Förutsägelse av Volvo V60-priser med hjälp av Multipel Linjär Regression

Girlie Razon

EC Utbildning

Kunskapskontroll – R Programmering

202404

Abstract

This study aimed to predict the prices of Volvo V60 cars using multiple linear regression models. The dataset was split into training, validation, and test sets, comprising 248, 53, and 51 observations, respectively. Two models were constructed and evaluated: a simple linear regression model with only the intercept term and a multiple linear regression model including predictor variables such as `Model_Year`, `Mileage`, `Fuel Type`, `Gearbox`, and `Horsepower`. The multiple linear regression model exhibited promising results, explaining approximately 84.3% of the variance in car prices. However, the significance of the `Gearbox` variable was marginal ($p = 0.0839$), indicating a need for further investigation. Additionally, a subset selection process identified the most influential variables, including `Model_Year`, `Mileage`, `Fuel Type`, and `Horsepower`. Model validation was conducted using the validation and test sets, and the root mean squared error (RMSE) was calculated for each set. The RMSE values were found to be 32,170.44 and 41,046.34 for the validation and test sets, respectively, indicating the model's ability to generalize to new data. In conclusion, the multiple linear regression model demonstrates potential for predicting Volvo V60 car prices. However, further refinement and validation are recommended to enhance the model's accuracy and robustness.

Förkortningar och Begrepp

MSE: Mean Squared Error

RMSE: Root Mean Square Error

AIC: Akaike Information Criterion

BIC: Bayesian Information Criterion

QQ: Quantile-Quantile

VIF: Variance Inflation Factor

API: Application Programming Interface

SCB: Statistiska Centralbyrån

PI: Prediction Interval

CI: Confidence Interval

EDA: Exploratory Data Analysis

Innehållsförteckning

Abstract 2
Förkortningar och Begrepp **MSE**: Mean Squared Error 3
1 Inledning 1
1.1 Bakgrund och Problemmotivering 1
1.2 Överallt Syfte 1
1.3 Problemformulering 1
1.4 Frågeställning 1
1.5 Omfattning 2
1.6 Översikt 2
2 Teori 3
2.1 Datainsamling 3
2.2 EDA (Exploratory Data Analysis) 3
2.3 Utforskande av korrelationer i car-datasetet: En Tilläggsanalys 4
2.4 Utvärderingsmått 5
2.4.1 Absoluta mått 5
2.4.2 Relativa mått 6
2.5 Regressionsmodell: Linjär Regressionsmodell 6
2.5.1 Enkel Linjär Regressionsmodell: Intercept-Only Modell | Noll Modell 6
2.6 Multipel Linjär Regressionsmodell 8
3 Metod 9
4 Resultat och Diskussion 10
5 Slutsatser 13
6 Teoretiska frågor 14
6.1 Fråga 1 14
6.2 Fråga 2 14
6.3 Fråga 3 14
6.4 Fråga 4 15
6.5 Fråga 5 15
6.6 Fråga 6 15
6.7 Fråga 7 16
7 Självutvärdering 17
Appendix A 18
Källförteckning 27

1 Inledning

Att förstå och prognostisera priserna på bilar är av avgörande betydelse för bilindustrin och konsumenterna. En korrekt prissättning kan påverka både företagens lönsamhet och konsumenternas köpbeslut. I synnerhet för Volvo V60-bilar, en populär modell på marknaden, är det av intresse att analysera vilka faktorer som påverkar dess priser. Genom att identifiera och förstå dessa faktorer kan företag och konsumenter fatta informerade beslut om bilköp och försäljning.

1.1 Bakgrund och Problemmotivering

Volvo V60 är en välkänd modell som har fått stor uppmärksamhet för sin kombination av prestanda, säkerhet och miljövänlighet. Modellen har genomgått flera uppdateringar och erbjuder olika motoralternativ för att möta kundernas behov och preferenser. Med den ökande efterfrågan på miljövänliga fordon och den ständigt föränderliga bilmarknaden är det av stort intresse att förstå vilka faktorer som påverkar prissättningen av Volvo V60-bilar.

1.2 Överallt Syfte

Syftet med denna rapport är att undersöka och förutsäga priserna på Volvo V60-bilar med hjälp av olika modeller för multipel linjär regression. Genom att analysera variabler som modellår, körsträcka, bränsletyp, växellådstyp och hästkrafter, syftar studien till att förstå vilka faktorer som har störst inverkan på bilpriserna.

1.3 Problemformulering

Det övergripande problemet som denna studie syftar till att lösa är att förutsäga Volvo V60-bilarnas priser baserat på olika påverkansfaktorer.

1.4 Frågeställning

För att uppfylla syftet med denna rapport kommer följande frågeställningar att besvaras:

1. Vilken inverkan har modellåret på priset på Volvo V60-bilar?
2. Hur påverkar körsträckan bilpriserna?
3. Finns det någon korrelation mellan bränsletyp och bilpriser?
4. Är växellådstyp en signifikant faktor för bilprissättning?
5. Hur relaterar hästkrafter till priset på Volvo V60-bilar?

1.5 Omfattning

Studien kommer att fokusera på att analysera priserna på Volvo V60-bilar på den svenska marknaden. Datainsamlingen och analysen kommer att begränsas till bilar av modellen Volvo V60 och kommer att omfatta ett antal relevanta variabler som kan påverka bilpriserna.

1.6 Översikt

Rapporten kommer att fortsätta med en översikt över de metodologiska tillvägagångssätten och analyserna som används för att uppnå studiens syfte och besvara frågeställningarna. Därefter presenteras resultaten av studien och diskuteras utifrån dess implikationer och slutsatser.

2 Teori

Detta avsnitt ger en översikt av den teoretiska bakgrunden som är relevant för att förstå sammanhanget av detta projekt.

2.1 Datainsamling

Vi på Grupp 2 samlade data från Blocket, en plattform för bilannonser där säljare och köpare möts. Tillsammans med Lidiia Kashevarova, Aikaterini Antoniou, Manna Mulanga, Jacob Andersson, Andreas Wendel och Nil Abukar samlade vi totalt 352 observationer från åren 2018 till 2023, med 50 till 71 observationer per år. Vårt mål är att förutse bilpriser och undersöka vilka faktorer som påverkar priserna. Detta kommer att uppnås genom att modellera och analysera Volvo V60-bilar med variabler såsom modellår, körsträcka, bränsletyp, växellådstyp, hästkrafter och pris. För att säkerställa effektiviteten definierade vi syftet med modellen och valde lämpliga fordonstyper. Genom att följa en strukturerad process genomförde vi en Proof of Concept (POC), inklusive att definiera projektidén, utveckla och testa prototypen samt samla feedback och insikter. Samarbetet inom gruppen var effektivt, med noggrann planering och tydlig kommunikation. För framtida förbättringar ser jag möjligheter till att öka mångfalden av åsikter och erfarenheter för att stimulera innovationen. Min styrka ligger i att organisera och kommunicera väl på chatten, men jag strävar efter att integrera olika perspektiv för att öka kreativiteten och innovationsförmågan ytterligare.

2.2 EDA (Exploratory Data Analysis)

För att förstå datan utfördes en explorativ dataanalys (EDA). Datan består av 352 rader och 6 kolumner. De sex kolumnerna inkluderar information om modellår, körsträcka, bränsletyp, växellådstyp, hästkrafter och pris på bilarna.

För att förbereda datan för analys kodades de kategoriska variablerna om till faktorer. Efter kodningen består datan av 352 observationer och 6 variabler.

