Regressionsmodell

För bilprisprediktion



Alvin Lagerson

EC Utbildning

Regressionsmodellering

2025

# Abstract

This study examines the possibility of predicting used electric vehicle prices using a linear regression model. The analysis reveals that factors such as mileage, drivetrain type, towing equipment, and car model significantly influence vehicle prices. Most variables remain highly significant even after correcting for potential heteroskedasticity, indicating strong and reliable effects. Some car models, however, do not exhibit statistically significant impacts.

Innehållsförteckning

[1 Inledning 1](#_Toc196657651)

[1.1 Bakgrund 1](#_Toc196657652)

[1.2 Syfte och frågeställning 2](#_Toc196657653)

[2 Teori 3](#_Toc196657654)

[2.1 Maskininlärningsmodeller 3](#_Toc196657655)

[2.1.1 Linjär regression 3](#_Toc196657656)

[2.1.2 Lasso 3](#_Toc196657657)

[2.2 Mått 3](#_Toc196657658)

[2.2.1 RMSE 3](#_Toc196657659)

[2.2.2 MAE 3](#_Toc196657660)

[2.2.3 MAPE 3](#_Toc196657661)

[2.3 Teoretiska antaganden för linjär regression 3](#_Toc196657662)

[2.3.1 Linjär relation 3](#_Toc196657663)

[2.3.2 Homoskedasticitet 3](#_Toc196657664)

[2.3.3 Multikollinjaritet 4](#_Toc196657665)

[2.3.4 Residualernas normalfördelning 4](#_Toc196657666)

[2.3.5 Modellens korrekthet 4](#_Toc196657667)

[2.3.6 Ingen autokorrelation 4](#_Toc196657668)

[2.4 Skrapning 4](#_Toc196657669)

[3 Metod 5](#_Toc196657670)

[3.1 Datainsamling 1 5](#_Toc196657671)

[3.2 Datainsamling 2 5](#_Toc196657672)

[3.3 Bibliotek 5](#_Toc196657673)

[3.4 EDA 5](#_Toc196657674)

[3.4.1 Urval 5](#_Toc196657675)

[3.4.2 Visualisering 6](#_Toc196657676)

[3.5 Förberedelser 7](#_Toc196657677)

[3.6 Träning och test 7](#_Toc196657678)

[3.7 Kontroll 7](#_Toc196657679)

[3.8 Förbättringar 8](#_Toc196657680)

[3.9 Slutliga modellen 8](#_Toc196657681)

[4 Resultat och Diskussion 9](#_Toc196657682)

[4.1 Testresultat av olika modeller 9](#_Toc196657683)

[4.2 Diskussion 9](#_Toc196657684)

[5 Slutsatser 11](#_Toc196657685)

[5.1 Frågeställning 11](#_Toc196657686)

[6 Teoretiska frågor 12](#_Toc196657687)

[7 Självutvärdering 14](#_Toc196657688)

[Appendix A 15](#_Toc196657689)

[8 Tabell, och figur och källförteckning 16](#_Toc196657690)

[8.1 Figurförteckning 16](#_Toc196657691)

[8.2 Tabellförteckning 16](#_Toc196657692)

[8.3 Källförteckning 16](#_Toc196657693)

# Inledning

## Bakgrund

Ett resultat av att ett samhälle blir rikare är att en mindre andel av gemene mans lön går till det basala i ens levnad. I takt med att Sverige har blivit rikare med tiden har en mindre andel av lönen lagts på mat och bostad. En effekt av detta är att kapitalvaror ökar. En av dessa varor är bilar och sen 2002 fram till 2020 ökade antal bilar med ungefär en miljon i Sverige vilket är en ökning på ca 25%, se Figur 1 nedan. Detta trots att befolkningen bara ökat med 16% under samma tid (SCB, 2025).  
En bild som visar text, Graf, skärmbild, linje

AI-genererat innehåll kan vara felaktigt.

Figur 1, Antal bilar över tid.

