Regressionsanalys



Leonardo Sjöberg

EC Utbildning

2024-04

# Abstract

This project aims to identify predictors of electric vehicle prices using linear regression analysis. By analyzing various factors such as milage, horsepower and color we seek to establish a predictive model for these prices. Ultimately, this project strives to contribute to the understanding and growth of the electric vehicle market.

# Förkortningar och Begrepp

RMSE – Root Mean Square Error

VIF - Variance Inflation factor

BIC - Bayesian Information Criterion

# 

Contents

[Abstract 2](#_Toc165059387)

[Förkortningar och Begrepp 3](#_Toc165059388)

[1 Inledning 1](#_Toc165059389)

[2 Teori 2](#_Toc165059390)

[2.1 Dataanalys 2](#_Toc165059391)

[2.2 Linjär Regressionsmodell 3](#_Toc165059392)

[2.3 Model och variabel val 3](#_Toc165059393)

[2.3.1 Hypotesprövning och variabel val 3](#_Toc165059394)

[2.3.2 Evaluerings Metriker 4](#_Toc165059395)

[2.4 Cross validation 5](#_Toc165059396)

[3 Metod 6](#_Toc165059397)

[3.1 Datainsamling och analys av data 6](#_Toc165059398)

[3.2 Linjär Regressionsanalys 6](#_Toc165059399)

[3.2.1 Antaganden 6](#_Toc165059400)

[3.2.2 Val av oberoende variabler 6](#_Toc165059401)

[3.3 Prediktions evaluering 7](#_Toc165059402)

[3.3.1 Cross validation och prediktering 7](#_Toc165059403)

[4 Resultat och Diskussion 8](#_Toc165059404)

[4.1 Data utforskning 8](#_Toc165059405)

[4.2 Linjär Regressionsanalys 9](#_Toc165059406)

[4.2.1 Antaganden 9](#_Toc165059407)

[4.2.2 Outliers och High Leverage 10](#_Toc165059408)

[4.2.3 Modell och variabel val 10](#_Toc165059409)

[4.3 Prediktions Förmåga 11](#_Toc165059410)

[4.3.1 Cross Validation 11](#_Toc165059411)

[4.3.2 Prediktion med Testdata 11](#_Toc165059412)

[5 Slutsatser 12](#_Toc165059413)

[6 Teoretiska frågor 13](#_Toc165059414)

[6.1 Del (1) Data insamling 13](#_Toc165059415)

[7 Självutvärdering 14](#_Toc165059416)

[Appendix A 15](#_Toc165059417)

[Källförteckning 23](#_Toc165059418)

# Inledning

En stor del av vårt samhälle beror på transport. För många behövs det för att kunna ta sig till och från jobb, få verksamheter att fungera genom att få leveranser med viktiga produkter och för allmänt ta sig till en annan plats för att kunna koppla av och komma från vardagen i form av en semester.

Med många nya tekniker som framkommer till bilens utveckling så som utvecklingen av elbilar, självkörande tekniker och allmänt minskningen på förbränning av drivmedel i mål att minska utsläpp för att nå miljömål, så ser visar det vara en minskning i personbilar i trafik även om ny teknik kommer för att få bilen som ett vardagligt alternativ i stället för kollektivtrafik. Enligt SCB (2024) ser vi i figur (1.1) att personbilar i trafik håller på att minska efter sin kraftiga stigning. Detta tyder på att folk köper allt mindre bilar då personbilarna i trafik minskar. Med denna minskning så medför det eventuellt en förändring i pris på dessa personbilar som redan finns i trafiken.

Ett nytt sätt att köra bil är genom att använda elektricitet som drivmedel och dessa bilar som använder denna metod kallas för elbilar. Dessa bilar är relativt nya i trafiken och har inte varit ute lika länge som bensin och dieselbilar. Dessa elbilar kan kategoriseras väldigt detaljerad och dessa kategorier kan ha en påverkan på priset, men vilka som har påverkan och inte påverkan kan vara en svår fråga att besvara utan att kolla djupare på frågan.

Därför kommer denna rapport att kolla på insamlade data från blocket (Blocket.se) med olika elbilar, göra en regressionsanalys där hypotesprövning kommer att utföras för att hitta de faktorer som påverkar priset på elbilar samt testa modellens styrka för att prediktera värdet på elbilar med faktorerna vi har valt.

Syftet med denna rapport är att finna de faktorer som har inverkan på elbilars pris, samt se hur väl modellen kan förutspå priset på elbilar baserat på dess funktioner. För att uppfylla syftet så kommer följande frågeställningar att besvaras:

1. Vilka faktorer har störst inverkan på priset på en elbil.
2. Hur väl kan vi förutspå priset på en elbil baserat på dess funktioner.