Genom att summera datan ser vi att den sträcker sig från modellår 2018 till 2023. Körsträckan varierar från 567 till 48 582 miles. Bränsletypen inkluderar diesel, bensin och hybrid. Växellådan kan vara automatisk eller manuell. Hästkrafterna sträcker sig från 150 till 463. Priserna varierar från 129 900 till 599 900. För att komplettera EDA-processen delades datan upp i train dataset, validation dataset och test dataset. Dimensionerna för varje dataset är som följer:

- Train dataset: 248 rader och 6 kolumner
- Validation dataset: 53 rader och 6 kolumner
- Test dataset: 51 rader och 6 kolumner

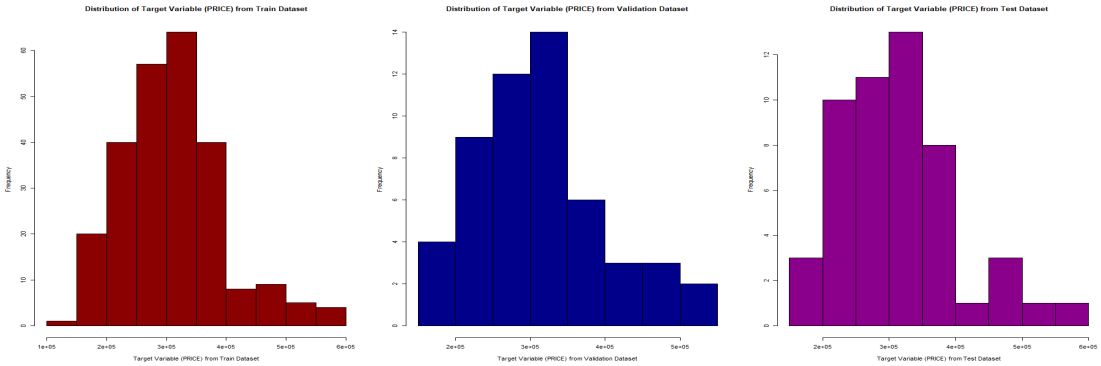

Figure 1: Distribution of Target Variable (Price) from Train, Validation & Test datasets

2.3 Utforskande av korrelationer i car-datasetet: En Tilläggsanalys

Korrelationerna i car_data-datasetet granskas för att förstå hur variablerna relaterar till varandra. Genom att analysera dessa korrelationer kan mönster och samband upptäckas som hjälper till att identifiera de viktigaste faktorerna för att förutsäga bilpriser. Korrelationen kan vara positiv, negativ eller neutral, och dess styrka indikeras av korrelationskoefficientens värde. Till exempel visar "correlation plot" nedan analysen att modellår och pris har en positiv korrelation på 0.72, medan körsträcka och pris har en negativ korrelation på -0.64. Detta ger en bättre förståelse för datasetets struktur och vägleder i skapandet av en effektiv modell för att analysera och förutsäga bilpriser.

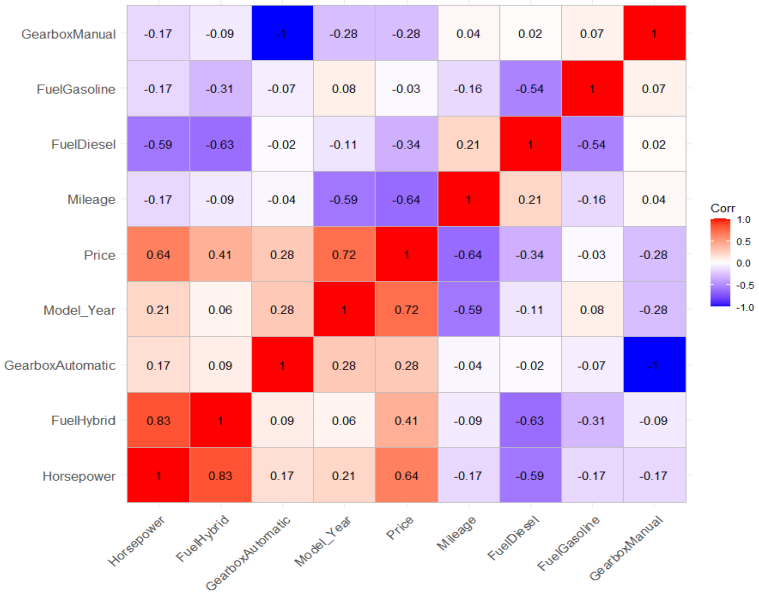

Figure 2: Correlation Plot between the target variable (y) & the feature variables (x)

2.4 Utvärderingsmått

Bedömningsmetoder är strategier som används för att utvärdera prestandan hos en regressionsmodell. Dessa metoder ger insikt i hur effektiv modellen är och hur exakta dess prognoser är. Följande mätvärden tillämpades i detta projekt.

2.4.1 Absoluta mått

Absoluta mått är mått som mäter avståndet eller skillnaden mellan två värden utan att ta hänsyn till riktningen av skillnaden. Det är ett sätt att kvantifiera avståndet mellan två punkter utan att bry sig om vilken punkt som är större eller mindre än den andra.

i. Root Mean Squared Error (RMSE) är kvadratroten av MSE, vilket ger ett mått på genomsnittlig absolut förutsägelsefel. Formel är:

$$\textbf{RMSE} = \sqrt{\frac{l}{n}\sum_1^n (y_i - \hat{y}_i)^2}$$

ii. Bayesian Information Criterion (BIC), är ett modellvalskriterium som används för att välja den bästa modellen bland ett antal alternativ. Modeller med lägre BIC-värden föredras vanligtvis eftersom de antyder en bättre balans mellan modellens passning till data och dess komplexitet. BIC straffar modeller med fler parametrar, vilket hjälper till att undvika överanpassning och främjar enklare och mer generaliserbara modeller. Formeln är:

$$\textbf{BIC} = -2 * \log(L) + k * \log(n)$$

- ***L***, är den maximala sannolikheten
- ***k,*** är antalet parametrar
- ***n,*** är antalet observationer

iii. Sigma är en absolut måttenhet som används för att mäta standardavvikelsen eller spridningen av en variabel i en datamängd. Standardavvikelsen σ beräknas enligt följande formel:

$$\boldsymbol{\sigma} = \sqrt{\frac{\sum(\boldsymbol{x_i} - \overline{\boldsymbol{x}})^2}{\boldsymbol{N}}}$$

Där $\boldsymbol{x_i}$ representerar varje observation, $\overline{\boldsymbol{x}}$ är medelvärdet av observationerna och $\boldsymbol{N}$ är antalet observationer.

2.4.2 Relativa mått

Relativa mått är mått som tar hänsyn till förhållandet eller proportionen mellan olika värden eller variabler. De jämför storleken på en variabel med en annan för att ge insikt om deras relativa betydelse eller förändring över tid.

iv. Determinationskoefficienten R-squared (R^2), visar hur stor andel av variationen i den beroende variabeln (Y) som kan förklaras med sambandet från den oberoende variabeln (X). Formeln är:

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

- ***TSS*** visar den totala variationen.
- ***RSS*** visar den oförklarade variationen.
- Täljaren ***(TSS - RSS)*** visar den förklarade variationen.

v. Adjusted R-squared (Adj R^2), är en justerad version av R^2 som tar hänsyn till antalet prediktorer i modellen. Om man tränar olika regressionsmodeller då skulle man välja den med högst ”Adjusted R^2”. Formeln är:

$$\boldsymbol{R^2_{adj}} = 1 - (1 - R^2)\frac{n-1}{n-P-1}$$

vi. Akaike Information Criterion (AIC), är ett mått som balanserar modellens passform med dess komplexitet, vilket är användbart vid jämförelse av modeller. Formeln enligt (Rebecca Bevans, t.o.m. 2020–2023) är:

$$\boldsymbol{AIC = 2K - 2ln\,(\hat{L})}$$

Där ***k*** är antalet uppskattade parametrar i modellen och $\boldsymbol{\hat{L}}$ är det maximerade värdet av sannolikhetsfunktionen för modellen.

