Statistik 2 2008, 1.4.08

Övning 1a

Hämta följande datafil: http://www.abo.fi/fak/mnf/mate/jc/statistik2/sedlar.txt
Filen innehåller sammanlagt 200 observationer på 6 olika typer av mätvärden för äkta och förfalskade sedlar. Den första variabeln indikerar om en sedel är äkta (värdet = 1) eller förfalskad (värdet = 2). Starta R och sedan R commander med kommandot:
library('Rcmdr')
Data kan importeras via menyn Data-Import...-from text file. Definiera den första variabeln som en faktor via Data-Manage...-Convert...to factor. Utför sedan en principalkomponentanalys (PCA) med variablerna V2-V7 via Statistics-Dimensional...-PCA. Kryssa inte för Analyze correlation matrix, men nog för Scree plot och Add components to data. Efter analysen visas egenvektorerna i output-fönstret. För att studera effekten av varierande mätningsskala, transformera variablerna V2,V3,V4,V7 genom att dividera deras värden med 10 (Data-Manage-Compute new variable). Upprepa sedan principalkomponentanalysen och jämför de nya värden på egenvektorerna med de tidigare erhållna värden. Gör ett spridningsdiagram (Graphics-Scatterplot matrix) av varje par av de tre första principalkomponenterna. Använd sedelkategorin (äkta/falsk) som gruppering i diagrammen. Hur urskiljs de två grupperna?

Övning 1b

Hämta följande datafil: http://www.abo.fi/fak/mnf/mate/jc/statistik2/geneticdata.txt
Filen innehåller genetiska profiler för 97 individer som är samplade från fyra olika geografiska områden. Den första kolumnen indikerar området (konvertera den till en faktor som ovan). Var och en av de övriga kolumnerna (85 stycken) indikerar om individen bär en viss mutation i arvsmassan. Utför en principalkomponentanalys som ovan och gör ett spridningsdiagram med de relevanta komponenterna (använd området som gruppering). Finns det skillnader mellan områden?

Övning 1c

Hämta följande datafil: http://www.abo.fi/fak/mnf/mate/jc/statistik2/uscrime.dat
Filen innehåller uppgifter om brottsfrekvenser gällande olika typer av brott (V3-V9) i delstaterna i USA. De två första variablerna indikerar delstatens areal respektive populationsstorlek. De två sista variablerna indikerar regioner (för V10 enligt följande: Northeast = 1, Midwest = 2, South = 3, West = 4). Gör en principalkomponentanalys med och utan den första variabeln. Blir det några skillnader? Kan geografiska regioner urskiljas med hjälp av principalkomponenterna?