Regressionsmodell

Blocket data

Umut Arslan

EC Utbildning

2020426

Abstract

The study's objective is to develop predictive regression models that can precisely forecast the prices of pre-owned electric vehicles, specifically targeting those that have logged a minimum of 500 km. Data, including nine potential explanatory variables, were compiled from blocket.se. After careful consideration, only five of these variables were selected for the final analysis. The model that demonstrated the greatest accuracy, as determined by the Root Mean Square Error was the one utilizing the Generalized Additive Model algorithm.

Innehållsförteckning

Förkortningar och begrepp

RMSE = Root Mean Square Error

GAM = Generalized Additive Model

MSE = Mean Squared Error

GLM = Generalized Linear Models

1 Inledning

Rapportens syfte är att utforska och jämföra olika regressionsmodellers förmåga att förstå sambandet mellan hur prissättningen sker givet diverse variabler. Fokuset ligger på begagnade elbilar baserat på data insamlad manuellt från blocket.se. Inom området för artificiell intelligens och maskininlärning, där maskiner tränas att efterlikna mänskligt tänkande och lärande, utgör regressionsanalys en grundläggande teknik. Genom att tillämpa denna teknik strävar vi efter att utrusta maskiner med förmågan att självständigt tolka och dra slutsatser från stora datamängder. Detta möjliggör inte bara automatisering av system utan även optimering av processer som kan förbättra livskvaliteten och effektivisera verksamheter inom både företag och myndigheter (Heidlund, 2018). Denna rapport visar förhoppningsvis hur traditionella statistiska metoder och maskininlärningsmodeller kan samverka för att skapa precisa prediktiva modeller, där målet är att identifiera den modell som med lägst RMSE givet den beroende variabeln Pris och givet de oberoende variabler som Miltal, Modellår, Biltyp, Hästkrafter och Märke. Utöver detta så är en ytterligare anledning att hjälpa mig som individ att köpa den billigaste elbilen givet parametrarna ovan, detta på grund utav den ökande tillväxten av elbilar på senare år, se Figure 1.

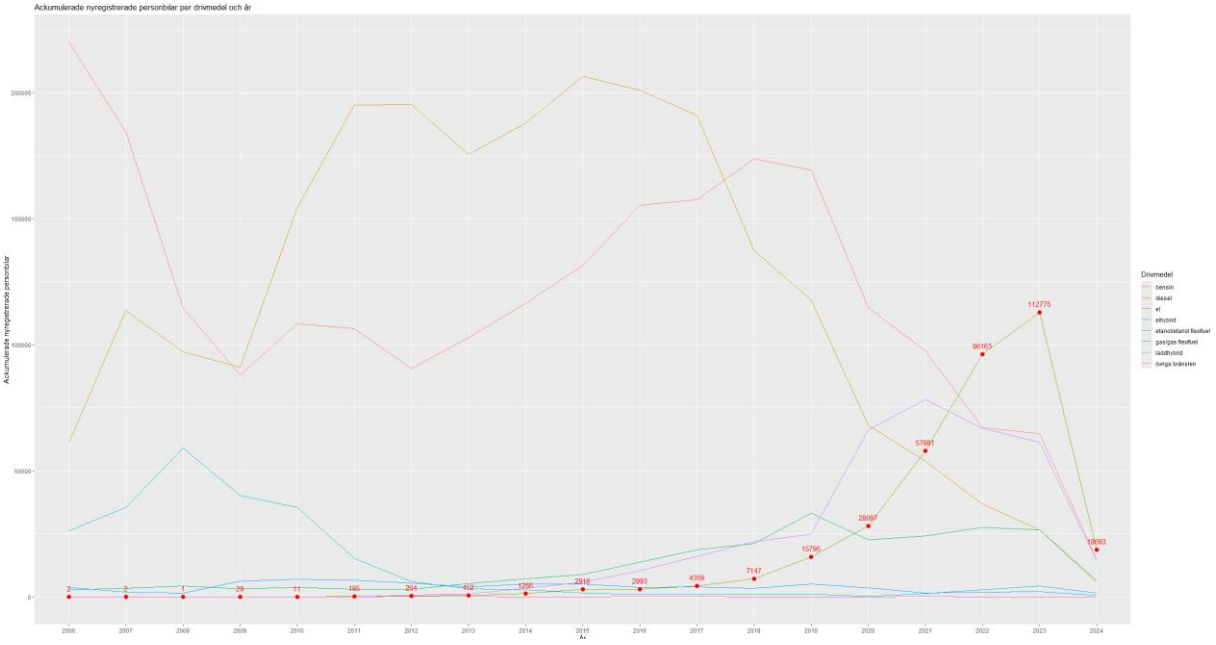


Figure 1 - Data taget från SCB API - https://api.scb.se/OV0104/v1/doris/sv/ssd/TK/TK1001/TK1001A/PersBilarDrivMedel

