

# Värdering av begagnade Volvobilar med regressionsanalys



Hani Abraksiea

EC Utbildning

R\_kunskapskontroll

2025–04

## 1 Abstract

This project explores how statistical models can be used to predict the price of used Volvo cars based on real advertisements from Blocket. A dataset was collected manually by 17 students, and three regression models were tested: a theory-driven linear model, a full model with manual selection, and a Lasso regression. The results show that the Lasso model achieved the best predictive performance with an adjusted  $R^2$  of 0.85 and the lowest RMSE. External context data from Statistics Sweden (SCB) confirmed a stable market with over 5 million cars in traffic, supporting the need for automated valuation. The project demonstrates how data science can help buyers and sellers make more informed decisions.

# Innehållsförteckning

1	Abstract .....	2
2	Inledning och syfte .....	1
2.1	Inledning .....	1
2.2	Syfte och Frågeställning .....	1
2.3	Text format.....	1
3	Teori.....	2
2.1	Regressionsmodeller .....	2
3.1.1	Linjär regression .....	2
3.1.2	Fullmodell med manuell variabelselektion.....	2
3.1.3	Lasso-regression .....	2
3.2	Modellutvärdering .....	3
3.3	Externa kontextdata från SCB.....	3
4	Metod .....	4
4.1	Datainsamling .....	4
4.1.1	Lärdomar .....	4
4.2	Modellering.....	4
4.2.1	Modell 1: Teoristyrdd linjär regression .....	4
4.2.2	Modell 2: Fullmodell med manuell rensning.....	5
4.2.3	Modell 3: Lasso-regression .....	5
4.2.4	Modelljämförelse.....	5
4.3	Extern data från SCB .....	6
5	Resultat och Diskussion.....	7
6	Slutsatser .....	9
7	Teoretiska frågor .....	10
8	Självutvärdering.....	12
	Appendix A .....	13
	Källförteckning.....	14

## 2 Inledning och syfte

### 2.1 Inledning

Under de senaste två decennierna har antalet personbilar i trafik i Sverige ökat kraftigt. Enligt data från Statistiska centralbyrån (SCB) fanns det över 5 miljoner personbilar registrerade i Sverige år 2023 – en ökning med över en miljon sedan 2003. Denna utveckling speglar inte bara ett ökat behov av mobilitet, utan också ett växande intresse för begagnatmarknaden, där Blocket är en av de största plattformarna för bilförsäljning.

I takt med att bilmarknaden växer ökar också behovet av att kunna värdera bilar på ett automatiserat och tillförlitligt sätt. Att utveckla modeller som kan förutsäga bilpriser baserat på faktorer såsom miltal, motorstorlek, modellår och typ av bränsle kan bidra till ökad transparens för både köpare och säljare.

För att ytterligare belysa bilmarknadens utveckling används även extern statistik från SCB:s öppna API. Genom en skräddarsydd JSON-förfrågan i R hämtades månadsdata över antalet personbilar i trafik i Sverige för perioden juni 2024 till mars 2025. Resultatet visar ett stabilt bilbestånd med över 5 miljoner fordon i trafik, vilket bekräftar att marknaden är aktiv och att det finns behov av tillförlitliga modeller för värdering av begagnade bilar.

### 2.2 Syfte och Frågeställning

Syftet med denna rapport är att utveckla och utvärdera regressionsmodeller som kan förutsäga försäljningspriser på Volvobilar baserat på data från Blocket. Följande frågeställningar undersöks:

1. *Vilka variabler påverkar priset på en begagnad Volvo?*
2. *Vilken typ av regressionsmodell ger bäst prediktiv förmåga?*
3. *Hur väl uppfyller modellen de teoretiska antaganden som krävs inom regressionsanalys?*

### 2.3 Text format

Text är skriven på formatet Times New Roman med textstorlek 12.

