Prissättningsmodell för Volvobilar

Linjär regressionsanalys

Adrian Krsmanovic

EC Utbildning

R kunskapskontroll

2024-04

Abstract

The objective of this project was to create a model using multiple linear regression that could predict the list price for a Volvo car, it also aimed to perform inference to interpret what variables affected the price and quantify the effect. The work consisted of steps such as data collection, data cleaning, model testing and evaluation. Further improvements were made to the model by using interaction terms and transformations. Overall, the project was a success and a model with <0.95 Adjusted R squared was achieved.

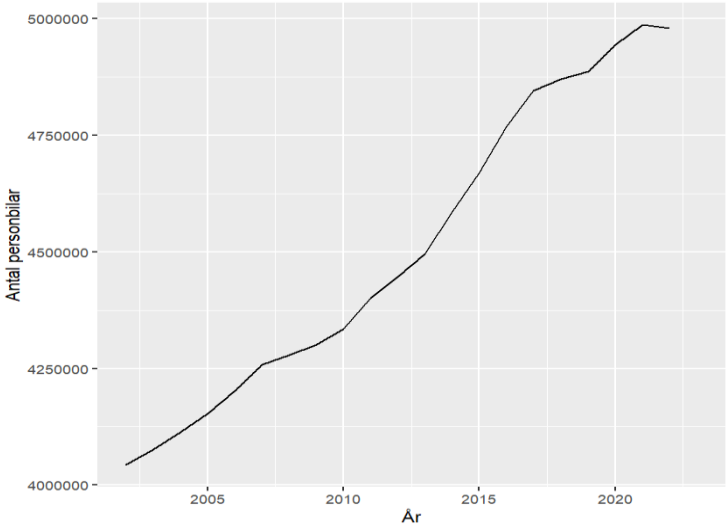
Innehållsförteckning

1 Inledning

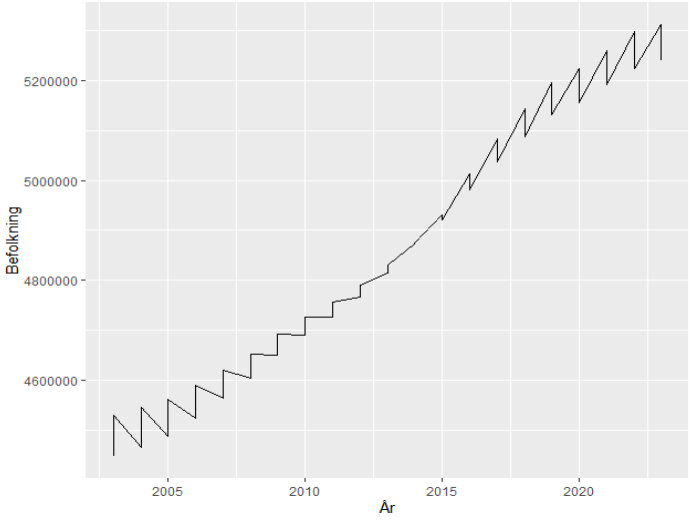
Linjär regression är en metod som tillämpas inom många olika fält, medicin, finans, marknadsföring m.m. Metoden är simpel i jämförelse med andra modeller som används för att prediktera värden och det är en av anledningarna till att den är populär. Även fast man ibland kan uppnå bättre resultat med mer komplexa modeller och metoder så tillåter linjär regression en djupare förståelse och bättre tolkning av parametrar och deras påverkan på den oberoende variabeln som ska predikteras. En marknad som i Sverige har växt markant under 2000-talet är bilmarknaden. Detta kan bero på flera anledningar som till exempel att befolkning har ökat vilket skapar ett större behov för fler bilar. Utbudet och tillgängligheten av elektriska bilar har ökat under senare år vilket även kan antas vara en bidragande faktor till växten av bilmarknaden. Tillväxten av bilmarknaden skapar ett ökat behov för mekaniker, bilförsäljare och andra yrken. Det skapar även ett behov av kompetens inom data och analys. Maskininlärning är ett fält som kan nyttjas inom bilmarknaden, maskininlärning kan beskrivas som förmågan för en dator att lära sig av data utan explicit programmering, ” Machine Learning is the science (and art) of programming computers so they can learn from data.” (Géron, 2019, p. 2). Detta projekts syfte är att visa processen från att samla in data om Volvobilar till att skapa en modell som predikterar värdet på dessa. Värdet på en bil kan inte bestämmas utifrån en specifik sak utan snarare baserat på en kombination av variabler, därmed kommer det i detta projekt tillämpas multipel linjär regressions vilket möjliggör prediktion av priset baserat på flera variabler. Statistisk inferens kommer även att utföras där syftet är att tolka vilka variabler som påverkar priset och hur stor påverkan dessa har.
För att uppnå syftet med projektet kommer följande frågor besvaras:

1. Skapa en modell med Adjusted R^2 över 0.9.
2. Är el bilar dyrare än hybrid, diesel och bensinbilar?
3. Har biltyp en påverkan på pris?
4. Är antaganden uppfyllda vilket möjliggör statistik inferens?

Figur 2 data hämtad via API från SCB över antal personbilar



Figur 1.2 data hämtad via API från SCB över befolkningsmängd



2 Teori

Nedan kommer nödvändig teori som har applicerats i projektet att beskrivas.