1.1 Målsättning

Målet är att ta reda på vilken regressionsmodell, begränsat utvalda, som presterar bäst på blocket-datat, där RMSE är måttstocken. Även vilken begagnad elbil som är billigast att köpa.

2 Teori

2.1 Multipel Regression

Multipel linjär regression är ett användbart verktyg för att hitta linjära samband mellan en responsvariabel och flera förklaringsvariabler. Det betyder inte nödvändigtvis ett orsakssamband, så där får man vara försiktig (James, Witten, Hastie, & Tibshirani, 2023).

Den teoretiska modellen ser ut som följande:

$$Y = B_0 + B_1X_1 + B_2X_2 + \ldots + B_iX_i + \varepsilon_i$$

Där Y är responsvariabel, B_0 är interceptet medan resterande betaparametrar är lutningsparametrar. X_i parametrarna är förklaringsvariabler/oberoende variabler. ε_i Är feltermen/störningstermen. Där B_0 är startvärdet/basvärdet medan resterande betaparametrar mäter ändringen i responsvariabeln för varje enhet adderad, givet att resterande variabler hålls konstanta.

2.2 Ridge Regression

Ridge regression är ett användbart verktyg för att skatta data med hög korrelation mellan förklaringsvariablerna, multikollinearitet/kollinearitet (McDonald, 2009). Algoritmen inför någon form av bias för att minska variansen i datat, detta ska teoretiskt minska medelfelet, MSE.

2.3 Lasso Regression

Lasso regression är ett användbart verktyg för att identifiera variabler och minimera prediktionsfelen. Detta görs genom att straffa parametrarna vilket krymper koefficienterna mot noll, de variabler som krymps mot noll utesluts från modellen (Ranstam & Cook, 2018).

2.4 Random forest regression

Kortfattat så skapar denna algoritm en samling av beslutsträd och itererar på dessa för att förbättra prediktionen och undvika överanpassning.

2.5 Generalized Additive Model

GAM är en ”likeliehood-baserad” regressionsmodell och en utveckling av GLM.

Skillnaden mellan dessa är GAM tillåter för en mer flexibel modellering genom att använda sig av smoothing funktioner för att lyckas fånga upp icke-linjära samband, modellen kan även utöver detta fånga upp linjära samband (Hastie & Tibshirani 1986).

3 Metod

Denna studie baserades på fordon annonserade på blocket.se. Totalt representerades 409 fordon i datamängden, som innehöll 10 variabler: Miltal, Modellår, Biltyp, Drivning, Hästkrafter, Färg, Märke, Modell, Län och Pris. För att identifiera de mest signifikanta förklaringsvariablerna för variabeln Pris, användes multipel regression som en utgångspunkt. Genom att testa olika kombinationer av variabler, fastställdes Miltal, Modellår, Biltyp, Hästkrafter och Märke som de mest relevanta.

Även om forward/backward stepwise selection var tillgängliga, valdes en manuell variabelselektion för denna analys. Data delades upp i tränings- och testset för vissa modeller, Ridge, Lasso samt GAM medan resterande använde hela datasettet. De två resterande modellerna var Multipel Regression samt Random Forest Regression. Varje modells prediktiva förmåga bedömdes genom att jämföra deras RMSE. Den modell som uppvisade det lägsta RMSE-värdet valdes som den mest lämpliga för prisförutsägelser.

4 Resultat

jämförelsen av regressionsmodellerna visade att prestandan var relativt likvärdig över samtliga modeller. Trots detta utmärkte sig Generalized Additive Model (GAM) med den lägsta RMSE, vilket indikerar en bättre förmåga att förutsäga pris på begagnade bilar jämfört mot resterande modeller. Detaljerade resultat presenteras i Tabell 1.

