**Elbilar till försäljning på Blocket**

Analys av avgörande variabler och skapande av regressionsmodell för automatisk prissättning.



Camilla Månsson

EC Utbildning

Kunskapskontroll R programmering

202404

Abstract  
This report conducts regression analysis on electric car sales data from Blocket, an online classifieds platform, focusing specifically on those for sale by companies in Skåne, Sweden. The study aims to develop predictive models for pricing while analyzing variables to identify the significant ones. The findings reveal that mileage and horsepower are indeed significant predictors of pricing. However, limitations arise due to insufficient data distribution and quality, underscoring the importance of careful data selection and model refinement for accurate predictions in future analyses. The findings highlight the need for further research to address these challenges and improve the accuracy of predictive models.

Innehållsförteckning

[1 Inledning 1](#_Toc165986182)

[1.1 Syfte och frågeställning 2](#_Toc165986183)

[2 Teori 3](#_Toc165986184)

[2.1 Linjär regression 3](#_Toc165986185)

[2.1.1 Enkel linjär regression 3](#_Toc165986186)

[2.1.2 Multipel linjär regression 3](#_Toc165986187)

[2.2 Processen för regressionsanalys 4](#_Toc165986188)

[2.3 Modellutvärdering 5](#_Toc165986189)

[2.3.1 P-värde 5](#_Toc165986190)

[2.3.2 VIF 5](#_Toc165986191)

[2.3.3 Adjusterd R-squared 6](#_Toc165986192)

[2.3.4 Bayesian Information Criterion (BIC) 6](#_Toc165986193)

[2.3.5 Root Mean Square Error (RMSE) 6](#_Toc165986194)

[2.3.6 Validering av prediktionsfel 6](#_Toc165986195)

[3 Metod 7](#_Toc165986196)

[3.1 Dataset 7](#_Toc165986197)

[3.2 Datainsamling 7](#_Toc165986198)

[3.2.1 Avgränsning 7](#_Toc165986199)

[3.3 Databearbetning 7](#_Toc165986200)

[3.4 Explorativ dataanalys 8](#_Toc165986201)

[3.5 Träning av modeller 10](#_Toc165986202)

[3.6 Modellutvärdering 10](#_Toc165986203)

[4 Resultat och Diskussion 11](#_Toc165986204)

[5 Slutsatser 13](#_Toc165986205)

[Källförteckning 14](#_Toc165986206)

# Inledning

I dagens samhälle står vi inför en omvälvande förändring inom fordonsindustrin, drivet av ett eskalerande behov av att minska koldioxidutsläppen och övergå till hållbara transportalternativ. En central del av denna förändring är EU:s ambitiösa mål om att från och med 2035 förbjuda försäljning av nya bensin- samt dieselbilar och istället främja marknaden för elbilar. Detta initiativ är en del av en bredare strategi för att göra transportsektorn koldioxidneutral senast år 2050 (Europaparlamentet, 2022).

Elbilsmarknaden i Sverige har växt betydligt under de senaste åren. Mellan 2020 och 2023 har den nästan tredubblats i storlek. Trots denna tillväxt utgör elbilar fortfarande en liten del av det totala antalet bilar på vägarna. Vid årsskiftet mellan 2022 och 2023 var endast 4 procent av alla registrerade bilar eldrivna. Det är dock intressant att notera att andelen nyregistrerade elbilar år 2022 utgjorde hela 32 procent av det totala antalet nya bilar. Dessutom har antalet privatpersoner som väljer elbil ökat avsevärt. Enligt Statistiska centralbyrån är majoriteten av elbilsägare män, med 72 procent, jämfört med kvinnor (SCB, 2023).

Det totala antalet bilar i trafik i Sverige har också ökat under de senaste åren vilket syns i figur 1 nedan.

Figur 1. Det totala antalet bilar i trafik i Sverige år 2014 till 2022. Uppgifterna är hämtade från Statistikmyndigheten SCB:s hemsida.

