

Prediktiv analys jämförelse med två regressionsmodeller — förutsäger priser på elbilar



ECUTBILDNING

Quang Tri Tran

EC Utbildning

Kunskapskontroll, R-programmering

202404

Abstract

This research project integrates data exploration, selective variable analysis, and correlation assessment, culminating in the application of statistical methods. Throughout the study, a consistent thread of statistical inference runs, providing a robust argument for our findings. We also address potential issues, including outliers, non-linearity, and non-normality, and describe our handling strategies. With the rise of electric vehicle in the market we will use this opportunity to focuses on developing a predictive model to predict car prices based on collected data from a Swedish website [Blocket.se]. Following questions will be answered throughout the project about regression models, significant factors, correlations, and regression analysis. By answering these questions, we aim to create a comprehensive regression modelling approach that can be used for car price predictions. The results will benefit both buyers and sellers in the automotive market.

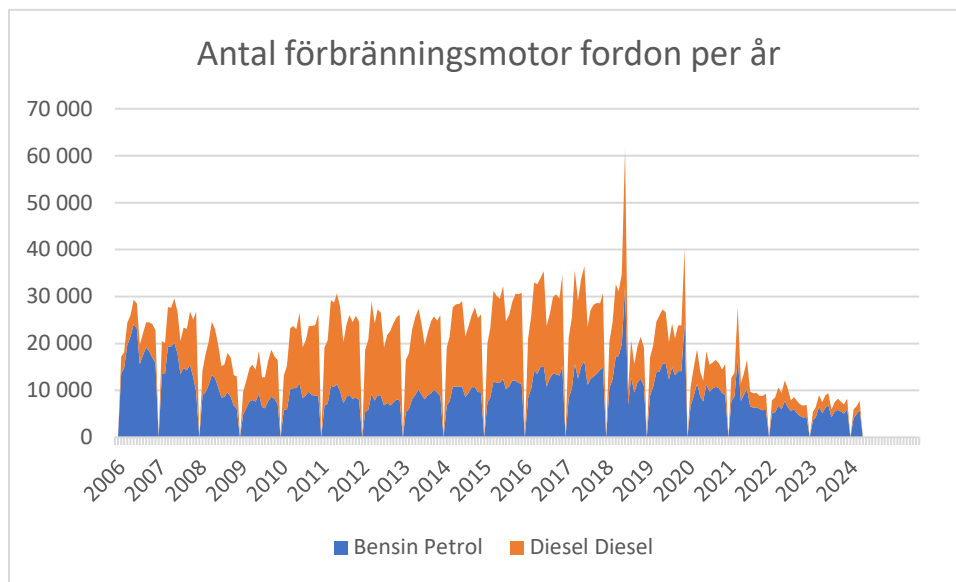
Keywords: data exploration, variable analysis, correlation, statistical inference, outliers, non-linearity, non-normality.

Innehållsförteckning

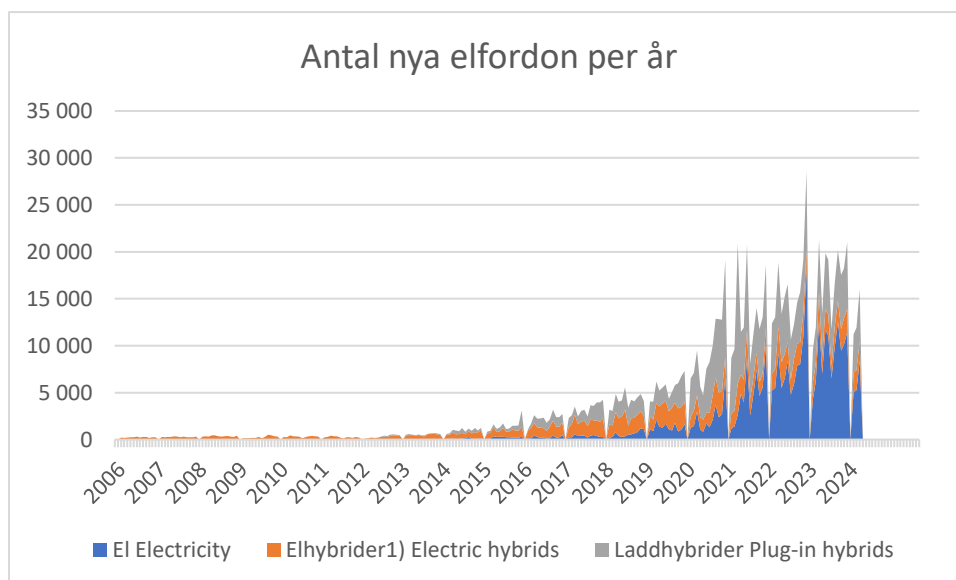
Abstract	2
1 Inledning.....	1
1.1 Syfte och frågeställning.....	2
2 Teori.....	3
2.1 R-bibliotek.....	3
2.2 EDA (Exploratory Data Analysis).....	3
2.3 Residuals analysis.....	4
2.4 Regressionsmodeller.....	4
2.4.1 Linjär regression	4
2.4.2 Multi linjär regression.....	4
2.5 Utvärderings metod	5
2.5.1 RMSE.....	5
2.5.2 MAPE	5
2.5.3 R2	5
3 Metod	6
3.1 Del 1: Insamling av data.....	6
3.2 Del 2: importera & utforska data	6
3.3 Del 3: Linjär regression modellering	6
3.3.1 Träning & test data	7
3.4 Del 4: Multipla linjär regression modellering.....	7
4 Resultat och Diskussion	8
4.1 Diskussion	10
5 Slutsatser	12
6 Självtvärdering.....	13
Appendix A	14
Källförteckning.....	16

