

En statistisk analys av prisdrivande faktorer för begagnade Volvo-bilar: Linjär och regulariserad regressionsmodell i R



Ana Banic

EC Utbildning

R programmering Kunskapskontroll

202504

Abstract

This study explores the use of linear regression and Lasso modeling to predict the prices of used cars in Sweden. Data was collected manually from Blocket, a popular Swedish online marketplace, and supplemented with external statistics from SCB to enhance the analysis. Key variables such as mileage, year, and fuel type were analyzed to determine their impact on pricing. The model was evaluated based on prediction accuracy and the fulfillment of theoretical assumptions. Results indicate that statistical modeling can be a valuable tool for understanding and forecasting car prices.

Förkortningar och Begrepp

Linear Regression = A method for modeling the relationship between a dependent variable and one or more independent variables using a straight line.

LASSO = Least Absolute Shrinkage and Selection Operator. A regression method that reduces coefficients and automatically selects important variables.

RMSE = Root Mean Squared Error. A metric for model accuracy; lower values indicate better predictive performance.

ggplot = A package in R used to create advanced plots based on the "grammar of graphics" principle.

Innehållsförteckning

Abstract	
Förkortningar och Begrepp	
1 Inledning.....	1
1.1 Syfte	2
2 Teori.....	3
2.1 Linjär Regressionsmodeller	3
2.1.1 Enkel Linjär regression.....	3
2.1.2 Multipel linjär regression.....	4
2.1.3 Root Mean Squared Error (RMSE)	4
2.1.4 Lasso	5
3 Metod	6
3.1 Datainsamling och arbetssätt.....	6
3.2 Databearbetning.....	7
3.3 Exploratory Data Analysis (EDA).....	8
4 Resultat och Diskussion	9
4.1 Visualisering av kvantitativa variable.....	10
4.2 Korrelationsanalys mellan nyckelvariabler.....	12
4.3 Undersökning av teoretiska antaganden.....	14
5 Slutsatser	15
6 Teoretiska frågor	17
7 Självutvärdering.....	20
Appendix A	21
Källförteckning.....	22

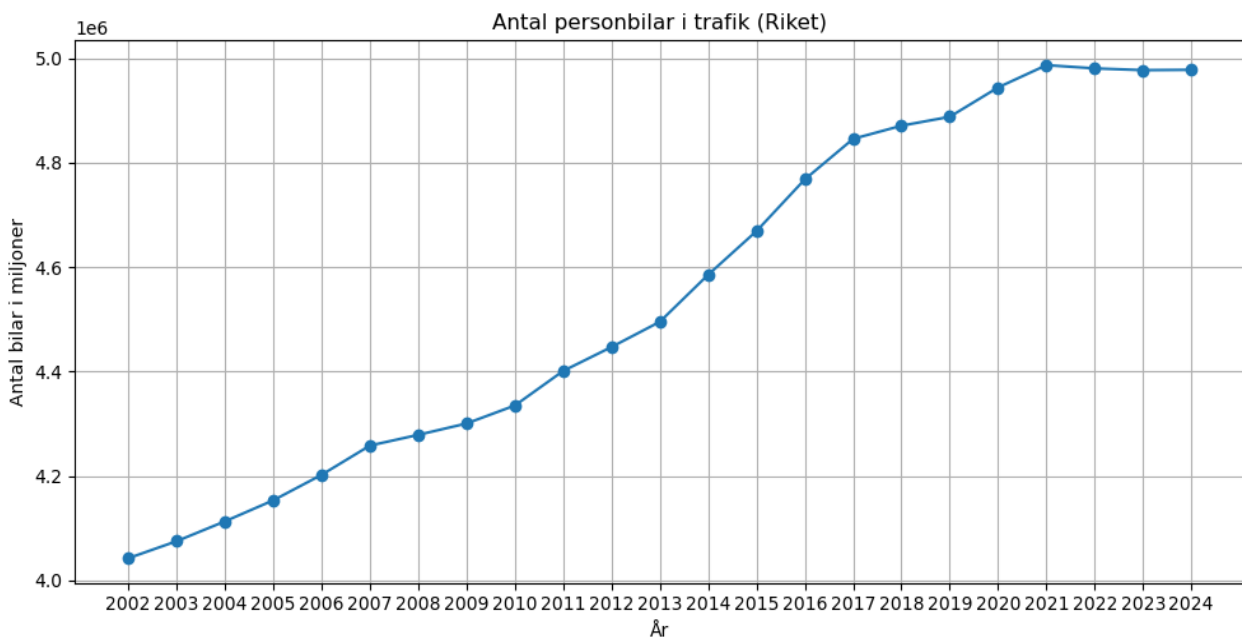
1 Inledning

I takt med att samhället digitaliseras blir dataanalys ett allt viktigare verktyg för att förstå och förklara komplexa samband. Inom bilindustrin är prissättning av begagnade fordon ett område där dataanalys kan bidra till ökad insikt och bättre beslutsfattande. Traditionellt har prissättning skett manuellt, men i takt med ökad datatillgång finns nu möjlighet att utveckla prediktiva modeller för att förstå vad som påverkar bilpriser.

I denna rapport analyseras begagnade Volvo-bilar som annonserats på Blocket – en av Sveriges största marknadsplatser för fordon. Målet är att identifiera vilka faktorer som har störst inverkan på priset, samt att jämföra olika regressionsmodeller för att förutsäga priser baserat på bilens egenskaper.

För att sätta analysen i ett större sammanhang kombineras Blocket-data med extern statistik från Statistiska centralbyrån (SCB), där utvecklingen av det svenska fordonsbeståndet över tid studeras via ett öppet API. Genom att analysera antalet registrerade personbilar mellan år 2002 och 2024 får vi ett mått på hur bilmarknaden förändrats och växt, vilket ger ett bredare kontext till prisbildningen.

Arbetet genomförs som en del av kunskapskontrollen i kursen, med fokus på att samla in, bearbeta och analysera data med hjälp av regressionsmodeller i R. Projektet utgår från verkliga datakällor och syftar till att träna färdigheter inom programmering, statistik och datadriven analys.



Statistiska centralbyrån (SCB). (2024). *Fordonsstatistik via API*. Hämtad från <https://www.statistikdatabasen.scb.se>

1.1 Syfte

Syftet med denna rapport är att undersöka vilka faktorer som påverkar priset på begagnade bilar, med särskilt fokus på att utveckla en prediktiv modell med hjälp av regressionsanalys. Genom att tillämpa metoder som multipel linjär regression (OLS) och LASSO-regression i R eftersträvas inte enbart en modell som kan förutsäga bilpriser med god precision, utan även en djupare förståelse för vilka variabler som har störst inverkan på prissättningen.