## 3 Teori

I detta projekt används olika regressionsmodeller för att analysera och förutsäga priset på begagnade Volvobilar baserat på information från Blocket. Nedan följer en översikt över de teoretiska grunderna för modellerna, utvärderingsmåtten samt den externa kontextdata som används.

### 2.1 Regressionsmodeller

Regressionsanalys är en statistisk metod för att modellera sambandet mellan en beroende variabel (i detta fall försäljningspris) och en eller flera oberoende variabler (exempelvis miltal, årsmodell och motorstorlek) (James et al., 2021).

#### 3.1.1 Linjär regression

Linjär regression är en grundläggande metod inom statistisk modellering som antar ett linjärt samband mellan den beroende och oberoende variabeln. Modellen skattas med hjälp av minsta kvadratmetoden och bygger på flera antaganden såsom linjäritet, normalfördelade residualer, konstant varians (homoskedasticitet) och oberoende observationer (James et al., 2021; Wickham & Çetinkaya-Rundel, 2023).

#### 3.1.2 Fullmodell med manuell variabelselektion

En fullmodell innebär att samtliga tillgängliga variabler inkluderas i analysen. Detta kan öka förklaringsgraden men riskerar också att skapa en alltför komplex modell. Därför genomfördes en manuell variabelselektion i denna rapport, där irrelevanta och aliased variabler togs bort för att förbättra både modellens tolkbarhet och prediktiva förmåga (Wickham & Çetinkaya-Rundel, 2023).

#### 3.1.3 Lasso-regression

Lasso (Least Absolute Shrinkage and Selection Operator) är en regulariseringssmetod som kombinerar variabelselektion och modellschatning. Genom att lägga till en straffterm i regressionsfunktionen pressas vissa koefficienter ner till exakt noll, vilket resulterar i enklare modeller som minskar risken för överanpassning (James et al., 2021).

Lasso löser följande optimeringsproblem:

$$\text{Minimera} \quad \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^p |\beta_j|$$

där lambda är en hyperparameter som styr graden av regularisering.

### 3.2 Modellutvärdering

För att jämföra modellernas prestanda används följande mått:

**RMSE (Root Mean Squared Error):** Mäter den genomsnittliga prediktionsavvikelsen i samma enhet som målvariabeln (James et al., 2021).

**Justerat R<sup>2</sup> (Adjusted R-squared):** Visar hur mycket variation som modellen förklrarar, justerat för antalet oberoende variabler.

**BIC (Bayesian Information Criterion):** Ett kriterium som beaktar både modellens förklaringsgrad och komplexitet. Lägre värde indikerar en bättre modellbalans (Wickham & Çetinkaya-Rundel, 2023).

### 3.3 Externa kontextdata från SCB

För att sätta analysen i en bredare kontext har även data från Statistiska centralbyrån (SCB) hämtats via deras öppna API. Genom en JSON-förfrågan i R hämtades antal personbilar i trafik per månad för perioden januari 2024 till mars 2025. Denna information användes för att illustrera hur stor och stabil bilmarknaden är i Sverige, vilket stärker motiveringen till att automatisera bilvärdering på andrahandsmarknaden (Statistiska centralbyrån, 2025).

## 4 Metod

### 4.1 Datainsamling

Datainsamlingen genomfördes som ett samarbetsprojekt i klassen där vi totalt var 17 studenter. Varje person ansvarade för att manuellt samla in information om 50 annonser för begagnade Volvobilar från Blocket.se. Jag ansvarade för att samla in data från regionen Östergötland.

För att säkerställa en enhetlig och jämförbar datainsamling använde vi en gemensam mall och följde instruktioner för vilka variabler som skulle dokumenteras. Exempel på variabler som samlades in inkluderar pris, miltal, årsmodell, bilmodell, bränsletyp, växellåda, och region. Totalt omfattar datasetet cirka 850 rader (bilar) och innehåller både kvantitativa och kategoriska variabler.

