Projekt i programmeringsspråket R

Regressionsanalys för att förutsäga bilpriser

Alia Atawna

EC Utbildning

Kunskapskontroll - R

2024-04

Abstract

This study utilizes multiple regression analysis to predict used car prices using data from Blocket.se, focusing on factors like mileage, model year, and brand. The model explains about 45.72% of the variance in car prices, suggesting effectiveness in assessing market trends but highlighting the need for additional variables for greater accuracy. Limitations include the dataset's regional specificity and potential non-generalizability. Future work could enhance precision by incorporating broader economic factors and advanced modeling techniques.

Innehållsförteckning

1 Inledning

Begagnade bilar utgör en betydande del av bilmarknaden i Sverige, och prissättningen av dessa fordon kan variera kraftigt beroende på en mängd faktorer som märke, modell, ålder, miltal och skick.

Marknaden för begagnade bilar är också dynamisk och påverkas av ekonomiska svängningar, teknologiska framsteg och förändringar i konsumenternas preferenser. Till exempel kan införandet av nya miljölagar påverka efterfrågan på vissa typer av bilar. Det är därför relevant att regelbundet analysera denna marknad för att förstå aktuella trender. Denna rapport bygger på en dataset sammanställd från annonser på www.blocket.se, vilket inkluderar data om begagnade bilar från olika län i Sverige. Datauppsättningen innehåller variabler som pris, miltal, modellår, bränsletyp, växellåda, och andra bilattribut, vilka alla är kritiska för att förstå och förutsäga bilpriser.

Syftet med denna rapport är är att utveckla en förståelse för de faktorer som påverkar prissättningen av begagnade bilar i Sverige och att bygga en prediktiv modell genom multipel regressionsanalys som kan förutsäga priset på en begagnad bil baserat på dess attribut. För att uppfylla syftet så kommer följande frågeställning att besvaras:

1. Hur väl kan en multipel regressionsmodell förutsäga priset på en begagnad bil baserat på dess attribut?

	C	D	E	F	G	H	I
1	CarName_ModelYear	CarName_Engine	CarName_Miles	CarName_gears	CarName_Price	CarName_Region	CarName_Dealer
2	2015	Diesel	14926	Automat	189800	Stockholm	Riddermark Bil, Veddesta - Järfälla
3	2017	Bensin	10976	Automat	168900	Stockholm	Riddermark Bil, Veddesta - Järfälla
4	2017	Diesel	7857	Automat	539900	Stockholm	Riddermark Bil, Veddesta - Järfälla
5	2018	Bensin	4225	Automat	349700	Stockholm	Riddermark Bil, Veddesta - Järfälla
6	2016	Diesel	17391	Automat	249900	Västmanland	Riddermark bil Västerås
7	2017	Bensin	7493	Automat	339900	Västerbotten	Niemi Bil Skellefteå
8	2017	Diesel	12273	Automat	289900	Västerbotten	Niemi Bil Umeå
9	2018	Bensin	9649	Automat	249900	Västerbotten	Niemi Bil Umeå
10	2017	Diesel	14620	Automat	359900	Norrbotten	Niemi Bil AB - Spantgatan
11	2019	Bensin	6567	Automat	209900	Göteborg	Moberg Bil AB - Göteborg
12	2019	Diesel	16075	Automat	214900	Göteborg	Moberg Bil AB - Göteborg
13	2018	Diesel	7666	Manuell	219900	Göteborg	Moberg Bil AB - Göteborg
14	2017	Bensin	8185	Manuell	84900	Stockholm	Svenska Bilgruppen i Haninge AB
15	2016	Diesel	11900	Automat	199900	Stockholm	Svenska Bilgruppen i Haninge AB

Figur 1. Data insamlad från Blockets hemsida

2 Teori

2.1 Linjär regression

Linjär regression är en grundläggande teknik inom maskininlärning och statistik för att modellera samband mellan en beroende variabel och en eller flera oberoende variabler. Målet med linjär regression är att upptäcka och kvantifiera det linjära sambandet mellan den beroende variabeln Y och de oberoende variablerna $X_1, \dots, X_p$, vilket beskrivs av modellen:

$$Y \approx \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

Här representerar β_n koefficienterna som visar effekten av varje oberoende variabel X_p på Y och ϵ är en felterm som fångar avvikelser från den perfekta linjära modellen. (James, Witten, Hastie & Tibshirani, 2014)

