Regressionmodelling in R and statistical analysis with data from blocket

Joakim Salomonsson

EC Utbildning

R programmering

2025–04

Abstract

A regression modelling exercise where I familiarize myself with the typical R programmering flow to perform a statistical data analysis. This report specifically explores the use of linear and robust regressions models to predict car prices based on data that was collected together with peers from the website Blocket. I tested several assumptions of regression models such as non-linearity, heteroskedasticity and multicollinearity and more. Robust regression was applied to handle outliers and non-normal residuals.

Innehållsförteckning

1 Inledning

Syftet med denna rapport är att presentera mitt arbete och dess resultat inom kursen R programmering. Som examination för denna kurs har jag bland annat genomfört följande:

- Svarat på teoretiska frågor
- Hämtat extern data från SCB både manuellt samt via API
- Samlat in data från Blocket i grupp
- Tvättat data från Blocket i Excel samt R
- Skapat regressionsmodeller i R
- Analyserat, tolkat och presenterat data utifrån ett statistiskt perspektiv

Jag kommer att besvara följande frågeställningar:

1. Vilken regressionsmodell passar bäst utifrån vår insamlade data?
2. Vilka variabler är mest signifikanta?

Regressionsmodellen byggdes och utvecklades ursprungligen med hela datamängden som kan ses i scriptet *RegressionModelDevelopment5.0.R*, men i efterhand insåg jag vikten av att utvärdera prediktiv prestanda på osedd data. Därför delades datan i en tränings- och testmängd i ett kompletterande steg och i ett separat R-script *Prediktivmodell_datasplit1.1.R* där jag har följt liknande steg som jag hade i min initiala modellutveckling. Det gör att modellen bättre kan bedömas utifrån generaliseringsförmåga. De prediktiva resultaten och felmåtten redovisas från den ursprungliga modellen på hela datan i avsnitt 4.6 samt från modellen validerad med träning/test i avsnitt 4.7

Den statistiska inferensen har gjorts på modellen som tränades på hela datasettet. Detta innefattar bland annat hypotesprövning, p-värden och konfidensintervall för regressionskoefficienterna.

2 Teori

2.1 Regressionsmodeller

2.1.1 Linjär Regressionsmodell

En linjär regressionsmodell är en statistisk modell som används för att prediktera ett värde Y baserat på en eller flera förklarande variabler X. Den kan uttryckas som:

$$Y \approx \beta 0 + \beta 1 X$$

(Linjär regression, James et al., 2023)

2.1.2 Robust Regression

I situationer där antagandet om homogen varians (homoskedasticitet) inte uppfylls, kan robust regression användas. Den är mindre känslig för både outliers och heteroskedasticitet, och ger mer tillförlitliga skattningar när standardmetoder som OLS påverkas negativt (FasterCapital, n.d.).

2.2 API

Ett API (Application Programming Interface) är ett gränssnitt som möjliggör kommunikation mellan olika system och mjukvarukomponenter. Det används av utvecklare för att integrera färdig funktionalitet eller tjänster i egna applikationer. Ett API kan liknas vid ett "Lego"-system, där varje modul (API:et) fungerar som en byggsten (Reddy, 2011).

2.3 Teoretiska antaganden

2.3.1 Icke-linjärt förhållande mellan den beroende variabeln och de oberoende variablerna

Linjär regressionsanalys bygger på antagandet att sambandet mellan den beroende variabeln och varje oberoende variabel är linjärt. I praktiken är detta inte alltid fallet, och vid icke-linjära relationer kan modellens förklaringskraft och precision minska avsevärt (James et al., 2023, s. 91-93). För att hantera sådana situationer kan olika transformationer användas:

- Log-transformation
  Ett vanligt sätt att hantera icke-linjäritet är att log-transformera variabler. Det innebär att en variabel ersätts med dess logaritm, exempelvis bas 10, bas 2 eller den naturliga logaritmen. Detta kan ofta leda till ett mer linjärt förhållande mellan variabler och minska heteroskedasticitet (James et al., 2023).
- Polynom
  En annan metod för att modellera icke-linjära samband är att inkludera polynomtermer, exempelvis kvadratiska eller kubiska termer. Detta tillåter modellen att böjas längs med datamönstret och kan därmed ge en bättre anpassning till verkliga samband mellan variabler (James et al., 2023, s. 92-93).

2.3.2 Korrellerade residualer - Ej oberoende residualer

Om residualer är korrelerade kan standardfelen underskattas, vilket leder till att konfidensintervall blir för snäva och därmed missvisande (James et al., 2023, s. 94).

- Durbin-Watson test
  För att undersöka huruvida residualerna är korrelerade används ofta Durbin-Watson testet. Ett värde nära 2 tyder på att residualerna är oberoende, medan ett värde ≤ 1 tyder på positiv autokorrelation (Chatterjee & Simonoff, 2013). Testet är särskilt användbart vid tidsberoende data.

2.3.3 Icke-konstant varians på residualerna (Heteroskedasticitet)

Ett annat grundantagande i regressionsmodeller är att residualerna har konstant varians (homoskedasticitet). Om detta inte uppfylls kan det påverka standardfel, p-värden och hypotesprövning (James et al., 2023, s. 94-96).

- Heteroskedasticitet
  När residualernas varians förändras med nivån på de förklarande variablerna föreligger heteroskedasticitet. Detta kan ofta hanteras genom att log-transformera den beroende variabeln (James et al., 2023), eller genom att använda robust regression, som är mindre känslig för sådana avvikelser (Rousseeuw et al., 2005).
- Breusch-Pagan test
  Ett vanligt test för att upptäcka heteroskedasticitet är Breusch-Pagan testet, som mäter huruvida variansen i residualerna är systematiskt relaterad till förklaringsvariablerna (Cook & Weisberg, 1983).

2.3.4 Normalfördelning

Normalfördelningen är central inom statistisk analys och är särskilt relevant i regressionssammanhang. En normalfördelning beskriver hur observationer är fördelade runt ett medelvärde och representeras ofta av en symmetrisk klockkurva (Wikipedia, n.d.-a).

- Shapiro-Wilk test
  För att undersöka om residualerna är normalfördelade används Shapiro-Wilk testet. Ett lågt p-värde tyder på att residualerna avviker från normalfördelning (Shapiro & Wilk, 1965).

2.3.5 Outliers

En outlier är en datapunkt där värdet på den beroende variabeln Y avviker kraftigt från det predikterade värdet givet dess X-värde. Outliers kan ha stor påverkan på modellens koefficienter och bör därför identifieras och hanteras korrekt (James et al., 2023, s. 97).

2.3.6 “High leverage”punkter

Till skillnad från outliers är en high leverage-punkt en datapunkt med ett extremt värde på någon av de oberoende variablerna, vilket kan ge stort inflytande på modellens anpassning. (James et al., 2023, s. 98)

- Cook's Distance
  Ett vanligt mått för att upptäcka datapunkter med stort inflytande är Cook's Distance. Om ett värde överstiger 4/n betraktas punkten ofta som problematisk (Mendenhall & Sincich, 1996).

2.3.7 Kollinearitet/multikollinearitet

Enligt James et al. (2023, s. 100) uppstår kolinearitet när två eller flera förklarande variabler är starkt korrelerade med varandra. Detta kan leda till att deras respektive effekter blir svåra att särskilja, vilket i sin tur försvårar tolkningen av regressionsmodellen.

- VIF (Variance Inflation Factor)
  Ett vanligt sätt att upptäcka kolinearitet är att beräkna VIF-värden. Ett VIF nära 1 tyder på att ingen kolinearitet förekommer, vilket är önskvärt. Värden över 5-10 indikerar dock hög multikolinearitet, vilket kan skapa instabilitet i skattningarna och leda till att små förändringar i datan kraftigt påverkar koefficienterna (James et al., 2023, s. 102).

