Regressions Analys



Jesper Kokkonen

EC Utbildning

Kunskapskontroll, R Programmering

202404

# Abstract

Denna studie utforskars sambandet mellan prissättning på andrahandsbilar i Sverige med Bilmärke, Bilmodel, Modelår, Diesel/Bensin, Miltal, Automat/Manuell, Region och Företag med tillämpning av regressionsanalys. Genom att använda data insamlad från Blocket.se och extern information om antalet personbilar i trafik från SCB.

Innehållsförteckning

[Abstract 2](#_Toc165030870)

[1 Inledning 1](#_Toc165030871)

[2 Teori 2](#_Toc165030872)

[2.1 Modeller och dataset 2](#_Toc165030873)

[Linear Model 2](#_Toc165030874)

[Lasso Regression 2](#_Toc165030875)

[Dataset 2](#_Toc165030876)

[3 Metod 3](#_Toc165030877)

[3.1 Databas och korrelation 3](#_Toc165030878)

[3.2 Val av modell 3](#_Toc165030879)

[4 Resultat och Diskussion 4](#_Toc165030880)

[5 Slutsatser 5](#_Toc165030881)

[6 Teoretiska frågor 6](#_Toc165030882)

[7 Självutvärdering 8](#_Toc165030883)

[Appendix A 9](#_Toc165030884)

[Källförteckning 10](#_Toc165030885)

# Inledning

Rapportens fokus kommer vara kring regression analys av andrahandsbilmarknaden i Sverige. Där en regressionsmodellering på insamlad data görs, data samlas in via dataskrappning från blocket.se och extern data för fordon i trafik från SCB hämtas in manuellt.

# Teori

## Modeller och dataset

## Linear Model

Lasso-regression är en klassificeringsalgoritm som använder shrinkage i enkla och sparsamma modeller. I Shrinkage krymps datavärden mot en central punkt som medelvärdet. Lasso-regression är en regulariserad regressionsalgoritm som utför regularisering som lägger till ett straff som är lika med det absoluta värdet av koefficienternas storlek. (geeksforgeeks.org, 2023)

## Lasso Regression

Linjär regression är en statistisk metod för att modellera sambandet mellan en beroende variabel och en given uppsättning oberoende variabler. Det förutspås att en rät linje kan användas för att approximera förhållandet. (geeksforgeeks.org, 2023)

## Dataset

Datasetet för fordon i trafik mellan 2014-2023 är inhämtat från SCB:s hemsida. Datasetet för andrahandsbilar samlades in via webscraping och innehåller 7 100 celler och 9 kolumner.

Datasetets attribut är som följer:

"CarName\_Brand" "CarName\_Model" "CarName\_ModelYear" "CarName\_Engine" "CarName\_Miles" "CarName\_gears" "CarName\_Price" "CarName\_Region" "CarName\_Dealer"

