Linjär regressionsmodell i R

Optimerad prissättning av begagnade bilar



Arina Godman

EC Utbildning

202404

# Abstract

In this report, we gathered data on secondhand cars from Blocket.se as part of a collaborative project. This dataset served as the foundation for constructing and training a regression model aimed at optimizing the pricing of used cars. Leveraging techniques such as stepwise forward selection of variables, our linear regression model attained an impressive 79.41% accuracy (justerat ), with an RMSE representing only 20% of the mean real price in the test dataset.

Additionally, we augmented our analysis by incorporating external data obtained from the Swedish Central Statistical Bureau (SCB) through an API.

Innehållsförteckning

[Abstract 2](#_Toc164935159)

[1 Inledning 1](#_Toc164935160)

[2 Teori 3](#_Toc164935161)

[2.1 Regressionsmodeller 3](#_Toc164935162)

[2.1.1 Uppskattning av koefficienterna 3](#_Toc164935163)

[2.1.2 Utvärdering av modellens noggrannhet 3](#_Toc164935164)

[2.2 Val av variabelset 4](#_Toc164935165)

[2.2.1 Sunt förnuft 4](#_Toc164935166)

[2.2.2 Stepwise subset selection 4](#_Toc164935167)

[3 Metod 5](#_Toc164935168)

[3.1 Data 5](#_Toc164935169)

[3.1.1 Datainsamling or rensning 5](#_Toc164935170)

[3.1.2 Outliers och korrelation 5](#_Toc164935171)

[3.1.3 Kategoriska data och Dummies 6](#_Toc164935172)

[3.2 Modellering och modellutvärdering 7](#_Toc164935173)

[3.3 API för inhämtning av extern data 7](#_Toc164935174)

[4 Resultat och Diskussion 8](#_Toc164935175)

[4.1 Undersökning av potentiella problem 8](#_Toc164935176)

[4.2 Modellval 9](#_Toc164935177)

[4.2.1 lm\_4 9](#_Toc164935178)

[4.2.2 lm\_5 9](#_Toc164935179)

[4.2.3 lm\_6 9](#_Toc164935180)

[4.3 Modellutvärdering 10](#_Toc164935181)

[4.4 Begränsningar 10](#_Toc164935182)

[5 Slutsatser 11](#_Toc164935183)

[6 Teoretiska frågor 12](#_Toc164935184)

[7 Självutvärdering 14](#_Toc164935185)

[8 Grupparbete 14](#_Toc164935186)

[Källförteckning 16](#_Toc164935187)

# Inledning

I den nuvarande ekonomiska situationen, där priserna stiger, blir det allt vanligare att välja att köpa begagnade bilar i stället för att investera i en helt ny bil. Enligt data från SCB (Statistiska centralbyrån) har antalet nyregistrerade bilar hos Transportstyrelsen minskat varje år sedan 2021 men antalet personer som får körkort är ungefär samma varje år enligt information från Transportstyrelsen.se.

En bild som visar text, linje, Graf, diagram

Automatiskt genererad beskrivning

Figur 1. Antal nyregistrerade bilar per år

En bild som visar skärmbild, Graf, linje, text

Automatiskt genererad beskrivning

Figur 2. Antal nyregistrerade bilar per månad för år 2022, 2023 och 2024

Att förlora pengar vid försäljning av en bil och osäkerheten kring vilket pris som är rimligt för både säljare och köpare är vanliga bekymmer. Att förlita sig på magkänsla är inte alltid den mest tillförlitliga strategin.

Syftet med detta projekt är att utveckla en regressionsmodell för att optimera prissättningen av begagnade bilar. Genom att tillämpa avancerade analysmetoder strävar jag efter att ge både säljare och köpare insikter som kan bidra till en rättvis och ekonomiskt fördelaktig affär. För att uppnå syftet kommer jag att genomföra följande steg:

1. Insamling och rening av data som ska användas för att träna modellen.
2. Identifiering av ett optimalt variabelset.
3. Säkerställande av modellens tillförlitlighet och möjligheten att genomföra statistiska inferenser.
4. Utveckling av en modell som kan uppnå en noggrannhet på 80% (justerat R2).

# Teori

## Regressionsmodeller

Linjär regression är en grundläggande teknik inom maskininlärning som används för att modellera och analysera relationer mellan en eller flera oberoende variabler och en kontinuerlig beroende variabel. Syftet med linjär regression är att hitta det linjära sambandet mellan Y och X.

