Regression Analysis

Bilpriser på Blocket



Siarhei Thor Fedatsenka

EC Utbildning

R - programming

202404

# Abstract

Denna studie utforskar sambandet mellan medelinkomst per län och prissättning på Sveriges andrahands bilmarknad genom regressionsanalys. Genom att använda data insamlad från Blocket.se och extern information om medelinkomster från SCB, syftar forskningen till att klargöra om inkomstnivåer driver upp priserna på andrahands bilar. Med hjälp av en OLS linjär modell visar resultaten att den slutgiltiga modellen har god prediktionsförmåga (justerad R^2 på 0.85), men att medelinkomsten per län inte har någon signifikant effekt på bilpriserna. Detta antyder att andra faktorer än inkomstnivåer spelar en mer avgörande roll för prisbildningen på andrahands bilmarknaden. Dem viktigaste variabler som påverkar priset är fortfarande åldern på modellen, miltal och hästkrafterna.

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

Innehållsförteckning

Abstract 2

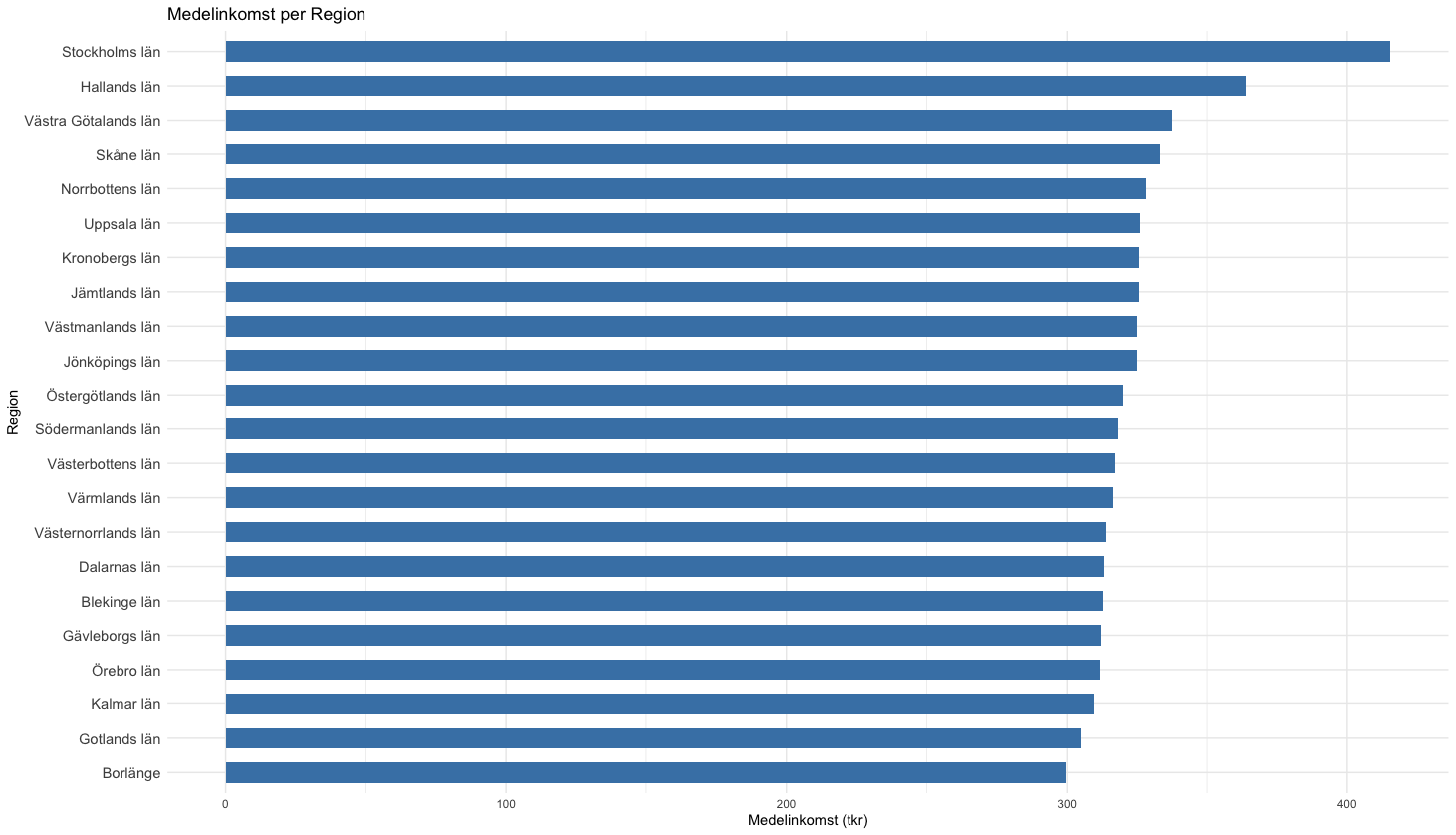
1. Inledning 1
2. Teori 2
3. Metod 4
4. Resultat och Diskussion 7
5. Slutsatser 9
6. Teoretiska frågor 10
7. Självutvärdering 12

