Stockholms universitet

Regressionsanalys av priser på Volvo personbilar

Lina Lion

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2019:21
Matematisk statistik
Juni 2019

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Matematisk statistik
Stockholms universitet
Kandidatuppsats **2019:21**
http://www.math.su.se/matstat

Regressionsanalys av priser på Volvo personbilar

Lina Lion*

Juni 2019

Sammanfattning

I det här arbetet är vi intresserade av att analysera vilka faktorerar som påverkar priset på begagnade bilar, vi inskränker oss till Volvo som är det vanligaste bilmärket.

Med hjälp av R programmet lyckas vi samla ihop ett dataset som består av 1297 skilda utannonserade begagnade bilar av märket Volvo på Bilias webbsida. Datasetet består av totalt 10 förklarande variabler (hästkrafter, drivhjul, förbrukning, utsläpp, toppfart, ägare, tillverkningsår, växellåda, bränsle och miltal) samt priset som responsvariabeln. Den statistiska metoden som används på datamaterialet är multipel linjär regression och transformationer av variabler.

Resultatet av den statistiska analysen som erhålls visar två olika modeller som uppfyller kraven för regressionsanalys. Som nästa steg används stegvis variabelselektion för att få fram modellen som är bäst anpassad till data. Slutmodellen i denna analys är den med logaritmerad responsvariabeln (pris) samt förklarande variabler som hästkrafter, utsläpp av koldioxid per kilometer, fyrhjulsdriven, bilens ålder, automat växellåda samt mil/år.

Baserad på den statistiska analysen kan vi även dra slutsatsen att bilens ålder påverkar priset som mest, förändringen är -8.54

Bilar av märket Volvo är indelade i olika modeltyper (ex V70, XC70 osv..). Därför finner vi det intressant att undersöka hur mycket modelltypen egentligen är avgörande vid prissättningen. Därmed inkluderas förklarande variabeln modelltyp till datasetet och förklarande variabler som beskriver egenskaper hos bilen (som ex hästkrafter, bränsle, drivhjul, förbrukning, utsläpp, toppfart, växellåda) exkluderades. Analoga statistiska undersökningar som förklaras ovan tillämpas på det nya datasetet med modelltyp som förklarande variabel.

Det slutgiltiga resultatet är att den statistiska modellen med modelltyp och den statistiska modellen utan modelltyp ger mycket likgiltiga resultat. Slutsatsen blir därmed att prissättningen på en begagnad bil av märket Volvo sker främst efter modellernas egenskaper snarare efter modellbeteckningen. Detta för att resultatet tyder på att man kan prediktera priset lika väl med modellbeteckningen som egenskaperna.

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige.
E-post: lina_lion7@hotmail.com. Handledare: Martin Sköld och Tom Britton.

Förord

Arbetet som presenteras omfattar ett självständigt arbete på 15 hp i matematisk statistisk som motsvarar ett kandidatarbete och har utförts vid matematiska institutionen vid Stockholms Universitet.

Först och främst vill jag rikta ett stort tack till mina två handledare Martin Sköld samt Tom Britton vid Stockholms Universitet. Jag vill tacka de för deras värdefulla rådgivning och support under examensarbetets gång. Jag vill även passa på och tacka min familj för deras stöd genom hela utbildningen.

Innehållsförteckning

1 Inledning **5**

1.1 Frågeställningar 5

2 Teori **6-9**

2.1 Regressionsteori 6-9

2.1.1 Skattning av regressionskoefficienten 6-7

2.1.2 Hypotestest 7

2.1.3 p-värde 7-8

2.1.3 Villkor-linjär regression 8-9

2.2 Variabeltransformationer 9

2.2.1 Logtransformering 9

2.2.2 Dummyvariabler 9

2.3 Multikollinjäritet 10

2.3.1 Variansinflationsfaktor 10

2.4 Modellval 10-12

2.4.1 Förklaringsgrad och justerad förklaringsgrad 10-11

2.5 Stegvis variabelselektion 11-12

3 Datamaterial **12-14**

3.1 Kort om Webbsidorna 12

3.2 Variabelförklaring 13

3.3 Dataöversikt 14

4 Statistisk analys **14-22**

4.1 Bearbetning av data 14

4.2 Enkel linjär regression 15

4.3 Multipel linjär regression 16-20

4.4 Modelljämförelse 20

4.5 Förklarande variabel-*Modelltyp* 20-22

5 Resultat **22-23**

6 Diskussion **23-24**

6 Referenslista ..... **25** $2{,}2 \cdot 10^{-16}$

1 Inledning

Huvudsyftet är att undersöka vilka faktorer som är mest signifikanta vid prissättningen av en begagnad bil. För att kunna utföra analysen väljer vi att begränsa oss till enbart bilar av märket Volvo. Allmänt är Volvos personbilar uppdelade efter olika modelltyper som t.ex. V60, S40 o.s.v. . Vi kommer i detta arbete även att undersöka vilken betydelse modellbeteckningen egentligen har när priset på en utannonserad Volvo sätts.

1.1 Frågeställningar

Frågeställningar som ska besvaras i detta arbete är följande:

I. Vilka faktorer är mest signifikanta vid prissättning av en begagnad bil?
II. Är faktorn *modelltyp* signifikant vid prissättning av en Volvo och vilken betydelse har den?

För att kunna besvara dessa frågeställningar används statistiska verktyg som multipel linjär regression samt transformation av variabler. Transformationerna som tillämpas i detta arbete är logtransformering samt dummyvariabler. Variabelselektion används därefter för att komma fram till den slutliga modellen som förklarar responsvariabeln (*pris*) mest.

2 Teori

2.1 Regressionsteori

Allmänt finns det olika typer av regressionsanalys där linjär regressionsanalys är den mest väsentliga för detta arbete. Det finns två olika kategorier av linjär regression: enkel linjär regression samt multipel linjär regression. Huvudpoängen med regressionsanalys är att hitta ett samband mellan en responsvariabel och en eller flera olika förklarande variabler. Det finns två olika kategorier av linjär regression: enkel linjär regression (1) samt multipel linjär regression (2). *(Sundberg studentkompendium, 2016, s. 50-51)*

Enkel linjär regression;

$$y_i = \alpha + \beta x_i + \varepsilon_i \,, \qquad i = 1, \dots, n. \tag{1}$$

Multipel linjär regression;

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_m x_{mi} + \varepsilon_i, \qquad i = 1, \dots, n. \tag{2}$$

Variabelförklaring för linjär regression:

$$\begin{cases} y \rightarrow \text{responsvariabeln} \\ \alpha \rightarrow \text{intercept} \\ \beta \rightarrow \text{regressionskoefficient} \\ x \rightarrow \text{förklarande variablar} \\ \varepsilon \rightarrow \text{residual } \{\text{slumpmässiga, oberoende och likafördelade}\}. \end{cases}$$