3 Metod

3.1 Insamling och användande av API samt PX-WEB API

För att genomföra den första delen av kunskapskontrollen kring extern data använde jag PX-WEB API från Statistiska centralbyrån (SCB), specifikt för personbilar i Stockholm under perioden 2015–2024. Efter att ha bekantat mig med SCB:s API:et skapade jag ett R-script som skickade en POST-förfrågan till:
https://api.scb.se/OV0104/v1/doris/en/ssd/TK/TK1001/TK1001A/PersBilarA

Förfrågan inkluderade en JSON-query där jag specificerade årtal och region. API:et returnerade data i JSON-format, som sedan konverterades till .xlsx för visualisering i Excel.

3.2 Insamling av data från blocket

Nästa steg innebar att samla in data manuellt från annonser på blocket.se. Eftersom det rekommenderades att arbeta i grupp, genomfördes datainsamlingen tillsammans i klassen.

3.2.1 Gruppdeltagare

Alvin, Arash, Ana, Emad, Gayathree, Hani, Katarina, Joakim, Michael, My, Peter, Per, Sharmin, Rana, Tahira, Tural och Zakariyae

3.2.2 Tillvägagångssätt

Varje deltagare samlade in ca 50 observationer, fördelat på olika regioner. Varje observation motsvarade en bilannons, och informationen (t.ex. pris, modell, miltal) samlades in i en gemensam Excel-fil.

3.2.3 Lärdomar

Manuell datainsamling ledde till flera utmaningar, inklusive:

- Olika format (t.ex. blandning av små/stora bokstäver)
- Mänskliga fel (fel kolumn, inkonsekventa kategorier)
- Stavfel i variabelvärden (t.ex. “SUB”)

För att hantera detta krävdes omfattande datastädning och standardisering.

3.3 “Tvättning” av datan från blocket

Eftersom datainsamlingen skedde manuellt av flera personer, krävdes omfattande datarensning och standardisering för att säkerställa att modellen kunde byggas på en ren och strukturerad datamängd.

3.3.1 “Grov tvättning” i Excel

Den initiala datatvätten gjordes i Excel, där tydliga felaktigheter och inkonsekvenser rättades till. Exempel på åtgärder inkluderar:

- Färg: Tog bort parenteser
- Bränsle: Ändrade "Hybrid" till "Miljö/Hybrid", samt "Diesel" (felstavning)
- Biltyp: Ändrade stavfel t.ex. “SUB till SUV”
- Pris och miltal: Tog bort tusentalsavgränsare (mellanslag)
- Modell: Tog bort mellanrum (ex. “XC40 244”) och korrigerade modellnamn
- Drivning: Enhetlig stavning

- Modellår: Justerade felaktiga format (t.ex. årtal på en rad)
- N/A: Tog bort tomma/icke ifyllda poster

Dessutom valdes biltyperna “Cab” och “Coupé” bort, eftersom dessa endast förekom i färre än tre observationer, vilket senare orsakade problem i regressionsmodellen.

3.3.2 Granulär tvättning i R

Efter den grova tvätten i Excel genomfördes mer detaljerad rensning i R för att skapa en modellklar datastruktur:

- Skapade ny variabel Age baserat på bilens första registreringsdatum
- Tog bort irrelevanta kolumner såsom URL, märke (endast Volvo), färg och datum i trafik
- Exkluderade variabeln Motorstorlek eftersom många observationer saknade detta
- Log-transformerade Försäljningspris och Miltal
- Använde polynom av grad 2 för ålder
- Tog bort Region – användes enbart som gruppering vid insamling
- “Lumpade” modeller med färre än 10 observationer till kategorin “Other”
- Rensade bort outliers samt high leverage-punkter med hjälp av Cook’s Distance och standardiserade residualer

3.4 Regressionsmodellering

Jag inledde analysen genom att skapa en linjär regressionsmodell i R med hjälp av funktionen lm(). Därefter undersökte jag i mitt script huruvida de teoretiska antagandena för regressionsanalys var uppfyllda, i enlighet med beskrivningen i kunskapskontrollen.

Utifrån resultaten anpassade jag modellen successivt. I den slutgiltiga modellen använde jag funktionen rlm() från paketet MASS, som implementerar en robust regressionsmetod (Huber M-estimator) för att minska påverkan från outliers, heteroskedasticitet och avvikelser från normalfördelning.

För att kunna utvärdera modellens förmåga att prediktera delades datan i en träningsdel (80%) och testdel (20%) i ett separat R script efter att jag hade utvecklat modellen. Modellen tränades på träningsdatan och utvärderades med testdatan. Se avsnitt 4.7.

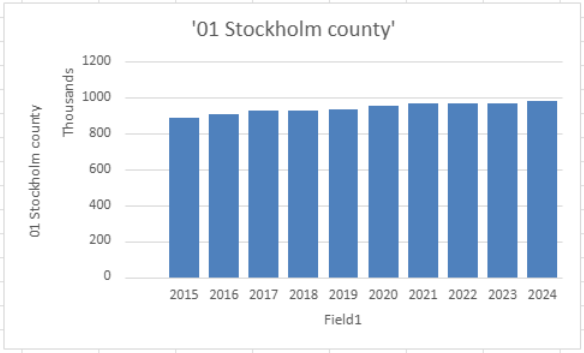
4 Resultat och Diskussion

Detta kapitel presenterar analys, tolkning och visualiseringar av data från både SCB och Blocket. Fokus ligger på resultatet från regressionsmodellen och dess statistiska tolkning.

4.1 Visualisering av extern data hämtad från SCB

Nedan visas en visualisering av antal registrerade personbilar i Stockholms län mellan åren 2015-2024. Data hämtades via API från Statistiska centralbyrån (SCB). Syftet med denna del var att öva på hantering av externa API:er, och visualiseringen visar en tydlig men relativt stabil ökning av bilbeståndet under perioden.

