

Innehållsförteckning

[Abstract 4](#_gjdgxs)

[1](#_30j0zll) Introduktion 4

[1.1](#_1fob9te) Objectives 5

[1.2](#_3znysh7) Frågor rapporten ska besvara 5

[2](#_2et92p0) Theory 6

[2.1](#_tyjcwt) Logik 6

[2.2](#_3dy6vkm) Lasso 6

[2.3](#_1t3h5sf) Metriks 7

[2.4](#_4d34og8) Träning, Validera och Test av data 7

[3](#_2s8eyo1) Method 9

[3.1](#_17dp8vu) Skapande av modell 9

[3.1.1](#_3rdcrjn) Insamling av data 9

[3.1.2](#_26in1rg) Hantering av data 10

[3.1.3](#_lnxbz9) Fetaure selection 10

[3.1.4](#_35nkun2) Data leakage 10

[3.1.5](#_44sinio) Val av modell funktion 11

[4](#_2jxsxqh) Resultat och Diskussion 12

[4.1](#_z337ya) Data Exploration 12

[4.2](#_3j2qqm3) Utvärdering av modell 12

[4.2.1](#_1y810tw) Modellens fit 12

[4.2.2](#_2xcytpi) Cross validation för att hitta optimala lambda 14

[4.2.3](#_1ci93xb) Cross validation för att hitta optimala lambda 14

[4.2.4](#_3whwml4) Prediktion med test data 14

[4.3](#_2bn6wsx) Diskussion, svar på frågor. 15

[4.3.1](#_3as4poj) Är en av prediktorerna användbar för att prediktera? 15

[4.3.2](#_49x2ik5) Val av variabler 15

[4.3.3](#_147n2zr) Modellens fit 16

[4.3.4](#_23ckvvd) Prediktioner 16

[5](#_ihv636) Appendix A 17

[5.1](#_1hmsyys) Figur 1: Bilar i trafik april 2024 17

[5.2](#_41mghml) Fordon i trafik och fabrikat producerat per år 17

[5.3](#_vx1227) Personbil bestånd Fabrikat: Volvo, Audi, BMW, Renault, Ford, Toyota 17

[Källa: Trafikanalys/SCB 18](#_3fwokq0)

[5.4](#_1v1yuxt) CV Error från sista modelles träning 18

[5.4.1](#_2u6wntf) Cross validation 18

[5.4.2](#_19c6y18) Cross validation 18

[5.5](#_3tbugp1) Referenser 19

# Abstract

1. Introduktion

Inom dataanalys är både prediktion och statistisk inferens centrala, men de tjänar olika syften och tillämpas på olika sätt. Prediktion, ofta använd inom maskininlärning, fokuserar på att skapa modeller som kan förutsäga framtida händelser eller utfall baserat på existerande data. Målet här är att minimera fel och förbättra modellens förmåga att generalisera till ny, osett data. Statistisk inferens, å andra sidan, strävar efter att dra slutsatser om en större population från ett prov av data och att förstå de underliggande relationerna mellan variabler. Detta innebär ofta att testa hypoteser eller uppskatta populationens parametrar, och det handlar mer om att förstå "varför" något händer snarare än bara att förutsäga "när" det kommer att hända.

Dagens bilhandel är just nu inte så förutsägbar som den var innan Covid, då bilhandeln svängde i takt med bostadsbranschen (Statistiken baseras på ungefär 20 000 försäljningar årligen och mäts med hjälp av BPI (Bilprisindex) av företaget bilpriser.se).

De trender som observerats på den begagnade bilmarknaden under den senaste pandemin och den följande ekonomiska osäkerheten har skiftat läget från att bilpriserna nådde en höjdpunkt för att sedan notera en betydande nedgång i slutet av 2023. KÄLLA Denna fluktuation påverkades av faktorer såsom lagändringar (Nyheter [Riksdagen](#hfutx0i7ks1k)) som berörde bränslepolitiken för bensin- och dieselbilar (Detta för att kontrollera drivmedelspriserna) och förändringar i internationell efterfrågan till följd av den svenska kronans fall ([Artikel Trafikanalys](#28h4qwu), [DI](#nmf14n)).

Dock är detta inget som märks av när man ser till antal registrerade personbilar, data från Statistiska centralbyrån, som ligger runt 4 977 163 sedan år 2020 (appendix: [Personbilar](#2grqrue)).

För de märkena som ingår i modellen så står de för 97% av alla registrerade bilar ([personbilbestånd fabrikat).](#2grqrue)

## Objectives

Det primära syftet är att förstå vad som påverkar prissättningen samt beskriva vilken effekt det blir, enligt min modell, på priset när man väljer att förändra någon variabel, såsom att gå upp ett steg i hästkrafter, eller välja en bil som har kört flera mil, där alla andra variabler hålls konstanta. Denna rapport tar sig an utvecklingen av en regressionsmodell för att förutsäga priser på begagnade bilar som annonseras på Blocket, Sveriges ledande online marknadsplats för begagnade bilar, samt drar slutsatser genom att undersöka värden som modellens noggrannhet och testa hypoteser för att skapa konfidensintervall och prediktionsintervall som ger information.

## Frågor rapporten ska besvara

1. [Finns det någon relation mellan någon av variablerna och priset (Y).](#qsh70q)
2. [Är alla prediktorer viktiga i relation till att förklara y/priset?](#1pxezwc)
3. [Hur väl passar modellen data?](#2p2csry)
4. [Hur bra är modell på att prediktera?](#3o7alnk)
5. Theory
   1. Logik

Bilpriser för begagnade bilar visar en säsongsmässig variation liknande den som ses på fastighetsmarknaden, där priserna tenderar att öka i maj, sjunka under sommaren, och sedan stiga igen i [oktober](#bg4m1kjxzqhk). Priset på en bil påverkas av flera faktorer där varumärket, hästkrafter och växellådstyp är direkt kopplade till produktionskostnaden. Drivmedelstyp påverkar också priset, beroende på bränslekostnadernas fluktuationer, vilket kan sänka bilens attraktivitet på marknaden. Körsträckan är en annan viktig faktor, där en bil generellt minskar i värde med cirka 10% så fort den tas i bruk. Genom att inspektera data kan man se att de dieselbilar som ingår oftast har kört flera mil, vilket är logiskt då de ofta används som företagsbilar för att eventuellt hålla nere kostnaderna för bränsle. Det är rimligt att förvänta sig att detta kan ha en effekt på modellen.