2.5 Regressionsmodell: Linjär Regressionsmodell

2.5.1 Enkel Linjär Regressionsmodell: Intercept-Only Modell | Noll Modell

Intercept-only-modellen är en enkel linjär regressionsmodell som endast inkluderar en intercept (konstant term) och inga prediktorvariabler. Denna modell representerar det genomsnittliga svaret för alla observationer, under antagandet att prediktorvariablerna har ett värde på noll. Resultatet från regressionsanalysen visar att en Intercept-only-modell har använts, vilket innebär att endast en konstant term ingår i modellen utan några

prediktorvariabler. Spridningen av residualerna sträcker sig från -184799 till 285201, med kvartiler som omfattar intervallet från -59824 till 44326. Koefficienten för intercept (konstanten) är 314699 med en standardavvikelse på 5543. Det höga t-värdet (56.77) och den låga p-värdet (<2e-16) indikerar att interceptet är statistiskt signifikant. Residualstandardfelet, som mäter spridningen av residualerna runt den förutsagda linjen, är 87290 med 247 frihetsgrader.

Intercept-only Modell formel: $\widehat{\boldsymbol{Y}} = \boldsymbol{\beta_0}$

- $\widehat{\boldsymbol{Y}}$ är det förutsagda värdet av responsvariabeln.
- $\boldsymbol{\beta_0}$ är interceptterman (konstanten).

Enligt Philip Lloyd, Specialist Calculus Teacher, Motivator and Baroque Trumpet Soloist, säger: *"INTERCEPT" hänvisar vanligtvis till de punkter där en graf korsar AXLARNA*.

Nedan är visualiseringen av Intercept-only-modellen enligt Philip Lloyd.

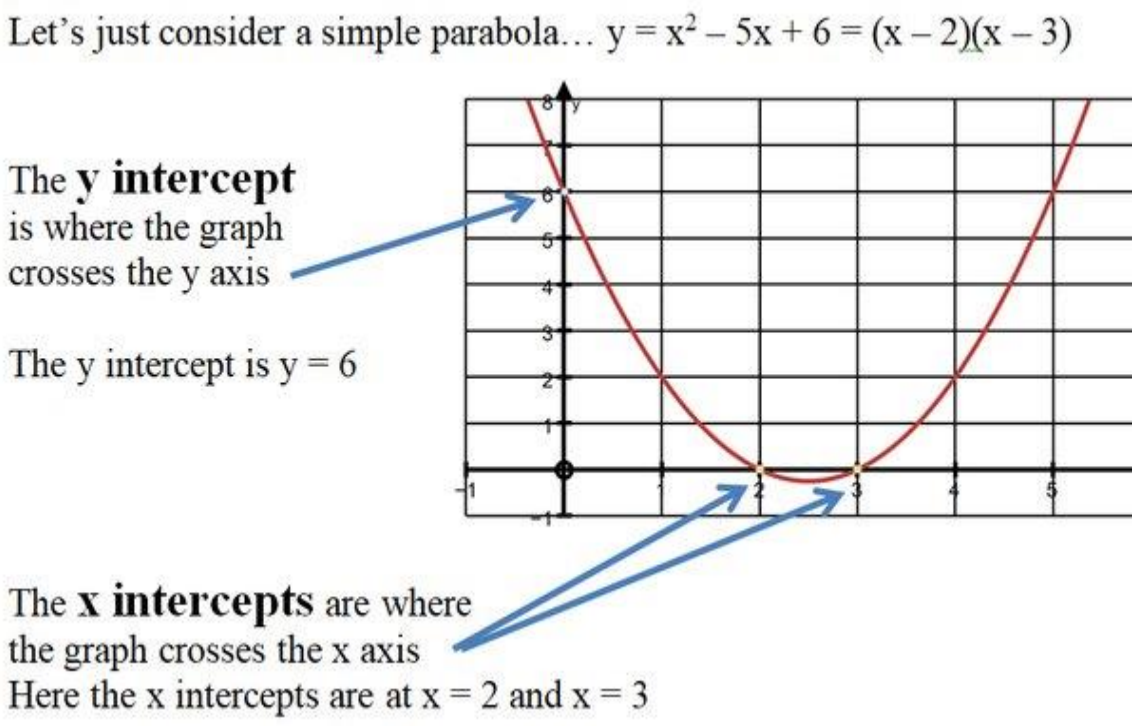

Figure 3: Simple Linear Regression: Intercept-only Model Visualization

2.6 Multipel Linjär Regressionsmodell

Multipel Linjär Regressionsmodell används för att förstå sambandet mellan flera oberoende variabler och en beroende variabel. Formeln är:

$$\boldsymbol{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \beta_p X_\rho + \varepsilon}$$

Där:

- $\boldsymbol{Y}$, är den beroende variabeln (bilpris).
- $\boldsymbol{\beta_0}$, är interceptet (konstanten).
- $\boldsymbol{\beta_1, \beta_2, \ldots, \beta_p}$ är koefficienterna för varje prediktorvariabel (som modellår, körsträcka, bränsletyp etc.).
- $\boldsymbol{X_1, X_2, \ldots, X_\rho}$ är de oberoende variablerna (pediktorvariablerna).
- $\boldsymbol{\varepsilon}$ är feletstermen (residualen).

3 Metod

I detta avsnitt beskrivs metodiken som användes för att utföra analysen av bilförsäljningsdata. Analysen innefattar dataförberedelse, explorativ dataanalys (EDA), modellutveckling, modellvalidering och modellinferens. Dataanalysen utfördes med hjälp av programmeringsspråket R och flera relevanta paket såsom `tidyverse`, `caret`, `ggplot2`, och `MASS`. Följande steg tillämpades för att genomföra analysen:

- Datainsamling: Datasetet erhölls genom att läsa in en Excelfil som innehöll bilförsäljningsdata.
- Dataförberedelse: Kategoriska variabler kodades om till faktorer för att underlätta analysen.
- Utforskande dataanalys (EDA): Fördelningen av målvariabeln och relationen mellan prediktorer och målvariabeln undersöktes med hjälp av histogram och scatterplot.
- Funktionskonstruktion: Dummyvariabler skapades för kategoriska kolumner för att inkludera dem i modellerna.
- Datadelning: Datasetet delades upp i tränings-, validerings-, och testuppsättningar för att kunna utvärdera modellprestanda.
- Modellutveckling: Flera linjära regressionsmodeller tränades på träningsdata, inklusive modeller med endast intercept och modeller med alla prediktorer. Funktionellt urval utfördes med bakåteliminering, framåteliminering och båda metoderna.
- Modelljämförelse: Prestandan hos modellerna jämfördes med avseende på relevanta prestandamått.
- Diagnostisk analys: Modellens förutsättningar och eventuella avvikelser undersöktes genom diagnostiska plottar.
- Modellvalidering: Den valda modellen utvärderades på en valideringsuppsättning för att bedöma dess generaliseringsförmåga.
- Modellinferens: Koefficienter från den valda modellen extraherades och hypotesprövning utfördes.
- Modelltestning: Den valda modellen testades på en separat testuppsättning för att bedöma dess prediktiva förmåga.
- Visualisering: Testdatamängden plottades med konfidens- och förutsägelseintervall för att ge en översiktlig bild av modellens prestanda.

