# Teoretiska frågor

1. Kort beskrivning av vad Quantile-Quantile (QQ) är.

Genom att grafiskt jämföra ett antal observationer mot en teoretisk fördelning, tex normalfördelningen, kan man svara på om observationerna följer den teoretiska fördelningen.

På x-axeln plottas kvantiler från den teoretiska fördelningen.

På y-axeln plottas kvantiler från observationerna.

Om punkterna ligger i en ungefär rak linje indikerar detta att observationerna följer den teoretiska fördelningen.

A graph of a line

Description automatically generated with medium confidence

1. Inom maskininlärning är fokus att utveckla modeller (algoritmer som automatiskt lär sig mönster och samband från data) som gör korrekta prediktioner eller tar beslut. Modellen kan vara avancerad och svår att tolka, men så länge dess prediktioner är korrekta finns inget behov av att förstå det bakomliggande.

Inom statistisk regressionsanalys används modeller för att både göra prediktioner och för att förstå de underliggande sambanden mellan variabler och hur mycket en variabel påverkar utfallet (om alls). Modellen behöver således vara enklare för att vi ska kunna tolka den.

Exempel på prediktion: Förutsäga huspriser.

Exempel på statistik inferens: Förstå hur huspriser påverkas av olika parametrar, tex boyta, geografisk plats och utsikt.

1. Konfidensintervall: Ett intervall inom vilket det okända, men fasta, sanna medelvärdet av den beroende variabeln (y) för en given uppsättning av oberoende variabler (x) förväntas finnas med en viss konfidensgrad.

Prediktionsintervall: Ett intervall inom vilket en ny observation av den beroende variabeln (y) för en given uppsättning av oberoende variabler (x) förväntas ligga med en viss konfidensgrad.

Prediktionsintervall är bredare än konfidensintervall eftersom enskilda observationer kan avvika mer än ett medelvärde.

1. 

β-parametrarna beskriver förhållandet mellan den beroende variabeln (y) och respektive oberoende variabel (x).

β0 kallas intercept och är det förväntade värdet på den beroende variabeln (y) när alla oberoende variabler (x) är lika med noll.

β1- βp kallas regressionskoefficienter och är påverkan från de oberoende variablerna (x) på den beroende variabeln (y), alltså förändringen i den beroende variabeln (y) när en oberoende variabel (x) ökar med en enhet, medan övriga hålls konstanta (varje β är justerat för effekterna av de andra oberoende variablerna i modellen). Ett stort respektive litet värde betyder att den oberoende variabeln (x) har en stor respektive liten påverkan på den beroende variabeln (y). Ett positivt respektive negativt värde betyder att en ökning av den oberoende variabeln (x) ökar respektive minskar värdet på den beroende variabeln (y).

1. Ja, det stämmer.

BIC är ett mått som kan hjälpa oss att utvärdera och jämföra modeller och alltså välja den bästa bland ett antal modeller med olika antal variabler. BIC består av två termer, en som belönar modeller som har en hög anpassning till data och en som straffar modeller som har en hög komplexitet, alltså många parametrar. Ett lägre värde på BIC indikerar att det är en bättre modell. I stället för att reservera en del av data för utvärdering, vilket minskar mängden data tillgänglig för att träna modellen, används hela datasetet för att skatta parametrarna och beräkna BIC (och liknande mått). BIC är särskilt användbart när vi har flera modeller och vill undvika överanpassning.

BIC används när man inte kan eller vill dela upp datasetet i tränings-, validerings-, och testdata, vilket kan vara praktiskt tex när ett dataset är litet.

BIC mäter inte hur väl en modell kan göra en prediktion. Om målet med modellen är korrekta prediktioner snarare än tolka förhållanden mellan variabler bör andra metoder användas.

1. Förklaring av algoritm för ”Best subset selection”.

Ta fram 2 upphöjt till p modeller som innehåller alla olika kombinationer av prediktorer.

1. Skapa en nullmodell, alltså en modell som inte innehåller några prediktorer. Medelvärdet från alla observationer används som prediktion för varje enskild observation.
2. Skapa 2pantal modeller som innehåller alla möjliga kombinationer av prediktorer. p är antalet prediktorer i datasetet.

För varje delmängd av modeller som innehåller lika många prediktorer väljs den bästa baserat på lägst RSS eller störst R2. Detta reducerar antalet modeller till p+1.

1. Välj den slutliga modellen som ska användas genom prediktionsfel på ett valideringsset, mått såsom AIC, BIC samt justerat R2 eller korsvalidering. RSS samt R2 skulle alltid ge den modell med flest variabler, varför de inte är bra mått att använda.
2. ”*All models are wrong, some are useful*”

Det går inte att skapa en modell som helt kan representera den komplexa och ofta slumpmässiga verkligheten. Modeller kan dock vara tillräckligt bra för att ge förståelse kring viktiga samband i data samt göra användbara prediktioner. En modell ska återge väsentliga egenskaper utan att kopiera alla detaljer. Det viktiga är att en modell är användbar för det aktuella syftet.