Statistisk analys och prisprediktion av begagnade bilar på den svenska marknaden

Geisol Yissel Urbina

EC Utbildning – Data Scientist

Kunskapskontroll i kursen R programmering

2025–04

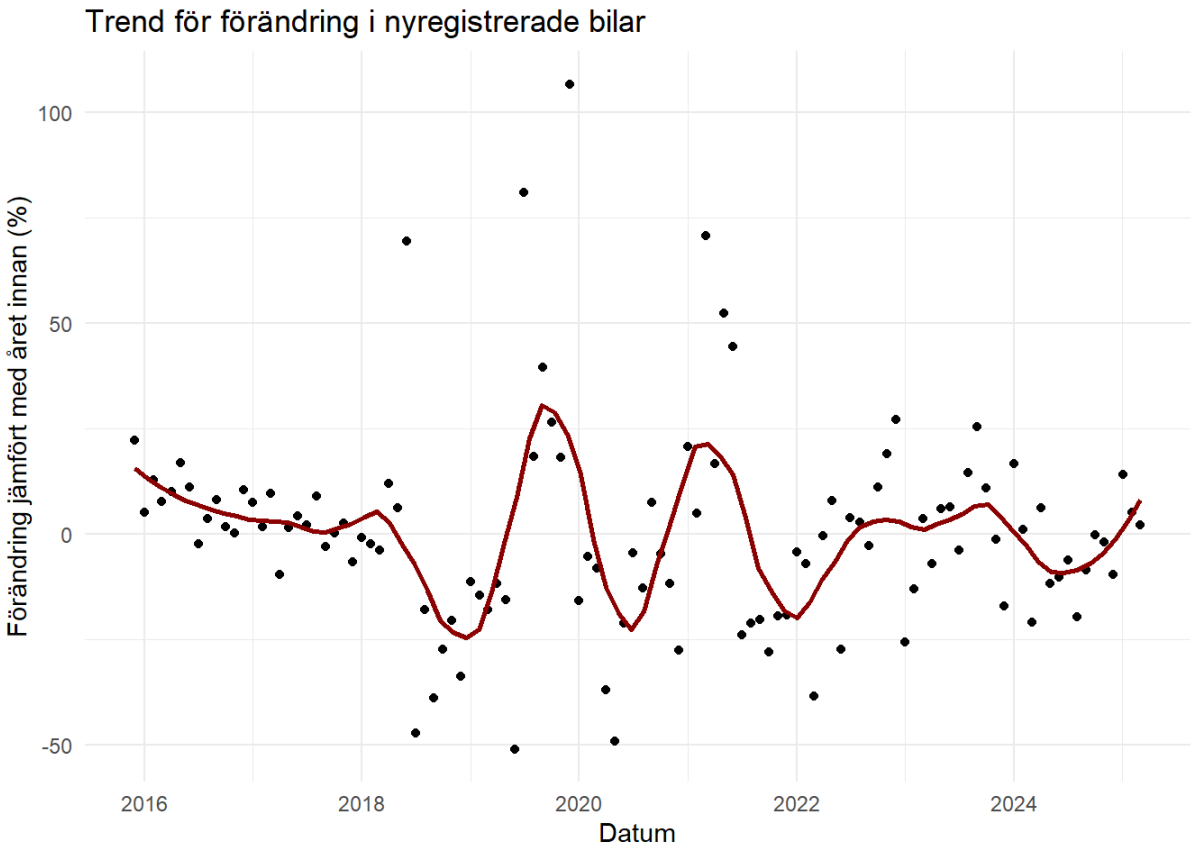
Abstract

This project predicts the market price of mid-range used cars in Sweden by combining linear regression inference and Random Forest regression. We automatically collected listings from Blocket, technical features from Transportstyrelsen, market trends from SCB, and insurance cost estimates from If. After cleaning, log-transforming and filtering outliers, linear models revealed model year and horsepower as the strongest driver of insurance and tax costs. Our tuned Random Forest achieved a log-RMSE of 0.228 (≈45 000 SEK) on cross-validation and 0.231 (≈47 000 SEK) on a hold-out test. The approach both confirms key pricing assumptions and uncovers discrepancies between listed and predicted values, offering a practical tool for value estimation.

Innehållsförteckning

1 Inledning

Under de senaste åren har den svenska begagnade bilmarknaden vuxit kraftigt och genomgått en snabb digitalisering. Webbplattformar som Blocket har minskat kraven för köp och försäljning, vilket har bidragit till ökad pristransparens och snabbare affärsprocesser. Enligt SCB:s statistik över nyregistrerade personbilar (Figur 1) tas omkring 150 000 nya fordon i trafik varje år. Dessa nyregistreringar blir till slut utbud på begagnatmarknaden och styr både utbud och prisnivåer. Figur 1 visar en tydlig förändring i nyregistrerade fordon. Mellan 2016 till 2018 minskade nyregistreringarna gradvis, för att sedan stiga kraftigt under 2019–2020, troligen en effekt av bonus-malis-systemets införande. Under pandemin 2020–2021 föll siffrorna kraftigt igen. Under perioden 2022–2024 ses en mer dämpad men fortsatt volatil utveckling, med en svag återhämtning under 2025. Dessa nya fordon blir succesivt utbud på begagnatmarknaden och påverkar direkt både prisnivåer och omsättningstakt.



Figur 1. Årlig procentuell förändring i nyregistrerade personbilar i Sverige (SCB).

1.1 Syfte

Det här projektet syftar till att genom både statistisk inferens och prediktiv modellering undersöka och förstå vilka faktorer som driver priset på begagnade bilar i Sverige. Målet är att identifiera mönster på bilmarknaden och bygga en modell som kan förklara och förutsäga med hög precision vad som påverkar värde på begagnad bilar. Förutom prisprognoser vill vi undersöka:

- Om elbilar har verkligen lägre försäkringskostnad än konventionella bilar.
- Hur tekniska egenskaper såsom motorstyrka och körsträcka påverkar fordonsskatten.
- Vilka faktorer förklarar varför vissa bilar kostar betydligt mer än andra.

För att nå dessa mål kombineras flera datakällor i en helt automatisk Python pipeline, vilket minimerade manuella fel och säkerställde en fullt repeterbar process. Från Blocket hämtades information om annonser som modellår, märke, körsträcka och andra grundläggande egenskaper. Med hjälp av bildigenkänning i Platerecognizer identifieras registreringsnummer, vilket gör det möjligt att koppla varje bil till data från Transporstyrelsen fordonsregister som hästkrafter, bränsle- och energiförbrukning, koldioxidutsläpp och skatt. Försäkringsbolaget If användes också för att få information om försäkringskostnaden via deras formulär. Dessutom användes statistik från Statistiska centralbyrån (SCB) för att skapa ett perspektiv på marknadens utveckling över tid. Vi analyserar särskilt trender i nyregistreringar av bilar under de senaste tio åren, vilket hjälper oss att förstår utbud och efterfrågan påverkas av exempelvis skatteändringar, pandemins effekter och förändringar i konsumentbeteende.

