Vad bestämmer priset på en bil?

Regressionsanalys av bilpriser

Jakob Rask

EC Utbildning

Kunskapskontroll – R-programmering

202404

Abstract

Data was gathered groupwise from car adverts on Blocket (an online marketplace).

The purpose of this report is using regression analysis to create a model that can predict car prices and from which we can make conclusions about the population.

The most significant variables are the age of the car, number of horsepower and milage.

Predictions performed on test data resulted in a RMSE of 47375,60 kr.

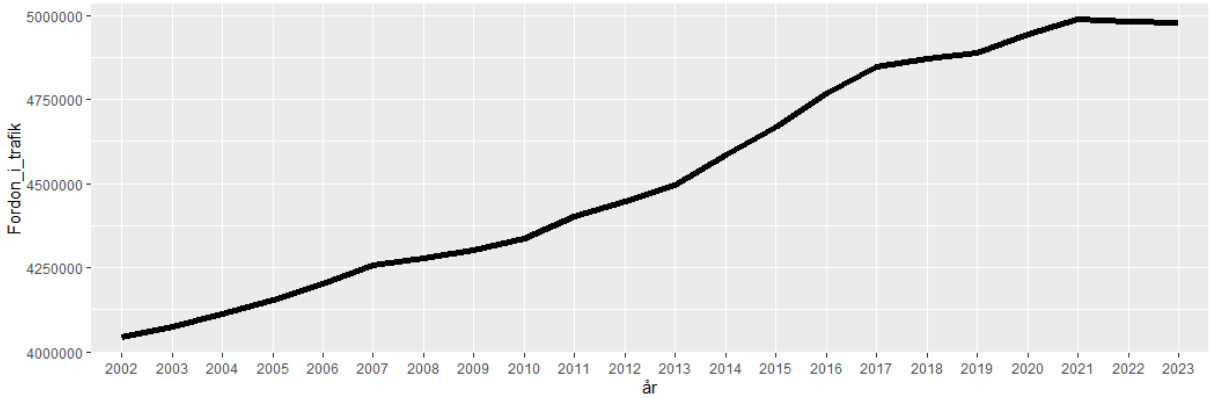
The price of a car is generally lower if the seller is a private person than if the seller is a company.

Innehållsförteckning

1 Inledning

Med regressionsanalyser kan vi på ett förenklat sätt förklara komplexa samband. Regressionsanalyser kan hjälpa oss att förstå hur olika faktorer påverkar den variabel vi vill undersöka. Ett exempel på regressionsanalys är att undersöka hur olika faktorer påverkar bilpriset och vilka av dessa faktorer som påverkar priset mest.

Trafikutvecklingen har ökat stadigt de senaste 20 åren, se figur nedan. En ökning av det totala antalet personbilar kan även få utbudet på andrahandsmarknaden öka. Detta medför ett större utbud på marknadsplatser som t.ex. Blocket (Blocket, 2024) vilket i sin tur genererar mer data avseende bilar och deras pris. Data som kan bli underlag för att göra regressionsanalyser avseende bilpriser.



Figur 1. Trafikutvecklingen avseende personbilar i Sverige sedan 2002 (SCB, 2024).

1.1 Syfte

Syftet med den här rapporten är att ta fram en regressionsmodell som på ett bra sätt visar förhållandet mellan den beroende variabeln och de oberoende variablerna, dvs i möjligaste mån återspeglar populationen/verkligheten. Samt går att använda för prediktering av priset på andra bilar av samma bilmärke.

1.2 Frågeställningar

Vilka faktorer har störst inverkan på priset?

Hur ser prediktionsintervallet ut jämfört med det sanna priset?

Skiljer sig prissättningen beroende på om säljaren är privatperson eller ett företag?

1.3 Avgränsningar

Insamlade data från bilannonser är begränsad till att endast vara av märket Volvo.

Bilar äldre än 25 år har valts bort för att göra bilar med olika bränsletyp mer jämförbara.

2 Teori

Multipel regressionsanalys

Multipel regressionsanalys innebär att man med hjälp av flera förklarande variabler analyserar variationen i den beroende variabeln (Körner, S. & Wahlgren, L. , 2015, s. 396).

Regressionslinje

Regressionslinjen är den linje som bäst beskriver de genomsnittliga sambandet mellan variablerna för en viss uppsättning observationspar (x, y) (Körner, S. & Wahlgren, L. ,2012, s. 161). Riktningskoefficienten för *x* anger hur mycket *y* ändras i genomsnitt då *x* ökar med 1 (Körner, S. & Wahlgren, L. , 2015, s. 370).

RMSE

RMSE visar prediktionernas medelavstånd (prediktionsfelet) till de sanna värdena och används som mått vid utvärdering av regressionsmodeller (Géron, 2019). RMSE beräknas med följande formel:

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

Outliers – för modellen

En outlier är en datapunkt vars y-värde ligger långt ifrån modellens predikterade y-värde. Outliers kan till exempel uppkomma på grund av felaktiga datavärden eller att modellen är bristfällig (James, G., Witten, D., Hastie, T. & Tibshirani, R., 2021).

3 Metod

Datainsamling har skett i en grupp om 5 personer. Gruppen samlade in data från totalt 750 bilannonser på Blocket och sammanställde detta i ett Excel-dokument. Gruppen valde att samla in data från enbart Volvobilar. Vid insamlingstillfället fanns ca 14500 annonser för Volvobilar (exkluderat leasing).

