Datainsamling (grupparbete)

1. Vem du har arbetat i grupp med?
   Magdalena W, Ian G, Eric F, David N och Robert W.
2. Hur har ni i gruppen arbetat tillsammans?
   Vi har strukturerat upp uppgiften så att alla förstod vad den innebar. Arbetsuppgifter som POC, webskrapning och manuell insamling av data delades upp.
3. Vad var bra i grupparbetet och vad kan utvecklas?
   Det var bra kommunikation under vägens gång och alla har varit engagerade. Dessutom var det ett bra klimat där alla pratade och hjälpte till. Det som kan utvecklas vidare var att återsamlas och sumera när arbetsuppgifter är klara. Det blev mer ostrukturerat ju längre vi kom i grupparbetet.
4. Vad är dina styrkor och utvecklingsmöjligheter när du arbetar i grupp?
   Mina styrkor är att ha kontakt med dom andra och skapa en bra klimat. Mina möjligheter att utvecklas handlar om att veta när jag ska ta mindre plats och låta andra tala. Det är ju trots allt från andra som jag kan lära mig mer.
5. Finns det något du hade gjort annorlunda? Vad i sådana fall?
   Det hade varit bra att sätta återsamlingspunkter (tid eller när aktivitet är avslutad) för att samla gruppen.

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: En QQ plot visar värden plottade mot kvantiler (ofta beräknade från normalfördelad distribution). Om resulterande samband är linjär, är datan 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?

Svar:

- **Maskininlärning: fokuserar på att hitta bästa generaliseringsförmågan till ny osedd data. Alltså att göra bästa prediktionen för ny osedd data. T ex prediktera hur mycket ett hus kommer att säljas för.**
- **Statistisk regressionsanalys: fokuserar på att hitta bästa passning till data med statistik. Här finns tydligare koppling mellan sannolikhet (t ex 95 % säkerhet) och prediktion. T ex förklarar även hur variabler påverkar huspriset (inferens), se koefficienterna. Eller intervall vari det sanna priset kommer ligga (prediktion).**

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

Svar: Prediktionsintervallet alltid större än konfidensintervallet, pga extra osäkerhetstermen epsilon. Exempel löner: Konfidensintervallet är snittlönen vid viss ålder och Prediktionsintervall lönen för en individ.

4. Den multipla linjära regressionsmodellen kan skrivas som: Y = $\beta 0 + \beta 1 x 1 + \beta 2 x 2 + ... + \beta p x p + \varepsilon$. Hur tolkas beta parametrarna?

Svar:

- **β0: intercept till funktionen, vad Y är när alla X_i är 0.**
- **β_i: Koefficienterna, effekten på Y av X_i när X_i ökar med en enhet, givet att alla andra variabler betraktas som konstanta (dvs partialderivata).**

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: Man kan estimera ”test error” från ”training error” mha t ex BIC, därför kan man utvärdera modellens ”generalization error” utifrån träningsdatan. Dela upp data i träning, val och test behövs därför inte. BIC tenderar att anta små värden för modeller med låg ”test error”.

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.

Svar: Skapa först en modell med 0 prediktorer. Sedan söker vi vilken modell som är bäst (lägst RSS eller högst R^2) om vi bara har en prediktor. Därefter söker man vilken modell som är bäst om vi ska ha två prediktorer. Osv… När vi är klara väljer vi bästa modellen (som vi redan känner till) utifrån det önskade antalet prediktorer.

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

Svar: definitionen av en modell är att den är en ”icke exakt beskrivning” av verkligheten. Per definition är alla modeller alltså mer/mindre felaktiga. Däremot kan modeller användas, mer eller mindre framgångsrikt, för att beskriva verkligheten.

Självutvärdering

1. Utmaningar du haft under arbetet samt hur du hanterat dem.
   Det har varit utmanande att arbeta med statistik utan att ha ett facit om vad som är rätt. Jag har envist försökt leta svar, och i många fall hittat dom. Det har varit lärorikt.

2. Vilket betyg du anser att du skall ha och varför.
   VG – med hänvisning till kurskriterierna.

3. Något du vill lyfta fram till Antonio?
   Statistisk Dataanalys (2:a upplagan), s 382.
   Tack för en utmanande kurs!