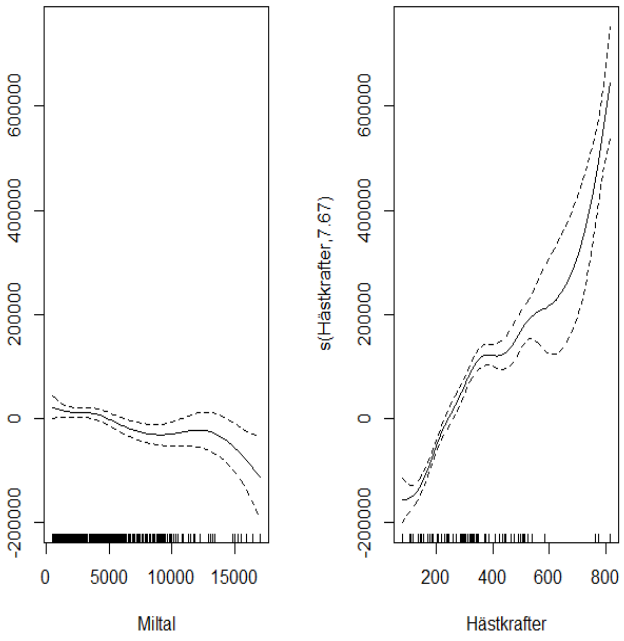
En närmare analys av de faktorer som påverkar priserna på begagnade bilar, som återfinns i Tabell 2, visar att högre miltal generellt leder till en prissänkning. Å andra sidan bidrar en ökning av hästkrafter till ett högre pris. När det gäller bilmodeller, visar resultaten att äldre modeller tenderar att vara billigare. Märkena Mercedes-Benz och Ford utgör undantag från denna trend, där de istället verkar bidra till ett högre pris.

Av biltyperna visas det att sedanmodeller är de näst mest prisvärda alternativen. Referenskategorierna för modellår, biltyp och märke är 2014, halvkombi respektive Audi. Dessa referensvärden fungerar som baslinje, värde lika med noll, för jämförelse med andra kategorier inom respektive variabel. Vi kan även se att 89,2% av den totala variansen i de observerade datan kan förklaras av modellen, vilket är högt.

RMSE_lm	RMSE_ridge	RMSE_lasso	RMSE_rf	RMSE_gam
66640.96	65566.94	64964.98	65515.54	56795.78

Tabell 1: RMSE för de 5 modellerna

Tabell 2 – slätade funktioner




```
Family: gaussian
Link function: identity

Formula:
Pris ~ s(Miltal) + Modellår + Biltyp + s(Hästkrafter) + Märke

Parametric coefficients:
                 Estimate Std. Error t value          Pr(>|t|)
(Intercept)      336678.7    46868.8   7.183 0.00000000000372 ***
Modellår2015     -41243.3    49840.1  -0.828          0.408475
Modellår2016       5817.4    48778.7   0.119          0.905133
Modellår2017      30921.2    47150.1   0.656          0.512355
Modellår2018        475.9    46626.8   0.010          0.991861
Modellår2019      -3790.8    44993.8  -0.084          0.932901
Modellår2020        517.1    44603.0   0.012          0.990757
Modellår2021      32598.1    45451.1   0.717          0.473692
Modellår2022      77800.8    45896.1   1.695          0.090881 .
Modellår2023     145286.3    47525.3   3.057          0.002396 **
Modellår2024     250628.4    50485.1   4.964 0.00000105010704 ***
Biltypkombi       61330.5    23769.1   2.580          0.010254 *
Biltypsedan       16511.7    13403.9   1.232          0.218777
Biltypsuv         54637.9     9994.2   5.467 0.00000008394485 ***
Märkebmw          -3836.1    17472.5  -0.220          0.826341
Märkeford         17797.4    16752.2   1.062          0.288746
Märkehyundai      -8667.3    16429.7  -0.528          0.598132
Märkekia         -47318.9    15664.1  -3.021          0.002694 **
Märkemb           51693.5    21736.8   2.378          0.017902 *
Märkemg          -67802.8    19067.2  -3.556          0.000425 ***
Märkenissan      -54293.9    18412.3  -2.949          0.003391 **
Märketesla       -70382.0    17262.7  -4.077 0.00005575366051 ***
Märkevolkswagen  -16878.8    16558.4  -1.019          0.308697
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                edf Ref.df      F               p-value
s(Miltal)     5.085  6.207  3.386               0.00263 **
s(Hästkrafter) 7.673  8.484 57.671 < 0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.882   Deviance explained = 89.2%
GCV = 3.6849e+09  Scale est. = 3.3627e+09  n = 409
```

Tabell 3 – Variabelpåverkan på priset.