A graph with lines showing the growth of the company's company

Description automatically generated with medium confidence

Figur 1.1: Graf på personbilar i trafik övertid.

# Teori

## Dataanalys

Den data som används kommer från Blockets hemsida, [www.blocket.se](http://www.blocket.se) (Blocket, 2024). Data setet innehåller 409 observationer med kolumnerna Miltal, Modellår, Biltyp, Drivning, Hästkrafter, Färg, Märke, Modell, Län och Pris. Följande lista beskriver kolumnerna mer i detaljs:

* Miltal beskriver hur långt elbilen har gått sammanlagt i antal mil. Det lägsta miltalet på elbilarna är 500 mil.
* Modellår beskriver vilket år bilen har tillverkats. I data setet har elbilarna ett modellår mellan 2014 och 2024.
* Biltyp beskriver bilens form och består av fyra olika former. Dessa former är kombi, halvkombi, sedan och SUV.
* Drivning syftar på om bilen har en fyrhjulsdrivning eller en tvåhjulsdrivning.
* Hästkrafter beskriver bilens kraft och skiljer sig mellan 83 - 815.
* Färg beskriver vad elbilen har för färg. Elbilarna kan anta 10 olika färger.
* Märke beskriver vem som har tillverkat bilen. Data setet innehåller 10 olika märken.
* Modell beskriver vilken modellen är från tillverkaren. Data setet innehåller 36 olika modeller.
* Län beskriver vilken plats elbilen säljs från i form av län. Data setet innehåller 52 olika län.
* Pris beskriver annonspriset på elbilen och skiljer sig mellan 89 000 – 890 000.

Alla kolumner i datan består inte av kvantitativa data utan består av kvalitativa data. De data som är kvalitativ är Biltyp, Drivning, Färg, Märke, Modell och Län. För att lösa detta problem behöver dessa kolumner att ändars om till dummyvariabler för att sedan kunna användas i regressionsanalysen. Detta sköter R i bakgrunden när den ska träna upp en linjär regressionsmodell.

A black car on a white background

Description automatically generated

Figur 2.1: Bild på blocket med pris och under fakta hittar vi andra punkter som kan samlas.

## Linjär Regressionsmodell

Linjär regression enligt James et al. (2023) är som de beskriver. Det är ett enkelt och simpelt sätt att försöka prediktera en kvantitativ beroende variabel Y med basis av oberoende variabel X. Det antas att det är en linjär relation mellan X och Y och kan skrivas matematiskt som

*.* (2.1)

β0 och β1 är okända parametrar men med hjälp av tränings data estimera dessa parametrar för en modell som är väldigt lik verkligen och som har förhoppningsvis inte så stort fel. Dessa koefficienter som är estimerade kan skrivas i formel på detta sätt

. (2.2)

(James et al., 2023, s.61)

När vi har mer än en oberoende variabel så måste vi utöka den linjära regressionsmodellen för att ackommodera för flera variabler. Enligt James et al. (2023) kan vi ge varje oberoende variabel en egen koefficient för varje p variabler vi har. Då får vi en regressionsmodell som ser ut så här

. (2.3)

står för brus i datan eller felet i själva modellen på grund av att estimeringarna inte är helt rätt eller att vi har för få koefficienter. Vi kan aldrig vara helt exakta med modell men vi kan alltid försöka göra en bra estimering av verkligheten. (James et al., 2023, s.72)

## Model och variabel val

Vid val av oberoende variabler och modeller så finns det olika metoder för att hitta den bästa modellen med rätt utvalda variabler. En metod som kommer att användas i detta projekt är Best subset selection. Enligt James et al. (2023) alla möjliga kombinationer av variabler och väljer den som ger minst fel med hjälp av RMSE. Alltså kollar den alla modeller med bara en variabel och ger sedan den som presterar bäst, sedan med 2 variabler och gör det med alla *p* variabler. (James et al., 2023, s.227) Den bästa modellen kommer best subset selection som kommer använda sig av hypotesprövning för att hitta de signifikanta variablerna.