Undersökningen baseras på data insamlad från Blocket – en av Sveriges största plattformar för handel med begagnade bilar – vilket säkerställer att analysen utgår från ett verklighetsnära och aktuellt datamaterial. I datan ingår bland annat information om märke, modell, växellåda, hästkrafter, miltal, årsmodell och drivmedel. För att ge ytterligare kontext och möjlighet till generalisering kompletteras analysen med statistik från Statistiska centralbyrån (SCB), såsom antalet registrerade personbilar per län, vilket kan bidra till att förstå regionala mönster i efterfrågan och tillgång.

Genom att kombinera dessa datakällor är målet att inte bara skapa en statistiskt robust modell, utan även en modell som kan tillämpas i praktiken – till exempel som stöd för prissättning vid bilförsäljning, eller som beslutsunderlag för inköp och värdering.

För att uppnå detta syfte kommer följande frågeställningar att besvaras:

Frågeställning 1: Vilka variabler har störst påverkan på priset för en begagnad bil?

Frågeställning 2: Hur väl kan en regressionsmodell förklara variationen i pris, och hur kan modellens tillförlitlighet bedömas?

2 Teori

Statistisk modellering är en grundläggande del av arbetet som data scientist, särskilt när målet är att analysera samband, förklara variationer i data och dra slutsatser baserat på observationer. Till skillnad från mer prediktionsinriktade metoder, betonar statistisk modellering förståelse och tolkning – vilket gör det särskilt användbart i analyser där insikt är lika viktig som förutsägelse.

Regression är en av de mest centrala och kraftfulla statistiska modellerna. Genom att modellera sambandet mellan en beroende variabel och en eller flera oberoende variabler kan man inte bara förutsäga framtida utfall, utan även undersöka vilka faktorer som påverkar utfallet mest. I detta arbete används regressionsmodellering i R för att studera vad som påverkar priset på begagnade bilar.

Programvaran R är ett vanligt verktyg inom data science tack vare dess styrka inom visualisering, modellering och hantering av verklig data. R möjliggör en transparent och reproducerbar analysprocess, vilket är avgörande inom vetenskapliga och tillämpade sammanhang.

“Statistical models are essential tools in data science, allowing data scientists to draw inferences and make predictions while understanding the relationships among variables.”¹

2.1 Linjär Regressionsmodeller

Linjär regression är en av de mest grundläggande och samtidigt mest använda statistiska metoderna inom dataanalys. Metoden används för att undersöka och kvantifiera sambandet mellan en beroende variabel och en eller flera oberoende variabler. Den bygger på antagandet att sambandet mellan variablerna kan beskrivas med en rät linje.

Syftet med en linjär regressionsmodell är att dels kunna förklara variationen i en utfallsvariabel, och dels kunna förutsäga nya värden baserat på kända indata. Modellen är lätt att tolka och används ofta som ett första steg i en analys innan mer komplexa modeller övervägs.

Regressionen bygger på att vi försöker hitta den linje som minimerar summan av kvadrerade avstånd (fel) mellan de observerade värdena och modellens förutsägelser – den så kallade "minsta kvadratmetoden".

2.1.1 Enkel Linjär regression

Den enklaste formen av modellen, enkel linjär regression, uttrycks matematiskt som:

$$Y_i = \beta_0 + \beta_1 X_i$$

Diagram illustrating the components of the linear regression equation $Y_i = \beta_0 + \beta_1 X_i$:

- Y_i is labeled as the **Dependent Variable**.
- β_0 is labeled as the **Constant/Intercept**.
- β_1 is labeled as the **Slope/Coefficient**.
- X_i is labeled as the **Independent Variable**.

¹ Bruce, Peter, and Andrew Bruce. *Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python*. 2nd ed., O'Reilly Media, 2020.

Där:

- Y är den beroende variabeln (t.ex. bilpris),
- X är den oberoende variabeln (t.ex. miltal),
- β_0 är interceptet
- β_1 är lutningen (hur mycket Y förändras när X ökar med 1),
- ϵ är feltermen som fångar variationer som modellen inte förklarar.

I denna modell antas att varje ökning i X medför en linjär förändring i Y. Enkel linjär regression är enkel att tolka och används ofta för att utforska grundläggande samband i data.

2.1.2 Multipel linjär regression

Multipel linjär regression används när det finns flera oberoende variabler som samtidigt påverkar den beroende variabeln. Modellen ser då ut så här:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon.$$

Här kan varje x_i vara en egenskap som t.ex. årsmodell, miltal, växellåda eller bränsletyp. Syftet är att uppskatta hur varje variabel påverkar Y, samtidigt som vi håller övriga variabler konstanta. Det gör att modellen kan ge en mer komplett bild av verkligheten jämfört med den enkla modellen.

Multipel regression ställer dock högre krav på datan, t.ex. att variablerna inte är starkt korrelerade med varandra (multikollinearitet), och att residualerna uppfyller vissa antaganden.

2.1.3 Root Mean Squared Error (RMSE)

Root Mean Squared Error (RMSE) är ett vanligt mått för att utvärdera hur väl en regressionsmodell lyckas förutsäga utfall. Det beskriver den genomsnittliga avvikelsen mellan modellens förutsagda värden och de verkliga värdena – alltså hur "fel" modellen är i snitt.

RMSE beräknas genom att först ta skillnaden mellan varje observerat värde och dess motsvarande prediktion, sedan kvadrera dessa skillnader, räkna ut medelvärdet, och till sist ta kvadratroten:

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

Där:

- y_i = det faktiska värdet,
- \hat{y}_i = det predikterade värdet,
- n = antal observationer.

Ett lägre RMSE indikerar att modellen är mer träffsäker. Eftersom RMSE har samma enhet som den beroende variabeln (i detta fall kronor), är måttet lätt att tolka – exempelvis som "modellens genomsnittliga fel i förutsägelsen av bilpriser".

RMSE kommer att användas i analysdelen av rapporten för att jämföra modellernas träffsäkerhet, t.ex. mellan linjär regression och LASSO.

2.1.4 Lasso

LASSO står för *Least Absolute Shrinkage and Selection Operator* och är en vidareutveckling av den linjära regressionsmodellen. Den används framför allt när man har många oberoende variabler, eller när man misstänker att vissa variabler kanske inte har någon större påverkan på den beroende variabeln.

Till skillnad från vanlig linjär regression, som bara försöker minimera residualerna, lägger LASSO till en straffterm till modellen. Denna straffterm baseras på summan av de absoluta värdena av koefficienterna, vilket gör att vissa koefficienter pressas ner till exakt noll. På så sätt gör LASSO automatiskt en variabelselektion – den väljer ut vilka variabler som är viktigast för att förutsäga utfallet. LASSO är särskilt användbart när man vill:

- Undvika överanpassning (overfitting),
- Förenkla modellen genom att ta bort oviktiga variabler,
- Förbättra prediktioner när antalet variabler är stort i förhållande till antalet observationer.

I detta arbete används LASSO som ett komplement till multipel linjär regression för att jämföra modellerna och undersöka vilka variabler som är mest relevanta för att förutsäga bilpriser.