Ökningen av antalet fordon i trafik återspeglas även i figur 2 som visar antalet bilar i trafik i Skåne. Det är också tydligt att det finns en större andel manliga ägare jämfört med kvinnliga.

Figur 2. Antalet bilar i trafik i Skåne år 2014 till 2023, uppdelat per ägarkategori. Uppgifterna är hämtade från Statistikmyndigheten SCB:s hemsida.

För att möta den växande trenden mot elbilar och den ökande mängden fordon i trafik kan det bli nödvändigt för företag att justera sina affärsmodeller och prissättningsstrategier för just elbilar. Genom att använda automatiserade prissättningsmodeller kan företag effektivt hantera och anpassa sin prissättningsstrategi till marknadens dynamik och de skiftande konsumentbeteendena. Denna studie syftar till att utforska dessa möjligheter genom att tillämpa regressionsanalys och statistisk inferens för att undersöka prissättningen av elbilar som säljs av företag i Skåne på annonssidan Blocket.

## Syfte och frågeställning

Syftet med denna studie är att försöka förstå vad som påverkar priset på elbilen och att skapa en regressionsmodell för att prediktera priset. För att uppfylla syftet så kommer följande frågeställningar att besvaras:

1. Kommer variabeln färg visa signifikans i prissättningsmodellen eller vilka variabler är signifikanta?
2. Kan den identifierade bästa modellen förutsäga priset med en felmarginal på mindre än 50 000 kr jämfört med det faktiska priset?
3. Bedöms vår insamlade data vara tillräcklig för att generera tillförlitliga prediktioner om elbilspriser?

# Teori

I detta avsnitt tas några grundläggande begrepp och metoder inom regressionsanalysen upp. Genom en kortare förklaring av dessa är målet en bättre förståelse för hur sambandet mellan variabler kan analyseras och tolkas. Men också hur modellers kvalitet och prestanda kan bedömas.

## Linjär regression

Linjär regression är trots modernare metoder fortfarande en användbar och vanligt förekommande statistisk inlärningsmetod. Det är en metod inom supervised learning som används för att prediktera kvantitativa svar (James, 2023).

### Enkel linjär regression

Enkel linjär regression går ut på att förutsäga en responsvariabel baserat på en enda förklarande variabel. Den antar att det finns en ungefärligt linjär relation mellan den förklarande variabeln och responsvariabeln. Matematiskt kan denna linjära relation skrivas som: Y ≈ β0 + β1X, där Y är responsvariabeln, X är den förklarande variabeln, β0  är interceptet och β1 är lutningen i den linjära modellen. Interceptet är det värde där linjen korsar y-axeln. För att beräkna uppskattningar av de okända parametrarna kan vi använda träningsdata och sedan använda uppskattningarna för att prediktera framtida värden av responsvariabeln baserat på givna värden av den förklarande variabeln (James, 2023).

### Multipel linjär regression

Multipel linjär regression är en statistisk metod som används för att förstå sambandet mellan flera förklarande variabler och en responsvariabel. Istället för att bara ha en förklarande variabel, som i enkel linjär regression, kan multipel linjär regression hantera flera variabler samtidigt. Modellen tar hänsyn till varje förklarande variabels individuella påverkan på responsvariabeln genom att tilldela separata koefficienter till varje variabel. På så sätt kan vi förstå hur varje förklarande variabel bidrar till förändringar i responsvariabeln, medan vi håller de andra variablerna konstanta. Den matematiska ekvationen för multipel linjär regression kan skrivas

Y = β0 + β1X1 + β2X2 + ... + βpXp + ε

* Y responsvariabeln.
* X1, X2, ..., Xp är de förklarande variablerna.
* β0 är interceptet, vilket är det förväntade värdet för Y när alla förklarande variabler är noll.
* β1, β2, ..., βp är koefficienterna som mäter den den förändring i förväntade värdet för Y för varje enhet förändring i motsvarande förklarande variabel, medan de andra variablerna hålls konstanta.
* ε är feltermen, som representerar den slumpmässiga variationen eller avvikelsen från den linjära relationen mellan de förklarande och responsvariablerna (James, 2023).