* 1. Lasso

Normer och Förhållandet mellan Koefficienter

Lasso använder ℓ1-normen som straff, där ℓ1-normen av en koefficient β ges av ∥β∥1 = ∑|βj |. Uttrycket ∥βˆL λ ∥1/∥βˆ∥1 handlar om att jämföra storleken på koefficienterna i två olika situationer:

∥βˆL λ ∥1 representerar summan av de absoluta värdena av koefficienterna när vi använder Lasso. Det visar hur stor "effekt" de utvalda variablerna har när vi tillämpar Lasso.

∥βˆ∥1 representerar summan av de absoluta värdena av koefficienterna utan någon regularisering alls (t.ex. när vi bara använder vanlig minsta kvadratmetod) (översatt från ‘An Introduction to Statistical Learning’, 6.2 Shrinkage Methods p. 243).

*bild 1*

*En bild som visar Teckensnitt, handskrift, text, vit

Automatiskt genererad beskrivning*

Förhållandet mellan dessa två (divisionen av den första med den andra) ger oss en uppfattning om hur mycket "inflytande" som har bevarats eller tagits bort genom att använda Lasso. Om förhållandet är nära 1 har Lasso inte tagit bort mycket information. Om förhållandet är lågt betyder det att Lasso har tagit bort mycket av koefficienternas storlek, vilket indikerar en stark variabelreduktion.

* 1. Metriks

De mätvärden som kommer att utvärderas är RMSE, R\_2, L1-straff och Lambda. Koefficienter representerar varje uppskattning den förväntade förändringen i den beroende variabeln (i detta fall log\_pris) för en enhetsförändring (skalad) i den oberoende variabeln, med hänsyn till andra variabler i modellen som inte ändras.

Model Fit:

Multiple R-squared representerar andelen av variansen i den beroende variabeln som förklaras av modellen. Modellen utför följande:

Shrinkage:

Lasso-modeller använder en straffparameter, λ, som kontrollerar storleken på koefficienterna i modellen. När λ ökar pressas fler koefficienter mot noll. Detta 'shrinkage' (krympning) av koefficienterna hjälper till att minska modellens komplexitet och överanpassning.

Sparsity:

Lasso tenderar att producera 'sparse models', det vill säga modeller där flera av koefficienterna är exakt noll. Detta fenomen beror på den absoluta straffningen (L1-straff) som Lasso använder, vilket gör att den är särskilt effektiv för att helt eliminera de minst viktiga variablerna från modellen.

Automatisk Feature Selection:

Eftersom vissa koefficienter ställs in på noll fungerar Lasso-modellen automatiskt som en metod för feature selection. När Lasso-modellen tränas väljer den inte bara koefficientvärden som minimerar förlustfunktionen (som kombinerar dataanpassning och straff) utan också effektivt väljer vilka features som ska inkluderas.

Modellval och Tuning:

I praktiken väljs den optimala storleken på λ vanligtvis genom en process som kallas korsvalidering, där olika värden av λ testas och modellens prediktiva förmåga på valideringsdata utvärderas. Vid skapandet av modellen används 'cv.glmnet' för att automatiskt genomföra denna process. Med detta får man det mest optimala värdet på λ som modellen sedan valideras på.

* 1. Träning, Validera och Test av data

Data delas i 2 olika sektioner, Träning och Test. Validering behövs ej då det används en cross validation med träningssetet.

Det sista testet utförs genom att ta två olika bilar från Blocket som modellen inte har sett tidigare och uppskattar priset på dessa två där alla variabler är de samma, förutom antal körda bil och modellen utvärderas på hur bra den predikterar priset på dessa två bilar då alla andra värden hålls konstanta.

1. Method

Processen följde enligt nedan:

Utforska variabler och samband: Undersökte om det finns något samband mellan variabler och priset. Genom att använda metoder som korrelationsanalys eller visualisering identifieras potentiella samband och initiala hypoteser.

Val av regressionsmodell: Val av lämplig regressionsmodell

Variabelval: Lasso-modellen väljer det, men de analyseras.

Modellanpassning och utvärdering: Anpassa modell, genom crosss-validaton .

Bedömning av modellens prestanda: Utvärdera modellens prestanda genom att titta på metriska som RMSE (Root Mean Squared Error) och R^2 (Coefficient of Determination). RMSE ger en uppfattning om hur väl modellen passar data, medan R^2 indikerar andelen varians i priset som förklaras av modellen.

Validering av modellen: Validering av regressionsmodell genom att testa den på oberoende dataset. Detta säkerställer att modellen generaliserar väl och är inte överanpassad till träningsdata.

Tolka resultat och slutsatser: Slutligen, tolkning av regressionsanalys och dra slutsatser om sambandet mellan variabler och priset. Identifiering vilka variabler som signifikant påverkar priset och hur mycket variation de förklarar.

* 1. Skapande av modell
     1. Insamling av data

Insamling av data skedde i grupp av 5 studenter som fick varsitt märke. Insamling gick till genom att ”skrapa” Blocket bilar med förinställd querystring, dvs <https://www.blocket.se/bilar/sok?filter=%7B%22key%22%3A%22car.details.makeModel.make%22%2C%22type%22%3A%22nested%22%2C%22values%22%3A%5B%22Toyota%22%5D%7D&filter=%7B%22key%22%3A%22car.details.vehicleModelYear%22%2C%22type%22%3A%22int%22%2C%22range%22%3A%7B%22start%22%3A%222014%22%2C%22end%22%3A%222024%22%7D%7D&filter=%7B%22key%22%3A%22deal.price%22%2C%22type%22%3A%22int%22%2C%22range%22%3A%7B%22start%22%3A%22150000%22%2C%22end%22%3A%22500000%22%7D%7D&filter=%7B%22key%22%3A%22car.details.chassi%22%2C%22type%22%3A%22string%22%2C%22values%22%3A%5B%22Cab%22%2C%22Coup%C3%A9%22%2C%22Halvkombi%22%2C%22Kombi%22%2C%22SUV%22%2C%22Sedan%22%5D%7D>