2.1 Linjär Regression

Linjär regression är en modell som används för att prediktera ett kvantativt värde Y baserat på en variabel X. ”it is a very straightforward approach for predicting a quantitative response Y on the basis of a single predictor variable X”, (Gareth et. al., 2023, p.61) Linjär regression gör vissa antaganden om datan,

1. Det finns ett linjärt samband mellan X och Y.
2. Residualerna är obereonde
3. Variansen är konstant för alla X(homoskedasticitet)
4. Residularena är normalfördelade

Det är viktigt att notera att antagandena aldrig förväntas vara totalt uppfyllda men det är något som eftersträvas. Formeln för linjär regression ser ut som nedan.

$$\gamma = \beta 0 + \beta 1X1 + \varepsilon$$

2.2 Multipel linjär regression

Multipel linjär regression bygger vidare på linjär regression men möjliggör prediktion för ett kvantativt värde baserat på två eller flera oberoende variabler. Multipel linjär regression gör likadana antaganden som simpel linjär regression. Formeln för multipel linjär regression ser likadan ut som den för simpel linjär regression med tillägget av fler beta koefficienter. (se formel nedan)

$$y = \beta 0 + \beta 1X1 + \beta 2X2 + .. + BpXp + \varepsilon$$

2.3 Interaktionseffekt

Interaktionseffekt appliceras när två variablers påverkan på Y antas ha ett samband. ” An interaction term is effectively a multiplication of the two features that you believe have a joint effect on the target”(Lewinson, 2023)

$$y = \beta 0 + \beta 1X1 + \beta 2X2 + \beta 3X1X2 + \varepsilon$$

2.4 Dummy encoding

Dummy encoding omvandlar kategoriska värden till numeriska värden vilket möjliggör användning av variablerna i en modell. Dummy encoding skapar dummy variables för varje unikt kategoriskt värde förutom ett som blir en baseline.

2.5 Mått och K-fold cross validation

2.5.1 Mått

RSS (residual sum of squares) mäter skillnaden mellan den observerade datan och regressionslinjen. Med andra ord kan det beskrivas som variation av Y som modellen inte kan förklara. ”It is the portion of variability your regression model does not explain, also known as the model’s error” (Frost, 2024)

$$RSS = \sum_{i=1}^{n} (yi - \hat{y}i)\text{^}2$$

R^2 (R squared) mäter hur stor andel av variationen i Y som kan förklaras av en variabel, ”R-Squared (R^2 or the coefficient of determination) is a statistical measure in a regression model that determines the proportion of variance in the dependent variable that can be explained by the independent variable”(Taylor, u.å). Genom att subtrahera ”Total sum of squares(TSS)” med ”Residual sum of squares(RSS)” återstår den förklarade variansen.

$$R^2 = \frac{TSS - RSS}{TSS}$$

Adjusted R squared (adjr^2) svarar på hur stor del av variansen som kan förklaras av modellen. Intuitionen bakom Adjusted R squared är att genom att justera för antalet variabler kan värdet minska om variabeln inte minskar RSS tillräckligt. ”The adjusted R-squared adjusts for the number of terms in the model. Importantly, its value increases only when the new term improves the model fit more than expected by chance alone”(Frost, 2017). Detta gör måttet till ett mer användbart mått vid multipel linjär regressions jämfört med R^2 som alltid förblir samma eller ökar vid tillägget av ytterligare variabler.

$$Adjusted\ R^2 = 1 - \frac{RSS/(n-d-1)}{TSS(n-1)}$$

AIC utvärderar modellens passning på träningsdata och lägger till ett straff för komplexiteten av modellen,” AIC works by evaluating the model’s fit on the training data and adding a penalty term for the complexity of the model” (Zajic, 2022). Tanken bakom AIC är att estimera ett test fel utifrån träningsdatan och reducera overfitting genom att straffa mer komplexa modeller.

$$AIC = \frac{1}{n}(RSS + 2d\hat{\sigma}^2)$$

BIC liknar AIC men i stället för 2 så består formeln av log(n), detta innebär att om n är större än 7 blir straffet högre för BIC än AIC. Resultatet av detta är att BIC tenderar att välja modeller med färre variabler ”The BIC statistic generally places a heavier penalty on models with many Variables. ” (Gareth et. al., 2023, p.234)

$$BIC = \frac{1}{n}(RSS + \log(n)d\hat{\sigma}^2)$$

RMSE (Root mean squared error) mäter medelavståndet mellan observationerna och prediktionerna och är ett vanligt förekommande mått. ” The root mean square error (RMSE) measures the average difference between a statistical model’s predicted values and the actual values” (Frost, 2023)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

2.5.2 K-fold cross validation

K-Fold cross validation är en metod som slumpmässigt delar upp datan i k antal grupper. Modellen tränas på k-1 grupper och valideras på den utelämnade. Detta görs k olika gånger och

varje gång är det en annan grupp som används för validering, MSE beräknas på valideringsgrupperna och ett medelvärde beräknas.
“The first fold is treated as a validation set, and the methodis fit on the remaining k − 1 folds. The mean squared error, MSE1, is then computed on the observations in the held-out fold”(Gareth et. al., 2023, p.203)