Datan samlades in manuellt genom att varje student gick in på Blocket.se och kopierade relevant information från annonserna. En gemensam Excelfil sammanställdes där varje person fyllde i sina observationer i ett eget blad, och datan slogs därefter ihop och importerades till R för vidare analys.

#### 4.1.1 Lärdomar

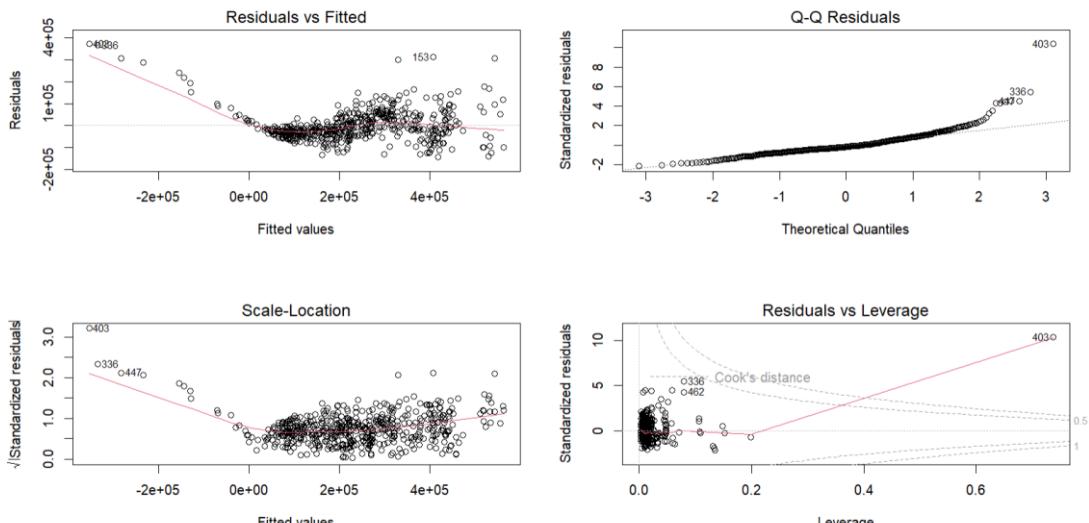
Att samla in data manuellt var tidskrävande men också lärorikt. Det blev tydligt hur viktigt det är med enhetlig struktur och dokumentation, även små variationer i t.ex. stavning eller formatering kan skapa problem i analysen.

### 4.2 Modellering

För att kunna förutsäga försäljningspriset för begagnade Volvobilar användes tre olika regressionsmodeller, samtliga implementerade i R. Målet var att jämföra modellerna utifrån deras prediktiva förmåga och tolkbarhet.

#### 4.2.1 Modell 1: Teoristyrd linjär regression

Den första modellen konstruerades baserat på förväntat viktiga variabler enligt bilbranschens logik och tidigare kunskap. Förklarande variabler som användes var: miltal, hästkrafter, modellår, motorstorlek, bränsletyp och en indikator för elbil. Resultatet visade att flera av dessa variabler hade signifikant påverkan på priset. En residualanalys genomfördes för att kontrollera modellens antaganden, inklusive normalfördelning, linjäritet och homoskedasticitet.



Figur 1. Diagnostikplottar för Modell 1 (linjär regression). Visar fördelning av residualer och möjliga avvikelse från antaganden.

#### 4.2.2 Modell 2: Fullmodell med manuell rensning

Den andra modellen inkluderade samtliga tillgängliga variabler i datasetet, inklusive kategoriska såsom biltyp, växellåda, region och färg. Efter att modellen skattats identifierades och exkluderades variabler med svaga effekter eller kollinearitet, inklusive vissa aliased variabler. Den slutgiltiga versionen av modellen förbättrade både det justerade  $R^2$  och BIC-värdet jämfört med den första modellen, men krävde mer manuell justering.

