Linjär regression på svenska bilmarknaden



Ian Gutierrez

EC Utbildning

Kunskapskontroll - R programmering för dataanalys

2024–04

# Abstract

The regression model built in this report studies the price of cars sold in Sweden. By analyzing the dataset me and my fellow students have gathered on the six most popular car brands in Sweden, the model will predict how different factors effect the price, such as brand, year model, fuel type and other characteristics. With the linear regression model and accuarcy metrics such as BIC and Mallow’s Cp the study aims to methodically predict, with reliable accuracy, the price variations among various car types.

The result of the study gives some insight to which factors effect the sales of cars sold in Sweden.

Innehåll

[Abstract 2](#_Toc165033777)

[1 Inledning 1](#_Toc165033778)

[2 Teori 2](#_Toc165033779)

[2.1 Regressionsmodell 2](#_Toc165033780)

[2.1.1 Multipel Linjär Regression 2](#_Toc165033781)

[2.2 Dataförberedelse 2](#_Toc165033782)

[2.2.1 Dummyvariabler 2](#_Toc165033783)

[2.3 Variabelval 2](#_Toc165033784)

[2.3.1 Forward Stepwise Selection 2](#_Toc165033785)

[2.4 Modellutvärdering 3](#_Toc165033786)

[2.4.1 RMSE (Root Mean Squared Error) 3](#_Toc165033787)

[2.4.2 MAE (Mean Absolut Error) 4](#_Toc165033788)

[2.5 Diagnostik av modellen 4](#_Toc165033789)

[2.5.1 Icke-linjäritet 4](#_Toc165033790)

[2.5.2 Korrelerade residualer 5](#_Toc165033791)

[2.5.3 Heteroskedasticitet 5](#_Toc165033792)

[2.5.4 Ej Normalfördelade Residualer 5](#_Toc165033793)

[2.5.5 Outliers 6](#_Toc165033794)

[2.5.6 ”High Leverage” Punkter 6](#_Toc165033795)

[2.5.7 Kollinearitet/Multikollinearitet 6](#_Toc165033796)

[3 Metod 7](#_Toc165033797)

[3.1 Datainsamling 7](#_Toc165033798)

[3.2 Dataförberedelse 7](#_Toc165033799)

[3.3 Variabelval 7](#_Toc165033800)

[3.4 Skapa Första Modellen 7](#_Toc165033801)

[3.5 Modellutvärdering 7](#_Toc165033802)

[4 Resultat och Diskussion 8](#_Toc165033803)

[4.1 Variabler & Koefficienter 8](#_Toc165033804)

[4.2 Jämförelse av Modeller 8](#_Toc165033805)

[4.3 Cook’s Distance 8](#_Toc165033806)

[4.4 Reflektion 8](#_Toc165033807)

[4.5 Nyregistrerade Bilar i Västra Götaland 9](#_Toc165033808)

[5 Slutsatser 10](#_Toc165033809)

[6 Referenser 11](#_Toc165033810)

# Inledning

Statistik har existerat i flera tusen år, men det är inte för än sent 1800-/tidigt 1900-tal som modern statistik började få sin grund med hjälp av statistiker så som Ronald Fisher och Karl Pearson. Idag när datorkraften har ökat avsevärt och programmering blivit mycket mer lättillgängligt så har statistiska analyser blivit mer tillgängliga för forskare och företag.

I detta arbete så kommer jag tillsammans med fem klasskamrater först samla ihop data ifrån blocket.se och sedan utföra en ”proof of concept” för att se att den data och parametrar som samlats in går att göra en modell på. Efter det så kommer en regressionsmodellering utföras, där målet är att på ett kontrollerat sätt minska antalet parametrar för att få en modell som kan med god träffsäkerhet prediktera priser på bilar mellan år 2014 och 2024.

# Teori

I denna del så kommer teorin bakom utförandet av en regressionsmodellering förklaras och gå in mer i detalj på hur man tolkar vad de olika resultaten.

## Regressionsmodell

En regressionsmodell är en metod för analys av sambandet mellan den beroende variabeln (i detta fall ”Pris”) och en eller fler oberoende variabler.

### Multipel Linjär Regression

I denna regressionsmodell så innehåller modellen mer än en oberoende variabel.

Y är den beroende variabeln, alltså det man försöker förutse.

är interceptet som när alla oberoende variabler är noll så är interceptet det förväntade medelvärdet för Y.