Formeln (2) kan skrivas om på matrisform

$$y = X\beta + \epsilon \,.$$

Där $y = (y_1 \dots y_n)^T, \beta = (\alpha, \beta_1 \dots \beta_m)^T, \epsilon = (\epsilon_1 \dots \epsilon_n)^T \;\; och\; X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{m1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & \cdots & x_{mn} \end{bmatrix}.$

Första kolumnen med enbart ettor i X föreställer interceptet α. *(Andersson, Lindensjö, Tyrcha, studentkompendium, 2018, s.19-20)*

2.1.1 Skattning av regressionskoefficienten:

Den okända regressionskoefficienten *(β)* skattas med minsta kvadratskattningen. Minsta-kvadrametoden går ut på att minimera summan av det kvadratiska avståndet mellan regressionslinjen och observationerna $\sum_{i=1}^{n}\left(y_i - \alpha - \sum_{j=1}^{m} x_{ji}\beta_j\right)^2$. *(Sundberg studentkompendium, 2016,s.19)*

Minsta kvadrat skattningen av β ges av:

$\hat{\beta} = (X^T X)^{-1} X^T Y$ där $\widehat{\beta} \sim N(\beta, \sigma^2 (X^T X)^{-1})$.

Om residualerna följer en normalfördelning med väntevärde noll och en konstant varians, kan vi skriva följande: *(Sundberg studentkompendium, 2016, s 22-23)*

$$E(\hat{\beta}) = \beta \tag{3}$$

$$Var(\hat{\beta}) = \sigma^2 (X^T X)^{-1}. \tag{4}$$

I detta fall är σ^2 en okänd parameter som skattas med det observerade data. En väntevärdesriktig skattning av σ^2 ges av följande:

$$\hat{\sigma}^2 = \frac{1}{n-m} \sum_{i=1}^{n} \left(y_i - \hat{\alpha} - \sum_{j=1}^{m} x_{ij} \hat{\beta}_j \right)^2 .$$

m föreställer antalet β-parameterar medan *n-m* anger antalet frihetsgrader. Frihetsgrader är antalet oberoende observationer som kan används vid skattningen av den okända parametern σ^2.

2.1.2 Hypotestester

Hypotestesterna som används i följande arbete går ut på att testa om förklarande variablerna har en signifikant inflytande på responsvariabeln *(y)*. Generellt kan man säga att testerna går ut på att skatta de olika regressionskoefficienterna *(β_j)*, samt testa om de är signifikant skild från noll. *(Sundberg studentkompendium, 2016, s. 59-60)*

Vi vill testa nollhypoteserna

$$H_0: \beta_j = 0$$

mot alternativa hypoteserna

$$H_1: \beta_j \neq 0$$

för varje *j*.

2.1.3 p-värde

Avsnitt 2.1.1 specifikt formeln (3) och (4) explicerar på att residualerna följer en normalfördelning med väntevärde 0 och konstant varians. Skattningen av den j:te regressionskoefficienten kommer i detta fall att följa en normalfördelningen på följande sätt:

$E(\hat{\beta}_j) = \beta_j$ och $Var(\hat{\beta}_j) = (\sigma^2 (X^T X)^{-1})_{jj}$.

$\sigma^2(X^TX)^{-1}{}_{jj}$ är den j: te diagonalelementen i matrisen som presenteras i *Andersson, Lindensjö och Tyrcha, studentkompendium, 2018, s.44*. Hypotestesten testas genom att tillämpa t-test. Syftet med testet är att göra en jämförelse mellan två medelvärden för att på så sätt kunna dra slutsats om de förekommer en signifikant olikhet.

Teststatistiskan följer en t-fördelning och skrivs på följande sätt:

$$T_j = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{((X^TX)^{-1})_{jj}}}.$$

Om $|T_j| \geq t_{\alpha/2}(n-m)$ förkastar nollhypotesen på signifikantnivån α

$$\begin{cases} n-m \rightarrow \text{antal frihetsgrader under } \mathrm{H}_0 \\ m \rightarrow \text{antal regressionskoefficinter } (\beta) \\ t_{\alpha/2}(n-m) \rightarrow \text{kristiska gränsen som läses ut ur tabell för t} - \text{fördelningens kvantiler.} \end{cases}$$

P-värdet används i detta arbete som ett mått på signifikansnivån för olika förklarande variabler i en modell. Måttet definieras som följande sannolikhet:

$$P(|T_j| \geq t_{\alpha/2}(n-m).$$

Vi väljer signifikansnivån (α = 0.05), nollhypotesen förkastas om p-värdet är mindre än den valda signifikantnivån. Att nollhypotesen förkastas innebär att regressionskoefficienten β_j för förklarande variabeln (x_j) är signifikant skild från 0. Detta innebär i sin tur att förklarande variabeln är signifikant för modellen. *(Andersson, Lindensjö och Tyrcha, studentkompendium, 2018 s 44).*

2.1.4 Villkor- linjär regression

Linjära regressionsmodeller används främst för att undersöka om en responsvariabeln kan förklaras linjärt av en eller flera andra förklarande variabler. Följande förutsättningar är grundläggande vid tillämpning av linjär regression. *(Andersson, Lindensjö, Tyrcha, studentkompendium, 2018, s.63).*

Linjäritet:
Innebär ett linjärt samband mellan responsvariabeln och alla förklarande variabler.

Ej perfekt multikollinjäritet:
Inget linjärt beroende mellan variablerna.

Normalfördelade residualer:
Residualerna kan beskrivas som avståndet mellan varje observerat värde och den skattade linjära linjen. *(Andersson, Lindensjö och Tyrcha, studentkompendium, 2018, s. 67-68).* Att residualerna följer en normalfördelning med väntevärde noll är en förutsättning för linjär regression. Residualerna är normalfördelade om de följer en rät linje i en normalfördelningsplot.