2.2 Multipel regressionsanalys

Multipel regressionsanalys, bygger vidare på detta koncept genom att inkludera två eller flera oberoende variabler för att modellera förhållandet till en beroende variabel. Målet här är att noggrant modellera hur den beroende variabeln kan förutsägas från de oberoende variablerna genom ekvationen:

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon$$

I detta fall representerar $\beta_0, \beta_1, \dots, \beta_n$ regressionskoefficienterna som indikerar hur starkt varje oberoende variabel bidrar till att förutsäga Y. ϵ representerar feltermen som fångar omodellerade effekter och slumpmässiga variationer. (James, Witten, Hastie & Tibshirani, 2014)

2.3 R^2, justerat R^2

R-kvadrat (R^2) är ett statistiskt mått som används för att bedöma hur väl en regressionsmodell passar till de observerade datan. Det representerar andelen varians i den beroende variabeln som förklaras av oberoende variabler i modellen, med värden som sträcker sig från 0 (ingen förklaring) till 1 (fullständig förklaring).

Det justerade R^2 tar hänsyn till antalet förklarande variabler i modellen, vilket hjälper till att undvika överanpassning genom att straffa onödig komplexitet, och gör det möjligt att jämföra modeller med olika antal parametrar rättvist. (James, Witten, Hastie & Tibshirani, 2014)

2.4 RMSE

Root Mean Square Error (RMSE) är ett mått på skillnaderna mellan värden som förutsägs av en modell och de faktiska observerade värdena. Den representerar kvadratroten av de genomsnittliga kvadrerade skillnaderna mellan förutsagda och observerade värden. RMSE beräknas enligt följande:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

Där n är antalet observationer och y_i är det faktiska observerade värdet, $\hat{y}_i$ är det förutsagda värdet av modellen. (James, Witten, Hastie & Tibshirani, 2014)

2.5 AIC

Akaike Information Criterion (AIC) är ett mått som används för att jämföra olika statistiska modeller baserat på deras kvalitet och komplexitet. AIC balanserar modellens anpassning till datan mot antalet parametrar den använder för att undvika överanpassning.

Formeln för AIC är AIC = 2*k* - 2ln(*L*), där är *k* antalet parametrar och *L* är modellens maximum likelihood. Ett lägre AIC värde indikerar en bättre modell. AIC är användbart för att välja den mest lämpliga modellen bland flera alternativ, där den prioriterar modeller som är både enkla och har god anpassning. (James, Witten, Hastie & Tibshirani, 2014)

2.6 BIC

Bayesian Information Criterion (BIC) är ett kriterium för modellval som används för att jämföra modeller baserat på deras passform och antal parametrar, med en strängare straff för komplexitet än Akaike Information Criterion (AIC).

BIC beräknas med formeln $\mathrm{BIC} = k\ln(n) - 2\ln(\widehat{L}).$ där *n* är antalet observationer, *k* antalet parametrar, och L maximum likelihood. BIC är särskilt användbart för stora datamängder eftersom det effektivt hjälper till att undvika överanpassning genom att prioritera enklare modeller med tillräcklig passform. (James, Witten, Hastie & Tibshirani, 2014)

3 Metod

3.1 Datainsamling

Det första steget var att samla in data om bilförsäljning från Blocket. Jag arbetade i grupp med Abdulrahman, Anton, Daniel, George, Goran, Jesper, John och Kawser. Inledningsvis mötte vi vissa utmaningar med att organisera arbetet, men efter några gruppmöten blev det bättre.

Varje gruppmedlem kunde bidra till arbetet på ett meningsfullt sätt. För att göra uppgiften mer hanterbar och pedagogisk bestämde vi att varje person skulle ansvara för att samla in data om ett mindre antal bilar av ett specifikt märke. Detta tillvägagångssätt möjliggjorde att vi kunde tillämpa en statistisk teori och skapa en "minimodell", som sedan kunde jämföras inom gruppen.

För att effektivisera insamlingen av data tog Daniel på sig uppgiften att utföra en så kallad "webscraping" från Blocket, varigenom han extraherade ungefär 7000 bilannonser. Denna data blev sedan basen för vår analys. Vi arbetade individuellt med att sammanställa och jämföra informationen baserat på den gemensamt överenskomna minimodellen. En av de mer intressanta observationerna från vår analys var hur olika bränsletyper, exempel med en Volvo V40, påverkade bilarnas prissättning.