Där de valda parametrarna är med, så som år från 2014-2024, pris från 150 000-500 000, alla typer av fordon utan yrkesfordon och familebuss. Då vi hade bestämt ett antal på 50 var först och tog beslutet att fylla på, så gjorde man här lite olika där vi flesta fyllde på med 200 till, men Audi blev 3 gånger så många, så för att det inte skulle bli fel pga. insamling av data så tog jag bort slumpmässigt bort rader, för att hålla mig till de beslut som tagits.   
Fördelningen ser ut som nedan.   
Medelpriset är ganska jämt fördelat där Volvo och Ford sticker ut något både i form av medelpris och Ford för standardavvikelse.  
Ev. dubbletter tas sedan bort i R studio, vilket resulterade i

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |

* + 1. Hantering av data

Skapande av faktorer och dummies:   
Kategoriska variabler som märke, drivmedel, växellåda, och biltyp omvandlas till numeriska värden för att kunna inkluderas i modellen. Detta görs genom att skapa dummyvariabler där varje kategori representeras av en egen variabel i modellen. Till exempel kan "drivmedel" som har kategorierna "bensin", "diesel", och "el" delas upp i tre dummyvariabler, som drivmedelbensin.

Skalning och log-formatering av pris:   
För att normalisera fördelningen av data och hantera extrema värden genomfördes en Log-transformering av pris.

Skalning av kontinuerliga variabler:  
 År, hästkrafter och miltal: Skalning av dessa variabler görs för att undvika problem med variabler som varierar över olika storleksordningar och få en gemensam skala och få lika stor vikt.

* + 1. Fetaure selection

[Se 3.1.5 val av modell.](#1ksv4uv)

* + 1. Data leakage

Data leakage inträffar när information från testdata oavsiktligt används för att träna modellen, vilket kan leda till överoptimistiska prestandaestimater och modeller som inte presterar lika bra på ny, osedd data.

För att undvika detta så delas data upp först i tränings-, validerings- och testset. Sen används information som medelvärde och standardavvikelse för att skala träningsdatan via funktionen scale\_vars, som är en anpassad R-funktion som används för att standardisera (skala) numeriska variabler i ett dataset.   
 Därefter används samma parametrar för att skala validerings- och testdatat.

Träningsdata: Används för att både träna modellen och, genom korsvalideringsprocessen, för att validera olika inställningar av modellparametrar (som lambda i Lasso). cv.glmnet utför detta genom att internt dela upp träningsdatan i flera delar, varav en del används som en "intern validerings set" under varje iteration av korsvalideringen. Detta innebär att varje subset av träningsdatan används både för träning och validering i olika omgångar av korsvalideringsprocessen.

Testdata: Används för att utvärdera modellens slutliga prestanda efter att modellen har tränats och validerats på träningsdatan. Testdata bör hållas helt separat och endast användas en gång, efter att alla beslut om modellinställningar har gjorts, för att få en opartisk uppskattning av modellens prestanda på ny, osedd data-

Slutgiltiga testet med 2 annonser från blocket :  
Används de skalningsparametrar från träningsdatan när den nya data förbereds för prediktion, vilka sparades i första steget under träningen.

* + 1. Val av modell funktion

Genom att ha testat ett flertal modeller så som lm(), med alla variabler, samt interaktions-variabler så var Glmnet Lasso(Least Absolute Shrinkage and Selection Operator) regression den som var mest robust i form av score men också att prediktera helt ny data från blocket.

Lasso hjälper till att effektivisera modellen genom att utföra både variabelurval och regularisering. Genom att införa en L1-penalty (absolutvärden av koefficienterna) tvingas vissa modellkoefficienter att vara exakt noll. Detta innebär att Lasso inte bara hjälper till att minimera överanpassning (overfitting) utan också identifierar de mest relevanta prediktorerna (variablerna) genom att utesluta irrelevanta variabler från den slutliga modellen.

Hantering av Multikollinearitet: Har inte observerat start multikollinearitet, låg ett flertal värden på ca 0.6, och Lasso hantera situationer där det finns korrelationer mellan prediktorerna.   
Modellår och Pris visade inte ge effekt som en interaktionsvariabel.

En bild som visar text, skärmbild, Teckensnitt, linje

Automatiskt genererad beskrivning

Förbättrad Prediktiv Prestanda: Du nämnde att Lasso predikterade bättre på helt ny data jämfört med andra modeller. Detta är en stark indikation på att Lasso har varit effektivt för att undvika överanpassning och har gett en modell som är generaliserbar till ny, osedd data. Detta är kritiskt för praktiska tillämpningar där modellen kommer att användas för att göra förutsägelser i "verkliga världen".

1. Resultat och Diskussion
   1. Data Exploration

Av att plotta residualer så granskades en del rader i datasetet som stack ut som extrema eller intressanta. Extrema värden var ouliers vara vissa fick vara kvar då det var el del bilar som stack iväg i pris av naturliga anledningar som nya el-bilar som hade ett lågt miltal. Dessa kunde man också se i cooks distance i form av grupperingar till höger i grafen

En bild som visar text, skärmbild, linje, diagram

Automatiskt genererad beskrivning

En del intressanta punkter dök upp med som att två bilar med samma pris, men andra variabler stämde inte, så att miltal var stor skillnad. Dessa annonser var borttagna från Blocket när det inspekterade och kan då vara ett fel i uppladdning av annonser, så alla dessa värden togs bort.

* 1. Utvärdering av modell
     1. Modellens fit

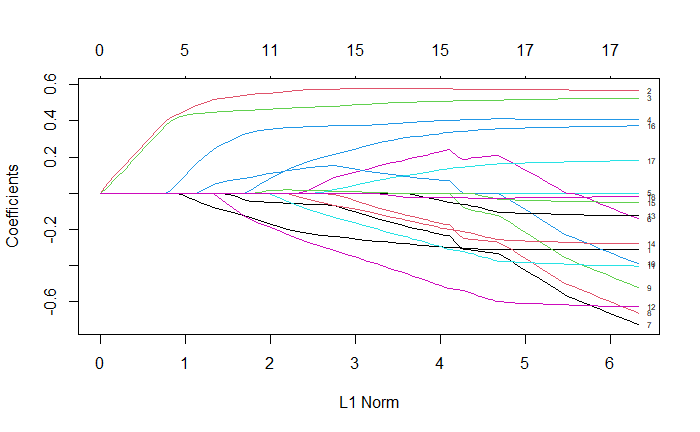
Vid första anpassning av modellen med träningsdatan så ser man tydligt hur de olika prediktorerna ändras beroende på straff, λ, som ges.   
Grafen nedan visar hur olika värden på λ påverkar modellens koefficienter.