Appendix A 13

Källförteckning 14

# Inledning

Rapporten konserterar sig kring regression analys av andrahandsbilmarknad i Sverige. Syftet är att en komplett regressionsmodellering på den insamlade datan. Datan samlas manuellt via blocket.se och genom dataskrappning med chrome extensions. Ett extern data från SCB hämtas via pxweb API.

Som en utökning av ett regressionanalys kommer vi att testa om medelinkomst per län påverkar prisdynamiken på andrahandsmarknad. Allmänt sätt högre inkomstnivåer är en av faktorerna som kan driva priserna uppåt, vi kommer att test om detta gäller för andrahandsbilar.

# Teori

Detta sektion beskriver viktigaste koncept kring regressionsanalys och pga ursprungliga källor för teori har varit på engelska under kursen, teori sektionen är på engelska.

1. Regression Analysis

Regression analysis is a statistical method used to examine the relationship between a dependent variable and one or more independent variables. The goal is to model the expected value of the dependent variable as a function of the independents, to predict or infer causation. In the context of car prices, regression can help identify which factors (e.g., make, model, age, mileage) most significantly impact the vehicle's price.

2. Ordinary Least Squares (OLS)

Ordinary Least Squares (OLS) is the most common method of estimating the parameters in a linear regression model. OLS minimizes the sum of the squared differences (residuals) between observed values and the values predicted by the linear model. This method provides the best linear unbiased estimates of the coefficients, assuming the residuals are normally distributed and homoscedastic.

3. Coefficient of Determination (R² and Adjusted R²)

The coefficient of determination, denoted as R², measures the proportion of variance in the dependent variable that is predictable from the independent variables. It is a key output of regression analysis that ranges from 0 to 1, where higher values indicate a better fit of the model to the data. Adjusted R² is a modified version of R² that adjusts for the number of predictors in the model, providing a more accurate measure in the context of multiple regression.

4. Bayesian Information Criterion (BIC)

The Bayesian Information Criterion (BIC) is used for model selection among a finite set of models; the model with the lowest BIC is generally preferred. It is particularly useful in regression analysis as it balances the goodness of fit of the model against its complexity by penalizing the number of parameters. This can help prevent overfitting when choosing among multiple models with different numbers of predictors.

5. Assumptions of OLS

To ensure the validity of the OLS estimates, several assumptions must typically be met:

a. Linearity: The relationship between the dependent variable and each independent variable must be linear.

b. Independence: Observations must be independent, without any correlation.

c. Homoscedasticity: The variance of error terms should remain constant across all levels of the independent variables.

d. Normality of Residuals: The residuals of the model should be normally distributed to ensure the validity of statistical tests.

e. No Perfect Multicollinearity: No independent variable should be a perfect linear function of another.

