Statistical learning

En modellering av bilpriser



Daniel Hemgren

EC Utbildning

R programmering för dataanalys

20240426

# Abstract

This report describes using blocket data of around 5000 car advertisements with info about the cars like mileage, engine horsepower, brand etc to build linear regression models mapping some or all these features on the price.

The major tasks performed in this report are two:

1. Two sets (one with 7 inputs and one with dimension reduction) are used to build multiple linear regression models used for inferential conclusions about car prices.
2. A simple model is built with based on ”trying to figure out exactly what moves car prices”

Innehållsförteckning

[Abstract 2](#_Toc165062815)

[1 Inledning 4](#_Toc165062816)

[1.1 Syfte 4](#_Toc165062817)

[1.1.1 Frågeställning 1 4](#_Toc165062818)

[1.1.2 Frågeställning2 4](#_Toc165062819)

[2 Teori 5](#_Toc165062820)

[2.1 Den Linjära regressionsmodellen 5](#_Toc165062821)

[2.1.1 Enkel Linjär regression 6](#_Toc165062822)

[2.1.2 Multipel Linjär regression 6](#_Toc165062823)

[3 Metod 7](#_Toc165062824)

[3.1 Data 7](#_Toc165062825)

[3.1.1 Datainsamling i grupp 7](#_Toc165062826)

[3.1.2 Data preprocessing av blocketdata 7](#_Toc165062827)

[3.1.3 SCB:s API pxweb() 8](#_Toc165062828)

[3.2 Modellkonstruktionsprocessen för Set1 och Set2 8](#_Toc165062829)

[8](#_Toc165062830)

[3.3 Resonerande modellering 11](#_Toc165062831)

[4 Resultat med diskussion 13](#_Toc165062832)

[4.1 Översikt av resultatet 13](#_Toc165062833)

[4.1.1 Resultat Frågeställning 1 13](#_Toc165062834)

[4.1.2 Resultat Frågeställning 2 14](#_Toc165062835)

[5 Slutsatser 15](#_Toc165062836)

[5.1 Frågeställningar 15](#_Toc165062837)

[5.1.1 Svar på Frågeställning 1 15](#_Toc165062838)

[5.1.2 Svar på Frågeställning 2 16](#_Toc165062839)

[6 Teoretiska frågor 17](#_Toc165062840)

[7 Självutvärdering 20](#_Toc165062841)

[8 Källförteckning 21](#_Toc165062842)

# Inledning

Statistisk inlärning handlar om att skapa förståelse för komplexa datasets. Den stora ökningen av tillgängliga data under de senaste decennierna har gjort att flera tekniker inom detta område används frekvent. Området har ett flertal tekniker som är ”supervised” med kända labels(t.ex bilpris) att mappa sina features mot. Det finns också ett flertal tekniker som är unsupervised som t.ex klustring. (Hastie, 2023)

## Syfte

Syftet med denna rapport är att göra en komplett regressionsmodellering av bilpriser på Blocket med huvudfokus på inferens och förståelse. De data som ska mappa bilpriserna är t.ex information i bilannonserna och data från SCB. Regressionsmodelleringen syftar till att skapa förståelse för hur bilpriser påverkas av olika faktorer. Ett val jag gör är att göra modeller på ”alla” bilar, då jag hellre skapar något ”sämre” modeller men som helst har en generaliserbarhet på ”alla” bilar även i andra sammanhang. Frågeställningarna är menade att vara skrivna på enkelt språk för att kunna föra ut resultatet till en allmän publik då regressionsanalys lämpar sig för detta och dess kraftfullhet ”i vanligt tal” kan försvinna i för stringent matematik. Min tanke är att Syfte-Slutsats ska kunna förstås av vem som helst då många har tankar om bilpriser andra delar av rapporten är inte tänkt att göra detta.

### Frågeställning 1

Vilka medeleffekter på bilpriset har de signifikanta oberoende variablerna i set1 ? resp set2?

### Frågeställning2

Skapa en regressionsmodell med utgångspunkt i ett resonemang om bilpriser och vad som borde påverka dem allra mest. Finns en ”gyllene indikator”? Dvs med endast en (eventuellt kombinerad) input?