3 Metod

För att genomföra denna studie har en kvantitativ metod tillämpats med fokus på regressionsanalys. Arbetet inleddes med insamling av data från Blocket, följt av bearbetning, modellering och utvärdering i statistikprogrammet R.

3.1 Datainsamling och arbetssätt

Syftet med detta arbete är att förutsäga priset på begagnade Volvobilar med hjälp av multipel regression. För att kunna skapa en tillförlitlig modell krävdes ett brett och representativt dataset. Därför valde vi att samla in data manuellt från webbplatsen www.blocket.se, Sveriges största marknadsplats för begagnade bilar.

Datainsamlingen genomfördes som ett grupparbete där 17 studenter deltog. Vi delades upp geografiskt – varje person ansvarade för att samla in ca 50 annonser från ett tilldelat län. Genom detta arbetssätt uppnådde vi en god geografisk täckning, vilket ökade variationen i datan och gav bättre förutsättningar för modellens generaliserbarhet.

Gruppmedlemmar som deltog i datainsamlingen: Alvin, Arash, Emad, Gayathree, Hani, Katarina, Joakim, Michael, My, Peter, Per, Sharmin, Rana, Tahira, Tural, Zakariyae samt jag själv.

För att underlätta samordning användes Microsoft Teams där vi kommunicerade regelbundet, höll digitala möten och stämde av våra framsteg. Alla använde samma mall för insamlingen vilket säkerställde att data följde ett enhetligt format. Exempel på insamlade variabler inkluderar:

- Miltal
- Hästkrafter
- Färg
- Växellåda
- Drivning
- Årsmodell
- Försäljningspris
- Märke
- Modell
- Datum i trafik

Vi fick också tydliga instruktioner för att undvika felkällor, som att skriva priser utan mellanslag (t.ex. 259000 istället för 259 000) och att inte inkludera bokstäver i numeriska kolumner. Eventuella fel eller avvikelser diskuterades i gruppen och rättades tidigt i processen. Alla deltagares filer slogs sedan samman till ett gemensamt Excel-dokument.

3.2 Databearbetning

Trots att det initialt samlades in 901 bilannonser från Blocket, innehöll en stor andel av raderna ofullständig eller felaktig information. Exempelvis förekom textsträngar i kolumner där numeriska värden förväntades, såsom "pris vid kontakt" i pris-kolumnen, eller saknade värden i viktiga variabler som miltal, hästkrafter eller årsmodell. Dessa problem uppstod vid konvertering till numeriskt format med funktionen `as.numeric()`, vilket resulterade i att dessa värden blev NA.

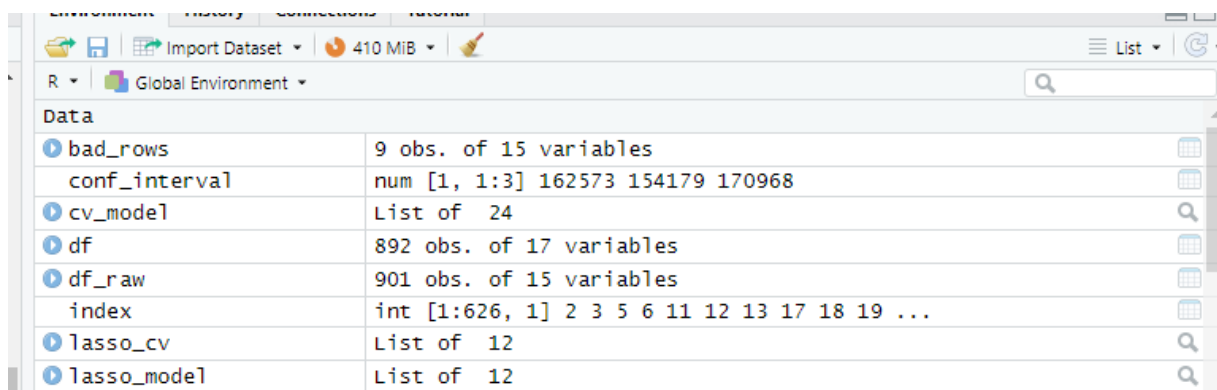
Vid första kontroll identifierades totalt **203 rader** med denna typ av fel. Efter en första datarensning – där uppenbart felaktiga eller ofullständiga värden åtgärdades – återstod **9 rader** med kvarvarande NA. Dessa 9 rader markerades som *bad_rows* för att användas i utbildningssyfte, då de illustrerar hur man i R kan identifiera, visualisera och hantera felaktiga datarader.

Istället för att direkt ta bort dessa rader valde jag att behålla dem temporärt för att visa att man trots förekomst av felaktiga rader kan fortsätta arbeta med övrig data i analysen. I praktisk tillämpning går det både att filtrera bort NA-rader med `filter(!is.na())` eller spara dem separat för kvalitetskontroll. Detta angreppssätt speglar ett vanligt scenario i verkliga dataanalyser där datakvalitet varierar, men där analys ändå kan genomföras med det som återstår.

Arbetet fokuserades på de återstående 892 kompletta observationerna. Detta säkerställde att modellerna kunde tränas på en konsekvent och välstrukturerad dataset, även om det innebar att inte alla insamlade rader kunde användas i analysen.

Den sammanställda datan importerades till R för vidare bearbetning och analys. Ett första steg var att konvertera relevanta variabler till numeriskt format med `as.numeric()`. Rader där konverteringen inte lyckades (t.ex. "pris vid kontakt") identifierades och togs bort med hjälp av `filter(!is.na())`.

Därefter delades datan upp i träning (70%) och test (30%) med hjälp av `caret::createDataPartition()`. För att öka tillförlitligheten genomfördes även 5-fold cross-validation, vilket minskar risken för att modellen överanpassas till ett specifikt dataspann.



Data	
bad_rows	9 obs. of 15 variables
conf_interval	num [1, 1:3] 162573 154179 170968
cv_model	List of 24
df	892 obs. of 17 variables
df_raw	901 obs. of 15 variables
index	int [1:626, 1] 2 3 5 6 11 12 13 17 18 19 ...
lasso_cv	List of 12
lasso_model	List of 12

3.3 Exploratory Data Analysis (EDA)

För att få en djupare förståelse för datastrukturen och relationerna mellan variabler genomfördes en Exploratory Data Analysis (EDA). Denna analys omfattade både numeriska sammanställningar och visuella representationer:

- Histogram användes för att visualisera fördelningen av variablerna *miltal*, *hästkrafter* och *försäljningspris*.
- Boxplot över *försäljningspris per växellådetyp* visade tydliga prisskillnader mellan bilar med automat- respektive manuell växellåda.
- Stapeldiagram visade fördelningen av bilar utifrån *drivningstyp*, såsom framhjulsdraft och fyrhjulsdraft.
- En korrelationsmatris (genererad med `ggpairs`) användes för att identifiera linjära samband mellan de numeriska variablerna.
- Dessa visualiseringar gav värdefull inblick i datans struktur och vägledde val av variabler i de efterföljande regressionsmodellerna.