Homoskedasticitet:

Residualerna måste även uppfyll kraven för en konstant varians, alltså

$$V(\varepsilon_j) = \sigma^2 \quad för\ alla\ j = 1, \dots, n.$$

Konstant varians innebär att variationen mellan slumpfelen/feltermerna ska vara konstant. Följer residualerna ett systematiskt mönster kallas det för heteroskedasticitet *(Andersson, Lindensjö och Tyrcha, studentkompendium, 2018, s. 67-68)*

2.2 Variabeltransformation

2.2.1 Logtransformering

Logtransformering är en central metod inom variabeltransformering. Transformeringen används främst i fallet då de förklarande variablerna och responsvariabeln inte följer ett linjärt samband. Metoden går ut på att logtransformera responsvariabeln och/eller en eller flera förklarande variabler. *(Sundberg studentkompendium, 2016, s. 82-84)*

Matematiskt kan detta betecknas på följande sätt;

$$log(y_i) = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_m x_{mi} + \epsilon_i.$$

$$y_i = \alpha + \beta_1 \log(x_{1i}) + \beta_2 \log(x_{2i}) + \cdots + \beta_m \log(x_{mi}) + \epsilon_i.$$

$$\log(y_i) = \alpha + \beta_1 \log(x_{1i}) + \beta_2 \log(x_{2i}) + \cdots + \beta_m \log(x_{mi}) + \epsilon_i.$$

2.2.2 Dummyvariabler

Dummyvariabler används för kategoriska variabler genom att ha variabler med värde 0 eller 1 för varje kategori. I fallet då den kategoriska variabeln enbart kan anta två värden som till exempel kön (kvinna eller man) då kan man skapa en dummyvariabel som antar värdet 1 och det är man och 0 om det är kvinna. Om den kategoriska variabeln antar flera värden än två variabler då skapas flera dummyvariabler. Anta ett exempel med ett föremål som kan anta 3 olika färger, svart, vit och grå. Dummyvariablerna som skapas för detta föremål blir då följande:

Dummyvariabel 1: 1 om svart, annars 0.

Dummyvariabel 2: 1 om vit, annars 0.

Färgen grå används som referens i detta fall som motsvarar att båda dummyvariablerna sätts till 0. På så sätt kan man skapa dummyvariabler för kategoriska variabler med flera kategorier. Om den kategoriska variabel antar *n* olika värden då skapas *(n-1)* dummyvariabler samt en referensvariabel. *(Sundberg studentkompendium, 2016, s. 82-83).*

2.3 Multikollinjäritet

Linjärt samband mellan två olika variabler kallas multikollinjäritet. Sannolikheten är stor att det statistiska resultatet blir missvisande ifall det råder multikollinjäritet mellan två variabler. Om två multikollinjära variabler förekommer i en och samma modell, kan båda visa på icke signifikant. Även om dessa två variabler i verkliga fallet har ett signifikant inverkan enskilda. För att kunna analysera multikollinjäritet används variansinflationsfaktor VIF.

2.3.1 Variansinflationsfaktor – VIF

Variansinflationsfaktor (VIF) är ett statistisk mått på multikollinjäritet. Ett högt värde på VIF tyder på kollinearitet *(Sundberg studentkompendium, 2016, s. 74).*

Specifikt kan man säga att VIF är variansen på det j: te regressionskoefficienten:

$$Var(\hat{\beta}_j) = \left(\frac{\sigma^2}{\sum_{i=1}^{n}(x_{im} - \bar{x}_m)^2}\right)\left(\frac{1}{1 - R_j^2}\right). \tag{4}$$

Den för faktorn i ekvation (4) är variansen av $\hat{\beta}_j$ för modellen $y_i = \alpha + \beta_j x_{ij} + \epsilon_i$, den andra faktorn kallas för VIF faktorn. Det är svårt att avgöra vilket värde på VIF som anger multikollinjäritet, den vanligaste gränsen brukar ligga mellan 5 oh 10.

R_j^2 betecknar förklaringsgraden (*se avsnitt 2.4.1*) som anger den variation som finns i x_j som förklaras av de övriga förklaringsvariablerna. Förklaringsgraden är noll endast om x_j är ortogonal mot de andra variablerna, i det fallet är VIF=1.

Från ekvation (4) kan man notera att inflationsfaktorn ökar när andel x_j som kan förklaras av de andra variablerna ökar. Med detta kan man säga att ett högt värde på VIF innebär multikollinjäritet.

2.4 Modellval

2.4.1 Förklaringsgrad och justerad förklaringsgrad

Förklaringsgrad (R^2) är ett mått på den totala variationen som förklarar responsvariabeln där måttet tillämpas främst vid modellval. *(Sundberg studentkompendium, 2016, s. 69)*

Förklaringsgraden beskrivs på följande sätt:

$$R^2 = \frac{Kvs(Regression)}{Kvs(Total)} = 1 - \frac{Kvs(residual)}{Kvs(totalt)}.$$

$$\begin{cases} \text{Kvs(regression)} = \sum(\hat{y}_i - \bar{y})^2 \\ \text{Kvs(residual)} = \text{residualkvadratsumman} = \sum(y_i - \hat{y}_i)^2 \\ \text{Kvs(totalt)} = \text{totala kvadratsumman} = \sum(y_i - \bar{y})^2 . \end{cases}$$

Där $\bar{y}$ betecknar medelvärdet av y och $\hat{y}_i$ anger skattade värdet av y_i.

Förklaringsgraden kan anta värden mellan 1 och 0. En modell som erhåller ett högt värde på R^2 föredras, detta för att ju högre R^2 desto mer förklarar modellen responsvariabeln. Fördelen med förklaringsgraden är att den är tillämplig vid modelljämförelse mellan statistiska modeller som enbart infattar samma datamaterial.

En nackdel med förklaringsgraden är att den är beroende av antalet variabler. Detta innebär att R^2 höjs vid inkludering av en variabel, även om den inkluderade variabeln i själva verket inte har någon inverkan på responsvariabeln.

Modifierad förklaringsgrad (R^2_{adj}) används för att undvika problematiken som uppstår av förklaringsgraden. Det är ett statistiskt mått på variationsminskning i den specifika modellen. R^2_{adj} tar hänsyn till antalet variabler i den given modell. *(Sundberg studentkompendium, 2016, s. 70)*

Matematiska formeln för justerade förklaringsgrad;

$$R^2_{adj} = 1 - \frac{\hat{\sigma}^2}{\hat{\sigma}_0^2},$$

där $\hat{\sigma}^2 = \sum(y_i - \hat{y}_i)^2 / (n - k - 1)$ och $\hat{\sigma}_0^2 = \sum(y_i - \bar{y})^2 / (n - 1)$ som variansskattningen när ingen *x*-variabel är inkluderad i modellen.

Sambanden mellan förklaringsgraden och den justerade förklaringsgraden definieras på följande sätt.

$$1 - R^2_{adj} = \frac{(1-R^2)(N-1)}{(N-m-1)}.$$

2.5 Stegvis variabelselektion

Stegvis variabelselektion är den statistiska tekniken som tillämpas vid val av förklarande variabler. Huvudpoängen med dessa tekniker är att succesivt inkludera eller/och utesluta en variabel i taget tills det att man enbart erhåller signifikanta förklaringsvariabler *(Sundberg studentkompendium, 2016, s. 73-74).* Generellt kan man säga att stegvis variabelselektion är uppdelat i tre olika metoder som beskrivs här nedan.