Koefficientens storlek: Varje linje representerar en variabel (prediktor) i modell, och höjden på linjen vid varje punkt representerar storleken (och tecknet) på den variabelns koefficient. Koefficientens värde ändras som en funktion av λ.

Koefficientbanor, från höger till vänster: Linjerna i diagrammet (koefficientbanorna). När λ ökar (och L1-normen minskar), börjar vissa linjer (koefficienter) plana ut vid noll, vilket indikerar att dessa variabler har tagits bort från modellen av Lasso på grund av deras relativt lägre betydelse eller bidrag till modellen.

Koefficienter vars banor snabbt går mot noll när λ ökar är mindre viktiga för modellens prediktiva förmåga, enligt Lasso

Variabler som förblir icke-noll: Variabler vars koefficientbanor förblir icke-noll även vid högre värden på λ är de som Lasso identifierar som viktiga för att förutsäga målvariabeln. Dessa är de kraftfullaste prediktorerna enligt modellen.



*graf 1*

**Ej Betydelsefulla predatorer**

Siffra 5:Växellådamanuell, 6:Biltyp\_Cupe,13: märketoyota, 15:märkevolvo, 18: Bränslemiljöbränsle/hybrid startar nära =0 och håller sig där när λ ökar indikerar att dessa prediktorer inte har någon betydande effekt på modellen. Dessa variabler är inte relevanta för att förutsäga målvariabeln under de nuvarande modellinställningarna och med given data. Lasso-straffet prioriterar inte dessa variabler alls, vilket antyder att de antingen är överflödiga eller inte tillför något prediktivt värde till modellen.

**Linjer som start med höga eller låga värden och går emot noll:**

Siffra 7:Biltyphalvkombi, 8:Biltypkombi, 9:Biltypsedan, 10:BiltypSUV, 11:märkeford, 17: bränsleel. Variabler vars koefficientbanor startar vid höga eller låga värden och sedan går mot noll (men inte nödvändigtvis når noll) när λ ökar visar att dessa variabler initialt har haft en stark påverkan på modellen. Deras stora koefficienter indikerar signifikant inflytande. Dock, som λ ökar och mer straff appliceras, minskar deras relativa betydelse.

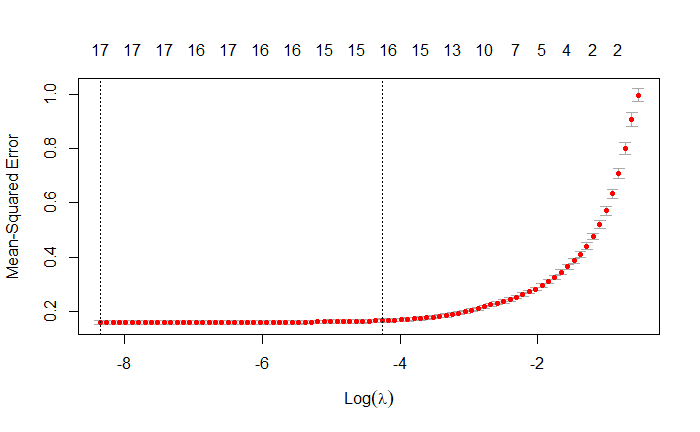
Och lite längre fram kommer, 12:märkereanult, 1:Miltal, 4:Växellådaautomat.

**Linjer som Kommer till Noll Längst till Vänster**

Linjer som närmar sig noll och förblir vid noll vid högre λ värden (längst till vänster på x-axeln) representerar variabler som kan bidra till modellen endast när det finns lite eller ingen regularisering. Desa predictor är hästkrafter och modellår.

Plot av CV

* + 1. Cross validation för att hitta optimala lambda



*graf 2*

Låg MSE: vid låg λ (-8 till -4): Om MSE är låg när λ är relativt litet (dvs. vid -8 till -4 på x-axeln), betyder det att modellen presterar väl med mindre straff. Modellen inkluderar fler variabler och kan fånga komplexa mönster i data.

MSE börjar öka vid -4 till -2: När λ ökar (rör sig från -4 till -2) och MSE börjar stiga, indikerar det att modellen börjar förlora sin förmåga att förutsäga korrekt. Detta sker eftersom högre λ tvingar fler koefficienter mot noll ([se tidigare graf](#4i7ojhp)), vilket potentiellt kan eliminera några viktiga variabler från modellen. Det kan leda till att modellen blir för enkel och inte kan fånga de nödvändiga mönstren i data, vilket resulterar i högre förutsägningsfel.

Kraftig ökning av MSE vid högre λ (nära -2): När λ är mycket högt, och MSE ökar kraftigt, innebär det att modellen har blivit för förenklad.

Optimala Lambda hamnar på

Optimalt Lambda: 0.0002825712

Vilket är en låg siffra, dvs högt L1 och de flesta prediktorer kommer att vara med i modellen. Deta värdet används för att träna den sista modellen.  
Koeffcienterna [listas här](#4f1mdlm).

* + 1. Cross validation för att hitta optimala lambda

Med det optimala Lambda så tränas modellen igen med all träningsdata.Koefficenterna skiljer sig en del åt, men båda testas i prediktion och valde att gå vidare med den senaste varianten som tränats med det optimala lambda.

* + 1. Prediktion med testdata

Score:

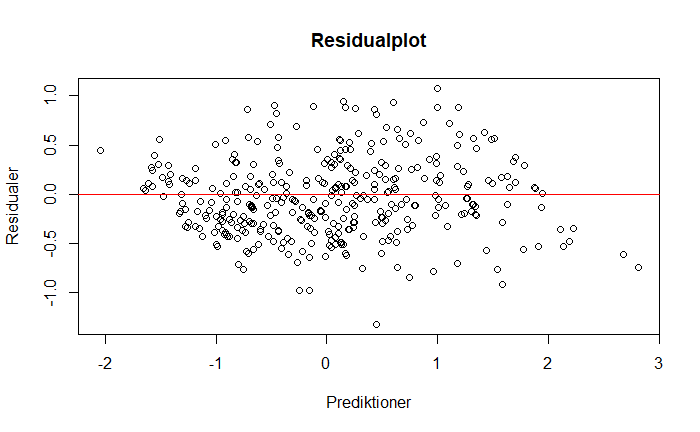
RMSE Rsquared MAE

0.3962144 0.8440479 0.3186498

RMSE: Ligger på 0.39 vilket är lågt och   
Rsuared: 84.4% av variansen i den beroende variabeln (priset) kan förklaras av oberoende variabler i din modell. Det värdet antyder att modellen fångar en stor del av datans beteende.

