Teoretiska frågor

Gruppuppgiften.

1. Vem du har arbetat i grupp med?  
   Erik Fredrixon, Filip Holmberg, Ian Gutierrez,  
   Magdalena Wallner och Robert Westman.
2. Hur har ni i gruppen arbetat tillsammans?  
   Gruppen fördelade arbetet av datainsamling likvärdigt där vi valde ett bilmärke var och samlade in 50 observationer till en början, efteråt gick vi igenom hur web-scraping fungerade och samlade därefter ihop ytterligare 200 observationer vardera. Vi hade regelbundna samtal och diskutioner för att hjälpa varandra och observera de andras resultat.
3. Vad var bra i grupparbetet och vad kan utvecklas?  
   Vår kommunikation fungerade bra men kan alltid utvecklas för det bättre.
4. Vad är dina styrkor och utvecklingsmöjligheter när du arbetar i grupp?  
   Mina styrkor i grupparbete är att jag gillar struktur samt att jag gärna för samtalet framåt så att frågor kan ställas och att ingen känner sig utanför. Jag ser att allt är en utvecklingsmöjlighet men framför allt att mina sociala kunskaper kan utvecklas.
5. Finns det något du hade gjort annorlunda? Vad i sådana fall?  
   Jag hade kunnat gett förslaget om fler samtalstider och bett om mer hjälp med   
   R-programmeringen.

## Teoretiska frågor

1. beskriv kortfattat vad en Quantile-Quantile (QQ) plot är.  
QQ-plot visar om data är normalfördelad genom att plotta observationerna från stickprovet mot dem teoretiska kvartilerna, om linjen i QQ-plotten är någorlunda rak är den data normalfördelad.

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?  
-Maskininlärning har mer fokus på prediktioner, men utesluter inte statistisk regressionsanalytiskt arbete eller statistisk inferens. I maskininlärning kan vi träna modeller att känna igen objekt, medan i statistisk regressionsanalys kan vi undersöka vilka oberoende variabler som påverkar värdet av bostäder.

3. Vad är skillnaden på ”konfidensintervall” och ”prediktionsintervall” för predikterade värden?  
-konfidensintervall används för att uppskatta osäkerheten i en populations parameter, medan prediktionsintervall används för att uppskatta osäkerheten i enskilda eller framtida observationer från samma population eller modell.

4. Den multipla linjära regressionsmodellen kan skrivas som:

𝑌 = 𝛽0 + 𝛽1𝑥1 + 𝛽2𝑥2+ . . . + 𝛽𝑝𝑥𝑝 + 𝜀 .

Hur tolkas beta parametrarna?  
-Beta parametrarna visar hur mycket dess tillhörande variabel X påverkar Y när X ökar med en enhet, alla andra variabler är konstanta. Detta kallas partialderivatan.

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?  
-Det stämmer att man inte behöver fördela data i träning, validering och test set då BIC estimerar test felet direkt från träningsdata.

6. Förklara algoritmen nedan för ”Best subset selection”  
A screenshot of a test

Description automatically generated  
(diskutera med gruppen)

- Logiken bakom best subset selection är att prova alla möjliga kombinationer av prediktorer och sedan välja den modell som har det minsta prediktionsfelet eller största förklaringsvärdet. I det första steget börjar man med noll-modellen som helt saknar prediktorer. Sedan testar vi att lägga till en prediktor till modellen (alla prediktorer prövas) och väljer den modell som har det minsta felet. Vi fortsätter på samma sätt med 2 prediktorer (väljer den kombination av 2 prediktorer som ger minst fel) osv tills vi har en modell som består av alla prediktorer. Nu har vi lika många modeller som prediktorer samt noll-modellen. Nu väljer vi den modell som har minst fel/högst förklaringsvärde.

7. Ett citat från statistikern George Box är: “All models are wrong, some are useful.”

Förklara vad som menas med det citatet.  
-Det är högst osannolikt att en modell kan prediktera 100% korrekt men om den kan prediktera 95% korrekt så är den användbar.

## Självutvärdering

1. Utmaningar du haft under arbetet samt hur du hanterat dem.   
   Denna kurs har haft många utmaningar, största utmaningen för mig var Själva R-Programmeringen, det fanns alltid något fel i min kod så det slutade alltid med att jag fick be om extra hjälp från antingen gruppen eller från min mentor om jag själv inte hittade lösningen.
2. Vilket betyg du anser att du skall ha och varför.   
   Jag anser att då jag inte lyckats få APIn att fungera och kunnat använda dess data skall jag betygsättas med ett G+
3. Något du vill lyfta fram till Antonio?

En väldigt intressant kurs för mig, jag hade önskat lite mer live kodning med dig på lektioner när ett nytt programmeringsspråk introduceras.