Backward elimination
Metoden innebär att man till en början har med alla förklarande variabler sedan succesivt eliminerar den variabeln som inte visas vara signifikanta. Har man flera variabler som är icke-signifikanta, elimineras den variabeln som är mest icke-signifikanta. Proceduren upprepas tills det att man har en modell med enbart signifikanta variabler.

Forward selection
Tekniken går ut på att man startar med enbart interceptet för att sedan stegvis lägga till en förklaringsvariabel i taget. Den första variabeln som läggs till är den som är mest signifikant, den andra som inkluderas i modellen är den som är näst mest signifikant, osv. Alla signifikanta variabler läggs till i modellen.

Stepwise regression
Stepwise regression kallas även för ”mellan metoden’’ av den anledningen att den är en blandning av de två föregående metoderna som beskrivs ovan. Precis som i Forward Selection börjar man med enbart interceptet och därefter succesivt inkluderar signifikantana variabler. Skillnaden däremot är att man utesluter den variabel som visas vara icke signifikanta under procedurens gång.

3 Datamaterial

Programmet R körs vid datainsamlingen, webbsidorna som används var *Biluppgifter* samt *Bilia*. Genom *Bilia* samlas registreringsnumret samt det utannonserade priset för respektive Volvo. Därefter används registreringsnumret i sökmotorn *Biluppgifter* med syfte att söka upp förklarande variablerna för respektive bil. Totalt fanns det 1809 bilar av märket Volvo utannonserade på Bilias webbplats under februari 2019, vi lyckades få ihop 1297 registreringsnummer. De resterande 512 observationer som saknade registreringsnummer kunde inte inkluderas i den statistiska undersökningen, vi valde att utesluta dessa.

3.1 Kort om Webbsidorna:

’’Biluppgifter *är Sveriges största oberoende fordonsupplysning, vi publicerar fordonsuppgifter som har sitt ursprung i vägtrafikregistret samt flera andra leverantörer. Grundinformationen hämtas och uppdateras regelbundet, oftast sker det varje vecka under tisdag natt.’’ (Billuppgifters webbplats, 2019)*

’’Koncernbolaget ***Bilia AB*** *har över 100 anläggningar och vi finns i Sverige, Norge, Tyskland, Luxemburg och Belgien. Vi är en av Europas största bilkedjor, vilket är en trygghet för dig som kund. Tack vare vår storlek kan vi teckna gynnsamma avtal med våra leverantörer, så att du får bra produkter till bra priser.’’(Bilias webbplats, 2019)*

3.2 Variabelförklaring

Tabell 1: Variabelförklaring.

Variabel	Förklaring	Webbsida
Pris (responsvariabeln)	Försäljningspriset på en Volvo i svenska kronor	Bilia
Hästkrafter	En annan beteckning för effekt. Mätts i hk och innebär 75 (kpm/s) kilopondmeter per sekund 1 kp =9,80665 newton vilket ger att 1 hk=735,499 watt (NE webbplats, 2019)	Biluppgifter
Drivhjul	Antal hjul som överför kraft mellan olika maskindelar i ett fordon. (NE webbplats, 2019)	Biluppgifter
Förbrukning	Antal liter bränsle fordonet förbrukar för varje 100 körda kilometer. Enhet: liter/100 km	Biluppgifter
Utsläpp	Utsläpp av koldioxid (CO_2) per körd kilometer. Enhet: g/km.	Biluppgifter
Toppfart	Maxfart för bilen. Enhet: km/h	Biluppgifter
Ägare	Antal registrerade ägare på personbilen.	Biluppgifter
Tillverkningsår	Årtalet bilar tillverkades	Biluppgifter
Växellåda	Överför kraft mellan motor och *drivhjul* i ett fordon. Väsentliga funktionen är att ändra växlingen beroende på kraftbehov och fart. (NE webbplats, 2019)	Biluppgifter
Bränsle	Ämne som genom förbränning omvandlas till användbar energi, som överförs till motorn.	Biluppgifter
Miltal	Antal körda mil.	Bilia
Modelltyp	Vilken modell bilen klassas som. Volva har bland annat V60, S40 o.s.v.	Biluppgifter

3.3 Dataöversikt

Tabell 2: Max- och Min.värde för respektive variabel.

Variabel	Min.värde	Max.värde
Pris (responsvariabeln)	84800 kr	839900 kr
Hästkrafter	114 hk	367 hk
Fyrhjulsdriven	2 hjul	4 hjul
Förbrukning	3,4 l/100km	10,5 l/100 km
Utsläpp	88 g/km	238 g/km
Toppfart	145 km/h	250 km/h
Ägare	0	8
Tillverkningsår	År 2010	År 2019
Växellåda	-	-
Bränsle	-	-
Miltal	0 mil	25390 miltal
Modelltyp	-	-

4 Statistisk analys

4.1 Bearbetning av data

Vi skapar följande variabler:

- **Ålder** =(säljår-tillverkningsår), säljår = 2019.

Följande dummyvariabler av de kategoriska förklarande variablerna skapas:

- **Fyrhjulsdriven:** 1 om fyrhjulsdriven, 0 om tvåhjulsdriven.
- **Automat:** 1 om automat, 0 om manuell.
- **Diesel:** 1 om diesel, 0 om bensin.

Tabell 3: Slutliga variabler efter bearbetning.

Variabel	Typ	Min.värde	Max.värde
Pris (responsvariabeln)	Numerisk	84800 kr	839900 kr
Hästkrafter	Numerisk	114 hk	367 hk
Fyrhjulsdriven	Dummy	0	1
Förbrukning	Numerisk	3,4 l/100km	10,5 l/100 km
Utsläpp	Numerisk	88 g/km	238 g/km
Toppfart	Numerisk	145 km/h	250 km/h
Ägare	Numerisk	0	8
Ålder	Numerisk	0 år	9 år
Automat	Dummy	0	1
Diesel	Dummy	0	1
Miltal	Numerisk	0 mil	25390 miltal

4.2 Enkel linjär regression

Enkel linjär regression tillämpas på respektive förklarande variabel enskilt, med variabel *pris* som responsvariabeln. Syftet med enkel linjär regression är att undersöka ifall respektive variabel är signifikanta, d.v.s. om varje enskilt variabel har en signifikant inverkan på responsvariabeln. I Tabell 4 nedan anges P-värdet och förklaringsgraden R^2.

Tabell 4: P- värde och R^2 för varje enskild variabel i enkel linjär regression.

