

# Vad påverkar priset på den begagnade bilen?

En statistisk analys över vilka egenskaper som påverkar privatsäljares prissättning av bilar på Blocket



Tova Thorén  
EC Utbildning  
Kunskapskontroll - R  
2024-04-25

## Abstract

With rising prices in the automotive market, consumers are turning to the used car market, driven by factors like inflation and supply disruptions. This study examines what affects the pricing of used cars in Sweden using data from Blocket. Through linear regression, we found that car age, mileage, horsepower, fuel type and transmission type significantly influence prices. However, mileage had only a small effect, indicating a complex relationship with age. The final model accurately predicts car prices for new Blocket listings with a 95% confidence level, providing insights into the dynamics of used car pricing.

## Innehållsförteckning

Abstract .....	2
1 Inledning.....	1
1.1 Problemformulering.....	2
1.1.1 Definitioner och avgränsningar .....	2
2 Teori.....	3
2.1 Linjär Regression .....	3
2.1.1 Skattning av regressionskoefficienterna.....	3
2.1.2 Modellanpassning och utvärdering .....	4
2.1.3 Hypotesprövning .....	5
2.1.4 Variabelselektion och regulariseringstekniker .....	5
2.1.5 Modellantaganden och potentiella problem.....	5
2.1.6 Variabeltransformering .....	6
2.1.7 Modellval och utvärdering.....	6
3 Metod .....	8
3.1 Datainsamling .....	8
3.2 Databearbetning .....	8
3.2.1 Hantering av saknade värden .....	9
3.2.2 Explorativ dataanalys (EDA).....	9
3.2.3 Logtransformering .....	10
3.2.4 Undersökning av kollinearitet (numeriska variabler) .....	11
3.2.5 Dataförberedning .....	12
4 Resultat och Diskussion.....	13
4.1 Träning av fem regressionsmodeller .....	13
4.2 Modellutvärdering och val av modell .....	15
4.3 Hypotesprövning och statistisk inferens.....	16
4.3.1 Prisprediktioner på nya Blocket-annonser .....	17
5 Slutsatser .....	18
Appendix A .....	19
Källförteckning.....	22

## 1 Inledning

Den som har varit på jakt efter en ny bil har förmodligen inte undgått de senaste årens skenande priser på marknaden. Enligt olika mätningar beräknas de största prisökningarna för nya bilar ha skett mellan 2018 och 2023, där bland annat effekterna av Covid-19, brist på råvarukomponenter och mer påkostad utrustning i bilarna läggs till som bidragande faktorer till utvecklingen (KVD, 2023). Samtidigt befinner vi oss i en lågkonjunktur med inflation och höga räntor, vilket minskar köpkraften hos många. Detta har resulterat i att allt färre har råd med nya bilar och i stället söker sig till begagnatmarknaden, som historiskt sett inneburit bilköp till förmånliga priser. Tidigare kunde man grovt uppskatta priset på en begagnad bil baserat på hur länge den använts, då en ny bil anses förlora en betydande del av sitt värde så fort den körts ut från bilhandlaren (Bilpriser.se, 2022).

Men med de senaste årens stigande nybilspriser och en ökad efterfrågan på begagnatmarknaden har utbudet minskat vilket har lett till prisökningar även för begagnade bilar. En svag svensk krona ses som ytterligare en förklaring till prisökningarna då bilexporten har ökat. Under de senaste åren har begagnade bilar sålts till priser som är lika höga som för nya bilar, och i vissa fall till och med högre. Enligt bilprisindex (BPI) var det genomsnittliga priset på begagnade bilar cirka 10% högre år 2023 jämfört med året innan (Bilpriser.se, 2023).

Prisökningarna kan även ses som en möjlig effekt i statistiken över antalet nyregistrerade bilar. Mellan 2019 och 2020 minskade antalet nyregistrerade bilar i Sverige med över 20%. Med undantag för en viss ökning 2021 fortsätter sedan en minskande trend i antalet nyregistreringar. Dessutom minskade antalet bilar i trafik marginellt mellan 2022 och 2023, vilket är första gången på över 20 år som den uppåtgående trenden bryts och antalet bilar i trafik minskar (SCB, 2024).

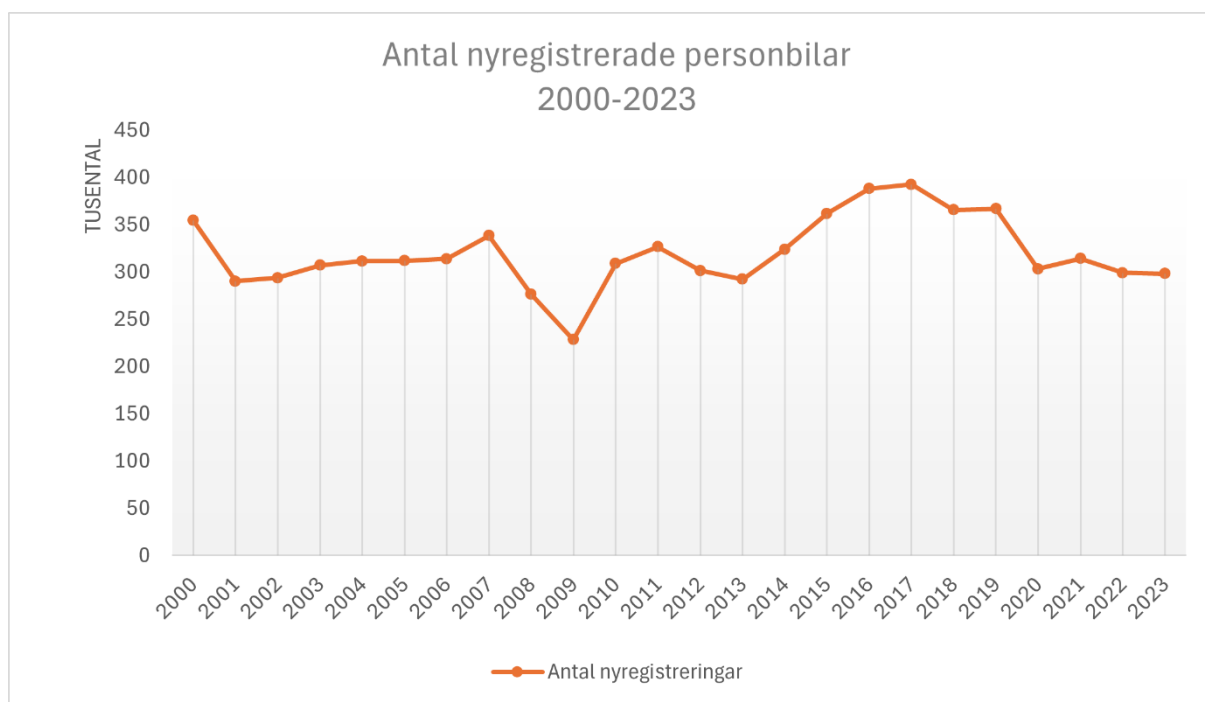


Bild 1. Graf över antalet nyregistrerade personbilar i Sverige mellan åren 2000–2023. Data hämtad via SCB:s API.

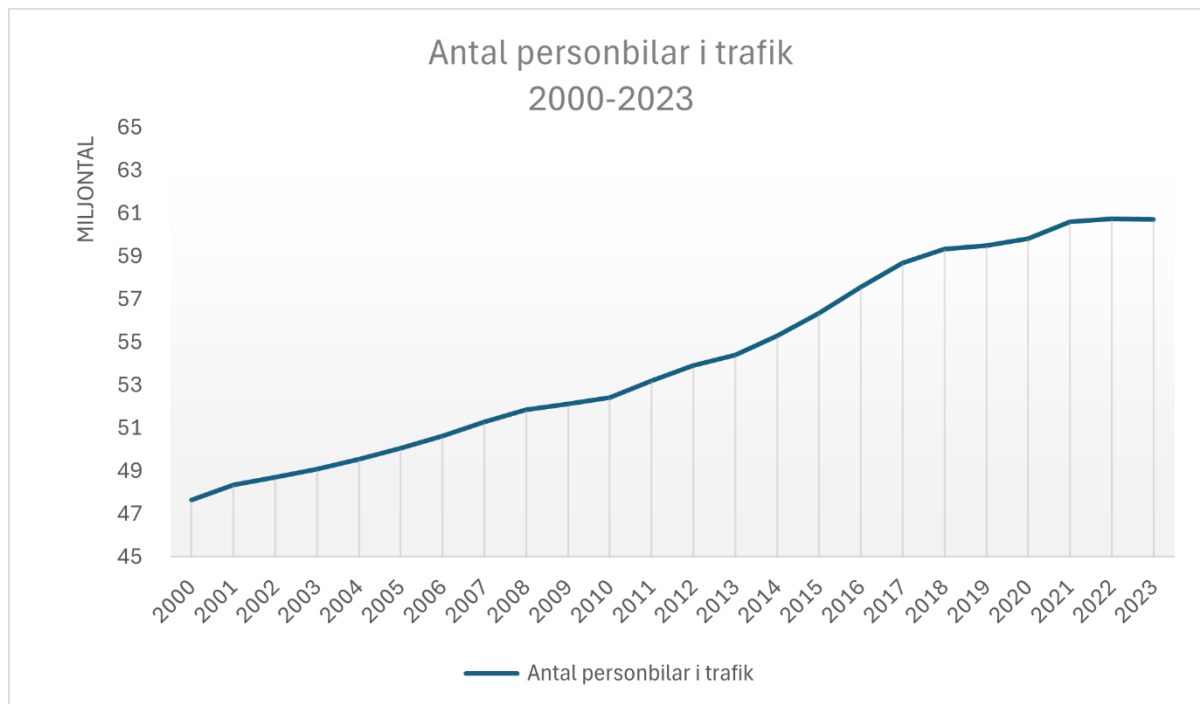


Bild 2. Graf över antalet personbilar i trafik i Sverige mellan åren 2000–2023. Data hämtad via SCB:s API.

