Prediktionsmodeller för Bilpris

Multipel linjär regression, Ridge regression och Lasso regression



Matilda Wilhelmsson

EC Utbildning

Kunskapskontroll R-programmering

202404

# Abstract

In this report we gather data on cars put up for sale on the Swedish website Blocket.se. We use this data to create three models (Linear, Ridge, Lasso) in R where we use the cars price as our dependent variable and use a variety of available independent variables to predict the car price. Our linear model had the smallest RMSE on the test data and is therefore chosen as the best model for predicting car price. Following variables were shown to be statistically significant for our model: Mileage, Model year, Horsepower, Fuel, Gear, Type of car and Brand of car.

**Skapas automatiskt i Word genom att gå till Referenser > Innehållsförteckning.**

Innehållsförteckning

[1 Inledning 1](#_Toc165208126)

[2 Teori 2](#_Toc165208127)

[2.1 Variabler som påverkar bilpris 2](#_Toc165208128)

[2.2 Exempel: Regressionsmodeller 2](#_Toc165208129)

[2.2.1 Multipel linjär regression 2](#_Toc165208130)

[2.2.2 Ridge regression 2](#_Toc165208131)

[2.2.3 Lasso regression 3](#_Toc165208132)

[2.3 Utvärdera och anpassa modeller 3](#_Toc165208133)

[2.3.1 k-Fold Cross-Validation 3](#_Toc165208134)

[2.3.2 Adjusted R-squared 3](#_Toc165208135)

[2.3.3 RMSE 4](#_Toc165208136)

[2.3.4 Outliers: leverage och Cook’s distance 4](#_Toc165208137)

[3 Metod 5](#_Toc165208138)

[3.1 Datainsamling 5](#_Toc165208139)

[3.1.1 Sållning bland datan på blocket 5](#_Toc165208140)

[3.2 Installation av nödvändiga paket 6](#_Toc165208141)

[3.3 Pipeline 6](#_Toc165208142)

[3.4 Träning av modeller 7](#_Toc165208143)

[3.4.1 Dela upp datan i träning och test 7](#_Toc165208144)

[3.4.2 Multipel linjär modell lm\_1 8](#_Toc165208145)

[3.4.3 Ridge-modell lm\_2 8](#_Toc165208146)

[3.4.4 Lasso\_modell lm\_3 8](#_Toc165208147)

[3.5 Jämföra 3 olika modeller 8](#_Toc165208148)

[3.6 Undersöka outliers 9](#_Toc165208149)

[4 Resultat 10](#_Toc165208150)

[5 Slutsatser 12](#_Toc165208151)

[6 Diskussion 13](#_Toc165208152)

[7 Teoretiska frågor 14](#_Toc165208153)

[Appendix A 16](#_Toc165208154)

[Källförteckning 19](#_Toc165208155)

# Inledning

Maskininlärning är ett område som ständigt erövrar ny mark. Nuförtiden kan maskininlärning används inom en mängd olika områden såsom bildigenkänning och självkörande bilar (IT-högskolan 2023). Maskininlärning används även för att prediktera priser på olika varor (Medium 2023). Prisprediktion kan hjälpa till att stabilisera ekonomin men även se till att privatpersoner betalar rätt pris för rätt vara.

För att underlätta för privatpersoner som tänker köpa eller sälja en bil så skulle maskininlärning kunna appliceras för att beräkna bilpris. Om man skapar en modell som tar in ett antal variabler som vi tror påverkar en bils pris och samlar ihop tillräckligt med data för att träna modellen så skulle vi kunna använda denna modell för att beräkna vad priset för en bil borde vara. Detta skulle automatisera en annars godtycklig och arbetskrävande manuell process.

Under en tjugoårsperiod ökade antalet bilar i trafik i Sverige från 4,042,790 stycken år 2002 till 4,980,543 år 2022 (SCB). Ökningen har varit konstant under perioden bortsett under 2021–2022 då bilar i trafik sjönk något. Det är tydligt att bilen spelar stor roll för den svenska befolkningen och den svenska ekonomin.