I formel ovan är Y den beroende variabeln och är de oberoende eller förklarande variablerna. är okända konstanter som representerar effekt av på beroende variabel Y (James, Witten, Hastie & Tibshirani, 2014, p.72). Regressionsmodellering utgår på att hitta uppskattningar förkoefficienter och kunna göra pålitliga prediktioner.

### Uppskattning av koefficienterna

För att göra prediktioner med linjär regression behöver vi först uppskatta koefficienterna eftersom dessa är okända. Vi använder data för att estimera dessa koefficienter. För varje observation i datasetet strävar vi efter att hitta uppskattning för koefficienterna att den linjära modellen passar datasetet väl. Målet är att hitta så att den linje som skapas är så nära som möjligt de givna datamängderna. För att mäta detta används ofta minsta kvadratmetoden, där målet är att minimera residualernas summa av kvadraterna (RSS) (James, Witten, Hastie & Tibshirani, 2014, p.62).

Uppskattningar på kan härledas från ovanstående formeln.

### Utvärdering av modellens noggrannhet

#### RSE och RMSE

RSE (Residual Standard Error) och RMSE (Root Mean Square Error) används för att utvärdera modellens noggrannhet genom att mäta skillnaden mellan de observerade värdena och de förutsagda värdena från modellen. RSE representerar det genomsnittliga felet mellan de faktiska och förutsagda värdena, medan RMSE tar roten ur RSE för att ge en indikation på hur mycket modellen avviker från de faktiska värdena i genomsnitt. Ju lägre värde på RMSE, desto bättre passar modellen data.

#### och justerat

-statistiken ger ett alternativt mått på passning och tar alltid värden mellan 0 och 1. visar vilken procent av variansen i datan kan förklaras med vår modell.

Justerat är ett mått som justerar träningsfelet för att kompensera för överanpassning i modellen. Det tar hänsyn till antalet parametrar och straffar överdriven komplexitet, vilket föredrar enklare modeller som ändå presterar väl. Genom att använda justerat kan vi jämföra modeller med olika antal parametrar utan att behöva använda olika data set för träning och utvärdering av modeller (James, Witten, Hastie & Tibshirani, 2014, p.235).

## Val av variabelset

Att välja det optimala subsetet av variabler för en regressionsmodell är en viktig del av modellutvecklingsprocessen. Genom att välja de mest relevanta variablerna kan vi förbättra modellens prediktiva förmåga samtidigt som vi undviker överanpassning genom att eliminera onödiga variabler. Detta hjälper till att skapa en modell som är mer robust och lättare att tolka. Det finns olika tekniker och strategier för att genomföra subset selection, varav varje har sina egna fördelar och begränsningar beroende på datamängden och modellens mål.

### Sunt förnuft

Att välja variabler utifrån sunt förnuft innebär att använda förståelse för problemområdet och relevanta teorier för att välja de variabler som förväntas ha en stark koppling till den beroende variabeln. Det kan innebära att inkludera variabler som har visat sig vara signifikanta i tidigare forskning eller som har en logisk förklaring för deras påverkan på resultatet. Genom att tillämpa sunt förnuft kan vi skapa en mer intuitiv och meningsfull modell som är mer följsam mot verkliga förhållanden.

### Stepwise subset selection

För att välja en optimal subset av variabler för regressionsmodellen används stegvis urvalsmetoder som alternativ till Best subset selection. Både forward och backward stepwise selection är attraktiva alternativ till best subset selection på grund av deras begränsade sökutrymme, vilket minskar risken för överanpassning och hög varians hos koefficientskattningarna. Vid forward stepwise selection börjar vi med en modell utan förklarande variabler och lägger till en variabel i taget tills alla variabler ingår. Backward stepwise selection börjar med en modell som innehåller alla variabler och tar sedan bort den minst användbara variabler i varje iteration. Dessa stegvisa urvalsmetoder är särskilt användbara när antalet variabler är stort och best subset selecion inte är praktiskt möjligt.

