

Bilpriser på blocket: säljarens perspektiv

Kunskapskontroll R



ECUTBILDNING

Anton Grigoriev

EC Utbildning, DS2023

Rapport

202404

Abstract

Collected data from blocket.se is analysed to be able to predict car prices set by sellers on the platform and to understand factors influencing these prices. 7000 prices analysed demonstrate that within 10-year-old cars age and model (price of the new car) are main factors, yet few minor patterns can be established.

Skapas automatiskt i Word genom att gå till Referenser > Innehållsförteckning.

Innehållsförteckning

Abstract	1
1 Inledning	1
2 Teori.....	2
2.1 Begagnadebilpriser, översikt.	2
3 Metod	2
3.1 Datainsamling (grupprapport)	2
3.2 Grupparbete	3
4 Resultat och Diskussion	4
4.1 Exploratory Data Analysis (EDA)	4
4.2 Pxweb API till Statistik, Sverige.	5
4.3 Priser i olika regioner.....	5
4.4 Gamla automat bilar	5
4.5 Privata ägare.....	5
4.6 Korrelation mellan miltal och bilåder	5
5 Slutsatser	6
6 Teoretiska frågor	7
7 Självutvärdering.....	8
Källförteckning	9

1 Inledning

Generellt sätt att sälja på blocket.se är att studera och prissätta eget varan i rad med massor liknande prylar som redan finns på blocket. Angående bilar man använder viss uppfattning att äldre och mer använda bilar är billigare, sätter priset i rad med andra liknande erbjudande och justerar priset för att sälja billigare men fortare, eller vänta längre och kanske få lite mera pengar.

Syftet med denna rapport är att fullborda krav för Kunskapskontroll i R kurs, för att uppfylla syftet så kommer följande frågeställning(ar) att besvaras:

1. Vilka faktorer förutom välkända spelar roll i prissättning på blocket? T.ex. region, privatägaren, etc.
2. Vilken roll spelar korrelation mellan faktorer som t.ex. ålder och miltal?
3. Tappar automatbilar priset fortare än manuell?

2 Teori

2.1 Begagnadebilpriser, översikt.

Priset på begagnad bil beror på nybilspris, samt möjlighet och kostnader för bil användning. T.ex. diesel kräver högre skatter, både viss ålder och miltal kräver dyrare ”medel” servis på bilen. Nyare modeller värdesättas högre p.g.a. nya funktioner och minskad bränsleförbrukning. Automatlådan möjliggör köring med automatkort, men det kan bli dyr att göra servis och reparera/byta gamla automatlådor.

I POC studiet på begränsad data (7-sits Toyota, Stockholm) jag kunde fastställa att bilen tappar i priset flera tusen kronor per år och runt tusen per hundra mil, dessutom dieselmotor är genomsnitt 10 tusen billigare, hybrider 25 tusen dyrare, automatlådan är dyrare. Faktumet att gamla automatbilar utgör ett dyrare klass visar att vi tittar på säljarens prissättningsperspektiv: köparen med manuell kort kommer knappast värdesätta gammal automat bil dyrare än bil med manuell lådan.

Man vet att gamla bilar som inte används kostar mindre än nya, i samma stund det finns tydligt korrelation mellan ålder och miltal. Alltså nya bilar med höga miltal, som taxibilar, är billiga. Detta ställer liknande fråga som ställdes för automatbilar: hur säljaren balanserar prissättning mellan ålder, miltal och miltal per år?

Med prisar mellan 20 och 500 tusen kronor man undrar om priset kommer att ändras avsevärd med region: priset av tågbiljett+bränsle+mat för att flytta bilen över långt distans knappast kommer bli över 2000.

I denna studiet jag utgör från hypotes att säljaren värdesätter sin bil helt utav blockets andra kända priser, med slumpfaktor som är rent personligt för säljarens omständigheter: hur fort/billigt vill han sälja och vilken uppfattning har han om bilens tillstånd, t.ex. synliga defekter, etc.

3 Metod

Jag använder Anaconda Jupiter Notebook (Anaconda, 2024).

3.1 Datainsamling (grupprapport)

Ni kommer göra en modell, vad är syftet med modellen och vilken data behövs för det?

Syftet med modellen är att prediktera bilpriser på begagnade bilar. Vi kommer att använda oss av annonser från blocket.se.