Denna metodik möjliggjorde en systematisk analys av bilförsäljningsdata och en grundlig utvärdering av linjära regressionsmodeller för att förutsäga bilpriser.

4 Resultat och Diskussion

Sammanfattningen av bildata visar att datasetet innehåller information om 352 bilar med variation i årsmodell, körsträcka, bränsletyp, växellåda, hästkrafter och pris. Årsmodellen sträcker sig från 2018 till 2023, med en median på 2021. Körsträckan varierar från 567 till 48,582 miles, med en median på 7496 miles. De tre vanligaste bränsletyperna är Diesel, Bensin och Hybrid. De flesta bilarna har en automatisk växellåda. Hästkrafterna varierar från 150 till 463, med en median på 198. Priset på bilarna sträcker sig från 129,900 till 599,900, med en median på 309,900.

För att förutsäga priset på bilarna tränades olika regressionsmodeller på träningsdatan, som bestod av 248 observationer, och utvärderades sedan på valideringsdatan med 53 observationer och testdatan med 51 observationer.

En enkel linjär regressionsmodell, `ml_model_0`, som bara inkluderar ett intercept, visade en förklaringsgrad (R^2) på 0. Evalueringen av modellen visade att den inte var tillräckligt komplex för att fånga upp alla variationer i priset.

En mer komplex modell, `ml_model_all`, som inkluderar alla tillgängliga prediktorer (årsmodell, körsträcka, bränsletyp, växellåda och hästkrafter), visade en förklaringsgrad på 0.843 på träningsdatan. Denna modell visade sig vara förbättrad jämfört med den enkla linjära modellen.

Genom att använda olika metoder för feature selection, såsom bakåteliminering, framåteliminering och en kombination av båda, erhölls liknande modeller som inkluderar årsmodell, körsträcka, bränsletyp, växellåda och hästkrafter som signifikanta prediktorer för priset på bilen.

De utvärderade modellerna visar liknande prestanda på tränings-, validerings- och testdatan, vilket tyder på att de är robusta och generaliserbara. Modellernas anpassade R^2 var cirka 0.839, vilket indikerar att de förklarar en betydande del av variationen i priset på bilarna. Den genomsnittliga kvadratroten av residualerna (RMSE) indikerar att modellerna har en medelavvikelse på cirka 34,523 kr från det sanna priset.

Resultaten av AIC (Akaike Information Criterion) och BIC (Bayesian Information Criterion) visar att de olika modellerna har liknande anpassningsförmåga, vilket bekräftar deras jämförbara prestanda. Sammantaget tyder resultaten på att de inkluderade prediktorerna är relevanta för att förutsäga priset på bilarna.

Genom att använda regressionsanalys har vi kunnat identifiera och kvantifiera faktorer som påverkar priset på begagnade bilar. Årsmodell, körsträcka, bränsletyp, växellåda och hästkrafter

har visat sig vara signifikanta prediktorer för priset. Dessa resultat kan vara värdefulla för att förstå bilmarknadens dynamik och för att göra mer precisa prissättningar vid försäljning av begagnade bilar.

Name	Model	AIC (weights)	AICc (weights)	BIC (weights)	R2	R2 (adj.)	RMSE	Sigma
ml_model_all	lm	5902.7 (0.250)	5903.3 (0.250)	5930.8 (0.250)	0.843	0.839	34523.250	35021.036
model_back	lm	5902.7 (0.250)	5903.3 (0.250)	5930.8 (0.250)	0.843	0.839	34523.250	35021.036
model_forward	lm	5902.7 (0.250)	5903.3 (0.250)	5930.8 (0.250)	0.843	0.839	34523.250	35021.036
model_both	lm	5902.7 (0.250)	5903.3 (0.250)	5930.8 (0.250)	0.843	0.839	34523.250	35021.036