2.6 Best subset selection

Best subset selection är en metod som kan användas för att hitta den bästa kombinationen av variabler. Metoden provar alla möjliga kombinationer av variabler och jämför resultaten.” To perform best subset selection, we fit a separate least squares regression best subset for each possible combination of the p predictors” (Gareth et. al., 2023, p.227)

2.7 Multikollinearitet och Variance inflation Factor(VIF)

Multikollinearitet uppstår när 2 eller fler oberoende variabler i en modell är korrelerade. Detta leder till problem eftersom de då inte är oberoende längre och inferensen blir opålitlig, ” This correlation is a problem because independent variables should be independent.” (Frost, 2017). VIF är ett mått för att mäta multikollineratit mellan variabler.

$$VIF_i = \frac{1}{1 - R_i^2}$$

2.8 Potentiella problem och tolkning

Linjär regression gör som tidigare nämnt vissa antaganden om datan, om antaganden inte är uppfyllda kan det resultera i att inferensen blir osäker och opålitlig. Undersökning av antaganden kan utföras genom statistiska grafer som är tillgängliga efter träningen av en modell. Nedan kommer dessa grafer visas och hur de kan tolkas.

2.8.1 Residuals vs Fitted

Denna plot visar om residualerna har ett icke linjärt samband, om det finns ett tydligt mönster för residualerna tyder det på ett icke linjärt samband. ”If you find equally spread residuals around a horizontal line without distinct patterns, that is a good indication you don’t have non-linear relationships. “(Bomae, 2015)

2.8.2 Normal Q-Q

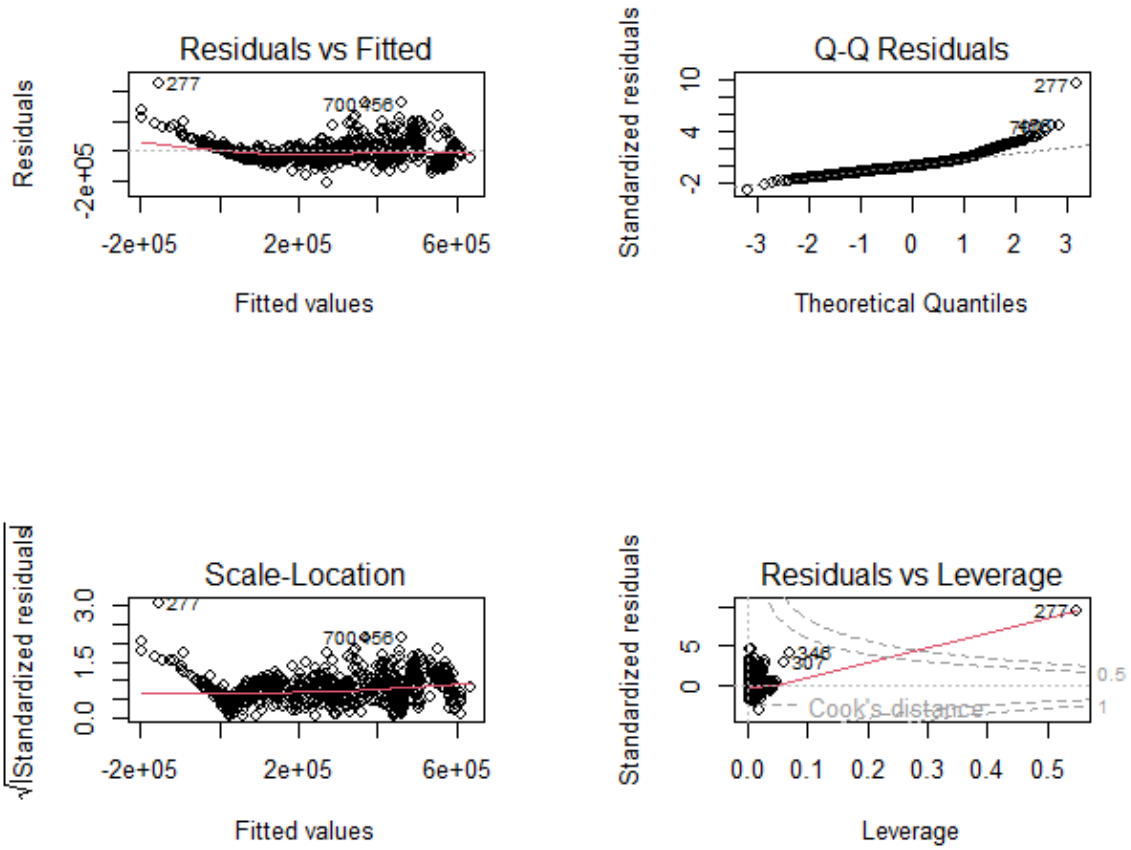
Denna plot visar om residualerna är normalfördelade, idealt ska punkterna bilda en rak linje. ”It’s good if residuals are lined well on the straight dashed line.” (Bomae, 2015)

2.8.3 Scale-Location

Denna plot visar om residualerna är jämnt spridda med prediktionerna, om residualerna sprider sig längre ifrån linjen åt höger tyder det på heteroskedasticitet. ”This is how you can check the assumption of equal variance (homoscedasticity). It’s good if you see a horizontal line with equally (randomly) spread points.” (Bomae, 2015)

2.8.4 Residuals vs Leverage

Denna plot används för att undersöka outliers och ”high leverage” punkter. Genom att kolla efter punkter som är utanför den punktmarkerade linjen som visar punkter med högt ”cooks distance score”. ” When cases are outside of the dashed lines (meaning they have high "Cook’s distance" scores), the cases are influential to the regression results.” (Bomae, 2015)



Figur 3 Exempel på plots för modeller

3 Metod

För att uppnå målen med denna rapport har stegen nedan utförts. Vissa steg kommer diskuteras djupare i kapitel 4.