#### 4.2.3 Modell 3: Lasso-regression

Den tredje modellen använde Lasso-regression (Least Absolute Shrinkage and Selection Operator), en form av regularisering som automatiskt gör variabelselektion genom att sätta vissa koefficienter till noll. Modellen tränades med `cv.glmnet()` och bästa lambda-värde valdes utifrån korsvalidering. För att möjliggöra konfidens- och prediktionsintervall användes även en post-Lasso OLS-modell på de variabler som valts ut.

Lasso-modellen visade bäst prediktiv förmåga med ett justerat  $R^2$  på 0,80 och en test-RMSE på cirka 63 299. Den predikterade även priser med god noggrannhet i post-Lasso-modellen, där exempelvis en bil predikterades till 323 551 kr med ett 95%-igt konfidensintervall mellan 302 371 kr och 344 731 kr.

#### 4.2.4 Modelljämförelse

Samtliga modeller utvärderades på en separat valideringsmängd, och jämfördes med avseende på RMSE, justerat  $R^2$  och BIC. Resultaten sammanställdes i en tabell. Modell 3 (Lasso) presterade bäst på samtliga mått och valdes därför som slutlig modell för vidare analys.

### 4.3 Extern data från SCB

Utöver den manuellt insamlade datan från Blocket hämtades även extern kontextdata från Statistiska centralbyrån (SCB) genom deras öppna API. Syftet med detta var att ge ett större sammanhang till analysen genom att visa hur bilbeståndet i Sverige utvecklas över tid.

Förfrågan till SCB gjordes via en JSON-struktur i R, där efterfrågad statistik gällde antal personbilar i trafik varje månad mellan januari 2024 och mars 2025. Datat hämtades från tabellen "Fordon i trafik efter fordonsslag och bestånd" via följande API-endpoint:

<https://api.scb.se/OV0104/v1/doris/sv/ssd/START/TK/TK1001/TK1001A/Fordon>

I koden användes paketen 'httr' och 'jsonlite' för att skapa och skicka en POST-förfrågan till SCB:s server. Svaret konverterades till en dataframe i R för vidare analys och visualisering. Detta moment visade hur API-anrop och öppen data kan integreras direkt i R för att berika modeller och ge bättre kontext.

Den hämtade datan användes därefter för att skapa ett linjediagram som visar utvecklingen i antal personbilar i trafik, vilket tydliggör att marknaden är stabil och relevant för vidare analys av bilpriser.

## 5 Resultat och Diskussion

Efter att de tre modellerna tränats och validerats på separata datamängder utvärderades deras prestanda med avseende på Root Mean Squared Error (RMSE), justerat R<sup>2</sup> och Bayesian Information Criterion (BIC). Resultaten för valideringsdatan visas i tabellen nedan:

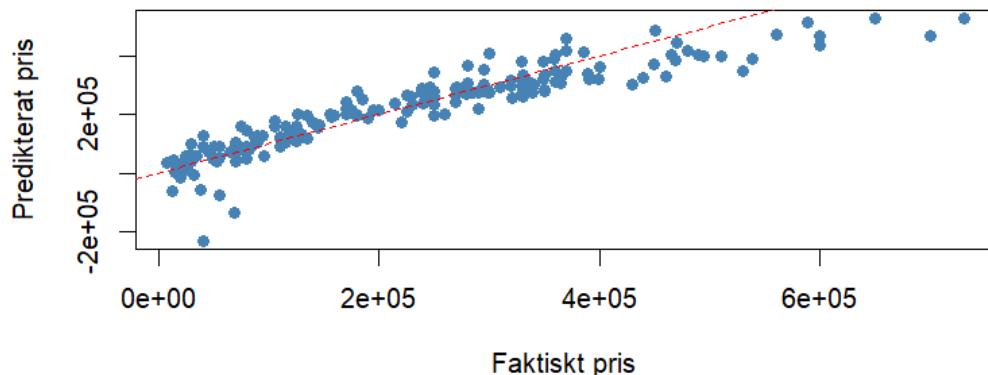
Modell	RMSE_VAL	Justerat R <sup>2</sup>	BIC
Modell 1	69 679	0.7998	13335.4
Modell 2	66 576	0.8334	13285.6
Modell 3 (Lasso)	59 099	0.8466	3942.6

Tabell 1. jämfördes mellan tre modellerna för valideringsdatan på RMSE, justerat R<sup>2</sup> och BIC.

