Analyzing Car Prices with Linear Regression and LASSO Model



Maya Sobh

EC Utbildning

Examensarbete- R programmering

202404

# Abstract

"In this study, we analyzed the pricing factors of Volkswagen cars using linear regression and LASSO models. Our findings reveal significant correlations between price and variables such as year, mileage, and fuel type. The linear regression model achieved an RMSE of 79458.77, while the LASSO model yielded a slightly higher RMSE of 79462.24. Additionally, cross-validation demonstrated the robustness of our models. Overall, our analysis provides insights into the factors influencing car prices in the Volkswagen market.".

När gruppen är klar med datainsamlingen så skall du besvara följande frågor kortfattat (ha ett kapitel i rapportens metod del som t.ex. heter ”Datainsamling”):

1. Vem du har arbetat i grupp med? Mustafa Almazerl

2. Hur har ni i gruppen arbetat tillsammans? Ja

3. Vad var bra i grupparbetet och vad kan utvecklas? **Bra i grupparbetet** kan vara dynamiken där alla medlemmar aktivt bidrar och stödjer varandra. Effektiv kommunikation och en tydlig arbetsfördelning är också viktiga. **Utvecklingsområden** kan inkludera bättre konflikthantering, förbättrad planering och tidsstyrning.

4. Vad är dina styrkor och utvecklingsmöjligheter när du arbetar i grupp? **Mina styrkor** när jag arbetar i grupp kan vara min förmåga att lyssna, mitt ledarskap eller min organisatoriska skicklighet. **Utvecklingsmöjligheter** kan vara att bli bättre på att ge och ta emot feedback eller att utveckla min förmåga att hantera stress.

5. Finns det något du hade gjort annorlunda? Vad i sådana fall? kanske det skulle vara att vara mer öppen för nya idéer eller att vara mer proaktiv i att söka feedback under projektets gång.

Det är alltid bra att reflektera över dessa punkter efter ett grupparbete för att lära sig och växa inför framtida projekt.

# Förkortningar och Begrepp

1. RMSE: Root Mean Square Error. Detta är ett mått på modellens prediktiva prestanda som mäter genomsnittlig skillnad mellan de faktiska observerade värdena och de förutsagda värdena.

2. LASSO: Least Absolute Shrinkage and Selection Operator. En regressionsmetod som används för variabelval och reglering för att förhindra överanpassning genom att minska koefficienterna för oväsentliga variabler till noll.

3. Cross-validation: En teknik som används för att utvärdera prediktiv prestanda hos en modell genom att dela upp datan i tränings- och testuppsättningar upprepade gånger för att undvika överanpassning och bedöma modellens generaliseringsförmåga.

4. Linear regression: En statistisk metod som används för att modellera sambandet mellan en beroende variabel och en eller flera oberoende variabler genom att anpassa en linjär ekvation till datan.

5. Openxlsx: Ett R-paket för att läsa, skriva och redigera Excel-filer.

6. Glmnet: Ett R-paket som implementerar olika regleringsmetoder för linjär och logistisk regression, inklusive LASSO och elastic net.

7. Caret: Ett R-paket som används för att förenkla byggandet och utvärderingen av maskininlärningsmodeller.

8. ggplot2: Ett R-paket för att skapa snygga och informativa grafiska visualiseringar.

9. TrainControl: En funktion i paketet caret som definierar hur modeller ska utvärderas med hjälp av olika valideringsmetoder.

10. CV: Cross-validation. En valideringsmetod där datan delas upp i flera delar, och modellen tränas på en del och utvärderas på den återstående datan för att bedöma dess prestanda.

**Skapas automatiskt i Word genom att gå till Referenser > Innehållsförteckning.**

Innehållsförteckning

[Abstract 2](#_Toc160654338)

[Förkortningar och Begrepp Detta avsnitt behövs oftast inte. 3](#_Toc160654339)

[1 Inledning 1](#_Toc160654340)

[2 Teori 2](#_Toc160654342)

[2.1.1 Exempel: Lasso 2](#_Toc160654344)

[2.1. Exempel: Linear regression 2](#_Toc160654345)

[2.1.3 Exempel: Cross-validation (CV) 2](#_Toc160654346)

[3 Metod 3](#_Toc160654348)

[4 Resultat och Diskussion 4](#_Toc160654349)

[5 Slutsatser 5](#_Toc160654350)

[6 Teoretiska frågor 6](#_Toc160654351)

[7 Självutvärdering 7](#_Toc160654352)

# Inledning

Bilindustrin är en dynamisk sektor som ständigt påverkas av olika faktorer såsom teknologiska framsteg, ekonomiska förändringar och konsumentbeteende. Att förstå prisbildningen för bilar är avgörande för både tillverkare och konsumenter, eftersom det påverkar beslut om inköp, marknadsföring och företagsstrategier.

