Regressionsmodellering på data från blocket.

Zakariyae Mokhtari

EC Utbildning

R-språk

2025–04

Abstract

This project explores car price prediction using two linear regression models. The first model includes only numerical variables such as mileage, model year, horsepower, and engine size. The second, improved model adds categorical predictors such as fuel type and transmission to better capture real-world pricing dynamics. To manage skewness and stabilize variance, a log transformation was applied to the target variable (price).

On the test set, the improved model achieved an RMSE of 57,132 SEK and an R^2 of 0.80, outperforming the basic model which reached an RMSE of 61,760 SEK and an R^2 of 0.77. Key coefficients include -0.000030 for mileage and +0.36 for manual transmission (relative to automatic), suggesting that increased mileage slightly decreases the price, while manual transmission is associated with a lower price compared to automatic.

These results, along with detailed assumption testing and diagnostics (e.g., residual analysis, multicollinearity, normality), are thoroughly discussed in the full report.

Innehållsförteckning

1 Inledning

Att förstå vilka faktorer som påverkar priset på begagnade bilar kan vara till stor nytta både för köpare, säljare och aktörer i bilbranschen. Genom att använda regressionsmodeller går det att identifiera samband mellan bilens egenskaper och dess försäljningspris, vilket kan hjälpa till att göra mer informerade beslut. I det här arbetet har två modeller tagits fram för att undersöka hur olika variabler påverkar prissättningen. Den första modellen bygger på numeriska variabler som miltal, modellår, motorstorlek och hästkrafter. Den andra modellen är en förbättrad version som även tar hänsyn till kategoriska variabler såsom bränsletyp och växellåda.

Syftet med denna rapport är att undersöka hur väl regressionsmodeller kan användas för att förutsäga försäljningspriset på bilar, och hur modellens resultat påverkas av vilka variabler som används.

För att uppfylla syftet kommer följande frågeställningar att besvaras:

1. Påverkar det modellens förmåga att förutsäga priset om kategoriska variabler inkluderas?
2. Vilken av modellerna presterar bäst enligt mått som RMSE och R^2?
3. Uppfyller modellerna de grundläggande statistiska antagandena för linjär regression?

För att bättre hantera snedfördelningen i prisvariabeln valdes att log-transformera försäljningspriset. Projektet har omfattat allt från datarensning till modellbygge, utvärdering och tolkning av resultaten med målet att visa hur statistiska metoder faktiskt kan användas i praktiken, till exempel vid värdering av begagnade bilar.

2 Teori

2.1 Regressionsmodeller

Regression är en statistisk metod som används för att analysera sambandet mellan en beroende variabel och en eller flera oberoende variabler. I detta projekt används linjär regression för att förutsäga bilars försäljningspris. Linjär regression bygger på antagandet att det finns ett linjärt samband mellan variablerna, vilket gör det möjligt att modellera och tolka hur förändringar i en variabel påverkar utfallet.

Två modeller har använts: en grundmodell med endast numeriska variabler, samt en förbättrad modell som även inkluderar kategoriska faktorer. Skillnaden mellan modellerna gör det möjligt att undersöka hur mycket mer information som tillförs genom att ta hänsyn till exempelvis bränsletyp eller växellåda.

Den beroende variabeln, försäljningspriset, har log-transformerats. Det innebär att modellen använder den naturliga logaritmen av priset – ett vanligt tillvägagångssätt när man arbetar med snedfördelade data. Log-transformation hjälper till att stabilisera variansen, minska påverkan av extremvärden och förbättra linjäriteten i modellen (Hospitality Institute, n.d.).

$$\log(\text{Pris}) = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \cdots + \beta_n \cdot X_n + \boldsymbol{\varepsilon}$$

Där log(pris) är den log-transformerade målvariabeln, X_1 till X_n representerar förklarande variabler som miltal, modellår, hästkrafter med flera, β-koefficienterna anger deras respektive påverkan på log (Pris), och ε är feltermen.

Den naturliga logaritmen används ofta i statistisk modellering, särskilt när data är snedfördelad. Den definieras som:

$$\ln(x) = \log_e(x) \quad \text{där } e \approx 2.718$$

2.2 Kategoriska variabler i regression

När man inkluderar kategoriska variabler i en regressionsmodell – som till exempel bränsletyp eller växellåda – måste dessa först omvandlas till ett numeriskt format. Det sker genom så kallad dummy-kodning, där varje kategori representeras som en separat variabel som får värdet 1 om bilen tillhör den kategorin, och 0 annars.

Exempelvis: en bil med bensinmotor får 1 för “Bränsle: Bensin” och 0 för övriga bränslekategorier. På så vis kan modellen beräkna skillnaden i pris mellan olika bränsletyper, där en kategori (oftast den vanligaste, till exempel diesel) fungerar som referens.

I detta projekt skedde dummy-kodningen automatiskt genom att relevanta kolumner markerades som faktorer i R. Det innebär att modellen själv skapade de nödvändiga binära variablerna vid modellbygget, vilket förenklade processen och minskade behovet av manuell kodning.

2.3 Antaganden vid linjär regression

När man bygger en linjär regressionsmodell finns det vissa osynliga regler som behöver följas för att resultaten ska gå att lita på. Dessa kallas statistiska antaganden, och de är viktiga för att modellen ska fungera som den ska – både när man tolkar sambanden och när man gör prediktioner.

I det här projektet har jag tagit hänsyn till fem sådana antaganden:

- **Linearitet (Linearity)**
  Antagandet att sambandet mellan de oberoende variablerna och den beroende variabeln är linjärt.

- **Normalfördelade residualer (Normally distributed residuals)**
  Feltermerna (residuals) ska vara ungefär normalfördelade. Testas med till exempel Shapiro-Wilk test och Q-Q plot.

- **Homoskedasticitet (Homoscedasticity)**
  Variansen i residualerna ska vara konstant över alla nivåer av de oberoende variablerna. Testas med till exempel Breusch-Pagan test.

- **Ingen autokorrelation (No autocorrelation)**
  Feltermerna ska vara oberoende av varandra. Testas med till exempel Durbin-Watson test.

- **Ingen multikollinaritet (No multicollinearity)**
  De oberoende variablerna ska inte vara starkt korrelerade med varandra. Mätning sker med VIF (Variance Inflation Factor).

Om dessa antaganden bryts kan tolkningen av modellen bli missvisande, även om den ser ut att prestera bra i form av R^2 eller RMSE (The Carpentries, n.d.).

2.4 Statistisk inferens i regression

Linjär regression är inte bara ett verktyg för prediktion, utan används även för att dra slutsatser om sambandet mellan variabler detta kallas statistisk inferens. Genom hypotesprövning kan vi avgöra om en viss variabel har en signifikant påverkan på utfallet (Patil, 2024).