Genom att samla alla data i samma pipeline får vi både detaljerad information om varje bil och en överblick över marknadstrenderna. Den kombination är avgörande för att förstå våra modeller och se vilka faktorer som verkligen styr priserna på begagnade bilar i Sverige.

2 Metod

2.1 Statistik centralbyrån (SCB)

Vi hämtade månadsstatistik över nyregistrerade bilar från SCB:s öppna API med hjälp av R-paketet **pxweb**. Sedan räknade vi ut hur mycket varje månad skiljde sig i procent jämfört med samma månad året innan, från december 2015-mars 2025. Resultatet lades in i en enkel tabell (data.frame) och plottade som en tidsserie för att identifiera tydliga toppar kopplade till skatteförändringar. Detta användes för att tolka effekter på andrahandspriser.

2.2 Datainsamling och dataförberedelse

Datainsamlingen har genomförts med hjälp av ett Python-skript med Selenium för webskrapning. Automationsverktyget gjorde det möjligt att hämta över 6000 annonser från Blocket. För att undvika problem med Blockets policy laddade vi ner Chrome för testning och ChromeDriver, använde undetected-chromedriver och justerade tidsintervallen mellan klick för att det skulle se mer mänskligt ut. Vi valde också att använda en VPN för att byta IP varje gång vi körde koden. I huvudsak gjorde koden att Chrome för testing öppnades och navigerade igenom alla resultatsidor och samlade in annonser. Den sparade länkarna till varje bil och hämtade därefter specifikationer som pris, färg, hästkrafter och bilder i individuella mappar. Redan insamlade annonser ignorerades, vilket gjorde datauppdateringen effektiv.

Det var viktigt att ha bilder på bilarna, och särskilt att registreringsskyltarna syntes. Alla annonser som saknade sådana bilder raderades. Detta leder oss till nästa steg, där vi använder en Python-baserad lösning för att automatiskt känna igen registreringsskyltar i bilannonsernas bilder. Genom att integrera Plate Recognizer's API skickade vi varje bild i respektive bils mapp till tjänsten, som returnerade det identifierade registreringsnumret. När registreringsnumret bedömdes som giltigt sparades det i filen **regnum.txt.** Därefter raderades alla bilmappar som antingen saknade bilder eller innehöll etiketter som inte ansågs giltiga. Med registreringsnumren kunde vi via Transportstyrelsens fordonsregister hämta tekniska specifikationer och kostnadsuppgifter såsom fordonskatt, bilmärke, bränsleförbrukning och CO_2-utsläpp. Dessa uppgifter sparades i respektive fordonsmapp som **textfiler.** För att undvika dubbletter hämtade skriptet endas data för de bilar där informationen saknades, vilket säkerställde att inga uppgifter sparades två gånger.

Insamlingen av försäkringskostnader för varje bil automatiserades genom att hämta prisinformation från If webbplats. För varje bil mapp som innehöll ett registreringsnummer navigerade skriptet till If:s formulärsida, fyllde i bilens registreringsnummer och ett personnummer och klickade på ”se ditt pris”. Om ett månatligt pris hittades sparades det som text i filen **forsakring.txt** i respektive bils mapp. Om priset inte gick att hämta tolkades som misslyckad och hela mappen raderades. På så sätt inkluderar vi månatliga försäkringskostnader i vår datauppsättning, en avgörande del av beräkningen av totala ägandekostnader. Sedan samlade vi alla **.txt** filer, extraherade nyckelvariabler och skrev ut dem i ett Excel-fil med en rad per annons för enkel analys.

Vissa data transformerades i Excel, exempelvis konverterades körsträckan från miles till kilometer. Efter hela processen återstod bara 568 observationer för analys.

2.3 Statistisk inferens

För att genomföra den statistiska inferensen använde vi R-kod. Vi har laddat ned paket som *tidyverse* för datahantering, vilket bland annat innehåller *dplyr* och *tidyr*, samt *ggplot2* som hjälpte oss att visualisera resultaten. Vi använde även *readxl* för att läsa in Excel-filen, *fastDummies* för att skapa dummy-variabler, vilket var viktigt för att kunna inkludera kategoriska variabler i analysen, samt paketet *car* som var viktigt för att använda VIF (variance inflation Factor) vid multikollineraritetsanalysen. Vi började med att log-transformera målvariablerna pris, helförsäkringskostnad och fordonskatt och viktiga förklarande variabler som hästkrafter, körsträcka och bränsleförbrukning, samt dummy-koda kategorivariabler som modellår och bränsletyp. Därefter byggde vi tre separata linjära modeller med funktionen *lm*, en för varje kostnadsvariabel. För varje modell tog vi fram koefficienter, p-värden och justerat R^2 via *summary*, och beräknade 95% kofidensintervall med *confint.* För att kontrollera multikollinaritet använde vi VIF från *car*-paket, där variabler med VIF över 5 exkluderades för att säkerställa stabila skattningar.

Slutligen granskade vi modellerna genom att plotta residualer och QQ-diagram med standarddiagnostik i R, vilket bekräftade att residualerna var någorlunda normalfördelade och homogenist spridda. Denna enkla med robusta process gav tydliga och tolkbara effektskattningar, samt lade grunden för vår vidare analys.

2.4 Prediktiv modellering

I den här delen använde vi prediktiv modellering för att skapa en modell som kan förutsäga priset på begagnade bilar, utifrån tekniska och ekonomiska data. Syftet var både att kunna räkna ut ett ungefärligt marknadspris och att hitta bilar som verkar vara ovanligt billiga eller dyra jämfört med vad modellen förväntar sig. Vi började med att installera de paket vi ansåg nödvändiga för vår analys som *dplyr*, *fastDummies* och *janitor* för datamanipulation, *rsample* för datauppdelning och korsvalidering, *rang*er för Random Forest, *purrr* för funktionsprogrammering, *yardstick* för prestationsmått, *tibble* för tabellhantering och *vip* för variabelimportance. Först förberedde vi data genom att utgå från vårt dataset (dataset_final.xlsx) och lägga till en unik Id-kolumn. Därefter log-transformerade vi alla målvariablerna (pris, fordonsskatt och helförsäkring), samt viktiga förklarande variabler som hästkrafter och mätarställning för att dämpa skevheter och skapade noll-ett-indikator för kategoriska fält som modellår och bränsletyp. Nästa steg var att beräkna 1- och 99-percentiler för varje log-transformerad variabel och filtrera bort extremvärden både före och efter uppdelningen av data.