Även om begagnatmarknaden har varit priskänslig de senaste åren, talar mätningar för att den uppåtgående trenden nu kan vara bruten. Under slutet av 2023 och början på 2024 har sjunkande priser påvisats, vilket kan ses som en långsam återgång till mer normala prisnivåer (KVD, 2023).

Den komplexa prissättningen på begagnade bilar kan förstås bättre genom att undersöka olika faktorer som påverkar den. En analys av effekter på prissättningen utifrån bilens egenskaper kan bidra med insikter som kan hjälpa konsumenter att fatta mer informerade beslut vid köp av begagnade bilar, samt bidra till en ökad förståelse för vad som styr prissättningen idag.

## 1.1 Problemformulering

Syftet med det aktuella projektet är att undersöka om – och i så fall vilka – egenskaper som påverkar prissättningen av begagnade bilar på Blocket. Tidigare kunde man grovt uppskatta priset på en bil utifrån dess ålder, går det att finna liknande trender utifrån det insamlade materialet? För att kunna besvara syftet har följande frågeställningar formulerats.

1. Kan bilens ålder förklara >50% av variationen i priset?
2. Om några, vilka egenskaper bidrar med en signifikant effekt på priset?
3. Går det att med hjälp av regressionsanalys prediktera det sanna priset för en begagnad bil på Blocket med en konfidens på 95%?

### 1.1.1 Definitioner och avgränsningar

Eftersom syftet är att undersöka effekten på prissättningen utifrån olika bilegenskaper så används linjär regressionsanalys. Då det är prissättningen på begagnatmarknaden som avses undersökas så har det insamlade materialet avgränsats till att gälla annonser från hemsidan Blocket. Det tåls därmed att poängtera att det som undersöks är det efterfrågade pris som säljaren sätter, vilket nödvändigtvis inte är detsamma som säljpriset.

## 2 Teori

I följande avsnitt ges en kort introduktion till linjär regression och vad det kan användas till. Vi går igenom hur den linjära regressionsalgoritmen arbetar, beskriver användningen av olika regulariseringstekniker och ger exempel på hur man kan tänka i termer av variabelselektion, val av modell och utvärdering i regressionsmodellering. Vidare behandlas även vanliga problem inom linjär regression som kan dyka upp under arbetets gång och exempel ges på hur man kan hantera dessa.

### 2.1 Linjär Regression

Linjär regression är en central metod inom statistik och maskininlärning, och kan användas för att undersöka samband och göra prediktioner mellan en responsvariabel  $Y$  i relation till en eller flera förklarande variabler  $X_1, X_2, \dots, X_p$ . Metoden bygger på en enkel algoritm som gör resultaten tolkningsbara och möjliga att dra slutsatser från med statistisk inferens (Kröner & Wahlgren, 2006).

För linjär regression antas att det finns ett ungefärligt linjärt samband mellan responsvariabeln och de förklarande variablerna, vilket matematiskt kan härledas till formeln för räta linjens ekvation. För att modellera detta samband behöver vi dock först skatta de okända parametrarna ( $\beta$ ) för linjens ekvation. Vi söker den linje som observationerna avviker så lite som möjligt från (regressionslinjen).

Formeln för multipel linjär regression (där mer än en förklarande variabel används) kan skrivas som

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \varepsilon$$

Här representerar  $\beta_0$  interceptet (punkten där regressionslinjen skär Y-axeln) och koefficienterna  $\beta_1, \beta_2, \dots, \beta_p$  representerar lutningen på regressionslinjen för respektive förklarande variabel, och kan tolkas som den effekt en enhets ökning på den förklarande variabeln har på responsvariabeln, givet att övriga variabler hålls konstanta. Epsilon ( $\varepsilon$ ) är den slumpmässiga feltermen i modellen (som kan förklaras av individuell variation) (James et al., 2023).

#### 2.1.1 Skattning av regressionskoefficienterna

För att skatta de okända parametrarna / regressionskoefficienterna ( $\beta$ ) används minsta kvadratmetoden (OLS). Metoden går ut på att beräkna de bästa skattningarna av koefficienterna genom att minimera summan av kvadraten på residualerna (RSS), vilket representerar skillnaden mellan det observerade värdet och regressionslinjen (James et al., 2023).

Formeln för RSS kan skrivas enligt nedan, där residualen ( $e = y - \hat{y}$ ).

$$RSS = e_1^2 + e_2^2 + \dots + e_n^2$$

Minsta kvadratmetoden för enkel linjär regression väljer det värde på  $\hat{\beta}_0$  och  $\hat{\beta}_1$  som minimerar RSS, där  $\bar{y}$  och  $\bar{x}$  är stickprovets medelvärde. Skattning av regressionskoefficienterna för multipel linjär regression har en något mer komplicerad formel, men metoden är densamma (James et al., 2023).

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

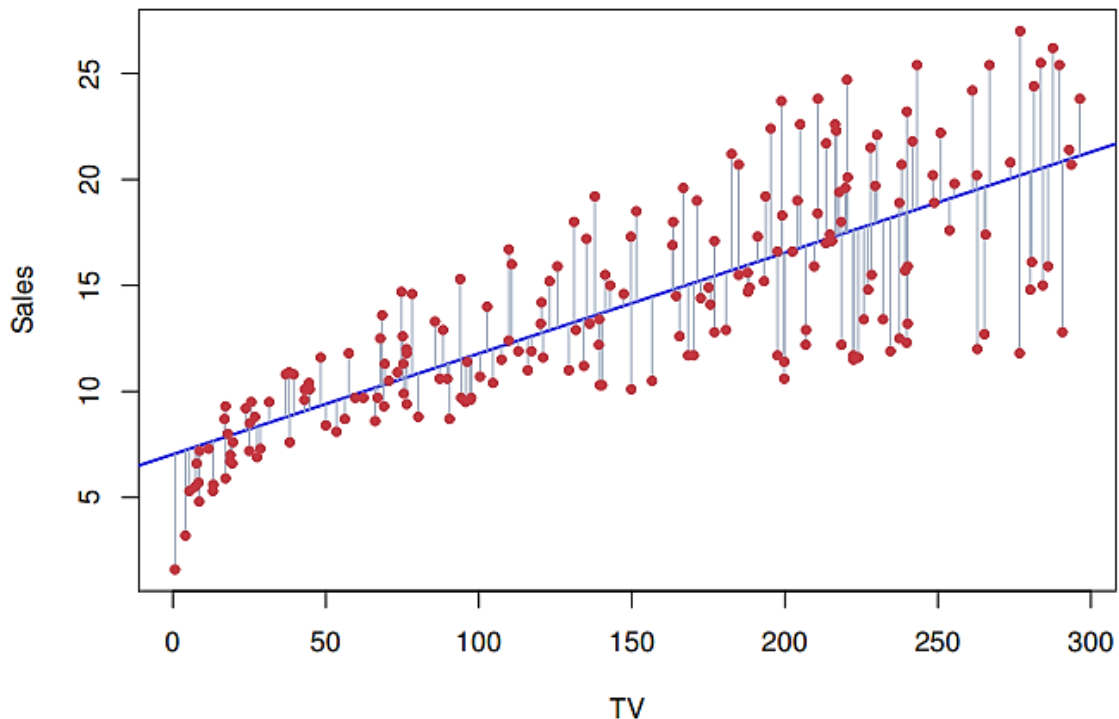


Bild 3. Diagram över försäljning regresserat på TV-marknadsföring. I diagrammet framgår hur regressionslinjen (i blått) anpassas efter residualerna (i grått) för observationerna (i rött). Hämtad från boken "An Introduction to Statistical Learning with Applications in R" (James et al., 2023 s.62).

### 2.1.2 Modellanpassning och utvärdering

Modellens förmåga att anpassa sig till den data som den tränats på utvärderas huvudsakligen med hjälp av två kvantitativa mått; den skattade standardavvikelsen av modellens felterm ( $RSE$ ) och determinationskoefficienten ( $R^2$ ) eller Adjusted  $R^2$  för multipel linjär regression (James et al., 2023).

Eftersom  $RSE$  mäter det skattade felet som modellen gör, vill vi ha ett så lågt  $RSE$  som möjligt. Formeln för  $RSE$  anges enligt nedan där  $RSS$  är residualernas kvadratsumma, som beskrevs tidigare.

$$RSE = \sqrt{\frac{1}{n-2}RSS}$$

Determinationskoefficienten  $R^2$ , eller Adjusted  $R^2$ , anger hur stor andel av variationen i responsvariabeln som kan förklaras utifrån sambandet med de förklarande variablerna. Adjusted  $R^2$  tar hänsyn till det ökade antalet variabler i multipel linjär regression. Till skillnad från  $RSE$  som mäter absoluta (fel)värden, så beräknas  $R^2$  i andelar – andelen av den förklarade variansen – och tar således alltid ett värde mellan 0 och 1. Till skillnad från  $RSE$  vill vi således på  $R^2$  ha ett så högt värde som möjligt (James et al., 2023).