Variabel	P-värde	R^2
Hästkrafter	$2{,}2 \cdot 10^{-16}$	0,5678
Fyrhjulsdriven	$2{,}2 \cdot 10^{-16}$	0,3895
Förbrukning	$2{,}2 \cdot 10^{-16}$	0,4119
Utsläpp	$2{,}2 \cdot 10^{-16}$	0,5083
Toppfart	$2{,}2 \cdot 10^{-16}$	0,4247
Ägare	$2{,}2 \cdot 10^{-16}$	0,2246
Ålder	$2{,}2 \cdot 10^{-16}$	0,4250
Automat	$2{,}2 \cdot 10^{-16}$	0,3421
Diesel	0,12	0,001867
Miltal	$2{,}2 \cdot 10^{-16}$	0,3045

.

Enligt det som presenteras i Tabell 4 kan vi påstå att alla variabler förutom dummyvariabeln *diesel* har en signifikant inflytande på responsvariabeln *pris*, på nivån 0,05. Variabeln *hästkrafter* förklarar responsvariabeln mest med R^2= 0,5678 medan variabeln *diesel* förklarar priset minst med R^2=0,001867.

Variansinflationsfaktor – VIF

Tabell 5: VIF-värde för respektive förklarande variabler.

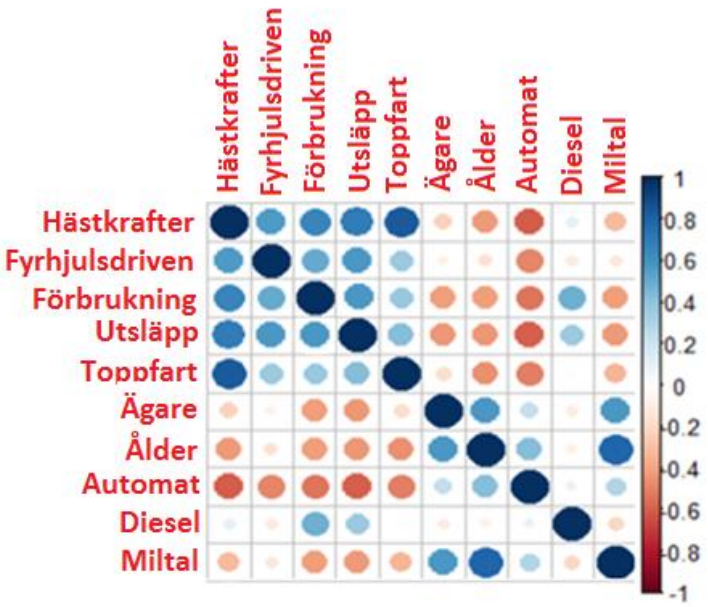
Variabel	VIF-värde
Hästkrafter	2,012448
Fyrhjulsdriven	1,332316
Förbrukning	1,190728
Utsläpp	1,290516
Toppfart	1,673726
Ägare	1,012192
Ålder	1,839235
Automat	1,529206
Diesel	1,011222
Miltal	1,998543

Tabell 5 anger variansinflationsfaktor för respektive variabel. Vi väljer sätt 5 som högsta värdet, detta innebär att överstiger VIF värdet 5 då kan man dra slutsatsen att multikollinjäritet råder mellan variabeln och responsvariabeln (*pris*). Resultatet i Tabell 5 visar att det inte existerar någon multikollinjäritet mellan förklarande variablerna och responsvariabeln. *Hästkrafter* med den högst VIF- värdet understiger värdet 5.

För att analysera hur de olika variablerna är korrelerade i förhållande till varandra undersöker vi multipel linjär regression och korrelationsmatris.

4.3 Multipel linjär regression

Korrelationsmatris



Figur1: Korrelationsmatris mellan förklarande variablerna.

Figur1 förklarar en stark korrelation mellan:

Toppfart ~ Hästkrafter

Vi antar att variablerna *toppfart* och *hästkrafter* ger likartad information och av denna anledning väljer vi att eliminera variabeln *toppfart* samt behålla variabeln *hästkrafter*. *Hästkrafter* förklarar priset mer än vad *toppfart* gör (se Tabell 4) därför väljer vi utesluta variabeln *toppfart*.

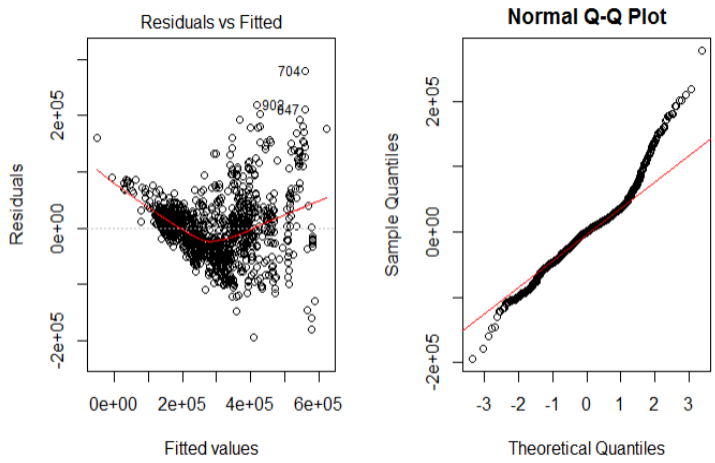
Miltal ~ Ålder.

Den starka korrelationen mellan variablerna *miltal* och *ålder* löser vi genom att skapa en kvotvariabel mellan *miltal* och *ålder*. Variabeln mil per år skapas som *mil/år*. Vi väljer därmed att eliminera bort variabeln *miltal* samt behålla variabeln *ålder*.

Grundmodellen

Multipel linjär regression för grundmodellen:

$$Pris_i = \alpha + \beta_1 Hästkrafter + \beta_2 Förbrukning + \beta_3 Utsläpp + \beta_4 Fyrhjulsdriven + \beta_5 Ägare + \beta_6 Ålder + \beta_7 Automat + \beta_8 Diesel + \beta_9 Mil/år + \epsilon_i.$$



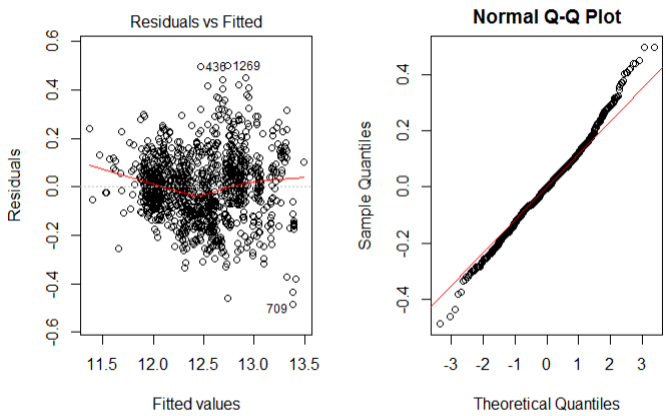
Figur 2: Residualplot och normalfördelningsplot för grundmodellen.