För varje variabel testas en hypotes:

$$H_0: \beta_i = 0 \quad \text{(ingen effekt)}$$

$$H_1: \beta_i \neq 0 \quad \text{(har effekt)}$$

Om p-värdet är lågt (ofta <0.05) förkastas nollhypotesen, vilket innebär att variabeln har en signifikant påverkan. Konfidensintervall rapporteras också för varje koefficient för att visa inom vilket intervall den verkliga effekten sannolikt ligger.

Utöver detta används inferenstester även för att granska om modellens antaganden är uppfyllda, som test för normalitet (Shapiro-Wilk), konstant varians (Breusch-Pagan) och autokorrelation *(Durbin-Watson).*

2.5 Koefficienterna i en regressionsmodell

Koefficienterna visar **hur mycket utfallsvariabeln förändras när en viss förklarande variabel ökar med 1 enhet**, *om allt annat hålls konstant*.

Exempel:

$$Pris = \beta 0 + \beta 1 \cdot Miltal$$

Om $\beta 1 = -20$, betyder det att priset i genomsnitt minskar med 20 kr för varje extra mil bilen gått.

- β0 är interceptet modellens skattning av priset när alla andra variabler är 0.
- Varje annan β anger den specifika effekten av sin tillhörande variabel.

Koefficienterna är alltså hjärtat i modellen de kvantifierar sambandet mellan variablerna och gör det möjligt att tolka och förklara modellens resultat. Enligt OpenStax (2023) ger koefficienterna i en linjär regressionsmodell en uppskattning av effekten som en enhets förändring i en oberoende variabel har på den beroende variabeln, givet att övriga variabler hålls konstanta *(OpenStax).*

3 Metod

Det här arbetet bygger på ett verkligt dataset med annonser för begagnade Volvobilar, hämtat från Blocket.se och sammanställt i en Excelfil. All analys har genomförts i R, där jag följde en systematisk process med datarensning, visualisering och modellering för att skapa och jämföra två olika regressionsmodeller.

Datasamlingen gjordes av en grupp på 17 personer där varje deltagare ansvarade för en region i Sverige. Målet var att samla in cirka 50 annonser var, vilket gav oss ett omfattande underlag med flera hundra bilar. Informationen samlades i en gemensam Excel-fil som vi delade via Microsoft Teams.

Vissa deltagare samlade in data helt manuellt, medan andra använde mer automatiserade metoder. Själv använde jag ett halvautomatiserat tillvägagångssätt med hjälp av web scraping-verktyget Hunderbit AI Web Scraper & Web Automation Agent, ett Chrome-tillägg som gjorde det lätt att plocka ut information som pris, årsmodell, miltal, motorstorlek, växellåda och bränsletyp direkt från webbsidan.

När vissa annonser hade bristfällig information, bytte jag ut dem manuellt för att säkerställa att datat var komplett inför analysen. I vissa fall hämtades kompletterande uppgifter även från externa och trovärdiga källor som Transportstyrelsen och Car.info.

3.1 Databehandling

Första steget var att städa upp datasetet, det innehöll en del oönskade tecken som "kr" och "mil" i numeriska kolumner. Dessa togs bort så att värdena kunde omvandlas till siffror och användas i analysen. Variablerna Bränsle och Växellåda kodades som faktorer för att markera dem som kategoriska. När modellen byggdes tog R hand om resten och skapade automatiskt dummyvariabler så att dessa kunde användas i regressionen på rätt sätt.

Jag tog även bort kolumner som var tomma eller bara fanns där av tekniska skäl, för att göra datasetet mer lättjobbat och fokuserat. Innan datan delades upp gjorde jag en första utforskande dataanalys (EDA) för att få en överblick över fördelningar, samband och eventuella outliers. Detta hjälpte mig att bättre förstå strukturen i datan och planera modelleringen.

Slutligen delades datan upp i ett tränings- och testset (80/20), vilket gjorde det möjligt att bygga modellen på ena delen och testa hur bra den faktiskt funkar på nya, osedda data.

3.2 Visualisering och EDA

För att få en översiktlig förståelse för datan genomfördes Exploratory Data Analysis (EDA) med hjälp av grafer skapade med `ggplot2`. Bland de viktigaste plottarna fanns:

Histogram över försäljningspris, se figure 1 visade att prisfördelningen var högt skev åt höger, vilket motiverade användning av log-transformation.

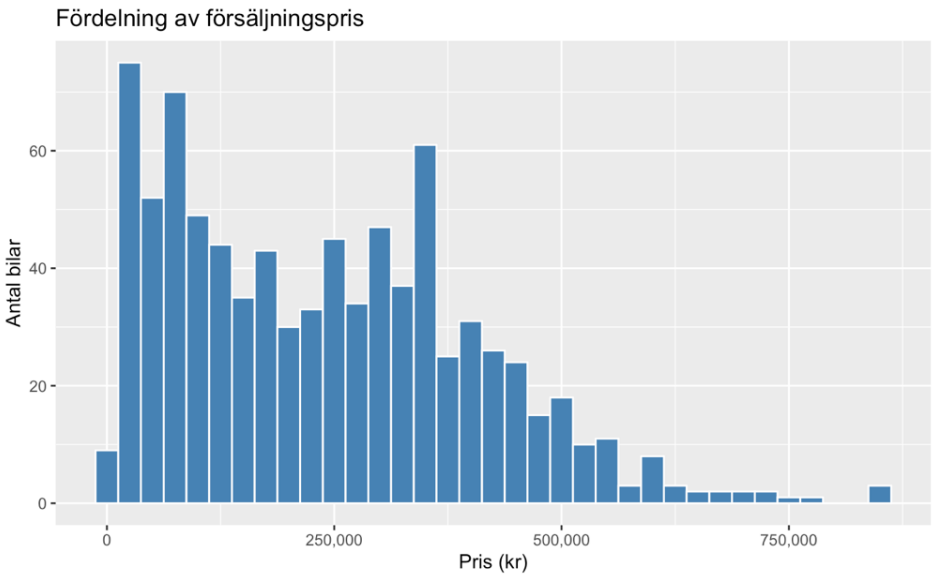
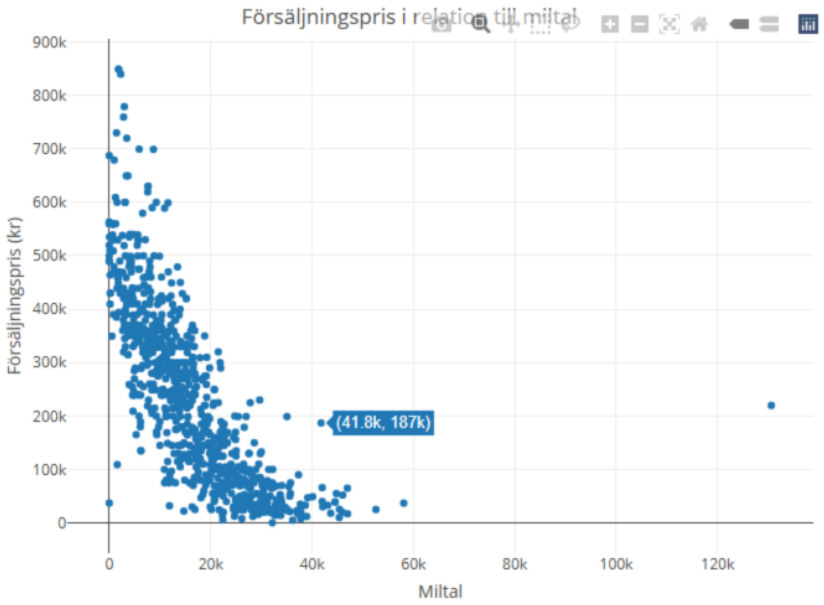