En bild som visar text, diagram, linje, Graf

Automatiskt genererad beskrivning

Figur 1: SCB data illustrerad i R.

Syftet med denna rapport är att undersöka vilka variabler som påverkar priset på en bil. Urvalet av variabler kommer att baseras på den samlade teoretiska kunskapen kring vilka variabler som påverkar bilpriser. Datan som används i rapporten är tagen från en känd svensk säljsida blocket.se (Blocket). Detta innebär att urvalet av variabler även kommer begränsas av de variabler som finns tillgängliga i annonserna på Blocket. För att uppfylla syftet så kommer följande frågeställning(ar) att besvaras:

1. Vilken modell har det minsta RMSE på testdatan?
2. Vilka av de följande faktorerna har en statistisk säkerställd påverkan på en bils pris: Märke, Bränsle, Växellåda, Miltal, Modellår, Biltyp, Drivning, Hästkrafter, samt Region?

# Teori

## Variabler som påverkar bilpris

Det finns en rad olika faktorer som påverkar en bils pris. Nedan kommer ett utvalt antal listas. Dessa faktorer kommer sedan inkluderas i olika regressionsmodeller och testas genom statistisk inferens för att avgöra om vi rent statistiskt kan se att de har någon påverkan på bilpriser. Faktorerna har valts ut då de är vanligt förekommande i diskussionen kring bilpriser samt fanns att insamla som data från Blocket.

* Märke
* Bränsle
* Växellåda
* Miltal
* Ålder
* Biltyp
* Drivning
* Hästkrafter (HK)
* Region (som bilen finns i)

## Exempel: Regressionsmodeller

I följande avsnitt kommer tre olika regressionsmodeller presenteras. Alla tre modeller kommer appliceras på träningsdatan. Sedan kommer modellerna utvärderas och den bästa kommer användas som den slutgiltiga förklaringsmodellen.

### Multipel linjär regression

Multipel linjär regression är utveckling av enkel linjär regression där fler prediktorer ingår i modellen. Varje prediktor har en egen koefficient som avgör hur prediktorn påverkar linjens lutning i modellen (G. James et. Al., s. 72–73). Precis som i enkel linjär regression uppskattas koefficienterna genom att använda sig av *Least squares metoden* som minimerar residualerna mellan modellens predikterade värden och de faktiska datapunkterna.

### Ridge regression

Till skillnad från multipel linjär regression är Ridge regression är en form av regression där prediktorernas koefficienter inte skattas med hjälp av *Least squares metoden* (G. James et. Al., s. 238). Koefficienternas skattas istället genom en algoritm som straffar små koefficienter och minskar de mot 0. På detta sätt så minimeras inflytandet av mindre viktigare koefficienter på modellen.

### Lasso regression

Lasso regression är den mer extrema varianten av Ridge regression. Istället för att minska mindre viktiga koefficienter så sätter Lasso regressionen de till at bli 0 och därför uteslutas ur modellen (G. James et. Al., s. 242. Detta innebär att till skillnad från Ridge så utför Lasso regression ett urval bland parametrarna och behåller endast de som anses viktiga nog. Det positiva med detta är att modellen oftast blir mindre och därför mer intuitiv än en Ridge-modell som fortfarande innehåller alla parametrar trots att de har en liten påverkan.

## Utvärdera och anpassa modeller

För att kunna välja den bästa modellen behöver vi utvärdera modellerna. Nedan följer några viktiga metoder och mått som kan användas för detta ändamål.