Data vi samlade in och analyserade inkluderade flera kritiska variabler: bilens märke och modell (CarName_name, CarName_Model), årsmodell (CarName_ModelYear), motortyp (CarName_Engine), miltal (CarName_Miles), växellåda (CarName_Gears), pris (CarName_Price), region där bilen såldes (CarName_Region), och återförsäljare (CarName_Dealer). Denna detaljrika datamängd gav oss möjlighet att djupdyka i bilmarknadens prisdynamik och ge en grundlig analys av faktorer som påverkar prissättningen av begagnade bilar på Blocket.

Genom detta projekt förbättrade vi inte bara våra tekniska färdigheter i datamanipulation och statistisk analys, utan vi utvecklade även vår förmåga att arbeta effektivt som en del av ett team. Reflektioner över projektets gång och de strategier vi tillämpade ger värdefulla insikter som kommer att vara till nytta i framtida samarbetsprojekt.

3.2 Dataförberedelse

Data laddades in i R för bearbetning. Initial datarensning utfördes som innefattade att identifiera och hantera saknade värden. Detta följdes av att kategorisera viktiga variabler och omvandla dem till faktorer för att underlätta analyser. De variabler som fokuserade på inkluderade:

- Märke och modell (CarName_name, CarName_Model)
- Årsmodell (CarName_ModelYear)
- Motortyp (CarName_Engine)
- Miltal (CarName_Miles)
- Växellåda (CarName_Gears)
- Pris (CarName_Price)
- Region (CarName_Region)
- Återförsäljare (CarName_Dealer)

```
> summary(cars)
 CarName_Brand      CarName_Model      CarName_ModelYear CarName_Engine     CarName_Miles    CarName_gears
 Length:7106        Length:7106        Min.   :2014      Length:7106        Min.   :     1   Length:7106
 Class :character   Class :character   1st Qu.:2016      Class :character   1st Qu.:  6300   Class :character
 Mode  :character   Mode  :character   Median :2018      Mode  :character   Median : 10054   Mode  :character
                                       Mean   :2018                         Mean   : 10698
                                       3rd Qu.:2020                         3rd Qu.: 14000
                                       Max.   :2024                         Max.   :263246

 CarName_Price      CarName_Region     CarName_Dealer
 Min.   :  34200    Length:7106        Length:7106
 1st Qu.: 174800    Class :character   Class :character
 Median : 229000    Mode  :character   Mode  :character
 Mean   : 250538
 3rd Qu.: 295000
 Max.   :1759000
>
```

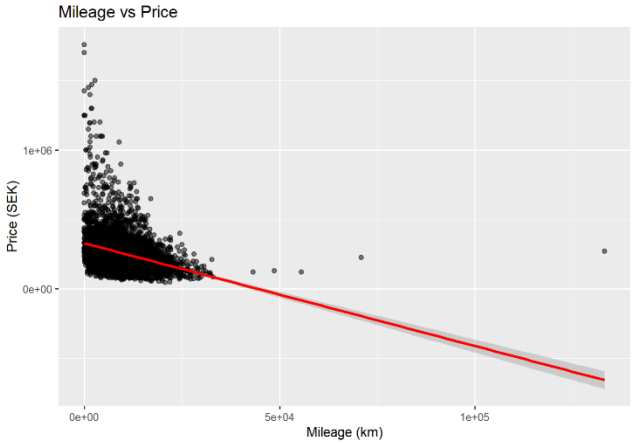
Figur 2. Importering av data i R

3.3 (EDA)

En utförlig explorativ dataanalys (EDA) genomfördes för att utvärdera datakvaliteten och förstå distributionen av nyckelvariabler. Olika visualiseringstekniker användes för att identifiera mönster och avvikelser i datan.

```
   CarName_Brand     CarName_Model CarName_ModelYear    CarName_Engine     CarName_Miles     CarName_gears
               0                 0                 0                 0                 0                 0
   CarName_Price    CarName_Region    CarName_Dealer
               0                 0                 0
>
```

Figur 3. Inga saknade värden kvar i datan



Figur 4. Visualisering av miltal vs pris

3.4 Regressionsanalys

Inför regressionsanalysen delades datan upp i tränings- och testset. Detta steg gjordes för att kunna validera modellens förmåga att generalisera till ny data. Multipel regressionsanalys utfördes där olika modeller konstruerades för att bedöma vilka variabler som var signifikanta prediktorer för bilpriset. Regressionsmodeller utvärderades baserat på deras R^2, justerade R^2, AIC och BIC-värden för att bedöma deras passform och effektivitet.