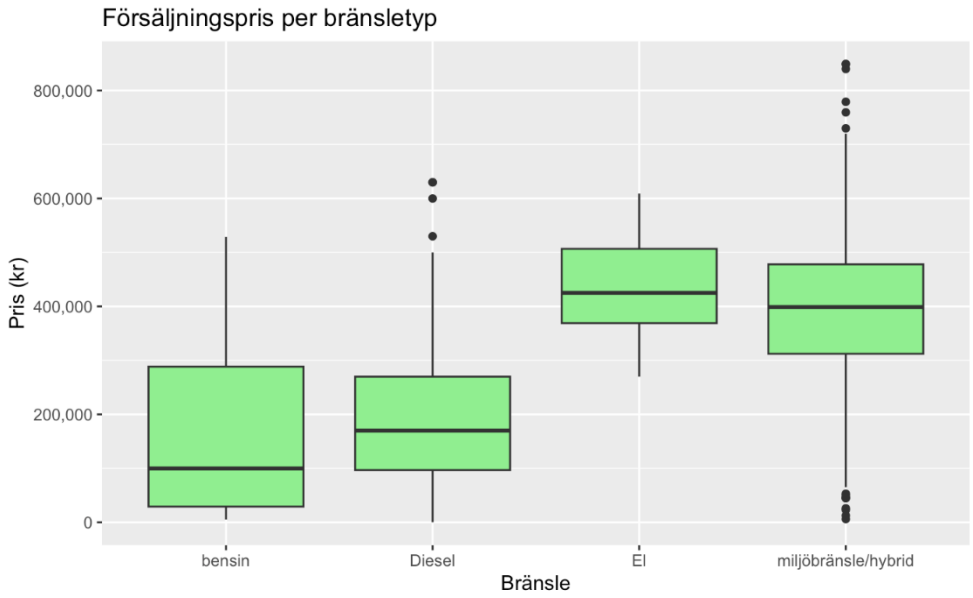
Figure 1. Histogrammet visar att försäljningspriserna är snedfördelade med en majoritet av bilarna i de lägre prisklasserna. Detta motiverade användning av log-transformation för att få en mer symmetrisk målvariabel.

Scatterplott mellan pris och miltal (figur 2) visade ett tydligt negativt samband, vilket också var väntat: bilar med högre miltal tenderar att kosta mindre. För att förbättra läsbarheten valdes en interaktiv visualisering med `plotly`, där användaren kan hovra över punkterna för att utforska varje observation mer i detalj. Detta gör det enklare att upptäcka mönster och outliers i data.



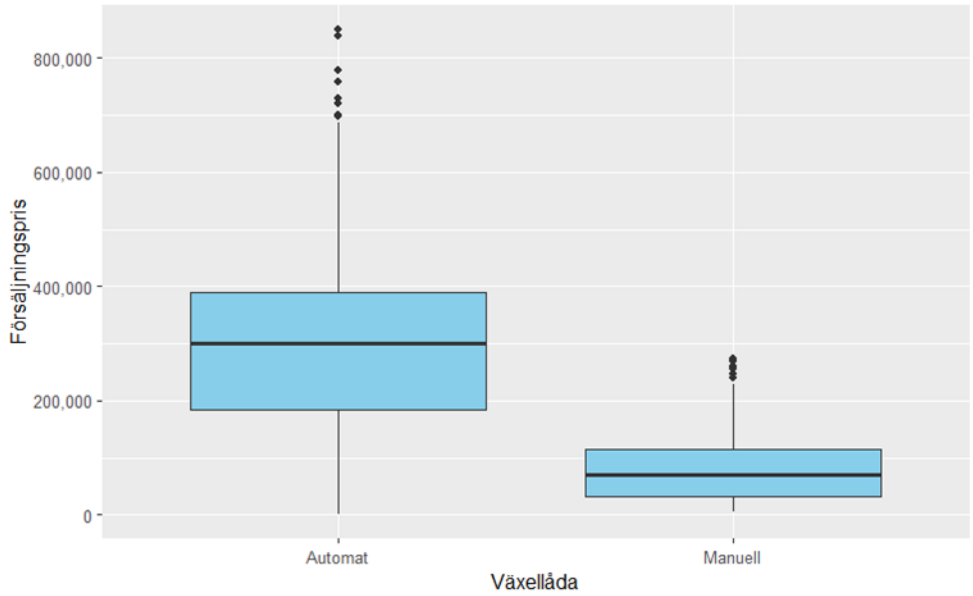
Figur 2. Interaktiv scatterplot skapad med plotly, som visar ett tydligt negativt samband mellan miltal och försäljningspris. Punkten kan hovras för att se detaljer om enskilda bilar.

Boxplots för bränsletyp och växellåda (Figur 3) visade tydliga prisskillnader mellan grupper. Till exempel låg priset för elbilar och hybridbilar generellt högre än för bensin och dieselbilar. Automatväxlade bilar hade i genomsnitt högre pris än manuella.



Figur 3. Försäljningspris för olika bränsletyper. Boxploten visar att elbilar och hybridbilar generellt har ett högre pris än bensin- och dieselbilar. Medianpriset för elbilar är tydligt högre, medan dieselbilar tenderar att ligga lägre i pris jämfört med övriga kategorier. Spridningen är störst bland hybridbilar, vilket kan tyda på större variation i modellutbud och utrustning.

Boxploten nedan (Figur 4) visar att bilar med automatisk växellåda i genomsnitt har betydligt högre försäljningspris än manuellt växlade bilar. Medianen är klart högre för automatbilar, och spridningen i pris är större (svansar), vilket kan bero på att automatväxlar oftare förekommer i nyare och mer utrustade bilmodeller.



Figur 4. Försäljningspris per växellådstyp. Automatbilar har högre medianpris och större variation än manuellt växlade fordon.

3.3 Modell 1 numeriska variabler

Den första modellen byggdes med enbart numeriska oberoende variabler: Miltal, Modellår, Hästkrafter och Motorstorlek. Eftersom försäljningspriserna var kraftigt snedfördelade log-transformerades målvariabeln innan modellen tränades, för att uppnå en mer normalfördelad residual.

Modellens prestanda utvärderades med följande nyckelmått:

- RMSE (Root Mean Square Error): visar genomsnittligt kvadratiskt fel i kronor.
- MAE (Mean Absolute Error): visar genomsnittligt absolut fel.
- R^2: anger hur stor andel av variationen i det log-transformerade priset som modellen kan förklara.