Vi delade datamängden stratifierat (80/20) baserat på log-pris, så att både träning- och testset speglade prisets fördelning. Utifrån de tidigare percentilgränserna tog vi bort outliers i respektive uppdelning. Vi testade först en enkel linjär modell, men dem var för stel och fick stora fel när priset påverkades av flera faktorer samtidigt. Så vi valde Random Forest eftersom den klarar av att hitta både icke-linjära samband och hur variablerna samspelar, vilket gav bättre prisprognoser.

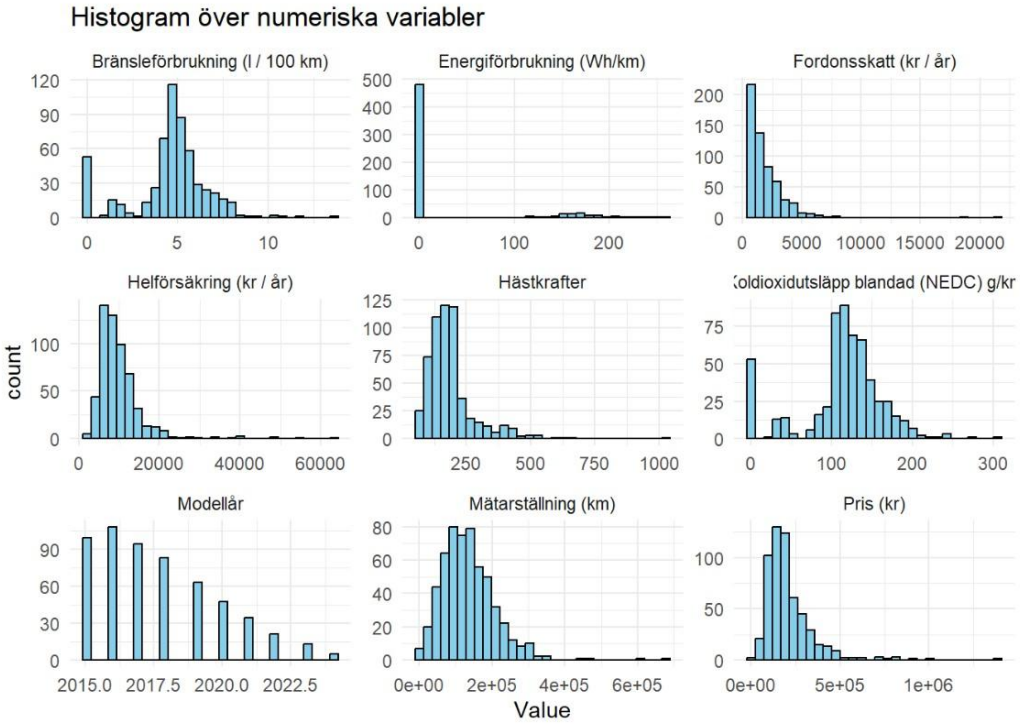
På träningsdata använde vi grid-search med en 5-faldig korvsvalidering för att hitta de hyperparametrar som gav lägst genomsnittligt log-RMSE. Vi testade antal träd, variabler per split, minsta nodstorlek och maxdjup. Med de bästa parametrarna tränade vi slutligen modellen på hela träningsdata och utvärderade log-RMSE, samt omvandlade prediktionerna tillbaka till kronor för att från RMSE i SEK på testdata.

För att förstå vilka funktioner som bidra mest till modellens noggrannhet visualiserade vi variabelimportance med vip-paket. Resultatet blev en stabil modell där tydligt visar vilka faktorer som påverkar priset mest.

I det sista steget genomförde vi en residualanalys där vi för varje bil beräknar avvikelsen mellan vekligt pris och förutspått pris, både i kronor och i procent. Därefter identifierar vi de fem bilar som är mest övervärderade och de fem som är mest undervärderade. Slutligen undersöker vi modellens variabelimportance med hjälp av paketet vip, där de tio viktigaste förklarande variablerna rangordnas efter hur mycket de bidrar till våra prisprognoser. Allt detta blir en hel process för att bygga, testa och förstå vår prismodell på ett enkelt sätt.

3 Resultat och Diskussion

3.1 Exploratory Data Analysis (EDA)



Figur 2. Histogram över centrala fordonsvariabler (endast numeriska).

När vi gjorde EDA ritade vi histogram för att se hur alla variabler var fördelade. Det visade sig att många numeriska variabler var högerskevda med långa svansar av extrema värden. Sådana sneda fördelning gör att några få stora observationer drar iväg resultaten. För att dämpa effekter och göra spridningen mer symmetrisk valde vi därför att log-transformera samtliga dessa variabler. Därefter ritade vi om histogrammen och kunde se att de log-transformerade värdena blev betydligt mer balanserade, vilket motiverade den fortsatta analysen på den log-skalan.

3.2 Korrelationsinspektion

	Variable	Correlation
	<chr>	<dbl>
1	Helförsäkring_log	1
2	Hästkrafter_log	0.641
3	Pris_log	0.491
4	Skatt_log	0.328
5	Koldioxidutsläpp blandad (NEDC) g/km	0.177
6	Mätar_log	0.175
7	Bränsleförb_log	0.0522

Tabell 1. Korrelationsanalysen.

Korrelationsanalysen visar att hästkrafter_log har det starkaste sambandet med den log-transformerade helförsäkringskostnaden (r ≈ 0.64), vilket innebär att bilar med fler hästkrafter i genomsnitt har högre försäkringskostnad. Därefter följer bilens pris (Pris_log) med en korrelation på cirka 0.49, vilket indikerar att dyrare fordon generellt också är dyrare att försäkra. Fordonsskatt (skatt_log) visar ett måttligt samband (r ≈ 0,33), medan koldioxidutsläpp och körsträcka (Koldioxidutsläpp blandad g/km och Mätar_log) båda ligger runt 0,18, vilket tyder på en svag men positiv effekt på försäkringskostnad. Slutligen har logaritmen av bränsleförbrukningen (Bränsleförb_log) nästan ingen korrelation (r ≈ 0,05), vilket tyder på att bränsleförbrukningen i sig påverkar försäkringskostnaden lite. Sammanfattningsvis är det främst motoreffekt och pris som driver försäkringspremien, medan övriga variabler spelar en mer begränsad roll.

3.3 Statistisk inferens för försäkringskostnad

En linjär regressionsmodell har använts för att förstå vad som påverkar priset på helförsäkring för bilar. Resultatet visar vilka egenskaper hos bilen som gör att försäkringen blir dyrare eller billigare.

