# Analys av prisfaktorer för begagnade bilar

## En statistisk studie i R


William Kabpimai

EC Utbildning

Kunskapskontroll – R Programmering

2025.11

## Abstract

This report analyses the pricing of used Volvo cars using a multiple linear regression model in R. The study investigates how mileage, model year, and horsepower affect the selling price. The results show that these three factors explain approximately 88% of the price variation. The model's assumption was verified through diagnostic plots, confirming that the findings are reliable for predicting car values.

# Innehållsförteckning

# 1 Inledning

Syftet med denna rapport är att förstå vad som styr priset på begagnade bilar på marknaden. Genom att använda data från tidigare studenters insamling har vi undersökt om faktorer som miltal och bilens ålder, kan förklara det pris en köpare får betala.

## 1.1 Problemställning

Hur mycket sjunker en bil i värde för varje mil den körs och kan vi skapa en modell med hög säkerhet gissa priset på en bil?

# 2 Teori

I detta arbete används multipel linjär regression. Det är en statistisk metod för att se sambandet mellan en beroende variabel (pris) och flera oberoende variabler (miltal, årsmodell och hästkrafter).

## 2.1 Grunden i linjär regression

Linjär regression handlar om att hitta ett matematiskt samband mellan variabler. I vårt fall vill vi se hur priset förändras när miltal eller ålder ändras. Modellen skapar en rät linje som bäst representerar alla datapunkter vi har. Vi mäter modellens styrka med något som kallas förklaringsgrad: $R^2$ Om ($R^2$) är nära 1 betyder det att modellen är väldigt pricksäker.

**Modellens delar:**

- **Intercept (startvärde):** Detta är vad modellen gissar att en bil kostar om alla andra värden som miltal vore noll.
- **Koefficienter (beta-värden):** Dessa siffror visar lutningen på linjen. Om koefficienten för miltal är negativ så betyder det att priset sjunker när miltalet ökar.
- **Residualer (felen):** Skillnaden mellan det pris modellen gissar och det pris bilen faktiskt har kallas för residualer. Vi vill att dessa fel ska vara så små som möjligt.

## 2.2 Skillnaden mellan att gissa och förklara

Vi skiljer på två saker som vår modell kan användas till:

1. **Prediktion:** Att använda modellen för att gissa vad en specifik bil kommer att kosta.
2. **Statistiskt inferens:** Att försöka förstå varför prisskillnaden uppstår och hur starkt sambandet mellan miltal och pris faktiskt är.

# 3 Metod

Metoden som valts är en kvantitativ analys utförd i programvaran R.

## 3.1 Datahantering i R

Först lästes Excel filen in med paketet “readxl”. Därefter gjordes en datastädning i skriptet med funktionen “drop_na ()” och “mutate(as.numeric(....)). Detta gjordes för att rensa bort ofullständiga annonser och säkerställer att R tolkar text datatyper som siffror. Utan städningen skulle inte den matematiska modellen kunna köras.

## 3.2 Statistisk modellering

Valet av statistisk metod föll på en multipel linjär regression. Denna metod valdes för att den tillåter oss att isolera effekten av en variabel exempelvis miltal, samtidigt som vi kontrollerar för andra faktorer som bilens ålder. I skriptet formulerades modellen enligt ekvationen: Pris – Miltal + Modellår + Hästkrafter. Genom att använda funktionen “summary()” genererades koefficienter som beskriver lutningen på regressionslinjen, samt ett statistiskt mått som p-värden och R2.

