Multipel linjär regression

Med R-programmering



Frida Kilby

EC Utbildning

Kunskapskontroll R-programmering

202404

# Abstract

Multiple linear regression is a powerful model to use for finding patterns in several random variables. It shows how the independent variables affect one dependent variable.

To collect data for creating a mode using multiple linear regression can be done in different ways. From the simple manually work to using and API from one of the many databases with online access.

In this report the steps from collecting data in different way to create, evaluate and use a multiple linear regression model is presented. Showing that more and correct collected data is essential to get a good model for prediction values.

Innehållsförteckning

[Abstract 2](#_Toc165025819)

[1 Inledning 1](#_Toc165025820)

[2 Teori 2](#_Toc165025821)

[2.1 Multipla Linjära Regressionsmodellen 2](#_Toc165025822)

[2.1.1 Arbetsgång för regressionsmodellering 2](#_Toc165025823)

[2.1.2 Information om modellen 2](#_Toc165025824)

[2.1.3 Justering av variabler 3](#_Toc165025825)

[2.1.4 Potentiella problem 3](#_Toc165025826)

[2.2 API 5](#_Toc165025827)

[3 Metod 6](#_Toc165025828)

[3.1 Datainsamling 6](#_Toc165025829)

[3.1.1 Datainsamling i grupp 6](#_Toc165025830)

[3.1.2 Extern data från SCB 7](#_Toc165025831)

[3.1.3 Datainsamling med API 7](#_Toc165025832)

[3.2 Regressionsmodellering 8](#_Toc165025833)

[3.2.1 Teoretiska frågor 8](#_Toc165025834)

[3.2.2 Arbetsgång regressionsmodellering 9](#_Toc165025835)

[4 Resultat och Diskussion 10](#_Toc165025836)

[4.1 Datainsamling 10](#_Toc165025837)

[4.2 Manuell data 10](#_Toc165025838)

[4.2.1 Hypotes 10](#_Toc165025839)

[4.2.2 Undersöka data 10](#_Toc165025840)

[4.2.3 Estimera regressionsmodeller 10](#_Toc165025841)

[4.2.4 Diagnostisera potentiella problem 10](#_Toc165025842)

[4.2.5 Åtgärda problem 11](#_Toc165025843)

[4.2.6 Jämföra modeller 12](#_Toc165025844)

[4.2.7 Tolka regressionsmodellen 12](#_Toc165025845)

[4.3 Dataskrapning 12](#_Toc165025846)

[4.3.1 Hypotes 12](#_Toc165025847)

[4.3.2 Undersöka data 12](#_Toc165025848)

[4.3.3 Estimera regressionsmodeller 12](#_Toc165025849)

[4.3.4 Diagnostisera potentiella problem 13](#_Toc165025850)

[4.3.5 Åtgärda problem 14](#_Toc165025851)

[4.3.6 Jämföra modeller 15](#_Toc165025852)

[4.3.7 Tolka regressionsmodellen 16](#_Toc165025853)

[4.4 Extern data 16](#_Toc165025854)

[4.4.1 Hypotes 16](#_Toc165025855)

[4.4.2 Undersöka data 16](#_Toc165025856)

[4.4.3 Estimera regressionsmodeller 16](#_Toc165025857)

[4.4.4 Diagnostisera potentiella problem 17](#_Toc165025858)

[4.4.5 Åtgärda problem 18](#_Toc165025859)

[4.4.6 Jämföra modeller 18](#_Toc165025860)

[4.4.7 Tolka regressionsmodellen 19](#_Toc165025861)

[5 Slutsatser 20](#_Toc165025862)

[6 Självutvärdering 21](#_Toc165025863)

[Källförteckning 22](#_Toc165025864)

# Inledning

Multipel linjär regression modellering används för att prediktera värden som beror på en eller flera oberoende variabler, exempelvis om en sjukdom kan bero på ålder, kön, yrke, tidigare sjukdomar eller annat. Det kan förutom att användas för att prediktera värden även användas för effektestimering, alltså se hur mycket en eller flera av de oberoende variablerna påverkar det som du vill ha reda på.

Vanliga användningsområden för multipel linjär regression är inom forskning och analys för att undersöka risken att drabbas av en specifik sjukdom, inom försäljning för att prediktera värdet på ett hus eller inom industri för att mäta påverkan av förändringar på slutprodukten i tid eller kvalitet. (Multipel Regression - science, 2024)

I den här rapporten ska data samlas in såväl manuellt som med hjälp av externa datakällor och analyseras med hjälp av multipel linjär regression. Det elementära i rapporten är både insamlingen av data samt skapandet och utvärderingen av modell för prediktering.

Syftet med rapporten är ge en grundläggande kunskap för flödet i modellskapande av multipla linjära regressioner samt tanke- och arbetssätt för insamling av relevant data på ett adekvat sätt.

För att uppfylla syftet så kommer följande frågeställning att besvaras:

1. Hur ser flödet för modell med multipel linjär regression ut
2. Hur ska data samlas in på ett relevant sätt
3. Hur görs statistisk inferens med modell skapad i R

# Teori

## Multipla Linjära Regressionsmodellen

Den multipla linjära regressionsmodellen:

B kallas för Betaparametrar och anger den genomsnittliga effekten på Y när xi ökar givet att alla andra variabler är fixa. Hur mycket som varje variabel påverkar.  
E är epsilon och är det som blir över, den slumpmässiga feltermen eller även kallat residual.

Multipel linjär regression används för att undersöka om det finns något statistiskt samband mellan den beroende variabeln (*Y*) och en eller flera av de oberoende variablerna (). (Multipel Regression - science, 2024)

### Arbetsgång för regressionsmodellering

1. Skapa en hypotes som ska undersökas
2. Undersök data som ska användas
3. Skapa regressionsmodell
4. Diagnostisera eventuella problem enligt rubrik 2.1.3
5. Åtgärda eventuella problem och gå tillbaka till punkt 3.
6. Jämför de skapade modellerna om fler än en
7. Tolka regressionsmodellens koefficienter, hypotesprövning eller utför prediktioner och analysera dessa.

### Information om modellen

#### Adjusterad R^2

#### Förklarar modellens varians i Y värdet, om Adjusterad R-squared är exempelvis 0,4 så kan 40 % av beroende variabelns förändring bero på modellens påverkan.

#### F-test

Används för att kontrollera om minst en oberoende variabel är användbar för att prediktera den oberoende variabeln. Detta görs genom hypotesprövning med f-test och p-värde. Om P-värdet för f-testet är under 5 procent kan vi förkasta nollhypotes och anta att minst en oberoende variabel kan användas för att prediktera den beroende variabeln. Det är viktigt att kontrollera p-värdet för f-test då detta beräknas för hela modellen oavsett antal variabler, till skillnad mot t-test som kontrollerar var variabel för sig.