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.62613 -0.13991  0.01238  0.14633  1.12691

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                 4.425059   0.355816  12.436  < 2e-16 ***
Hästkrafter_log             0.673330   0.043156  15.602  < 2e-16 ***
Modellår2016                0.130499   0.040766   3.201 0.001448 **
Modellår2017                0.084723   0.042999   1.970 0.049300 *
Modellår2018                0.175707   0.045889   3.829 0.000143 ***
Modellår2019                0.189512   0.051514   3.679 0.000257 ***
Modellår2020                0.122110   0.057830   2.112 0.035176 *
Modellår2021                0.105902   0.067657   1.565 0.118090
Modellår2022               -0.320774   0.080153  -4.002 7.14e-05 ***
Modellår2023               -0.767089   0.101016  -7.594 1.34e-13 ***
Modellår2024               -0.666266   0.144128  -4.623 4.72e-06 ***
Mätar_log                  -0.002024   0.016187  -0.125 0.900519
BränsleDiesel               0.191398   0.030503   6.275 7.09e-10 ***
BränsleEl                   0.148035   0.052617   2.813 0.005076 **
BränsleMiljöbränsle/Hybrid -0.039388   0.047873  -0.823 0.411004
Pris_log                    0.088860   0.034501   2.576 0.010267 *
---
```

Tabell 2. Tolkning av regressionskoefficienter med målvariabel helförsäkring.

Tabell 2 visar att försäkringspriset för bilar påverkas mest av hästkrafter, då en ökning på 10% i motorstyrka höjer premien med ungefär 7%. Bilar tillverkade mellan 2016 och 2020 är också generellt dyrare att försäkra än äldre bilar, och ju högre pris bilen har, desto högre blir försäkringskostnaden. Till exempel ger 10% högre bilpris ungefär 1% högre premie. Både diesel- och elbilar är dyrare att försäkra än bensinbilar; dieselbilar är cirka 21% dyrare och elbilar runt 16% dyrare än motsvarande bensinbilar. Det som däremot gör försäkringen billigare är om bilen är av den allra nyaste årsmodellen, 2022–2024, vilket sannolikt beror på

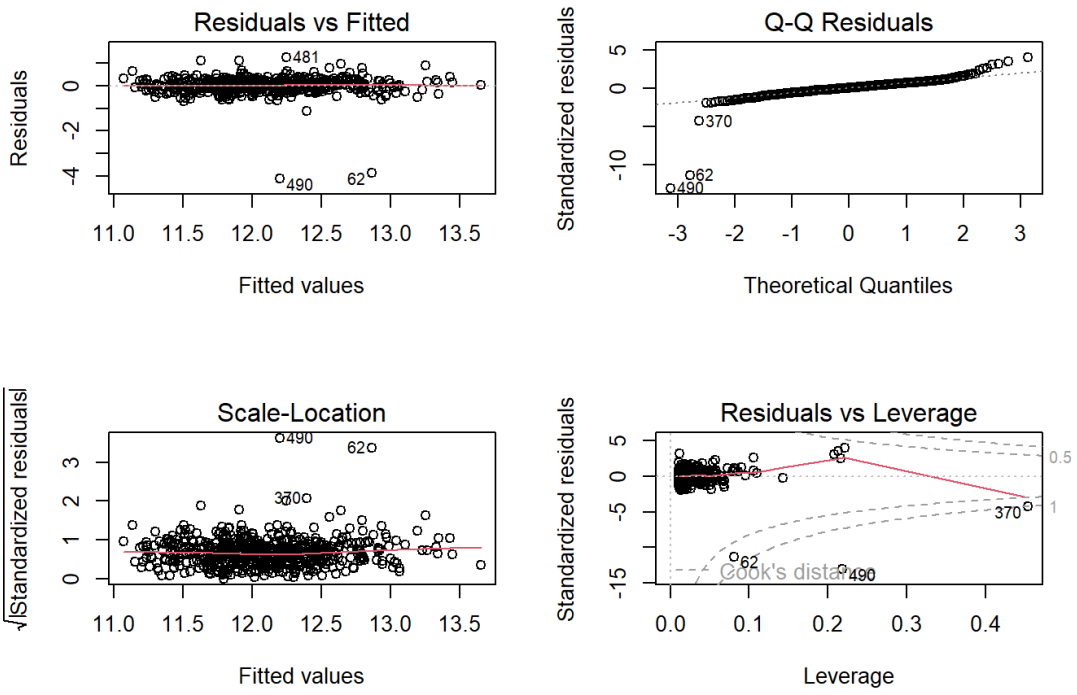
ny teknik och bättre säkerhet. Till exempel en bil från 2024 kan till exempel vara över 50% billigare att försäkra jämfört med en äldre bil. Körsträckan, alltså hur långt bilen har gått, verkar inte påverka försäkringspriset, och inte heller om bilen är en hybrid- eller miljöbil jämfört med en bensinbil. Modellen förklarar cirka 61% av skillnaderna i försäkringspris mellan olika bilar, vilket innebär att dessa faktorer har stor betydelse för hur försäkringspremierna sätts.

	2.5 %	97.5 %
(Intercept)	3.7261375286	5.12398106
Hästkrafter_log	0.5885590978	0.75810107
Modellår2016	0.0504226542	0.21057577
Modellår2017	0.0002605098	0.16918554
Modellår2018	0.0855680936	0.26584602
Modellår2019	0.0883236695	0.29070109
Modellår2020	0.0085155506	0.23570527
Modellår2021	-0.0269944559	0.23879814
Modellår2022	-0.4782159558	-0.16333214
Modellår2023	-0.9655113110	-0.56866605
Modellår2024	-0.9493729123	-0.38315816
Mätar_log	-0.0338211684	0.02977225
BränsleDiesel	0.1314817013	0.25131371
BränsleEl	0.0446815413	0.25138920
BränsleMiljöbränsle/Hybrid	-0.1334234613	0.05464811
Pris_log	0.0210903071	0.15662998

Tabell 3. Tabellen visar 95% Konfidensintervall för varje variabels koefficient i regressionsmodell.

Konfidensintervallen visar att motorstyrka (hästkrafter), modellår 2016–2020, de allra nyaste modellåren 2022–2024, bilens pris samt om bilen drivs på diesel eller el har en statistiskt säkerställd effekt på försäkringspriset. Det innebär att fler hästkrafter, nyare bil (2016–2020), högre pris samt diesel- eller eldrift tydligt ökar försäkringspremien, medan de allra nyaste bilarna (2022–2024) har signifikant lägre premier. Samtidigt visar resultaten att körsträcka och om bilen är hybrid eller miljöbil inte har någon tydlig effekt på priset.

Resultaten från VIF-analysen visar att det inte finns någon problematisk multikollinearitet mellan variablerna i vår regressionsmodell. Alla variabler har justerade GVIF-värden (GVIF^(1/(2*Df))) som ligger långt under gränsvärdet 5, vilket innebär att de inte är starkt korrelerade med varandra och att multikollineariteten är låg i modellen.



Figur 3. Diagnostikplotar för regressionsmodell med försäkringskostnad som målvariabel.

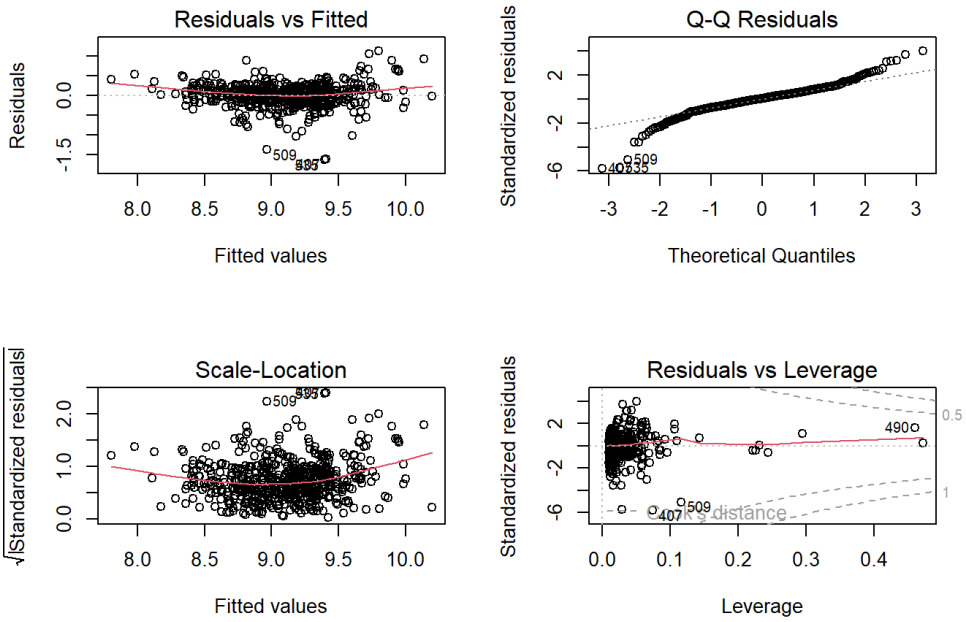
Diagnostikplottarna visar att modellen fungerar bra och uppfyller de viktigaste kraven för linjär regression. Residualerna är fördelade jämnt kring noll utan tydliga mönster, vilket tyder på att modellen inte har några allvarliga systematiska fel. QQ-plottens residualer följer till största delen normalfördelningslinjen, vilket innebär att normalitetsantagandet är acceptabelt uppfyllt även om några enstaka extrema värden sticker ut. Scale-Location visar att variansen hos residualerna är ungefär konstant, vilket betyder att homoskedasticitetsantagandet (lika spridning) är uppfyllt. Slutligen visar Residuals vs Leverage att de flesta observationer har låg påverkan på modellen, men det finns några få observationer med hög leverage som skulle kunna ha stor inverkan på resultatet.

3.4 Statistisk inferens för bilpris

Variabel	β estimat	95 % KI	p-värde	Tolkning*