## 3.3 Kvalitetssäkring och diagnostik

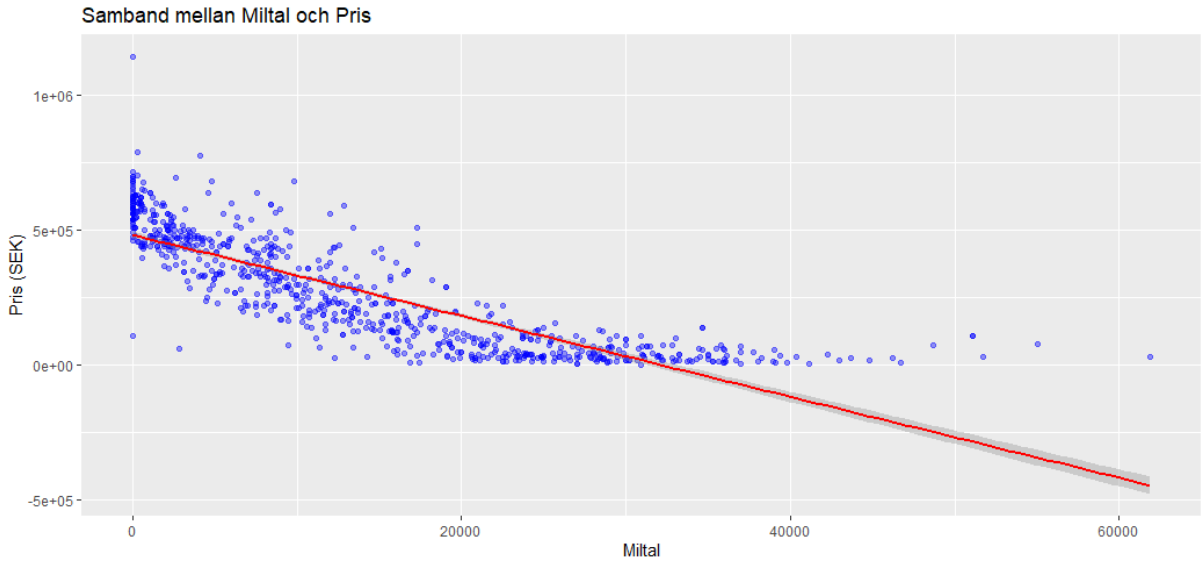
För att säkerställa att modellens resultat är pålitlig, genomfördes en diagnostisk kontroll av skillnaden mellan verkligt pris och modellens gissning. Detta gjordes visuellt i R med funktionen “plot(modell)”. Vi använder en QQ plot för att kontrollera att modellens fel är normalfördelad. Detta är ett krav för att vi ska kunna lita på resultaten och ifall våra slutsatser stämmer.

# 4 Resultat och Diskussion

När regressionen kördes med raden: lm(Försäljningspris - Miltal + Modellår + Hästkrafter) fick vi följande resultat:

## 4.1 Samband mellan miltal och pris

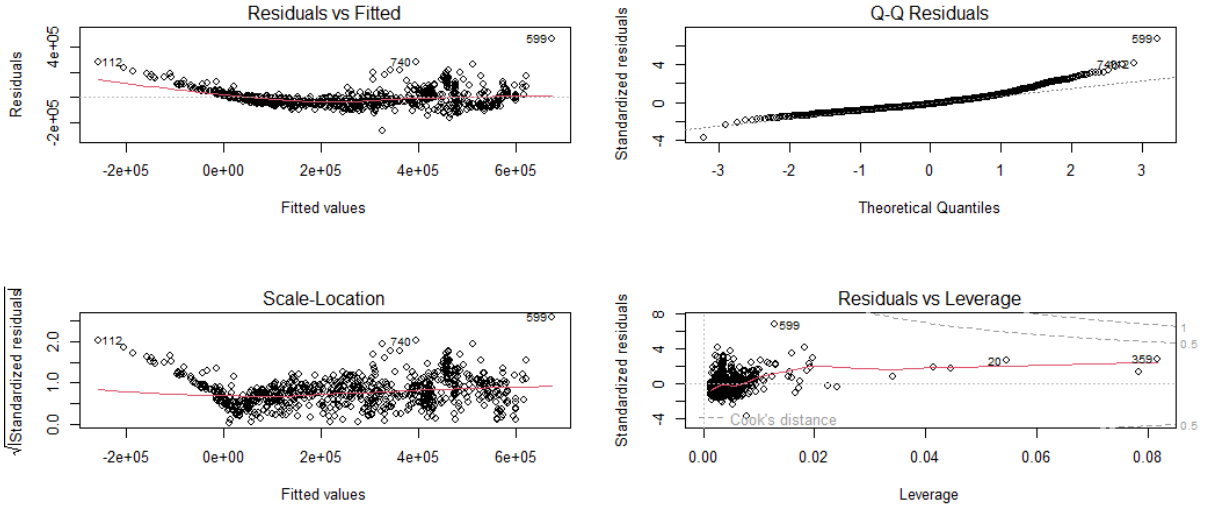
Innan regressionsmodellen skapades visualiserades sambandet mellan miltal och pris (se figur 1). Som förväntat visar grafen en negativ trend där priset sjunker i takt med att miltalet ökar. Den röda linjen representerar den genomsnittliga värdeminskningen.