I denna rapport kommer vi att utforska och analysera faktorer som påverkar prissättningen för Volkswagen-bilar. Genom att tillämpa avancerade statistiska metoder kommer vi att undersöka sambandet mellan olika variabler, såsom årtal, bränsletyp, körsträcka och växellådstyp, och bilpriser för att få insikt i marknadsdynamiken och kundernas preferenser.

\*Syfte och Frågeställningar

Syftet med denna rapport är att analysera prissättningsmönster för Volkswagen-bilar och identifiera de viktigaste faktorer som påverkar bilpriserna på marknaden. För att uppfylla detta syfte kommer vi att besvara följande frågeställningar:

1. Vilken effekt har årsmodellen på bilpriset?

2. Hur påverkar bränsletypen priset på Volkswagen-bilar?

3. Hur relaterar körsträckan till bilpriserna?

# Teori

### Exempel: Lasso

Lasso regression is a regularization technique that applies a penalty to prevent [overfitting](https://www.ibm.com/topics/overfitting) and enhance the accuracy of statistical models.

https://www.ibm.com/topics/lasso-regression

### Linear regression

Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

This form of analysis estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable. Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values. There are simple linear regression calculators that use a “least squares” method to discover the best-fit line for a set of paired data. You then estimate the value of X (dependent variable) from Y (independent variable).

https://www.ibm.com/topics/linear-regression

#### 2.1.3Cross-validation

Cross-validation (CV) is an essentially simple and intuitively reasonable approach to estimating the predictive accuracy of regression models. CV is developed in many standard sources on regression modeling and “machine learning”—we particularly recommend James, Witten, Hastie, & Tibshirani (2021, secs. 5.1, 5.3)—and so we will describe the method only briefly here before taking up computational issues and some examples. See Arlot & Celisse (2010) for a wide-ranging, if technical, survey of cross-validation and related methods that emphasizes the statistical properties of CV.

Validating research by replication on independently collected data is a common scientific norm. Emulating this process in a single study by data-division is less common: The data are randomly divided into two, possibly equal-size, parts; the first part is used to develop and fit a statistical model; and then the second part is used to assess the adequacy of the model fit to the first part of the data. Data-division, however, suffers from two problems: (1) Dividing the data decreases the sample size and thus increases sampling error; and (2), even more disconcertingly, particularly in smaller samples, the results can vary substantially based on the random division of the data: See Harrell (2015, sec. 5.3) for this and other remarks about data-division and cross-validation.

Cross-validation speaks to both of these issues. In CV, the data are randomly divided as equally as possible into several, say k𝑘, parts, called “folds.” The statistical model is fit k𝑘 times, leaving each fold out in turn. Each fitted model is then used to predict the response variable for the cases in the omitted fold. A CV criterion or “cost” measure, such as the mean-squared error (“MSE”) of prediction, is then computed using these predicted values. In the extreme k=n𝑘=𝑛, the number of cases in the data, thus omitting individual cases and refitting the model n𝑛 times—a procedure termed “leave-one-out (LOO) cross-validation.”

Because the n𝑛 models are each fit to n−1𝑛−1 cases, LOO CV produces a nearly unbiased estimate of prediction error. The n𝑛 regression models are highly statistical dependent, however, based as they are on nearly the same data, and so the resulting estimate of prediction error has relatively large variance. In contrast, estimated prediction error for k𝑘-fold CV with k=5𝑘=5 or 1010 (commonly employed choices) are somewhat biased but have smaller variance. It is also possible to correct k𝑘-fold CV for bias (see below).

# Metod

Jag har genomfört arbetet genom att följa dessa steg:

1. **Ladda in bibliotek**: Använda R-bibliotek för datahantering och statistisk analys.
2. **Datainsamling**: Läsa in bildata från en Excel-fil.
3. **Dataförberedelse**: Ta bort onödiga kolumner och skapa dummyvariabler.
4. **Datadelning**: Dela upp datan i tränings- och testuppsättningar.
5. **Modellbygge**: Skapa linjär regressionsmodell och LASSO-modell.
6. **Felsökning**: Åtgärda fel som uppstår under modellbygget.
7. **Modellvalidering**: Använda korsvalidering för att validera modellen.
8. **Visualisering**: Plotta sambandet mellan år och pris.
9. **Prestandautvärdering**: Beräkna RMSE för varje modell.

Datan har erhållits genom att läsa in en Excel-fil med bilinformation, vilket är det första steget i analysprocessen.