För att kontrollera modellens tillförlitlighet genomfördes ett antal diagnostiska tester:

- Shapiro-Wilk för att undersöka normalitet i residualerna,
- Breusch-Pagan för att testa homogen varians (homoskedasticitet),
- Durbin-Watson för eventuell autokorrelation,
- samt VIF (Variance Inflation Factor) för att kontrollera multikollinearitet mellan prediktorer.

3.4 Modell 2 med kategoriska variabler

I den andra modellen lade jag till två kategoriska variabler: *Bränsletyp* och *Växellåda*. De påverkar ofta bilens pris, så tanken var att modellen skulle bli bättre när de togs med. Eftersom variablerna redan var sparade som faktorer, skapade R automatiskt de så kallade dummyvariablerna som behövs för att använda dem i en regressionsmodell.

Precis som i modell 1 använde jag en log-transformering av försäljningspriset för att jämna ut fördelningen.

Innan jag kunde använda modellen för att förutsäga priser i testdatan, behövde jag se till att alla faktornivåer (t.ex. bränsletyper) fanns med även där. Om någon nivå bara fanns i testdatan men inte i träningsdatan, togs den observationen bort – annars skulle modellen krascha vid prediktion.

Koden nedan figur 5 visar hur modell 2 byggdes med både numeriska och kategoriska variabler. Innan prediktion justerades faktornivåerna i testdatan, och en filtrering genomfördes för att undvika prediktionsfel. Resultaten utvärderades sedan med RMSE, MAE och R^2

```r
options(scipen = 999)

model2 <- lm(logPris ~ Miltal + Modellår + Hästkrafter + Motorstorlek + Bränsle + Växellåda, data = train_data)
summary(model2)
options(scipen = 999)
# Säkerställ att faktornivåerna i test_data matchar train_data
test_data$Bränsle <- factor(test_data$Bränsle, levels = levels(train_data$Bränsle))
test_data$Växellåda <- factor(test_data$Växellåda, levels = levels(train_data$Växellåda))
# Filtrera bort test-observationer med faktornivåer som inte finns i träningen
test_data <- test_data %>%
  filter(!is.na(Bränsle), !is.na(Växellåda)) %>%
  filter(Bränsle %in% levels(train_data$Bränsle), Växellåda %in% levels(train_data$Växellåda))
# Prediktion och back-transformering
pred_log2 <- predict(model2, newdata = test_data)
pred2 <- exp(pred_log2)
pred2[pred2 < 0] <- 0
# Modellutvärdering

postResample(pred2, test_data$Försäljningspris)
```

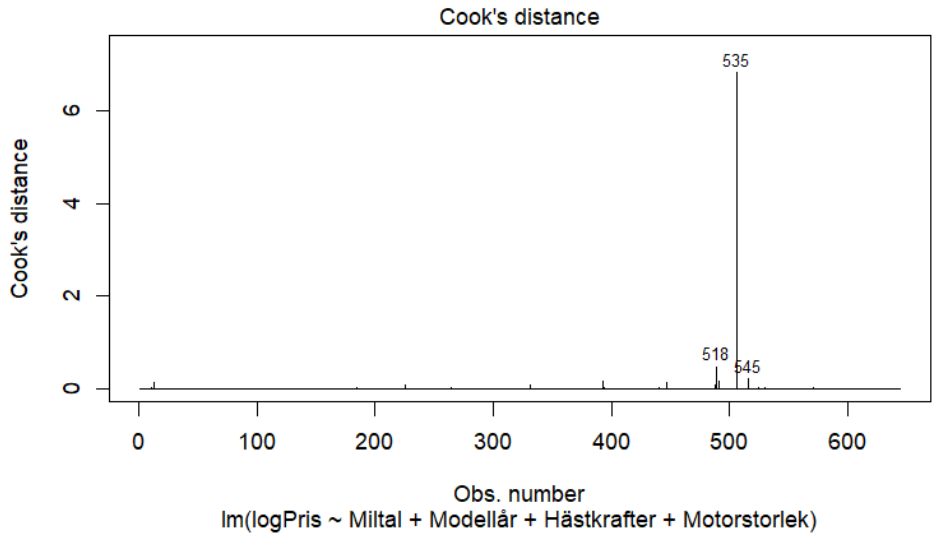
Figur 5. Kod för Modell 2 med både numeriska och kategoriska variabler. Faktornivåer anpassas före prediktion, och modellens prestanda utvärderas med RMSE, MAE och R^2.

3.5 Utökad diagnostik – Cook's Distance

Efter att modell 2 tränats ville jag ta reda på om vissa enskilda observationer hade oproportionerligt stort inflytande på modellens resultat. Det gjorde jag med hjälp av Cook's Distance (Figur 6), ett mått som visar vilka datapunkter som påverkar modellen mest. En tumregel är att observationer med ett värde över $\frac{4}{n}$ kan behöva granskas närmare.

Några observationer stack ut med höga värden, men eftersom de representerade realistiska bilar till exempel med låg körsträcka och hög effekt valde jag att behålla dem. Jag bedömde att de tillförde värdefull variation snarare än att snedvrida modellen.

För en lättförståelig genomgång av Cook's Distance, se Statorials (2021) på YouTube.



Figur 6. Cook's Distance för samtliga observationer i modellen. Enstaka punkter, som observation 535, har stort inflytande och kan påverka modellens resultat.

4 Resultat och Diskussion

Två regressionsmodeller har tränats och utvärderats för att förutsäga försäljningspriset på begagnade Volvobilar. Båda använde ett log-transformerat pris som målvariabel, men skiljde sig åt i vilka prediktorer som inkluderades. Log-transformeringen valdes eftersom prisfördelningen var tydligt högerskev, vilket bryter mot antagandet om normalfördelade residualer. Genom att logga priserna minskade effekten av extremvärden, variansen blev mer stabil och modellen kunde anpassas bättre – vilket förbättrade både tolkbarhet och prediktion.

I en log-linjär modell innebär varje koefficient att priset förändras med en viss procent snarare än ett fast belopp, vilket ger en mer realistisk bild av hur olika faktorer påverkar priset.

4.1 Modellprestanda

Modell 1 använde enbart numeriska variabler: miltal, modellår, hästkrafter och motorstorlek. Resultatet visade att miltal hade en tydligt negativ påverkan på priset (β = -0,000029992, p <0,001), medan modellår och hästkrafter visade positiv påverkan. Trots detta hade modellen en högre prediktionsavvikelse.

För att tolka detta mer konkret:

β (beta-koefficient: miltal):

- Det här är värdet på lutningen i regressionen för just *Miltal*.
- β=-0.000029992 betyder att för varje extra körd mil förväntas priset (log-transformerat) minska med 0.0045 % alltså väldigt lite per mil, men det blir mycket när miltalet är högt.

p < 0.001:

- Det är ett p-värde från hypotesprövningen.
- Det visar hur statistiskt säker effekten är.
- Ett så lågt p-värde (<0,001) betyder att det är extremt osannolikt att resultatet beror på slumpen så vi kan säga att miltal har en signifikant negativ effekt på priset.