Hästkrafter (log)	0,859	[0,783 ; 0,935]	<0,001	+1 % fler hk ⇒ ≈ 0,86 % högre pris
Modellår 2016	0,179	[0,081 ; 0,277]	<0,001	≈ 20 % dyrare än 2015
Modellår 2017	0,295	[0,194 ; 0,396]	<0,001	≈ 34 % dyrare än 2015
Modellår 2018	0,421	[0,316 ; 0,527]	<0,001	≈ 52 % dyrare än 2015
Modellår 2019	0,529	[0,412 ; 0,646]	<0,001	≈ 70 % dyrare än 2015
Modellår 2020	0,581	[0,449 ; 0,712]	< 0,001	≈ 79 % dyrare än 2015
Modellår 2021	0,625	[0,470 ; 0,781]	< 0,001	≈ 87 % dyrare än 2015
Modellår 2022	0,809	[0,627 ; 0,991]	<0,001	≈ 124 % dyrare än 2015
Modellår 2023	0,553	[0,313 ; 0,793]	<0,001	≈ 74 % dyrare än 2015
Modellår 2024	–0,007	[–0,357 ; 0,342]	0,97	Ingen effekt (få observationer)
Mätar_log	–0,053	[–0,091 ; –0,014]	0,008	+1 % mer mil ⇒ ≈ 0,05 % lägre pris
Bränsle Diesel	0,035	[–0,039 ; 0,109]	0,35	Ingen säker skillnad mot bensin
Bränsle El	–0,258	[–0,383 ; –0,132]	<0,001	≈ 23 % billigare än bensin
Bränsle Miljö/Hybrid	–0,095	[–0,211 ; 0,021]	0,11	Ingen säker skillnad mot bensin

Tabell 4. Tolkning av regressionskoefficienter med målvariabel bilpris.

Resultaten visar att bilens pris påverkas av motorstyrka, årsmodell, mätarställning och delvis av bränsletyp. Större motor gör bilen dyrare; en 10 % ökning i hästkrafter ger ungefär 8,6 % högre pris. Nyare bilar har också högre pris, och effekten ökar för varje modellår. Till exempel är en bil från 2022 betydligt dyrare än en äldre bil, och även årsmodeller från 2016 till 2021 har stegvis högre priser jämfört med referensåret. Mätarställning har en svag negativ effekt, vilket betyder att bilar med högre körsträcka blir billigare. När det gäller bränsletyp är elbilar signifikant billigare än bensinbilar, medan dieselbilar och hybridbilar inte skiljer sig i pris från bensinbilar. Modellen förklarar drygt 61 % av variationen i bilpriserna, vilket tyder på att de inkluderade faktorerna är viktiga för hur priset bestäms.



Figur 4. Diagnostikplottar för regressionsmodell med pris som målvariabel.

Figur 4 visar att modellen i stort sett uppfyller de statistiska kraven för linjär regression. I "Residuals vs Fitted" syns att residualerna ligger ganska jämnt fördelade kring noll, utan något tydligt mönster, vilket tyder på att modellen passar datan bra, även om det finns några punkter som sticker ut. QQ-plott visar att residualerna till stor del följer en rak linje, vilket innebär att normalitetsantagandet är ganska väl uppfyllt, även om det finns några avvikelser i svansarna. "Scale-Location" indikerar att spridningen hos residualerna är någorlunda jämn över olika förväntade värden, vilket är bra för modellens tillförlitlighet. Slutligen visar "Residuals vs Leverage" att de flesta observationer har låg påverkan på modellen, men några enstaka punkter med hög leverage kan påverka modellens resultat.

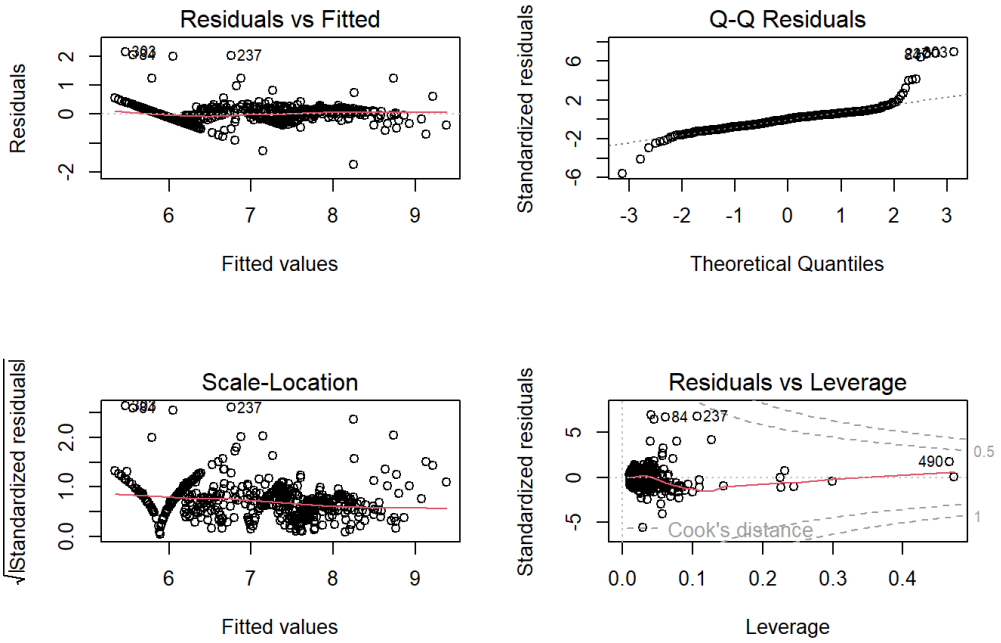
3.5 Statistisk inferens för fordonsskatt

Analysen visar att bilens skatt till stor del bestäms av vilken bränsletyp bilen har, mängden koldioxidutsläpp och hur ny bilen är. Framför allt har diesel- och elbilar betydligt högre skatt än bensinbilar, och effekten är mycket tydlig. Ju mer koldioxid bilen släpper ut, desto högre blir skatten. Nya bilar, särskilt de som är tillverkade efter 2019, har också högre skatt än äldre bilar. Bilens hästkrafter har viss betydelse, där fler hästkrafter kan ge något högre skatt, men den effekten är mindre säker. Faktorer som bilens pris, hur långt bilen har gått och om det är en hybridbil påverkar inte skatten tydligt enligt modellen. Modellen förklarar nästan 88 % av variationen i bilskatt, vilket innebär att dessa variabler ger en mycket bra förklaring till hur skatten bestäms för olika bilar.

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.73571 -0.17956  0.00722  0.14782  2.14602

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                        3.7474036  0.3921125   9.557  < 2e-16
Hästkrafter_log                    0.0844377  0.0480966   1.756  0.07972
Modellår2016                      -0.0086872  0.0444502  -0.195  0.84512
Modellår2017                      -0.0051490  0.0468984  -0.110  0.91261
Modellår2018                       0.0722257  0.0500423   1.443  0.14951
Modellår2019                       0.1370839  0.0561549   2.441  0.01495
Modellår2020                       0.3889436  0.0636855   6.107 1.92e-09
Modellår2021                       0.4169018  0.0744365   5.601 3.37e-08
Modellår2022                       0.6054214  0.0878560   6.891 1.52e-11
Modellår2023                       0.5350574  0.1107225   4.832 1.75e-06
Modellår2024                       0.4453881  0.1588154   2.804  0.00522
Mätar_log                         -0.0155264  0.0176585  -0.879  0.37964
BränsleDiesel                      1.0983733  0.0333746  32.910  < 2e-16
BränsleEl                          0.8811862  0.0912131   9.661  < 2e-16
BränsleMiljöbränsle/Hybrid         0.0799366  0.0632723   1.263  0.20699
Pris_log                           0.0571287  0.0382993   1.492  0.13637
`Koldioxidutsläpp blandad (NEDC) g/km`  0.0140908  0.0005157  27.325  < 2e-16
```