Trots att vi känt till effekten av oljeförbränning sen mitten av förra seklet har jakten på välfärd varit viktigare i hela världen än effekten av den och även så i Sverige. Nu är vi mitt i en rådande klimatkris som effekt av vår jakt på välfärd och man har börjat titta på hur vi kan upprätthålla den välfärd vi uppnått men med lägre utsläpp. En stor post i utsläpps-potten är transportsektorn och det har börjat satsas på elbilar för att minska denna. Antalet elbilar har ökat långsamt sen mätningarna började men sen ungefär 2020 har de ökat ganska kraftigt, se Figur 2 nedan. Detta medför att det finns ett ökat antal begagnade elbilar på marknaden.

En bild som visar linje, Graf, diagram, skärmbild

AI-genererat innehåll kan vara felaktigt.

Figur 2, Antal elbilar över tid.

## Syfte och frågeställning

Syftet med denna rapport är att skapa en regressionsmodell för att förstå vilka parametrar som påverkar prissättningen på elbilar nu när det börjar finnas ett stort antal olika modeller av elbilar till salu.

Går det att skapa en regressionsmodell som predikterar pris på bil med under 50 000 kr i RMSE och över 80% i förklaringsgrad, R²?

# Teori

## Maskininlärningsmodeller

### Linjär regression

Linjära regressioner gör prediktioner genom att beräkna en viktad summa av ingående variabler plus en konstant som kallas bias (eller intercept). Modellen beskrivs med formeln:

**

β₀ är skärningspunkten på y-axeln och β₁ … βₚ är de förväntade förändringarna i y när xₚ ökas med en enhet och övriga variabler hålls konstanta. För att träna modellen används ofta Mean Squared Error (MSE) som mått för att optimera parametrarna så att de bäst passar träningsdatan. Målet är att minimera MSE, vilket innebär att modellen anpassas för att minimera skillnaden mellan observerade och förutsagda värden. Root Mean Square Error (RMSE) är en annan relaterad måttskala. Den linjära regressionsmodellen använder enkel matematik men är grunden för många avancerade metoder inom maskininlärning (Geron, 2019).

### Lasso

Lasso-regression (Least Absolute Shrinkage and Selection Operator Regression) är en linjär regressionsmodellsmetod som minskar multikollinjaritet och väljer ut viktiga variabler genom att sätta oviktiga vikter till exakt noll. Det använder ℓ1-normen som regularisering, till skillnad från Ridge som använder ℓ2-normen. Detta gör modellen enklare och hjälper till att hantera många irrelevanta variabler. Den är bra för datamängder med många funktioner (Geron, 2019).

## Mått

### RMSE

När man väljer ett prestandamått för regression används ofta Root Mean Squared Error (RMSE), som mäter hur stora fel en modell typiskt gör i sina prediktioner, med högre vikt på stora fel. RMSE är kopplad till ℓ2-normen, vilket gör den känslig för outliers (Geron, 2019).

### MAE

Ett alternativ är Mean Absolute Error (MAE), baserat på ℓ1-normen, som är mindre känslig för outliers. Både RMSE och MAE jämför prediktioner med faktiska värden för att utvärdera modellens noggrannhet. RMSE föredras generellt när outliers är sällsynta (Geron, 2019).

### MAPE

​Mean Absolute Percentage Error (MAPE) är ett mått på prognosnoggrannhet som uttrycker fel i procent. Det beräknas som det genomsnittliga absoluta procentuella felet mellan förutsagda och faktiska värden.

## Teoretiska antaganden för linjär regression

### Linjär relation

En förutsättning för linjär regression är att sambandet mellan de oberoende och den beroende variabeln ska vara linjära. Det går att kolla med en enkel visualisering som scatterplot (Wermke, u.d.).

### Homoskedasticitet

Heteroskedasticitet innebär att variansen hos feltermerna i en regression är ojämn över olika nivåer av de förutsagda värdena. Detta kan observeras som ett fläkt- eller konliknande mönster i en residualplot (CFI, 2025). Residualerna bör ha konstant varians (homoskedasticitet) (Wermke, u.d.).

### Multikollinjaritet

De oberoende variablerna bör vara fria från starka samband med varandra. Ju mer variablerna är korrelerade desto större blir standardfelet och resultatet mindre precist. Variance inflation factor (VIF) kan användas för att kontrollera multikollinjaritet där värdet ska vara under 5 (Wermke, u.d.).

### Residualernas normalfördelning

Residualerna bör vara hyfsat normalfördelade för att linjär regression. Detta är särskilt viktigt vid små dataset, testas enkelt med histogram (Wermke, u.d.).