Vilken typ av fordon vill vi modellera? Exempelvis kan det vara problematiskt om hälften är exklusiva bilar såsom Ferrari och andra hälften vanliga bilar såsom Mazda.
Vi har valt 8 märken: Kia, Volkswagen, Volvo, Audi, BMW, Peugeot, Opel & Toyota

Säkerställ att datan ni samlar in går att läsa in i R och att det blir som ni tänker er. Gör alltså en ”Proof of Concept” (POC).

POC kommer göras på ett urval från alla märken

Vilken typ av data skall vi samla in?

Vi har valt att avgränsa oss till

Biltyp- Halvkombi, Kombi, Suv, Sedan

Modellår- 2014 - 2024

Märke- Kia, Volkswagen, Volvo, Audi, BMW, Peugeot, Opel & Toyota

Minsta pris - 20 000kr

Säljare - Privat & Företag

Växellåda- Manuell & Automat

Drivmedel- Bensin, & Diesel

Hur skall vi samla in datan på ett konsistent sett i gruppen?

En mall för gruppen att samla in data för POC

A	B	C	D	E	F	G	H	I
CarName_Brand	CarName_Model	CarName_ModelYear	CarName_Engine	CarName_Miles	CarName_gears	CarName_Price	CarName_Region	CarName_Dealer
1 Volkswagen	Passat		2015 Diesel	14926 Automat		189800 Stockholm		Riddermark Bil, Veddesta - Järfälla
2 Volkswagen	Golf		2017 Bensin	10976 Automat		168900 Stockholm		Riddermark Bil, Veddesta - Järfälla
4 BMW	X6		2017 Diesel	7857 Automat		539900 Stockholm		Riddermark Bil, Veddesta - Järfälla
5 BMW	430		2018 Bensin	4225 Automat		349700 Stockholm		Riddermark Bil, Veddesta - Järfälla
6 Audi	A6		2016 Diesel	17391 Automat		249900 Västmanland		Riddermark bil Västerås
7 Volvo	S60		2017 Bensin	7493 Automat		339900 Västerbotten		Niemi Bil Skellefteå
8 Volkswagen	Tiguan		2017 Diesel	12273 Automat		289900 Västerbotten		Niemi Bil Umeå
9 Volkswagen	Golf		2018 Bensin	9649 Automat		249900 Västerbotten		Niemi Bil Umeå
0 Audi	Q7		2017 Diesel	14620 Automat		359900 Norrbotten		Niemi Bil AB - Spångatan
1 Volvo	V40		2019 Bensin	6567 Automat		209900 Göteborg		Moberg Bil AB - Göteborg
2 Volvo	V90		2019 Diesel	16075 Automat		214900 Göteborg		Moberg Bil AB - Göteborg
3 Audi	A4		2018 Diesel	7666 Manuell		219900 Göteborg		Moberg Bil AB - Göteborg

Kan man göra några kontroller så datan är "rimlig"?

Genom att genomföra explorativ dataanalys för att identifiera outliners och felaktigheter. Eller exempelvis jämföra med SCB för att säkerställa att datan är realistiska

Hur mycket data skall vi samla in?

Vi har samlat in 7107 bilar hittills

3.2 Grupparbete

1. Vem du har arbetat i grupp med?

Abdulrahman, Alia, Daniel, George, Goran, Jesper, John och Kawser

2. Hur har ni i gruppen arbetat tillsammans?

Vi diskuterade flera gånger i Teams och i chatten skapade i Teams och delade insamlade filar.

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

Det var generellt bra och effektivt, men bara med ändamål att göra kunskapskontroll, kunde struktureras mera för ett större projekt, t.ex. att ha bättre kontroll på uppladdade data, som var inte helt användbar.

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

Jag kan tydligt förklara för andra, men ska utveckla samarbetsförmåga.

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

Vi kunde ha gjort lite gemensamt analys av insamlade data.