5 Slutsatser och Diskussion

Att jämföra olika regressionsmodeller har varit en komplex process, mycket på grund av den avancerade matematiska logiken som modellerna är baserade på. Den straffbaserade logiken i Ridge och Lasso regression, förhindrar modellerna att överanpassa genom att bestraffa stora koefficientvärden, är inte direkt intuitivt. Även om jag har en grundläggande förståelse för att dessa komplexa processer förbättrar modellernas förmåga att generalisera, så är det fortsatt svårt att faktiskt förstå på riktigt. När det kommer till GAM, ökar komplexiteten ytterligare, med flera lager av avancerade beräkningar som bidrar till modellens förmåga att hantera icke-linjära samband, interaktioner mellan variabler och även slätningsfunktioner/splines.

Förbiser vi dessa utmaningar så har i alla fall studien lyckats uppnå sitt primära mål, att identifiera den regressionsmodell, begränsad tid, som presterar bäst på blocket-datat enligt RMSE-kriteriet. När det gäller att hitta den billigaste begagnade elbilen, har analysen påvisat att överväga flera faktorer, inklusive miltal, hästkrafter och bilens årsmodell. Märken som Kia och MG visade sig erbjuda mer prisvärda alternativ, vilket kan vara av intresse. Jag har ej jämfört modellen mot verkliga Kia eller MG bilar för att se hur korrekt den faktiskt gissar.

Nedan så ser vi plottar från den Multipla linjära Regressionen. Denna figure fick mig att vilja testa en mer flexibel icke-linjär modell som GAM. Q-Q plotten hade ganska stora svansar men majoriteten av ”kroppen” är centrerad runt linjen. Det är riktigt oklart.

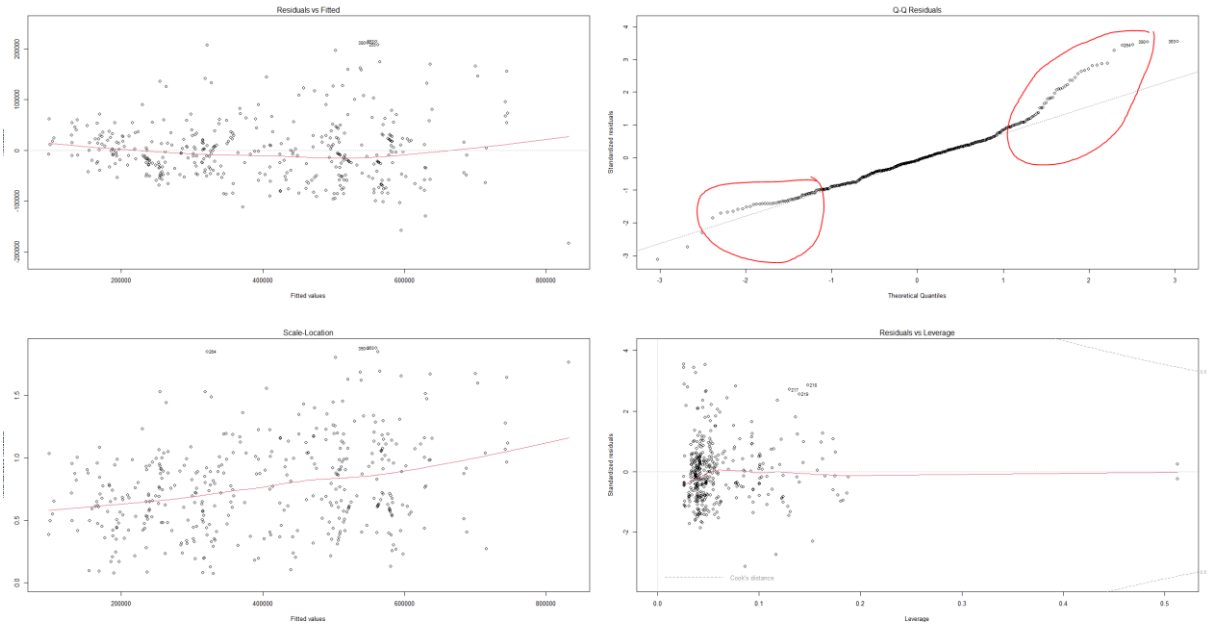


Figure 2

6 Teoretiska frågor

1. Q-Q plot används generellt för att kontrollera normalitetsantagande/linjäritet. Y-axeln plottas punkter från din egna data och x-axeln plottas data från den teoretiska fördelningen, gaussfördelning mest troligt.

