Regressionsanalys

Volvo v60 pris-analys



Andreas Wendel

EC Utbildning

Kunskapskontroll, R-programmering

2024-04

# Abstract

This report will give you an insight to how a Volvo v60 car price is valued using regression analysis. A Models will be tested and using those models you can see the statistical significance of the chosen variables and how much those independent variables will affect the dependent variable. Said model will also be able to predict new Volvo v60 prices.

Innehållsförteckning

[Abstract 2](#_Toc165048164)

[1 Inledning 1](#_Toc165048165)

[1.1 Frågeställningar 2](#_Toc165048166)

[2 Teori 3](#_Toc165048167)

[2.1 Insamling av data 3](#_Toc165048168)

[2.1.1 Data bredd 3](#_Toc165048169)

[2.1.2 Variabler från datan 3](#_Toc165048170)

[2.2 Logistisk modell 3](#_Toc165048171)

[2.2.1 betaparametrar, epsilon 4](#_Toc165048172)

[2.3 Variabel analys, Noll-hypotes 4](#_Toc165048173)

[2.4 Model evaluering 5](#_Toc165048174)

[2.4.1 Best subset selection 6](#_Toc165048175)

[2.4.2 BIC, Cp 7](#_Toc165048176)

[2.4.3 Cross validation 8](#_Toc165048177)

[2.4.4 RMSE 8](#_Toc165048178)

[3 Metod 9](#_Toc165048179)

[3.1 Insamling av data 9](#_Toc165048180)

[3.2 Undersökning av data 9](#_Toc165048181)

[3.2.1 Behövs variabler transformeras om? 9](#_Toc165048182)

[3.2.2 kolla om det finns några datapunkter som ”skadar” våran modellering 9](#_Toc165048183)

[3.2.3 Linjärt förhållande, heteroskedasticitet, normalfördelade residualer 10](#_Toc165048184)

[3.3 Statistiks inferens av variabler 12](#_Toc165048185)

[3.4 Model evaluering 13](#_Toc165048186)

[3.4.1 Best subset selection 13](#_Toc165048187)

[3.4.2 Cross-validering 14](#_Toc165048188)

[3.5 Testa bästa modellen på ny observation 15](#_Toc165048189)

[4 Resultat och Diskussion 16](#_Toc165048190)

[4.1 Val av data och variabler. 16](#_Toc165048191)

[4.1.1 Datakälla 16](#_Toc165048192)

[4.2 Transformering av data 16](#_Toc165048193)

[4.3 Oberoende variabeln och modellens prediktioner 17](#_Toc165048194)

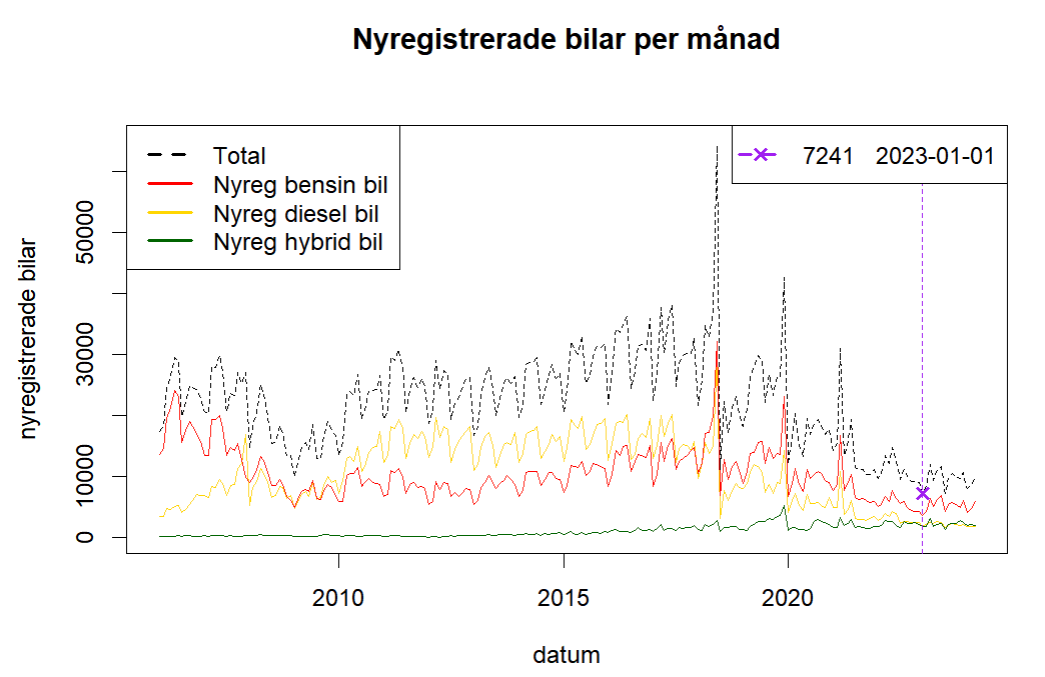
[5 Slutsatser 18](#_Toc165048195)

[Källförteckning 19](#_Toc165048196)

# Inledning

Under de senaste åren har världen kommit till en låg konjunktion, höga räntor, dyra varor och en dyrare vardag att leva i. Något man kan tänka sig ändras tillsammans riksbankens räntor är andel lån, majoriteten av människor finansiera sina bilköp och med höga räntor kommer högre kostnader för konsumenten. Om man kollar på Figur 1 så kan man se en drastisk minskning av nyregistrerade bilar från toppen av 2018–06 (63 473 bilar) till en ny botten av 2023–01 (7241) vilket tyder på att konsumenter väljer att behålla sina gamla bilar längre eller att man vänder sig åt begagnade bilar som kan köpas utan lån eller till mindre lån som inte lägger lika mycket press på konsumenten ekonomiskt.

Syftet med denna rapporten kommer vara att analysera dagens bilmarknad. En regressionsmodell kommer att byggas och analyseras samt även kunna prediktera nya bilars priser. Genom modellen kommer man kunna se signifikansen av vad som prissätter en begagnad bil samt hur dessa variabler påverkar priset.



Figur 1: Data från hela riket 2006-01 till 2024-03 från scv.se : Notera att datan börjar ifrån 2006–01 eftersom då började drivmedel(elhybrid) dokumenteras

## Frågeställningar

Man kommer också att besvara på följande frågor

1. Vilka variabler skall ingå i analysen.
2. Hur signifikanta är variablerna
3. Hur hanterar man datan
4. Vilken/vilka modeller hanterar våran analys bäst och varför
5. Hur väljer man en slutlig modell
6. Kan modellen prediktera nya priser

# Teori

## Insamling av data

Datainsamling kommer teoretiskt sätt bara att handla om att samla in relevant data. För att göra en modell på begagnade bilar så måste man samla in data om begagnade bilar. Relevant data kring detta område kan bara komma från bilar som har haft minst en ägare, dessutom kan miltalet på en bil inte vara noll, det skulle intyga att bilen inte har använts eller då är ny och kommer direkt från fabrik.

Denna relevanta datan kan komma från individer så som ens granne eller folk ute i trafiken, man kan samt kolla på annonser där folk lägger ut sina bilar.