3.1 Verktyg

För detta projekt användes R programmering i programmet RStudio, paket som användes var bland annat dplyr, MASS, ggplot2, alla paket som nyttjades finns i början av koden.

3.2 Insamling av data

Datainsamling utfördes gemensamt som en del av ett grupparbete. 250 observationer av Volvobilar samlades in manuellt från blocket, 14 variabler samlades in för varje observation. En Proof of concept utfördes på den insamlade datan och efter att ha uppnått godkända resultat utfördes resten av insamling. Den slutgiltiga datan bestod av 14 variabler och cirka 750 rader.

3.3 Data städning och EDA

Data städningen skedde i flera olika steg och nedan kommer de visas. En simpel EDA utfördes efter importering av data, men i linjär regressionsanalys sker steg mer iterativt och datan utforskades och undersöktes kontinuerligt under data städningen. Det finns därför inget tydligt separerbart avsnitt för just EDA som man ofta finner i Maskininlärning.

3.3.1 Dubblett hantering i Excel

Den fullständiga datan innehöll cirka 750 observationer när den var sammanställd. 25 dubblett rader lokaliserades och raderades genom skapandet av en ny kolumn som innehöll värdena i alla kolumner för den raden, med hjälp av konditionell formatering kunde dubbletter lokaliseras och markeras.

3.3.2 Data rensning i R

Genom funktionen ”readxl” i R kunde datan importeras från excel till en data frame i R. Totalt fanns det 15 variabler och 725 rader. För att reducera variablerna exkluderades kolumnerna ”Index”, ”Färg”, och ”Märke”. Kolumnen ”Datum i trafik” omvandlades till ”Dagar i trafik” genom att subtrahera dagens datum med datum i trafik för varje rad. Raderna som saknade ett datum i trafik fick värdet 0.

3.3.3 Data konvertering i R

Kolumnen ”Hästkrafter” blev efter importering till R konverterad till character format, för att återigen konvertera den till numerisk säkerställdes det först att kolumnen endast innehöll numeriska värden, detta gjordes genom funktionen ”gsub” och alla icke numeriska värden togs bort. Värdena kunde därefter konverteras till numeriska. Kolumnen ”Växellåda” innehöll ”Automat”, ”Manuell” och ”Automat\r\n”, den sistnämnda tyder på att vissa av raderna i excel innehöll en radbrytning efter ”Automat”. Genom att använda ”gsub” kunde ”\r\n” tas bort och kvar återstod endast ”Manuell” och ”Automat”.

3.3.4 Reducering av unika värden

Kolumnen ”Motorstorlek” innehöll 198 saknade värden, 189 av dessa var el bilar där motorstorlek inte mäts i kubikcentimeter som för vanliga förbränningsmotorer. Dessa ersattes med ”Electric_engine”. För att reducera de unika värdena i kolumnen dividerades de numeriska värdena med 100 för att sedan avrundas till närmaste tiotal, efter avrundningen dividerades de med 10 för att omvandla måttet till liter vilket är ett vanligt förekommande mått för motorstorlekar. Detta reducerade kolumnen till 11 värden. Kolumnen undersöktes med hjälp av en bar plot för att se antalet observationer för de olika värdena. För att ytterliga minska värdena

delades de upp i ”Smaller_engine” bestående av 1.5, 1.6, 1.8 och 1.9 liters motorer, ”Medium_engine” bestående av 2.0 liters motorer, ”Bigger_engine” bestående av 2.3, 2.4, 2.5, 3.0 och 3.2 liters motorer. Efter reducering fanns det 4 unika värden kvar i kolumnen.

3.3.5 Saknade värden

Genom att återigen titta på saknade värden för varje kolumn med hjälp av funktionen ”colSums” kunde 6 saknade värden i kolumnen ”Biltyp” upptäckas. Genom paketet ”dplyr” kunde dessa rader tas fram och 5 rader upptäcktes där biltyp var det enda saknade värdet för observationen. Eftersom dessa bilar endast finns i kombi och inte i andra biltyper kunde de saknade värdena åtgärdas. Kvar återstod 23 rader som inte logiskt kunde åtgärdas och de togs därmed bort. Genom att skapa en barplot över antalet bilar för varje biltyp med hjälp av paketet ”ggplot2” kunde det upptäckas att endast 4 rader sammanlagt bestod av typerna ”Familjebuss”, ”Cab” och ”Coupe”, dessa rader togs bort.

3.3.6 Dummy encoding

Efter att den initiala data städningen var klar och önskade kolumner var justerade för att innehålla färre unika värden utfördes en dummy encoding. Detta gjordes med hjälp av paketet ”car” i R vilket möjliggjorde simpel encoding där ett värde för varje kolumn blev till baseline. Den slutgiltiga datan bestod av 17 variabler och 708 rader.