# Teori

Den breda teorin bakom denna rapport är teorin om Statistical learning teorin handlar om att hitta funktionen (f) mellan olika sets av data. Det är framförallt verktyg för att förstå data (Hastie, 2023, p. 1)

Statistical learning kan vara supervised eller unsupervised. Supervised statistical learning går i grova drag ut på att skapa modeller för att uppskatta en output baserat på en eller flera inputs. Unsupervised statistical learning har inputs men inga definierade outputs ändå kan vi här lära strukturer och relationer i datat.

Den specifika teorin i denna rapport är teoribildningen runt linjär regressionsmodellering. Denna teori lämpar sig för syften där en hög ”explainability” är viktigt. Detaljerad förståelse för hur varje del av input påverkar output kan uppnås med dessa modeller, det kallas att skapa inferenser. (Hastie, 2023).

Regression kan också användas för att modellera diskreta outputs då kallas modeller logistic regression och använder ofta logistic function.

Huvudfokus i denna rapport ligger inom Linjär Regression som är en supervised metod för att skapa modeller mellan inputs och en kontinuerlig output. Detta kallas även att ”skapa en linjär modell mellan en output med kontinuerlig label och en eller flera input features”. Linjär regression kan verka ”för enkel” men är mycket kraftfull och har otroliga fördelar inom tolkning och förklaring. Linjär regression gör att vi undviker att modellen blir en ”black box” (https://en.wikipedia.org/wiki/Black\_box, n.d.)

Det är antagligen fullt möjligt att modellera bilpriser med tex ett deep neural network med features såsom: högupplösta bilder av mikroskador på lacken, digitala ljudupptagningar av frekvenser hos motorvibrationer (något t.ex Atlas Copco mäter för att identifiera fel i pumpar) etc. Modellen som blivit resultatet hade varit mycket exakt men hade varit mycket svår att förstå. Den typen av modell är inte målet i denna rapport därför används Linjär regression helst också i enkel form.

## Den Linjära regressionsmodellen

Linjär regression är en statistisk metod som används för att modellera det linjära sambandet mellan en eller flera oberoende variabler (features) och en kontinuerlig beroende variabel (target). Målet med linjär regression är att skapa en linjär ekvation som beskriver den förväntade genomsnittliga effekten av de oberoende variablerna på den beroende variabeln. Linjär regression bygger på flera förutsättningar, inklusive linjärt samband mellan variablerna, normalfördelning av residualerna och homoskedasticitet (konstant varians hos residualerna) och oberoende observationer. (Hastie, 2023).

Att beräkna den minsta kvadratsumman är det vanligaste sättet att passa in denna modell i datat (Hastie, 2023, p. 59).

### Enkel Linjär regression

I sin absoluta grundform så ser den ut som nedan, där Y är ”approximerately modeled” av de två beta termerna. Den första är interceptet och den andra blir lutningen på linjen. (Hastie, 2023, p. 61)

A black numbers and plus symbols

Description automatically generated with medium confidence

Det riktiga populationssambandet (som vi ej kan observera) ser ut enligt nedan:

A close-up of symbols

Description automatically generated

Vid flera olika samplingar kommer regressionerna att bli som de blåa nedan men medelvärdet av dem kommer ligga nära den icke observerbara populationsregressionen som är röd nedan.

A graph of a line

Description automatically generated with medium confidence

Vid 95% signifikans måste den fixa men okända Beta ligga inom 2 Standardfel från ”β hat” enligt intervallet för ”β hat” nedan: (I rapportens syftesdel är 99% signifikans vald vilket blir 3 standardfel)

A black text on a white background

Description automatically generated

### Multipel Linjär regression

Denna modell är en förlängning av den enkla regressionen där varje Beta får en egen lutning mot den beroende variabeln. De fixa men okända Beta koefficienterna estimeras på samma sätt som i enkel regression, där valt antal standardfel runt estimaten ger ett intervall för Beta på en viss signifikansnivå. +-2 Standardfel är på 95% signifikans och väljer man +-3 standardfel så är det t.ex på 99% signifikansnivå. ger I denna modell är det mycket viktigt att tolka Beta1 som medeleffekten på Y av en enhetsförändring av X1 när alla andra prediktorer är oförändrade.



Även här gäller att residualerna måste uppfylla kraven på att vara normalfördelade, homoskedastiska (konstant varians) och oberoende av varandra (ej autokorrelerade).

Multipel regression har en inneboende risk som innebär att med för många inputs (ibland fler än observationer) så kommer alltid signifikans att uppnås av några av ren slump (Hastie, 2023, p. 78).