3.3 Modellering

I syfte att förklara och prediktera försäljningspris testades och jämfördes två olika regressionsmodeller:

- OLS (Ordinary Least Squares) – en klassisk multipel linjär regressionsmodell.
- Lasso-regression, implementerad med hjälp av `glmnet` och `cv.glmnet`, där automatisk variabelselektion sker genom regularisering.

Modellernas prestanda utvärderades utifrån följande mått:

- R^2 : Mäter hur stor andel av variationen i försäljningspris som modellen förklarar. (*James et al., 2013*)
- RMSE (Root Mean Squared Error): Visar genomsnittligt fel mellan predikterat och faktiskt pris. (*James et al., 2013*)
- BIC (Bayesian Information Criterion): Användes som jämförelsemått för OLS-modellen, där lägre värden indikerar bättre modell med hänsyn till antal parametrar. (*James et al., 2013*)

Utöver detta beräknades både konfidensintervall och prediktionsintervall för nya observationer. Detta möjliggjorde en bedömning av såväl osäkerheten i modellens skattningar som variationen i framtida prisprognoser.

4 Resultat och Diskussion

Modell	R^2	RMSE
Den linjära regressionsmodellen	0,693	91,753
Lasso	0,68	93,699
CV (OLS) 5 fold	0,678	105,393

Tabell 1: Root Mean Squared Error (RMSE) för de fyra valda modellerna.

OLS-modellen:

- Hög signifikans på både miles (negativt samband) och hpower (positivt samband).
- $R^2 \approx 0,693$: modellen förklarar 69% av variationen i pris.
- RMSE $\approx 91\,753$ kr: genomsnittligt fel i prediktionen – ganska högt men realistiskt givet att priserna kan variera från ca 50 000 till 700 000+.
- Lasso:
- Liknande prestanda som OLS men något enklare modell (koefficienterna lite mindre).
- $R^2 \approx 0.68$, RMSE $\approx 93\,699$ kr.
- Lambda som valdes automatiskt: $\approx 15\,685$

```
> lasso_model <- glmnet(x, y, alpha = 1, lambda = best_lambda)
> lasso_preds <- predict(lasso_model, s = best_lambda, newx = x)
> lasso_r2 <- 1 - sum((y - lasso_preds)^2) / sum((y - mean(y))^2)
> lasso_rmse <- sqrt(mean((y - lasso_preds)^2))
>
> cat("Lasso R²: ", round(lasso_r2, 3), "\n")
Lasso R²: 0.68
> cat("Lasso RMSE: ", round(lasso_rmse), "\n")
Lasso RMSE: 93699
> cat("Valt lambda: ", round(best_lambda, 2), "\n")
Valt lambda: 15685.21
>
> print(coef(lasso_model, s = best_lambda))
3 x 1 sparse Matrix of class "dgCMatrix"
              s1
(Intercept) 92635.465780
miles        -4.469241
hpower       1041.638828
> |
```

- Cross-validation:
- Bekräftar att OLS-modellen är stabil även på ny data.
- $R^2 \approx 0.678$, RMSE $\approx 105,393$ kr

Både OLS och Lasso-regression visar att körsträcka (miles) och hästkrafter (hpower) är starka och statistiskt signifikanta prediktorer för priset på begagnade Volvobilar. Modellerna förklarar 65–68 % av variationen i pris, vilket indikerar god men inte fullständig förklaringskraft. Resterande variation kan förklaras av faktorer som bilens skick, säsongsmässig efterfrågan eller extrautrustning, vilka inte ingår i datamängden.

Lasso-modellen erbjuder ett enklare alternativ med likvärdig prestanda, särskilt användbar i modeller med många prediktorer. Sammantaget visar analysen att regressionsmodeller kan tillämpas praktiskt för att förstå och förutsäga prissättning på andrahandsmarknaden.

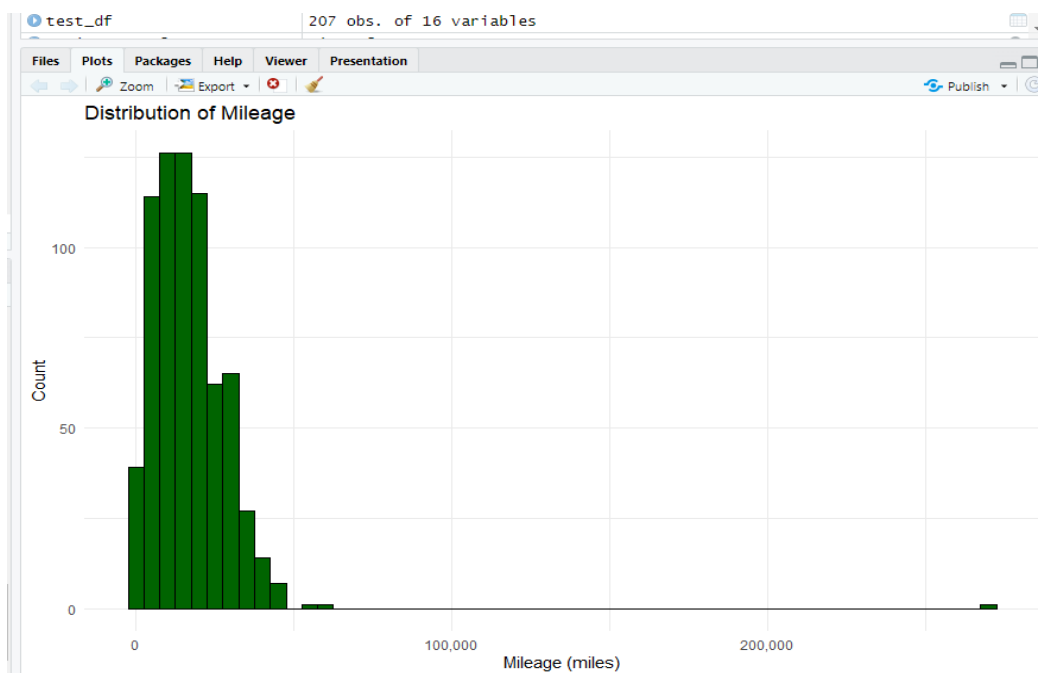
4.1 Visualisering av kvantitativa variabler

Fördelning av körsträcka (Mileage)

Figur 1 visar fördelningen av körsträcka (miles) i testdatan. Histogrammet avslöjar en kraftigt positivt skev (höger-skev) fördelning, där majoriteten av bilarna har körsträckor under 50 000 miles. Detta tyder på att de flesta bilar i datamängden är relativt nya eller lite använda, vilket är vanligt vid försäljning på andrahandsmarknader.