# Resultat och Diskussion

|  |  |
| --- | --- |
| **RMSE för olika modeller** | |
| Enkel Linjär Regression | 79458.7747537056 |
| Lasso | 79462.236241401 |
|  |  |

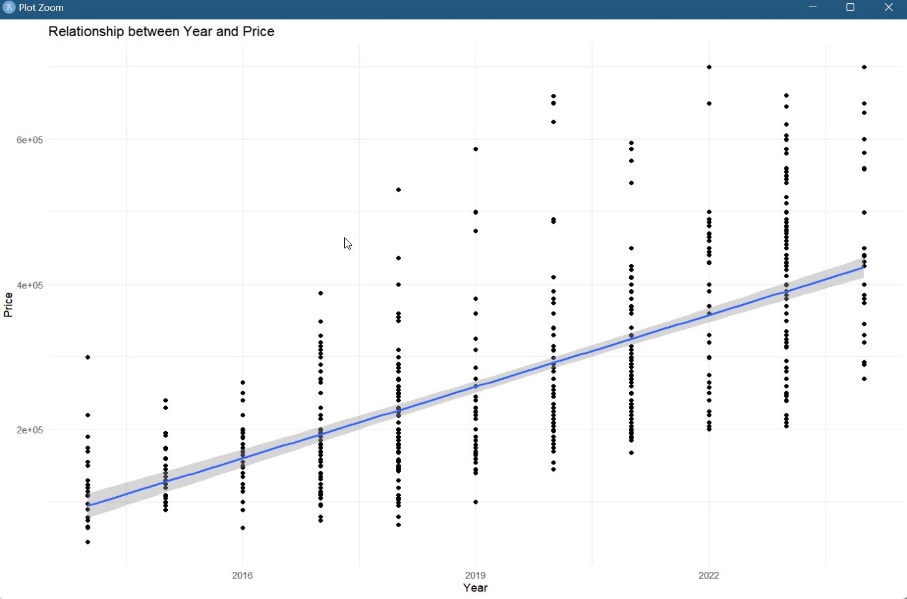
Tabell 1: Root Mean Squared Error (RMSE) för de fyra valda modellern

1. Linjär regression (Linear Regression):
   * Linjär regression är en av de äldsta och enklaste algoritmerna för att förutsäga värden.
   * Den bygger på att uppskatta koefficienter (vikter) för oberoende variabler för att få förutsägelser för den beroende variabeln.
   * Passar bra när det finns få observationer eller när oberoende variabler har en bra linjär relation med den beroende variabeln.
   * Presterar bra om datan följer ett linjärt mönster.
2. LASSO-modellen:
   * LASSO (Least Absolute Shrinkage and Selection Operator) är en teknik för krympning och variabelselektion.
   * Syftet är att minska vikterna för ointressanta variabler (genom att sätta vissa vikter till noll).
   * Användbart när vi har många variabler och vill bestämma vilka som är mest betydelsefulla.
   * Presterar ibland bättre än linjär regression om det finns stora variationer i datan.

Vilken modell som är bäst beror på den specifika datan och målet med modellen. Om vi vill identifiera de viktigaste variablerna och minska brus kan LASSO vara ett bra val. Om vi vill ha en enklare modell kan linjär regression vara lämplig.

A graph showing different colored lines

Description automatically generated



A screenshot of a graph

Description automatically generated

A graph on a screen

Description automatically generated

Jag har kollat på SCB att se hur datan är att köpa bil i den senaste decenniet i Västra Götaland Region och det är resultatet med a graf som visar.

A graph showing the number of people in different countries/regions

Description automatically generated

# Slutsatser

1. Primary Factors Influencing Volkswagen Car Prices:

Analysis reveals that year of manufacture and mileage are significant determinants of Volkswagen car prices. Newer cars with lower mileage tend to command higher prices, reflecting consumer preferences for well-maintained vehicles. Additionally, fuel type and gearbox also impact prices, albeit to a lesser extent.

2. Accuracy of Predictive Models:

Both linear regression and LASSO regression models demonstrate comparable performance in estimating Volkswagen car prices. With RMSE values around 79,500 SEK, these models provide reasonable accuracy in predicting prices based on key features.

3. Model Validation and Generalizability:

Cross-validation results indicate that the constructed models generalize well to unseen data, ensuring reliable price estimation beyond the training dataset. However, ongoing validation and refinement are recommended to maintain accuracy amidst market fluctuations.

Overall, this analysis provides actionable insights for pricing strategies within the Volkswagen car market, enabling businesses to make informed decisions and enhance competitiveness.