Därför används t.ex F statistic som anpassar sig efter antal prediktorer och därmed inte lider av detta problem.

informationskriterier som Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) eller justerad R-kvadrat används för att bestraffa för modellkomplexitet och välja den bästa modellen. Eller genom att dela upp data i tränings- och valideringsuppsättningar flera gånger, passa modeller på träningsuppsättningarna och utvärdera deras prestanda på valideringsuppsättningarna. Välj modellen med bäst genomsnittlig prestanda över valideringsuppsättningarna. Detta kräver datorkraft och kallas resamplingtekniker. (Hastie, 2023, p. 197)

# Metod

## Data

Datat som används är ca 5000 bilannonser från blocket som har samlats in med webscraper. Datat har rensats från uppenbara fel och observationer med luckor. Datat i en bilannons innehåller infomation om bilen med t.ex årsmodell, drivmedel, miltal, märke osv.

### Datainsamling i grupp

Första delen av datainsamlingen innebar att samla in ca 30 bilannonser var manuellt. Det var en bra övning för att se att t.ex olika format och tecken kan skapa problem/möjligheter. Redan i ett tidigt skede är det viktigt att tänka på hur data kan kodas om kategoriskt och om eventuella standardiseringar kan bli aktuella.

1. Vem du har arbetat i grupp med?

Xaioyong, William, Siarhei, Melike, Khaldoun, Nathalie, Frida, Dan

2. Hur har ni i gruppen arbetat tillsammans?

Vi delade upp själva insamlingen i disjunkta sets vi valde geografiska områden för detta syfte. För att sedan samla ihop oss efter insamlingen och besluta om att sätta ihop detta till ett helt set att jobba med.

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

Bra var olika uppgifter var lätt att dela upp och själva datainsamlingen föreslogs att en person gör en testinsamling som om den slog väl ut kan appliceras av de andra. Det är bra att ha med sig: testa det beslutade i liten skala för att eventuellt backa beslutet eller skala upp det hela.

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

Mina styrkor är att jag ifrågasätter på ett tidigt skede och testar ”reality check” frågor till gruppen, på det vi ska göra. Detta är också svagheten att denna tidiga ”utredningsfas” kan bli för omfattande och caset blir ”för stort” för den uppgift som ska genomföras.

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

Allt gick smidigt och egentligen löste vi det mycket bra.

### Data preprocessing av blocketdata

I Excel görs en EDA för att få en känsla för vilka data som ingår: Växellådstyp,Mätarställning, Årsmodell, Drivmedel, Märke,Motorvolym, Hästkrafter, Pris,Län. EDA:n visar hög korrelation mellan motorvolym och hästkrafter, dimensionsreducering med borttagande av motorvolym görs. De kategoriska kolumner som behöver kodas är växellådstyp, märke och drivmedel. Här görs en dummy encoding där automat, eldrift och premium(Mercedes-Benz, BMW,Ferrari, Lamborghini el. Maserati) blir egna binära kategorier. Län kodas i tre grupper :Stad, övrtätort, Glesbygd. Priset begränsas till under sek 1.200.000 då det är tydligt att många outliers finns bland de extremt höga priserna. Jag landar i ett set på ca 5000 observationer med 8 kolumner som är Hästkrafter, Mil, Ålder, Automat, Premium, Eldrift, Län och Pris. Redan i detta skede skapas dels detta bruttoset och ett dimensionsreducerat set där multicollinearitet är minimerad. I det dimensionsreducerade setet görs en sammanslagning av Mil och Ålder(då de har en viss correlation) efter standardisering och en summering görs av de binära variablerna Automat, Premium, Eldrift till en summa. (Om en bil har Automat och är Premiummärke och har Eldrift så har den t.ex värde 3 här). Nedan ses bruttosetet och det dimensionsreducerade setet.

* Set1: hk, mil, ålder, automat, premium, el, location, pris
* Set2: hk, mil+ålder, auto+premium+el,pris

Seten har en storlek på ca 5000 observationer och jag gör en 60-40 train-test split i tidigt skede för att kunna pröva modeller med resampling tekniker i train och för att testa slutresultatet i resultatdelen med ett helt orört stort testset för att undvika ev. ”otur” i ett för litet testset som kan ge ett för bra/för dåligt resultat beroende på det specifika setet. Med ca 5000 observationer har jag råd med ett så stort test set. En sak jag medvetet undviker i detta läge är att behålla enkelheten i förståelsen av resultatet genom att t.ex inte logaritmera priserna i detta skede. Förhoppningen är att med många inputs kanske det går att modellera faktiska kronor vilket är önskvärt ur enkelhetssynpunkt.

