Statistik 2, del 4:

Bivariata korrelationer kontra partiella korrelationer

Korrelation och partiell korrelation beskrivs kort här:

http://en.wikipedia.org/wiki/Correlation

http://en.wikipedia.org/wiki/Partial_correlation

Vi undersöker nu liknande beroendemönster i data som i modulen med interaktioner och korstabeller, men denna gång utifrån kontinuerliga data. Läs in filen Mattedata.sav. Filen innehåller kurspoäng i olika matematiska delområden för ett antal studenter. Skapa först bivariata korrelationer och matris spridningsdiagram för att se om variablerna verkar ha samband med varandra. Slutsatser?

Beräkna nu partiella korrelationer och jämför resultaten med de som erhölls tidigare. Slutsatser?

Den mest lämpliga analysen för dessa data kallas för covariance selection modeling som utgår från kontinuerliga mätvärden, men vi gör en approximation och matar in datat i B-Course (Mattedata.dat) som diskretiserar värden före modellval och –sökning. Hur blir slutsatserna här jämfört med parvisa och partiella korrelationer?

Interaktionsmodeller för korstabelldata

Då man observerar flera kategoriska variabler samtidigt, har man ofta intresse för vilka av variablerna verkar ha samband med varandra. Under kursen Statistik 1 demonstrerades vilka problem kan uppstå om man enbart analyserar variablerna parvist (t ex med 2-dimensionella korstabeller) och testar för beroenden. Exempelvis kan falska samband komma fram (Simpson's paradox) eller man kan även missa samband. En allmän klass av modeller för dylika data kallas log-linjära interaktionsmodeller. Namnet härstammar från en logaritmisk utveckling av sannolikhet för varje cell i den flerdimensionella korstabellen, där interaktionstermerna återspeglar samband mellan variabler. Ifall det inte finns någon interaktionsterm i modellen där variablerna A och B ingår, hävdar modellen att A och B är oberoende eller betingat oberoende av varandra. Ett visuellt sätt att betrakta interaktionsmodellerna är att rita upp en graf där variablerna är noder och dessa förbinds med länkar enbart då en interaktionsterm eller fler omfattar dem samtidigt. Graferna är ofta även riktade och då fokuserar man på betingade fördelningar givet de noder som är ens föräldrar. En introduktion till diverse grafiska modeller hittar man här:

http://www.cs.ubc.ca/~murphyk/Bayes/bnintro.html

Ett gratis onlineprogram som erbjuder möjligheten att lära interaktionsmodeller från data finns på:

http://b-course.cs.helsinki.fi/obc/

Ett annat gratisprogram är MIM som hittas på

http://www.hypergraph.dk/

Även SPSS erbjuder möjligheter till analys av interaktionsmodeller, men dessa är tyvärr mer begränsade, speciellt vad gäller presentation av modellerna och valet av lämpliga modeller med hjälp av sökalgoritmer.

Vi börjar med SPSS genom att analysera 2 x 2 tabellen i filen Murder2D.sav. Ett eventuellt samband kan testas med Chi^2-testet. Datat i tabellen är en sammanslagning av 3D datat i filen Murder3D.txt. Gör en log-linjär modellanalys (Log-linear – Model selection) och jämför resultatet med det tidigare erhållna.

Vi fortsätter med att analysera på motsvarande sätt data i filerna Survival2D.sav Survival3D.sav. Hur blir slutsatserna denna gång?

Sedan tittar vi på filen sexbias.sav och analyserar sambanden mellan variablerna (Log-linear – Model selection). Hur blir slutsatserna? En noggrannare titt på de 6 olika tabellerna för huvudämnena får man från menyn Cross-tabs (kryssa för Chi^2 testet). Vad upptäcker vi om sambandet mellan kön och antagning?

Vi upprepar nu analyserna ovan genom att använda B-Course istället. Input formatet är simpel textfil så vi använder filerna (murderdata2d.txt osv, samt sexbias.dat)

Läs nu in filen EconomicActivity.sav. Den omfattar 665 observationer på 8 dikotoma variabler. Menyn Log-linear analysis – Model selection kan användas för att få inblick i datats beroendestruktur. Definiera Saturated model, Range (1:2), dvs 256 celler i korstabellen, samt Option Association table (ta bort kryssen från övrig output för att det inte kommer för mycket stoff på en gång i utskriftsfönstret).

Utskriftsfönstret ger information om vilka interaktioner som bedömts vara signifikanta och vilka som uteslutits ur modellen. Problemet här detsamma som i multipel regressionsanalys, dvs p-värdenas storlek beror kraftigt på kontexten – alltså vilka andra termer som råkar finnas med i modellen då man testar en viss hypotes om att en interaktionsterm är lika med noll.

Efter att vi betraktat den hittade optimala modellens interaktionsstruktur, skall samma data matas i textform (econdata.txt) till B-Course.

Vi upprepar analyserna ovan med data över riskfaktorer för hjärtsjukdom (HeartData.sav & heartdata.txt) som omfattar 1841 observationer på 6 dikotoma variabler (alltså både med SPSS och B-Course).