En förutsättning vid tillämpning av multipel linjär regression är att residualerna ska ha en konstant varians och följa en normalfördelning. Den vänstra plotten i Figur 2 tyder på att residualerna inte är slumpmässiga kring noll, vilket i sin tur innebär att variansen inte är konstant. Baserad på den högra delen av Figur 2 kan vi även utesluta att residualerna följer en normalfördelning.

Logtransformering

Modell 1 Logtransformering av responsvariabeln (*pris*)

$$Log(Pris_i) = \alpha + \beta_1 Hästkrafter + \beta_2 Förbrukning + \beta_3 Utsläpp + \beta_4 Fyrhjulsdriven + \beta_5 Ägare + \beta_6 Ålder + \beta_7 Automat + \beta_8 Diesel + \beta_9 Mil/år + \epsilon_i$$

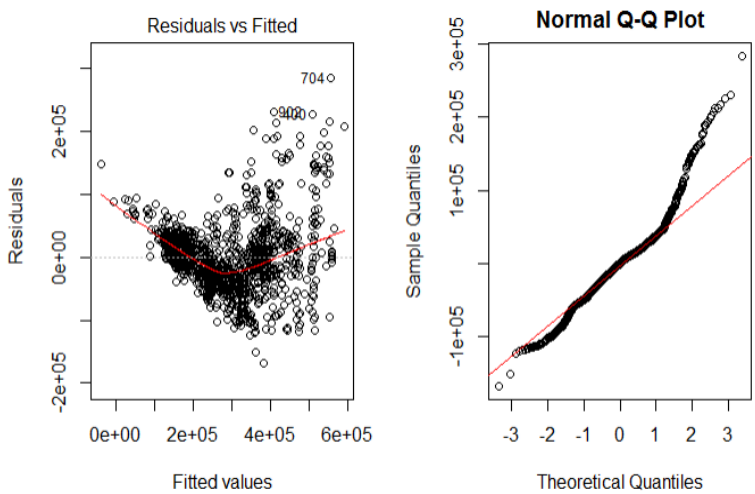


Figur3: Residualplot och normalfördelningsplot för modell 1.

Modell 2: Logtransformering av *hästkrafter*.

Hästkrafter erhåller den högsta förklaringsgraden (Tabell 4), följaktligen väljer vi att logtransformera just *hästkrafter*.

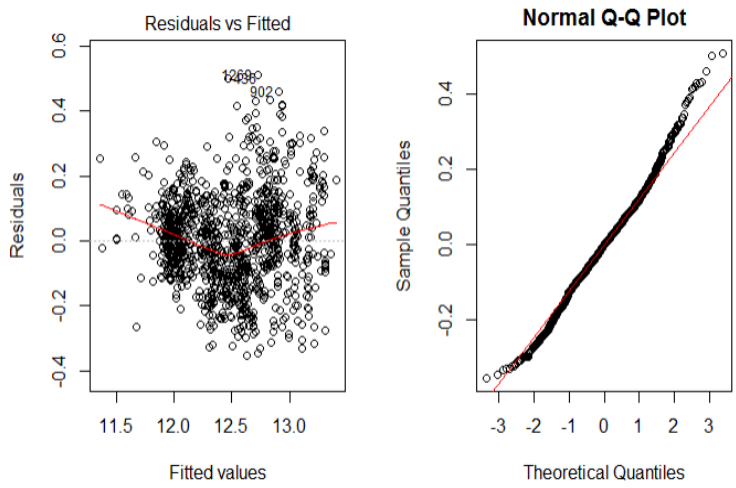
$$Pris_i = \alpha + \beta_1 Log(Hästkrafter) + \beta_2 Förbrukning + \beta_3 Utsläpp + \beta_4 Fyrhjulsdriven + \beta_5 Ägare + \beta_6 Ålder + \beta_7 Automat + \beta_8 Diesel + \beta_9 Mil/år + \epsilon_i.$$



Figur 4*:* Residualplot och normalfördelningsplot för Modell 2.

Modell 3: Logtransformering av *pris* och *hästkrafter*.

$$Log(Pris_{i)} = \alpha + \beta_1 Log(Hästkrafter) + \beta_2 Förbrukning + \beta_3 Utsläpp + \beta_4 Fyrhjulsdriven + \beta_5 Ägare + \beta_6 Ålder + \beta_7 Automat + \beta_8 Diesel + \beta_9 Mil/år + \epsilon_i.$$



Figur 5: Residualplot och normalfördelningsplot för Modell 3.

Residualploten och normalfördelningsploten som redovisas Figur3-5 visar på att modell 1 (Figur 3) samt modell 3 (Figur 5) uppfyller bäst villkoren för multipel linjär regression. Residualploten som representerar den vänstra delen av figuren visar att residualerna inte följer ett mönster utan de kan man tolka som slumpmässiga kring 0. Att residualerna är slumpmässiga kring noll innebär i sin tur att de följer en konstant varians. Q-Q ploten i både Figur 3 och 5 påpekar att residualerna följer den röda linjen, vilket framgår att de följer en normalfördelning.

Modell 2 som redovisas i Figur 5 tyder inte på att residualerna är slumpmässiga kring noll eller att de följer normalfördelning. Av denna anledning uteslutas modell 2 i kommande analys.

Variabelselektion

Variabelselektion (se avsnitt 2.5) tillämpas på modell 1 och modell 3. Backward elimination, Forward selection och Stepwise regression erhöll samma resultat. Resultatet av samtliga metoder blir att eliminera bort variablerna *förbrukning*, *ägare* och *diesel*.

Följande modeller får vi efter variabelselektion:

Modell 1 efter variabelselektion [=Modell1V]:

$$log(Pris_i) = \alpha + \beta_1 Hästkrafter + \beta_2 Utsläpp + \beta_3 Fyrhjulsdriven + \beta_4 Ålder + \beta_5 Automat + \beta_6 Mil/År + \epsilon_i.$$

Modell 3 efter variabelselektion [=Modell3V]:

$$log(Pris_i) = \alpha + \beta_1 \log(Hästkrafter) + \beta_2 Utsläpp + \beta_3 Fyrhjulsdriven + \beta_4 Ålder + \beta_5 Automat + \beta_6 Mil/År + \epsilon_i$$

4.4 Modelljämförelse

Tabell 6: Justerad förklaringsgrad för modellerna efter variabelselektion.

Modell	R^2_{adj}
Modell 1V	0,89789
Modell 3V	0,74988

Tabell 6 visar att modell 1V har ett högre R^2_{adj} än modell 3V vidare kan vi anta att modell 1V förklarar de observerade data bättre än modell 3V.