Data från SCB har inhämtats genom användande av API i R-koden.

3.1 Bearbetning och undersökning av data

Excel-dokumentet med insamlade data har rättats till avseende mindre inskrivningsfel och liknande. Saknade värden och dubbletter hanteras och data filtreras med avseende på ålder.

Genom att göra parvisa plottar kan vi tydligt se mönster mellan de numeriska variablerna, samt variablernas respektive relation till priset. Undersökning görs även för de kategoriska variablerna (Säljare, Bränsle, Växellåda, Biltyp, Drivning, Färg och Modell).

Data delas upp i träningsdata (80 % av antalet observationer) och testdata (20 %).

3.2 Modellval, anpassning och utvärdering

Utifrån analyserna av variablerna och deras eventuella samband väljs variabler för en regressionsmodell. För att identifiera de mest relevanta variablerna tar vi även hjälp av metoden ”best subset selection”.

Genom att testa olika modeller och diagnosticera för eventuella problem, förbättras modellen succesivt. Utvärdering sker avseende RMSE genom cross-validation på de modeller som klarat sig bäst.

Den modell som anses bäst används för att prediktera testdata och utvärderas avseende RMSE. Modellen utvärderas även med avseende på statistisk inferens.

3.3 Datainsamling – frågor

1. Vem du har arbetat i grupp med?

Datainsamlingen har utfört tillsammans med Melissa Hansson, Adrian Andersson Krsmanovic, Keikiet Pham och Robert Shaw.

2. Hur har ni i gruppen arbetat tillsammans?

Gruppen diskuterade syftet med datan, utifrån vad respektive person hade för avsikt att analysera. Insamlade data är jämnt fördelad över de fyra bränslekategorierna (Bensin, Diesel, El och Miljöbränsle/Hybrid) för att möjliggöra jämförande analyser.

3. Vad var bra i grupparbetet och vad kan utvecklas?

Bra diskussioner kring syftet med insamlade data samt tillvägagångssätt för insamling. Kan förbättra tillvägagångssättet så att risken för dubbletter i data minimeras.

4. Vad är dina styrkor och utvecklingsmöjligheter när du arbetar i grupp?

Jag lyssnar på övriga gruppmedlemmars tankar och bidrar med idéer kring det.

5. Finns det något du hade gjort annorlunda? Vad i sådana fall?

Att använda ett digitalt anteckningsverktyg (exempelvis Miro) hade underlättat. Det hade gjort det lättare för personer att förklara vad den menar, blir tydligare och minskar risken för missförstånd.

4 Resultat och Diskussion

4.1 Bearbetning av data

I Excel-dokumentet har en kolumn har lagts till där antal dagar i trafik beräknas som skillnaden mellan datumet för insamlandet och ”Datum i trafik” från annonserna. En bil hade fel ”Datum i trafik” (år 2022 för en årsmodell 2000) då den var registrerad i Norge först i ett flertal år. ”Datum i trafik” har därmed justerats till inköpsåret enligt annonsen (år 2000).

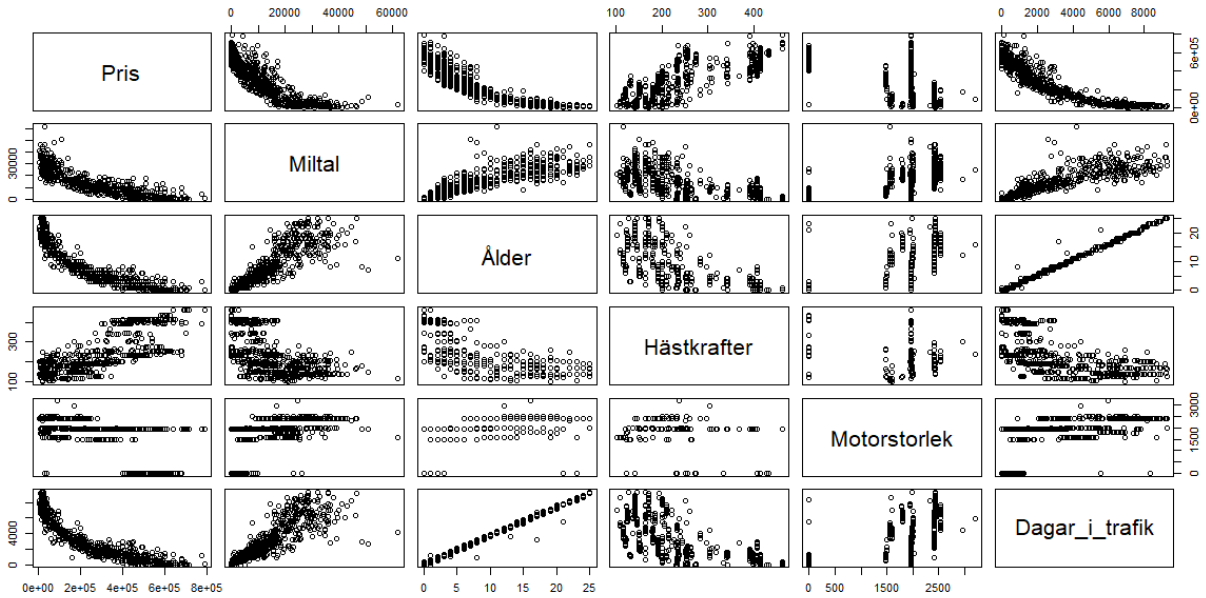
Dubbletter (som har uppstått då insamlandet från olika personer råkat överlappa varandra) tas bort för att inte skapa obalans bland observationerna. Insamlade data innehåller även saknade värden för en del observationer. Exempelvis saknar elbilar värde för variabeln ”Motorstorlek”, där sätts den till 0 tills vidare. De observationer som saknar värde för en eller flera variabler är få till antalet och tas bort. Datan filtreras för att exkludera bilar äldre än 25 år.