### k-Fold Cross-Validation

Istället för att använda ett valideringsdataset så kan vi använda oss av cross-validation för att utvärdera MSE (mean squared error) hos vår linjära modell (G. James et. Al., s. 203). En variant av cross-validation kallas k-fold. Metoden innebär att man delar in träningsdata i k-många folds. En fold åt gången få agera valideringsdata medan de andra agerar träningsdata. Detta upprepas tills alla folds har fått agera valideringsdata en gång. För varje fold får vi fram MSE. Modellens slutliga MSE är medelvärdet av k-MSE och är en uppskattning av testfelets MSE.

När det kommer till cross-validation för Ridge- och Lasso-modeller så ligger fokus på att hitta rätt ”tuning-parameter” eller lambda (G. James et. Al., s. 250). Lambda är det värde som båda modellerna använder för att reglera koefficienterna i modellerna. För varje fold testar vi olika värde på lambda. Vi kollar sedan vilken lamda som ledde till det minsta värdet på MSE och använder detta lambda i vår slutliga modell.

### Adjusted R-squared

För att förstå Adjusted R-squared behöver vi först förstå måttet R-squared. R-squared är ett mått på modellens förmåga att förklara variansen hos datan (G. James et. Al., s.70). Måttet kan anta värden mellan 0-1, t. ex. innebär att R-squared på 0.7 att modellen förklarar 70% av variansen hos datan.

Adjusted R-squared är som namnet antyder ett justerat mått på modellens förklaringskraft. Adjusted R-squared tar hänsyn till hur många variabler modellen har och hur mycket de enskilda variablerna bidrar till förklaringskraften (G. James et. Al., s.234-235). Detta innebär att istället för att alltid föredra fler variabler så inkluderar inte Adjusted R-squared fler variabler i modellen om dessa inte bidrar tillräckligt mycket till modellen. Detta är speciellt användbart i modeller med många variabler då dessa kan simplifieras.

### RMSE

Ett annat viktigt mått som vi kan använda för att utvärdera vår modells prediktioner överensstämmer med den observerade datan är RMSE. RMSE bygger på MSE (mean squared error) som är ett mått på standardfelet hos våra prediktioner (G. James et. Al., s. 29).

RMSE är helt enkelt roten ur MSE och gör att vi lättare kan tolka resultatet då vi får det i originalskalan genom att ta roten ur.

### Outliers: leverage och Cook’s distance

När vi har skapat en modell är det även viktigt att vi kollar på dess outliers. Outliers är punkter som står ut från övriga data genom att antagligen ha ett större eller mindre värde än de flesta andra värdena. Speciellt problematiska är outliers som har en hög leverage och ett högt Cook’s distance.

En outlier har hög leverage om den har ett ovanligt värde på x-variabeln (G. James et. Al., s. 98–99). Dessa punkter har en större påverkan på modellen överlag. För att testa om outlierns leverage och residual leder till en stor påverkan på modellen använder vi ett mått som kallas Cook’s distance(Christian Thieme 2021). Om en outlier har ett Cook’s-värde på över 3 gånger medelvärdet för Cook’s så har outliern en stor påverkan på modellen och man bör överväga om man ska ta bort den datapunkten.

# Metod

I kommande avsnitt redogörs för de övergripande metodvalen som gjorts i skapandet av denna rapport. För hela koden se Appendix.

## Datainsamling

Datan som använts i denna rapport har samlats in från den svenska köp- och säljsidan blocket.se (Blocket). Blocket lämpar sig väl för datainsamling av denna typ bl. a. för det stora antal annonser som finns, inte minst för bilar. Annonserna är även skapade på ett sätt där säljaren uppmanas fylla i viss generell information kring fordonet. Detta gör att de flesta annonser har liknande information och därför kan jämföras med varandra. Problemet kvarstår dock att alla säljare inte fyllt i all grundläggande information om bilen vilket skapar problem då dessa värden kommer vara saknade värden i vår databas. Hur detta problem hanteras återkommer vi till i ett senare avsnitt.