### Data bredd

Något att konstatera är också hur bredd man vill ha datan. Att jämföra vardagsbilar så som en Volvo till lyxbilar som en Porsche kan medföra stor varians i datan. En modell med hög varians kommer ha svårigheter att prestera och att medföra för mycket varierade data som kan leda till en dålig modell jämfört med en modell som specifikt undersöker exempelvis ett bilmärke.

### Variabler från datan

Att begränsa antal variabler kan vara bra för modellen eftersom det minskar variansen hos modellen men det medför också att modellen blir begränsad till just dem variablerna eller den typen av bilar. Dessutom så medföra det att man får ha ett minimumantal observationer för varje oberoende variabel för att modellen skall kunna hitta en bra betaparameter. Konstatera också att det finns variabler som inte är lika enkelt att dokumentera som specifika fysiska skador eller tekniska problem.

## Logistisk modell

Regression-analys handlar om att bygga modeller från datan och sedan kunna dra statistiska slutsater.

Man kan först konstatera att eftersom modellen kommer ha mer än 1 variabel så kommer man ha en multipel linjär regressionsmodell. Det betyder att man kommer ha en flervariabelsfunktion i modell se ekvation 1. Simpelt sätt betyder det att man kommer ha ett Y-värde (kallat beroende variabel) som påverkas av flera X värden (kallat oberoende variabler). Modellens jobb är att genom observationer hitta en regressionslinje mellan alla observationer och då samtidigt hitta koefficienter till varje variabel.

Djupare in kan man säga att regressionsmodellen hanterar funktionen som en partial derivata där modellen hittar den genomsnittliga effekten(koefficienten) för varje variabel när alla andra variabler är konstanta, modellen gör detta i iteration tills alla variabler har en koefficient som gör att modellen kan träna in observationerna för att sedan prediktera in nya värden.

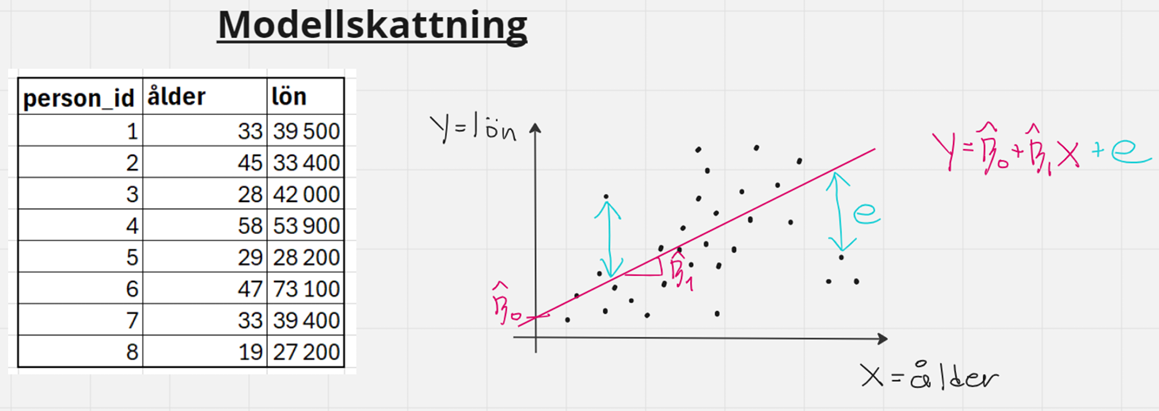
Ekvation 1

### betaparametrar, epsilon

Viktigt att medföra att alla observationer kommer medföra ett delta ifrån de värden som regressionsmodellen ger se Figur 2, här ser man en visualisering av en linjär modell notera att B\_1 har en hatt det tyder på att det är modellens skattning av betaparametern(koefficienten) och inte den ”riktiga”, att få den riktiga parametern är möjligt om man till exempel gör modellen till en polynomfunktion men det kommer medföra en modell som bli ”overfitted” vilket medför att modellen kommer ha svårt att prediktera nya observationer eftersom dem nya observationerna kommer ändra den ”riktiga” betaparametern, därför vill man ha en skattning som funkar bra på nya observationer. En betaparameter tolkas som ökningen av y när x(variabeln) ökar. Eftersom man inte har den ”riktiga” parametern så kommer regressionslinjen inte att täcka alla observationer. För att hitta en skattade betaparametern så använder man sig av RSS ”Residual sum of squares”. Kolla på Figur 2 igen så kan man se att epsilon tolkas som skillnaden från observationerna till modellens värden.

RSS handlar om att summera alla kvadratiska epsilon (kvadratisk eftersom man inte vill ha negativa värden) och sedan välja den betaparametern som har minst RSS se ekvation 2.

Ekvation 2 ŷ = modellens predikterade beroende variable



Figur 2: Linjär modell där y(lön) påverkas av x(ålder) och person\_id är flera observationer

## Variabel analys, Noll-hypotes

Eftersom variablernas betaparametrar är en skattning av den ”riktiga” så kan utnyttja statistisk inferens för att exempelvis skapa konfidensintervall eller genomföra hypotesprövningar. Det medför att man kan undersöka betaparametrarana och se hur dem påverkar modellen.

Notera att statistik inferens handlar om att göra intelligenta gissningar med hjälp av ett slumpmässigt urval och med detta urvalet kan man då exempelvis göra skattningar av våran betaparameter.

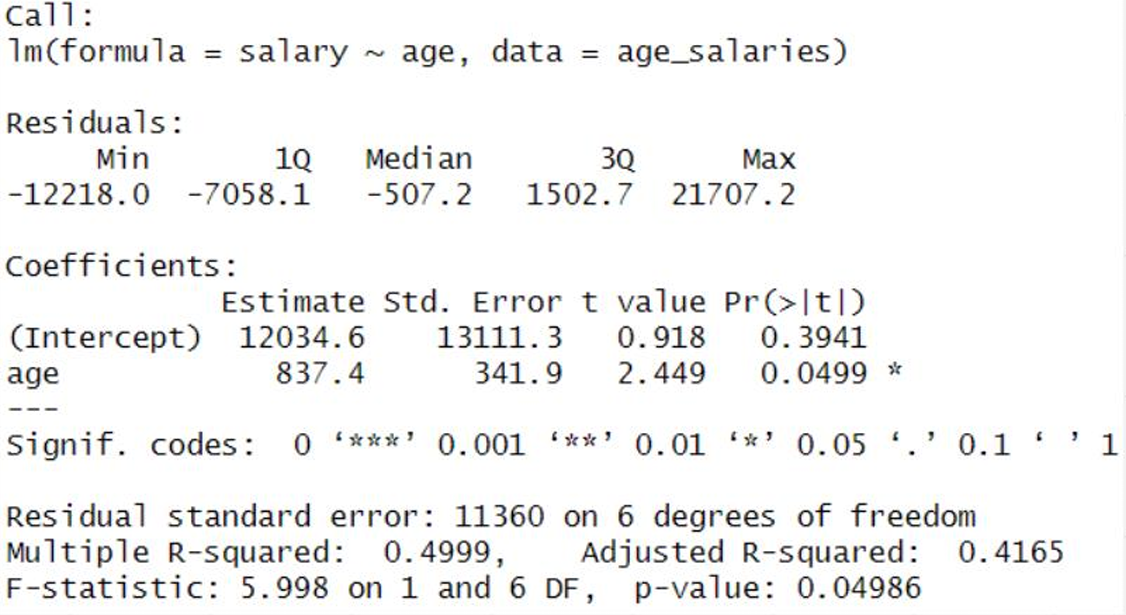
I detta fallet är urvalet observationerna och skattningen är våran betaparameter

man kan till exempel se konfidens intervallet för en betaparameter, vilket kan visa sannolikhet (ofta 95%) för vilket intervall som betaparametern kan anta. Till exempel om vår variabel är ålder och vårat y är lön så kan man säga att våran betaparameter för variabeln ålder kan med 95% sannolikhet anta ett värde i ett intervall ex 5000–15000. Det intervallet är en ”gissning” från hela populationen utifrån vårat urval (observationer). Detta funkar dock bara om residualerna(epsilon) antas vara normalfördelade då statistik inferens utgår ifrån de antagandet.