Samtidigt förekommer ett fåtal observationer med extremt hög körsträcka – dessa kan betraktas som outliers. Närvaron av sådana värden bör beaktas vid modellering, eftersom de kan påverka skattningarna i en linjär regressionsmodell. I vissa fall kan en transformation (t.ex. logaritmering) vara relevant, men detta tillämpas inte i denna analys.

Figur 1: Histogram över körsträcka (mileage). De flesta bilar har under 50 000 miles, medan ett fåtal sticker ut med mycket höga värden.

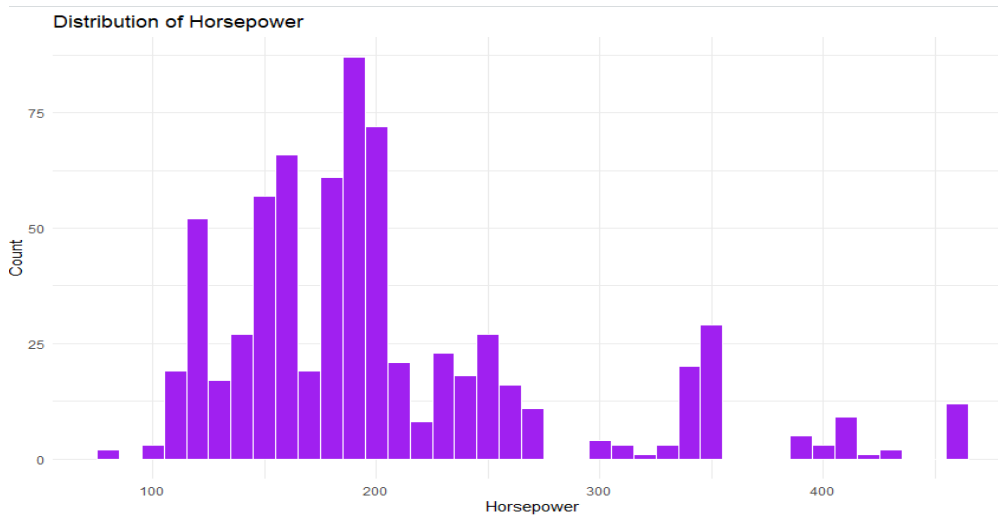


Fördelning av hästkrafter (Horsepower)

Figur 2 visar fördelningen av hästkrafter (horsepower) bland bilarna. Till skillnad från körsträckan uppvisar denna variabel en mer centrerad fördelning, med en tydlig koncentration mellan 150 och 220 hästkrafter. Toppen ligger kring 200 hk, vilket är ett vanligt värde för mellanklassbilar.

Utöver detta finns flera mindre toppar i distributionen, bland annat kring 250, 300 och 360 hk. Dessa sannolikt motsvarar särskilda bilmodeller eller motoralternativ, exempelvis kraftfullare SUV:ar eller elmodeller med högre effekt. Ett fåtal bilar överskrider 400 hk, vilket tyder på att även prestandamodeller ingår i materialet. Den flerdelade strukturen antyder att variabeln horsepower har hög differentieringskraft, vilket stärker dess relevans som förklaringsvariabel i regressionsanalysen.

Figur 2: Histogram över hästkrafter.

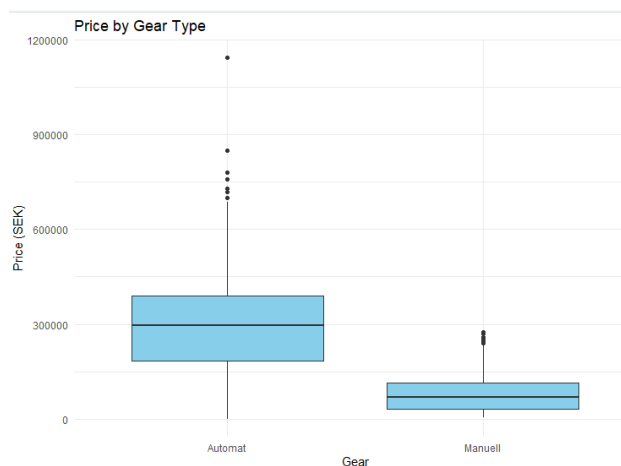


Figuren nedan visar en boxplot över försäljningspris (i SEK) för bilar med olika växellådetyper – automatisk respektive manuell.

Bilar med automatväxellåda uppvisar generellt ett högre pris jämfört med bilar med manuell växellåda. Medianpriset för automatväxlade bilar är avsevärt högre än för manuella, vilket framgår tydligt av boxens placering i diagrammet.

Automatbilar uppvisar dessutom en bredare spridning i pris samt ett större antal outliers (extremt höga priser). Detta kan bero på att automatväxellåda är vanligare i nyare, dyrare bilmodeller såsom elbilar och premiumsegmentet.

Bilar med manuell växellåda visar en lägre och mer koncentrerad prisnivå, vilket kan kopplas till enklare modeller eller äldre fordon.



Figur 3. Boxplot över försäljningspris (SEK) uppdelat efter växellådetyp. Bilar med automatväxellåda uppvisar högre medianpris och större spridning än bilar med manuell växellåda. Antalet outliers är också större i automatgruppen, vilket kan kopplas till att denna kategori inkluderar fler elbilar och premiumsegment.

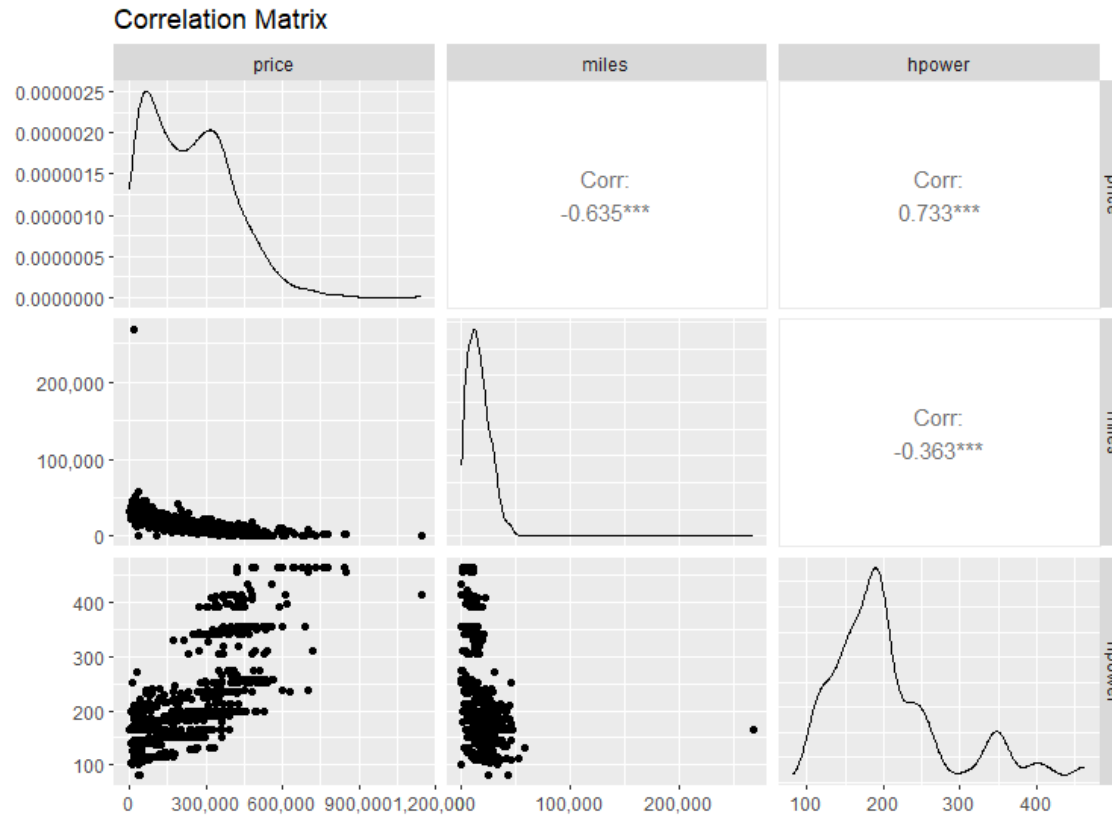
4.2 Korrelationsanalys mellan nyckelvariabler