Totalt har data från 10 083 bilannonser samlats in. Dubbletter har undvikits genom att jämföra det unika ID som varje annons får vi sitt skapande. Totalt består vårt dataset av 15 kolumner där 1 kolumn består av annonsernas ID. Denna kolumn räknas i denna rapport inte med som en variabel vilket innebär att vi efter insamling har 14 olika variabelvärden.

Följande variabler och dess värden har hämtats från blocket: Märke, Modell, Bränsle, Växellåda, Miltal, Modellår, Biltyp, Drivning, Hästkrafter, Färg, Motorstorlek, Datum i trafik, Region samt Pris. Som vår beroende variabel (y) i vår modell kommer vi att ha Pris eftersom prediktion av en bils pris är det vi är intresserade av. Övriga 13 variabler kommer få utgöra våra oberoende variabler (x) i modellen då vi tror att dessa påverkar bilpriset.

### Sållning bland datan på blocket

Vi har valt att begränsa vilken data vi tagit in från Blocket av två huvudsakliga anledningar. Den första är att vissa data skulle innebära en stor risk för outliers och andra statistiska problem. Den andra anledningen är helt enkelt att omfattningen av datan hade blivit för stor för denna uppsats.

Först och främst har vi valt att endast ta med bilar som har modellår 2000 eller senare. Detta för att undvika att vi får med samlar-bilar som kan ha ett högt värde trots att de har en hög ålder och höga miltal. Vi har även valt att endast ha med bilar som säljs av privatpersoner då bilar som säljs av företag skulle kunna vara nyproducerade och denna uppsats fokuserar på bilar som säljs i andrahand. Vi har valt att inte inkludera yrkesfordon då dessa kan ha ovanliga komponenter som gör de till speciellt värdefulla och detta skulle göra att de står ut från privatbilar. Till sist har vi valt att begränsa prisklassen till bilar som säljs för 20 000 till 500 000 kronor. Detta för att undvika skrotbilar som kan vara trasiga eller ha andra problem trots att de kanske inte har ett högt miltal etc. Vi begränsar även för de dyrare bilarna då de helt enkelt inte är så många. Dessa bilar skulle kunna ha ett högt leverage och på så sätt göra modellen osäker när det kommer till att prediktera nya bilar i denna prisklass.

## Installation av nödvändiga paket

Följande paket har installerats då deras metoder har använts i koden.

En bild som visar text, skärmbild, Teckensnitt, nummer

Automatiskt genererad beskrivning

Figur 2: Installera packet i R

## Pipeline

Vi börjar med processera datan i en så kallad pipeline.

En bild som visar text, Teckensnitt, skärmbild

Automatiskt genererad beskrivning

Figur 3: Pipeline

* Första steget i pipelinen är att ladda in datan i R-studio.
* Vi filtrerar sedan på pris då det verkar som att några annonser som föll utanför våra ramar ändå kom med.
* Sedan använder vi först **mutate** för att ta bort white-space bland värden som uppstått under insamlingen av datan.
* Vi använder **mutate** igen för att ersätta de celler som har ”NA” i text-format och ersätter genom **if.na** det med vanliga NA-värdet.
* Vi använder mutate för att ta bort mellanrum mellan siffror i kolmunen Miltal.
* Vi gör om följande variabler till factors: Modell, Bränsle, Växellåda, Biltyp, Drivning, Färg och Region. Detta eftersom de är kategoriska variabler.
* Vi gör om följande variabler till integers: Modellår, Miltal, HK, Motorstorlek och Pris. Eftersom dessa variabler har numeriska värden.
* Vi gör sedan (rad 34) om variabeln datum.i.trafik till en datumvariabel. Sedan (rad 36) sållar vi bort annonser som har ett datum innan 1999 eftersom några annonser som faller utanför våra ramar ändå kom med i scrapern.
* Sedan (rad 35) så gör vi om variabeln Modellår till en ny variabel, Ålder.
* Vi väljer att ta bort vissa variabler som kom med i scrapern. Vi tar bort motorstorlek eftersom det inte finns något teoretiskt stöd för att detta skulle påverka bilpriset. Vi tar bort variabeln Modell eftersom det helt enkelt finns för många olika modeller i datan för att de ska gå att jämföra med varandra. Vi rensar Datum.i.trafik/Datum\_i\_trafik eftersom dessa variabler är väldig lik variabeln Ålder(Modellår). Vi tar även bort variabeln färg då vissa ovanliga färger gör modellen svår att testa.
* Till slut tar vi bort alla rader som har saknade värden. Detta påverkar inte mängden data avsevärt. Vi har efter detta kvar 9 449 stycken annonser vilket är tillräckligt för att vi ska kunna skapa fungerande modeller.