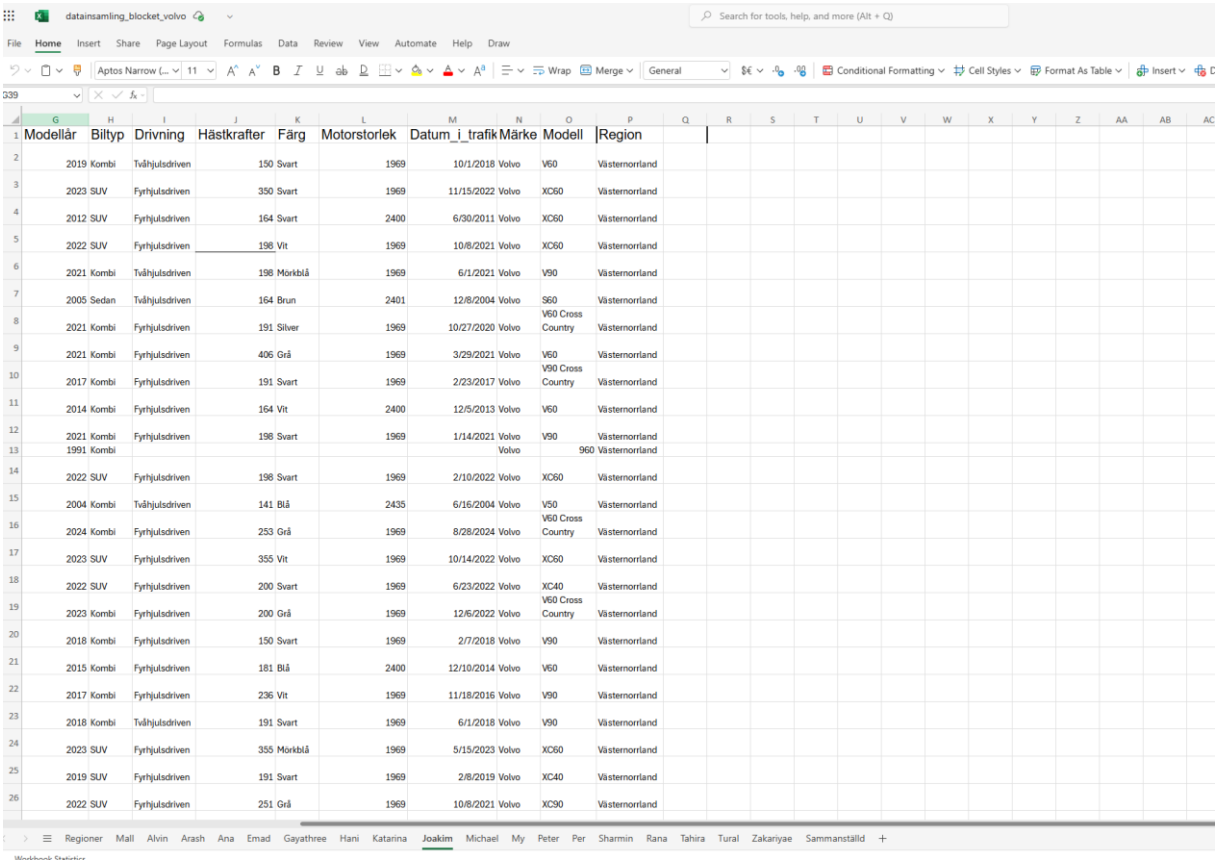
Passenger cars in use, number by region, type of ownership, observations and year											
		Number									
		2015	**2016**	**2017**	**2018**	**2019**	**2020**	**2021**	**2022**	**2023**	**2024**
01 Stockh	**total**	887057	909434	929570	931771	935865	953391	965686	965722	967916	983506
Corrected 2017-03-01: Figures for type of ownership ´Total´ for year 2016 have been corrected.											
2018-03-12 Data for 2017 is corrected.											
Latest update:											
Number:											
20250203 15:13											
Source:											
Transport Analysis											
Contact:											
Number:											
Annika Johansson, Statistics Sweden											
+46 010-479 68 25											
annika.johansson@scb.se											
Magnus Nyström, Statistics Sweden											
+46 010-479 63 73											
magnus.nystrom@scb.se											
Units:											
Number:											
number											
Data type:											
Number:											
Official statistics											
Database:											
Statistics database 2.3											
Internal reference code:											
TK1001AB											



Figur 1: Skärmdump av datan som hämtades via API från SCB.

4.2 Visualisering av den insamlade datan som gjordes i grupp

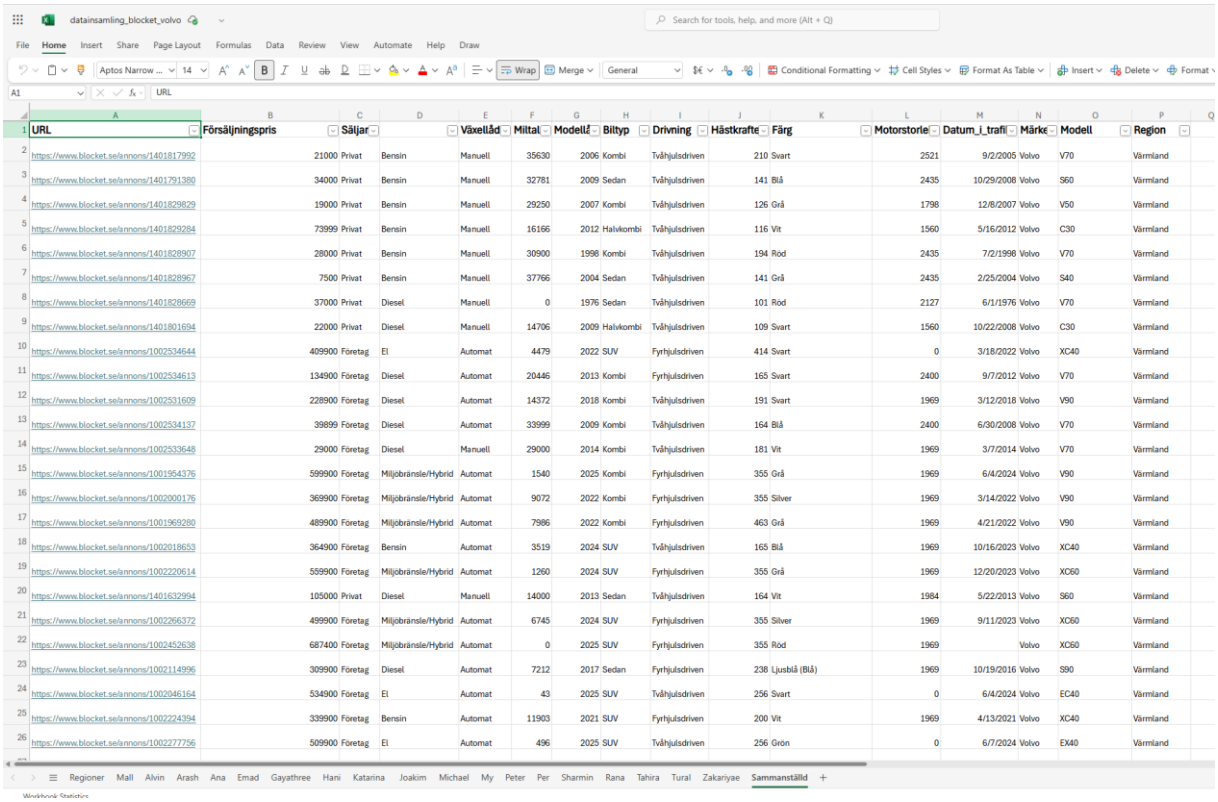
I denna del visas utdrag från det insamlingsarbete som genomfördes manuellt av kursdeltagare. Insamlingen gjordes i grupper, där varje deltagare ansvarade för ett visst antal bilannonser. Nedan visualiseras både en individuell datainsamling (min del) och den sammanslagna datamängden.



Modellår	Biltyp	Drivning	Hästkrafter	Färg	Motorstorlek	Datum_i_trafik	Märke	Modell	Region
2019	Kombi	Tvåhjulsdriven	150	Svart	1969	10/1/2018	Volvo	V60	Västernorrland
2023	SUV	Fyrhjulsdriven	350	Svart	1969	11/15/2022	Volvo	XC60	Västernorrland
2012	SUV	Fyrhjulsdriven	164	Svart	2400	6/30/2011	Volvo	XC60	Västernorrland
2022	SUV	Fyrhjulsdriven	198	Vit	1969	10/8/2021	Volvo	XC60	Västernorrland
2021	Kombi	Tvåhjulsdriven	198	Mörkblå	1969	6/1/2021	Volvo	V90	Västernorrland
2005	Sedan	Tvåhjulsdriven	164	Brun	2401	12/8/2004	Volvo	S60	Västernorrland
2021	Kombi	Fyrhjulsdriven	191	Silver	1969	10/27/2020	Volvo	V60 Cross Country	Västernorrland
2021	Kombi	Fyrhjulsdriven	406	Grå	1969	3/29/2021	Volvo	V60	Västernorrland
2017	Kombi	Fyrhjulsdriven	191	Svart	1969	2/23/2017	Volvo	V90 Cross Country	Västernorrland
2014	Kombi	Fyrhjulsdriven	164	Vit	2400	12/5/2013	Volvo	V60	Västernorrland
2021	Kombi	Fyrhjulsdriven	198	Svart	1969	1/14/2021	Volvo	V90	Västernorrland
1991	Kombi						Volvo	960	Västernorrland
2022	SUV	Fyrhjulsdriven	198	Svart	1969	2/10/2022	Volvo	XC60	Västernorrland
2004	Kombi	Tvåhjulsdriven	141	Blå	2435	6/16/2004	Volvo	V50	Västernorrland
2024	Kombi	Fyrhjulsdriven	253	Grå	1969	8/28/2024	Volvo	V60 Cross Country	Västernorrland
2023	SUV	Fyrhjulsdriven	355	Vit	1969	10/14/2022	Volvo	XC60	Västernorrland
2022	SUV	Fyrhjulsdriven	200	Svart	1969	6/23/2022	Volvo	XC40	Västernorrland
2023	Kombi	Fyrhjulsdriven	200	Grå	1969	12/6/2022	Volvo	V60 Cross Country	Västernorrland
2018	Kombi	Fyrhjulsdriven	150	Svart	1969	2/7/2018	Volvo	V90	Västernorrland
2015	Kombi	Fyrhjulsdriven	181	Blå	2400	12/10/2014	Volvo	V60	Västernorrland
2017	Kombi	Fyrhjulsdriven	236	Vit	1969	11/18/2016	Volvo	V90	Västernorrland
2018	Kombi	Tvåhjulsdriven	191	Svart	1969	6/1/2018	Volvo	V90	Västernorrland
2023	SUV	Fyrhjulsdriven	355	Mörkblå	1969	5/15/2023	Volvo	XC60	Västernorrland
2019	SUV	Fyrhjulsdriven	191	Svart	1969	2/8/2019	Volvo	XC40	Västernorrland
2022	SUV	Fyrhjulsdriven	251	Grå	1969	10/8/2021	Volvo	XC90	Västernorrland