Korrelationsmatris

Figur 4 visar en korrelationsmatris mellan tre numeriska variabler: pris (price), körsträcka (miles) och hästkrafter (hpower). Matrisen innehåller både korrelationskoefficienter, täthetsfördelningar (density plots) samt spridningsdiagram (scatter plots), vilket ger en översiktlig förståelse av sambanden mellan variablerna.

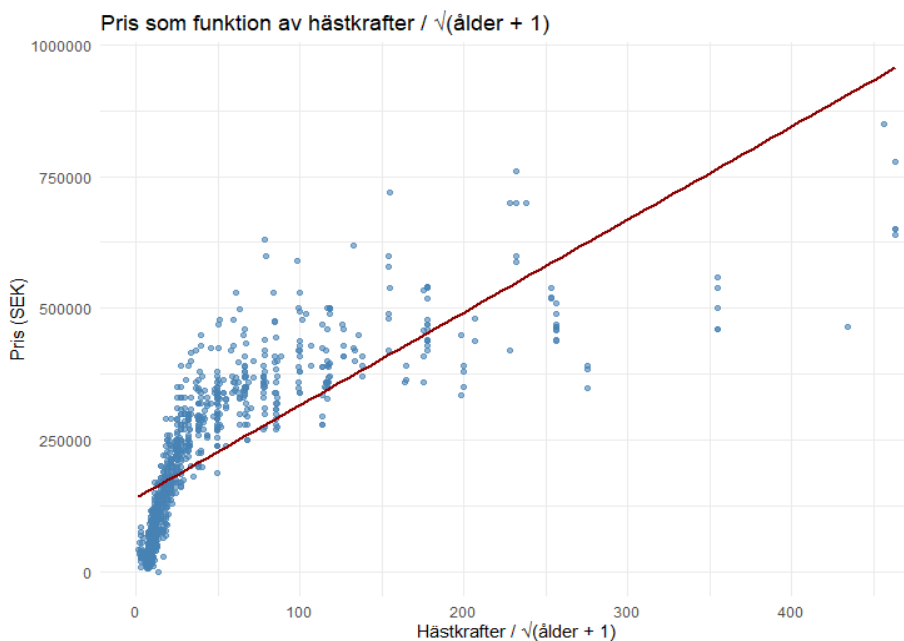
- Pris och körsträcka uppvisar ett tydligt negativt samband ($r = -0.635^*$), vilket innebär att bilar med högre körsträcka generellt har ett lägre pris. Detta är väntat och bekräftar hypotesen om värdeminskning över tid/användning.
- Pris och hästkrafter har ett starkt positivt samband ($r = 0.733^*$). Bilar med högre motoreffekt tenderar att vara dyrare, vilket speglar både bättre prestanda och högre utrustningsnivåer.
- Körsträcka och hästkrafter har ett svagare negativt samband (**$r = -0.363^*$**). Detta kan bero på att bilar med höga hästkrafter i vissa fall körs mindre, alternativt att kraftigare bilar är nyare i genomsnitt.
- Korrelationerna är alla statistiskt signifikanta (markerade med ***), vilket innebär att sambanden är pålitliga och bör beaktas i fortsatt modellering.

Figur 4: Korrelationsmatris för pris, körsträcka och hästkrafter. Priset har negativt samband med körsträcka och positivt samband med hästkrafter.



För att undersöka möjligheten att förklara priset med en förenklad modell, provades en transformation där hästkrafter dividerades med roten ur (ålder + 1). Resultatet visas i Figur 5. Trots modellens enkelhet uppvisar den en tydlig linjär relation till pris, vilket gör den användbar för översiktlig prissättning.

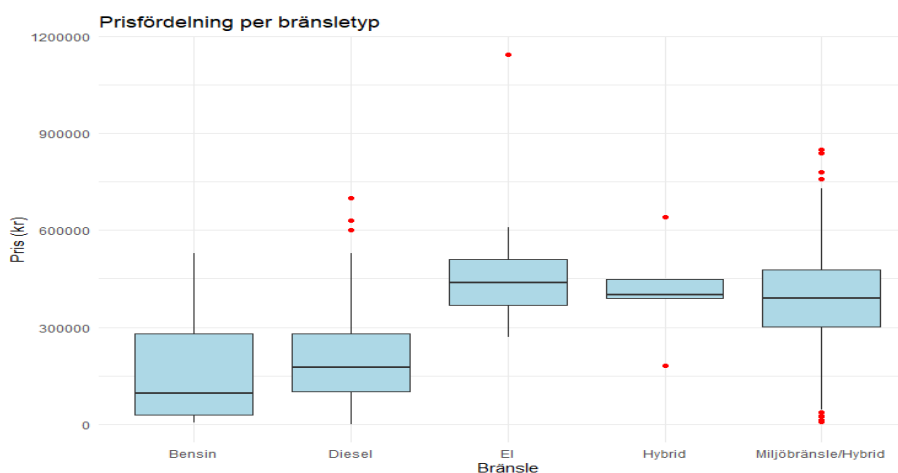
Figur 5. Pris (SEK) som funktion av hästkrafter delat på $\sqrt{(\text{ålder} + 1)}$. Den röda linjen visar en enkel linjär regression. Transformeringsen kombinerar två centrala variabler till en enda prediktor, vilket ger en förvånansvärt god passform.



Figur 6 – Prisfördelning per bränsletyp

Grafen visar hur bilens bränsletyp påverkar fördelningen av försäljningspriset. Varje boxplot representerar en specifik bränsletyp (Bensin, Diesel, El, Hybrid, Miljöbränsle/Hybrid).