Residualer: Ser random ut. De linjer som blir till höger och vänster i sluttande riktning är på grund av att vi satte strikta gränser kring min och max pris.

QQ-plot: Punkterna följer en rät linje, vilket indikerar att residualerna är normalfördelade. Detta är ett tecken på att modellen är väl anpassad och att de statistiska testerna som förutsätter normalitet är giltiga. Man kan se en viss högerställd svans, vilket betyder att vi har färre höga priser, och vi ser också att det går lite över linjen för små värden, vilket visar att vi inte har så många extrema låga värden. Fördelningen är skiftad till vänster, men det justerades då vi skalade numeriska värden.

Dessa mätvärden tillsammans tyder på att modellen är tillförlitlig och presterar väl övergripande.  
En bild som visar text, diagram, linje, Graf

Automatiskt genererad beskrivning

*graf 3*

*graf 4*

* 1. Diskussion, svar på frågor.

* + 1. Är en av prediktorerna användbar för att prediktera?

Ja, det ser vi i [plotten av L1](#4i7ojhp) och koefficienterna, där hästkrafter och årsmodell är viktiga i modellen, men även miltal och växellådaautomat.

Detta är naturligt då hästkrafter och årsmodell är något som påverkar priset positivt, då flera hästkrafter är kraftigare motor, samt automatväxel och är dyrare att producera. Detta krävs ingen modell att för att kunna avgöra, utan detta tillhör teorin kring bilens värde som många spekulerar i.

* + 1. Val av variabler

Eftersom modellen i den här rapporten skaps via Lasso så kan vi inte titta på F-värden som man vanligtvis gör (BOKEN p 89), utan detta görs automatisk. Men både kring teorin samt ovan nämnda plott så kan man se vilka variabler som är viktiga. I detta fall då man fick ett lågt lambda så anser modellen att ganska många av variablerna är viktiga att ha med.

Jag valde att ta ett sidospår här och jämföra med de första modellerna jag skapade som jag inte tar upp i den här rapporten och det kan nämnas då att för den modellen med 3 prediktorer (Miltal, Hästkrafter, Modellår), som skapades med R’s lm() och en annan modell med samma prediktorer som Lasso-modellen, så fick den sistnämnda bättre AIC (1424, jämför med 2154), vilket styrker Lasso-modellens val med sitt låga Lambda att många prediktorer krävs just för den här modellen. De flesta fick \* stjärnor förutom Bränslemiljöbränsle/hybrid, biltypcoupe och MärkeVolvo, som fick noll.

Den modellen med fler parametrar hade också lägre residuala standardfel och högre R-värdena (både Multiple och Adjusted R-squared), vilket indikerar att den mer förklarar en större del av variansen i datan jämfört med den enklare modellen.

Rent teoretiskt hade jag förväntat mig att just den bränsletypen skulle få större betydelse, samt Volvo, för att de är mer populära val.

När jag tittade närmare på Miljömärke och hur ofta det förekom med Toytota, genomförde ett Chi-test så fick jag p-värde på 2,22\*10^ -128, vilket nära 0, så visar det sig att det finns ett starkt statistiskt samband mellan dessa variabler, vilket inte är så förvånande då Toyota har länge satsat på den typ av bränsle. Detta kan vara en anledning.

Tittade också på volvo och bränsle och såg att Volvo har hög andel dieseldrivna bilar. Just att volvo som är känt för sin säkerhet inte har fler av ex el eller hybrid är något som är intressant och man hade kunnat undersöka närmare. Kanske ser inte populationen ut som vårt dataset.

Får ingen specill indikation när jag testar lm-modellen med hjälp av VIF(), då alla hamnar under 1.8.

Så, med detta har jag förrtått att de flesta prediktorer är viktiga för modellen.

#### 

|  | **Inte Hybrid** | **Hybrid** |
| --- | --- | --- |
| Inte Toyota | 1536 | 273 |
| Toyota | 76 | 258 |

* + 1. Modellens fit

Då lambda var relativt lågt så finns det en risk för att överanpassa till träningsdatan, men Två av de vanligaste värdena man tittar på här är RMSE och R2. Då båda modellerna, fit och cv\_lasso ger liknande värden så tolkas det som att det är en stabil modell och att den kan förutsäga ny, osedd data väl.

RMSE (fit): 0.3961069

R-squared (fit): 0.8426741

RMSE (cv\_lasso): 0.3961069

R-squared (cv\_lasso): 0.8426741

* + 1. Prediktioner

Jag hade velat plotta prediktionsintervall och confidenseintervall men det var en utmaning iom lasso-modellen här krävs mer kunskap kring R och Bootstrap.

Nya annonser lades in, där alla parametrar var samma förutom miltal. Då modellen inte tar ställning till Utrustning så tas detta inte upp och kan ju påverka priset i annonsen som modellen inte kan ta ställning till.

Den modell som testades var: Toytoa Corolla 1.8 Hybrid 5dr 122hk, som var värden som var konstanta medan miltal ändras.

En bild som visar text, skärmbild, linje, Teckensnitt

Automatiskt genererad beskrivning

En bild som visar text, Teckensnitt, skärmbild

Automatiskt genererad beskrivning

Data skalades och hanterades på samma sätt som träning och testdata och testades i modellen vilket gav priset: 249 993,1. Något över det utsatta priset, men den här bilen är ute på Gotland kan påverka.

Nästa bil har gått 5050 och predikteras till priset 207 720 i stf 199 900.