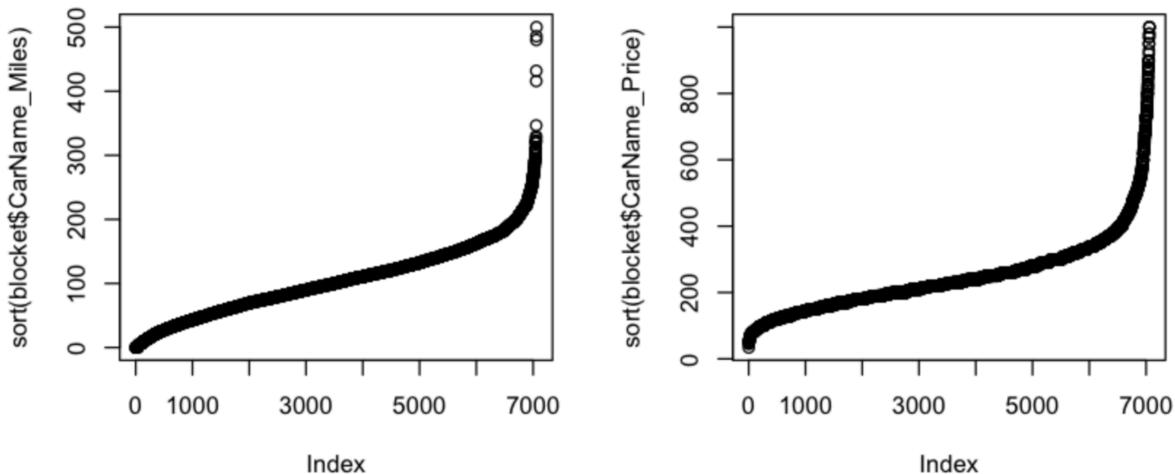
4 Resultat och Diskussion

Data från blocket.se var samlade med webbskrapning (Morris AB, 2021) med bara märke, modell, modellår, bränsle, miltal och pris, samt. växellåda, region och försäljare. Det sista fältet samlades som N/A för privata personer. Jag har valt att preparera data igenom

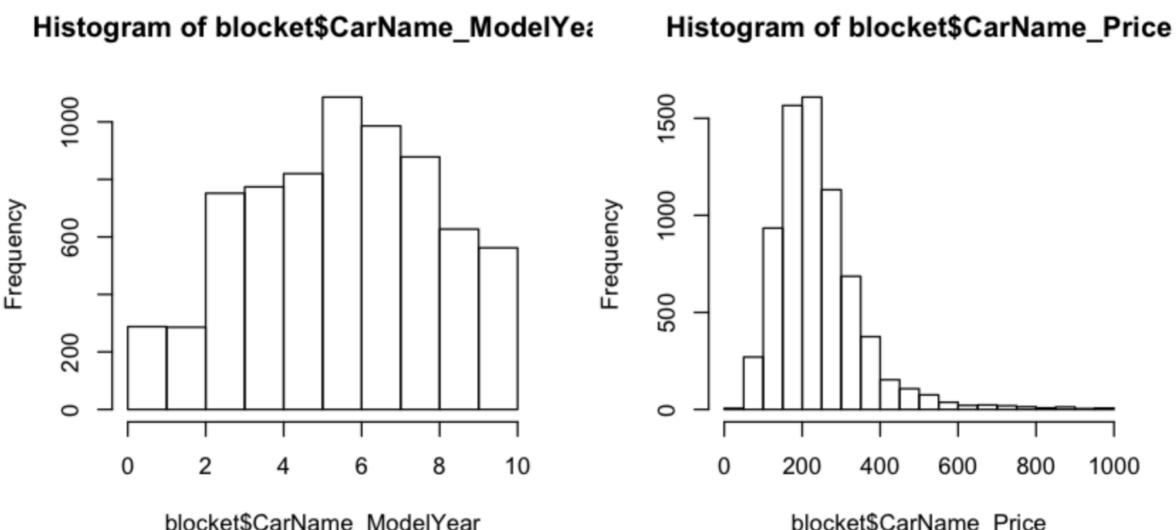
- a) normalisera modellår till ålder, miltal till 100 mil, priset till 1000 kronor (tkr).
- b) Försäljare representeras som "privat" eller "firma"
- c) Extra dyra (över 1000 tkr) och extra mycket körd (över 50 000 mil) bilar stryks som avvikelse. Den sista är på grund av att vi har samlat bara relativt nya bilar.

I POC studie hittade jag bilar av samma modell från tillverkningsår 2020 till 2000. I webbskrapningsdata bilar är ungefär normal distribuerat från 2024 tillbaka, men bara hälften av Gaussian syns i 10 års period. De flesta bilar som kördes över 50 000 mil ($500 * 100$ mil) är just bilar från 2014 – 2016. Effekt att ta bort dyra bilar är att stabilisera regiondistribution, med jämn fel ± 6 tkr.