### Hypotesprövning och variabel val

För att se om det finns någon relation mellan den beroende variabeln **Y** och den oberoende variabeln **X** kan man använda sig av hypotesprövning. Den mest vanliga hypotesprövningen enligt James et al. (2023) är testa noll hypotesen mot den alternativa hypotesen. I detta fall vill man testa om det inte finns någon relation mellan variablerna vilket är noll hypotesen, mot att det finns en relation mellan variablerna vilket är den alternativa hypotesen. Detta kan skrivas matematisk som

,

Vilket motsvarar noll hypotesen, och sedan

,

Som motsvarar den alternativa hypotesen. För att testa noll hypotesen måste vi fastställa att vår estimering av beta-koefficienten är tillräckligt långt från noll så att vi kan vara säkra på att det inte är lika med 0. Är beta-koefficienten tillräckligt långt från värdet noll kan vi förkasta noll hypotesen och anta att där finns en relation mellan **X** och **Y**. Man använder sig av t-värdet för att testa noll hypotesen där

Ett stort värde på **t** indikerar på att där finns en relation mellan **X** och **Y** och att noll hypotesen kan förkastas, medan ett litet värde på t indikerar på att där inte finns någon relation mellan **X** och **Y** vilket tyder på att noll hypotesen inte kan förkastas. (James et al., 2023, s.67)

### Evaluerings Metriker

Dessa metriker har använt genom hela projektet för att evaluera modellers prestanda samt för att finna eventuella multikollinearitet mellan oberoende variabler. Följande Metriker kommer att beskrivas för hur de har använts och vad de mäter.

* R- Squared enligt James et al. (2023) mäter hur mycket av variansen i den beroende variabeln beror på dem oberoende variablerna. Det mäts genom att ta den totala summan av kvadrater och subtraherar det med summan av residualerna i kvadrat för att då få fram totala summan rätt i kvadrat, som sedan divideras med totala summan av kvadrater för att få ett värde mellan 0–1. De kan skriva som

(James et al., 2023, s.70) Denna Metrik kommer att användas för att evaluera modellers prestanda och välja ut den bästa.

* C(p) enligt James et al. (2023) försöker estimera test Medel kvadrat felet genom följande ekvation

.

Det estimeras genom att ta summan av residualerna i kvadrat adderat med ett straff då träningsfelet tenderar att underskatta test felet, som sedan divideras med antalet observationer (James et al., 2023, s.233). Metriken används för att jämföra modellers prestanda och man söker efter det som ger minst värde.

* BIC står för Bayesian Information Criterion och är ganska lik till C(p) värdet. Enligt James et al. (2023) är den större skillnaden att den tar ett större straff på modeller som har många variabler vilket gör att den gynnar modeller med färre variabler. Formeln för BIC följer

(James et al., 2023, s.243)

Denna metrik kommer att användas för att mäta modellers prestanda och modeller med låg BIC eftersöks.

* VIF står för Variance Inflation factor. Enligt James et al. (2023) används VIF för att mäta kollinearitet hos de oberoende variablerna som används i modellen. VIF är förhållandet mellan variansen av den estimerade beta koefficienten på hela modellen dividerat med variansen på den estimerade beta koefficienten på modellen med bara den koefficienten. Ett värde på ett betyder att det inte finns någon kollinearitet. Formeln följer

och kommer användas för att hitta eventuella kollinearitet mellan variabler. Variabler med VIF över fem kommer att undersökas och eventuellt tas bort. (James et al., 2023, s.102)

* RMSE står för Root Mean Squared Error. Enligt James et al. mäter det modellens medelfel genom att ta medelvärdet av det sanna värdet subtraherat med det predikterade värdet i kvadrat för att få positiva värden, som sedan tas kvadratroten av för att få ett värde som är lätt tolkat. Formeln följer

och används för att kolla modellens prediktionsfömåga. (James et al., 2023, s.29)

## Cross validation

K-fold cross validation är en metod för att estimera en modells prestanda utan att använda sig av testdata. Enligt James et al. (2023) delar man upp sitt dataset i k olika grupper där man tränar upp modellen på k-1 grupper och använder den uteblivna gruppen för att validera modellen. man byter sedan ut validerings datan med en annan grupp och tränar om modellen och utför detta k gånger. Sedan tar man medelvärdet på alla modellers metriker som RMSE och R-squared för att se hur modellen presterar på valideringsdatan (James et al., 2023, s.203). Detta är ett bra sätt att värdera sin modell och se på ett ungefär hur väl den presterar osedd data.

# Metod

För att nå svaren i frågeställningarna har praktiskt arbete genomförts och detaljer kan hittas i kod (Appendix A).

## Datainsamling och analys av data

Denna del kollar på datainsamling och dataanalys. Det har genomförts på följande sätt,

1. [www.blocket.se](http://www.blocket.se) använts med filter med 500 mil och upp, biltyp som är kombi, halvkombi, sedan och SUV, samt att drivmedel är el.
2. Samla data på de olika bilarna i en excel fil. Kolumner som samlas in är Pris, Miltal, Modellår, biltyp, drivning, hästkrafter, färg, Märke, modell och län.
3. Ladda upp excel filen i R.