Formeln för  $R^2$  kan anges enligt nedan, där  $TSS$  är den totala kvadratsumman.

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

### 2.1.3 Hypotesprövning

En viktig del i linjär regression är att dra statistisk inferens från resultaten genom hypotesprövning. Man vill oftast undersöka om det finns ett signifikant samband mellan responsvariabeln och åtminstone en av de förklarande variablerna. När man kör en regressionsmodell i R så gör programmet automatiskt hypotesprövningen åt oss genom att beräkna F-statistik. I grova drag så beskriver F-statistik förhållandet mellan två varianser och vi hade förväntat oss ett värde nära 1 om nollhypotesen är sann. Tillsammans med F-statistik beräknas även p-värdet, vilket kan användas för att bestämma om vi kan förkasta nollhypotesen eller inte. Det är vanligt att man tittar på p-värdet snarare än F-statistik eftersom det tillåtna värdet på F-statistik till viss del är beroende av antalet observationer och variabler i stickprovet (James et al., 2023).

### 2.1.4 Variabelselektion och regulariseringstekniker

Efter att ha identifierat signifikanta variabler är nästa steg att hantera modellkomplexiteten. Det finns en risk med att gå efter de individuella p-värdena när antalet förklarande variabler är många, och det finns därför andra sätt att undersöka de olika variablernas "viktighet" – i form av så kallad variabelselektion. Detta kan göras med variabelselektionstekniker såsom "best subset selection", "forward selection" eller "backward selection". I grova drag identifieras den kombination av variabler som presterar bäst och genom att filtrera ut mindre bidragande variabler kan man både förbättra modellens predikteringsförmåga och minska modellkomplexiteten. En tumregel är att alltid sträva efter en så enkel modell som möjligt (James et al., 2023).

Två andra metoder som används flitigt som alternativ till den vanliga linjära regressionsmodellen är Lasso regression och Ridge regression, som båda använder sig av regulariseringstekniker. Vad det innebär i praktiken är att man adderar en extra strafffunktion ( $\lambda$ ) till minsta kvadratmetoden. Till skillnad från metoderna för variabelselektion så tränas Lasso och Ridge modellerna med samtliga variabler men utför en indirekt variabelselektion genom att krympa koefficienterna. Lasso kan selektera variabler genom att krympa koefficienterna till exakt 0, vilket innebär att de inte längre används i modellen. Regularisering verkar effektivt på att minska modeller med hög varians och kan motverka överanpassning av data (James et al., 2023).

### 2.1.5 Modellantaganden och potentiella problem

När vi arbetar med linjära regressionsmodeller så förutsätter modellen att vissa antaganden om datan är uppfyllda. Om något eller någon av dessa antaganden bryts så kan vi inte lita på den statistiska inferens och/eller prediktioner som modellen gör. Modellen antar följande

- Linjärt förhållande mellan responsvariabeln och de förklarande variablerna
- Icke-korrelerade residualer, felen är fördelade oberoende av varandra
- Homoskedasticitet, residualernas standardavvikelse  $\sigma_\epsilon$  är lika stor för alla nivåer
- Normalfördelade residualer

I praktiken anses dessa antaganden aldrig vara helt uppfyllda, men det finns samtidigt många metoder för att undvika problemen och åtgärda dem om antaganden skulle anses vara brutna. Ett bra verktyg för att upptäcka potentiella problem är genom residualanalyser. I R finns inbyggda residualanalyser som kan avslöja hur väl (eller dåligt) modellen representerar datan.

Genom att granska residualerna kan vi upptäcka mönster i datan som modellen inte kan förklara. Om det linjära antagandet är uppfyllt så finns det exempelvis inget tydligt mönster i det diagram som visar residualerna i relation till de tränade värdena, utan residualerna antar en jämn och slumpmässig spridning. Om det däremot finns en tydlig U-form så indikerar det på att datan följer ett icke-linjärt



samband. Ett vanligt sätt att hantera detta på är att använda icke-linjära transformationer av de förklarande variablerna, så som  $\log X$ , roten ur  $X$  och  $X^2$  (James et al., 2023).

Det går även att undersöka antagandet om normalfördelade residualer, genom att titta på ett QQ-diagram. Ett annat problem som kan uppstå är att residualernas standardavvikelser inte är konstanta, dvs det finns en ojämn spridning av residualerna som modellen inte kan förklara. Det kallas även för heteroskedasticitet och motsatsen, det vi vill uppnå, är homoskedasticitet vilket innebär att residualernas standardavvikelser är lika stora för alla nivåer. En möjlig lösning till heteroskedasticitet är att logtransformera responsvariabeln, vilket resulterar i en större krympning av stora responsvärden (Kim, 2015).

Outliers kan också vara ett problem, och innebär att det observerade värdet är långt ifrån det predikterade värdet. Residualanalyser kan användas för att identifiera outliers, eller "high-leverage" punkter. Även om de inte påverkar skattningen av regressionslinjen särskilt mycket så kan de ha stor negativ effekt på  $RSE$  och  $R^2$ . Det tåls dock att poängtera att outliers även kan indikera på att det finns ett mönster eller ett samband som modellen inte kan förklara, dvs att det saknas en eller flera förklarande variabler (James et al., 2023).

Slutligen är det även viktigt för den statistiska analysen att det inte finns multikollinearitet, vilket innebär att två eller flera av de oberoende variablerna korrelerar med varandra. Om det finns indikationer på multikollinearitet i modellen så går det inte att uttala sig om den individuella effekten på  $Y$  för respektive variabel, vilket i sin tur även leder till en minskad styrka i hypotesprövningar.

#### 2.1.6 Variabeltransformering

Som nämnt kan det vara en idé att logtransformera responsvariabeln som ett sätt att åtgärda heteroskedasticitet (Kim, 2015). I regel kan det vara bra att logtransformera en variabel om den är kraftigt positivt snedfördelad (dvs färre stora värden). Poängen med variabeltransformation är att ändra skalan på observationerna så att de kan beskrivas som normalfördelad data. Det är vanligt att kvantitativa mått så som inkomst, pris och population växer exponentiellt, vilket resulterar i en log-normalfördelning. I sådana fall kan logtransformering vara lämpligt. När responsvariabeln logtransformeras, indikerar det vanligtvis ett logaritmiskt eller multiplikativt samband, snarare än det vanliga additiva samband som regressionsmodeller förutsätter. Detta innebär att en enhetsförändring i den oberoende variabeln tolkas som en procentuell förändring i responsvariabeln (Ford, 2018).

Vidare hanterar regressionsmodeller kategoriska variabler genom att automatiskt skapa dummyvariabler för faktorer. Detta sker som standard när modellen möter kategorisk data och kräver således inte manuell transformation innan modellen anpassningen (James et al., 2023).

#### 2.1.7 Modellval och utvärdering

I valet av vilken modell vi ska använda behövs kvantitativa mått för att utvärdera modellens förmåga att generalisera till ny, osedd data. Vid träning av modellerna använde vi mått såsom det relativa kvadratfelet,  $RSE$  för att uppskatta träningsfelet och bedöma hur väl modellen anpassar sig till datan.

Ett av de vanligaste måtten för att bedöma modellens förmåga att generalisera till ny data är utifrån medelkvadratfelet,  $MSE$ .  $MSE$  är ett mått på hur nära modellens förutsägelser ligger jämfört med de faktiska värdena; ett lägre  $MSE$  indikerar att modellen gör mer precisa förutsägelser. För att kunna tala om prediktionsfelet i samma enhet som responsvariabeln så används i stället  $RMSE$ , som är kvadratroten av  $MSE$  (James et al., 2023).

För att kunna utvärdera modellen på ett tillförlitligt sätt kan den totala datamängden delas upp i tre olika delar; träning (60%), validering (20%) och test (20%). Modellen tränas då på träningsdata, modellval sker baserat på prestanda på valideringsdata och den slutliga modellen utvärderas sedermera på testdata, för att säkerställa att den inte överanpassas till träningsdata. Alternativt kan man använda andra kvantitativa mått så som Mallows  $C_p$ , AIC, BIC och Adjusted  $R^2$  (James et al., 2023).

En fördel med att dela upp datan är emellertid att vi får en direkt skattning av modellens prestanda på osedd data och gör på så vis mindre antaganden om den sanna modellen.













### 3 Metod

I följande avsnitt beskrivs arbetsgången, och de metodval som gjorts, från det att data samlades in till den färdiga förberedda data som modellerna sedermera tränades på.

#### 3.1 Datainsamling

Datainsamlingen har genomförts genom en kombination av webbskrapning från Blockets hemsida och manuell datainsamling. Arbetet utfördes i en grupp om fem personer, där en person ansvarade för webbskrapningen och övriga gruppmedlemmar samlade in data manuellt från 30 bilannonser vardera. För att undvika dubletter vid den manuella inhämtningen så fick varje person ett eget bilmärke att arbeta med, och vid inhämtning via webbskrapning så filtrerades de annons-ID bort som redan inhämtats manuellt.