Något annat man kan se är signifikans hos variabler. Genom att utföra en nollhypotes så kan man se sannolikheten för om en variabels betaparameter är noll eller med andra ord inte påverkar vårat y.

Man använder sig av en testfunktion se ekvation 3, som är t-fördelad med n-2 frihetsgrader. Om H\_0 är sann kan man förkasta noll-hypotesen och säga att det är statistiskt signifikant att x variabeln påverkar y. Man får även ut ett p-värde som ger nivån av signifikant men p-värdet säger egentligen den sannolikhet att förkasta en sann noll-hypotes. Exempel om x variabeln får ett p-värde på 1% (gränsen är oftast 5%) så kan man förkasta noll-hypotesen, det betyder egentligen att det finns 1% sannolikhet att man förkasta en sann noll-hypotes vilket betyder att det finns 1% chans att betaparametern faktiskt är noll trots att det är väldigt låg sannolikhet, så låg att man kallar det ej signifikant. Se figur 3 här är ett kodexempel där modellen säger att det är p-värdet<5% och då kan man förkasta noll-hypotesen och säga att ålder faktiskt påverkar lön

Ekvation 3



Figur 3: information om statistik inferens med modellen y(lön)=x(ålder). intercept kan tolkas som värdet när alla x är = noll alltså om ålder är 0 så är ens lön = 12034, konstatera att man troligtvis inte har någon lön när man inte är född vilket kan vara ett potentiellt problem för modellen. Variabeln ålder är precis under 5% och blir då signifikant, man kan då förkasta noll hypotesen och säkert säga att ålder påverkar lönen och enligt modellen så påverkas lönen med 837 per ökning ålder

## Model evaluering

Att evaluera en modell handlar om att välja den modell som ger bäst resultat, resultat kan jämföras på många sätt men normalts sätt handlar det om att ha så låg epsilon som möjligt, ett problem blir dock att ju mer variabler man har, ju mindre kommer residualerna att bli, det betyder dock inte att ha flera variabler kommer leda till en bättre modell. En bra modell skall kunna prediktera observationer till ett resonabelt värde oftast inom ett lågt intervall och inte långt ifrån det verkliga observationerna.

man kan evaluera modeller genom flera mått. De två vanligaste metoderna är att använda sig utav RSE ”residual standard error” som uppskattar standardavvikelsen för residualerna se ekvation 4. Simpelt sett så kan denna formeln visa hur mycket modellens prediktions värden avviker i snitt från de sanna värdena men det är viktigt att konstatera att prediktioner görs på samma data som modellen tränas på, alltså inte ny data som modellen inte har sätt innan.

Ett annat sätt är att evaluera en modell är med Determinationskoefficienten också känd som R^2 se ekvation 5. Dessa mått använder sig också utav RSS men dessutom av TSS som står för ”total sum of squares” denna formeln kommer visa hur stor andel (värde mellan 0 och 1) av variation den beroende variabeln (Y) kan förklaras av end oberoende variabeln (x).

Notera att R^2 kommer att vara lika stor eller större om man medför flera variabler eftersom modellen kommer ha lika små eller mindre residualer på träningsdatan vilket måttet evaluerar, det betyder dock inte att residualerna på testdatan kommer ha lika små fel när man har så många variabler. Istället använder man sig av Adjuster R^2 se ekvation 6, vilket medför ett typ av straff för att ha flera variabler. Tanken med adj-R^2 är att efter korrekt antal variabler så kommer ytligare variabler bara medföra liten ökning i RSS vilket straffar adj-R^2 då p ökar.

ekvation 4: n=observationer, p=antal variabler

, ekvation 5: ȳ = medelvärdet den beroende variabeln

ekvation 6: notera för att adjusted-R^2 skall öka så måste minskningen hos 1-R^2 kompensera för ökningen hos n-1/n-p-1 när p(variabler) ökar.

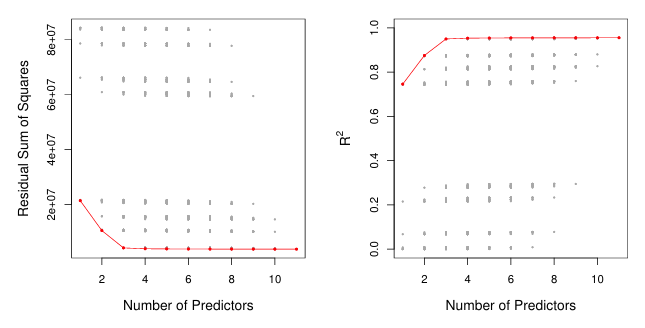
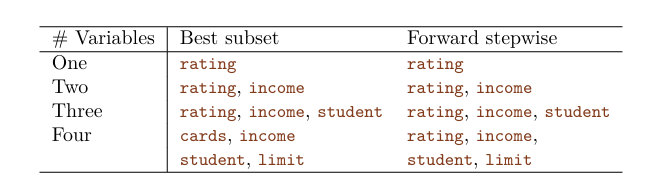
### Best subset selection

Ett sätt att evaluera modeller med olika antal variabler kan vara genom att använda sig av en metod som heter best subset selection. Best subset selection handlar om att stegvis testa och evaluera flera modeller beroende på hur många variabler man har. Steg för steg börjar man med att

1. Tar fram Modell M\_0 som kallas för noll modellen som innehåller inga oberoende variabler, denna modellen predikterar bara observationernas medelvärde.
2. I nästa steg så testar man modeller stegvis för varje iteration av variabler man kallar dem predictors. Låt säga att man har 10 predictors kallas p, man börjar då med att göra 10 modeller som innehåller varsin predictor och sedan väljer den model som ger bäst resultat utifrån bland annat RSE och adj-R^2 och kallar den för M\_1
3. I nästa steg kommer man att göra samma sak men väljer att ha 2 predictors i varje model, Detta medför att man kommer ha = 45 kombinationer, man väljer sedan den modellen som ger bäst resultat.
4. Dessa steg görs om med kombinationer där k är antal predictors i given modell.
5. Tillslut kommer du få p+1 antal modeller (glöm inte noll modellen)

Tillslut kan man se påverkan hos modellen givet vilka och hur många oberoende variabler man använder i modellen se Figur 4. Notera att best subset selection kommer medföra att det tränas 2^p antal modeller som skall evalueras, det betyder att vid väldigt höga antal variabler p så kan det krävas väldigt mycket datorkraft.