3.3.7 Plotting, outliers och korrelation

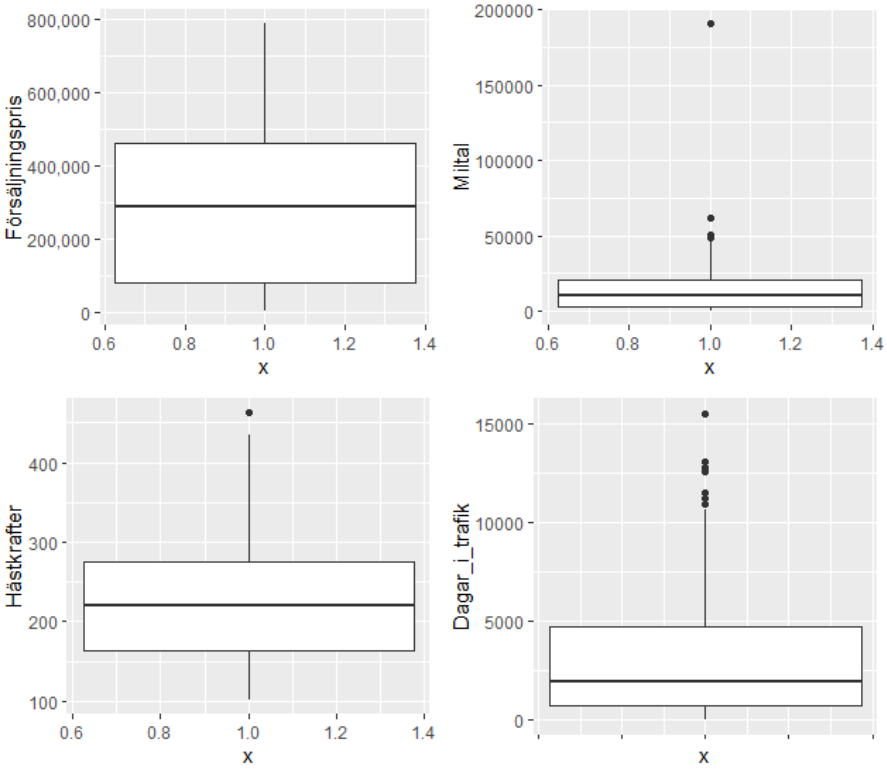
Först skapades en heatmap för att undersöka korrelationen mellan de olika variablerna. Det upptäcktes att variablerna ”Bränsle_El” och ”Motorstorlek_Electric_engine” hade en korrelation på 1, därför togs ”Bränsle_El” bort. För att undersöka sambandet mellan olika variabler och försäljningspriser skapades scatter plots för ”Hästkrafter”, ”Miltal”, ”Modellår” och ”Dagar i trafik”. För att undersöka outliers skapades box plots. En tydlig outlier upptäcktes för variabeln miltal, vid närmare undersökning konstaterades det att det höga miltalet berodde på en felskrivning, det justerades och värdet konverterades till mil som resten av datan. ”Dagar_i_trafik” innehöll även outliers men vid närmare undersökning var det tydligt att dessa var äldre bilar vilket förklarar deras höga värden för dagar i trafik.

3.4 Modeller

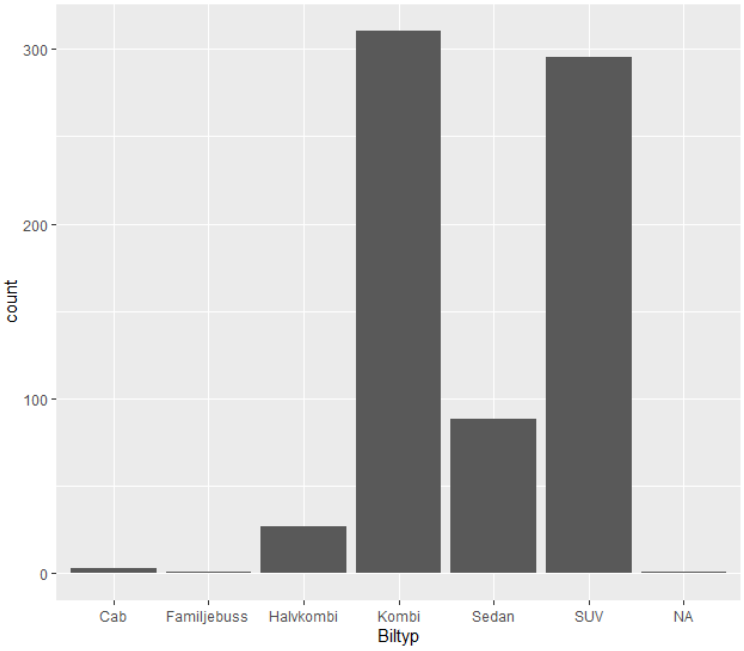
För detta projekt användes endast multipel linjär regressions med hjälp av funktionen ”lm” i R. En modell med alla variabler tränades för att undersöka multikollinearitet mellan variabler, hög kollineratitet mellan ”Biltyp_SUV” och ”Biltyp_Kombi” noterades och dessa variabler slogs ihop till ”Biltyp_Big”. Hög kollineratitet upptäcktes mellan dagar i trafik och modellår, ”Dagar i trafik” exkluderades därmed från modellerna. En modell tränades med alla kvarstående variabler. En ny modell tränades endast med de signifikanta variablerna. Slutligen tränades en tredje modell med variabler som logiskt kan antas påverka priset på en bil. Best subset selection utfördes även men de bästa variablerna var de som användes i modell 2.