# Metod

## Data

### Datainsamling or rensning

Min arbetsgrupp och jag har manuellt samlat in data från Blocket.se och därefter lagt in det i ett Excel-dokument. Informationen som samlats in representerar begagnade "familjebilar" i prisklassen 100 000 - 500 000 kr från följande län: Dalarna, Stockholm, Östergötland, Jönköping, Kalmar, Gotland, Skåne och Västra Götaland. Sammanlagt har vi 701 observationer. Följande variabler samlades in:

|  |  |  |
| --- | --- | --- |
| * Pris (i svenska kronor) * Stad * Län * Bränsle * Växellåda | * Miltal * Modellår * Biltyp * Drivning * Hästkrafter | * Färg * Motorstorlek * Datum i trafik * Märke * Modell |

För det första utfördes grundläggande rensning i Excel. Under denna process upptäcktes att i varje kolumn var minst ett värde som hade mellanslag efter själva data, vilket resulterade i att dessa data lästes som separata kategorier i R eller numeriska data lästes in som kategoriska.

Andra problem och dess åtgärder kan du se nedan.

|  |  |  |
| --- | --- | --- |
| **Variabel** | **Problem** | **Lösning** |
| Pris | Innehöll mellanslag efter siffror | Hittade dessa värden och tog bort mellanslag |
| Bränsle | 3 olika namn för en kategori ”Miljöbränsle/Hyb” | Ersatt dessa värden med en ”Miljobransle\_Hyb” |
| Miltal/Datum i trafik / Modellår | Innehöll saknade värden och 0 för bilar med modellår 2024 och 2023 | Tog bort dessa observationer då min modell är till för prissättning av begagnade billar och inte helt nya |
| Färg | Innehöll nyanser av färger så som ”Mörkgrå”, ”Ljusbrun” osv | Hade standardiserat färgnamnen genom att ersätta nyanser som "Mörkblå" och "Ljusblå" med den generella termen "Blå". |
| Motorstorlek | Saknade värden för El-bilar | Ersatt saknade värden med 0 |
| Märke/Modell | Innehöll samma Märke/Modell men med olika stavningar | Ersatt dessa värden med en standard. T.ex. Kia CEED/KIA Cee’d / KIA ceed blev Kia Ceed |

Tabell 1. Rensning av data

Efter rensning av datan återstod 679 observationer för analys och modellträning, medan 22 observationer hade saknade värden och därför uteslöts från analysen. För att gå vidare till nästa steg har jag tagit beslut att bli av med variabler: Län, Stad, Datum i trafik och Motorstorlek då jag anser att de är mindre viktiga när det gäller prissättning av en bil.

### Outliers och korrelation

Som nästa steg genomfördes en grundläggande kontroll av numerisk data i R med hjälp av pairs-funktionen. Från visualiseringen kunde en outlier observeras, vilken var en bil med modellår 1970. Även om det kan finnas en rimlig förklaring till varför denna bil har ett högt värde, såsom samlarobjektstatus, så går detta emot syftet med modellen som är att prissätta begagnade familjebilar. Därför beslutades det att ta bort alla outliers med hjälp av z-poäng och tröskeln 3.

En bild som visar text, diagram

Automatiskt genererad beskrivning

Figur 3. Pairplot av numerisk data med outliers

Vid nästa steg undersökte jag korrelation mellan numeriska variabler vilket visade ett starkt linj’ärt samband mellan Miltal och Modellår. För att motverka multikollinearitet har jag valt att behålla bara Modellår eftersom det var mer kollerrad med bilens Pris än Miltal.

En bild som visar text, träd

Automatiskt genererad beskrivning

Figur 4. Pairplot av numerisk data utan outliers

### Kategoriska data och Dummies

Vid nästa steg genomförde jag omvandlingen av kategorisk data till dummyvariabler med hjälp av funktionen fastDummies. Denna funktion genererar automatiskt dummyvariabler för varje kategorisk variabel och tar bort en kategori för att undvika multikollinearitet. Trots att funktionen lm i R automatiskt konverterar kategoriska variabler till dummyvariabler, valde jag att använda fastDummies för att ha möjlighet att manuellt utföra val av variabelset i nästa steg.

## Modellering och modellutvärdering

Vid detta steg delade jag upp datasetet i tränings- och testset med hjälp av funktionen sample. Därefter utförde jag modellering av linjära regressionsmodeller med funktionen lm i R. Totalt konstruerades sex modeller med olika variabeluppsättningar: två modeller skapades med hjälp av Forward Stepwise Selection och Backward Stepwise Elimination.