#### Best Subset Selection

Best subset selection är en algoritm som används för att välja den bästa uppsättningen variabler från en större mängd möjliga variabler för att bygga en regressionsmodell. Målet är att välja de variabler som ger bäst resultat. Den första modellen M0 är en nullmodell som inte innehåller några variabler men predikterar medelvärdet för varje observation. För varje möjligt antal variabler dvs från 1 till det totala antalet variabler p passas alla möjliga kombinationer av prediktorer in i modellerna. Den kombination som har minst RSS eller störst R2 väljs som den bästa modellen för det aktuella antalet variabler och betecknas som Mk. Den bästa modellen av M0,…,Mp väljs sedan med hjälp av en metod såsom cross-validation eller med hjälp av AIC, BIC eller adjusted R2 på validideringsdelen av datasetet.

En bild som visar text, skärmbild, Teckensnitt, nummer

Automatiskt genererad beskrivning

Figur 3. Algoritmen för Best subset selection.

## Processen för regressionsanalys

I en typisk arbetsgång för regressionsanalys inleds processen ofta med formuleringen av en hypotes där målet är att undersöka sambandet mellan variabler och förstå den potentiella påverkan de har på resultatet. Efter att hypotesen eller det förutsagda variabeln har identifierats går nästa steg ut på att inspektera och förstå den tillgängliga datan. Detta kan innebära att utföra explorativ dataanalys för att få en känsla för datan och identifiera eventuella mönster eller avvikelser. Efter att datan har utforskats är nästa steg att utveckla och estimera regressionsmodeller. Detta involverar vanligtvis att välja lämpliga oberoende variabler och att använda en regressionsmetod för att anpassa modellen till datan. När regressionsmodellerna har skapats är det viktigt att diagnostisera eventuella problem eller avvikelser som kan påverka modellens validitet eller prediktionsförmåga. Detta kan inkludera att undersöka följande punkter (Progomet, n.d.):

1. Icke-linjära relationer: Den linjära regressionsmodellen förutsätter att sambandet mellan den beroende variabeln och de oberoende variablerna är linjärt. Om detta antagande inte är rimligt kan modellens pålitlighet ifrågasättas. Residualerna (residual = skillnaden mellan det observerade värdet och det predikterade värdet) kan visualiseras i förhållande till de förutsagda värdena för att upptäcka eventuella icke-linjära mönster. Om sådana mönster observeras kan icke-linjära transformationer av de oberoende variablerna övervägas.
2. Korrelerade residualer: Om residualerna är korrelerade indikerar det att de inte är oberoende av varandra, vilket bryter mot ett av regressionsmodellens antaganden.
3. Heteroskedasticitet: Om variansen hos residualerna inte är konstant över olika nivåer av de oberoende variablerna kan det leda till problem med modellens precision och validitet.
4. Ej normalfördelade residualer: För statistisk inferens är det viktigt att residualerna följer en normalfördelning. Detta kan undersökas med hjälp av histogram och QQ-plot.
5. Avvikande observationer (outliers): Observationer där det verkliga värdet avviker markant från det förutsagda värdet kan påverka modellens resultat betydligt och bör granskas noggrant.
6. Observationer med högt inflytande (high leverage): Punkter med ovanligt höga eller låga värden på de oberoende variablerna kan ha ett betydande inflytande på modellens passform och bör identifieras och undersökas.
7. Kollinearitet/Multikollinearitet: När två eller flera oberoende variabler är starkt korrelerade kan detta leda till problem med modellens stabilitet och tolkning av effekterna av varje variabel (Progomet, n.d.).