3.5 Utvärdering och förbättringar

Modellerna utvärderades och jämfördes med måtten Adjusted R^2. AIC, BIC samt RMSE. Den bäst presterande modellen valdes och undersöktes närmare. Diagnostiska plotsen som presenterades i teorin tydde på flera problem och antaganden som inte var uppfyllda, interaktionseffekter mellan variabler och transformeringar utfördes för att förbättra modellen. Dessa kommer att presenteras djupare i kapitel 4.



Figur 5 Boxplots för x variabler



Figur 4 bar plots för Biltyp kolumn

4 Resultat och Diskussion

I detta kapitel kommer resultat för projektet att visas och diskuteras. Vissa val förknippat med data rensning och modellförbättring kommer att diskuteras och motiveras.

4.1 Datainsamling och rensning

Datainsamling utfördes som tidigare nämnt i ett gemensamt grupparbete. För att möjliggöra olika typer av analyser för gruppens medlemmar beslutades det att samla in 14 olika variabler. Detta innebar att variabler behövde reducerades eftersom många av variablerna innehöll flera unika värden vilket i sin tur innebar ett stort antal variabler efter dummy encoding. För att begränsa antalet variabler exkluderades från start ”Märke” eftersom datan endast innehöll Volvobilar, ”Färg” eftersom Volvo primärt säljer familjebilar antogs detta inte ha lika stor påverkan på priset som till exempel det har på sportbilar. Variabeln ”Modell” exkluderades också även fast det kan antas ha en stor påverkan på pris innebar de 36 unika värdena för stor risk för overfitting. Ett stort antal variabler innebär också att variabel selektionen inte blir lika pålitlig och därför togs vissa variabler bort innan modellträning och vissa kolumner blev indelade i färre värden för att minska dummy variables. Kolumnen motorstorlek innehöll 13 unika värden vilket reducerades till 3. Detta visade sig vara lönsamt eftersom variablerna reducerades stort och ett bra resultat uppnåddes ändå.

4.2 Modeller och utvärdering

4.2.1 Modeller

Initialt tränades och utvärderas 3 olika modeller med olika variabler.

1. Modell 1 innehöll alla variabler förutom dagar_i_trafik på grund av hög multikollineratiet.
2. Modell 2 innehöll endast de variabler som tydde på signifikans för modellen.
3. Modell 3 innehöll variabler som logiskt antas ha en påverkan på priset.

Best subset selection utfördes där grafen visade på att 7 var det optimala antalet variabler, de 7 variabler som gav bäst resultat var samma variabler som användes i modell 2 och därmed behövde inte en ny modell tränas och utvärderas.

4.2.2 Utvärdering

Modell 2 uppnådde bäst resultaten på BIC samt RMSE men lite sämre på adjr2 och AIC jämfört med modell 1, Detta är logiskt då modell 2 innehåller färre variabler vilket ofta innebär ett lägre värde på BIC. Modell 2 valdes därför att undersökas närmare då färre variabler simplifierar modellen vilket i detta fall är önskvärt då det underlättar inferens.

Mått:	**Adjusted R^2**	**BIC**	**AIC**	**RMSE**
Modell 1	0.8878379	17863.87	17795.43	76570.54
Modell 2	0.8859050	17842.69	17801.62	75963.47
Modell 3	0.8833306	17864.03	17818.41	76426.22

Tabell 1: Mått för de olika modellerna

4.3 Problem

Genom att använda ”plot()” funktionen i R får man tillgång till de olika diagnostiska plots som nämndes i kapitel 3. Nedan kommer upptäckter av potentiella problem presenteras.

4.3.1 Icke linjärt samband

Residuals vs Fitted plotten visade ett tydligt mönster vilket tyder på ett icke linjärt samband för residualerna. (se figur 6)

4.3.2 Icke normalfördelade residualer

QQ plotten visade en viss avvikning från linjen i den övre svansen vilket tyder på icke normal fördelade residualer, denna avvikning ansågs inte orsaka ett större problem men var värt att notera. (se figur 6)

4.3.3 Heteroskedasticitet

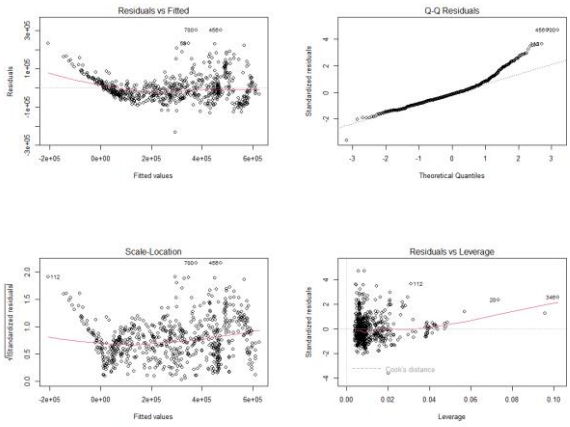
Scale-Location plotten visade tydliga tecken på heteroskedasticitet och variansen för residualerna tycktes öka. (se figur 6)

4.3.4 Outliers och high leverage punkter

Residuals vs Leverage plotten visade vissa potentiella outliers men då det var planerat att utföra transformationer avvaktades det med att hantera dessa. (se figur 6)

