Regressionsanalys

Volvo v60 bilpriser



Jacob Andersson

EC Utbildning

R programmering - Regressionsanalys

2024-15

# Abstract

This study presents an analysis of the factors influencing the pricing of Volvo V60 cars and the development of a regression model for price prediction. Utilizing data on various car attributes and prices, statistical methods including regression analysis and variable selection techniques were employed to identify significant predictors of Volvo V60 car prices and develop an accurate prediction model.

The results indicate that factors such as the age of the car, mileage, fuel type, gearbox and model year did impact the pricing of Volvo V60 cars. Furthermore, the developed regression model demonstrates strong performance in predicting Volvo V60 car prices, with high adjusted R-squared values and low AIC and BIC values, indicating a well-fitted model.

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

Innehållsförteckning

[Abstract 2](#_Toc165033739)

[1 Inledning 1](#_Toc165033740)

[2 Teori 2](#_Toc165033741)

[2.1 Regressionsanalys 2](#_Toc165033742)

[2.2 Variabelurval 2](#_Toc165033743)

[2.3 Modellutvärdering och Prediktionsförmåga 2](#_Toc165033744)

[3 Metod 3](#_Toc165033745)

[3.1 Insamling och Förberedelse av Data 3](#_Toc165033746)

[3.2 Regression och Variabelurval 3](#_Toc165033747)

[3.3 Utvärdering av Modellens Prediktionsförmåga 3](#_Toc165033748)

[3.4 Visualisering och Presentation av Resultat 3](#_Toc165033749)

[3.5 Datainsamling SCB 3](#_Toc165033750)

[4 Resultat och Diskussion 4](#_Toc165033751)

[4.1 Signifikanta Prediktorer för Bilpris 4](#_Toc165033752)

[4.2 Prediktionsmodellens Prestanda 4](#_Toc165033753)

[4.3 Visualisering av Resultat 4](#_Toc165033754)

[4.4 Sammanfattning och Implikationer 4](#_Toc165033755)

[5 Slutsatser 6](#_Toc165033756)

[5.1 Faktorer som påverkar prissättningen av Volvo V60-bilar 6](#_Toc165033757)

[5.2 Prestanda av regressionsmodellen för prissättning av Volvo V60-bilar 6](#_Toc165033758)

[6 Teoretiska frågor 7](#_Toc165033759)

[7 Självutvärdering 9](#_Toc165033760)

[Appendix A 10](#_Toc165033761)

[Källförteckning 13](#_Toc165033762)

# Inledning

Denna rapport presenterar en analys av faktorer som påverkar prissättningen av bilar samt en utforskning av modeller för prognostisering av pris. Genom att använda en kombination av statistiska verktyg och regressionsanalys, undersöks och analyseras sambandet mellan olika faktorer och deras påverkan på priset. Datan som används i denna rapport har samlats in från annonser från blocket och behandlats med hjälp av programmeringsspråket R.

Efter identifiering av annonser, inleds en utförlig dataladdning och förbehandling för att säkerställa att datan är lämplig för analys. Därefter utförs en regressionsanalys för att undersöka variablernas betydelse för prissättningen av bilar. Vi använder både forward stepwise selection och backward stepwise selection för att identifiera de mest signifikanta prediktorerna för priset.

Genom att tillämpa lämpliga statistiska modeller och visualiseringstekniker, bearbetas resultatet av analysen och modellernas prognosförmåga bedöms. Avslutningsvis presenteras en summering av resultaten.

Denna rapport syftar till att bidra till en djupare förståelse för prissättning och olika faktorer som kan ha en påverkan på priset.

Frågeställningar som kommer besvaras i denna rapport är:

1. Vilka faktorer har störst påverkan på prissättningen av Volvo V60-bilar?
2. Hur väl presterar regressionsmodellen för att förutsäga prissättningen av Volvo V60-bilar?

# Teori

Prisprognostisering av bilar är en viktig del av fordonsindustrin och påverkar både konsumenter och producenter. Genom att förutse priset på bilar kan företag fatta välgrundade beslut gällande produktion, marknadsföring och försäljning. Denna teoretiska del av rapporten ger en översikt av de teoretiska och metodologiska ramverk som används för att förstå och prognostisera priset på bilar.