# Metod

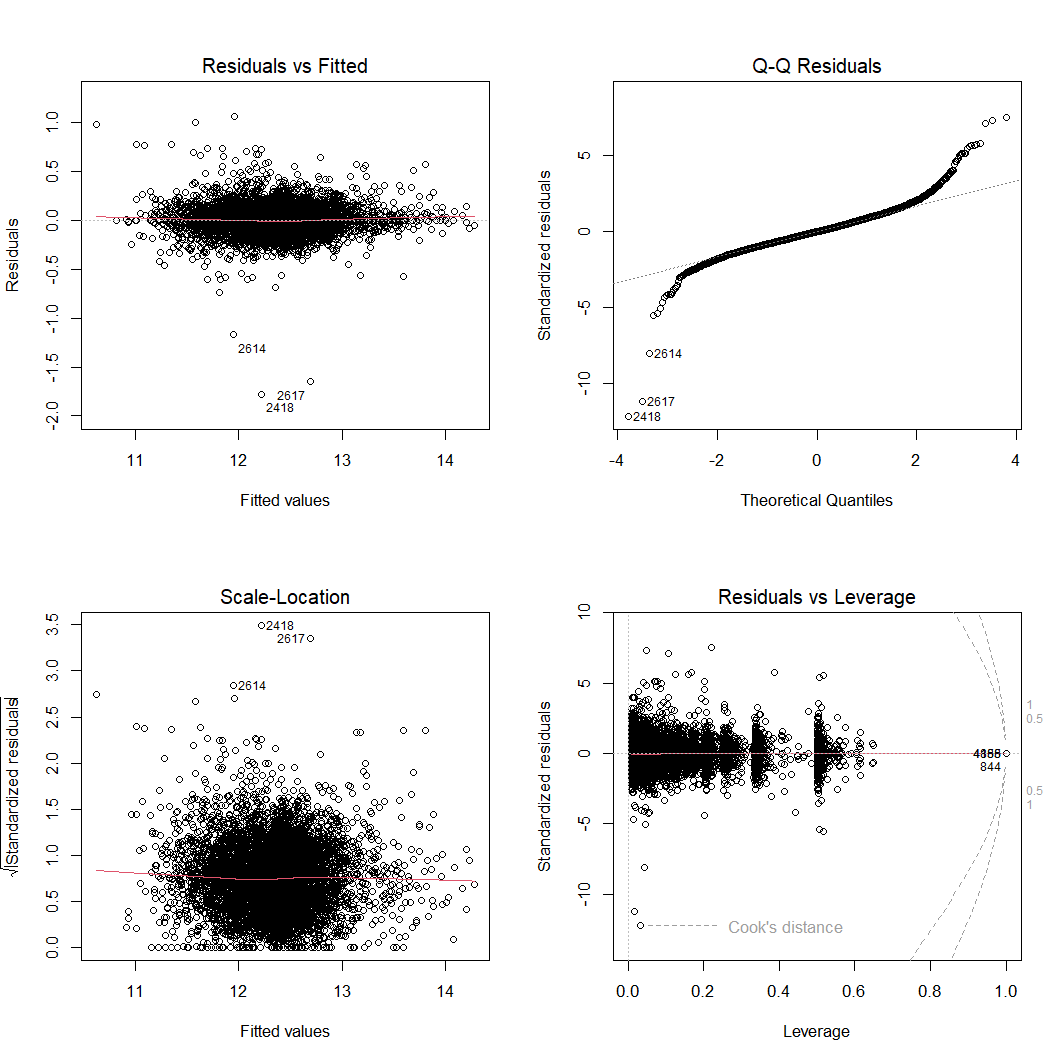
## Databas och korrelation

Statistiska Centralbyrån tillhandahåller statistik inom en rad olika ämnen samt årtal som är tillgänglig för alla att ta del av. Data för fordon i trafik i hela riket från 2014–2023 laddades ner. Data för fordon samlades in via webscraping, därefter filtrerades diverse fel som uppstod av webscrapningen. Efter insamlade data sökte jag efter korrelationer mellan olika kolumner för att se om någon av dem skulle kunna kombineras.

## Val av modell

Jag valde att undersöka tre olika modeller Lasso Regression och två olika Linear Model. Jag började först med att använda all data i en Linear Model därefter tänkte jag vad som skulle vara logiskt kring vilka kolumner skulle ge det bästa resultatet. Efter Linear Model gjordes ett försök till att använda Lasso Regression.

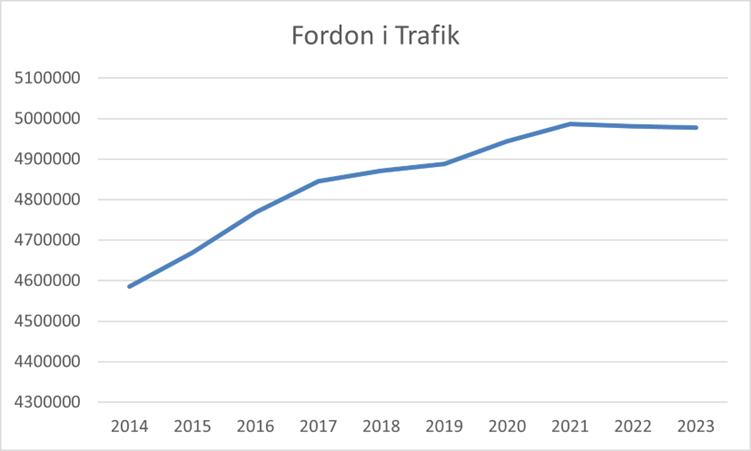
Modellen som valdes var Linear Model med all data inkluderad.



Diagnostik plot för vald modell visar inga större avvikelser

# Resultat och Diskussion

Enligt statistik från SCB har fordon i trafik ökat med ca 400 000 fordon mellan 2014-2023 med detta i åtanke och det rådande världsläget med hög inflation och högre räntor till följd av den hög inflationen så antar jag att fordons priserna för begagnade bilar har haft en ökning det senaste året.



När det kommer till korrelation så har korrelationskoefficienten mellan "CarName\_ModelYear" och "CarName\_Miles" är den ungefär -0,495. Detta betyder att när "CarName\_ModelYear" ökar (vilket indikerar äldre bilar), tenderar "CarName\_Miles" att minska (vilket indikerar färre körda mil). Korrelation innebär inte alltid en orsakssamband, men om vi antar att detta är korrekt så skulle detta kan en av förklarningarna vara att det nya bilar som sälja på blocket har används flitigt under bara några få år jämför med äldre bilar kan ha stått parkerat och inte har används lika flitigt.