Dataanalysen har genomförts på följande sätt,

1. Kolla data kolumnerna så att de har rätt datatyp.
2. Inspektera datan så att de inte finns några saknade värden, så att alla värden överensstämmer med varandra och att där finns tillräckligt med observationer för olika värden.
3. De värden med få observationer samlas eventuellt in till andra. Till exempel att ljusgrå och mörkgrå samlas under gruppen grå i stället.
4. Inspektera linjäritet mellan de olika variablerna för att se om det finns kollinearitet och linjäritet mellan oberoende variabel och beroende variabel.
5. Dela upp setet i tränings och testdata.

## Linjär Regressionsanalys

Vanlig linjär regression analys genomfördes på datasetet och dessa steg tog för att uppnå en analys på rätt sätt. Först har antaganden kollats för att se att en linjär regressionsanalys passar för detta problem. Sedan har rätt variabler valts för att passa en modell som presterar bäst på datan. Sedan har Cross Validation utförts för att se hur väl datan presterar, för att senare evalueras på testdata för att verkligen se hur väl modellen presterar. Denna del kommer att använda träningsdata.

### Antaganden

Antaganden har testats och kollats med olika plots som till exempel QQ-plot för att se raden av normalitet. Följande steg utfördes.

1. Skapa en linjär regressionsmodell med alla utvalda oberoende variabler mot den beroende variabeln Pris.
2. Kolla att andtagenden som Linjäritet, Normalitet, att variansen på fel är konstant, att medelfelet är 0 och kolla eventuella outliers och high leverage punkter i data setet.

### Val av oberoende variabler

Vid val av oberoende variabler har best subset selection utförts för att hitta den bästa modellen, samt se hur många variabler och vilka som behövs för att skapa den bästa modellen. Dessa steg har utförts.

1. Använd tränings data för att använda best subset selection för att hitta den bästa modellen.
2. Inspektera C(p), BIC, Adjusted R-Square för att hitta den bästa modellen med de bästa variablerna.

## Prediktions evaluering

För att evaluera slutmodellen har Cross validation och prediktering på testdata. Metrikerna RMSE och R-Squared användes för att kolla modellen predikteringsförmåga.

### Cross validation och prediktering

Dessa steg har utförts för att evaluera vår modell.

1. Använd tränings datasetet för att evaluera modellen.
2. Cross validation används för att kolla modellens prestanda på valideringdata.
3. Använd testdata för att prediktera testdata.
4. Analysera deras R-Squared och Root Mean Squared Error.

# Resultat och Diskussion

## Data utforskning

Den insamlade datan från blocket var utforskad och analyserad, och från den utforskningen hitta några problematik med datan som vart tvungen att ändras. Följande problem blev funna,

* Stort antal dummies som påverkar curse of dimensionality.
* Dummies med för få observationer.
* Kollinearitet mellan variabler.

Dessa problem har lösts genom att använda tekniken Binning för att samla ihop Dummies med få observationer, samt har vissa variabler tagits bort för att minska antalet dummies samt lösa problemet med kollinearitet.

De variabler med dummies där dummies har samlats ihop är Färg där nyanser av vissa färger har samlats ihop då de har väldigt få observationer. Tanken bakom detta är att sänka variansen och eventuell överanpassning i utbyte mot en ökad bias. Till exempel på färger som samlas ihop är ljusgrå och mörkgrå blir grå. Då dessa redan är grå så bör det inte ha någon stor påverkan på själva modellen.

Där finns två variabler med alldeles för många unika observationer. Detta kommer skapa ett stort antal dummies vilket kommer göra att modellen har för få observationer per variabler. Där finns två variabler som har stort antal dummies vilket är Län och Modell. Dessa tillsammans skapar över hundra dummies vilket kommer ge mindre än fyra observationer per variabel och vi eftersöker tio och över. Dessa variabler kommer att tas bort för att minska variansen för en liten ökning i bias. Tanken med att ha med Län är för att se om priset ökar ju mer norrut man kommer då det kan finnas mindre utbud i norr än i söder. Tanken med Modell är att se hur modell av bil kan påverka mer priset längre än bara märket, men då det finns många olika modeller kommer det skapa problem för modellen då det kan skapa en överanpassning.

Där finns en eventuell kollinearitet mellan variablerna Modellår till både Miltal och Hästkrafter. Detta kan skapa problem för modellen då de kan använda en variabel som predikor när den inte egentligen har någon signifikans för modellen då den är en del av en annan variabel. I figur 4.0 ser man en tendens till en linjär minskning i miltal i förhållande till modellår, samt en ökning i hästkrafter i förhållande till Modellår. Denna observation kan förklaras som att ju äldre en bil är desto mer har den kört vilket är vanligt för bilar att de används över åren. Sedan ser vi att hästkraften ökar generellt med åren vilket beror på att bilarna blir bara bättre och kraftigare med åren vilket kan reflekteras i åren. Även om VIF-värdena ligger under 5 kan man resonera fram till varför det finns kollinearitet mellan dessa variabler och därför kommer Modellår att tas bort.

