R-programmering för dataanalys

Regressionsanalys för att prediktera värdet på begagnad Volvobilar

Lisa Påhlsson

EC Utbildning

2025-04

Abstract

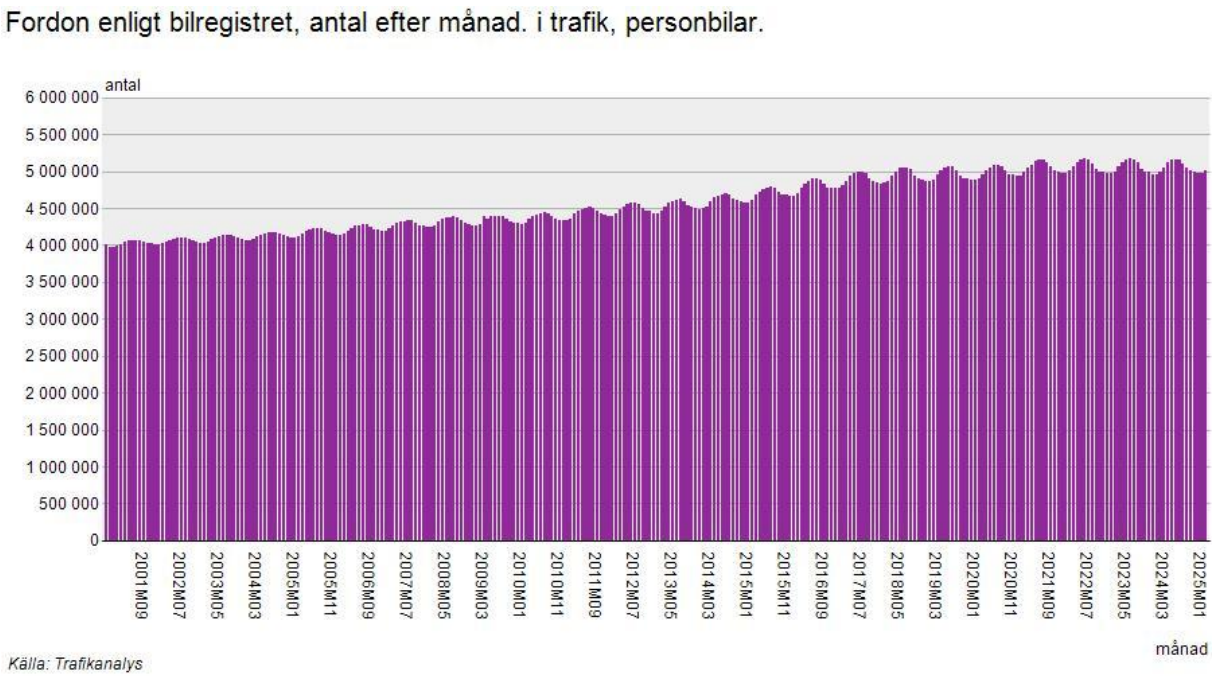
Predictive Modeling of Used Volvo Car Prices Using Regression Analysis

This report develops a predictive model to estimate used Volvo car prices, highlighting key influencing factors. Through correlation analysis and multiple regression modeling, mileage, model year, horsepower, and fuel type were identified as the strongest predictors. After evaluating models using AIC and adjusted R^2, the optimal model was selected, demonstrating strong generalization ability and pricing accuracy. The results indicate that the final model explains a substantial portion of the price variation in both the training and test datasets, suggesting robust predictive performance.

Innehållsförteckning

1 Inledning

Under de senaste två decennierna har antalet personbilar i trafik i Sverige ökat stadigt. Enligt statistik från Trafikanalys (via SCB (2025)) har antalet personbilar ökat från ca 4 miljoner år 2001 till över 5 miljoner år 2025. Denna utveckling visar inte bara på ett ökande bilägande utan skapar också ett större utbud på begagnatmarknaden.



Figur 1: Bilden visar antalet registrerade personbilar i trafik i Sverige per månad, baserat på data från SCB (2025). Grafen indikerar en stadig ökning av antalet bilar i trafik från 2001 till 2023, vilket understryker den växande begagnatmarknaden och behovet av effektiva prissättningsmodeller.