, är regressionskoefficienterna. Om exempelvis ändras med en enhet och alla andra oberoende variabler är likadana så är den förväntade förändringen i den beroende variabeln .

är de oberoende variablerna.

är det slumpmässiga felet som inte kan förklaras av de oberoende variablerna.

Ekvation 1 Ekvation för multipel regression

(James, Witten, Hastie, & Tibshirani, 2023, ss. 71-72)

## Dataförberedelse

Denna del handlar om de metoder som använts för att förbereda datan innan man börjar byggandet av själva modellen.

### Dummyvariabler

Dummyvariabler är ett effektivt sätt att använda sig att kategoriska värden genom att omvandla en eller fler kategorier till ja eller nej (0/1). Om man vill skapa dummyvariabler så kommer man få k – 1 kategorier där k är antalet unika kategoriska värden, så om man exempelvis har ”röd, grön, blå” som kategoriska värden så kommer man få 2 dummy variabler. (H.B., 2024)

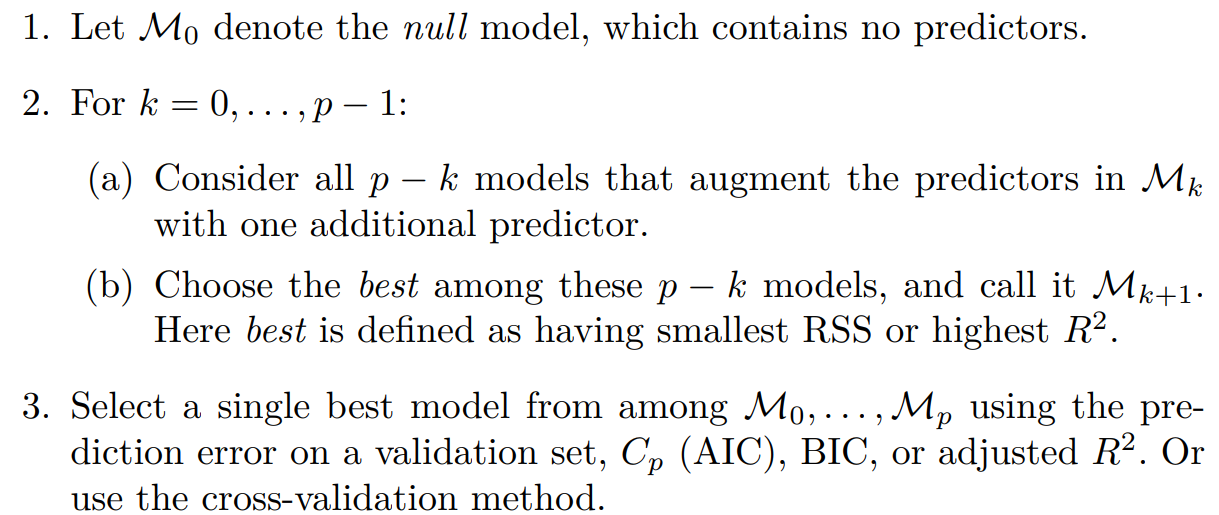
## Variabelval

Vid variabelval är det viktigt att man förstår de värden man får ut eller exempelvis hur subset selection fungerar så man väljer den som passar datan bäst.

### Forward Stepwise Selection

Forward stepwise selection är ett effektivt substitut för best subset selection då istället för att söka igenom alla möjliga kombinationer av variabler så börjar forward stepwise selection med noll variabler, sedan söker den efter den variabel som presterar bäst själv, sedan söker den efter den variabel som presterar bäst med den/de tidigare variabler till den har valt alla variabler.

(James, Witten, Hastie, & Tibshirani, 2023, ss. 229-231)



Algoritm 1 Forward stepwise selection

#### Mallow’s Cp

Mallow’s Cp används för att testa olika uppsättningar av prediktionsrvariabler för att hitta den bästa kombinationen av prediktionsrvariabler.

Ekvation 2 Mallow's Cp

(James, Witten, Hastie, & Tibshirani, 2023, ss. 233 - 235)

#### Adjusted

Adjusted utför samma syfte som Mallow’s Cp men den stora skillnaden är att adjusted tar hänsyn till antalet prediktorer i modellen och staffar modellen för onödiga tillägg av prediktorer.