4.4 Åtgärder

För att åtgärda problemen provades olika metoder, först tränades en ny modell med en interaktionsterm ”Hästkrafter * Motorstorlek_Electric_engine”. Detta förbättra modellen lite med ett högre adjr2 och minskade på mönstret i den första plotten smått. För att försöka förbättra modellen ännu mer provades 3 olika transformationer. Log transformation av den oberoende variabeln vilket inte gav någon tydlig förbättring. En box-cox transformation vilket är en teknik som transformerar en variabel så att datan ska efterlikna en normal fördelning, detta gav väldigt bra resultat och åtgärdade både det icke linjära sambandet samt heteroskedasticitet. Vid användning av box cox transformation blir tolkningar och inferens mer komplicerade och svårtolkade, därför provades slutligen en ”roten ur” transformation med hjälp av funktionen ”sqrt” i R på den beroende variabeln och samtliga kvantativa oberoende variabler. Detta gav liknande resultaten som box-cox transformationen och är lättare att tolka. I Residuals vs Leverage plotten finns en tydlig ”high influence point” som ligger utanför det prickade området, den syns även i de andra plotterna där den ligger långt ifrån andra punkter. Vid närmare undersökning upptäcktes det att observationen har ett pris på cirka 62 000 även fast det är en 2022 modell och endast har gått 2840 mil, detta tyder på en felskrivning då det är ett orimligt pris och observationen togs bort.



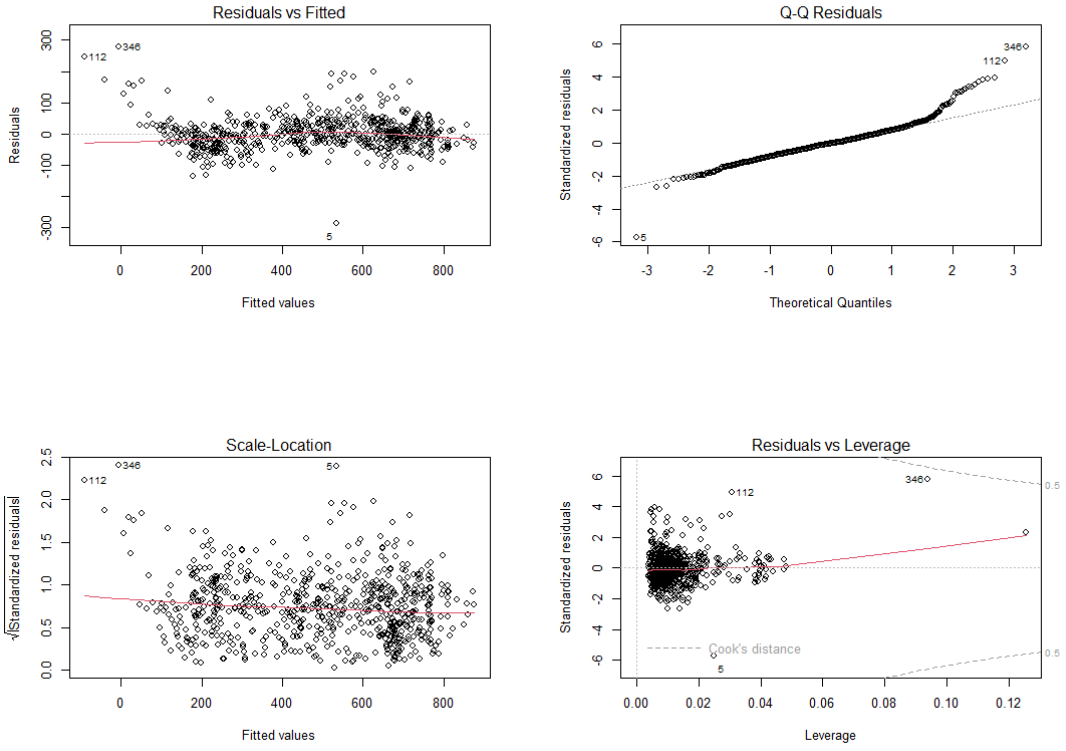
Figur 7 Diagnostical plots för modell 2

4.4.1 Utvärdering efter förbättringar

Eftersom ”roten ur” transformationen gav bra resultat och underlättar inferens signifkant jämfört med Box-cox transformationen beslutades det att ”roten ur” är en mer lämpligt transformation att utföra. Modellen utvärderas och uppnådde 0.9497 adjusted R^2 vilket är högre än föregående modeller. Genom cross validation mättes RMSE för modellen, eftersom variablerna är transformerade krävs en backtransformation för att kunna få RMSE på samma skala som föregående modeller. Detta gjordes genom att spara prediktionerna och observationerna från cross validation, kvadrera de och beräkna RMSE. Resultaten blev 49221 vilket är en förbättring jämfört med föregående modeller.

4.5 Inferens

Slutligen utfördes ett test av inferens där en ny observation hämtades från blocket och ett konifidensintervall samt prediktionsintervall utfördes. Ett prediktionsintervall beräknades för observationen och det prediktera priset var fel med cirka 15000 kr. Inferens utfördes även efter att modellen hade tränats genom funktionen ”summary” i R. Detta skapar automatiskt nollhypoteser för alla oberoende variabler och man direkt tolka ifall en variabel har en statistisk signifikant påverkan på den beroende variabeln genom ett p värde. Alla variabler förutom ”biltyp_big” och ”biltyp_sedan” uppnådde resultat som tydde på att man kan förkasta nollhypotesen att variabeln inte har en påverkan på pris.