Genom att noggrant diagnostisera dessa aspekter kan modellens tillförlitlighet och precision förbättras och det kan säkerställas att slutsatserna är välgrundade och generaliserbara. Vid upptäckt av problem är det viktigt att försöka åtgärda dessa för att förbättra modellerna. Efter åtgärder genomförs en ny kontroll för att säkerställa att problemen har lösts (Progomet, n.d.).

När modellerna anses tillfredsställande görs vanligtvis en jämförelse mellan dem om det finns flera alternativ. Efter att en modell har valts för tolkning kan dess koefficienter granskas, hypoteser prövas och konfidensintervall analyseras. Dessutom kan prediktioner utföras och analyseras för att förstå modellens prestanda (Progomet, n.d.).

Det är viktigt att komma ihåg att även om regressionsmodellen kan vara kraftfull och informativ, har den sina begränsningar. Det är avgörande att förstå och kommunicera dessa begränsningar för att undvika förhastade slutsatser av modellens resultat (Progomet, n.d.).

## Modellutvärdering

Modellutvärdering är en avgörande del av regressionsanalysen som syftar till att bedöma och jämföra prestandan hos olika regressionsmodeller. Genom att noggrant utvärdera modellernas förmåga att förklara variationen i data och deras förmåga att generalisera till nya observationer kan forskare och analytiker fatta välgrundade beslut och dra pålitliga slutsatser.

### P-värde

P-värdet är en viktig statistisk mätare som används för att bedöma om en observerad effekt är statistiskt signifikant. Det representerar sannolikheten att observera ett resultat lika extremt eller mer extremt än det faktiska resultatet under antagandet att nollhypotesen är sann. Ju lägre p-värde desto mer avvikande är resultatet från nollhypotesen. Ett lågt p-värde indikerar att sannolikheten för att den observerade effekten uppstår av slumpen är minimal, vilket stärker förtroendet för att det faktiskt finns en verklig skillnad eller effekt i den population som studeras.

### VIF

VIF är ett användbart verktyg för att upptäcka multikollinaritet bland de oberoende variablerna i en regressionsmodell. När VIF-värdena är höga för vissa variabler indikerar det att dessa variabler är starkt korrelerade med andra variabler i modellen, vilket kan göra det svårt att skilja deras individuella effekter och kan minska modellens stabilitet. Att identifiera och hantera multikollinaritet är viktigt för att säkerställa att modellen ger korrekta och pålitliga resultat.

### Adjusterd R-squared

Adjusted R-squared är ett mått som ger en indikation på hur väl modellen passar data med hänsyn till antalet variabler och bidrar till en rättvis jämförelse mellan modeller. Ju högre värde ju bättre modell.

### Bayesian Information Criterion (BIC)

Bayesian Information Criterion (BIC) är ett mått som används för modellval. BIC tar hänsyn till både modellens passning till data och dess komplexitet. Lägre BIC-värden indikerar en bättre balans mellan modellens passning och komplexitet.

### Root Mean Square Error (RMSE)

Root Mean Square Error (RMSE) är ett mått på modellens förutsägelsefel. Det representerar genomsnittlig avvikelse mellan de observerade och förutsagda värdena och kan användas för att utvärdera modellens noggrannhet.

### **Validering av prediktionsfel**

För att validera modellens prediktionsförmåga kan man använda sig av ny data. Genom att jämföra förutsagda värden med verkliga observationer i en separat testuppsättning kan man bedöma modellens förmåga att generalisera till nya data.

# Metod

I detta avsnitt beskrivs den metod som användes för att genomföra regressionsanalysen. Nedan följer varje steg i processen, från att välja och förbereda datan till analyser och evaluering av modeller. All kod som användes i analysen finns tillgänglig i en separat R-fil.

## Dataset

Datasetet som används består av egeninsamlad data från annonssidan Blocket och avser elbilar. Datasetet består av 467 observationer.