I takt med att antalet begagnade bilar ökar blir behovet av effektiva och rättvisa prissättningsmodeller allt viktigare. Traditionellt har prissättning skett manuellt och baserats på subjektiva bedömningar. Genom att använda statistiska metoder som regressionsanalys kan man automatisera prissättningen på ett effektivare sätt.

1.1 Syfte och frågeställning

Syftet med denna rapport är att skapa en prediktiv modell för att kunna förutspå priset på begagnade Volvobilar och därigenom visa hur datadrivna metoder kan effektivisera prissättningen på en växande marknad.

För att uppfylla syftet kommer följande frågeställningar att besvaras:

1. Vilka faktorer påverkar priset på en begagnad Volvobil?
2. Hur väl kan en regressionsmodell förutsäga priset baserat på dessa faktorer?

2 Teori

2.1 Regressionsanalys

I denna studie används regressionsanalys som en statistisk metod för att undersöka sambandet mellan olika variabler och priset på begagnade Volvobilar.

Regressionsanalys är en grundläggande metod inom statistisk modellering för att kvantifiera samband mellan variabler (James et al., 2013).

2.1.1 Enkel och multipel regression

En enkel linjär regression används när man undersöker sambandet mellan en beroende variabel och en enda oberoende variabel. Formeln för en enkel linjär regression är:

$$Pris = \beta_0 + \beta_1 * X + \epsilon$$

där:

- Pris är det förväntade priset på bilen
- X är en oberoende variabel (t.ex. bilens ålder)
- β_0 är interceptet (pris när X = 0)
- β_1 är lutningen (hur mycket priset förändras för varje enhet förändring i X)
- ϵ är feltermen.

I denna rapport används multipel linjär regression då flera faktorer samtidigt påverkar bilens pris. Modellen blir därmed:

$$Pris = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \ldots + \beta_n * X_n + \epsilon$$

2.1.2 Antagande i linjär regression

För att resultaten från en regressionsmodell ska vara tillförlitliga behöver vissa statistiska antaganden vara uppfyllda (James et al., 2013):

- Linearitet: Sambandet mellan beroende och oberoende variabler ska vara linjärt.
- Oberoende: Observationerna ska vara oberoende av varandra.
- Homoskedasticitet: Variansen hos residualerna (felen) ska vara konstant över alla värden av de oberoende variablerna.
- Normalfördelning av residualerna: Feltermerna bör vara ungefär normalfördelade.

Om dessa antaganden inte är uppfyllda kan modellen behöva justeras, exempelvis genom transformation av variabler.

2.2 Prediktiv modellering och värdet av automatiserad prissättning

Att skapa prediktiva modeller för bilpriser är kan vara värdefullt både för företag och privatpersoner. Genom att använda historisk data och statistiska metoder kan man göra en modell för att förutspå priser. Detta är särskilt viktigt i en marknad där antalet bilar i trafik ökar, vilket har visats tidigare i denna rapport, eftersom konkurrensen och variationen i utbudet ökar allt mer.

3 Metod

3.1 Datainsamling

Datasetet som används i denna analys har ursprungligen samlats in från Blocket.se, det består av en lista över begagnade Volvo-bilar till salu. Datasetet innehåller variabler såsom märke, modell, försäljningspris, modellår, miltal, motorstorlek, hästkrafter, drivmedel, växellåda, färg, biltyp, drivning och ägare.

3.2 Databearbetning

För att säkerställa att datasetet var analysklart genomfördes flera steg av förbehandling

Först sågs Excel-filen över för att hantera följande:

- En kolumnöversikt för att identifiera datamängdens struktur och säkerställa att relevanta variabler inkluderades.
- Kolumnen färg innehåll många varierande färgkombinationer. Färgvariabler med snarlika benämningar, såsom ljusgrå och grå, grupperades för att minska kategorisk variation.
- Även andra kolumner hade mindre fel som korrigerades.