Den data som hämtades in från bilannonserna var utifrån den information som uppgetts i de faktarutor som säljaren ska fylla i och som är fristående från fritextbeskrivningen.

Fakta			
 Bränsle <b>Diesel</b>	 Växellåda <b>Manuell</b>	 Miltal <b>24 606</b>	 Modellår <b>2017</b>
 Biltyp <b>Kombi</b>	 Drivning <b>Fyrhjulsdriven</b>	 Hästkrafter <b>150 Hk</b>	 Färg <b>Silver</b>
 Motorstorlek <b>1968 cc</b>	 Datum i trafik <b>2016-11-23</b>	 Märke <b>Volkswagen</b>	 Modell <b>Passat</b>

För att motverka en allt för stor spridning i den data som samlades in, definierades ett antal sökkriterier. Följande kriterier/avgränsningar användes vid den manuella såväl som den automatiserade inhämtningen:

- Endast privatsäljare
- Modellår >= 2000
- Prisintervall: 20 000 kr – 500 000 kr
- Inga yrkesfordon
- Ingen leasing

#### 3.2 Databearbetning

Efter att ha hämtat och laddat in det insamlade materialet genomfördes en översiktlig granskning för att undersöka struktur, kvalité och form på datan. Bland annat framgick att datan innehöll ett antal dubletter, saknade värden och hade en relativt obalanserad distribution inom samtliga kategorier. En anledning till detta kan vara för att materialet avgränsades till privatsäljare, vilket troligtvis leder till en större variation i hur data fylls i – exempelvis identifierades flertalet felaktiga datainmatningar inom miltal där man i stället har uppgett värdet i kilometer. Det finns även många egenskaper med saknade värden där man valt att inte fylla i faktarutorna men i stället uppgett informationen i fritext.

Innan distributionsanalyser och liknande kunde genomföras transformerades datan om till datatyperna numeriska och faktorer – initialt var samtlig data definierad som text. På så sätt förbereds även datan inför regressionsanalys, eftersom modellerna förutsätter numeriska värden och per automatik omvandlar faktorer till dummyvariabler.

### 3.2.1 Hantering av saknade värden

Totalt sett fanns det 3 277 saknade värden i materialet, fördelat enligt nedan.

Id	Märke	Modell	Bränsle
0	82	107	82
Växellåda	Miltal	Modellår	Biltyp
82	82	82	87
Drivning	HK	Färg	Motorstorlek
444	401	480	852
Datum i trafik	Region	Pris	
413	0	83	

För att hantera detta togs först observationer med saknade priser bort (vid sökning på annons ID framgick att annonserna var bortplockade från Blocket – och kan således förklaras av en bugg i webbskrapningen). När dessa observationer hade plockats bort försvann också samtliga saknade värden i märke, bränsle, växellåda, miltal och modellår.

Som framgår i tabellen ovan fanns det ett stort antal saknade värden i motorstorlek och när detta granskades vidare framgick att det till stor del kan förklaras av elbilarna. Motorstorlek plottades även i relation till hästkrafter vilket visade på ett linjärt samband och bekräftade misstanken om att de mäter ungefär samma sak. Så i stället för att riskera att förlora en stor del av elbilarna i vidare analyser och för att undvika kollinearitet mellan motorstorlek och hästkrafter, så valdes motorstorlek att plockas bort från materialet.

Därefter raderades samtliga återstående saknade värden och det återstod då drygt 7 500 observationer i materialet.

### 3.2.2 Explorativ dataanalys (EDA)

Efter att ha tvättat och bearbetat datan genomfördes en EDA för att vidare undersöka samband mellan variablerna, datafördelning och identifiera potentiella outliers eller oväntade observationer som kan påverka regressionsanalysen. Några av de insikter som framkom under EDA är följande:

- Samtliga numeriska variabler är positivt skevfördelade, dvs de har få höga värden.
- 75% av bilarnas miltal ligger under 22 652, medan maxvärdet är 473 954.
- Responsvariabeln (Pris) ser ut att följa en lognormalfördelning.
- Fördelningen inom de kategoriska variablerna är relativt obalanserad, exempelvis är elbilar underrepresenterade i relation till diesel och bensinbilar. Detsamma gäller för cab, coupé och familjebuss i relation till andra biltyper. Det gäller även i relation till priset. Exempelvis är elbilarnas priser normalfördelade medan det finns en positiv skevfördelning bland framför allt bensin och dieselbilarna, vilket kan ha att göra med distributionen.
- Det finns en stor spridning i märken och modeller där vissa märken är dominerande, så som Volvo, BMW, Volkswagen, Audi och Mercedes. Den stora variationen i modeller kan förklaras av en faktisk variation i modeller men det skiljer sig även i hur man uppger modellnamnet.
- Det finns 3 517 unika värden för "Datum i trafik".
- Det finns tendenser till ett icke-linjärt samband mellan ålder och pris, vilket kan tyda på att det finns ett exponentiellt samband.

Följande steg togs därefter, i ett sätt att förbättra kvaliteten på datan:

- En ny variabel "Ålder" skapades utifrån datum i trafik. Då kunde även två outliers (felaktig datainmatning där år < 2000 uppgetts som datum i trafik) identifieras och raderas.

- Materialet avgränsades till de vanligaste (och mest förekommande i datan) biltyperna sedan, SUV, kombi och halvkombi.
- Miltal avgränsades till mer än 0 och max 35 000 för att inte riskera att få med bilar som använts till exempelvis taxikörning eller där man har matat in data felaktigt. I snitt ligger miltalet för en vanlig personbil på cirka 1000 mil/år.
- Pris avgränsades till mer än 20 000 och mindre än 500 000, vilket var de initiala filteravgränsningarna, men där andra observationer följt med i webbskrapningen.
- Märke begränsades till de som har mer än 30 observationer och modeller exkluderades.

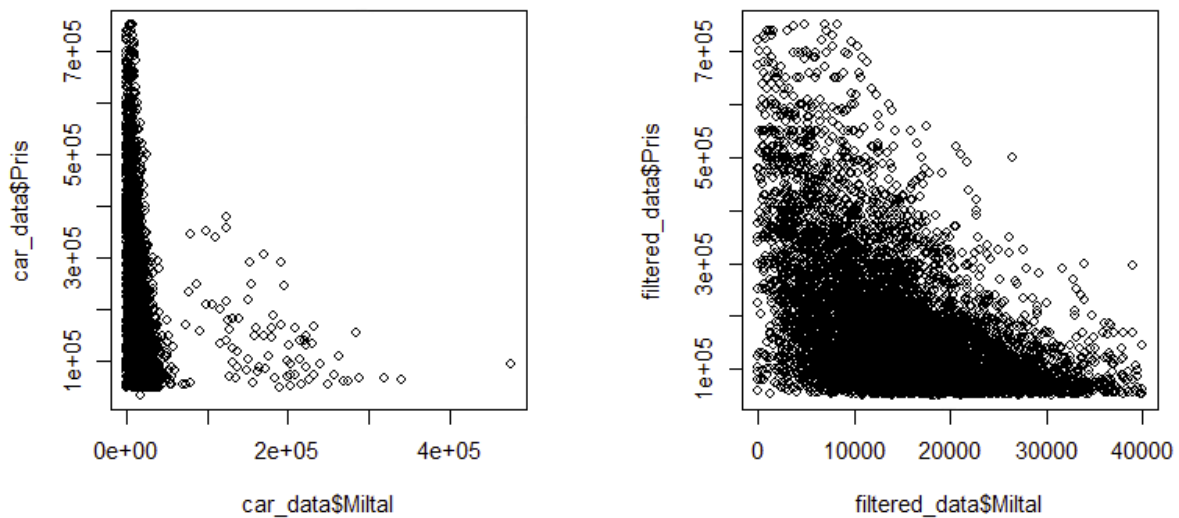


Bild 4. Samband mellan miltal och pris. Till vänster är innan avgränsningarna och till höger efteråt.

### 3.2.3 Logtransformering

Pris och hästkrafter logtransformerades för att rätta till icke-linjära samband i datan.

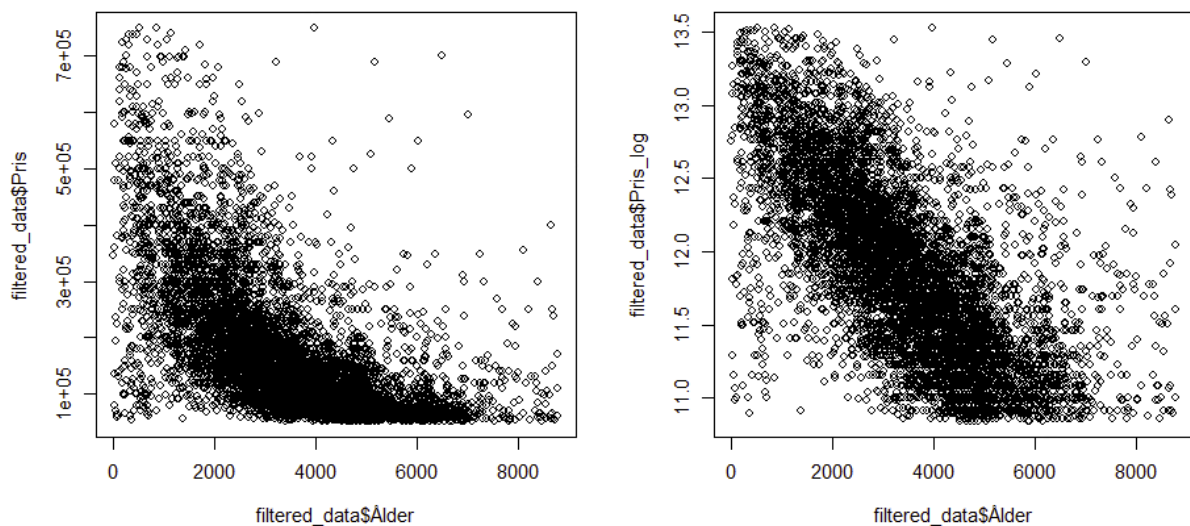


Bild 5. Samband mellan ålder och pris. Till höger är sambandet mellan ålder och det logaritmerade priset.