### SCB:s API pxweb()

Pxweb är SCB:s API (pxweb: R tools for PXWEB API. URL: http://github.com/ropengov/pxweb) och det är ett mycket bra sätt att hämta data från SCB och andra myndigheter såsom de andra nordiska ländernas statistikmyndigheter. Jag hämtar en serie med antal bilar/1000 invånare för alla län i Sverige. Dessa län kodas på samma sätt som län i blocketdatat i tre kategorier: Stad, Glesbygd, Övrtätort. De tre kategorierna fylls med data från Pxweb med antal bilar/1000 invånare i denna geografiska kategori. Jag kallar kategorin ”bildensitet” och ersätter location i Set1 med detta värde.

* + Set1: hk, mil, ålder, automat, premium, el, bildensitet, pris (medelpris: SEK 251.000)
  + Set2: hk, mil+ålder, auto+premium+el, pris (medelpris: SEK 251.000)

## Modellkonstruktionsprocessen för Set1 och Set2

Bruttosetet (Set1) med 7 inputs ger vid en första training följande koefficienter:

# A screenshot of a computer code Description automatically generated

En första training summary med de 7 inputs i Set1.

Nollhypotesen:

 

Där nollhypotesen(ingen påverkan på output) kan förkastas med en mycket hög F-statistic. Att bilen har eldrift uppvisar en viss negativ medeleffekt på priset i detta tidiga skede kan ses som överraskande men den är inte signifikant vid 99% signifikans Bildensitet har ingen signifikant påverkan på priset vid 99% signifikans och även vid lägre signifikansgrader.

A collage of graphs and diagrams

Description automatically generated

Diagnostiska plots för residualerna efter en första training.

Antagandet om linearitet får anses hålla i Residuals vs Fitted då residualerna lägger sig runt linjen och saknar egen form. Q-Q och Scale-Location visar att antagndet om normalfördelade residualer med homoscedasticitet i huvudsak gäller.Cook’s distance (max 0,2) visar att ingen observation påverkar koefficienterna extremt mycket. En tidig preprocessing i Excel visade att många extrema ”outlier” problem löstes med att modellera endast bilpriser upp till sek 1.200.000 vilket visar sig stämma här.Modellen uppfyller i stora drag de krav som ställs på en multipel linjär regression. Några koefficienter är inte helt signifikanta vid 99% signifikans och därför testar jag att skapa subsets för att se om modellen kan förbättras med regsubsets metoden (https://www.rdocumentation.org/packages/leaps/versions/3.1/topics/regsubsets, n.d.) som är en kombination av forward och backward selection där det bästa subset blir enligt

* BIC: "hk”, "mil", "ålder", "automat", and "premium".
* adjusted R-squared : "hk", "mil", "ålder", "automat", "premium", and "el".

I valet mellan en större och en mindre modell så är ofta en mindre modell att föredra så därför går jag på första alternativet med 5 inputs. Alternativet ser ut enligt nedan:

A screenshot of a computer

Description automatically generated

Efter användning av regsubsets funktionen så återstår 5 inputs av 7 från Set1.

Hur anpassad är modellen till datat? Finns någon risk för overfitting? För att få en bild av bias variance tradeoff så görs resampling med 5 folds:

A screenshot of a computer

Description automatically generated

Resampling av 5 folds.

Tydligt är att outliers påverkar RMSE då modellen uppvisar RMSE i dessa 5 resamplingar mellan 69222-76862 vilket ska jämföras med medelprisnivån 251000. MAE är ej lika påverkad av outliers och har ett absolut medelfel på 45172-48771 för dessa 5 resamplingar.Variansen i pris förklaras mellan 69,3%-76,7% i dessa 5 resamplingar det visar att det är en modell som är robust över olika dataset.

Set2:

På samma sätt(förutom att skapa subsets) trainar jag det dimensionsreducerade Set2 och gör diagnostik av residualer. Här har alla beta koefficienter hög signifikans direkt.

A screenshot of a computer code

Description automatically generated

Första training av Set2.