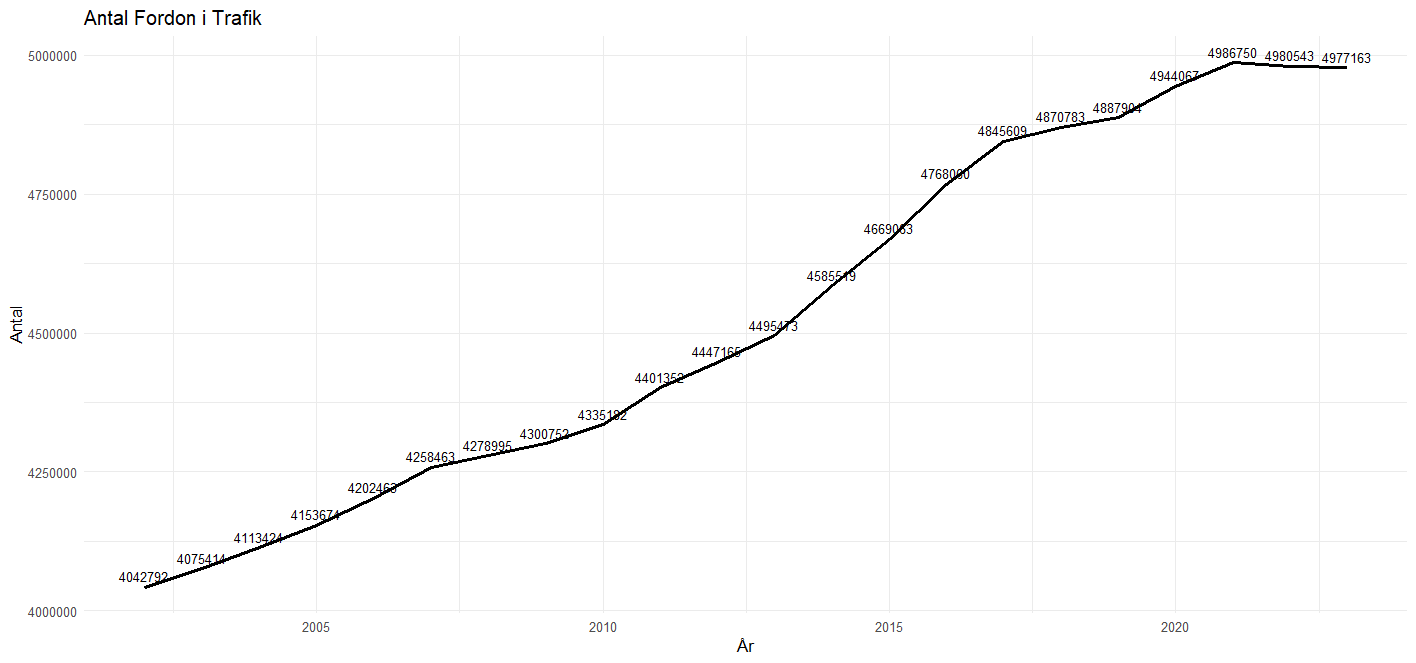
Väljer man att ändra på årsmodell så är det svårare då miltal generellt är olika och måste justeras efter et.   
Procent fel ligger på ca 9-10% och kan förklaras av modellen eller att vi inte tagit ställning till extra utrustningen.

Då pris har ett medelvärde på 7600 mil och en standardavvikelse 5288, skalad coefficient på

-0.31, så blir det rent matematiskt : −0.31054108×5371.54 mil minskning per mil, vilket blir 1655 kr värdminskning per mil. Detta är inte riktigt sanningen då man vet att värdeminskning från helt ny till till att man använder den är mycket kraftigare, dvs är förmodligen inte helt linjär från start.

1. Appendix A
   1. https://www.riksdagen.se/sv/aktuellt/aktuelltnotiser/2023/nov/23/utskott-sager-ja-till-sankt-reduktionsplikt-for\_cms7f419a62-c5fb-46e4-bd92-50109ac0cb75sv/

* 1. Figur 1: Bilar i trafik april 2024



*Figur 1 KÄLLA: SCB FORDON I TRAFIK (API)*

*https://api.scb.se/OV0104/v1/doris/sv/ssd/START/TK/TK1001/TK1001A/FordonTrafik*

* 1. Fordon i trafik och fabrikat producerat per år

*Figur 2 Källa:SCB + Mobilty Sweden, på uppdrag av SCB (https://mobilitysweden.se/statistik/Fordonsbestand)*

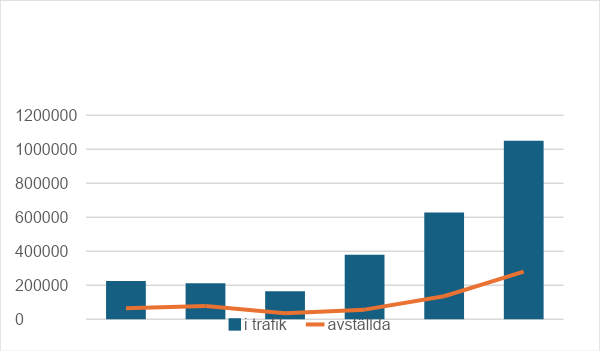
En bild som visar text, diagram, skärmbild, Graf

Automatiskt genererad beskrivning

* 1. Personbil bestånd Fabrikat: Volvo, Audi, BMW, Renault, Ford, Toyota

Tabell 1

Källa: Trafikanalys/SCB



* 1. CV Error från sista modelles träning

* + 1. Cross validation

[1] "CV Error: 0.158279557609279"

19 x 1 sparse Matrix of class "dgCMatrix"

s1

(Intercept) -2.713549e-01

Miltal -2.843089e-01

Hästkrafter 5.801080e-01

Modellår 5.040473e-01

Växellådaautomat 3.922674e-01

Växellådamanuell -8.456802e-15

Biltypcoupé 1.947156e-01

Biltyphalvkombi -1.930763e-01

Biltypkombi -1.317408e-01

Biltypsedan .

Biltypsuv 8.733832e-02

Märkeford -2.540683e-01

Märkerenault -4.589468e-01

Märketoyota -9.789312e-03

Märkevolkswagen -1.583137e-01

Märkevolvo .

Bränslediesel 3.115135e-01

Bränsleel 1.034856e-01

Bränslemiljöbränsle/hybrid -2.368552e-02

19 x 1 sparse Matrix of class "dgCMatrix"

s0

(Intercept) 0.26805097

Miltal -0.31054108

Hästkrafter 0.56974972

* + 1. Cross validation

19 x 1 sparse Matrix of class "dgCMatrix"

s0

(Intercept) 0.26805097

Miltal -0.31054108

Hästkrafter 0.56974972

Modellår 0.52558039

Växellådaautomat 0.40677006

Växellådamanuell .