### Modellens korrekthet

Modellen måste vara korrekt specificerad, vilket innebär att alla relevanta variabler inkluderas och irrelevanta variabler exkluderas. Felaktig modellering kan leda till partiska eller ineffektiva estimat (Araz Rawshani, 2025).

### Ingen autokorrelation

Residualerna bör inte vara korrelerade med varandra. Detta kan testas med Durban-Watson test där ett värde runt 2 är bra medan 0 eller 4 inte är bra (Wermke, u.d.).

## Skrapning

Skrapning innebär att man med ett skript kopierar data från webbsidor eller från ett programs användargränssnitt (IT-ord, 2018). Genom att granska ”bakom” slutresultatet av en webbsida går det att hitta olika element som har en given uppgift på sidan eller innehåller en viss data, dessa kan tilldelas en variabel och sparas ner i till exempel en dataframe eller csv-fil. Det går även att plocka ner hela sidor i en ”html-soppa” där man kan söka efter data som om det vore en txt-fil och kopiera data som ligger efter ett givet stycke text (Lagerson).

# Metod

## Datainsamling 1

Vid insamling av data var vi en större grupp som arbetade tillsammans. Vi organiserade oss i en kanal på Teams där vi diskuterade upplägget. Vi begränsade urvalet till att enbart ta med Volvo och att varje medlem skulle hämta data från minst 50 annonser var och sen sammanställa all data.

Jag skulle nu med facit i hand begränsat mer för att sålla ut ointressant data. Som till exempel auktioner som annonserar orimligt låga priser men även äldre bilar som kan variera kraftigt i pris då de dels kan vara skrotbilar som slumpas bort, dels välskötta veteran- och entusiastbilar som kostar mer än vad en ny bil kostar.

De som var med i gruppen förutom jag var: Ana, Arash, Emad, Gayatrhree, Hani, Joakim, Rana, Katarina, Michael, My, Per, Peter, Sharmin, Tahira, Tural och Zakariyae. Det positiva med att jobba i grupp är att varderas arbete blir mindre, det negativa är att alla inte gör exakt likadant. Mer förarbete och exempel skulle nog göra insamlingen bättre. Min styrka är generellt att jag kan Excel bra och hanterar datoranvändning bra men vad jag behöver jobba med är att jag kan vara otålig.

I det här fallet började jag titta på alternativ i väntan på övriga och nyfikenheten ledde mig in på skrapning.

## Datainsamling 2

Den andra setet data samlade jag in genom att med Python först skrapa fram annonsadresser för att sen genom ett Python-skript öppna varje annons och hämta data och lägga i en csv-fil. Hämtningen skedde i flera steg då jag vid flera tillfällen upptäckte att jag glömt eller missat något. Sen blev efterbearbetningen av datan ganska stor då jag haft för korta pauser i skriptet i väntan på laddning vilket gjorde att vissa data inte hämtades korrekt och fick hämtas manuellt i efterhand. Jag fokuserade min hämtning på enbart elbilar och uppdaterade i efterhand med batteristorlek på alla bilar, vilket bara fanns i vissa annonser.

## Bibliotek

Biblioteken jag använt i Python är bs4, selenium, webdriver\_manager, time, re, requests och csv. I R-kodningen har jag använt en lång rad olika bibliotek. Bland dessa finns ggcorrplot, tidyverse, skimr, ggplot2, caret, dplyr, janitor, scales, lmtest, car, broom och patchwork.