Modeller:

1. lm\_1 – modell som innehåller alla variabler förutom: Län, Stad, Motorstorlek och Miltal
2. lm\_2 – modell som inte innehåller variabler med standardavvikelse = 0
3. lm\_3 – modell utan multikollinearitet
4. lm\_4 – modell utan multikolinearitet och high leverage punkter
5. lm\_5 – lm\_4 med variabelset skapat av forward stepwise selection
6. lm\_6 – lm\_4 med variabelset skapat av backward stepwise elemination

Genom att använda justerat R² valde jag den bästa modellen och beräknade dess RMSE. Jag använde också modellen för att prediktera priset på min bil, som jag köpte i november 2023, och skapade konfidens- och prediktionsintervall för denna förutsägelse.

För varje modell utförde jag också diagnostiska plotar för att säkerställa modellens tillförlitlighet.

## API för inhämtning av extern data

För att hämta extern data från SCB (Statistiska Centralbyrån) använde jag Statistikdatabasens API (PxWebApi 1.0) och programmeringsspråket Python. Processen involverade att ansluta till SCBs databas och genomföra sökningar (querying) för att hämta data i realtid. Jag valde att använda json-kodning eftersom det var enklast.

För att analysera datan använde jag bibliotek pandas och matplotlib.pyplot och seaborn för att plota datan.

# Resultat och Diskussion

## Undersökning av potentiella problem

En regressionsmodell (lm\_1) med alla variabler utom Län, Stad, Motorstorlek och Miltal var utgångspunkten för min undersökning. Från diagnosploten nedan ser vi att residualerna är ungefär normalfördelade och att det inte finns något tydligt icke-linjärt samband mellan residualerna och de anpassade värdena. Residuals vs Leverage plot visar att det kan finns extrema punkter i vår data men vid rensning av data har vi blivit av med extrema punkter.

. Dock visade modellen att tre variabler - "Marke\_Volvo", "Marke\_Dacia" och "Marke\_Chrysler" - hade en standardavvikelse på 0. För att undersöka multikollinearitet med hjälp av funktionen vif var jag tvungna att ta bort variablerna "Marke\_Volvo", "Marke\_Dacia" och "Marke\_Chrysler" och träna om modellen (lm\_2), vilket resulterade i exakt samma diagnosplot och justerat.

En bild som visar text, diagram

Automatiskt genererad beskrivning

Tabell 1. Diagnosplott för lm\_1 och lm\_2

Undersökningen av multikollinearitet (lm\_2) med hjälp av vif-funktionen visade att det fanns multikollinearitet mellan variablerna "Biltyp\_Halvkombi", "Biltyp\_Kombi" och "Biltyp\_SUV". Vidare analys av korrelationen mellan dessa variabler visade på ett starkt linjärt samband mellan "Biltyp\_Kombi" och "Biltyp\_SUV". För att hantera detta beslutade jag att ta bort variabeln "Biltyp\_Kombi". Efter att ha gjort detta visade nästa steg att multikollinearitet inte längre fanns i modellen (lm\_3). Däremot visade det sig att det fanns två punkter med cook's distans lika med 1 som inte plottades ovanför i diagnosploten. Efter vidare undersökning har jag beslutat att ta bort dessa extrema punkter (observationer).

På nästa sida kan ni se diagnosplot för lm\_4, modellen som är fri från multikollinearitet, påverkas inte av extrema punkter och innehåller inga variabler med en standardavvikelse som är lika med 0. Enligt ploten finns det ett svagt icke-linjärt samband mellan anpassade värdena och residualer. Men när vi undersöker plot anpassade värden vs roten ur standardiserade residualer set det ut som sambandet är icke-existerande. Om man undersöker påverkan av manipulationen på justerat då är skilnaden ganska liten.

En bild som visar text, diagram

Automatiskt genererad beskrivning

Figur 5. Diagnosplot lm\_4

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| lm\_1 | lm\_2 | lm\_3 | lm\_4 | lm\_5 | lm\_6 |
| 0,8153 | 0,8153 | 0,8152 | 0,7951 | 0,7951 | 0,7983 |

Tabell 2. Justerat R^2 per modell

## Modellval

Modellval genomfördes bland modeller: lm\_4, lm\_5 och lm\_6.