Efter detta gjordes följande hantering i R studio i samband med att EDA utfördes:

- I syfte att fokusera på de mest relevanta variablerna för prisförutsägelse exkluderades *Märke*, *Index*, *Datum_i_trafik* och *År_i_trafik* (denna var en duplicering av Datum_i_trafik men bara med år för jämförelse men blev ej aktuell att använda).
- Variabeln "Motorstorlek" omvandlades från text till numerisk form.
- Saknade värden hanterades genom att ersätta saknade värden i "Hästkrafter" och "Motorstorlek" med respektive medianvärde. Saknade värden i kategoriska variabler som "Växellåda", "Biltyp", "Drivning", "Färg" och "Modell" ersattes med "Okänt", dessa var väldigt få.
- Kategoriska variabler konverterades till faktorer för att kunna användas i regressionsanalysen.

Utöver detta framkom att försäljningspriserna var kraftigt högerfördelade och därför log-transformerades prisvariabeln ("Försäljningspris") till "log_pris" för att bättre uppfylla linjära regressionsantaganden om normalfördelade residualer.

3.3 Modellering

I denna analys användes endast en uppdelning i tränings- och testdata, utan en separat valideringsmängd, eftersom modellvalet baserades på AIC-värdet. AIC (Akaike Information Criterion) är ett informationskriterium som tar hänsyn till både modellens anpassning och komplexitet, vilket minskar behovet av en separat valideringsmängd för modellselektering. Testdatan användes sedan för att utvärdera modellens generaliseringsförmåga.

Modellerna jämfördes med hjälp av AIC, BIC (Bayesian Information Criterion) och justerat R^2-värde. Den modell med lägst AIC-värde valdes för vidare analys. I denna modell identifierades variablerna Miltal, Modellår, Hästkrafter och Bränsle som de mest signifikanta faktorerna för prisprediktion.

3.4 Modellutvärdering

Den slutgiltiga modellen utvärderades genom:

- Residualdiagnostik, inklusive residualplots och histogram, användes för att verifiera att modellens grundläggande antaganden om linjäritet och homogen varians var uppfyllda.
- Beräkning av 95 % konfidensintervall för regressionskoefficienterna
- Beräkning av prediktionsförmåga på testdata genom RMSE (Root Mean Squared Error) och R^2-värde.
- Analys av multikollinearitet med hjälp av VIF (Variance Inflation Factor)

3.5 Extern data

Extern statistik från SCB (Trafikanalys) visar en stadig ökning av antalet personbilar i trafik i Sverige, vilket stärker behovet av datadrivna prissättningsmodeller på den växande begagnatmarknaden.

4 Resultat och Diskussion

För att identifiera den bästa modellen jämfördes fem olika modeller utifrån sina AIC-värden. AIC och BIC användes som jämförelsemått eftersom de tar hänsyn både till modellens förklaringsgrad och komplexitet. Lägre AIC/BIC indikerar en bättre balans mellan god förklaringsförmåga och låg komplexitet, vilket bidrar till att undvika överanpassning.

Modellen som inkluderade variabeln Bränsle tillsammans med basvariablerna (Miltal, Modellår och Hästkrafter) hade lägst AIC (546,8) och valdes därför för vidare analys. Vissa kategoriska variabler med många nivåer, såsom Färg och Modell, exkluderades från analysen på grund av tekniska utmaningar vid faktorisering och för att hålla modellen hanterbar och tolkbar. Initialt verkade Modell påverka priset men eftersom variabeln behövde grupperas för att vara praktiskt användbar förlorades detta samband vid fortsatt analys.

Vid analys av den valda modellen visade samtliga basvariabler signifikanta samband med det log-transformerade priset ($p < 0{,}001$). Bränsletyp påverkade också priset.