## EDA

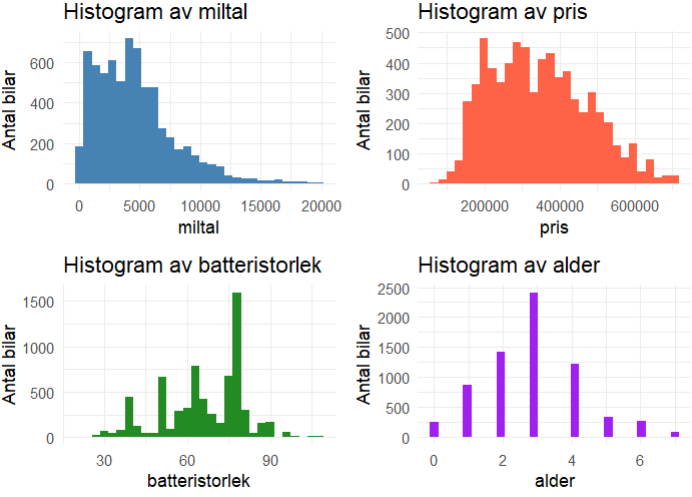
Skrapningen resulterade i 10 000 annonser som genom viss bearbetning och sållning i Excel kom ner till ungefär 9 100. Datasetet består av unika nyckeln, URL, bränsle, växellåda, miltal, modellår, kaross, drivning, märke, modell, pris, dragkrok och batteristorlek. URL, bränsle och växellåda togs bort redan innan import till R då de inte var relevanta. Eftersom bearbetningen tog mycket tid så lärde jag känna datan ganska väl där, mest genom filtreringar. Vid inladdningen till R så gjorde jag en kortare eda även där. De numeriska kolumnerna var rätt men jag gjorde om de som skulle vara faktorer direkt i början sen ändrade jag årsmodell till ålder.

### Urval

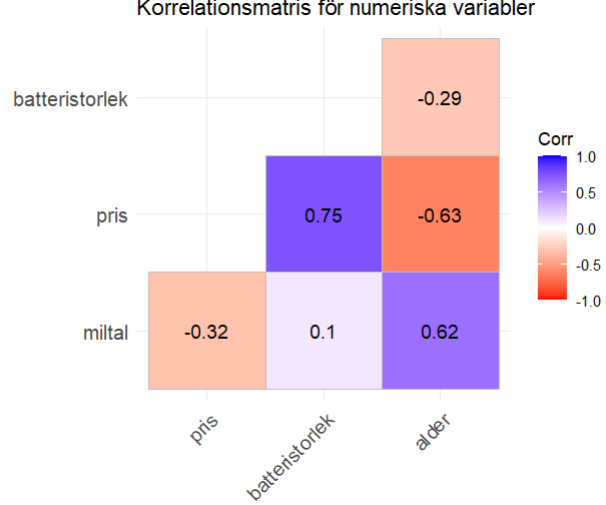
Genom en NA-genomsökning fanns 16 rader som saknade pris, men detta var snabbt fixat i Excel. Sen ritade jag upp de numeriska graferna i histogram, se Figur 1. Här är det filtrerat med bilar mellan 100 mil och 20 000 mil. Anledningen är att nya bilar inte har något värde i en modell som ska prediktera begagnade bilars pris och att genomsnittsåldern på bilar i Sverige ligger på 20 000 mil. Även om det finns många bilar över det miltalet är det svårt att prediktera pris så det finns ett golv för hur lite en elbil kan bli värd då till exempel batteriet har ett stort värde fortfarande Som exempel finns en Tesla model S till salu som gått 84 000 mil som kostar 135 000kr. Pris är satt till 650 000 kr då bilarna över där främst är speciella lyxbilar. Ålder satte jag till under 8 år, främst för att det innan dess fanns väldigt få elbilar, se Figur 2. Till sist begränsades batteristorlek till över 20 kWh då de under detta främst är mopedbilar som har dessa små batterier. Eftersom EDA inte är en linjär process så är detta vad jag kom fram till efter många vändor och plottar. Efter dessa urval fanns 6 832 rader data kvar.

### Visualisering

Det visade sig finnas flest 3 år gamla bilar, vilket är rimligt då leasingar oftast går på 3 år. Antalet nya bilmodeller ökar också vilket gör att det finns många bilar som gått väldigt lite.



Figur 3, Histogram av miltal, pris, batteristorlek och ålder.



Figur 4, Korrelationsmatris.

Korrelationsmatrisen, Figur 4, visar att miltal har ett negativt samband med pris, vilket innebär att högre miltal oftast leder till lägre pris. Pris är starkt positivt korrelerat med batteristorlek, vilket betyder att bilar med större batteristorlek vanligtvis är dyrare. Generellt sett framträder starka kopplingar mellan bilens ålder, miltal och pris, medan batteristorleken också påverkar priset men mindre tydligt relaterar till ålder.