### lm\_4

lm\_4 är modell utan multikolinearitet, heteroskedastisitet och extrema punkter, som består av 47 variabler. Av dessa variabler har 16 signifikanta koefficienter med en p-värde mindre än 0.05. Justerat R^2-värdet är 0.7941, vilket indikerar att modellen förklarar cirka 79.41% av variationen i responsvariabeln. Den residuala standardfelet är 41390, vilket visar den genomsnittliga avvikelsen mellan de observerade värdena och de förutsagda värdena. F-testet visar att modellen är signifikant (p-värde: < 2.2e-16), vilket innebär att modellen som helhet är användbar för att förutsäga responsvariabeln.

### lm\_5

lm\_5 är en modell som byggdes på variabelset som var utvald med hjälp av forward stepwise selection som tillämpades till modell lm\_4. Det gav exakt samma output som lm\_4.

### lm\_6

Det är en modell som tränades på variabelset som var utvald med stepwise backward elimination. Modellen trännades på 24 varabler och 22 av dem har signifikanta koefficienter i modellen. Den residuala standardfelet är 40970 vilket är lite lägre än för modeller lm\_4/lm\_5 och justerat är något högre än för modeller lm\_4 och lm\_5. Dessutom innehåller den nästan hälften av antal variabler i modeller lm\_4 och lm\_5 vilket är bra för att motverka överanpassning.

Utifrån det överväger jag att välja modell lm\_6 för vidare undersökning och prediktioner.

## Modellutvärdering

Modellen Lm\_6 är konstruerad med en träningsuppsättning som inkluderar följande variabler: Modellar, Hastkrafter, Vaxellada\_Manuell, Biltyp\_Coupe, Biltyp\_Halvkombi, Biltyp\_SUV, Drivning\_Tvahjulsdriven, Farg\_Brun, Farg\_Gron, Farg\_Vit, Marke\_Fiat, Marke\_Ford, Marke\_Hyundai, Marke\_Jeep, Marke\_Land\_Rover, Marke\_Mercedes\_Benz, Marke\_Mitsubishi, Marke\_Nissan, Marke\_Porsche, Marke\_Renault, Marke\_Seat, Marke\_Suzuki, Marke\_Toyota och Marke\_Volkswagen.

Beräknad Root Mean Square Error (RMSE) för denna modell är 50837,4, vilket utgör 20% av det genomsnittliga priset i test-datasetet. Detta resultat anses vara tillfredsställande och i linje med de angivna målen för projektet.

Vid en explorativ användning har modellen applicerats för att förutsäga priset för mitt personliga bilinköp som genomfördes i november 2023. Prediktionen blev 138 465 kr med konfidensintervall (106 188; 170 741.9) och prediktionsintervall (50 933.49 och 225 996.5). Det riktiga priset som jag köpte billen för är 119 000 kr vilket ligger inom konfidens- och prediktionsintervallen.

## Begränsningar

Modellens begränsningar är främst förknippade med de data som användes för att utveckla den. Dessa data avser främst "vanliga familjebilar" med ett prisområde mellan 100 000 och 500 000 kronor. Det innebär att modellen endast kan förutsäga priser för bilar inom detta specifika prisspann. En annan begränsning är att datan samlades in från enskilda län i Sverige. Trots försök att samla in data från olika socioekonomiska grupper och områden med varierande befolkningsstorlek, kan modellen ändå vara tränad på ett icke-representativt urval. Detta kan påverka modellens förmåga att generalisera till andra regioner eller olika typer av bilar utanför det angivna prisspannet. Det är viktigt att vara medveten om dessa begränsningar när man applicerar modellen i praktiken.

# Slutsatser

I projektets resultat och diskussion har jag utforskat flera aspekter och potentiella problem av regressionsmodellering för att optimera prissättningen av begagnade bilar. Genom att noggrant analysera och manipulera modeller har jag kommit fram till slutsatser som kommer att bidra till en mer effektiv prissättningsstrategi för begagnade bilar.

Efter att ha utfört en noggrann analys av flera modeller har jag valt att fokusera på modell lm\_6 för vidare undersökning och prediktioner. Denna modell, tränad på ett välvalt variabelset och utvecklad med hjälp av stepwise backward elimination, visade sig vara robust och effektiv med en hög justerad och en låg residual standard error. Dessutom innehåller den endast 24 variabler, vilket minskar risken för överanpassning och ökar dess generaliseringsförmåga.