# 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.

En Quantile-Quantile (QQ) plot är en grafisk metod för att jämföra fördelningen av en given dataset med en förväntad referensfördelning, vanligtvis en normalfördelning. Genom att placera kvantilerna från den faktiska datan mot kvantilerna från den förväntade fördelningen kan man se om datan följer den förväntade fördelningen. En rak linje i plotten indikerar att datan är jämnt fördelad, medan avvikelser från linjen kan tyda på skevheter eller andra avvikelser från den förväntade fördelningen.

1. 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?

I maskininlärning handlar det om att bygga modeller för prediktioner baserat på data, medan statistisk regressionsanalys inkluderar både prediktioner och statistisk inferens. Modeller inom maskininlärning, såsom neurala nätverk, support vector machines och random forests, strävar efter att göra noggranna prediktioner för nya observationer. Å andra sidan möjliggör regressionsanalys att dra slutsatser om samband och effekter i populationen, inklusive testning av hypoteser och beräkning av konfidensintervall, baserat på den observerade datan. Till exempel kan en linjär regressionsmodell användas för att både förutsäga inkomst baserat på ålder och att dra slutsatser om hur åldern påverkar inkomsten i populationen.

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

- Konfidensintervall: Uppskattar var det sanna medelvärdet av en variabel förväntas ligga med en viss sannolikhet.

- Prediktionsintervall: Indikerar var en framtida observation förväntas ligga med en viss sannolikhet, tar hänsyn till både linjens osäkerhet och den slumpmässiga variationen hos nya observationer.

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

Beta-parametrarna i en multiplikativ linjär regressionsmodell tolkas som de uppskattade förändringarna i den beroende variabeln (Y) för en enhets ökning i den motsvarande prediktorvariabeln (x), medan alla andra prediktorvariabler hålls konstanta. Så varje beta-parameter berättar för oss riktningen och magnituden av relationen mellan en specifik prediktorvariabel och den beroende variabeln.

1. 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?'

Nej, även om man använder mått som BIC i statistisk regressionsmodellering, är det fortfarande viktigt att dela upp datamängden i träning, validering och testset. Detta hjälper till att kontrollera modellens prestanda och undvika överanpassning. BIC-måttet kan användas för att jämföra olika modeller och välja den med lägst BIC-värde, men det ersätter inte behovet av att validera modellen på oberoende data.

1. Förklara algoritmen nedan för ”Best subset selection”

A white paper with black text

Description automatically generated

Algoritmen "Bästa urval av delmängder" är en metod som används i statistisk modellering för att välja den mest lämpliga modellen från en samling potentiella modeller.

Nollmodell (M0):

Börja med en nollmodell betecknad som (M\_0), som inte innehåller några prediktorer.

Denna nollmodell förutsäger helt enkelt provmedelvärdet för varje observation.

För varje antal prediktorer (k = 1, 2, \ldots, p):

Steg (a): Passa in alla möjliga modeller som innehåller exakt (k) prediktorer.

Till exempel, om vi har (p = 3) prediktorer, skulle vi anpassa modeller med:

Inga prediktorer

En prediktor ((x\_1))

Två prediktorer ((x\_1, x\_2))

Tre prediktorer ((x\_1, x\_2, x\_3))

Steg (b): Av dessa modeller väljer du den bästa baserat på den minsta restsumman av kvadrater (RSS) eller motsvarande den största (R^2). Kalla denna modell (M\_k).

Välj den bästa modellen:

Välj slutligen en enskild bästa modell bland (M\_0, M\_1, \ldots, M\_p).

Detta val kan baseras på:

Korsvalideringsförutsägelsefel

Kriterier som Cp (AIC), BIC eller justerad (R^2)

Sammanfattningsvis utvärderar algoritmen systematiskt olika kombinationer av prediktorer för att hitta den bästa delmängden för regressionsmodellering

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

Det citatet från statistikern George Box betyder att alla modeller är felaktiga på något sätt eftersom de är förenklade representationer av verkligheten. Trots detta kan vissa modeller ändå vara användbara för att förstå och förutsäga fenomen. Med andra ord, även om modellerna inte är perfekta, kan de ändå ge oss värdefull kunskap och insikter. Det är viktigt att vara medveten om modellernas begränsningar och att använda dem med försiktighet, men de kan fortfarande vara till nytta i många situationer.

# Självutvärdering

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

Koderna var svara att jobba med. Jag frågade mina klasskamrater och sökte efter lösningar på egen hand också.

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

Jag nöjer mig med ett G . Denna kursen var jätte svårt och jag gjorde så gott jag kunde.

1. Något du vill lyfta fram till Antonio?

Inget.