Även Set2 resamplas på 5 folds för att få bild av känsligheten för nya datasets.

A screenshot of a computer

Description automatically generated

Resampling med 5 folds utförs på Set2.

Här ses att Set2 förklarar variansen i bilpriserna i samma utsträckning som Set1. Mean R2 för dessa folds är något lägre än för Set1 . RMSE kommer upp i hela 83324 i ett fold här och det visar att denna modell har en något större känslighet för nya data. Det går egentligen lite emot tänket att detta dimensionsreducerade Set2 borde vara less ”overfitted” och ha högre bias mindre variance. En första tanke är att dessa 2 Sets landar generellt sett väldigt lika i sin mapping av datat. Skillnaden är mindre än jag hade förväntat mig. Denna diskussion utvecklas i Resultatdelen.

## Resonerande modellering

För att skapa en kraftfull och enkel (enkelhet är styrkan med linjär regression tycker jag) modell tänkte jag tillbaka vad som verkligen påverkar bilpriser och vad jag tror om detta. Jag gick tillbaka till EDA och funderade. I ursprungssetet (Set1) med 7 inputs är det egentligen bara hästkrafter, miltal och ålder som uppvisar rimligt hög korrelation med priset med 0,73, -0,44 resp. -0,51. Mil och ålder korrelarar högt sinsemellan men även med en hopslagning(multiplikation) av dessa så får den resulterande variabeln lägre korrelation med pris än vad ålder har själv med -0,45. En kvot mellan dessa korrelerar mycke lågt med priset på 0,09.

Priser är ofta lognormala (de kan t.ex inte i normalfallet gå under 0) och därför är det naturligt att logaritmera priserna från blocket. Jag undvek detta med grundseten för att om möjligt hitta enklast möjliga modell, i nuläget väljer jag att testa logaritmering då modellen jag söker nu är en mindre modell som har en bättre förklarbarhet i och med detta.

Det logaritmerade priset korrelerar mot variablerna med samma mönster som det vanliga priset, hästkrafter, ålder och miltal högst och i den ordningen. Intressant är att hästkrafter nu korrelerar sämre med 0,62 mot tidigare 0,73. Tydligt är att hästkrafter och ålder korrelarar tydligast mot pris och åt olika håll. En variant blir att standardisera båda och ta standardiserade hästkrafter – standardiserad ålder som oberoende variabler i modellen.

En snabb kontroll av modellen visar att den är överraskande bra trots att det är en enkel regression. En kontroll av residualer och outliers visar att modellen blir ännu bättre om de äldsta bilarna som är äldre än 20år tas bort även bilar under 10000kr plockas bort. Totalt sett är det mindre än 50 observationer som försvinner, precis som innan så kapas bilar över 1.200.000kr också. Modellen har en korrelation med det logaritmerade priset på 0,89. Detta trots att alla typer av bilar och drivmedel, växellådstyper osv. Ingår. Med en viss förvåning tittar jag på grafen i Excel mellan log pris och differensen mellan hästkrafter och ålder med ca 5000 bilannonser från blocket.

A graph with blue dots

Description automatically generated

Denna plot i Excel med en oberoende(om än kombinerad av två) variabel förvånade mig med sitt tydliga linjära samband.

Jag tar in detta i R för att borra lite djupare, actual mot fitted values på logPris ser som väntat bra ut:

A graph with a line and dots

Description automatically generated with medium confidence

Plot mellan Actual och fitted log Pris

Ett train set på ca3000 av de 5000 bilannonserna visar att max och minresidualen i setet är hög och bestämmer bilpriset fel ca exp(-1,45) till exp(1,42) vilket är ”en fjärdedel upp till fyra gånger för stort” detta är väntat som extrem inom 3000 observationer. Totalt sett förklaras variansen i bilpriset mycket bra med R2 på 0,795.

Summary av den enkla regressionen


Den bästa training jag gjort för att förklara bilpriser med en input.

# Resultat med diskussion

Resultatdelen syftar till att titta på hur modellerna fungerar på orörda testset och fördjupa diskussionen runt varför detta resultat ses och vad det kan användas till. Prediktion görs i frågeställning2.