Addressing Violations: If these assumptions are violated, adjustments may include transforming variables, using techniques like Generalized Least Squares (GLS) or robust standard errors, or applying regularization methods such as Ridge or Lasso.

6. Significance Testing and Confidence Intervals

Regression analysis also involves hypothesis testing and confidence intervals to determine the statistical significance of each coefficient, helping to infer whether a particular variable has a meaningful impact on the dependent variable. A low p-value (< 0.05) typically indicates that the coefficient is significantly different from zero.

# Metod

**Dataset 1.**

Data består av två dataset. Dataset 1 samlades manuellt via blocket websida insamlingen sked med 8 personer fördelat på olika län och filtrerat efter **elbilar** sålda av **företag.** Följande attributs har samlats in:

"Miltal" "Modellår" "Biltyp" "Drivning" "Hästkrafter" "Färg" "Datum.i.trafik" "Märke" "Modell" "Pris" "Län"

Total antal datapunkter: 214. Detta dataset har inte används för rapporten och är i del av gruppsarbete.

**Dataset 2.**

Dataset 2 samlades via webscraping och innehåller kring 10000 datapunkter. Data har blivit rensat och preprocessed för att få samma struktur. Efter rensningen har dataset följande attribut:

"Company" "Location" "Year" "Fuel" "Miles" "Gear" "Price" "Name" "Brand" "Model" "Engine.Volume" “Horsepower”

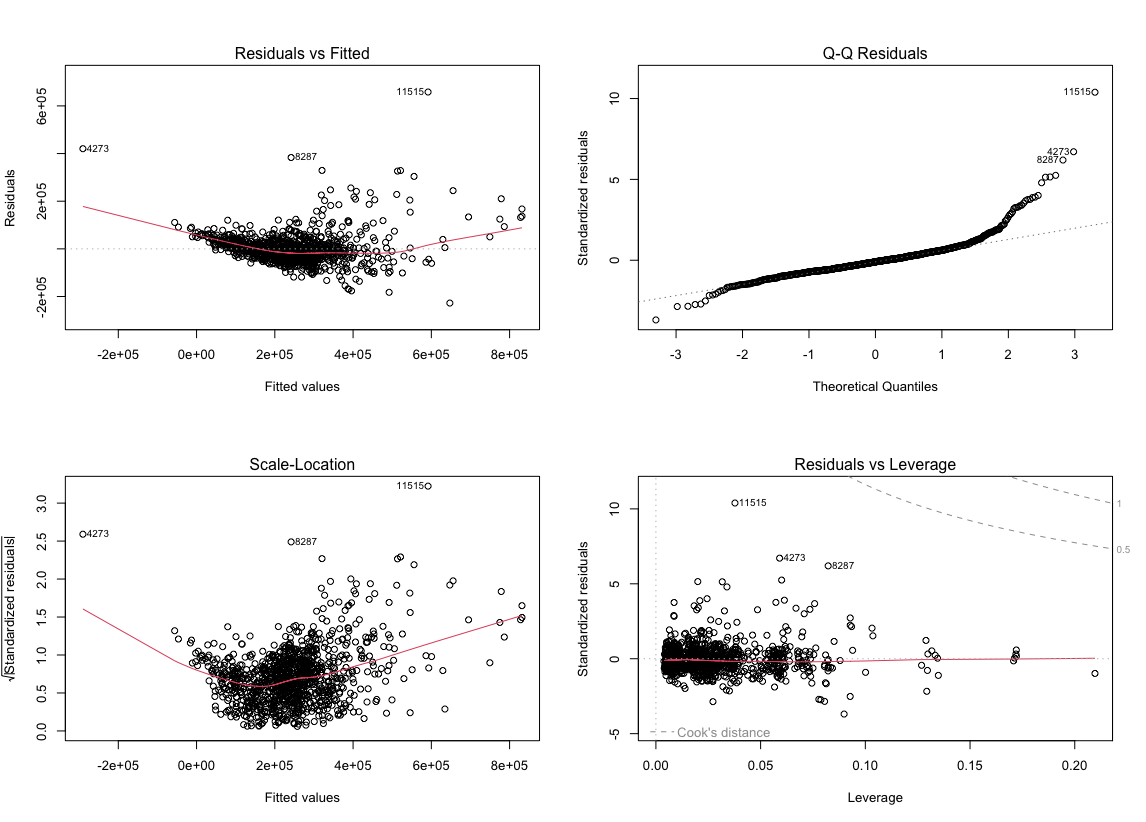
Total antal datapunkter: 1047. För detta studie har bara dataset 2 och extern data används.