Även om flera av variablerna Figur 7. i modellen visade sig vara statistiskt signifikanta som till exempel *Modellår* och *Hästkrafter* valde jag att fokusera lite extra på miltal. Det är en variabel som folk direkt kan relatera till. Man vet att en bil med många mil på nacken oftast är billigare, och det handlar inte bara om siffror utan om slitage, ålder och förväntad livslängd.

Det var också den variabel som tydligast gick åt det håll man intuitivt förväntar sig, och sambandet var dessutom mycket starkt och statistiskt säkert (p <0,001). Därför kändes det naturligt att använda just miltal för att ge ett konkret exempel på hur modellen fungerar och hur resultaten kan tolkas i praktiken.

```
Coefficients:
                Estimate     Std. Error t value             Pr(>|t|)
(Intercept)  -135.556302012   7.240406572 -18.722 < 0.0000000000000002 ***
Miltal         -0.000029992   0.000002318 -12.938 < 0.0000000000000002 ***
Modellår        0.073124960   0.003578115  20.437 < 0.0000000000000002 ***
Hästkrafter     0.002099220   0.000276380   7.595    0.000000000000109 ***
Motorstorlek    0.000113767   0.000060420   1.883               0.0602 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figur 7. Regressionsresultat för Modell 1. Miltal har en tydlig negativ och signifikant effekt på priset.

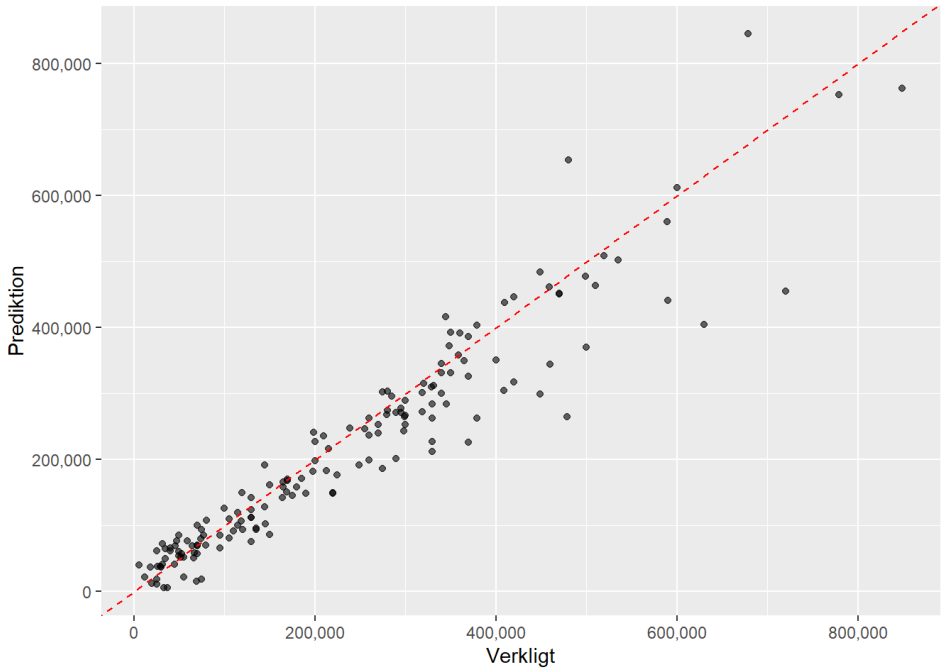
Effekten av miltal på försäljningspriset visade sig vara både tydlig och statistiskt säker. Den skattade koefficienten (β = -0,000029992) hade ett mycket litet standardfel (SE = 0,000002318), vilket betyder att modellen är ganska säker på just den uppskattningen. Ju mindre standardfelet är, desto mer kan man lita på värdet.

Dessutom var t-värdet för miltal -12.94, vilket är väldigt långt från noll. Det visar att sambandet mellan miltal och pris är starkt i alla fall i statistisk mening. Med andra ord det finns mycket starka bevis för att fler körda mil faktiskt sänker priset, och det är inte något som bara råkade hända av slumpen.

I **Modell 2** lade jag till två nya variabler: Bränsletyp och Växellåda. De har ofta stor betydelse i verkligheten till exempel är automatväxlade bilar generellt dyrare än manuella. Och ja, modellen blev faktiskt bättre av det. RMSE, som mäter hur mycket modellen i genomsnitt gissar fel, minskade från 61 760 kr till 57 132 kr. Även MAE, som visar genomsnittligt fel utan att kvadrera det, gick ner från 41 610 kr till 36 671 kr (se Tabell1).

Modellen fick också ett högre R^2-värde från 0.88 till 0.90. Det betyder att modellen nu förklarar 90 % av variationen i priserna, vilket är riktigt bra. Men som alltid med R^2 gäller det att inte stirra sig blind ibland kan ett högt värde bero på att modellen överanpassat sig till just den data den tränades på.

För att få en känsla för hur bra Modell 2 faktiskt gissar priser, jämförde jag de verkliga försäljningspriserna med de som modellen räknade fram. Som man kan se i Figur 8 ligger många punkter nära den röda linjen, vilket visar att modellen ofta är rätt ute.



Figur 8. Samband mellan verkliga och förutsagda priser i Modell 2. Den röda linjen visar perfekt förutsägelse. Punkter nära linjen tyder på hög träffsäkerhet.

Modell	RMSE(Kr)	R^2	MAE (kr)
Modell 1 (endast numeriska)	61 760	1	41 610
Modell 2 (med faktorer)	57 132	1	36 671

Tabell 1. Sammanställning av modellprestanda på testdata

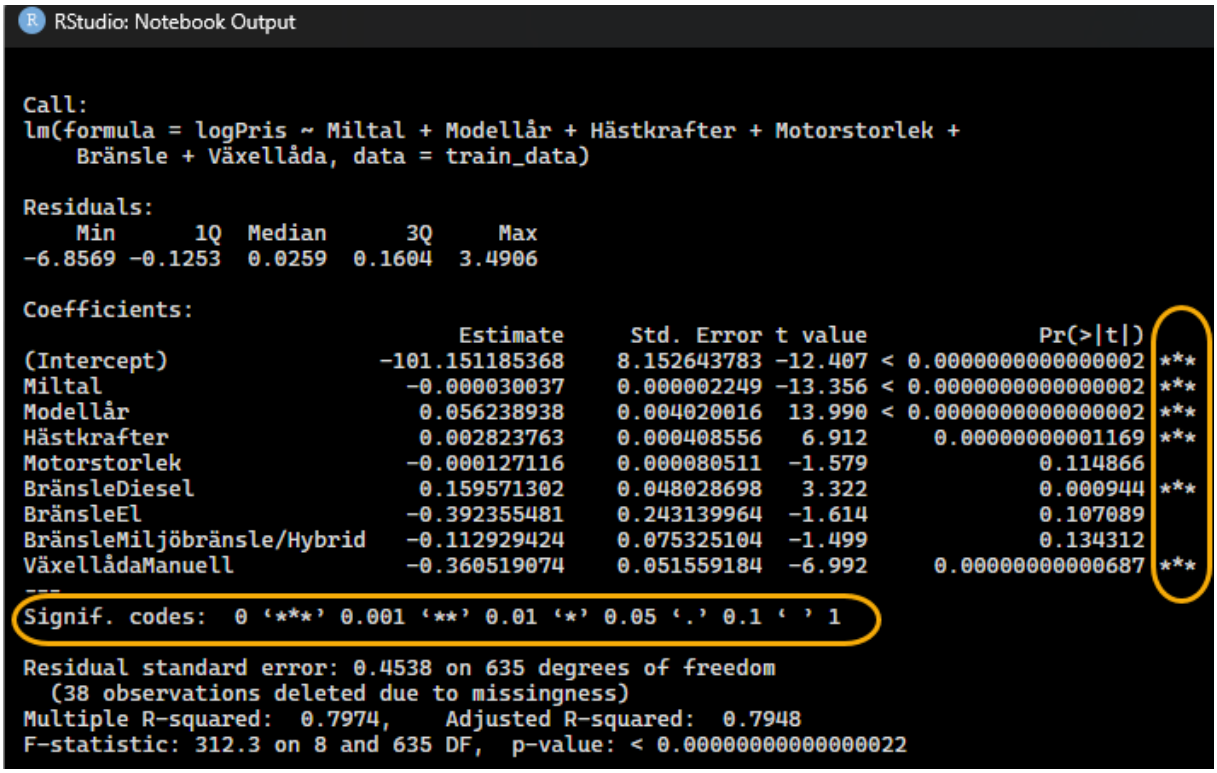
4.2 Statistisk inferens och residualdiagnostik

Statistisk inferens

Statistisk inferens visade att flera av modellens variabler har en signifikant effekt på försäljningspriset. Bland annat:

- **Miltal:** Negativ effekt, $p < 0.001$
- **Modellår:** Positiv effekt, $p < 0.001$
- **Växellåda (Manuell):** Negativ effekt, $\beta = -0{,}361$, $p < 0{,}001$
- **Bränsle (Diesel):** Positiv effekt, $p \approx 0{,}009$
- **Bränsle (El & Hybrid):** Inte signifikanta ($p > 0.1$)

De låga p-värdena för miltal, modellår, växellåda och dieselbränsle innebär att vi med hög statistisk säkerhet kan säga att dessa faktorer påverkar priset. El och hybridbränsle visade däremot ingen signifikant effekt i denna modell, vilket kan bero på att de förekommer i färre annonser eller samvarierar med andra faktorer som redan förklarar variationen i priset.