Biltypcoupé -0.12200677

Biltyphalvkombi -0.70711916

Biltypkombi -0.64346639

Biltypsedan -0.50212242

Biltypsuv -0.37103843

Märkeford -0.40295463

Märkerenault -0.62831022

Märketoyota -0.12487540

Märkevolkswagen -0.28010525

Märkevolvo -0.05192801

Bränslediesel 0.37198833

Bränsleel 0.17978372

Bränslemiljöbränsle/hybrid -0.01964774

Sålda begagnade bilar  
  
En bild som visar text, diagram, Teckensnitt, Graf

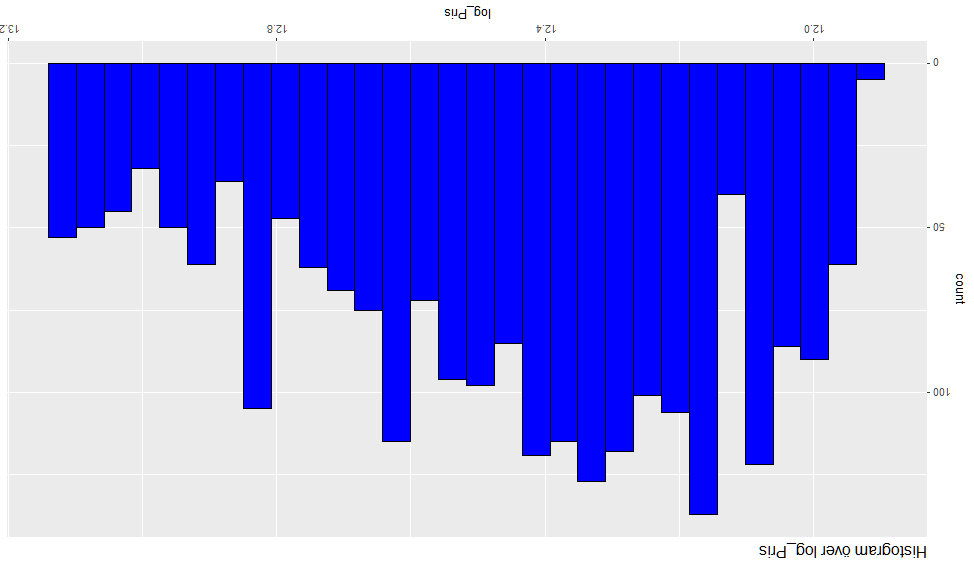
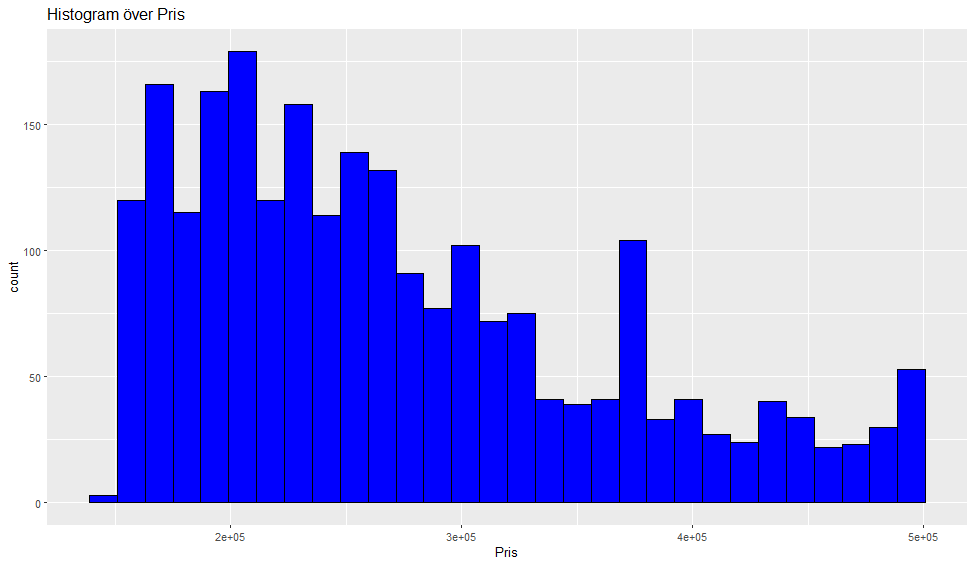
Automatiskt genererad beskrivning

* 1. Referenser

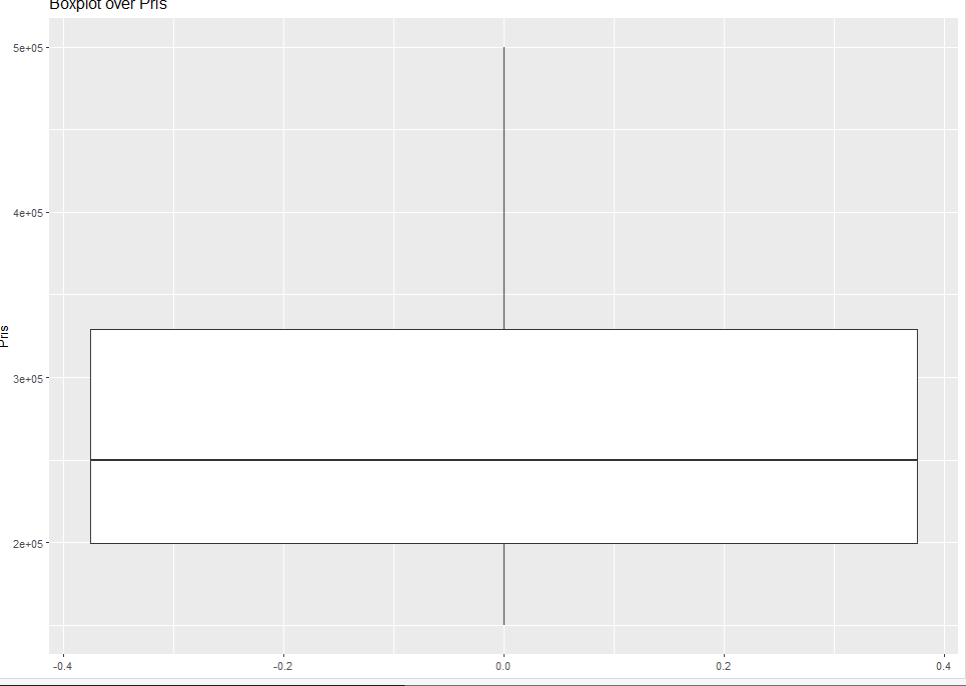
1. Sänkt reduktionsplit för bilar <https://www.riksdagen.se/sv/aktuellt/aktuelltnotiser/2023/nov/23/utskott-sager-ja-till-sankt-reduktionsplikt-for_cms7f419a62-c5fb-46e4-bd92-50109ac0cb75sv/>
2. Billigare bränsle, dieselpriser stiger: <https://www.di.se/nyheter/billigare-bransle-da-stiger-priset-pa-dieselbilarna/>
3. Ökad export pga kronas värdeminskning: <https://www.trafa.se/vagtrafik/rekordar-for-exporten-av-begagnade-personbilar-14513/>).