**Extern data: SCB medelinkomst per Län.**

Datan kring medelinkomst har samlats via PXWEB API för 2022.

**Basmodel.**

För regressionsmodelering OLS linear modell har används. Kategoriska variabler som “Brand", "Model" , ”Engine.Volume" , “Name”, "Company" har inte inkluderats pga kolinieritet eller låg varians inom kategorier som ledde till overfitting. Basmodell har tränats i R:

model\_l = lm(Price ~.-Brand-Model-Engine.Volume-Name-Company, data = data\_l\_clean)

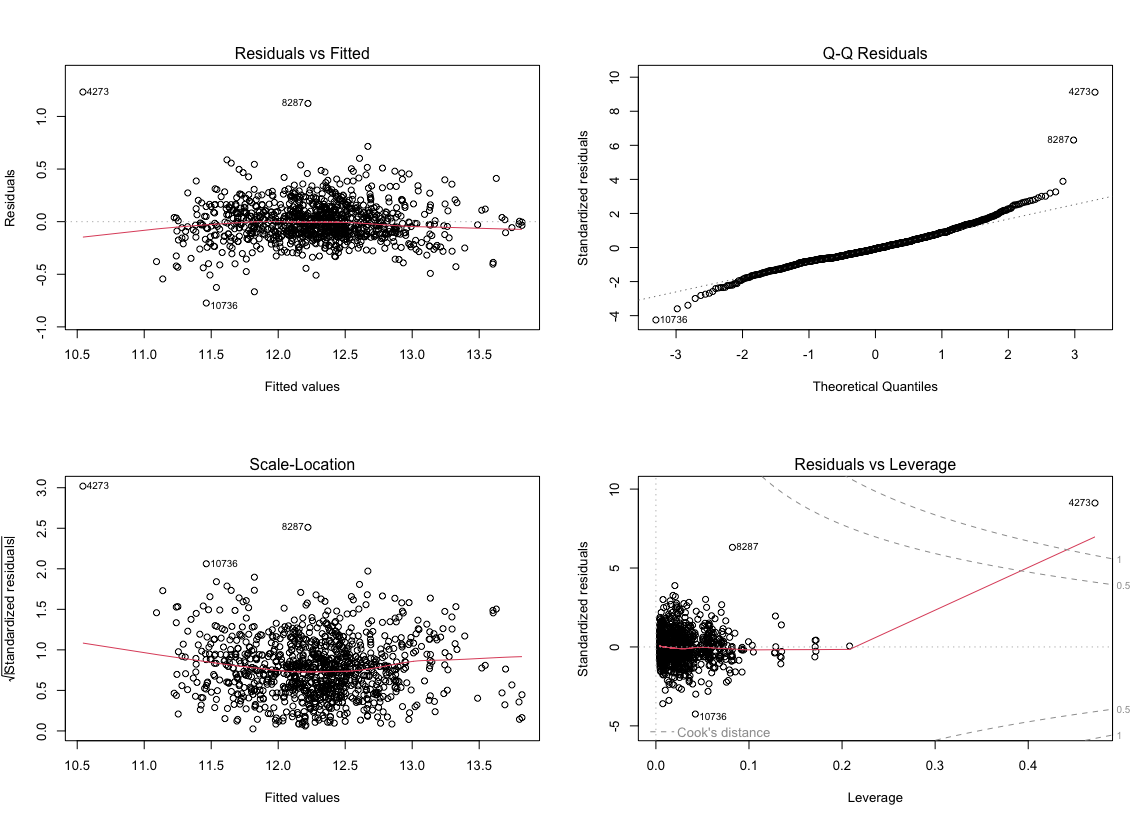
**Basmodel utvärdering.**

Baserad på diagnostiska plots bas modellen har flera problem:

| Problem | Åtgärd |
| --- | --- |
| Residuals vs Fitted tyder på en vis icke linearitet | Interactions variabler: +Horsepower:Year+Miles:Year |
| Scale-Location tyder på relativt starkt Heteroskadasitet | Log(y) |
| Q-Q visar icke normal fördelat residual | Efter lösning på första två behövs inte åtgärdas |
| Outliers 11515, 42732 | Inget problem med data, lämnat |

**Slutligt model.**

model\_l\_log = lm(I(log(Price)) ~.-X-Brand-Model-Engine.Volume-Name-Company-Year+Miles:Age, data = data\_l\_clean)

Diagnostiska plots visar inga stora avvikelser och modellen kan användas för statistisk inferens. ****

**Medelinkomst värden.**

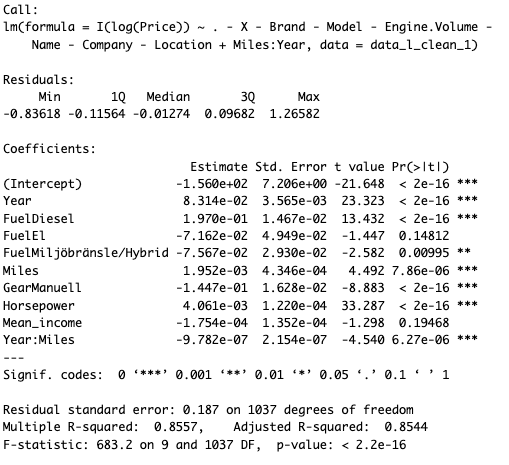
Information från extern källa har blivit integrerat som kontinuerlig variabel istället för kategorisk, så:

Stockholm = 415.4,

Dalarna = 313.3,

Skåne = 333.2 …

# Resultat och Diskussion

Slutlig modell har adj R2 på 0.85 som tyder på stark prediktionsförmåga.