4.5 Förklarande variabel *Modelltyp*

Som ett intressant vidare analys, fann vi det intresseväckande att undersöka hur mycket modelltypen för en specifik bil egentligen tilltalar vid prissättningen. Av denna anledning elimineras förklarande variabler som anger bilen egenskaper och istället tillämpar vi multipel linjär regression på variabler som presenteras i Tabell 7.

Tabell 7: Variabler för data med *Modelltyp.*

Variabler	Typ
Pris (responsvariabeln)	Numerisk
Ålder	Numerisk
Ägare	Numerisk
Mil/År	Numerisk
Modelltyp	Kategorisk

Vi skapar följande dummyvariabler av *Modelltyp*

- V70: 1 om *Modelltyp* är V70, annars 0.
- V60: 1 om *Modelltyp* är V60, annars 0
- XC40: 1 om *Modelltyp* är XC40, annars 0.
- XC60: 1 om *Modelltyp* är XC60, annars 0.
- V40: 1 om *Modelltyp* är V40, annars 0.

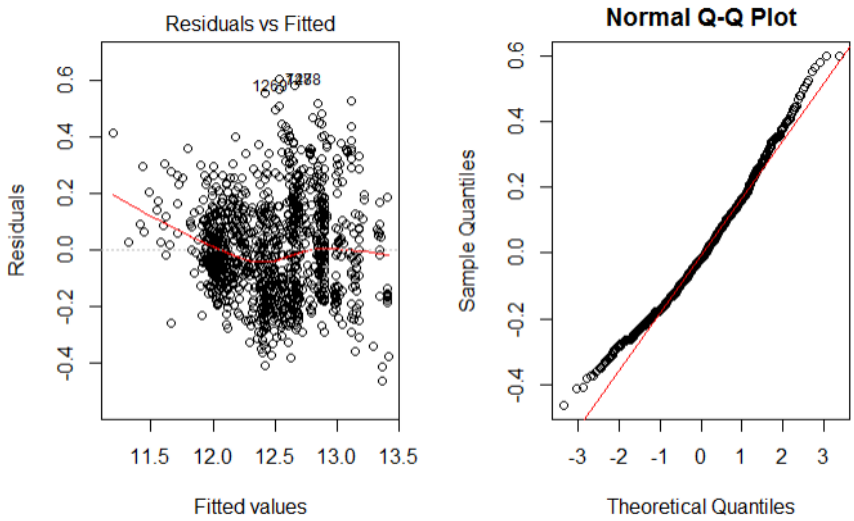
- XC90: 1 om *Modelltyp* är XC90, annars 0.
- S90: 1 om *Modelltyp* är S90, annars 0.
- V90: 1 om *Modelltyp* är V90, annars 0.
- XC70: 1 om *Modelltyp* är XC70, annars 0.
- S60: referensvariabel, om alla variablerna ovan är 0 är *Modelltyp* S60

Tekniken som används i avsnitt 4.2–4.3 tillämpas på data med *Modelltyp* vilket resulteras i följande modell 4.

Modell 4:

$$\log(Pris_i) = \alpha + \beta_1 \text{V70} + \beta_2\ \text{V60} + \beta_3\ \text{XC40} + \beta_4 \text{XC60} + \beta_5\ \text{V40} + \beta_6\ \text{XC90} + \beta_7\ \text{S90} + \beta_8\ \text{V90} + \beta_9\ \text{XC70} + \beta_{10}\ \text{Ä}gare + \beta_{11}\text{Å}lder + \beta_{12} Mil/\text{å}r + \epsilon_i$$

Residualplot och normalfördelningsplot



Figur6: Residualplot och normalfördelningsplot för Modell 4.

Modelljämförelse

Jämförelse mellan modell 4 (Figur 6) och modell 1V (Tabell 6).

Tabell 8: Justerad förklaringsgrad för modell 4 samt modell 1V.

Modell	$\mathbf{R^2_{adj}}$
Modell 4	0,8965
Modell 1V	0,89789

Med en mycket liten marginal kan man utifrån Tabell 8 dra slutsatsen att modell 1V förklarar data bättre än modell 4, annars är modellerna enligt analysen ovan mycket likgiltiga.

5 Resultat

Statistiska analysen som genomförs i detta arbete resulterar i att vi kan dra slutsatsen att Modell 1V och Modell 4 är ganska likvärdiga. Med en liten marginal är Modell 1V den bättre modellen. Av denna anledning, väljer vi att djupare undersöka Modell 1V härnäst:

Modell 1V

Med logaritmerad responsvariabeln erhålls följande:

$$log(Pris_i) = \alpha + \beta_1 Hästkrafter + \beta_2 Utsläpp + \beta_3 Fyrhjulsdriven + \beta_4 Ålder + \beta_5 Automat + \beta_6 Mil/\text{År} + \epsilon_i.$$

Icke logaritmerad responsvariabeln erhålls följande:

$$Pris_i = e^{\hat{\alpha}} \cdot e^{\hat{\beta}_6 \text{Hästkrafter}} \cdot e^{\hat{\beta}_3 \text{Utsläpp}} \cdot e^{\hat{\beta}_3 \text{Fyrhjulsdriven}} \cdot e^{\hat{\beta}_4 \text{Ålder}} \cdot e^{\hat{\beta}_5 \text{Automat}} \cdot e^{\hat{\beta}_6 \text{Mil/År}} \cdot \epsilon'. \quad (5)$$

Notera att $\epsilon' = e^{\hat{\epsilon}}$ är log normalfördelad med väntevärde $e^{\sigma^2/2}$ och varians $\left(e^{\sigma^2} - 1\right)e^{\sigma^2}$.

Formeln (5) med icke logaritmerad responsvariabeln innehåller multiplikativa faktorer. De multiplikativa faktorer kommer att medföra att responsvariabeln förändras multiplikativt med e^{β_i} för varje enhetsförändring i respektive förklarande variabel. I Tabell 9 nedan redovisas hur mycket den icke logaritmerade responsvariabeln förändras procentuellt. Den procentuella förändringen erhålls genom att dra bort en enhet från $e^{\beta_i x_i}$ för att sedan multiplicera med 100.

Tabell 9: Resultat av modell 1V.

