Datainsamling

Manuellt, Scraping och API



David Norman

EC Utbildning

R-programmering

2024–04

# Abstract

Collection of external data on cars and their attributes from the Blocket. Collection of data took place both manually and with the help of Web-scraper tool and collection amount of 300 - 2727 number of cars. Linear regression modeling explains 71% of the variation in the data with an adjusted R^2 of 0.7159.

An external collection of data from SCB explains that the popularity of fossil-fueled cars tends to drop in a span of 10 years, while electric cars and hybrid cars tends to rise in popularity, this could help predict car prices more precise.

# Förkortningar och Begrepp R-Studio namn på variabler.

Bildata – namnet på excel filen med data från blocket.

mdl – namnet på modellen för Linjär regression.

nybil - namnet på en ny bil data från blocket.

Innehållsförteckning

[1 Inledning 1](#_Toc165194408)

[2 Teori 2](#_Toc165194409)

[2.1 Linjär regressionsmodell 2](#_Toc165194410)

[2.2 Adjusted R-Squared 2](#_Toc165194411)

[2.3 Web scraper 2](#_Toc165194412)

[2.4 API 2](#_Toc165194413)

[3 Metod 3](#_Toc165194414)

[3.1 Manuell datainsamling. 3](#_Toc165194415)

[3.2 Web-scraping. 3](#_Toc165194416)

[3.3 R-studio 3](#_Toc165194417)

[3.4 API 3](#_Toc165194418)

[4 Resultat och Diskussion 4](#_Toc165194419)

[5 Slutsatser 5](#_Toc165194420)

[Appendix A 6](#_Toc165194421)

[6 Referenser 7](#_Toc165194422)

# Inledning

Datainsamling är något som existerat och varit viktigt sedan människans begynnelse, allt från nedskrivning av solens rörelse till web-scraping. I dagens samhälle så är mycket väldigt beroende av just data, skapandet av rapporter, nyfödsel, ny produktion och framför allt reklam.

Datainsamling har ett stort användningsområde såsom att kontrollera antalet nyregistrerade bilar och med den data kunna analysera och se vilken bränsletyp är mer populär och därmed mer relevant i en prisprediktion av bilar.

Nedan kan data analyseras och slutsatsen att bensin och dieseldrivna bilar har sjunkit i popularitet de senaste tio åren enligt SCB, och att ladd hybrid och elbilar har blivit mer populära medan elhybrider verkar ha mindre utveckling med åren.

*Extern data från SCB nyregistrerade personbilar efter region, drivmedel och månad.*

Syftet med denna rapport är att undersöka och diskutera vilken typ av datainsamling är bättre, för att uppfylla syftet så kommer följande frågeställningar att besvaras:

1. Web-scraping eller manuell insamling vad är bäst?
2. Restriktioner på insamlingen av blocketdata
3. Vilka faktorer har en statistisk signifikant påverkan på bilpriset i modellen

# Teori

## Linjär regressionsmodell

Enkel linjär regression är en okomplicerad linjär regressionsmetod för att prediktera kvantitativt svar Y på basis av en enskild prediktionsvariabel. S-61 (James, Witten, Hastie, & Tibshirani, 2023)

Enkel linjär regression är användbar vid prediktioner på basis av en enskild prediktionsvariabel, men i praktiken finns oftast mer än en prediktionsvariabel i detta fall valdes Multipel linjär regression metoden för att prediktera priset för bilar.

## Adjusted R-Squared

Till skillnad från Cp, AIC, BIC, för vilka ett litet värde indikerar en modell med ett lågt testfel, ett stort värde på Adjusted indikerar en modell med ett litet testfel, att maximera Adjusted motsvarar att minimera . S- 235, (6.4) (James, Witten, Hastie, & Tibshirani, 2023)

Intuitionen bakom Adjusted är att när alla korrekta variabler har inkluderats i modellen och ytterligare noise-variabler läggs till kommer endast en mycket liten av RSS att ske. S- 235, (6.4) (James, Witten, Hastie, & Tibshirani, 2023)

## Web scraper

Web scraper startade som en Google chrome extension 2013 och skapades för att göra webb data tillgänglig för alla med sitt verktyg. Verktyget samlar ihop data och ger exportmöjligheten till CSV, XLSX eller JSON format. (SIA, u.d.)