3.5 Modellträning och prisförutsägelse

De utvecklade modellerna tränades och användes sedan för att förutsäga bilpriser. De jämfördes med förutsagda priserna och med det faktiska priser från datamängden för att utvärdera modellernas prediktiva noggrannhet.

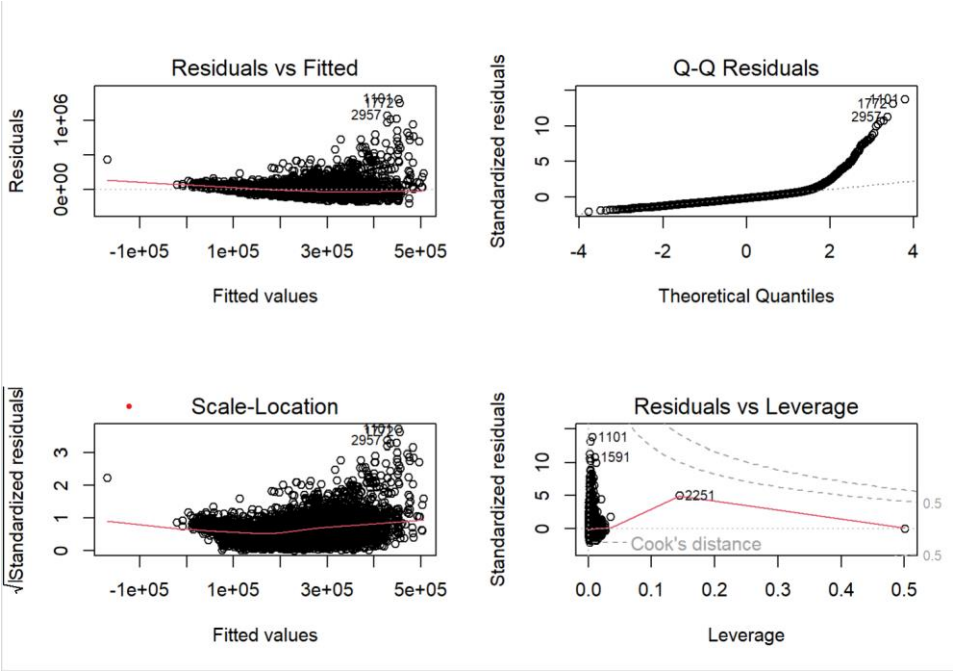
3.6 Konfidens- och prediktionsintervall

Konfidens och prediktionsintervall beräknades för de förutsagda priserna för att ge en uppskattning av osäkerheten i modellernas prognoser.

4 Resultat och Diskussion

I denna regressionsanalys användes en multipel regressionsmodell för att identifiera faktorer som signifikant påverkar prissättningen av begagnade bilar i Sverige. Modellen inkluderade variabler som miltal, modellår, märke, motor, växellåda och region. Resultaten visade att miltal och modellår är starka prediktorer för bilpriser, där nyare och mindre använda bilar tenderar att vara dyrare.

Diagnostiska plotter användes för att utvärdera modellens passform och identifiera eventuella dataavvikelser. Från 'Residuals vs Fitted' plotten observerades en relativt jämn spridning av residualer, vilket indikerar en god passform för modellen. Dock observerades en del outliers, vilket pekar på extrema värden som kan ha påverkat modellens prediktiva förmåga. 'Residuals vs Leverage' plotten visade att inga enskilda observationer hade oproportionerligt stort inflytande på modellens parameteruppskattningar, vilket är positivt för modellens robusthet.



Figur 5. Plott visualiseringar

En ytterligare modell, där interaktionstermer mellan märke, motor och växellåda inkluderades, konstruerades för att utforska mer komplexa samband. Denna modell visade lägre AIC och BIC värden, vilket tyder på en förbättrad modellanpassning jämfört med den initiala modellen utan interaktionstermer.

Under modellutvecklingen upptäcktes ingen uppenbar multikollinearitet bland de oberoende variablerna, vilket verifierades genom att kolla variansinflationsfaktorn (VIF). Detta bekräftar att modellen inte lider av allvarlig multicollinearitet som skulle kunna förvränga de uppskattade effekterna av oberoende variabler.