(figur 1: Spridningsdiagram som visar hur priset (Y-axeln) påverkas av bilens miltal (X-axeln).

## 4.2 Diagnostik plottar

- **Miltal:** Resultaten visar en negativ koefficient på cirka –6,36. Det betyder att för varje mil bilen körs så sjunker priset med ungefär 6,40 kr när vi avrundar det.
- **Modellår:** För varje år "nyare" bilen är så ökar priset med cirka 9460 kr.
- **Förklaringsgrad:** Värdet blev 0,8836. Det innebär att hela 88,4 % av prisskillnaderna kan förklaras av de variabler vi valt.




(Figur 2: Diagnostiska grafer för att kontrollera regressionsmodellens statistiska antaganden.)

**Residuals vs Fitted:** Denna visar om det finns mönster som modellen missat. Här ser vi en lätt böjning, vilket tyder på att sambandet i verkligheten kan vara något mer komplext än en helt rak linje.

## 4.3 Diskussion om modellens pålitlighet

Genom koden "plot(modell)" genererades en QQ plot. Eftersom punkterna i grafen följer den diagonala linjen kan vi konstatera att modellens fel är normalfördelade. Detta bevisar att metoden vi har valt är statistisk korrekt för denna typ av data.

# 5 Slutsatser

Slutsatsen är att miltal är den största faktorn för värdeminskningen men att bilens ålder och kraft också spelar stor roll. Modellen vi byggt i R är mycket träffsäker och kan användas för att göra bra värderingar av begagnade bilar.

Studien visar att miltalet har en linjär negativ påverkan på priset, för varje mil bilen rullar minskar värdet med genomsnitt 6,40 kr. Samtidigt ser vi att bilen ålder är en kritisk faktor, där ett nyare modellår i snitt addera 9460 kr till värdet jämfört med året innan. Hästkrafter visade sig också vara en stark drivkraft för priset, vilket tyder på att köpare på marknaden är villiga att betala en premie för starkare motorer.

Sammanfattningsvis bekräftar de diagnostiska testerna, särskilt QQ ploten att modellen är stabil. Även om faktorer som bilens servicehistorik och utrustningsnivå inte fanns med i vår data (vilket förklarar de resterande procenten av variationen) ger denna regressionsmodell ett kraftigt verktyg för att objektivt värdera en bil och förstå marknadstrender.

# 6 Teoretiska frågor

1. Vad är en Quantile-Quantile (QQ) plot?

QQ-plotten är ett diagram som hjälper dig att snabbt se om dina data, exempelvis om felen i din statistiska modell följer normalfördelningen. Om det ser bra ut så ligger punkterna nära en rak linje. Om det ser dåligt ut så avviker punkterna från den raka linjen. Syftet är att kontrollera ett antagande för statistiska tester.

2. Skillnaden mellan prediktion och statistisk inferens.

**Prediktion:** handlar om att gissa det korrekta värdet på ett utfall och fokusen ligger på resultatet, exempelvis att exakt förutsäga priset på en bil baserat på dess attribut (miltal och ålder).

**Statistisk inferens:** handlar om att förstå och förklara sambandet mellan variablerna. Där fokusen ligger på relationen, exempelvis att fastställa om miltal eller ålder har en statistiskt negativ effekt på bilpriset.

**Slutsats:** prediktion fokuserar på vad (korrekt utfall), medan inferens fokuserar på varför (samband och effekt).

3. Vad är skillnad på "konfidensintervall" och "prediktionsintervall" för predikterade värden?

Båda intervallen används för att kvantifiera osäkerheten kring en prediktion, men de mäter osäkerheten för olika saker.