Figur 9. Regressionsresultat för Modell 2. Miltal, modellår, hästkrafter, växellåda och dieselbränsle har signifikant effekt på priset. El och hybridbränsle visar ingen signifikant påverkan i denna modell.

4.3 Residualdiagnostik

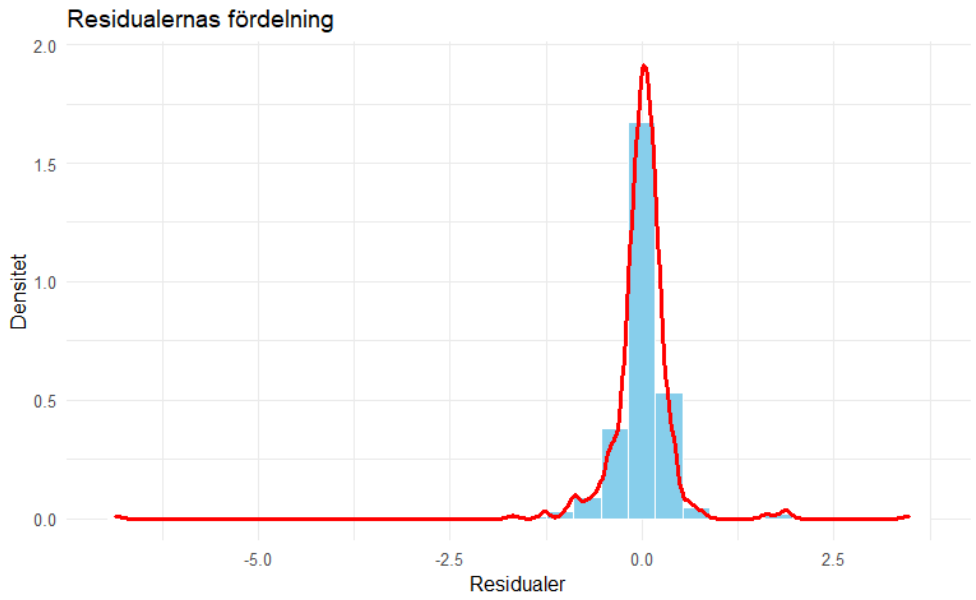
För att kolla hur bra modellen verkligen funkar gjordes några tester på residualerna – alltså skillnaden mellan modellens gissningar och de verkliga priserna.

- **Shapiro-Wilk-testet** visade att residualerna inte riktigt följer en normalfördelning. Det är inte helt oväntat i verkliga data, och bekräftades också i Q-Q-diagrammet.
- **Breusch-Pagan-testet** visade att variansen i residualerna inte var jämn – med andra ord, modellen gissar lite olika bra beroende på vilket pris det handlar om.
- **Durbin-Watson-testet** däremot visade att det inte fanns någon systematisk upprepning i felen (ingen autokorrelation), vilket är bra.

Även om allt inte stämde perfekt med antagandena, fanns det inga allvarliga avvikelser som gjorde modellen opålitlig. Modellen verkar ändå fungera stabilt.

Jag kollade också efter multikollinaritet – alltså om några av variablerna i modellen överlappar varandra för mycket. Alla VIF-värden låg klart under gränsen på 5, vilket betyder att det inte finns någon sådan risk i det här fallet.

Histogrammet se Figur 10. över residualerna visar hur felen i modellen är fördelade. Även om fördelningen inte är helt symmetrisk och normal, ligger de flesta värden nära noll, vilket är ett gott tecken på att modellen ändå presterar stabilt.



Figur 10. Histogram över residualerna. De flesta ligger nära noll, men fördelningen är inte helt normal.

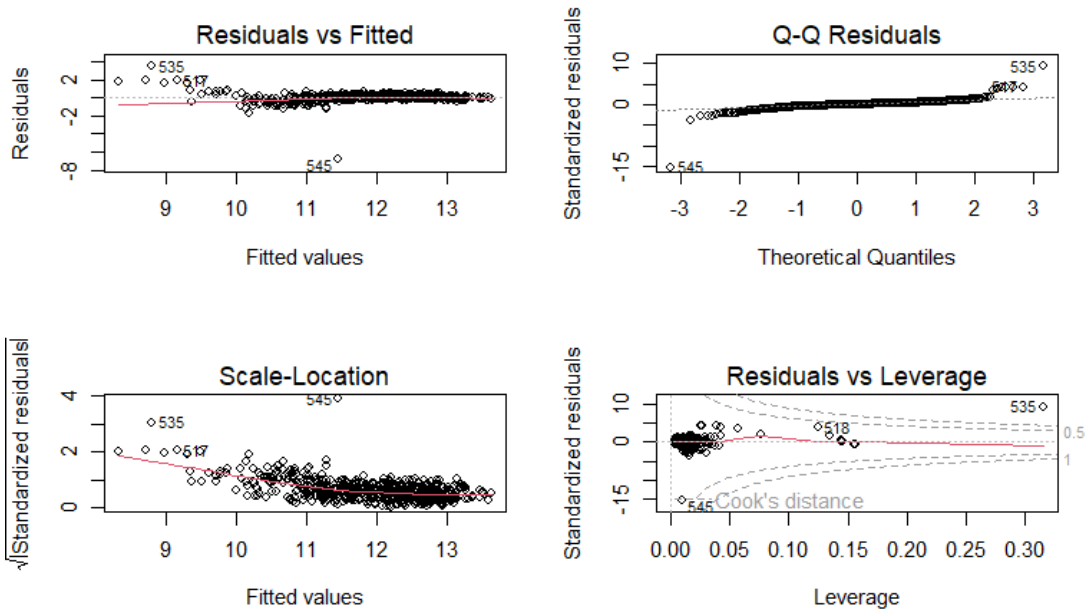
För att få en bättre bild av hur väl Modell 2 följer antagandena i linjär regression användes fyra vanliga diagnostikplottar (Figur 11). De visar:

- hur residualerna fördelar sig i förhållande till de förutsagda värdena,
- om residualerna verkar normalfördelade (Q-Q-plot),
- om variationen i residualerna är konstant (Scale-Location),
- och om det finns observationer som sticker ut mycket och påverkar modellen mer än andra (Residuals vs Leverage).

I figuren syns några avvikande punkter, främst i de extrema observationerna. Men överlag finns inga tydliga mönster eller allvarliga brott mot antagandena som kräver åtgärd. Modellen verkar därmed vara tillräckligt robust för att kunna tolkas och användas.

För att fördjupa min förståelse använde jag en extern guide som tydligt förklarar betydelsen av residualanalys. En formulering därifrån sammanfattar det på ett bra sätt:

"Checking residuals is a way to discover new insights in your model and data!" (University of Virginia Library, 2015)



Figur 11. Diagnostikplottar för Modell 2. Används för att visuellt granska linjäritet, variansstabilitet, normalfördelning och inflytelserika punkter.

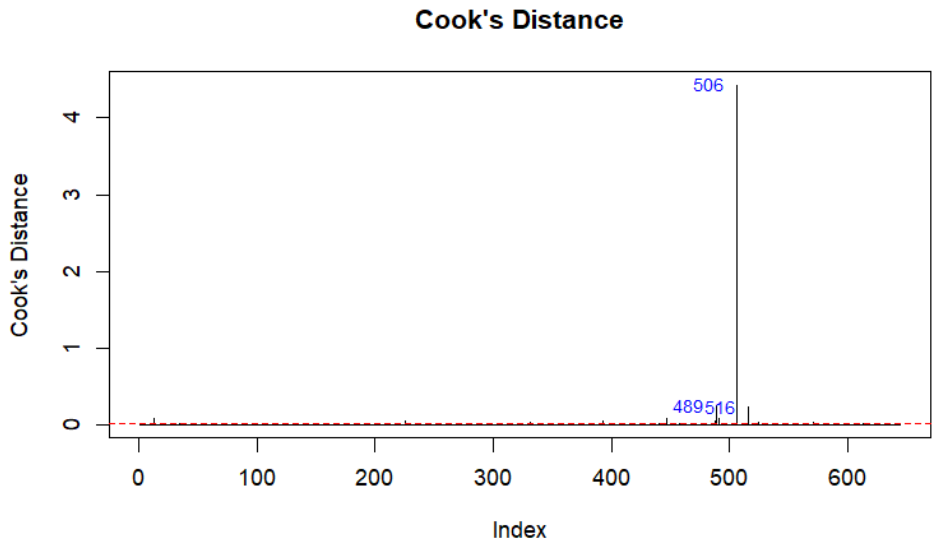
4.4 Multikollinaritet (VIF)

Samtliga prediktorer låg under den kritiska gränsen 5. För kategoriska variabler justerades GVIF med faktorns frihetsgrader, vilket ger:

Variabel	$GVIF^{(1/(2*Df))}$
Miltal	1.36
Modellår	1.67
Hästkrafter	1.88
Motorstorlek	1.57
Bränsle	1.32
Växellåda	1.30

Tabell 2. Justerade GVIF-värden (GVIF^(1/(2×Df))) för modellens prediktorer. Alla värden ligger långt under gränsen 5, vilket tyder på låg multikollinaritet.

Slutligen användes **Cook's Distance** för att identifiera observationer som har ovanligt stort inflytande på modellens resultat. Figur 11 visar att tre observationer – nummer **506**, **489** och **516** – stack ut tydligt genom att ligga över den kritiska gränsen.



Figur 12. Cook's Distance för Modell 2. De tre observationerna med störst inflytande är markerade.

När jag granskade dessa bilar (se Figur 13) mer i detalj visade det sig att de representerar bilar med ovanliga kombinationer, som mycket högt pris trots låg körsträcka eller ovanligt hög effekt. Eftersom sådana bilar faktiskt kan förekomma på marknaden som specialutrustade eller nya modeller valde jag att behålla dem i modellen. De ansågs alltså inte vara felaktiga, bara lite ovanliga.