## Datainsamling

Datainsamlingen utfördes som ett samarbetsprojekt, där varje deltagare ansvarade för att manuellt samla in observationer från Blocket.se. Varje gruppmedlem samlade mellan 90 och 100 observationer. De relevanta uppgifterna från varje annons inkluderade annons-ID, märke, modell, miltal, hästkrafter, modellår och färg. För att säkerställa en enhetlig insamling klipptes uppgifterna in i ett gemensamt Excel-dokument.

### Avgränsning

I gruppen togs ett gemensamt beslut om att begränsa vår datainsamling till endast elbilar som fanns till försäljning i Skåne och som såldes av företag. Vi valde att inte ta med bilar som hade ett miltal under 50 mil.

## Databearbetning

Efter att vi hade samlat in alla observationer fortsatte vi arbetet individuellt. Trots att databearbetningen enkelt kan utföras i R, valde jag initialt att använda Excel. I excel hanterade jag saknade värden, som dock var få, och redigerade blanksteg som smugit sig in. Jag valde även att ta bort annons-id kolumnen i den fil som sedan laddades in i R, då jag inte fann den informationen relevant för uppgiften.  
  
Under databearbetningsfasen konverterades variablerna i R till lämpliga typer för att underlätta hanteringen och analysen av datan. Detta inkluderade omvandling av variabler till numeriska eller faktorvariabler beroende på deras egenskaper och syfte. Dessutom skapades en kategori för modellår för att underlätta förståelsen och analysen av datan.

## Explorativ dataanalys

För att tydliggöra hur den insamlade datan såg ut genomfördes en explorativ dataanalys. Under detta steg skapades olika visualiseringar för att utforska datan och identifiera eventuella problem. Histogram användes för att undersöka fördelningen av numeriska variabler som pris, miltal och hästkrafter. Det blev uppenbart att y-variabeln ”Pris” inte följde en normalfördelning vilket ledde till att jag senare valde att logaritmera den för att uppnå normalfördelning. Denna förändring framgår tydligt i figur 4

## 

A graph of a car

Description automatically generated

Figur 4. Histogram över Y-variabeln ”Pris” före respektive efter logaritmering. Bild från egen visualisering i R.

Dessutom användes stapeldiagram för att visualisera fördelningen av antal bilar per färg och märke. Efter detta togs beslutet att utföra viss förändring av datan för att hantera sällsynta observationer. Detta inkluderade att gruppera färger respektive märken med få observationer under en gemensam kategori. Märken med färre än 30 observationer grupperades tillsammans under ”övriga märke”. På samma sätt grupperades sällsynta färger under ”övriga färger”.

Variabeln ”modell” valdes att ta bort helt då det var så extremt många olika modeller och väldigt få observationer per modell.

Scatter plots användes för att undersöka sambandet mellan variabler som pris och hästkrafter samt pris och miltal. I figur 5 syns resultatet. Punkterna följer ganska bra en uppåtgående linje vilket indikerar ett positivt linjärt samband mellan hästkrafter och pris. Det innebär att bilar med högre hästkrafter tenderar att ha högre pris. Den blå linjen är en linjär regressionslinje som har en positiv lutning som bekräftar det positiva sambandet. Det finns dock en del spridning runt   
  
regressionslinjen och det i sin tur kan vara en indikation på att det finns andra faktorer som påverkar priset på bilen.  
  
Vad gäller sambandet mellan miltal och pris följer det en nedåtgående linje vilket visar ett negativt linjärt samband. Det innebär att bilar med högre miltal tenderar att ha lägre pris. Även här finns en del spridning runt regressionslinjen vilket visar på att det troligtvis finns andra faktorer som påverkar priset.

A graph of different types of graphs

Description automatically generated with medium confidence

Figur 5. Sambandet mellan hästkrafter och pris samt miltal och pris. Egen visualisering skapad i R.

Boxplots och violinplot användes för att undersöka fördelningen av pris för olika kategorier som märke, modellår och färg.