Tabell 5. Regressionskoefficienter för målvariabeln fordonskatt. Varje koefficient har testats med tvåsidiga t-test (H_0: $\beta = 0$) och 95 % konfidensintervall.



Figur 5. Diagnostikplottar för regressionsmodell med fordonsskatt som målvariabel.

Diagnostikplottarna visar att modellen fungerar ganska bra, men det finns vissa saker att vara uppmärksam på. Felen (residualerna) är inte helt slumpmässiga och några värden ligger utanför det normala, vilket antyder att modellen kanske inte fångar alla mönster i datan. QQ-plottan visar att de flesta värden följer normalfördelningen, men det finns några tydliga avvikelser i kanterna, alltså några extrema värden. Variationen i felen verkar öka något för högre förväntade värden, vilket kan betyda att modellen inte är helt stabil över hela skalan. Slutligen har de flesta observationer liten påverkan på modellen, men ett fåtal sticker ut och kan påverka resultatet mycket.

3.6 Prediktiv modellering av bilpriser (Random Forest)

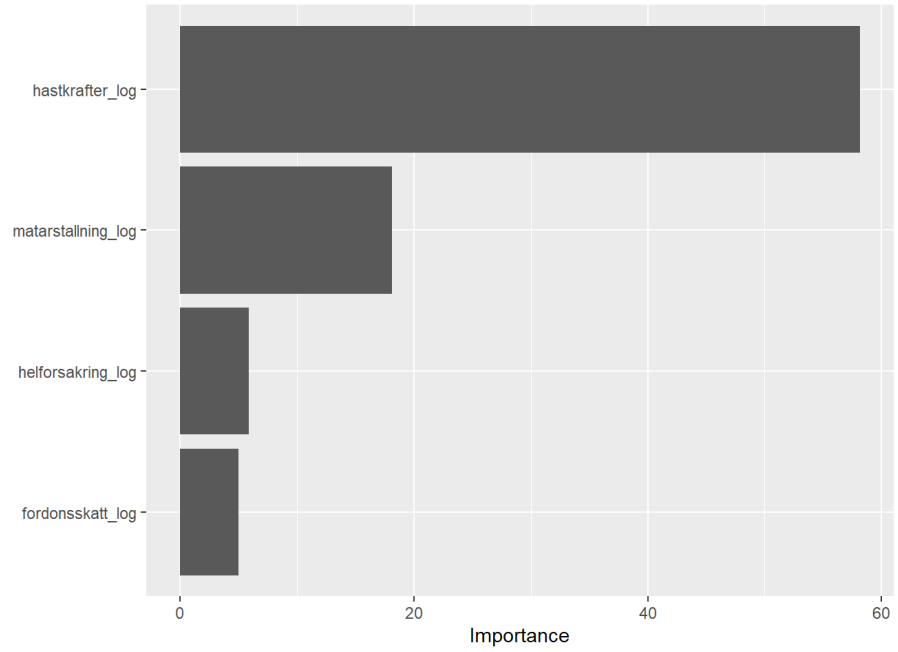
Efter log-transformering av samtliga numeriska prediktorer och målet, samt borttagning av observationer utanför det 1- och 99-percentilen, tränades en Random Forest-regressionsmodell för att prediktera logaritmerade bilpriser. Datan stratifierades och splittrades i 80 % träning och 20 % test, och optimering av hyperparametrar genomfördes med grid search med kombination med 5-faldig korsvalidering. Den slutliga modellen tränades på samtliga träningsdata med de optimala hyperparametrarna.