#### Statistisk signifikans – p-värde

För att se om det finns något samband mellan Y och X kan en nollhypotes etableras. Nollhypotes kan vara att det inte finns någon statistisk signifikans mellan X och Y vilket då innebär att lutningen för X är lika med noll. Det finns då inget samband mellan den beroende och den oberoende variabeln.

Detta kan utvärderas med hjälp av t-test, vilket visas som sannolikheten för att nollhypotes är sann. Är denna under 5 % så förkastar vi nollhypotes och därmed finns statistisk signifikans mellan Y och X. (Multipel Regression - science, 2024)

### Justering av variabler

#### Dummy variabler

Linjär regression fungerar enbart på numeriska värden. Detta innebär att när det finns olika klasser, regioner, modeller eller liknande behöver dessa göras om till dummy variabler för att kunna användas i modellerna. Dummy variabler görs antingen i ordinal eller nominal skala, ordinal skala används för variabler som inte har en inbördes ordning. Då skapas det k minus 1 stycken nya variabler. Finns det inbördes ordning används nominal skala och variabeln sätts till 1,2,3…, k.

#### Interaktionseffekt

För variabler som tillsammans kan ha påverkan på den beroende variabeln kan deras interaktionseffekt användas. Då läggs interaktionseffekten till tillsammans med variablerna i dess ursprungsform enligt hierarki principen.

### Potentiella problem

#### Linjärt förhållande, linjäritet

Genom att visualisera residualerna i en figur går det uttyda om det finns ett linjärt förhållande mellan den beroende och de oberoende variablerna för en modell. Det ska i figuren inte synas något tydligt mönster i residualerna. Är det så att det inte finns något linjärt samband i modellen kan inte en linjär regressionsmodell användas. Avhjälpning för det här problemet kan vara att transformera en eller flera av de oberoende variablerna med ex. x^2, X^0,5 eller log(x).

#### Korrelerade residualer

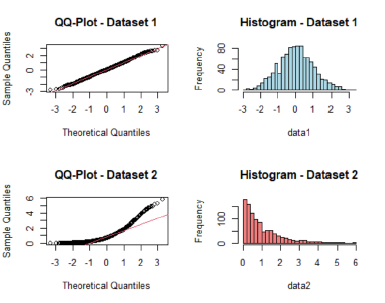
Vid beräkning av standardavvikelse mellan de predikterade värdena och betakoefficienterna så förutsätter vi att residualerna är oberoende. Om detta inte är sant så skattas standardavvikelsen, konfidensintervallet, prediktionsintervallet och p-värdena lägre än vad de bör vara. Vi kan då inte lita på beräkningarna som modellen gör. Detta sker ofta inom tidsserier, men kan även ske utanför sådan typ av data. För att lösa problemet är det av vikt att fundera kring hur data samlas in så att det exempelvis inte finns släktskap mellan olika observationer för att minska korrelation mellan dessa.

#### Icke konstant varians, heteroskedacitet

Vid beräkning av standardavvikelse för predikterade värden och betakoefficienterna så antas att variansen är konstant för de olika residualerna. Detta kallas för Homoskedacitet. Är det istället så att varians varierar mellan de olika residualerna blir inte konfidensintervall, prediktionsintervall och hypotestest korrekta. Detta kan avhjälpas genom att transformera den beroende variabeln med en konkav funktion, dock behöver modellens prediktering sedan transformeras tillbaka i slutändan. Ett sätt att kontrollera detta är att se om data i diagnostisk figur nere till vänster är relativt rak, är det så att data är formad mer som en tratt så finns indikationer på heteroskedacitet.

#### Normalfördelning residualer

Kontroll av residualernas normalfördelning kan kontrolleras med QQ-plot. Är data normalfördelad så följer data linje i denna figur. Sticker data iväg i en mer kurvad form är data exponentialfördelad. Det kan då vara en god idé att identifiera ev. outliers, eller kontrollera om det går att transformera den beroende variabeln eller transformera av de oberoende variablerna.



Figur 1 från Antonios föreläsningsanteckningar

#### Outliers

Om en punkt är långt ifrån det sanna värdet kan detta vara en outlier. För att identifiera outliers används figur som visar studentized residuals. Punkter med mer än 3 i värde är möjliga outliers, om en outlier består av felaktigt inmatad data är det självklart att denna ska raderas. Men de kan också uppkomma för att modellen är felaktigt specificerad vilket avhjälps genom att skapa en annan modell eller ändra skapad modell.

#### High Leverage punkter, Influencal points

Dessa punkter är enkla att identifiera med bara en variabel, men med flera variabler kan det vara precis tvärt om. Detta är predikteringar som skiljer sig väldigt mycket från observationens värde. Punkterna kontrolleras med Leverage statistic. Är detta värde större än 2 gånger ((p+1)7n) så beräknas det generellt som en high leverage punkt. Detta kontrolleras enklast genom figur som visar Leverage och Studentized residualer, vilket visar både om den är en outlier (har högt Studentized värde) och har hög leverage. Om så är fallet har den observationen stor påverkan på modellen och är således mycket problematisk.

#### Kollinearitet, multikollinearitet

Kontrolleras med hjälp av funktionen vif() som står för variance inflation factor. För denna ska värden vara under 5 (eller ibland under 10) för att inte multikolllinearitet ska finnas. Om variabler har över värdet fem kan dessa variabler påverka varandra och ha linjärt samband (perfekt multikollinearitet), det kan då finnas vinning i att kombinera dessa.

En bild som visar text, diagram, linje, Graf

Automatiskt genererad beskrivning

Figur 2 Diagnostic plots, figurer för att kontrollera modells eventuella problem

## API

Ett API är ett applikationsprogrammeringsgränsnitt. Det är ett sätt för program och system att prata med en applikation och används som ett översättningslager för att det ska gå enklare att skicka signaler eller kod till och från en applikation.

# Metod

## Datainsamling

### Datainsamling i grupp

Gruppen för datainsamling bestod av:

* Frida Kilby
* William Blennow
* Melike Koyuncu
* Siarhei Thor Fedatsenka
* Dan Heikenberg
* Daniel Hemgren
* Xianyong Yang
* Natalie Dobrovolska
* Khaldoun Agha

Data samlades in på två olika sätt från webbplatsen blocket.se som är en digital marknadsplats. De två sätten som data samlades in på var manuellt och genom dataskrapning. Manuellt samlades ungefär 30 poster per person in och genom dataskrapning samlades ungefär 10 000 poster in. För den manuella skrapningen gjordes ett urval på El-bilar från företag till fast pris och varje person hämtade poster från en bestämd region. För dataskrapningen gjordes inget urval. Riktlinjerna gjordes tillsammans vid första träffen och alla var delaktiga i att sätta upp efter vilka kategorier och hur vi skulle arbeta.