Figur 2: Skärmdump av datan som samlades in i grupp från Blocket, specifikt min del

Datainsamlingen låg till grund för den fortsatta datarensningen, transformationen och regressionsmodelleringen som beskrivs i senare avsnitt.



URL	Försäljningspris	Säljar		Växellåd	Miltal	Modellå	Biltyp	Drivning	Hästkrafte	Färg	Motorstorle	Datum_i_trafi	Märke	Modell	Region
https://www.blocket.se/annons/1401817992	21000	Privat	Bensin	Manuell	35630	2006	Kombi	Tvåhjulsdriven	210	Svart	2521	9/2/2005	Volvo	V70	Värmland
https://www.blocket.se/annons/1401791380	34000	Privat	Bensin	Manuell	32781	2009	Sedan	Tvåhjulsdriven	141	Blå	2435	10/29/2008	Volvo	S60	Värmland
https://www.blocket.se/annons/1401829829	19000	Privat	Bensin	Manuell	29250	2007	Kombi	Tvåhjulsdriven	126	Grå	1798	12/8/2007	Volvo	V50	Värmland
https://www.blocket.se/annons/1401829284	73999	Privat	Bensin	Manuell	16166	2012	Halvkombi	Tvåhjulsdriven	116	Vit	1560	5/16/2012	Volvo	C30	Värmland
https://www.blocket.se/annons/1401828907	28000	Privat	Bensin	Manuell	30900	1998	Kombi	Tvåhjulsdriven	194	Röd	2435	7/2/1998	Volvo	V70	Värmland
https://www.blocket.se/annons/1401828967	7500	Privat	Bensin	Manuell	37766	2004	Sedan	Tvåhjulsdriven	141	Grå	2435	2/25/2004	Volvo	S40	Värmland
https://www.blocket.se/annons/1401828669	37000	Privat	Diesel	Manuell	0	1976	Sedan	Tvåhjulsdriven	101	Röd	2127	6/1/1976	Volvo	V70	Värmland
https://www.blocket.se/annons/1401801694	22000	Privat	Diesel	Manuell	14706	2009	Halvkombi	Tvåhjulsdriven	109	Svart	1560	10/22/2008	Volvo	C30	Värmland
https://www.blocket.se/annons/1002534644	409900	Företag	El	Automat	4479	2022	SUV	Fyrhjulsdriven	414	Svart	0	3/18/2022	Volvo	XC40	Värmland
https://www.blocket.se/annons/1002534613	134900	Företag	Diesel	Automat	20446	2013	Kombi	Fyrhjulsdriven	165	Svart	2400	9/7/2012	Volvo	V70	Värmland
https://www.blocket.se/annons/1002531609	228900	Företag	Diesel	Automat	14372	2018	Kombi	Tvåhjulsdriven	191	Svart	1969	3/12/2018	Volvo	V90	Värmland
https://www.blocket.se/annons/1002534137	39899	Företag	Diesel	Automat	33999	2009	Kombi	Tvåhjulsdriven	164	Blå	2400	6/30/2008	Volvo	V70	Värmland
https://www.blocket.se/annons/1002533648	29000	Företag	Diesel	Manuell	29000	2014	Kombi	Tvåhjulsdriven	181	Vit	1969	3/7/2014	Volvo	V70	Värmland
https://www.blocket.se/annons/1001954376	599900	Företag	Miljöbränsle/Hybrid	Automat	1540	2025	Kombi	Fyrhjulsdriven	355	Grå	1969	6/4/2024	Volvo	V90	Värmland
https://www.blocket.se/annons/1002000176	369900	Företag	Miljöbränsle/Hybrid	Automat	9072	2022	Kombi	Fyrhjulsdriven	355	Silver	1969	3/14/2022	Volvo	V90	Värmland
https://www.blocket.se/annons/1001969280	489900	Företag	Miljöbränsle/Hybrid	Automat	7986	2022	Kombi	Fyrhjulsdriven	463	Grå	1969	4/21/2022	Volvo	V90	Värmland
https://www.blocket.se/annons/1002018653	364900	Företag	Bensin	Automat	3519	2024	SUV	Tvåhjulsdriven	165	Blå	1969	10/16/2023	Volvo	XC40	Värmland
https://www.blocket.se/annons/1002220614	559900	Företag	Miljöbränsle/Hybrid	Automat	1260	2024	SUV	Fyrhjulsdriven	355	Grå	1969	12/20/2023	Volvo	XC60	Värmland
https://www.blocket.se/annons/1401632994	105000	Privat	Diesel	Manuell	14000	2013	Sedan	Tvåhjulsdriven	164	Vit	1984	5/22/2013	Volvo	S60	Värmland
https://www.blocket.se/annons/1002266372	499900	Företag	Miljöbränsle/Hybrid	Automat	6745	2024	SUV	Fyrhjulsdriven	355	Silver	1969	9/11/2023	Volvo	XC60	Värmland
https://www.blocket.se/annons/1002452638	687400	Företag	Miljöbränsle/Hybrid	Automat	0	2025	SUV	Fyrhjulsdriven	355	Röd	1969		Volvo	XC60	Värmland
https://www.blocket.se/annons/1002114996	309900	Företag	Diesel	Automat	7212	2017	Sedan	Fyrhjulsdriven	238	Ljusblå (Blå)	1969	10/19/2016	Volvo	S90	Värmland
https://www.blocket.se/annons/1002046164	534900	Företag	El	Automat	43	2025	SUV	Tvåhjulsdriven	256	Svart	0	6/4/2024	Volvo	EC40	Värmland
https://www.blocket.se/annons/1002224394	339900	Företag	Bensin	Automat	11903	2021	SUV	Fyrhjulsdriven	200	Vit	1969	4/13/2021	Volvo	XC40	Värmland
https://www.blocket.se/annons/1002277756	509900	Företag	El	Automat	496	2025	SUV	Tvåhjulsdriven	256	Grön	0	6/7/2024	Volvo	EX40	Värmland

Figur 3: Skärmdump av datan som samlades in i grupp från Blocket, specifikt den sammanställda delen

4.3 Undersökning av teoretiska antaganden

4.3.1 Icke-linjärt förhållande mellan den beroende variabeln och de oberoende variablerna

Vid undersökning av sambandet mellan pris och prediktorerna upptäckte jag att antagandet om linjäritet inte höll, särskilt för variablerna ålder och miltal. Dessa uppvisade ett icke-linjärt samband med försäljningspriset, vilket är rimligt då till exempel priset tenderar att minska snabbt vid låg ålder men planar ut för äldre bilar.