## Regressionsanalys

Regression är en statistisk metod som används för att undersöka sambandet mellan en eller flera oberoende variabler och en kontinuerlig beroende variabel. Inom prisprognostisering av bilar kan regression användas för att uppskatta hur olika faktorer såsom bilens ålder, miltal, märke, modell, bränsletyp och andra attribut påverkar dess pris.

Genom att utforma och anpassa lämpliga regressionsmodeller kan man förutse priserna på bilar baserat på deras egenskaper. Det finns olika typer av regressionsmodeller som kan användas, inklusive linjär regression, polynomial regression, och andra mer avancerade tekniker som till exempel ridge regression och elastic net regression. I denna undersökning används en multipel linjär regression

## Variabelurval

Ett viktigt steg inom prisprognostisering är att välja de mest relevanta variablerna att inkludera i modellen. Variabelurval handlar om att identifiera de prediktorer som bäst förklarar variationen i priset på bilar. Forward stepwise selection och Backward stepwise selection är två metoder för variabelurval som används för att systematiskt välja in eller ta bort variabler från modellen baserat på deras signifikansnivå och bidrag till modellens prediktionsförmåga.

## Modellutvärdering och Prediktionsförmåga

När en regressionsmodell har utformats är det viktigt att utvärdera dess prediktionsförmåga. Modellutvärdering kan inkludera metoder som att beräkna R^2, adjusted r squared, Cp (Mallow´s Cp), BIC (Bayesian Information Criterion), och andra mått på modellens passform och prediktionsnoggrannhet.

Genom att noggrant utvärdera modellens prestanda kan man bedöma dess förmåga att förutsäga priset på bilar och identifiera eventuella områden där modellen kan förbättras.

Denna teoretiska bakgrund ger en grundläggande förståelse för de metoder och begrepp som används för att prognostisera priset på bilar. I följande avsnitt tillämpas dessa teoretiska koncept på analysen av bilpriser med hjälp av statistiska verktyg och programmeringsspråket R.

# Metod

## Insamling och Förberedelse av Data

Data som används i denna studie samlades in från en källa som tillhandahåller information om olika attribut hos bilar samt deras priser. Informationen omfattar variabler såsom ålder på bilen, miltal, hästkrafter, bränsletyp och växellåda.

För att förbereda datan för analys genomfördes en serie förbehandlingssteg. Detta inkluderade anpassning av datan för att hantera eventuella saknade värden eller anomalier. Därefter genomfördes en eventuell omkodning av kategoriska variabler till numeriska värden för att underlätta analysprocessen. Slutligen delades datan upp i tränings- och testuppsättningar för att möjliggöra utvärdering av modellens prediktionsförmåga.

## Regression och Variabelurval

För att undersöka sambandet mellan bilattribut och priset på bilar användes linjär regression. Den beroende variabeln var logaritmen av bilpriset för att uppnå en mer normalfördelad responsvariabel och minska effekterna av utstickande observationer.

Variabelurval genomfördes med hjälp av både Forward stepwise selection och Backward stepwise selection. Dessa metoder användes för att systematiskt identifiera de mest signifikanta prediktorerna för priset på bilar baserat på deras bidrag till modellens förklaringskraft och statistiska signifikansnivå.

## Utvärdering av Modellens Prediktionsförmåga

För att utvärdera modellens prediktionsförmåga användes flera mått såsom adjusted R squared, Cp, BIC och andra relevanta mått på modellens passform och prediktionsnoggrannhet.

## Visualisering och Presentation av Resultat

Resultaten av analysen presenteras genom användning av olika visualiseringsmetoder såsom grafiska illustrationer av regressionsmodeller, variabelurvalsprocessen och jämförelser av olika modellers prediktionsförmåga. Dessutom presenteras resultaten i form av tabeller och beskrivande statistik för att underlätta förståelsen och tolkningen av resultaten.