Utöver denna pipeline justerar jag även kategorierna i variabeln Märke. Märken med färre än 50 observationer i datasetet klumpas ihop i en ny kategori som döps till Other. Detta då de annars kan orsaka problem när vi i nästa steg delar upp datan i ett träningsset och testset.

## Träning av modeller

### Dela upp datan i träning och test

Innan vi börjar träna våra modeller så behöver vi dela upp datan i ett träningsset och ett testset. Träningssättet använder vi när vi tränar och använder cross-validation på modellerna. Testsetet använder vi senare.

En bild som visar text, skärmbild, Teckensnitt, nummer

Automatiskt genererad beskrivning

Figur 4: Dela upp datan i träningsset och testset

Vi tar bort kolumnerna Id och row\_nbr då dessa inte fyller någon funktion i den fortsatta analysen.

### Multipel linjär modell lm\_1

Vi börjar med att skapa vår multipla linjära modell. Vi sätter Pris som vår beroende variabel och testar den mot alla de oberoende variablerna i träningsdatasetet.

En bild som visar text, Teckensnitt, skärmbild, linje

Automatiskt genererad beskrivning

Figur 5: Multipel linjär modell

### Ridge-modell lm\_2

Vi tränar vår Ridge-modell med hjälp av paketet trainControl som gör det möjligt för oss att bestämma alpha-värde. Med alpha-värde satt till 0 så tränar vi en Ridge-modell. Vi kör även cross-validation med 5 folds för att vår modell ska få det bästa lambdavärdet.

En bild som visar text, Teckensnitt, skärmbild, linje

Automatiskt genererad beskrivning

Figur 6: Ridge-modell

### Lasso\_modell lm\_3

Vi tränar även vår Lasso-modell med hjälp av trainControl. Vi sätter nu alpha-värdet till 1 och tränar därför en Lasso-modell. Vi kör cross-validation med 5 folds för att få det bästa lambdavärdet.

En bild som visar text, Teckensnitt, skärmbild, linje

Automatiskt genererad beskrivning

Figur 7: Lasso-modell

## Jämföra 3 olika modeller

Nu jämför vi hur våra tre modeller presterar på testdatan. Vi kollar på RMSE för att bestämma vilken modell som är bäst på att prediktera testdatan.

En bild som visar text, skärmbild, Teckensnitt, nummer

Automatiskt genererad beskrivning

Figur 8: Testning av modeller

## Undersöka outliers

Efter att vi testat modellerna kollar vi närmre på den modell som presterat bäst. Vi börjar med att kolla på residualerna. Det är viktigt att residualerna är normalfördelade samt att ingen av de överstiger gränsvärdet för Cook’s distance. Vi kollar även på Adjusted R-squared för att få en uppfattning om hur mycket av variansen i datan som förklaras av modellen.

# Resultat

|  |  |
| --- | --- |
| **RMSE för olika modeller** | |
| lm\_1 | 46 093 .41 |
| lm\_2 | 46 789 .82 |
| lm\_3 | 46 106 .11 |

Tabell 1: Root Mean Squared Error (RMSE) för de fyra valda modellerna.

När vi applicerar modellerna på testdatan får vi fram ovanstående RMSE för respektive modell. Två observationer kan göras. Vår multipla linjära modell lm\_1 presterar bäst och har lägst RMSE. Skillnaden mellan modellerna är liten, de verkar prestera ungefär lika bra.

