R programmering

En linjär regressionsmodellering av insamlat data och data från Statistiska Centralbyrån (SCB) för år 2021



Nil Abukar

EC Utbildning

R programmering

2024-04-26

# Abstract

This report aims by conducting a linear regression analysis to investigate the relationship between fuel (bränsle) and horsepower (hästkrafter) in vehicles from gathered data from Blocket and Statistiska Centralbyrån (SCB) from 2021. Data has been gathered from Blocket between 2018-2023 over sales of vehicals however this report aims to focus on the year 2021. By using packages such as tidyverse and readxl and caret in the R programming language allow an understanding of the relationship between fuel and horsepower in vehicles. Moreover it evaluates the ability of a linear regression model to predict horsepower based on fuel. The result shows that the linear regressionmodel fits the data well and that fuel is a significant predictor for explaining the variation in horsepower.

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

Innehållsförteckning

[Abstract 2](#_Toc165058337)

[1 Inledning 1](#_Toc165058338)

[2 Teori 2](#_Toc165058339)

[2.1 Linjär regressionsmodell 2](#_Toc165058340)

[2.2 Tidyverse 2](#_Toc165058341)

[2.3 Explorativ dataanalys (EDA) 3](#_Toc165058342)

[3 Metod 4](#_Toc165058343)

[4 Resultat och Diskussion 5](#_Toc165058344)

[5 Slutsatser 8](#_Toc165058345)

[6 Teoretiska frågor 9](#_Toc165058346)

[7 Självutvärdering 11](#_Toc165058347)

[Källförteckning 12](#_Toc165058348)

# Inledning

Syftet med denna rapport är en linjär regressionanalys över insamlade data från Blocket och SCB över bilar från 2021, främst sambandet mellan hästkrafter och bränsle. För att uppfylla syftet med rapporten kommer följande frågeställningar att besvaras.

Syftet med denna rapport är komparativ studie mellan Random Forest Classifier och Support Vector Machine, för att uppfylla syftet så kommer följande frågeställning(ar) att besvaras:

1. Finns det ett samband mellan bränsle och hästkrafter?
2. Hur anpassar sig den linjära regressionsmodellen till datan?

# Teori

I denna del av rapporten kommer att handla om teoretiska koncept gällande valda metoder för min data. En linjär regressionsmodell har tillämpats i kodningen av insamlade data från Blocket och SCB för för att förstå om det finns ett samband mellan bränsle och hästkrafter i fordon för år 2021. I R programmering även paket så som Tidyverse, Readxl (Excel) och Caret (cross validation)

## Linjär regressionsmodell

Linjär regressionsmodell är en statistisk modell för att analysera sambandet mellan en responsvariabel, (ofta kallad Y) och en eller flera variabler samt deras interaktioner (ofta kallad X). Det är ven en av de mest grundläggande statistiska modellerna som finns och dess resultat kan tolkas av nästan alla (Porras, 2022). De primära formerna av linjär regression är enkel linjär regression och multipel linjär regression (Linear Regression in R ~ a Guide With Examples, 2023). I en enkel linjär regression finns en oberoende och en beroende variabel. Modellen uppskattar lutningen och skärningspunkten för den bästa passade linjen vilket representerar relationen mellan variablerna. Lutningen representerar förändringen i den beroende variabeln för varje enhetsförändring i den oberoende variabeln, medan skärningspunkten representerar det predikterade värdet av den beroende variabeln när den oberoende variabeln är noll. En multipel linjär regression är när det finns flera oberoende variabler som används för att prediktera en beroende variabel. Till skillnad från enkel linjär regression används multipel linjär regression för att modellera sambandet mellan flera variabler mellan de oberoende och beroende variablerna (Mali, 2024).

Det används i prediktiv analys av data där prediktionen av en rak linje som kan användas för att få en förstå sambandet. Linjär regressions syfte är att identifiera linjen för att minimera skillnaderna mellan observerade datapunkter och linjens förväntade värden (GeeksforGeeks, 2023).