1 Inledning

Stigande olje- och gaspriser har gjort så att framsteg inom batteriteknik riktas nu större uppmärksamhet mot batteridrivna elektriska fordonskoncept. Förbränningsmotorn blev den mer kostnadseffektiva lösningen på grund av den förbättrade tillgången till olja som sedan stimulerade massproduktion av gasfordon. Denna utveckling ledde till att det batteridrivna fordonet försvann mer eller mindre som ett alternativt drivmedel under år 1930 (Chan, 2007). Efter förbränningsmotorns dominans under det senaste århundradet upplever nu olika typer av elfordon ett ökat allmänintresse i nuvarande tid.



figur 1. Visa antal nyregistrerade bensin och diesel bilar per år. (SCB databas)



figur 2. Visa antal nyregistrerade el och hybrid bilar per år. (SCB databas)

Med stigande popularitet innebär det att marknaden är i högtid med elbilar runt om i världen och även i Sverige (se figur 2). Det riktar sig då åt att tänka sig vad som bestämmer prissättningen på elfordon eller är det mest kundbeteende. Användningen av maskininlärning underlättar arbeten för denna undersökning då maskininlärning är en spännande disciplin som har möjliggjort stora framsteg inom AI och teknik. Genom att använda data och algoritmer kan vi skapa intelligenta system som kan

lära sig och anpassa sig över tid, inklusive att förutsäga bilpriser med precision. Maskininlärning har därmed revolutionerat många områden och bilindustrin är inget undantag.

Till detta arbete kommer det att utforskas insamlingen av data från elbilar, använda maskininlärning för att analysera historiska data och förutsäga framtida bilpriser. Låt oss dyka in i elbilarnas spännande värld!

1.1 Syfte och frågeställning

Syftet med denna rapport är att skapa en komplett regression modellering utifrån insamlade data från blocket.se som är kapabel att utföra bilpris prediktioner. För att uppfylla syftet så kommer följande frågeställning(ar) att besvaras:

1. Utveckla två regressionsmodeller för att förutsäga prissättningen utifrån en eller flera potentiella faktorer.
2. Kan en enfaktoriell modell prestera bättre än en flerfaktoriell modell?
3. Vilka bilfaktorer är signifikanta för priset i detta fall?
4. För att förstå samband hur variabler relaterar till varandra måste det veta om korrelation. Vilken par bilfaktorer har tydligt samband med varandra?

2 Teori

2.1 R-bibliotek

Bibliotek som används i arbetet;

1. Dplyr
Dplyr är ett verktyg för datahantering i R. Den bygger på "grammar of data manipulation" och används för att filtrera, transformera och sammanfoga data.
2. Ggplot2
Detta paket bygger på "grammar of Graphics" och är en metod att skapa anpassningsbara grafer.
3. Readxl
Enkelt verktyg för att importera in och läsa Excel filer.
4. gridExtra
gridextra paketet ger möjlighet att anpassa layouten för flera grafer på en sida.
5. Pxweb
Är ett paket för att hämta data från statliga myndigheter, kommuner, utomsvenska statistikbyråer och internationella organisationer som presenterar statistik. (API)
6. Zoo
Zoo används för att hantera och analysera "tids" data eller tidsseriedata.
7. Olsrr
Hjälper till att skapa, analysera och visualisera linjära regressionsmodeller samt bedömning av modelanpassning och variabelurval.
8. Moments
Ett paket som beräknar skevhet, kurtosis och andra moment för en given datamängd. Det ger en enkel metod för att få statistiska moment och relaterade egenskaper från data.

För mer information om biblioteken som används och syntaxen, de finns i dokumentations hemsidan

[<https://www.rdocumentation.org/>].

2.2 EDA (Exploratory Data Analysis)

Utförande av EDA skapa djupare förståelse om data och identifiera mönster och trender. Det är en viktig fas inom statistik och datanalys. Det innebär att noggrant undersöka och karakterisera data innan vidare steg till mer avancerade analyser (Velleman et.al, 2023) .

- Dataset översikt och deskriptiv statistik:
 - Börjar med att förstå data relativt till antalet observationer, antal och typer av variabler, förekomsten av duplicerade rader och andelen saknade värden.
 - Deskriptiv statistik ger oss en översikt över data, inklusive medelvärden, medianer, standardavvikelser och andra mått.
- Visualisering och utforskning av variabler:
 - Genom att använda diagram och grafer kan data visualiseras och identifiera mönster.
 - Spridningsdiagram, histogram och boxplot är vanliga verktyg för att utforska samband och fördelningar.