Ekvation 3 Adjusted R^2

(James, Witten, Hastie, & Tibshirani, 2023, ss. 233 - 235)

#### BIC

Även BIC fungerar liknande som Mallow’s Cp och adjusted men även här har man lagt till ett straff för antalet prediktorer, skillnaden är att adjusted straffar tillägg av onödiga prediktorer medan BIC straffar för tillägg av prediktorer oavsätt om de är bra eller dåliga. BIC är mer trolig att välja en modell med färre prediktorer än adjusted , speciellt när man har få observationer.

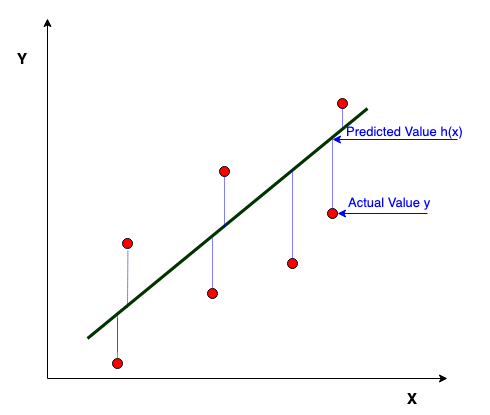
Ekvation 4 BIC

(James, Witten, Hastie, & Tibshirani, 2023, ss. 233 - 235)

## Modellutvärdering

### RMSE (Root Mean Squared Error)

RMSE är ett mått på den genomsnittliga skillnaden mellan de obersverade värdena och de predikterade värdena. (Gunjal, 2021)



Figur 1 RMSE

### MAE (Mean Absolut Error)

MAE mäter den genomsnittliga absolutskillnaden mellan det predikterade värdet och det faktiska värdet. Fördelen med MAE är att den är inte lika känslig för extrema värden eller ”outliers”. (Ahmed, 2023)

Ekvation 5 MAE

## Diagnostik av modellen

När modellen är evaluerad så är nästa steg att söka efter de sju problem man kan stöta på när man skapar en regressionsmodell: Icke-linjärt förhållande mellan den beroende och de oberoende variabler, korellerade residualer, heteroskedasticitet, ej normalfördelade residualer, outliers, ”High leverage” punkter och kollinearitet/multikollinearitet.

### Icke-linjäritet

Ett av de problemen man kan stöta på är att man ha icke linjärt samband mellan den beroende och de oberoende variablerna. Detta måste man undersöka, exempelvis genom att jämföra den beroende mot de oberoende variablerna, som i exemplet i figur 2, där vi har ett negativt linjärt samband. (James, Witten, Hastie, & Tibshirani, 2023, ss. 93 - 94)

En bild som visar text, skärmbild, Electric blue, linje

Automatiskt genererad beskrivning

Figur 2 Negativ linjärt samband

### Korrelerade residualer

Om residualer är korrelerade så finns det ett systematiskt mönster i hur de skiljer sig får de observerade värdena. Detta är ett problem då modellen kan dra felaktiga slutsatser genom att missa något viktigt samband eller inte tar hänsyn till viktiga prediktorer. (James, Witten, Hastie, & Tibshirani, 2023, ss. 94 - 95)

#### Residualer

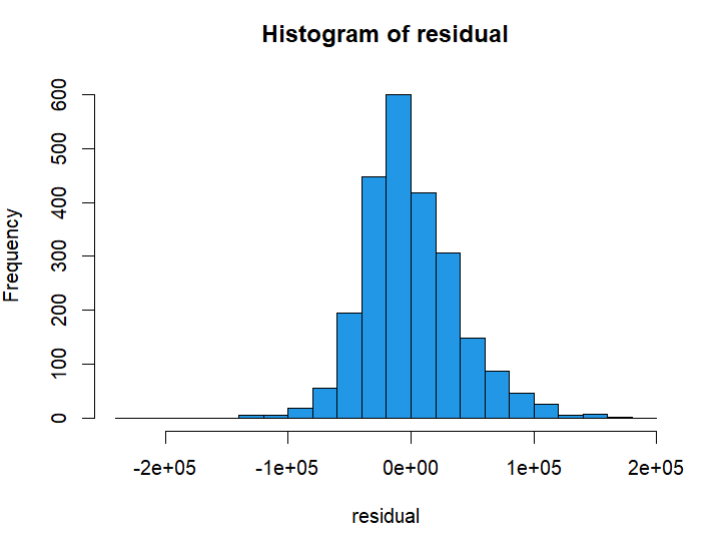
När man drar en linje genom ett antal observationer så är differensen mellan de observerade och de predikterade värden residualer. Varje observation har sin egen residual och skillnaden mellan regressionslinjen och det observerade värdet är alltså residualen. (Frisk, 2018)