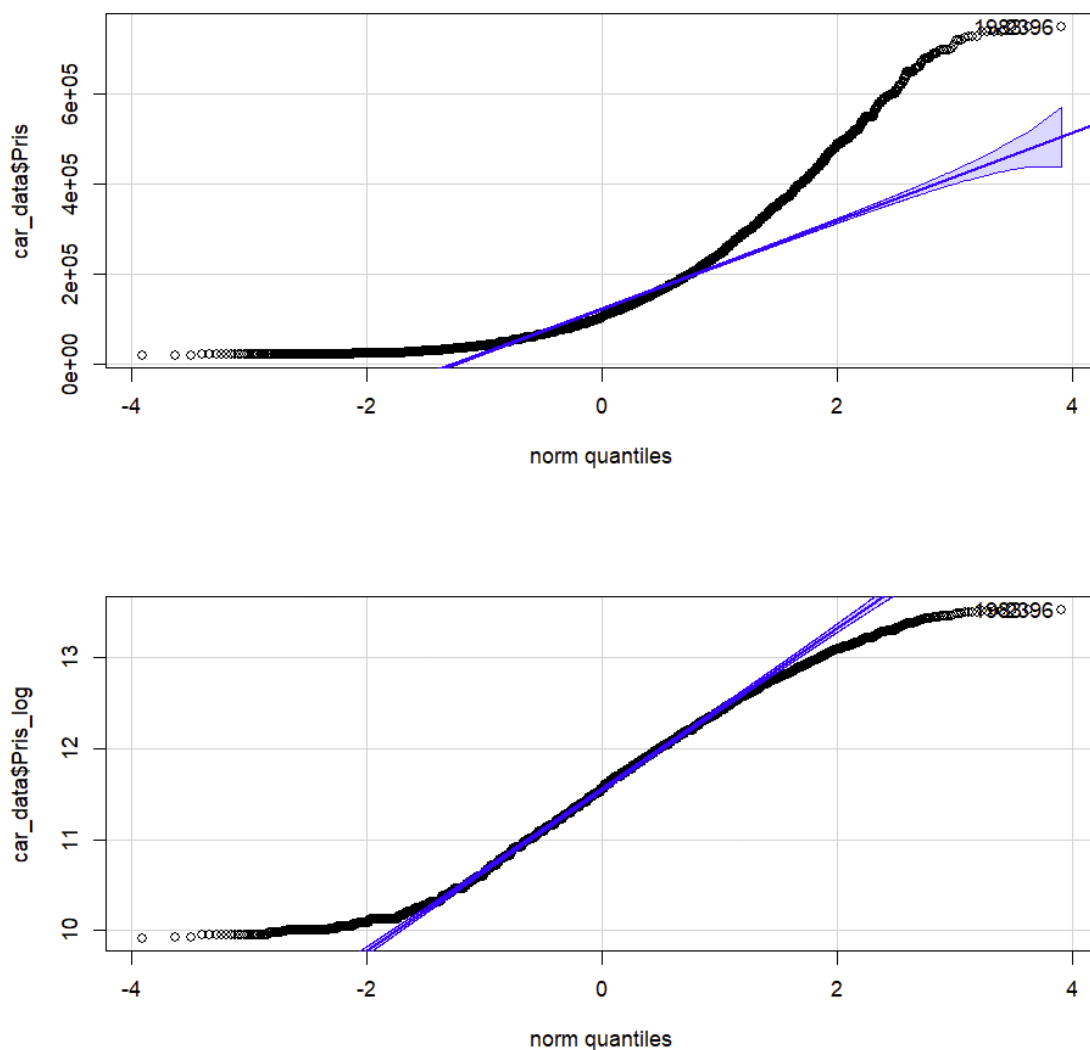


Bild 6. QQ-plot visar datafördelningen i Pris, innan och efter logtransformering. I det nedre diagrammet syns att det logaritmerade priset ligger betydligt närmre en normalfördelning.

### 3.2.4 Undersökning av kollinearitet (numeriska variabler)

Därefter undersöktes relationen mellan de numeriska variablerna.

	Miltal	Modellår	HK	Pris	Ålder
Miltal	1.00000000	-0.5519551	-0.07724163	-0.5486525	0.5574540
Modellår	-0.55195512	1.0000000	0.17440391	0.6896730	-0.9337500
HK	-0.07724163	0.1744039	1.00000000	0.6524245	-0.2329831
Pris	-0.54865248	0.6896730	0.65242451	1.0000000	-0.6987679
Ålder	0.55745397	-0.9337500	-0.23298312	-0.6987679	1.0000000

Insikter från korrelationsmatrisen enligt nedan.

Samband mellan pris och de oberoende variablerna:

- Det finns ett starkt positivt samband mellan pris och de två oberoende variablerna HK (hästkrafter) och modellår. Det kan tolkas som att ju fler hästkrafter en bil har desto högre är prissättningen och på samma sätt prissätts nyare modeller dyrare.
- Det finns ett starkt negativt samband mellan pris och ålder, vilket tolkas som att ju äldre bilen är desto lägre prissättning.

Samband mellan de oberoende variablerna (gulmarkerade):

- Det finns en hög korrelation mellan flertalet oberoende variabler. Exempelvis finns det ett positivt samband mellan ålder (och modellår) och miltal som indikerar på att ju äldre bilen är desto fler miltal har den gått, vilket är ett rimligt antagande. Det finns även ett naturligt negativt samband mellan modellår och ålder – ju äldre modell (året modellen lanserades) desto längre tid (år) har bilen varit i trafik.

Detta är insikter som kommer vara viktiga att ha i åtanke vid modelleringen, särskilt sambanden mellan de oberoende variablerna vilket kan leda till multikollinearitet. Eftersom modellår och ålder verkar mäta ungefär samma sak, exkluderades modellår från fortsatta analyser.

### 3.2.5 Dataförberedning

Slutligen förbereddes datan inför modellering och utvärdering genom att delas upp i tre olika set – träning, validering och test.

## 4 Resultat och Diskussion

I följande avsnitt presenteras resultaten från regressionsmodellerna tillsammans med en diskussion om valet av variabler och modellprestanda. Slutligen redovisas även resultaten från den färdiga modellens prediktioner på helt nya Blocket-annonser och en tolkning av modellens koefficienter.

### 4.1 Träning av fem regressionsmodeller

Sammanlagt tränades fem linjära regressionsmodeller på träningsdata. I den första modellen inkluderades samtliga prediktorer, men vid analys av residualerna upptäcktes att modellantaganden inte kunde uppfyllas.