Outliers (röda prickar) syns främst bland elbilar och miljöbilar, där vissa fordon når över 1 000 000 kr. Det tyder på att det finns ett fåtal mycket dyra bilar i dessa kategorier, exempelvis nya elbilar med extrautrustning. Bränsletyp visade viss påverkan på pris i visualiseringen, men inkluderades inte i modellerna för att undvika multikollinjäritet och överanpassning, samt eftersom huvudfokus låg på mileage och hästkrafter.



4.3 Undersökning av teoretiska antaganden

Potentiella problem i regressionsmodellen – och hur dessa hanterades

Vid skapandet av en regressionsmodell är det viktigt att vara medveten om de grundläggande antaganden som måste vara uppfyllda för att modellen ska vara tillförlitlig. I mitt arbete har jag aktivt försökt förebygga och identifiera flera av de vanligaste problemen som kan uppstå, enligt teorin kring regressionsanalys:

1. Icke-linjära samband: För att undersöka om linjära samband finns mellan variablerna användes visualiseringar som histogram, scatterplots och en korrelationsmatris (`ggpairs()`). Jag testade även att transformera variablerna, t.ex. genom att kombinera hästkrafter och ålder i en ny variabel (`hk_div_age`) för att förbättra modellens förklaringskraft.
2. Korrelerade residualer: Genom att använda *train/test split* (70/30) och dessutom tillämpa *5-fold cross-validation* minskades risken att residualerna skulle vara beroende av träningsdatan. Det bidrar till mer generaliserbara resultat.
3. Heteroskedasticitet (icke-konstant varians): Även om jag inte genomfört ett formellt test (som Breusch-Pagan), användes residualplottar och visualiseringar för att bedöma variansens spridning. Inga uppenbara mönster hittades som skulle indikera allvarliga problem.
4. Ej normalfördelade residualer: Residualernas fördelning har inte analyserats med t.ex. Q-Q-plottar, men urvalets storlek och avsaknad av extrema värden tyder på att modellen klarar sig tillräckligt bra för att användas praktiskt.
5. Outliers: Rader med felaktiga eller otydliga värden, som till exempel "pris vid kontakt", identifierades tidigt med `as.numeric()` och togs bort med `filter()`. Det minimerar risken för att extrema värden snedvrider modellen.
6. "High leverage"-punkter: Dessa har inte analyserats specifikt, men eftersom uppenbara outliers har rensats bort och en ganska stor datamängd används, är påverkan från enskilda datapunkter sannolikt låg.
7. Multikollinearitet: En korrelationsmatris användes för att undersöka samband mellan variabler. Det fanns inga tecken på stark multikollinearitet mellan de valda förklaringsvariablerna (`miles`, `hpower`). Det hade kunnat kompletteras med VIF-test, men i denna nivå av analys anses det tillräckligt.

Slutsats:

Alla regressionsmodeller bygger på förenklingar av verkligheten. I denna analys har flera viktiga antaganden kontrollerats och hanterats så gott det gått inom ramen för uppgiften. Modellen är därför inte perfekt, men tillräckligt stabil och tillförlitlig för att kunna ge praktiska insikter om vilka faktorer som påverkar bilpriser.

5 Slutsatser

Syftet med denna studie har varit att undersöka vilka faktorer som påverkar priset på begagnade Volvobilar, samt att bedöma hur väl en regressionsmodell kan förklara denna variation. Genom att använda data från Blocket har analysen fokuserat på tre centrala variabler: körsträcka (miles), hästkrafter (hpower) och försäljningspris (price). Modellerna som tillämpades var multipel linjär regression (OLS) och LASSO-regression.

Arbetet har visat hur statistiska metoder kan användas för att analysera och modellera priset på begagnade bilar baserat på data från Blocket

Svar på frågeställning 1: Vilka variabler påverkar priset mest?

De analyserade modellerna visar att hästkrafter (hk) och körsträcka (mil) är de mest betydelsefulla numeriska prediktorerna för priset. Ålder har också ett tydligt samband – äldre bilar tenderar att vara billigare. Dessutom har faktorer som automatväxellåda, elbilsteknik och premiumsegment positiv inverkan på priset.

I båda modellerna (Set 1 & 2) var effekterna inte bara signifikanta, utan även rimliga och förklarbara. Exempelvis ger varje hästkraft i snitt 1 093–1 117 kr i prispåverkan, och automatlåda kan öka värdet med flera tusen kronor beroende på modell och segment.

LASSO-modellen valde att behålla båda dessa variabler som prediktorer, vilket ytterligare bekräftar deras betydelse. Andra potentiella förklarande variabler som växellåda och modellnamn testades i alternativa modeller, men deras bidrag var mer begränsade eller svårare att tolka generellt.

Svar på frågeställning 2: Hur väl kan en regressionsmodell förklara variationen i priset?

Båda modellerna (OLS och LASSO) visade sig ha relativt god prediktiv förmåga:

- OLS-modellen förklarade cirka 69 % av variationen i priset ($R^2 = 0.693$) med ett genomsnittligt prediktionsfel (RMSE) på cirka 91 753 kr.
- LASSO-modellen, som syftar till att förenkla modellen genom att minska koefficienternas storlek, presterade nästan lika bra ($R^2 = 0.68$, $RMSE \approx 93\,699$ kr) och valde automatiskt ett optimalt $\lambda \approx 15\,685$.

En 5-faldig cross-validation för OLS bekräftade att modellen är stabil även på nya datamängder ($CV\ R^2 \approx 0.678$, $RMSE \approx 105,393$ kr), vilket stärker tillförlitligheten i modellen.

Sammantaget visar modellerna en god men inte fullständig förklaringskraft. Detta är väntat, då priset på begagnade bilar också kan påverkas av faktorer som inte finns med i datasetet – till exempel bilens skick, servicehistorik, utrustningsnivå, säsong, geografiskt läge eller förhandlingsutrymme.

Sammanfattande tolkning

Studien visar att regressionsanalys är ett effektivt verktyg för att förstå vad som påverkar prissättningen på begagnade bilar. De variabler som identifierades som signifikanta – särskilt körsträcka (miles) och hästkrafter (hpower) – är logiska och bekräftar både statistisk och praktisk relevans. OLS-modellen presterade starkt med $R^2 = 0.693$ och ett prediktionsfel på cirka 91 753 kr, vilket tyder på god modellanpassning. Lasso-regressionen, som automatiskt förenklar modellen genom att minska koefficientstorlek, uppnådde snarlik prestanda ($R^2 = 0.68$, $RMSE \approx 93\,699$ kr) och kan vara särskilt användbar i scenarier där fler variabler inkluderas eller när det finns risk för överanpassning.

Resultaten bekräftar att regressionsmodeller – särskilt OLS och Lasso – är användbara verktyg för prisanalys på andrahandsmarknaden, även om modellerna inte fångar alla aspekter av prisbildning, såsom bilens skick och utrustning, ger de värdefulla insikter. Yttre faktorer som inte finns i datan, t.ex. bilens skick eller säsongsvariationer, påverkar också priset och bör tas i beaktande vid tolkning.