Max-värdet för ”miltal” är 190724 mil, vilket verkar orimligt högt (ca 10 gånger högre än värdet för tredje kvartilen). En trolig orsak är felangivelse i annonsen, att kilometertalet har angivits istället för miltal. Värdet divideras därför med 10 (avrundat till ett heltal), vilket bättre stämmer överens med andra bilar med motsvarande ålder.

4.2 Undersökning av data

Genom att göra parvisa plottar kan vi tydligt se mönster mellan de numeriska variablerna och få en uppfattning om hur väl de korrelerar.

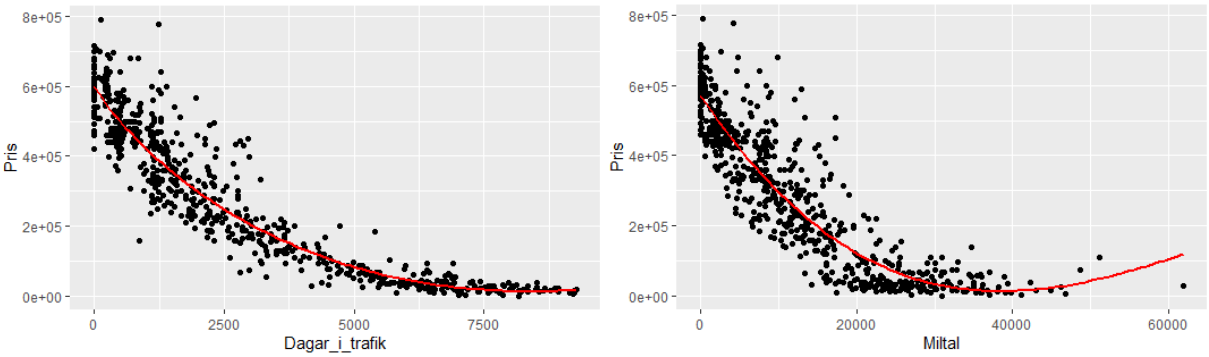
Högre miltal och högre ålder (även antal dagar i trafik) påverkar priset negativt. Ett högre antal hästkrafter verkar påverka priset positivt. Avseende motorstorlek syns inget uppenbart samband (denna faktor blir heller inte representativ då elbilarnas motorstorlek är satt till 0).



Figur 2. Parvis plot av de numeriska variablerna

Vi ser bland annat att miltal, ålder och antal dagar i trafik har liknande mönster när de jämförs med priset. Ålder och dagar i trafik har nästan perfekt korrelation. I fortsatta analyser används därför bara dagar i trafik.

Relationerna mellan priset och antal dagar i trafik respektive miltal har båda ett exponentiellt avtagande mönster, se figur nedan.



Figur 3. Relationen mellan priset och antal dagar i trafik (t.v.) respektive miltal (t.h.).

Vi undersöker även de kategoriska variablerna (Säljare, Bränsle, Växellåda, Biltyp, Drivning, Färg och Modell). Se även figurer i Appendix A.

När vi lägger till färg för om säljaren är privatperson eller företag, framkommer ett visst mönster. De bilar som säljs av privatpersoner är generellt sett äldre, har högre miltal och mindre hästkrafter. Detta är faktorer som också har påverkan på priset. Jämförelse av regressionslinjer för respektive kategori visar att bilar sålda av företag generellt sett har högre pris.

Utifrån bränsletyp kan vi se att elbilar generellt sett har lägre ålder och lägre miltal, vilket gör att de ligger högt i pris. Jämförelse av regressionslinjer för respektive kategori visar att bensin- och dieseldrivna bilar generellt sett ligger lägre i pris än elbilar och miljöbränsle/hybrid.

Manuellt växlade bilar återfinns generellt sett bland bilar som är äldre och har gått fler mil. Manuellt växlade bilar har även generellt sett mindre hästkrafter.