Genom att tillämpa dessa metoder och tekniker kunde vi undersöka och analysera sambandet mellan bi lattribut och priset på bilar samt utveckla en prognosmodell som kan användas för att förutsäga bilpriser med tillfredsställande precision

## Datainsamling SCB

En översiktlig datainsamling gjordes på antalet bilar som är registrerade i Sverige. Resultaten visar på en generell ökande trend, mellan åren 2018 till 2022, men en liten nedåtgång senaste året.

# Resultat och Diskussion

Efter att ha genomfört en omfattande analys av data om bilattribut och priser samt tillämpat statistiska metoder för variabelurval och regressionsanalys, presenteras nedan resultatet av studien.

## Signifikanta Prediktorer för Bilpris

Genom Forward stepwise selection och Backward stepwise selection identifierades flera signifikanta prediktorer för priset på bilar. Bland de mest betydelsefulla faktorerna inkluderades bilens ålder, miltal och hästkrafter. Bränsletyp och växellåda visade också ha viss betydelse för att förklara variationen i bilpriserna, men hade inte lika stor effekt som dom förstnämnda.

## Prediktionsmodellens Prestanda

Den utvecklade regressionsmodellen visade sig ha god prediktionsförmåga för bilpriser, vilket återspeglas i höga justerade R-kvadratvärden och låga residual standard error.

Det justerade R-kvadratvärdet på 0.8651 indikerar att cirka 86.51% av variationen i bilpriserna kan förklaras av de oberoende variablerna i modellen. Detta tyder på att modellen är väl anpassad till data och har en stark förklaringskraft.

Residual standard error, som är 0.09912, representerar det genomsnittliga avståndet mellan de observerade bilpriserna och de förutsagda bilpriserna från modellen. Ett lägre värde på residual standard error indikerar att modellen ger mer exakta förutsägelser om bilpriserna.

Dessutom visar F-statistiken en signifikant modell (p-värde: < 2.2e-16), vilket tyder på att de inkluderade variablerna tillsammans har en signifikant effekt på förklaringen av variationen i bilpriserna.

Denna kombination av högt adjusted R squared, lågt residual standard error och signifikant F-statistic stärker tilltron till modellens förmåga att pålitligt förutsäga priser för Volvo V60-bilar.

## Visualisering av Resultat

Resultaten av analysen visualiserades genom användning av grafiska illustrationer såsom regressions plottar, variabelurvalsgrafer och jämförelse av olika modellers prediktionsnoggrannhet. Dessa visualiseringar bidrog till en tydligare förståelse av sambanden mellan bilattribut och priser samt modellernas prestanda i att förutsäga priserna på bilar.

## Sammanfattning och Implikationer

Sammanfattningsvis visar resultaten av studien att det finns tydliga samband mellan olika bilattribut och priset på bilar. Genom att utforma och tillämpa lämpliga regressionsmodeller kan bilpriserna förutsägas med tillfredsställande precision. Dessa insikter kan vara värdefulla för fordonsindustrin, konsumenter och andra intressenter och kan användas för att fatta informerade beslut gällande prissättning, marknadsföring och försäljning av bilar. Fortsatt forskning och utveckling inom området för prisprognostisering av bilar kan ytterligare förbättra förståelsen och förmågan att förutsäga bilpriser i framtiden.

# Slutsatser

## Faktorer som påverkar prissättningen av Volvo V60-bilar

Baserat på regressionsanalysen och variabelurvalet har det visat sig att flera faktorer har betydande påverkan på prissättningen av Volvo V60-bilar. Bland dessa faktorer är bilens ålder, miltal och hästkrafter de mest signifikanta. Dessutom har specifika tekniska attribut och karaktärsdrag hos Volvo V60, såsom bränsletyp och växellåda visat sig vara av betydelse för att förklara variationen i bilpriserna.

## Prestanda av regressionsmodellen för prissättning av Volvo V60-bilar

Den utvecklade regressionsmodell visar sig ha god prestanda när det gäller att förutsäga prissättningen av Volvo V60-bilar. Modellen uppvisar höga justerade R-kvadratvärden och låga BIC-värden, vilket indikerar en väl anpassad modell med god förklaringskraft. Genom att jämföra modellen med andra alternativa modeller visar att vald regressionsmodell är relativt pålitlig i sin prediktionsförmåga och kan användas med tillförsikt för att förutsäga prissättningen av Volvo V60-bilar.