För att förhindra det kan man välja använda sig av forward stepwise selection vilket istället för att testa alla kombinationer så testar men ytligare modeller genom att stegvis lägga till nästa bästa predictor, det betyder att man testar p-k modeller vilket medför att man kommer träna totalt 1+p(p+1)/2 modeller. Det kan dock leda till att man inte testar dem optimala modellerna se Figur 5



Figur 5: Skillnaden mellan Best subset och Forward Stepwise selection med datasetet Credit. Notera att Best subset hat hittat den bästa 4 predictor modellen eftersom den har testat alla kombinationer jämfört med Forward stepwise

Figur 4: Resultat från best subset selection med 10 oberoende variabler, notera att efter 3 predators så slutar modellen prestanda att öka

### BIC, Cp

BIC och Cp är två andra sätt att evaluera modeller som också dyker upp hos best subset selection. Båda två beräkningar uppskattar modellens test errors så som Adj-R^2 gör fast istället använder variansen av fel hos epsilon se ekvation 7. Simpelt sätt så om σhatt^2 är en o-bias skattning av σ^2 så kan man förmoda att Cp är en o-bias skattning av test MSE ”mean squared error”. Det betyder att Cp tenderar att anta små värden när en modell har låga test fel.

ekvation 7: där d = predictors i modellen

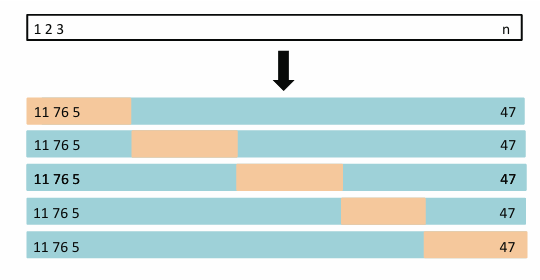
Likt Cp så fungerar BIC på liknande sätt. BIC ”Bayesian information criterion” tenderar också att ta små värden när en modell har låg test error se ekvation 8. Här är istället koefficienten 2 utbytt mot en log(n) där n är antal observationer. Detta medför att när n>7 så kommer log(n)>2 vilket fungerar som ett straff mot modeller som har många predictors när man har många observationer.

ekvation 8: notera att log() är den naturliga logaritmen.

### Cross validation

För att testa en modell på data som modellen inte har blivit tränad på är ett bra sätt att evaluera sin modell. I vanliga fall kan man använda ca 80% av datan för att träna en modell och den resterande delen för att evaluera en modell.

Genom att använda sig a cross validation som är en metod där man delar upp datan i k antal delar (vanligtvis 5,10) och sedan tränar modellen på k-1 delar av datasetet och testar på den sista delen.

Detta görs om i iterationer till modellen har tränats på alla k kombinationer se Figur 6. Detta gör så att man kan få ett medelvärde på alla måtten beroende på vilka delar som modellen har tränats och testats på

Figur 6: Här är datasetet updelat i 5 fold (k), modellen tränas på 4 folds (blå) och tests på den sista (beiga) folden. Detta görs sedan om tills alla folds har vått vara test datan. Tillslut får man ett medelvärde av testfel

### RMSE

RMSE “root mean squared error” se ekvation 9, kommer ifrån MSE med syftet att dra roten ur MSE för att få en mer begriplig tolkning. MSE se ekvation 10 beräknar medelvärdet för alla kvadratiska prediktions fel. Problemt blir dock om man använder dessa mått på träningsdatan så kommer man får bra resultat men det betyder inte att man kan garantera en modell som predikterar bra på ny data, därför kan man använda RMSE tillsammans med Cross validation för att få en uppskattning av hur bra en modell predikterar på test data. Jämfört med exemplevis BIC som ger en skattning av test error så kan RMSE användas för att få ett begripligt mått då man får felet i samma enhet som Y

Ekvation 9: notera att jämfört med RSE så har man inte ett p värde då man inte jämför prediktioner beroende av predictors, man drar roten ur för att skillnaden i fel skall var i samma enhet som den beroende variabeln

Ekvation 10: Beräknar de kvadratiska felet hos prediktioner

# Metod

## Insamling av data

Insamling av data gjorde i grupp. I gruppen diskuterade man om vad för variabler man ville ha med i observationerna. Man kom fram till att förhålla sig till ett bilmärke samt en modell vilket var Volvo v60. Dessutom delade man upp observationerna från modeller vars bilar var skapade från år 2018-2023. Källan för data kom ifrån en annonsörhemsida[[1]](#endnote-1) som enkelt gav information om variablerna man valt att använda.

Man valde att samla information om bilens modellår, miltal, hästkrafter, bränsletyp, samt vilken typ av växellåda. Detta blev dem oberoende variablerna och man samlade även in priset som blev den beroende variabeln.

Insamlingen gjordes för hand in i ett Excel dokument där varje individ ansvarades för att samla minst 50 observationer av var sitt modellår, med andra ord att varje observation för modell år hade minst 50 observationer. Tillslut fick man in 352 observationer.

## Undersökning av data

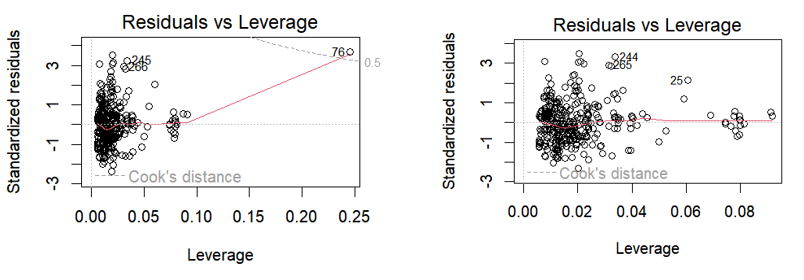
Man stegvis undersökte datan.

### Behövs variabler transformeras om?

För är det bra att om alla variabler är numeriska, man kan då snabbt konstatera att man har 2 oberoende variabler som inte är numeriska utan istället är kategoriska dessa är bränsletyp och växellåda. Nu kan man konstatera om dessa kategoriska variabler har en rangordning till sig, det skulle kunna gå att argumentera för det men man har valt att inte klassificera dem som en rangordning och då måste man transformera den kategoriska variabeln till ”dummyvariabler”.

Kort sagt så omvandlar man en kategori till en typ av check. Man har bränsletyp Bensin, Diesel och elhybrid det betyder att man kan ha en variabel som kollar om bilen är en bensin exempel x\_bensin = 1 och då en annan variabel som kollar om bilen är diesel exemple x\_diesel = 0, om bilen är hybrid så kommer båda variabel kontrollerna returnera 0, 0. Det betyder att man får kategorier-1 ”dummyvariabler”. Det betyder också att man får koefficienter för om en bil är bensin, diesel eller hybrid. Hybrid som inte har en variable kommer då att bli ”standard” och inte ändra på regressionslinjen.

### kolla om det finns några datapunkter som ”skadar” våran modellering

Först kan man kolla på om det ingår några outlier i datan. En outlier är en observation som av någon anledning står ut från dem resterande observationerna, ibland kan en outlier medföra problem hos modellen då den kan drastiskt ändra regressionlinjen då residualen från outliern kan vara väldigt drastisk. Kolla på figur 7 så kan man se att observation 76 sticker ut något enormt.