Bra i grupparbetet var att alla fick komma till tals, vi hade en naturlig ledare som initialt styrde upp arbetet men inga beslut togs utan att alla fick göra sin röst hörd. Vi kom på så vi snabbt igång med datainsamlingen och kunde snabbt arbeta oss igenom den manuella insamlingen. Utvecklingsområden för grupparbetet är kanske att arbeta i något mindre grupper, det blev snabbt en ganska stor grupp för att lösa en liten uppgift. Detta fick dock sin naturliga indelning när en del visade intresse för dataskrapning och en annan del visade mer intresse för det manuella arbetet.

Mina styrkor är nog att jag arbetar på att hålla tillbaka tankar tills att se om de redan dyker upp och sedan fylla på med sådant som jag har erfarenhet av. Ett utvecklingsområde är att kanske inte vara för rak i en ny grupp då detta kan uppfattas som att tålamodet är ganska kort

Något som skulle kunna göras annorlunda är att test att samla data på de avgränsningar som är satta innan arbetet fortgår. På så vis ser alla hur data enklast samlas in och kan arbeta vidare individuella snabbare.

Under datainsamlingen fördes diskussioner kring gränsdragningar och urval.  
Gruppen valde tidigt att göra urval för elbilar och genom region för att undvika att samla dubbletter, insamlingen gjordes under några dagar i april vilket enbart ger en ögonblicksbild i det stora hela avseende försäljningen av elbilar i just dessa regioner.

Diskussion fördes även om möjligheten att knyta den data som vi samlar in mot data från SCB, detta verkar dock vara väldigt svårt i och med att vi får en ögonblicksbild och SCB levererar mer data över tid. Många dialoger fördes kring vilka frågeställningar som rapporten skulle kunna anta för att använda bägge dataset för en modell i regressionssyfte.

Avgränsningar avseende elbilar skapade problem i och med att den vanliga faktarutan på blocket då inte innehåller lika avancerad information som den gör för andra drivmedel, vi försökte samla data från transportstyrelse ruta också men då denna ruta verkar vara tillval hade inte alla bilar information som kunde hämtas i denna.

Ingen större säkerställning på att data kan läsas in i R gjordes, men då vi hade resurser i form av en deltagare som använde python för att modifiera data till en csv fil kände gruppen sig trygg med att inläsning i R kommer att kunna ske.

Efter nedladdning av komplett data från gruppens insamling rensades manuellt insamlad data med hjälp av Excel. Detta då filen består av 255 observationer med 12 variabler och på så sätt är enkel att överskåda.

Data från dataskrapning som gruppkollega har laddat ner har gåtts igenom med hjälp av Python laddades ner i form av CSV-fil. Denna lästes in i PowerBi för vidare rensning, detta för att inläsning i Excel medförde svårigheter med svenska tecken. När data rensats och formaterats i PowerBi kopierades tabell in i Excel för vidare inläsning i R.

**En bild som visar text, skärmbild, nummer, Teckensnitt

Automatiskt genererad beskrivning**

Figur 3 Utklipp från Excel på manuell hantering av data

### Extern data från SCB

För att från SCB samla liknande data som skett i den manuella insamlingen valdes data från databasen Transporter och kommunikation, Fordonsstatistik, Fordonsstatistik, Nyregistrerade personbilar efter län och kommun samt drivmedel från SCB (Statistikdatabasen, 2024). Urval i databasen gjordes enligt tabell nedan och data laddades ner direkt från webbsidan i Excel- och CSV-format. Nedladdning av data skedde även genom API och direkt in som dataframe i R, läs mer om det i kapitel 3.3.

|  |  |  |
| --- | --- | --- |
| **Regioner** | **Drivmedel** | **Tid** |
| 01 Stockholm | Bensin | Från 2006M1 till 2024M03 |
| 03 Uppsala | Diesel |  |
| 04 Södermanland | El |  |
| 09 Gotland | Elhybrid |  |
| 12 Skåne | Laddhybrid |  |
| 18 Örebro | Etanol/etanol flex |  |
| 20 Dalarna | Gas/gas flexifuel |  |
| 1480 Göteborg | Övriga bränslen |  |

Tabell 1 Valt urval på SCB

### Datainsamling med API

För att kunna ladda ner data via API sparades den sökning som gjordes via SCB:s hemsida i kod för att passa Json i Python. Extraheringen av data gjordes därefter med R via paketet pxweb. (Demonstration of the pxweb R package, 2024) (Christian Lindell, 2024) Data extraherades dels med hjälp av interactive\_pxweb() vilket är en funktion som hjälper till att hämta data med olika val genom SCB:s dataset. Data hämtades också direkt med kod, med hjälp av funktionen pxweb\_get().

## Regressionsmodellering

### 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-Quantlie plot är en plot som visar sample quantlies mot theoretical quantlies. Denna plot används för att identifiera om data är normalfördelad eller inte. För normalfördelad data är denna plot nästan en rät linje med observationerna tätt samlade runt linjen, medan för andra set av observationer så visas extremvärden på ett tydligt sätt genom att vara över eller under linjen i QQ-ploten och då är data inte 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?  
   Det är ett korrekt sätt att tolka det. I regressionsanalys tränas modeller på samma sätt som i maskinlärning, dock används ofta samma modeller för att även utföra statistisk inferens så som hypotesprövning, konfidensintervall eller prediktionsintervall på prediktioner från en tränad modell.
3. Vad är skillnaden på ”konfidensintervall” och ”prediktionsintervall” för predikterade värden?  
   Konfidensintervallet visar intervallet för snittet. Medan prediktionsintervallet visar intervallet för en individ inkl. epsilon (osäkerheten). Detta gör att konfidensintervallen är smalare än prediktionsintervallet för samma predikterade värden.
4. Den multipla linjära regressionsmodellen kan skrivas som: 𝑌=𝛽0+𝛽1𝑥1+𝛽1𝑥2+...+𝛽𝑝𝑥𝑝 +𝜀 .   
   Hur tolkas beta parametrarna?

Betaparameterarna är den genomsnittliga effekten på Y när xi ökar med en enhet givet att alla andra variabler är fixa. Vad som blir över är epsilon som även kallas osäkerheten, residual eller den slumpmässiga feltermen.