För att hantera detta tillämpade jag följande transformationer

- Log-transformation av Försäljningspris och Miltal, vilket hjälpte till att linjärisera förhållandet och stabilisera variansen.
- Polynom av grad två för Ålder *(*Age), för att modellera den krökta relationen mellan ålder och pris mer flexibelt.

Dessa åtgärder förbättrade modellens förklaringsgrad och säkerställde att linjäritetsantagandet uppfylldes i den slutgiltiga modellen.

4.3.2 Korrellerade residualer - Ej oberoende residualer

För att undersöka om residualerna var korrelerade använde jag Durbin-Watson testet. Testet gav värdet DW = 2.156 med ett p-värde = 0.9876. Detta indikerar att det inte finns någon signifikant autokorrelation i residualerna, vilket innebär att antagandet om oberoende residualer är uppfyllt.

4.3.3 Icke-konstant varians på residualerna (Heteroskedasticitet)

Inledningsvis bröts antagandet om homogen varians (homoskedasticitet), vilket framgick tydligt av Breusch-Pagan testet:
Före log-transformation: BP = 298.87, df = 42, p < 2.2e-16
Efter log-transformation av Försäljningspriset: BP = 171.2, df = 23, p < 2.2e-16

Även om testet fortfarande indikerar signifikant heteroskedasticitet, har variationen minskat avsevärt. Eftersom jag dessutom tillämpar robust regression, som är mer tålig mot icke-konstant varians, bedöms detta antagande som tillräckligt hanterat i den slutliga modellen.

4.3.4 Ej normalfördelade residualer

I ett tidigt skede av analysen visade residualerna tydliga avvikelser från normalfördelning. Som en åtgärd log-transformerade jag Försäljningspriset, vilket minskade vissa avvikelser. Trots detta fortsatte nya extremvärden att uppstå efter rensning. Därför valde jag att tillämpa robust regression, som är särskilt effektiv vid hantering av icke-normalt fördelade residualer och outliers.

Q-Q plott för residualerna (se bild nedan) visar avvikelser i nedre svansen, särskilt för datapunkterna 410 och 479. Vid granskning kunde jag inte identifiera några fel i dessa observationer. Givet att datasetet är tillräckligt stort, och att robust regression reducerar påverkan från enstaka avvikare, bedöms detta antagande som tillräckligt hanterat.

Shapiro-Wilk testet gav:
W = 0.9587, p-värde = 1.278e−14
vilket bekräftar att residualerna inte är normalfördelade på ett statistiskt signifikant sätt. Den praktiska påverkan på modellen är dock begränsad på grund av robust regressions användning.

4.3.5 Outliers

För att hantera outliers identifierade jag observationer med standardiserade residualer större än |3|. Dessa observationer togs bort då de bedömdes ha oproportionerlig påverkan på modellens koefficienter.

Utöver detta använde jag robust regression, vilket automatiskt minskar vikten av observationer med stora residualer, även om de inte formellt klassificeras som outliers. Genom denna metod kombineras datarensning med en modell som är mindre känslig för extremvärden vilket ger en mer stabil och tillförlitlig modell.

4.3.6 “High leverage” punkter

För att identifiera observationer med "high leverage” använde jag Cook's Distance som mått. Observationer som hade Di > 4/n betraktades som potentiellt inflytelserika och togs bort från modellen för att minska deras påverkan.

Robust regression minskar även vikten av observationer som är “high leverage” vilket stärker modellens stabilitet och tillförlitlighet.

4.3.7 Kollinearitet/multikollinearitet

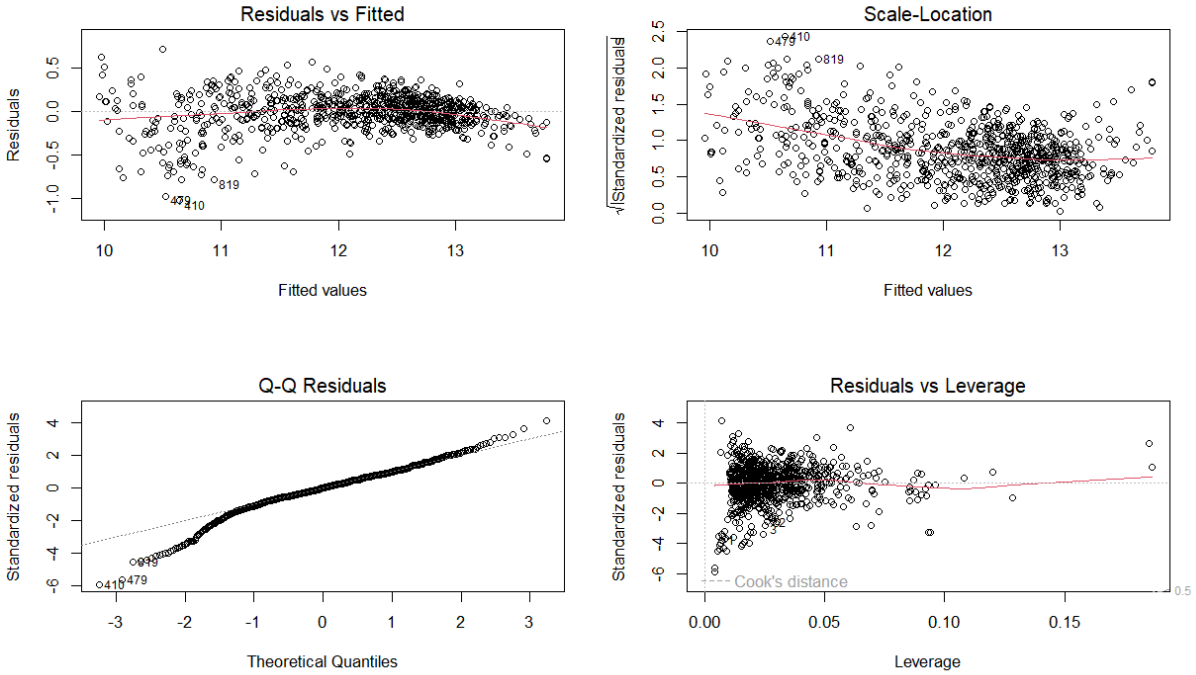
Vid undersökning kring kollinearitet upptäckte jag att antagandet om låg multikollinearitet bröts vid flera tillfällen, vilket framkom genom höga VIF-värden. För att åtgärda detta genomfördes följande steg:

- Variabeln Biltyp justerades genom att ta bort nivåerna Cab och Coupé, då dessa hade för få observationer för att ge statistiskt tillförlitliga skattningar.
- Variabeln Region uteslöts eftersom den främst användes för gruppindelning vid datainsamling och inte bidrog till prediktion av pris. Dessutom bidrog den till aliaserade koefficienter och ökade multikollinearitеten.
- Faktorn Modell förenklades med hjälp av fct_lump() i R, vilket grupperade ovanliga nivåer till en gemensam kategori (Other). Detta minskade överanpassning och instabilitet i modellens koefficienter.
- Efter dessa justeringar låg samtliga VIF-värden under gränsen 5, vilket indikerar att multikollinearitet inte längre utgör ett problem i den slutgiltiga modellen.

Slutgiltig VIF score:

```
> vif(model_robust)
                  GVIF Df GVIF^(1/(2*Df))
poly(Age, 2)   7.221144  2        1.639274
log_Miltal     2.883121  1        1.697976
Hästkrafter    4.679092  1        2.163121
Säljare        1.579324  1        1.256712
Bränsle        8.899467  3        1.439552
Växellåda      1.829046  1        1.352422
Biltyp       144.931420  3        2.291890
Drivning       2.397589  1        1.548415
Modell       472.868812 10        1.360621
```

Tabell 1: VIF score



Figur 4: Datapunkter för 410 och 479 var t.ex. flaggade i Q-Q plotten med stora negativa residuala effekter, men efter att inspekterat datapunkterna kunde jag ej se några fel. Robust regression minskar deras påverkan.

