**Statistik över nyregistrerade bilar och prisfaktorer**



Ryan Leyor

EC Utbildning

R kunskapskontroll

2025–04

# 

# Abstract

This report is based on analysis of newly registered cars in Sweden using official SCB data, combined with regression modelling based on second-hand volvo car prices collected from Blocket. The SCB data was used to visualize and highlight trends in fuel types between 2018 and 2025.   
Blocket data was analyzed through a linear regression model, where price was modeled as a function of mileage, model year, fuel type and gearbox.

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

Innehållsförteckning

[Abstract 2](#_Toc197367032)

[1 Inledning 1](#_Toc197367033)

[2 Teori 2](#_Toc197367035)

[2.1 Exempel: Regressionsmodeller 2](#_Toc197367036)

[2.1.1 Exempel: Lasso 2](#_Toc197367037)

[2.1.2 Exempel: Ridge 2](#_Toc197367038)

[2.1.3 Exempel: Elastic Net 2](#_Toc197367039)

[2.2 Exempel: Neurala Nätverk 2](#_Toc197367040)

[3 Metod 3](#_Toc197367041)

[4 Resultat och Diskussion 4](#_Toc197367042)

[5 Slutsatser 5](#_Toc197367043)

[6 Teoretiska frågor 6](#_Toc197367044)

[7 Självutvärdering 7](#_Toc197367045)

[Appendix A 8](#_Toc197367046)

[Källförteckning 9](#_Toc197367047)

# Inledning

Statisk analays är viktigt för att undersöka och förstå förändringar och utvecklingar. Genom linjär regression kan vi undersöka samband mellan olika variabler och förklara hur faktorer som ålder, typ och använding påverkar exempelvis priset på något vara och se dess trender.  
Som i detta fall, bilar. Där många faktorer spelar stor roll på förändringar.

Bilmarknaden förändras snabbt i Sverige. Allt från flera El bilar, höjda bränslepriser och möjlipåverkan. För att förstå detta bättre och kunna dra slutsater, kommer frågan – kan man se denna förändring i faktisk data.

I denna rapport kombineras data från SCB:s statistik över nyregistrerade bilar och data över begagnade Volvo bilar från plattformen Blocket. För att besvara frågor som hurvida konsumenter i Sverige har utvecklats när det gäller köp av bilar och prissättning av Volvo bilar över en viss tid.

# Teori

## Linjär regressionmodell

Linjär regression är en s kraffull modell som används för att analysera en relation mellan variabler. Det viktiga i sådan modell är att en av variabler är fungerar som beroende variabel som i detta fel pris på fordonet och övriga ska vara oberoende variabler som exempel *miltal , modellår och drivmedel*. Syftet med detat blir då att kunna se och förklara hur de oberoende variabler påverkar den beroende variabeln.

Formen för linjär regressionsmodell är:

där *Y* är oborende variabel som är baserat på en ell fler X1….Xn som är oberoende variabler. är intreceptet där linjen skär Y-axel (när X är 0). som modellen skapar. βn är kofficinter (Lutning) som visar hur mycket varje X påverkar Y och ε är felterm , variation i Y som inte förklaras av X.

### Antaganden

För att den linjära regressionen ska fungera på rätt sätt måste vissa antaganden vara uppfyllda. Som exempel måste det finnas ett *linjärt samband* mellan dn boeroende och oberoende variablen.

Antagandet som också måste uppfyllas är *Homoskedasticitet*, där felens varians ska vara kontant över hela data mängden om inte, blir den statistiska inferensen som görs på modellen felaktig.

Ett annat antagande är *oberoende residualer* vilket helt enkelt tyder å att modellens fel inte ska vara beroende av varandra.

Sen finns det även normalfördelade residualer, att feltermerna borde följa ungefärlig normalfördelning vilket låter oss använda modellen för hypotesprövning och statiskt inferens.

Sist men inte minst, Ingen kollinearitet. Att inga av dem oberoende variablerna ska vara korrelerade med varandra.

## Multikollinearitet

I detta arbete användes funktionen vif() för att undersöka förekomsten av multikollinearitet. Ett VIF-värde över 5 eller 10 kan indikera att en variabel är för starkt korrelerad med andra variablerna

## Variabelurval

Från paketet leaps kan man använda funktionen regsubsets() som jämför olika kombinationer av variabler. Vilket i det här fallet användes för att göra det möjligt att välja bästa kombinationen av variabler.

# Metod

## Datainsamling

I arbetet har två olika dataset använts: Nyregistrerade bilar från Statistiska Centralbyrån (SCB) och ett från webbsidan Blocket där annonser om bilar till salu finns samlade.

### SCB Data

För att undersöka trender i vad för typer av nyregistrerade bilar i jämförelse med drivmedel finns så användes SCB:s API där med hjälp av paketet pxweb i R studio. Väl inne på SCB:s API användes följande filtreringar: Översikt över samtliga Län i Sverige, på tidspannet 2018-2025 samt drimedelstyperna Bensin, Diesel och EL-Bilar.   
  