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?   
   Det är korrekt Hassan. Om du använder dig av BIC så behöver du inte använda ett test eller validerings set. Detta för att test error skattas från traning error. Det finns ytterligare några metoder att använda som är AIC eller Adjusted R^2.
2. Förklara algoritmen nedan för ”Best subset selection”



1. I algoritmen för best subset selections görs först en modell som beräknar medelvärdet för varje observation, denna kallas M0.
2. Därefter görs beräkningar på varje variabel ensam och bästa modellen identifieras, därefter två variabler tillsammans och bästa modellen identifieras. Detta stegs görs tills k=p, vilket innebär att om vi har 4 oberoende variabler så görs detta i 4 steg. Dessa modeller kallas Mk och identifieras som bästa med hjälp av RSS och R^2.
3. När bästa modellen identifierats för varje antal av variabler 1,2,3,4 så väljs bästa modellen bland M0 till Mp mer hjälp av AIC, BIC eller Adjusterad R^2.
4. Ett citat från statistikern George Box är: “All models are wrong, some are useful.”Förklara vad som menas med det citatet.  
   Med det citatet menas att det inte finns någon perfekt modell. Utan att vi ska fokusera på om någonting kan bli applicerbart i vardagen på ett användbart sätt istället för att debattera om varje svar är korrekt i alla lägen.

### Arbetsgång regressionsmodellering

För regressionsmodellering kommer flera modeller att skapas, detta då de olika dataseten innehåller olika variabler med olika innehåll. Dataseten kommer nedan att kallas Manuell för setet bestående av 255 manuellt insamlade observationer, Dataskrapning för setet innehållande strax över 10 000 observationer och Externt dataset för datasetet innehållandes 12 000 observationer nedladdade från SCBs sida.

Arbetsgång för regressionsmodellering är:

1. Hypotes
2. Undersöka data
3. Estimera regressionsmodeller
4. Diagnostisera potentiella problem
5. Åtgärda problem
6. Jämföra modeller
7. Tolka regressionsmodellen

# Resultat och Diskussion

## Datainsamling

Manuell datainsamling är väldigt tidskrävande och sätter stor tillit till den som arbetar med insamlingen. Den möjliggör dock att kunna hämta extra data som finns i text eller flikar som inte dataskrapning kommer åt. Dataskrapning är ett väldigt bra sätt att samla mycket data snabbt, dock erhölls många tomma rader vilket i just linjär regression gör att modellerna inte kan tränas på ett bra sätt på de delar av insamla data. Nedladdning av data via API och SCB:s hemsida var väldigt smidigt dock var denna data tidsbaserad vilket gjorde den svår att arbeta med i linjär regressionsanalys.

## Manuell data

### Hypotes

Manuella, en elbil i Stockholm är dyrare än en elbil i övriga län.

### Undersöka data

I dataset med manuellt insamlad data finns 255 rader och 12 olika variabler. Av dessa variabler är 4 numeriska, varav en variabel är Priset vilket för hypotesen kommer att vara den beroende variabeln. Övriga variabler med numeriska värden kommer vara de oberoende variablerna, dessa är Miles, ModelYear och Horsepower. För att få med Location och kunna testa hypotesen görs Location om till Dummy Variables. Då några av variablerna innehåller tomma rader sätts de manuellt till värde 1 för Miles då dessa ser ut att vara nya bilar och för Horsepower sätts saknade värden för Kia EV6 SUV och EV9 SUV manuellt till värde från annan lika post. De insatserna minskade tomma rader från 11 till 5 st.

En bild som visar text, skärmbild, Teckensnitt, nummer

Automatiskt genererad beskrivning

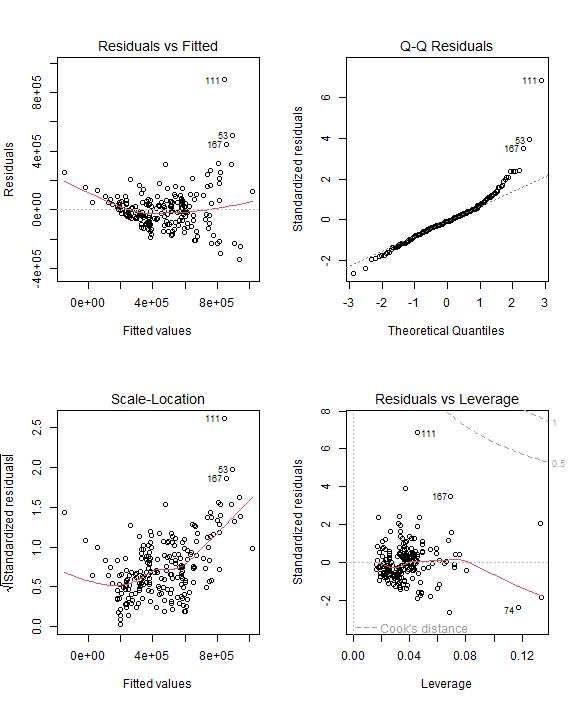
Figur 5 Utklipp över manuellt inläst data

### Estimera regressionsmodeller

En modell skapas med pris som beroende variabel och Miles, ModelYear och Horsepower som oberoende variabler. Samma modell skapas men Dummy Variabler för Location läggs till.

### Diagnostisera potentiella problem

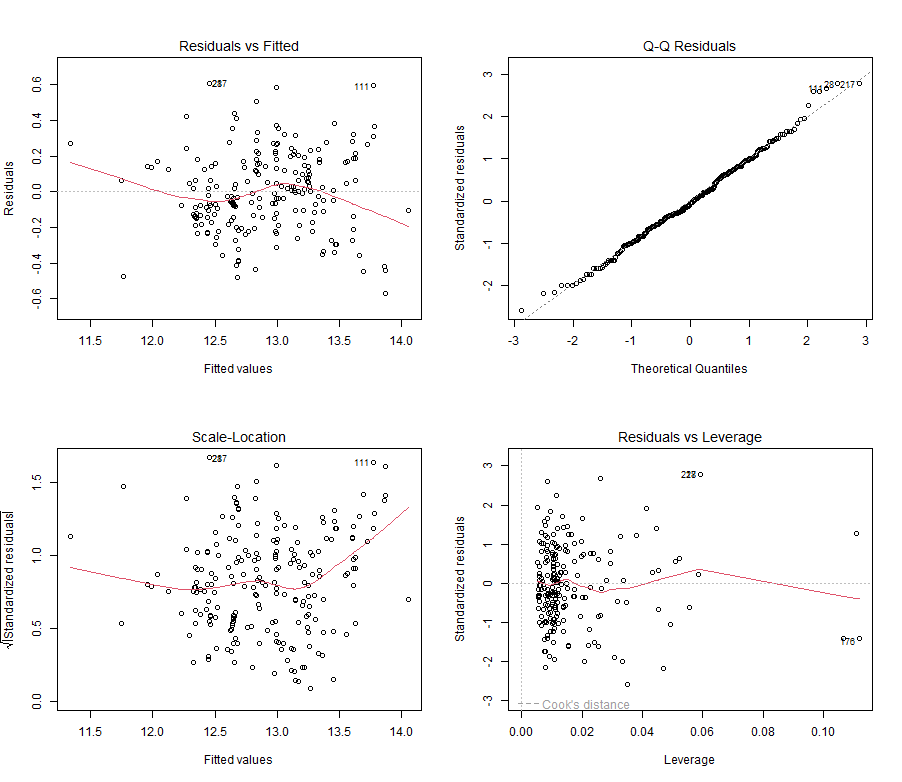
Med hjälp av diagnostic-plots kontrolleras modellens data. I första modellen ser sambandet ser ut att vara icke-linjärt enligt första bilden då linjen ser mer ut som en parabol. Och nedre vänstra bilden visar en tendens av hetreoskedacitet, Q-Q plotten visar att det finns viss normalfördelning av data. Det finns några punkter som sticker ut som outliers men ingen som verkar vara Influencal point. Diagnostical plots ser i stort sett lika ut för modell med Locations inlagt.