### Heteroskedasticitet

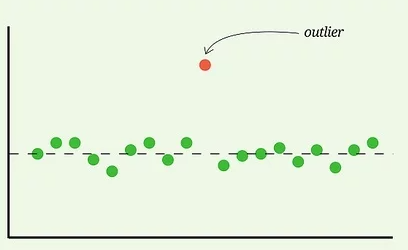
När man beräknar standardavvikelsen för betakoefficienter och de predikterade värden så ska variansen för residualerna vara konstanta. Om detta inte är fallet så kan man inte lita på inferensen och konfidensintervallet, prediktionsintervallet och hypotesprövningen bli felaktigt. Detta kallas för heteroskedasticitet. (Taboga, 2021)

### Ej Normalfördelade Residualer

Residualerna i en multipel linjär regression skall vara normalfördelade, om det inte stämmer så kan olika typer av fel uppstå i modellen. För att anta att felen är slumpmässiga så kan man plotta felens distrubition och jämföra med en normalfördelning. (Wallin & Jangenstål, 2015)



Figur 3 Exempel på normalfördelade residualer



Figur 4 Exempel på en outlier

### Outliers

En outlier är ett observerat värde som är långt ifrån det predikterade värdet och i värsta fall kan ha stark påverkan på den slutgiltiga modellen. Detta bör man undersöka och se om det bara är ett inmatningsfel i sitt dataset eller någon annan relativt simpel lösning. Om man inte hittar någon lösning på det så kan det gå bra att bara ta bort en outlier. (James, Witten, Hastie, & Tibshirani, 2023, ss. 97 - 98)

### ”High Leverage” Punkter

När ett obeserverat värde är en ”high leverage” punkt så betyder det att den enskilda punkten har en stark påverkan på den slutgiltiga modellen. (James, Witten, Hastie, & Tibshirani, 2023, ss. 98 - 100)

### Kollinearitet/Multikollinearitet

När man pratar om kollinearitet/multikollinearitet så menar man att två eller fler variabler har ett linjärt förhållande. Problemet med att ha variabler som har hög multikollinearitet är att regressionsmodellen har det svårt att förstå påverkan de individuella variablerna har på den beroende variabeln. (James, Witten, Hastie, & Tibshirani, 2023, ss. 100 - 103)

# Metod

Nu är det dags att gå igenom de olika metoder som använts för att, förbereda datan, välja variabler och bygga modellen m.m.

## Datainsamling

Datan som använts kommer från blocket.se och i vår grupp valde vi att ta varsitt märke av de sex mest sålda bilmärken i Sverige mellan 2014 till idag och mellan 150 000 kr och 500 000 kr. Vi bestämde oss först för att samla ca 50 bilar per person (totalt 300) men efter vår proof of concept så ändrade vi oss till att hämta ca 300 till per person men vissa hämtade lite fler vilken blev att vi slutade med ett dataset på ca 2700 bilar.

## Dataförberedelse

Under data förberedelsen så transformeras först de dataklasser som var ”characters” till ”factor”, sedan ta bort dubbletter, i datasettet finns ca 2700 bilar och risken för att någon har fått en dubblett är hög då vi har använt en webbskrapare. Efter det så tas den variabel anses inte ha så stor påverkan på modellen, vilket i detta fall var ”Drivning”. När den variabeln som anses inte ha så stor påverkan tagits bort så görs de kategoriska variablerna om till dummyvariabler.

## Variabelval

Nu när datan är förberedd så utförs en forward stepwise selection. Anledningen till att göra forward stepwise selection istället för best subset selection är för att det finns nästan 20 variabler och ca 2700 observationer vilket skulle ta best subset selection väldigt lång tid. Med hjälp av Mallow’s Cp, adjusted och BIC så tas det fram plots på vilka variabler som de uppskattar kommer prestera bäst och tar bort de variablerna som uppskattas inte behövas.

## Skapa Första Modellen

Innan man skapar modellen så delas data upp i träning- och testdata som är delade 70 – 30. När modellen är klar så är det dags att utvärdera!