Modellens prediktionsförmåga utvärderades dels med korsvalideringens genomsnittliga RMSE på log-skalan, dels på en oberoende testmängd. För att få en intuitiv tolkning beräknades även RMSE i SEK genom att multiplicera det logaritmiska RMSE med medelvärdet för predikterat pris i testdata.

Dataset	**RMSE (log-skala)**	**RMSE (SEK)**
Train (CV, K=5)	0,2313	44 297
Test	0,2353	56 829

Tabell 6. RMSE_SEK = mean(prediktionspris) × RMSE_log. CV = korsvalidering

Modellen uppvisar god generaliserbarhet: skillnaden mellan RMSE för korsvalidering (0,232) och testet (0,239) är minimal, vilket tyder på låg risk för överanpassning. Ett RMSE på 0,239 i log-skala innebär att de predikterade priserna i genomsnitt avviker med cirka 57 000 kr från de faktiska transaktionspriserna i testdata. Variabelimportansen indikerar att modellår, hästkrafter-log och mätarställning-log är de faktorer som bidrar mest till modellens förklaringskraft. Residualanalysen avslöjar dessutom både över- och underprissatta bilar, något som kan användas för att flagga avvikande objekt för vidare analys eller affärsmässiga beslut.



Figur 6. Variabelimportans från Random-Forest-modellen som predikterar log-pris

Hästkrafter dominerar modellens beslut, följt av körsträcka. Försäkrings- och skattenivåer spelar också in men i mindre grad. Detta ger stöd för att priset på en begagnad bil i första hand drivs av dess prestanda (hästkrafter) och skick (körsträcka), medan årliga kostnader som försäkring och skatt har sekundär betydelse i prediktionen.

3.6.1 Residualanalys – överprissatta vs underprissatta bilar

Efter att Random-Forest-modellen tränats och testats beräknades residualerna på testdata, det vill säga skillnaden mellan faktiskt pris och modellens predikterade pris för varje bil.

```
# A tibble: 5 × 7
     id fordonsbenamning     handelsbeteckning   pris pred_sek diff_sek diff_pct
  <int> <chr>                <chr>              <dbl>    <dbl>    <dbl>    <dbl>
1   481 KIA EV9              EV9               735000  421114.  313886.     74.5
2   333 BMW X5 XDRIVE45E     X5 XDRIVE45E      560000  355138.  204862.     57.7
3   269 MERCEDES-BENZ V-KLA… V-KLASSE          469000  337449.  131551.     39.0
4   360 SUBARU OUTBACK       OUTBACK           305000  184374.  120626.     65.4
5   478 VOLVO                XC60              270000  160831.  109169.     67.9
```

Tabell 7. Fem överprissatta bilar, där pred_sek = prognos i kr, diff_sek =skillnad i kr, och diff_pct = skillnad i procent.

Tabellen visar de fem mest överprissatta bilarna i testdata genom att jämföra varje bils annonspris med modellens beräknade marknadsvärde. Kia EV9 toppar listan: den annonseras för 735 000 kr men modellen värderar den till cirka 421 000 kr, ett överpris på drygt 314 000 kr (+75 %). Även BMW X5 och Mercedes-Benz ligger högt över modellvärdet, med pris-påslag på 58 % respektive 39 %. Subaru och Volvo XC60 är 65 % respektive 68 % dyrare än vad modellen bedömer som rimligt. Dessa stora positiva residualer indikerar antingen verklig överprissättning eller att bilarna har unika egenskaper som modellen inte fångar till exempel ovanlig utrustning eller mycket låg tillgång på marknaden. För köpare signalerar listan objekt där priset bör ifrågasättas eller förhandlas, medan säljare får en tydlig indikation på att deras utropspris ligger betydligt över vad marknaden sannolikt är villig att betala.

```
# A tibble: 5 × 7
     id fordonsbenamning     handelsbeteckning   pris pred_sek diff_sek diff_pct
  <int> <chr>                <chr>              <dbl>    <dbl>    <dbl>    <dbl>
1   390 MITSUBISHI OUTLANDER MITSUBISHI OUTLA… 208000  290057.  -82057.    -28.3
2   327 VOLVO V60            V60               329900  411162.  -81262.    -19.8
3   280 MERCEDES-BENZ AMG C… AMG CLA 45        285000  365814.  -80814.    -22.1
4    64 TESLA MODEL 3        MODEL 3           270000  347833.  -77833.    -22.4
5    97 NISSAN               NISSAN NAVARA     239000  309774.  -70774.    -22.8
```

Tabell 8. Fem underprissatta bilar, där pred_sek = prognos i kr, diff_sek =skillnad i kr, och diff_pct = skillnad i procent.

Tabellen visar de fem mest underprissatta bilarna i testdatan, det vill säga annonser där priset ligger betydligt under det marknadsvärde modellen beräknar. Mest anmärkningsvärd är en Mitsubishi som säljs för 208 000 kr trots att modellen värderar den till cirka 290 000 kr – ett potentiellt fynd med ett prisgap på drygt 82 000 kr (-28 %). Volvo V60, Mercedes, Tesla Modell 3 och Nissan Navara ligger alla 70 000–81 000 kr (-20 % till -23 %) under sina respektive modellpriser. Sådana negativa residualer tyder på att bilarna antingen är aggressivt prissatta för att säljas snabbt eller har egenskaper som modellen inte fångar, till exempel hög körsträcka, skador eller bristande servicehistorik. Residualanalysen demonstrerar hur modellen kan användas för att identifiera attraktiva affärsmöjligheter och avvikelser i prissättningen på begagnatmarknaden.

4 Slutsatser

Våra frågeställningar besvarades med hjälp av både inferentiell och prediktiv analys.