2. Vid maskininlärning tenderar man inte att fokusera på orsakssamband, dock är orsakssamband nästan en omöjlighet att bevisa, men snarare hur bra modellen gissar rätt på okända data. Detta kan även göras med ”statistisk” regressionsanalys men det mer generella användningsområdet för regression är att med empirisk 95% säkerhet kunna säga att det exempelvis finns ett samband mellan Pris och Hästkrafter.

3. Konfidensintervall säger med 95% säkerhet inom vilket intervall det sanna medelvärdet ligger för hela populationen medan ett prediktionsintervall säger med 95% säkerhet inom vilket intervall den enskilda stokastisk variabel ligger mellan enligt Dybowski och Roberts (2001).

4. B_0 är interceptet medan resterande betaparametrar är lutningsparametrar. X_i parametrarna är förklaringsvariabler/oberoende variabler. ε_i Är feltermen/störningstermen. Där B_0 är startvärdet/basvärdet medan resterande betaparametrar mäter ändringen i responsvariabeln för varje enhet adderad, givet att resterande variabler hålls konstanta.

   ” We interpret βj as the average effect on Y of a one unit increase in Xj , holding all other predictors fixed. In the advertising example, the model becomes (James, Witten, Hastie, & Tibshirani, 2023)”

5. Du behöver inte använda dig utav träning, test och valideringsdata, du kan i stället använda dig utav metoder som AIC, BIC, Adjusted R^2 och C_p. Man jämför alltså modellernas BIC mot varandra, där lägsta generellt är bäst (James, Witten, Hastie, & Tibshirani, 2023). De metoderna nämda ovan straffar modellerna på olika sätt, där C_p och BIC liknar varandra mest teoretiskt.

6. Modellen testar alla kombinationer av förklaringsvariablerna som går och väljer ut den modellen som har bäst AIC, BIC, Adjusted R^2, C_p eller cross-validering

7. Jag antar att George menar på att ingen modell är skräddarsydd för ens specifika data, därav så är ingen modell rätt men de kan ändå vara användbara givet antaganden osv.

7 Självutvärdering

1. Utmaningar du haft under arbetet samt hur du hanterat dem.
   Vad som klassas som normalfördelad data, jag vet alla 5-6 punkter men det är svårt m.h.a bilder att själv tolka 😊

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

3. Något du vill lyfta fram till Antonio?
   Roligt men svår kurs.

Källförteckning

Heidlund, M. (2018). IT-konsulters positionering inom artificiell intelligens nu och i framtiden (Dissertation). Retrieved from https://urn.kb.se/resolve?urn=urn:nbn:se:miun:diva-34693

Ekenstedt, C., & Holmström, G. (2018). Artificiell intelligens och maskinlärning i finansbranschen (Dissertation). Retrieved from https://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-224326

Dybowski, R., & Roberts, S. J. (2001). Confidence intervals and prediction intervals for feed-forward neural networks. Cambridge University Press. Retrieved from http://www.cambridge.org/catalogue/catalogue.asp?isbn=0511339941.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2023). An Introduction to Statistical Learning with Applications in R (2nd ed.) Retrieved from https://www.statlearning.com/

McDonald, G. C. (2009). Ridge regression. Wiley Interdisciplinary Reviews: Computational Statistics, 1(1), 93-100. Retrieved from https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wics.14

Ranstam, J., & Cook, J. A. (2018). LASSO regression. Journal of British Surgery, 105(10), 1348-1348. Retrieved from https://academic.oup.com/bjs/article/105/10/1348/6122951

R Core Team. (u.å.). gam: Generalized Additive Models. R Project. Retrieved from https://search.r-project.org/CRAN/refmans/mgcv/html/gam.html

Hastie, T., & Tibshirani, R. (1986). Generalized Additive Models. Retrieved from https://pdodds.w3.uvm.edu/files/papers/others/1986/hastie1986a.pdf