4.1 Exploratory Data Analysis (EDA)



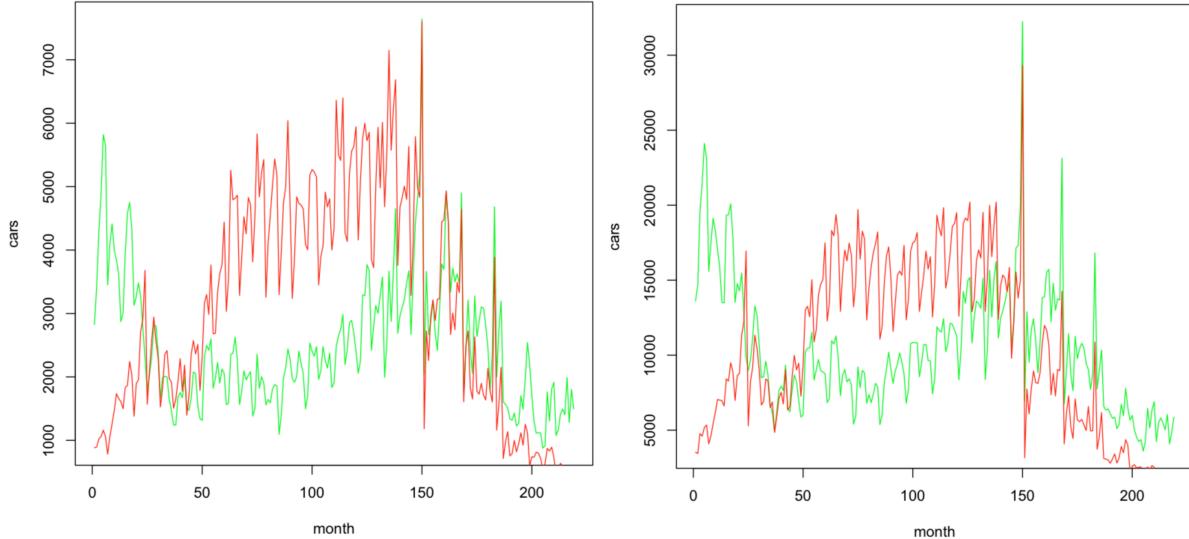
Figur 1: Sorterade miltal i 100 mil och priser i tkr.



Figur 2: Ålder och pris i tkr.

4.2 Pxweb API till Statistik, Sverige.

POC studie visade, att dieseldrivna 7-sits Toyotor i Stockholm är 10 ± 15 tkr billigare än bensindrivna, fas resultaten är ganska osäker. Enkel regressionsanalys för hela Sverige visar motsatta 6 ± 2 tkr dyrare dieselbilar. För att förstår skillnaden använder jag pxweb API (Magnusson, 2019) och scb.se data (Statistics Sweden, 2024). Min hypotes var att dieseldrivna bilar är populära på landet, men inte i Stockholm där de är hotad med begränsningarna. Det visar att jag har fel: det finns visst deficit av dieseldrivna bilar och de, som behöver köra mycket är beredda att betala mer.



Figur 3: Nya *bensin* och *diesel* personbilar registrerade i Sverige (vänster) och Stockholm (höger).

4.3 Priser i olika regioner

Medelpriser i olika regioner verkar följa vanligt mönster: lite högre i Stockholm och Malmö, högre norr där man vanligtvis tror att bilar rostar mindre p.g.a. man använder mindre reagenser mot snö (vanligt vidskepelse (AutoPower.se 2024)), mycket högre priset på Gotland (ön). Trots detta, data visar inte att detta är signifikaanta avvikelse – osäkerhet är alltid mycket större eller lika med regionalt prisavvikelse, även på Gotland 21 ± 23 tkr medelprishöjning visar, att man fortfarande kan hitta bilar med pris utan ö-tillägg.

4.4 Gamla automat bilar

Om jag delar bilar efter automat/manuell växellådor, då är ålderparametrar -20.5 ± 0.4 tkr för automat och -9.2 ± 0.4 tkr för manuell bil, som visar att även säljare värderar begagnade automatbilar så att de åldrar och tappar i priset fortare.

4.5 Privata ägare

Det är statistiskt signifikant, att privata ägare säljer sina bilar 3 ± 2 tkr billigare. Detta är också intressant att privata ägare värderar manuella bilar 6 ± 2 tkr billigare än bilfirmor, som visar att de snarare värderar automatbilar som mera värdefulla.

4.6 Korrelation mellan miltal och bilålder

Parameter selektion visar, att det är mest viktig att veta bilmodell/märke (dyra Audi och BMW sticker ut) och modellår. Miltal för bilar inom 10 år spelar inte för stor roll, dessutom är den korrelerat med bilålder.