## Förberedelser

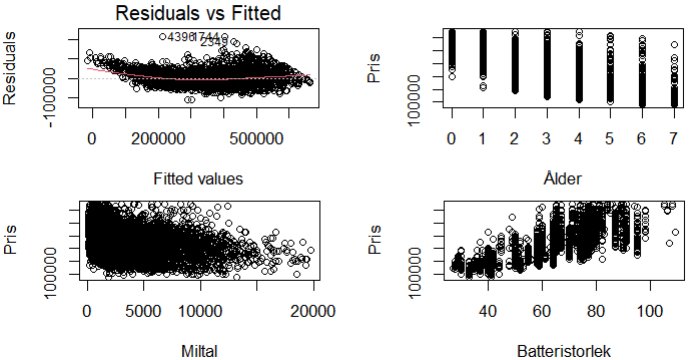
Förutom angivna filter som tidigare nämnts så delads datan upp i träning och test, 70 och 30 med reproducerbarhet. Sen filtrerade jag för att säkerställa att de kategoriska nivåer (märke och modell) som finns i träningsdatan också matchas med testdatan.

## Träning och test

Första regressionsmodellen gick så där men det var mest skrivfel och andra ”buggar”. Väl fungerande presterade modellen inte så bra, men det är svårt att uppskatta vad som är rimligt utan direkt jämförelse. I slutet efter hela undersökningen så kom jag på att filtreringen var gjord på ”fel plats” så redan den första modellen presterade mycket bättre än vad jag först trodde.

## Kontroll

För att kontrollera att modellen uppfyllde de teoretiska antaganden som en linjär regression ska uppfylla så genomfördes vissa tester, däribland dwtest, bptest, cor, vif och ett antal grafiska plottar. Dessa visade att det fanns vissa problem, bland annat multikollinjaritet där märke och modell och till viss del kaross förklarade samma sak. Nedanstående, Figur 5, tyder på en ickelinjaritet.



Figur 5, Test av linjaritet.

Efter genomfört Durbin-Watson test tydde resultatet på att det inte finns någon tydlig autokorrelation i residualerna. Däremot verkar modellen ha ett högt BPtest-tal vilket tyder på att modellen bryter mot homoskedasticitet. Modellens residualer verkar normalfördelade men ena svansen tyder på en del extrema värden, se Figur 6.

En bild som visar diagram, text, Graf, linje

AI-genererat innehåll kan vara felaktigt.

Figur 6, Residualernas normalfördelning

## Förbättringar

På grund av rådande ovan nämnda problem tog jag bort märke och kaross från modellen för att försöka lösa multikollinjaritet då de till stor del förklaras av bilmodell och logtransformerade pris och miltal för att försöka lösa ickelinjariteten och heteroskedasticiteten. Svansarna på residualfördelningen, residualerna med hög inverkan tog jag bort.

Efter att en ny modell testats verkar multikollinjariteten vara ett mindre problem än innan, dock inte helt löst. Ickelinjariteten ser bra ut och svansarna verkar bra, se Figur 7.

En bild som visar text, linje, Graf, diagram

AI-genererat innehåll kan vara felaktigt.

Figur 7, Residualernas fördelning, linjaritet

Däremot finns fortfarande problemet med heteroskedasticiteten kvar då BP-värdet fortsatt är högt.

## Slutliga modellen

Efter ytterligare utvärdering behövde jag ta bort batteristorlek för att minska multikollinjariteten från höga nivåer, och även ta bort ålder för att eliminera multikollinjariteten helt. För att hantera heteroskedasticiteten valde jag att använda robusta standardfel via Huber-White-sandwich-estimatorn, vilket justerar standardfelen utan att behöva ändra själva modellen. Det hade varit intressant att djupare diagnosticera och identifiera mönstret bakom heteroskedasticiteten, men tiden räckte tyvärr inte till denna gång.

# Resultat och Diskussion

## Testresultat av olika modeller

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Modell | RMSE | R² | MAE | MAPE |
| Första | 33 050 | 0,93 | 24 958 |  |
| Städad | 31 848 | 0,93 | 22 123 |  |
| Slutliga | 41 885 | 0,89 | 28 960 | 8,75 % |

Tabell 1, Resultat av de olika modellerna.