Ett sätt att testa kvaliteten på modellens anpassning är att titta på residualerna eller skillnaderna mellan de verkliga och predikterade värdena (Porras, 2022). I den oberoende variabeln antar residualerna att felaktigheterna (residualerna) saknar samband mellan varandra. Detta innebär att felaktigheten i en observation bör vara oberoende av en annan felaktighet. I homoscedasticity innebär att variansen av residualerna är konstanta över alla nivåer av den oberoende variabeln. Spridningen av residualerna bör vara ungefär lika för alla värden av X. Detta antagande kan kontrolleras genom att skapa en plot av residualer mot anpassade värden. Normalitet av residualerna fokuserar på antaganden att residualerna följer en normal distribution (GeeksforGeeks, 2023a).

## Tidyverse

Tidyverse är ett sammanhängande system av paket för datahantering, utforskning och visualisering. Tidyverse-pakten är avsedda att göra statistiker och datavetare mer produktiva genom att guida dem genom arbetsflöden som underlättar kommunikation och resulterar i reproducerbara arbetsprodukter. Grundläggande handlar tidyverse om kopplingar mellan verktygen som gör arbetsflödet möjligt. Tidyverse paketen och funktionerna är resultatet av försök-och-fel experiment som har pågått under flera år för att hitta en minimal uppsättning av funktioner som är tillräckliga för att möjliggöra det kanoniska data science arbetsflödet. Efter flera försök av att identifiera och konstruera den mest fördelaktiga uppsättningen har tidyverse mognat till sin nuvarande form.

Det finns dock begräsningar med tidyverse som beror på faktorer som utanför designers kontroll medan andra kan vara avsiktliga. Begräsningarna av den första typen kan uppstå på grund av bristande överenskommelse om huruvida vissa data kan eller bör pressas in i en ’’rektangulär’’ datastruktur. Behovet av nära samordning mellan tidyverse-paketen producerar vissa begränsningar av den andra sorten. Det finns många högkvalitativa R-paket som är av stor nytta för data scientist, men baserade på designmål som skiljer sig från de i tidyverse (Views, 2017).

## Explorativ dataanalys (EDA)

Explorativ dataanalys (EDA) används av data scientist för att undersöka och analysera datamängder och sammanfatta deras huvudsakliga egenskaper, ofta genom att använda metoder för data visualisering. EDA hjälper till att bestämma hur man bäst manipulerar datakällor för att få de svar man behöver, vilket gör det enklare för datavetare att upptäcka mönster, avvikelser, testa en hypotes eller kontrollera antaganden *(What Is Exploratory Data Analysis? | IBM, n.d.).*

EDA används främst för att se vad data kan avslöja bortom den formella modelleringen eller hypotesproövning och ger en bättre förståelse för variablerna i datamängden och relationer mellan dem. Det kan också hjälpa till att avgöra om de statistiska teknikerna som övervägs för dataanalys är lämpliga *(What Is Exploratory Data Analysis? | IBM, n.d.).*

EDA är en grundläggande kreativ process där nyckeln till att ställa kvalitativa frågor för att kunna generera en större kvantitet av frågor. Varje fråga som ställs blir man exponerad till en ny aspekt av sin data och detta ökar chanserna till göra nya och flera upptäck (Grolemund, n.d.)

# Metod

En linjär regressionsanalys har utförts för att undersöka sambandet mellan två variabler bränsle och hästkrafter. Detta har utförts med hjälp av två excel filer därav en är data över bilar som är upp för försäljning på Blocket och en extern data över fordonstatistik från SCB .

Vid datainsamlingen från Blocket som består av kolumnnamn såsom modellår 2018-2023, bränsle (hybrid = 0, Bensin=1, Diesel=2), växellåda (Automat=0, Manuelll= 1), hästkrafter, pris och länk till annonserna. Den externa data över fordonstatiskt avser fordon, personbilar, i trafik och kommun och år. Den externa data har fokuserats på antal bilar som har varit i trafik under 2021 likväl datainsamlingen från blocket.

Sverige har sedan juli 2018 ett statligt styrmedel, det vill säga reduktionsplikt, som ska verka för att minska växthusgasutsläppen från vägtrafiken. Reduktionsplikten fastställer hur mycket växthusgasutsläpp från drivmedelsprodukter ska minskas, vilket i praktiken innebär en främjade av biodrivmedel och en gradvis utfasning av fossila bränslen såsom råolja (Statistiska Centralbyrån, 2022). Företag ges frihet att välja olika biodrivmedel med varierande grad av växthusgasutsläpp, och hur mycket av dessa biodrivmedel som ska blandas in i bensin och diesel. För att ge konsumenterna tydlig information om drivmedlens sammansättning och ursprung finns det krav på att redovisa utsläppsdata, samt information om vilka råvaror som ingår i drivmedlet och var det kommer ifrån.