Figur 7: notera att observation 76 har väldigt hög leverage vilket betyder att den påverkar regressionlinjen något enormt

Genom att ta bort den observationen får man en modell som kommer prestera bättre. Eftersom man skall ha en anledning till att ta bort den observationen så ska man undersöka den.

Observation 76 är en 2016 bil som gått 48 582 mil dessutom är det en automatdiesel bil, priset för denna bilen är 129900 kr annonserat och problemet uppstår här när modellen har predikterat att bilen kostar 17282 kr vilket leder till en residual som är långt ifrån det ”riktiga” priset. Man kan tolka detta på olika sätt antigen så kanske modellen saknar variabler eller så har annonsören evaluerat sin bil till ett extremt pris som inte är rimligt, det kan vara möjligt att annonsören har skrivit in fel miltal på bilen men det är något man inte kan kontrollera. Man har istället valt att ta bort den observationen eftersom residualen är enorm och bidrar till en stor påverkan hos regressionlinjen.

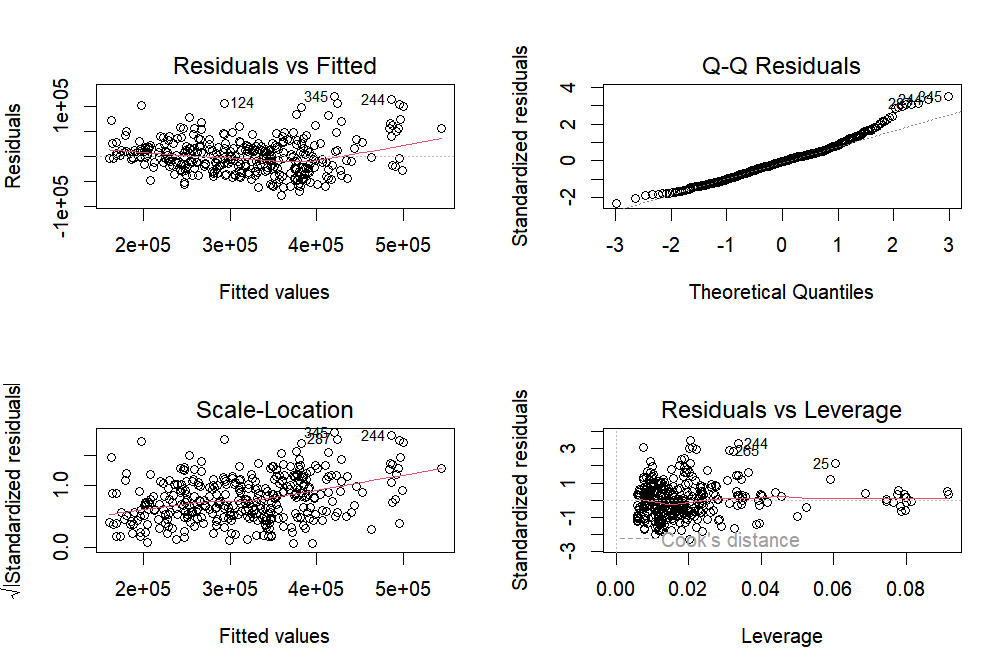
### Linjärt förhållande, heteroskedasticitet, normalfördelade residualer

Tillslut har man 3 saker till att kolla men det går att undersöka hyfsat snabbt. Först måste man kolla om det faktiskt är ett linjärt förhållande mellan den beroende variabeln och de oberoende variablerna, sedan nämndes det att residualerna måste vara normalfördelade när man gör statistik inferens. Och tillslut kollar man om residualerna är heteroskedasticska vilket betyder att variansen hos residualerna är icke-konstanta

Modellen antar ett linjärt förhållande mellan Y=b\_0+b\_1X… och om det förhållandet inte är linjärt så kan man inte lita på modellen i varken prediktioner eller inferens. Man kan undersöka detta genom att kolla på alla residualer.

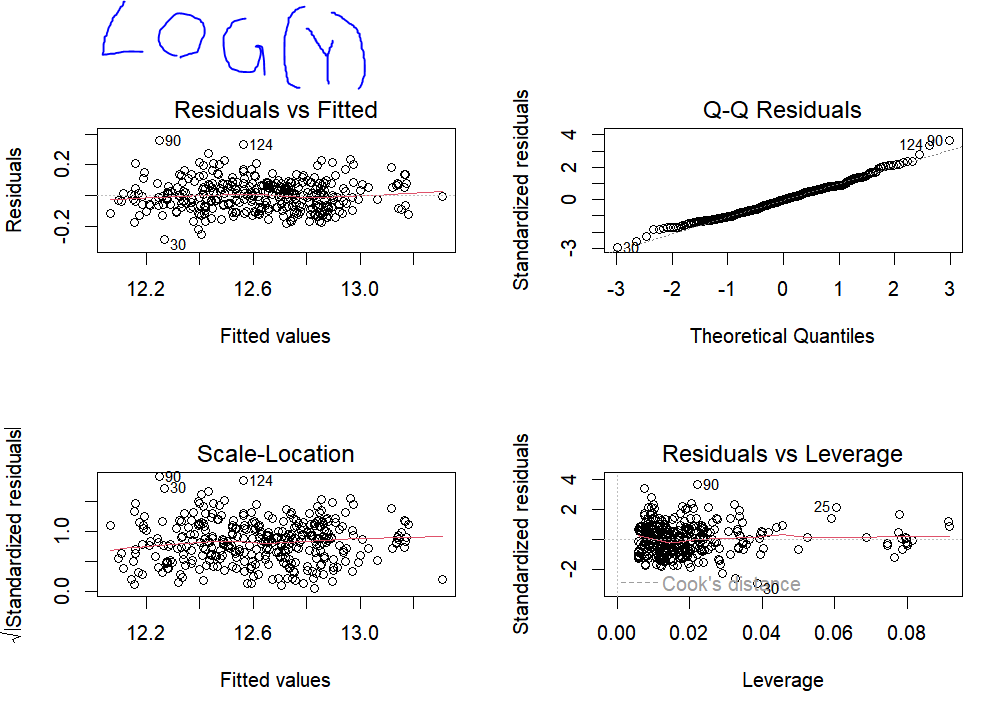
Man kan kolla om residualerna är normalfördelade genom att undersöka fördelningen visuellt hos residualerna. Man kan också använda sig av en så kallad Q-Q (quantile-quantile) plot vilket kort sagt jämför residualernas fördelning till en standardiserad normalfördelning och ser om värdena rör sig linjärt med varandra.

Heteroskedasticitet kan man kolla genom att visualisera och se om residualerna på något sätt ändrar sin variansen. Konsekvensen för att ha heteroskedastiska residualer kan leda till att konfidensintervallen, prediktions intervallen och samt hypotesteste blir felaktiga.

Se på Figur 8 så kan man visualisera alla dessa problem och bedöma om man bör göra någon typ av åtgärd. Om man kollar på våran Q-Q plot kan man se att det finns några extrema värden åt höger.