5 Slutsatser

Enkelt linjär regression stämmer med blocket.se priser med mycket bra precision

Multiple R-squared: 0.8467, Adjusted R-squared: 0.8424

och regression för log(pris) med

Multiple R-squared: 0.9009, Adjusted R-squared: 0.8981

tillsammans med faktum att miltal ansvarar för lågt andel av prisförlust visar att bilar inom 10 år tappar priset med ålder (så fort som de rullar ut från försäljningshallen (vibilagare.se, 2014)). Äldre bilar kommer att tappa priset längsammare och mera med miltal och slitage. För att sätta priset på begagnad bil på blocket.se räcker att utgå från nypris (eller märke/modell) och ålder.

Automatbilar tappar i värdet fortare, fast inte så fort hos privata ägare, som är genomsnitt billigare än företagsförsäljare. Dieselbilar är genomsnitt lite dyrare, och med lite lycka kan man hitta passande begagnad bil i vilken region som helst.

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.

QQ plot visar hur nära till normalt distribuerat data sekvens igenom att jämföra den med idealt normalt distribuerat data sekvens. Desto närmare till raka linjen ligger QQ plot desto mera normalt distribuerade är data.

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?

Fokus på prediktioner innebär att systematiskt information som X ger om Y, $Y = f(X) + \epsilon$ används som en svart låda. Statistisk regressionsanalys har målet är att skapa en funktion som bäst passar observerade data, då kan denna funktion användas både för prediktioner, men också för att bättre förstå sambandet mellan Y och X.

3. Vad är skillnaden på "konfidensintervall" och "prediktionsintervall" för predikterade värden? "Konfidensintervall" är intervallet kommer att innehålla det verkligt okända värdet för paramatern, d.v.s. "quantify the uncertainty surrounding the average". "Prediktionsintervall" visar hur Y skiljer sig från \hat{Y} , d.v.s. större intervall, som innehåller både felet i $f(X)$ och avvikelse av ett individuellt värde från populationen.

4. Den multipla linjära regressionsmodellen kan skrivas som:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \epsilon.$$

Hur tolkas beta parametrarna?

Beta parametrarna kan tolkas som partiala derivater $\partial Y / \partial x_i$.

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?

BIC innehåller term som straffar överkomplifierade modeller (overfitting) och på så sett väljer modell utan validering. Trots detta, direkt validering och cross-validering är bättre än indirekt BIC metod, om man har tillräckligt datorkapacitet.

6. Förklara algoritmen nedan för "Best subset selection"

Man går igenom alla möjliga subset, strukturerad brute force-metoden, där man går från medelvärdet (konstant) till modell med maximalt antal parametrar och väljer den bästa av alla.

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

Statistiska modeller saknar alltid verklighetens komplexitet men kan vara användbara ändå. Modell är approximation, som simplifierar, gör enklare att förstå komplexa saker.

7 Självutvärdering

1. Utmaningar du haft under arbetet samt hur du hanterat dem.

Bara två av gruppens handplockade filer var användbara.

2. Vilket betyg du anser att du skall ha och varför.

VG för jag hade uppfyllt kriteria.

3. Något du vill lyfta fram till Antonio?

Det känns att kursen har lite för mycket överlappning med liknande kurs i Python och MI.

Källförteckning

Anaconda (2024) <https://www.anaconda.com/>

"AutoPower.se - Forum". www.autopower.se. 23 april 2024. Läst 26 april 2024.

James, Gareth; Witten, Daniela; Hastie, Trevor; Tibshirani, Robert (2013). ["An Introduction to Statistical Learning"](#). Springer Texts in Statistics. doi:[10.1007/978-1-4614-7138-7](https://doi.org/10.1007/978-1-4614-7138-7). ISSN 1431-875X. <http://dx.doi.org/10.1007/978-1-4614-7138-7>. Läst 26 april 2024. Mans Magnusson, Markus Kainu, Janne Huovari, and Leo Lahti (rOpenGov). pxweb: R tools for PXWEB API. URL:<http://github.com/ropengov/pxweb>, 2019.

Statistics Sweden (2024). "New registered passenger cars by region, fuel, observations and month." [Data accessed 2024-04-26 00:33:12 using pxweb R package 0.9.1], <URL:<https://api.scb.se/OV0104/v1/doris/en/ssd/TK/TK1001/TK1001A/PersBilarDrivMedel>>.

Morris AB ["Web scraping ur ett juridiskt perspektiv - Morris Law"](#). 21 maj 2021. Läst 26 april 2024

["Bilarna som tappar mest i värde i Vi Bilägare"](#). www.vibilagare.se, 2014. Läst 26 april 2024.