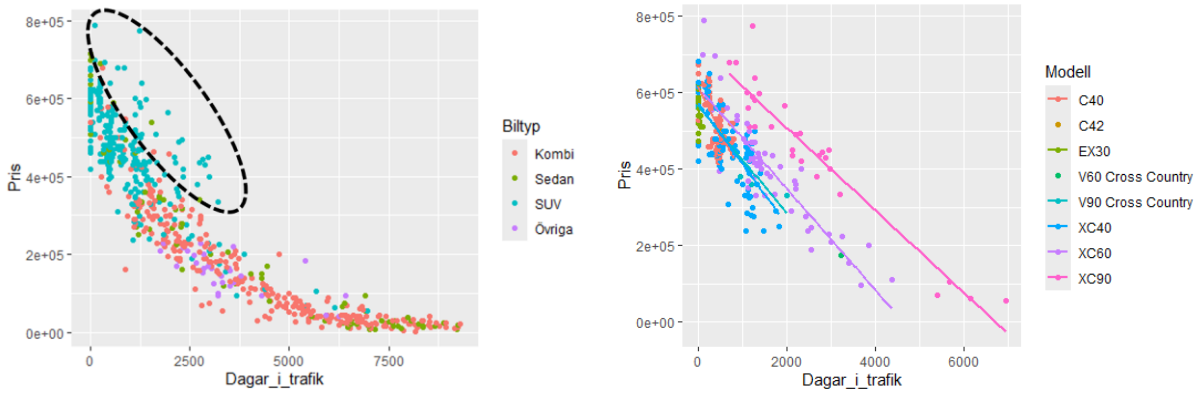
Kombi och Sedan är biltyper som finns med bland både nya och gamla bilar, medan SUV främst återfinns bland de nyare bilarna.

Fyrhjulsdrivna bilar ligger generellt sett något högre i pris, men det kan hänga samman med andra variabler.

Färg ser inte ut att ha något generellt samband med priset.

Modelltyp har delvis ett samband med priset, men det kan även hänga samman med andra variabler. Det finns ett tydligt samband mellan modelltyp och biltyp.

Merparten av bilarna inom biltypen SUV är samlade till vänster i grafen, dvs har ett högt pris och är relativt nya (få antal dagar i trafik). Då spridningen av priset inom kategorin SUV är ganska stor kollar vi närmare på hur det ser ut för de aktuella modelltyperna. Det visar sig att merparten av de bilar som generellt sett har ett högre pris utifrån antalet dagar i trafik, är av modell XC90 (se figur t.v. nedan). Därmed kan modell XC90 vara en intressant variabel att ha med i modellen.



Figur 4. Observationer inom biltyp SUV med högre pris (t.v.). Observationerna utgörs av modell XC90 (t.h.)

4.3 Modellval och modellanpassning

Ett första antagande är att bilens ålder och hur många mil den har gått spelar stor roll för priset. Andra intressanta parametrar kan antas vara bilmodell och drivmedel (bränsletyp).

För att få det mer bekräftat används metoden ”best subset selection” som ger svar på vilken/vilka variabler som ingår i den bästa modellen, beroende på antalet variabler som används.

Enligt metoden framkommer att variabler såsom antal dagar i trafik (ålder), hästkrafter, miltal, modelltyp XC90 och bränsletyp el är av störst betydelse för modellen.

Genom att göra en plot av modellen får vi fram olika grafer som kan användas för att diagnosticera eventuella problem med modellen.

4.3.1 Modell 1

En kontroll av F-testet och dess p-värde visar att åtminstone en av de beroende variablerna är användbara för att prediktera den beroende variabeln.

För den första modellen ser vi tecken på att modellen inte är linjär, utan har en tendens av kurvatur. Modellen visar också tecken på heteroskedasticitet (att variansen hos residualerna inte är konstant). Det framgår även att residualerna avviker från normalfördelningen samt att modellen ger upphov till möjliga outliers.

4.3.2 Modell 2

Då vi tidigare såg ett exponentiellt avtagande mönster för relationen mellan ”Pris” och ”Dagar i trafik” respektive ”Miltal”, prövar vi att transformera dessa variabler.

Modell 2 ger ett bättre resultat men viss heteroskedasticitet kvarstår.

4.3.3 Modell 3

För att försöka lösa problemet med heteroskedasticiteten prövar vi att transformera y-variabeln. Genom att ta kvadratroten av y-variabeln får vi ett mer acceptabelt resultat. Dock förekommer det en möjlig outlier som verkar ha ett lägre pris än genomsnittet utifrån de valda variablerna.

4.3.4 Modell 4

För att försöka skapa en modell som är bättre anpassad för denna möjliga outlier prövar vi med interaktioner med olika variabler. Resultatet blir generellt sett bättre men outliern finns fortfarande kvar.

4.3.5 Modell 5

En sista modell skapas med polynomfunktioner för att försöka åtgärda problemet med outliers. Modellen lyckas anpassa sig bättre så att inga outliers finns kvar.

4.4 Modellutvärdering

De tre modeller som visade på acceptabla resultat utvärderas med avseende på RMSE, som här fås genom K-fold cross-validation, se tabell nedan.

RMSE för valda modeller	
Modell 3 – transformation av y och x	37.64501
Modell 4 – transformation + interaction	35.81634
Modell 5 – transformation med poly	37.57875