- Bedömning av datakvalitet:
 - Undersök eventuella outliers, skevhet, avvikelser och möjliga fel i data.
 - Detta hjälper till att förstå om data är tillförlitligt och om det finns några problem som behöver åtgärdas.

Data från blocket ([blocket.se]) innehåller 409 datapunkter, 10 faktorer som kolumn och är från 2014-2024. En grov insamling av data innehåller vanligtvis fel i formattyp eller saknade värden i datan.

Data från Svenska statistiska centralbyrån (SCB) innehåller 344925 datapunkter, 4 kolumner och är från 2006-2024. Data källa SCB API program.

2.3 Residuals analysis

Residuals är skillnaden mellan de observerade värdena och de förutsagda värdena från en statistisk modell (ex. linjär regression och multi linjär regression). I fallet med enkel linjär regression är residualen för varje observation:

$$\text{Residuals} = \text{observerat värde} - \text{predikerad värde}$$

Residualsanalys är ett kraftfullt verktyg för att utvärdera modellens prestanda och förstå hur väl den passar data. Genom att observera residuals beteende kan man undersöka skevheten är godtycklig eller om normalitet existera i distributionen. Att inkludera residualsanalys i rapporten läggs det hänsyn till modellens noggrannhet och robusthet (James, 2021).

2.4 Regressionsmodeller

2.4.1 Linjär regression

Linjär regression är en statistisk teknik som används för att undersöka sambandet mellan EN oberoende variabel och en beroende variabel. Den oberoende variabeln representerar prediktorn och den beroende variabeln är kriteriet. Formeln visas nedan;

$$Y = \alpha + \beta x + \varepsilon .$$

Y är den beroende variabelns som påverkas av X som är den oberoende variabeln. Alfa parametern är intercepten med y-axeln och beta är koefficienten/lutningen (James, 2021).

2.4.2 Multi linjär regression

En uppföljare av linjär regression men mer komplex i strukturen. Här undersöks sambandet för FLERA oberoende variabler och en beroende variabel. Följer samma principer som linjära regressionen (James, 2021).

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \varepsilon .$$

β_0 : Baslinjenivån för Y när alla oberoende variabler är noll. intercepten

$\beta_1, \beta_2, \dots, \beta_p$: Förändringen i Y för varje enhetsändring i x_p , medan andra variabler hålls konstanta.
Regressionskoefficienter

x_1, x_2, \dots, x_p : dessa är oberoende variabler/prediktorer som förklarar variationen i Y.

ϵ : Slumpmässig variation som inte kan förklaras av de oberoende variablerna. Representerar residuals.

2.5 Utvärderings metod

För att bedöma prestandan hos en modell används det följande utvärderingsmått: RMSE, MAPE och R^2 . Utvärderingsmått kommer förklaras kort nedan.

2.5.1 RMSE

- Mäter den genomsnittliga avvikelsen mellan de äkta värden och förutsagda värden.
- Lägre RMSE indikerar bättre modellprestanda.

- Formel: $RMSE = \sqrt{\frac{\sum_{i=1}^N |y(i) - \hat{y}|^2}{N}}$

2.5.2 MAPE

- Mäter den genomsnittliga procentuella avvikelsen mellan de äkta värden och förutsagda värden
- Lägre MAPE indikerar mindre procentuell avvikelse från de äkta värden.

- Formel: $MAPE = \frac{\sum \frac{|A-F|}{A} * 100}{N}$

2.5.3 R^2

- Mäter andelen variation i den beroende variabeln som förklaras av modellen.
- Hög R^2 indikerar god anpassning av modellen till data.

- Formel: $R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$

3 Metod

Hur har du genomfört ditt arbete? Exempelvis, hur har datan erhållits?

3.1 Del 1: Insamling av data

första delen av arbetet involverade att samla egna data från blocket ([blocket.se]) i grupp. Innehållet samlades in manuellt och sparat i Excel. Följande villkorar bestämdes ur grupp och är viktiga att ha med i datainsamlingen:

Specifika insamlingsområden

1. Bilmärken – hyundai, audi, nissan, volkswagen, tesla, kia, mg, mb, ford, bmw.
2. Bränsle – El/eletrisk.
3. Biltyp – suv, sedan, kombi, halvkombi

Allmänna insamlingsområden

4. Miltal
5. Modellår
6. Drivning
7. Hästkrafter
8. Modell
9. Pris

3.2 Del 2: importera & utforska data

En datautforskning utfördes för att förstå data innehållet,

1. Importera blocket-databasen sparad i Excel in i Rstudio.
2. Utför str() och summary() funktionen för att få fram struktur och sammanfattning av dataramen.
3. Identifiera och hantera fel i datan.
4. Visualisera om korrelation finns bland variablerna. (se figur 3)
5. Variablerna med godtycklig korrelation väljs ut och analyseras vidare.

Den utvalda variabeln "Hästkrafter" väljs ut och har visat korrelation med variabeln "Pris".