1. Artikeln från Trafikanalys belyser att 2023 var ett rekordår för exporten av begagnade bilar från Sverige, med omkring 150 000 fordon exporterade, en ökning med 39% från 2022. Majoriteten av de exporterade bilarna var bensin- eller dieseldrivna, och många av dem var relativt nya (0–5 år gamla). En viktig faktor till denna ökning är den låga värderingen av svenska kronan, vilket gjorde svenska bilar mer överkomliga för utländska köpare. Det var även en ökning av exporten av el- och hybridbilar. För mer detaljer kan du läsa hela artikeln på [Trafikanalys' hemsida](<https://www.trafa.se/vagtrafik/rekordar-for-exporten-av-begagnade-personbilar-14513/>).

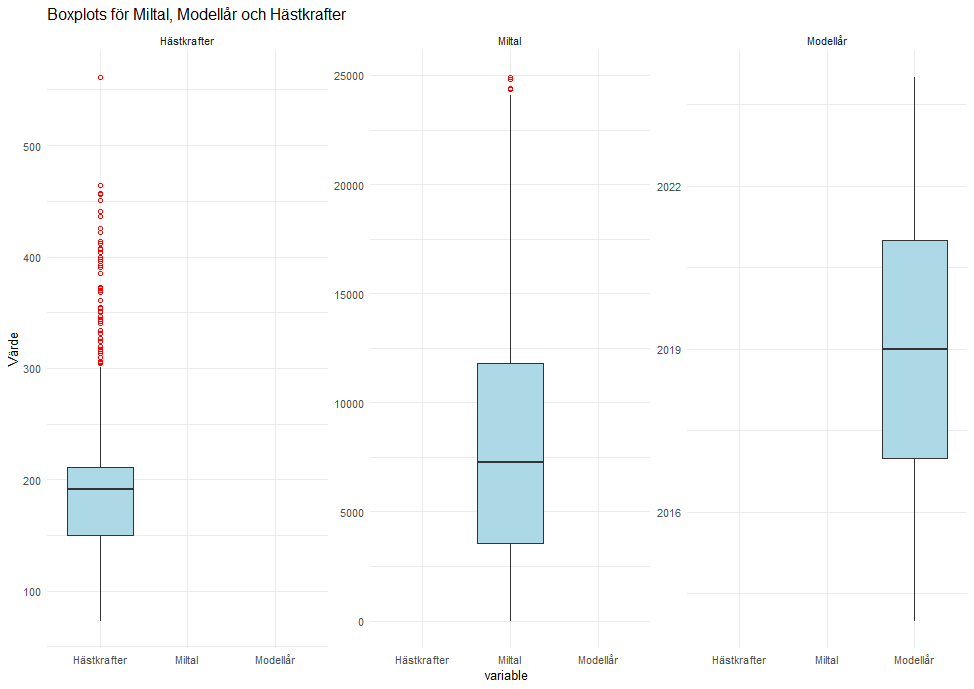
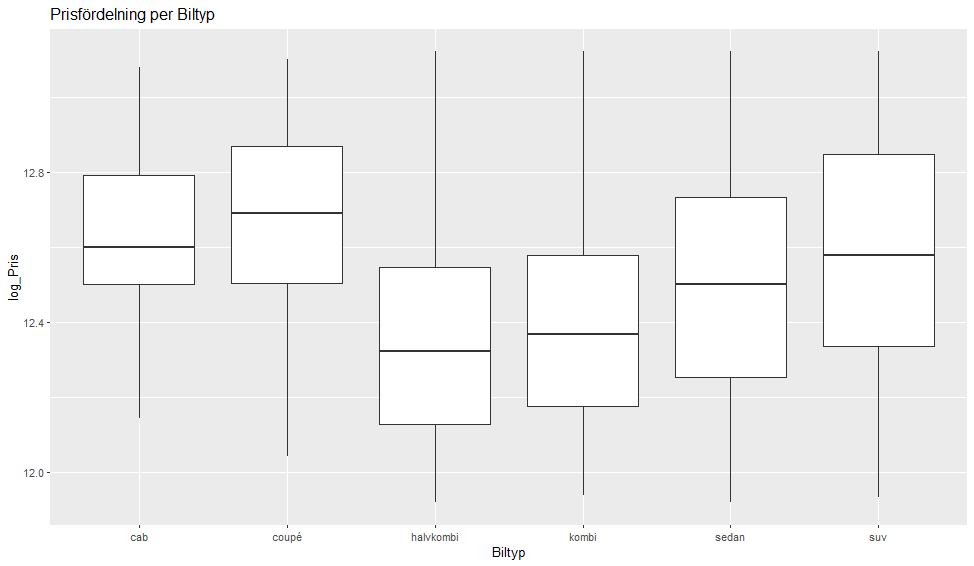
1. Man skriver också på Dagens Industri att dieselpriserna går upp då man infört slopad reduktionsplikt och ett mindre utbud på marknaden (av min undersökning är det ca --% som är tillsalu, vilket bekräftar DI) <https://www.di.se/nyheter/revansch-for-dieselbilen-priserna-stiger-huggsexa/>



Plottar alla Pris och variabler.

 En bild som visar skärmbild, text, Rektangel, linje

Automatiskt genererad beskrivning

  
En bild som visar text, skärmbild, diagram, Rektangel

Automatiskt genererad beskrivning