**Konfidensintervall:** mäter osäkerheten kring det genomsnittliga predikterade värdet för alla observationer med samma variabler (X). Det mäter hur säkra vi är på regressionslinjens position. Intervallet är snävare eftersom det bara tar hänsyn till osäkerheten i skattningen av medelvärdet.

**Prediktionsintervall:** mäter osäkerheten på ett individuellt nytt predikterat värde för en enda observation. Det mäter var en specifik ny datapunkt kommer att hamna. Intervallet är bredare eftersom det inkluderar både osäkerheten i medelvärdet och den naturliga slumpmässiga variationen.

**Slutsats:** konfidensintervallet mäter var medelvärdet av en grupp sannolikt ligger och intervallet är smal eftersom den bara mäter osäkerheten i linjens position. Medan prediktionsintervallet mäter var en enskild ny datapunkt sannolikt kommer att hamna och intervallet är brett eftersom den måste täcka linjens position plus den naturliga slumpvariationen.

4. Hur tolkas beta-parametrarna i den multipla linjära regressionsmodellen?

Beta-parametrarna kallas för regressionskoefficienter. De beskriver sambandet mellan varje oberoende variabel (x) och den beroende variabeln (Y).

**Interceptet**: beta-noll, enkelt sagt: Det är startvärdet för (Y). Definitionen är det förväntade värdet på (Y) om alla andra variabler (x) är noll.

**Koefficienterna:** beta-k anger hur mycket (Y) ändras när en av (X) variablerna ändras. Definition är den förväntade förändringen i (Y) när den specifika variabeln (xk) ökar med en enhet och vi låtsas att alla andra (X) variabler inte rör sig.

**Slutsats:** Koefficienterna (beta-k) visar alltså den unika effekten varje variabel har på utfallet (Y).

5. Behövs träning/validering/test set om man nyttjar BIC?

Att använda mått som BIC är att de innehåller en inbyggd "straffterm" för modellens antal variabler. Istället för att dela upp datan i träning och test, använder BIC hela datamängden men justerar resultatet matematiskt för att motverka överanpassning. Det är ett sätt att statistiskt uppskatta hur väl modellen skulle prestera på ny data utan att ha ett separat test-set.

6. Algoritmen för "Best subset selection".

Det är en metod för att hitta den bästa kombinationen av variabler till sin modell. Man börjar med en noll-modell som inte har några prediktioner, den gissar bara på medelvärdet. Efter det itererar vi (k = 1 till p) för varje antal variabler skapar man alla möjliga kombinationer av modellen med exakt (k) variabler. Sedan väljer man den lokalt bästa (Mk): för varje grupp, väljer man ut den modellen som har lägst fel (RSS) eller högst (R^2). Slutligen väljer man en enda vinnande modell bland alla utvalda (M0.....Mp) genom att använda mått som exempelvis BIC eller korsvalidering för att hitta den mest balanserade modellen.

7. Vad kan "All models are wrong, some are useful" innebära?

En modell är en förenkling av verkligheten. Man kan inte fånga varje detalj eller slumpmässig faktor som påverkar exempelvis bilpriset perfekt. Därför kan ingen modell exakt reflektera verkligheten. Men trots att de inte är perfekta, kan de vara väldigt användbara för att ta beslut, förstå trender eller göra gissningar. Syftet är att skapa ett verktyg som hjälper oss att förstå saker och ting bättre.

# 7 Självutvärdering

1. Vad tycker du har varit roligast i kunskapskontrollen?

- Att se hur Excel datan kunde förvandlas till en graf där man ser hur priset faller i takt med miltalen.

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

- Den störst utmaningen var att få R att förstå Excel filen. Jag löste det genom att lära mig använda rätt bibliotek och genom att tvätta datan så att alla värden blev numeriska.

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

- Tanken med uppgiften var att lösa ett "Godkänt". Jag tänker att arbetet ser tunn ut och har inte hunnit lägga på fler aspekter såsom hur bränsletyp också påverkar priset.
  Men jag har lyckats skapa en fungerande R-skript som läser in och tvättar data, utfört en multipel linjär regression med respektabelt resultat och verifierat det med en QQ plot.

# 8 Källförteckning


- Data: Insamlad av tidigare studenter från Blocket (Volvo annonser).