A graph with numbers and a number of numbers

Description automatically generated with medium confidenceA graph of numbers and lines

Description automatically generated with medium confidence

Figur 4.0: Grafer på Miltal och Hästkrafter över Modellår som visar eventuell kollinearitet.

## Linjär Regressionsanalys

### Antaganden

Vi använder nu de variabler som vi valt ut för att träna upp en modell och vi ser direkt problem med våra antaganden enligt Figur (4.1).

* Datan verkar inte helt normalfördelad och väldigt höger skevad.
* Variansen av fel är inte konstant.

Att dessa antaganden inte är mötta skapar stora problem då vi inte kommer kunna prediktera rätt samt vi inte kommer kunna utföra någon statistisk inferens då datan inte är normalfördelad. Ett sätt att försöka lösa detta problem är att transformera vår beroende variabel. Genom att använda kvadratroten av Priset ser vi i figur (4.2) att vår data följer ett linjärt samband, vi får ett mer normal fördelat data set, samt att variansen av fel blir mer konstant. Detta gör dock vår modell mer svårtolkad då vi letar efter kvadratroten av priset och inte det vanliga priset. En transformation kan också hjälpa till med normalfördelnings och variansproblemet. Det minskar skevheten drastiskt då det påverkar större värden mer än lägre, samt att de stabiliserar variansen.

Från figur (4.2) ser vi att vissa punkter påverkar grafen negativt. I Residuals vs fitted som kollar linjäritet ser vi att den avtar i slutet. Det beror på att där är bara punkter under grafen som påverkar detta, men detta tyder inte på att data är ickelinjär för om de tre sista punkter tas bort så kommer linjen att bli rak och visa på linjäritet. Detsamma kan sägas för Scale-location. Grafen visar att variansen ökar i slutet men det beror mer på att vi har för få värden som representerar de bilar med högt pris som då gör att vissa punkter drar upp linjen när det egentligen inte behöver typa på ändring i varians för populationen.

Med hjälp av shapiro wilks test som visar att vi har normalfördelad data samt att medelvärdet i fel är 0 så kan vi konstatera att alla antaganden är mötta.

A group of images of a graph

Description automatically generated with medium confidence

Figur 4.1: Fyra olika grafer för att evaluera olika antaganden innan kvadratrots transformation av priset.

A collage of images of a person's body

Description automatically generated

Figur 4.2: Fyra olika grafer för att evaluera olika antaganden efter kvadratrots transformation av priset.

### Outliers och High Leverage

Potentiella outliers och high levarage punkter har undersökts men har konstaterats att även om dessa punkter må vara lite problematiska, så sticker inte dessa punkter ut när de undersöks med deras features. En potentiell outlier är punkt 260 i de olika graferna och undersöker vi det vidare så finner man inga särskilda problem med punkten och där finns där inga riktiga skäl att ta bort den, även om den kan vara en outlier.

### Modell och variabel val

När antagandena är mötta så kommer best subset selection att utföras för att hitta den bästa modellen bland alla modeller. Tabell (4.3) Visar de bästa modellerna för de olika antalen prediktorer. För Bestämma den bästa modellen kommer den med bästa värden på de olika metriker att väljas ut. Som vi ser i tabellen är det då den modellen med fyra prediktorer som presterar bäst ut av de alla. De

|  |  |  |  |
| --- | --- | --- | --- |
| Model (Variabler) | Adjusted R^2 | C(p) | BIC |
| Hästkrafter | 0,4919 | 460,2404 | 8597,6372 |
| Hästkrafter, Märke | 0,6697 | 190,0046 | 8499,6955 |
| Miltal, Hästkrafter, Märke | 0,7567 | 60,5461 | 8404,5534 |
| Miltal, Biltyp, Hästkrafter, Märke | 0,7894 | 14,5869 | 8371,4907 |
| Miltal, Biltyp, Hästkrafter, Färg, Märke | 0,7881 | 25,5241 | 8416,1029 |
| Miltal, Biltyp, Drivning, Hästkrafter, Färg, Märke | 0,7892 | 25,0000 | 8419,1711 |

Tabell 4.3: Resultatet av Best Subset Selection och resultatet från de olika metrikerna.