## API

Application Program Interface förkortas API. API är ett protokoll som används för program, system och applikationer på ett enkelt sätt ska kunna kommunicera med varandra.

Ex. ”API:er finns även i en dator. Exempelvis kommunicerar ditt ordbehandlingsprogram med ditt operativsystem när du vill skriva ut något på din skrivare. Kommunikationen sker då via operativsystemets API.” - (E-Identitent, u.d.)

# Metod

## Manuell datainsamling.

Data av bilpriser samlades in manuellt från Blocket.se där vi filtrerade

* Pris mellan 150 000 – 500 000 valdes för att få någorlunda fungerande bilar.
* Familjebussar och arbetsfordon togs bort från filtrering då de generellt är dyrare än andra personbilar.
* Leasingbilar togs bort från sökningen för att säkerställa ett fast pris och inte månadspris.
* Dem tio vanliga bilmärkena valdes för att begränsa sökningen.
* Efter att detta filtrerades började det manuella arbetet, där vi tog och kopierade
* pris, bränsletyp, växellåda, miltal, biltyp, hästkrafter, färg och bilmärke.

Detta fördes sedan över till Excel där vi transformerade och fastställde data för att arbeta med i R-Studio, sammanlagt fick vi ihop 300+ bilar via den manuella datainsamlingen.

## Web-scraping.

Data av bilpriser samlades in med hjälp av web-scraping tillägg ”Web Scraper” på Firefox.  
filtreringar från den manuella sökningen lades till i web scraper, med hjälp av webscraper fick vi ihop data för 2727 bilar som sedan användes i R-studio.

## R-studio

I R-studio användes biblioteken ”readxl”, ”car” och ”caret”.

Vi börjar med att ladda in filen och fästa den i variabeln ”bildata”, därefter undersöker vi data för att få fram vad för datatyp vi tittar på samt namnen på datan. Efter inspektion börjar vi med att transformera data till numerisk eller faktorisk data för att kunna applicera den i en linjär modell.  
Sedan letar vi efter duplicerade rader för att säkerställa att vi inte har dupletter av bilar.  
Vi delar sedan in data i träning och testuppsättningar på 80% och 20%, sedan kontrolleras storleken på de olika sätten.

Vi skapar en linjär regressionsmodell som heter ”mdl” där vi väljer de viktigaste parametrarna  
och inspekterar modellen vi gör en plott på modellen för att se de olika residualerna och för att hitta outliers, vi ser på plotten att bilar med lågt miltal oftast blir outliers.   
När vi undersöker outliers ser vi att ingen utav dem är över gränsvärdet för cook’s distance och därför utgör dem ingen fara för modelleringen.  
Modellen testas sedan mot testsättet.

Vi skapar sedan en prediktion på en ny randomiserad bil. och ser att modellen passar inom ramarna.

## API

Data samlades in från statistiska central byrån med hjälp av deras API länk och biblioteket i R-studio ”pxweb”, ett interaktivt sätt att välja parametrar av data man vill ha. Sedan sparades en json fil som konverteras till en dataframe och läses in i R-studio.

# Resultat och Diskussion

Residuals:

Min 1Q Median 3Q Max

-223871 -24375 -4268 20307 189286

*Residuals data från R-studio*

Det finns en stor spridning på residualerna.

Residual standard error: 47350 on 2164 degrees of freedom

Multiple R-squared: 0.7181, Adjusted R-squared: 0.7159

F-statistic: 324.2 on 17 and 2164 DF, p-value: < 2.2e-16

*RSE, MR-S, AR-S, F-S och P värde från R-Studio*

Modellen förklarar till 71% av variationen i datan, då p-värdet är lågt kan man dra slutsatsen att modellen är signifikant.

vif(mdl)

GVIF Df GVIF^(1/(2\*Df))

Växellåda 1.297447 1 1.139055

Bränsle 4.025878 3 1.261276

Miltal 1.763378 1 1.327922

Biltyp 1.644356 5 1.050992

Drivning 2.136800 1 1.461780

Hästkrafter 2.027089 1 1.423759

Märke 2.778640 5 1.107601

*vif resultat från R-Studio*

Då GVIF är under 10 och under 5 (vanliga gränsdragningar för gvif) så är multikollinearitet mellan variablerna inte ett stort problem för modellen och samtliga variabler kan behållas.