Figur 6 Diagnostics plots för Manuell modell 1

### Åtgärda problem

Heteroskedacitet kan åtgärdas genom att logaritmera y-variabeln. Test av detta visar att variansen mellan residualerna blir mindre, även potentiella outliers blir eliminerade.



Figur 7 Diagnostic plots över Manuell modell efter logaritmering

### Jämföra modeller

|  |  |  |  |
| --- | --- | --- | --- |
| **Modell** | **Adjusted R-Squared** | **P-värde F-test** | **BIC** |
| MD3 | 0,7946 | 2.2 e-16 | -13,14618 |
| MD4 | 0,8008 | 2.2 e-16 | 1,564457 |

Tabell 2 Jämförelse av data mellan modeller

Modellerna rätt lika avseende adjusterad R-squared och P-värde, BIC skiljer sig däremot mellan modellerna. Att värdet för BIC blir negativt beror på att den modellen även förutser negativa värden. Det finns två olika sätt att beräkna BIC och i det här fallet används den beräkningen som kan ge negativa värden. MD 4 väljs för att testa hypotesen.

### Tolka regressionsmodellen

MD 4 valdes för att testa hypotesen; en elbil i Stockholm är dyrare än en elbil i övriga län.  
När ny data lades in i form av 6 bilar från 2022 som gått 3000 mi har 120 hästkrafter och som var och en tillhör varje location så erhölls följande prediktioner och inferens:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Konfidensintervall | | | | | Prediktionsintervall | | |
| Manuellt | fit | lwr | upr | **diff** | lwr | upr | **diff** |
| Gotland | 328 752,4 | 300 993,6 | 359 071,1 | **58 077,5** | 210 692,3 | 512 966,7 | **302 274,4** |
| Dalarna | 313 706,5 | 294 787,4 | 333 839,9 | **39 052,5** | 201 940,1 | 487 331,6 | **285 391,5** |
| Uppsala | 299 974,3 | 275 929,7 | 326 114,1 | **50 184,4** | 192 422,2 | 467 641,2 | **275 219,0** |
| Stockholm | 293 961,2 | 268 767,7 | 321 516,2 | **52 748,5** | 188 343,1 | 458 807,4 | **270 464,3** |
| Göteborg | 289 276,6 | 266 721,1 | 313 739,5 | **47 018,4** | 185 641,7 | 450 765,8 | **265 124,1** |
| Örebro | 281 853,8 | 263 147,8 | 301 889,5 | **38 741,7** | 181 261,6 | 438 270,0 | **257 008,4** |

Tabell 3 Resultat för inlagd data

Som konfidensintervallet visar så är el-bilar dyrast att köpa på Gotland och därefter Dalarna. Det kan förmodligen bero på urval och efterfrågan. Stockholm ligger enligt denna modell på en fjärdeplats och således är inte vår hypotes sann. En elbil är inte dyrare i Stockholm än i övriga län.

## Dataskrapning

### Hypotes

Dataskrapning, en bil oavsett bränsle i Stockholm är dyrare än en bil i övriga län.

### Undersöka data

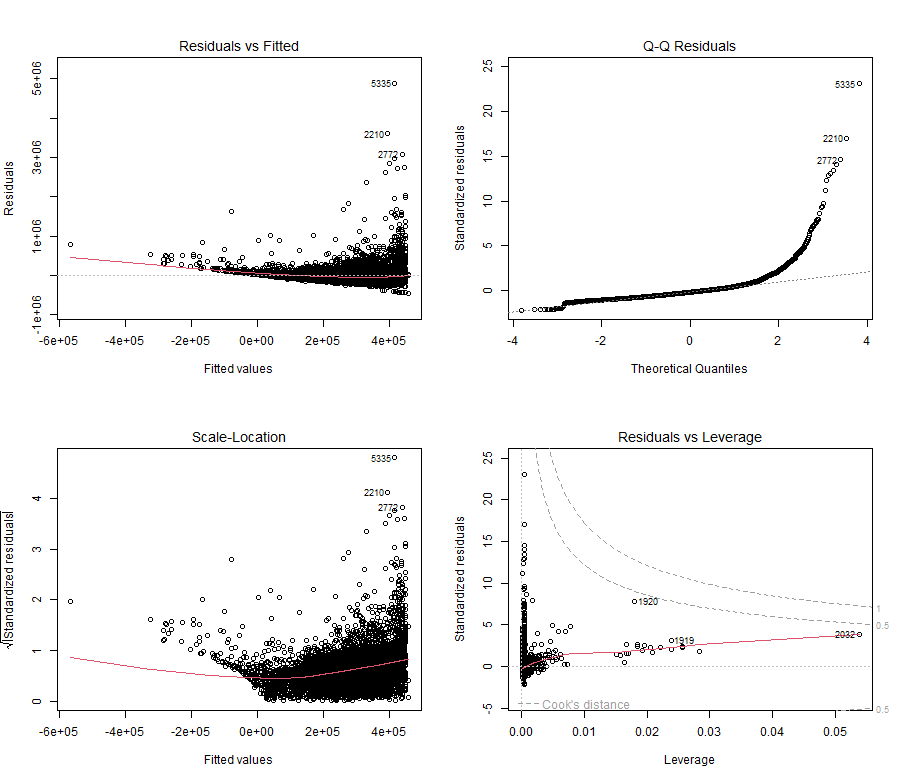
I dataset med dataskrapning finns 11525 observation med 8 variabler. Av dessa variabler var Year, Miles och Price numeriska. Förutom dessa ska Location göras om till Dummy Variabler för att kunna testa hypotesen. Platserna i datasetet är fler än i manuella datasetet, så för att få en viss koppling mellan hypoteserna tas län som inte är lika det manuella datasetet bort. Kvar blir då 7315 observationer.