3.3 Del 3: Linjär regression modellering

Utvald variabeln "Hästkrafter" blir den oberoende variabeln och "Pris" den beroende variabeln.

1. Variabelanalys utfördes på utvald variabel för att ytterligare analysera dess egenskaper och fördelning. Samma process gjordes med den beroende variabeln.
2. Genom Shapiro-testet observerades att det finns en icke-normal fördelning, därmed en sökning efter skevhet utfördes. Sökning gav måttlig positiv skevhet, åtgärden är kvadratrotstransformation. Samma transformation gjordes till den beroende variabeln.
3. Anpassa oberoende variabel och beroende variabel i en linjär regressionsmodell. (se figur 8 och figur 9 för före och efter transformation av regressionsmodellen)
4. Utvärdera modellens prestanda (se tabellerna i Resultat och Diskussion).

3.3.1 Träning & test data

Eftersom modellen är nu justerat och tränad med hela databasen, gäller det nu i nästa steg att dela datan i två delar, en träning dataset och en test dataset. Tränings dataseten i arbetet innehåller 80 procent av original data, resterande 20 procenten finns i test dataseten. Original datan innehåller 409 datapunkter, uppdelningen ger: 327 datapunkter i träning, 82 datapunkter i test.

3.4 Del 4: Multipla linjär regression modellering

Multipla linjär regression använder sig av flera oberoende variabler i stället för en.

Många försök och justeringar gjordes även att använda flera oberoende variabler, men normalitet kunde inte uppnås. Kvadratrotstransformation applicerades på hela datan för att närma sig normalfördelning.

1. Databasen efter kvadratrotstransformation delas upp i träning och test dataset.
2. Modellen tränas och testas.
3. Utvärdera modellen (se tabellerna i Resultat och Diskussion).

4 Resultat och Diskussion

Grafer och visualiseringar på resultat finns i Appendix A. Nedan visas endast resultat från modellernas bästa prestation efter många iterationer och justeringar.

RMSE resultat	
Enkel Linjär Regression	93.47
Multipla linjär Regression	61.20

Tabell 1: Root Mean Squared Error (RMSE) för de två valda modellerna.

MAPE resultat	
Enkel Linjär Regression	0.1293
Multipla linjär Regression	0.08

Tabell 2: Mean Absolute Percentage Error (MAPE) för de två valda modellerna.

R2 resultat	
Enkel Linjär Regression	0.5794
Multipla linjär Regression	0.7956

Tabell 3: R-Squared (R2) för de två valda modellerna.

Korrelation mellan Pris och oberoende variabler	
Miltal	-0.2267105
Hästkrafter	0.6996439
Modellår	0.5876044
Drivning	0.5453824

Tabell 4: korrelation. Mäter styrkan och riktningen av samband mellan två variabler.

```
Residuals:
    Min       1Q   Median       3Q      Max
-161.093  -33.707    4.483   36.014  179.321

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.364e+05  8.343e+03 -16.347  <2e-16 ***
Hästkrafter  2.685e+01  1.600e+00  16.776  <2e-16 ***
Miltal       -8.533e-02  9.720e-02  -0.878    0.381
Modellår     3.039e+03  1.856e+02  16.376  <2e-16 ***
Drivning     -1.629e+01  1.960e+01  -0.831    0.407
---
signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 61.35 on 322 degrees of freedom
Multiple R-squared:  0.7956,    Adjusted R-squared:  0.7931
F-statistic: 313.4 on 4 and 322 DF,  p-value: < 2.2e-16
```

figur 5. Sammanfattning av prestationsförmåga från multipla linjär regression i träningsdataset.

Residuals:

	Min	1Q	Median	3Q	Max
	-209.728	-62.530	1.307	62.211	214.728

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	154.574	22.576	6.847	3.78e-11 ***
Hästkrafter	29.446	1.392	21.160	< 2e-16 ***