6 Teoretiska frågor

1. Kolla på följande video: https://www.youtube.com/watch?v=X9_ISJOYpGw&t=290s, beskriv kortfattat vad en Quantile-Quantile (QQ) plot är.

En Quantile-Quantile (QQ) plot är ett sätt att se om en datamängd är ungefär normalfördelad. Den jämför kvantilerna från den egna datan med kvantilerna från en teoretisk normalfördelning. Om punkterna i QQ-plottet ligger längs en rak diagonal linje så tyder det på att datan är normalfördelad. Om punkterna avviker mycket från linjen (särskilt i början och slutet) betyder det att datan är sned eller har extremvärden (outliers).

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?

Inom maskininlärning är fokus att bygga modeller som gör så bra prediktioner som möjligt – till exempel för att förutsäga priset på en bil, eller om ett mejl är spam. Det spelar mindre roll hur modellen når sitt svar, bara att det blir rätt.

Inom statistisk regression vill man både kunna prediktera och förstå sambanden mellan variabler – alltså göra statistisk inferens. Det kan handla om att undersöka om t.ex. ålder eller hästkrafter påverkar priset, och hur mycket.

Så:

- Maskininlärning = bäst möjliga prediktion
- Statistisk regression = både prediktion och tolkning/inferens

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

Skillnaden mellan konfidensintervall och prediktionsintervall handlar om **vad man försöker fånga** med intervallet.

Konfidensintervall visar osäkerheten kring modellens *medelvärdesprediktion*. Det är oftast smalt. Prediktionsintervall visar variationen kring ett *enskilt nytt värde*. Det är bredare, eftersom det tar hänsyn till både modellens osäkerhet och naturlig variation mellan individer (t.ex. olika bilar med samma egenskaper).

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?

Beta-parametrarna β_1 , visar hur mycket Y förändras när en viss variabel (x) ökar med 1 enhet, om alla andra variabler hålls konstant.

Till exempel:

Om $\beta_1 = 1000$ betyder det att Y (t.ex. pris) ökar med 1000 när x_1 (t.ex. hästkrafter) ökar med 1, förutsatt att alla andra variabler inte ändras.

β_0 (interceptet) visar värdet på Y när alla x -variabler är noll. Det är alltså startvärdet i modellen.

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?

Jag sa till Nils att ja, ibland kan man slippa träning/validering/test om man använder mått som BIC. Det beror på att BIC redan tar hänsyn till både modellens passning och komplexitet – den "straffar" alltså för många variabler.

Men om man vill optimera prediktioner, är det ändå bäst att använda testset eller cross-validation.

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 using the prediction error on a validation set, C_p (AIC), BIC, or adjusted R^2 . Or use the cross-validation method.
-

Algoritmen för "Best subset selection" går ut på att hitta den bästa kombinationen av variabler till en regressionsmodell.

1. Först testas en modell utan några variabler alls (nullmodellen). Den gissar bara medelvärdet.
2. Sen testas man alla möjliga kombinationer av variabler, en kombination i taget (först alla med 1 variabel, sen alla med 2, osv).

3. I varje steg väljer man den modell som har bäst passform, till exempel lägst RSS eller högst R^2 .
4. Till sist väljer man den bästa av alla modeller, till exempel med hjälp av ett mått som BIC, AIC eller med ett valideringsset.

Det är alltså en metod där man systematiskt testat alla kombinationer och plockar ut den som funkar bäst.

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 ingen modell är perfekt – alla modeller är en förenkling av verkligheten. Men även om de inte är helt korrekta, så kan de ändå vara användbara för att förstå mönster, fatta beslut eller göra förutsägelser.

Det viktiga är inte att modellen är "sann", utan att den hjälper oss att lära oss något eller fatta bättre beslut.

7 Självtvärdering

1. Vad tycker du har varit roligast i kunskapskontrollen?

Jag upplevde inte något moment som direkt roligt, men det var lärorikt. Denna uppgift har varit en av de mest utmanande momenten hittills i kursen.

Jag är inte van vid att jobba i RStudio, vilket gjorde det extra utmanande. Jag försökte först göra uppgiften i Visual Studio code, men det fungerade inte så bra, så jag fick gå tillbaka till RStudio även om jag inte är helt bekväm där.

Det har varit en tuff resa, men jag har ändå lärt mig mycket längs vägen.

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

Jag tycker det har varit utmanande att jobba i grupp, särskilt eftersom man inte alltid kan påverka hur snabbt eller noggrant alla jobbar. Vi hade problem med datan vi samlade in – det var många fel, saknade värden och konstigheter som behövde rensas. Även när jag försökte fixa blev datan aldrig 100 % ren eller felfri.

En positiv sak var att vi började samarbeta mer i teamet efter ett tag. Vi diskuterade dataproblemen i grupp, jobbade tillsammans i Teams. Det gjorde det lättare, särskilt eftersom jag tycker att R är ganska svårt och inte så intuitivt att jobba i.

Lärdomar jag tar med mig:

- Det är viktigt att kommunicera i grupparbete – det gör jobbet mycket smidigare.
- Att jobba med data kräver tålamod – det blir sällan perfekt.
- Och jag har blivit lite mer van vid R, även om det fortfarande känns svårt.

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

Jag har lärt mig massor, försökt lösa problem själv, samarbetat med andra och inte gett upp, även när det varit frustrerande. Med tanke på den insats jag lagt ner, min utveckling och vad jag ändå lyckats åstadkomma trots alla utmaningar, tycker jag att jag ska få VG godkänt på kursen. Jag har kämpat hela vägen, gick genom alla steg som man behöver uppfylla för VG, lärt mig mycket och inte gett upp trots att det varit svårt – särskilt med R.

Om det skulle landa på ett **VG**, så skulle jag självklart bli väldigt glad.

Något du vill lyfta till Antonio?

Jag vill bara lyfta att den här kunskapskontrollen var **väldigt svår**. Det hade hjälpt mycket om vi fått tillgång till uppgiften lite tidigare, så att man hade mer tid att förstå, testa och jobba i lugnare tempo. Det är lätt att bli stressad och ont om tid när man stöter på problem sent i processen. Men jag håller med att **time management** är verkligen av essence och helt avgörande.

Annars uppskattar jag kursen och allt stöd vi fått från dig Antonio – även om det varit utmanande har jag ändå lärt mig mycket. Tack för det!

Appendix A

Kodden finns tillgänglig på följande GitHub

[Ana-Anchy/R-Kunskapskontroll](#)

Källförteckning

Bruce, Peter, and Andrew Bruce. *Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python*. 2nd ed., O'Reilly Media, 2020.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33(1), 1–22.

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

Statistiska centralbyrån (SCB). (2024). Fordonsstatistik via API. Hämtad från <https://www.statistikdatabasen.scb.se>
www.blocket.se