# Teoretiska frågor

1. Beskriv kortfattat vad en Quantile-Quantile (QQ) plot är.

En quantile-quantile plot är ett diagram som kan användas för att undersöka huruvida residualerna är normalfördelade eller ej, vilket är ett antagande vi gör vid framtagande av konfidens- och prediktionsintervall samt hypotesprövning. Diagrammet jämför den teoretiska fördelningen (spridningen) av ett data set om det vore normalfördelat, med den verkliga spridningen av datan. Om den verkliga datan är normalfördelat, bör den följa QQ-plot kan därför användas för att visualisera fördelningen av datan vi undersöker.

1. 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?

Prediktioner handlar kortfattat om att göra säkra ”gissningar” på ny data utifrån den modell vi tränat. T ex, Om en person är 55 år gammal och har eftergymnasial utbildning predikterar vi att hen har 45 000 kr i lön.

Statistisk inferens kan förklara sambanden, t ex med hur mycket ökar lönen om man vidareutbildar sig? Vilka faktorer är avgörande för att öka lönen? Kanske är inte ålder eller utbildning signifikant över huvud taget. Finns det ett linjärt samband mellan valda variabler (ålder och utbildning) eller är sambandet mer komplicerat?

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

I exemplet i svaret på fråga 2 skulle ett konfidensintervallet skapa ett intervall på den genomsnittliga lönen för en viss ålder och en viss utbildningsnivå. Prediktionsintervallet inkluderar feltermen epsilon, och är istället ett intervall för en vad en specifik individ har vid en viss ålder och utbildningsnivå. Eftersom prediktionsintervallet även inkluderar epsilon är det bredare än konfidensintervallet.

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

B0 är interceptet, dvs värdet på Y när X = 0. B1 är den genomsnittliga ökningen eller minskningen i Y när X1 ökar ett ”steg” och alla andra X är oförändrade. B2 är den genomsnittliga ökningen eller minskningen i Y när X2 ökar ett ”steg” och alla andra X är oförändrade, osv för alla B upp till Bp.

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 det stämmer. BIC, Cp och AIC är alla mått som kan användas för att uppskatta ”test error” från ”training error”, utan att använda sig av validerings- och testset. Målet är ett så litet värde av samtliga mått. Även Adjusted R2 kan användas för att uppskatta hur väl en modell presterar, men då vill vi ha ett så högt mått som möjligt.

6. Förklara algoritmen nedan för ”Best subset selection”

1. M0 utgör startpunkten för metoden. Den innehåller inte några prediktorer.

2. För p prediktorer och där k = 1, 2, 3 … p tränar man modellen med alla möjliga kombinationer av predikatorer för k prediktorer. T ex om vi har p=5 prediktorer, tränar man alla möjliga modeller som innehåller k =1 prediktor, sen alla möjliga modeller för k = 2 prediktorer, sedan k = 3 osv. Den bästa modellen, det vill säga den modell med högst R2 eller lägst RSS, för varje k väljs ut och kallas Mk.

Det vill säga, för en prediktor väljs bästa M1 ut, för två prediktorer väljs bästa M2 ut osv upp till Mp.

3. Den bästa modellen av M0, M1, M2…. Mp väljs ut genom användningen av exempelvis prediction error på ett validerings set alternativt Cp, BIC, eller Adjusted R2- värdet eller genom cross-validation.

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

Med det menas att våra antaganden om linjär regression är i praktiken aldrig uppfyllda. Vi kommer aldrig kunna nå perfekta modeller, med helt uppfyllda antaganden, men de kan likväl vara användbara. Vi får dock vara försiktiga med hur vi förhåller oss till dem, då vi aldrig kan påstå att de är 100 % sanna.

# Självutvärdering

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

Den största utmaningen under arbetet har varit tidsåtgången till datainsamlingen, databearbetningen och arbetsprocessen för att träna en så användbar modell som möjligt. Det har känts stressigt att hinna med för att få så mycket tid som möjligt med modelleringen av linjära modeller, men det har varit ett gott samarbete med gruppen som har underlättat.

1. Vilket betyg du anser att du skall ha och varför.   
   G, då jag anser att jag uppfyllt de kriterier.
2. Något du vill lyfta fram till Antonio?

**Datainsamling**

1. Vem du har arbetat i grupp med?

Mustafa, Parsan, Ming och (Mohamed)

1. Hur har ni i gruppen arbetat tillsammans?

Redan under första mötet hade Mustafa redan börjat titta på möjligheten med webscraping. Ett försök till webscraping hade redan gjorts från hans håll men misslyckats. Vi kom därför överens om som grupp att han skulle göra ett nytt försök med de metodval som vi kom överens om under gruppmötet, vilket följer: bilar som säljs i Västra Götaland. Inga andra filtreringar gjordes initialt då vi alla ville ha möjligheten att titta igenom datan och göra justeringar i Excel i efterhand. Vi hade dock långa diskussioner och övervägningar fram och tillbaka innan vi kom fram till detta, där alla var aktiva med inputs och synpunkter.

Dagen efter första mötet hade Mustafa lyckats med webscraping, varpå jag och Mustafa hjälptes åt att städa datan då överflödig information kom med i kolumnerna. Ming och Parsan var delaktiga under detta moment genom delad skärm så alla kunde komma med synpunkter på hur datan skulle ”städas upp”.

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

Alla deltog aktivt i gruppdiskussionerna och var flexibla kring när gruppmöten kunde hållas.

Till nästa gång kan vi vara mer strukturerade med hur vi fördelar arbetsuppgifter inom gruppen.

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

Styrkor: Inte rädd för att göra min röst hörd, lyssnar gärna in och försöker få alla delaktiga

Utvecklingsmöjligheter: Blir lätt otålig om man fastnar och vill komma med lösningar (är inte nödvändigtvis de bästa lösningarna, ibland är det bättre att tänka långsamt)

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

Strukturerat upp fördelningen av arbetsuppgifter bättre