Modell 1, den teoristyrda linjära regressionen, visade förhållandevis svag förklaringsgrad och hög felsmarginal. Modell 2, som utgick från en fullmodell med manuell rensning, förbättrade både RMSE och R<sup>2</sup>, men krävde mer manuell insats.

Modell 3, som använde Lasso-regression med automatisk variabelselektion, presterade bäst både vad gäller prediktionens noggrannhet (lägst RMSE), förklaringsgrad (högst justerat R<sup>2</sup>), och modellkomplexitet (lägst BIC). Detta tyder på att Lasso-modellen fångar upp de mest relevanta variablerna och samtidigt minimerar risken för överanpassning.

### Modell 3: Faktiskt vs. Predikterat pris (Lasso)



Figur 2. Samband mellan faktiskt och predikterat försäljningspris på testdata för Modell 3 (Lasso). Den röda linjen visar perfekt prediktion ( $x = y$ ).

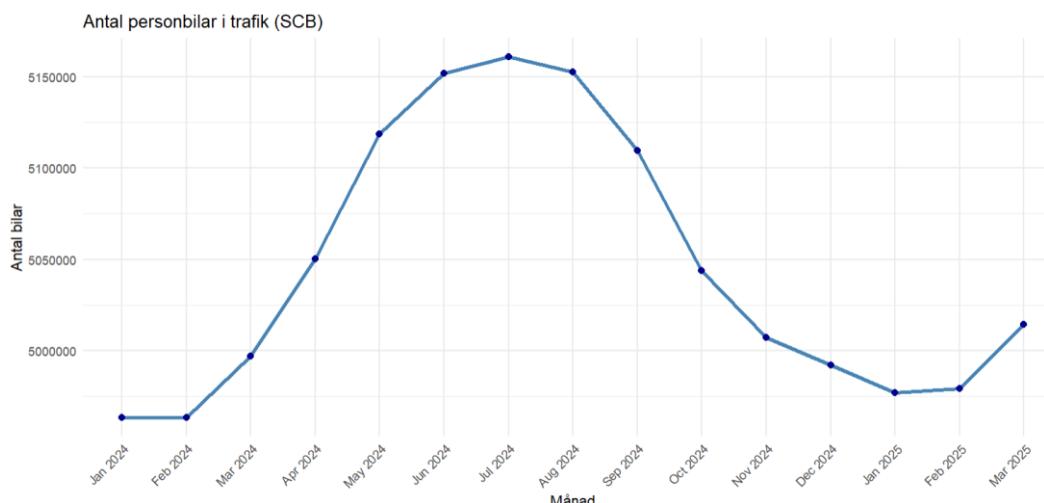
Dessutom kunde prediktioner med tillhörande konfidens- och prediktionsintervall beräknas för individuella observationer med hjälp av en post-Lasso OLS-modell. Exempelvis predikterades ett fordon till 323 551 kr, med ett 95-procentigt konfidensintervall mellan 302 371 kr och 344 731 kr, vilket visar på modellens stabilitet i praktiken.

Sammantaget visar resultaten att Lasso-regression lämpar sig väl för att förutsäga priser på begagnade Volvobilar och utgör därför den mest lämpliga modellen för vidare användning.