4.4 Analys av regressionsmodellen

Innan vi gick över till en robust regressionsmodell hade vi ett R^2 på 0.936. Efter att ha bytt till rlm kunde vi räkna ut ett "pseudo" R^2 och då fick vi ett liknande värde: 0.9399906

Se nedan tabell för koefficienter:

Variabel	Estimate	Std. Error	t value	p-value
(Intercept)	12.76777	0.13461	94.84946	0.00E+00
poly(Age, 2)1	-16.17129	0.40797	-39.63852	3.22E-192
poly(Age, 2)2	4.00663	0.26676	15.01942	3.46E-45
log_Miltal	-0.09588	0.01064	-9.00835	1.46E-18
Hästkrafter	0.00107	0.00018	5.92921	4.50E-09
SälljarePrivat	-0.15961	0.02004	-7.96378	5.61E-15
BränsleDiesel	0.12404	0.02083	5.95395	3.89E-09
BränsleEl	-0.15785	0.04481	-3.52285	4.51E-04
BränsleMiljöbränsle/Hybrid	-0.03581	0.02903	-1.23357	0.2177
VäxellådaManuell	-0.13037	0.02067	-6.30679	4.67E-10
BiltypKombi	0.06168	0.06046	1.02005	0.308
BiltypSedan	0.03755	0.06589	0.56989	0.5689
BiltypSUV	-0.0569	0.07084	-0.80324	0.4221
DrivningTvåhjulsdriven	-0.10363	0.02085	-4.97066	8.14E-07
ModellV40	0.06094	0.07935	0.76801	0.4427

ModellV50	-0.37623	0.06856	-5.48758	5.44E-08
ModellV60	0.00922	0.05831	0.15813	0.8744
ModellV60 Cross Country	-0.05421	0.06487	-0.83573	0.4035
ModellV70	-0.1471	0.06119	-2.40376	0.0165
ModellV90	0.07984	0.06081	1.31291	0.1896
ModellXC40	0.2392	0.06658	3.59245	3.47E-04
ModellXC60	0.26064	0.07007	3.71995	2.13E-04
ModellXC90	0.55097	0.07593	7.25675	9.26E-13
ModellOther	0.14341	0.05112	2.80533	0.0051

Tabell 2: Koefficienter och effekter i regressionsmodellen

4.4.1 Signifikanta variabler

Nedanstående variabler valdes för deras statistiska signifikans ($p < 0.001$) och relevans för att förstå prisvariationer, såsom ålder och miltal.

- Age(Ålder)

Bilens ålder har ett icke-linjärt samband med försäljningspriset, vilket innebär att en linjär modell riskerar att ge en missvisande bild. För att hantera detta inkluderades en polynomterm av grad två för variabeln ålder. Detta möjliggör en böjd modellanpassning som bättre speglar verkligheten (se figur nedan).

De skattade koefficienterna för polynomkomponenterna var:

poly(Age, 2)1: -16.1713

poly(Age, 2)2: 4.0066

Tolkningen av detta är att nya bilar tappar mycket i värde initialt, medan värdeminskningen planar ut för äldre bilar. Denna icke-linjära relation är typisk för andrahandsmarknaden och modellen fångar detta samband väl.

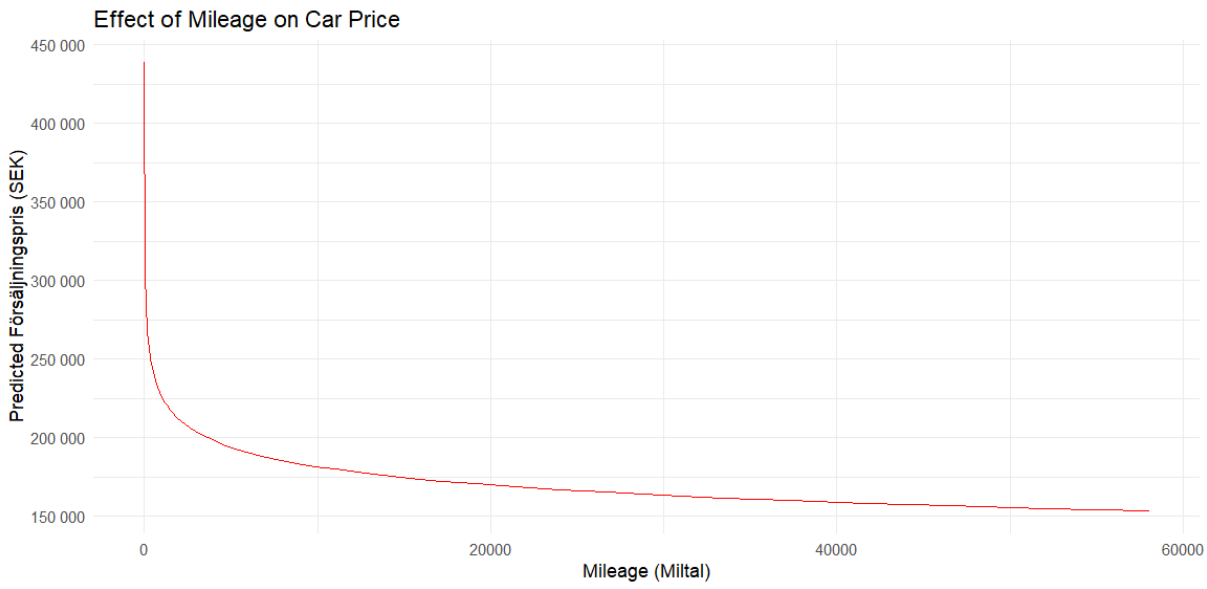
Figur 5: Graf över bilens ålder och dess påverkan på pris

- Miltal:

Även Miltal uppvisar ett icke-linjärt samband med försäljningspriset. Det är sällan så att varje extra kilometer påverkar priset lika mycket oavsett nivå vilket också syns tydligt i grafen nedan.

För att modellera detta mer korrekt tillämpades en log-transformation av miltalet, vilket resulterar i en log-linjär relation mellan variabeln och priset.

Enligt den skattade modellen innebär detta att en 1 % ökning av miltal är associerad med en 0.0959 % minskning i pris. Denna effekt är i linje med förväntningar och förstärker modellens förklaringsvärde.

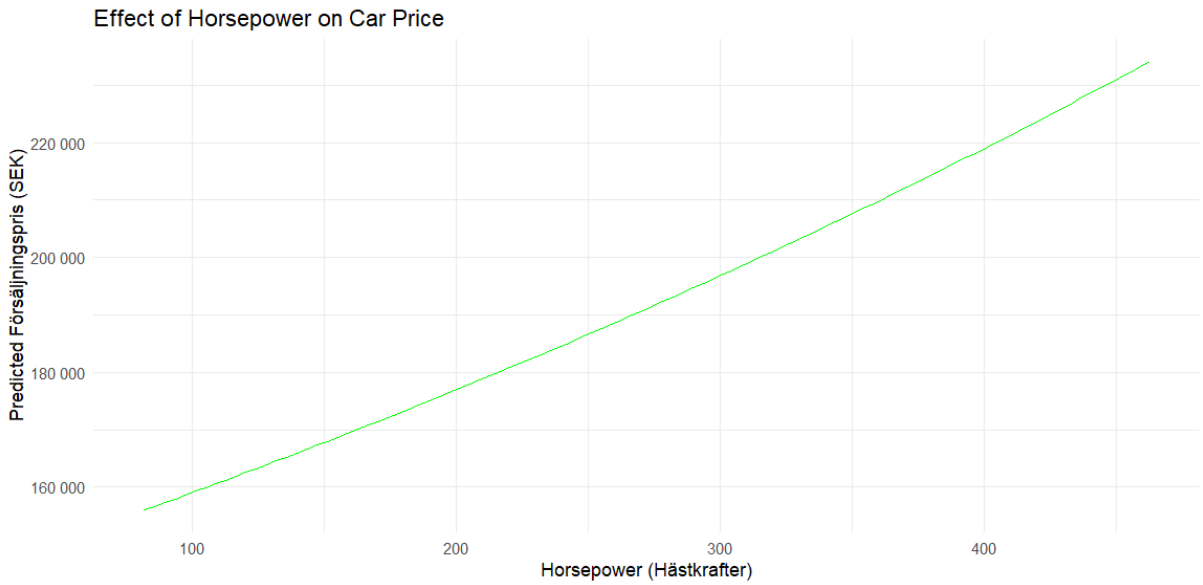


Figur 6: Graf över bilens miltal och dess påverkan på pris

- Hästkrafter:

Variabeln Hästkrafter visar ett positivt samband med log-priset. Enligt modellen innebär detta att en ökning med en hästkraft är lika med en ökning i log(pris) med cirka 0.0011 enheter.