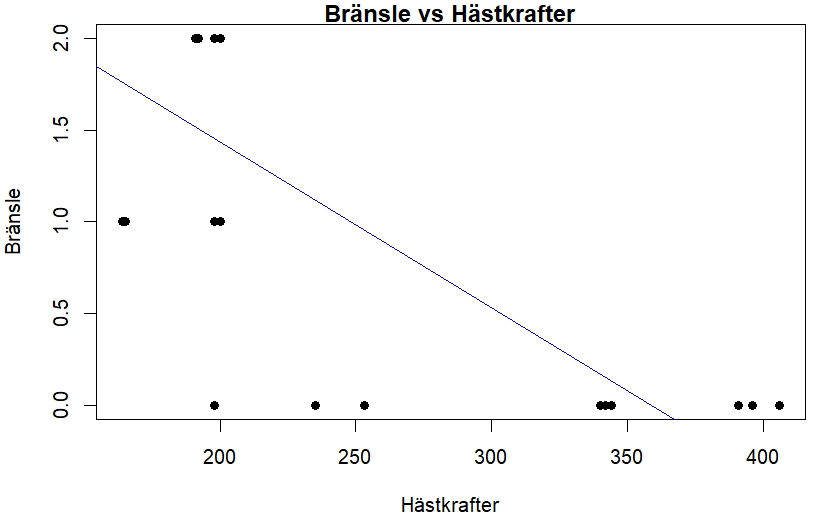
Vid årsskiftet 2021/2022 steg priset på diesel på tankstationer till över 20 kronor per liter, medan bensinpriset nådde cirka 19 kronor per liter i slutet av januari. Priset på drivmedel består av flera komponenter, inklusive moms, punktskatt och produktkostnad. Produktkostnaden påverkas av olika faktorer, där råolja är den mest betydande vid tillverkning av bensin och diesel. Prisutvecklingen för råolja påverkar i hög grad drivmedelspriserna och styrs av storpolitik, geopolitiska händelser samt den globala ekonomins utveckling, vilket ibland kan variera kraftigt (Statistiska Centralbyrån, 2022).

I denna rapport valdes år 2021 eftersom det rådde en storspridning av sjukdomen covid-19 som påverkade hela världen däribland många länder fick stänga ner sina samhällen för att minska på spridningen av sjukdomen. Den 11 mars 2020 meddelade World Health Organization (WHO) att covid-19 är numera en pandemi efter att ha upptäckt ett antal utbrott av fall den 31 december 2019 i Wuhan, Kina. I Sverige fanns det inga så kallade’’lockdowns’’ utan restauranger, biografer och övriga allmänna platser fick begränsas upp mot antalet person och öppettider (Folkhälsomyndigheten, 2023). Många länder valde dock stänga ner sina samhällen och hade strikta restriktioner där många inte fick lämna sina hem förutom om deras arbete var samhällsviktiga, till exempel sjukvårdspersonal. Pandemin hade en stor påverkan på världsekonomin och det har lett till att många länder har en hög inflation. Vidare har inflationen i världen drivits av flera saker utöver inflation men också stigande energi- och råvarupriser (Torstensson, S. 2021).

Med hjälp av linjär regression, EDA, visualiseringar och cross validation kommer datan från Blocket och SCB att studera om det finns ett samband mellan bränsle och hästkrafter från 2021 samt antal bilar i trafik i Sverige. Vidare ska det studeras huruvida har pandemin och världsekonomin haft en påverkat samt miljöpolitiken.

# Resultat och Diskussion

Min kod utför en omfattande linjär regressionsanalys av två excel filer vid namn datainsamling\_grupparbete (data) och 2021\_bilar (scb\_data). Modellen visar en att för varje enhet typ av bränsle minskar hästkrafter med 72.498 enheter. Detta antyder på att olika bränsletyper kan ha olika påverkan på motoreffekten hos fordon. Det har gjorts en signifikanstest som visar att både interceptet och koefficienten för bränsle är signifikanta (p<0.001), vilket indikerar att de har en betydande effekt på hästkrafter.