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -3.482e+07 1.015e+06 -34.304 < 2e-16 \*\*\*

Växellådamanuell -2.312e+04 2.716e+03 -8.515 < 2e-16 \*\*\*

Bränslediesel 2.596e+04 2.609e+03 9.950 < 2e-16 \*\*\*

Bränsleel 3.918e+04 3.831e+03 10.225 < 2e-16 \*\*\*

Bränslemiljöbränsle/hybrid 2.204e+03 2.855e+03 0.772 0.4402

Miltal -4.747e+00 2.669e-01 -17.786 < 2e-16 \*\*\*

Modellår 1.736e+04 5.028e+02 34.534 < 2e-16 \*\*\*

Biltypcoupé -2.239e+04 1.140e+04 -1.964 0.0497 \*

Biltyphalvkombi -7.710e+04 9.896e+03 -7.791 1.02e-14 \*\*\*

Biltypkombi -7.169e+04 9.955e+03 -7.202 8.16e-13 \*\*\*

Biltypsedan -5.668e+04 1.046e+04 -5.418 6.69e-08 \*\*\*

Biltypsuv -4.990e+04 9.953e+03 -5.014 5.77e-07 \*\*\*

Drivningtvåhjulsdriven -1.641e+04 2.414e+03 -6.797 1.38e-11 \*\*\*

Hästkrafter 7.881e+02 1.838e+01 42.876 < 2e-16 \*\*\*

Märkeford -2.774e+04 3.001e+03 -9.243 < 2e-16 \*\*\*

Märkerenault -4.323e+04 3.178e+03 -13.604 < 2e-16 \*\*\*

Märketoyota -5.791e+03 3.499e+03 -1.655 0.0981 .

Märkevolkswagen -2.459e+04 2.788e+03 -8.818 < 2e-16 \*\*\*

Märkevolvo -5.336e+02 2.675e+03 -0.199 0.8419

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

*Coefficients från R-Studio*

Samtliga variabler är signifikanta förutom Miljöbränsle/hybrid, Toyota och Volvo.

# Slutsatser

* Web-scraping eller manuellt vad är bäst?  
  Web-scraping är en väldigt bekväm metod att samla in data i stora mängder, dock fanns det ett antal problem med det. Mycket utav den data som blev insamlad saknade värde på dem filtreringar som gjordes vilket sedan var tvunget att manuellt raderas från excel filen. Detta går säkerligen att fixa med vidare arbete och erfarenhet.  
  Till skillnad från Web-scraping så var den manuella datainsamlingen väldigt tidskrävande och repetitivt, men det säkerställde också att datan kunde användas direkt efter insamlingen då insamlingen blev i det format som behövdes.  
  Att säga att scraping är bättre än manuell datainsamling är svårt att säga, allt beror på storleken och hängivenheten av individen som samlar in datan.
* Restriktioner på insamlingen.

I vår datainsamling användes filtrering av bilar och bilmärken samt priser, att sätta ett prisintervall mellan 150 000 och 500 000 var i vår dåvarande tanke ett sätt att säkerställa fungerande bilar med seriösa säljare och att det är i ungefär det priset en ”normal” svensk familj med två inkomster skulle tänka sig betala för en bil.

Detta gav oss ett bra resultat på prediktionerna men också gav oss ett problem. Om en individ sökte en bil under vårt prisintervall skulle inte modellen fungera liknande för om individen sökte en bil över vår intervall. Med detta sagt så anser jag att just prisintervalls restriktioner inte skall göras på en prediktions modell

* Vilka faktorer har en statistisk signifikant påverkan på bilpriset i modellen.  
  Faktorerna som har en signifikant påverkan på bilpriset är samtliga variabler bortsett från

Bränsle faktorn - Miljöbränsle/hybrid.

Märkes faktorn - Toyota och Volvo.

# Referenser

E-Identitent. (u.d.). *Vad är API*. Hämtat från Svensk E-Identitet: https://e-identitet.se/news/vad-ar-api/

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

SIA, W. G. (u.d.). *Webscraper*. Hämtat från https://www.webscraper.io/