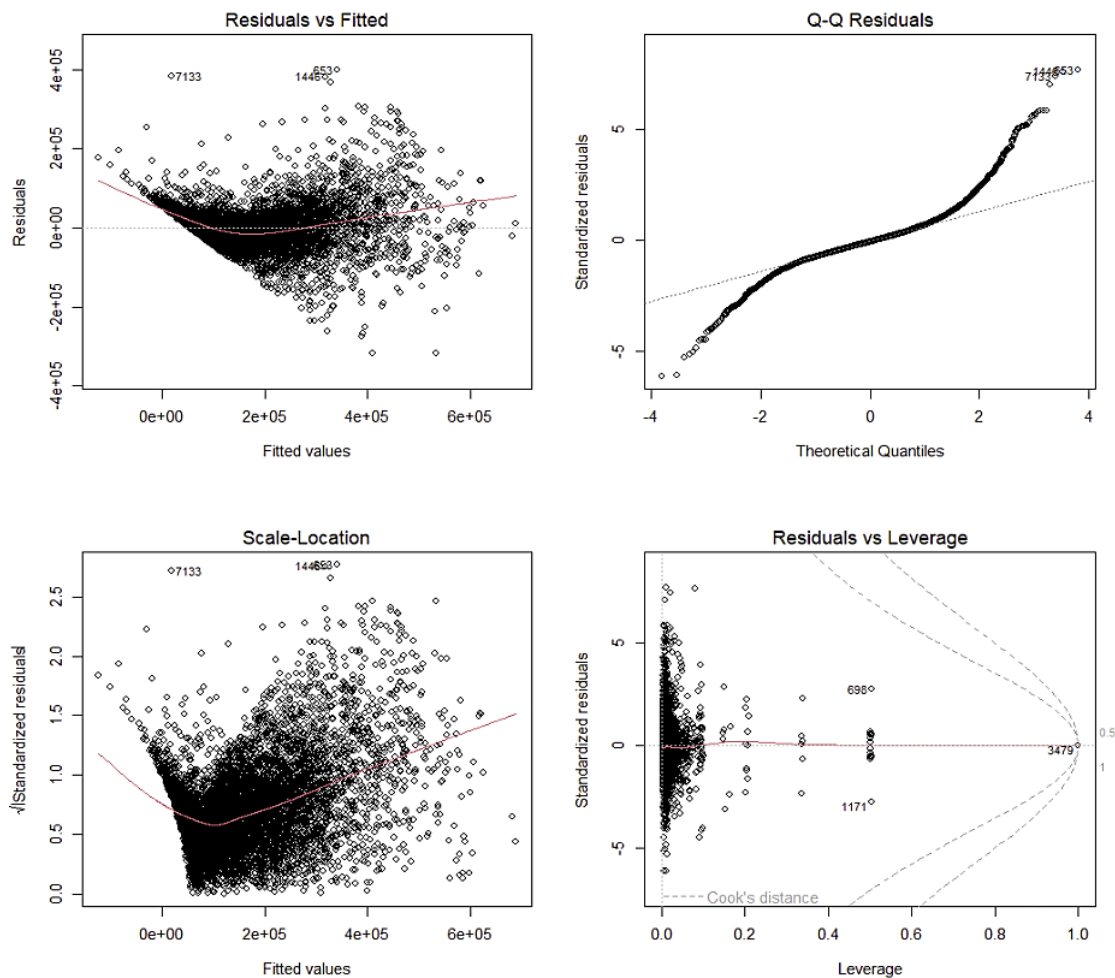


Bild 7. Residualanalys av modell 1. Modellantaganden anses inte uppfyllda.

Diagramanalysen visar tydliga U-formade mönster i residualerna, vilket indikerar att det finns ett icke-linjärt samband som inte fångas av modellen. Dessutom visade residualerna en stor spridning i variansen (heteroskedasticitet) och vissa avvikelser från en normalfördelning. Vid beräkning av VIF observerades hög korrelation mellan bilens märke och andra oberoende variabler.

Med anledning av brutna modellantaganden förkastades modell 1 och i stället tränades en ny modell där den beroende variabeln och prediktorvariabeln hästkrafter logtransformerades, medan andra variabler användes i sin ursprungliga skala. För att undvika multikollinearitet exkluderades även märke.



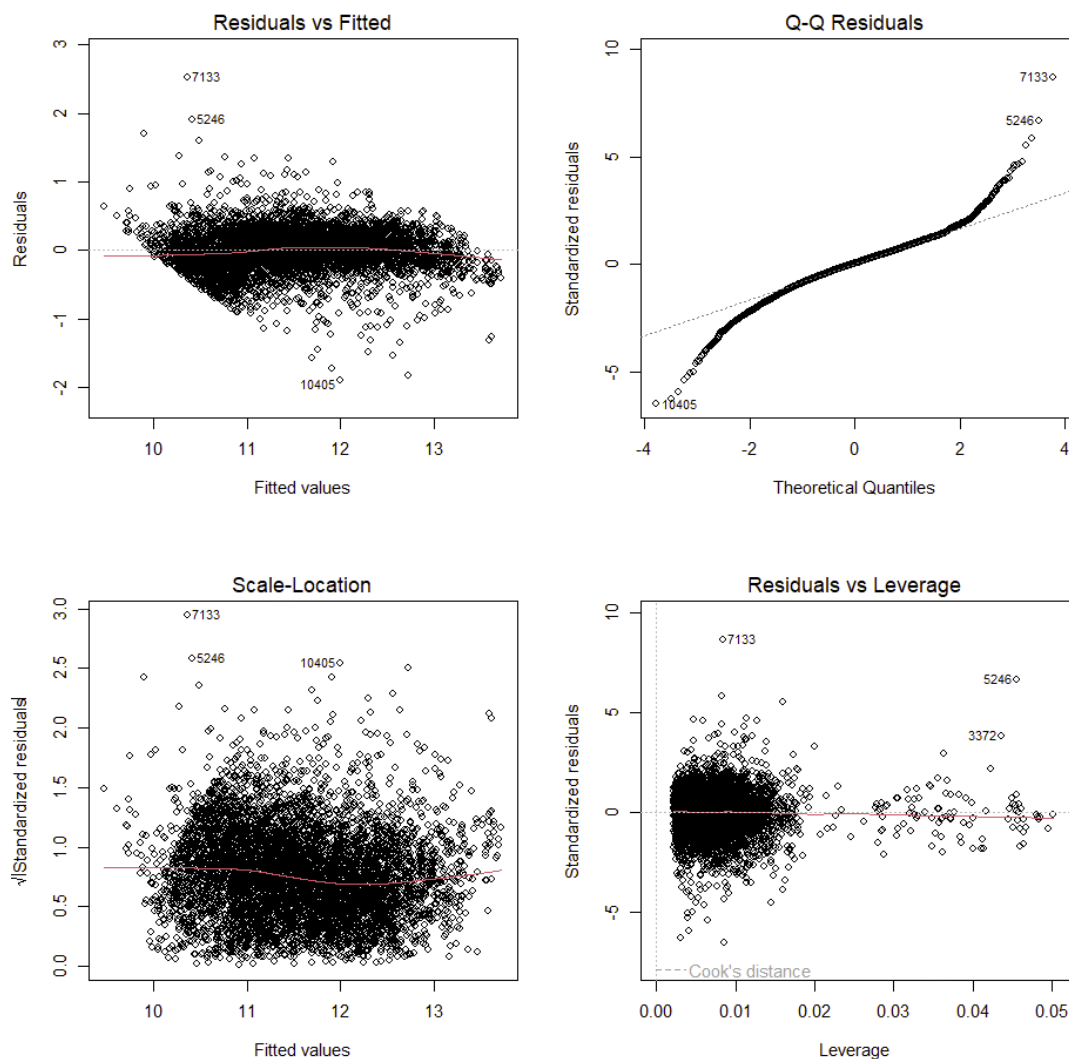


Bild 8. Residualanalys av modell 2. Modellantaganden anses uppfylla.

Som framgår av diagramanalysen ovan fångades det icke-linjära sambandet i den andra modellen genom att logtransformera den beroende variabeln. Detta resulterade i en jämnare spridning av residualerna och effekter av potentiellt avvikande observationer har även minskat. Trots vissa indikationer på bristande normalfördelning i residualerna, ansågs modellantaganden nu vara uppfyllda.

I QQ-diagrammet framgår att svansarna avviker från normalfördelningen, vilket kan tolkas som att datan har en leptokurtisk fördelning (Kim, 2015). Majoriteten av observationerna ligger nära medelvärdet och följer en normalfördelning där men det finns det några observationer i svansarna som är avvikande och drar ut dem. För att undersöka detta exkluderades ålder från modellen vilket ledde till att svansarna blev betydligt smalare och mindre extrema. Samtidigt minskade den förklarade variansen drastiskt när ålder plockades bort och RSE ökade. Utifrån domänskunskap vet vi också att bilens ålder historiskt sett haft ett stort förklaringsvärde på priset och av den anledningen behövs variabeln inför vidare modellering. Sammanfattningsvis kan en tolkning av denna observation vara att det finns en stor variation i hur äldre bilar prissätts – vilket i sin tur kan ha att göra med faktorer som inte inkluderas i den aktuella modellen så som samlarvärde eller servicebehov.

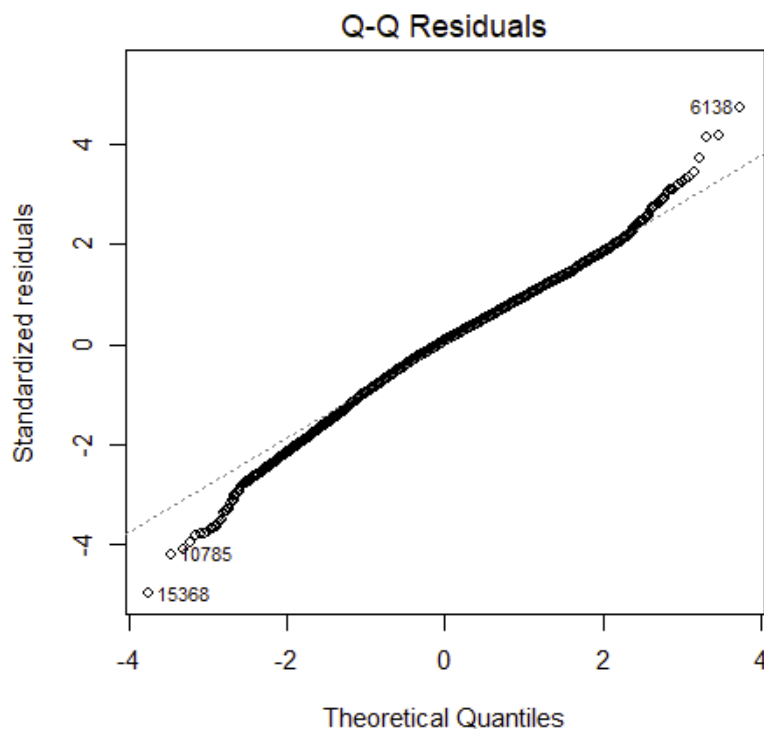


Bild 9. Q-Q-diagram över residualernas fördelning efter att ålder har exkluderats från modellen.