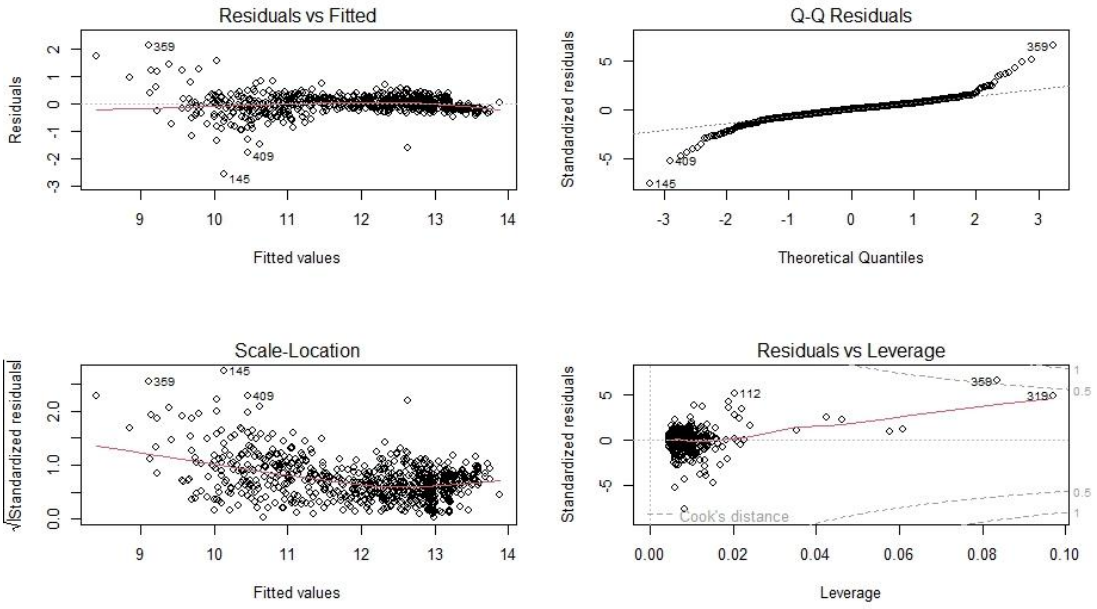
Konfidensintervallen för modellens koefficienter visade att:

- Miltal hade ett smalt negativt intervall (-0,0000411 till -0,0000341) vilket bekräftar en liten men statistiskt signifikant negativ effekt på försäljningspriset.

- Modellår hade ett positivt intervall (0,0746 till 0,0865) vilket innebär att nyare bilar tenderar att säljas till högre priser.
- Hästkrafter påverkade också priset positivt (0,0022 till 0,0029).
- Bränsletyp:
    - Dieselbilar hade ett positivt intervall (0,2636 till 0,4084) vilket tyder på ett stabilt högre pris jämfört med bensinbilar.
    - El- och hybridbilar hade intervall som inkluderade noll, vilket innebär att deras prisnivåer inte skilde sig signifikant från bensinbilar på 95 %-konfidensnivå. Det innebär att det inte finns tillräcklig statistisk evidens för att hävda en systematisk prisskillnad för dessa drivmedelstyper.

Sammanfattningsvis indikerar konfidensintervallen att Miltal, Modellår, Hästkrafter och Bränsletyp (Diesel) är robusta prediktorer för bilens försäljningspris.

Modellen förklarade en stor del av variationen i försäljningspriset (Adjusted R^2 = 0,9168). Residualanalysen (se Figur 2) visade inga tydliga systematiska avvikelser, vilket tyder på att antagandena om homogen varians och linjäritet är rimligt uppfyllda.
Dock observerades viss svansbildning i Q-Q-plottens ytterkanter, vilket tyder på mindre avvikelser från normalitet i residualerna. Detta kan påverka precisionen i skattningarna av konfidensintervall, men bedöms ha begränsad praktisk betydelse för modellens prediktionsförmåga.

