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.
Det är en graf som jämför fördelningen av vår data mot en normalfördelning. När punkterna ligger efter en diagonal linje så är vår data normalfördelad men om punkterna avviker från linjen så kan vi behöva transformera datan då våra modeller förväntar sig att datan den tränas på är 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?
Ja det stämmer. Maskininlärning handlar om att bygga prediktiva modeller som ger oss svar när vi stoppar in ny data. Statistisk regressionsanalys handlar om att skatta relationer mellan en beroende variabel och en eller fler oberoende variabler och med det skapa prediktiva modeller och dra slutsatser. Exempelvis så kan vi undersöka sambandet mellan ålder och lön och dra slutsatser om hur starkt sambandet är.

3. Vad är skillnaden på ”konfidensintervall” och ”prediktionsintervall” för predikterade värden?
Konfidenseintervall handlar om snittet för en viss grupp medans prediktionsintervall handlar om en individ och inkluderar osäkerheten I detta. Det betyder att prediktionsintervallet är alltid större än konfidensintervallet.

4. Den multipla linjära regressionsmodellen kan skrivas som:
$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + \epsilon$
Hur tolkas beta parametrarna?
β_p för p≥1 tolkas som: ”Vad är effekten på Y när x_p ökar med en enhet, givet att alla andra variabler är fixa.”

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?
Ja BIC är ett sätt att skatta testfeltet genom att lägga till ett straff på träningsfelet för att kompensera för den bias som uppstår.

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.

Best subset selection skapar en modell för varje enskild oberoende variabel och kombinationer av oberoende variabler och jämför dem med varandra för att hitta den bästa modellen. Bästa innebär den modell med lägst fel.

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

Ingen modell kan fånga all komplexitet I verkliga världen men dem kan ändå vara användbara. Alltså vi behöver inte förklara alla aspekter för att kunna ge ett bra estimat av verkligheten.

Självutvärdering

1. **Utmaningar du haft under arbetet samt hur du hanterat dem.**
   Jag har så svårt med att ta mig tid för att skriva en sån här rapport. Jag försökte lägga upp min tid bättre än förra kursen men det blev bara lite bättre. Ska inför nästa rapport aggera som om det är inlämning samma vecka som vi får uppgiften för att förhoppningsvis få bukt på mig själv.

2. **Vilket betyg du anser att du skall ha och varför.**
   G, jag känner inte att jag gjort denna uppgift med säkerhet. Har känt mig ganska lost I arbetsflödet för annalysen.

3. **Något du vill lyfta fram till Antonio?**