Därefter tränades en tredje modell där icke-signifikanta variabler så som färg (förutom vit) och region togs bort. Resultaten visade ingen märkbar skillnad i modellprestanda efter att dessa variabler exkluderats, vilket talar för att dem inte bidrar med någon signifikant effekt på priset givet de andra variablerna i modellen.

Den fjärde modellen tränades med de fem bäst presterande variablerna enligt "best subset selection": milital, ålder, hästkrafter (logaritmerad), bränsle och växellåda.

Den femte och sista modellen tränades med enbart det logaritmerade priset regresserat på bilens ålder, detta för att kunna undersöka en av de frågeställningar som sattes upp för projektet.

## 4.2 Modellutvärdering och val av modell

Modellval baserades delvis på prestanda på valideringsdata och delvis utifrån strävan efter minskad modellkomplexitet. Modellerna utvärderades på valideringsdata med hjälp av RMSE och Adj  $R^2$ .

Modell	RMSE (log)	RMSE (original)	Adj $R^2$	Antal prediktorer
2	0.2710031	ca 31% fel	87,96%	42
3	0.2715352	ca 31% fel	87,88%	11
4	0.2748828	ca 32% fel	87,54%	7
5	0.4387832	ca 55% fel	67,40%	1

Modell 4 valdes ut för vidare utvärdering på testdata. Den var tillräckligt enkel och med en marginell skillnad i prestanda trots att den enbart inkluderade 1/6 av prediktorerna jämfört med de som användes i modell 2. Sammanfattningsvis är det knappt någon märkbar skillnad i prestanda mellan de tre första modellerna vilket tyder på att ett fåtal av prediktorerna kan anses stå för den totala andelen förklarad varians i modellerna. Det verkar inte behövas mer än de 5 variabler (7 inkl. dummysnivåer) som används i modell 4 för att förklara en stor andel av variationen i priset. Trots att

bilens ålder som ensam prediktor (modell 5) verkar kunna förklara en stor andel själv så uppskattar modellen göra mer än hälften fel på valideringsdata, vilket inte anses tillräckligt bra.

Vid utvärdering på testdata av modell 4 uppmättes ett RMSE på 0.2844675 i logaritmisk skala, vilket motsvarar cirka 33% fel. Eftersom RSE på träningsdata var liknande den på testdata, och högre än den på valideringsdata, verkar det som att modellen inte överanpassar till träningsdata. Eftersom RSE är baserat på träningsdata och RMSE på validerings- och testdata är de två måtten inte direkt jämförbara, men de kan användas som indikatorer för att förstå modellens prestanda.

Eftersom det inte finns indikationer på att modellen överanpassar data och baserat på det knappa antalet prediktorer som används, fanns det ingen anledning att använda ytterligare regularisering på datan (så som ridge och lasso).

#### 4.3 Hypotesprövning och statistisk inferens

När den valda modellen tränades om på hela datasetet framgick att en variabel, miljöbränsle/hybrid, inte var signifikant. Övriga variabler visade en stark signifikant effekt på priset. Resultatet framgår av tabellen nedan, dock är koefficienterna (förutom HK\_log) relaterade till det logaritmerade priset och behöver därför först exponentieras för att kunna tolkas som den procentuella effekt de har på priset.

```
Residuals:
    Min       1Q   Median       3Q      Max
-2.02870 -0.14029  0.01996  0.17095  2.77054

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   8.891e+00  4.739e-02  187.622  <2e-16 ***
Miltal        -2.932e-05  5.062e-07  -57.916  <2e-16 ***
Ålder         -8.213e-02  8.379e-04  -98.015  <2e-16 ***
HK_log        8.001e-01  9.034e-03   88.562  <2e-16 ***
BränsleDiesel 1.830e-01  7.284e-03   25.117  <2e-16 ***
BränsleEl     -1.622e-01  1.602e-02  -10.125  <2e-16 ***
BränsleMiljöbränsle/Hybrid -5.742e-03  1.093e-02  -0.525    0.599
VäxellådaManuell -1.798e-01  7.295e-03  -24.650  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2831 on 9283 degrees of freedom
Multiple R-squared:  0.8763,    Adjusted R-squared:  0.8762
F-statistic: 9394 on 7 and 9283 DF,  p-value: < 2.2e-16
```

Eftersom hästkrafter och pris är i samma logaritmerade skala så kan sambandet tolkas som att 1% ökning av bilens hästkrafter (oavsett bilens ålder, miltal osv) leder till en 0,8% ökning av priset. Som exempel på hur övriga koefficienter ska tolkas så exponentieras koefficienten för ålder,  $\exp(-0.08213)$  vilket blir ungefär 0.92. Det innebär att för varje ett års ökning på bilens ålder förväntas priset minska med cirka 8% ( $1-0.92$ ).

Det kanske mest överraskande resultatet i modellen är att elbilar leder till ett lägre pris i relation till bensinbilar. Detta skulle dock delvis kunna bero på den ojämna prisdistributionen bland bensinbilar och underrepresentationen av elbilar i materialet, vilket framgick under EDA. Något annat som var lite oväntat är att miltal har en marginell effekt på priset med en koefficient som är väldigt nära noll. Eftersom det finns en viss korrelation mellan bilens ålder och antal mil den kört så kan man tänka sig att åldern spelar ut effekten av miltal på priset. Detta är dock intressant med tanke på att både miltal och pris har setts som viktiga faktorer att tänka på vid köp av en begagnad bil, men det skulle kunna finnas en viss skillnad i när de olika faktorerna blir intressanta. En gammal bil som har gått lägre mil

verkar i detta fall vara värd mindre än en nyare bil som har gått många mil. Samtidigt kan denna effekt också vara ett resultat av att miltal under databearbetningen begränsades till 40 000.

Vi har bland annat valt att använda variabler så som miltal och hästkrafter för att prediktera priset på en begagnad bil. Om man tänker efter så är det inte särskilt förvånande att äldre bilar prissätts billigare än motsatsen och att bilar med fler hästkrafter generellt sett är dyrare. Det är egentligen inte särskilt upplysande information. I stället för att enbart utgå från variabler som direkt relaterar till egenskaper hos bilen (ålder, miltal, bränsle osv), hade man kunnat använda någon form av marknadsmässig indikator (efterfrågan, inflation) för att undersöka hur prissättningen påverkas. Det skulle dock innebära att man samlar in data över en längre tid för att kunna utvärdera effekterna på prisutvecklingen. Samtidigt kan man också tänka sig att det finns fler egenskaper som leder till variationer i priset som inte kan förklaras av de aktuella prediktorerna, så som servicebehov eller samlarvärde som nämndes tidigare i relation till ålderns påverkan på residualernas normalfördelning. Dessutom går det inte att undvika den underliggande variation som förekommer i datan till följd av inkonsekvent och/eller felaktig datainmatning av privatsäljare. Så även om resultaten är signifikanta och modellantaganden är uppfyllda, bör vi vara försiktiga med att dra för häftiga slutsatser kring begagnatmarknaden i stort.

#### 4.3.1 Prisprediktioner på nya Blocket-annonser

Men för att ändå testa hur väl modellen predikterar priset på nya bilannonser så valdes tre nya annonser ut från Blocket. Skärmdumpar över de annonser som valdes ut återfinns som bilagor.

När samtliga egenskaper (miltal, ålder, hästkrafter, bränsle och växellåda) matats in för de nya bilarna så fick modellen prediktera priset genom att uppskatta ett konfidensintervall inom var det sanna värdet med 95% konfidens beräknas ligga.

Det faktiska priset	Det predikterade priset	Konfidensintervall (95%)
55 000	55 220	54 347—56 106
255 000	254 019	250 421 – 257 668
119 000	117 216	115 253 – 119 213

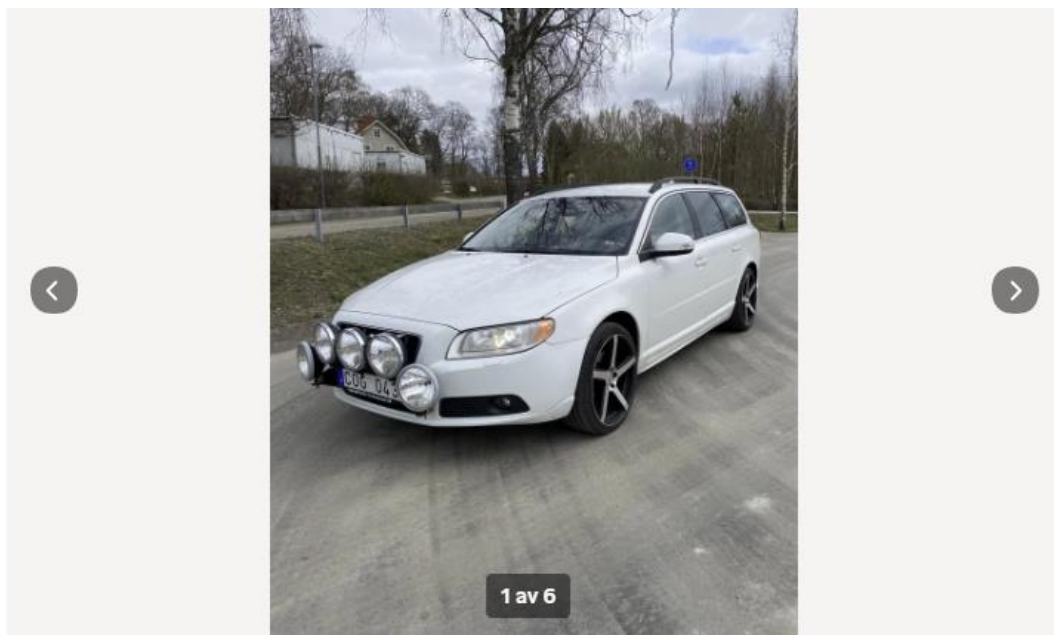
Sammanfattningsvis kan alltså konstateras att modellen lyckades fånga det sanna värdet vid samtliga tre tillfällen, vid uppskattning av det genomsnittliga priset för bilen med 95% konfidens.