### Estimera regressionsmodeller

Modeller skapas med beroende variabeln pris. Första modellen innehåller Year och Miles som oberoende variabler. Andra modellen innehåller utöver de två första variablerna även Location som Dummy Variabel, där varje län är en egen variabel.

### Diagnostisera potentiella problem

Diagnostic plot för modell med Year och Miles som oberoende variabler visar tecken på heteroskedacitet och att data är inte normalfördelad. Det finns även tendenser på att vissa punkter kan vara outliers i och med att de har värde långt över 3. Dock verkar ingen punkt vara en influencal punkt. Modell innehållande locations ser ut på ett liknande sätt.



Figur 8 Diagnostic plots för dataskrapad data modell 1

Kontroll av multikollinearitet för bägge modellerna visar att i modellen med location som dummies så finns viss multikollinearitet för Stockholm, Göteborg och Örebro då dessa värden är högre än 5 eller 10 som är en vanligt förekommande gräns.

|  |  |  |
| --- | --- | --- |
| **Variabler** | **DD1** | **DD2** |
| **Miles** | 1,480764 | 1,489457 |
| **Year** | 1,480764 | 1,502876 |
| **Stockholm** |  | 10,921334 |
| **Gotland** |  | 1,073556 |
| **Göteborg** |  | 5,618959 |
| **Skåne** |  | 6,804787 |
| **Södermanland** |  | 3,021406 |
| **Uppsala** |  | 3,542749 |
| **Örebro** |  | 3,521508 |

Tabell 5 Resultat för vif() funktion av variabler i modellerna med dataskrapad data

### Åtgärda problem

En kontroll av punkter som kan vara outliers på skalan standardized residuals visar att detta är bilar med högre pris än övriga bilar. Detta då de observationerna är nya och av dyrare modell, Mercedes-Benz och Ferrari. Dessa kan således inte tas bort utan behövs för rimlighet i data. För bilar med hög Leverage visar det sig vara bilar med priset 1, av vilken den bil som stack ut mest dessutom var riktigt gammal. Dessa bör egentligen ha ett högre värde då kontakt förmodligen ska ske med säljare för att erhålla faktiskt pris på dessa, vilket innebär att data inte är korrekt och ska således raderas. Radering sker genom att dessa sätts till NaN och påverkar 4 observationer.

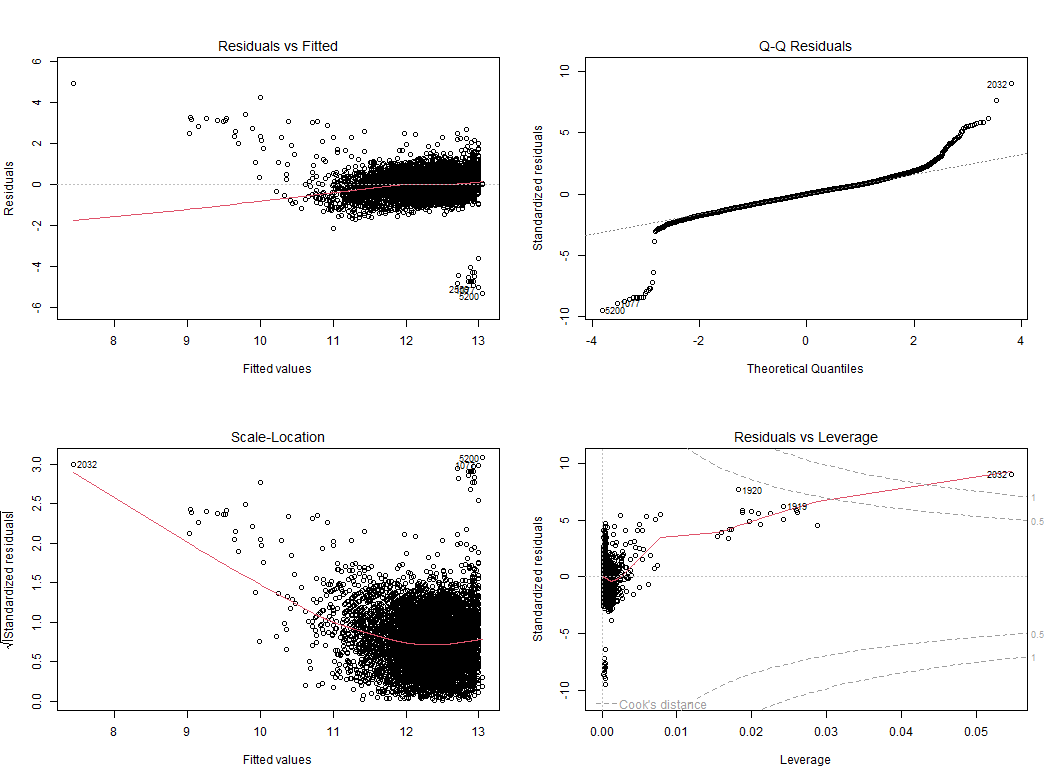
Därefter logaritmeras den beroende variabeln för att se om detta kan påverka heteroskedaciteten, vilket verkar lösa problemet. Punkter som fortsatt ser ut att vara influencal points är gammal bil med högt pris.

En korrelationsmatris för de oberoende variablerna skapades men visade inte någon tydlig korrelation mellan en eller flera variabler så detta problem kvarstå utan åtgärd.

En bild som visar text, Teckensnitt, nummer, skärmbild

Automatiskt genererad beskrivning

Efter borttagning av felaktig data och logaritmering av Y så ser diagnostical plot för modellerna bättre ut. De outliers som finns kvar är influential points och är de riktigt gamla bilarna som har ett högt pris. Dessa får vara kvar då de är korrekt inmatad data och således ska påverka modellen.



Figur 9 Diagnostic plots över logaritmerad modell

DD4 modell ser ut på ett liknande sätt, men där är inte punkt 2032 lika influerande som i modell DD3.  
En bild som visar text, diagram, linje, skärmbild

Automatiskt genererad beskrivning

Figur 10 Diagnostic plots för logaritmerad modell med locations

### Jämföra modeller

|  |  |  |  |
| --- | --- | --- | --- |
| **Modell** | **Adjusted R-Squared** | **P-värde F-test** | **BIC** |
| DD3 | 0,4088 | 2.2e-16 | 12303,19 |
| DD4 | 0,4118 | 2.2e-16 | 12321,4 |

Tabell 6 Jämförelse mellan modellerna

Modellerna får ett rätt lika såväl adjusterat R-squared värde som BIC, modell DD4 väljs för att använda till hypotesen.