## Översikt av resultatet

De två olika modelleringsmetoderna nedan under de olika frågeställningarna visar på olika styrkor och svagheter. I Frågeställning 1 skapas modeller som lämpar sig för att göra inferenser av speciella bilattribut. I den andra frågeställningen så skapas modeller som mycket enkelt kan säga priset på en bil baserat på få inputs. Båda modelleringsmetoderna kan vara värdefulla i olika sammanhang.

### Resultat Frågeställning 1

Vilka medeleffekter på bilpriset har de signifikanta oberoende variablerna i set1 ? resp set2?

Resultatet på testsetet för Set1 ses nedan. Även med testsetet så har alla betakoefficienter hög signifikans. R2 och F-statistic har värden på samma nivåer som vid träning.

A screenshot of a computer code

Description automatically generated

Tabell 1: Set1

Denna modell kan teoretiskt ge ett negativt pris. En hårdkodning är gjord för detta då värdet blir ”skrotvärde” 1000kr i dessa fall. Med ca 142.000kr som utgångsläge så är en hästkraft värd ca 1100kr en körd mil gör att bilen tappar 7kr och ett år ger minskning med ca 7.500kr. Automatväxellåda är värd förvånande höga 25.000kr och premiummärke är värt ca 17.000kr allt annat lika.

Resultatet på testsetet för Set2 ses nedan. Även med testsetet så har alla betakoefficienter hög signifikans. Dimensionsreduceringen i detta set ger en modell med ungefär samma R2 och F-statistic som ovan men med en sämre möjlighet till inferenser av individuella bilattribut vilket lämpar sig bättre med Set1.

A screenshot of a computer screen

Description automatically generated

Tabell 2: Set2

### Resultat Frågeställning 2

Skapa en regressionsmodell med utgångspunkt i ett resonemang om bilpriser och vad som borde påverka dem allra mest. Finns en ”gyllene indikator”? Dvs med endast en (eventuellt kombinerad) input?

Resultatet blev att Hästkrafter-ålder efter att båda blivit standardiserade är en bra enkel regression. Även på ett helt nytt set med ca 2000 bilannonser så blir summary enligt nedan:

A close-up of a computer screen

Description automatically generated

A diagram with a line and dots

Description automatically generated with medium confidence

Detta får anses som den ”gyllene indikatorn” som förklarar bilpriser med en input och fungerar även på detta testset med en RMSE på testsetet på 76450kr.

Denna modell kan tolkas som att om både hästkrafter och ålder är på sina medelvärden (186 hk och 6år här) så är medelpriset på en bil exp(12,29) vilket är 217.500kr. För varje standardavvikelse(85hk) som hästkrafter flyttar uppåt eller ålder flyttar nedåt (3,4år) så ökar priset med exp(0,31) vilket är ökning med 36%. T.ex bil med 271hk (186+85) som är 6år gammal kan medelpredikteras till 217.500\*1,36 vilket blir 295800kr. Väldigt grovt och oberoende av märke etc. Men också väldigt enkelt och kanske ”useful” om vi tänker på Box: ”all models are wrong but some are useful”

Efter mycket undersökande av dataseten så hittade jag nedanstående modell i Excel som inte är den bästa men den absolut enklaste för att förklara bilpriser med en variabel.

Medelpriset på en bil blir 58261kr + 2479kr\*(Hästkrafter/(roten ur (ålder+1))). Modellen har en R2 på 0,74 fullt jämförbart med de avancerade modellerna testade tidigare i förmågan att förklara variansen i bilpriser. Men med styrka i att vara lätttolkad och inte för att göra inferenser om värdet av t.ex automatväxellåda el färger eller andra attribut på bilen. Är detta ett exempel på en ”gyllene indikator”? Den är mycket enkel men jag tycker att den tidigare med standardiserade värden är mer konsekvent över alla bilar och har bättre värden helt enkelt även om den är något mer svårtolkad.

A graph with blue dots

Description automatically generated

Den enklaste modellen: Pris = 58261kr +2479kr\*(hk/√(ålder+1))

# Slutsatser

Det har varit ett intressant och roligt arbete där modeller tränats på blocketdata och där jag varit nyfiken på vad som verkligen påverkar bilpriser.

## Frågeställningar

De tre frågeställningarna i Syftesdelen kommer här gås igenom för att se om frågeställningarna kunde besvaras.

### Svar på Frågeställning 1

Vilka medeleffekter på bilpriset har de signifikanta oberoende variablerna i set1 ? resp set2?