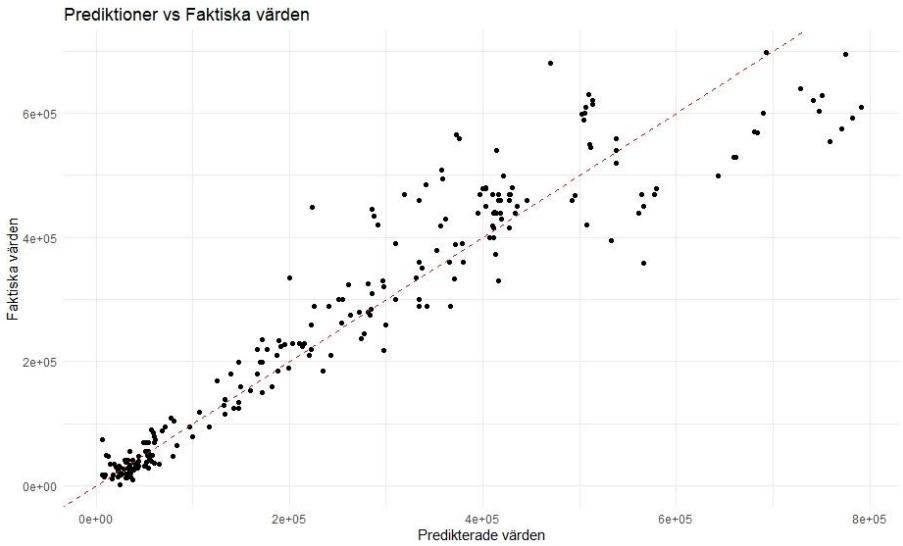


Figur 2: Bilden visar diagnostiska plottar för den slutgiltiga regressionsmodellen. Residualer vs. Fitted-plott indikerar att ingen tydlig icke-linjär trend förekommer. Q-Q-plott visar att residualerna generellt följer en normalfördelning, med vissa mindre avvikelser i ytterkanterna. Scale-Location-plott tyder på homogen varians i residualerna, medan Residuals vs. Leverage-plott visar att inga datapunkter har oproportionerligt stor påverkan på modellen.

RMSE	R^2
66 703 kr	0,900

Tabell 1: Root Mean Squared Error (RMSE) och R^2 på testdata

RMSE-värdet på cirka 66 700 kronor indikerar att modellen i genomsnitt gör ett fel på denna summa vid prisprediktioner och modellen förklarade cirka 90 % (R^2 = 0,90) av variationen i de faktiska priserna på testdata. Prediktionsplotten (se Figur 3) visar att modellen generellt skattar priserna väl, även om vissa avvikelser finns för enstaka observationer. Eftersom test-R^2 (0,900) inte skiljer sig dramatiskt från tränings-Adjusted R^2 (0,9168) tyder det på att modellen inte är överanpassad utan generaliserar väl även till ny data. Den lilla skillnaden mellan tränings- och testresultaten indikerar att modellen är stabil och har god prediktionsförmåga.



Figur 3: Figuren visar en jämförelse mellan de predikterade och faktiska försäljningspriserna. De flesta värden ligger nära identitetslinjen, vilket tyder på att modellen har god prediktionsförmåga. Vissa avvikelser observeras dock, särskilt för bilar med högre priser, vilket kan indikera en något högre osäkerhet i detta segment.

5 Slutsatser

Syftet med denna studie var att skapa en prediktiv modell för att förutspå försäljningspriset på begagnade Volvobilar och visa hur datadrivna metoder kan effektivisera prissättningen på en växande marknad.

1. Vilka faktorer påverkar priset på en begagnad Volvobil?
Analysen visade att följande variabler hade en signifikant påverkan på priset:

- **Miltal**: Ett högre miltal minskade försäljningspriset.
- **Modellår**: Nyare årsmodeller såldes till högre priser.
- **Hästkrafter**: Bilar med fler hästkrafter hade högre priser.
- **Bränsletyp**: Dieselbilar såldes generellt dyrare än bensinbilar, medan el- och hybridbilar inte visade en signifikant prisskillnad.

2. Hur väl kan en regressionsmodell förutsäga priset baserat på dessa faktorer?
Den bästa modellen (baserad på lägst AIC) inkluderade Miltal, Modellår, Hästkrafter och Bränsletyp som förklarande variabler. Modellen uppnådde ett justerat R^2 på 0,9168, vilket innebär att cirka 92 % av variationen i försäljningspriset kan förklaras av modellen.
Vid testning på ny data visade modellen ett R^2 på 0,90 och ett RMSE på cirka 66 700 kronor vilket indikerar att modellen har god förmåga att förutsäga priset på begagnade Volvobilar med relativt hög precision.