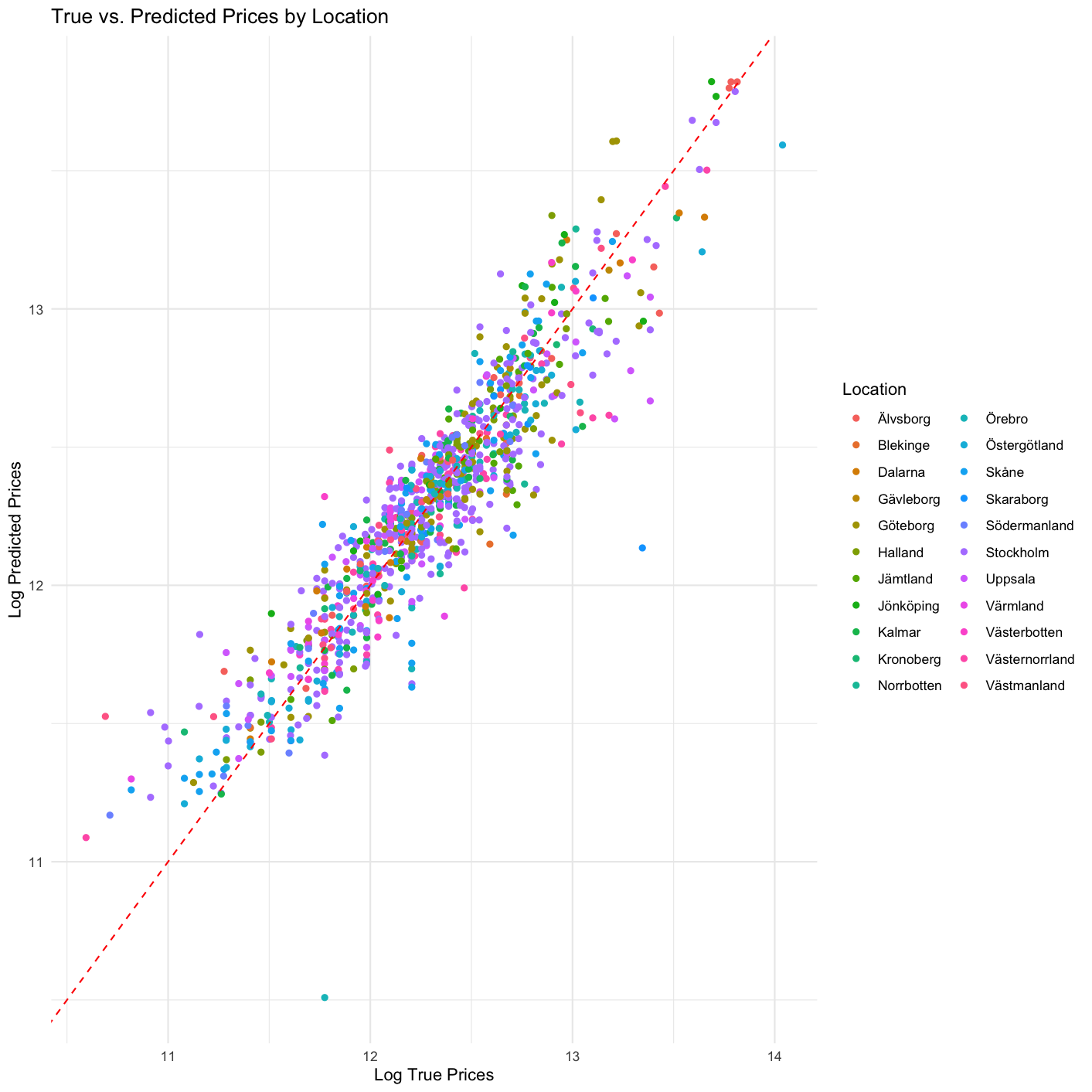
Tabell 1: Model summary.

Pga av log transformation av y, det går inte att tolka koefficienterna direkt utan modellen predikterar log av priset. I princip koefficienterna kan tolkas som ett procentuellt förändring av beroende variabel med ökning av oberoende variabel med 1 enhet.

Figur 2: Logtransformation.

Viktig att nämna interaktionsvariabel Year:Miles som gör att Years och Miles kan inte vara förklarad var för sig men ska tolkas i kombination med varandra. På detta sätet Year:Miles interaktionen tyder på att det finns negativt samband av extra enheten på priset 0.00009%, samtidigt miles har positiv påverkan och years negativt detta kan tolkas att miles har inte lika stark inflytande som variabel years.

FuelEl verkar inte har statistiskt signifikans pga för få datapunkter. Det vore intressant att forska samband mellan typ av bränsle med utökat dataset.

Medel inkomst av olika län har inget statistisk signifikans i modellen. Detta kan tyda på att det inte finns någon samband på hur mycket folk känner och priser på andrahands bilar. En av förklaringar kan vara market uniformitet eller informationstillgänglighet till exempel, men även här kan det vara att det inte finns tillräckligt mycket träning datapunkter för att dra några slutsatser kring det.

Figur 3: True vs Predicted, ln scale, inga synliga klustringar på att län påverkar priset.

# Slutsatser

Baserat på regressionsanalys av datasetet verkar medelinkomster per län inte ha någon signifikant påverkan på bilpriser. Åldern, antal körda kilometer, antal hästkrafter samt typ av bränsle har däremot stark prediktiv förmåga på bilpriset. Logtransformation av priset, samt introduktion av interaktionseffekten förbättrat R2 score och löst problem med heteroskadasitet och icke normalitet av residualer.