Figur 8: Summering om predikterade värden och residualer från en modell med alla variabler och borttagen observation 76.

Om man kollar på residualer mot fitted values grafen så kan man också argumentera att det inte är ett helt linjärt förhållande mellan Y = b\_0+b\_1X… . Det linjära förhållandet är inget extremt men man har valt att testa och göra en transformation för att möjligen lösa normalfördelningen och det linjära förhållandet.

Man har då testat att transformera den beroende variabeln till den naturliga logaritmen av Y. Det betyder också att om man skall testa ny data så måste man transformera dem observationerna till Ln samt sedan transformera tillbaka dem för att få tillbaka det riktiga priset. Se Figur 9 här kan man se en bättre normalfördelning samt ett mer linjärt samband. Trots att det ser bättre ut har man valt att ha 2 modeller en med LN transformation och en utan.

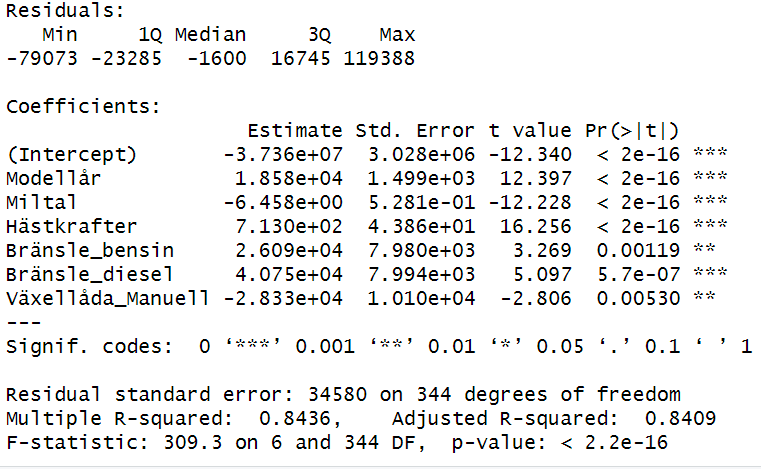
Figur 9: Regressionmodell där den beroende variabeln y är transformerard med den naturliga logaritmen LN(y). notera att log(y) är en engelska standarden för ln(). Alla variabler ingår

## Statistiks inferens av variabler

Nu när man har 2 modeller så kan man testa hur predictors eller oberoende variabler påverkar den beroende variabeln.

Eftersom man har en multipel linjär regressionsmodell så kan man göra ett f-test som kollar vilka variabler som är statistiskt signifikanta från att vara noll, alltså ett noll hypotestest. Om man kollar på den första modellen kallad lm\_1 se figur 10. I f-testet kan man se att alla variablerna är signifikanta och man kan förkasta noll-hypotesen för alla variabler. Man kan dessutom läsa betaparametern för alla variabler och se hur dem påverkar den beroende variabeln. Man kan nu säga att det är statistiskt signifikant att ett högre miltal på bilen kommer minska priset samt som att en ny årsmodell kommer att vara dyrare och mera hästkrafter också kommer öka priset. Notera att data antecknar vilket modellår bilen gjordes och inte hur gammal bilen är utifrån när den konstruerades.

Notera också att bränsletyp hybrid kommer att tolkas som standarden från regression linjen och om en bil exempelvis är dieseldriven så kommer priset ökas jämfört mot en hybrid bil, det kan tolkas som att hybrid bilar är billigare än bensin och dieselbilar. Om man gör samma undersökning för Lm\_2 som är samma modell fast med transformerad beroende variabel så kommer man får liknande resultat men istället motsvarar betaparametern till en naturliglogaritm koefficient.



Figur 10: F-test på oberoende variabler med modell LM\_1 (pris ~. data=bilar)

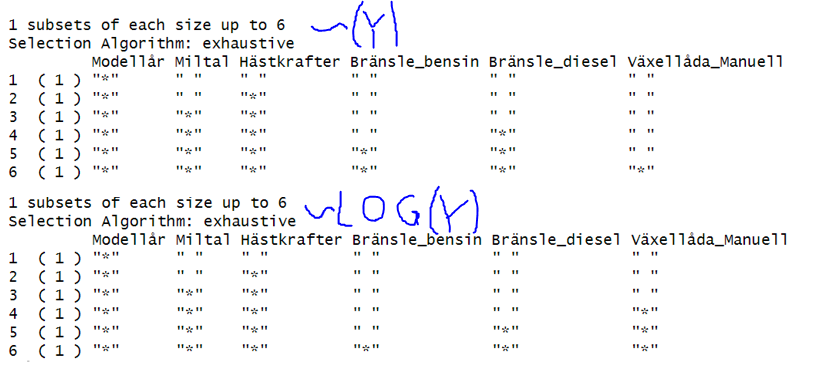
## Model evaluering

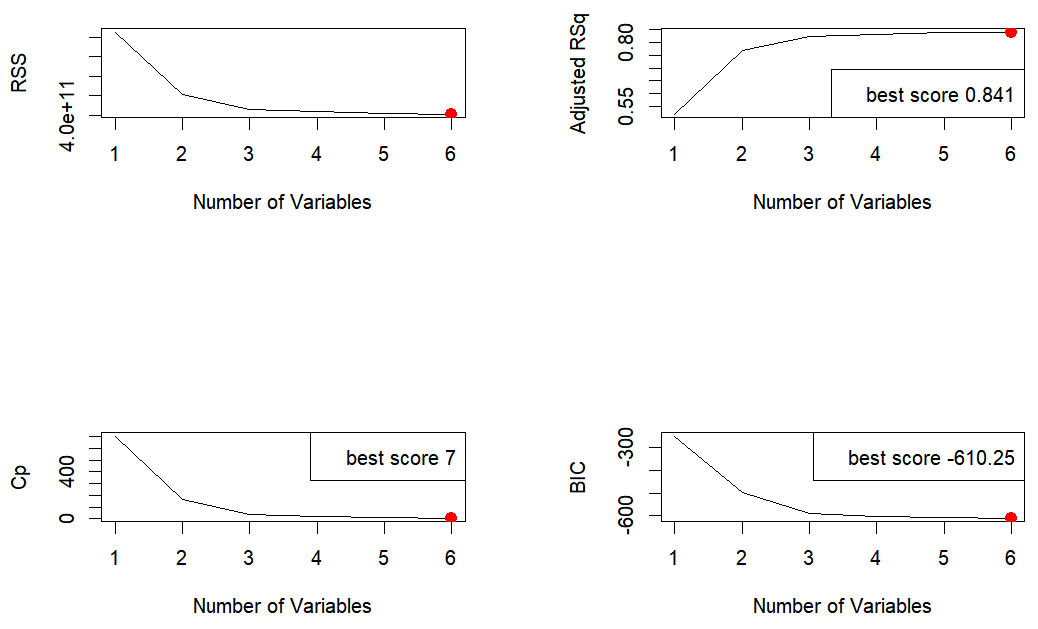
Man kan nu testa hur olika modeller presterar beroende på vilka predictorer man använder sig av.

Man har valt att testa modeller som använder sig av den normal beroende variablen y och modeller som använder sig a den transformerade beroende variabeln. Man har valt att använda sig av best subset selection metoden då den testar alla kombinationer av predictorer samt att man har max 6 beroende variabler vilket inte kräver så mycket datorkraft.