### Tolka regressionsmodellen

DD4 valdes för att testa hypotesen; en bil oavsett bränsle i Stockholm är dyrare än en bil oavsett bränsle i övriga län. När ny data lades in i form av 6 bilar som gått 3000 mil, år 2022 och som var och en tillhör varje location så erhölls följande data:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Konfidensintervall | | | | | Prediktionsintervall | | |
|  | fit | lwr | upr | **diff** | lwr | upr | **diff** |
| Stockholm | 364 758,0 | 356 757,3 | 372 938,0 | **16 180,7** | 121 991,7 | 1 090 634,7 | **968 643,0** |
| Göteborg | 361 329,2 | 348 212,4 | 374 940,1 | **26 727,7** | 120 796,7 | 1 080 814,3 | **960 017,6** |
| Örebro | 343 789,5 | 236 269,6 | 362 250,2 | **125 980,6** | 114 861,2 | 1 028 991,5 | **914 130,3** |
| Uppsala | 343 504,0 | 326 252,2 | 361 668,0 | **35 415,8** | 114 770,1 | 1 028 099,2 | **913 329,1** |
| Skåne | 339 960,6 | 328 608,1 | 351 705,3 | **23 097,2** | 113 664,0 | 1 016 797,1 | **903 133,1** |
| Södermanland | 332 751,0 | 313 858,2 | 352 781,0 | **38 922,8** | 111 138,7 | 996 261,5 | **885 122,8** |
| Gotland | 304 440,5 | 224 424,0 | 412 986,1 | **188 562,1** | 97 685,6 | 948 799,2 | **851 113,6** |
| Dalarna | 295 500,2 | 271 714,8 | 321 367,7 | **49 652,9** | 98 534,1 | 886 194,7 | **787 660,6** |

Tabell 7 Resultat för inlagd data

Enligt modellen är en bil från 2022 och som kört 3000 mil dyrast i Stockholm oavsett bränsletyp. Konfidensintervallet för just Stockholm är rätt smalt så modellen verkar säker på sin sak. I detta fall kan vi bekräfta vår hypotes på den data som samlades in för just den dagen.

## Extern data

### Hypotes

SCB-set, kommer elbilar att vara den mest förekommande nyregistrerade bilen under 2024.

### Undersöka data

Externt dataset består av 12264 observation med 4 variabler. Detta dataset skiljer innehållsmässigt emot övriga dataset då data är tidsbaserad och varje post anger hur många bilar som var nyregistrerade för en viss tidsperiod. För att testa hypotes om en typ av drivmedel så görs drivmedel om till Dummy Variabler. Även tidsramen som är år + månad tillsammans görs om så att summering kan ske på år, summering sker därefter även på region för att visa alla regioner tillsammans.

En bild som visar text, skärmbild, Teckensnitt, nummer

Automatiskt genererad beskrivning

En bild som visar text, skärmbild, Teckensnitt, nummer

Automatiskt genererad beskrivning

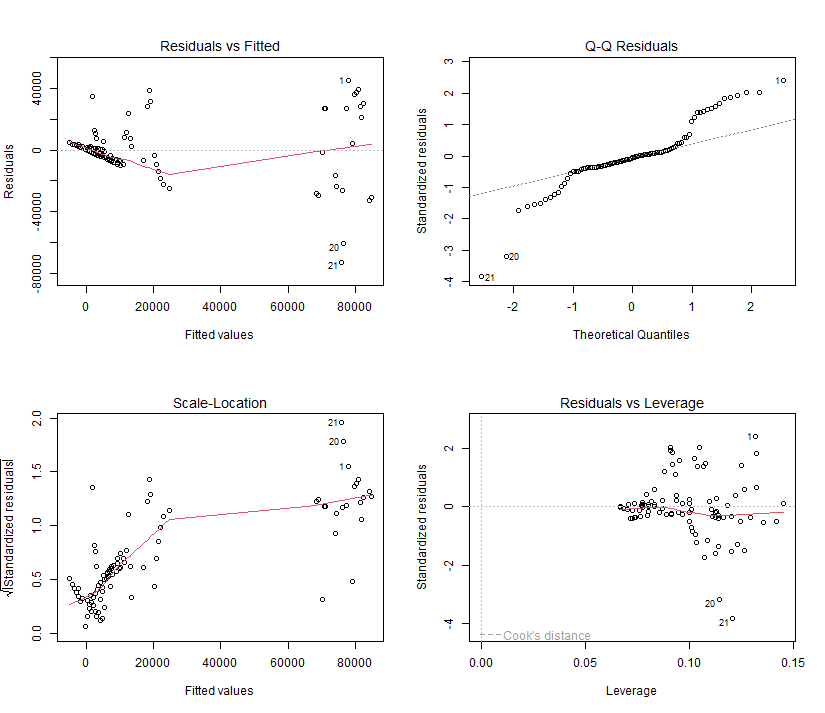
Figur 11 Före och efter uppdelning på år och månad

### Estimera regressionsmodeller

För att skapa en modell görs data om till tränings, validerings och test-set. På träningsset skapas en modell med bilar som beroende variabel och bränsletyp, år och region som oberoende variabler.

### Diagnostisera potentiella problem

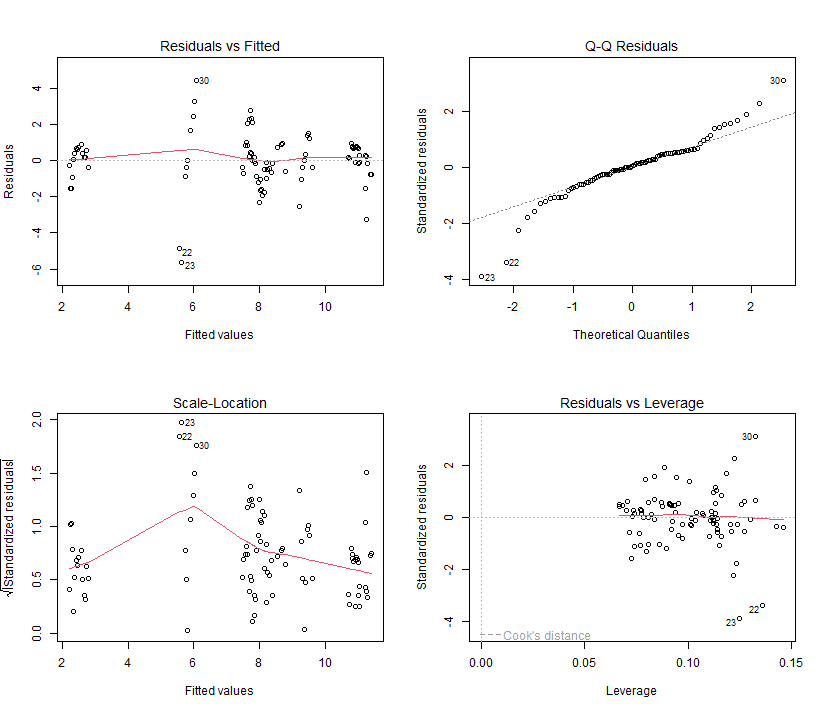
Det ser ut som att det finns linjärt samband med År, Bränsletyp och Region som de oberoende variablerna för att kunna prediktera antal nyregistrerade bilar. Data ser inte ut att vara normalfördelad och det finns lite tecken på heteroskedacitet.