## 5 Slutsatser

Sammanfattningsvis kan följande slutsatser dras i relation till syftet med projektet och de frågeställningar som analysen ämnade besvara.

1. Vid regression av logaritmen av bilpriset mot bilens ålder upptäcktes att åldern ensam kunde förklara över 67% av den variation i bilpriserna som observerades.
2. Baserat på resultaten har påvisats att bilens ålder, miltal (dock med en marginell effekt), typ av växellåda, bränsle (diesel/el/bensin) och antal hästkrafter har en signifikant effekt på det logaritmerade bilpriset. Vidare observerades även att vissa märken, typ av drivning och vissa biltyper verkade ha en signifikant effekt enligt andra modeller som utforskades.
3. Den slutgiltiga regressionsmodellen lyckades prediktera det sanna bilpriset för nya bilannonser från Blocket med en konfidens på 95%. Med andra ord, modellen kunde ge ett konfidensintervall inom vilket det sanna värdet av bilpriset förväntades ligga.

## Appendix A



🕒 Inlagd: idag 16:19

📍 [Nyköping \(hitta.se\)](#)

📖 1 Spara

### Volvo V70 D3 Momentum Euro 5

**55 000 kr**

[I samarbete med Lendo: Beräkna lånekostnad direkt](#)

#### Fakta



Bränsle  
**Diesel**



Växellåda  
**Manuell**



Miltal  
**32 888**



Modellår  
**2011**



Biltyp  
**Kombi**



Drivning  
**Tvåhjulsdriven**



Hästkrafter  
**164 Hk**



Färg  
**Vit**



Motorstorlek  
**1984 cc**



Datum i trafik  
**2011-01-12**

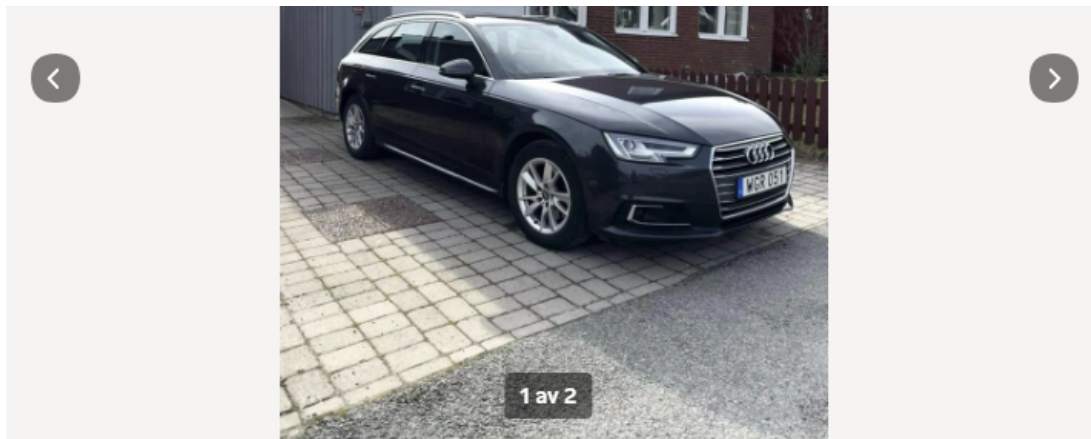


Märke  
**Volvo**



Modell  
**V70**

^ Visa mindre fakta



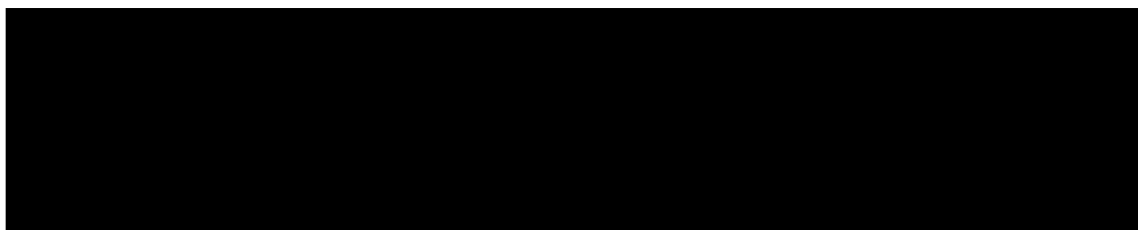
🕒 Inlagd: idag 16:21

📍 Botkyrka (hitta.se)

📖 Spara

## Audi A4 Avant 40 TFSI S TronicAmbition,Proline, Sport Euro 6 **255 000 kr**

⚙️ I samarbete med Lendo: Beräkna lånekostnad direkt



👤 Aktiv på Blocket ➡️ Svarar snabbt

### Fakta

📄 Bränsle <b>Bensin</b>	🔗 Väckellåda <b>Automat</b>	📅 Miltal <b>6 409</b>	📅 Modellår <b>2018</b>
🚗 Biltyp <b>Kombi</b>	⚙️ Drivning <b>Tvåhjulsdreven</b>	🕒 Hästkrafter <b>191 Hk</b>	🖌️ Färg <b>Grå</b>
⚙️ Motorstorlek <b>1984 cc</b>	📅 Datum i trafik <b>2018-09-18</b>	🛡️ Märke <b>Audi</b>	📖 Modell <b>A4</b>

^ Visa mindre fakta





🕒 Inlagd: igår 16:17

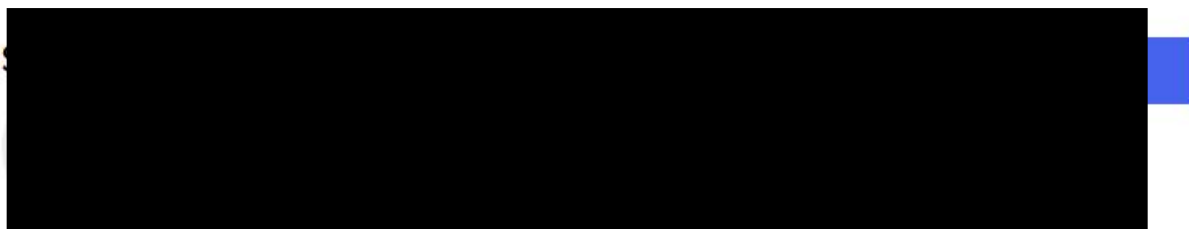
📍 Ängelholm (hitta.se)

📖 2 Spara

## Volvo V60 Cross Country D4 AWD Geartronic Summum Euro 6

### 119 000 kr

[I samarbete med Lendo: Beräkna lånekostnad direkt](#)



👉 Svarar snabbt

### Fakta



Bränsle  
Diesel



Växellåda  
Automat



Miltal  
31 458



Modellår  
2016



Biltyp  
Kombi



Drivning  
Fyrhjulsdriven



Hästkrafter  
191 Hk



Färg  
Svart



Motorstorlek  
2400 cc



Datum i trafik  
2016-01-05



Märke  
Volvo



Modell  
V60 Cross Country

⤴ Visa mindre fakta



## Källförteckning

- Bilpriser.se. (2022). Varför är begagnade bilar så dyra numera? Hämtad från <https://www.bilpriser.se/varderingsguiden/varfor-ar-begagnade-bilar-sa-dyra-numera/>
- Bilpriser.se. (2023). Begagnatmarknaden våren 2023. Hämtad från <https://www.mynewsdesk.com/se/kvdbil/pressreleases/begbilsrapport-priserna-paa-begagnade-bilar-fortsaetter-uppaat-3246584>
- Ford, C. (2018). Interpreting Log Transformations in a Linear Model. Hämtad från <https://library.virginia.edu/data/articles/interpreting-log-transformations-in-a-linear-model>
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2023). *An Introduction to Statistical Learning with Applications in R*. (2nd ed.). Springer.
- Kim, B. (2015). Understanding Diagnostic Plots for Linear Regression Analysis. Hämtad från <https://library.virginia.edu/data/articles/diagnostic-plots>
- Körner, S., & Wahlgren, L. (2006). *Statistisk dataanalys*. Studentlitteratur AB, Lund.
- KVD. (2023). Begagnad Bilrapport: Priserna på begagnade bilar fortsätter uppåt. Hämtad från <https://www.mynewsdesk.com/se/kvdbil/pressreleases/begbilsrapport-priserna-paa-begagnade-bilar-fortsaetter-uppaat-3246584>
- KVD. (2023). Efter alla prisrekord – nu sjunker priset på begagnade bilar. Hämtad från <https://www.kvd.se/artiklar/livet-med-bil/efter-alla-prisrekord-nu-sjunker-priset-pa-begagnade-bilar>
- Statistiska centralbyrån. (2024). Fordonsstatistik. Fordon enligt bilregistret efter fordonsslag och bestånd. Hämtad via SCB:s API.