Figure 4: Comparison of Model Performance Indices Results

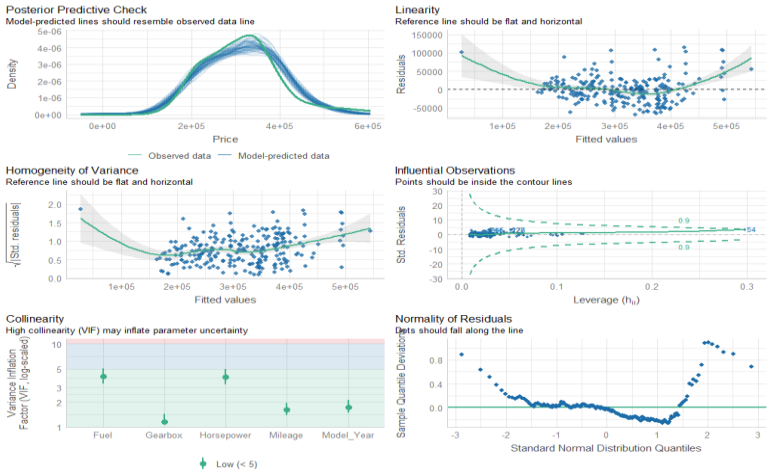

Figure 5: Diagnostic Analysis & Statistical Inference

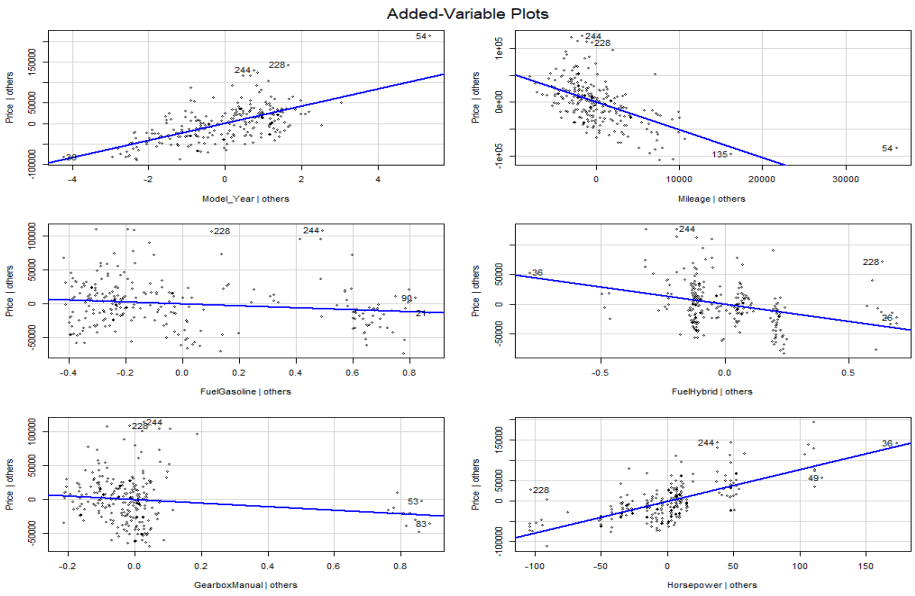

Figure 6: Visualization of the chosen mode

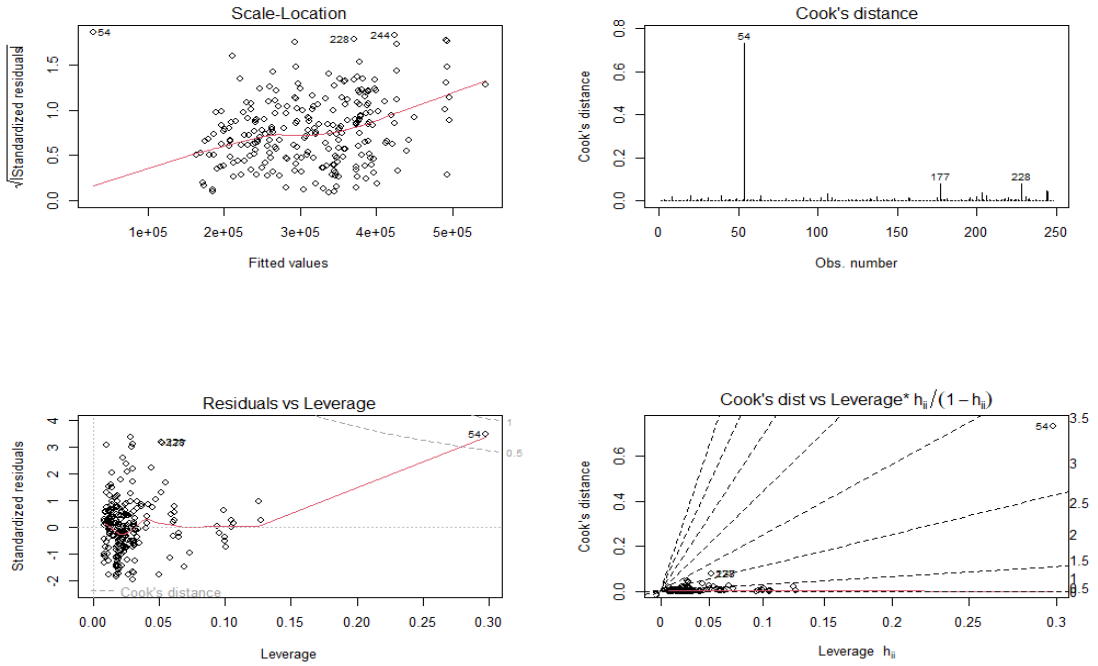

Figure 7: Other Diagnostic Analysis

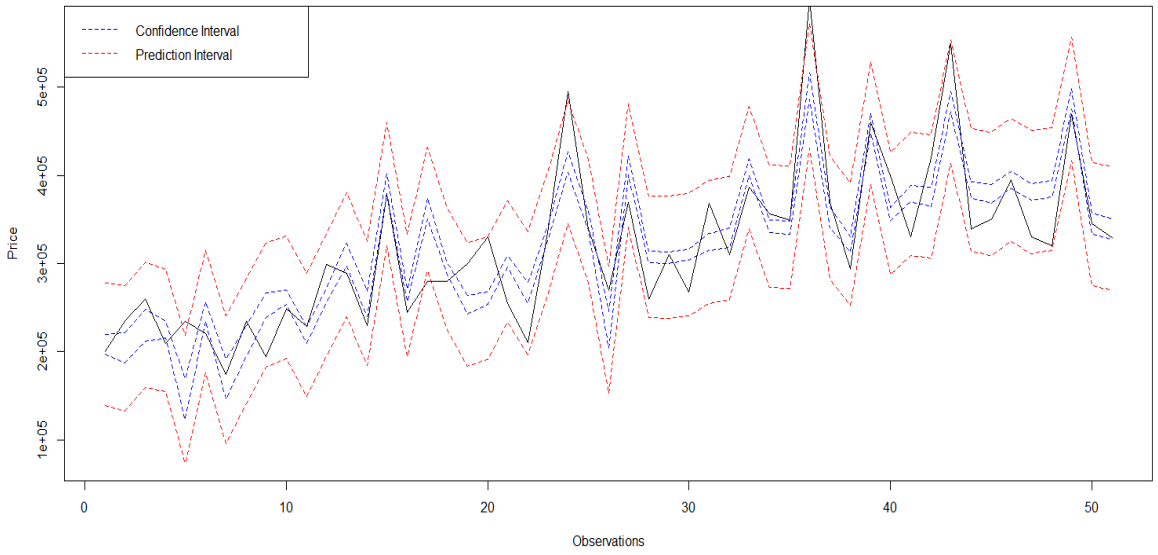

Figure 8: Confidence Interval & Prediction Interval

5 Slutsatser

Modellåret har en positiv korrelation med priset på Volvo V60-bilar. För varje ökning av modellåret med ett enhet förväntas priset öka med cirka 21,060 SEK, när andra faktorer hålls konstanta.

Körsträckan har en negativ påverkan på bilpriserna. För varje ökning av körsträckan med en enhet förväntas priset minska med cirka 5,184 SEK, när andra variabler beaktas.

Bränsletypen har en signifikant effekt på bilpriserna. Bilar med hybridbränsletyp förväntas ha lägre priser, med en minskning av cirka 58,090 SEK jämfört med bensinbilar. Dieselbilar har också lägre priser, med en minskning av cirka 14,120 SEK jämfört med bensinbilar.

Växellådstypen påverkar också bilpriserna. Manuella växellådor förväntas ha lägre priser än automatiska växellådor, med en minskning av cirka 25,660 SEK.

Hästkrafter har en positiv korrelation med bilpriserna. Varje ökning av hästkrafter med en enhet förväntas priset öka med cirka 779 SEK.

När vi kopplar dessa slutsatser till SCB-data kan vi observera att fördelningen av nyregistrerade bilar i olika bränsletyper varierar mellan regioner. Till exempel kan vi se att elhybridbilar tenderar att ha högre genomsnittliga priser än bensin- och dieselbilar. Detta stödjer våra resultat som visar att bränsletypen påverkar bilpriserna. Genom att jämföra prisfördelningen för bilar i olika län kan vi också upptäcka eventuella regionala skillnader som kan påverka våra slutsatser om bilprissättning.

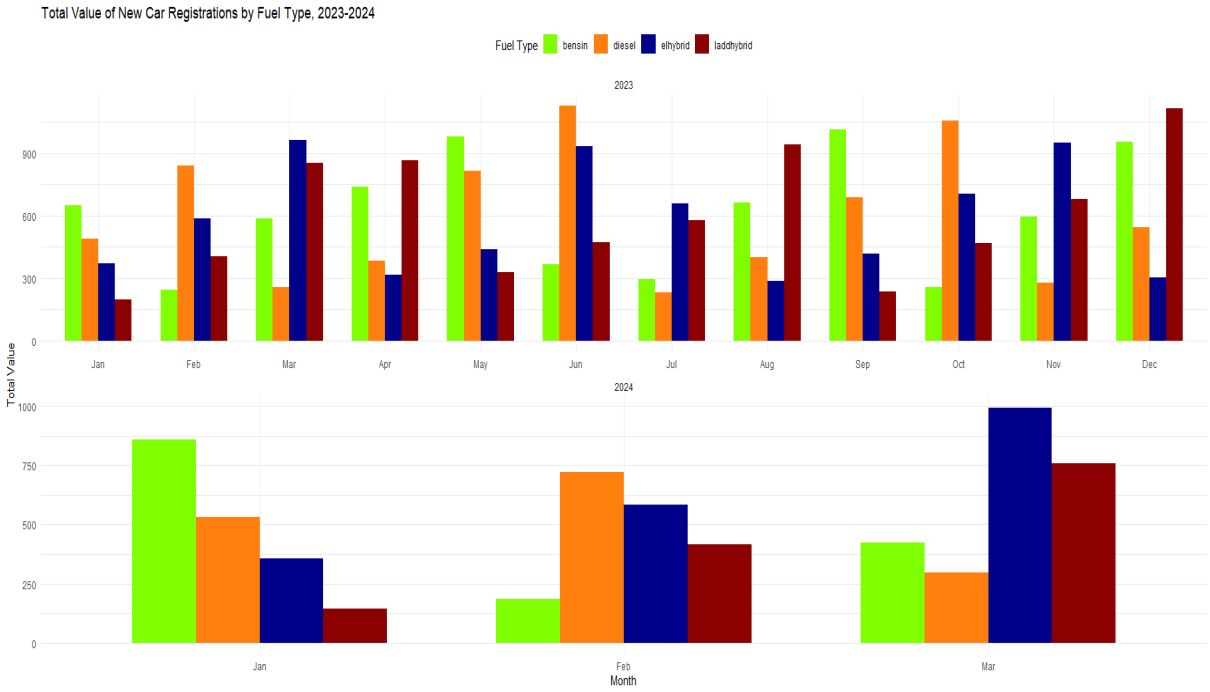

Figure 9: SCB Statistikdatabasen; Nyregistrerade personbilar efter region, drivmedel och månad