Modellens förmåga att prediktera rätt pris skulle kunna förbättras ytterligare. Med hjälp av fler variabler som till exempel utrustningsnivå finns det ännu mer precision att hämta.

# 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.**

*Svar: En QQ-Plot används för att se om vår data när normalfördela, om vår data simulerar en rak linje så är vår data normalfördelad*

**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?**

*Svar: I maskininlärning är huvudfokus på prediktioner medan statistisk regressionsanalys inkluderar både prediktioner och statistisk inferens, vilket innebär att man försöker förstå och tolka sambandet mellan variabler i datan.*

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

*Svar: Konfidensintervall baseras på ett snitt tillexempel snittlön för en viss ålder medan prediktionsintervall baseras på lönen för en specifik individ. Prediktionsintervallet är bredare eftersom vi har med (Epsilon) i åtanke.*

**4. Den multipla linjära regressionsmodellen kan skrivas som:**

**Y = 𝛽0 + 𝛽1𝑥1 + 𝛽1𝑥2+ ...+ 𝛽𝑝𝑥𝑝 +𝜀 .**

**Hur tolkas beta parametrarna?**

*Svar:* 𝛽0 *är intercenpet där linjen skär y-axeln,* **𝛽**1, **𝛽2, 𝛽p** är koefficienterna som visar hur mycket förändring i Y vi förväntar oss för varje enhets förändring i den respektive oberoende variabeln.

**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?**

*Svar: Det stämmer* *när du jämför olika modeller med BIC, strävar du efter att hitta den modell som har lägst BIC-värde. Ett lägre BIC indikerar att modellen har en bättre balans mellan att passa data väl och att vara tillräckligt enkel för att undvika överanpassning.*

**6. Förklara algoritmen nedan för ”Best subset selection”**

En bild som visar text, skärmbild, Teckensnitt, nummer

Automatiskt genererad beskrivning

*Svar: Vi börjar med att generera alla möjliga kombinationer av variabler som vi kan använda i vår modell. För varje kombination av variabler passar vi en linjär regressionsmodell och beräknar ett prestandamått. Vi väljer den kombination av variabler som ger den högsta prestandan enligt det valda prestandakriteriet.*

7. Ett citat från statistikern George Box är: “All models are wrong, some are useful.”

Förklara vad som menas med det citatet.

*Svar: Att ingen modell perfekt representerar verkligheten. Modellerna är en förenkling av verkligheten men kan ge en bra uppskattning.*

# Självutvärdering

1. **Utmaningar du haft under arbetet samt hur du hanterat dem.**Svårt att väva in datan från SCB och visste inte hur jag skulle väva in det bra i rapprten, blev mer som en egen del i rapporten.
2. Vilket betyg du anser att du skall ha och varför.   
   G, Gjorde inte VG delen
3. Något du vill lyfta fram till Antonio?

# Appendix A

En bild som visar text, skärmbild, linje, Graf

Automatiskt genererad beskrivning  
*Figur 1: Hur modellen predikterar bilpriserna, röd linje är predikterade värdet, blå prickar är observationen*

En bild som visar text, diagram, linje, skärmbild

Automatiskt genererad beskrivningEn bild som visar text, diagram, linje, Graf

Automatiskt genererad beskrivning*Figur 2: Datans normalfördelning innan och efter logaritmen användes*

En bild som visar text, skärmbild, diagram, Rektangel

Automatiskt genererad beskrivning  
*Figur 3: Här ser vi R^2 som visar Antalet variabler vi väljer samt hur bra modellen presterar*

En bild som visar text, skärmbild, Teckensnitt, nummer

Automatiskt genererad beskrivning  
*Figur 4: Visar hur bra modellen har predikterat bilpriset*

En bild som visar text, skärmbild, diagram, nummer

Automatiskt genererad beskrivning  
*Figur 4: Visar Antalet bilar i Sverige mellan 2018-2022 och hur trenden går.*

En bild som visar text, Teckensnitt, skärmbild, linje

Automatiskt genererad beskrivning  
*Figur 1: Residualerna som ligger jämt fördelat runt noll*

# Källförteckning