Set1: Medeleffekter på bilpris

Med utgångspunkt i 142.100kr så ger varje hästkraft +1093kr varje mil ger -7kr varje år ger -7493kr automatväxellåda ger +25090kr premiummärke ger +16790kr

A white background with black text

Description automatically generated

Set2: Medeleffekter på bilpris

Med utgångspunkt i 34.000kr så ger varje hästkraft +1117kr varje standardavikelse det standardiserade mil+ålder ger 77.000kr och varje del av automatväxellåda, premiummärke och eldrift ger +10574kr i Pris.

A white background with black text

Description automatically generated

### Svar på Frågeställning 2

Skapa en regressionsmodell med utgångspunkt i ett resonemang om bilpriser och vad som borde påverka dem allra mest. Finns en ”gyllene indikator”? Dvs med endast en (eventuellt kombinerad) input?

A diagram with a line and dots

Description automatically generated with medium confidence

Detta får anses som den ”gyllene indikatorn” i en variabel med RMSE på 76450kr på testsetet.

En möjlig konkurrent som vinner i enkelhet är Excel modellen nedan: Den har en bra fit men är oprövad som prediktor.

A graph with blue dots

Description automatically generated

Den enklaste modellen: Pris = 58261kr +2479kr\*(hk/√(ålder+1))

# 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.

Ett verktyg för att se om ett dataset är fördelat som en specifik fördelning ofta normalfördelningen. Quantilerna i fördelningen(Quantile=delar med samma sannolikhet) kommer ha ett linjärt samband med datasetet om så är fallet.

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?

Maskininlärning kan vara svårare att förklara och därför är det vanligare med prediktioner som kommer lite ur en ”black box”. Det är svårt att säga vad en viss pixel är värd i MNIST t.ex. Helhetsprediktionen av vilken siffra det är blir fokus i detta exempel. Regressionsanalys har ofta riktiga värden som input och därför kan man ofta värdera en viss input specifikt och t.ex testa nollhypotesen att den specifika inputen inte påverkar output. Om nollhypotesen har en väldigt liten sannolikhet så kan vi t.ex säga i fallet Bilpriser att ett år nyare bil t.ex påverkar priset med +10000kr, två år nyare bil +20000kr osv. vid linjär modellering och är ett exempel på inferens.

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

Konfidensintervall handlar om osäkerheten i t.ex regressionskoefficienter och prediktionsintervallet tar hänsyn till detta och variationen i observationerna. Konfidensintervallet kan ses som en delmängd av prediktionsintervallet. Konfidensintervallet handlar om att säga något om en population och prediktionsintervallet handlar om det och den naturliga variationen i individuella observationer.

4. Den multipla linjära regressionsmodellen kan skrivas som:

� = 𝛽0 + 𝛽1𝑥1 + 𝛽1𝑥2+ ...+ 𝛽𝑝𝑥𝑝 +𝜀 .

Hur tolkas beta parametrarna?

Den första tolkas som ett intercept och är värdet av den beroende variabeln när alla de oberoende variablerna är 0. De andra beta parametrarna kan tolkas som förändringstakt (lutning) för den beroende variabeln vid en enhetsförändring i en oberoende variabel när de andra oberoende variablerna är konstanta. Om t.ex Beta1 är 3,2 så ändras den beroende variabeln 3,2 enheter vid en enhetsförändring av x1. Den exakta tolkningen är alltid beroende av variablernas skala.

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?

Träningsset (Training set): Detta är datamängden som används för att träna modellen. Modellen anpassas till detta dataset genom att beräkna parametrar som minimerar felet(vanligen minimering av kvadrerade fel) mellan de observerade och förutsagda värdena.

Valideringsset (Validation set): Efter att modellen har tränats, används valideringssetet för att utvärdera modellens prestanda och välja de bästa hyperparametrarna (t.ex. antal variabler, grad av polynom) eller andra modellspecifika inställningar. Valideringssetet används för att göra justeringar i modellen för att förbättra dess prestanda. (When using resampling techniques such as cross-validation, having a dedicated validation set becomes less critical, as the data is repeatedly split into training and validation subsets during the resampling process. However, even in these cases, it is still beneficial to reserve a separate test set for final model evaluation to provide an unbiased estimate of its generalization performance.) (Hastie, 2023)

Testset (Test set): Slutligen används testsetet för att bedöma modellens prestanda och generalisering till nya, oberoende data. Testsetet bör vara oberoende av både tränings- och valideringsdatamängderna och används för att ge en objektiv bedömning av hur väl modellen presterar på nya data.