Nu kollar vi närmre på vår bästa modell, lm\_1. Vi kollar först på residualerna.

En bild som visar text, diagram, skärmbild, skiss

Automatiskt genererad beskrivning

Figur 9: Analys av residualer i lm\_1

Som vi kan se på grafen längst upp till vänster verkar det inte finnas någon heteroskedasicitet hos residualerna. De verkar variera jämnt över olika y-värden. Längst upp till höger ser vi en QQ-plot som visar att våra residualer är approximativt normalfördelade. Detta är viktigt då de statistiska metoder vi använt har detta som ett krav. Längst ner till vänster ser vi samma plot som uppe till vänster fast med standardiserade residualer. Längst ner till höger ser vi hur hög leverage våra residualer har samt om någon av residualerna överstiger gränsvärdet för Cook’s distance. Leverage är lågt för samtliga och ingen residual överstiger Cook’s distance.

Sammantaget så kan modellen med detta godtas för vidare statistisk analys av koefficienter och Adjusted R-squared. Det kollar vi vidare på i bilden nedan.

En bild som visar text, skärmbild, Teckensnitt, dokument

Automatiskt genererad beskrivningEn bild som visar text, meny, dokument, skärmbild

Automatiskt genererad beskrivning

Figur 10: Tabell över koefficienter i lm\_1 samt Adjusted R-squared.

Eftersom vissa variabler såsom Märke och Region innehåller många kategorier får vi ut en gedigen lista över koefficienter för varje kategori. Det vi kan avläsa från de icke-kategoriska variablerna är att variablerna Miltal, Modellår och HK är signifikanta. Störst påverkan på bilpriset av dessa har Modellår där priset ökar med 12 200 kronor för varje år (t.ex. om en bil är från 2013 eller 2014).

Bland de kategoriska variablerna med färre kategorier så är (Bränsle) Diesel, Miljöbränsle/Hybrid signifikanta. Variabeln VäxellådaManuell (baseline är Automat) är signifikant och visar på en prisminskning på 11 420 kronor på grund av att växellådan är manuell istället för automat.

Bland Biltyper (där Cab är baseline-modell) är samtliga biltyper signifikanta. Samtliga biltyper har negativa koefficienter vilket innebär att alla är mindre värda jämfört med Cab-modellen.

Variabeln DrivningTvåhjulsdriven är signifikant med en negativ koefficient på 9 174 kronor. Den jämförs mot en fyrhjulsdriven bil.

Bland de många bilmärken är följande signifikanta: Citroen, Fiat, Ford, Hyundai, Kia, Land Rover, Mazda, Mercedes-Benz, Mitsubishi, Nissan, Opel, Peugeot, Renault, Saab, Seat, Toyota samt Volkswagen. Dessa märken har alltså en signifikant påverkan på bilpriset i vår modell.

Inom variabeln Region så är ingen kategori signifikant. Det verkar alltså inte finnas någon större skillnad i bilpris beroende på vilken region bilen finns i.

Vår Adjusted R-squared för modellen är 0,8043. Detta innebär att vår modell till 80 procent förklarar variansen i datan. Övrig varians är oförklarad.

# Slutsatser

Här nedan besvaras frågeställningarna som presenterades i början av rapporten.

1. Vilken modell har det minsta RMSE på testdatan?

* Den multipla linjära modellen, lm\_1, har lägst RMSE på testdatan.

1. Vilka av de följande faktorerna har en statistisk säkerställd påverkan på en bils pris: Märke, Bränsle, Växellåda, Miltal, Modellår, Biltyp, Drivning, Hästkrafter, samt Region?