Tabell 1 visar att den slutliga modellen har ett Mean Absolute Percentage Error (MAPE) på cirka 8,75 %, vilket innebär att det genomsnittliga prediktionsfelet är knappt 9 % av det faktiska bilpriset. En MAPE under 10 % anses generellt indikera en mycket bra prediktionsförmåga för många tillämpningar. Trots att RMSE och R² försämrades något i jämförelse med tidigare modeller, har modellen nu en högre trovärdighet tack vare att de teoretiska antagandena för linjär regression i högre grad är uppfyllda. Sammantaget innebär detta att modellen är praktiskt användbar och statistiskt tillförlitlig för att förutsäga bilpriser inom det undersökta spannet.

## Resultattolkning och hypotesprövning

De allra flesta variabler är statistiskt signifikanta på 1%-nivån (p < 0.01), vilket visar på mycket starka samband mellan den beroende variabeln och exempelvis bilmodell, antal mil körda, samt utrustning såsom dragkrok och fyrhjulsdrift.

Miltal har en negativ och mycket signifikant effekt (Estimate: -0.148, p < 0.001). Varje ökning med i miltal innebär i snitt en minskning försäljningspriset.

Fyrhylsdrift ger en positiv och stark effekt (Estimate: 0.137, p < 0.001). Även dragkrok har en positiv effekt (Estimate: 0.011, p < 0.01), men relativt liten i storlek.

Nollhypotesen (att koefficienterna är noll) kan förkastas för de flesta variabler (p < 0.01), vilket innebär att dessa variabler har en signifikant effekt på den beroende variabeln.

För vissa bilmodeller (till exempel bilmodellerna Jaguar I-Pace och Maxus T90) kan nollhypotesen inte förkastas, vilket antyder att de inte bidrar signifikant till modellen.

## Analys

Vid analys av den slutliga regressionsmodellen undersöktes huruvida de teoretiska antagandena för linjär regression var uppfyllda. Multikollinjariteten minimerades genom att ta bort variabler som bidrog till höga VIF-värden (bland annat batteristorlek och ålder). Ickelinjariteten och heteroskedasticitet hanterades delvis genom transformation av målvariabeln (logaritmering av pris och miltal). Viss heteroskedasticitet kvarstod enligt Breusch-Pagan-testet men ska vara okej efter coeftest användes för beräkning av robusta standardfel. Normalfördelningen av residualerna bedömdes vara acceptabel vid visuell inspektion av residualplotar och outliers samt high leverage points är hanterade. Ingen autokorrelation i residualerna påvisades. Sammantaget kan modellen anses vara tillräckligt robust för att dra slutsatser ifrån, men viss försiktighet bör iakttas vid tolkning på grund av små avvikelser från perfekt homoskedasticitet.

## Diskussion

Detta projekt syftade till att bygga en prediktionsmodell för bilpriser baserat på ett stort dataset. Under arbetets gång framkom flera viktiga insikter. Inledningsvis visade modellen mycket goda resultat, men efter närmare granskning upptäcktes brister kopplade till multikollinjaritet och heteroskedasticitet, vilket kan underminera tillförlitligheten i resultaten. Genom att noggrant filtrera variabler, ta bort outliers och justera för starkt korrelerade variabler kunde dessa problem mildras. Detta resulterade i en modell som bättre uppfyller de teoretiska antagandena för linjär regression.

Trots att den slutliga modellen fick något sämre värden för RMSE och R² jämfört med den första modellen, förbättrades dess statistiska kvalitet och generaliserbarhet. Projektet illustrerar den viktiga balansen mellan modellens prediktiva styrka och att följa regressionsanalysens teoretiska krav.

Ett viktigt förbättringsområde för framtida arbete vore att undersöka mer avancerade modeller, exempelvis robust regression eller maskininlärningsmodeller, för att ytterligare hantera icke-normalitet och heteroskedasticitet. Det hade även varit värdefullt att analysera mönstret i residualerna djupare och anpassa modellen efter eventuella icke-linjära relationer.

Sammanfattningsvis visar projektet vikten av en noggrann modellutvärdering och hur små justeringar kan göra stor skillnad för tillförlitligheten av prediktiva modeller.

## Tankar och vidare projekt

Eftersom det totala urvalet är så stort kan man med fördel göra en modell som tittar på enskilda märken/modeller för att få en mer precis prediktion. I början av själva regressionsmodelleringen testade jag att göra en modell som bara kollade Tesla Model 3 som fick 26 105 i RMSE.