6 Teoretiska frågor

Besvara följande teoretiska 7 frågor:

6.1 Fråga 1

Beskriv kortfattat vad en Quantile-Quantile (QQ) plot är.

En Quantile-Quantile (QQ) plot är en grafisk metod för att jämföra den empiriska fördelningen av ett dataset med en teoretisk fördelning, vanligtvis en normalfördelning. På QQ-plotten placeras de kvantiler som observerats i datasetet på x-axeln och de förväntade kvantilerna för den valda teoretiska fördelningen på y-axeln. Om de punkter som representerar datasetets kvantiler ligger längs en rak linje indikerar det att datasetet följer den teoretiska fördelningen. Abweichungen från linjen indikerar avvikelser från den teoretiska fördelningen.

6.2 Fråga 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?

I maskininlärning är fokus främst på att utveckla algoritmer och modeller som kan göra prediktioner baserade på data. Dessa prediktioner kan vara av olika slag, som att förutsäga priser, klassificera bilder eller rekommendera filmer. Statistisk regressionsanalys, å andra sidan, handlar inte bara om att göra prediktioner utan också att förstå och dra slutsatser om sambandet mellan variabler i data. Till exempel kan en regressionsanalys av sambandet mellan rökning och lungcancer ge insikt om hur starkt sambandet är och om det är signifikant.

6.3 Fråga 3

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

Ett konfidensintervall är en uppskattning av osäkerheten kring en parameter i populationen, såsom medelvärdet av en variabel. Det ger en intervall av värden inom vilka vi tror att den sanna parametern ligger med en viss sannolikhet.

Ett prediktionsintervall, å andra sidan, tar inte bara hänsyn till osäkerheten kring parametern utan också osäkerheten kring de individuella observationerna. Det ger oss en intervall av värden inom vilka vi förväntar oss att en ny observation kommer att falla med en viss sannolikhet. Prediktionsintervall är bredare än konfidensintervall eftersom de också tar hänsyn till den inhemska variationen hos individuella observationer.

6.4 Fråga 4

Den multipla linjära regressionsmodellen kan skrivas som:

$$\boldsymbol{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \beta_p X_\rho + \varepsilon}$$

Hur tolkas beta parametrarna?

Beta-parametrarna i den multipla linjära regressionsmodellen representerar förändringen i responsvariabeln (Y) för varje enhetsförändring i den respektive oberoende variabeln (X), medan alla andra variabler hålls konstanta. β_0 representerar det förväntade värdet på Y när alla oberoende variabler är noll. β_1, β_2, ..., β_p representerar förändringen i Y för varje enhetsförändring i de respektive oberoende variablerna X_1, X_2, ..., X_p.

6.5 Fråga 5

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?

Nej, det stämmer inte. Även om metoder som BIC (Bayesian Information Criterion) kan användas för modellvalidering och utvärdering av modellkomplexitet, ersätter de inte behovet av träning, validering och testset. Träning, validering och testset används för att utvärdera modellens prestanda på oberoende data och för att undvika överanpassning. BIC kan hjälpa till att välja den bästa modellen baserat på komplexitet och passning till data, men det eliminerar inte behovet av att testa modellen på nya data för att bedöma dess generaliseringsförmåga.

6.6 Fråga 6

Förklara algoritmen nedan för ”Best subset selection”

Algoritmen "Best subset selection" är en metod för att välja den bästa uppsättningen prediktorer (variabler) för att bygga en modell som bäst förutsäger en responsvariabel. Här är en förklaring av algoritmen:

Steg 1: Nullmodell

Börja med en nullmodell som inte innehåller några prediktorer. Denna modell förutsäger endast medelvärdet av responsvariabeln för varje observation.

Steg 2: Utforska alla möjliga prediktoruppsättningar

För varje ***k*** (från 1 till ***p***, där ***p*** är antalet prediktorer):

- Passa alla möjliga modeller som innehåller exakt ***k*** prediktorer.
- Välj den bästa modellen bland dessa baserat på något kriterium, till exempel minsta kvadratsumma av residualerna **(RSS)** eller högsta ***R ^ 2*** - värde.

Steg 3: Välj den bästa modellen

Välj den bästa modellen från alla de valda modellerna i steg 2. Detta kan göras med hjälp av ett valideringsset eller informationskriterier som **AIC** (Akaike Information Criterion) eller **BIC** (Bayesian Information Criterion).

6.7 Fråga 7

Ett citat från statistikern George Box är: “All models are wrong, some are useful.” Förklara vad som menas med det citatet.

George Box menade att alla modeller, oavsett hur noggrant de är konstruerade, är förenklade representationer av verkligheten och kommer därför alltid att ha brister eller avvikelser. Trots detta kan vissa modeller fortfarande vara användbara för att förstå och förutsäga fenomen i verkligheten, så länge de ger någon form av insikt eller information som kan vara till nytta.

7 Självutvärdering

Jag stötte på utmaningar med att förstå och använda vissa funktioner, tolka och implementera komplexa statistiska begrepp samt sammanfatta information på ett koncist sätt. För att lösa utmaningarna med att förstå och använda funktionerna sökte jag dokumentation och bad om hjälp från kollegor. För att hantera de komplexa statistiska begreppen använde jag ytterligare litteratur och online-resurser. När det gäller att sammanfatta information arbetade jag noggrant med att förenkla och strukturera mina sammanfattningar för att göra dem mer koncisa och begripliga. Jag anser att jag förtjänar betyget G baserat på min insats och engagemang under arbetet. Trots utmaningarna har jag arbetat hårt för att förstå och tillämpa de statistiska principerna på ett korrekt sätt och har levererat kvalitativt arbete. Tack Antonio för allt stöd och vägledning under kursen! Jag uppskattar verkligen din hjälp och tålamod.