4.1 Justerat R^2 och modelldiagnostik

Justerat R^2 för den slutliga modellen var 0.4572, vilket indikerar att modellen förklarar nästan 46% av variansen i bilpriser. Detta är en tillfredsställande nivå för en ekonomisk modell där många externa faktorer kan påverka priset. F-testet i vår regressionsanalys visade att modellen är statistiskt signifikant

4.2 Modellval och Utvärdering

Modellval skedde genom jämförelse av flera modeller med och utan interaktionstermer. Modellen med interaktionstermer valdes för vidare analys baserat på dess statistiska indikatorer och förmåga att förklara en större del av variansen i responsvariabeln. Denna modell inkluderade färre variabler men med starkare prediktiv förmåga.

5 Slutsatser

Denna studie har visat att det är möjligt att noggrant förutsäga priser på begagnade bilar med hjälp av tillgängliga data om bilens egenskaper och historik. För bilhandlare och köpare kan insikterna från denna analys användas för att bättre förstå marknadspriser och göra mer informerade köp eller försäljningsbeslut.

För framtida forskning rekommenderas att inkludera ytterligare variabler som kan tänkas påverka bilpriser, såsom bilens tillstånd, exakta specifikationer och ekonomiska faktorer som räntor eller bränslepriser, för att ytterligare förbättra modellens noggrannhet och tillförlitlighet.

Denna analys är begränsad till data som samlats in från en specifik webbplats och kan därför inte nödvändigtvis generaliseras till andra marknader eller regioner utanför Sverige. Vidare kan den icke linjära dynamiken och de potentiella interaktionseffekterna mellan variablerna kräva mer avancerade statistiska tekniker eller maskininlärningsmetoder för att fullständigt fånga de underliggande sambanden.

6 Teoretiska frågor

1. **Kolla på följande video: https://www.youtube.com/watch?v=X9_ISJ0YpGw&t=290s , beskriv kortfattat vad en Quantile-Quantile (QQ) plot är.**
   En Quantile-Quantile (QQ) plot är ett statistiskt verktyg som används för att jämföra två sannolikhetsfördelningar genom att plotta deras kvantiler mot varandra. Om fördelningarna är lika kommer punkterna på QQ-plotten att ligga ungefär på linjen y = x. Denna typ av diagram är särskilt användbar för att kontrollera om data uppfyller antagandet om normalfördelning.

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?**
   I maskininlärning ligger fokuset på att bygga modeller som effektivt kan förutsäga framtida utfall baserat på tidigare data. Här är målet att maximera prediktionsnoggrannheten, ofta utan att gå djupare in i de underliggande sambanden mellan variablerna. I statistisk regressionsanalys, kan man både göra prediktioner och bedriva statistisk inferens. Statistisk inferens används för att förstå och förklara sambanden mellan olika variabler, vilket hjälper till att dra mer ingående slutsatser om data. Detta innebär att man inte bara förutsäger utfall, utan också förstår hur olika faktorer påverkar varandra och kan därmed fatta mer välgrundade beslut. (Bobbitt , 2021)

3. **Vad är skillnaden på ”konfidensintervall” och ”prediktionsintervall” för predikterade värden?**
   Konfidensintervall ger ett intervall där en populations parameter, som medelvärde, sannolikt ligger, baserat på urvalsdata. Prediktionsintervall anger istället ett intervall där framtida observationer förväntas falla med en viss sannolikhet, och tar hänsyn till både osäkerheten i skattningen och den naturliga variationen i data. Kort sagt, konfidensintervall handlar om skattning av populationsparametrar, medan prediktionsintervall handlar om att förutse framtida dataobservationer. (Statology, n.d.)

4. **Den multipla linjära regressionsmodellen kan skrivas som: $Y = \beta 0 + \beta 1x1 + \beta 1x2 + \ldots + \beta pxp + \varepsilon$. Hur tolkas beta parametrarna?**
   I den multipla linjära regressionsmodellen representerar β parametrarna effekten av varje oberoende variabel *xi* på den beroende variabeln Y. Varje β**i** (där *i* varierar från 1 till *p*) indikerar förändringen i *Y* för en enhets ökning i xi, med alla andra variabler hållna konstanta. Interceptet β**0** representerar det förväntade värdet av Y när alla x är lika med noll. (Guber, n.d.)