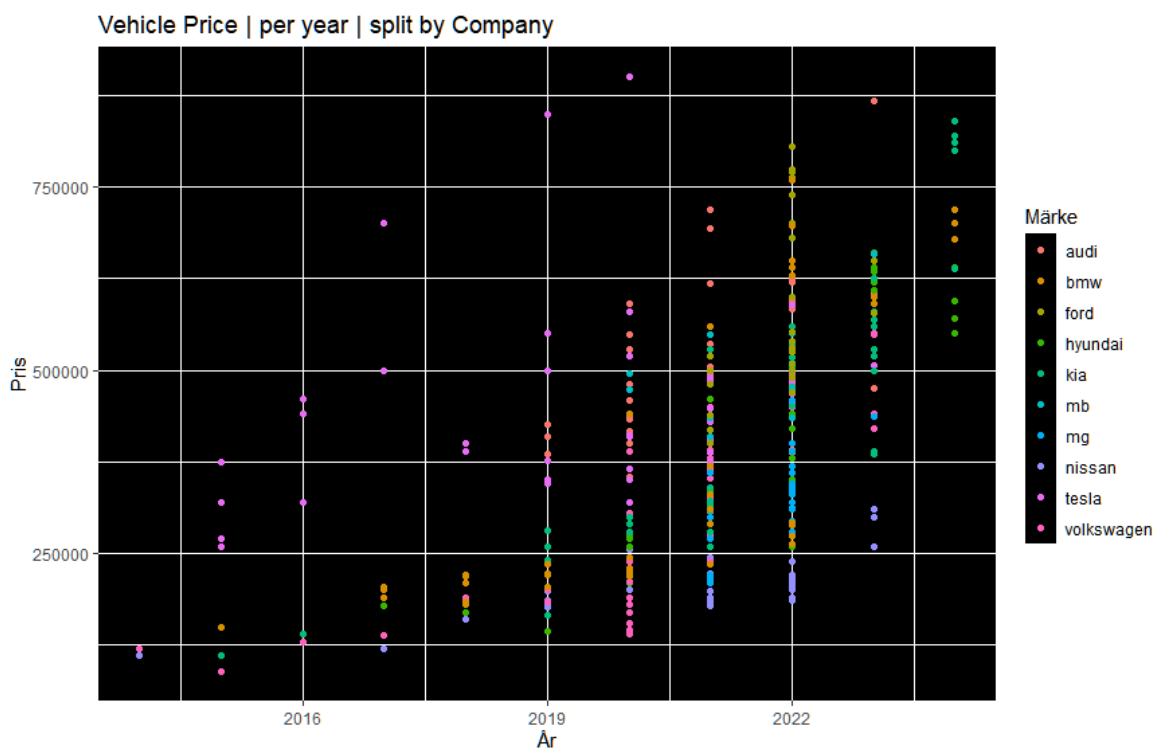
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 87.6 on 325 degrees of freedom

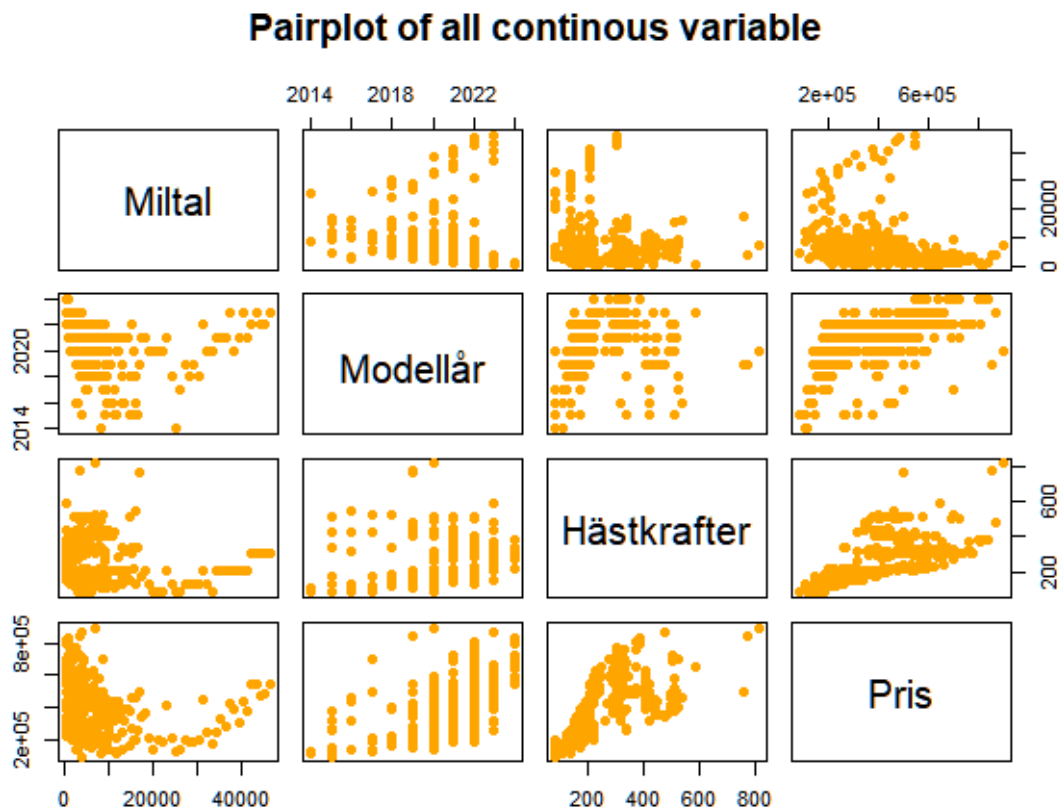
Multiple R-squared: 0.5794, Adjusted R-squared: 0.5781

F-statistic: 447.8 on 1 and 325 DF, p-value: < 2.2e-16

figur 6. Sammanfattning av prestationsförmåga från enkel linjär regression i träningsdataset.



figur 4. Grafen visar datapunkter från olika bilföretag och hur elbil priset ökar med tiden.



figur 3. Grafen visar alla numeriska kontinuerliga variabler i databasen från blocket och korrelation kan identifieras i bilden.