Appendix A

```
# DATA ----------------------------------------------------------------------------------------------------------------

# Load Dataset

car_data <-
read_excel("C:/Users/girli/OneDrive/Desktop/R_Exercises/R_Kunskapskontroll/Volvo_V60_Data_2018_2023/car_data.xlsx")

rmarkdown::paged_table(car_data)

glimpse(car_data)

summary(car_data)

# Encode categorical variables

car_data$Fuel <- as.factor(car_data$Fuel)

car_data$Gearbox <- as.factor(car_data$Gearbox)

dim(car_data)

str(car_data)

summary(car_data)


# EDA - Exploratory Data Analysis -------------------------------------------------------------------------------------

# Check the distribution of the target variable: "Price"

hist(car_data$Price, main = "Distribution of Target Variable (PRICE)", xlab = "Target Variable (PRICE)",
col = "chartreuse")

# Scatterplot matrix

pairs(Price ~ Model_Year + Mileage + Horsepower + Fuel + Gearbox, data = car_data)

# Create a scatter plot between "Horsepower" and "Price" by "Fuel" and "Gearbox"

ggplot(car_data, aes(x = Horsepower, y = Price, color = Fuel)) +

  geom_point(color = "darkred") +

  labs(title = "Scatter Plot: Car Price vs. Horsepower by Fuel Type", x = "Horsepower", y = "Car Price")
+

  theme_minimal() +

  geom_smooth(method = lm)

ggplot(car_data, aes(x = Horsepower, y = Price, color = Gearbox)) +

  geom_point(color = "darkviolet") +

  labs(title = "Scatter Plot: Car Price vs. Horsepower by Gearbox Type", x = "Horsepower", y = "Car
Price") +

  theme_minimal() +
```

```
  geom_smooth(method = lm)
# Create a scatter plot between "Mileage" and "Price" by "Fuel" and "Gearbox"
ggplot(car_data, aes(x = Mileage, y = Price, color = Fuel)) +
  geom_point(color = "darkmagenta") +
  labs(title = "Scatter Plot: Car Price vs. Mileage by Fuel Type", x = "Horsepower", y = "Car Price") +
  theme_minimal() +
  geom_smooth(method = lm)
ggplot(car_data, aes(x = Mileage, y = Price, color = Gearbox)) +
  geom_point(color = "darkgreen") +
  labs(title = "Scatter Plot: Car Price vs. Mileage by Gearbox Type", x = "Horsepower", y = "Car Price")
+
  theme_minimal() +
  geom_smooth(method = lm)
# Create dummy variables for categorical columns
car_data_dummy <- car_data %>%
  mutate(
    FuelHybrid = as.numeric(Fuel == "Hybrid"),
    FuelGasoline = as.numeric(Fuel == "Gasoline"),
    FuelDiesel = as.numeric(Fuel == "Diesel"),
    GearboxAutomatic = as.numeric(Gearbox == "Automatic"),
    GearboxManual = as.numeric(Gearbox == "Manual")
  )
# Split data into training, validation, and test sets
set.seed(123)
train_index <- createDataPartition(car_data$Price, p = 0.7, list = FALSE)
validation_index <- createDataPartition(car_data[-train_index, ]$Price, p = 0.5, list = FALSE)
car_data_train <- car_data[train_index, ]
car_data_validation <- car_data[-train_index, ][validation_index, ]
car_data_test <- car_data[-train_index, ][-validation_index, ]
dim(car_data_train)
dim(car_data_validation)
dim(car_data_test)
#Check the distribution of target variable from Train Dataset
```

```
par(mfrow = c(1, 3))

hist(car_data_train$Price, main = "Distribution of Target Variable (PRICE) from Train Dataset", xlab =
"Target Variable (PRICE) from Train Dataset", col = "darkred")

# Check the distribution of target variable from Validation Dataset

hist(car_data_validation$Price, main = "Distribution of Target Variable (PRICE) from Validation
Dataset", xlab = "Target Variable (PRICE) from Validation Dataset", col = "darkblue")

# Check the distribution of target variable from Test Dataset

hist(car_data_test$Price, main = "Distribution of Target Variable (PRICE) from Test Dataset", xlab =
"Target Variable (PRICE) from Test Dataset", col = "darkmagenta")

# Select only numeric and dummy variable columns

numeric_dummy_data <- select(car_data_dummy, -Fuel, -Gearbox)

# Create a correlation plot

ggcorrplot::ggcorrplot(cor(numeric_dummy_data), hc.order = TRUE, lab = TRUE)

 corr <- cor((numeric_dummy_data))

 corr


# MODELING: Train, Validate, Test ----------------------------------------------------------------------------------

# Train: Multiple Linear Regression

# Model 1:

ml_model_0 <- lm(Price ~ 1, car_data_train) # Intercept, represents the mean value of mpg for all
cars represented by the ones included in this data set.

summary(ml_model_0)

# Model 2:

ml_model_all <- lm(Price ~ ., data = car_data_train)

summary(ml_model_all)

# Feature Selection

# Backward Elimination

model_back <- step(ml_model_all,direction = "backward", trace = 0)

summary(model_back)

# Forward Selection

model_forward <- step(

  ml_model_0, direction = "forward", scope = list(lower = ml_model_0,

  upper = ml_model_all),trace = 0

)
```

```
summary(model_forward)
# Both models
model_both <- step(
  ml_model_0, direction = "both", scope = list(lower = ml_model_0, upper = ml_model_all),
  trace = 0
)
summary(model_both)
# Model Comparison (Problem01: Issue about compare_performance function)
compare_performance(ml_model_all,model_back,model_forward,model_both)
# Subset Selection & Diagnostic Analysis: Linear Regression Assumption --------
check_model(ml_model_all)
# Model Using Scaled Data
# Scaling data: Use the pipe operator
num_data <- car_data_train %>%
  select(is.numeric) %>%
  sapply(scale)
fac_data <- car_data_train %>% select(where(is.factor))
# Find the smaller number of rows between num_data and fac_data
min_rows <- min(nrow(num_data), nrow(fac_data))
# Combine the matrices
car_scale <- cbind(num_data, fac_data)
model_scale <- lm(Price ~ ., car_scale)
summary(model_scale)
check_model(model_scale)
# Model Validation
# Predictions on validation set
ml_pred_val <- predict(ml_model_all, newdata = car_data_validation)
# Calculate Mean Squared Error (MSE)
ml_rmse_val <- sqrt(mean((ml_pred_val - car_data_validation$Price)^2))
cat("Multiple Linear Regression RMSE on Validation Set:", ml_rmse_val, "\n")
#Visualize the chosen model
summary(ml_model_all)
avPlots(ml_model_all)
```

```
# Subset Selection and Diagnostic Analysis ------------------------------------------------------------------------

# Perform subset selection using regsubsets

regfit <- regsubsets(Price ~ ., data = car_data_train, nvmax = 6)  # Choose the maximum number of predictors

# Summary of subset selection results

summary(regfit)

# Plot with heatmap

par(mfrow = c(2, 2))

plot(regfit, scale = "adjr2", main = "Subset Selection Heatmap")

# Residual plot to check for homoscedasticity

plot(ml_model_all, which = 1)

# Normal Q-Q plot to check for normality of residuals

plot(ml_model_all, which = 2)

# Cook's distance plot to check for influential points

cooksd <- cooks.distance(ml_model_all)

plot(cooksd, pch = 20, main = "Cook's Distance Plot")

abline(h = 4/length(ml_model_all$coefficients), col = "red")  # Adjust based on the model

# summary(cooksd)

# Other diagnostic plots

par(mfrow = c(2, 2))

plot(ml_model_all, which = 3)

plot(ml_model_all, which = 4)

plot(ml_model_all, which = 5)

plot(ml_model_all, which = 6)

acf(resid(ml_model_all))

threshold <- 0.05

high_cooks_indices <- which(cooksd > threshold)

# Check if there are any observations with high Cook's distances

if (length(high_cooks_indices) > 0) {

  # Subset data for observations with high Cook's distances

  high_cooks_data <- data.frame(Index = high_cooks_indices, cooksd = cooksd[high_cooks_indices])

  # Plot the relationship between index and Cook's distances
```

```
  ggplot(high_cooks_data, aes(x = Index, y = cooksd)) +
    geom_point() +
    labs(title = "Relationship between Index and Cook's Distances", x = "Index", y = "Cook's Distance")
} else {
  print("No observations with high Cook's distances found.")
}
# Calculate leverage values
leverage <- hatvalues(ml_model_all)
# Plot leverage values
plot(leverage, main = "Leverage Values", xlab = "Observation Index", ylab = "Leverage")
# Identify observations with high leverage
high_leverage_indices <- which(leverage > (2 * (ncol(car_data) - 1) / nrow(high_cooks_data)))
# Check if there are any observations with high leverage
if (length(high_leverage_indices) > 0) {
  print("Observations with high leverage:")
  print(high_leverage_indices)
} else {
  print("No observations with high leverage found.")
}
# Check for duplicates and sort
high_leverage_indices <- unique(sort(high_leverage_indices))
# Display high leverage observations
print(high_leverage_indices)
# Further investigate these observations (e.g., review data, calculate influence measures)
# For example, calculate Cook's distance
cooksd <- cooks.distance(ml_model_all)
cooksd_high_leverage <- cooksd[high_leverage_indices]
# Assess influence of high leverage points
influential_threshold <- 4 / nrow(car_data)
influential_indices <- which(cooksd_high_leverage > influential_threshold)
# Display influential observations
print(influential_indices)
```

```
# MODEL INFERENCE ----------------------------------------------------------------------------------------------------

# Extract coefficients and perform hypothesis testing

coef_summary <- summary(ml_model_all)

cat("Coefficient Summary:\n")

print(coef_summary)

# Predict on the test dataset

test_predictions <- predict(ml_model_all, newdata = car_data_test)

# test_predictions

summary(test_predictions)


# Test the Chosen Model ----------------------------------------------------------------------------------------------

# Evaluate test predictions

test_rmse <- sqrt(mean((test_predictions - car_data_test$Price)^2))

cat("Test RMSE for the chosen model:", test_rmse, "\n")

# Calculate confidence interval

conf_interval <- predict(ml_model_all, newdata = car_data_test, interval = "confidence", level = 0.95)

summary(conf_interval)

# Calculate prediction interval

pred_interval <- predict(ml_model_all, newdata = car_data_test, interval = "prediction", level = 0.95)

summary(pred_interval)

# Create data frames for intervals

conf_data <- data.frame(car_data_test, lower_bound = conf_interval[, "lwr"], upper_bound =
conf_interval[, "upr"])

pred_data <- data.frame(car_data_test, lower_bound = pred_interval[, "lwr"], upper_bound =
pred_interval[, "upr"])


summary(conf_data)

summary(pred_data)

# Plot the test data with confidence and prediction intervals

par(mfrow = c(1, 1))

plot(car_data_test$Price, type = "l", ylim = c(min(pred_interval), max(pred_interval)), xlab =
"Observations", ylab = "Price")

lines(conf_data$lower_bound, col = "blue", lty = 2)
```

```
lines(conf_data$upper_bound, col = "blue", lty = 2)

lines(pred_data$lower_bound, col = "red", lty = 2)

lines(pred_data$upper_bound, col = "red", lty = 2)

legend("topleft", legend = c("Confidence Interval", "Prediction Interval"), col = c("blue", "red"), lty =
2)


# API: www.statistikdatabasen.scb.se  ----------------------------------------------------------------------------

# Define the API endpoint URL

api_url <- "https://www.statistikdatabasen.scb.se/sq/148752"

# Make a GET request to the API endpoint

response <- httr::GET(url = api_url)

# Check if the request was successful

if (response$status_code == 200) {

 # Parse the JSON data from the API response

 api_data <- httr::content(response, "text", encoding = "UTF-8")

 api_data <- jsonlite::fromJSON(api_data)

 # Extract relevant information from the parsed data

 keys <- names(api_data$dimension$Tid$category$label)

 year <- substr(keys, 1, 4)

 month <- substr(keys, 6, 7)

 value <- api_data$value

 fuel_type <- api_data$dimension$Drivmedel$category$label

 # Convert the character vector to a factor

 fuel_type <- factor(unlist(fuel_type))

 # Create a data frame

 data_df <- data.frame(year = as.integer(year),

                       month = as.integer(month),

                       fuel_type = fuel_type,

                       value = value)

 # Convert month to factor with appropriate labels

 month_labels <- c("Jan", "Feb", "Mar", "Apr", "May", "Jun", "Jul", "Aug", "Sep", "Oct", "Nov", "Dec")

 data_df$month <- factor(data_df$month, levels = 1:12, labels = month_labels)

 # Plot the data with a grouped bar plot
```

```
  ggplot(data_df, aes(x = month, y = value, fill = fuel_type)) +
    geom_bar(stat = "identity", position = "dodge", width = 0.7) +
    labs(x = "Month", y = "Total Value", fill = "Fuel Type", title = "Total Value of New Car Registrations by Fuel Type, 2023-2024") +
    scale_fill_manual(values = c("bensin" = "chartreuse", "diesel" = "#ff7f0e", "elhybrid" = "darkblue", "laddhybrid" = "darkred")) +
    facet_wrap(~year, scales = "free", ncol = 1) +
    theme_minimal() +
    theme(legend.position = "top")  # Move legend to the top
} else {
  # Handle error
  stop("Error: API request failed.")
}
summary(data_df)
str(data_df)
print(data_df)
coef_summary_api <- summary(data_df)
cat("Coefficient Summary:\n")
print(coef_summary_api)
```

