Statistik 2 2010, 3.-9.5.2010

Stansens PC-klass ASA-huset.
Schema:

mån	**ti**	**ons**	**to**	**fre**
9.15-12.00 13.15-15.00	9.15-12.00 13.15-15.00	10.15-13.00	10.15-12.00 13.15-16.00	10.15-12.00 13.15-16.00

Under kursens gång bekantar man sig med ett antal statistiska metoder som utnyttjas flitigt på många forskningsområden. Den gemensamma nämnaren för dessa metoder är att man betraktar samtidigt två eller fler variabler. Det finns dock avsevärda skilnnader mellan metoderna, eftersom somliga är enbart explorativa i sin natur medan andra utnyttjas i första hand för att jämföra specifika hypoteser eller teorier som förklaringsmodeller för den variation vi ser i data. Följande statistiska metoder kommer att betraktas under kursen:

1. Principalkomponentanalys
2. Multipel regression
3. ANOVA, MANOVA och interaktionstermer
4. Logistisk regression
5. Interaktionsmodeller för korstabelldata
6. Mixed models
7. Flerdimensionell skalning
8. Klusteranalys
9. Faktoranalys

Alla datorövningar kommer att utföras med SPSS och vissa även med online programvara. Övningarna varvas med beskrivningar av metoderna. Syftet med beskrivning av de underliggande statistiska idéerna är inte att återge en tekniskt exakt definition, utan man försöker i första hand avmystifiera metodiken.

Material del I

Principalkomponentanalys

Principalkomponentanalys används till mycket inom data-analys, men dess huvudsakliga syfte är oftast att hitta/urskilja intressanta mönster i flerdimensionella data. En allmän beskrivning med länkar till vidare material finns på http://en.wikipedia.org/wiki/Principal_component_analysis.

Öppna datafilen SedlarData.sav och gör en principalkomponentanalys (Dimension reduction/Factor). Sätt först antalet komponenter till 6, och sedan till 3 och 2. Kom ihåg att spara komponenterna. Gör ett spridningsdiagram för datat utifrån den erhållna principalkomponenterna och använd variabeln Grupp som markör. Kan du urskilja de två grupperna?

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). Scree plot visar hur mycket varje komponent förklarar av variationen i data.

I följande övning betraktar vi genetiska data där varje case är en individ (97 totalt) samplats från en av fyra geografiska regioner (GenetiskData.sav) Den första kolumnen i datamatrisen indikerar regionen och var och en av de övriga kolumnerna (85 stycken) indikerar om individen bär en viss mutation i arvsmassan. Försök utföra principalkomponentanalys som ovan med samtliga mutationsindikatorer (V2-V86). Vad händer? Det finns fem indikatorer som är konstanta i datamaterialet (V20, V55, V63, V70, V71). Tag bort dem från listan av variabler som sätts in i principalkomponentanalysen och utför sedan analysen (Scree plot & scores). Gör ett 3D spridningsdiagram med 3 principalkomponenter och undersök om det verkar finnas skillnader i genetiska profiler mellan regionerna.

I följande artikeln av Novembre et al i Nature hittar man ett exempel på dylik användning av principalkomponentanalys i ett genetiskt sammanhang, där man har ett stort antal samplade individer (ca 3,000) och variabler (ca 500,000):

http://www.nature.com/nature/journal/v456/n7218/full/nature07331.html

Multipel linjär regression

Multipel regression används mycket allmänt för att skapa prediktioner för en eller flera responsvariabler av intresse, givet ett antal pontentiella prediktorvariabler. Metodiken beskrivs utförligt i boken Applied Multivariate Statistical Analysis (2003) av W. Härdle och L. Simar.
Läs in följande datafil HousingData.sav. Filen innehåller sammanlagt 506 observationer på 14 variabler som karaktäriserar olika distrikt i Boston-området i USA. Variablerna i datamaterialet betraktas i detalj på s. 44-52 i Härdle & Simar.
Rita låddiagram (boxplot), samt spridningsdiagram (scatterplot matrix) för samtliga variabler. Studera hur de observerade värdena fördelar sig.