Sammanfattningsvis visar resultaten att en noggrant utvald regressionsmodell kan vara ett effektivt verktyg för att förutsäga bilpriser och därmed bidra till en mer datadriven och effektiv prissättningsstrategi på begagnatmarknaden.

6 Teoretiska frågor

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

QQplots används främst för att visuellt jämföra en datamängds fördelning med en referensfördelning. I regressionsanalyser används QQ-plots ofta för att kontrollera om residualerna är normalfördelade. Det är också möjligt att använda QQ-plots för att jämföra två datamängder men huvudsyftet är att jämföra en datamängds fördelning mot en förväntad teoretisk fördelning.

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?

Ja, det stämmer. Maskininlärning fokuserar främst på prediktioner med målet att skapa modeller som kan förutse ett utfall. I statistisk regressionsanalys kan man göra både och vilket innebär att man kan analysera hur starkt sambandet är mellan variabler och avgöra om resultaten är statistik signifikant, dvs att dra slutsatser om en population baserat på ett stickprov. Man kan säga att vid maskininlärning är resultatdriven medan statistisk regressionsanalys är för att utforska och förstå sambanden med resultatet.

Exempel:

- Maskininlärning kan ett företag använda för att förutsäga framtida försäljningssiffror baserat på tidigare köpbeteenden och trender. Målet är att skapa en modell som är noggrann.
- Statistisk regressionsanalys är fördelaktigt av exempelvis forskare att använda för att förutspå sambandet mellan utbildningsnivå och inkomst. Då kan man både göra prediktioner av vad en person tjänar i snitt och utföra statistisk inferens för att avgöra om sambandet är statistiskt signifikant eller beror på slumpen.

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

Ett konfidensintervall för ett predikterat värde beskriver osäkerheten kring modellens genomsnittliga förväntade värde, för en viss kombination av X-variabler. Ett prediktionsintervall är bredare och inkluderar både osäkerheten i modellen och slumpvariationen hos enskilda observationer.

4. **Den multipla linjära regressionsmodellen kan skrivas som: $Y = \beta 0 + \beta 1x1 + \beta 2x2 + \ldots + \beta pxp + \varepsilon$. Hur tolkas beta parametrarna?**

Varje betaparameter(β_1, β_2 osv) visar hur mycket det förväntade värdet på Y förändras om den tillhörande x-variabeln ökar med 1 enhet under förutsättning att alla andra variabler är konstanta.

β_0 är interceptet, vilket är modellens förväntade värdet på Y när alla x-variabler är 0.

5**. Din kollega Nils 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 Nils?**

Nej, mått som BIC och AIC justerar för modellens komplexitet och kan användas för att jämföra modeller utan att dela data. Men de mäter inte hur bra modellen generaliserar till ny data. För att kontrollera modellens prediktionsförmåga i praktiken behövs fortfarande träning, validering och test set.

6. Förklara algoritmen nedan för ”Best subset selection”

Best subset selection innebär att man testar alla möjliga kombinationer av prediktorer för att hitta den modell som presterar bäst enligt ett kriterium (t.ex. lägst AIC, BIC eller högst R^2). Denna metod kan dock bli tung beräkningsmässigt om antalet prediktorer växer.

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

Ingen modell beskriver verkligheten exakt, egentligen är alla modeller förenklaringar. Däremot kan modeller ändå vara användbara eftersom de hjälper oss att förstå samband, dra slutsatser eller göra förutsägelser.

7 Självutvärdering

1. **Utmaningar du haft under arbetet samt hur du hanterat dem.**
   Motivation och ork framför allt, jag har knappt hanterat dem heller då orken inte finns att samla ihop allt riktigt för tillfället. Hoppa det går att ta sig igenom nästa kurs också för det var med nöd och näppe jag ens fick ihop detta känns det som. Men har lovat mig själv att hänga i terminen ut så får sommaren avgöra om det går rädda upp motivation och ork att plugga vidare.

2. **Vilket betyg du anser att du skall ha och varför.**
   G då jag tycker jag förstått de grundläggande begreppen.