```r
# Visa de tre mest inflytelserika observationerna
data[c(506, 489, 516), ]
```

A tibble: 3 × 15

Försäljningspris <dbl>	Säljare <chr>	Bränsle <fctr>	Växellåda <fctr>	Miltal <dbl>	Modellår <dbl>	Biltyp <chr>
687400	Företag	Miljöbränsle/Hybrid	Automat	0	2025	SUV
348800	Företag	Miljöbränsle/Hybrid	Automat	5674	2020	Kombi
534900	Företag	El	Automat	43	2025	SUV

3 rows | 1-7 of 15 columns

A tibble: 3 × 15

Drivning <chr>	Hästkrafter <dbl>	Färg <chr>	Motorstorlek <dbl>	Datum_i_trafik <S3: POSIXct>	Märke <chr>	Modell <chr>	Region <chr>
Fyrhjulsdriven	355	Röd	1969	<NA>	Volvo	XC60	Värmland
Fyrhjulsdriven	304	Mörkblå	1969	2019-09-06	Volvo	V60	Värmland
Tvåhjulsdriven	256	Svart	0	2024-06-04	Volvo	EC40	Värmland

3 rows | 8-15 of 15 columns

Figur 13. Cook's Distance för Modell 2. De tre observationerna med högst inflytande på modellen (489, 506 och 516) är markerade.

5 Slutsatser

Syftet med denna rapport var att undersöka hur väl regressionsmodeller kan användas för att förutsäga försäljningspriset på begagnade Volvobilar, och hur valet av prediktorer påverkar modellens prestanda.

Resultaten visar tydligt att modellen blir mer träffsäker när kategoriska variabler som bränsletyp och växellåda inkluderas. Modell 2, som innehåller både numeriska och kategoriska variabler, presterade bättre än Modell 1 enligt både RMSE och MAE. Det innebär att modellen gav mer tillförlitliga prisestimat.

Utifrån analysen kan de tre frågeställningarna nu besvaras:

1. **Påverkar det modellens förmåga att förutsäga priset om kategoriska variabler inkluderas?**
   Ja, träffsäkerheten ökade när dessa variabler lades till, vilket visar att de har betydelse för bilens värde.

2. **Vilken av modellerna presterar bäst enligt mått som RMSE och R^2?**
   Modell 2 presterade bäst. Den hade både lägre RMSE (57 132 kr) och MAE (36 671 kr) jämfört med Modell 1.

3. **Uppfyller modellerna de statistiska antagandena för linjär regression?**
   Ja, för det mesta. Modellen höll sig inom ramarna för vad som brukar krävas inga allvarliga avvikelser upptäcktes. Det fanns vissa tecken på att residualerna inte var helt normalfördelade och att variationen i felen inte var helt jämn (heteroskedasticitet), men inget som direkt gör modellen opålitlig. Inga tecken på autokorrelation hittades, och VIF-värdena visade att variablerna inte överlappar för mycket, vilket är ett gott tecken.

Sammanfattningsvis visar det här arbetet att linjär regression, särskilt när man väljer sina variabler noggrant och använder log-transformering, kan ge bra förståelse för hur bilpriser sätts. Samtidigt visar analysen av residualerna att det finns plats för förbättring till exempel genom att lägga till fler variabler eller testa mer avancerade modeller.

5.1 Praktisk tillämpning

För att visa hur modellen fungerar i praktiken förutsågs priset på en bil med följande specifikationer: (se Figur 14)
Modellen beräknade ett pris på **158 933 kr**, medan det faktiska försäljningspriset var **179 900 kr**. Avvikelsen var relativt liten och visar att modellen är användbar även i verkliga scenarier.