Interceptet, tillika brytpunkten, för hästkrafter är 325.013 och kofficienten för bränsle är – 72.498. Detta visar på att linjens lutning korsar y-axeln och indikerar på när ett fordon är hybrid kan det förväntas att hästkrafterna är 325.013. Koefficienten visar för varje ökning det vill säga huruvida det är hybrid, bensin eller diesel minskar hästkrafterna med 72.498 enheter. Detta föranleder till att för båda koefficienterna är p värdet mycket låg därav är de signifikanta och att det finns en statistisk signifikant linjär relation mellan bränsle och hästkrafter.

Modellen visar att det standardavvikelsen för residualerna på 48.18 ger en indikation på att spridningen av restvärdena kring den linjära regressionen. Ju lägre detta värde är desto bättre passar modellen datan. Dessutom utifrån R-squared och adjusted R-squared visar att proportionen av variansen i den beroende variabeln (hästkrafter) som förklaras av den oberoende variabeln (bränsle) med en R-square värde på 0.6553. I detta fall med en 65.53 % av variationen i hästkrafter av variationen i bränsle. Med hänsyn till den adjusted R-squared på 0.6502 tar det hänsyn till antalet prediktorer i modellen och ger ett mer konserverat mått på modellens passform.

Detta visar på att det de facto finns en signifikant linjär relation mellan bränsle och hästkrafter. Modellen förklarar cirka 65.53 % av variationen i hästkrafterna baserat på variationen i bränslet och har en låg residual standard error, vilket tyder på att den passar data väl. Det visar även att det har en god förmåga att förutsäga hästkrafterna baserat på bränsle med hänsyn till komplexiteten av att fler prediktorer läggs till. Det kan dock även ifrågasättas huruvida det passar datan väl med en hög variation i hästkrafter och en låg residual standard error.

En bild som visar text, skärmbild, linje, diagram

Automatiskt genererad beskrivning

Däremot visar ploten residuals vs fitted values en tydlig trend längs den horisontella axeln eftersom de är slumpmässigt fördelade över de förutsedda värdena. Det visar heller inte en jämn spridning runt noll längs den vertikala axeln vilket innebär att variationen i residualerna inte bör öka eller minska systematiskt över de förutsedda värdena.

En bild som visar text, diagram, linje, Graf

Automatiskt genererad beskrivning

I tabellen normal Q-Q plot ligger punkterna inte i en fullständig rät linje som ger stöd för att antaganden för modellen är uppfyllda och att den är lämplig för att göra prediktioner. Det visar att modellen har vissa brister eller att den inte fångar hela variationen i data på ett adekvat sätt. Det kan leda till systematiska avvikelser så som att modellen indikerar på att det inte kan förklara visa aspekter av data som kan beror på icke-linjära samband eller att viktiga prediktorer saknas i modellen. En annan aspekt är att det finns avvikande punkter från den förväntade normalfördelningslinjen som indikerar närvaron av outliers eller influentiella observationer som kan påverka modellens prestada. Det är därmed viktigt att identifiera och undersöka sådana observationer för att avgöra om de bör inkluderas eller uteslutas från analysen. I detta fall saknas det skäl att utesluta detta från analysen eftersom det medför en vidare diskussion för framtida studier.

En bild som visar text, skärmbild, diagram, linje

Automatiskt genererad beskrivning

Resultatet av modellen visar en residual analys plot för att undersöka homoscedasticity eller konstanta varianser i residualerna över olika nivåer av förutsagda värden (fitted values). Ploten visar att absolutvärden av residualerna är slumpmässigt fördelade runt en horisontell linje vid en konstant nivå, vilket indikerar att variansen i residualerna är de facto konstanta.

Scb\_datan har visat att den har varit ofullständig eller organiserad när det har vävts in i modellen. Det har vidare inte stoppat från att frågeställningarna ska besvaras om det finns ett samband mellan hästkrafter och bränsletyp och hur väl datan anpassar sig till en linjär regressionsanalys. Detta kommer att diskuteras vidare i nästa kapitel i denna rapport.