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

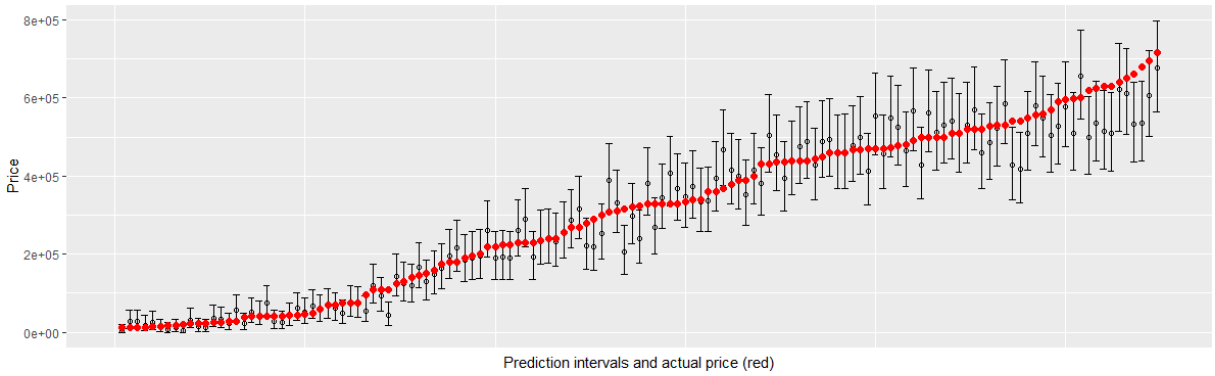
Eftersom y-variabeln är transformerad går det inte att tolka resultatet av RMSE korrekt men det går ändå att se att modell 4 får bäst resultat (lägst RMSE).

4.5 Utvärdering av den valda modellen

Modellen använder variablerna: hästkrafter, modelltyp XC90, säljartyp privatperson samt en interaktion mellan miltal och antal dagar i trafik. Y-variabeln (priset) är transformerat med kvadratroten (sqrt(y)).

En prediktion av testdata genomförs med den valda modellen, och resultatet visar ett RMSE på 47375,6 (efter att y-variabeln transformerats tillbaka). Resultatet av RMSE tolkas som att det predikterade priset i genomsnitt skiljer sig +/- 47375,60 kr från det sanna priset.

Genom att jämföra det sanna priset från testdata med ett 95% prediktionsintervall får vi en bra bild över intervallbredderna och hur väl intervallen täcker de sanna priset, se figur nedan.



Figur 5. 95% prediktionsintervall för testdata, med de sanna priserna som röda prickar.

Genom att studera på koefficienterna för de ingående variablerna kan vi utläsa att miltal och antal dagar i trafik har negativ inverkan på priset. Att en bil är av modelltyp XC90 har en positiv inverkan på priset. Bilar som säljs av privatpersoner har generellt sett ett lägre pris än bilar sålda av företag.

5 Slutsatser

Frågeställningarna från kapitel 1.2 kan vi nu ge svar på:

Vilka faktorer har störst inverkan på priset?

De faktorer (variabler) som har störst betydelse för priset är miltal, hästkrafter och antal dagar i trafik (ålder).

Hur ser prediktionsintervallet ut jämfört med det sanna priset?

Prediktionsintervallet jämfört med det sanna priset för varje observation i testdata framgår i Figur 5.

Skiljer sig prissättningen beroende på om säljaren är privatperson eller ett företag?

Ja, priset för en bil som säljs av ett företag ligger generellt sett högre än om säljaren är en privatperson.

Resultatet visar hur svårt det är att ”korrekt” prissätta en bil utifrån ett begränsat antal variabler. Det finns många andra aspekter som också kan spela roll för priset såsom skicket på bilen eller extern påverkan i form av trender eller hög-/lågkonjunktur. Men det går att skapa modeller som kan estimera ett generaliserat pris för bilar, om än med förhållandevis brett prisintervall. Med mer och bättre data går det troligtvis att skapa bättre modeller, som visserligen inte kommer vara helt korrekta men i alla fall användbara.

6 Teoretiska frågor

1. Beskriv kortfattat vad en Quantile-Quantile (QQ) plot är.
   En QQ-plot är en graf (visualisering) som används för att bedöma hur väl normalfördelade de observerade värdena är.

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?
   Med statistisk inferens menas att utifrån ett slumpmässigt urval kunna dra slutsatser om hela populationen. Det kan t.ex. vara en skattning av medellängden för män respektive kvinnor. Det kan även innefatta hypotesprövning, t.ex. ”Är män längre än kvinnor?”.

3. Vad är skillnaden på ”konfidensintervall” och ”prediktionsintervall” för predikterade värden?
   Konfidensintervall: Ett intervall som (med en viss konfidensgrad) förväntas täcka det genomsnittliga värdet. Exempelvis genomsnittslönen för en viss ålder, $Y = \beta_0 + \beta_1 X$
   Prediktionsintervall: Ett intervall som (med en viss konfidensgrad) förväntas täcka det sanna värdet av en predikterad observation. Exempelvis lönen för en viss individ, $Y = \beta_0 + \beta_1 X + \epsilon$

4. Den multipla linjära regressionsmodellen kan skrivas som: $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon$. Hur tolkas beta parametrarna?
   β_0 är interceptet, dvs där linjen skär y-axeln.
   β_j för $j \geq 1$ tolkas som: ”effekten på Y när x_j ökar med en enhet, givet att alla andra variabler är fixa.” Det vill säga om x_1 ökar med 1 så förändras Y med β_1.

5. Din kollega Hassan 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 det stämmer. Mått som till exempel BIC estimerar felet för testdata utifrån träningsdata.

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.

Börjar med att träna alla p antal modeller som har 1 predictor (X-variabel), sedan alla p över 2 antal modeller som har 2 predictors, osv fram till k antal predictors. Resulterar i bästa modellen för respektive k antal predictors (dvs bästa modellen med 1 predictor, bästa med 2 predictors, osv).