# Teoretiska frågor

Kolla på följande video: https://www.youtube.com/watch?v=X9\_ISJ0YpGw&t=290s , beskriv

kortfattat vad en Quantile-Quantile (QQ) plot är.

*QQ plot är ett 2 dimensional graph med y axis som standartiserade data och x axis är ett respektive värde för sample omfördelat i quintiles enligt normalfördellning.*

2. Din kollega Karin frågar dig följande: ”Jag har hört att i Maskininlärning så är fokus på

prediktioner medan man i statistisk regressionsanalys kan göra såväl prediktioner som

statistisk inferens. Vad menas med det, kan du ge några exempel?” Vad svarar du Karin?

*Karin har i princip rätt, ett enkel exempel är om vi vill prediktera huspriser. Med regressionanalys kan vi både skapa produktioner med linear model och förklara relationer mellan oberoende variabler, även effekt av enstaka oberoende variabler på beroende variabel med en viss konfidensnivå. ML-modell fokuserar främst på ett skapa så noggranna prediktioner som möjligt, där relation mellan variablerna är inte i fokus.*

3. Vad är skillnaden på ”konfidensintervall” och ”prediktionsintervall” för predikterade värden?

*Konfidensintervall ger en uppskattning av var vi förväntar oss att det genomsnittliga värdet av den beroende variabeln ligger för en given uppsättning av oberoende variabler. Detta intervall speglar osäkerheten kring uppskattningen av medelvärdet.*

*Prediktionsintervall ger en uppskattning av var en enskild framtida observation av den beroende variabeln förväntas falla, givet en specifik uppsättning av oberoende variabler. Detta intervall tar hänsyn till både osäkerheten i uppskattningen av medelvärdet och den naturliga variationen (eller spridningen) av individuella observationer.*

4. Den multipla linjära regressionsmodellen kan skrivas som:

𝑌 = 𝛽0 + 𝛽1𝑥1 + 𝛽1𝑥2+ . . . + 𝛽𝑝𝑥𝑝 + 𝜀 .

Hur tolkas beta parametrarna?

𝛽*0 - är ett värde för intercept, om alla värden för X är 0, i många fall b0 kan talkas som ett utgångsvärde värde för regression.* 𝛽𝑝 *är koefficient som tyder på förändring av Y med 1 enhet av Xp, om alla förutsättningar för regression är uppfyllda detta koefficienter kan direkt tolkas som ett effekt av förändring av Xp på Y.*

5. Din kollega Hassan frågar dig följande: ”Stämmer det att man i statistisk

regressionsmodellering inte behöver använda träning, validering och test set om man nyttjar

mått såsom BIC? Vad är logiken bakom detta?” Vad svarar du Hassan?

*BIC är ett kriterium för modellval som tar hänsyn till modellens komplexitet. Det straffar komplexa modeller för undvika overfitting. Om huvudsyftet är ett statistiskt inferens så räcker det med att kolla på BIC, AIC, adj. R2 i modell sammanfattning. Om datan räcker det skadar aldrig att använda test set, framförallt om modellen ska användes för prediktioner.*

6. Förklara algoritmen nedan för ”Best subset selection”

*Det här är ett brute force algoritm för att hitta bästa kombination av oberoende variabler för den bästa prestanda för ett utvarderingsmått som R2, BIC/AIC mm. Metoden är opraktisk med en stor antal variabler.*

# Självutvärdering

1. Utmaningar du haft under arbetet samt hur du hanterat dem.   
   Scope av arbete och hantering av normalitet av residualer. Tolkningen av interaktionseffekter är inte altid lätt att resonera kring.
2. Vilket betyg du anser att du skall ha och varför.   
   VG, kriterier för VG är uppfyllda. Jag har inte kört någon ridge eller lasso, då jag hade inget behov för det.
3. Något du vill lyfta fram till Antonio?

Kanske köra lite fler olika regressionanalyser för att få bättre känsla för statistisk inferens.

# Appendix A

# Källförteckning

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning: With Applications in R. Springer Science & Business Media.