# Slutsatser

Modellen visar att det finns en signifikant linjär relation mellan bränsle och hästkrafter. Det visar olika bränsletyper kan påverka motoreffekten på olika sätt, vilket bekräftas av signifikanta intercept- och koefficentvärden. Modellen visar också att för varje enhet av bränsletyp minskar hästkrafter med cirka 72 enheter vilket visar att sambandet är statistiskt signifikant.

Genom att analysera spridningen av restvärdena kring den linjära regressionen och utvärdera R-squared-värdet bedöms hur väl modellen passar. Trots vissa brister i modellen så som icke uppfyllda antaganden och avvikande punkter i Q-Q plotten visar residual analysen att variansen i residualerna är konstant.

Avseende Sveriges statliga styrmedel, reduktionsplikt som infördes 2018 för att minska växthusgasutsläppen från vägtrafiken. Korrelationen mellan den linjära regressionsanalysen och hur pandemin påverkade världsekonomin visar hur olika faktorer påverkar bränslepriser och därmed också användningen av olika bränsletyper. Modellen betonar vikten att förstå sambandet mellan bränsletyp och fordonseffekt och ekonomiska/politiska faktorer för att utforma effektiva åtgärder för att minska växthusgasutsläppen och främja hållbar transport. Genom att integrera både vetenskaplig forskning och politiska initiativ kan vi arbeta mot en mer hållbar och miljövänlig fordonindustri.

Sammanfattningsvis, finns det utrymme för förbättringar och vidare undersökningar som ger denna linjära regressionsanalys en värdefull grund för att förstå sambandet mellan bränsletyp och hästkrafter hos fordon. Vidare forskning kan fokusera på att förbättra modellen genom att inkludera ytterligare prediktorer, använda sig av en multipel linjär regressionsanalys och hantera brister i data för att få en mer heltäckande bild av detta samband.

# Teoretiska frågor

1. En Quantile-Quantile (QQ) plot är en grafisk metod som bedömer huruvida observationer är normalfördelade. Vikten att bedöma normalitet är viktigt eftersom många statistiska samverkan antar att en provtagning sker från en normalfördelad population.
2. Både maskininlärning och statistisk regressionsanalys gör prediktioner dock skiljer deras syfte åt. I maskininlärning är syftet att utveckla modeller som kan göra prediktioner baserat på data. Det är främst om att optimera prediktionerna för att minimera fel i förutsägelserna. Å andra sidan inom statistisk regressionsanalys utöver prediktionerna fokuserar det på att förstå relationer mellan variabler och dra slutsatser om populationen baserat på provdata. Detta innefattar att bedöma signifikansnivåer, konfidensintervall och andra statistiska mått för att göra inferenser om populationen. Sammanfattningsvis inom maskininlärning tenderar det att vara mer inriktad på att optimera prediktioner medan statistisk regressionsanalys inkluderar både prediktion och statistisk inferens för att förstå sambandet och göra generaliserbara slutsatser om populationen.
3. Konfidensintervall är en intervalluppskattning av en parameter i en population och ger en idé om osäkerheten i vår uppskattning av parameterns värde. Det representerar intervallet där man med en vis konfidensgrad förväntar sig att det sanna värdet av parametern ligger baserat på provdata. Prediktionsintervall är en intervalluppskattning av ett framtida observationsmässigt värde från en population. Det används för att bedöma osäkerheten i framtida prediktioner och bedöma hur väl modellen generaliserar till nya data. Till skillnad från konfidensintervall så representerar intervallet med en vis konfidensgrad en förväntning av var det framtida värdet kommer att ligga baserat på modellens prediktiva förmåga och osäkerhet. Det tar även hänsyn till både osäkerheten i modellparameterna och variabiliteten i data.
4. Beta-parametrarna representerar olika aspekter av sambandet mellan de oberoende variablerna och den beroende variabeln Y. 𝛽0 som även kallas för intercept är detta konstanten eller förväntade värdet av Y när alla oberoende variabler är noll. Det representerar den genomsnittliga effekten på Y när alla oberoende variabler är noll. Koefficienterna för oberoende variabler tillika 𝛽1𝑥1 + 𝛽1𝑥2+ . . . + 𝛽𝑝𝑥𝑝 visar hur mycket den förväntade värdet av Y förändras när den oberoende variabeln ökar med ett enhetsvärde medan alla andra variabler hålls konstanta. Varje 𝛽 koefficient representerar den marginella effekten av den motsvarande oberoende variabeln på Y förutsatt att alla andra variabler är konstanta.
5. Det stämmer inte. Vid användning av BIC, som är ett mått på modellkomplexitet, straffar modeller med flera parametrar. BIC tillför jämförelse av modeller och väljer den bästa modellen avseende komplexitet och passform till datan. Det ersätter därmed inte behovet att dela upp datan i träning, validering och test set. Syftet med att använda träning, validering och test set är att säkerställa att modellen generaliserar väl till nya data och att resultaten är kraftfull med tillförlitliga modellutvärderingar.
6. Formeln best subset selection identifierar den bästa modellen i träningsdatan för varje delmängd. Detta göras för att minimera problem från en av 2p möjliga modeller till en av p+1 möjliga modeller. För att välja den bästa modellen behöver man välja bland dessa p + 1 alternativen. Detta måste genomföras med försiktighet eftersom RSS av dessa p + 1 modeller minskar montont och R2 ökar montont samtidigt som antalet inkluderande funktioner i modellerna ökar.

