Statistik 2 hemtentamen

I hemtentamen för kursen Statistik 1 konstaterades det följande: Det viktiga i statistikinlärningen är att tillämpa kunskaperna i praktiken – helst på ens eget ämnesområde. Detta bidrar till att skapa en god intuition för vad statistiska begrepp and analyser innebär i en kontext som är bekant för en själv.

Under kursen Statistik 2 är syftet att ta ett steg vidare från Statistik 1 och få inblick i ett antal populära multivariata statistiska metoder som används över många tillämpningsområden. Tyvärr finns det ingen allmän bok om multivariata metoder på svenska, men boken 'Statistisk dataanalys' av Svante Körner & Lars Wahlgren (2006, Studentlitteratur) erbjuder en bra icke-teknisk beskrivning bl a av variansanalys, linjär regressionsanalys samt logistisk regressionsanalys. Kurshemsidan erbjuder också en del länkar till bakgrundsmaterial om multivariata statistiska analyser.

I hemtentamen skall deltagarna kunna kritiskt tillämpa statistisk metodik på diverse datamaterial för att ta ställning till frågor och hypoteser som är typiska i multivariata analyser. Resultat från dessa analyser samt egna reflexioner kring dem sammanställs i en rapport som skickas till föreläsaren per mejl för bedömning. Eventuella icke-godkända rapporter returneras till deltagarna med beskrivning av de ändringar som måste utföras för att studieprestationen skall godkännas.

För samtliga uppgifter i hemtentaman uppmanas deltagarna själva att skaffa lämpliga datamaterial. Om inga lämpliga data hittas, kan man även använda sig av de data som nämns specifikt i själva uppgifterna.

Att identifiera lämpliga datamaterial är en viktig del av inlärningsprocessen, eftersom datamaterialets karaktär påverkar kraftigt vilka analyser är meningsfulla att utföra. Samtidigt bekantar man sig med typiska frågeställningar och mätmetoder som används inom ett visst ämne. Data kan hämtas t ex från internet, från vetenskapliga publikationer (helst på ens eget ämnesområde), från forskarna på eget ämne eller från längre hunna studerande på samma ämne som redan håller på med graduprojekt. Vid sållning av vetenskaplig litteratur är Google Scholar och Web of Science utmärkta hjälpmedel. Flera deltagare kan givetvis dela på ett och samma datamaterial, men rekommendationen är att skaffa data som reflekterar ens egna intressen inom studierna.

Vid rapportering av resultaten bör datakälla och materialets karaktär beskrivas för samtliga uppgifter. Själva datafilerna behöver ej bifogas med rapporten. Notera att data från en och samma undersökning/experiment kan användas vid flera uppgifter ifall att det är lämpligt med tanke på frågeställningen. Rapporterna för hemtentamen bör returneras senast 30. september 2011.

Uppgift 1:

Multipel linjär regression (MLR). Beskriv i korthet vad MLR går ut på och tillämpa den på ett datamaterial med en responsvariabel samt två eller fler förklarande variabler (kallas även

kovariat/prediktorer). Bedöm hur väl modellen lämpar sig för datat och beskriv eventuella ändringar/alternativa modeller som förbättrar prediktionerna och/eller utesluter onödiga förklarande variabler. Ifall du inte har tillgång till egna intressanta data, använd HousingData.sav som finns i denna mapp:

http://web.abo.fi/fak/mnf//mate/jc/miscFiles/dataDel1.zip

Datamaterialet beskrivs i detalj i boken Applied Multivariate Statistical Analysis (2003) av W. Härdle & L. Simar, se http://web.abo.fi/fak/mnf//mate/jc/statistik2/mvapdf.pdf. En naturlig responsvariabel i detta data är HomeValue (värdet på fastighet/lägenhet). Identifiera och anpassa en lämplig linjär regressionsmodell för prediktion av värdet och undersök modellens användbarhet.

Uppgift 2:

Variansanalys (ANOVA) med minst två förklarande faktorer. Beskriv syftet med dylik ANOVA och förklara hur interaktioner mellan faktorer fungerar i ANOVA modeller. Tillämpa ANOVA på ett datamaterial med minst två faktorer och undersök om en interaktion mellan faktorerna bör inkluderas i modellen. Ifall du inte har tillgång till egna intressanta data, använd exempelvis datat om den minnestudie som beskrivs av Dr. Karl Wuensch här: http://core.ecu.edu/psyc/wuenschk/spss/ANOVA2-SPSS.doc
Själva SPSS datafilen finns här: http://core.ecu.edu/psyc/wuenschk/spss/ANOVA2.sav

Uppgift 3:

Logistisk regressionsanalys. Beskriv syftet med logistisk regressionsanalys och tillämpa den på ett datamaterial. Ifall du inte har tillgång till egna intressanta data, använd exempelvis data från NHL, Premier League, (eldyl, de finns på nätet direkt tillgängliga) för att skapa en statistisk prediktionsmodell för att ett visst lag vinner en match. Alternativt kan du analysera en ytterligare datafil (DropOut) av Dr. Karl Wuensch, den kan laddas ned här: http://core.ecu.edu/psyc/wuenschk/spss/spss-Data.htm
Studien handlar om prediktion av de fall där en elev hoppar av skolan.

Uppgift 4:

Analys av sambandsmönster bland variabler. Beskriv varför det är nödvändigt att analysera samband mellan variabler så att de betraktas samtidigt och ej enbart parvis. Analysera sambandsmönster exempelvis med gratisprogrammet B-course (http://b-course.cs.helsinki.fi/obc/). Ett mycket aktuellt data är t ex svar av kandidater i riksdagsvalet på de frågor som ställdes i YLEs Vaalikone. Journalisten Jens Finnäs har gjort ett skript som plockar fram datat från nätet, se här:
http://dataist.wordpress.com/2011/03/25/yle-vaalikone-data-updated/
Se även kommentarerna av journalisten Christoffer Gröhn här:
http://val2011.ratata.fi/post/255982/
Ifall du är intresserad av att analysera just detta data och har svårigheter att använda skriptet,

kan någondera journalisten kanske hjälpa dig på spåret. Ifall du inte vill använda dig av det politiska datat, kan du exempelvis analysera sambanden bland olika brottstyper som behandlades under kursens del 7 (UScrime.sav).

Uppgift 5:

Principalkomponentanalys. Ge en beskrivning av vad man exempelvis kan åstadkomma med principalkomponentanalys för multivariata data och tillämpa den på ett datamaterial. Ifall du inte har tillgång till egna intressanta data, använd exempelvis HousingData.sav från uppgift 1.

Uppgift 6:

Faktoranalys. Beskriv några syften med faktoranalys och tillämpa den på lämpligt datamaterial. Ifall du inte har tillgång till egna intressanta data, använd exempelvis data EPQ av Dr. Karl Wuensch, den kan laddas ned här: http://core.ecu.edu/psyc/wuenschk/spss/spss-Data.htm
Datat handlar om attityder kring etik, se även:
http://core.ecu.edu/psyc/wuenschk/SPSS/ItemAnalysis-SPSS.doc