Vid utvärdering av modellen fann jag att dess beräknade Root Mean Square Error (RMSE) var 50837,4, vilket motsvarar 20% av det genomsnittliga priset i test-datasetet. Denna nivå av precision anses vara tillfredsställande och ligger i linje med projektets mål.

Som en praktisk tillämpning testade jag modellen genom att förutsäga priset för mitt personliga bilinköp, vilket visade sig vara i linje med de förutsägda konfidens- och prediktionsintervallen.

Slutligen, medveten om modellens begränsningar, särskilt dess fokus på "vanliga familjebilar" och insamlade data från specifika regioner i Sverige, är det viktigt att förstå att dess användning kan vara mer begränsad utanför detta prisspann och geografiska område. Trots detta utgör modellen ett värdefullt verktyg för prissättning av begagnade bilar inom sitt tillämpningsområde.

Sammanfattningsvis har detta projekt framgångsrikt utvecklat och utvärderat en regressionsmodell för prissättning av begagnade bilar, vilket ger insikter och verktyg för att underlätta rättvisa och ekonomiskt fördelaktiga affärer för både säljare och köpare.

# 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.

Svar: Man placerar observationerna längs y-axeln baserat på deras kvantil i det givna urvalet och längs x-axeln baserat på deras förväntade kvantil i standard normalfördelning. QQ-plotten visar hur väl de två fördelningarna överensstämmer. Ju närmare punkterna ligger längs diagonal linje som startar i punkt (0,0), desto närmare överensstämmer fördelningarna.

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?

Svar: Ja, du hörde rätt. I maskininlärning ligger huvudfokus ofta på att utveckla modeller som kan göra prediktioner med hög noggrannhet. I statistisk regressionsanalys är målet inte bara att göra prediktioner utan också att förstå och dra slutsatser om relationen mellan variabler i data.  
Till exempel, om vi undersöker sambandet mellan kön, ålder och inkomst, kan en maskininlärningsmodell användas för att göra en prediktion om en persons inkomst baserat på deras kön och ålder med hög noggrannhet. Å andra sidan kan en regressionsanalys utföra statistiska inferenser för att fastställa om ålder och kön har en signifikant påverkan på inkomstnivån i populationen och för att uppskatta hur mycket inkomsten förväntas öka eller minska för varje års ökning i ålder.

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

Svar: Konfidensintervall representerar ett intervall av värden som med X% sannolikhet innehåller det sanna medelvärdet för en responsvariabel baserat på specifika värden på en eller flera förklarande variabler. Prediktionsintervall representerar ett intervall av värden som med X% sannolikhet innehåller det sanna värdet för en responsvariabel för en enskild ny observation baserat på specifika värden på en eller flera förklarande variabler. Prediktionsintervall är alltid bredare än konfidensintervall, eftersom de inkluderar både felet i uppskattningen för f(X) och osäkerheten kring hur mycket en enskild punkt kommer att skilja sig från populationens regressionsplan.

4. Den multipla linjära regressionsmodellen kan skrivas som: 𝑌 = 𝛽0 + 𝛽1𝑥1 + 𝛽1𝑥2+ . . . + 𝛽𝑝𝑥𝑝 + 𝜀 . Hur tolkas beta parametrarna?

Svar: Beta-parametrarna i en multipel linjär regressionsmodell representerar de förväntade förändringarna i Y för varje enhets förändring i de förklarande variablerna X, medan alla andra variabler hålls konstanta.

β0 är det förväntade värdet på Y när alla X är noll.

β1, β2, ..., βp representerar de förväntade förändringarna i Y för varje enhets förändring i de respektive Xn, medan alla andra X hålls konstanta.

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?

Svar: Ja, det stämmer. Logiken bakom BIC och likdanade metoder är att de gör en justering av träningsfelet för att ta hänsyn till den överanpassning som kan finnas i modellen. Genom att inkludera en justering för antalet parametrar eller komplexiteten hos modellen, straffar dessa mått överdriven komplexitet och föredrar enklare modeller som ändå presterar väl. Därför behöver vi nödvändigtvis inte dela ut dataset för att välja den bästa modellen.

6. Förklara algoritmen nedan för ”Best subset selection”

En bild som visar text, skärmbild, Teckensnitt, nummer

Automatiskt genererad beskrivning

Svar: 1. Man beräknar urvals medelvärde – M0.