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?**
   När man använder mått som BIC i statistisk regressionsmodellering, kan man ibland undvika att dela upp data i tränings, validering och test set. BIC hjälper till att kvantifiera en modells passform samtidigt som den straffar för ökad komplexitet, vilket kan motverka överanpassning. Detta innebär att BIC tillhandahåller ett sätt att bedöma modellens kvalitet utan att nödvändigtvis behöva oberoende data för validering. Dock kan traditionell

uppdelning fortfarande vara fördelaktig för att testa modellens generaliserbarhet på oberoende data. (Johnson, n.d.)

6. Förklara algoritmen nedan för ”Best subset selection”

Algorithm 6.1 *Best subset selection*

1. Let $\mathcal{M}_0$ denote the *null model*, which contains no predictors. This model simply predicts the sample mean for each observation.
2. For $k = 1, 2, \ldots p$:
   (a) Fit all $\binom{p}{k}$ models that contain exactly k predictors.
   (b) Pick the best among these $\binom{p}{k}$ models, and call it $\mathcal{M}_k$. Here *best* is defined as having the smallest RSS, or equivalently largest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using using the prediction error on a validation set, C_p (AIC), BIC, or adjusted R^2. Or use the cross-validation method.

1. **Nollmodell:** Börja med en nollmodell, som inte inkluderar några prediktorer. Denna modell förutspår helt enkelt medelvärdet för varje observation.
2. **Passa modeller:** För en sekvens av *k= 1* till *p* prediktorer, passa alla möjliga modeller som innehåller exakt *k* prediktorer.
   - För varje *k*, bestäm den bästa modellen, betecknad *Mk*, som har den minsta residualsumman av kvadrater (RSS) eller ekvivalent det största R^2.
3. **Välj bästa modell:** Välj den bästa modellen *M*0, *M*1,...,*Mp* genom att använda antingen det minsta prediktionsfelet på ett valideringsset, *Cp* (AIC), BIC eller justerat R^2, eller använd korsvalideringsmetoden.

Logiken bakom denna algoritm är att överväga alla möjliga kombinationer av prediktorer för att bestämma vilken kombination som ger bäst passform enligt det valda kriteriet, såsom AIC, BIC eller justerat R^2. Detta hjälper till att ta itu med problemet med modellval genom att hitta en modell som balanserar avvägningen mellan passform och komplexitet. (James et al.2023)

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

George Boxs citat "All models are wrong, some are useful" betyder att även om ingen statistisk modell exakt återspeglar verkligheten fullt ut, kan de ändå ge värdefulla insikter och vara praktiska verktyg för beslutsfattande och förutsägelse, så länge vi är medvetna om deras begränsningar.

7 Självutvärdering

1. **Utmaningar du haft under arbetet samt hur du hanterat dem.**
   En av de utmaningar jag stötte på under arbetet var att genomföra regressionsanalysen och tolka resultaten.

2. **Vilket betyg du anser att du skall ha och varför.**
   Jag anser att betyget G är lämpligt för mig, då jag tror arbetet uppfyller de grundläggande kriterierna.

3. **Något du vill lyfta fram till Antonio?**
   Kursens fokus på självständigt lärande har varit givande. Det har utmanat mig att utveckla min förmåga att lösa problem och tillämpa teoretiska kunskaper praktiskt, vilket kommer vara en värdefull erfarenhet för min framtida karriär.

Källförteckning

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2014). *An Introduction to Statistical Learning* (2nd ed., corrected printing). Springer. Hämtad 10 april, 2024 från https://hastie.su.domains/ISLR2/ISLRv2_corrected_June_2023.pdf.download.html

Bobbitt, Z. (2021, October 8). Inference vs. Prediction: What's the Difference? Statology. Hämtad 12 april, 2024 från https://www.statology.org/inference-vs-prediction/

Statology. (n.d.). Confidence Interval vs. Prediction Interval: What's the Difference? Hämtad 12 april, 2024 från https://www.statology.org/confidence-interval-vs-prediction-interval/

Guber, J. (n.d.). How to Interpret Regression Coefficients. Statology. Hämtad 12 april, 2024 från https://www.statology.org/how-to-interpret-regression-coefficients/

Johnson, J. (n.d.). How can the AIC or BIC be used instead of the train/test split? StatsExchange. Hämtad 12 april, 2024 från https://stats.stackexchange.com/questions/360690/how-can-the-aic-or-bic-be-used-instead-of-the-train-test-split