## Prediktions Förmåga

### Cross Validation

När den bästa modellen är vald kommer den att testas med hjälp av cross validation samt testas helt på testdata för att evaluera modellens prediktionsförmåga. Med hjälp av 10-fold cross validation ser vi i tabellen (4.4) att resultaten som är givna för de olika metrikerna är väldigt bra och att vidare utveckla modellen för en bättre inte är särskilt nödvändigt, därför kan vi gå vidare till att kolla testdata.

### Prediktion med Testdata

Resultatet av att använda testdata är väldigt likt resultaten från Cross Validation vilket är ett bra tecken på att överanpassning inte har sket utan att modellen har generellt sätt en väldigt bra prediktionsförmåga. För att kolla om modellen verkligen predikterar bra kan vi dividera RMSE med medelvärdet av priset för att få ett normaliserat RMSE där värden nära 0 tyder på en bra prediktionsförmåga. Med ett värde på cirka 0.2 kan vi konstatera att vi har en bra modell som ger prediktioner med bra säkerhet. I Tabell 4.4 ser vi värden som har kommit fram av att testa modellen. Den viktigaste är prediktionerna i originalpris som har tagits fram genom att ta prediktionerna i kvadrat. De andra värdena är för att säkerställa att modellen är bra nog för att sedan testas på testdata för att evaluera modellen på osedd data.

|  |  |  |  |
| --- | --- | --- | --- |
| Validering metod | R-Squared | RMSE | Normaliserad RMSE |
| Cross Validation | 0,8152 | 60,12 | 0.0948591 |
| Prediktion (Transformerat Pris) | 0,8263 | 54,01 | 0.0852015 |
| Prediktion (Original Pris) | 0,8124 | 71 513 | 0.1708714 |

Tabell 4.4: Metriker för Cross Validation och prediktion i transformerat och originalpris.

# Slutsatser

Följande slutsatser kan dras från frågeställningarna som från början av rapporten,

1. Ut av de variabler som har använts kan vi konstatera att Miltal, Biltyp, Hästkrafter, Märke är de variabler som påverkar priset
2. Med hjälp av dessa funktioner kan vi få fram en modell som har en relativt säker prediktionsförmåga.

# Teoretiska frågor

1. QQ-plot används vid statistisk inferens för att se om datan som används är normalfördelat. Man ordnar datan i storleksordning och plottar dessa mot motsvarande kvantiler från en standard normalfördelning. Data är närmalfördelan om punkterna från datan ligger på eller nära en rak linje.
2. Maskininlärning fokuserar mestadels på att försöka prediktera och förutsäga framtida data baserat på historisk data. Detta kan till exempel vara att prediktera huspris med hjälp av sorlek, läge och antal rum. Medan statistisk inferens fokuserar på att dra slutsatser om en population baserat på ett urval data. Till exempel att man försöker dra slutsatser på vilka faktorer ut av storlek, läge och antal rum som påverkar huspriset.
3. Skillanden är att konfidensintervall är intervallet där en populationsparameter som medelvärde, varians eller standardavikelse kommer att befinna sig inom med en viss konfidensgrad som bestämms innan beräkningen. Prediktionsintervall är ett intervall som som prediktioner kommer att falla inom med en viss sannolikhet, givet vad vi vet från ett urval av data.
4. Beta parametrarna är lutningskoefficienterna för varje oberoende variabel, samt interceptet beta noll som ger värdet av Y då alla x är lika med noll. Beta koefficienterna visar hur mycket Y kommer att ändras då man ökar den oberoende variabeln med en enhet.
5. Det stämmer till viss del då dessa mått är gjorda för att försöka mäta modellers prestanda som och ta hänsyn till modellöveranpassning och modellkomplexitet genom att straffa sådana modeller genom att ge dem till exempel ett högre BIC-värde. SÅ även om vi inte delar upp datan i olika set, så sker fortfarande en viss modelvalidering med hjälp av dessa metriker. Dock ger inte dessa metriker hela bilden av modellens prestanda som om man validerar modellen med testdata, vilket ger en bättre bild av modellens prestanda.
6. Best subset selection fungerar genom att hitta den bästa modellen med alla varianter av prediktorer. Man testar först alla varianter av modeller med bara en variabel och väljer den modellen som har minst RSS eller störst R-squared. Sedan gör man samma sak med två variabler och fortsätter tills alla modeller med k variabler har testats. Sedan väljer man den bästa modellen ut av de modellerna med olika mängder variabler med hjälp av metriker som C(p) (AIC), BIC och Adjusted R-squared eller med hjälp av cross validation. Denna algoritm kommer leda till att man hittar den bästa modellen.
7. ’All models are wrong’ syftar på att ingen modell fånga alla aspeker av verkligheten vilket då gör dem felaktiga på det sättet. Medan ’some are useful’ syftar på att de fortfarande kan ge en värdefull insikt på genom att förstå samband mellan variabler, förutspå framtida händelrser och ge informerade beslut.