6. Förklara algoritmen nedan för ”Best subset selection”

A screenshot of a test

Description automatically generated

1.Skapa M0 utan predictors som bara predikterar medelvärdet av varje observation och fungerar som ”baseline”

2. (a)För varje värde av 𝑘, passa alla möjliga kombinationer av modeller som innehåller exakt 𝑘 prediktorer av de totala 𝑝 tillgängliga prediktorerna.

(b) Bland modellerna med exakt 𝑘 prediktorer, välj den med minsta RSS eller motsvarande den störstaR2 . Modellen med lägst RSS indikerar det bästa avvägandet mellan passformens kvalitet och modellens komplexitet.

3.Efter att ha erhållit de bästa modellerna för varje värde av 𝑘, välj en enda bästa modell från 𝑀0,𝑀1,...Mp genom:

* informationskriterier som Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) eller justerad R-kvadrat för att bestraffa för modellkomplexitet och välj modellen med det bästa kriterievärdet.
* Eller genom: Dela upp data i tränings- och valideringsuppsättningar flera gånger, passa modeller på träningsuppsättningarna och utvärdera deras prestanda på valideringsuppsättningarna. Välj modellen med bäst genomsnittlig prestanda över valideringsuppsättningarna. Detta kräver datorkraft.

Algortimen är konceptuellt lockande men det är en ”brute force” metod och redan över p=20 så blir det miljoner möjliga modeller. (>2^20)

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

Ingen modell är till fullo verkligheten men den kan ändå vara väldigt användbar. Ett exempel är en mycket grovt skissad karta med bara en flod och ett berg den visar otroligt lite av verkligheten men om du kommer till floden så kan du gå rätt väg till berget om du vet var norr ligger.

# Självutvärdering

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

Detta är mer iterativt än machine learning och jag har studerat datat otroligt mycket i denna uppgift ,mer än modelleringen i tid räknat. men jag tycker det är intressant och är ofta nyckeln till bra modeller tror jag.

Jag tycker att jag har ett bra upplägg på rapporten men hade gärna lämnat in den till någon som kan ”göra den snygg” också, designen blir liksom inte bra i sådana här mallar förrän jag är mycket van med just den mallen.

R studio var lite av en utmaning i början för att det var nytt. Nu tycker jag förståelsen för vissa programmeringsbegrepp djupnat när jag sett det i R och Python.

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

Jag tycker att jag förstår området och har löst de utmaningar jag stött på och är värd VG.

Jag har fått en ny bild av Regressionsmodellering och insett att det är en mer kraftfull metod än jag förstått tidigare från ekonometriska regressioner. Förståelsen för Linjär regression tycker jag verkligen att jag har nu.

1. Något du vill lyfta fram till Antonio?

Det här var riktigt roligt och jag har vänt dessa bildata ”ut o in” väldigt mycket under sista veckorna.

Jag har även beräknat andra moment i normalfördelningen som Kurtosis,skew och annat för att hitta genvägar men tog inte med det hela då det blir för mycket, men det hjälpte min egen förståelse av datat.

# Källförteckning

Box, G. (u.d.).

Geron, A. (u.d.). *(Hands-On\_Machine\_Learning\_with\_Scikit-Learn-Keras-and-TensorFlow-2nd-Edition-Aurelien-Geron.pdf).*

Hastie, T. (2023). An Introduction to ststistical learning 2nd.

*http://yann.lecun.org*. (u.d.).

*https://en.wikipedia.org/wiki/Black\_box*. (u.d.). Hämtat från https://en.wikipedia.org/wiki/Black\_box: https://en.wikipedia.org/wiki/Black\_box

*https://www.rdocumentation.org/packages/leaps/versions/3.1/topics/regsubsets*. (u.d.).

Mans Magnusson, Markus Kainu, Janne Huovari, and Leo Lahti (rOpenGov). (u.d.). Mans Magnusson, Markus Kainu, Janne Huovari, and Leo Lahti (rOpenGov).

pxweb: R tools for PXWEB API. URL: http://github.com/ropengov/pxweb. (u.d.).

Statistics Sweden (2024). “Personbilar i trafik efter region, ägarkategori,. (u.d.).