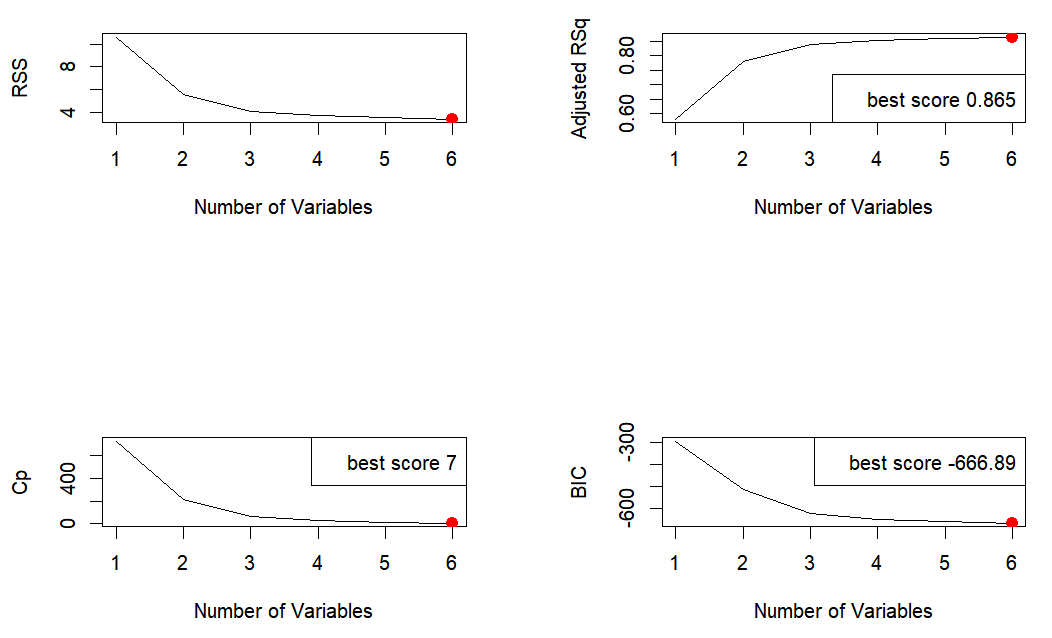
### Best subset selection

Genom att använda Best subset selection så kan man få fram vilka variabler som presterar bäst. Om man kollar på Figur 11 så kan man se skillnaden på hur best subset väljer predictors beroende på om man använder sig av en transformerad beroende variabel eller inte. Notera att fram till k=3 så använder båda modellerna samma predictors men efter k>3 så anser den log() transformerade modellen att växellåda ger bättre resultat än bränsletyp. Om man kollar på evaluerings mått som best subset ger en se Figur 12 och Figur 13 så kan man se att båda modellerna presterar bäst med alla predictors trots deras straff i mått samt att den transformerade modellen gav bäst resultat i alla måtten.



Figur 11: Skillnad på resultat av modeller givet k oberoende variabler i modellen.

Figur 12: Best subset resultat med ej transformerad beroende variabel



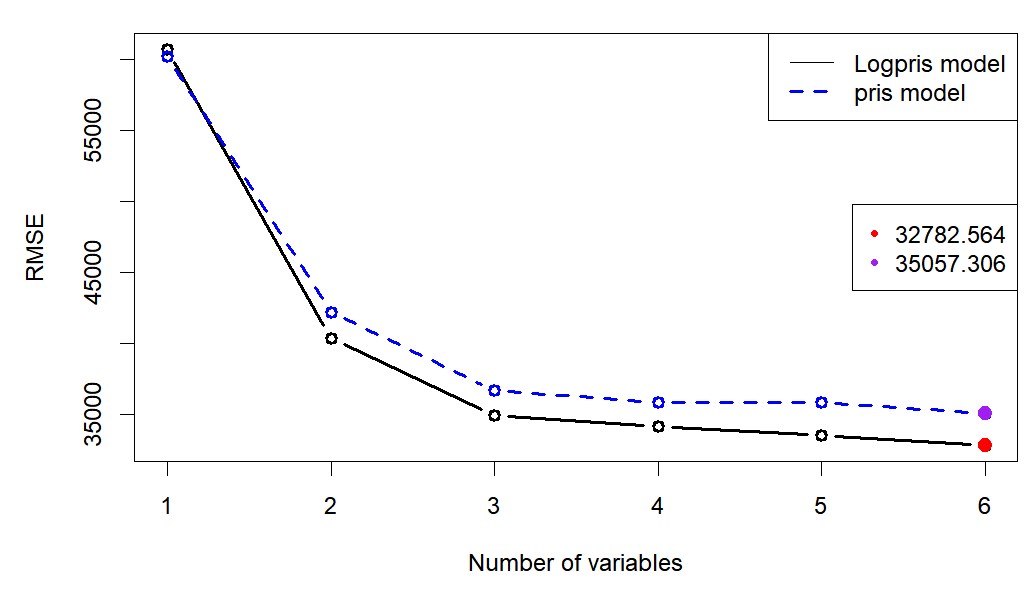
Figur 13: Best subset selection med log(y) transformeda beroende variabel

### Cross-validering

Med best subset kunde man konstatera att den bästa modellen använder sig av alla predictors och ger en olika mått för att evaluera modeller. Problemet är dock att dessa mått är en estimering av test felen på modellen och för att få något typ av läsbart resultat så måste man testa modellen på data som modellen inte har sett innan.

Med hjälp av Cross-validering så kan man även få ut test fel på modellerna och jämföra modellerna test fel mot varandra. Genom att kombinera Best subset selction med Cross validering så kan man få fram test fel på hos alla k modeller. Man har valt att använda sig av 10 folds vilket betyder att man kommer få ett test fel medelvärde ”RMSE” mellan 10 olika tränade modeller för varje kombination av .

Enligt graferna se Figur 14 så kan man se att den modellen vars beroende variabel är transformerad har presterat bäst överlag och därför har man valt att slutligen använda sig av den modellen som innehåller alla variabler samt har transformerat den beroende variabeln



Figur 14: RMSE resultat av best subset selection med hjälp av Cross validation. Den lägsta RMSE är logpris modellen med alla variabler som ger ett medelfel på 32782 kr per varje prediktion

## Testa bästa modellen på ny observation

Den valda modellen skall också kunna prediktera nya observationer så man har valt att gå in på samma källa som man hämtade datan från och plockat data från en slumpmässig observation.

Man tog 2 slumpmässiga observationer från annonsör 1[[2]](#endnote-2) och annonsör 2[[3]](#endnote-3). Bil 1 är en 2022 hybrid automat v60 med 340 hästkrafter som brukat 3600 mil denna bilen annonseras för 379 900kr, bil 2 är en 2020 diesel automat v60 med 150 hästkrafter som brukat 8801 mil och annonseras för 239 000 kr.

Om modellen predikterar dessa observationer så kommer modellen fram till att bil 1 skall kosta 420 502 kr och att bil 257 149 vilket är ca 10% och 7% ifrån det riktiga priset. Om man kollar på prediktions intervallet med 95% så får man ett intervall mellan 345 467kr och 511 835kr vilket täcker det sanna priset.