Bearbetning gjordes med hjälp av dplyr för att gruppera och summera antalet bilar per år och drivmedelstyp.

Sedan skapasdes en visualisering med ggplot2 för att visa utvecklingen över tid av el-, bensin- och diseldrivna bilar.

Resultatet sparades därefter som en CSV-fil för vidare användning i rapportens visualisering. (*Figur 1/2*)

Det som är att ta hänsyn till är att eftersom år 2025 inkluderades i datan, blir det väldigt djup sjunkning för samtliga drivmedel efter år 2024, då datan är endast uttaget på 3 månader i år 2025.

## Blocket Data (Regression)

Datan från Blocket består av en sammanställning av tidigare studenters insamlingar av begagnade Volvo bilar. Variabler som försäljningpris, miltal, modellår, drivmedel och växellåda användes i analysen.

Datan öppnades i excel för att granskas och hitta brister. Tomma celler ersattes med ordet ”Okänd”. Sedan upptäcktes även miltalsvärden som var ”-”, dessa omformaterades till nollor och sedan ersattes även med Okänd.

Felmeddelanden som brutna formler och externa länkar dök även upp som rensades.

Efter inläsning av datan i Rstudio påbörjade stegen för omkodning samt rensnig. Kategoriska variabler omvandlades till faktorer och saknade värden hanterade.

Flera lm modeller med olika kombinationer testades för att se ifall dess plot skulle få bättre/olika resultat, vilket det inte egentligen gjorde.

Där beslutades att testa med hjälp av regsubsets() från paketet leaps , välja bästa kombinationen av variabler där även VIF-Värden användes för att kontrollera modellens antaganden och att modellen inte drabbades av multikollinearitet.

Slutliga modellen visade bättre resultat *(figur 3)* där antaganden för linjär regression var uppfyllda jämfört med tidigare versioner. Det som hjälpte dessa resultat var log-transformering som lades till i slutliga modellen.

# Resultat och Diskussion

En bild som visar text, skärmbild, Teckensnitt, nummer

AI-genererat innehåll kan vara felaktigt.

Figur 1: SCB – Nyregistrerade bilar över samtliga län i Sverige baserat på drivmedel

En bild som visar text, diagram, linje, Graf

AI-genererat innehåll kan vara felaktigt.

Figur 2: Graf över nyregistrering av bilar i Sveriges läner.

En bild som visar text, diagram

AI-genererat innehåll kan vara felaktigt.

Figur 3: diagnostikgrafer från en linjär regressionsmodell.

# Slutsatser

Utifrån SCB:s data ser man tydligt att trenden går från bensin och diesel mot elbilar. Även om jag inte har valt att ta med hybridbilar i min analys, ser man ändå en klar förändring i val av drivmedel igenom åren.  
  
Efter att ha testat flera modeller och jämfört olika plottar så kunde jag se att resultatet förbättrades ganska mycket när jag valde bättre kombinationer av variabler. Framför allt märktes det tydligt i plottarna att modellen blev mer stabil och förutsägbar.

En möjlig förklaring till varför vissa bilar inte riktigt följer mönstret kan vara att bilar ofta tappar mycket värde direkt efter köp även om de inte körts så långt. Det är en känd trend att en bil med nästan samma miltal men olika ålder kan ha ganska olika priser eller i vissa fall till och med samma pris.

Det påverkar såklart hur modellen tolkar sambandat mellan pris och miltal.

Residual-plotten tillexempel visar att den inte är perfekt, men att den blivit mycket bättre än tidigare versioner.

Något jag anser vara viktigt i dem resultaten är att det är svårt att se en trend då mycket varieras i resultaten vilket ger mig beslutet att man inte kan förutse något i förväg, som något trend frammåt.

# 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.*

Används för att kontrollera om datan i en regressionsmodell är normalfördelad. Om punkterna ligger nära en rak linje kan det tyda på att fördelningen är normal. Om datan avviker från linjen betyder det att residualerna inte följer normal fördelning vilket kan skapa ett problem.

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?*

Linjär regression kan absolut användas både inom statistisk och maskininlärning men syftet kan skilja sig. Maskininlärning kanske används mer i syftet att göra förutsägelser , tillexempel hur mycket en bil kommer kosta.  
Men i statistisk regressionsanalys försöker man se vilka samband variabler har, som priset påverkas av miltal på bilen.

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

Ett konfidensintervall utgår från medelvärdet, medan prediktionsintervallet är vad vi predicerar för en framtida observation.

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

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

*Hur tolkas beta parametrarna?*

är intreceptet där linjen skär Y-axel (när X är 0). som modellen skapar. βn är kofficinter (Lutning) som visar hur mycket varje X påverkar Y och ε är felterm , variation i Y som inte förklaras av X.

