tråd och reflekterande kring datan. Jag hoppas att detta också märks i mina svar på de teoretiska frågorna.

För mig är det svårt att säga att jag inte når upp till VG även om jag upplever att jag presterat på en högre nivå i andra kurser vi har läst.

1. Något du vill lyfta fram till Antonio?

Inget mer än det jag framfört i tidigare reflektioner. Jag hade nog sett att vi lagt ner mer tid på hur vi hanterar data innan vi ska modellera den, som med det mesta (och har varit i tidigare kurser) så är repetition en väldigt bra inlärningsmetod som också skapar förståelse för min del. Så det vill jag väl belysa och önska i framtida kurser.

# Appendix A

En bild som visar text, skärmbild, dokument, Teckensnitt

Automatiskt genererad beskrivning

Figur 5 : Fullständiga koefficienter för modellen

# Appendix B

En bild som visar text, diagram, karta

Automatiskt genererad beskrivning

En bild som visar text, diagram, karta

Automatiskt genererad beskrivning

En bild som visar text, diagram, skiss

Automatiskt genererad beskrivning

Figur 6: Utveckling av modeller efter databehandling

# Källförteckning

Statista (2024). <https://www.statista.com/statistics/263264/top-companies-in-the-world-by-market-capitalization/> (hämtat 2024-04-23)

SCB (2024). Nyregistrerade personbilar efter län och kommun samt drivmedel. Månad 2006M01 - 2024M03. <https://www.statistikdatabasen.scb.se/pxweb/sv/ssd/START__TK__TK1001__TK1001A/PersBilarDrivMedel/> (hämtat 2024-04-23)

James G, Witten D, Hastie T , Tibshirani R. (2023). *An Introduction to Statistical Learning with Applications in R Second Edition.* Springer.