Figur 8 Diagnostik plots efter transformation

5 Slutsatser

5.1 Frågeställning

Det slutgiltiga resultatet av projektet ansågs vara lyckat,

1. En modell med Adjusted R^2 över 0.9 uppnåddes.
2. Det är statistiskt säkerställt att El bilar är dyrare än andra typer av bilar.
3. Det går inte med statistik signifikans att påstå att biltyp har en påverkan på pris.
4. Antaganden var inte helt uppfyllda men efter justeringar och transformationer lyckades antaganden bli tillräckligt uppfyllda för att inte orsaka problem.

5.2 Problem och lösningar

Under arbetet uppstod det flera problem.

1. Stort antal variabler och unika kategoriska värden.
2. Multikollineratitet mellan variabler.
3. Antaganden för linjär regressions var inte uppfyllda för datan.

Dessa problem kunde lösas via olika metoder och åtgärder vilket i slutändan gav bra resultat då en modell med bra prestation skapades och inferens gick att utföra vilket var målet med arbetet.

5.3 Nästa steg

Eftersom tid var begränsat utfördes inte alla steg som potentiellt kunde förbättrat modellen. Det viktigaste och mest uppenbara hade varit att samla in fler observationer. Detta hade inte bara tillåtit modellen att tränas mer utan även möjliggjort inkludering av till exempel modell kolumnen som i detta arbete var tvungen att exkluderas. Större data set hade även möjliggjort för fler variabler vilket hade gjort variabel selektion möjlig med flera alternativ vilket potentiellt hade kunnat förbättra analysen. Datan som samlades in beslutades gemensamt vara tillräckligt eftersom det möjliggjorde bra analyser och modeller och som sagt fanns det inte mer tid för datainsamling. Vidare förbättring av modellen hade också gjorts möjlig om ytterligare variabler hade samlats in, dessa hade kunnat vara till exempel utrustning, om en bil har R-design eller inte och geografisk plats. Dessa är faktorer som kan antas ha en ytterligare påverkan på priset men eftersom tiden var begränsat inkluderades inte dessa.

Appendix A

```
Coefficients:
                                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                                   -4.589e+04  2.112e+03 -21.729  < 2e-16 ***
sqrt(Miltal)                                  -1.512e+00  7.107e-02 -21.275  < 2e-16 ***
sqrt(Modellår)                                 1.029e+03  4.704e+01  21.863  < 2e-16 ***
sqrt(Hästkrafter)                              2.118e+01  1.038e+00  20.402  < 2e-16 ***
Biltyp_Sedan                                   2.053e+01  1.184e+01   1.734   0.0833 .
Motorstorlek_Electric_engine                   2.098e+02  3.408e+01   6.155 1.26e-09 ***
Motorstorlek_Medium_engine                     3.032e+01  6.042e+00   5.018 6.63e-07 ***
Biltyp_Big                                     1.725e+01  1.048e+01   1.646   0.1002
sqrt(Hästkrafter):Motorstorlek_Electric_engine -1.142e+01  1.925e+00  -5.934 4.65e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 50.51 on 699 degrees of freedom
Multiple R-squared:  0.9479,    Adjusted R-squared:  0.9473
F-statistic:  1590 on 8 and 699 DF,  p-value: < 2.2e-16
```

Figur 9 Summary av slutgiltig modell

```
Predicted Price Lower prediction Interval Upper prediction Interval
                 452797.0                  330996.0                  593641
.4
```

Figur 10 Prediktionsintervall för obersvation med priset 469900

Källförteckning

Bomae, K. (2015). *Understanding Diagnostic Plots for Linear Regression Analysis*

https://library.virginia.edu/data/articles/diagnostic-plots

Lewinson, E. (2023). *A Comprehensive Guide to Interaction Terms in Linear Regression.*

https://developer.nvidia.com/blog/a-comprehensive-guide-to-interaction-terms-in-linear-regression/

Frost, J. (2024). *Residual Sum of Squares (RSS) Explained*

https://statisticsbyjim.com/regression/residual-sum-of-squares-rss/#comments

Frost, J. (2017) *How to Interpret Adjusted R-Squared and Predicted R-Squared in Regression Analysis*

https://statisticsbyjim.com/regression/interpret-adjusted-r-squared-predicted-r-squared-regression/#comments

Gareth, J., Witten, D., Hastie, T., Tibshirani, R., (2023). *An introduction to Statistical Learning*.

Gèron, A. (2019). *Hands-on Machine learning with Scikit-Learn Keras & TensorFlow. Canada: O´Reilly.*

Mans Magnusson, Markus Kainu, Janne Huovari, and Leo Lahti (rOpenGov). pxweb: R tools for PXWEB API. URL: http://github.com/ropengov/pxweb

Taylor, S. (u.å). *What is R-Squared*

https://corporatefinanceinstitute.com/resources/data-science/r-squared/

Zajic, A. (2022). *What is Akaike Information Criterion(AIC)?*

https://builtin.com/data-science/what-is-aic