## Träning av modeller

För att utveckla modeller för att förutsäga priset på bilar användes flera linjära regressionsmodeller. Innan modellerna skapades delades datasetet upp i träning, validering och test. Den första modellen, ”Logisk”, inkluderade egenvalda variabler såsom modellårskategori, miltal och hästkrafter. Den andra modellen, ”Full”, utgick från en fullständig modell med alla variabler medan den tredje modellen, ”Best subset”, byggdes med hjälp av best subset regression. För varje modell utfördes även viss datarening såsom att hantera multikollinearitet, borttagning av outliers och urval av endast signifikanta variabler för att förbättra modellens prediktionsförmåga.

## Modellutvärdering

För att utvärdera modellernas prestanda användes olika metriker såsom Adjusted R^2 och BIC samt senare även RMSE. Dessa metriker användes för att jämföra modellerna och identifiera den som bäst passar datan och ger mest tillförlitliga prediktioner. Dessutom genomfördes en utvärdering av modellerna på en valideringsuppsättning för att bedöma deras generaliserbarhet och prediktionskraft på nya data. Därefter valdes den bästa av de tre som utvärderades på den osedda testdatan.

# Resultat och Diskussion

I detta avsnitt presenteras resultaten av modellutvärderingen samt en diskussion om de observerade resultaten och tillvägagångssätten som användes under analysen.

I tabell 1 nedan visas resultatet efter utvärdering av modellerna på valideringsdata.

|  |  |  |  |
| --- | --- | --- | --- |
| **Utvärdering av modellerna på valideringsdata** | | | |
| **Modell** | **RMSE** | **Adjusted R-squared** | **BIC** |
| Logisk | 144084.3 | 0.6550690 | -118.6065 |
| Full | 133812.3 | 0.7458778 | -193.2287 |
| Best Subset | 136162.8 | 0.6883512 | -123.8792 |

Tabell 1: RMSE, Adjusted R-squared samt BIC för de tre olika modellerna.

Modellen "Full" valdes eftersom den uppnådde bäst resultat och när priser predikterades på testdata med den valda modellen blev RMSE 180263.3. Det är värt att notera att valideringsprediktionerna var bättre än på testdata. En huvudorsak till detta kan vara den begränsade storleken på datasetet samt variationen i observationer mellan validerings- och testsetet.

Våra observationer är få och ganska spridda. Det finns betydande prisskillnader och inte tillräckligt med observationer inom olika prisklasser eller med olika märken eller egenskaper. Denna brist på mångfald kan påverka modellens förmåga att generalisera korrekt till nya data. Istället för att dela upp datasetet kanske en cross-validation-metod hade varit lämpligare. Detta skulle möjliggöra en bättre användning av den befintliga datan och därmed förbättra modellens prestanda.  
  
Under analysen undersöktes även konfidensintervall och prediktionsintervall i förhållande till de faktiska priserna för observationerna. Trots denna analys kunde jag inte urskilja några tydliga mönster mer än att modellen inte predikterar priserna så bra som önskat, se figur 6.

A graph with a red line

Description automatically generated

Figur 6. Sambandet mellan det predikterade priset och det faktiska priset. Prickarna avser respektive observation. Den röda linjen representerar en perfekt matchning mellan det predikterade priset och det faktiska priset.

När det kommer till behandlingen av variabler som märke och färg, fattades beslutet att kombinera dem på grund av det begränsade antalet observationer för vissa kategorier. En alternativ strategi skulle ha varit att helt utesluta dessa variabler även om det skulle ha resulterat i ännu färre observationer. Detta är en aspekt som kan granskas närmare för att bedöma dess inverkan på modellens prestanda. Vidare kan hanteringen av variabeln modellår, där den delades upp i ny och medelålder, ifrågasättas. Beslutet fattades utan noggrann motivering och grundades enbart på intuition vilket kan innebära att det finns utrymme för förbättringar och överväganden kring alternativa tillvägagångssätt.