3. **Något du vill lyfta fram till Antonio?**

   -

Appendix A

```
# Laddar bibliotek

library(tidyverse)

library(readxl)

library(knitr)

library(car)


# Läser in data

bil_data <- read_excel("data_insamling_volvo_blocket.xlsx")


# Tar bort irrelevanta kolumner

bil_data <- bil_data %>% select(-c(Märke, Index, Datum_i_trafik, År_i_trafik))


# Hanterar saknade värden

bil_data <- bil_data %>%

  mutate(

    Hästkrafter = ifelse(is.na(Hästkrafter), median(Hästkrafter, na.rm = TRUE),
Hästkrafter),

    across(c(Växellåda, Biltyp, Drivning, Färg, Modell), ~replace_na(.x, "Okänt")),

    Motorstorlek = parse_number(Motorstorlek)

  )

bil_data$Motorstorlek[is.na(bil_data$Motorstorlek)] <- median(bil_data$Motorstorlek, na.rm
= TRUE)


# Omvandlar kategoriska variabler till faktorer

bil_data <- bil_data %>%

  mutate(across(c(Säljare, Bränsle, Växellåda, Biltyp, Drivning, Färg, Modell), as.factor))


# Skapar ny modellvariabel för vanligaste modellerna

topp_modeller <- names(sort(table(bil_data$Modell), decreasing = TRUE)[1:7])

bil_data$Modell_topp <- ifelse(bil_data$Modell %in% topp_modeller,
as.character(bil_data$Modell), "Övrig")


# Skapar logaritmerad prisvariabel

bil_data$log_pris <- log(bil_data$Försäljningspris)


# Dela in i tränings- och testdata
```

```
set.seed(123)
train_index <- sample(1:nrow(bil_data), size = 0.7 * nrow(bil_data))
train_data <- bil_data[train_index, ]
test_data <- bil_data[-train_index, ]

# Definierar modeller
bas_formula <- log_pris ~ Miltal + Modellår + Hästkrafter
modeller <- list(
  lm(bas_formula, data = bil_data),
  lm(update(bas_formula, . ~ . + Motorstorlek), data = bil_data),
  lm(update(bas_formula, . ~ . + Bränsle), data = bil_data),
  lm(update(bas_formula, . ~ . + Motorstorlek + Bränsle), data = bil_data),
  lm(update(bas_formula, . ~ . + Motorstorlek + Bränsle + Drivning), data = bil_data)
)
modellnamn <- c("Bas", "+Motorstorlek", "+Bränsle", "+Motorstorlek+Bränsle",
"+Motorstorlek+Bränsle+Drivning")

# Jämför modeller (AIC/BIC)
jämför_modeller <- data.frame(
  Modell = modellnamn,
  AIC = sapply(modeller, AIC),
  BIC = sapply(modeller, BIC),
  Adjusted_R2 = sapply(modeller, function(m) summary(m)$adj.r.squared)
) %>% arrange(AIC)

# Väljer bästa modell
bästa_modell <- modeller[[which.min(sapply(modeller, AIC))]]

# Analyserar bästa modellen
summary(bästa_modell)
vif(bästa_modell)
confint(bästa_modell)

# Modellvalidering på testdata
train_model <- lm(formula(bästa_modell), data = train_data)
test_data$predikt_log_pris <- predict(train_model, newdata = test_data)
test_data$predikt_pris <- exp(test_data$predikt_log_pris)
```

```
# Beräknar RMSE och R²
rmse <- sqrt(mean((test_data$Försäljningspris - test_data$predikt_pris)^2))
r2 <- cor(test_data$Försäljningspris, test_data$predikt_pris)^2


# Visualiserar prediktioner
ggplot(test_data, aes(x = predikt_pris, y = Försäljningspris)) +
  geom_point() +
  geom_abline(intercept = 0, slope = 1, color = "red", linetype = "dashed") +
  labs(x = "Predikterade värden", y = "Faktiska värden", title = "Prediktioner vs Faktiska
värden") +
  theme_minimal()
```

Källförteckning

SCB (2025). *Antal personbilar i trafik i Sverige*. https://www.statistikdatabasen.scb.se/pxweb/sv/ssd/START__TK__TK1001__TK1001A/Fordon/ [2025-04-25]

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning*. New York: Springer.