Efter projektets slut och stor genomtanke av hela arbetsflödet kan jag inte låta bli att lite smått ifrågasätta fulla värdet. Tanken från början var att skapa en modell som kan prediktera priset på begagnade elbilar, men samtidigt är dataunderlaget egentligen en databas med osålda bilar. Så vad modellen gör är att uppskatta vad annonserat pris skulle vara snarare än försäljningspriset. Vi kan inte veta om dessa bilar blir sålda för annonserat pris. Vissa bilar kanske blir sålda för tiotusentals, kanske hundratusentals kr mindre än annonserat beroende på hur väl säljaren känner marknaden, få bilar säljs dock för mer. Det verkligt intressanta och meningsfulla projektet hade varit att få ta del av Blockets slutpriser, det sista annonserade priset för varje såld bil de senaste 6 månaderna. Det hade verkligen kunnat skapa ett värde.

# Slutsatser

## Frågeställning

1. Går det att skapa en regressionsmodell som predikterar pris på bil med under 50 000 kr i RMSE och över 80% i förklaringsgrad, R²?

Jag skapade en modell som uppnådde 41 885 i RMSE och en R² på 0,89, modellen förklarar 89 % av variationen i bilpriserna.

# Teoretiska frågor

*Beskriv kortfattat vad en Quantile-Quantile (QQ) plot är.*

En QQ-plot är ett diagram där man ställer upp sin data i storleksordning och jämför fördelningen i sin data med en teoretisk fördelad, vanligen normalfördelad, data. Om punkterna ligger längst en rak linje följer din data den teoretiska data du jämför med. En böjning uppåt eller nedåt påvisar att datan har mer eller mindre spridning i ändarna och en S-form tyder på en skevhet, alltså att datan inte är symmetrisk. Avvikelser i början och slutet är extremvärden/outliers.

*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 är huvudmålet att skapa en modell som predikterar utfall på okända data där fokuset är just prediktionen, inte anledningen till att det händer. Fokuset är alltså att gissa rätt snarare än att förstå vilka variabler som påverkar. Regressionsanalys använd både för att göra prediktioner och att förstå sambanden mellan variabler. Här är du intresserad av vad som påverkar vad och om effekten är statistiskt säkerställd.

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

Konfidensintervall beskriver medelvärdet av y givet x, till exempel intervallet där medelvärdet av försäljningen ligger givet kostnaden för annonsering. Prediktionsintervall beskriver ett framtida utfall för y givet x, till exempel vad man borde kunna förvänta sig av framtida försäljning nästa gång man annonserar för x kr.

*Den multipla linjära regressionsmodellen kan skrivas som:*

**

*Hur tolkas beta parametrarna?*

β₀ är skärningspunkten på y-axeln och β₁ … βₚ är de förväntade förändringarna i y när xₚ ökas med en enhet och övriga variabler hålls konstanta.

*Din kollega Nils 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?*

Inom maskininlärning tränar du modeller med träningsdata, validerar sen för att välja den bästa modellen och sen testar modellens generaliseringsförmåga på ny osedda data. BIC och AIC försöker å sin sida uppskatta modellens generaliseringsförmåga direkt genom att du jämför modeller med olika antal variabler och väljer den med lägst BIC/AIC. Du behöver inte nödvändigtvis dela upp datan, till exempel om fokus ligger på inferens eller modellförståelse. Har man mycket data och modellen främst ska användas för prediktion eller att man ska utvärdera hur bra modellen funkar på ny data kan det vara bra att dela upp datan.

*Förklara algoritmen nedan för ”Best subset selection”*

Man börjar med nollmodellen, den enklaste modellen, där man inte använder några förklarande variabler utan förutsäger bara medelvärdet av y för alla observationer. Därefter går man vidare stegvis och testar alla möjliga modeller med 1 variabel sen alla med 2 variabler och så vidare upp till totala antalet tillgängliga variabler. För varje antal variabler väljs modellen som är bäst, vanligtvis den med lägst RSS eller högst R². Till sist jämför man dessa ”vinnarmodeller” för varje nivå av k och väljer den modell som presterar bäst totalt sett, baserat på ett valideringsmått, exempelvis AIC, BIC, justerat R² eller valideringsfel/korsvalidering.