Koefficient	P-värde	Estimat	$e^{estimat}$	Förändring
Hästkrafter	$1{,}3 \cdot 10^{-11}$	$2{,}762 \cdot 10^{-3}$	1,0028	0,28 %
Fyrhjulsdriven	$2{,}2 \cdot 10^{-16}$	$4{,}846 \cdot 10^{-2}$	1,0496	4.96 %
Utsläpp	$3{,}2 \cdot 10^{-5}$	$-2{,}811 \cdot 10^{-4}$	0,9997	– 0,028 %
Automat	$1{,}2 \cdot 10^{-14}$	$3{,}214 \cdot 10^{-3}$	1,003	3 %
Ålder	$5{,}2 \cdot 10^{-8}$	$-8{,}925 \cdot 10^{-2}$	0,9146	– 8.54 %
Mil/år	$4{,}5 \cdot 10^{-6}$	$-3{,}756 \cdot 10^{-3}$	1,00376	– 0,37 %

Tabell 9 sammanställer resultatet av slutmodellen (Modell 1V) där det låga p-värdet för respektive variabel tyder på att alla variabler är signifikanta på nivån 0,05. Från Tabell 9 kan vi även läsa den procentuella förändringen i priset när det ökar med en enhet för respektive koefficient. Vi kan t.ex. notera att antalet *hästkrafter*, *fyrhjulsdriven* samt *automat växellåda* har en positiv förändring på responsvariabeln (*pris*).

Det som framgår av Tabell 9 kan tolkas som att en *fyrhjulsdriven* bil är ungefär 4,96 % dyrare än en tvåhjulsdriven. Det noteras även att en bil med *automat växellåda* är i genomsnitt 3 % dyrare än en bil som har en *manuell växellåda*. Vi kan även påpeka att för varje ökad enhet i *hästkrafter* ökar priset med drygt 0,28 %. Utifrån Tabell 9, kan man även notera en negativ inverkan på priset för följande variabler: *utsläpp av koldioxid*, *ålder* på bilen samt antal körda mil per år.

Utsläpp av koldioxid har enligt Tabell 9 den minsta påverkan på priset och detta kan tolkas som att när antalet gram koldioxid per körd mil ökar med en enhet minskar priset med drygt 0,028 procentenheter. *Ålder* på bilen har däremot den största negativa effekten på priset av en begagnad bil av märket Volvo. Det innebär att priset blir 8,54 procentenheter mindre än föregående år. Antalet utsläppta gram av koldioxid per körda kilometer påverkar priset med – 0,37 %.

5 Diskussion

Huvudsyftet med detta självständiga arbete är att undersöka vilka förklarande faktorer som har ett statistiskt samband på priset för en Volvo som är utannonserad på Bilia. Modell 1 som presenteras i Figur 3 samt Modell 3 som presenteras i Figur 5 är två reducerade modeller som uppfyller villkoren för linjär regression.

Stegvis variabelselektion på Modell 1 och Modell 3 resulterar i att förklarande variabler som *förbrukning*, *ägare* samt dummyvariabel *diesel* inte är signifikanta för prissättning av en Volvo. Av denna anledning väljer vi att eliminera dessa variabler från Modell 1 och Modell 3.

Från Tabell 6 kan vi dra slutsatsen att Modell 1V (Modell 1 efter variabelselektion) förklarar ungefär 89,8% av den totala variationen i priset. Resultatet innebär att modellen förklarar en stor variation för de observerade data.

En annan aspekt av den statistiska analysen som tillämpas i detta arbete är att undersöka datamängden utan förklarande variabler som står för bilens egenskaper (t.ex. *bränsl*e, *hästkrafter* osv) där variabeln *modelltyp* istället inkluderas. Ändamålet med detta är att undersöka hur mycket modellbeteckningen egentligen förklarar priset. Modell 4 som presenteras i Figur 6 är den mest anpassade modellen inom multipel linjär regression med *modelltyp* som dummy förklarande variabler.

Slutligen utförs en modelljämförelse mellan modell 4 och modell 1V. Syftet med denna jämförelse är att undersöka om priset på en Volvo sätts efter modellbeteckningen eller efter bilens egenskaper. Tabell 8 introducerar justerade förklaringsgraden för respektive modell där vi utifrån denna kan dra slutsatsen att dessa två modeller är väldigt likvärdiga. Slutsatsen är att priserna är satta efter modellernas egenskaper snarare än efter modellbeteckning. Detta eftersom resultatet tyder på att vi kan prediktera priset lika väl med egenskaperna som med modellbeteckningen.

Vidare väljer vi att fördjupa oss i modell 1V för att se hur respektive förklarande variabel påverkar priset. Tabell 9 består av den procentuella förändringen i priset när en förklarande variabel förändras med en enhet. Slutsatsen av denna analys är att *fyrhjulsdrivna* bilar är dyrare än tvåhjulsdrivna. Vi kan även konstatera att personbilar med *automat växellåda* är dyrare än bilar med *manuell växellåda*.

Enligt Tabell 9 har faktorn *ålder* den största inverkan på priset där det framgår att för varje år sänks priset procentuellt med ungefär 9 procentenheter. Att priset sänks för varje år är inget överraskande däremot kan man ifrågasätta om priset verkligen följer en konstant förändring genom alla åren. Generellt brukar priset på bilar sjunka mer de första 1-5 åren. Det man eventuellt hade kunnat göra är att dela upp förklarande variabeln *ålder* i två undergrupper. En variabel som motsvarar ’’nya bilar’’ för alla bilar som är 5 år eller yngre och en annan variabel för ’’äldre bilar’’ för alla bilar äldre än 5 år. På detta sätt kan man undersöka hur det skiljer sig mellan åren.

Resultatet av arbetet tyder på att vi har lyckats få en fram en modell som täcker en stor variation av priset på en begagnad bil. Alternativa felkällor i detta arbete är bl.a. vid datainsamling. Enbart 1297 objekt av totalt 1809 möjliga objekt kunde analyseras i arbetet, vilket innebär att drygt 28.3 % av tillgänglig data inte har analyserats.

7. Referenslista

- ANDERSSON, LINDENSJÖ och TYRCHA, J. 2018. Notes in Econometrics. Stockholm. Department of Mathematics, Stockholm University.
- SUNDBERG, R., 2014 Lineära Statistska Modeller Stockholm: Department of Mathematics, Stockholm University
- Billuppgifters webbplats, 2019, hämtad 2019-03-03, https://biluppgifter.se/om
- Bilias webbplats, 2019, hämtad 2019-03-03 ,https://www.bilia.se/sok-bland-vara-bilar/
- NE webbplats, 2019, hämtad 2019-03-04, https://www.ne.se/uppslagsverk/encyklopedi/l%C3%A5ng/v%C3%A4xell%C3%A5da
- NE webbplats, 2019, hämtad 2019.03-05, https://www.ne.se/uppslagsverk/encyklopedi/l%C3%A5ng/bil
- SCB webbplats, 2019, hämtad, 2019-03-20, https://www.scb.se/hitta-statistik/statistik-efter-amne/transporter-och-kommunikationer/vagtrafik/fordonsstatistik/