```
# prediktion av en ny bil:
```{r}
ny_bil ← data.frame(
  Miltal = 6125,
  Modellår = 2018,
  Hästkrafter = 153,
  Motorstorlek = 1969,
  Bränsle = factor("Bensin", levels = levels(train_data$Bränsle)),
  Växellåda = factor("Manuell", levels = levels(train_data$Växellåda))
)

log_pred ← predict(model2, newdata = ny_bil)
pris_pred ← exp(log_pred)
pris_pred
# Det verkliga priest är 179 900 kr
```

       1
158932.9
```

Figur 14. Prediktion av pris för en ny bil med hjälp av Modell 2, som inkluderar både numeriska och kategoriska variabler

Det här projektet har inte bara visat hur regression fungerar i teorin, utan också hur kraftfullt det kan vara i praktiken och hur mycket insikt man faktiskt kan få bara genom att rota in i datan.

6 Självutvärdering

1. Vad tycker du har varit roligast i kunskapskontrollen?

Det jag gillade mest var att se hur allt hänger ihop att börja med ett rått dataset och steg för steg bygga upp något som faktiskt säger något meningsfullt. Det kändes kul att kunna tolka resultaten och förstå hur olika bilfaktorer påverkar priset på riktigt. Det blev som att lösa ett pussel, fast med kod och statistik.
Det var också roligt att samarbeta med andra klasskamrater i insamlingen av data, även om det inte var en del av själva kunskapskontrollen. Det gav ändå en känsla av att projektet var på riktigt – som ett gemensamt jobb vi alla bidrog till.

2. Hur har du hanterat utmaningar? Vilka lärdomar tar du med dig till framtida kurser?

Det har varit lite av en berg och dalbana, särskilt när vissa tester inte gav tydliga svar eller när jag fastnade i tolkningen av resultaten. Men jag lärde mig att inte ge upp direkt att ta en paus, fråga om hjälp, googla och testa olika vägar. Framför allt har jag insett hur viktigt det är att förstå varför man gör saker inte bara hur.

3. Vilket betyg anser du att du ska ha och varför?

Jag har jobbat med det här projektet med målet att uppnå kriterierna för ett VG och jag hoppas att jag har lyckats. På bara några veckor har jag hunnit bekanta mig med ett helt nytt programmeringsspråk. Även om jag inte kan R ordagrant än, så känner jag att jag förstått grunderna och har kommit en bra bit på vägen.
Jag har också lärt mig mer om datainsamling, modellbygge och hur man tolkar resultat – även om just tolkningen är något jag fortfarande vill bli säkrare på. Det är inget man lär sig över en natt, utan något som växer fram med tiden.
Det som känns bra är att jag inte är rädd för det längre, jag ser faktiskt fram emot att fortsätta lära mig, kanske under LIA-perioden eller på egen hand i små projekt, till exempel på Kaggle. Jag känner att jag är på väg åt rätt håll, och det motiverar mig.

4. Något du vill lyfta till Antonio?

Jag vill tacka för ditt engagemang det har verkligen märkts att du vill att vi ska förstå, inte bara bli klara. Dina tips och förklaringar har hjälpt mycket och jag har känt mig trygg att fråga och vågat prova själv.

Appendix A

Koden: RPubs - Model 1 & Model 2 Volvo-bilar

Källförteckning

Hospitality Institute. (n.d.). *Log Transformation for Regression Linearity*. Retrieved from https://hospitality.institute/MHA1002/log-transformation-regression-linearity/

The Carpentries. (n.d.). *Regression assumptions*. Retrieved from https://carpentries-incubator.github.io/high-dimensional-analysis-in-python/05-Regression-assumptions/index.html

OpenStax. (2023). Introductory Business Statistics *(2nd ed.). Kapitel 13.5: Interpretation of Regression Coefficients: Elasticity and Logarithmic Transformation. https://openstax.org/books/introductory-business-statistics-2e/pages/13-5-interpretation-of-regression-coefficients-elasticity-and-logarithmic-transformation*

Patil, K. (2024, September 1). *Linear Regression vs. Statistical Inference: Understanding Key Differences, Assumptions, and Applications*. LinkedIn. Retrieved from https://www.linkedin.com/pulse/linear-regression-vs-statistical-inference-key-ketan-patil-0zsef/

Statorials. (2021). *Cook's Distance: Identifying Influential Data Points in Regression Analysis* [Video]. YouTube. https://www.youtube.com/watch?v=zJcj8ZdjbYw

University of Virginia Library. (2015). *Understanding diagnostic plots for linear regression analysis*. Hämtad från https://library.virginia.edu/data/articles/diagnostic-plots