1. Har elbilar lägre försäkringskostnad än konventionella bilar?
   Nej. Regressionsmodellen visar att elbilar ligger i genomsnitt ≈ 15 % högre i helförsäkringspremie än motsvarande bensinbilar ($p < 0{,}01$). Dieselbilar är ännu dyrare (+ 21 %), medan hybridbilar inte skiljer sig signifikant från bensin. Elbilar är alltså inte billigare att försäkra i det analyserade datasetet.
2. Hur påverkar tekniska egenskaper såsom motorstyrka och körsträcka fordonsskatten?
   Koldioxidutsläpp och drivmedel dominerar mest. Skattmodellen visar stark, positiv effekt av CO_2-utsläpp samt kraftiga påslag för diesel- och elbilar. Hästkraft-log har endast en svag, gränssignifikant effekt och körsträcka- log ingen tydlig effekt alls. Skatten styrs alltså i första hand av utsläppsprofil och bränsletyp – inte av ren motoreffekt eller hur långt bilen gått.
3. Vilka faktorer förklarar varför vissa bilar kostar betydligt mer än andra?
   De tre viktigaste drivkrafterna, bekräftade av både linjär regression och Random Forest är:
   - Modellår: nyare bilar värderas högre, årsmodeller 2016–2022 ger successivt större prispåslag.
   - Hästkraft: visar en 10 % ökning i hästkrafter höjer priset med ca 8–9 % och
   - Körsträcka: högre mätarställning sänker priset.

5 Teoretiska frågor

1. En Quantile-Quantile (QQ) plot är ett diagram som jämför fördelningen av data mot en teoretisk fördelning. I en QQ-plot plottas kvantilerna från data mot motsvarande kvantiler från den teoretiska fördelningen. Om punkterna i plottet följer en rak linje innebär det att dina data har en liknande fördelning som den teoretiska. QQ-plots används ofta för att undersöka om data är normalfördelade.

2. Inom maskininlärning ligger focus på att skapa modeller som kan göra bra prediktioner, alltså att förutsäga nya utfall. Medan statistisk regressionsanalys vill man däremot inte bara prediktera, utan också förstå sambandet mellan variabler, det vill säga göra statistisk inferens. Till exempel i vårt projekt har vi använt en maskininlärningsmodell för att förutsäga priset på en bil baserat på egenskaper som modellår, miltal, märke och ålder. Målet är att göra prisprognoser så exakta som möjligt, även om man inte kan tolka hur variablerna påverkar utfallet. Medan i statistisk inferens har vi använt regressionsanalys, där vi kunde undersöka vilka variabler påverkade försäkringskostnad. I det här fallet var hästkraft och pris.

3. Skillnad mellan konfidensintervall och prediktionsintervall är att konfidensintervall ger ett intervall för medelvärdet av den beroende variabeln för en given x förväntas ligga. Medan prediktionsintervall ger ett intervall där ett enskilt nytt värde på den beroende variabeln för en given x förväntas hamna. Prediktionsintervallet är alltid breddare är konfidensintervallet, eftersom det tar hänsyn till både osäkerheten i medelvärdet och variationen mellan individuella observationer.

4. Beta-parametrar (β_p) i en multipel linjär regressionsmodell beskriver hur mycket vi förväntar oss att det beroende variabeln Y ska förändras om just den oberoende x-variabeln ökar med 1 enhet, samtidigt som alla andra variabler hålls oförändrade.

5. BIC (Bayesian Information Criterion) är ett mått som används för att jämföra olika statistiska modeller. BIC hjälper oss att välja den bästa modellen genom att ta hänsyn till hur bra modellen prestera med våra data och hur många parametrar modellen har. Då finns det inget behov av att dela upp datan eftersom den balanserar passform och komplexitet. Men det är fortfarande viktigt träning, validering och test för att säkerställa att modellen fungerar bra på nya data.
6. Best subset selection är en metod för att välja den bästa uppsättningen variabler i en regressionsmodell. Logaritmen förklara steg för steg *best subset selection.* Först börjar man med nollmodellen, utan prediktorer. Sedan testar man alla möjliga kombinationer av prediktorer, till exempel alla modeller med 1 p (variabel), sen alla med 2p, och så vidare. För varje antal prediktorer, väljer man den modell med lägst RSS eller motsvarande R^2. Till sist väljer man den bästa modellen baserat på ett externt kriterium (BIC, AIC eller justerat R^2).
7. Citatet betyder att alla modeller är fel på något sätt eftersom de inte kan visa hela verkligheten. Men de är användbara för att förstå och förutsäga saker.

6 Självutvärdering

Jag tyckt väldigt mycket om kursen R-programmering! Kunskapskontrollen fick mig att lämna min komfortzon, och jag har lärt mig otroligt mycket. Vi bestämde oss för att arbeta tillsammans med Maria Lagerholm i det här projektet; vi har samarbetat hela vägen, från datainsamlingen till analyserna och modellering. Vi har delat upp arbetsuppgifterna och haft kontinuerlig kontakt, med många möten för att kontrollera koden och jämföra resultat. Vi tog det slutgiltiga modellbeslutet tillsammans. Att arbeta i team har varit väldigt roligt, men också utmanande. Jag har lärt mig mycket av Maria Lagerholm och också av själva processen att göra projektet.

Appendix A

Du hittar koden och detaljerad dokumentation för detta projekt på GitHub:

https://github.com/GeisolUrbina/R_programmering

Källförteckning

- Video:" Linjär Regression" by A. Prgomet – Education Topics Explained https://www.youtube.com/watch?v=NcxMuCG6FS8&list=PLgzaMbMPEHEyLy3NJ8tZqHBzoVcZlowX4&index=2
- SCB Statistisk databasen - Nyregistrerade personbilar, antal efter region, drivmedel och månad https://www.statistikdatabasen.scb.se/pxweb/sv/ssd/START__TK__TK1001__TK1001A/PersBilarDrivMedel/table/tableViewLayout1/
- R for Data Science https://r4ds.had.co.nz/introduction.html
- Blocket https://www.blocket.se/
- Plate recognizer https://platerecognizer.com/
- Transportstyrelsen https://www.transportstyrelsen.se/sv/vagtrafik/fordon/fordons-agaruppgift/
- If försäkring https://www.if.se/privat/partner/nordea/forsakringar/fordon/bil
- R Interface to PXWEB APIs [dataset] by Måns Magnusson https://github.com/rOpenGov/pxweb
- R for Data Science – R Markdown (27- 27.7) https://r4ds.had.co.nz/r-markdown.html
- Data camp - Variance Inflation Factor (VIF): Addressing Multicollinearity in Regression Analysis https://www.datacamp.com/tutorial/variance-inflation-factor?dc_referrer=https%3A%2F%2Fwww.google.com%2F
-