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

Alla modeller är kraftiga förenklingar av verkligheten, ingen modell kan förklara allt. Däremot kan modeller hjälpa till att förstå samband och vilka parametrar som har stor inverkan samt hjälpa till att uppskatta framtida utfall. Det viktiga är inte att modellen visar verkligheten utan att den är tillräckligt bra för att hjälpa oss fatta beslut eller dra slutsatser.

# Självutvärdering

1. Vad tycker du har varit roligast i kunskapskontrollen?

Egentligen tyckte jag det var roligast med skrapningen, det kändes väldigt hands on på något sätt. Men ju lägre jag kom in i projektet ju roligare och intressant blev själva regressionsmodelleringen.

1. Hur har du hanterat utmaningar? Vilka lärdomar tar du med dig till framtida kurser?

Jag har haft en del bulor efter vägen. Bland annat var skrapningen något jag aldrig gjort tidigare men jag är envis och la orimligt mycket tid på den delen. Jag försökte bland annat ta hjälp av folk jag känner som kunde förklara begrepp och hur webbsidornas kodning fungerar för att sen förstå hur jag ska angripa skrapning. I slutet av skrapningen så löste jag problem jag hade i början men som var för sent att göra om, vilket ju ändå tyder på att processen gett mig något. Detsamma gäller ju hela projektet, om jag hade börjat med projektet idag hade jag använt tiden mycket effektivare och dessutom genomfört projektet mer systematiskt och med större förståelse, men det är ju lite av poängen med ett sådant projekt, att komma ut visare på andra sidan.

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

Jag gick ut med målet att satsa på VG. I jakten på betyg så fick jag ju inte till någon bra tidsplan utan jag la orimligt mycket tid på datainsamlingen, vilket jag känner har påverkat slutresultatet negativt. Det har dock gett mig personligen bra, om än halvt irrelevanta i sammanhanget, kunskaper. Jag hoppas det jag gjort räcker, men jag hade egentligen velat göra analysen mer genomarbetad.

1. Något du vill lyfta fram till Antonio?

Det skulle möjligtvis vara att ha projektet igång sen starten av kursen, där de teoretiska uppgifterna kan vara något man gör redan i början för att förstå bättre och sen även ha mer tid åt kodning och analys senare.

# Appendix A

Koden för projektet och skrapningen finns på följande länk: <https://github.com/Al-Vinci/Regressionsmodell>

# Tabell, och figur och källförteckning

## Figurförteckning

[Figur 1, Antal bilar över tid. 1](#_Toc196648623)

[Figur 2, Antal elbilar över tid. 1](#_Toc196648624)

[Figur 3, Histogram av miltal, pris, batteristorlek och ålder. 6](#_Toc196648625)

[Figur 4, Korrelationsmatris. 6](#_Toc196648626)

[Figur 5, Test av linjaritet. 7](#_Toc196648627)

[Figur 6, Residualernas normalfördelning 7](#_Toc196648628)

[Figur 7, Residualernas fördelning, linjaritet 8](#_Toc196648629)

## Tabellförteckning

[Tabell 1, Resultat av de olika modellerna. 9](#_Toc196648630)

## Källförteckning

Araz Rawshani, A. R. (2025). *Science.nu*. Hämtat från Science.nu: https://science.nu/

CFI. (2025). *CFI*. Hämtat från https://corporatefinanceinstitute.com/resources/data-science/heteroskedasticity/

Geron, A. (2019). *Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow.* O'Reilly Media, Inc.

Grolemund, H. W. (2016). *R for Data Science.*

*IT-ord*. (2018). Hämtat från https://it-ord.computersweden.se/ord/skrapa/

Lagerson, A. (u.d.). Sammanfattning av mina erfarenheter, sorry för dålig källa.

Prgomet, A. (Regissör). (2023). *Support Vector Machines* [Film].

SCB. (2025). *SCB - Sveriges befolkning*. Hämtat från https://www.scb.se/hitta-statistik/sverige-i-siffror/manniskorna-i-sverige/sveriges-befolkning/

Wermke, W. (u.d.). *Uppsala universitet*. Hämtat från https://media.medfarm.uu.se/play/attachmentfile/video/4787/video.pdf