* Våra numeriska variabler Miltal, Modellår och Hästkrafter är signifikanta. För variabeln Bränsle är kategorierna Diesel och Miljöbränsle/Hybrid signifikanta. DrivningTvåhjulsdriven är signifikant. VäxellådaManuell är signifikant. Samtliga kategorier i variabeln Biltyp är signifikanta. Följande kategorier inom variabeln Märke är signifikanta: Citroen, Fiat, Ford, Hyundai, Kia, Land Rover, Mazda, Mercedes-Benz, Mitsubishi, Nissan, Opel, Peugeot, Renault, Saab, Seat, Toyota samt Volkswagen.

# Diskussion

I denna rapport har vi byggt och testat modeller som predikterar bilpris. Trots att vi har inkluderat många olika märken och biltyper så har vi försökt säga något generellt om en bils pris. Detta kan vara problematiskt eftersom olika märken och typer av bilar har vitt skilda prisspann. Vi har försökt att kontrollera för denna variation genom att använda ett stort dataset. För att förbättra modelleringen ur denna synpunkt så skulle man kunna ha ett ännu större dataset så att varje bilmärke har tillräckligt många datapunkter för att vi ska kunna säga något om märkets påverkan i sig.

Antalet kategoriska variabler var stort i vår modell. En modell med många kategoriska variabler är svårare att träna och dra slutsatser ifrån. Som ovan nämnts kan många kategorier leda till att vissa kategorier endast har ett fåtal datapunkter. För att undvika detta problem skulle man i en modifierad modell kunna ta bort variabler såsom Märke och Region. Region hade inga signifikanta kategorier i vår modell så där borde påverkan på modellen inte vara stor om vi exkluderar variabeln Region. Det blir dock mer problematiskt att utesluta variabeln Märke då många av variabelns underkategorier var signifikanta för modellen. Vissa bilmärken är helt enkelt mer populära och kommer därför ha ett högre pris än en jämförbar bil av ett mindre populärt märke. Denna effekt skulle vi då inte kunna se om vi exkluderade variabeln märke.

Slutligen så är det viktigt att reflektera över den typ av data vi har valt att samla in. Datan från Blocket är det som säljarna själva lagt in. Det är inte orimligt att anta att vissa data är felaktig, så som miltal, hästkrafter etc. Genom att samla in en större mängd data så hoppas vi dock att endast en mindre del av datan är felaktig och inte påverkar modellen avsevärt. Viktigt att påpeka här är även att priset på bilen är priset som säljaren själv har satt som första pris. Vi får inte se vilket pris bilen till slut såldes för. Det kan tänkas att priset i många fall kanske blev något lägre än säljarens pris, vissa bilar kanske även fick flera bud, där priset blev högre än ursprungspriset. I en framtida modell hade det varit önskvärt att samla in data på just slutpriset för att få en bättre uppfattning om bilens värde i dagens marknad.

# 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 om vår data är normalfördelad. Den gör detta genom att plotta kvartilerna från stickprovet mot de teoretiska kvartilerna. Om resultatet blir en någorlunda rak linje är datan approximerat normalfördelad. Detta är viktigt då många statistiska metoder förutsätter att datan är någorlunda normalfördelad för att kunna ge användbara resultat.

**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?**

- Det stämmer att maskininlärning har fokus på prediktioner medan statistisk regressionsanalys har fokus på statistisk inferens. I maskininlärning vill vi träna en maskin att kunna göra prediktioner, t. ex. få maskinen att kunna känna igen ansikten på bilder. I regressionsanalys är vi intresserade av vilka oberoende variabler som på påverkar den beroende variabeln, t. ex. vilka faktorer (utbildning, inkomst, boende etc.) som påverkar nivån av kriminalitet i olika stadsdelar i storstäder.

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

- Konfidensintervall predikterar med vald säkerhet 90/95-procent etc. det sanna värdet av en parameter. Ett prediktionsintervall predikterar med vald säkerhet det sanna värdet för en ny observation som tillförs modellen. Ett prediktionsintervall är alltid bredare än ett konfidensintervall eftersom det även består av en felterm 𝜀.