Anpassa en regressionsmodell, där variabel HomeValue är den beroende variabeln och variablerna CrimeRate, NonretailBusiness samt OldHouses är prediktorer. Vilka variabler verkar koppling till huspriset enligt p-värden? Jämför dessa resultat med spridningsdiagrammen. Undersök modellanpassningen med hjälp av grafisk diagnostik (histogram för residualer, normal probability plot). Verkar modellen lämplig?

Anpassa en ny modell, där endast de variabler som hade signifikanta regressionskoefficienter (på 5%-nivån) tas med som prediktorer tillsammans med de övriga variablerna i datat. Har situation ändrats från den tidigare modellen, dvs är de tidigare prediktorerna fortfarande signifikanta när nya variabler inkluderades?

Leta efter en modell med forward och backward metoderna och granska resultaten. Blir det några skillnader?

I många situationer är det nödvändigt att transformera variabler innan de sätts in i linjär regressionsmodell, t ex för att stabilisera variansen och för att åstadkomma mer symmetriska fördelningar. Gör en logaritmisk transformation på variablerna 1,3,5,6,8,9,10 och 14 och upprepa sedan analysen ovan. Verkar modellen ha en bättre anpassning (betrakta residualer) efter transformationen?

I datafilen GeneticTraitsData.sav finns samma genetiska information som betraktades redan under principalkomponentanalysen, men därutöver finns det observationer på fyra kontinuerliga fenotyper (Trait1-4). Vi använder linjär regression skilt för varje fenotyp och sätter in de genetiska indikatorerna som förklarande variabler (prediktorer) och letar reda på vilka som antyds ha en koppling till fenotyperna.

Kör linjär regression med forward och backward sökningen och notera vilka variabler som blir signifikanta prediktorer. Residualanalys avslöjar att modellanpassningen är förhållandevis god. För varje fenotyp upptäcks det två eller fler prediktorer som är signifikanta på 5%-nivån. Däremot vet man (eftersom det fenotypiska datat är simulerat) att endast V21 har en riktig effekt på fenotyp Trait4, alla andra identifierade prediktorer med signifikant p-värde är falska positiva fynd. Därför bör korrektion för multipla test användas i dylika sammanhang, t ex se följande länkarna:

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

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

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 online program 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).

ANOVA, MANOVA och interaktionstermer

Enkel variansanalys (ANOVA) och MANOVA (fler än en responsvariabel) är populära statistiska metoder för analys av variation hos kontinuerliga eller ordinala responsvariabler då man betraktar dem givet en eller fler förklarande faktorer som kan påverka responsvariablernas beteende. Det är vanligt att datat i detta sammanhang är experimentellt (förklarande faktorerna kan väljas och deras värden kan manipuleras), men analyserna görs flitigt även för icke-experimentella data. Följande icke-teknisk beskrivning av ANOVA och interaktionstermer belyser modellernas natur på ett utmärkt sätt:

http://skeetersays.blogspot.com/2008/08/demystifying-statistics-on.html

Allmän introduktion till olika typer av ANOVA hittar man på:

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

och en tillämpad beskrivning av MANOVA i ett psykometriskt sammanhang finns på:

http://ibgwww.colorado.edu/~carey/p7291dir/handouts/manova1.pdf

Detaljerna på dessa modelltyper kommer att diskuteras under föreläsningen.

Läs nu in filen motiv.sav. Uppgifterna beskriver studierprestationer under försöket med två olika typer av undervisnings. Studiemotivation är en ytterligare variabel. Anpassa en variansanalysmodell (General linear model – Univariate) där achievement är responsen, teach en fixed faktor och motiv en stokastisk faktor. Alternativt kunde man sätta in motiv som kovariat eftersom dess kategorier är ordnade. Definiera modellen så att den har en interaktionsterm mellan de två förklarande termerna. Hur blir slutsatserna?

Läs in filen AQUES.sav. Vi anpassar nu en multivariat linjär modell (General linear model – Multivariate) där reaktionstiderna m.a.p. bägge händerna är responsvariablerna, kön och tidigare rökningsvanor är fixed faktors, samt sätter in åldern och de tre pulsmätningarna som kovariat. Hur blir slutsatserna? Matris-spridningsdiagram och box-plottar hjälper här för att betrakta data visuellt.

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?