Figur 12 Diagnostic Plot från extern data modell

### Åtgärda problem

Vid försök att åtgärda heteroskedasitet genom att logaritmera y-variabeln så går inte detta då det finns värden för y som är noll. Dessa värden tas bort genom att sätta dem till NaN. Vid logaritmering av y-variabel så försvinner nu hetroskedaciteten.



Figur 13 Diagnostic plot för extern data efter logaritmering

Då data kommer från tidsserie vilket är svårt att arbeta med så kommer data även att behandlas i Excel för hypotesprövning. I Excel summeras antal registrerade personbilar per år och regioner och ett diagram skapas för jämförelse av bränsletyper.

### Jämföra modeller

Vid jämförelse mellan modellerna med hjälp av validerings-setet erhölls följande resultat.

|  |  |  |  |
| --- | --- | --- | --- |
| Model | RMSE | Adjusted R-Squared | BIC |
| ed3 | 18929,6 | 0,6857184 | 2098,087 |
| ed6 | 14121,36 | 0,7661964 | 368,633 |

Andra modellen med logaritmerad y-variabel kommer att användas tillsammans med diagram från Excel för vidare arbete.

### Tolka regressionsmodellen

På kurvan i diagram skapat i Excel så syns en tydlig trend på el-bilar för de senaste åren, så enligt den data som är insamlad från SCB tyder mycket på att el-bilar kommer vara den mest frekvent nyregistrerade bränsletypen under 2024. Vilket bekräftar vår hypotes.

En bild som visar linje, Graf, diagram, text

Automatiskt genererad beskrivning

För att kunna använda modellen skapad i R för samma prediktering så tränas modellen om på hela datasetet samt predikteras ny data i form av 8 poster alla med år 2024 och en för varje bränsletyp. Vilket ger nedanstående resultat, som inte alls anger EL som drivmedel för 2024, summering verkar också ganska hög jämtemot sista åren enligt Excels diagram. Det kan helt enkelt bero på att modellen inte tränas i tidsskala på år utan på fullständiga data som den är och således tar hänsyn till årsföljd vid prediktering. Konfidensintervall och prediktionsintervall ger indikation på att modellen har stort intervall för flertalet av bränsletyperna

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Konfidensintervall | | | | | Prediktionsintervall | | | |
|  | fit | lwr | upr | **diff** | fit | lwr | upr | **diff** |
| El | 1 117 | 470 | 2 656 | **2 186** | 1 117 | 39 | 31 551 | **31 512** |
| Bensin | 111 079 | 46 722 | 264 086 | **217 364** | 111 079 | 3 933 | 3 136 639 | **3 132 706** |
| Diesel | 99 157 | 41 707 | 235 742 | **194 035** | 99 157 | 3 511 | 2 799 990 | **2 796 479** |
| Elhybrid | 10 065 | 4 234 | 23 929 | **19 695** | 10 065 | 356 | 284 222 | **283 866** |
| Laddhybrid | 14 279 | 5 557 | 36 692 | **31 135** | 14 279 | 495 | 411 776 | **411 281** |
| Etanol | 3 627 | 1 525 | 8 624 | **7 099** | 3 627 | 120 | 102 431 | **102 311** |
| Gas | 3 776 | 1 588 | 8 979 | **7 391** | 3 776 | 133 | 106 652 | **106 519** |
| Övriga | 23 | 10 | 55 | **45** | 23 | 0 | 659 | **659** |

Tabell 9 Resultat från modell med extern data

# Slutsatser

Vid all datainsamling är det av yttersta vikt att det finns en plan bakom insamlandet innan detta sker. Det är således bra att ha en grundläggande kunskap och tänkt igenom hur vald hypotes ska uppfyllas innan data börjar samlas in och väljas ut.

Flödet för modeller med multipel linjär regression har i sin helhet gåtts igenom flera gånger under skrivandet av den här rapporten. Flödet börjar således att sätta sig, det är dock enkelt att glömma att göra kontroll av modell och åtgärdande av problem på alla skapade modeller innan utvärdering av modellerna sker. Detta kan ge väldigt olika resultat för både den valda modellen men även för övriga testade modeller.

I den här rapporten användes konfidensintervall och predikteringsintervall som mått på modellernas statistiska inferens. För den manuellt insamlade data blev konfidensintervall för predikteringarna ganska högt, det är inte så konstigt då modellen enbart innehöll 255 olika observationer. För data insamlad via dataskrapning blev konfidensintervallet lite mer snävt vilket gör att modellen bör ha en rätt bra träffsäkerhet i sina predikteringar.

Det hade varit smart att skapat modellen för dataskrapning med hjälp av tränings- validerings- och test-set för att på ett bättre sätt kunna utvärdera den modellen i efterhand.

# Självutvärdering

1. Utmaningar du haft under arbetet samt hur du hanterat dem.   
   Under den här kursen hamnade jag sent direkt. Det gjorde att jag inte hade möjlighet att ens försöka komma tillbaka utan fick skjuta på mycket av lärandet till fasen med kunskapskontroll. Detta gjorde att många polletter trillade ner sista dagarna och dagarna innan inlämning har varit väldigt intensiva.
2. Vilket betyg du anser att du skall ha och varför.   
   Ett G vore välkommet efter den kraftsamling som ändå utförts för att skapa någon slags förståelse för uppgiften.
3. Något du vill lyfta fram till Antonio?  
   Det kändes som en kort teoridel i den här kursen mot en lite längre kunskapsdelsövning. Jag hade uppskattat längre teoridel med kanske annan typ av kunskapskontroll för den här kursen då vi precis skrivit rapport i kursen innan. Och det upplevs som att mycket tid läggs på själva rapportskrivande och mindre på i det här fallet programmering i R.

# Källförteckning

Christian Lindell. (den 19 04 2024). *Ladda hem data från SCB med R del 1 - pxweb*. Hämtat från dittochdata.netlify.app: https://dittochdata.netlify.app/2018/06/08/ladda-hem-data-fr%C3%A5n-scb/

*Demonstration of the pxweb R package*. (den 19 04 2024). Hämtat från Youtube: https://www.youtube.com/watch?v=lWMS22XCwrE

*Multipel Regression - science*. (den 22 04 2024). Hämtat från forskning & utveckling: https://science.nu/amne/multipel-regression/

*Statistikdatabasen*. (den 19 04 2024). Hämtat från https://www.statistikdatabasen.scb.se/