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

**𝑌 = 𝛽0 + 𝛽1𝑥1 + 𝛽2𝑥2+ . . . + 𝛽𝑝𝑥𝑝 + 𝜀 .**

**Hur tolkas beta parametrarna?**

- Beta-parametrarna visar hur mycket dess tillhörande variabel X påverkar Y när X ökar med en enhet och alla andra variabler hålls konstanta. Detta kallas även för partialderivatan.

**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?**

* Det stämmer att man inte behöver använda olika set, träning, validering, test om man använder BIC. BIC estimerar testfelet direkt från träningsfelet hos träningsdatan.

**6. Förklara algoritmen nedan för ”Best subset selection”** En bild som visar text, skärmbild, Teckensnitt, nummer

Automatiskt genererad beskrivning

- Logiken bakom best subset selection är att prova alla möjliga kombinationer av prediktorer och sedan välja den modell som har det minsta prediktionsfelet eller största förklaringsvärdet. I det första steget börjar man med noll-modellen som helt saknar prediktorer. Sedan testar vi att lägga till en prediktor till modellen (alla prediktorer prövas) och väljer den modell som har det minsta felet. Vi fortsätter på samma sätt med 2 prediktorer (väljer den kombination av 2 prediktorer som ger minst fel) osv tills vi har en model som består av alla prediktorer. Nu har vi lika många modeller som prediktorer samt noll-modellen. Nu väljer vi den modell som har minst fel/högst förklaringsvärde.

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

**Förklara vad som menas med det citatet.**

* En modell är en förenkling av verkligheten. Vi kommer aldrig kunna skapa en modell som är helt korrekt i alla situationer. Detta betyder inte att alla modeller är oanvändbara. Skapar vi en modell som är tillräckligt bra så kan den vara användbar för att uppskatta hur ett verkligt samband mellan variabler ser ut.

# Appendix A

En bild som visar text, skärmbild, Teckensnitt

Automatiskt genererad beskrivningEn bild som visar text, skärmbild, Teckensnitt, nummer

Automatiskt genererad beskrivningEn bild som visar text, skärmbild, Teckensnitt, nummer

Automatiskt genererad beskrivningEn bild som visar text, skärmbild, Teckensnitt, nummer

Automatiskt genererad beskrivningEn bild som visar text, Teckensnitt, linje, skärmbild

Automatiskt genererad beskrivningEn bild som visar text, Teckensnitt, linje, skärmbild

Automatiskt genererad beskrivningEn bild som visar text, skärmbild, Teckensnitt, nummer

Automatiskt genererad beskrivning

# Källförteckning

G. James, W. Daniela, H. Trevor, T. Robert. *An Introduction to Statistical Learning: with Applications in R.* Andra upplagan. 2023. Webblänk: [islr\_2\_ed.pdf](file:///C:\Users\Matil\Downloads\islr_2_ed.pdf)

IT-högskolan. *Vad är maskininlärning?* 2021. Webblänk: [Vad är maskininlärning? | IT-Högskolan (iths.se)](https://www.iths.se/vad-ar-maskininlarning/)

Medium. *Machine Learning in Commodity Price Forecasting: Revolutionizing the Markets.* 2023. Webblänk: [Machine Learning in Commodity Price Forecasting: Revolutionizing the Markets | by Price Vision | 𝐀𝐈 𝐦𝐨𝐧𝐤𝐬.𝐢𝐨 | Medium](https://medium.com/aimonks/machine-learning-in-commodity-price-forecasting-revolutionizing-the-markets-8b2e5d103356)

Thieme Christian. *Towards Data Science: Identifying Outliers in Linear Regression — Cook’s Distance.* 2021. Webblänk: [Identifying Outliers in Linear Regression — Cook’s Distance | by Christian Thieme | Towards Data Science](https://towardsdatascience.com/identifying-outliers-in-linear-regression-cooks-distance-9e212e9136a)