## Del (1) Data insamling

1. Vem du har arbetat i grupp med? Umut, Kamila, Kicki, Matthew, Quan och Ali

2. Hur har ni i gruppen arbetat tillsammans? Vi har fördelat arbetet lika med samma uppgift fast med olika data att samla in.

3. Vad var bra i grupparbetet och vad kan utvecklas? Att vi fick fram all data som vi behövde och kom fram till en något som vi alla tyckte var intressant att eftersöka. Något som kan utvecklas är att vara mer specifika i uppgifterna som skall göra så allt blir rätt.

4. Vad är dina styrkor och utvecklingsmöjligheter när du arbetar i grupp? Att prata mer när jag är säker om något styrkor kan jag inte komma på.

5. Finns det något du hade gjort annorlunda? Vad i sådana fall? Att vara mer specifik vi insamling av viss data.

# Självutvärdering

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

Den stora utmaningen har varit att jag inte har varit särskillt säker på det jag gjort även om det egentligen är helt rätt metod och inget fel i det jag gör.

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

Ett VG är aldrig fel men G funkar ocskå.

1. Något du vill lyfta fram till Antonio?

Nej

# Appendix A

data <- read.csv("C:/Users/leona/EC-Data science/R/Kunskapskontroll/ec\_r-main/ec\_r-main/Cars\_data\_feed\_2.csv", header = TRUE, sep = ",")

feed <- data[,1:10]

# Print the feed

head(feed, 5)

tail(feed, 5)

# Get the number of empty cells in each column and print the result

empty\_cells <- sapply(feed, function(x) sum(is.na(x)))

print(empty\_cells)

# Get the type of each column and print the result

column\_types <- sapply(feed, class)

print(column\_types)

# For each of the columns 3, 6 - 9 of feed, remove the leading, middle, and trailing whitespaces

feed[,c(3,6:9)] <- sapply(feed[,c(3,6:9)], function(x) gsub("^\\s+|\\s+$", "", x))

# Convert all characters in the columns 3, 6 - 9 of feed to small letters

feed[,c(3,6:9)] <- sapply(feed[,c(3,6:9)], function(x) tolower(x))

# Remove the square brackets from the cells of the column 6 of feed

feed[,6] <- gsub("\\[|\\]", "", feed[,6])

# Convert the fourth column of feed to character

feed[,4] <- as.character(feed[,4])

# Remove all spaces in the cells of the last column of feed and convert them all to integers

feed[,10] <- as.integer(gsub(" ", "", feed[,10]))

# Print out each unique value and its frequency in the first column of feed

print(table(feed[,3]))

feed[,6][2]

x <- 0

### Binning av färger

for (färg in feed[,6]) {

x <- x + 1

feed[,6][x] <- switch (färg,

'ljusgrå' = 'grå',

'mörkblå' = 'blå',

'mörkgrå' = 'grå',

'ljusgrön' = 'grön',

'mörkgrön' = 'grön',

färg)

}

print(table(feed[,6]))

print(feed)

# Get the number of unique values in each column and print the result, each one in a new line

unique\_values <- sapply(feed, function(x) length(unique(x)))

print(unique\_values)

feed.data <- data.frame(feed)

print(feed.data)

# Set the seed for reproducibility

set.seed(123)

# Calculate the number of rows to select for the test set

n <- nrow(feed.data)

k <- round(0.20 \* n)

# Randomly select 'k' rows for the test set

test\_indices <- sample(1:n, size = k)

test\_set <- feed.data[test\_indices, ]

# The remaining data will be used for the training set

train\_set <- feed.data[-test\_indices, ]

# Print the train and test sets

print(table(train\_set[,8]))

print(test\_set)

# Boxplot för att se outliers, kolla normalitet, linjäritet, correlationer.

install.packages("data.table")

library(data.table)

plot(feed)

pairs(feed[,c(1,2,5,10)])

plot(feed$Modellår, feed$Hästkrafter, main="Hästkrafter över Modellår", xlab="Modellår", ylab="Hästkrafter")

plot(feed$Modellår, feed$Miltal, main="Miltal över Modellår", xlab="Modellår", ylab="Miltal")