I procentuell tolkning motsvarar detta att en extra hästkraft ökar det förväntade försäljningspriset med ungefär 0.11 %.



Figur 7: Graf över bilens hästkrafter och dess påverkan på pris

- SäljareTyp (Privat vs Företag):

Modellen inkluderar en dummyvariabel för säljartyp, där Privat säljare jämförs med referenskategorin Företag. Koefficienten för Privat är -0.1596 i log-prismodellen.

Denna koefficient kan tolkas som att bilar som säljs av privatpersoner är i genomsnitt cirka 14.75 % billigare än motsvarande bilar som säljs av företag, allt annat lika.

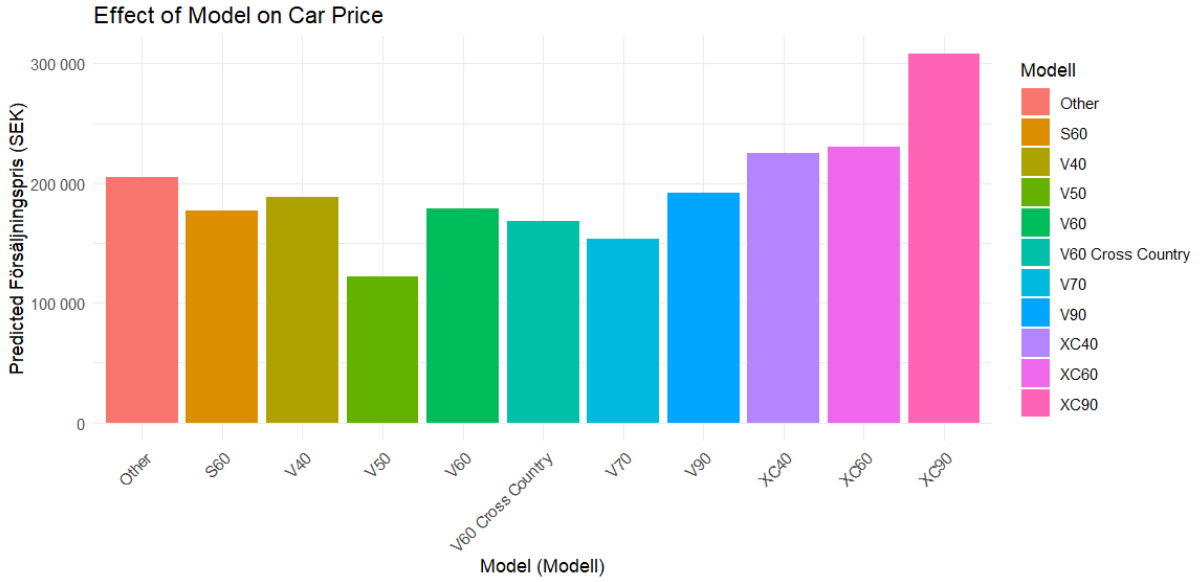
Detta beräknas enligt:

100×(e^-0.1596-1) = -14.75%

Figur 8: Graf över säljare typ och dess påverkan på pris

- Modell:

Regressionsresultaten visar att modellerna XC90, XC60 och XC40 är signifikant dyrare än övriga bilmodeller i datamängden. Deras positiva koefficienter i log-prismodellen tyder på att dessa SUV-modeller har ett högre förväntat försäljningspris, även efter kontroll för andra variabler såsom ålder, miltal och motorstyrka.



Figur 9: Graf över bilens modell och dess påverkan på pris

4.5 Hypotesprövning och konfidensintervall

För varje variabel i modellen har ett t-test använts för att pröva följande hypoteser:
- H0: Koefficienten är lika med 0 (ingen effekt)
- H1: Koefficienten är inte lika med 0 (det finns en effekt)

Vid ett p-värde < 0.05 förkastas nollhypotesen och vi säger att variabeln är signifikant. Nedan visas ett urval av modellens resultat:

Variabel	Värde	Std. Error	p-värde	95% CI	Effekt på pris
poly(Age, 2)1	-16.1713	0.4080	<0.001	[-16.971, -15.371]	Bidrar till stark prisnedgång för yngre bilar men som avtar med åldern
poly(Age, 2)2	4.0066	0.2668	<0.001	[3.483, 4.530]	Minskar prisnedgången för äldre bilar
log_Miltal	-0.0959	0.0106	<0.001	[-0.1167, -0.0751]	Minskar priset med 0.0959% per 1% ökning i Miltal
Hästkrafter	0.0011	0.0002	<0.001	[0.0007, 0.0015]	ca 0.11% per hästkraft
SäljarePrivat	-0.1596	0.0200	<0.001	[-0.1988, -0.1204]	−14.75% jämfört med företag

Tabell 3: Hypotesprövning och 95% konfidensintervall

T-testerna baseras på robusta standardfel från rlm, vilket ger tillförlitliga resultat trots heteroskedasticitet.

4.6 Prediktionsjämförelse och felmått på hela datasettet

För att utvärdera modellens prediktionsförmåga har jag jämfört det predikterade försäljningspriset med det faktiska försäljningspriset för ett urval av bilar. Modellen ger en blandning av över- och underskattningar, vilket är väntat i en verklig datamiljö.

4.6.1 Resultat av jämförelse mellan faktiska och predikterat pris

Nr	Försäljningspris (SEK)	Predikterat pris (SEK)	Skillnad	Tolkning
1	388900	454085	+65 185	Överskattning
2	199900	254473	+54 573	Överskattning
3	319900	300872	-19 028	Underskattning
4	289800	240875	-48 925	Underskattning
5	54900	60243	+5 343	Överskattning
6	389000	366980	-22 020	Underskattning
7	44800	48228	+3 428	Överskattning
8	459800	409648	-50 152	Underskattning
9	299900	237077	-62 823	Underskattning
10	419800	472853	+53 053	Överskattning

Tabell 4: Jämförelse mellan predikterat pris och faktiskt pris

4.6.2 Felmått på hela datasettet

Jag beräknade tre vanliga felmått för hela datamängden och eftersom jag först inte valt att dela upp datan i tränings samt testdel så är detta en indikation på modellens anpassning till datan och en potentiellt inte lika realistisk bild av modellens prestation gentemot ny data. Se tabell nedan för felmått på träningsdata:

- Mean Absolute Error (MAE): 31 725 kr, genomsnittligt absolut fel per observation.
- Root Mean Squared Error (RMSE): 49 597 kr, förstärker effekten av stora fel och visar att vissa observationer har relativt stora avvikelser.
- Mean Absolute Percentage Error (MAPE): 16.84 %, modellen har i genomsnitt ett fel på ca 17 % av det verkliga priset.