4.1 Diskussion

I den inledande fasen av rapporten, där insamling av data från blocket och utforskning av datans innehåll. Beaktas det att datainsamlingen hanterades i grupp och utifrån blockets hemsida med specifika krav på fordonen. Med bild från Svenska statistiska centralbyrån kan det synas hur elbilar stiger i trend (se figur 2) och prisökningen går att se i figur 4.

Härefter pågick arbetet individuellt. Blocket datan utforskades genom att analysera strukturen i dess ram, vilket innebär att leta sig in bland numeriska och kategoriska värden samt att kontrollera varje rad och kolumn för att vara korrekta med både värden och format. Det uppkom saknade värden, versala bokstäver och parenteser som behövde städas upp.

Totalt städades 1203 data observationer från datan, majoriteten av dem 1203 data observationerna togs bort och resterande fick korrigeras om. Prestationsförmågan hos regressionsmodeller hade varit mycket dåligt om rensning av data inte gjorts i förväg då kodning språket R är ett väldigt känsligt språk och kan hindra regressionsmodellerna att fungera korrekt om datan inte är justerat på rätt sätt.

En annan faktor som skulle påverka prestandan hos modellerna är felaktigheter i data som extremvärden (outlier), icke-normalitet, icke-linjär. Extremvärden kan tas bort om de anses onödiga eller fel i data. Icke-normalitet kan identifieras genom att plotta histogram, qq-plot eller använda tester som shapiro-test. Även om icke-normalitet hittas betyder det inte att det är felaktigt, eftersom det beror på vilken typ av databas detta är. Om den bedöms korrekt bör icke-normaliteten finnas kvar i data som potentiell faktor, för när mer data samlas in kan det vara vettigt varför det beter sig så eller potentiellt kan det vara starten på någon ny rörelse.

Utforskningen av datan ger oss en grundläggande förståelse för variablerna och deras relationer. Korrelationsanalysen hjälper oss att välja relevanta variabler för vidare analys (se figur 3 och tabell 4) och sedan anpassa in i regressionsmodellerna. Eventuella fel och brister i datan går ibland inte att undvika och därmed kommer snedvridning i modellerna.

Två regressionsmodeller initierades och testades i denna rapport. De tränade först med att använda alla data så att justeringar kan göras innan modellerna inmatas i träning dataset och test dataset processen. Enkel linjär regression kunde justeras och anpassas till en normalfördelning medan flera försök till multipla linjära regressionen inte kunde anpassas till en normalfördelning. Detta betyder inte att multipel regression inte är giltig, men man bör vara bara medveten och försiktig om tolkningen på antaganden. Eftersom forskning anser att normalfördelning är ett av grundantagandena för att arbeta med linjär regression.

i sådana situationer är det viktigt att överväga syftet med analysen. Om prediktiv kraft är viktigare än normalitetskrav, när man utför slutledning (t.ex. hypotestestning) bör man vara mer försiktig med normalitetsantagandet. Analysen på modellerna är baserat på deras prestanda på tre viktiga mått:

- Root Mean Squared Error (RMSE)
- Mean Absolute Percentage Error (MAPE)
- R-Squared (R^2)

Resultatet från tabellerna (tabell 1, tabell 2 och tabell 3) kan diskuteras som:

1. En lägre RMSE-värde indikera bättre passform för data. Multipla linjär regression har lägre RMSE än enkel linjär regression. Det antas att den flerfaktoriella modellen ger mer exakta förutsägelser.
2. Lägre MAPE innebär att den genomsnittliga procentuella avvikelsen från de faktiska värdena är mindre. Detta stärker tidigare antaganden om bättre prestanda för den flerfaktoriella modellen.
3. Högre R^2 -värden betyder att större andel av variansen i den beroende variabelns kan förklaras av den oberoende variabeln. Multipla linjär regression > enkel linjär regression. Stöder ytterligare om bättre passform för den flerfaktoriella modellen.

5 Slutsatser

Följande slutsatser kan dras angående uppsatta mål, uppgifter och forskningsfrågor i detta arbete.

1. Två regressionsmodeller skapades. En enkel linjär regression som enfaktoriell modell som endast tog en oberoende variabel för sina prediktioner, och en multipla linjär regression skapades som flerfaktoriell modell för flera oberoende variabler.
2. Kan en enfaktoriell modell prestera bättre än en flerfaktoriell modell?

I detta fall utifrån tabell 1, tabell 2 och tabell 3 så presterar multipla linjär regression i över lag bättre än enkel linjär regression om normalitetskrav inte tas till hänsyn. Om det övervägs att normalitetskrav är viktigare än prediktionskraft så blir det till att välja att enfaktoriellmodell passar bättre. I verkliga världen finns det sällan bara en faktor, det handskas om flera faktorer dagligen i arbetslivet därmed kräver det mer arbete för att förbättra flerfaktoriella modeller än enfaktoriella modeller. Detta är då synpunkten trots icke-normaliteten multi linjär regression har men många resultat i arbetet stödjer vår slutsats.