1. *Din kollega Nils 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?*

Det stämmer till viss del. När man använder BIC behöver man inte dela upp data i träning, validering och test som man gör i maskininlärning. BIC automatiskt försöker hitta en balans mellan hur bra modellen passar datan och hur enkel den är. Därför kan BIC identifiera vilken av regrissionsmodeller som bäst.   
Men om man kör träning, validering och test kan vara bra ändå, eftersom det minskar risken för för overfitting och visar direkt hur modellen preseterar på ny data.

1. Förklara algoritmen nedan för ”Best subset selection”

En bild som visar text, skärmbild, Teckensnitt, nummer

Automatiskt genererad beskrivning

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

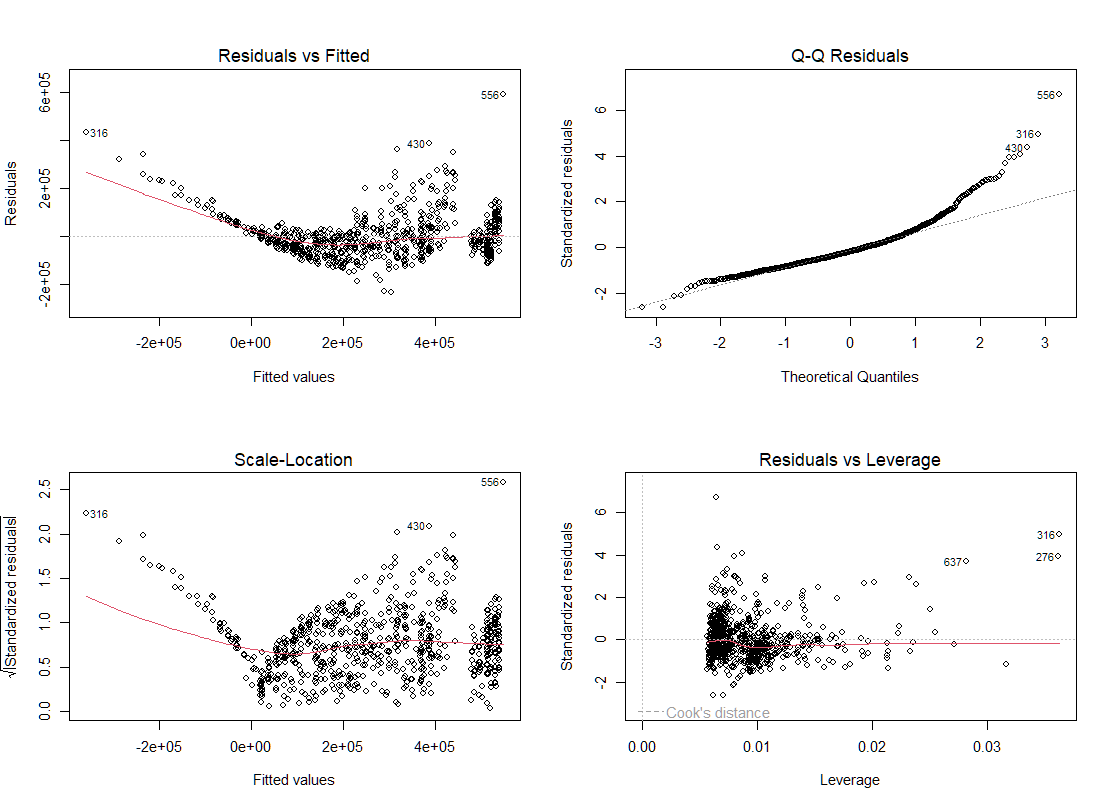
*Förklara vad som menas med det citatet.*

Man kan inte skapa en perfekt modell alla modeller bygger på något som inte faktiskt fungerar i verkligen. Men även fast den detta är faktumet så är vissa modeller är ändå alltid användbara och man behöver bara kunna hur man använder dem.. så sluta sträva efter ett perfekt modell(?)

# Självutvärdering

1. Vad tycker du har varit roligast i kunskapskontrollen? Det är alltid jätte roligt att få reslutat av koden och fatta vad man egentligen håller på med.
2. Hur har du hanterat utmaningar? Vilka lärdomar tar du med dig till framtida kurser? Utmaningarna är att programering är som att lära sig ett nytt språk och man kan enkelt glömma bort det. Men Jag kommer förstå innebörden för att klara mig igenom vardagen som om det vore ett nytt språk jag lär mig..
3. Vilket betyg anser du att du ska ha och varför? G, då jag har verkligen inte gjort mitt bästa.

# Appendix A



Figur 4: Modeller innan databehandling och slutlig resultat (Volvo bilar i blocket datan)

En bild som visar text, diagram, Graf, linje

AI-genererat innehåll kan vara felaktigt.

Figur 5: graf innan databehandling av SCB nuregistrerade bilar

# Källförteckning

R for data Science

<https://r4ds.hadley.nz/>

SCB nyregristrerade personbilar efter län

https://www.statistikdatabasen.scb.se/pxweb/sv/ssd/