Felmått	Värde	Tolkning
MAE	31 725 kr	Genomsnittligt absolut fel
RMSE	49 597 kr	Felmått som förstärker stora avvikelser
MAPE	16.84 %	Genomsnittligt procentuellt fel

Tabell 5: Felmått för hela datamängden

Dessa felmått visar att modellen presterar relativt väl, särskilt givet variationen i bilpriser. Dock bör man vara medveten om att vissa enskilda prediktioner kan avvika kraftigt från verkligheten, vilket reflekteras i skillnaden mellan MAE och RMSE.

4.7 Utvärdering av modell när data delas upp i träningsdel samt testdel

4.7.1 Felmått för modell på testdata

Felmått	Värde	Tolkning
MAE	34069 kr	Genomsnittligt absolut fel
RMSE	46189 kr	Felmått som förstärker stora avvikelser
MAPE	26.64 %	Genomsnittligt procentuellt fel

Felmåtten för testdelen visar något högre värden jämfört med analysen på hela datan (se avsnitt 4.6), vilket är förväntat eftersom modellen inte tidigare sett testdatan. Detta visar att modellen har god, men inte perfekt, generaliseringsförmåga. Skillnaden mellan MAE och RMSE understryker även att vissa enskilda observationer avviker kraftigt från de predikterade värdena.

En MAPE på 26,64 % innebär att modellens prediktioner i genomsnitt avviker med cirka 26,64 % från det faktiska försäljningspriset på testdatan. För bilprisprediktion kan detta anses vara en acceptabel nivå av fel i vissa sammanhang, särskilt eftersom bilpriser påverkas av många faktorer som inte fångas i datan, såsom bilens skick, marknadstrender eller förhandlingsutrymme. För en köpare eller säljare kan dock en avvikelse på drygt 26 % vara betydande till exempel, för en bil med ett verkligt

pris på 300 000 kr kan det predikterade priset variera med upp till 80 000 kr, vilket kan påverka beslutsfattandet. För att förbättra modellens praktiska användbarhet skulle ytterligare variabler, såsom bilens skick eller servicehistorik, kunna inkluderas, och en korsvalideringsteknik som k-fold cross-validation skulle kunna användas för att säkerställa en mer robust generalisering till osedd data.

5 Slutsatser

- Vilken regressionsmodell passar bäst utifrån vår insamlade data?

Den modell som bäst hanterar avvikelser från klassiska antaganden (såsom heteroskedasticitet, outliers och icke-normalfördelade residualer) är robust regression med Hubor-estimator. Den slutgiltiga modellen har hög förklaringsgrad (justerat $R^2 \approx 0.94$) och uppfyller samtliga regressionsantaganden i tillräcklig grad.

- Vilka variabler är mest signifikanta?

De mest signifikanta förklarande variablerna i modellen är:
Ålder, Miltal, Hästkrafter och Säljartyp (Privat/Företag). Dessa variabler har alla p-värden < 0.001 och tydligt tolkbara effekter på log-transformationen av priset.

En avslutande slutsats är att modellen i praktiken fungerar som ett prediktionsverktyg för att uppskatta pris på Volvobils-annonser och inte ett faktiskt slutgiltigt försäljningspris, eftersom vi inte vet om bilen såldes eller inte, eller till vilket pris. Modellen uppvisar en felmarginal på 26.64 % (MAPE), vilket är acceptabelt givet variationen i annonser kring bilpriser.

En förbättring vore att samla in det faktiska slutpriset för varje annons (om tillgängligt), vilket skulle möjliggöra en direkt jämförelse mellan predikterat och faktiskt försäljningspris.

6 Teoretiska frågor

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

En Quantile-Quantile plot är en visualisering som kan användas för att undersöka ifall ett visst dataset följer en normalfördelning.

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, rätt förenkelt skulle jag hålla med om det påståendet som Karin säger ovan. I Maskininlärning så är det centrala att kunna göra prediktioner på antingen ett regressionsproblem eller ett klassifikationsproblem och fokus ligger ofta på att försöka skapa en så bra modell som möjligt som efterliknar världen och verkligheten medan analys av modellen är inte lika centralt. I statistisk regressionsanalys används det till både att göra prediktioner och att analysera datan och försöka tolka de olika parametrar och variabler. T.ex. i mitt arbete ovan kan vi prediktera ett pris på en Volvo men vi har också insett, tolkat och analyserat hur de olika variablerna påverkar vår modell.

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

Ett konfidensintervall beskriver med vilken procentuell säkerhet (oftast 95%) som vi kan definiera ett genomsnitt. Ett prediktionsintervall beskriver med vilken procentuell säkerhet vi kan prediktera ett nytt värde.

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

$\beta 0$ är interceptet, dvs det värdet Y startar på. $\beta 1$ är lutningskoefficient, med andra ord den genomsnittliga förändringen i Y när X ökar med en enhet.

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?

Ja eftersom i statistisk regressionsmodellering används ofta hela datasettet för att träna en modell. Vid en sådan situation passar det att tillämpa BIC främst för att undvika överanpassning men även för att BIC hjälper till att välja rätt modell.

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 $k = 1, 2, \ldots p$:
   (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 $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using using the prediction error on a validation set, C_p (AIC), BIC, or adjusted R^2. Or use the cross-validation method.

Algoritmen testar olika prediktorer/variabler för att hitta den bästa modellen utifrån AIC, BIC eller R^2

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

Statistiska modeller kan användas som ett verktyg, hjälpmedel till oss för att fatta beslut, hitta mönster, göra prediktioner osv. Men det är sällan (om ens någonsin?) en modell har en 100% accuracy eller tillförlitlighet och med det i åtanke så är alla modeller felaktiga. En modell är en förenklad version av världen så som vi ser den.

7 Självutvärdering

1. Vad tycker du har varit roligast i kunskapskontrollen?

Det roligaste är att se modellens resultat och hur de olika variablerna påverkar det predikterade priset. Det är väldigt intressant att se och det känns verkligen som att det är verklighetsförankrad och något man kan sätta in i och använda i praktiken.

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

Utmaningar har varit att skriva en väl-formulerad rapport som inte blir överflödig. Det är svårt när man sätter sig in i ett nytt ämne och skriva en rapport om det. Men samtidigt är processen väldigt bra för att man lär sig väldigt mycket under rapport-skrivandes stund. Detta gäller även den teoretiska biten för statistisk. Det är väldigt mycket att lära sig om hur man ska tolka datan och vad som är viktigt.

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

Jag anser att jag förtjänar G för att jag har klarat alla kursens delmoment.

4. Något du vill lyfta till Antonio?

Tack igen för en väldigt spännande kurs, ser fram emot sista kursen för läsåret.

Appendix A

https://github.com/salojoakim/R_kunskapskontroll

Källförteckning

- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2023). *An Introduction to Statistical Learning* (2nd ed.). Springer. https://www.statlearning.com/
- Rousseeuw, P. J., & Leroy, A. M. (2005). *Robust regression and outlier detection*. Wiley.
- FasterCapital. (n.d.). *Robust regression to tackle heteroskedasticity: A reliable approach*. https://fastercapital.com/content/Robust-Regression-to-Tackle-Heteroskedasticity--A-Reliable-Approach.html
- Reddy, M. (2011). *API design for C++*. Elsevier Science.
- Chatterjee, S., & Simonoff, J. S. (2013). *Handbook of regression analysis*. John Wiley & Sons.
- Cook, R. D., & Weisberg, S. (1983). Diagnostics for heteroskedasticity in regression. *Biometrika, 70*(1), 1–10. https://doi.org/10.1093/biomet/70.1.1
- Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika, 52*(3–4), 591–611. https://doi.org/10.1093/biomet/52.3-4.591
- Mendenhall, W., & Sincich, T. (1996). *A second course in statistics: Regression analysis* (5th ed.). Prentice-Hall.
- Wikipedia contributors. (n.d.). *Normalfördelning*. Wikipedia. Retrieved April 18, 2025, from https://sv.wikipedia.org/wiki/Normalf%C3%B6rdelning