7. Ett citat från statistikern George Box är: “All models are wrong, some are useful.” Förklara vad som menas med det citatet.
   En modell är en generalisering av verkligheten och kommer inte vara korrekt i alla avseenden. Men vissa modeller kan vara tillräckligt bra för att vara användbara.

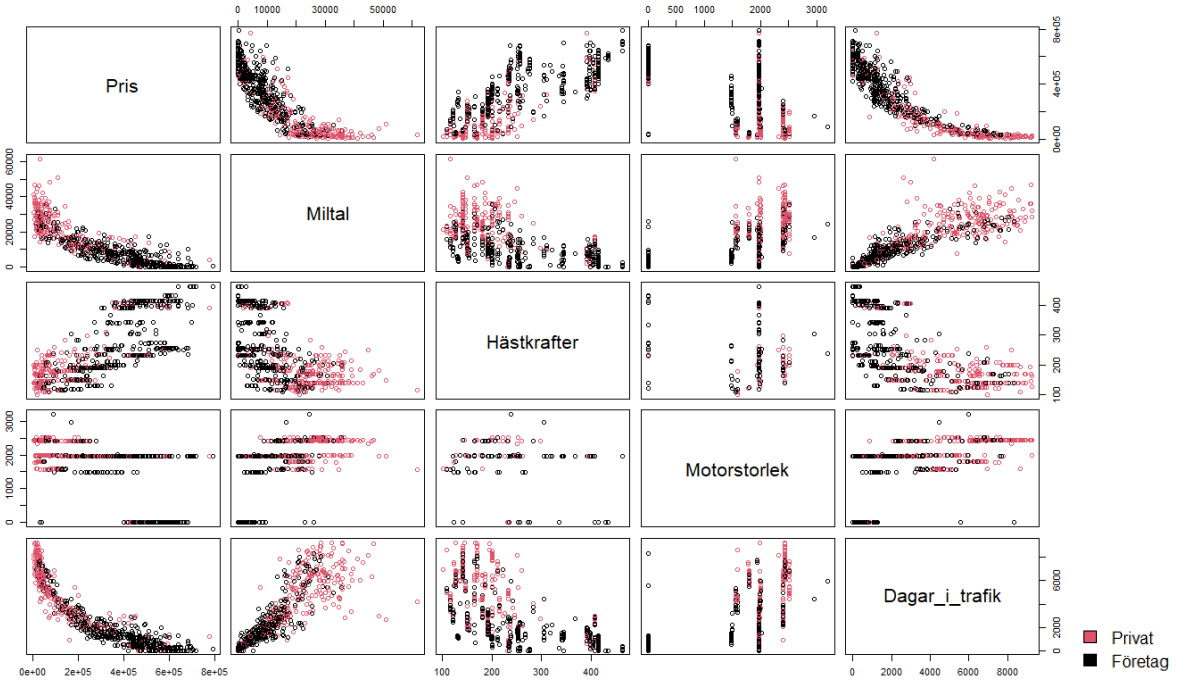
7 Självutvärdering

1. Utmaningar du haft under arbetet samt hur du hanterat dem.
   En utmaning har varit tidsdispositionen. Det är lätt att det går åt mycket tid för koden och undersöka data och testa olika modeller, men att rapporten kommer andra hand. Det jag har lärt mig nu är att jag behöver sätta egna ”deadlines” för olika moment samt att arbeta parallellt med både kod och rapport för att strukturen ska bli bättre.

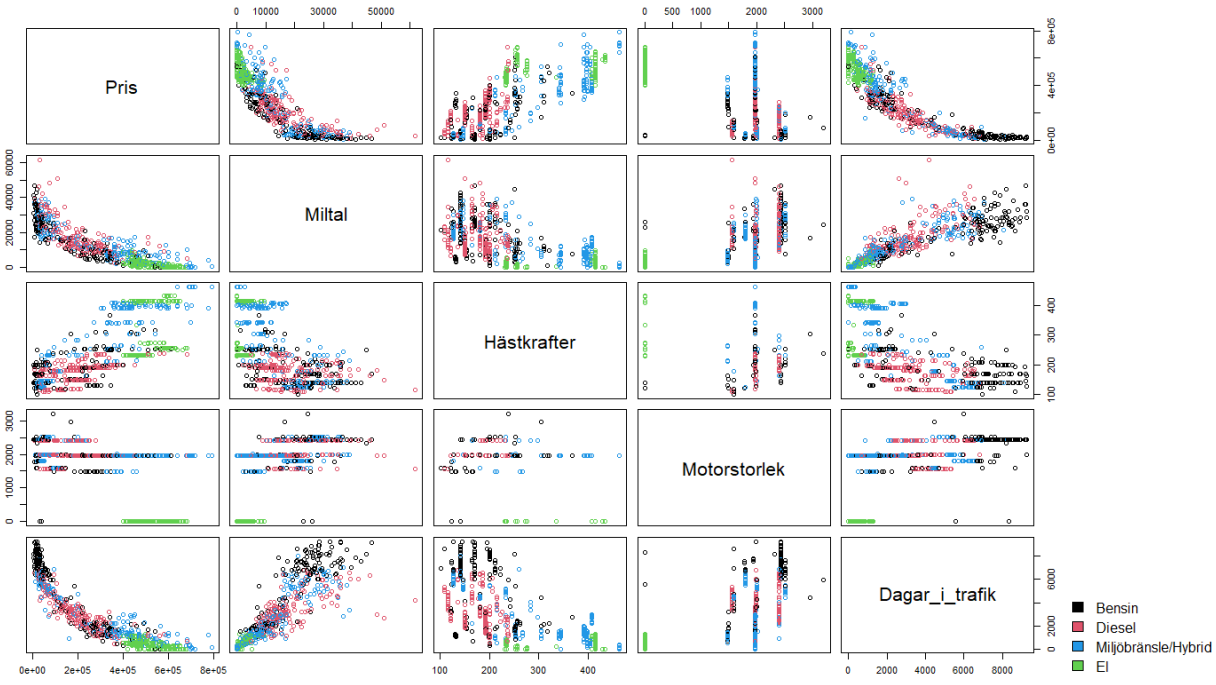
2. Vilket betyg du anser att du skall ha och varför.
   Möjligtvis G, lite oklart om detta når upp till kraven på VG. Beror på om kvaliteten är tillräckligt hög.