Resultaten av Linear Model med alla valda kolumner fick en adjusted R^2 på 0.88, Linear Model med logiskt valda kolumner fick en adjusted R^2 på 0.57 och Lasso Regression fick en adjusted R^2 på 0.44. Valet blev därav Linear Model med alla valda kolumner. När det kom till att validera det olika modelleran på validerings data så upp stod det ett problem där jag inte lyckade att implementera min Lasso Regression, men eftersom Lasso Regressionen inte hade någon större R^2 värde antog jag att den ändå inte skulle blivit den bästa modellen.

# Slutsatser

Baserat på regressionsanalys av datasetet av andrahandsbilar har ålder, antal körda mil, manuell/automat, bilmärke samt modell en stark prediktiv förmåga på bilpriset. Region och bensin/ diesel bilar hade inte någon signifikant påverkan på bilpriser.

# Teoretiska frågor

**Vilken typ av fordon vill vi modellera? Exempelvis kan det vara problematiskt om hälften är exklusiva bilar såsom Ferrari och andra hälften vanliga bilar såsom Mazda.**

Vi har valt 8 märken: Kia, Volkswagen, Volvo, Audi, BMW, Peugeot, Opel & Toyota

**Säkerställ att datan ni samlar in går att läsa in i R och att det blir som ni tänker er. Gör alltså en ”Proof of Concept” (POC).**

POC kommer göras på ett urval från alla märken

**Vilken typ av data skall vi samla in?**

Vi har valt att avgränsa oss till

Biltyp- Halvkombi, Kombi, Suv, Sedan

Modellår- 2014 - 2024

Märke- Kia, Volkswagen, Volvo, Audi, BMW, Peugeot, Opel & Toyota

Minsta pris - 20 000kr

Säljare - Privat & Företag

Växellåda- Manuell  & Automat

Drivmedel- Bensin, & Diesel

**Hur skall vi samla in datan på ett konsistent sett i gruppen?**

En mall för gruppen att samla in data för POC

A screenshot of a computer

Description automatically generated

**Kan man göra några kontroller så datan är ”rimlig”?**

Genom att genomföra explorativ dataanalys för att identifiera outliners och felaktigheter. Eller exempelvis jämföra med SCB för att säkerställa att datan är realistiska

**Hur mycket data skall vi samla in?**

Vi har samlat in 7100 bilar hittills

(1) Teoretiska frågor Besvara följande teoretiska 7 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.

**Quantile-Quantile (QQ) plot är ett grafiskt verktyg som används i statistik för att jämföra fördelningen av en provdatauppsättning med en känd teoretisk fördelning, så som en normalfördelning.**

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?

**Regressionsanalys gör det möjligt för oss att förklara relationer med hjälp av en linjär modell och belysa sambanden mellan oberoende variabler, såväl som effekten av varje oberoende variabel på en specificerad konfidensnivå. Maskininlärningsmodeller skapas främst för att få en så noggrann prediktion som möjligt där relation mellan variablerna inte är i fokus**

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

**Konfidensintervall för predikterade värden visar exempelvis snittlönen för en 20 åring.**

**Prediktionsintervall visar exempelvis den predikterade lönen för en enskild 20 åring.**

4. Den multipla linjära regressionsmodellen kan skrivas som: 𝑌 = 𝛽0 + 𝛽1𝑥1 + 𝛽1𝑥2+ . . . + 𝛽𝑝𝑥𝑝 + 𝜀 . Hur tolkas beta parametrarna?

**Beta parametrarna skattas genom att välja de parametervärden som minimerar RSS**

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?

**Vi hade sämre datorer förr. Måtten var praktiska och användbara för modellutvärdering i enklare sammanhang.**

6. Förklara algoritmen nedan för ”Best subset selection”

**För att göra Best subset selection, använder vi en separat minsta kvadraters regression bästa delmängdsurval för varje möjlig kombination av p-prediktorerna.**

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

**Citatet syftar på att användbara insikter kan framföras från modeller som inte är en perfekt representation av de fenomen de modellerar.**

# Självutvärdering

1. Utmaningar du haft under arbetet samt hur du hanterat dem.   
   Jag har haft utmaningar när det gäller tillämpning av min lasso regression, jag försökte att hantera det genom att söka på nätet men jag fann inget som kunde hjälpa mig
2. Vilket betyg du anser att du skall ha och varför.   
   G, jag tycker att jag har uppfyllt de kriterier för att kunna få betyget G
3. Något du vill lyfta fram till Antonio?

Nej

# Appendix A

# Källförteckning

Geeksforgeeks.org, 2023 <https://www.geeksforgeeks.org/lasso-regression-in-r-programming/>

Geeksforgeeks.org, 2023 <https://www.geeksforgeeks.org/simple-linear-regression-using-r/>