2. a)För varje antal förklarande variabler (X) från 1 till det totala antalet p, skapa alla möjliga kombinationer av X1…p. Till exempel, om vi har 5 förklarande variabler totalt, kommer vi att skapa modeller med alla möjliga kombinationer när vi har bara 1, 2, 3, 4 och 5 förklarande variabler.

b)För varje antal förklarande variabler väljer vi den bästa modellen. Dvs bland alla modeller som består av 1 förklarande variabel väljs en modell - M1, bland alla modeller som består av 2 förklarande variabler väljs den bästa modellen - M2 osv. Detta görs med hjälp av RSS och R^2.

3. Bland alla våra modeller M0, M1 osv Mn väljer man den bästa modellen med hjälp av validationset, cross-validation eller AIC/BIC/R-adj-2.

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

Svar: Jag tror att det som menas är att en modell aldrig kommer att vara exakt som verkligheten. Modell är alltid en förenkling av verkligheten och det blir aldrig 100% rätt. Däremot även sådana ”felaktiga” modeller man vara användbara för sina ändamål.

# Självutvärdering

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

Jag trodde att R Studio var jobbig så jag började skriva projektet i Jupiter Notebook. Det var omöjligt att installera olika bibliotek från Notebooken, så att man behövde göra det genom terminal i Anaconda eller länt till CRAN eller även ladda ner binaries och installera från zip-fil. Sedan var det många konflikter mellan packet och det resulterade i att när jag var helt klar med projektet och skulle börja skriva min rapport gick det inte att köra koden i Jupiter Notebook för att den flaggade som inga av bibliotek var installerade trots att de var det. Så, jag orkade inte sitta med detta längre och bara kopierade koden till R Studio.

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

Jag bryr mig inte, men jag tror att VG för att jag har gjort alla VG-bitar.

1. Något du vill lyfta fram till Antonio?

Vore bättre med flera mindre uppgifter än en gigantisk kunskapskontroll.

# Grupparbete

1. Vem du har arbetat i grupp med?

Shriya Walia, Filip Östlund, Muhammad Mahmudur Rahman, Isabella Frid, Shangchanhui Feng, Wissam Rateb, Turzo Khan.

1. Hur har ni i gruppen arbetat tillsammans?

Shriya Walia, Filip Östlund, Muhammad Mahmudur Rahman, Isabella Frid, Shangchanhui Feng och jag samlade data enligt mall som vi kom överens om i förväg. Vi satt tydliga mål och deadline. VI har också kommit överens om vilken data vi samlar in och var och en av oss samlade data från ett län som denna person har valt.

Wissam Rateb och Turzo Khan kom senare när vi var redan klara med datainsamlingen.

Vi har diskuterad i gruppen också olika aspekter av kunskapskontroll, t.ex. API.

1. Vad var bra i grupparbetet och vad kan utvecklas?

Det var bra att alla var klara vid deadline och att vi diskuterade mycket olika grejer.

Det var däremot oklart vilket språk som vi använder i teamet: kommunikation har skett på svengelska vilket var inte skönt och skapade förvirring (i alla fall för mig). Vi satt inte heller tydliga riktlinjer för t.ex. hur vi kännetecknar saknade värden (NA, NULL, 0).

1. Vad är dina styrkor och utvecklingsmöjligheter när du arbetar i grupp?

Jag tror att jag är bra på att se utvecklingsmöjligheter för grupparbete :D och att sätta tydliga mål, riktlinjer och hitta på ett optimalt arbetssätt. Jag tycker om att lyssna på andras åsikter och lära mig hur andra arbetar, så att jag kan utvecklas i min roll.

1. Finns det något du hade gjort annorlunda? Vad i sådana fall?

Jag skulle nog vara mer aktiv och bestämd när det gäller diskussioner om gemensamt arbetssätt och prata även om små saker (som saknade värden) högt.

# Källförteckning

Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow (Second Edition)*. Sebastopol, CA: O’Reilly Media, Inc.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2023). An introduction to statistical learning: with Applications in R. Springer.

Transportstyrelsen. (n.d.). Körkort. Hämtad från https://www.transportstyrelsen.se/sv/vagtrafik/statistik/korkort/Statistik-over-korkortsinnehavare-efter-kon/