För att förbättra resultatet och modellens prestanda finns det flera potentiella åtgärder att överväga. Trots att diagnostic plots användes utforskades inte alla potentiella problemområden noggrant. Det är möjligt att det finns andra aspekter som kan förbättras för att skapa en mer effektiv modell. En ökning av datamängden skulle möjliggöra ett mer representativt och varierat dataset vilket i sin tur skulle förbättra modellens förmåga att generalisera korrekt. Vidare kan utforskning av mer avancerade modelleringsmetoder bidra till att bättre hantera datan och därigenom förbättra prediktionsförmågan. Slutligen kan ytterligare undersökningar genomföras för att identifiera eventuella brister i datan och modellen samt för att utveckla mer robusta och tillförlitliga prediktionsmodeller.

# Slutsatser

Under arbetet med modelleringen har det blivit uppenbart att balansen mellan att finjustera variabler och koda kan vara en utmaning. Det är svårt att fastställa en definitiv gräns för när modellen är tillräckligt bra för uppgiften och detta kan vara särskilt utmanande för nybörjare som måste navigera genom en mängd olika alternativ.

För att besvara de ursprungliga frågeställningarna:

1. **Signifikans av färgvariabeln:** Resultaten visar att färgvariabeln inte är signifikant för att förutsäga priset på bilar. Däremot har både miltal och hästkrafter visat sig vara betydelsefulla.
2. **Prediktion av pris under 50 000 kr fel:** Tyvärr kunde inte den bästa modellen förutsäga priset med tillräcklig precision inom en felmarginal på 50 000 kr baserat på den insamlade datan.
3. **Tillräcklighet av insamlad data för prediktioner:** Även om antalet observationer är tillräckligt, är fördelningen av data av avgörande betydelse. Det är svårt att definitivt bedöma om den insamlade datan ensam är tillräcklig för att generera tillförlitliga prediktioner om elbilspriser. Men det blir tydligt att kvaliteten på insamlad data är avgörande för att uppnå tillfredsställande prediktioner. Det är viktigt att noggrant överväga datakvalitet och distribution vid insamling av data och att undvika för många bortval och selekteringar i förväg. Vid framtida datainsamling är det viktigt att noggrant överväga dessa faktorer för att säkerställa att modeller byggs på en stabil grund.

Sammanfattningsvis var detta en intressant uppgift som visade på vikten av att förstå data och dess inneboende egenskaper samt hur små variationer kan påverka resultatet av modelleringen.

# Källförteckning

**Böcker:**   
  
James, G., Witten, D., Hastie, T., & Tibshirani, R. (2023). An Introduction to Statistical Learning with Applications in R (2nd ed.).

**Webbkällor:**

Bildata Blocket. (n.d.). Hämtad 13 April 2024 från <https://www.blocket.se>.

Europaparlamentet. (2022, 19 Oktober). "EU:s förbud mot försäljning av nya bensin- och dieselbilar." Hämtad från <https://www.europarl.europa.eu/topics/sv/article/20221019STO44572/eu-s-forbud-mot-forsaljning-av-nya-bensin-och-dieselbilar>.

Progomet, A. (n.d.). Video Linjär Regression [Video]. Hämtad 26 April 2024 från <https://www.youtube.com/watch?v=NcxMuCG6FS8>.

Statistikdatabasen. (n.d.). Hämtad 17 April 2024 från <https://www.statistikdatabasen.scb.se>.

Statistiska centralbyrån. (n.d.). "API för Statistikdatabasen." Hämtad April 2024 från <https://www.scb.se/vara-tjanster/oppna-data/api-for-statistikdatabasen>.

Statistiska centralbyrån. (n.d.). "Tredubbling av elbilar på två år." Hämtad April 2024 från <https://www.scb.se/hitta-statistik/redaktionellt/tredubbling-av-elbilar-pa-tva-ar2/>.