3. Vilka bilfaktorer är signifikanta för priset i detta fall?

Det har visat sig att Hästkrafter och Modellår statistisk signifikans för priset medan Miltal och Drivning är ej signifikanta för priset. Det menas att förutsäga pris med faktorer som Hästkrafter och Modellår inte beror på slumpen, därmed går det att förkasta noll hypotesen och acceptera alternativ hypotesen.

4. För att förstår samband hur variabler relaterar till varandra måste det veta om korrelation. Vilken par bilfaktorer har tydligt samband med varandra?

Det finns tydliga samband mellan olika bilfaktorer och priset. Den som visar tydligast korrelation med pris är Hästkrafter på 0.6996439. Det indikera ett måttligt positivt samband då maximalt samband har värdet 1.

Den svagaste korrelationen är mellan pris och miltal på -0.2267105, ett svagt negativt samband vilket indikerar att om miltal ökar så minskar priset.

Multipla linjära regressioner kräver mer arbete, och rekommenderas undersöka ytterligare för att hantera normalitetsproblem. Vi bör fortsätta utforska fler variabler och överväga andra modeller (t.ex., polynomial regression eller andra icke-linjära metoder) för att förbättra prediktionerna.

6 Självutvärdering

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

Många olika möjligheter finns att göra detta arbete samt som många utmaningar.

Utmaningarna hanterades med att få insikt genom mycket googla, ser hur andra angriper liknande problem i andra hemsidor och videos, och sedan bygger min uppfattning kring frågan och börja angripa utmaningen igen. Och repetera processen om jag fastnar igen.

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

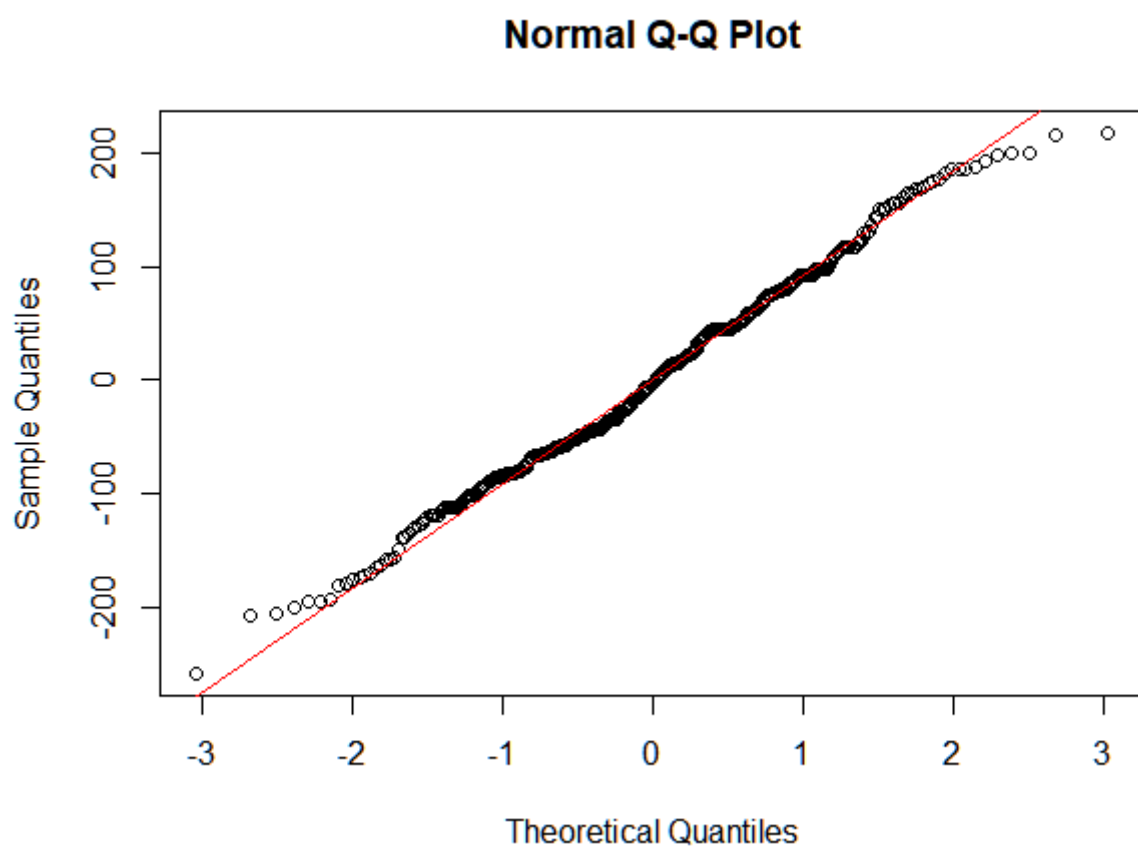
Till denna uppgift siktar jag på ett VG. Från början till slut har jag följt och hanterat hur en komplett regression modellering beskriver sig att vara i princip. I mitt arbete har jag kopplat ihop datautforsknings analys, vidare till selektiv variabelanalys samt relatera det med korrelationsanalys innan anpassning till statistiska metoder. En röd tråd med statistisk inferens som argument löper genom arbetet. Jag har även beskrivit potentiella problem såsom outliers, icke-linjaritet och icke-normalitet samt hur det hanterades.