3. Något du vill lyfta fram till Antonio?

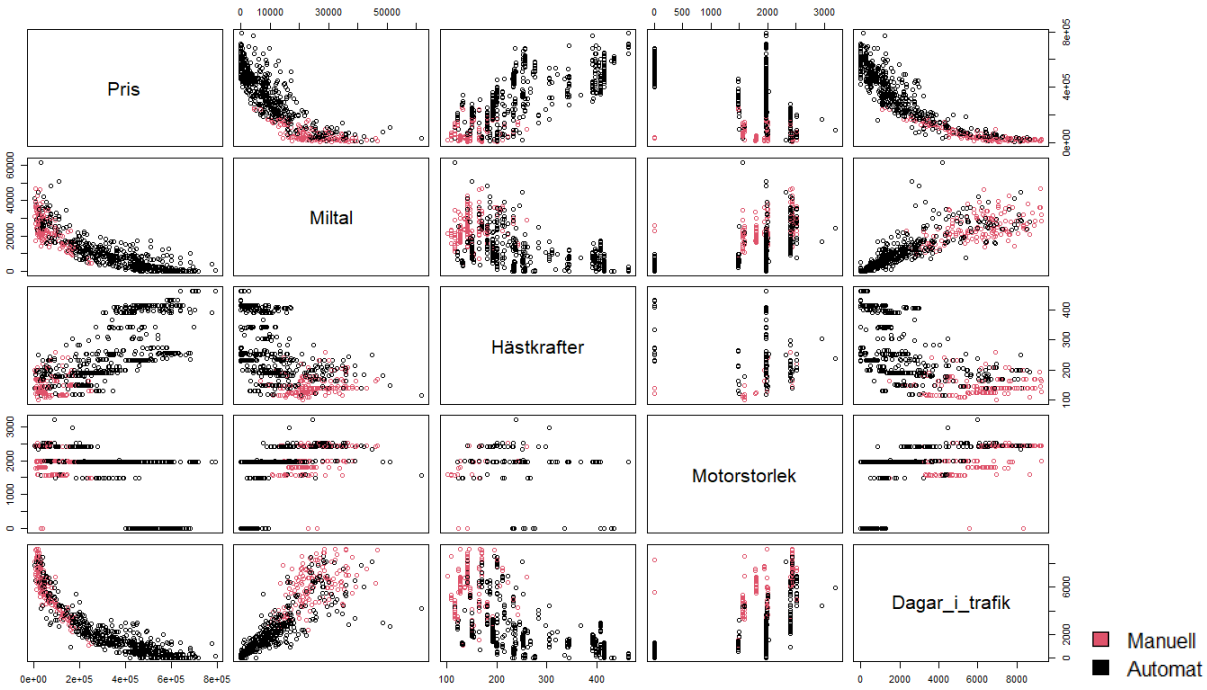
Appendix A



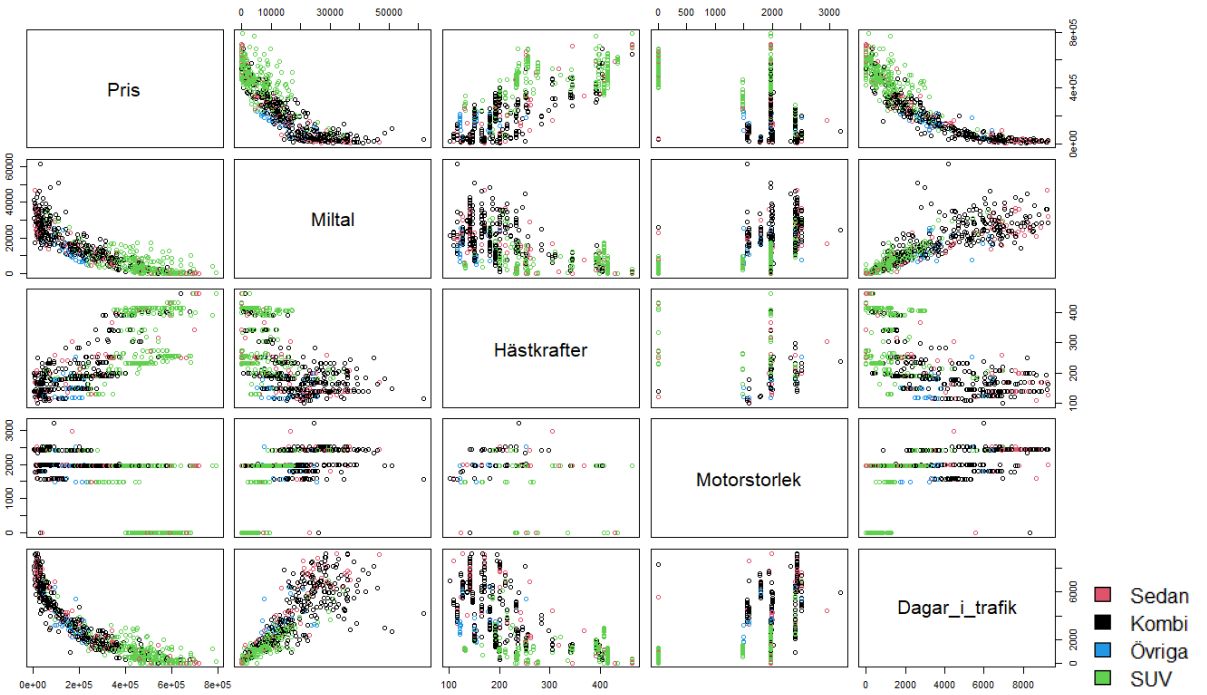
Figur A1. Parvis plot avseende Säljartyp



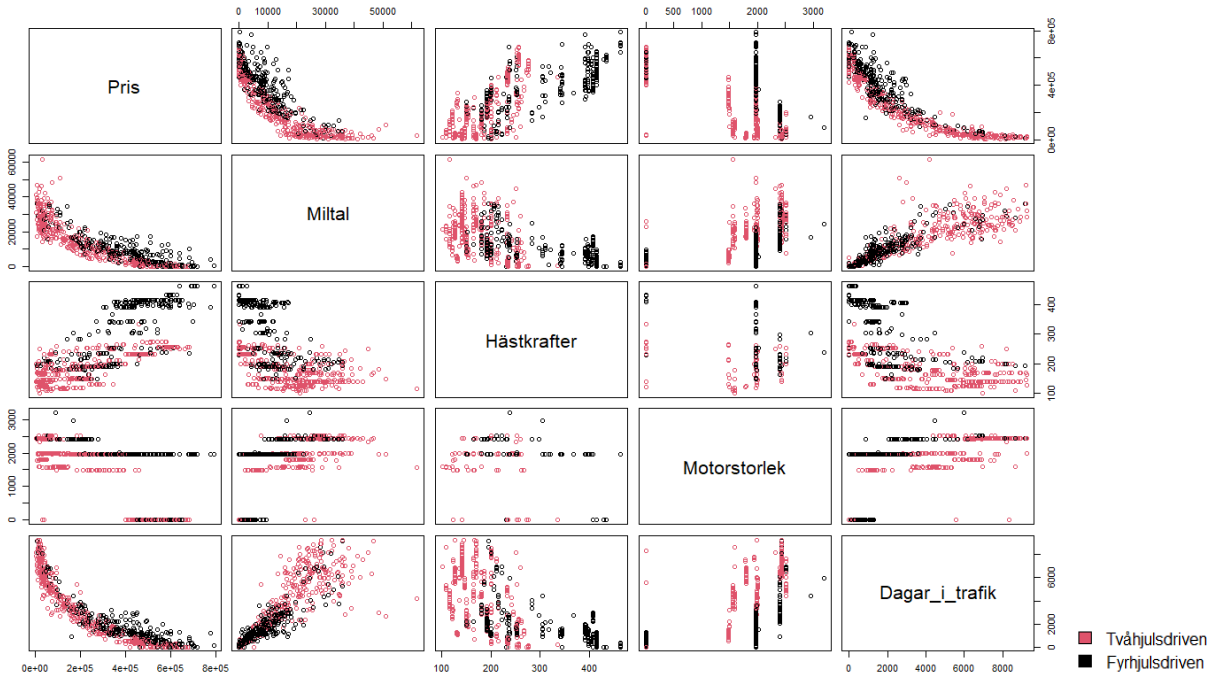
Figur A2. Parvis plot avseende Bränsletyp



Figur A3. Parvis plot avseende typ av växellåda



Figur A4. Parvis plot avseende biltyp



Figur A5. Parvis plot avseende typ av drivning

Källförteckning

Blocket. (2024). Hämtat från https://blocket.se

Géron, A. (2019). *Hands-on Machine Learning with Scikit-Learn, Keras & Tensorflow: Concepts, Tools and Techniques to Build Intelligent Systems (2nd ed.)*. Sebastopol: O'Reilly Media, Inc.

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2021). *An introduction to Statistical Learning – with Applications in R (2nd ed.)*. Retrieved from https://www.statlearning.com/

Körner, S. & Wahlgren, L. (2012). *Praktisk statistik (4:e uppl.)*. Lund: Studentlitteratur.

Körner, S. & Wahlgren, L. (2015). *Statistisk dataanalys (5:e uppl.)*. Lund: Studentlitteratur.

SCB. (2024). Statistikdatabasen. Hämtat från https://www.statistikdatabasen.scb.se/pxweb/sv/ssd/