Utöver modellerna användes även extern data från SCB för att visa hur bilmarknaden utvecklas över tid. Denna data, hämtad via API, visar att antalet personbilar i trafik i Sverige är stabilt över 5 miljoner under perioden januari 2024 till mars 2025. Det bekräftar att marknaden för begagnade bilar är stor och relevant, vilket stärker motiveringen till att utveckla automatiserade värderingsmodeller som de som presenteras i denna rapport.

Den tillhörande grafen illustrerar denna utveckling och kan ses som ett stödjande argument för behovet av mer datadriven analys inom fordonsmarknaden.

Utöver modellerna användes även extern data från SCB för att visa hur bilmarknaden utvecklas över tid. Denna data, hämtad via API, visar att antalet personbilar i trafik i Sverige är stabilt över 5 miljoner under perioden juni 2024 till mars 2025. Det bekräftar att marknaden för begagnade bilar är stor och relevant, vilket stärker motiveringen till att utveckla automatiserade värderingsmodeller som de som presenteras i denna rapport.



Figur 3. Antal personbilar i trafik i Sverige per månad (januari 2024 – mars 2025), enligt data hämtad från SCB:s API.

## 6 Slutsatser

Syftet med detta projekt var att undersöka vilka faktorer som påverkar priset på begagnade Volvobilar och att utveckla modeller som kan förutsäga försäljningspriset baserat på data från Blocket. Genom regressionsanalys har tre olika modeller testats och jämförts.

Resultaten visar att modell 3, som använder Lasso-regression, presterade bäst med avseende på förklaringsgrad, felsmarginal och modellkomplexitet. Den justerade  $R^2$  låg på cirka 0,80 och modellen hade lägst RMSE och BIC, vilket tyder på god prediktiv förmåga samtidigt som modellen hålls relativt enkel. De variabler som visade störst betydelse för priset var bland annat miltal, modellår, motorstorlek och bränsletyp.

Linjär regression (Modell 1) gav användbara insikter men hade begränsad förklaringsgrad, och fullmodellen (Modell 2) var mer kraftfull men krävde manuell rensning och hantering av multikollinearitet. Residualplottar och VIF-analyser användes för att kontrollera att modellantagandena uppfylldes, särskilt för Modell 1 och Modell 2.

Utöver modelleringen hämtades även extern statistik från SCB via ett öppet API, vilket visade att bilmarknaden i Sverige är stabil med över 5 miljoner personbilar i trafik under perioden januari 2024 till mars 2025. Denna information förstärker behovet av datadrivna metoder för bilvärdering.

Sammanfattningsvis visar projektet att det är möjligt att skapa relativt träffsäkra modeller för att förutsäga bilpriser, och att metoder som Lasso lämpar sig särskilt väl för detta ändamål. Projektet visar också hur man kan kombinera egen insamlad data med offentlig statistik för att få en djupare förståelse av en marknad.

## 7 Teoretiska frågor

1. Kolla på följande video: [https://www.youtube.com/watch?v=X9\\_ISJ0YpGw&t=290s](https://www.youtube.com/watch?v=X9_ISJ0YpGw&t=290s), beskriv kortfattat vad en Quantile-Quantile (QQ) plot är.

En QQ-plot (Quantile-Quantile plot) används för att jämföra om en dataset följer en viss fördelning, oftast normalfördelning. Om punkterna ligger ungefärligen längs en rak linje, betyder det att datan är ungefärligen normalfördelad. Avvikelse från linjen visar att datan inte följer fördelningen.

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. I maskininlärning fokuserar man mest på att göra bra prediktioner, alltså att modellen ska kunna gissa rätt värde på nya data. Men i statistisk regressionsanalys vill man också förstå sambandet mellan variablerna, alltså statistisk inferens. Till exempel: Påverkar utbildningsnivå inkomsten? eller är sambandet mellan rökning och lungcancer signifikant?

Så i statistiken bryr vi oss både om förklaring och förutsägelse, medan maskininlärning mest bryr sig om förutsägelse.