# Resultat och Diskussion

## Val av data och variabler.

Man kan ju konstatera att man har valt ett få antal oberoende variabler som skall ha ett samband mellan den beroende variabeln (pris) eftersom man nådde resultat som sa att alla 6 variabler ledde till en bättre modell så kan man argumentera för det kan behövas att samla in fler variabler i observationerna. Man valde att ta just 6 stycket variabler så man inte skulle behöva samla in för många observationer man måste också konstatera att det kanske inte finns tillräckligt med data för att täcka tillräckligt med observationer. Dessutom finns det andra problem med datan man kan ha använt sig av.

Om man kollar på kategoriska variabler så som biltillverkare så kan man sluta upp med nästan 50+ dummyvariabler och om du också tar in bilmodell så man då kommer va tvungen så kommer du få ytligare x antal dummyvariabler pär varje biltillverkare, det blir snabbt något extremt mycket. Teoretiskt sätt kan man säga att vissa biltillverkare medför en dyrare evaluering och att den då är rangordnad vilket gör variabel hanteringen enklare men hur skall du kunna rangordna 2 lyxbilar som Lamborghini och Ferrari?

Något kort att konstatera är också vilka variabler som är onödiga eller som man inte kan få tag på. Många kan tänka att färg på bilen inte spelar roll men för vissa individer så spelar det stor roll men normal färg som svart kanske inte påverkar bilen så mycket men en onormalfärg som gul kan påverka mycket, samt även exklusiva bilar som har speciella färger. Det kan vara väldigt svårt att evaluera en variabel som är väldigt personlig.

### Datakälla

Insamling av data från annonser är troligtvis det simplaste sättet att hitta data på men det kan också vara en nackdel. Konstatera att din granne inte vill sälja sin bil så han väljer att ge dig en evaluering av sin bil som han tror den är värd baserat på x antal variabler så som miltal, år och mer. Om man nu kollar på en annonsörs perspektiv så kan den individen att välja så evaluera sin bil till ett pris vars individen är bekväm med att sälja den. Annonsören kan välja att sälja bilen av många anledningar så som för vinst eller för att rädda en ekonomisk kris vilket kan leda till att annonsören inte evaluerar priset till vad bilen är värd utan vad annonsören vill ha för bilen.

## Transformering av data

Man valde att transformera den beroende variabeln Y (pris) eftersom datan gav en problem så som ej normalfördelad, eller ej linjärt förhållande. Man kan argumentera att residualerna inte och det linjära förhållandet inte va så extrema att man behövde transformera datan och därför valde man att behålla den gamla Y variabeln och jämföra med en transformerad Y variabel.

Något annat att konstatera är att man kunde transformera datan på något annat sätt. Det finns den värld där man istället skulle kunna sätta in en polynomfunktions variabel i modellen som kanske hade kunnat ge samma eller bättre resultat men eftersom att log transformering fungerade och gav bra resultat valde man att fortsätta med den.

## Oberoende variabeln och modellens prediktioner

Man nämnde förut att man kan evaluera ett bilpris av många anledningar som inte direkt är knyttet till alla de oberoende variablerna, Det finns anledningar till att pris sätta ett objekt högt eller lågt av olika anledningar och i många fall så kan det vara individen som bara inte vet exakt hur mycket bil egentligen är värd, detta blir då en mänsklig faktor som man inte kan använda som oberoende variabel. Man kan dock också argumentera att eftersom man har så pass många observationer så kommer man statistiskt hitta ett medelvärde för alla prissättningar och att modellen då kan tolkas på ett sätt som rättfärdigt evaluerar en bil. Detta kan då bli en indikator till om en bil går för en premie eller rabatt. Om man exempelvis använder sig av bil 1 och 2 från del 3.5 så kan man argumentera för att båda bilar går för rabatterat pris mot vad dem egentligen är värda. Det är dock ett ambitiöst antagande och om man skulle ta ett sånt antagande med denna modellen så finns det risk för att faktiskt göra förlust på ett köp om modellen har fel.

# Slutsatser

Man har bygget en modell som kan prediktera Volvo v60 priser. Man har också statistiskt undersökt hur dem valda variablerna påverka den beroende variabeln Pris och man kan då säga att det är statistiskt signifikant att alla variablerna påverka priset. Det finns potential för att utöka modellen och undersöka om andra variabler påverkar priset samt om de är signifikanta. Denna undersökningen kan bidra till att man gör mer ekonomiska beslut när man väljer att köpa eller sälja en Volvo v60 det är dock viktigt att konstatera att modellen har sina begränsningar och att priser också beror på den ekonomiska makrosituationen. I dessa tider men höga räntor så kan det finnas svårigheter för att sälja sin begagnade bil vilket kan drastiskt ändra på priset för begagnade bilar och då kan modellen ge dåliga resultat efter som att observationerna kommer ifrån detta året och inte från en längre period.

Om man dock konstaterar våran frågeställning från 1.1

1. Vilka variabler skall ingå i analysen.

Man valde att begränsa sig till 6 variabler vilket alla gav en signifikant påverkan hos den beroende variabeln Y (pris).

1. Hur signifikanta är variablerna

Alla variablerna va signifikanta med minst 2 stjärnor vilket betyder att det är mindre än 1% chans att förkasta en sann noll-hypotes hos variablerna, man kan till och med konstatera att den variabeln med högst p-värde låg på 0.00119 vilket betyder att de är 0.119% chans för en sann noll-hypotes

1. Hur hanterar man datan

Efter några moment valde man att testa att transformera den beroende variabeln y med den naturliga logarimten detta därför att det fanns små indikationer till att residualerna inte blev normalfördelade och att förhållandet mellan Y och X\_n inte var exakt linjärt. Det fanns dock möjlighet till att argumentera för att normalfördelningen och förhållandet var tillräckliga för att fungera och då valde man att ha 2 modeller en med transformation och en utan för att jämföra.

Dessutom fanns det en outlier i datasetet som man valde att ta bort för att förbättra modellen. Den outliern kan konstateras vara så dåligt evaluerad av annonsören att det förstörde modellen, alltså att bilen var något otroligt överprisat

1. Vilken/vilka modeller hanterar våran analys bäst och varför

Modellen med transformerad y variabel samt som använde sig av alla x variabler gav bästa resultat efter fler tal tester.

1. Hur väljer man en slutlig modell.

Den slutliga modellen blev vald eftersom den gav bäst resultat hos alla mått. Det slutliga testet RMSE med Cross validation gav även ett resultat med ett fel test på 32 782kr vilket var acceptabelt.

1. Kan modellen prediktera nya priser

Modellen klarar att prediktera nya observationer med ca 10-7% fel på dem testade observationerna. Med RMSE kan man även se att genomsnitt så kommer modellen prediktera fel med ca 32 782kr. Man kan dock argumentera att dem testade observationerna kan ha varit bilar som också är under evaluerad av annonsören.

# Källförteckning

1. <https://blocket.se/> [↑](#endnote-ref-1)
2. <https://www.blocket.se/annons/1001307003> [↑](#endnote-ref-2)
3. <https://www.blocket.se/annons/1001216362> [↑](#endnote-ref-3)