#Modellår verkar ha en Kolliniaritet med mital och hästkrafter

## Eftersom vi har alldelles för många features så har jag valt att ta bort

## Dessa för att inte det ska påverka modellen på ett dåligt sätt

rownames(train\_set) <- NULL

train\_set <- train\_set[c(-271, -291, -9), ]

print(train\_set[207,])

library('olsrr')

library('QuantPsyc')

fit <- lm(Pris..Y. ~ . - Län - Modell - Modellår, data=train\_set)

fit\_transformed <- lm(sqrt(Pris..Y.) ~ . - Län - Modell - Modellår, data=train\_set)

fit2 <- lm(Pris..Y. ~. - Län - Modell - Modellår, data=train\_set)

fit <- lm(sqrt(Pris..Y.) ~ Miltal + Hästkrafter + Märke + Biltyp, data=train\_set)

plot(fit)

plot(fit\_transformed)

ols\_plot\_resid\_hist(fit)

ols\_test\_normality(fit)

summary(fit)

print(fit$coefficients)

par(mfrow=c(2,2))

plot(fit)

vif(fit)

cooks.distance(fit)

# Print the ordered leverage values

print(ordered\_hats)

plot(cooks.distance(fit))

#Checking dimensionality p<<n

coef <- fit$coefficients

n\_coef <- length(coef)

n\_obs <- nrow(train\_set)

print(n\_obs/n\_coef)

# Best subset selection

Bestfit\_p <- ols\_step\_best\_subset(fit\_transformed, penter=0.05)

Bestfit\_p

plot(Bestfit\_p)

#

par(mfrow=c(1,1))

plot(regfit.full, scale='Cp')

coef(regfit.full, which.min(reg.summary$cp))

library(caret)

vif\_values <- vif(fit)

# Print VIF values

print(vif\_values)

model <- train(sqrt(Pris..Y.) ~ Miltal + Hästkrafter + Märke + Biltyp,

method='lm',

data=train\_set,

trControl = trainControl(

method = 'cv',

number = 10,

verboseIter = TRUE

)

)

print(model)

predictions <- predict(fit, test\_set)

print(predictions)

R\_sq <- R2(predictions, sqrt(test\_set$Pris..Y.))

RMSE <- RMSE(predictions, sqrt(test\_set$Pris..Y.))

MAE <- MAE(predictions, sqrt(test\_set$Pris..Y.))

R\_sq <- R2(predictions, test\_set$Pris..Y.)

RMSE <- RMSE(predictions, test\_set$Pris..Y.)

MAE <- MAE(predictions, test\_set$Pris..Y.)

print(c(R\_sq, RMSE, MAE))

mean((test\_set$Pris..Y.))

print(sqrt(test\_set$Pris..Y.))

pred\_error\_rate <- RMSE / mean((test\_set$Pris..Y.))

pred\_error\_rate <- RMSE / mean(sqrt(test\_set$Pris..Y.))

print(pred\_error\_rate)

#### SCB

library(pxweb)

data <- pxweb\_interactive("https://api.scb.se/OV0104/v1/doris/sv/ssd/START/TK/TK1001/TK1001A/PersBilarA")

px\_data\_frame <- as.data.frame(data, column.name.type = "text", variable.value.type = "text")

px\_data\_frame

plot(px\_data\_frame)

# PXWEB query

pxweb\_query\_list <-

list("Region"=c("00"),

"Agarkategori"=c("060"),

"ContentsCode"=c("TK1001AB"),

"Tid"=c("2002","2003","2004","2005","2006","2007","2008","2009","2010","2011","2012","2013","2014","2015","2016","2017","2018","2019","2020","2021","2022","2023"))

# Download data

px\_data <-

pxweb\_get(url = "https://api.scb.se/OV0104/v1/doris/sv/ssd/START/TK/TK1001/TK1001A/PersBilarA",

query = pxweb\_query\_list)

# Convert to data.frame

px\_data\_frame <- as.data.frame(px\_data, column.name.type = "text", variable.value.type = "text")

# Get pxweb data comments

px\_data\_comments <- pxweb\_data\_comments(px\_data)

px\_data\_comments\_df <- as.data.frame(px\_data\_comments)

# Cite the data as

pxweb\_cite(px\_data)

px\_data\_frame

x <- px\_data\_frame$år

y <- px\_data\_frame$`Personbilar i trafik`

par(mfrow=c(1,1))

plot(x, y, type='l', main="Personbilar i trafik per År (i 1000)", xlab="År", ylab="Personbilar i Trafik")

# Källförteckning

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

Blocket. (2024). *Blocket Elbilar.* [Cars\_data\_feed\_2.csv]. Blocket. www.blocket.se

Statistiska Centralbyrån (2024). *Personbilar i trafik efter region, ägarkategori, tabellinnehåll och år.*

SCB. <https://api.scb.se/OV0104/v1/doris/sv/ssd/START/TK/TK1001/TK1001A/PersBilarA>