3. Något du vill lyfta fram till Antonio?

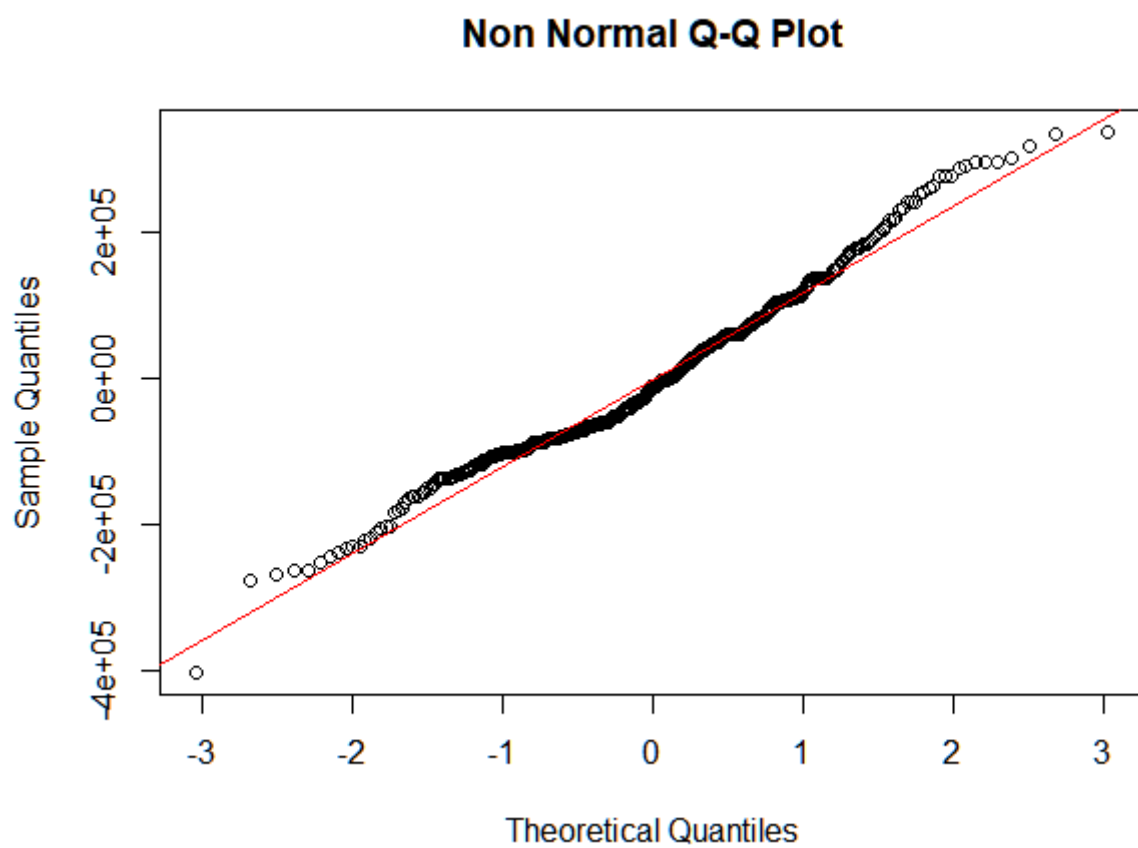
Keep up the good work, Antonio!

(Annars om det går bra att fråga här, vad händer med en individ om praktik inte hittas? Söker man om utbildningen då eller finns de nå extra chans till nästa termin? T_T)

Appendix A



figur 8. Mer anpassad till normalitet efter kvadrattransformationen för enkel linjär regressionsmodell.



figur 9. Grafen visar icke-normalitet då punkterna ej följer röda linjen, för enkel linjär regression.

Källförteckning

Chan, C.C. (2007) 'The State of the Art of Electric, Hybrid, and Fuel Cell Vehicles', Proceedings of the IEEE, Proc. IEEE, 95(4), pp. 704–718. doi:10.1109/JPROC.2007.892489.

Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow : concepts, tools, and techniques to build intelligent systems (Second edition). O'Reilly.

James, G. (2021) An Introduction to Statistical Learning. with Applications in R. 2nd ed. 2021. Edited by D. Witten, T. Hastie, and R. Tibshirani. Springer US (Springer Texts in Statistics).

Rajashekara, K. (1993) 'History of electric vehicles in General Motors', Conference Record of the 1993 IEEE Industry Applications Conference Twenty-Eighth IAS Annual Meeting, Industry Applications Society Annual Meeting, 1993., Conference Record of the 1993 IEEE, p. 447. doi:10.1109/IAS.1993.298962.

Velleman, P.F. and Hoaglin, D.C. (2023) 'Exploratory data analysis', in APA handbook of research methods in psychology: Data analysis and research publication., Vol. 3, 2nd ed. Washington, DC: American Psychological Association (APA Handbooks in Psychology®), pp. 53–73. doi:10.1037/0000320-003.

<https://www.rdocumentation.org/>