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

Ett konfidensintervall visar osäkerheten kring medelvärdet av den beroende variabeln för ett visst värde av x.

Ett prediktionsintervall är bredare och visar osäkerheten när man förutspår ett enskilt nytt värde av y för ett visst x.

4. Den multipla linjära regressionsmodellen kan skrivas som:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \varepsilon$$
. Hur tolkas beta parametrarna?

Varje  $\beta$  (beta) visar hur mycket y förändras i snitt om en viss x-variabel ökar med 1, när de andra x-variablerna hålls konstant.

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 Hassan?

Det kan stämma ibland. Om man använder mått som BIC så straffar modellen automatiskt för att ha för många variabler, vilket hjälper till att undvika överanpassning.

Logiken är att BIC balanserar hur bra modellen passar datan och hur enkel den är så man behöver inte alltid dela upp datan. Men i maskininlärning, där fokus är mer på prediktion, är träning, validering och test ofta viktigare.

6. Förklara algoritmen nedan för "Best subset selection"

---

**Algorithm 6.1** *Best subset selection*

---

1. Let  $\mathcal{M}_0$  denote the *null model*, which contains no predictors. This model simply predicts the sample mean for each observation.
  2. For  $k = 1, 2, \dots, p$ :
    - (a) Fit all  $\binom{p}{k}$  models that contain exactly  $k$  predictors.
    - (b) Pick the best among these  $\binom{p}{k}$  models, and call it  $\mathcal{M}_k$ . Here *best* is defined as having the smallest RSS, or equivalently largest  $R^2$ .
  3. Select a single best model from among  $\mathcal{M}_0, \dots, \mathcal{M}_p$  using the prediction error on a validation set,  $C_p$  (AIC), BIC, or adjusted  $R^2$ . Or use the cross-validation method.
- 

Algoritmen testar alla möjliga kombinationer av variabler för att hitta den bästa modellen.

1. Börjar med en modell utan några variabler (bara medelvärdet).
2. För varje antal variabler kkk:
  - Testar alla modeller med exakt kkk variabler.
  - Väljer den modell som passar bäst (lägst RSS eller högst  $R^2$ ).
3. Väljer sedan den bästa modellen totalt med hjälp av t.ex. BIC, AIC, justerad  $R^2$  eller valideringsdata.

Poängen är att hitta en modell som både är bra på att förklara data och inte för komplicerad.

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

Citatet betyder att inga modeller är perfekta, de förenklar verkligheten. Men vissa modeller kan ändå vara användbara för att förstå, förklara eller förutsäga saker även om de inte är helt sanna. Modeller är verktyg och inte exakta kopior av verkligheten.

## 8 Självutvärdering

1. Vad tycker du har varit roligast i kunskapskontrollen?  
Att samla in data med hjälp av andra samt att lära mig R och hur det fungerar med olika fönster i R-studio.
2. Hur har du hanterat utmaningar? Vilka lärdomar tar du med dig till framtida kurser?  
Inom diskussion med andra och bra samarbete.
3. Vilket betyg anser du att du ska ha och varför?  
VG då mitt arbete uppfyller kriterierna för VG betyg.
4. Något du vill lyfta till Antonio?  
Tack för en spännande kurs och allt du har lärt oss.

## Appendix A

Källkoden för projektet finns tillgänglig på GitHub:

 GitHub-repository [https://github.com/HaniAbraksiea/ds24\\_R/tree/main](https://github.com/HaniAbraksiea/ds24_R/tree/main)

## Källförteckning

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning: With Applications in R* (2nd ed.). Springer. <https://www.statlearning.com/>

Statistiska centralbyrån. (2025). \*Fordon i trafik efter fordonsslag och bestånd\*. Hämtad från <https://www.statistikdatabasen.scb.se>

Wickham, H., & Çetinkaya-Rundel, M. (2023). *R for Data Science* (2nd ed.). O'Reilly Media. <https://r4ds.hadley.nz/>