## Modellutvärdering

Först inspekterades de värden för , RMSE och MAE för att se ungefär hur modellen presterar. Sedan genomgås de sju steg som nämns i 2.5.1 – 2.5.7. Efter varje steg så betraktas eventuella problem, och om sådana identifieras försöks dessa lösas. Om problemen inte kan lösas direkt, övervägs alternativ som att utesluta observerade värden eller variabler. Till exempel i steget ”outliers” identifieras och utesluts tre observerade värden. När samtliga sju steg har genomförts skapas en ny modell, vilken genomgår samma sju steg. När tillfredsställande resultat uppnås, anses modellen vara färdigställd.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Resultat för modeller** | | | | | |
| **Modeller** | **Antal variabler** | **Adjusted** | **BIC** | **RMSE** | **MAE** |
| lin\_mod | 19 | 0,8047 | 57213 | 39367 | 29934 |
| lin\_mod\_2 | 11 | 0,7897 | 57288 | 38924 | 30571 |

Tabell 1 Resultat för de två modeller i regressionsanalysen

# Resultat och Diskussion

## Variabler & Koefficienter

De tre variabler som hade störst påverkan på modellen var ”Bränslediesel”, ”Bränsleel” och ”Märketoyota”. Med mer tid så hade detta blivit undersökt då mina kollegor och även mig själv trodde att ”Miltal” och ”Modellår” skulle ha störst påverkan.

## Jämförelse av Modeller

Reduktionen av variabler inte haft den effekten som förväntades. Men som man kan se så är båda modellerna relativt lika i sin prediktionsförmåga och det som skiljer de åt är antalet variabler vilket i var målet med denna rapport.

## Cook’s Distance

Med plotten för cook’s distance så kan man se något intressant, det finns två kluster, en i mitten (markerat i grönt) av plotten och en längst till höger (markerat i blått). Det grönmarkerade klustret består av endast coupébilar medan det blåmarkerade klustret består av endast cabbar. Inga av dessa observationer ligger dock utanför gränsvärdet så de har inte blivit borttagna från datan.

## Reflektion

Om mer tid fanns så hade det varit intressant att pröva fler modeller och utvärdera fler kombinationer av variabler som får bäst prediktionsförmåga.

En bild som visar text, diagram, skärmbild, cirkel

Automatiskt genererad beskrivning

Figur 5 Markerat i grönt: coupé. Markerat i blått: cab.

# Slutsatser

Målet med denna rapport var att sänka antalet variabler på ett kontrollerat sätt och få en modell med god prediktionsförmåga på bilar från 2014 och 2024.

Med dummyvariabler så hade modellen från början 19 variabler (inkl. den beroende variabeln) medan den slutgiltiga modellen hade 11 (inkl. den beroende variabeln). Detta mål var satt för att minska risken för overfitting. Dock så måste man vara vaksam så att man inte tar bort för många variabler och har en modell som underfittar i stället.

Även om lin\_mod har bättre och MAE så kan man anse att lin\_mod\_2 är den bättre modellen eftersom den har färre variabler och därför bör ha bättre generaliseringsförmåga.

# Referenser

Ahmed, M. W. (den 24 08 2023). *medium.com*. Hämtat från https://medium.com/@m.waqar.ahmed/understanding-mean-absolute-error-mae-in-regression-a-practical-guide-26e80ebb97df

Dawson, C. (den 26 01 2021). *Medium.com*. Hämtat från https://medium.com/swlh/understanding-multiple-linear-regression-e0a93327e960

Frisk, E. (2018). *www.statistiskordbok.se/*. Hämtat från https://www.statistiskordbok.se/ord/residual/

Gunjal, S. (2021). *kaggle.com*. Hämtat från https://www.kaggle.com/discussions/general/215997

H.B., B. (den 23 04 2024). *https://stattrek.com/*. Hämtat från https://stattrek.com/multiple-regression/dummy-variables

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2023). *An Introduction to Statistical Learning with Applications in R.*

Taboga, M. (2021). *www.statlect.com*. Hämtat från "Heteroskedasticity", Lectures on probability theory and mathematical statistics. Kindle Direct Publishing. Online appendix.: https://www.statlect.com/glossary/heteroskedasticity

Wallin, S., & Jangenstål, S. (den 22 05 2015). *www.diva-portal.org*. Hämtat från https://www.diva-portal.org/smash/get/diva2:813986/FULLTEXT01.pdf