KU LEUVEN kulak

Voorbeeld examenvragen

Statistische modellen en data-analyse

---

1. De dataset `Europa.txt` bevat het percentage tewerkgestelden in de verschillende industrieën in Europese landen in 1979. In totaal zijn er 9 categorieën (variabelen). Een belangrijke opmerking is dat de data dateren van tijdens de Koude Oorlog. Bekijk goed welke de 9 categorieën zijn die hier beschouwd worden.

   (a) Construeer een zo eenvoudig mogelijke en duidelijke voorstelling van deze gegevens waarbij geen belangrijke informatie verloren mag gaan. Motiveer je keuzes.

   (b) Onderzoek op basis van deze voorstelling of multivariate normaliteit voor deze gegevens kan verondersteld worden.

2. De data `Pollution.txt` bevat gegevens van 9 variabelen die men opgemeten heeft in 60 SMSA's (Standard Metropolitan Statistical Area: gebieden die allen aan bepaalde voorwaarden voldoen). De gemeten variabelen zijn:

   - `JULT`: de gemiddelde temperatuur in juli (in graden Fahrenheit)
   - `SO`: de relatieve pollutiecapaciteit van sulfurdioxide, SO_2
   - `HUMID`: het percentage relatieve vochtigheid, jaarlijks gemiddelde om 13 uur
   - `MORT`: het totale leeftijdsaangepaste sterftecijfer (het aantal sterfgevallen per 100 000)

   Bouw een regressiemodel dat het gemiddelde totale sterftecijfer zo goed mogelijk verklaard in functie van de gemiddelde temperatuur in juli, de relatieve pollutiecapaciteit van sulfurdioxide en het percentage relatieve vochtigheid.

   (a) Rapporteer uw finale model en leg uit hoe je te werk gegaan bent om tot dit model te komen.

   (b) Interpreteer zorgvuldig elk van de geschatte regressiecoëfficiënten in dit model en geef weer hoe goed dit regressiemodel is.

   (c) Als het model nog tekorten vertoont, geef dan aan hoe de analyse nog verder verbeterd zou kunnen worden.

3. We beschouwen de dataset `visvangst.txt` waarin 158 vissen van 7 verschillende soorten gevangen en gemeten werden. Alle vissen werden gevangen in een meer in de buurt van Tampere in Finland. De 7 verschillende soorten werden nadien ook nog ruwer geklasseerd door hen te groeperen in 2 categorieën die bestaan uit soorten die nauw verwant zijn met elkaar. De volgende variabelen werden gemeten:

   - `soort`: verschillende soorten vis genummerd van 1 t.e.m. 7
   - `gewicht`: gewicht van de vis (in gram)
   - `lengte`: lengte van de neus tot het einde van de staart (in cm)
   - `categorie`: indeling van de vissen in categorie 1 (= 0) en categorie 2 (= 1)

   (a) Ga na of er verschillen bestaan tussen het gemiddelde gewicht van de 7 verschillende soorten vis en geef dan aan welke soorten een verschillend gewicht vertonen.

   (b) Stel een model op dat toelaat om vissen in te delen in de twee mogelijke categorieën. Hoe goed slaagt dit model erin om beide categorieën te onderscheiden?

4. De dataset `bakery.txt` bevat voedingsinformatie van 25 producten van "Brumby's bakery" in Bundaberg in november 2003. Meer precies zijn de volgende variabelen gekend (telkens per 100g van het product): de energie in kJ (`Energy`), de proteïnen in g (`Protein`), de vetten in g (`Fat`), de totale carbohydraten in g (`Total.Carbo`), de suikers in g (`Sugars`), de vezels in g (`Fiber`), het sodiumgehalte in mg (`Sodium`) en het potassiumgehalte in mg (`Potassium`).

(a) Tracht de producten op te delen in groepen van vergelijkbare producten op basis van deze variabelen.

(b) Geef de gekozen opdeling. Geef aan hoe je tot dit resultaat gekomen bent en motiveer waarom je dit resultaat verkiest.