# Självutvärdering

1. Utmaningar du haft under arbetet samt hur du hanterat dem.

* Det har varit att komma på rimliga frågeställningar som har varit intressanta för mig och försöka smala av data för att inte göra det för stort och komplicerat samt att modellen får ytterligare brister än vad som har framkommit.

1. Vilket betyg du anser att du skall ha och varför.

* Jag anser att jag bör få G då jag har nått samtliga kriterier för att nå G i betyg.

1. Något du vill lyfta fram till Antonio?

* Tack för denna kurs. Det har varit väldigt givande och intressant som alla andra kurser. Tack för att du lyssnar och har förståelse. Din stöttning är enormt viktigt och motiverar en ännu mer. Så tack!

# Källförteckning

Bevans, R. (2023, June 22). Linear Regression in R | A Step-by-Step Guide & Examples. Scribbr. <https://www.scribbr.com/statistics/linear-regression-in-r/#:~:text=Linear%20regression%20is%20a%20regression,total%20error%20of%20the%20model>.

Porras, E. M. (2022, December 5). Linear regression in R tutorial. <https://www.datacamp.com/tutorial/linear-regression-R>

GeeksforGeeks. (2023, November 15). Simple linear regression in R. GeeksforGeeks. <https://www.geeksforgeeks.org/simple-linear-regression-using-r/>

Linear Regression in R ~ A guide with examples. (2023, August 25). <https://www.bachelorprint.com/statistics/linear-regression-in-r/#:~:text=Linear%20regression%20is%20a%20foundational,of%20a%20linear%20regression%20visual>.

Mali, K. (2024, January 23). Everything you need to Know about Linear Regression! Analytics Vidhya. <https://www.analyticsvidhya.com/blog/2021/10/everything-you-need-to-know-about-linear-regression/>

GeeksforGeeks. (2023a, November 7). Linear Regression Assumptions and Diagnostics using R. GeeksforGeeks. <https://www.geeksforgeeks.org/linear-regression-assumptions-and-diagnostics-using-r/>

Views, R. (2017, June 8). What is the tidyverse? <https://rviews.rstudio.com/2017/06/08/what-is-the-tidyverse/>

IBM, What is Exploratory Data Analysis? | IBM. (n.d.). <https://www.ibm.com/topics/exploratory-data-analysis>

Folkhälsomyndigheten, *När hände vad under pandemin*? (2023, June 13). <https://www.folkhalsomyndigheten.se/smittskydd-beredskap/utbrott/aktuella-utbrott/covid-19/nar-hande-vad-under-pandemin/>

Statistikmyndigheten SCB. *Drivmedelspriserna på rekordnivåer*. (n.d.). <https://www.scb.se/hitta-statistik/temaomraden/sveriges-ekonomi/fordjupningsartiklar_Sveriges_ekonomi/drivmedelspriserna-pa-rekordnivaer/>

Torstensson, S. (2021), *Fortsättningen på det svängiga läget i världsekonomin - Ekonomifakta*. (n.d.). Ekonomifakta. <https://www.ekonomifakta.se/Artiklar/2021/november/fortsattningen-pa-det-svangiga-laget-i-varldsekonomin/>