Källförteckning

Philip Lloyd, N. Z. (2023). Specialist Calculus Teacher, Motivator and Baroque Trumpet Soloist.

Hämtat från https://qr.ae/psuEeY

William Irvin Lewis, (2024). Hämtat från https://www.motortrend.com/reviews/2024-volvo-v60

Malcolm Forster, Elliot Sober, (2011). Akaike Information Criterion, ScienceDirect

Hämtat från https://www.sciencedirect.com/topics/economics-econometrics-and-finance/akaike-information-criterion

Rob J Hyndman, AIC Calculations. Hämtat från https://robjhyndman.com/hyndsight/lm_aic.html

Sachin Date, India. Time Series Analysis, Regression, and Forecasting.

Hämtat från https://timeseriesreasoning.com/contents/akaike-information-criterion/

Marcus Collard, S.E. (2022). Thesis, Price Prediction for Used Cars. A Comparison of M.L. Regression

Hämtat från https://www.diva-portal.org/smash/get/diva2:1674070/FULLTEXT01.pdf

Stack Exchange, (2020). Cross Validated.

Hämtat från https://stats.stackexchange.com/questions/429526/why-is-the-intercept-in-multiple-regression-changing-when-including-excluding-re/429608#429608

Kaushik Jagini, India. (2020). Why do we need an Intercept in regression models?

Hämtat från https://medium.com/swlh/why-do-we-need-an-intercept-in-regression-models-76485a98d03c

Motor 1, (2023). Volvo V60. Hämtat från https://www.motor1.com/volvo/v60/

Shwetha Acharya, India. (2021). What are RMSE and MAE? A simple guide to evaluation metrics.

Hämtat från https://towardsdatascience.com/what-are-rmse-and-mae-e405ce230383

JSON-stat. JSON-stat Format. Hämatat från Full Spec. JSON-stat

Onyejiaku Theophilus Chidalu, What is the as.integer() function in R?

Hämtat från https://www.educative.io/answers/what-is-the-asinteger-function-in-r

SCB Statistikdatabasen, S.E. (2024). Hämtat från https://www.statistikdatabasen.scb.se/sq/148752